Eléments de statistique pour citoyens d'aujourd'hui et managers de demain

HEC formation fondamentale L3 — Prof. : Hedi Benamar, Aurélien Garivier, Gilles Stoltz — Année universitaire 2010–11

Table des matières

Avant-propos : Objectifs et règles du jeu

1. Présentations

1.1. Vos enseignants.

Vos enseignants pour les cours magistraux sont
- Hedi Benamar, doctorant en finance à HEC Paris,
- Aurélien Garivier, chercheur au CNRS,

et pour les sessions de travaux pratiques,
- Gilles Stoltz, professeur affilié à HEC Paris et chercheur au CNRS.

Le site web du cours `http://www.hec.fr/stoltz` (rubrique : Statistiques – L3) est très complet et vous permettra par exemple de récupérer les différents polycopiés en cas d'absence ; les exemplaires restants étant par ailleurs également remisés au bureau de promotion L3, chez Béatrice Poivre. Le site comprend également les fichiers de données SPSS pour les sessions de travaux pratiques, ainsi que les corrigés des quizz que nous ferons au cours du semestre.

1.2. Vous, les étudiants. Ce cours est taillé pour être parfaitement accessible aux étudiants issus des voies EC/E, EC/S et B/L. En effet, Aurélien Garivier et Gilles Stoltz ont été / sont interrogateurs au concours d'entrée à l'Ecole normale supérieure, voie B/L et connaissent parfaitement le programme de cette voie ; et les programmes de EC/E et EC/S sont encore plus exigeants et ambitieux au niveau mathématique. Ainsi, si vous décrochez à un moment donné, ce sera par manque de travail, pas par manque d'aptitude ni parce que le cours serait trop difficile !

Pour les étudiants issue de la voie A/L, cela risque cependant d'être plus compliqué, car vous n'avez pas vu les résultats fondamentaux nécessaires (notion de variable aléatoire, indépendance, loi des grands nombres, théorème de la limite centrale). Votre objectif sera de pouvoir comprendre et mettre en œuvre les recettes statistiques – et par compréhension, on entend compréhension de loin, pas nécessairement dans le détail. N'hésitez pas à vous signaler à votre enseignant en début de semestre. Des cours de soutien spécifiques pourront être mis en place si le besoin s'en fait sentir.

Remarque 0.1. Pour les étudiants de EC/S et EC/E : nous avons observé les années précédentes une faiblesse de votre formation en probabilités et statistique (par rapport à celle en analyse et algèbre). Cela tient notamment à la formation initiale de vos enseignants (les plus jeunes d'entre eux, cependant, ont souvent passé l'agrégation avec l'option de probabilités et statistique).

FIGURE 1. Je vous souhaite de ne pas mettre les pieds dans une entreprise comme celle-ci. A vrai dire, il tiendra beaucoup à vous qu'il n'en soit pas ainsi !

Par ailleurs, le programme officiel de la voie EC/S est ambitieux, trop ambitieux : en gros, le contenu de nos cinq premières séances de cours (sur douze) devrait déjà être connu, mais nous savons que ce n'est que partiellement le cas, et que même lorsque les mathématiques correspondantes sont connues, les interprétations, essentielles en statistique, ne le sont pas. Or, ces interprétations, c'est votre valeur ajoutée de futurs anciens élèves d'une école de commerce... Vous pourrez donc avoir l'impression de redites, mais ce sera sans doute une impression trompeuse. Par conséquent, ne vous reposez surtout pas sur vos lauriers !

1.3. Le cours. Depuis cette année 2010–11, vous bénéficiez tous, quelle que soit votre filière d'origine, d'un cours de statistique de niveau assez avancé, associé à un (très) gros polycopié. Pour vous, ce cours de statistique ne ressemblera pas à l'application de recettes de cuisine sortie de nulle part. Au contraire, vous aurez un aperçu des raisonnements mathématiques y conduisant et vous serez dès lors davantage conscients de la logique et des limites de chaque méthode. De plus, nous ferons le lien entre les mathématiques de classes préparatoires et les mathématiques universitaires nécessaires. Cela vous demandera de fournir un réel travail, ce qui pourra vous sembler agaçant en ces mois où vos pensées se tournent plus vers la fête. Mais cela vous donnera une plus grande assurance face au traitement des chiffres pour le reste de votre carrière.

En termes de liens avec l'entreprise, l'objectif principal est en fait d'éviter la situation de la figure 1. Plus sérieusement, nous voyons deux intérêts majeurs à ce cours pour votre vie future – intérêts qui lui donnent d'ailleurs son nom.

- Un intérêt en termes de management : un premier poste typique est celui de chef de produit, et cela implique de mener des études marketing. Selon la taille de l'entreprise, vous aurez à écrire l'enquête, à la conduire et à en exploiter les résultats ou vous devrez superviser les opérations précédentes. Même dans le second cas (peu courant !), vous savez bien qu'on ne peut superviser efficacement (sans se faire rouler par ses subordonnés) que quelque chose que l'on maîtrise un peu soi-même. A noter : les cours de marketing approfondiront et développeront les techniques vues dans ce cours d'un point de vue *business* ; ils omettront en revanche les mathématiques sous-jacentes.

– Un intérêt citoyen, un élément-clé de votre formation humaine générale – citons Herbert George Wells (écrivain britannique, 1866–1946) : "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." En effet, nous verrons que les chiffres ne parlent pas d'eux-mêmes, et même pire, qu'on peut les manipuler et se faire manipuler. Cependant, un statisticien averti arrive à détecter cela.

... Sans compter que ce cours vous servira dès cette année ou la suivante pour vos missions JE ou pour traiter des données lors de missions humanitaires : si, si, vous n'imaginez pas les mails que nous pouvons recevoir quelques mois après le cours une fois que tout est malheureusement oublié et que vos prédécesseurs veulent remettre leurs connaissances à plat... !

2. Objectifs et déroulement du cours, règles du jeu

Ce qui suit ici pourra être adapté par chacun des enseignants. En tout état de cause, ces règles du jeu seront précisées lors du premier cours.

2.1. Déroulement pratique des cours, évaluation.

Déroulement pratique :

– 1h30 de cours magistral, tous les mercredis après-midis (13h–14h30 ou 14h40–16h10), en groupes d'environ 40 étudiants.
– Pour chaque étudiant, une séance de TPs de 1h30 ; elles seront effectuées en demi-groupes d'environ 20 étudiants, vers fin avril ou début mai (les séances seront inscrites dans votre calendrier Intranet).

Remarque 0.2. Pour les TPs, vous pourrez utiliser soit les ordinateurs des salles où nous serons (ils seront équipés de SPSS 18), soit votre propre ordinateur, à la condition expresse d'avoir installé au préalable SPSS 18 (voir `https://intranet.hec.fr/mi/` pour l'installation de ce dernier). Contactez les moyens informatiques de l'Ecole en cas de souci ; à chacun ses tâches : en aucun cas nous ne serons vos techniciens d'installation SPSS.

Evaluation : la note finale est déterminée à partir

– des notes obtenues aux quizz de 30 minutes (sans documents) ; la présence aux quizz est obligatoire, l'absence à un quizz vaut la note F ;
– des compte-rendus de TPs (effectués seul ou en binôme, mais pas à plus de deux) ; la présence en TPs est obligatoire, l'absence [1] à la séance de TP à laquelle vous avez été affecté entraîne la note F ;
– de l'examen final (avec documents) ;
– à tout cela s'ajoute un bonus / malus éventuel déterminé en fonction de la présence physique aux cours (mesurée par une feuille d'émargement passée dans les rangs) et de la participation orale (ou écrite : cf. par exemple le signalement par email des coquilles situées dans le polycopié).

1. La seule tolérance est en cas de *permutation* d'un étudiant avec un autre, annoncée à l'avance par email ; nous sommes obligés d'appliquer cette règle stricte à cause de débordements qui sont arrivés les années précédentes, avec des TPs à 12 étudiants, et d'autres à 35 !

La distribution des notes finales est contrainte depuis cette année par des règles uniformes entre les différents groupes et cours de l'Ecole HEC (par exemple, taux maximal de notes A fixé à 20 %, taux maximal de notes A + B + C fixé à 70 %).

REMARQUE 0.3 (Documents aux quizz et à l'examen). Tous les polycopiés de cours seront autorisés à l'examen. En revanche, *aucun* document n'est autorisé lors des quizz, mais une calculatrice y sera souvent nécessaire.

REMARQUE 0.4 (Annales des examens). A la fin de ce polycopié, vous trouverez six sujets d'examen (examens principaux et de rattrapage des années 2007–08, 2008–09 et 2009–10).

2.2. Comportement en cours. Cela nous désole d'ajouter un tel paragraphe à cet avant-propos ; c'est le fruit de débordements passés. Ainsi :

- Il va sans dire qu'il est interdit d'utiliser son téléphone portable en cours, de même que tout ordinateur portable. Seule une calculatrice est utile, voire indispensable (notamment pour les quizz).
- En ce qui concerne toute absence prévisible et légitime (la détermination du caractère légitime nous revenant de droit) : merci de nous en informer par mail au préalable, notamment s'il s'agit d'un TP ou en cas de quizz.
- L'arrivée en retard à un quizz réduit d'autant la durée de composition de ce dernier : aucun temps supplémentaire ne sera accordé.

2.3. Format des polycopiés, méthode de travail.

2.3.1. *Format des polycopiés.* A chaque unité de cours correspond une partie de ce polycopié, divisée elle-même en trois chapitres :

1. la version rédigée du cours que nous écrivons au tableau (pratique pour les absents ou pour ceux qui auraient été mal réveillés) ; elle se lit comme un roman et votre mission sera de la relire au plus une fois ;
2. des compléments de cours facultatifs pour les étudiants motivés qui veulent aller plus loin (on ne les verra pas en cours, ils ne seront pas au programme de l'examen) ;
3. des exercices, avec corrections : nous en ferons quelques-uns en cours mais la plupart d'entre eux seront à faire chez vous, dans un premier temps sans regarder la correction ; à votre demande expresse, nous pourrons revenir sur certains d'entre eux en cours si la correction écrite est trop laconique à votre goût.

Enfin :

4. Il est également à noter que le polycopié se clôt par une série de fiches de synthèse rappelant, pour chaque partie, les notions à retenir pour la suite du cours et pour l'examen. (Il est donc *a priori* inutile de ficher le cours, puisque nous l'avons fait pour vous.)
5. Des annales des quizz sont disséminées dans le polycopié, dans des parties à part ; les annales des examens sont quant à elles groupées en fin de polycopié.

Effectuez les quizz pour vous entraîner et méditez bien la correction que nous en avons faite : notre expérience nous montre que la plupart des erreurs que nous y soulignons se répercutent malgré tous les efforts de promotion en promotion (notamment en ce qui concerne la rédaction) !

REMARQUE 0.5 (De notre flexibilité). Ce découpage de chaque partie en trois chapitres et une fiche de synthèse est issu des souhaits des étudiants des années précédentes et permet de satisfaire tous les goûts : ceux qui aiment les exercices, ceux qui veulent des fiches de synthèse, ceux qui préfèrent un cours rédigé et reprécisant les articulations essentielles, ceux qui veulent aller plus loin, etc. Si vous avez des suggestions de présentation, d'ajout de contenu, etc., nous sommes à votre disposition pour en discuter et les mettre en œuvre.

REMARQUE 0.6 (Clés pour le management). Nous n'oublions pas que nous sommes dans une école de commerce (et pas en faculté de mathématiques). Les formules de mathématiques, vous les oublierez, nous ne nous faisons aucune illusion. Mais au fil des différents cours, nous vous écrirons au début de la fiche de synthèse, en toutes lettres, les points-clés culturels à méditer. La culture : ce qu'il reste quand on a tout oublié ! Mais qui aura formé votre esprit à ne pas penser l'étude des chiffres de manière binaire (vrai ou faux) et vous aura montré tous les écueils et difficultés de cette science délicate.

2.3.2. *Méthode de travail suggérée.* Nous vous proposons après chaque cours d'en relire la version rédigée et de faire les exercices. Il n'est pas nécessaire de prendre des notes en cours ; il vaut peut-être mieux se concentrer sur le tableau et suivre en direct les enchaînements (qui vont un peu vite parfois : arrêtez-nous !).

Posez vos questions au fil du trimestre, et n'attendez pas d'être dépassés. Les quizz vous permettront de vous tester et voir s'il faut redresser la barre. Dans tous les cas, un travail sérieux et régulier (environ 1h30 à 2h) par semaine sera nécessaire. Le cours est en format réduit (douze séances de 1h30, à comparer aux 24 séances de 1h30 en économie ou en droit) ; mais c'est parce qu'une partie du travail d'assimilation doit être faite chez vous.

2.4. Contenu pédagogique du cours. Les objectifs pédagogiques, que nous allons préciser dans le premier chapitre de rappels, sont les suivants :

- faire le lien entre des données et un modèle mathématique ;
- exploiter cette modélisation et appliquer la bonne méthode statistique pour arriver à une quantification (fourchette de valeurs, degré de crédibilité d'une hypothèse) ;
- interpréter la quantification obtenue en termes stratégiques et savoir quel crédit lui apporter ;
- être conscient que l'on ne peut pas toujours conclure ou que l'on se trompe parfois.

En fait, la science statistique est difficile, au sens où étant donné des observations (issues d'un certain aléa), on veut effectuer des énoncés déterministes. Ainsi :

REMARQUE 0.7 (La statistique se trompe parfois). Souvenez-vous de l'exemple suivant, que nous étudierons plus en détails : lors des municipales 2008, les prévisionnistes ont d'abord donné, pour la mairie du 5ème arrondissement, Lyne Cohen-Solal gagnante (à 20h45), avant d'annoncer la victoire de Jean Tiberi (à 21h15). La fin du dépouillement à 22h a confirmé la victoire de ce dernier.

REMARQUE 0.8 (La statistique ne peut pas toujours conclure ; il faut quantifier les impressions). Parfois, on manque de données et on ne peut pas se prononcer avec certitude. Par exemple, si je vous dis qu'on a lancé une pièce et que 40 % du temps, on a obtenu pile, et que je vous demande si la pièce est biaisée, que répondrez-vous ? Tout dépend du nombre de lancers effectués : avec 10 lancers (4 fois pile, 6 fois face), on ne peut rien dire. Et avec 100 lancers (40 fois pile, 60 fois face) ? Et 1 000 lancers (400 fois pile, 600 fois

face) ? On quantifiera ensemble dans ce cours le seuil n du nombre d'observations à partir duquel cette proportion de 40 % indique de manière dite statistiquement significative que la pièce est biaisée.

Principes à retenir et que nous développerons :
- Les chiffres ne parlent pas d'eux-mêmes (et nous les ferons accoucher !).
- La statistique permet de quantifier les impressions, et de passer d'un vague sentiment subjectif à un indicateur objectif.
- La statistique procure cette quantification et aide un être humain à prendre une décision (elle éclaire la décision, mais ne la prend pas).

La fiche de présentation synthétique du cours est reproduite à la figure 2.

ELEMENTS DE STATISTIQUE

Intervenants : Hedi BENAMAR / Aurélien GARIVIER / Gilles STOLTZ

Présentation

In God we trust, all others bring data. [Edwards Deming, professeur de statistique américain, 1900-1993]

En gestion, la statistique intervient comme un outil pour soutenir la prise de décision. N'a de bons arguments que celui ou celle dont le raisonnement est appuyé par des statistiques bien choisies et bien présentées (c'est la statistique descriptive). Ne peut définir de stratégie pertinente ou ne peut prévoir efficacement les quantités à produire, acheter ou vendre, que celui ou celle qui sait modéliser les données et en extraire les lois des phénomènes en jeu (c'est la statistique inférentielle).

Objectifs pédagogiques

Je ne crois aux statistiques que lorsque je les ai falsifiées moi-même. [Winston Churchill, homme politique britannique, 1874-1965]

Le cours vise principalement à introduire et faire méditer les concepts fondamentaux et méthodes élémentaires de la statistique pour permettre un apprentissage autonome ultérieur de méthodes complémentaires. On veut développer le sens critique nécessaire lors de la mise en œuvre et de l'interprétation d'un traitement statistique (par exemple, le résultat d'un test d'hypothèses).

Pour cela, on introduira et utilisera d'une part un cadre mathématique rigoureux pour la modélisation de phénomènes aléatoires, qui consolide celui vu en classes préparatoires ; d'autre part, un retour constant aux données sera effectué, par leur traitement sous le logiciel d'analys statistique SPSS.

Déroulement du cours

12 séances de cours magistral de 1h30
1 séance de travaux pratiques sur machine en demi-groupes

Contenu (progression cours par cours)

1. Rappels de calcul des probabilités, panorama de la démarche statistique
2. Modélisation statistique (le modèle statistique)
3. Notions d'estimateur, de quantiles
4. Estimations par intervalles, intervalles de confiance
5. Estimations par intervalles, suite / Introduction aux tests
6. Méthodologie des tests d'hypothèses et tests de comparaison à une valeur de référence
7. Tests de comparaison de deux populations
8. Tests du chi-deux
9. Régression linéaire simple (avec une variable explicative quantitative)
10. Régression linéaire multiple (avec plusieurs variables explicatives quantitatives)
11. Régression linéaire multiple, suite
12. Séance de révisions et d'études de cas

Evaluation

La note finale est déterminée par les quizz, la séance de travaux pratiques, l'examen final, ainsi que par la présence et la participation en cours.

Bibliographie (ouvrages disponibles à la bibliothèque)

G. Saporta. *Probabilités, analyse des données et statistique*. Technip, 2006.
M. Laviéville. *Statistique et probabilités : Rappels de cours et exercices corrigés*. Dunod, 1998.

FIGURE 2. Fiche synthétique de présentation du cours.

Première Partie

Panorama de la démarche statistique

Version rédigée du cours

Le but de ce chapitre est d'illustrer quelques concepts fondamentaux sur des exemples ne mettant en jeu que les lois les plus simples, essentiellement celles de Bernoulli.

1. Des données à une modélisation

En statistique, comme dans la vraie vie, on se pose des questions, et on essaie d'y répondre. On considère deux situations différentes conduisant à une modélisation similaire,

- la répétition d'une expérience (on lance une pièce plusieurs fois pour déterminer si elle est équilibrée) ;
- la considération d'un échantillon au sein d'une population (on interroge un certain nombre de personnes prises au hasard pour déterminer le sentiment général, par exemple lors d'un sondage électoral).

1.1. Telle pièce est-elle équilibrée ? On lance n fois une pièce et on inscrit les résultats dans un tableau. On code pile par 1 et face par 0.

Lancer numéro	1	2	3	4	5	6	7	8	...
Résultat	0	0	1	0	1	1	1	0	...

On note x_1, x_2, x_3, ... les résultats successifs. Par exemple, on obtient pile au troisième tirage : $x_3 = 1$.

Le mathématicien se place aux instants avant les tirages et considère ceux-ci comme la répétition d'une expérience aléatoire : il note X_1, X_2, ... les variables aléatoires correspondantes. Elles représentent la valeur (aléatoire) du lancer avant que celui-ci n'ait lieu. Les résultats x_t sont alors les réalisations des variables aléatoires X_t, on les appelle les valeurs observées ou les données. Les X_t sont elles-mêmes appelées, par extension, les observations. Notez bien la différence entre l'utilisation des majuscules pour les variables aléatoires et les minuscules pour leurs réalisations (les données).

S'il s'agit de la même pièce et qu'on ne modifie pas la manière dont on lance, alors on peut dire que les X_t sont indépendantes et identiquement distribuées. Leur loi commune est une loi de Bernoulli, de paramètre inconnu noté p_0.

Remarque 1.1. On connaît donc la forme de la loi commune des observations (une loi de Bernoulli), mais on ignore son paramètre. L'objet de la statistique est de dire des choses sur ce paramètre.

Ici, se poser la question de savoir si la pièce est équilibrée revient à se demander si $p_0 = 1/2$.

A retenir pour la suite : pour répondre à cette question, le traitement mathématique commence par considérer le modèle statistique formé par l'ensemble des lois de Bernoulli : $\mathcal{B}(p)$, où $p \in [0,1]$. La vraie loi (inconnue) $\mathcal{B}(p_0)$ régissant les observations se trouve bien dans ce modèle. On dispose d'un nombre fini n d'observations $X_1, X_2, \ldots, X_n$. On fera ensuite usage des données obtenues pour répondre à des questions posées sur p_0.

1.2. Jean Tibéri vs. Lyne Cohen-Solal. A 20h45, le soir des élections municipales de 2008, les médias annoncent la victoire de Lyne Cohen-Solal à la mairie du cinquième arrondissement. A 21h15, ils font marche arrière et annoncent celle de Jean Tibéri. Nous verrons dans ce cours pourquoi et comment cela est possible. En gros, sur 100 résultats d'élections annoncés avant la fin du dépouillement total des voix, quelques-uns, de l'ordre de 5, sont faux.

Il y avait 25 043 suffrages exprimés, comme on s'en est rendu compte après dépouillement total à 22h. A 20h45, les statisticiens avaient utilisé des premiers résultats partiels (les feuilles de centaine). Admettons qu'ils aient eu connaissance de 10 telles feuilles. Trois candidats étant en lice, ils avaient donc, bulletins nuls ou blancs déduits, des valeurs observées $x_1, \ldots, x_{978}$ disons. Par exemple, 463 voix pour Lyne Cohen-Solal, 409 voix pour Jean Tibéri et 106 voix pour Philippe Meyer. Lors du dépouillement final, on a en trouvé respectivement 11 044, 11 269 et 2 730.

Pour faire leur prédiction à 20h45, les statisticiens modélisent les valeurs observées comme étant les réalisations de variables aléatoires $X_1, \ldots, X_{978}$ indépendantes et identiquement distribuées selon la distribution finale des voix, celle de 22h, que l'on peut représenter comme un triplet (p_{CS}, p_T, p_M), et dont on n'a su qu'à ce moment-là que sa valeur était $(0.441, 0.45, 0.109)$.

En effet, on tire un petit nombre de valeurs dans un grand ensemble, c'est comme si l'on tirait successivement avec remise et que chaque tirage donnait des résultats selon la proportion générale. On appelle ces X_t un échantillon tiré dans la population.

La distribution finale (p_{CS}, p_T, p_M) est inconnue à 20h45, et l'objet de la statistique est de l'inférer à partir de la connaissance des X_t, pour $t = 1, \ldots, 978$. Dit autrement, on part de l'échantillon pour en déduire une meilleure connaissance de la population. Le modèle statistique est ici formé par l'ensemble des probabilités sur trois éléments,

$$\left\{(p_1, p_2, p_3) \in (\mathbb{R}_+)^3 : \ p_1 + p_2 + p_3 = 1\right\}.$$

REMARQUE 1.2. Ici, on ne s'intéresse pas tant aux valeurs de p_1 et p_2 qu'à savoir si $p_1 > p_2$ ou $p_1 < p_2$.

2. Un exemple du travail du mathématicien : la construction de fourchettes

Le mathématicien développe des méthodes théoriques mettant en jeu les X_t et étudie par exemple comment les utiliser au mieux pour construire des fourchettes (appelées intervalles de confiance en statistique) autour des paramètres inconnus des lois sous-jacentes. Ensuite, on remplace les X_t par leurs réalisations x_t et on en déduit un résultat.

2.1. Deux outils bien connus pour commencer. Vous avez vu en classes préparatoires la loi des grands nombres. Elle modélise le fait que lorsque l'on lance un grand nombre de fois une pièce équilibrée, on a à-peu-près pile la moitié du temps.

THÉORÈME 1.1 (Loi des grands nombres). *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes et identiquement distribuées admettant une espérance* μ. *Alors la moyenne empirique converge vers l'espérance,*

$$\overline{X}_n \overset{\text{not.}}{=} \frac{1}{n}(X_1 + \ldots + X_n) \xrightarrow{\mathbb{P}} \mu \ .$$

La convergence dans le théorème est en probabilité. On en rappelle la définition formelle ci-dessous pour mémoire, mais vous pourrez l'oublier, car elle tue toute intuition. On retiendra simplement qu'en un sens, $\overline{X}_n$ doit être proche de μ à partir d'un certain rang. Ce sens est, tout à fait précisément, que pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big\{|\overline{X}_n - \mu| > \varepsilon\Big\} \longrightarrow 0 \ .$$

DÉFINITION 1.1. *Lorsqu'une variable aléatoire* X *admet un moment d'ordre deux, i.e.,* $\mathbb{E}\,X^2 < +\infty$, *elle admet également une espérance et on définit sa variance comme*

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}\,X)^2\right] = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 \ .$$

Son écart-type est $\sigma(X)$, *la racine carrée de la variance.*

REMARQUE 1.3. Vous aurez remarqué que nous notons désormais l'espérance μ et la variance σ^2, et non plus, comme vous l'avez peut-être fait en classes préparatoires, m et v. C'est une question d'habitude... En fait, les lettres romaines seront utilisées pour les moyennes $\bar{x}_n$ et écarts-types s_n calculés sur les valeurs observées $x_1, \ldots, x_n$. Les lettres grecques seront quant à elles employées lors de l'énoncé du modèle statistique sous-jacent aux $X_1, \ldots, X_n$. La figure 3 essaie d'imprimer cette distinction en vous.

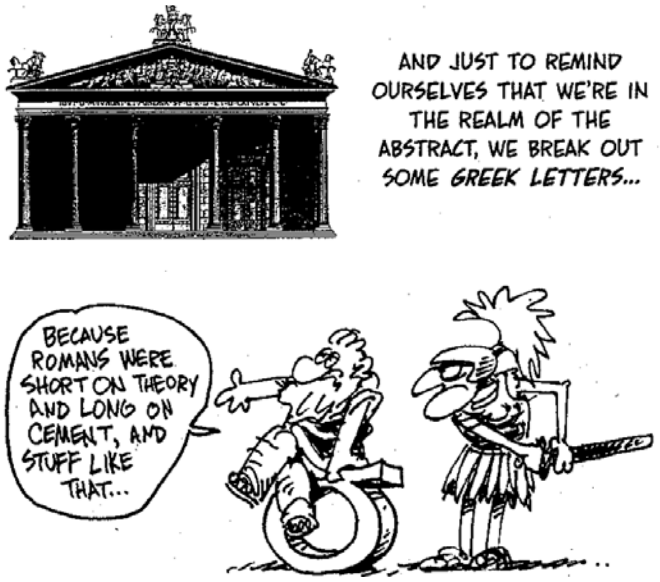


FIGURE 3. Les lettres grecques sont pour l'énoncé du modèle théorique et les lettres romaines, pour le traitement des données concrètes.

Lorsque $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées et que leur loi commune admet un moment d'ordre deux, un calcul simple montre que la variance de $\overline{X}_n$ vaut

$$\sigma^2(\overline{X}_n) = \frac{1}{n^2}\Big(\sigma^2(X_1) + \ldots + \sigma^2(X_n)\Big) = \frac{1}{n}\,\sigma^2(X_1) \ .$$

On a alors, en appliquant l'inégalité de Chebychev-Markov, le contrôle suivant des écarts de $\overline{X}_n$ par rapport à son espérance μ.

PROPOSITION 1.1 (Application de l'inégalité de Chebychev-Markov). *Soient des variables aléatoires* $X_1, \ldots, X_n$ *indépendantes et identiquement distribuées, de loi commune admettant une espérance* μ *et une variance* σ^2. *Alors*

$$\mathbb{P}\Big\{|\overline{X}_n - \mu| > \varepsilon\Big\} \leqslant \frac{\sigma^2}{n\varepsilon^2} \ .$$

REMARQUE 1.4. L'inégalité de Chebychev-Markov entraîne donc la loi des grands nombres pour des variables aléatoires admettant un moment d'ordre deux.

2.2. Application : construction d'intervalles de confiance. Repartons de l'exemple de la pièce et supposons qu'en 100 lancers, on ait eu 42 piles et 58 faces, soit une moyenne observée

$$\overline{x}_{100} = \frac{1}{100}(x_1 + \ldots + x_{100}) = 0.42 \ .$$

On pense donc avec raison que le vrai paramètre p_0 se situe sans doute autour de 0.42 mais on voudrait quantifier les incertitudes. Un nombre seul ne suffit pas ! Est-ce 42 % plus ou moins 0.1 % ou plus ou moins 10 % ? Dans le premier cas, on peut conclure que la pièce est biaisée, on ne peut en revanche être formel dans le second cas.

On fait donc appel à la théorie, en passant, dans un premier temps, par les X_t. Celles-ci sont de loi commune une loi de Bernoulli de paramètre p_0, et admettent donc l'espérance commune $\mu = p_0$ et la variance commune $\sigma^2 = p_0(1-p_0)$. La proposition précédente implique que

$$\mathbb{P}\Big\{|\overline{X}_n - p_0| > \varepsilon\Big\} \leqslant \frac{p_0(1-p_0)}{n\varepsilon^2} \ ,$$

soit qu'avec probabilité au moins 95 % (un niveau standard de confiance), on a

$$p_0 \in \Big[\overline{X}_n - \varepsilon_n,\ \overline{X}_n + \varepsilon_n\Big] \ ,$$

où ε_n est tel que

$$\frac{p_0(1-p_0)}{n\varepsilon_n^2} \leqslant 5\,\% \ .$$

On voudrait calculer ε_n, or tel qu'écrit ci-dessus, il dépend de p_0, que l'on ne connaît pas. Se rappelant que $x(1-x) \leqslant 1/4$, il suffit en particulier de choisir ε_n tel que

$$\frac{1}{4n\varepsilon_n^2} = 5\,\% \ , \qquad \text{soit} \quad \varepsilon_n = \sqrt{\frac{1}{4n \times 5\,\%}} = \frac{2.24}{\sqrt{n}} \ .$$

On conclut donc de cette étude théorique qu'avec probabilité au moins 95 %, le vrai paramètre p_0 se situe dans l'intervalle, construit uniquement sur les observations,

$$\left[\overline{X}_n - \frac{2.24}{\sqrt{n}}\ ,\ \overline{X}_n + \frac{2.24}{\sqrt{n}}\right] \ .$$

L'aléa (le fait qu'on ne soit sûr qu'à 95 % et pas 100 %) est causé par les événements atypiques, comme par exemple, obtenir au moins 90 fois pile sur 100 lancers, alors que la pièce est équilibrée. Cet événement a une probabilité faible mais non nulle, et quand on somme les probabilités de tous ces événements atypiques, que l'on exclut, on arrive à 5 %. Une autre remarque est qu'en grossissant l'intervalle, on peut en proposer un qui soit valide, à 99 % par exemple ; il suffit de remplacer les 5 % par 1 % dans la formule de détermination de ε_n. Enfin, pour être sûr à 100 % de son assertion, il est nécessaire

de proposer l'intervalle [0, 1], mais c'est ridicule : c'est comme si l'expérience statistique n'avait pas servi !

La morale, c'est qu'en statistique, pour avancer, il faut prendre des risques, accepter de ne pas avoir toujours raison. L'intervalle que l'on propose peut être faux si par malchance on a subi un événement atypique lors de la création de l'échantillon. Ainsi va la vie... et c'est qu'illustrera la figure 4 plus loin.

Application : Les valeurs observées $x_1, \ldots, x_{100}$, de moyenne $\overline{x}_{100} = 0.42$, conduisent à l'intervalle

$$\left[\overline{x}_{100} - \frac{2.24}{\sqrt{100}}\ ,\ \overline{x}_{100} + \frac{2.24}{\sqrt{100}}\right] = [0.196,\ 0.644] = [19.6\,\%,\ 64.4\,\%]$$

pour la valeur p_0. Pour l'instant, on ne peut pas encore conclure que la pièce est biaisée, même si on pourrait avoir tendance à le croire. Il faut continuer à lancer !

Remarque 1.5. Attention, s'il y a effectivement une probabilité 95 % au moins pour que p_0 appartienne à l'intervalle aléatoire construit sur les X_t, cette probabilité est 0 ou 1 concernant l'intervalle réalisé $[19.6\,\%,\ 64.4\,\%]$: en effet, p_0 appartient ou pas à un tel intervalle déterministe, il n'y a pas d'autre issue. Pour éclairer cette distinction, pensez à un tirage du Loto : étant donnée une grille, avant le tirage, il lui est associé une (faible) probabilité de gain du gros lot ; après le tirage, la grille est soit gagnante, soit perdante. L'indication de la probabilité *a priori* de 95 % dans le cas des intervalles de confiance nous rassure quant à la validité de ce que l'on obtient après l'expérience sans toutefois que ce sentiment rassurant puisse être une certitude.

Les intervalles obtenus ici sont un peu plus grands que nécessaires. Cela vient de deux choses. D'une part, par la majoration de la variance $p_0(1-p_0)$ par $1/4$: c'est d'autant plus grave que p_0 est proche de 0 ou de 1 (de 0 % ou de 100 %). D'autre part, par l'inégalité de Chebychev-Markov elle-même, qui est assez grossière. Une solution souvent plus efficace, quoiqu'asymptotique, est de recourir au théorème de la limite centrale.

2.3. Théorème de la limite centrale, énoncé et commentaires.

Théorème 1.2 (Théorème de la limite centrale). *Soient des variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées, de loi commune admettant un moment d'ordre deux, d'espérance et de variance communes notées μ et σ^2. Alors, on a la convergence en loi*

$$\frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu) \ \rightharpoonup\ \mathcal{N}(0,\,1)\ .$$

Rappelons tout d'abord la définition de la convergence en loi (vers une loi à densité : ce qui suit ne vaut pas en l'état pour la convergence vers une loi chargeant des points). C'est la convergence simple des fonctions de répartition. Ainsi, on a dans le théorème que pour tout réel x,

$$\mathbb{P}\left\{\frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu) \leqslant x\right\} \overset{\text{not.}}{=} F_n(x) \ \longrightarrow\ \mathbb{P}\{Z \leqslant x\} = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}\,dz \overset{\text{not.}}{=} \Phi(x)\ ,$$

où Z suit une loi normale standard $\mathcal{N}(0,\,1)$. Il s'en déduit évidemment, par soustraction, que pour tout intervalle $[a,\,b]$, on a également

$$\mathbb{P}\left\{a \leqslant \frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu) \leqslant b\right\} \ \longrightarrow\ \mathbb{P}\{a \leqslant Z \leqslant b\} = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}\,dx\ .$$

C'est cette seconde formulation que vous avez dû voir en classes préparatoires.

REMARQUE 1.6. Distinguez bien les symboles $\to$ (convergence "déterministe"), $\xrightarrow{\mathbb{P}}$ (convergence en probabilité) et $\rightharpoonup$ (convergence en loi).

REMARQUE 1.7. Avant d'intégrer cette noble école, vous aviez sans doute noté X_n^* la variable aléatoire intervenant dans le théorème de la limite centrale ; or, dans la version ci-dessus, on a

$$X_n^* = \frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu)\ ;$$

certes, vous aviez probablement utilisé la définition

$$X_n^* = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

ou par toute autre formulation similaire, mais persuadez-vous bien que ces définitions sont toutes équivalentes, cela procède d'un simple calcul. Je vous demanderais d'oublier désormais toutes ces notations X_n^*, etc. Pour des raisons qui vous apparaîtront claires et limpides plus tard, il vaut mieux s'intéresser à la moyenne empirique $\overline{X}_n$, recentrée, et la multiplier par le facteur $\sqrt{n}/\sigma$. Efforcez-vous donc d'énoncer le théorème de la limite centrale comme je l'ai fait plus haut.

On rappelle quelques règles de manipulations des lois normales. Si X suit une loi normale $\mathcal{N}(\mu,\, \sigma^2)$ alors $aX + b$ suit une loi normale $\mathcal{N}(b + a\mu,\, a^2\sigma^2)$. Le théorème de la limite centrale indique donc que $\overline{X}_n$ suit (approximativement) une loi normale de moyenne μ et de variance σ^2/n,

$$\overline{X}_n \overset{(d)}{\approx} \mathcal{N}\left(\mu,\, \frac{\sigma^2}{n}\right)\ .$$

(Evidemment, on a une égalité en distribution pour tout n lorsque les X_t sont elles-mêmes gaussiennes.) Autrement dit, la loi de $\overline{X}_n$ admet un pic gaussien en μ de plus en plus prononcé.

2.4. Construction de fourchettes à l'aide du théorème de la limite centrale. Si l'on repart de l'exemple des lancers de la pièce, alors le théorème donne

$$\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}(\overline{X}_n - p_0) \ \rightharpoonup\ \mathcal{N}(0,1)$$

d'où, en prenant $a = u$ et $b = -u$ pour u bien choisi,

$$\mathbb{P}\left\{\frac{\sqrt{n}}{\sqrt{p_0(1-p_0)}}|\overline{X}_n - p_0| \leqslant u\right\} \longrightarrow \mathbb{P}\{|Z| \leqslant u\} = 95\,\%\ ,$$

où Z suit une loi $\mathcal{N}(0,1)$. On appelle u un quantile de la loi normale ; nous verrons ultérieurement comment en trouver la valeur. Ici, on admet pour l'instant que $u = 1.96$. Comme

$$\frac{\sqrt{n}}{p_0(1-p_0)}|\overline{X}_n - p_0| \leqslant 1.96$$
$$\iff\ p_0 \in I_n \overset{\text{not.}}{=} \left[\overline{X}_n - \frac{1.96\,\sqrt{p_0(1-p_0)}}{\sqrt{n}},\ \overline{X}_n + \frac{1.96\,\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right]\ ,$$

on pourrait en déduire (dans un instant d'égarement) que l'intervalle de confiance contenant, asymptotiquement, le vrai paramètre p_0 est I_n. Mais I_n dépend lui-même de p_0 ! On

va donc proposer un intervalle $\widehat{I}_n$ le contenant et ne dépendant plus de p_0, par exemple, en utilisant que $1.96\sqrt{x(1-x)} \leqslant 1.96/2 \leqslant 1$,

$$I_n \subset \widehat{I}_n \overset{\text{not.}}{=} \left[\overline{X}_n - \frac{1}{\sqrt{n}},\ \overline{X}_n + \frac{1}{\sqrt{n}}\right] .$$

REMARQUE 1.8. Une convention utile pour la suite : on mettra un petit chapeau $\widehat{\ }$ à toutes les quantités qui ne dépendent que des observations. Notez bien la différence entre I_n, qui ne nous sert que d'un point du vue théorique, et $\widehat{I}_n$, qu'on peut calculer entièrement en injectant les valeurs observées dans sa formule.

Application : La précision d'une fréquence empirique est ainsi (pour un niveau de confiance asymptotiquement à 95 %) de $\pm 1/\sqrt{n}$. Pour l'exemple de la pièce, le vrai paramètre est donc de $42\,\% \pm 10\,\%$. On ne peut donc pas encore exclure le cas où elle serait équilibrée. Remarquez cependant que l'intervalle calculé est plus petit que celui précédemment obtenu par l'inégalité de Chebychev-Markov (mais il a une garantie théorique un peu plus faible, parce qu'asymptotique).

REMARQUE 1.9 (A propos des sondages d'opinion). On les mène souvent par téléphone, en essayant d'interroger un échantillon "représentatif" (i.e., dont les éléments sont tirés au hasard dans la population mais en se fixant par avance une répartition géographique et socio-professionnelle). L'échantillon est généralement composé de 1 000 personnes. Quelle est la précision attendue ? $1/\sqrt{1\,000}$, soit 3 %. Peut-être mieux, puisque l'échantillon se veut représentatif. Mais dans tous les cas, que penser de ces unes de journaux télévisés annonçant, en 2007, que les taux de projets de vote en faveur de tel ou tel candidat à la présidentielle ont évolué de 1 % depuis le dernier sondage ? Rien de bon : une si petite variation est sans doute uniquement causée par l'aléa de tirage de l'échantillon !

2.5. Deux dessins pour conclure sur les intervalles de confiance. La figure 4 montre que sur 100 réalisations d'un intervalle de confiance à 95 %, environ 5 % ne contiennent pas le vrai paramètre ! (Ici, on s'est fixé le paramètre p_0 et on a fait comme si on ne le connaissait pas.)

La figure 5 vous montre le genre d'énoncés (pas nécessairement romantiques, on est d'accord) que peut être amené à effectuer un statisticien.

3. Tests d'hypothèses : Pourquoi deux résultats contradictoires peuvent être simultanément vrais !

Les tests illustrent combien le traitement statistique est politique. Un même jeu de données peut amener à des conclusions (ou plutôt, des absences de conclusion) radicalement opposées, selon le préjugé que l'on cherche à infirmer.

Un test part de deux hypothèses H_0 et H_1.
- H_0 est l'hypothèse de laquelle on ne veut se départir que si on a une excellente raison de le faire ; c'est par exemple une hypothèse de prudence (le caractère dangereux de tel médicament, telle théorie scientifique actuellement tenue pour vraie).
- H_1 est l'hypothèse alternative, celle vers laquelle on se tournerait si H_0 était révélée fausse par le test.

EXEMPLE 1.1. Considérons un nouvel exemple. On se demande s'il est opportun de changer le nom d'une enseigne (par exemple, "Camaïeu Homme" vers "Jules", cela est arrivé

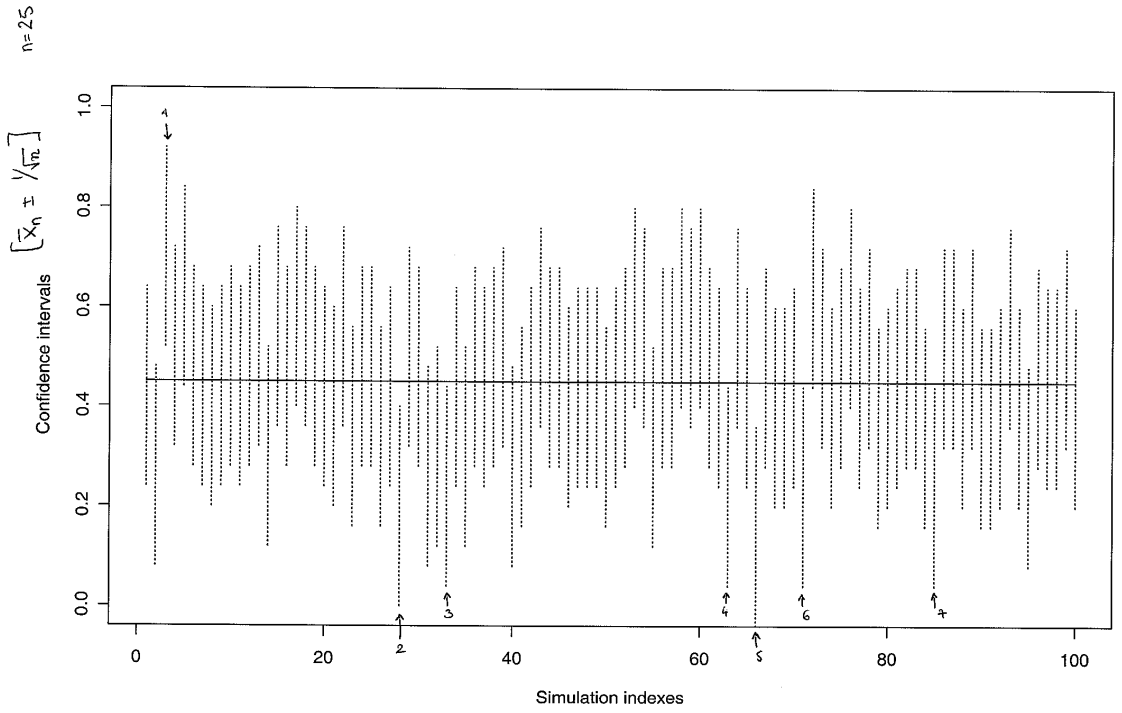


FIGURE 4. 100 répétitions de la même expérience : on tire informatiquement $X_1, \ldots, X_{25}$ indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre $p_0 = 0.45$ et on construit à chaque fois $\widehat{I}_{25}$; on s'est trompé 7 fois sur les 100.

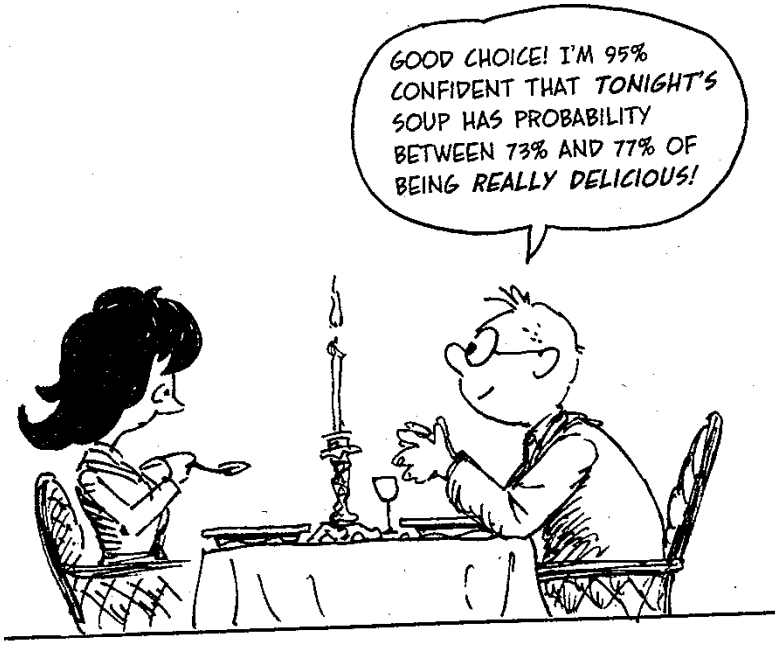


FIGURE 5. La statistique donne des intervalles de confiance. C'est utile, mais pas nécessairement romantique.

en l'an 2000 ; ou AGF vers Allianz en 2009) ou d'un produit. Bien sûr, un tel changement a un coût non négligeable (il faut refaire de nombreuses en-têtes et faire une campagne de communication), mais si le changement plaît aux consommateurs, cela pourra avoir un impact positif sur les ventes. On interroge un panel de consommateurs ; chacun indique s'il préfère la nouvelle (1) ou l'ancienne (0) dénomination. On note $x_1, \ldots, x_{293}$ les 293 réponses exploitables (autres que "je n'ai pas d'avis") obtenues en interrogeant des clients aux caisses des magasins pendant un samedi partout en France. Parmi elles, il y a eu 156 réponses en faveur du nouveau nom.

Pour les mêmes raisons que plus haut, on peut modéliser le problème en partant de ce que l'on a affaire à des observations $X_1, \ldots, X_{293}$ indépendantes et identiquement distribuées, selon une loi de Bernoulli de paramètre noté p_0.

REMARQUE 1.10. Ici, contrairement au cas du lancer de la pièce, on a une interprétation statistique de p_0 : c'est la vraie proportion (inconnue) de la population qui a une opinion positive en faveur du nouveau nom. On a tiré un échantillon de la population, à qui on a posé la question. A partir des réponses de l'échantillon, on veut tirer des conclusions sur p_0.

Un directeur du marketing prudent partira de $H_0 : p_0 \leqslant 1/2$ (qui correspond à H_0 : "les clients préfèrent l'ancienne dénomination"), à cause du coût au changement. L'hypothèse alternative serait $H_1 : p_0 > 1/2$. On peut calculer comme précédemment un intervalle de confiance symétrique autour de $\overline{x}_{293} = 156/293 = 53.2\,\%$; on obtient, par exemple par théorème de la limite centrale et avec un niveau de confiance asymptotique à 95 %, la réalisation suivante, au vu des données : l'intervalle $[47.4\,\%, 59.0\,\%]$. (Refaites le calcul pour vous entraîner !)

La conclusion est que les données, exploitées via la construction de l'intervalle de confiance, ne contredisent pas gravement H_0, et qu'on conserve donc cette dernière : il n'y a pas d'argument statistiquement significatif contre une absence de changement de dénomination. Les données ne contredisent pas assez gravement l'hypothèse prudente de laquelle le directeur du marketing est parti et cette hypothèse est conservée, malgré le sentiment que dégage la valeur de $\overline{x}_{293}$.

Les tests, comme le montrent les compléments, ne sont pas réductibles aux intervalles de confiance et ont leur logique propre, qui peut dérouter au premier abord (et qui explique pourquoi deux faits contradictoires peuvent être tenus pour vrais simultanément : quand pour aucun d'entre eux on n'a assez d'arguments statistiques pour le rejeter).

Compléments pour étudiants avancés

Ce qui suit est un complément sur la théorie des tests. Il pourra vous sembler abscons ou difficile à comprendre. Tout devrait aller mieux après le cours formel sur les tests, d'ici quelques semaines : la présentation rapide et le paradoxe reproduits ci-dessous sont un simple galop d'essai pour vous !

4. Théorie des tests, suite

A l'exemple 1.1, pour effectuer le test de $H_0 : p_0 \leqslant 1/2$ au vu de $H_1 : p_0 > 1/2$, il serait plus efficace de procéder comme suit. On a intuitivement envie de rejeter H_0 lorsque la statistique $\overline{X}_n$ dépasse un certain seuil, qu'il nous faut simplement quantifier. Si H_0 est vraie, alors dans le pire des cas, $p_0 = 1/2$ et il s'agit de déterminer le seuil maximal que $\overline{X}_n$ atteint à cause de ses fluctuations aléatoires autour de $1/2$. Le théorème de la limite centrale indique que, Z suivant une loi normale $\mathcal{N}(0,1)$,

$$\mathbb{P}\left\{\frac{\sqrt{n}}{\sqrt{1/4}}\left(\overline{X}_n - \frac{1}{2}\right) \leqslant u\right\} \longrightarrow \mathbb{P}\{Z \leqslant u\} = 95\,\%$$

pour $u = 1.65$ (là encore, on admet pour l'instant la détermination de cette valeur). Cela veut dire qu'avec probabilité asymptotique de 95 %, on a

$$\overline{X}_n \leqslant \frac{1}{2} + \frac{1.65}{2\sqrt{n}}\,, \quad \text{soit, pour } n = 293, \quad \overline{X}_n \leqslant 54.8\,\%\,.$$

Puisque $\overline{x}_n = 53.2\,\%$, les données sont encore dans la limite des observations typiques lorsque $p_0 = 1/2$ et on ne rejette pas H_0.

Exercice 1.1. Un directeur du marketing manipulateur ou peu prudent aurait pris $H_0' : p_0 > 1/2$ et $H_1' : p_0 \leqslant 1/2$. Il aurait rejeté H_0' si $\overline{X}_n$ était descendu en-dessous d'un certain seuil. Montrez, par un calcul similaire à celui qui précède, que ce seuil vaut 45.2 %. En conséquence de quoi, il aurait lui aussi conservé son hypothèse H_0'.

Remarque 1.11. Notez donc que les tests conservent à la fois H_0 et H_0', et ce, sur le même jeu de données, alors que ces deux hypothèses sont la négation l'une de l'autre ! Le choix de l'hypothèse de départ est ainsi hautement politique, puisqu'un test tend à la conserver et à ne s'en départir que pour d'excellentes raisons, lorsque les données la contredisent gravement. Les tests traitent H_0 et H_1 de manière dissymétrique. Cela vous explique pourquoi association de consommateurs et industriels peuvent tirer des conclusions radicalement différentes à partir des mêmes observations. En vérité, tant qu'on conserve l'hypothèse de départ, on n'a pas vraiment progressé. Comme le montre l'exercice suivant, c'est peut-être qu'on la conserve faute de mieux. En revanche, rejeter une hypothèse de départ et passer à l'hypothèse alternative est signe que l'on a appris quelque chose sur les données. A retenir : en statistiques, une hypothèse est vraie tant qu'elle n'a pas été gravement contredite ! C'était la première minute citoyenne...

EXERCICE 1.2. Montrez, avec les techniques précédentes, que, dans le premier test, H_0 est en revanche rejetée lorsque l'on interroge 2 930 clients et que 1 560 d'entre eux se prononcent en faveur du nouveau nom. Notez que la proportion de réponses positives est pourtant toujours de 53.2 % ! Cela illustre le fait qu'un test est d'autant plus informatif que l'échantillon est grand.

On verra dans le cours à venir sur les tests qu'en réalité, plutôt que de dire simplement qu'on garde ou non H_0, on pourra quantifier notre attachement à elle et en fournir un degré de crédibilité. (Souvenez-vous toujours que l'objet de la statistique mathématique est la quantification des impressions.) En sorte, à partir d'observations contenant un certain aléa, on arrive à dire combien on est confiant qu'une certaine assertion déterministe (formée par H_0) est vraie ou non.

Exercices

EXERCICE 1.3. Ecrivez l'application du théorème de la limite centrale à des variables aléatoires indépendantes et identiquement distribuées selon une loi exponentielle, puis selon une loi géométrique.

EXERCICE 1.4. Calculer les intervalles de précision autour des scores prédits pour Lyne Cohen-Solal et Jean Tibéri par la méthode fondée sur l'inégalité de Chebychev-Markov. Etes-vous en mesure de prédire le gagnant ? Que pouvez-vous dire de Philippe Meyer ?

EXERCICE 1.5. Reprenez les questions de l'exercice précédent avec la méthode fondée sur le théorème de la limite centrale. Les résultats en sont-ils changés ? Comment expliquer alors la confusion des résultats tels que clamés à 20h45 puis amendés à 21h15 ?

EXERCICE 1.6 (Un article de la presse scientifique). Lisez l'article reproduit à la figure 6 ("Euro coin accused of unfair flipping", 4 janvier 2002, tiré du *New Scientist*) : écrivez le modèle statistique lié à la situation étudiée et tâchez de retrouver par le calcul la preuve et/ou la valeur de l'ensemble des résultats énoncés.

EXERCICE 1.7 (Pour étudiants avancés). Traitez les exercices 1.1 et 1.2.

Euro coin accused of unfair flipping

Click to Print

17:39 04 January 2002
From New Scientist Print Edition. Subscribe and get 4 free issues.
Debora MacKenzie

The introduction of the Euro, the largest currency switch in history, has proceeded with few problems - until now. Polish statisticians say the one Euro coin, at least in Belgium, does not have an equal chance of landing "heads" or "tails". They allege that, when spun on a smooth surface, the coin comes up heads more often.

The observation is not to be taken lightly on a sports-mad continent where important decisions can turn on the flip of a coin. But the accusation of bias has been countered by statistical analysis from, of all places, Euro-sceptic Britain. The UK is one of only three EU countries that have not adopted the common currency.

Tomasz Gliszczynski and Waclaw Zawadowski, statistics teachers at the Akademia Podlaska in Siedlce, received Belgian Euro coins from Poles returning from jobs in Belgium and immediately set their students spinning them. Gliszczynski says spinning is a more sensitive way of revealing if a coin is weighted than the more usual method of tossing in the air.

The students had already spun the Polish two-zloty piece more than 10,000 times to show it was biased. But for the Belgian Euro, they have so far managed only 250 spins.

Of these, 140, or 56.0 per cent, came up heads. Glyszczynski attributes such assymetry to a heavier embossed image on one side of the coin. All Euros have a national image on the "heads" side and a common design on the "tails". Belgium portrays its portly king, Albert, on the heads side.

Not significant

But Howard Grubb, an applied statistician at the University of Reading, notes that, "with a sample of only 250, anything between 43.8 per cent and 56.2 per cent on one side or the other cannot be said to be biased".

This is because random variation can produce such scatter even if the coin is truly unbiased. With a larger number of spins, such randomness would even out and results would approach 50:50.

The range of 6.2 per cent on either side of 50 per cent is expected to cover the results, even with a fair coin, in 95 of every 100 experiments. Nonetheless, Grubb cautions, the Polish result is at the outside of this range, and would be expected in only about 7 of every 100 experiments with a fair coin, leaving a glimmer of hope for their hypothesis. Clearly, more research is needed.

Gliszczynski plans to continue his experiments - aimed mainly at teaching his students statistics - with the German Euro, which has an eagle on its heads side, and present them at a conference in February.

New Scientist carried out its own experiments with the Belgian Euro in its Brussels office. Heads came up five per cent less often than tails. This looks like the opposite of the Polish result but in fact - in terms of statistical significance - it is the same one.

FIGURE 6. Controverses autour d'une étude statistique : les données ne permettent pas de conclure. Tiré de *New Scientist*, 4 janvier 2002.

PARTIE 1 Panorama de la démarche statistique / Corrigés

Exercice 1.

- Loi exponentielle $\mathcal{E}(\lambda)$, $\lambda \in \mathbb{R}_+^*$: espérance $1/\lambda$ et variance $1/\lambda^2$

d'où $\lambda\sqrt{n}\left(\bar{X}_n - 1/\lambda\right) \longrightarrow \mathcal{N}(0,1)$

ou $\sqrt{n}\left(\bar{X}_n - 1/\lambda\right) \longrightarrow \mathcal{N}(0, 1/\lambda^2)$

- Loi géométrique $\mathcal{G}(p)$, $p \in]0,1[$ (sur $\mathbb{N}^*$) : espérance $1/p$ et variance $\frac{1-p}{p^2}$

d'où $\frac{p}{\sqrt{1-p}}\sqrt{n}\left(\bar{X}_n - 1/p\right) \longrightarrow \mathcal{N}(0,1)$

ou $\sqrt{n}\left(\bar{X}_n - 1/p\right) \longrightarrow \mathcal{N}\left(0, \frac{1-p}{p^2}\right)$

Exercice 2.

- Méthode : il faut se ramener à des variables aléatoires iid selon une loi de Bernoulli de paramètre le paramètre d'intérêt.

- On rappelle le modèle : $X_1, \dots X_{978}$ iid $\sim (p_{CS}, p_T, p_M)$
Rappel : avant 22h, la valeur de ce triplet est inconnue.

Alors $Y_t = \mathbb{1}_{\{X_t \text{ a voté CS}\}} \sim \text{Ber}(p_{CS})$

$Y_1, \dots Y_{978}$ iid $\sim \text{Ber}(p_{CS})$

et $\left[\bar{Y}_{978} \pm \frac{2.24}{\sqrt{978}}\right]$ est un intervalle de confiance à 95% sur p_{CS}

Vu $\bar{y}_{978} = \frac{463}{978} \simeq 0.473$ ($= 47.3\%$), on obtient la réalisation :

$$\left[\ 0.473 \pm 2.24/\sqrt{978}\ \right] = \left[\ 40.2\ \%\ ;\ 54.4\ \%\ \right]$$

De même, la réalisation d'un intervalle de confiance à 95% sur p_T est

$$\left[\ \frac{409}{978} \pm 2.24/\sqrt{978}\ \right] = \left[\ 34.7\%;\ 48.9\%\ \right]$$

et pour p_M, on obtient :

$$\left[\ \frac{106}{978} \pm \frac{2.24}{\sqrt{978}}\ \right] = \left[\ 3.7\%;\ 18.0\%\ \right]$$

On regarde quels intervalles sont disjoints (ou pas)

- Conclusions ? Philippe Meyer est élu, mais pour les deux autres candidats, on ne peut pas encore conclure : on n'a pas encore assez de données ! Avec plus de données, la taille des intervalles sera plus resserrée.

↳ Les chiffres ne parlent donc vraiment pas d'eux-mêmes !

Exercice 3

On va obtenir des intervalles plus petits (méthode par TCL plus efficace) : essentiellement, on remplace le facteur 2.24 par 1 dans les formules ci-dessus et on obtient les réalisations :

- pour p_S : $\left[\ 0.473 \pm 1/\sqrt{978}\ \right] = \left[\ 44.1\%;\ 50.5\%\ \right]$
- pour p_T : $\left[\ 0.418 \pm 1/\sqrt{978}\ \right] = \left[\ 38.6\%;\ 45.0\%\ \right]$

Ils ne sont toujours pas disjoints mais ils le seraient si on prenait des intervalles de confiance à 90% (quantile 1.65 au lieu de 1.96) plutôt qu'à 95% :

$$44.5\% = 0.418 + \frac{1.65}{2\sqrt{978}} < 0.473 - \frac{1.65}{2\sqrt{978}} = 44.7\%$$

C'est ce qu'ont dû faire les sondeurs, augmentant ainsi leur probabilité de se tromper ... ce qui est ennuyé !

Exercice 4.

Répétition d'une expérience avec succès ou échec :

$X_1, \dots X_{250}$ iid $\sim Ber(p_0)$

avec $p_0 \in]0,1[$

Interprétation : p_0 est le paramètre d'équilibrage ($p_0 = 1/2$ si et seulement si la pièce est équilibrée)

Face (heads) $= 1$, Pile (tails) $= 0$

Données : $x_1, \dots x_{250}$

avec $\bar{x}_{250} = \frac{140}{250} = 0.56$

On veut tester $H_0 : p_0 = 1/2$ contre $H_1 : p_0 \neq 1/2$

TECHNIQUE #1 (celle fondée sur les intervalles de confiance) :

La réalisation d'un intervalle de confiance à 95% sur p_0 est

$$[\bar{x}_{250} \pm \underbrace{1/\sqrt{250}}_{\text{voire, pour être plus précis : } \frac{1.96}{2\sqrt{250}} \simeq 0.062 = 6.2\%}] = [56.0\% \pm 6.2\%] = [49.8\%, 62.2\%]$$

↳ on ne peut pas encore exclure $p_0 = 1/2$, on conserve H_0 faute mieux.

Il aurait fallu avoir plus de données. Conclusion : répétez l'expérience en réalisant davantage de lancers.

TECHNIQUE #2 (celle de l'article du New Scientist, cf compléments du cours)

Sous H_0, $\bar{X}_{250}$ tend (par loi des grands nombres) à être

proche de $1/2$.

Avec probabilité $\simeq 95\%$, $\bar{X}_{250} \in \left[1/2 \pm \frac{1.96}{2\sqrt{250}} \right]$

(Cela découle du TCL par des calculs similaires à ceux effectués dans les compléments.)

$= [43.8\% ; 56.2\%]$

Or $\bar{x}_{250} = 56.0\%$ est dans cet intervalle de valeurs typiques, on n'a donc aucune raison de rejeter H_0.

Mais avec probabilité $\simeq 93\%$ $\bar{X}_{250} \in \left[1/2 \pm \frac{1.81}{2\sqrt{250}} \right]$

$= [44.3\% ; 55.7\%]$

et cette fois, $\bar{x}_{250}$ n'est plus dans l'intervalle.

(Note : on verra plus tard comment le 93% et le 1.81 sont liés ; admettez le lien pour l'instant.)

L'article souligne cependant que baisser le niveau de confiance de 95% à 93% est dangereux.

Exercice 5. (Pour étudiants avancés)

- Premier sous-exercice : Sous H'_0, le pire cas (limite) est toujours $1/2$, et le seuil est obtenu comme

$$1/2 - \frac{1.65}{2\sqrt{293}} = 45.2\%$$

$\bar{x}_{293}$ étant au-dessus de ce seuil, on conserve H'_0.

- Second sous-exercice : Seuil cette fois à

$$1/2 + \frac{1.65}{2\sqrt{2930}} = 51.5\%$$

... et on lance le changement de nom !

et $\bar{x}_{2930} = 53.2\%$ est au-dessus : on rejette $H_0 : p_0 \leq 1/2$ en faveur de $H_1 : p_0 > 1/2$

Deuxième Partie

Recueil, représentation et modélisation de données

Version rédigée du cours

Objectifs. Il va s'agir de comprendre et de mettre en application le mot d'esprit suivant :

> *In God we trust, all others bring data.*
> Edwards Deming (universitaire américain, consultant pour l'industrie, 1900–93)

En entreprise (pour lancer un produit) ou dans les déclarations gouvernementales (pour montrer l'impact des réformes), on n'est crédible que lorsque l'on s'appuie sur des données.

Ce chapitre explique comment, à partir de données recueillies [2] par exemple lors d'enquêtes, définir un modèle statistique permettant une analyse mathématique. On verra notamment que cela revient à postuler la forme de la loi sous-jacente aux observations et à la connaître à quelques paramètres près. L'étude ultérieure portera justement sur l'inférence de ces paramètres.

1. Les différents types de données

On explique tout d'abord comment lire un tableau de données et comment démarrer la formalisation mathématique. On verra plus tard comment bien recueillir des données.

On commence par un exemple. Dans le fichier `hourlywagedata.sav` (disponible sur le site web du cours), on a regroupé les salaires horaires d'infirmières américaines. Les données se présentent comme indiqué à la figure 7 : sous la forme d'une matrice, donc. Chaque ligne correspond à une infirmière différente. C'est la convention dans les fichiers de données : on utilise une ligne par groupe de valeurs observées.

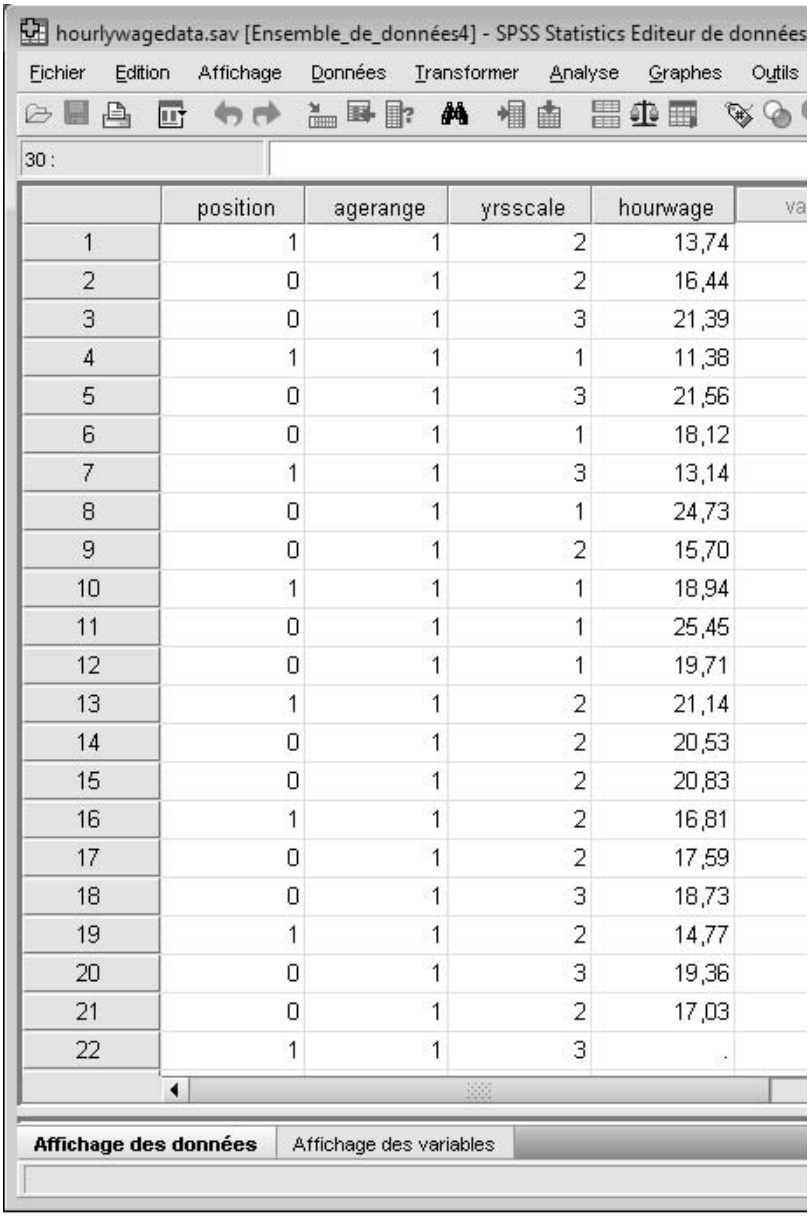
On remarque que la matrice ne comporte que des nombres.

La minute SPPS 2.1. En cliquant sur l'onglet `Affichage des variables`, on découvre le sens de certains nombres et on peut afficher ce que l'on va appeler ci-dessous la table de correspondance.

Ainsi, on voit que la première colonne indique si l'infirmière travaille à l'hôpital (0) ou en cabinet (1) ; la deuxième donne sa tranche d'âge (1 si elle a entre 18 et 30 ans, 2 entre 31 et 45 ans, 3 entre 46 et 65 ans) ; la troisième nous renseigne sur son ancienneté dans le métier (on lit un nombre entre 1 et 6, selon qu'elle exerce depuis moins de 5 ans, entre 6 et 10 ans, etc.) ; enfin, la quatrième colonne donne son salaire horaire. Au final, on peut écrire la table de correspondance indiquée au tableau 1.

Remarque 2.1. Il n'y a pas de variable pour décrire le sexe ! On semble donc n'avoir interrogé que des femmes. Et les infirmiers, alors ?

2. Ce qui distingue la statistique de la voyance ? Le fait que la première a besoin de données pour se prononcer. Récemment, un ancien élève ayant le projet de lancer une start-up est venu me voir, il voulait que je l'aide à modéliser le marché auquel il comptait s'attaquer. Avait-il des données, issues de coups de sonde préliminaires ? Que nenni, mais j'étais mathématicien, alors forcément j'allais savoir effectuer la modélisation, dans son esprit. Ne répétez pas sa méprise !



	position	agerange	yrsscale	hourwage
1	1	1	2	13,74
2	0	1	2	16,44
3	0	1	3	21,39
4	1	1	1	11,38
5	0	1	3	21,56
6	0	1	1	18,12
7	1	1	3	13,14
8	0	1	1	24,73
9	0	1	2	15,70
10	1	1	1	18,94
11	0	1	1	25,45
12	0	1	1	19,71
13	1	1	2	21,14
14	0	1	2	20,53
15	0	1	2	20,83
16	1	1	2	16,81
17	0	1	2	17,59
18	0	1	3	18,73
19	1	1	2	14,77
20	0	1	3	19,36
21	0	1	2	17,03
22	1	1	3	.

FIGURE 7. Les 22 premières lignes d'un fichier de données en contenant 3 000 ; notez l'indication d'une donnée manquante en ligne 22 (symbole point).

On peut noter les données de la ligne i par (p_i, a_i, y_i, h_i). Par exemple, $p_5 = 0$, $a_5 = 1$, $y_5 = 3$, $h_5 = 21.56$. Le fichier comporte 3 000 lignes, mais on voit que 89 cases de salaires ne sont pas renseignées. (Cela peut être causé par la difficulté à lire des renseignements récoltés sur papier ou par des refus de réponse à l'enquête.) Il reste donc 2 911 quadruplets (p_i, a_i, y_i, h_i) totalement exploitables.

LA MINUTE SPPS 2.2. Pour compter les valeurs manquantes, on utilise `Analyse / Statistiques descriptives / Effectifs`.

On va commencer par indiquer le type et l'étendue des données. Le salaire est une donnée dite quantitative, parce qu'elle mesure quelque chose (en l'occurence, le salaire horaire). Ici, il est difficile de dire si la mesure effectuée est discrète ou continue : on saute

Nom de la variable	Description	Significations	Valeurs
`position`	Type d'institution	Hôpital	0
		Cabinet	1
`agerange`	Age	Entre 18 et 30 ans	1
		Entre 31 et 45 ans	2
		Entre 46 et 65 ans	3
`yrsscale`	Ancienneté dans le métier	Inférieure à 5 ans	1
		Entre 6 et 10 ans	2
		Entre 11 et 15 ans	3
		Entre 16 et 20 ans	4
		Entre 21 et 35 ans	5
		Plus de 36 ans	6
`hourwage`	Salaire horaire		

TABLE 1. Table de correspondance pour les données de la figure 7.

de centime en centime, la mesure est donc de facto discrète, mais au vu du grand nombre de valeurs possibles et de la finesse de la grille, c'est presque une mesure continue ! L'étendue de ces données (i.e., l'ensemble des valeurs possibles) est, disons l'intervalle $[4, +\infty[$ (où 4 serait le salaire horaire minimum garanti, par exemple, s'il existe).

D'autres exemples de données quantitatives qu'on aurait pu recueillir pour ces infirmières seraient leur âge (quand il n'est pas codé en catégories comme ici), le nombre d'enfants à leur foyer, le nombre de kilomètres qu'elles parcourent pour se rendre à leur hôpital, etc.

Les trois autres séries données du tableau 1 sont, elles, qualitatives : elles indiquent une catégorie. Les deuxième et troisième séries sont dites qualitatives ordinales : elles font référence à des catégories que l'on peut classer entre elles. Leurs étendues respectives sont $\{1, 2, 3\}$ et $\{1, 2, 3, 4, 5, 6\}$. La première série de données, celle qui indique le type du lieu de travail, est nominale : on a deux catégories, mais sur lesquelles on n'a pas d'ordre évident (est-il mieux de travailler en cabinet ou à l'hôpital ?). Pour ces trois séries de données, le lien entre la valeur de la variable et l'information sous-jacente (lieu d'exercice, âge, temps d'expérience) est réalisé par ce que l'on a appelé ci-dessus une table de correspondance.

Dans d'autres registres, d'autres exemples de variables qualitatives ordinales seraient la position dans une fratrie ou le rang dans un classement de performance de vendeurs, de même que toute variable catégorisant une variable quantitative sous-jacente. Quant aux variables qualitatives nominales, on peut penser à des variables décrivant la marque du véhicule que l'on conduit, la région dans laquelle on est né, etc.

On résume la discussion précédente dans le tableau 2.

LA MINUTE SPPS 2.3. Dans l'onglet `Affichage des variables`, on voit également, dans la dernière colonne, intitulée `Mesure`, le type de chacune des séries de données.

Données			
Qualitatives		Quantitatives	
Nominales	Ordinales	Discrètes	Continues
Marque de voiture	Rang d'un classement	Nombre d'enfants	Salaire
Sexe	Niveau d'éducation	Nombre de ventes	Taille
Statut conjugal	Degré de satisfaction		

TABLE 2. Résumé de la discussion sur les différents types de données. Note : les données qualitatives doivent être codées par des entiers (1, 2, 3, etc.) lors de leur traitement sous un logiciel statistique.

2. Le recueil des données et son but

On va expliquer maintenant
- comment bien recueillir des données ;
- pourquoi on veut recueillir des données.

On verra que les deux questions sont fortement liées.

2.1. Pourquoi recueillir des données ? On veut étudier les caractéristiques d'une population donnée. Dans l'exemple précédent, il s'agit de l'ensemble des infirmières des Etats-Unis, pour qui l'on veut connaître le salaire horaire moyen. Comme il est trop coûteux en temps et en argent d'interroger absolument toutes les infirmières, on en tire un échantillon. (On verra plus bas qu'une bonne manière de choisir un échantillon est de le tirer au hasard.) On résume ce principe d'impossibilité quasi-physique de connaissance de la population par l'image de la figure 8.

On constitue donc un échantillon, composé dans l'exemple introductif de 3 000 infirmières, on en interroge ses membres et on reporte les réponses dans le tableau de données. Evidemment, telle ou telle réponse est inexploitable, et c'est ce qui donne les 2 911 groupes de données complètes de la figure 7.

REMARQUE 2.2. D'une manière générale, à cause de ces réponses inexploitables, il faut souvent interroger plus de gens que de réponses exploitables attendues. Pire, si l'on veut sélectionner un échantillon dans une sous-population bien particulière (par exemple, les parents ayant au moins trois enfants), il faudra interroger beaucoup plus de personnes que de réponses exploitables attendues, en commençant par leur demander s'ils ont, ou non, trois enfants au moins. Pensez aux sondeurs ou aux enquêteurs de rue : ils commencent souvent par vous poser une ou deux questions afin de savoir dans quelle sous-population vous ranger (selon votre âge, vos revenus, votre statut de locataire ou de propriétaire).

2.2. Comment bien recueillir des données ? Au hasard ! Ne prolongeons pas le suspense : il s'agit de collecter les individus de l'échantillon au hasard (voir figure 9), par tirage aléatoire uniforme dans la population. Dès que la population est suffisamment grande et/ou que l'échantillon qui est en tiré est suffisamment petit, alors on peut supposer que les données sont la réalisation de variables aléatoires indépendantes et identiquement

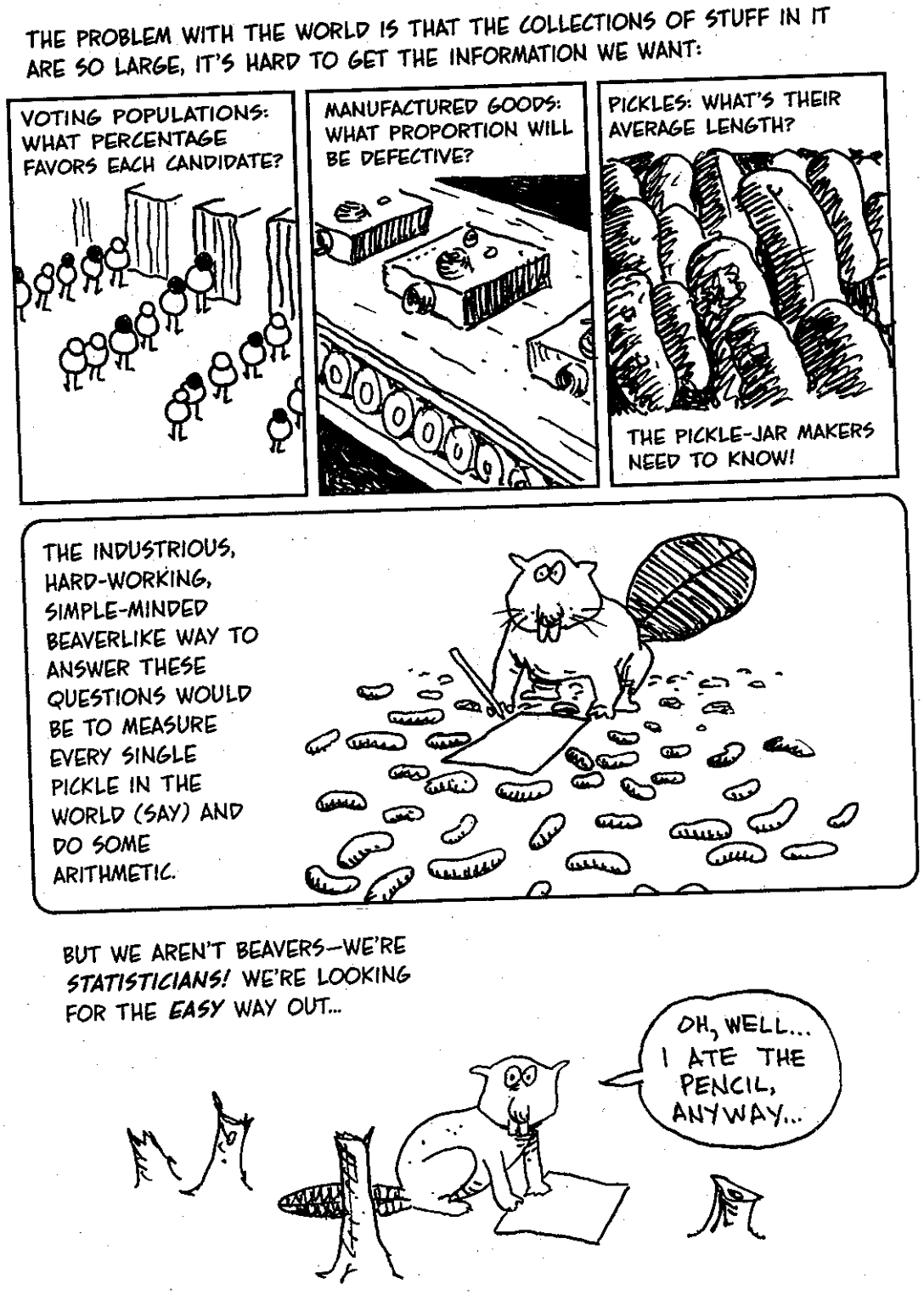


FIGURE 8. Vous n'êtes pas des petites bêtes industrieuses : en conséquence de quoi, vous ferez désormais des sondages plutôt que des décomptes exhaustifs sur des populations immenses.

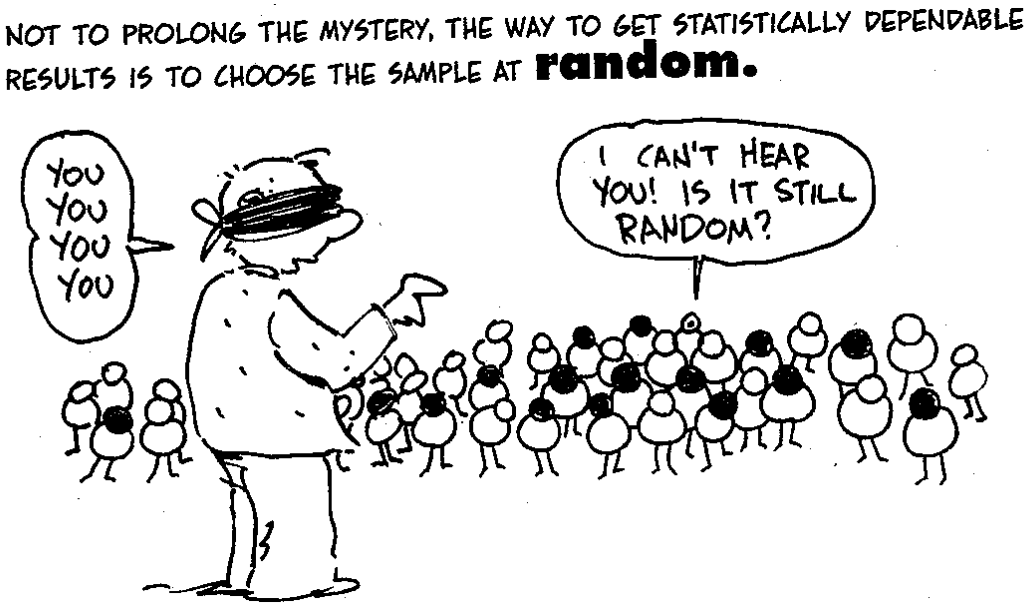


FIGURE 9. La bonne manière de collecter des données : interroger au hasard !

distribuées : cela procède de l'approximation d'une suite de tirages sans remise par une suite de tirages avec remise.

EXEMPLE 2.1. Les données de salaire $h_1, \ldots, h_{2911}$ des infirmières peuvent être considérées comme la réalisation d'un échantillon $H_1, \ldots, H_{2911}$ de variables aléatoires indépendantes et identiquement distribuées. La loi commune est celle qui gouverne le salaire dans la population ; elle est inconnue et c'est justement l'objet de ce cours que de pouvoir tirer des choses sur elle à partir des données observées.

REMARQUE 2.3. On rappelle que lorsque l'on passe des données observées à la modélisation par des variables aléatoires, on passe des symboles en minuscules $h_1, \ldots, h_{2911}$ à ceux en majuscules, $H_1, \ldots, H_{2911}$.

Le caractère d'observations indépendantes et identiquement distribuées provient donc de la méthode aléatoire employée pour récolter les données. Si l'on a effectivement interrogé des infirmières dans de nombreux lieux et services différents, toutes choisies au hasard dans des lieux eux-mêmes choisis au hasard, tout ira bien. (Il faut pour cela disposer d'une liste nationale de toutes les infirmières et d'un ordinateur y tirant des noms uniformément au hasard.) Si en revanche l'on ne s'est contenté que d'un sondage aléatoire dans un ou deux hôpitaux, alors l'échantillon risque de n'être représentatif que de ces hôpitaux, et pas de l'ensemble de ceux du pays. On ne pourra alors rien dire de garanti statistiquement sur la loi d'intérêt (le salaire horaire moyen de toutes les infirmières américaines). Il faut aussi faire attention à la manière d'interroger : l'indépendance entre les H_j provient également, d'une part, du fait qu'on interroge chaque infirmière chacune à son tour (sinon, en grand groupe, les dires des uns influencent ceux des autres) et d'autre part, du fait qu'on n'en prend pas trop dans le même endroit (en un même lieu, il y a des échelles de salaire locales ; ou alors la promotion de l'une, donc le meilleur salaire de l'une, empêche la progression des autres, etc.).

EXEMPLE 2.2 (Les sondages téléphoniques). On interroge 1 000 personnes au hasard au téléphone quant à la notoriété d'un produit, en tirant des numéros à 10 chiffres au hasard dans l'annuaire et en les composant automatiquement. Alors, les observations pourront bien être modélisées par des variables aléatoires indépendantes et identiquement distribuées. Le seul point d'attention serait ici la commune distribution : c'est plutôt la notoriété moyenne accordée par ceux qui ont une ligne de téléphone fixe que l'on essaie d'évaluer ainsi. Pour mémoire, une bonne partie des jeunes de 18 à 30 ans ne dispose que de téléphones portables, et pas de lignes fixes ! On ne peut donc les interroger par sondage... et cela commence à devenir un véritable souci pour nos amis sondeurs, qui n'ont plus de moyens naturels (et aléatoires) de contacter cette population. Un annuaire des téléphones portables leur serait d'un grand secours ! Avec un peu d'humour, on peut aussi imaginer que ceux qui possèdent plusieurs lignes faussent également l'échantillonnage (voir la figure 10).

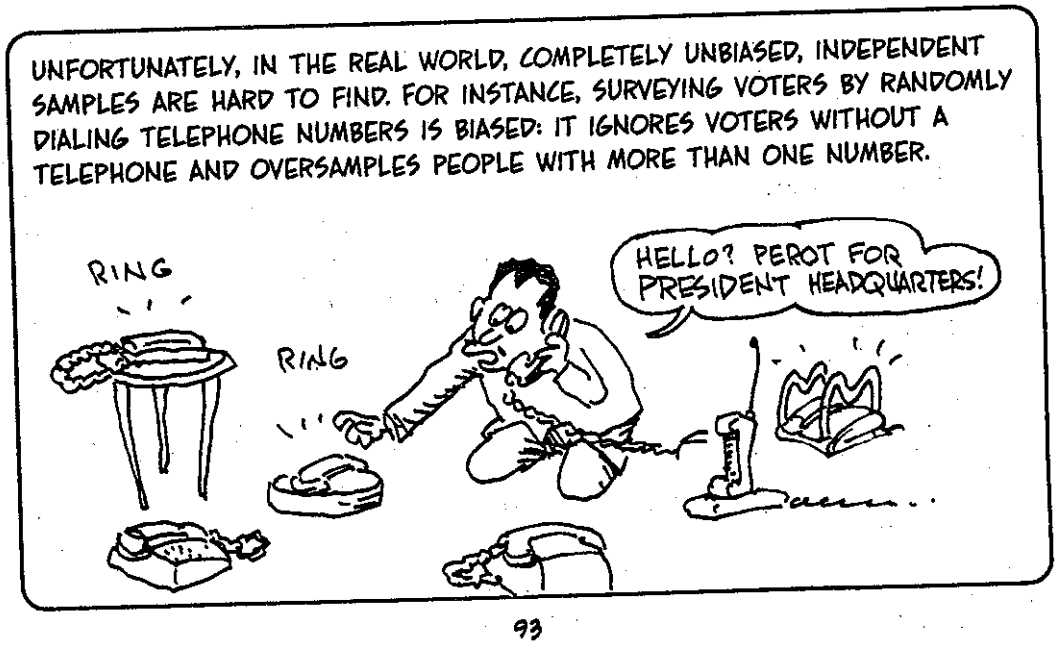


FIGURE 10. Il est difficile d'obtenir un échantillon aléatoire bien représentatif lors d'un sondage téléphonique aléatoire : on n'a pas accès à l'opinion de ceux qui n'ont pas de téléphone et l'opinion de ceux qui ont plusieurs lignes compte davantage !

EXEMPLE 2.3 (Heures et jours). Si l'on interroge des gens dans un supermarché, alors d'une part, les observations risquent d'être moins indépendantes (les clients qui se connaissent peuvent se passer le mot), et surtout, risquent d'être distribuées selon une distribution qui dépend fortement du jour et de l'heure. Aux Galeries Lafayette, le jeudi soir, c'est le soir des célibataires, munis de leur panier violet ; dans les petits supermarchés de ville, le lundi soir est celui des gens aisés partis en week-end à la campagne et qui font les courses à leur retour, ou des jeunes qui ont fait la fête tout samedi et tout dimanche ; dans les hypermarchés, le samedi, surtout avec la nouvelle semaine de quatre jours, c'est le moment des familles ; quant aux minutes qui suivent l'ouverture, c'est, chaque matin,

le défilé des retraités... Pour harmoniser tout cela, il faut venir souvent et n'interroger qu'un nombre de clients raisonnables (pendant quinze minutes toutes les quatre heures par exemple ?).

C'est, d'ailleurs, la même chose pour les sondages téléphoniques : si l'on appelait les gens la journée, on ne tomberait que sur les retraités et les chômeurs. C'est bien pour ça que les sondeurs visent les alentours de 20h...

EXEMPLE 2.4 (Biais de motivation). La figure 12 illustre un biais classique dans la constitution d'un échantillon, causé par le degré de motivation : on interroge les x premières personnes volontaires. Ce n'est pas grave dans une enquête de rue si elle est faite sans dédommagement et que l'on tente sa chance avec tout le monde ; ça l'est si l'on a une grande affiche promettant par exemple de pouvoir participer à une tombola si l'on prend le temps de répondre au questionnaire (on cible les membres de la population aimant les jeux de hasard) ou si les passants (personnes âgées, jeunes, familles avec enfants) ont des proportions de refus ou d'acceptation de participation à l'enquête différents (on risque de voir les jeunes filer et de sur-interroger les personnes âgées, qui ont davantage envie de parler). Ainsi, dans les enquêtes de satisfaction pour un produit donné (pour ma part, suite à l'achat de ma Fiat 500), les sondeurs relancent plusieurs fois les non-répondants ; en effet, ceux qui sont mécontents tiennent plus souvent à le dire que ceux qui sont contents du produit acheté, et ils profitent du questionnaire qui leur est adressé pour s'exprimer. L'objectif du sondeur est d'obtenir des taux de réponse égaux dans les deux sous-populations, parmi les contents et les mécontents.

EXEMPLE 2.5 (Biais de manipulation). La figure 11 présente une situation peut-être pas si fictive que cela ! Le sondeur pourrait en effet toujours manipuler les résultats en sélectionnant les sondés sur leur apparence ou sur leur appartenance à telle ou telle catégorie, tout en jurant par ailleurs avoir constitué un échantillon représentatif de la population générale...

FIGURE 11. Un dessin tiré du *Canard enchaîné* (le contexte étant la sélection de figurants lors des déplacements présidentiels souvent petits et encartés UMP).

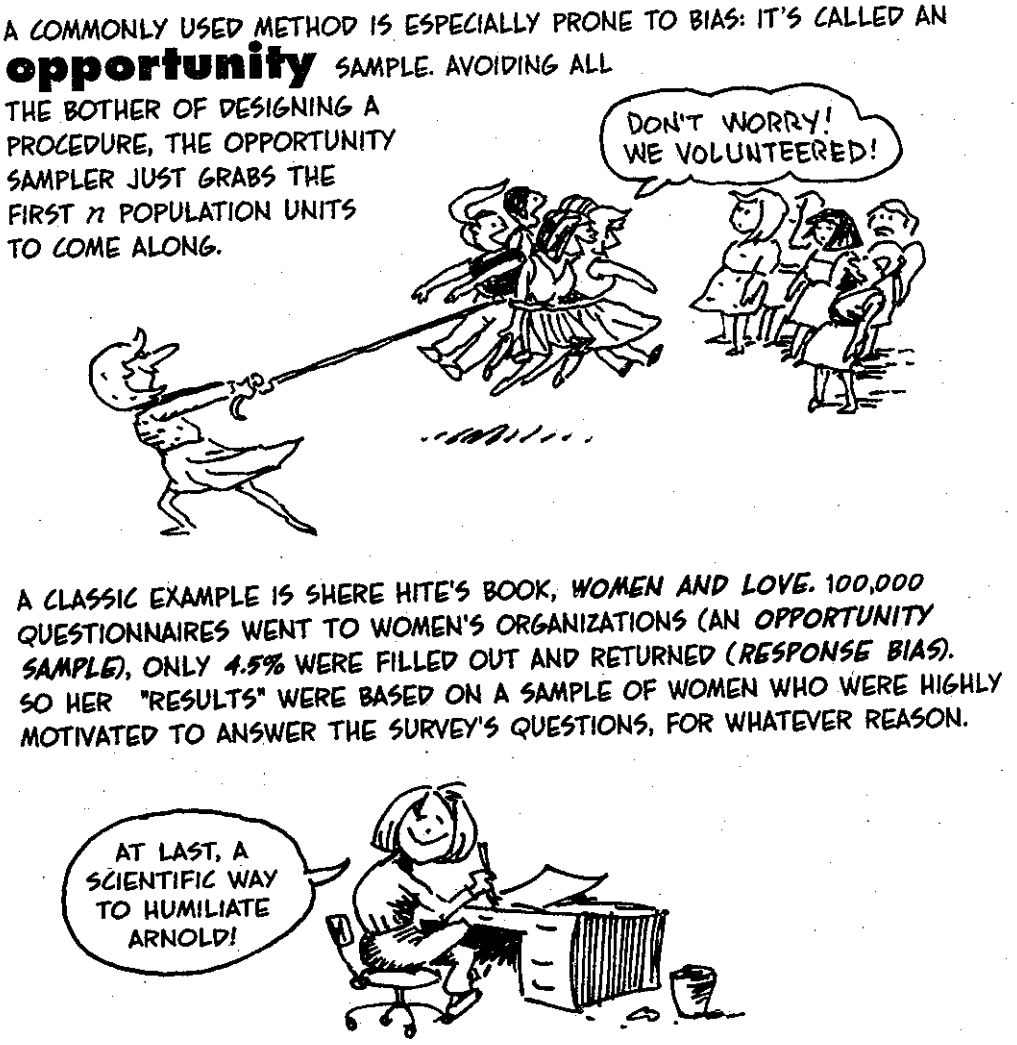


FIGURE 12. Le sondeur part à la quête des sondés, il n'attend pas que ces derniers viennent à lui : cela fausserait la représentativité de l'échantillon.

REMARQUE 2.4. Tous les détails sur le bon recueil des données (qui interroger, comment rédiger le questionnaire, etc.) vous seront donnés en cours de marketing. Dans notre cours, nous nous efforcerons de justifier le caractère indépendant et identiquement distribué des observations en expliquant dans chaque cas que le recueil des données a été bien fait, que tel ou tel écueil a été évité, et que l'on a bien touché uniformément au hasard la population d'intérêt.

REMARQUE 2.5 (Une autre méthode de sondage). La méthode de sondage par quotas, utilisée par exemple dans les enquêtes d'opinion, repose sur une constitution raisonnée de l'échantillon. En partant du fait que les variables qui vont être analysées dépendent d'autres caractères connus de la population (par exemple la catégorie socioprofessionnelle), on tâchera de respecter dans l'échantillon les mêmes proportions de chacune des catégories dans la population entière. Ensuite, on chargera chaque enquêteur d'interroger un nombre donné d'individus de chaque catégorie. Empiriquement, les résultats sont sans doute plus

satisfaisants que ceux de la méthode aléatoire, mais mathématiquement, l'analyse de la précision des sondages par quotas est délicate.

3. Modélisation mathématique : les lois classiques et le cas général

Ce qui précède nous a amené de données $x_1, \dots, x_n$ (où, en pratique, on connaît bien sûr la valeur de n) à une modélisation de ces données comme la réalisation d'un échantillon $X_1, \dots, X_n$ de variables aléatoires indépendantes et identiquement distribuées selon une certaine loi, inconnue.

On distingue deux cas :
- la forme de la loi commune est connue à un ou deux paramètres près, que l'on voudrait connaître eux aussi (nous donnerons des exemples ci-dessous) ;
- on n'a aucune idée de la forme de cette loi et on voudrait juste connaître son espérance (ou sa variance).

Notons que dans ce dernier cas, l'espérance de la loi est la moyenne d'un certain caractère sur toute la population : c'est par exemple la taille moyenne de tous les cornichons existant dans le monde, celle que voulait calculer le castor de la figure 8.

On introduit alors la notion de modèle statistique.
- Quand on peut indiquer la forme de la loi, les modèles établis ci-dessous seront de la forme : $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon telle loi, de paramètre $\theta_0 \in \Theta$. Le vrai paramètre θ_0 est inconnu, on sait simplement que c'est un élément de Θ. Les techniques des chapitres suivants permettront d'estimer θ_0 ; mais dans la description du modèle, on se doit de préciser l'ensemble des valeurs Θ que le paramètre pourrait prendre.
- Quand on n'a aucune idée sur la forme de la loi, on ne s'intéressera généralement qu'à l'espérance μ_0 de la loi commune, et, lorsqu'elle est définie, à sa variance σ_0^2. Ici encore, on indice par un zéro les quantités correspondant à la population (et donc inconnues). On écrira : $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une certaine loi, d'espérance μ_0 (et de variance σ_0^2).

3.1. La loi de Bernoulli. Elle modélise le comportement d'un quantité qui ne peut prendre que deux valeurs : homme ou femme, oui ou non, etc.

Loi de Bernoulli $\mathcal{B}(p)$	
Etendue	$\{0, 1\}$
Paramètre	$p \in \Theta = [0, 1]$
Nom	Fréquence p
Densité	$\mathbb{P}\{X = 1\} = p$ et $\mathbb{P}\{X = 0\} = 1 - p$
Espérance et variance	p et $p(1-p)$, respectivement

Exercice 2.1. On voudrait savoir si les étudiants en cours de scolarité ont lu le réquisitoire *J'ai fait HEC et je m'en excuse* de Florence Noiville. On conduit un sondage

par téléphone sur le campus d'HEC, en tirant au hasard 100 numéros dans la liste téléphonique des résidences. On obtient 94 réponses exploitables (6 mauvais coucheurs ayant raccroché le téléphone sans écouter la question parce qu'ils étaient pressés) : 51 sondés ont lu le livre et 43 ne l'ont pas lu. Ecrivez le modèle statistique correspondant.

CORRECTION 2.1. Voici la solution :

1. Population visée : les étudiants HEC en cours de scolarité et vivant sur le campus.
2. Table de correspondance : on note dans la suite 1 lorsque le livre a été lu, 0 lorsqu'il ne l'a pas été.
3. Données : $x_1, \dots, x_{94}$, appartenant à l'ensemble $\{0, 1\}$.
4. Description des données : on observe une moyenne sur l'échantillon de
$$\overline{x}_{94} = \frac{1}{94} \sum_{i=1}^{94} x_i = \frac{51}{94} \approx 54.3\,\% \ .$$
5. Modélisation : vu le tirage aléatoire, vu l'étendue $\{0, 1\}$, on peut considérer que les données sont issues de la réalisation de $X_1, \dots, X_{94}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre $p_0 \in [0, 1]$.
6. Interprétation : p_0 est la proportion des étudiants HEC en cours de scolarité et vivant sur le campus ayant lu le livre.

REMARQUE 2.6 (Attention !). La proportion p_0 est inconnue ! La quantité 54.3 % est une estimée de p_0, mais avec les moyens qui sont les nôtres pour l'instant, on ne sait pas encore quantifier combien cette estimée est potentiellement proche ou pas de la vraie valeur inconnue p_0.

Par ailleurs, p_0 n'est pas à interpréter comme la probabilité qu'un étudiant pris au hasard ait lu le livre ; une telle interprétation est probabiliste, or, ce qui nous intéresse ici, c'est le point de vue statistique : la connaissance d'une certaine fréquence moyenne sur la population.

3.2. La loi normale. Elle est la loi d'observations quantitatives qui résultent de la combinaison de nombreux effets ; c'est le théorème de la limite centrale qui explique pourquoi elle est si fréquente.

Loi normale $\mathcal{N}(\mu, \sigma^2)$	
Etendue	$\mathbb{R}$ (mais en pratique, $[\mu - 3\sigma, \mu + 3\sigma]$ avec probabilité $1 - 3\,‰$)
Paramètres	$(\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$
Noms	Espérance μ et variance σ^2 (écart-type σ)
Densité	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$ pour $x \in \mathbb{R}$

EXEMPLE 2.6 (Retour à l'exemple des salaires des infirmières). La figure 13 représente la répartition des données $h_1, \dots, h_{2911}$ de salaires horaires déclarés par les infirmières sondées. On a également tracé la densité de la loi normale (de paramètres estimés sur les données). On conclut que graphiquement, il ressort l'impression que la loi commune des

salaires $H_1, \ldots, H_{2911}$ est normale, de paramètres μ_0 et σ_0 inconnus (mais estimés par 20 et 4 environ). μ_0 représente le salaire moyen des infirmières sur l'ensemble des Etats-Unis.

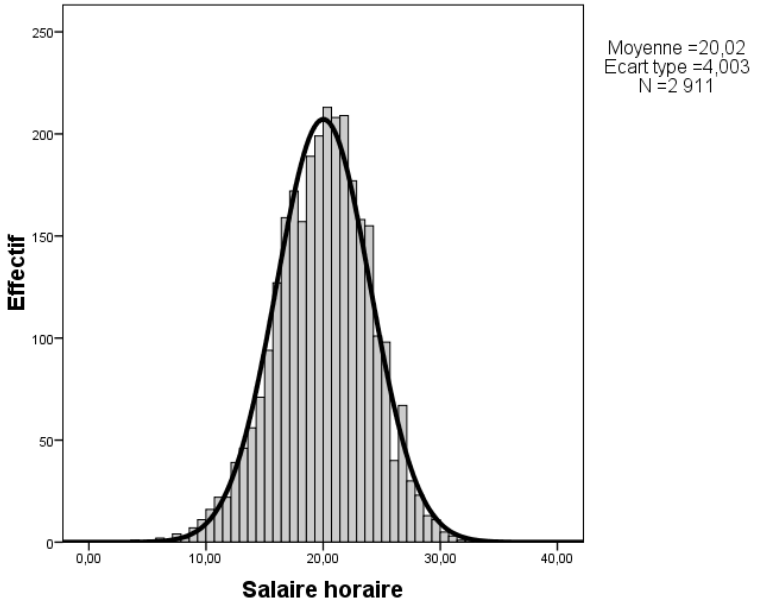


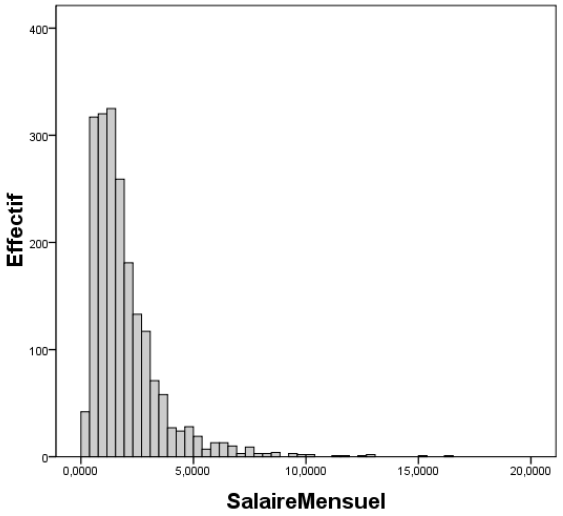
FIGURE 13. Histogramme des salaires horaires déclarés par l'échantillon d'infirmières sondées.

LA MINUTE SPPS 2.4. Pour obtenir l'histogramme de la figure 13, il suffit de cliquer sur `Graphes / Boîtes de dialogue [...] / Histogramme` (et de cocher la case demandant si l'on veut que la courbe de la densité gaussienne soit tracée).

REMARQUE 2.7 (Tests de normalité). L'ajustement de la répartition des valeurs observées à une loi normale nous a semblé tout à fait raisonnable : c'est un argument subjectif. On verra plus tard comment quantifier la qualité de cet ajustement, dans le chapitre sur les tests : par les tests de Kolmogorov-Smirnov (à adapter pour tenir compte d'une estimation préalable) et de Shapiro-Wilk. La mise en œuvre de ces deux tests formera un argument objectif.

Voici également une justification théorique que la loi des salaires à l'intérieur d'une profession soit normale : le salaire, à profession donnée, dépend de l'histoire personnelle de chacun, de sa formation initiale, des relations qu'il a ou n'a pas, de ses talents de négociateur lors de son embauche, etc. Or une telle somme de petits phénomènes aléatoires conduit à une loi normale : c'est ce que dit le théorème de la limite centrale.

On peut citer comme autres exemples courants d'occurrences de la loi normale :
- les erreurs de mesures physico-chimiques ;
- la pluviométrie annuelle, qui est la somme de 365 pluviométries journalières, et qui dépend des conditions climatiques générales, qui elles-mêmes sont aléatoires comme combinaisons de nombreux facteurs ;
- la taille d'une population, puisque la taille dépend des gènes, mais aussi de l'environnement, et notamment, de l'alimentation ;



Statistiques

SalaireMensuel

N	Valide	2000
	Manquante	0
	Moyenne	1,965100
	Médiane	1,533100
	Ecart-type	1,6062763
	Variance	2,580
	Minimum	,1484
	Maximum	16,3780
Centiles	80	2,749860
	90	3,660740
	95	4,977220

FIGURE 14. Statistiques descriptives des données simulées de salaires mensuels nets (en milliers d'euros).

– le rendement d'un champ, qui dépend de l'exposition de la parcelle, de la pluviométrie, de la qualité de la terre.

3.3. La loi log-normale. Par définition, Y suit une loi log-normale si on peut l'écrire comme e^X, où X suit une loi normale. Le tableau de paramètres, etc., de la loi log-normale se déduit donc de celui de la loi normale.

REMARQUE 2.8 (Test de log-normalité). Des tests de log-normalité se déduisent bien sûr des tests de normalité. Partant de données x_1, x_2, ..., il suffit de voir si les transformées $\ln x_1$, $\ln x_2$, ... peuvent être dites distribuées selon une loi normale (voir la remarque 2.7).

La loi log-normale apparaît dans les problèmes où il y a un facteur d'échelle, et notamment dans la distribution des salaires dans un échantillon inter-professionnel. (On prend donc des salariés de toutes professions, cette fois.) Les mieux payés sont vraiment (exponentiellement) mieux payés que les employés de base. Cela découle, là encore, du théorème de la limite centrale : on négocie souvent les augmentations de salaire en facteurs multiplicatifs, de la forme $1 + a \approx e^a$, où a est de l'ordre de quelques pourcents en cas de conservation de poste et est plus grand lorsqu'il s'agit d'une promotion ; et à chaque année t correspond une telle augmentation a_t, aléatoire et assez indépendante des précédentes augmentations. Des versions généralisées du théorème de la limite centrale peuvent alors être appliquées.

LA MINUTE SPPS 2.5. Chargez le fichier `Salaires.sav` (disponible sur le site web du cours) et étudiez-le sous SPSS avec `Analyse / Statistiques descriptives / Effectifs`. Calculez la moyenne et la médiane, le minimum et le maximum, ainsi que quelques quantiles[3] (les déciles à 80 % et 90 %, puis le quantile à 95 % par exemple), faites afficher un histogramme des données. Vous devez obtenir quelque chose de similaire à la figure 14.

3. On rappelle que le quantile à 90 % est un nombre $q_{90\,\%}$ tel que 90 % des données soient inférieures à $q_{90\,\%}$ et 10 % soient supérieures à $q_{90\,\%}$; en particulier, on appelle médiane le quantile à 50 %.

REMARQUE 2.9 (Médiane). La médiane est le nombre tel que la moitié des observations lui soit supérieure et l'autre moitié lui soit inférieure (quantile à 50 %.) En pratique, il suffit de classer les observations, et lorsque l'on en a un nombre pair, de prendre la moyenne des deux observations les plus centrales, et en cas de nombre impair, prendre simplement l'observation centrale. L'avantage de la médiance par rapport à la moyenne, c'est qu'elle est moins sensible aux observations extrêmes, elle décrit de manière plus robuste une certaine tendance centrale. Lorsque l'on parle du pouvoir d'achat, il vaut donc mieux raisonner en termes de médiane que de moyenne, messieurs les hommes politiques !

Ce fichier de données (simulées) est assez typique des salaires français ; on remarque notamment l'écart entre la moyenne et la médiane : la moyenne des salaires mensuels nets est aux alentours de 1 900 euros, mais la médiane se situe vers 1 500 euros. Cela est causé par les quelques salaires élevés qui tirent la moyenne vers le haut, tandis que la médiane est tirée vers le bas par les salaires minimum garantis légalement. En revanche, les déciles

❸ **Distribution des salaires annuels nets de tous prélèvements**

en euros courants

Déciles	Ensemble		Hommes		Femmes	
	2003	2004	2003	2004	2003	2004
D1	11 744	12 055	12 218	12 511	11 114	11 430
D2	13 158	13 466	13 739	14 018	12 335	12 680
D3	14 464	14 753	15 124	15 409	13 405	13 745
D4	15 875	16 166	16 598	16 892	14 571	14 893
Médiane	**17 497**	**17 802**	**18 322**	**18 622**	**16 002**	**16 310**
D6	19 494	19 813	20 463	20 805	17 748	18 073
D7	22 128	22 498	23 460	23 850	19 951	20 299
D8	26 340	26 788	28 286	28 769	23 005	23 425
D9	34 841	35 513	38 119	38 832	28 877	29 436
D9/D1	3,0	2,9	3,1	3,1	2,6	2,6

Champ : salariés à temps complet du secteur privé et semi-public.
Lecture : en 2004, 10 % des salariés à temps complet du secteur privé et semi-public gagnent un salaire annuel net inférieur à 12 055 euros, 20 % un salaire inférieur à 13 466 euros.
Source : DADS, Insee (fichier 2004 semi-définitif).

FIGURE 15. Données de salaires annuels présentées par l'INSEE (la source est la DADS : déclaration annuelle des données sociales, celle effectuée par les entreprises aux organismes paritaires).

(à 80 % et à 90 %) sont trop élevés sur ces données simulées. Ainsi, d'après l'INSEE (voir le tableau de la figure 15), les 10 % de salariés les mieux payés gagnaient en 2004 plus de 35 513 euros nets annuels, soit environ 2 960 euros nets par mois. Le décile à 90 % est donc autour de 3 430 euros nets actuels environ, si l'on tient compte d'une évolution moyenne de ces hauts salaires de 2.5 % par an depuis ce moment, pour compenser l'inflation (c'est sans doute plus : les écarts ont tendance à se creuser entre les très bas et les très hauts salaires). Pour information, le décile à 80 % en 2004 (le seuil tel que seuls 20 % des salariés gagnent plus) pourra vous surprendre : il est à 26 788 euros nets annuels, soit environ 2 230 euros nets mensuels ("seulement ?", pourrez-vous penser, ou pas).

Pour mieux refléter le ressenti de la majorité des interrogés, on introduit les boîtes à moustaches. Leur principe, ainsi que leur application aux données considérées dans ce paragraphe, est expliqué à la figure 16. On retiendra notamment que le ressenti des 50 % les plus centraux de la population est décrit par la boîte centrale.

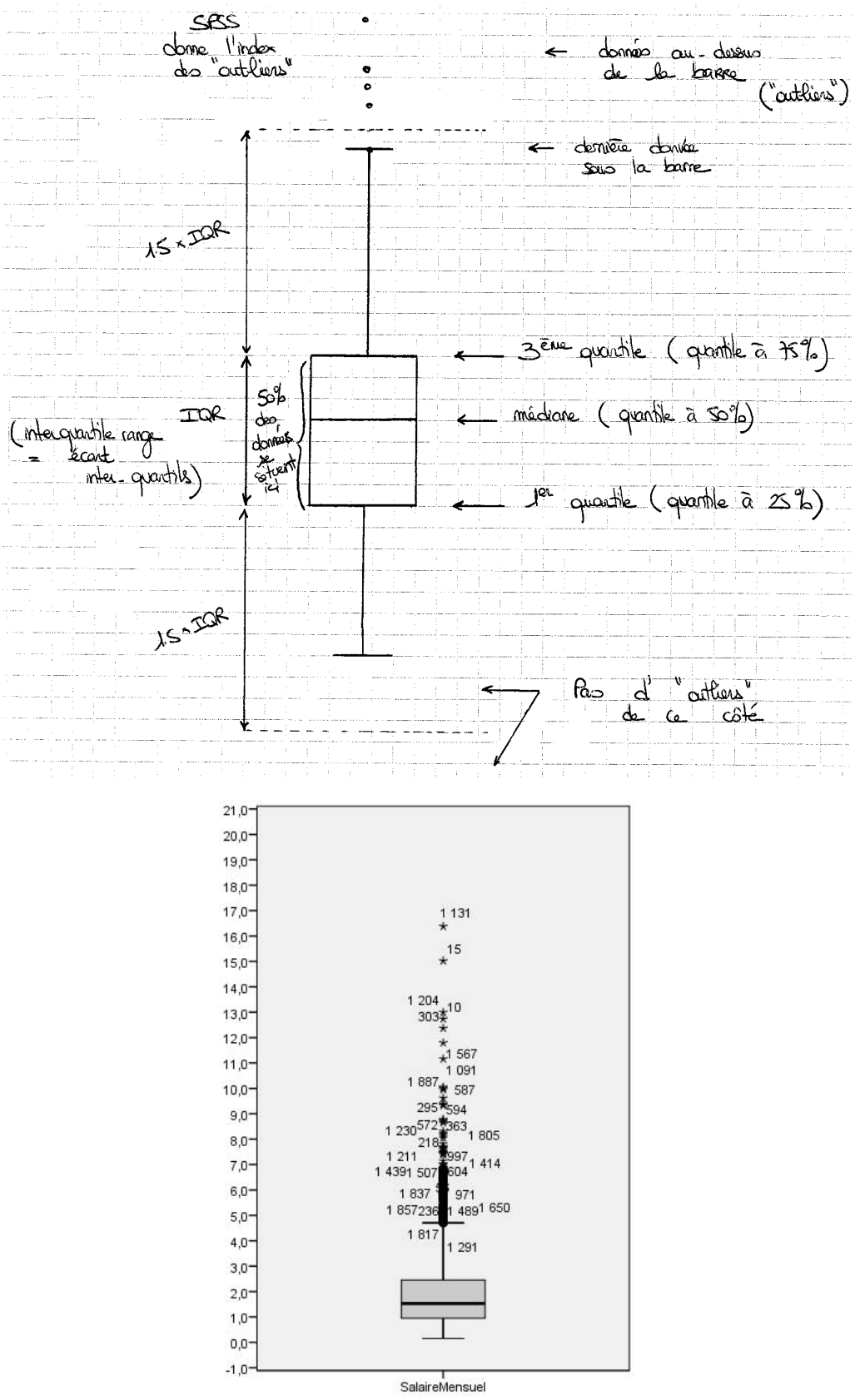


FIGURE 16. Boîtes à moustaches : principe et application.

Compléments pour étudiants avancés

4. Deux autres lois parfois utiles en modélisation

La loi de Poisson. Elle modélise les succès issus d'un grand nombre d'essais individuels, comme, par exemple, le nombre de contrats (fenêtres, opérations de défiscalisation) vendus par un télévendeur ou le nombre de réponses à un spam. Dans les deux cas, les différentes réponses aux requêtes étant indépendantes les unes des autres, et en supposant un tirage au hasard des n personnes sollicitées, le nombre de succès est *a priori* donné par une loi binomiale $\mathcal{B}in(n, p)$ de paramètres n et p, où p est le taux moyen de réponses qu'obtient un vendeur. Or, on sait qu'on a la convergence en loi

$$\mathcal{B}in(n, p_n) \rightharpoonup \mathcal{P}(\lambda) \qquad \text{lorsque } \; np_n \to \lambda \; ,$$

ce qui justifie que l'on puisse approximer la loi de ce nombre de succès par une loi de Poisson. Cela forme l'approximation dite binômiale-Poisson, ou « loi des petits nombres. » (En pratique, on le fait quand $n \geqslant 30$ et $1 \leqslant np_n \leqslant 10$.)

	Loi de Poisson $\mathcal{P}(\lambda)$
Etendue	$\mathbb{N}$
Paramètres	$\lambda \in \Theta = \mathbb{R}_+^*$
Noms	Espérance λ (et variance $\sigma^2 = \lambda$)
Densité	$\mathbb{P}\{X = k\} = e^{-\lambda}\lambda^k/k!$ pour $k \in \mathbb{N}$

Exemple 2.7 (Comparaison de performances de vendeurs). Un service d'évaluation des ressources humaines veut comparer deux télévendeurs d'opérations de défiscalisation (ceux qui vous demandent, juste après vous avoir salué, si vous payez plus ou moins de 3 000 euros d'impôts). Sur un mois, on note chaque jour leurs performances, et on suppose qu'on sait que les performances sont indépendantes des jours (pas meilleures le vendredi que le lundi par exemple). On a des valeurs observées $x_1, \ldots, x_{21}$ pour le premier et $y_1, \ldots, y_{19}$ pour le second (qui a pris deux jours de congés). On peut modéliser les premiers résultats comme étant la réalisation de $X_1, \ldots, X_{21}$, qui sont 21 variables aléatoires indépendantes et identiquement distribuées selon une certaine loi de Poisson, de paramètre λ_x ; et faire de même pour la seconde série, pour parvenir à $Y_1, \ldots, Y_{19}$ et au paramètre λ_y. Le but du traitement statistique sera alors de voir si les paramètres inconnus sous-jacents λ_x et λ_y sont différents ou non. Si c'est le cas, on parle de paramètres statistiquement différents, et cela prouvera qu'un vendeur est meilleur que l'autre.

Remarque 2.10. Des raisons similaires expliquent pourquoi chacun de nous reçoit un nombre poissonien de lettres, d'appels téléphoniques ou de mails chaque jour : nous avons

un grand nombre de correspondants potentiels, qui ont chacun une probabilité faible de nous contacter un jour donné.

REMARQUE 2.11 (Test d'ajustement à une loi de Poisson). Ici encore, on peut tester l'ajustement à une loi de Poisson, en recourant au test du χ^2, qui sera étudié dans la partie 10. En pratique, on soupçonne avoir affaire à une loi de Poisson lorsque la moyenne des valeurs observées est proche de leur variance.

La loi exponentielle. Elle est beaucoup utilisée dans les études médicales ou dans les études de fiabilité, pour modéliser les durées de survie (à une affection grave, comme un cancer) ou celles avant la prochaine panne (prochain pneu crevé sur une voiture, prochaine défaillance d'une machine sur une chaîne industrielle). Son intérêt réside dans le fait qu'elle est dite sans mémoire : si X suit une loi exponentielle, alors pour tous temps t_1, $t_2 > 0$,

$$\mathbb{P}\big\{X > t_1 + t_2 \;\big|\; X > t_1\big\} = \mathbb{P}\{X > t_2\}\,.$$

Le taux de panne est constant : la survenue ou non d'une panne (ou d'un décès) à l'instant présent ne dépend pas du nombre de pannes passées (ou du temps de survie actuel). Cette modélisation n'est donc pas valable durant les phases de rodage (mise en place d'un nouveau protocole thérapeutique) ou d'usure (une voiture tend, après un certain nombre d'années et de kilomètres, à voir ses différents éléments mécaniques lâcher les uns après les autres).

Des observations suivant la loi exponentielle peuvent faire penser à la loi des séries, à cause de ce phénomène d'absence de mémoire. On peut pendant un temps long n'avoir aucune panne puis subitement, deux ou trois pannes coup sur coup.

Loi exponentielle $\mathcal{E}(\lambda)$	
Etendue	$\mathbb{R}_+$
Paramètres	$\lambda \in \Theta = \mathbb{R}_+^*$
Noms	Espérance $1/\lambda$ (et variance $2/\lambda^2$)
Densité	$f(x) = \lambda e^{-\lambda x}$ pour $x > 0$ et $f(x) = 0$ pour $x \leqslant 0$

REMARQUE 2.12 (Test d'ajustement à une loi exponentielle). Ici encore, on peut tester l'ajustement de données à une loi exponentielle, en recourant à une version du test de Kolmogorov-Smirnov avec estimation préalable (voir les compléments de la partie 9).

Il existe d'autres lois usuelles... On pourrait, pour chacune des lois usuelles, binomiale, uniforme, géométrique, etc., décrire le contexte dans lequel elle apparaît naturellement, en donner un exemple concret, et rappeler sa forme... mais de fait, nous croiserons surtout des lois de Bernoulli ou des lois normales, lorsque la forme est connue, et à défaut, nous nous contenterons d'estimer la moyenne de population μ_0.

Exercices

La version rédigée du cours contient l'exercice 2.1, qui est suivi d'un corrigé présentant un modèle de rédaction, que l'on rappelle ci-dessous.

La démarche de modélisation

1. Définir la population effectivement visée.
2. Rappeler la table de correspondance (comment on code les données qualitatives proposées par l'énoncé, lorsqu'il y en a) ;
3. Préciser les données (leur donner un nom, rappeler leur nombre, définir leur étendue, i.e., l'ensemble des valeurs pouvant être prises).
4. Décrire les données (rappeler les statistiques d'échantillon disponibles).
5. Modéliser les données, au vu de la méthode de recueil utilisée : souvent, sous la forme de réalisations de variables aléatoires indépendantes et identiquement distribuées, selon une loi dont on précisera si on en connaît la forme (normale, Bernoulli) ou pas.
6. Préciser les paramètres d'intérêt de cette loi et surtout, les interpréter en termes de comportement de la population (en une phrase, donner le sens et le but de la connaissance de ces paramètres). Ces paramètres sont inconnus ; il s'agit souvent de la moyenne de certaines quantités sur l'ensemble de la population.

EXERCICE 2.2. Traitez la question 1 de l'exercice II de l'examen principal de 2008 (l'exercice sur l'assurance dédiée aux étudiants).

EXERCICE 2.3. Effectuez les modélisations correspondant aux données présentées dans l'examen de rattrapage de 2008 (tant pour le montant des achats pour Noël 2008 que lors de l'étude en 2007 et 2008 de la concurrence liée à Internet).

EXERCICE 2.4. Traitez la question 1 de l'exercice I de l'examen principal de 2007 (l'exercice sur la société de vente par correspondance).

EXERCICE 2.5. Modélisez la situation et les données introduites à l'exercice II de l'examen de rattrapage de 2007 (l'exercice sur l'optimisme des Français et le montant de leurs dépenses de loisirs).

Exercice 1 (cf. Exo II de l'examen principal 2008, question 1)

- Population ciblée : les étudiants assurés (déjà assurés)
- A chacun d'eux on demande si oui (1) ou non (0) il a eu un accident responsable au cours de l'année, ainsi que le montant à charge pour son assurance en cas de réponse positive
- Données :

  Existence d'un accident responsable : $x_1, \dots x_{1472}$ $(\in \{0,1\})$

  Montant des frais à charge : $y_1, \dots y_{256}$ $(\in \mathbb{R}_+)$ ↑ attention, on n'a que 256 montants !

- Statistiques d'échantillon :

  d'une part, $\bar{x}_{1472} = \frac{256}{1472} \cong 17.4\%$

  d'autre part, $\bar{y}_{256} = 1865\,€$ et $s_{y,256} = 524\,€$ (écart-type de $y_1, \dots y_{256}$)

- Modélisation : sondage aléatoire par téléphone, donc :

  $x_1, \dots x_{1472}$ réalisations de $X_1, \dots X_{1472}$ iid $\sim$ Ber (p_0)

  $y_1, \dots y_{256}$ réalisations de $Y_1, \dots Y_{256}$ iid selon une certaine loi d'espérance μ_0 et d'écart-type σ_0

- Interprétation : → p_0 est la vraie fréquence, sur l'ensemble des étudiants, de ceux qui ont eu un accident responsable l'année écoulée (paramètre inconnu, à moins d'interroger un à un les millions d'étudiants français)

  → μ_0 est le montant moyen que ces derniers ont coûté à leur assurance (et σ_0, l'écart-type de ces frais de réparation) [l'ensemble des étudiants ayant eu un accident responsable]

Exercice 2 (cf. examen de rattrapage 2008)

* Etude 1 : Montant des achats pour Noël 2008

- Population visée : les clients de week-end de Vélizy 2
- Table de correspondance inutile, on ne pose que des questions à réponses quantitatives (revenus du foyer rapportés au nombre d'adultes, budget cadeaux)
- Données $(x_1, y_1), \dots (x_{172}, y_{172}) \in (\mathbb{R}^+)^2$

où $\begin{cases} x_i = \text{revenus} \\ y_i = \text{budget cadeaux} \end{cases}$ du foyer du i-ème sondé

- $\bar{x}_{172} \simeq 1386 €$, $\bar{y}_{172} \simeq 376 €$

médiane des $x_1, \dots x_{172}$: $m_{x,172} = 1180 €$, des $y_1 \dots y_{172}$: $m_{y,172} = 300 €$

écart-type des $x_1 \dots x_{172}$: $s_{x,172} \simeq 695 €$, des $y_1 \dots y_{172}$: $s_{y,172} \simeq 260 €$

- Vu le sondage aléatoire : les données sont la réalisation de $(X_1, Y_1), \dots (X_{172}, Y_{172})$ iid selon une certaine loi sur $(\mathbb{R}^+)^2$

Attention ! Cette loi n'est pas une loi-produit : les revenus X influencent le budget cadeaux Y.

- Paramètres d'intérêt : p.ex., espérances μ_X et μ_Y de la première et de la seconde marginale de cette loi.

μ_X est le revenu moyen de l'ensemble des adultes fréquentant Vélizy 2, μ_Y leur budget cadeaux moyen.

Bien sûr, μ_X et μ_Y sont <u>inconnus</u> (il y a trop de clients pour les interroger tous). Ils le resteront, mais on verra comment les estimer.

Etude 2 : Etude de la concurrence d'Internet.

- Population visée : toujours la même, le clients (de week-end) de Velizy 2
- Table de correspondance : $\begin{cases} 1 & \text{si complément de courses sur Internet} \\ 0 & \text{sinon} \end{cases}$
- Données 2007 : $x_1, \dots x_{193} \quad \in \{0,1\}$
  Données 2008 : $y_1, \dots y_{172} \quad \in \{0,1\}$
- Statistiques : $\bar{x}_{193} = \frac{35}{193} \simeq 18.1\%$

  $\bar{y}_{172} = \frac{41}{172} \simeq 23.8\%$
- Modélisation : vu qu'il s'agit de deux sondages aléatoires établis dans le temps : $x_1, \dots x_{193}$ et $y_1, \dots y_{172}$ sont la réalisation de $(X_1, \dots X_{193})$ et $(Y_1, \dots Y_{172})$ où les X_j sont iid $\sim Ber(p_x)$, les Y_k sont iid $\sim Ber(p_y)$ et où les X_j sont indépendants des Y_k.
- Paramètres d'intérêt : p_x et p_y, qui désignent respectivement les proportions de l'ensemble des clients de Velizy 2 qui en 2007 et 2008 ont fait ou allaient faire un complément de courses de Noël sur Internet.

  Note : p_x et p_y sont inconnues mais on va les estimer.

Exercice 3 (cf. Exercice I de l'examen principal 2007, question 1)

- Population : les 50 000 clients du fichier clients
- Table de correspondance : on notera 1 en cas de commande et 0 sinon ; le montant de la commande est une donnée quantitative et n'a pas besoin d'être codé.
- Données : $x_1, \dots x_{1000} \in \{0,1\}$ représentant le fait que les clients ont acheté ou non suite à la nouvelle promotion

  et $y_1, \dots y_{170} \in \mathbb{R}_+$ le montant des commandes (quand il y a eu commande) (avant remise)
- Statistiques descriptives :

$$\bar{x}_{1000} = \frac{170}{1000} = 17\%, \qquad \bar{y}_{170} = 73 €, \qquad s_{y,170} = 8 €$$

- Modélisation : Vu l'envoi au hasard de la promotion :

  $X_1, \dots X_{1000}$ iid $\sim$ Ber(p_0)

  $Y_1, \dots Y_{170}$ iid selon une certaine loi sur $\mathbb{R}_+$
- Paramètres : p_0, le nouveau taux de commande si on généralisait la promotion à l'ensemble du fichier clients (on voudra le comparer à 13%)

  μ_0, l'espérance de la loi commune des Y_j, serait le nouveau montant moyen des commandes si on généralisait la promotion (sera à comparer à l'ancien montant, en tenant compte de l'évolution de la marge) et σ_0 l'écart-type correspondant

p_0, μ_0 et σ sont inconnus mais on va les estimer. (Ils seront connus une fois la promotion généralisée.)

Exercice 4 (cf. Exercice II de l'examen de rattrapage 2007)

- Population : l'ensemble des Français (ayant une ligne de téléphone fixe)
- Table de correspondance :
  - * Montant mensuel par foyer des achats hors produits de nécessité : donnée quantitative
  - * Age : 20-40 → 1 ; 40-60 → 2 ; 60 et + → 3
  - * Optimisme : Optimiste → 1 ; Pas optimiste → 2 ; Sans opinion → 3
- Données : $(x_1, a_1, o_1), \dots (x_{1837}, a_{1837}, o_{1837})$

où $\left\{ \begin{array}{l} x_j = \text{montant mensuel ....} \\ a_j = \text{âge} \\ o_j = \text{optimisme} \end{array} \right\}$ du j-ème sondé

et $x_j \in \mathbb{R}^+$, $a_j \in \{1,2,3\}$ et $o_j \in \{1,2,3\}$

- Statistiques : $\bar{x}_{1837} = 598$ € avec $s_{x,1837} = 254$ €

  Pour les âge et l'optimisme, voir le tableau de l'énoncé

- Vu le sondage aléatoire par téléphone, les (x_j, a_j, o_j) sont la réalisation de $(X_1, A_1, O_1), \dots (X_{1837}, A_{1837}, O_{1837})$ iid selon une certaine loi sur $\mathbb{R}^+ \times \{1,2,3\}^2$

  Attention ! Cette loi n'est sans doute pas une loi-produit (X_1, A_1 et O_1 sont liées et non indépendants p.ex.).

- Paramètre : p.ex. μ_0, 1ère marginale de la loi commune, et le montant moyen pour l'ensemble de Français (inconnu).

Troisième Partie

Interlude : deux quizz sur la modélisation

Premier énoncé (sujet posé en 2009)

Aucun document autorisé

Quizz 1 – Eléments de statistique mathématique – 2009

Question de cours

Enoncez le théorème de la limite centrale.

Utilisez votre esprit critique !

Que penser de la votation citoyenne sur le statut de la Poste ? Voici les résultats et commentaires proclamés par `http://www.appelpourlaposte.rezisti.org/`, le site officiel de la consultation :

> 2 123 717 personnes se sont rendues aux urnes pour donner leur avis dans une dizaine de milliers de points de vote. Le résultat du vote est sans appel. Plus de 90 % des électeurs disent "non" au changement de statut de la Poste et à l'ouverture de son capital (donc, à sa privatisation).

Déterminez, en quelques mots, la portée statistique de ces résultats (ce que l'on peut en conclure, ou pas).

Exercice de modélisation (issu d'une discussion avec un ancien élève)

Un ancien élève est venu me voir. Il avait fondé son entreprise de cours d'anglais à distance : chaque semaine, un ensemble d'exercices, lectures et vidéos arrive par email aux abonnés. Ces derniers paient pour ce service un montant mensuel et sont libres d'interrompre le service à tout moment, sans délai de résiliation. Il voulait modéliser le comportement de ceux qui s'abonnent pour la première fois. Il avait donc envoyé une offre alléchante en mass-emailing, à 1 000 internautes tirés au hasard dans un long listing (plus de 25 000 noms) d'anciens clients du *Wall Street Institute* et avait suivi ces derniers pendant plusieurs mois. Voici les données qu'il avait obtenues : 140 clients avaient souscrit à l'offre, dont

- 25 s'étaient désinscrits au bout d'un mois,
- 42 l'avaient fait au bout de deux mois,
- 29, au bout de trois mois,
- 44 étaient encore abonnés au début du quatrième mois.

Modélisez ce problème, selon le schéma en six points vu en cours. (Ecrivez au dos de cette feuille.)

Premier corrigé (sujet posé en 2009)

Gilles Stoltz

Quizz 1 – Eléments de statistique mathématique – 2009

Question de cours

Enoncez le théorème de la limite centrale.

Soient $X_1, \dots X_n$ des variables aléatoires indépendantes et identiquement distribuées selon une loi admettant un moment d'ordre deux, d'espérance et de variance notées μ_0 et σ_0^2. Alors

$$\frac{\sqrt{n}}{\sigma_0}\left(\overline{X}_n - \mu_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

(où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$)

Utilisez votre esprit critique !

Que penser de la votation citoyenne sur le statut de la Poste ? Voici les résultats et commentaires proclamés par `http://www.appelpourlaposte.rezisti.org/`, le site officiel de la consultation :

> 2 123 717 personnes se sont rendues aux urnes pour donner leur avis dans une dizaine de milliers de points de vote. Le résultat du vote est sans appel. Plus de 90 % des électeurs disent "non" au changement de statut de la Poste et à l'ouverture de son capital (donc, à sa privatisation).

Déterminez, en quelques mots, la portée statistique de ces résultats (ce que l'on peut en conclure, ou pas).

1) Il y a un fort biais (de motivation) : l'échantillon n'est nullement représentatif de l'ensemble des Français.
2) La portée statistique est quasi-nulle : tous les gens motivés et concernés (= la population visée, en fait) ont voté et on a dépouillé : il n'y a plus d'incertitude, on connaît ici le comportement de la population visée.
3) Cependant, il y a une portée politique forte : sans doute aurait-on à trouver les 4,4 millions d'électeurs nécessaires pour réclamer un référendum d'initiative populaire.

Exercice de modélisation (issu d'une discussion avec un ancien élève)

Un ancien élève est venu me voir. Il avait fondé son entreprise de cours d'anglais à distance : chaque semaine, un ensemble d'exercices, lectures et vidéos arrive par email aux abonnés. Ces derniers paient pour ce service un montant mensuel et sont libres d'interrompre le service à tout moment, sans délai de résiliation. Il voulait modéliser le comportement de ceux qui s'abonnent pour la première fois. Il avait donc envoyé une offre alléchante en mass-emailing, à 1 000 internautes tirés au hasard dans un long listing (plus de 25 000 noms) d'anciens clients du *Wall Street Institute* et avait suivi ces derniers pendant plusieurs mois. Voici les données qu'il avait obtenues : 140 clients avaient souscrit à l'offre, dont

- 25 s'étaient désinscrits au bout d'un mois,
- 42 l'avaient fait au bout de deux mois,
- 29, au bout de trois mois,
- 44 étaient encore abonnés au début du quatrième mois.

Modélisez ce problème, selon le schéma en six points vu en cours. (Ecrivez au dos de cette feuille.)

1. Population visée : – les 25 000 membres du fichier clients de WSI
– au même, l'ensemble des internautes ayant un besoin de formation en anglais.

2. Table : – 1ère série de données : { 0 pour abonnement / 1 pour offre restée lettre morte

– 2ème série de données : 1, 2, 3, +
1, 2, 3 : nombre de mois payés si ≤ 3
+ : reste au moins 4 mois

variable qualitative ordinale obtenue par catégorisation de la variable donnant le temps effectivement passé.

3. Données : $x_1, \dots x_{1000} \in \{0,1\}$ pour les souscriptions
$y_1, \dots y_{140} \in \{1,2,3,+\}$ pour les temps d'abonnement

4. Statistiques : • $\bar{x}_{1000} = 14.0\,\%$

• pour les y_j, on fait un tableau de fréquences :

	1	2	3	+
140 données :	17.9%	30.0%	20.7%	31.4%
	↑	↑	↑	↑
	= 25/140	= 42/140	= 29/140	= 44/140

⚠ S'agissant d'une variable qualitative, la notion de moyenne n'a pas de sens !

5. On a tiré un échantillon de 1000 clients au hasard, ce qui conduit à la modélisation : $X_1 \dots X_{1000}$ iid $\sim$ Ber(p), $p \in [0,1]$

et $Y_1, \dots Y_{140}$ iid $\sim \nu$ où ν est une loi sur $\{1,2,3,+\}$: $\nu = (\nu_1, \nu_2, \nu_3, \nu_+)$
avec $\nu_j \geq 0$, $\nu_1+\nu_2+\nu_3+\nu_4 = 1$

6. p = taux de réponses positives à l'offre que l'on enregistrerait si on lançait la campagne sur l'ensemble du fichier clients

ν = loi d'évolution de la durée d'abonnement que l'on enregistrerait alors.

RÉSULTATS 96 étudiants inscrits au cours

BRAVO !

A	B	C	D	E	autres (absents, excusés...)
22	39	21	8	0	6

Remarque : pour avoir A il n'y avait pas besoin de bien traiter $y_{95} \dots y_{140}$ et d'introduire $\bar{Y}_n$.

Question de cours :

- Le symbole pour la convergence en loi est $\xrightarrow{\mathcal{L}}$, et non $\xrightarrow{\mathcal{L}}$ ou $\xrightarrow{\mathcal{L}}$
- $X \sim \mathcal{U}(0,1)$ pour dire "X suit la loi $\mathcal{U}(0,1)$" ; ne pas utiliser $\hookrightarrow$
- La variable aléatoire en jeu est $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu_0)$

  et non pas $\frac{X_1 + \dots X_n - n\mu_0}{\sigma\sqrt{n}}$ (juste mais peu lisible et peu interprétable).

Utilisez votre esprit critique :

- Certains ont conclu : "Une grande majorité des votants ont voté contre..."
  ↳ Assertion sans aucune valeur ajoutée !
- Attention, la population ciblée est difficile à qualifier. On ne touche pas de manière représentative l'ensemble des Français, à cause du biais de motivation.
- Une fois dépouillés le bulletin il n'y a plus de statistiques inférentielles à faire, il suffit d'une statistique descriptive : le taux en faveur du non.

- Certains ont dit que le problème était que l'échantillon était trop petit : or il est grand, très grand ! C'est plutôt son manque de représentativité qui posait problème.

Exercice de modélisation :

- Il ne faut pas oublier l'interprétation des paramètres.
- Pour p_0 :
  - ce n'est pas la probabilité qu'une personne donnée s'abonne (c'est une interprétation mauvaise car probabiliste)
  - ce n'est pas la proportion de contacts à s'inscrire : on connaît cette dernière, c'est $\overline{x}_{1000}$
  - c'est la proportion des 25 000 membres du fichier clients (= la population) qui s'inscriraient si on leur proposait tous l'offre.
- Bien distinguer clairement les données x_j de leur modélisation par des variable aléatoires X_j
  Ce n'est pas la loi de ~~Bernouilli~~, c'est la loi de Bernoulli.
- J'ai souvent vu confondus la population (25 000 clients) et l'échantillon (1000 clients tirés au hasard).
- Les durées d'abonnement, vu les données proposées, étaient des variables <u>qualitatives</u> (ordinales). Pour elles, la notion de moyenne n'a pas de sens, on ne peut définir $\overline{y}_{140}$. C'est pour cela que j'ai défini une catégorie "+", avec elle, on n'est pas tenté de faire des moyennes.
- Si on avait eu des données plus précises, on aurait pu prendre $\mathbb{N}^*$ comme

étendue pour les $y_1, \dots y_{140}$ et la modéliser comme la réalisation de $Y_1, \dots Y_{140}$ iid selon une certaine loi sur $\mathbb{N}^*$ d'espérance notée μ_0 ; et interpréter μ_0 comme le nombre moyen de mois que les clients issus du fichier WEB resteront abonnés.

- Solution alternative élégante proposée par certains : au lieu d'avoir $x_1, \dots x_{1000}$ et $y_1, \dots y_{140}$ avoir : $z_1, \dots z_{1000}$

avec la table :

0	si non-adhésion
1	si 1 mois d'adhésion
2	2
3	3
+	si adhésion ≥ 4 mois.

Dans ce cas le paramètre aurait été une probabilité ψ sur $\{0, 1, 2, 3, +\}$: $\psi = (\psi_0, \psi_1, \psi_2, \psi_3, \psi_+)$.

Second énoncé (sujet posé en 2008)

Aucun document autorisé

Question de cours

Enoncez le théorème de la limite centrale.

Question subsidiaire et facultative (à ne traiter que s'il vous reste du temps) : Lorsque les observations sont gaussiennes, que pouvez-vous dire du résultat du théorème de la limite centrale ?

Un premier exercice de modélisation

(Une histoire inspirée par une vieille rumeur.) Il y a longtemps de cela, les étudiants d'HEC se sont vus proposer la construction sur le campus ou d'une église ou d'une piscine. Pour savoir s'ils devaient faire une campagne active de lobbying, les gros bras du bureau des sports se sont postés à la cantine et ont interrogé tous les étudiants qui passaient par là. Ils ont obtenu 135 réponses, dont 63 en faveur de la piscine.

Modélisez ce problème : précisez la population, indiquez les données, proposez une modélisation et indiquez le ou les paramètre(s) d'intérêt.

Un second exercice de modélisation

Vous êtes nutritionniste au Ministère de la Santé et vous intéressez au nombre moyen de fruits et légumes que consomment les Français chaque jour. Vous commandez une enquête, qui indique une consommation moyenne de 4.37 de fruits et légumes par jour, pour les 1 068 personnes ayant accepté (ou su) répondre.

Modélisez ce problème (mêmes commentaires que ci-dessus).

Second corrigé (sujet posé en 2008)

Quizz #1 – CORRIGÉ.

Question de cours : $X_1, \dots X_n$ iid selon une loi admettant une espérance μ et une variance σ^2. Alors

$$\frac{\sqrt{n}}{\sqrt{\sigma^2}}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Question subsidiaire : Lorsque la loi commune est gaussienne, $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ et donc $\frac{\sqrt{n}}{\sqrt{\sigma^2}}(\bar{X}_n - \mu) \sim \mathcal{N}(0,1)$
A tout rang fini n, la variable aléatoire suit déjà la loi $\mathcal{N}(0,1)$.

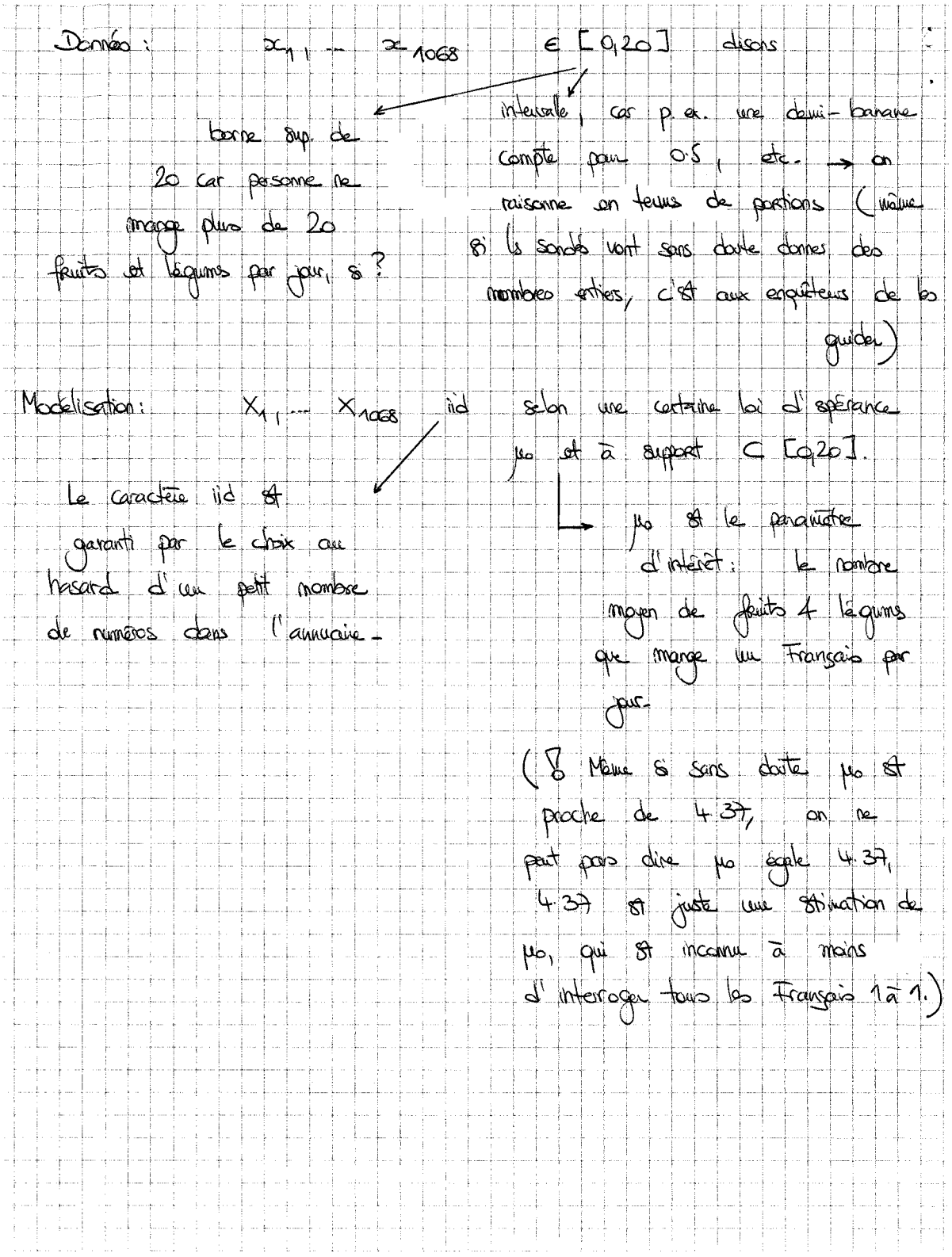
Eglise vs. Piscine : Population : les étudiants mangeant à la cantine ($\approx$ les étudiants du campus)

Données : $x_1, \dots x_{135} \in \{0,1\}$ (0 = Piscine, 1 = Eglise)
$\bar{x}_{135} = \frac{72}{135} = 53.3\%$

Modélisation : $X_1, \dots X_{135}$ iid $\sim Ber(p_0)$ où p_0 est le paramètre d'intérêt : la proportion d'étudiants mangeant à la cantine et en faveur de la piscine. p_0 est inconnu, mais sans doute proche de 53.3%.

Le caractère iid est assuré par le fait qu'on interroge peu d'étudiants et qu'on le fait au hasard.

Fruits & légumes : Population : l'ensemble des Français (en fait, l'ensemble des Français ayant un téléphone fixe)


Données : $x_1, \dots x_{1068} \in [0,20]$ disons
borne sup. de 20 car personne ne mange plus de 20 fruits et légumes par jour, si ?
intervalle, car p. ex. une demi-banane compte pour 0.5, etc. → on raisonne en terms de portions (même si les sondés vont sans doute donner des nombres entiers, c'est aux enquêteurs de les guider)
Modélisation : $X_1, \dots X_{1068}$ iid selon une certaine loi d'espérance μ_0 et à support $\subset [0,20]$.
Le caractère iid est garanti par le choix au hasard d'un petit nombre de numéros dans l'annuaire.
μ_0 est le paramètre d'intérêt : le nombre moyen de fruits + légumes que mange un Français par jour.
(! Même si sans doute μ_0 est proche de 4.37, on ne peut pas dire μ_0 égale 4.37, 4.37 est juste une estimation de μ_0, qui est inconnu à moins d'interroger tous les Français 1 à 1.)

Commentaires.

Question de cours :

- on note μ (et pas m ni E) l'espérance et σ l'écart-type ;
- j'ai encore vu des assertions

$$\frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \longrightarrow \mathcal{N}(0,1) ;$$

elles sont juste mais sont peu interprétables statistiquement ;

- notation pour la convergence en loi : $\longrightarrow$ et non pas $\rightarrow$;

$\sim$ signifie "suit la loi" ; je ne sais pas ce que vous entendez par "$\hookrightarrow$" ou "$\rightsquigarrow$" : convergence ou égalité des distributions ?

Exercices de modélisation.

Pour les deux :

- il faut justifier le caractère iid des X_i ;
- certains confondent population (= ensemble des personnes d'intérêt) et échantillon (= les sondés)
- Une fois qu'on a indiqué les paramètres, il faut les interpréter ; ainsi le p_0 du 2ème exercice n'est pas « la probabilité qu'un élève demande 1 piscine » : c'est vrai, certes, mais c'est une interprétation probabiliste ; l'interprétation statistique est « la proportion de la population du campus en faveur de la piscine ».
- p_0 est inconnu et donc a priori différent de 63/135 ; de même on ne peut dire que

p_0 vaut 4.37. N'ayez pas peur de dire que les paramètres sont inconnus ! L'objet de la statistique est justement de dire des choses sur eux.

Exercice 2: Beaucoup ont écrit $X_1 + \dots X_{135} \sim \text{Bin}(135, p_0)$ mais je préfère on reste à l'étape précédente:

$X_1, \dots X_{135}$ iid $\sim \mathcal{B}(p_0)$.

$X_1 + \dots X_{135}$ n'a pas de sens statistique, mais $\bar{X}_{135}$ on a un.

Exercice 3: – Si on écrit que l'étendue est $\mathbb{N}$, il ne faut pas proposer une modélisation gaussienne !

(– Note: j'ai accepté que vous disiez que l'étendue était $\mathbb{N}$.)

Quatrième Partie

Estimation ponctuelle et quantiles des lois usuelles

Version rédigée du cours

Résumé : Des parties précédentes, il faut essentiellement retenir que dans la plupart des cas étudiés dans ce cours, on dispose de valeurs observées que l'on modélise comme la réalisation de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi. On s'intéresse à certains paramètres de cette loi : par exemple, son espérance μ_0, son écart-type σ_0, etc. Ces paramètres sont inconnus, à moins de faire une étude exhaustive de la population, ce qui est souvent trop coûteux (en temps ou en argent).

Objectif : Dans cette partie, on explique comment estimer ces paramètres. On prépare par ailleurs le terrain des cours suivants en introduisant deux nouvelles lois, la loi de Student et celle du χ^2, on définit la notion de quantile, et on apprend à lire dans une table les quantiles des lois ainsi introduites, ainsi que ceux de la loi normale standard.

Attention : Cette partie est beaucoup plus théorique que toutes les autres de ce cours ; elle introduit les définitions mathématiques nécessaires pour la suite. Nous n'en verrons que de brefs éléments en cours, à charge pour vous de lire en détails ce qui suit (une fois n'est pas coutume).

1. Notions d'estimateur et d'estimée

On fixe dans ce qui suit un modèle $X_1, \ldots, X_n$ de variables aléatoires indépendantes et identiquement distribuées selon une loi $\mathbb{P}_{\theta_0}$ appartenant à l'ensemble $\left\{\mathbb{P}_\theta,\ \theta \in \Theta\right\}$. Il correspond aux données $x_1, \ldots, x_n$.

1.1. De la théorie...

Définition 4.1 (Estimateur). *Un estimateur est toute variable aléatoire construite uniquement à partir des observations $X_1, \ldots, X_n$. En particulier, il ne doit pas dépendre de quantités inconnues, telles que θ_0 ou $\mathbb{P}_{\theta_0}$.*

Remarque 4.1. Une convention utile est qu'on note les estimateurs par les quantités qu'ils estiment, surmontées de petits chapeaux $\widehat{\ }$. Ainsi, dans un modèle de Bernoulli, lorsque le modèle est l'ensemble des $\mathcal{B}(p)$, avec $p \in [0,1]$, on note par $\widehat{p}$ les estimateurs de p_0, et parfois même $\widehat{p}_n$ pour rappeler la taille n de l'échantillon. Dans tous les modèles, on notera $\widehat{\mu}$ (ou $\widehat{\mu}_n$) les estimateurs de l'espérance commune μ_0 des observations de l'échantillon, et $\widehat{\sigma}$ (ou $\widehat{\sigma}_n$), ceux de l'écart-type σ_0.

Exemple 4.1. Dans le modèle de Bernoulli, on pourrait proposer les quantités suivantes comme estimateurs du vrai paramètre de fréquence p_0 :

$$\widehat{p}_n = \overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)\ , \qquad \text{ou} \quad \widehat{p}_n = X_1\ , \qquad \text{ou} \quad \widehat{p}_n = 0.5\ .$$

On sent évidemment que le premier est le meilleur estimateur et que les deux autres sont très mauvais.

REMARQUE 4.2 (Estimateur vs. « bon » estimateur). S'il est facile de définir la notion d'estimateur, il est en revanche beaucoup plus difficile de dire ce qu'est un bon estimateur ! Il n'y a pas de notion universelle de bon estimateur, mais je vous proposerai ci-dessous la liste, non exhaustive, de quelques qualités qu'un estimateur peut posséder :

- le caractère sans biais ;
- la consistance ;
- la normalité asymptotique.

L'objet de la recherche en statistique est alors, entre autres, d'exhiber des estimateurs possédant ces qualités dans des modèles plus compliqués que ceux fondés sur des variables aléatoires indépendantes et identiquement distribuées (avec davantage de dépendance entre les observations ou en situations d'observations incomplètes, etc.). Comme toujours, l'utilisateur est éventuellement rassuré par ces garanties, mais en réalité, il se préoccupe surtout de la valeur que va prendre l'estimateur sur les données.

1.2. ... A la pratique !

DÉFINITION 4.2 (Estimée). *Une estimée est la réalisation d'un estimateur sur les données* $x_1, \ldots, x_n$. *Autrement dit, l'estimée est la valeur que l'on peut calculer en remplaçant les* X_j *par les* x_j *dans la définition de l'estimateur correspondant.*

1.3. Ce que l'on veut estimer. On peut vouloir estimer θ_0, comme on le décrit en préambule du chapitre, ou une fonction de $\mathbb{P}_{\theta_0}$, comme l'espérance μ_0 ou la variance σ_0. Par souci de généralité, nous notons dans la suite $g(\theta_0)$ cette quantité objet de l'étude. On parlera alors d'estimateurs de $g(\theta_0)$.

2. Première qualité éventuelle d'un estimateur : le caractère sans biais

DÉFINITION 4.3 (Estimateur sans biais). *Un estimateur* $\widehat{g}_n$ *de* $g(\theta_0)$ *est dit sans biais lorsque, quel que soit le vrai paramètre* θ_0,

$$\mathbb{E}[\widehat{g}_n] = g(\theta_0) \ ;$$

il est dit biaisé sinon.

Interprétation : Que l'espérance de l'estimateur soit égale à l'objectif de l'estimation $g(\theta_0)$ nous fait espérer que la plupart du temps, l'estimateur lui-même soit proche de $g(\theta_0)$. Cela découle, par exemple, de l'inégalité de Chebychev-Markov (proposition 1.1), qui, je vous le rappelle, contrôle les déviations de $\widehat{g}_n$ autour de son espérance : autant que ce dernier ait pour espérance $g(\theta_0)$, l'objectif à estimer !

EXEMPLE 4.2. La moyenne empirique $\widehat{\mu}_n = \overline{X}_n$ est un estimateur sans biais de l'espérance μ_0, de même que l'estimateur $\widehat{\mu}'_n = X_1$.

EXEMPLE 4.3. On suppose que la loi commune des X_j admet un moment d'ordre deux, que l'on note

$$m_2(\theta_0) = \mathbb{E}\left[X_1^2\right] \ .$$

L'estimateur

$$\widehat{m}_{2,n} = \frac{1}{n}\sum_{j=1}^{n} X_j^2$$

est un estimateur sans biais de $m_2(\theta_0)$.

Exercice 4.1. On suppose ici encore que la loi commune admet un moment d'ordre deux et on définit de manière naturelle un estimateur de sa variance

$$\sigma_0^2 = m_2(\theta_0) - \mu_0^2 \qquad \text{par} \qquad \widehat{m}_{2,n} - \left(\overline{X}_n\right)^2 .$$

1. Montrer que c'est un estimateur biaisé de σ_0^2 et en déduire un estimateur non biaisé, noté $\widehat{\sigma}_n^2$.
2. Montrer que l'on peut écrire ce dernier sous la forme (que l'on retiendra pour la suite)

$$\widehat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}_n\right)^2 .$$

Correction 4.1. On calcule, en développant le carré de la somme,

$$\begin{aligned}\mathbb{E}\left[\left(\overline{X}_n\right)^2\right] = \frac{1}{n^2}\,\mathbb{E}\left[\left(X_1 + \ldots + X_n\right)^2\right] &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + 2 \sum_{1 \leqslant i < j \leqslant n} \mathbb{E}\left[X_i X_j\right]\right) \\ &= \frac{1}{n^2}\left(n\, m_2(\theta_0) + 2\,\frac{n(n-1)}{2}\,\mu_0^2\right) = \frac{1}{n}\, m_2(\theta_0) + \frac{n-1}{n}\,\mu_0^2 ,\end{aligned}$$

où l'on a utilisé l'indépendance et l'identique distribution des X_t à la deuxième égalité. On a alors

$$\mathbb{E}\left[\widehat{m}_{2,n} - \left(\overline{X}_n\right)^2\right] = m_2(\theta_0) - \left(\frac{1}{n}\, m_2(\theta_0) + \frac{n-1}{n}\,\mu_0^2\right) = \frac{n-1}{n}\left(m_2(\theta_0) - \mu_0^2\right) = \frac{n-1}{n}\,\sigma_0^2 .$$

On en conclut que l'estimateur

$$\widehat{\sigma}_n^2 = \frac{n}{n-1}\left(\widehat{m}_{2,n} - \left(\overline{X}_n\right)^2\right)$$

est un estimateur sans biais de la variance σ_0^2. On montre maintenant qu'il peut également s'écrire sous l'autre forme proposée. Pour cela, il suffit de prouver que

$$n\left(\widehat{m}_{2,n} - \left(\overline{X}_n\right)^2\right) = \sum_{j=1}^{n} \left(X_j - \overline{X}_n\right)^2 .$$

Or, en développant,

$$\begin{aligned}\sum_{j=1}^{n} \left(X_j - \overline{X}_n\right)^2 = \sum_{j=1}^{n} \left(X_j^2 - 2X_j\,\overline{X}_n + \left(\overline{X}_n\right)^2\right) &= \left(\sum_{j=1}^{n} X_j^2\right) - 2\left(\sum_{j=1}^{n} X_j\right)\overline{X}_n + n\left(\overline{X}_n\right)^2 \\ &= n\,\widehat{m}_{2,n} - 2n\,\overline{X}_n\,\overline{X}_n + n\left(\overline{X}_n\right)^2 = n\left(\widehat{m}_{2,n} - \left(\overline{X}_n\right)^2\right) ,\end{aligned}$$

ce qui conclut l'exercice.

La minute SPPS 4.1. Lorsque l'on calcule une variance (par exemple, par `Analyse / Statistiques descriptives / Descriptives`), c'est bien la formule de variance débiaisée qui est utilisée. Dit autrement, la valeur que l'on lit dans le tableau produit par SPSS est une estimée de la variance sur les valeurs observées, calculée à partir de l'estimateur sans biais $\widehat{\sigma}_n^2$ introduit plus haut. Je m'en suis convaincu par l'expérience reportée à la figure 17.

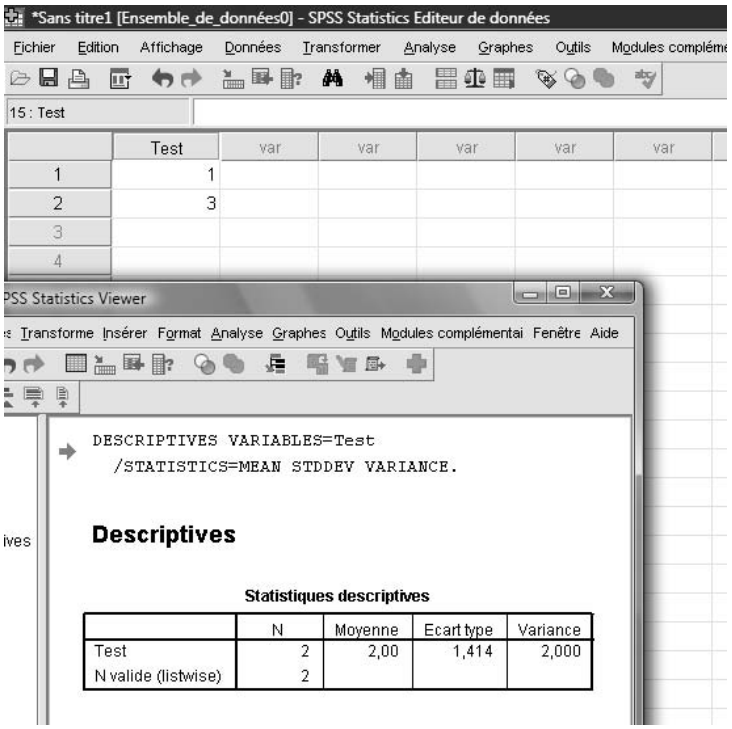


FIGURE 17. Calcul sous SPSS d'une estimée de la variance, sur les deux données $x_1 = 1$ et $x_2 = 3$.

REMARQUE 4.3. Comment estimer l'écart-type σ_0 ? Il semble raisonnable de considérer $\sqrt{\widehat{\sigma}_n^2}$. Malheureusement, cet estimateur est en général biaisé : il est difficile de relier

$$\sigma_0 = \sqrt{\mathbb{E}\left[\widehat{\sigma}_n^2\right]} \qquad \text{et} \qquad \mathbb{E}\left[\sqrt{\widehat{\sigma}_n^2}\right] \ .$$

En fait, on peut prouver en toute généralité que la première quantité est plus grande que la seconde ; mais l'écart entre les deux dépend fortement du modèle, de sorte qu'il n'existe pas cette fois-ci de manière universelle de débiaiser.

REMARQUE 4.4. La figure 18 compare les performances des versions biaisée et débiaisée de l'estimateur de la variance dans un modèle gaussien. On remarque que le débiaisement est utile surtout lorsque la taille d'échantillon n est petite et que sa mise en œuvre est moins cruciale lorsque n est grande.

Pour des tailles d'échantillon n plus grandes, on préférera s'intéresser à la consistance des estimateurs, que l'on définit maintenant.

3. Deuxième qualité éventuelle d'un estimateur : la consistance

Rigoureusement parlant, la consistance ne peut être la propriété que d'une suite d'estimateurs.

DÉFINITION 4.4 (Estimation consistante). *Une suite* $(\widehat{g}_n)$ *d'estimateurs de* $g(\theta_0)$ *est dite consistante lorsque*

$$\widehat{g}_n \xrightarrow{\mathbb{P}} g(\theta_0) \ .$$

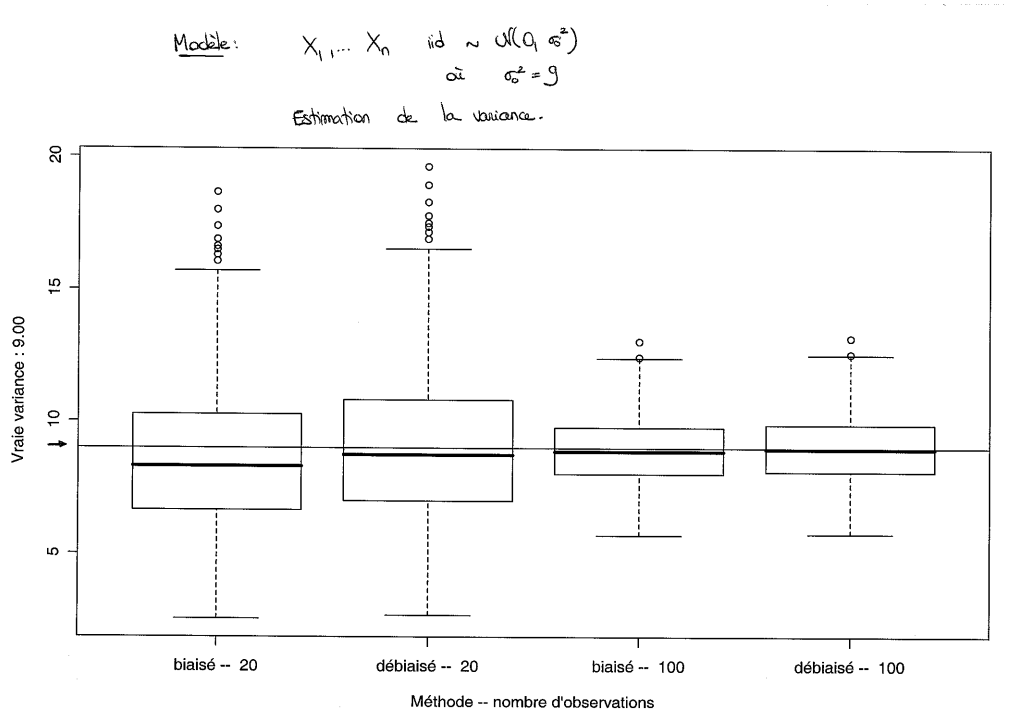


FIGURE 18. Comparaison des performances des versions biaisée et débiaisée de l'estimateur de la variance dans un modèle gaussien, par utilisation de boîtes à moustaches.

Interprétation : On a la garantie qu'à un rang n assez grand et avec grande probabilité, $\widehat{g}_n$ soit proche du paramètre d'intérêt $g(\theta_0)$. Contrairement au critère de biais, le critère de consistance ne vaut pas seulement en espérance, mais avec grande probabilité.

REMARQUE 4.5. La consistance est évidemment une vue de l'esprit, un outil d'évaluation théorique. En pratique, la taille d'échantillon n est ce qu'elle est ! On peut tout au plus planifier de la prendre suffisamment grande si l'on n'a pas encore fini la phase de recueil des données (en un sens qui sera quantifié par la troisième qualité éventuelle, la normalité asymptotique, voir ci-dessous, mais pas par la propriété de consistance, qui ne met en jeu aucune vitesse de convergence).

EXEMPLE 4.4. La loi des grands nombres assure que la suite des moyennes empiriques $\widehat{\mu}_n = \overline{X}_n$ estime l'espérance μ_0 de manière consistante (lorsque cette dernière existe).

En fait, la loi des grands nombres est souvent l'outil fondamental pour prouver une consistance, surtout quand on l'associe au résultat de la proposition 4.1 ci-dessous [4], qui décrit des propriétés de stabilité de la convergence en probabilité.

PROPOSITION 4.1 (Stabilité de la convergence en probabilité). *La convergence en probabilité passe aux fonctions continues d'un nombre fini de variables. Par exemple, dans le cas d'une fonction continue de deux variables* $(y, z) \mapsto g(y, z)$*, si par ailleurs*

4. Il ne nous serait pas trop difficile de prouver cette proposition, en utilisant notamment qu'une fonction continue de plusieurs variables sur un compact est absolument continue ; il suffit alors de choisir un compact idoine ; mais ce serait pénible (c'est une preuve en δ et ε !), coûteux en temps et inutile pour les sciences de gestion, de sorte qu'on admet le résultat énoncé.

on a deux suites de variables aléatoires (Y_n) *et* (Z_n) *convergeant en probabilité respectivement vers des variables aléatoires* Y *et* Z*,*

$$Y_n \xrightarrow{\mathbb{P}} Y \text{ et } Z_n \xrightarrow{\mathbb{P}} Z \ , \qquad \text{alors} \qquad g(Y_n, Z_n) \xrightarrow{\mathbb{P}} g(Y, Z) \ .$$

En particulier,

$$Y_n + Z_n \xrightarrow{\mathbb{P}} Y + Z \qquad \text{ou} \qquad Y_n Z_n \xrightarrow{\mathbb{P}} YZ \ .$$

EXEMPLE 4.5. On suppose ici que la loi commune des variables indépendantes $X_1, \ldots, X_n$ admet un moment d'ordre deux. Par application de la loi des grands nombres et de la proposition 4.1, on peut alors proposer comme suite d'estimateurs consistants de la variance σ_0^2 la suite des

$$\widehat{m}_{2,n} - (\overline{X}_n)^2 \ .$$

Comme $n/(n-1) \to 1$, a également alors, par une nouvelle application de la proposition 4.1, que la suite des $\widehat{\sigma}_n^2$ est consistante pour l'estimation de σ_0^2. Enfin, en passant aux racines carrées, on obtient donc deux suites d'estimateurs consistants de l'écart-type σ_0.

REMARQUE 4.6. Les techniques employées dans l'exemple précédent, i.e., loi des grands nombres combinée à la proposition 4.1, forment ce que l'on appelle la méthode des moments. On la présente plus en détails dans les compléments.

On conclut ce paragraphe par un dernier exemple.

EXEMPLE 4.6 (Loi de Bernoulli). Lorsque la loi commune des observations est de Bernoulli (de paramètre inconnu p_0), comment estimer la variance

$$\sigma_0^2 = p_0(1-p_0) = g(p_0) \ ?$$

Comme $\widehat{p}_n = \overline{X}_n$ est une suite d'estimateurs consistants de p_0, on a envie de considérer, vu la proposition 4.1, les estimateurs

$$g(\widehat{p}_n) = \widehat{p}_n(1-\widehat{p}_n) = \overline{X}_n(1-\overline{X}_n) \ .$$

En fait, puisque pour des variables de Bernoulli, $X_j = X_j^2$, l'expression à laquelle on vient de penser coïncide exactement avec l'expression générale

$$\widehat{m}_{2,n} - (\overline{X}_n)^2$$

de la version biaisée de l'estimateur de la variance. Mais à cause de son écriture naturelle, c'est souvent elle que l'on utilise, sans facteur de débiaisement.

4. Quantiles d'une loi

4.1. Définition théorique. Dans la suite, nous n'aurons presque toujours qu'à déterminer les quantiles de lois admettant une densité f strictement positive sur $\mathbb{R}$ ou $\mathbb{R}_+$. Les définitions suivantes ne considèrent donc que ce cadre. On énoncera une remarque précisant la définition des quantiles empiriques (sur des données).

DÉFINITION 4.5 (Fonction de répartition). *Etant donnée une variable aléatoire* X*, de loi admettant une densité* f*, on définit leur fonction de répartition* F *comme, pour tout* $x \in \mathbb{R}$*,*

$$F(x) \overset{\text{not.}}{=} \mathbb{P}\{X \leqslant x\} = \int_{-\infty}^{x} f(t)\,dt \ .$$

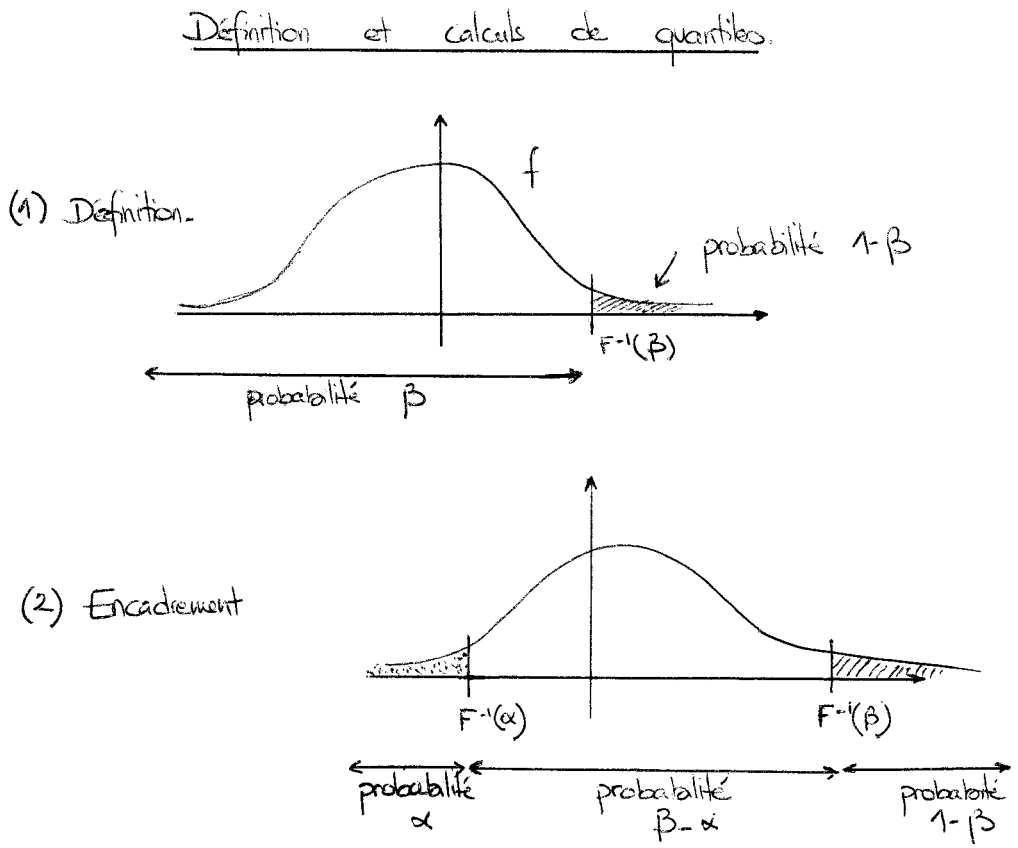


FIGURE 19. Illustration graphique de la définition des quantiles, pour une loi chargeant tout $\mathbb{R}$: la loi normale.

DÉFINITION–THÉORÈME 4.1 (Fonction quantile). *Avec les notations précédentes, lorsque la loi charge tout* $\mathbb{R}$*, i.e.,* $f > 0$ *sur* $\mathbb{R}$*, la fonction de répartition* F *réalise une bijection croissante* $\mathbb{R} \to]0, 1[$*. On appelle fonction quantile son inverse* F^{-1}*. Pour tout* $\beta \in]0, 1[$*, on définit le* β*–quantile (ou quantile d'ordre* β*) comme* $F^{-1}(\beta)$*.*

En particulier, on a, avec les notations précédentes,

$$\mathbb{P}\Big\{X \leqslant F^{-1}(\beta)\Big\} = \beta \qquad \text{et} \qquad \mathbb{P}\Big\{X \geqslant F^{-1}(\beta)\Big\} = 1 - \beta \ .$$

On illustre cela à la figure 19. En particulier, la médiane est le quantile à 50 %.

REMARQUE 4.7. Certaines lois ne chargent que $\mathbb{R}_+$, comme par exemple la loi exponentielle (ou les lois du χ^2, que nous allons voir bientôt). Pour elles, on adapte la définition précédente de la sorte. Soit une loi telle que $f > 0$ sur $\mathbb{R}_+^*$ et $f = 0$ sur $\mathbb{R}_-^*$ (la valeur en 0 est sans objet). La fonction de répartition associée F réalise ici une bijection $\mathbb{R}_+ \to [0, 1[$; on appelle fonction quantile son inverse F^{-1}, et pour tout $\beta \in [0, 1[$, on définit le β–quantile comme $F^{-1}(\beta)$. La différence essentielle par rapport au cas précédent concerne le traitement du cas $\beta = 0$. On a en effet de nombreuses valeurs x telles que

$$\mathbb{P}\{X \leqslant x\} = 0 \ ;$$

parmi elles, on ne retient que $x = F^{-1}(0) = 0$.

4.2. Application aux lois usuelles. On utilise les notations suivantes pour les lois usuelles.

- Pour la loi normale standard $\mathcal{N}(0,1)$, les β–quantiles sont notés z_β.
- On introduira formellement au cours prochain la loi de Student $\mathcal{T}_k$. Elle admet un paramètre entier k, appelé nombre de degrés de liberté. On note ses quantiles $t_{k,\beta}$, où k désigne le nombre de degrés de liberté et β l'ordre du quantile.
- De même pour les lois du χ^2, désignées chacune par χ^2_k, selon son nombre de degrés de liberté k : on note ses quantiles $\chi^2_{k,\beta}$ ou $c_{k,\beta}$. (On verra que la loi du χ^2 ne charge que $\mathbb{R}_+$.)

REMARQUE 4.8 (Relations de symétrie). La loi normale standard et les lois de Student sont symétriques par rapport à 0. En conséquence, pour tous β et k,

$$z_{1-\beta} = -z_\beta \qquad \text{et} \qquad t_{k,1-\beta} = -t_{k,\beta} \ .$$

4.3. Calcul pratique.
En pratique, et selon les outils à disposition,

- on exploite un logiciel sur ordinateur (c'est la méthode moderne) ;
- ou on lit une table de quantiles (c'est la méthode ancienne, qui date du temps où les calculs sur ordinateur n'étaient pas accessibles à tous, et que nous retiendrons pour l'examen et les quizz, où les ordinateurs sont interdits).

Recours à des tables. Lorsque l'on lit une table, il faut bien vérifier si c'est la fonction quantile F^{-1} ou la fonction de répartition F qui est tabulée. Le graphique en haut de la table, de même que le titre en gras, l'indiquent. Ainsi, sur les tables disponibles à la fin de ce polycopié, vous pouvez lire, dans l'ordre :

1. les valeurs de la fonction de répartition F de la loi $\mathcal{N}(0,1)$;
2. les valeurs de la fonction quantile F^{-1} de la loi $\mathcal{N}(0,1)$;
3. les valeurs des fonctions quantiles F^{-1} des lois χ^2_k ;
4. les valeurs des fonctions quantiles F^{-1} des lois $\mathcal{T}_k$.

Pas de mystère, il faut un peu de pratique pour savoir lire les tables. Il est donc essentiel et très important que vous fassiez les exercices que je propose, seuls, de manière concentrée et sans regarder la solution.

Calcul informatique. Sous SPSS, il est possible en utilisant des fonctions de la forme `IDF.xxx`, où `xxx` est le nom de la loi (par exemple, `IDF.NORMAL` ou `IDF.T`). C'est également possible sous `Excel`. En fait, SPSS est tellement automatisé que la plupart du temps, il est absolument inutile de calculer soi-même les quantiles, SPSS le fait pour l'utilisateur sans que celui-ci le demande, afin de déterminer directement un intervalle de confiance du niveau requis ou pour donner la P-valeur d'un test. (Les cours suivants expliqueront cela en détails.)

Cas de données $x_1, \ldots, x_n$. On peut aussi définir ici la notion de quantile empirique d'ordre β, en généralisant la définition de la moyenne empirique : il suffit de trouver un seuil (en une donnée ou entre deux données) tel qu'une fraction au moins β des données lui soient inférieures et une fraction au moins $1-\beta$ des données lui soient supérieures.

Compléments pour étudiants avancés

5. Troisième qualité éventuelle d'un estimateur : la normalité asymptotique

Rigoureusement parlant, la normalité asymptotique ne peut être, comme la consistance, la propriété que d'une suite d'estimateurs.

Définition 4.6 (Normalité asymptotique). *Une suite* $(\widehat{g}_n)$ *d'estimateurs de* $g(\theta_0)$ *est dite asymptotiquement normale, à vitesse* $\sqrt{n}$ *et de variance asymptotique* σ^2_{g,θ_0}*, lorsque*

$$\sqrt{n}\,(\widehat{g}_n - g(\theta_0)) \quad \rightharpoonup \quad \mathcal{N}\left(0,\, \sigma^2_{g,\theta_0}\right) \ .$$

Interprétation : On peut prouver que la normalité asymptotique entraîne la consistance. C'est une propriété plus précise qui indique que la fluctuation de l'estimateur autour de l'objectif à estimer est approximativement normale :

$$\widehat{g}_n \overset{(d)}{\approx} \mathcal{N}\left(g(\theta_0),\ \frac{\sigma^2_{g,\theta_0}}{n}\right) \ .$$

La figure 20 illustre cela. Les résultats de normalité asymptotique nous seront fort utiles lors de la construction d'intervalles de confiance asymptotiques, au cours suivant.

Exemple 4.7. Le théorème de la limite centrale indique (dès lors que le moment d'ordre deux existe) que la suite des moyennes empiriques $\widehat{\mu}_n = \overline{X}_n$ estime l'espérance μ_0 de manière asymptotiquement normale, à vitesse $\sqrt{n}$ et avec σ_0^2 comme variance asymptotique.

De la même manière que la convergence en probabilité passe aux fonctions continues, la convergence en loi passe aux fonctions continues, tandis que la normalité asymptotique passe, d'une certaine manière, aux fonctions C^1. La seconde assertion de la proposition suivante, associée au théorème de la limite centrale, est l'ingrédient fondamental pour prouver des propriétés de normalité asymptotique ; cependant, je ne vous la propose que pour votre culture, nous n'en aurons pas besoin dans le cadre de ce cours.

Proposition 4.2 (Propriétés de la convergence en loi).

1. *Si on a la convergence en loi* $Y_n \rightharpoonup Y$ *et si* ψ *est une fonction continue, alors on a encore* $\psi(Y_n) \rightharpoonup \psi(Y)$.
2. *Si la suite de variables aléatoires* (Y_n) *est asymptotiquement normale, telle qu'il existe* y *et* σ^2_y *avec*

   $$\sqrt{n}(Y_n - y) \quad \rightharpoonup \quad \mathcal{N}(0, \sigma^2_y) \ ,$$

   et si ψ *est une fonction* C^1*, alors* $(\psi(Y_n))$ *est également asymptotiquement normale :*

   $$\sqrt{n}\,(\psi(Y_n) - \psi(y)) \quad \rightharpoonup \quad \mathcal{N}\left(0,\, \psi'(y)^2\,\sigma^2_y\right) \ .$$

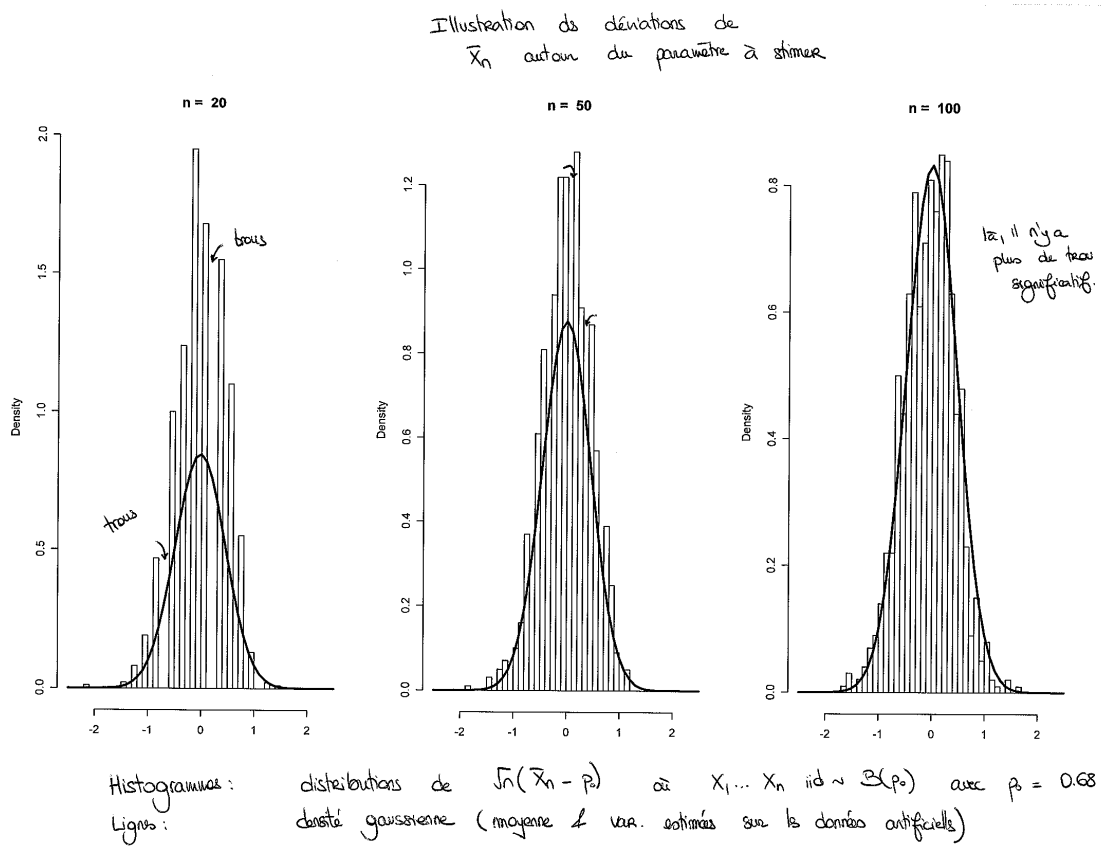


FIGURE 20. Dans un modèle de Bernoulli, histogrammes des déviations $\sqrt{n}(\overline{X}_n - p_0)$ de $\overline{X}_n$ autour du paramètre d'intérêt p_0, en fonction de la taille d'échantillon n.

6. Cas particulier : estimation d'une tendance centrale

On a introduit dans cette partie un vocabulaire général pour l'estimation : intéressons-nous à un cas particulier fréquent, l'estimation d'une tendance centrale. Par tendance centrale, on entend comportement moyen, mais bien sûr, il peut s'agir d'estimer la moyenne ou la médiane, selon le contexte et la loi considérés.

Les propriétés de l'estimateur de la moyenne empirique ont été vues plus haut : il est sans biais, consistant, et, sous réserve de l'existence d'un moment d'ordre deux, asymptotiquement normal. C'est souvent le meilleur, en théorie. Le problème, c'est que sur des données, il n'y a souvent qu'un ajustement imparfait à la modélisation théorique espérée (la pratique n'est pas aussi belle que la théorie) : par exemple, il y a souvent quelques données atypiques ("outliers" : bien plus petites ou bien plus grandes que les autres), voire incorrectes.

Les réalisations $\overline{x}_n$ de la moyenne empirique $\overline{X}_n$ sont sensibles à ces données atypiques et on va mentionner quelques estimateurs plus robustes (moins sensibles à elles).

6.1. Estimateur de la médiane empirique. Une première alternative est la médiane. Lorsque la distribution est symétrique autour d'une valeur, comme c'est le cas par exemple pour la loi normale, moyenne et médiane coïncident. Estimer l'une revient à estimer l'autre. A l'inverse, lorsque la distribution est dissymétrique, la médiane peut refléter

davantage le comportement moyen ressenti, comme nous l'avions vu cours précédent, sur l'exemple des salaires inter-professionnels.

Or, il se trouve que l'estimateur de la médiane empirique $\widehat{M}_n$ est un bon estimateur de la médiane de la loi commune. Il nous faut définir plus précisément ces deux quantités.

DÉFINITION 4.7 (Médiane d'une loi). *On fixe une loi et on prend une variable aléatoire* X *distribuée selon cette loi. Sa médiane est tout nombre* m *tel que*

$$\mathbb{P}\{X \leqslant m\} \geqslant 1/2 \qquad \text{et} \qquad \mathbb{P}\{X \geqslant m\} \geqslant 1/2 \ .$$

La médiane existe toujours mais n'est pas nécessairement unique (voir la figure 21).

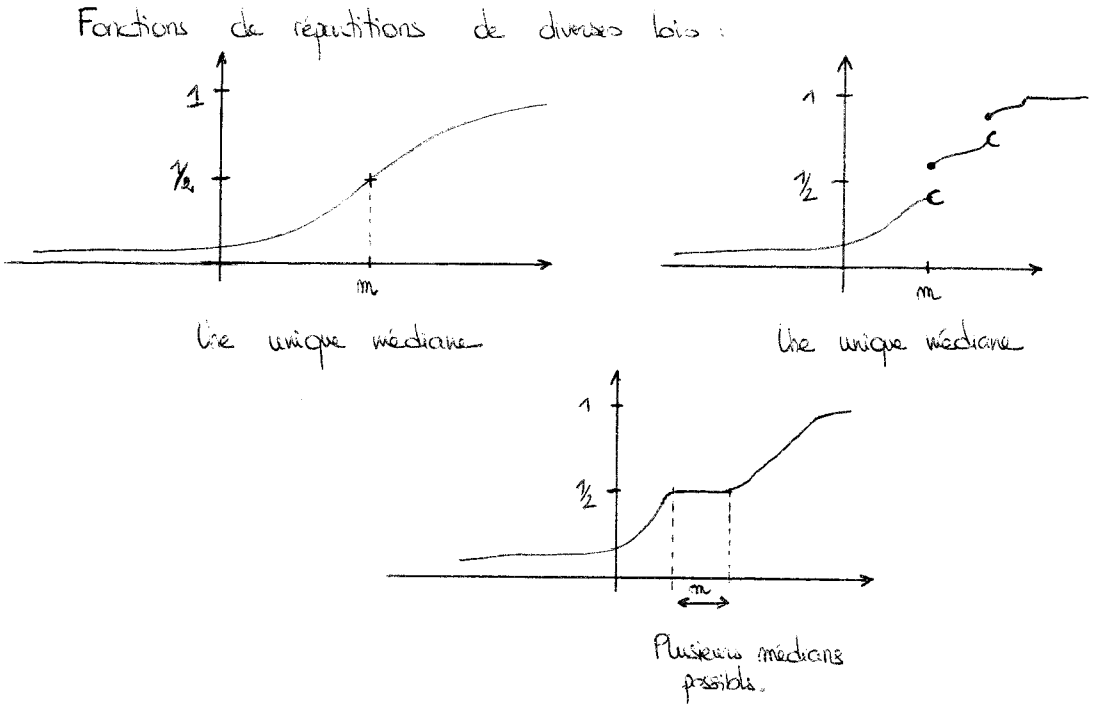


FIGURE 21. Détermination de médianes (uniques ou non) pour différentes lois.

Cependant, dans le cadre de la section 4, la médiane est unique : c'est le quantile à 50 %.

DÉFINITION 4.8 (Médiane empirique). *On fixe* n *observations* $X_1, \ldots, X_n$*, que l'on réordonne :*

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)} \ .$$

Par convention, la médiane empirique $\widehat{M}_n$ *de* $X_1, \ldots, X_n$ *est définie comme*

$$\widehat{M}_n = \begin{cases} X_{(k+1)} & \text{lorsque } n = 2k+1 \text{ est impair ;} \\ \dfrac{X_{(k)} + X_{(k+1)}}{2} & \text{lorsque } n = 2k \text{ est pair.} \end{cases}$$

$\widehat{M}_n$ *est telle que la moitié au moins des observations* X_t *lui soient plus petites, et qu'au moins la moitié des mêmes observations lui soient plus grandes. (Le lien avec la définition précédente est expliqué à la figure 22.)*

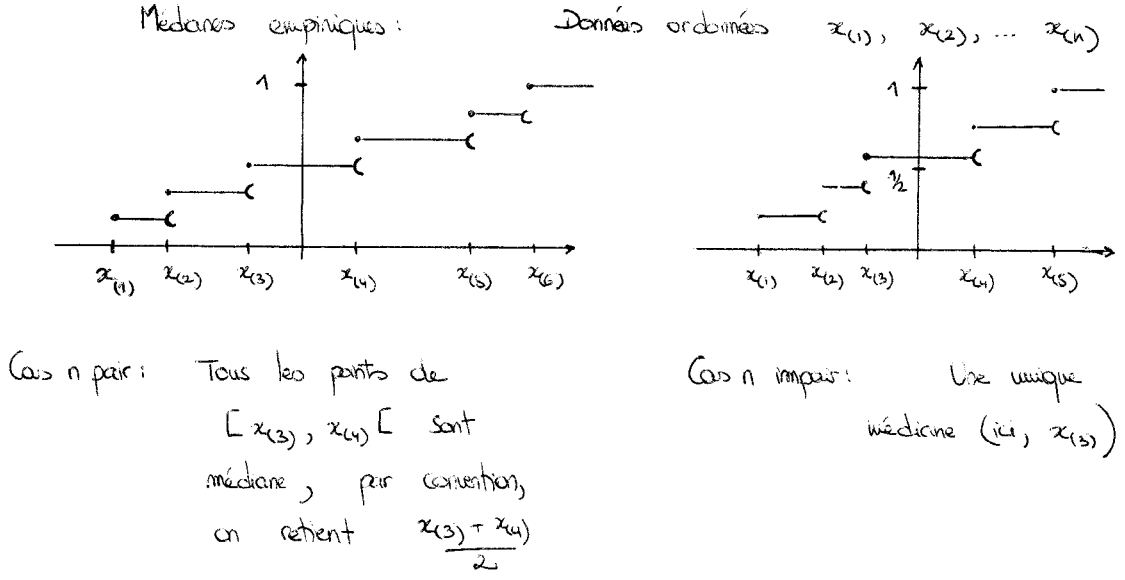


FIGURE 22. Lien entre médianes empiriques et médianes d'une loi.

On a le résultat théorique suivant. (On y suppose l'unicité de la médiane pour chaque loi du modèle car on ne connaît pas θ_0 ; il suffirait évidemment d'avoir cette unicité pour la vraie loi sous-jacente.)

THÉORÈME 4.1. *Soient des variables aléatoires* $X_1, \dots, X_n$ *indépendantes et identiquement distribuées selon une loi* $\mathbb{P}_{\theta_0}$ *prise dans le modèle* $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$. *Si la médiane* m_θ *de chacune des lois* $\mathbb{P}_\theta$ *est unique, alors l'estimateur de la médiane empirique* $\widehat{M}_n$ *est consistant, i.e.,*

$$\widehat{M}_n \xrightarrow{\mathbb{P}} m_{\theta_0} .$$

REMARQUE 4.9. Le théorème ci-dessus est un exemple utile d'estimateur consistant pour une quantité (la médiane) ne pouvant être exprimée comme une moyenne ou une variance. On a même une propriété de normalité asymptotique sous des hypothèses supplémentaires.

L'estimateur $\widehat{M}_n$ est donc un bon estimateur, bien que souvent un peu moins bon en théorie et en pratique que la moyenne empirique $\overline{X}_n$, lorsque moyenne et médiane coïncident (voir la figure 23).

Il se montre en revanche robuste : il est peu sensible aux données atypiques ("outliers"), ce qui est une qualité assez essentielle.

6.2. Autres estimateurs de la tendance centrale.

LA MINUTE SPPS 4.2. La figure 24 reprend quelques lignes parmi celles obtenues en lançant `Analyse / Statistiques descriptives / Explorer` et en sélectionnant les `estimateurs-M` après sélection du bouton `Statistiques`.

On voit tout d'abord la moyenne, puis la moyenne calculée sur 95 % des observations, en enlevant les 2.5 % d'observations les plus petites et les 2.5 % d'observations plus grandes (histoire de se débarrasser des valeurs atypiques). Si ces deux valeurs diffèrent beaucoup,

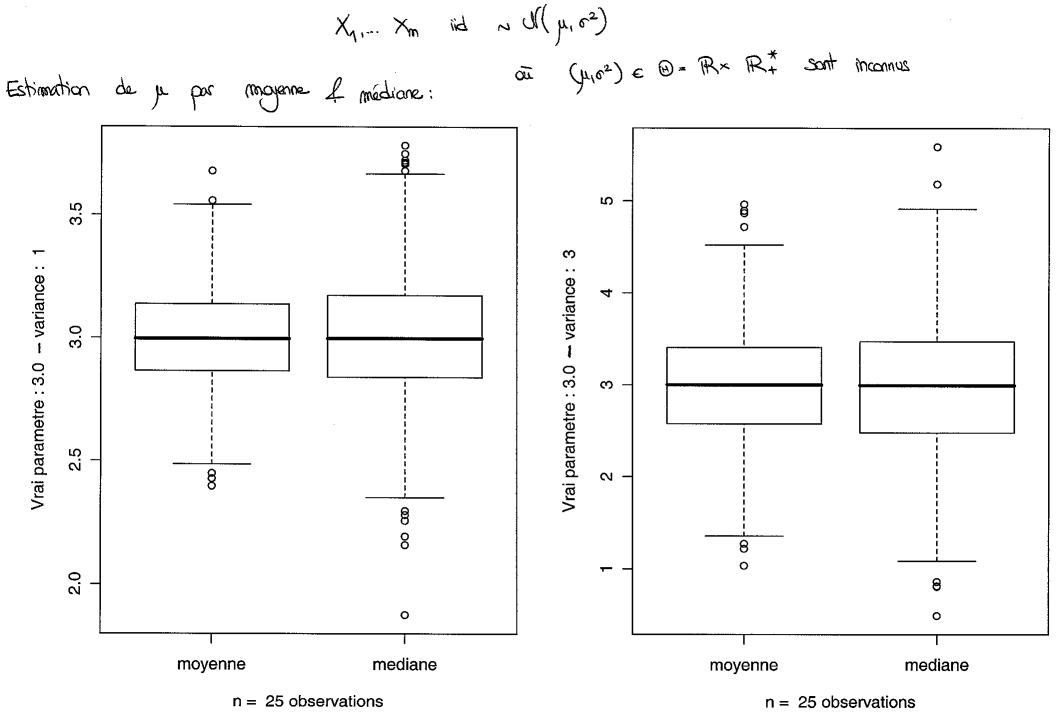


FIGURE 23. Comparaison des performances des estimateurs de la moyenne et de la médiane empiriques, sur des données simulées issues de lois normales de paramètres respectifs $\mu_0 = 3$ et $\sigma_0^2 = 1$ (gauche) et $\mu_0 = 3$ et $\sigma_0^2 = 3$ (droite).

[Ensemble_de_données] Salaires.sav

Descriptives

		Statistique	Erreur standard
SalaireMensuel	Moyenne	1,965100	,0359174
	Moyenne tronquée à 5%	1,769574	
	Médiane	1,533100	

M-estimateurs

	M-estimateur de Huber[a]	Tukey[b]	M-estimateur de Hampel[c]	Andrews[d]
SalaireMensuel	1,610143	1,499708	1,593716	1,496129

a. La constante de pondération est 1,339.

b. La constante de pondération est 4,685.

c. Les constantes de pondération sont 1,700, 3,400 et 8,500...

d. La constante de pondération est 1,340*pi.

FIGURE 24. Calcul d'estimées de la tendance centrale sur l'exemple des salaires inter-professionnels.

c'est le signe qu'il y a des valeurs atypiques tirant la moyenne vers le haut ou le bas. Vient ensuite la médiane.

Enfin, suit un tableau présentant des alternatives robustes aux estimateurs de la moyenne et de la médiane empiriques. Je ne les présente pas en détails, mais essayez de trouver ce qu'en dit la documentation de SPSS (version anglaise) :

M-estimators. Robust alternatives to the sample mean and median for estimating the location. The estimators calculated differ in the weights they apply to cases. Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, and Tukey's biweight estimator are displayed.

REMARQUE 4.10. Lorsque la loi sous-jacente est symétrique autour de 0, la suite des moyennes calculées sur 95 % des observations est également consistante. Ce n'est plus le cas en général en l'absence de symétrie.

7. La méthode des moments

7.1. Présentation générale. Dans tout ce qui suit, on considère encore un modèle statistique $X_1, \ldots, X_n$ formé de variables aléatoires indépendantes et identiquement distribuées. On note, pour $k = 1, 2, \ldots$ et sous réserve d'existence,

$$m_k(\theta_0) = \mathbb{E}\left[X_1^k\right]$$

le k–ième moment de la loi commune des observations de l'échantillon. Par la loi des grands nombres, on l'estime de manière consistante par

$$\widehat{m}_{k,n} = \frac{1}{n}\left(X_1^k + \ldots + X_n^k\right) .$$

(Evidemment, on a $\widehat{m}_{1,n} = \overline{X}_n$.) Si le paramètre d'intérêt $g(\theta_0)$ peut s'écrire comme

$$g(\theta_0) = \psi\big(m_1(\theta_0), \ldots, m_k(\theta_0)\big)$$

pour une certaine fonction ψ continue et un entier k, alors on propose la suite d'estimateurs définie par

$$\widehat{g}_n = \psi\left(\widehat{m}_{1,n}, \ldots, \widehat{m}_{k,n}\right) .$$

Cette suite est consistante pour l'estimation de $g(\theta_0)$, comme l'assure la proposition 4.2.

Interprétation : La méthode des moments consiste donc à remplacer les moments par leurs estimateurs empiriques.

On note que c'est exactement ainsi qu'ont été formés les estimateurs de l'exemple 4.5. Cependant, comme on l'a vu à l'exercice 4.1, la méthode des moments conduit ainsi à une suite d'estimateurs biaisés de la variance.

REMARQUE 4.11. Une version multi-dimensionnelle du théorème de la limite centrale montre que les estimateurs par moments étudiés plus haut sont asymptotiquement normaux – sous réserve d'existence des moments : si l'estimateur met en jeu des moments d'ordre inférieur ou égal à k, alors il suffit que la loi commune des observations admette un moment d'ordre 2k pour que la normalité asymptotique soit établie.

7.2. Premier exemple simple : la loi de Poisson. Dans les cas les plus simples, le ou les paramètres du modèle sont donné(s) par la moyenne et/ou la variance, et les techniques précédentes s'appliquent donc aisément.

Ainsi, pour la loi $\mathcal{P}(\lambda_0)$, i.e., la loi de Poisson de paramètre λ_0, on a $\mu_0 = \sigma_0^2 = \lambda_0$. La méthode des moments propose donc les estimateurs

$$\widehat{\lambda}_n = \widehat{m}_{1,n} = \overline{X}_n \qquad \text{et} \qquad \widehat{\lambda}'_n = \widehat{m}_{2,n} - \left(\overline{X}_n\right)^2 .$$

Lequel est le meilleur ? Un coup d'œil à la figure 25 semble indiquer que c'est la moyenne empirique plutôt que la variance empirique.

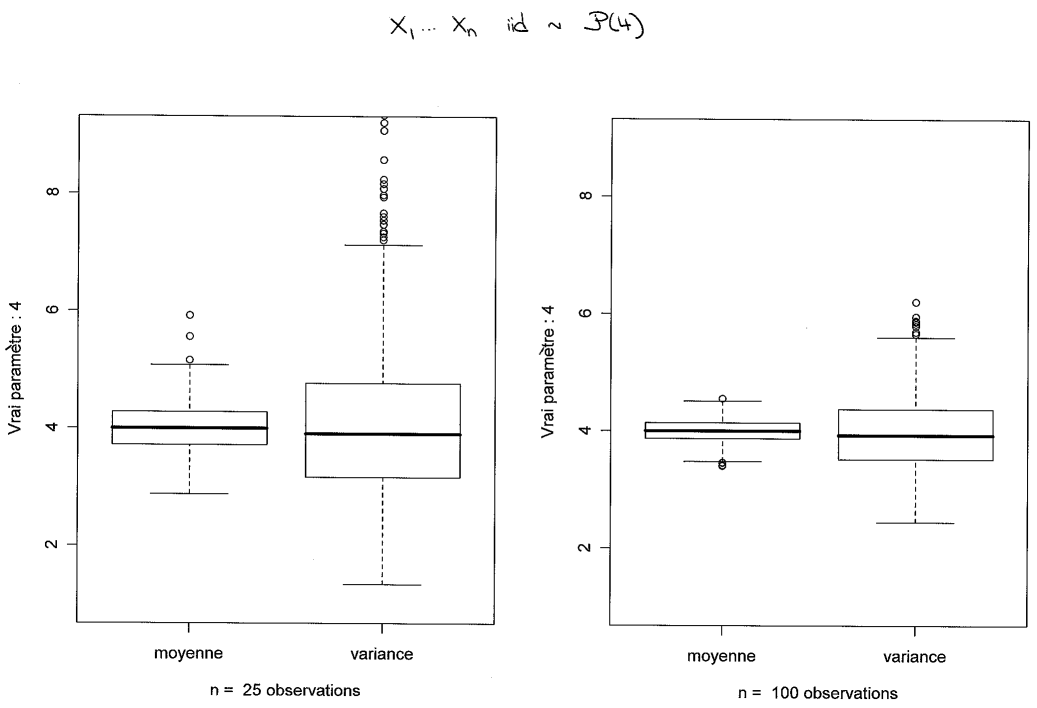


FIGURE 25. Représentations du comportement des estimateurs de la moyenne empirique et de la variance empirique dans un modèle de Poisson de paramètre $\lambda_0 = 4$.

Des résultats de normalité asymptotique vont en fait nous permettre de trancher en faveur du premier. C'est là un exemple montrant combien la réflexion théorique peut être importante en pratique.

En effet, l'estimateur de la moyenne empirique vérifie, par théorème de la limite centrale, que

$$\sqrt{n}\left(\widehat{\lambda}_n - \lambda_0\right) \quad \rightharpoonup \quad \mathcal{N}(0, \lambda_0)$$

tandis qu'on peut montrer, par des versions multi-dimensionnelles du théorème de la limite centrale et du résultat de la proposition 4.2 (hors du programme de ce cours), que

$$\sqrt{n}\left(\widehat{\lambda}'_n - \lambda_0\right) \quad \rightharpoonup \quad \mathcal{N}\left(0,\, \lambda_0 + 2\lambda_0^2\right) \ .$$

On rappelle que l'estimateur le meilleur des deux est celui de variance asymptotique la plus faible (en effet, c'est celui le plus ramassé autour de λ_0, celui dont le pic gaussien a la base la plus étroite). Le traitement mathématique montre ainsi que dans le cadre d'un modèle de Poisson, on préfère l'estimateur de la moyenne empirique $\widehat{\lambda}_n$ à celui de la variance empirique $\widehat{\lambda}'_n$.

7.3. Second exemple simple : la loi exponentielle.

EXERCICE 4.2 (Loi exponentielle). On considère le modèle statistique $X_1, \dots, X_n$ de variables aléatoires indépendantes et identiquement distribuées selon une loi exponentielle $\mathcal{E}(\lambda)$, où $\lambda \in]0, +\infty[$. On note λ_0 le vrai paramètre (inconnu).

1. Montrer, en intégrant par parties, que $m_1(\lambda_0) = 1/\lambda_0$ et $m_2(\lambda_0) = 2/\lambda_0^2$.
2. En déduire deux estimateurs par moments de λ_0.

3. Montrer qu'au moins un de ces estimateurs est asymptotiquement normal et préciser sa variance asymptotique.

7.4. Notion d'estimateur efficace.

REMARQUE 4.12 (Une remarque culturelle). Voici, à titre culturel uniquement, une allusion rapide à un résultat fondamental de statistique (à destination de ceux qui seraient curieux, là encore, de savoir ce en quoi consiste la recherche en statistique). Dans la définition de la normalité asymptotique, le paramètre de variance asymptotique σ^2_{g,θ_0} joue un rôle crucial car il mesure la qualité de l'estimation : on veut considérer des estimateurs de variance asymptotique minimale. Se posent les problèmes de 1. calculer cette valeur minimale et 2. d'exhiber des estimateurs l'atteignant (dits, par définition, efficaces). Le point 1. est réglé par la minoration dite de Cramer–Rao. Quant au point 2., il est généralement résolu par une autre méthode de construction d'estimateurs, dite du maximum de vraisemblance (mais que nous ne verrons pas dans ce cours).

EXEMPLE 4.8. Dans le modèle de Poisson, on peut montrer que la variance minimale pour l'estimation de λ_0 est λ_0 : l'estimateur de la moyenne empirique $\widehat{\lambda}_n$ est efficace.

7.5. Deux exemples plus complexes illustrant les limites de la méthode des moments. Nous donnons deux exemples où les estimateurs proposés par la méthode des moments sont parfois déraisonnables. Cela illustre que la méthode des moments n'est pas la panacée. Elle est facile à mettre en œuvre, certes, mais ne procure pas toujours un résultat satisfaisant. D'ailleurs, d'une manière générale, aucune autre méthode d'estimation n'est parfaite, chacune a ses qualités et défauts.

7.5.1. *Où l'on illustre le vice mathématique !*

Les mathématiciens aiment bien construire des contre-exemples tordus, ne correspondant pas forcément à une réalité, simplement pour le plaisir de montrer que telle ou telle intuition est fausse.

Soit le modèle suivant. $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi uniforme sur $\{0, 1, \ldots, \theta_0\}$, où l'on sait seulement que $\theta_0 \in \mathbb{N}$. Il s'agit donc d'une loi uniforme sur les $\theta_0 + 1$ premiers entiels naturels, où θ_0 est un entier inconnu, à estimer.

EXEMPLE 4.9. J'ai eu beaucoup de mal à trouver un exemple convaincant. On pourrait par exemple penser au vrai nombre θ_0 de voitures immatriculées par an. Un an après un changement de la méthode d'attribution des plaques (et pour peu que la nouvelle méthode permettre de relier rapidement la plaque au rang de la demande d'attribution), on peut effectuer un sondage visuel à un carrefour et obtenir des observations $X_1, X_2, \ldots$, et ce, afin d'évaluer θ_0, le nombre effectif d'immatriculations. (Dans le même genre d'idées, on peut penser, en temps de guerre, à l'évaluation du nombre de chars que possède l'adversaire, permise grâce à ces observations de terrain pour que les dits chars disposent d'un numéro de série visible et donnant une indication sur leur ordre de fabrication... ce qui n'est pas garanti, les militaires du renseignement pensant quand même à cacher le maximum d'informations !)

Comme $\mu(\theta_0) = \theta_0/2$, on est tenté de proposer

$$\widehat{\theta}_n = 2\,\overline{X}_n \ ,$$

qui est consistant. Mais il est clair qu'on a nécessairement $\theta_0 \geqslant \max\{X_1, \ldots, X_n\}$. Or il se peut que $2\overline{X}_n$ soit strictement plus petit que cette valeur, auquel cas on sait pertinemment qu'on pourrait estimer mieux en prenant le maximum. En réalité, une autre méthode, dite du maximum de vraisemblance, conduirait à ce meilleur estimateur,

$$\widehat{\theta}_n = \max\{X_1, \ldots, X_n\} .$$

7.5.2. *Loi binomiale.*

EXEMPLE 4.10. Considérons un cours qui n'aurait lieu que les vendredis matins (pas les jours d'affluence maximale, n'est-ce pas ?). Supposons qu'un étudiant, ignorant par ailleurs le nombre total k_0 d'étudiants dans le groupe, observe la présence ou l'absence au cours pendant, disons, dix séances : il note $n_1, \ldots, n_{10}$ le nombre de présents. On suppose que chaque étudiant décide indépendamment des autres de se lever ou pas le vendredi matin, et qu'il a le courage de le faire avec une probabilité p_0 commune à tous. Ainsi, on peut modéliser les données comme la réalisation des variables aléatoires indépendantes et identiquement distribuées $N_1, \ldots, N_n$ (avec $n = 10$ ici), de loi commune la loi binomiale $\mathcal{B}\text{in}(k_0, p_0)$, de paramètres k_0 et p_0 inconnus et à estimer.

REMARQUE 4.13. Si l'on connaissait k_0, il serait facile d'estimer p_0 : ce serait comme si l'on avait affaire à un $(k_0\, n)$–échantillon de loi de Bernoulli de paramètre p_0.

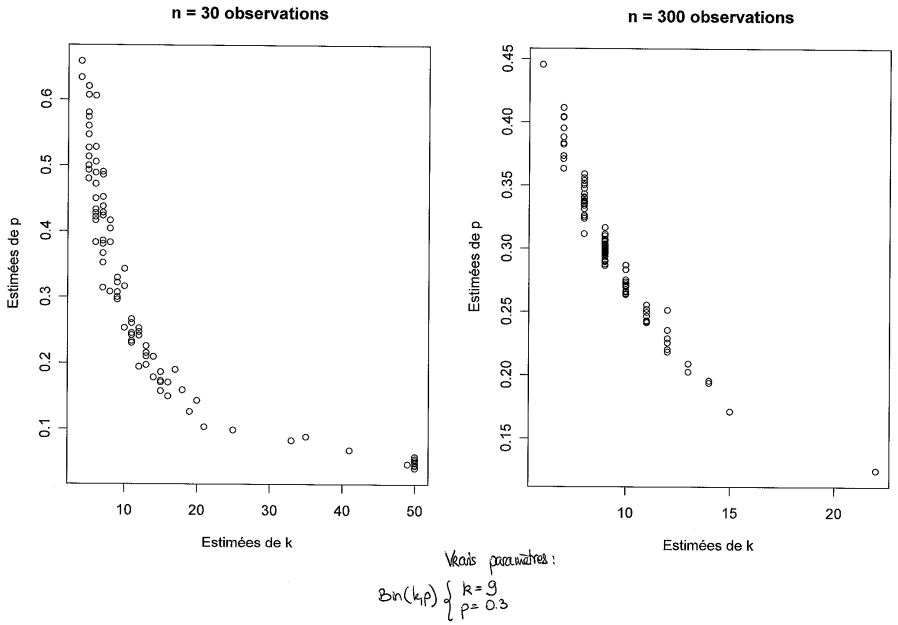


FIGURE 26. Différentes valeurs possibles pour le couple d'estimateurs exhibé dans l'exercice 4.3 ; utilisation d'échantillons de tailles $n = 30$ et $n = 300$, les vrais paramètres étant $k_0 = 9$ et $p_0 = 0.3$.

EXERCICE 4.3. On veut employer la méthode des moments pour estimer k_0 et p_0.

1. Calculez $m_1 = m_1(k_0, p_0)$ et $m_2 = m_2(k_0, p_0)$.
2. Aboutissez alors, après quelques manipulations, au système d'équations suivant,

$$\begin{cases} m_1 &= k_0\, p_0 \\ m_2 &= m_1^2 + m_1(1 - p_0) . \end{cases}$$

3. Résolvez-le pour trouver finalement, avec les notations ci-dessus,

$$\widehat{k}_n = \frac{\left(\widehat{m}_{1,n}\right)^2}{\widehat{m}_{1,n} + \left(\widehat{m}_{1,n}\right)^2 - \widehat{m}_{2,n}} \qquad \text{et} \qquad \widehat{p}_n = \frac{\widehat{m}_{1,n} + \left(\widehat{m}_{1,n}\right)^2 - \widehat{m}_{2,n}}{\widehat{m}_{1,n}} \ .$$

4. Expliquez pourquoi ces deux estimateurs sont consistants.
5. Au vu de la figure 26, que pensez-vous de la qualité de ces estimateurs ? Vous semblent-ils bons ou non ?

Note : en fait, il se peut même que les deux estimateurs exhibés ci-dessus prennent des valeurs négatives, sur des valeurs observées choisies avec suffisamment de vice par un mathématicien, alors que l'on sait que les vrais paramètres sont forcément positifs. Cela montre, encore une fois, les limites de la méthode des moments dans les cas complexes !

Exercices

Exercices qu'il est obligatoire de résoudre sérieusement et seul

Exercice 4.4 (Loi normale $\mathcal{N}(0,1)$).

1. Lire sur la table idoine les quantiles à 60 % et à 40 %. Lire ensuite les quantiles à 98.2 % et à 27.6 %.
2. Indiquer les probabilités qu'une variable aléatoire distribuée selon une loi normale soit : plus petite que 1.82, plus grande que -0.63, et comprise entre -2.13 et 1.98.
3. Enfin, calculer u tel que
$$\mathbb{P}\{|N| \leqslant u\} = 80\,\%$$
où N suit une loi $\mathcal{N}(0,1)$.

Indication : dans tout l'exercice, utiliser la symétrie de la loi normale par rapport à 0.

Exercice 4.5 (Lois de Student $\mathcal{T}$).

1. Déterminer le quantile à 2.5 % de la loi de Student à 9 degrés de liberté.
2. Préciser ensuite u tel que
$$\mathbb{P}\{|T| \leqslant u\} = 60\,\%$$
où T suit une loi de Student à 17 degrés de liberté.
3. Encadrer enfin la probabilité qu'une variable aléatoire T' distribuée selon la loi de Student à 5 degrés de liberté soit supérieure ou égale à 1.

Indication : ici aussi, utiliser la symétrie des lois de Student par rapport à 0.

Exercice 4.6 (Lois du χ^2). Soit S une variable aléatoire distribuée selon la loi du χ^2 à 7 degrés de liberté.

1. Déterminer a et b tels que
$$\mathbb{P}\{S \geqslant a\} = 95\,\% \qquad \text{et} \qquad \mathbb{P}\{S \leqslant b\} = 95\,\%\,.$$
2. Déterminer ensuite $\mathbb{P}\{S \in [a,b]\}$.
3. Encadrer enfin $\mathbb{P}\{S \leqslant 2\}$.

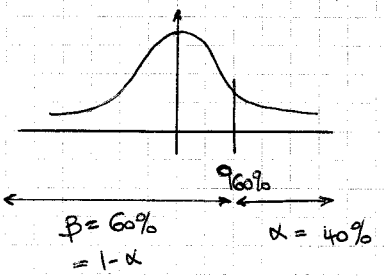
Exercices totalement facultatifs et à faire le cas échéant uniquement pour son épanouissement personnel !

Il s'agit des exercices 4.2 et 4.3, formulés dans les compléments de cours.

Exercice 1 : Loi normale $\mathcal{N}(0,1)$

(1) • On utilise la table #2 "loi normale : quantiles" → les lignes donnent les deux premiers chiffres de α, et les colonnes le troisième chiffre

• Quantile à 60% : situation de la forme



qui correspond donc à $\alpha = 40\%$, soit :

$$q_{60\%} = 0.2533$$

• Quantile à 40% : par symétrie, $q_{40\%} = -q_{60\%} = -0.2533$

• Quantile à 98.2% : correspond à $\alpha = 1.8\% = 0.018$

soit $q_{98.2\%} = 2.0969$

• Quantile à 27.6% : on calcule d'abord le quantile à $100\% - 27.6\% = 72.8\%$ (soit $\alpha = 27.6\% = 0.276$) :

$q_{72.8\%} = 0.5948$ puis

$$q_{27.6\%} = -q_{72.8\%} = -0.5948$$

(2) • On utilise cette fois la table #1 "loi normale : fonction de répartition" → les lignes donnent les deux premiers chiffres de u et les colonnes, le troisième.

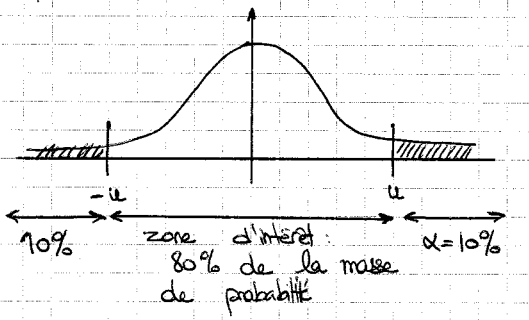
On note dans la suite N une variable aléatoire $\sim \mathcal{N}(0,1)$.

• $\mathbb{P}\{N \leq 1.82\} = F(1.82) = 0.9656 = 96.56\%$

- $\mathbb{P}\{N \geq -0.63\} = \mathbb{P}\{N \leq 0.63\}$ par symétrie
  $= F(0.63) = 0.7357 = 73.57\,\%$

- $\mathbb{P}\{-2.13 \leq N \leq 1.98\} = \mathbb{P}\{N \in [-2.13, 0]\} + \mathbb{P}\{N \in [0, 1.98]\}$
  $= \left(\mathbb{P}\{N \leq 2.13\} - \tfrac{1}{2}\right) + \left(\mathbb{P}\{N \leq 1.98\} - \tfrac{1}{2}\right)$ (par symétrie)
  $= F(2.13) + F(1.98) - 1$
  $= 0.9834 + 0.9761 - 1$
  $= 0.9595 = 95.95\,\%$

(3) Ici, le plus simple est de raisonner sur un dessin :

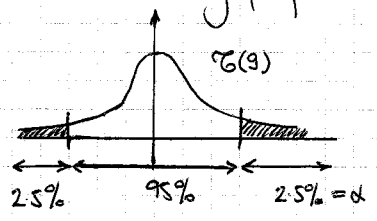


On reconnait $u = q_{90\%}$ (ou, c'est pareil, $-u = q_{10\%}$)
et sur la table #2, avec $\alpha = 10\%$, on lit :

$$q_{90\%} = 1.2816$$

Exercice 2: (Lois de Student $\mathcal{T}$) On utilise la table #4 "loi de Student: quantiles"

(1) Par symétrie, on établit le graphique suivant:

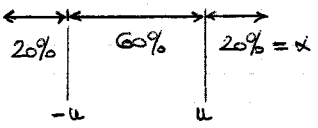


$t_{9,\,2.5\%} = -\,t_{9,\,97.5\%}$ où l'on lit sur la table (avec $\alpha = 2.5\%$)

$t_{9,\,97.5\%} = 2.262$

et donc finalement: $t_{9,\,2.5\%} = -2.262$

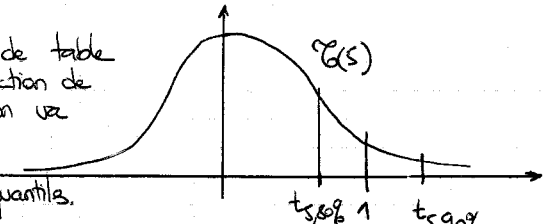
(2) On refait le même graphique, mais avec



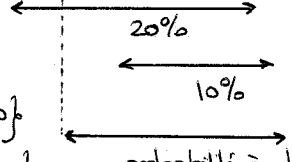
et l'on reconnait $u = t_{17,\,80\%}$ soit (avec $\alpha = 20\%$)

$u = t_{17,\,80\%} = 0.863$

(3) On n'a pas de table de la fonction de répartition, on va procéder par encadrement par des quantiles.



Le raisonnement reproduit sur le graphique ci-dessous:



$\mathbb{P}\{T' \geq 1.476\} \leq \mathbb{P}\{T' \geq 1\} \leq \mathbb{P}\{T' \geq 0.920\}$

$= \mathbb{P}\{T' \geq t_{5,90\%}\}$ $= \mathbb{P}\{T' \geq t_{5,80\%}\}$

$= 10\%$ $= 20\%$

soit $\mathbb{P}\{T' \geq 1\} \in [10\%, 20\%]$

Exercice 3: (Loi du χ^2) On utilise la table #3 "loi du χ^2: quantils"

Soit $S \sim \chi^2(7)$

(1) • Par définition, $\mathbb{P}\{S \leq \chi^2_{7,95\%}\} = 95\%$ (correspond à $\alpha = 5\%$)

soit $b = \chi^2_{7,95\%} = 14.07$

• Pour a, une définition équivalente est:

$\mathbb{P}\{S \geq a\} = 95\% \iff \mathbb{P}\{S \leq a\} = 5\%$

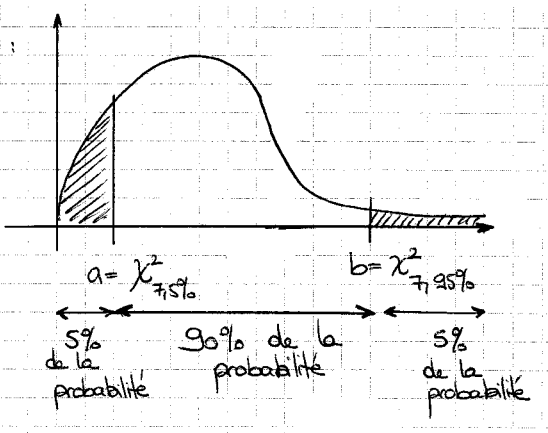
et l'on reconnait $a = \chi^2_{7,5\%}$.

Par lecture (en prenant $\alpha = 95\%$), il vient $a = \chi^2_{7,5\%} = 2.17$

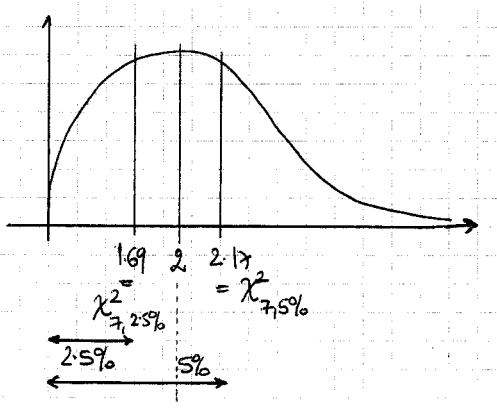
(2) $\mathbb{P}\{S \in [a;b]\} = \mathbb{P}\{S \leq b\} - \mathbb{P}\{S \leq a\}$

$= 95\% - 5\% = 90\%$,

Ce que l'on retrouve graphiquement:



(3)



Même principe qu'au (3) de l'exercice précédent:

$\mathbb{P}\{S \leq 2\} \in [2.5\%, 5\%]$

Exercice (facultatif) — **Loi exponentielle et méthode des moments.**

(1) Si $X \sim \mathcal{E}(\lambda)$ alors sa densité est donnée par

$$x \longmapsto \begin{cases} 0 & \text{sur } \mathbb{R}_- \\ \lambda e^{-\lambda x} & \text{sur } \mathbb{R}_+^* \end{cases}$$

$$m_1(\lambda) = \mathbb{E}X = \int_0^{+\infty} x \lambda e^{-\lambda x} dx$$

$$\underset{\text{IPP}}{=} \left[-x e^{-\lambda x} \right]_0^{+\infty} + \int_0^{+\infty} e^{-\lambda x} dx$$

soit $m_1(\lambda) = 1/\lambda$

et $$m_2(\lambda) = \mathbb{E}X^2 = \int_0^{+\infty} x^2 \cdot \lambda e^{-\lambda x} dx$$

$$\underset{\text{IPP}}{=} \left[-x^2 e^{-\lambda x} \right]_0^{+\infty} + \underbrace{2 \int_0^{+\infty} x e^{-\lambda x} dx}_{= \frac{2}{\lambda} \mathbb{E}[X] = 2/\lambda^2}$$

soit $m_2(\lambda) = 2/\lambda^2$

(2) On peut donc proposer $\hat{\lambda}_n = 1/\bar{X}_n$ et $\hat{\lambda}'_n = \sqrt{2/\hat{m}_{2,n}}$.

(3) Si $X \sim \mathcal{E}(\lambda_0)$ alors $\mathrm{Var}\, X = m_2(\lambda_0) - \left(m_1(\lambda_0)\right)^2 = \dfrac{1}{\lambda_0^2}$

Le théorème de la limite centrale assure alors que

$$\sqrt{n}\left(\bar{X}_n - 1/\lambda_0\right) = \sqrt{n}\left(\bar{X}_m - m_1(\lambda_0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/\lambda_0^2)$$

En utilisant les propriétés de la convergence en loi (et la fonction $\mathcal{C}^1$ $\Psi : x \in \mathbb{R}_+^* \mapsto 1/x$, avec $\Psi'(x) = -1/x^2$), il vient :

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda_0\right) = \sqrt{n}\left(1/\bar{X}_n - \lambda_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Psi(1/\lambda_0)\frac{1}{\lambda_0^2}\right) = \mathcal{N}(0, \lambda_0^2)$$

La suite $(\hat{\lambda}_n)$ est bien asymptotiquement normale.

Exercice (facultatif) : Loi binomiale et méthode des moments.

(1) On note X une variable aléatoire suivant la loi $\mathrm{Bin}(k_0, p_0)$.

$m_1(k_0, p_0) = \mathbb{E}X = k_0 p_0$

et $m_2(k_0, p_0) = \mathbb{E}X^2 = \mathbb{E}\left[(Y_1 + \dots Y_{k_0})^2\right]$ où $Y_1, \dots Y_{k_0}$ iid $\sim \mathrm{Ber}(p_0)$

$= k_0 \underbrace{\mathbb{E}Y_1^2}_{= \mathbb{E}Y_1 = p_0} + k_0(k_0-1) \underbrace{\mathbb{E}Y_1 Y_2}_{= (\mathbb{E}Y_1)^2 = p_0^2 \text{ par indépendance}}$

soit $m_2(k_0, p_0) = k_0 p_0 + k_0(k_0-1) p_0^2$

$= k_0 p_0 (1-p_0) + k_0^2 p_0^2$.

(2) Découle de (1).

(3) • $m_2 = m_1^2 + m_1(1-p_0)$ entraîne $1 - p_0 = \dfrac{m_2 - m_1^2}{m_1}$

puis $p_0 = \dfrac{m_1 + m_1^2 - m_2}{m_1}$

• ensuite, $m_1 = k_0 p_0$ entraîne $k_0 = \dfrac{m_1}{p_0} = \dfrac{m_1^2}{m_1 + m_1^2 - m_2}$

vu ce qui précède.

• en remplaçant m_1 et m_2 par leurs estimateurs empiriques $\hat{m}_{1,n}$ et $\hat{m}_{2,n}$, on obtient finalement les expressions proposées.

(4) Cela procède de la loi des grands nombres (qui implique que $\hat{m}_{1,n}$ et $\hat{m}_{2,n}$ sont consistants respectivement pour $m_1(k_0, p_0)$ et $m_2(k_0, p_0)$), ainsi que de la stabilité de la convergence en probabilité par passage aux fonctions continues.

(5) Ils semblent très mauvais : ils sous-estiment k_0, et p_0 n'est estimé raisonnablement que dans le cas où l'on a 300 données.

Cinquième Partie

Estimation par intervalles

Version rédigée du cours

Résumé : La partie précédente, plus théorique que la moyenne des parties de ce cours, a introduit le concept de l'estimation (ponctuelle : par une valeur) et a étudié quelques propriétés qu'il est désirable pour un estimateur d'avoir. Elle a également défini les quantiles d'une loi.

Objectif : En pratique, si, comme en estimation ponctuelle, on ne propose qu'une seule valeur pour le paramètre d'intérêt, il n'y a aucune chance que l'on propose la vraie valeur, même si l'on est en sans doute proche. Il vaut donc mieux indiquer une fourchette de valeurs possibles, tout un intervalle : ni trop gros, pour qu'il soit assez informatif, ni trop petit, pour qu'on soit raisonnablement sûr qu'il contienne la vraie valeur. Nous allons voir que typiquement, un $(1/\sqrt{n})$–voisinage d'un estimateur réalise ce compromis (où n est, comme toujours, la taille d'échantillon). Je vais expliquer comment on construit ces voisinages, mais pour la suite, je ne vous demanderai que d'apprendre leur expression, pas de savoir refaire les preuves.

Motivation et objectifs stratégiques

Ce chapitre va vous aider à comprendre la remarque très ironique suivante :

> Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.
> Georges Elgozy (économiste français, 1909–1989)

Vous ne présenterez plus vos résultats comme avant, et verrez qu'il est souvent ridicule de présenter des pourcentages avec de nombreuses décimales, que cela marque un manque de recul profond et vous décrédibiliserait [5] face à vos orateurs. Il faut en effet toujours préciser une marge d'erreur, l'analyse des données ne donne qu'une idée plus ou moins vague de la vérité, pas la vérité en elle-même.

Exemple 5.1 (Chirac, Le Pen et Jospin ; ou de l'usage pas assez systématique des intervalles de confiance). Les sondeurs précisent généralement des marges d'erreur (voir par exemple la figure 27, mais les journalistes ne la reprennent pas.

A une exception notable près : à 20h, les soirs d'élection, de plus en plus de pincettes sont prises pour présenter les estimations des résultats. Parfois, même un peu trop : le 21 avril 2002, pas à 20h pile, mais quelques minutes plus tard, il est précisé quelque chose comme $17.0\,\% \pm 0.5\,\%$ de votes en faveur de Jean-Marie Le Pen et $16.5\,\% \pm 0.5\,\%$ pour Lionel Jospin, une annonce pas encore tout à fait claire du second candidat du second tour. Dans l'analyse détaillée des résultats, mais pas dans l'estimation à 20h, les journalistes, surpris par un résultat que les sondages ne prévoyaient pas, prennent le maximum de précautions

5. Inversement, vous pourrez moucher tous ceux qui annoncent des chiffres trop précis, sans justifier de la précision garantie : lancez la citation précédente d'un ton détaché, en réunion, pendant que votre collègue fait sa présentation de chiffres avec peu de recul et aucune garantie de précision. Tout le monde vous prendra, à peu de frais, pour un dieu des statistiques !

et proposent des intervalles avec une intersection non-vide. Ce n'est évidemment pas la même chose que d'annoncer $17.0\,\% \pm 0.1\,\%$ et $16.5\,\% \pm 0.1\,\%$, car là, le classement est clair. On voit bien sur cet exemple que les estimées $17.0\,\%$ et $16.5\,\%$ sans intervalles de confiance associés sont quasiment inutiles ! (Au final, les scores seront de $16.86\,\%$ et $16.18\,\%$.)

Les journalistes n'avaient en revanche pris aucune sorte de précautions lors de la campagne de 2002, lorsqu'ils commentaient les sondages d'opinion. Depuis cet événement, qui a jeté le doute sur ces sondages, plutôt que de parler de marges d'erreur (à $\pm 3\,\%$), ils accompagnent chaque résultat de sondage de la formule : « Nous vous rappelons que ce sondage ne reflète qu'un instantané de l'opinion et ne préjuge pas du résultat de l'élection. » Peut mieux faire... Quand le grand public aura-t-il le droit à une information complète avec marges d'erreur ?

1. Le minimum de vocabulaire pour commencer

Comme expliqué dans les chapitres précédents, on suppose qu'on a déjà modélisé les valeurs observées $x_1, \ldots, x_n$ comme étant la réalisation de $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi commune appartenant à un certain ensemble $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$. La vraie loi $\mathbb{P}_{\theta_0}$ est inconnue et on s'intéresse à l'estimation d'une quantité $g(\theta_0)$ dérivée de cette loi (par exemple, son espérance).

Définition 5.1. *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est la donnée d'un couple d'estimateurs $\widehat{g}_n \leqslant \widehat{g}'_n$, à partir de qui on construit*

$$\widehat{I}_n = \left[\widehat{g}_n,\ \widehat{g}'_n\right] .$$

Définition 5.2 (Niveau). *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est dit de niveau au moins égal à $1-\alpha$, où $\alpha \in [0,1]$, si, quelle que soit la valeur de θ_0,*

$$\mathbb{P}\{g(\theta_0) \in \widehat{I}_n\} \geqslant 1-\alpha .$$

Exemple 5.2. On a vu dans la partie 1 que dans un modèle de Bernoulli (de paramètre p_0), l'intervalle

$$\left[\overline{X}_n - \frac{2.24}{\sqrt{n}}\ ,\ \overline{X}_n + \frac{2.24}{\sqrt{n}}\right]$$

(résultant de l'application de l'inégalité de Chebychev-Markov) est un intervalle de confiance de niveau $95\,\%$ pour le paramètre p_0.

Définition 5.3 (Niveau asymptotique). *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est dit de niveau asymptotique au moins égal à $1-\alpha$, où $\alpha \in [0,1]$, si quel que soit θ_0, pour tout $\varepsilon > 0$, il existe un rang N tel que pour tout $n \geqslant N$,*

$$\mathbb{P}\{g(\theta_0) \in \widehat{I}_n\} \geqslant 1-\alpha-\varepsilon .$$

Note : c'est le cas en particulier lorsque l'on peut garantir

$$\lim_{n\to\infty}\ \mathbb{P}\{g(\theta_0) \in \widehat{I}_n\} = 1-\alpha .$$

La définition 5.3 est cependant plus générale, au sens où il n'y est pas nécessaire qu'une limite existe.

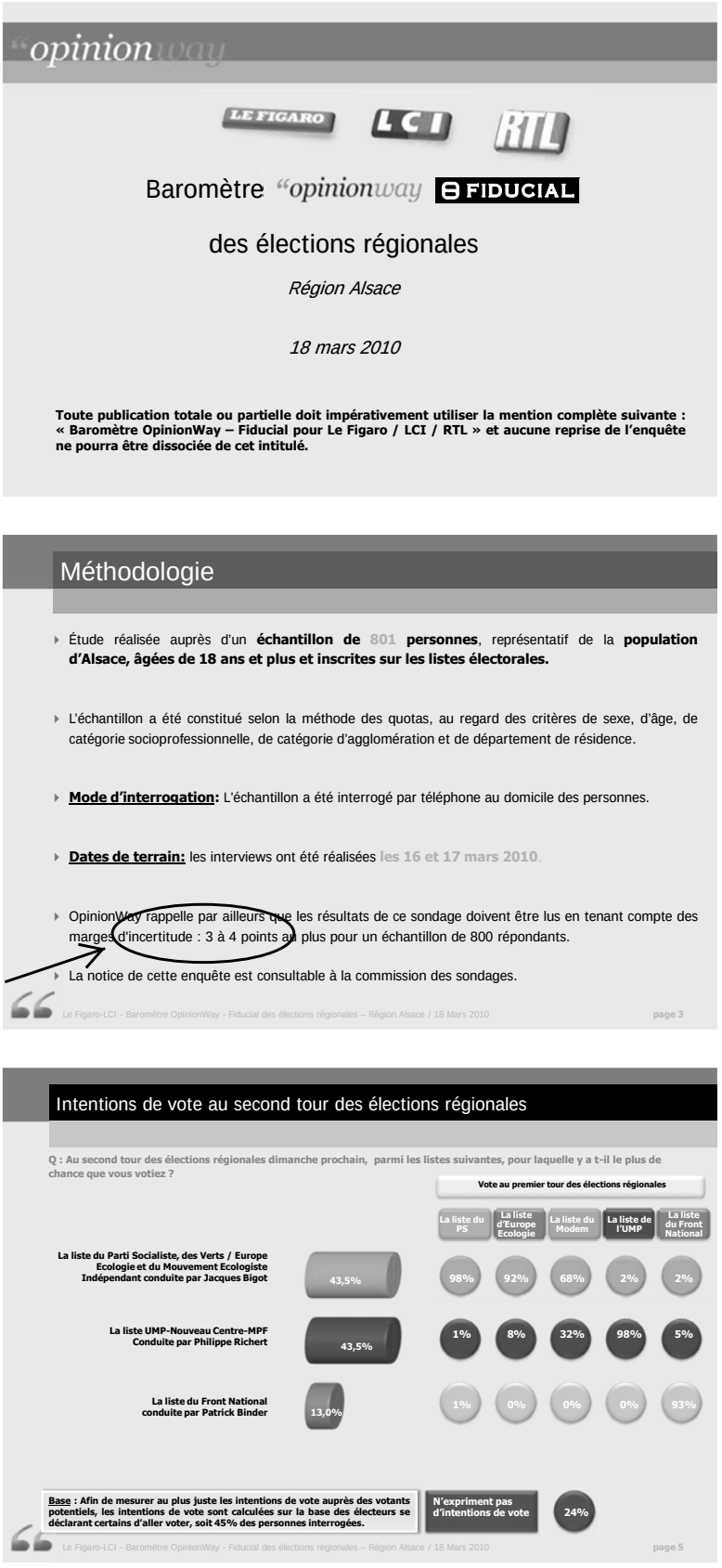


FIGURE 27. Un exemple de résultat de sondage fourni par un institut avec indication de la marge d'erreur (d'incertitude) possible.

EXEMPLE 5.3. On a vu dans la partie 1 que dans un modèle de Bernoulli (de paramètre p_0), l'intervalle

$$\left[\overline{X}_n - \frac{1}{\sqrt{n}} \ , \ \overline{X}_n + \frac{1}{\sqrt{n}}\right]$$

(résultant de l'application du théorème de la limite centrale et d'une majoration de la variance) est un intervalle de confiance asymptotique de niveau 95 % pour le paramètre p_0.

REMARQUE 5.1. Il est beaucoup plus difficile de dire ce qu'est un « bon » intervalle de confiance (alors que pour les estimateurs ponctuels, on avait de nombreux critères de qualité : caractère sans biais, consistance, etc.). Intuitivement, on veut qu'il soit de largeur la plus petite possible tout en garantissant le niveau. Ainsi, si $g(\theta_0)$ est un réel, on peut toujours trouver un intervalle de confiance de niveau $1-\alpha$, en prenant par exemple tout $\mathbb{R}$ (voir l'illustration procurée par la figure 28, pour une proportion p_0). Mais ce n'est pas bien utile de prendre un ensemble aussi gros : on n'exploite pas les données ! Ce qui suit va montrer comment, justement, les exploiter intelligemment.

FIGURE 28. Il est facile de briller à peu de frais... Un peu d'efforts (lisez la suite) et vous saurez faire mieux que ce mauvais détective !

2. Intervalles de confiance asymptotiques sur la moyenne

2.1. Résultat théorique-clé : énoncé et éléments de preuve. On se place dans un modèle où toutes les lois possibles, et en particulier, la vraie loi sous-jacente, admettent un moment d'ordre deux. On rappelle quelques notations issues des parties précédentes. Un estimateur sans biais de la variance $\sigma_0^2 = \sigma^2(\theta_0)$ est défini par

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 \ .$$

THÉORÈME 5.1. *Pour des variables aléatoires* $X_1, \ldots, X_n$ *indépendantes et identiquement distribuées selon une loi admettant un moment d'ordre deux et d'espérance notée* $\mu_0 = \mu(\theta_0)$*, on a*

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\left(\overline{X}_n - \mu_0\right) \rightharpoonup \mathcal{N}(0,1) \ .$$

Nous donnons ci-dessous un éclairage sur ce résultat et des éléments de preuve, uniquement pour la culture. La démonstration du théorème ne sera pas à retenir, mais simplement son énoncé. En revanche, il faudra comprendre ci-dessous la preuve du corollaire 5.1 à partir du résultat du théorème, et même savoir la retrouver.

REMARQUE 5.2 (Eclairage). Le théorème de la limite centrale assure que

$$\sqrt{\frac{n}{\sigma_0^2}}\left(\overline{X}_n - \mu_0\right) \rightharpoonup \mathcal{N}(0,1) \ .$$

Le théorème 5.1 montre que cette convergence est encore vraie lorsque la variance (inconnue) est estimée par l'estimateur de la variance empirique.

ELÉMENTS CULTURELS. La preuve formelle du théorème 5.1 repose sur le lemme suivant, que nous ne démontrerons pas. Si nous voulions le faire, il faudrait que vous ayez d'abord une meilleure caractérisation (équivalente) de la convergence en loi, par exemple via la convergence simple des espérances de fonctions continues bornées des variables aléatoires en jeu, plutôt que via la convergence simple de leurs fonctions de répartition. Vous sentez bien que cela nous prendrait trop de temps, alors je vous propose d'admettre le lemme qui suit.

LEMME 5.1 (de Slutzky). *Soient deux suites de variables aléatoires* (Y_n) *et* (Z_n)*, une variable aléatoire* Y *et une constante* z *telles que les convergences suivantes (respectivement, en loi et en probabilité) ont lieu :*

$$Y_n \rightharpoonup Y \qquad \textit{et} \qquad Z_n \xrightarrow{\mathbb{P}} z \ ;$$

alors on a les convergences en loi

$$Y_n + Z_n \rightharpoonup Y + z \qquad \textit{et} \qquad Y_n Z_n \rightharpoonup zY \ .$$

DÉMONSTRATION (DU THÉORÈME 5.1). Le théorème de la limite centrale s'applique et donne

$$Y_n = \sqrt{\frac{n}{\sigma_0^2}}\left(\overline{X}_n - \mu_0\right) \rightharpoonup Y \sim \mathcal{N}(0,1) \ .$$

Par ailleurs, on a vu dans la partie 4 que

$$\widehat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma_0^2 \ , \qquad \text{d'où} \qquad Z_n = \sqrt{\frac{\sigma_0^2}{\widehat{\sigma}_n^2}} \xrightarrow{\mathbb{P}} 1$$

(on a utilisé ici le passage de la convergence en probabilité aux fonctions continues, voir la proposition 4.1). Le lemme de Slutzky conclut alors la preuve. □

2.2. Conséquence : formule générale des intervalles de confiance asymptotiques sur la moyenne. Le théorème 5.1 admet la conséquence suivante.

COROLLAIRE 5.1. *On note z_β le β–quantile de la loi $\mathcal{N}(0,1)$. Alors les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ pour la moyenne μ_0 :*

$$\left[\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \overset{\text{not.}}{=} \left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] ;$$

$$\left]-\infty,\ \overline{X}_n + z_{1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] ;$$

et

$$\left[\overline{X}_n - z_{1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ +\infty\right[.$$

DÉMONSTRATION. On ne fait le raisonnement que pour un des trois intervalles, les arguments étant similaires pour les deux autres. On a

$$\mu_0 \in \widehat{I}_n \overset{\text{not.}}{=} \left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \iff -z_{1-\alpha/2} \leqslant \sqrt{\frac{n}{\widehat{\sigma}_n^2}}\left(\overline{X}_n - \mu_0\right) \leqslant z_{1-\alpha/2} ,$$

de sorte que par le théorème 5.1,

$$\mathbb{P}\left\{\mu_0 \in \widehat{I}_n\right\} \longrightarrow \mathbb{P}\left\{-z_{1-\alpha/2} \leqslant Z \leqslant z_{1-\alpha/2}\right\} = 1-\alpha ,$$

où Z suit une loi normale $\mathcal{N}(0,1)$. L'égalité de la dernière probabilité à $1-\alpha$ procède des mêmes techniques que celles que nous avons employées à l'exercice 4.4 : la loi normale est symétrique par rapport à 0, ainsi,

$$z_{\alpha/2} = -z_{1-\alpha/2} ;$$

puis, par inclusion d'événements,

$$\mathbb{P}\left\{z_{\alpha/2} \leqslant Z \leqslant z_{1-\alpha/2}\right\} = \mathbb{P}\left\{Z \leqslant z_{1-\alpha/2}\right\} - \mathbb{P}\left\{Z < z_{\alpha/2}\right\} = (1-\alpha/2) - \alpha/2 = 1-\alpha .$$

(Faites un dessin pour mieux suivre le raisonnement ci-dessus ; il sera similaire à la figure 19.) □

REMARQUE 5.3 (De l'importance d'estimer la variance...). D'aucuns pourraient vouloir tirer de la convergence en loi donnée par le théorème de la limite centrale (notée $Y_n \rightharpoonup Y$ dans les éléments culturels de la page précédente) que, par exemple, l'intervalle

$$I_n = \left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_0^2}{n}}\right]$$

est un intervalle de confiance asymptotiquement de niveau 95 %. En fait, le niveau est bien celui annoncé, mais I_n n'est pas un intervalle de confiance ! Il dépend en effet de la variance inconnue σ_0^2. (Remarque : c'est pour cela que je n'ai pas mis de petit chapeau à I_n !) Il faut estimer cette variance inconnue, et c'est en substituant à σ_0^2 son estimateur dans l'expression de I_n que nous pouvons retrouver l'intervalle symétrique correspondant du corollaire 5.1.

REMARQUE 5.4 (Application concrète des formules). Face à des données obtenues comme la réalisation d'un échantillon, et pour encadrer la valeur μ_0 de l'espérance de la loi commune (inconnue), il faut 1. se fixer un niveau de confiance $1-\alpha$ et 2. déterminer la forme de l'intervalle de confiance.

- Pour le point 1., le niveau standard de confiance est de 95 %, sauf indication contraire.
- Pour 2., cela dépend du problème posé et de l'objectif poursuivi : il faut utiliser son bon sens. Le corollaire 5.1 propose trois intervalles de confiance, l'un symétrique autour de la moyenne empirique, et les deux autres majorent et minorent respectivement l'espérance de la loi sous-jacente. Le premier intervalle est dit bilatère, les deuxième et troisième sont unilatères.

EXERCICE 5.1. On veut connaître le budget moyen consacré par les Français pour leurs vacances. On tire 1 500 personnes au hasard dans l'annuaire, et on note les réponses $x_1, \ldots, x_{978}$ des 978 sondés qui acceptent ou sont capables [6] de participer à l'enquête. On obtient un budget moyen de 945 euros par sondé, avec un écart-type de 309 euros. Que peut-on en déduire sur le véritable montant moyen consacré par l'ensemble des Français aux vacances ?

CORRECTION 5.1. On effectue la résolution en supposant que l'on a bien obtenu par ce sondage aléatoire un échantillon représentatif de la population (en négligeant le fait que ce sont peut-être ceux qui ont le plus de temps pour répondre au téléphone qui ont le plus de temps tout court et partent le plus souvent en vacances, ou le fait que ceux qui n'ont pas d'argent pour se payer le téléphone n'en ont sans doute pas non plus pour prendre des congés).

On modélise tout d'abord les données $x_1, \ldots, x_{978} \in \mathbb{R}_+$ comme la réalisation de variables aléatoires $X_1, \ldots, X_{978}$ indépendantes et identiquement distribuées selon une loi admettant un moment d'ordre deux et d'espérance μ_0. Cette espérance forme notre paramètre inconnu : c'est le véritable montant moyen consacré par l'ensemble des Français aux vacances. Ici, les statistiques d'échantillon (les valeurs réalisées de $\overline{X}_{978}$ et $\widehat{\sigma}_{978}$) sont $\overline{x}_{978} = 945$ et $s_{x,978} = 309$.

On peut appliquer le résultat du corollaire 5.1. On n'a pas de raison de ne pas prendre un intervalle bilatère. On prend un niveau de 95 %, à qui correspond le quantile $z_{97.5\,\%} = 1.96$. La taille d'échantillon valant $n = 978$ (et non pas 1 500 !), il vient donc l'intervalle de confiance au niveau approximativement 95 %

$$\left[\overline{X}_{978} \pm z_{97.5\,\%}\sqrt{\frac{\widehat{\sigma}^2_{978}}{978}}\right] ,$$

qui admet pour réalisation

$$\left[945 \pm 1.96\sqrt{\frac{309^2}{978}}\right] = [945 \pm 20] = [925,\ 965] .$$

(Remarquez que vu la taille de l'intervalle, il serait ridicule de préciser des centimes. On arrondit donc vers les extrémités : l'écart de ± 19.37 euros passe donc plutôt à ± 20 euros.)

On verra plus loin des exemples de situations où le recours à des intervalles unilatères est naturel.

2.3. Cas particulier : estimation par intervalle d'une proportion. On se place ici dans le cas d'un modèle de Bernoulli, dans lequel on sait que $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi $\mathcal{B}(p_0)$ appartenant à la famille

6. Pas facile de pouvoir calculer précisément l'ensemble de ses dépenses pour les vacances et week-ends, surtout quand on part souvent !

$\{\mathcal{B}(p),\ p \in [0,1]\}$. Le vrai paramètre p_0 est inconnu et on cherche à l'estimer par un intervalle.

On a vu à l'exemple 4.6 que dans le cas d'un modèle de Bernoulli, l'estimateur biaisé de la variance prenait une expression simple,

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \overline{X}_n(1-\overline{X}_n)\ .$$

Or, les éléments culturels ci-dessus montrent que la seule propriété que l'on utilise de l'estimateur de la variance est son caractère consistant. Ainsi, dans le cas d'une proportion, on tend à remplacer $\widehat{\sigma}_n^2$ par $\overline{X}_n(1-\overline{X}_n)$, qui a une expression plus simple à calculer et plus agréable à l'œil. Comme $\mu_0 = p_0$, il vient

$$\sqrt{\frac{n}{\overline{X}_n(1-\overline{X}_n)}}\left(\overline{X}_n - p_0\right) \ \rightharpoonup\ \mathcal{N}(0,1)$$

et les mêmes techniques que dans la preuve du corollaire 5.1 permettent de tirer des intervalles de confiance asymptotiques à partir de cette convergence en loi.

Corollaire 5.2. *On note z_β le β–quantile de la loi $\mathcal{N}(0,1)$. Alors les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ pour la proportion p_0 :*

$$\left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]\ ;$$

$$\left[0,\quad \overline{X}_n + z_{1-\alpha}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right]\ ;$$

$$\textit{et}\qquad \left[\overline{X}_n - z_{1-\alpha}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},\quad 1\right]\ .$$

Exercice 5.2. Une banque se demande si elle n'aurait pas accordé trop de prêts immobiliers à des clients pas assez solvables. Pour se rassurer, elle voudrait avoir une majoration du taux de défaillance des prêts en cours et imagine la procédure suivante : sélectionner des clients ayant un prêt en cours, regarder le montant de leur échéance, leur salaire moyen (puisqu'il est domicilié à la banque, c'était une condition nécessaire pour l'obtention du prêt), la valeur actuelle de leur bien immobilier (moins une décote de 20 % pour se prémunir contre une mévente), et déterminer si le client pourra continuer à rembourser son prêt jusqu'à l'échéance, et sinon, le cas échéant, si saisir son bien en cours de prêt et le revendre couvrirait la dette.

Elle tire alors au hasard des clients dans son portefeuille et fait analyser leur situation. L'analyse durant une heure environ par client et deux analystes étant mobilisés pendant deux semaines, 140 cas peuvent être traités. (On voit ici qu'il serait déraisonnablement coûteux d'étudier l'ensemble des milliers de prêts en cours !) Sur ces 140 cas, seuls 6 présentent un risque sérieux de défaillance. Quel est alors son verdict quant au taux maximum de défaillance attendu ?

Correction 5.2. On note $x_1, \ldots, x_{140}$ les résultats de l'analyse : x_j vaut 0 si la dette du j–ème client est solvable et 1 s'il est atteint par un risque sérieux de défaillance. Les clients étant tirés au hasard dans le fichier de la banque, qui en comporte beaucoup, on

peut bien modéliser le résultat de l'expérience comme la réalisation des variables aléatoires $X_1, \ldots, X_{140}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0. Le paramètre inconnu p_0 de cette loi est justement le taux de défaillance de l'ensemble des prêts en cours. Il s'agit ici de le majorer pour se rassurer : on propose pour cela un intervalle de confiance unilatère ; par ailleurs, on va prendre un niveau de confiance élevé, par exemple à 99 %.

On vient de voir qu'avec probabilité tendant vers 99 % lorsque $n \to \infty$,

$$p_0 \in \left[0,\ \overline{X}_n + z_{99\,\%}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right] .$$

Le calcul pratique, avec $n = 140$, $\overline{x}_{140} = 6/140$ et $z_{99\,\%} = 2.326$, donne la réalisation $[0,\ 8.3\,\%]$ de l'intervalle théorique rappelé. (Là encore, on évite de mettre trop de chiffres après la virgule !)

La conclusion (stratégique, à rapporter aux instances dirigeantes) est que la majoration du taux de défaillance attendue est de 8.3 %. Par ailleurs, on peut préciser que l'on a obtenu cette majoration comme la réalisation d'un intervalle de confiance de niveau élevé : 99 %.

Remarque 5.5 (Attention à l'interprétation !). N'oubliez jamais la remarque 1.5 formulée dans la première partie : il y a une probabilité 0 ou 1 que le vrai taux p_0 (tel que mesuré des années plus tard, à la liquidation de tous les prêts en cours) soit dans l'intervalle $[0,\ 8.3\,\%]$. En revanche, il y a une probabilité (environ [7]) $0.99 = 99\,\%$ que p_0 soit dans l'intervalle théorique

$$\left[0,\ \overline{X}_{140} + z_{99\,\%}\sqrt{\frac{\overline{X}_{140}(1-\overline{X}_{140})}{140}}\right] .$$

Il faut bien distinguer les variables aléatoires utilisées pour la modélisation et les formules théoriques de leurs réalisations.

Remarque 5.6 (Marge d'erreur des sondages). Pour un sondage d'opinion, on interroge typiquement 1 000 personnes. Quelle est la marge d'erreur que l'on obtient sur le résultat ? Vous avez peut-être déjà entendu qu'elle était autour de 3 % : retrouvons [8] ce résultat par l'analyse mathématique. A un niveau fixé de 95 % (où $z_{97.5\,\%} = 1.96$), la demi-largeur du premier intervalle de confiance du corollaire 5.2 est alors donnée, dans le pire des cas, par

$$\max_{x\in[0,1]} 1.96\sqrt{\frac{x(1-x)}{1\,000}} \approx \frac{1}{\sqrt{1\,000}} \approx 3\,\% .$$

On a utilisé ici que $x \in [0,1] \mapsto x(1-x)$ admet $1/4$ pour maximum.

3. Planification de sondages

Jusqu'à présent, on partait des observations, on se fixait un niveau de confiance $1-\alpha$, et on en déduisait un intervalle de confiance. La taille de ce dernier dépendait de n, α, et de l'estimée de la variance ; lorsque ce dernier était bilatère, sa demi-largeur indiquait la précision ε de l'estimation.

7. La probabilité est 99 % à la limite quand $n \to \infty$; ici, n est suffisamment grand, $n = 140$, de sorte que l'on est proche de 99 %.

8. Nous l'avions déjà fait au premier cours, mais comme la pédagogie est l'art de la répétition...

On se pose désormais le problème inverse. Si, avant de commencer l'expérience statistique, on se fixe une précision ε et un niveau de confiance α, combien d'éléments n recueillir dans l'échantillon ? La banque de l'exercice 5.2 ne s'était pas posé ce problème, elle avait recueilli autant d'observations qu'elle le pouvait, mais dans d'autres cas, et notamment ceux où il est encore plus coûteux d'en obtenir, il vaut mieux réfléchir à leur nombre à l'avance.

Cette situation est illustrée par la figure 29.

3.1. Planification : cas de l'estimation d'une proportion. On reprend les notations du paragraphe 2.3. Dans ce cas, on a vu que les intervalles de confiance (bilatères) sont de la forme

$$\widehat{I}_n = \left[\overline{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \right] .$$

La précision dépend donc des observations à venir via $\overline{X}_n$! On va majorer cette précision de la même manière qu'à la remarque 5.6 et on choisira n tel que ce majorant soit plus petit que ε. Toujours en utilisant que $x(1-x) \leqslant 1/4$, il vient

$$\widehat{I}_n \subset \widehat{J}_n \overset{\text{not.}}{=} \left[\overline{X}_n \pm z_{1-\alpha/2} \frac{1}{2\sqrt{n}} \right] .$$

$\widehat{J}_n$ est également un intervalle de confiance, asymptotiquement de niveau $1-\alpha$; il est plus large que $\widehat{I}_n$, mais sa précision (sa demi-largeur) est indépendante des observations. Il suffit de prendre, à α et ε fixés, n suffisamment grand pour que

$$z_{1-\alpha/2} \frac{1}{2\sqrt{n}} \leqslant \varepsilon , \qquad \text{soit} \qquad n \geqslant \left(\frac{z_{1-\alpha/2}}{2\varepsilon} \right)^2 .$$

En particulier, lorsque $\alpha = 5\,\%$, il s'agit de prendre $n \geqslant 1/\varepsilon^2$.

Remarque 5.7 (Planification vs. exploitation d'une expérience). Evidemment, après avoir conduit l'expérience, on peut délaisser l'intervalle $\widehat{J}_n$ pour $\widehat{I}_n$: la précision sera alors d'autant meilleure que $\overline{X}_n$ est éloigné de $1/2$. On retiendra pour la suite que pour planifier, on considère des intervalles du type de $\widehat{J}_n$, mais que si l'on dispose des données, on se tourne vers ceux du type de $\widehat{I}_n$. L'exercice suivant illustre cela.

Exercice 5.3 (Les communes qui veulent avoir des raisons de ne pas investir). Des élus de l'opposition requièrent la construction d'un nouvel équipement public, par exemple, un cinéma. La majorité diligente une étude de faisabilité : elle veut tout d'abord déterminer le taux de fréquentation attendu et commandite pour ce faire un sondage. Elle aimerait une estimation de ce taux à $\pm 2\,\%$: de combien de foyers n les services municipaux doivent-ils récupérer une opinion ? (Attention, cela suppose évidemment un nombre plus grand n' de coups de fil !) Le directeur des services aurait besoin de ce nombre pour établir le planning de travail jusqu'au prochain conseil municipal.

Trois semaines plus tard, le sondage a été effectué. On suppose que la moyenne des promesses de fréquentation au moins occasionnelle se situe à $14.2\,\%$. Que peut-on dire de la moyene qu'on aurait eue si on avait pu interroger tous les habitants un par un ?

Correction 5.3. La précision est de $\varepsilon = 0.02$ et si l'on conserve le niveau asymptotique habituel de $95\,\%$, il s'agit de récupérer

$$n \geqslant \frac{1}{\varepsilon^2} = \frac{1}{0.02^2} = 2\,500$$

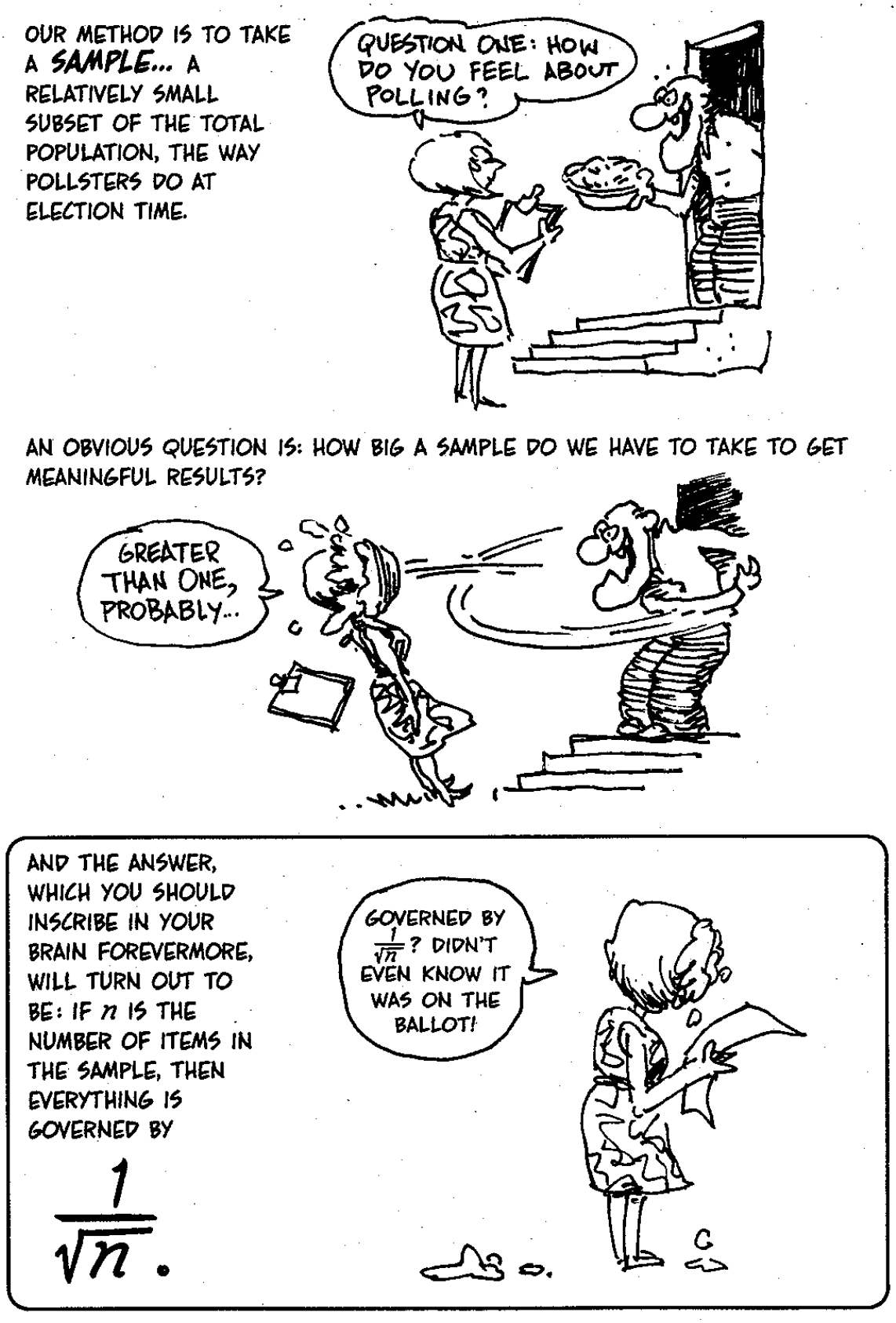


FIGURE 29. Planification de sondages et mauvais jeux de mots associés.

réponses, selon les calculs précédents. Evidemment, cela obligera à interroger autour de $n' = 3\,000$ habitants (foyers). Pour une petite ville, de l'ordre de 20 000 habitants, cela est beaucoup et fausse l'approximation habituelle de tirage avec remise pour obtenir la modélisation en termes de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli. Mais pour une ville de taille plus raisonnable (plus de 50 000 habitants), ce chiffre est convenable. Dans tous les cas, le sondage prendra du temps et sera coûteux, parce que l'échantillon demandé est de grande taille (plus que la taille typique de 1 000 sondés).

On suppose ensuite qu'on a réalisé le sondage et obtenu 2 500 réponses, dont 355 ($= 2\,500 \times 14.2\,\%$) positives, soit des données $x_1, \ldots, x_{2500} \in \{0, 1\}$, telles que

$$\overline{x}_{2500} = \frac{355}{2\,500} = 14.2\,\% \ .$$

On les modélise comme la réalisation de $X_1, \ldots, X_{2500}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0. Ce dernier s'interprète comme le vrai taux de fréquentation si le cinéma était construit. L'intervalle de confiance pour p_0, approximativement à 95 %, est donné par

$$\left[\overline{X}_{2500} \pm 1.96\sqrt{\frac{\overline{X}_{2500}(1-\overline{X}_{2500})}{2500}}\right]$$

et sa réalisation sur les données vaut

$$\left[14.2\,\% \pm 1.96\sqrt{\frac{\overline{x}_{2500}(1-\overline{x}_{2500})}{2\,500}}\right] = \left[14.2\,\% \pm 1.96\sqrt{\frac{0.142(1-0.142)}{2\,500}}\right]$$
$$= [14.2\,\% \pm 1.4\,\%] \ .$$

On s'attend à ce que la proportion p_0 de fréquentations futures déclarées qu'on aurait obtenue si on avait pu interroger tous les habitants un par un soit dans cet intervalle, car c'est la réalisation d'un intervalle de confiance à 95 %.

EXERCICE 5.4 (Les communes qui veulent investir à raison). Dans une autre ville, la majorité municipale veut construire un nouvel équipement public, et sait à l'avance (grâce à l'expérience de villes voisines) que pour qu'il soit rentable, il faut qu'au moins 10 % de la population le fréquente. Peut-on ici faire mieux et interroger moins de gens qu'à l'exercice précédent, étant donné l'objectif chiffré ?

CORRECTION 5.4 (Attention, exercice difficile pouvant être sauté !). La solution analytique générale est laissée à votre sagacité : voici seulement une solution particulière. L'affirmation que nous allons prouver est la suivante :

> Avec 1 000 réponses au sondage seulement, on peut garantir une précision de 1.1 % pour la zone de résultats la plus cruciale, tandis que pour les autres résultats, la conclusion sera clairement que l'équipement sera rentable.

En effet, avec des calculs et notations similaires à ceux de l'exercice précédent (si ce n'est que l'on ne dispose que de données $x_1, \ldots, x_{1000}$ ici), on distingue alors deux cas :

- si la moyenne d'échantillon $\overline{x}_{1000}$ est plus grande que 13 %, alors l'intervalle de confiance qui sera construit sur elle (à 95 %) sera de demi-largeur inférieure à 3 % (cf. remarque 5.6) et donc ne contiendra pas le seuil de 10 % ; on conclura à la rentabilité ;

– sinon, on a $\overline{x}_{1000} \leqslant 13\,\%$ et en particulier, $x \mapsto x(1-x)$ étant croissante sur $[0,\, 1/2]$, la précision est majorée par

$$1.96\sqrt{\frac{\overline{x}_{1000}(1-\overline{x}_{1000})}{1\,000}} \leqslant 1.96\sqrt{\frac{0.13\,(1-0.13)}{1\,000}} \approx 1.1\,\%\ .$$

En conclusion, par avance, on sait, avant même de réaliser le sondage, qu'il n'est pas nécessaire d'interroger plus de 1 000 personnes.

3.2. Planification : cas général. Ici, contrairement au cas des fréquences, on n'a pas de majoration a priori de la variance. Dans le cas précédent, le terme de variance dans la précision de l'intervalle était toujours borné par $1/4$, grâce à l'inégalité $x(1-x) \leqslant 1/4$, et on pouvait majorer par avance la précision indépendamment des observations,

$$z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \leqslant z_{1-\alpha/2}\,\frac{1}{2\sqrt{n}}\ .$$

Dans le cas général, il s'agirait de majorer, à l'avance, avant de conduire l'expérience,

$$z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\ ,$$

or, on n'a aucune idée de la valeur qui va être réalisée pour $\widehat{\sigma}_n^2$! Pas même son ordre de grandeur. Il faut ici réaliser une première estimation sur un petit échantillon de taille t (par exemple, $t \approx 30$), en déduire une estimée s_t de l'écart-type et résoudre en n, à α et ε fixés,

$$z_{1-\alpha/2}\sqrt{\frac{s_t^2}{n}} \leqslant \varepsilon\ , \qquad \text{soit} \quad n \geqslant s_t^2\left(\frac{z_{1-\alpha/2}}{\varepsilon}\right)^2\ .$$

n donne alors la taille du nouvel échantillon de données à recueillir, notées $x_{t+1}, \ldots, x_{t+n}$. On exploite alors ces n nouvelles données uniquement ou l'ensemble des $t+n$, selon qu'il y a indépendance ou non entre le nouvel et l'ancien échantillon. Remarquez bien qu'il n'est pas tout à fait garanti que la nouvelle estimée de la variance s_n (construite sur les n nouvelles données) ou s_{t+n} (construite sur les données complètes) soit proche de l'estimée préliminaire s_t (construite sur les t premiers coups de sonde). Cela l'est cependant si t avait déjà été pris suffisamment grand, par consistance de l'estimateur de la variance : c'est pourquoi on recommande $t \approx 30$. De toute façon ici, on fait ce qu'on peut et il faut s'en satisfaire...

Exercice 5.5. Les gérants d'un marché couvert se demandent quel est le montant moyen des achats par client ; ce sera une partie de leur argumentation montrant aux commerçants installés dans la halle que le montant de la redevance à verser est tout à fait raisonnable, vu le chiffre d'achat. Chaque client volontaire présente aux enquêteurs l'ensemble de ses tickets et ils en font la somme ; ce n'est pas évident, certains tickets se perdent dans les poussettes de marché, les allées sont étroites et le flux incessant des clients ne se prête pas à de longs palabres. Alors, pour planifier les choses le week-end suivant, au bout de 30 clients (ayant dépensé en moyenne 47.32 euros, avec un écart-type sur les données de 12.05 euros) inconfortablement interrogés, le gérant se demande quel nombre n de clients il lui faudrait au total pour avoir une précision de ce montant moyen à ± 1 euro près. Il réalise alors le sondage de $n-30$ clients supplémentaires en dégageant un espace plus calme dans le marché et en embauchant sept enquêteurs pendant une demie-journée, et obtient une moyenne (y compris les 30 premiers interrogés) de 48.51 euros,

avec un écart-type observé de 11.37 euros. Modéliser la situation en donnant le paramètre d'intérêt, ainsi que les estimations qui lui sont fournies par chacun des deux sondages successifs.

CORRECTION 5.5. On se place dans toute la correction au niveau de confiance habituel de 95 %. On dispose des données $x_1, \dots, x_{30}$ d'une part, puis $x_{31}, \dots, x_n$ d'autre part, à valeurs dans $\mathbb{R}_+$. Vu l'interrogation au hasard, on peut modéliser ces données comme la réalisation des variables aléatoires indépendantes et identiquement distribuées $X_1, \dots, X_n$. On note μ_0 l'espérance de cette loi commune : c'est le montant moyen des achats par client du marché, et il forme notre paramètre d'intérêt. Le premier sondage, sur 30 clients, fournit la réalisation suivante de l'intervalle de confiance symétrique général :

$$\left[\overline{x}_{30} \pm z_{97.5\,\%}\sqrt{\frac{s^2_{x,30}}{30}}\right] = \left[47.32 \pm 1.96\,\frac{12.05}{\sqrt{30}}\right] = [47.32 \pm 4.31] \approx [43.0,\, 51.7]\ .$$

La précision que l'on s'attend à obtenir avec n données au total est donnée, vu l'estimée de l'écart-type, par

$$1.96\,\frac{12.05}{\sqrt{n}}$$

et on veut qu'elle soit plus petite que 1 euro. On choisit donc $n \geqslant (1.96 \times 12.05)^2$, par exemple $n = 558$. (Ce qui signifie que chacun des sept enquêteurs devra interroger environ 75 clients : c'est beaucoup !)

On obtient alors, après réalisation du sondage complet, l'intervalle de confiance

$$\left[\overline{x}_{558} \pm z_{97.5\,\%}\sqrt{\frac{s^2_{x,558}}{558}}\right] = \left[48.51 \pm 1.96\,\frac{11.37}{\sqrt{558}}\right] = [48.51 \pm 0.94] \approx [47.5,\, 49.5]\ .$$

La précision est effectivement celle que l'on attendait (elle est même un peu meilleure parce que l'estimée de la variance a un peu diminué ; mais cela aurait très bien pu être le contraire...).

4. Comment estimer la moyenne lorsque la taille d'échantillon est petite ?

Les résultats précédents sont asymptotiques et reposent sur le théorème de la limite centrale. On n'en croit les résultats en pratique que lorsque la taille d'échantillon n est suffisamment grande, disons $n \geqslant 30$. Lorsque ce n'est pas le cas, on a une solution, mais uniquement dans le cas du modèle gaussien (qui est cependant très répandu, comme nous l'avons vu au second cours ; en particulier, cela ne vaut pas dans le cas d'un modèle de Bernoulli, où il s'agit d'estimer une proportion ou une fréquence).

Cette solution repose sur la loi dite de Student [9], que l'on introduit maintenant. Lorsque les observations $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2, alors $\overline{X}_n \sim \mathcal{N}(\mu_0, \sigma_0^2/n)$ et donc

$$\frac{\overline{X}_n - \mu_0}{\sqrt{\sigma_0^2/n}} = \sqrt{\frac{n}{\sigma_0^2}}\,(\overline{X}_n - \mu_0) \ \sim\ \mathcal{N}(0,1)\ .$$

9. Vous aimerez, j'en suis sûr, le contexte dans lequel cette dernière a été trouvée : William Sealey Gosset (1876–1937) était chimiste à la brasserie Guinness à Dublin, puis ensuite à Londres. C'est pour le contrôle de qualité qu'il a été conduit à s'intéresser à l'échantillonnage, et surtout aux petits échantillons. Il a publié ses travaux sous le pseudonyme de Student. C'est lui qui a mis en évidence la loi dont il est question ici.

(A gauche, il s'agit de la même variable aléatoire que celle considérée au théorème de la limite centrale.) On ne peut déduire de cette assertion un intervalle de confiance sur μ_0, car il faut encore se débarrasser de la dépendance en σ_0^2. On l'estime ici encore. Et justement, estimer la variance σ_0^2 par l'estimateur non biaisé $\widehat{\sigma}_n^2$ conduit à la loi de Student.

DÉFINITION–THÉORÈME 5.1. *Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2. Alors la loi de*

$$\frac{\overline{X}_n - \mu_0}{\sqrt{\widehat{\sigma}_n^2/n}} = \sqrt{\frac{n}{\widehat{\sigma}_n^2}}\,(\overline{X}_n - \mu_0)$$

est indépendante de μ_0 et σ_0^2 ; on l'appelle la loi de Student à $n-1$ degrés de liberté, et on la note $\mathcal{T}_{n-1}$.

DÉMONSTRATION. (N'est donnée que pour la culture ; peut être sautée.) Il s'agit de voir pourquoi la loi de la variable aléatoire considérée ne dépend ni de μ_0 ni de σ_0^2. On introduit les variables aléatoires (non observées, mais cela n'a pas d'importance)

$$X_1' = \frac{X_1 - \mu_0}{\sigma_0}, \ \ldots, \ X_n' = \frac{X_n - \mu_0}{\sigma_0} \ ;$$

elles sont indépendantes et identiquement distribuées selon la loi $\mathcal{N}(0,1)$. On note

$$\overline{X'}_n \qquad \text{et} \qquad \widehat{\sigma'}^2_n$$

les estimateurs de la moyenne et de la variance empiriques construits sur les X_j' et on note, par un calcul immédiat, que

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2}}\,(\overline{X}_n - \mu_0) = \sqrt{\frac{n}{\widehat{\sigma'}^2_n}}\,\overline{X'}_n \ ,$$

ce qui conclut la preuve. □

COROLLAIRE 5.3. *On note $t_{n-1,\beta}$ le β–quantile de la loi $\mathcal{T}_{n-1}$. Alors, sous les hypothèses de la définition–théorème précédente, les intervalles suivants sont des intervalles de confiance de la moyenne μ_0, non asymptotiques et exactement de niveau $1-\alpha$,*

$$\left[\overline{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ \overline{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \overset{\text{not.}}{=} \left[\overline{X}_n \pm t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \ ;$$

$$\left]-\infty,\ \overline{X}_n + t_{n-1,1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \ ;$$

et

$$\left[\overline{X}_n - t_{n-1,1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ +\infty\right[\ .$$

La preuve de ces résultats est totalement similaire à ceux du corollaire 5.1, au remplacement des lois normales et de leurs quantiles par ceux de la loi de Student.

REMARQUE 5.8. Les intervalles proposés sont non asymptotiques, dans ce cadre de modèle gaussien : c'est une amélioration considérable par rapport au modèle général. Ce n'est cependant pas la panacée, parce qu'en pratique, bien malin qui a affaire à des données exactement gausiennes... En revanche, la distribution des valeurs observées est

souvent raisonnablement proche d'une loi normale, de sorte que les intervalles de confiance ci-dessus sont d'un niveau raisonnablement proche de $1-\alpha$.

REMARQUE 5.9. On peut voir (par exemple sur les tables) que, à $\beta \geqslant 0.5$ fixé, $t_{k,\beta}$ est une suite décroissante avec k, de limite z_β. Les intervalles de confiance construits à partir de la loi de Student sont donc plus gros que ceux construits sur la loi gaussienne. Cela se comprend bien intuitivement : si la taille d'échantillon est petite, il y a des chances pour que l'estimation de la variance ne soit pas tout à fait bonne, et on compense cela en augmentant la taille de l'intervalle de confiance. On retrouve ces considérations sur la figure 30, qui présente des tracés des fonctions densité et de répartition de lois de Student, de différents degrés de liberté, ainsi que leur limite, la loi normale standard (en fait, $\mathcal{T}_k$ est quasiment indistinguable de la loi $\mathcal{N}(0,1)$ lorsque $k \geqslant 30$). On dit que la loi de Student admet des queues plus épaisses que la queue de la loi normale. Par précaution, SPSS utilise les quantiles des lois de Student quelle que soit la valeur de la taille d'échantillon n.

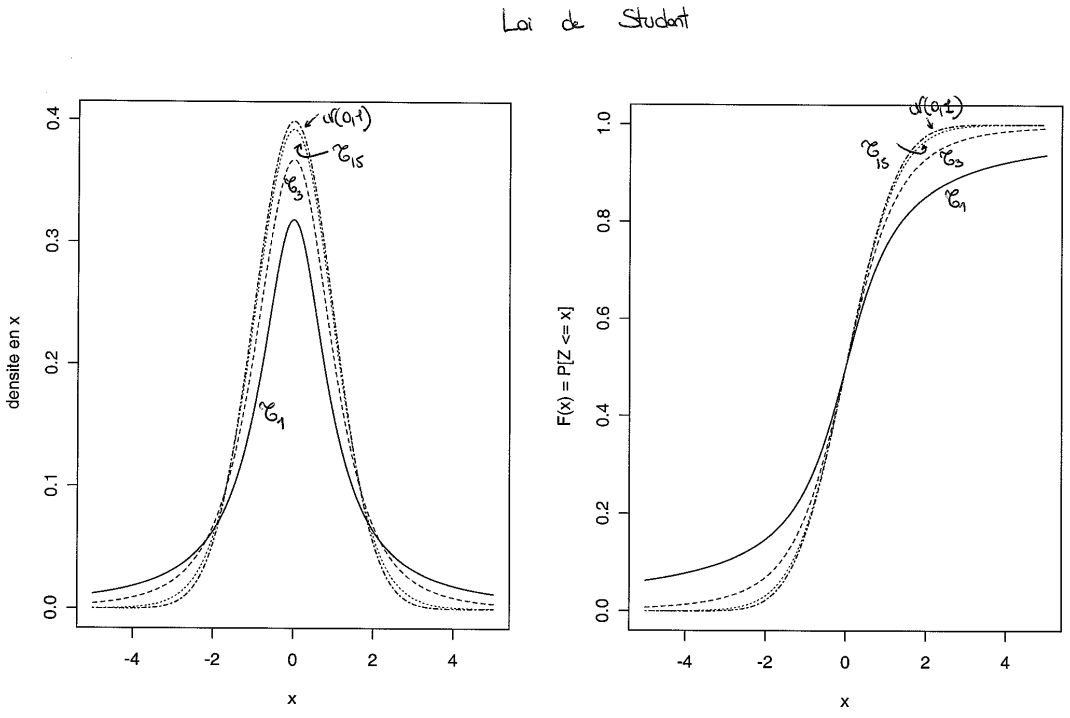


FIGURE 30. Densités (à gauche) et fonctions de répartitions (à droite) des lois de Student, pour différents degrés de liberté, et pour leur loi limite, la loi normale standard $\mathcal{N}(0,1)$.

5. Intervalles de confiance simultanés

Ici, il s'agit d'estimer par intervalle deux paramètres à la fois, disons μ_0 et μ_0'. On propose deux intervalles $\widehat{I}$ et $\widehat{J}$ tels que, par exemple,

$$\mathbb{P}\{\mu_0 \in \widehat{I}\} = 1-\alpha \qquad \text{et} \qquad \mathbb{P}\{\mu_0' \in \widehat{J}\} = 1-\beta$$

et on se demande ce que l'on peut garantir sur

$$\mathbb{P}\left\{\mu_0 \in \widehat{I} \ \text{ et } \ \mu_0' \in \widehat{J}\right\}.$$

Deux situations se présentent, selon qu'il y a indépendance ou non. La première est la moins fréquente, mais c'est celle que vous suggèrent vos anciens réflexes de calcul devant l'évaluation d'une intersection d'événements.

L'intérêt est alors qu'on a un intervalle de confiance sur toute quantité de la forme $g(\mu_0, \mu_0')$: il est donné par

$$\left[\min_{\widehat{I}\times\widehat{J}} g, \quad \max_{\widehat{I}\times\widehat{J}} g\right].$$

La plupart du temps, $g(x, y) = xy$ ou $g(x, y) = x - y$.

5.1. Expériences indépendantes. On dispose de données $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$ obtenues indépendamment. On suppose qu'on peut les modéliser comme la réalisation, d'une part, des variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ (à qui correspond le paramètre μ_0), et d'autre part, des variables aléatoires indépendantes et identiquement distribuées $Y_1, \ldots, Y_m$ (à qui correspond le paramètre μ_0'). De plus, les $X_1, \ldots, X_n$ sont indépendantes des $Y_1, \ldots, Y_m$.

Dans ce cas, $\widehat{I}$ étant construit sur les $X_1, \ldots, X_n$ uniquement, et $\widehat{J}$ sur les $Y_1, \ldots, Y_m$, on obtient bien, par indépendance,

$$\mathbb{P}\left\{\mu_0 \in \widehat{I} \text{ et } \mu_0' \in \widehat{J}\right\} = \mathbb{P}\{\mu_0 \in \widehat{I}\}\ \mathbb{P}\{\mu_0' \in \widehat{J}\} = (1-\alpha)(1-\beta) = 1 - (\alpha+\beta) + \alpha\beta\ .$$

Exercice 5.6 (Les infirmières). Une étude plus fine des salaires des infirmières américaines semble montrer que les salaires dépendent du type de poste (à l'hôpital ou en cabinet). Parmi les 2 911 salaires horaires validement recueillis, 1 945 concernent des infirmières travaillant à l'hôpital : on les note $x_1, \ldots, x_{1945}$; et 966 des infirmières en cabinet : on note leurs salaires $y_1, \ldots, y_{966}$. Avec SPPS (`Analyse / Statistiques descriptives / Explorer`), on construit des intervalles de confiance (de niveau asymptotique 99.5 %) sur les salaires moyens μ_0 et μ_0' de chacun des deux types de poste (voir figure 31) :

$$[20.45,\ 20.90] \qquad \text{et} \qquad [18.27,\ 19.11]\ .$$

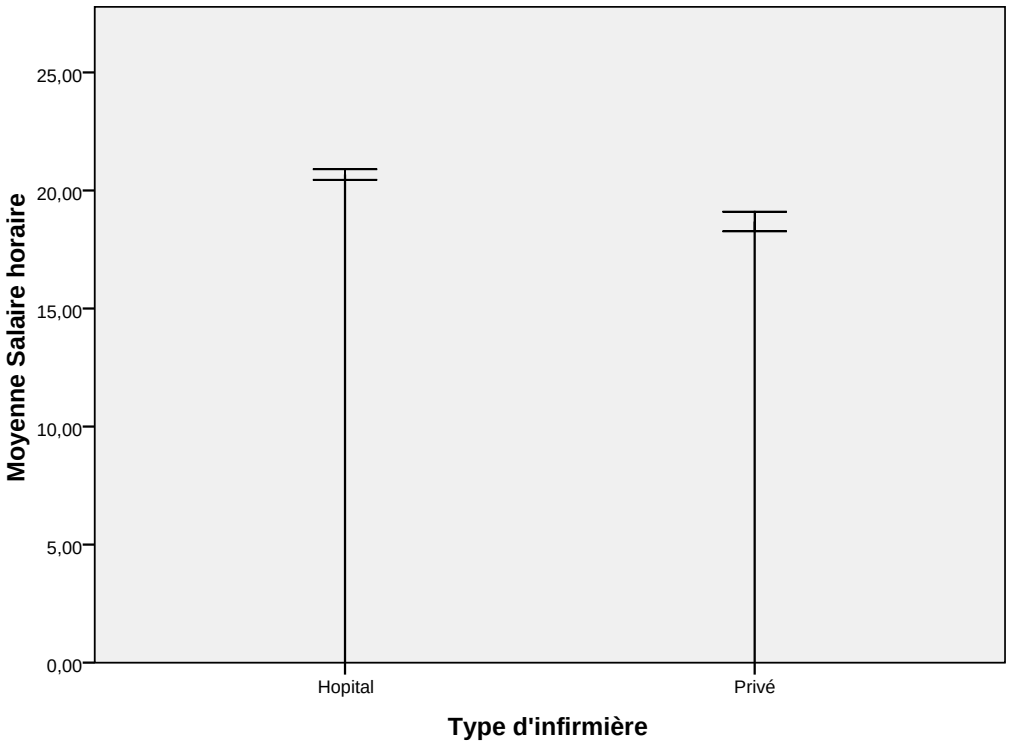
On se pose alors les questions suivantes :

1. En déduire (la réalisation d')un intervalle de confiance de niveau à préciser sur la différence de salaire $\mu_0 - \mu_0'$.
2. Que conclure ? Peut-on dire que les salaires sont significativement différents ?
3. Obtenir la représentation graphique de la figure 31 (utiliser les commandes `Graphes / Générateur de diagrammes`).

Correction 5.6. Ici, on est bien dans le cadre de deux expériences indépendantes, car on semble s'être fixé par avance la taille de chacun des deux sous-échantillons à interroger : 2 000 infirmières travaillant à l'hôpital et 1 000 infirmières travaillant en clinique, et l'on a sans nul doute mené les deux sondages de manière indépendante, en tirant des infirmières au hasard dans les deux sous-listes. Avec des notations désormais bien habituelles, on obtient alors que

$$\widehat{I} = \left[\overline{X}_{1945} \pm z_{99.75\,\%}\sqrt{\frac{\widehat{\sigma}^2_{X,1945}}{1945}}\right] \qquad \text{et} \qquad \widehat{J} = \left[\overline{Y}_{966} \pm z_{99.75\,\%}\sqrt{\frac{\widehat{\sigma}^2_{Y,966}}{966}}\right]$$

sont, respectivement pour μ_0 et μ_0', des intervalles de confiance de niveau approximativement égal à 99.5 %. L'énoncé a donné leurs réalisations respectives : [20.45, 20.90] et



Bâtons de variation : 99,5% IC

Récapitulatif du traitement des observations

		Observations					
		Valide		Manquante		Total	
	Type d'infirmière	N	Pourcent	N	Pourcent	N	Pourcent
Salaire horaire	Hopital	1945	97,3%	55	2,8%	2000	100,0%
	Privé	966	96,6%	34	3,4%	1000	100,0%

Descriptives

	Type d'infirmière			Statistique	Erreur standard
Salaire horaire	Hopital		Moyenne	20,6764	,07927
		Intervalle de confiance à 99.5% pour la moyenne	Borne inférieure	20,4536	
			Borne supérieure	20,8991	
			Ecart-type	3,49582	
	Privé		Moyenne	18,6859	,14763
		Intervalle de confiance à 99.5% pour la moyenne	Borne inférieure	18,2706	
			Borne supérieure	19,1013	
			Ecart-type	4,58852	

FIGURE 31. Deux intervalles de confiance à 99.5 % sur le salaire horaire moyen des infirmières selon qu'elles travaillent à l'hôpital ou en cabinet. La colonne "Erreur standard" propose la valeur de l'estimée de la variance divisée par la racine carrée de la taille d'échantillon : $s_n/\sqrt{n}$. Par exemple, $3.49582/\sqrt{1945} = 0.7927$.

[18.27, 19.11]. On a simultanément $\mu_0 \in \widehat{I}$ et $\mu_0' \in \widehat{J}$ avec probabilité approximativement égale à

$$(1-\alpha)^2 = 99\,\% + 0.0025\,\% \approx 99\,\% \,, \qquad \text{où} \quad \alpha = 0.5\,\% \,.$$

En particulier, c'est le niveau au moins garanti pour l'intervalle de confiance

$$\widehat{I} - \widehat{J} = \left[\overline{X}_{1945} - \overline{Y}_{966} \pm \left(z_{99.75\,\%} \sqrt{\frac{\widehat{\sigma}^2_{X,1945}}{1945}} + z_{99.75\,\%} \sqrt{\frac{\widehat{\sigma}^2_{Y,966}}{966}} \right) \right]$$

sur $\mu_0 - \mu_0'$. Ce dernier admet pour réalisation (attention à bien croiser !) :

$$[20.45 - 19.11,\ 20.90 - 18.27] = [1.34,\ 2.63] \,.$$

Comme la valeur 0 n'appartient pas à la réalisation de cet intervalle de confiance de niveau 99 % sur la différence de salaires $\mu_0 - \mu_0'$, on en conclut que de manière très significative statistiquement, les salaires moyens sont différents selon le type de poste occupé.

5.2. Utilisation des mêmes données pour les deux intervalles, donc absence d'indépendance : la méthode de Bonferroni. Dans le cas général, il n'y a pas d'indépendance entre les événements $\mu_0 \in \widehat{I}$ et $\mu_0' \in \widehat{J}$, souvent parce qu'ils sont construits à partir des mêmes données. On va ici parvenir au niveau garanti $1-(\alpha+\beta)$. Par rapport à la technique précédente, on perdra donc un facteur additif $\alpha\beta$, qui cependant, est souvent négligeable en pratique : lorsque $\alpha = \beta = 5\,\%$ par exemple, il vient $\alpha\beta = 0.25\,\%$!

On utilise le calcul général suivant, qui vaut dans toutes les circonstances, et même en cas d'absence d'indépendance entre les événements $\mu_0 \in \widehat{I}$ et $\mu_0' \in \widehat{J}$:

$$\begin{aligned} \mathbb{P}\Big\{ \mu_0 \in \widehat{I} \ \text{ et } \ \mu_0' \in \widehat{J} \Big\} &= 1 - \mathbb{P}\Big\{ \mu_0 \notin \widehat{I} \ \text{ ou } \ \mu_0' \notin \widehat{J} \Big\} \\ &\geqslant 1 - \Big(\mathbb{P}\big\{ \mu_0 \notin \widehat{I} \big\} + \mathbb{P}\big\{ \mu_0' \notin \widehat{J} \big\} \Big) = 1 - 2\alpha \,. \end{aligned}$$

Remarque 5.10. En pratique, on prendra souvent des niveaux $\alpha = \beta = 2.5\,\%$ pour $\widehat{I}$ et $\widehat{J}$, de sorte que le fait que les deux intervalles de confiance contiennent chacun le vrai paramètre respectif μ_0 et μ_0' arrive avec probabilité au moins 95 %, ce qui est le niveau usuel.

Dans la section dévolue aux exercices du présent polycopié, on mettra en œuvre cette technique générale sur les exercices suivants : exercice II (question 4) de l'examen principal 2008 ; exercice I (question 5) de l'examen principal 2007.

Compléments pour étudiants avancés

6. Intervalles de confiance sur la variance

On peut calculer des intervalles de confiance sur autre chose que la moyenne, par exemple sur la variance dans le cas du modèle gaussien.

DÉFINITION 5.4. *Si* $Z_1, \ldots, Z_k$ *sont indépendantes et identiquement distribuées selon une loi* $\mathcal{N}(0,1)$*, alors on appelle la loi de* $Z_1^2 + \ldots + Z_k^2$ *la loi du* χ^2 *à* k *degrés de liberté, et on la note* χ_k^2.

En particulier, l'espérance d'une variable aléatoire distribuée selon une loi χ_k^2 est k. La loi du χ^2 est tabulée (voir les tables de quantiles à la fin de ce polycopié). On reproduit à la figure 32 des tracés des fonctions densité et de répartition de lois du χ^2, de différents degrés de liberté.

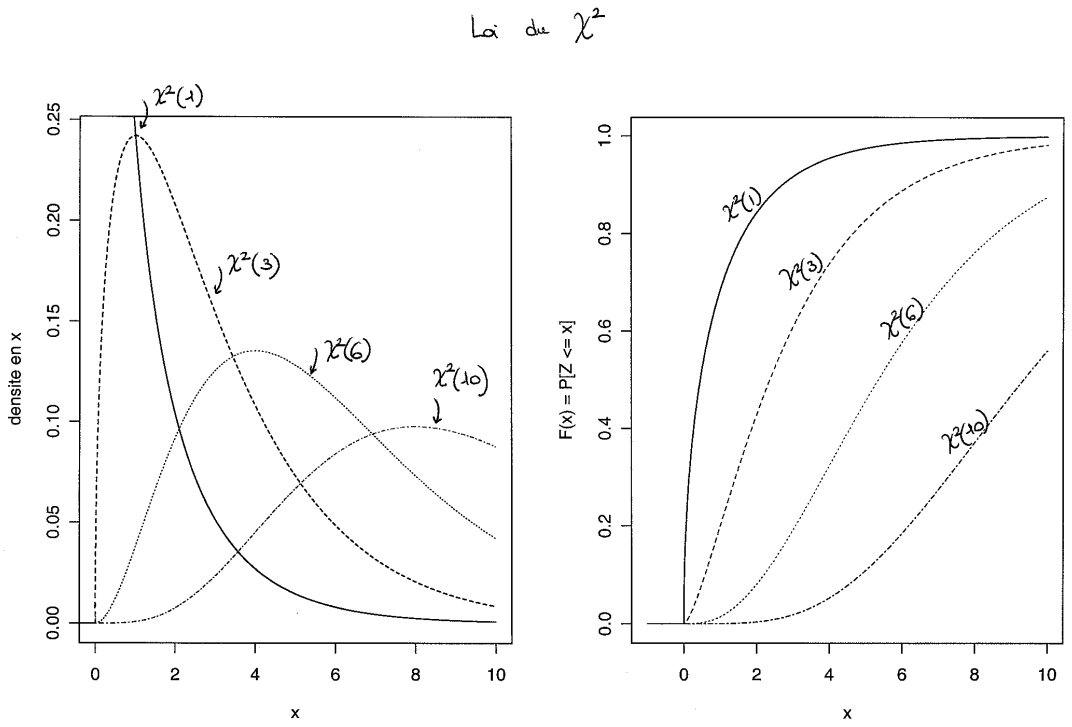


FIGURE 32. Densités (à gauche) et fonctions de répartitions (à droite) des lois du χ^2, pour différents degrés de liberté.

Lorsque $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2, alors, en considérant les variables aléatoires (non observées) $X'_t = (X_t - \mu_0)/\sigma_0$, indépendantes et identiquement distribuées, elles, selon une loi normale

standard $\mathcal{N}(0,1)$, on montre que

$$\frac{1}{\sigma_0^2}\sum_{j=1}^{n}(X_j-\mu_0)^2 \ \sim\ \chi_n^2 \ .$$

On note également cela, de manière équivalente,

$$\sum_{j=1}^{n}(X_j-\mu_0)^2 \ \sim\ \sigma_0^2\chi_n^2 \ .$$

En pratique, on ne connaît pas μ_0, et lorsqu'on l'estime, on perd un degré de liberté, comme le montre la proposition suivante (admise).

PROPOSITION 5.1. *Lorsque $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2, alors*

$$\frac{1}{\sigma_0^2}\sum_{j=1}^{n}\left(X_j-\overline{X}_n\right)^2 \ \sim\ \sigma_0^2\chi_{n-1}^2 \ .$$

On note également cela, de manière équivalente,

$$\widehat{\sigma}_n^2 \ \sim\ \frac{\sigma_0^2}{n-1}\chi_{n-1}^2$$

où $\widehat{\sigma}_n^2$ est l'estimateur sans biais de la variance.

Cette proposition entraîne immédiatement, par un simple jeu d'écriture de la définition des quantiles et quelques manipulations algébriques, le corollaire suivant.

COROLLAIRE 5.4. *On note $c_{n-1,\beta}$ le β−quantile de la loi χ_{n-1}^2. Alors, si les observations sont indépendantes et identiquement distribuées selon une loi gaussienne de paramètres μ_0 et σ_0, les intervalles suivants sont des intervalles de confiance de la variance σ_0^2, exactement de niveau $1-\alpha$:*

$$\left[\frac{(n-1)\,\widehat{\sigma}_n^2}{c_{n-1,1-\alpha/2}},\ \frac{(n-1)\,\widehat{\sigma}_n^2}{c_{n-1,\alpha/2}}\right] \ ; \qquad \left[0,\ \frac{(n-1)\,\widehat{\sigma}_n^2}{c_{n-1,\alpha}}\right] \ ; \qquad \textit{et} \qquad \left[\frac{(n-1)\,\widehat{\sigma}_n^2}{c_{n-1,1-\alpha}},\ +\infty\right[\ .$$

On utilise le second intervalle de confiance quand on soupçonne que la loi commune des données admet une dispersion plutôt faible, et que l'on veut quantifier cette impression. A l'inverse, le troisième intervalle sert pour quantifier les impressions de grande dispersion des données.

Dans ce cours, nous ne discuterons que peu de l'estimation de la variance. Cela arrivera juste un peu lors des tests de comparaison de deux populations dans la partie 9 (quand on devra faire un pré-test d'égalité des variances).

Exercices

Six exercices issus du cours

Effectuez les exercices 5.1 à 5.6, dont l'énoncé et la correction détaillée se trouvent dans la version rédigée de ce cours.

Un exercice de synthèse

Sa correction est détaillée et montre une rédaction-type.

EXERCICE 5.7 (Video killed the radio star). Une radio généraliste s'inquiète de sa perte d'audience. Une enquête faite auprès de 1 186 Français choisis au hasard dans l'annuaire et ayant accepté de répondre montre que 239 d'entre eux ont déclaré écouter au moins de temps en temps la station. Pour assurer son avenir, elle voudrait améliorer sa connaissance des habitudes de ses plus jeunes auditeurs : parmi les sondés, 102 étaient étudiants, et 30 l'écoutaient au moins de temps en temps. La direction de la prospective va diligenter un nouveau sondage destiné uniquement aux étudiants.

1. Modéliser le problème (pour chacun des deux sondages), en précisant la population, les données et de le modèle ; indiquer en particulier les paramètres d'intérêt respectifs.
2. Déduire du premier sondage une précision et un intervalle de confiance pour les paramètres d'intérêt de chacun des deux sondages.
3. Combien de personnes faut-il interroger au cours de la seconde enquête, si le degré de confiance retenu est de 95 % et la précision désirée, 3 % ?
4. A l'issue du second sondage, il a été constaté 33.8 % d'auditeurs. Donner une estimation et un intervalle de confiance du paramètre faisant l'objet de l'étude (avec un degré de confiance de 95 %).
5. Peut-on affirmer que l'audience du segment étudiant a augmenté d'une enquête à l'autre ? (Les deux enquêtes sont séparées de six mois et la station a revu entretemps sa grille de programmes.)

Cinq exercices issus des annales

La correction proposée est plus laconique que celle des exercices précédents ; je donne essentiellement les réponses attendues. Attention, il vous appartient de rédiger davantage !

EXERCICE 5.8. Traitez l'exercice II de l'examen principal de 2008. (Note : j'ai ajouté à la fin du corrigé une question subsidiaire.)

EXERCICE 5.9. Répondez aux questions 1., 2., 4., 5., 6. de l'exercice I (vente par correspondance) de l'examen principal de 2007.

EXERCICE 5.10. Traitez les trois questions du paragraphe "Modélisation et première estimation" de l'examen de rattrapage de 2008.

EXERCICE 5.11. Répondez à la question 1. de l'exercice II de l'examen de rattrapage de 2007.

EXERCICE 5.12. Considérez l'énoncé de l'exercice II (sur les somnifères) de l'examen principal de 2007. Sans répondre aux questions qui y sont posées, et en admettant la modélisation gaussienne suggérée, exhibez un intervalle de confiance de niveau 95 % sur le paramètre Δ. Que peut-on en conclure ?

Exercice 1. Video killed the radio Star.

(1) **Premier sondage :** la population étudiée est l'ensemble des Français (ils sont tous auditeurs potentiels)

on dispose de données $x_1, \dots x_{1186} \in \{0,1\}$ (où l'on note $x_j = 1$ si le j-ème sondé écoute la radio au moins de temps en temps)

la statistique dont on dispose est le taux d'audience moyen sur l'échantillon :

$$\bar{x}_{1186} = \frac{239}{1186} = 20.2\,\%$$

on modélise (via le tirage au hasard) ces données comme la réalisation de $X_1, \dots X_{1186}$ iid $\sim \mathrm{Ber}(p)$, où $p_0 \in [0,1]$

p_0 est le vrai taux d'audience sur l'ensemble des Français.

Second sondage : la population étudiée sera l'ensemble des étudiants français

on disposera de données $y_1, \dots y_n$ (où n est à déterminer), avec $y_j \in \{0,1\}$ et la même convention que précédemment

ces données seront la réalisation de $Y_1, \dots Y_n$ iid $\sim \mathrm{Ber}(q_0)$, où $q_0 \in [0,1]$, pour peu que l'on ne tire bien les sondés étudiants au hasard

q_0 est le vrai taux d'audience sur l'ensemble des étudiants Français (P est a priori différent de p_0)

Premier sondage, suite : Des données $x_1, \dots x_{1186}$ on extrait $y_1, \dots y_{102}$ correspondant aux étudiants et on modélise ces données comme la réalisation de $Y_1, \dots Y_{102}$ iid $\sim Ber(q_0)$

on dispose par ailleurs de la statistique

$$\bar{y}_{102} = \frac{30}{102} = 29.4\%$$

(2) Pour p_0 : on considère l'intervalle de confiance asymptotique, de niveau 95% :

$$\left[\bar{X}_{1186} \pm q_{97.5\%} \sqrt{\frac{\bar{X}_{1186}(1-\bar{X}_{1186})}{1186}} \right]$$

dont la réalisation est

$$\left[\bar{x}_{1186} \pm 1.96 \sqrt{\frac{\bar{x}_{1186}(1-\bar{x}_{1186})}{1186}} \right]$$

$$= \left[20.2\% \pm 1.96 \sqrt{\frac{0.202(1-0.202)}{1186}} \right]$$

$$= [\, 20.2\% \pm 2.3\% \,]$$

Pour q_0 : on considère l'intervalle de confiance asymptotique, de niveau 95% :

$$\left[\bar{Z}_{102} \pm q_{97.5\%} \sqrt{\frac{\bar{Z}_{102}(1-\bar{Z}_{102})}{102}} \right]$$

dont la réalisation est

$$\left[\bar{y}_{102} \pm 1.96 \sqrt{\frac{\bar{y}_{102}(1-\bar{y}_{102})}{102}} \right]$$

$$= \left[29.4\% \pm 1.96 \sqrt{\frac{0.294(1-0.294)}{102}} \right]$$

$$= [\, 29.4\% \pm 8.9\% \,]$$

(3) - Cf. cours, on cherche à résoudre ici $1/\sqrt{n} \leq 3\%$ (majoration de la variance par $1/4$)

soit $n \geq 1112$

- Cependant, ici, grâce à l'intervalle sur q_0, on peut majorer plus efficacement la variance, par

$$\max_{x \in [29.4\% \pm 8.9\%]} x(1-x) = 0.236$$

de sorte que la précision attendue est (au niveau 95%)

$$3 \sigma_{97.5\%} \sqrt{\frac{0.236}{n}} = \frac{0.95}{\sqrt{n}} \quad (\leq 3\%)$$

ce qui donne $n \geq \left(0.95/0.03\right)^2 = 1003.$

- Dans la suite, on prendra, pour la simplicité des propos, $n = 1000$.

(4) On dispose donc de $y_1, \dots y_{1000}$ avec $\bar{y}_{1000} = 38.3\%$

Pour q_0 : l'intervalle de confiance asymptotique, de niveau 95%, est $\left[\bar{Y}_{1000} + 3\sigma_{97.5\%} \sqrt{\frac{\bar{Y}_{1000}(1-\bar{Y}_{1000})}{1000}} \right]$

dont la réalisation est $\left[\bar{y}_{1000} \pm 3\sigma_{97.5\%} \sqrt{\frac{\bar{y}_{1000}(1-\bar{y}_{1000})}{1000}} \right]$

$= [38.3\% \pm 3.0\%]$

↑ On constate qu'on a effectivement la précision attendue.

(5) les deux intervalles exhibés

$[29.4\% \pm 8.9\%] = [20.5\%, 38.3\%]$

et $[38.3\% \pm 3.0\%] = [35.3\%, 41.3\%]$

ne sont pas disjoints : il n'est absolument pas certain que l'augmentation du taux d'audience observé sur les échantillons donne une augmentation statistiquement significative du taux d'audience dans la population (l'ensemble des étudiants français). Cela peut être simplement dû aux aléas de sondage.

Note : Par méthode de Bonferroni (ou indépendance), les deux intervalles proposés sont la réalisation d'un couple d'intervalles valant avec niveau 90% seulement.

Exercice 2. Cf. exercice II de l'examen principal 2008.

Attention ! La rédaction ci-dessous est trop laconique, j'attends de vous davantage de détails. Cf. l'exercice 1 pour une rédaction-modèle.

(1) Voir un exercice de la partie 2 ; bref rappel des notations :

Existence d'un accident responsable : $x_1, \dots x_{1472} \in \{0,1\}$

réalisations de : $X_1 \dots X_{1472}$ iid $\sim Ber(p_0)$

Montants des frais à charge : $y_1, \dots y_{256} \in \mathbb{R}_+$

réalisations de : $Y_1, \dots Y_{256}$ iid selon une certaine loi d'espérance μ_0

Les paramètres d'intérêt sont p_0 et μ_0 : taux d'accident responsable et, le cas échéant, montant des frais à charge, pour la population formée par l'ensemble des étudiants assurés de France.

(2) Pour p_0 : estimateur $\overline{X}_{1472}$, estimée $\overline{x}_{1472} = \frac{256}{1472} \approx 17.4\%$

Pour μ_0 : estimateur $\overline{Y}_{256}$, estimée $\overline{y}_{256} = 1865$.

(3) 1er cas : l'entrepreneur cherche à entrer sur ce marché sans risque, il prendra des intervalles unilatères majorant les paramètres d'intérêt (et retiendra ces majorants pour effectuer son étude de rentabilité) :

Pour p_0 : $\left[0, \overline{X}_{1472} + 3_{95\%} \sqrt{\frac{\overline{X}_{1472}(1-\overline{X}_{1472})}{1472}}\right]$ est un intervalle de confiance asymptotique de niveau 95% ; il admet pour réalisation

$$\left[0 ; 0.174 + 1.65 \sqrt{\frac{0.174(1-0.174)}{1472}}\right] = [0, 19.0\%]$$

Pour μ_0 : $\left[0, \ \bar{Y}_{256} + 3{,}95\% \sqrt{\frac{\hat{\sigma}^2_{Y,256}}{256}} \right]$

est un intervalle de confiance asymptotique de niveau 95%, qui admet pour réalisation (où $s^2_{y,256} = 524^2$) :

$$\left[0, \ 1865 + 1.65 \frac{524}{\sqrt{256}} \right]$$

$$= [0, \ 1919]$$

2nd cas : C'est le rêve de sa vie, il est sûr de se lancer, mais voudrait ajuster sa trésorerie et savoir précisément le montant des fonds à lever. Il veut simplement mieux connaître son marché et prendra donc des intervalles bilatères.

On obtient les réalisations suivantes d'intervalles de confiance asymptotiques de niveau 95% chacun :

- pour p_0 : $[17.4\% \pm 2.0\%]$
- pour μ_0 : $[1865 \pm 65]$

3ème cas : Un ami de l'entrepreneur voulant lui montrer qu'il fait l'erreur de sa vie lui fournirait des intervalles minorant les paramètres d'intérêt, histoire de lui montrer combien toute cette affaire est dangereuse ; ces intervalles admettant respectivement pour réalisation : $[15.8\%, 100\%]$ et $[1811, +\infty[$.

(4) (On ne fait ici que les calculs du 2nd cas ; les autres cas sont similaires.) Il s'agit d'estimer $\mu_0 p_0$.

Les intervalles de confiance sur p_0 et μ_0 valent simultanément avec probabilité $\geq 90\%$; de ces intervalles, on déduit par multiplication la réalisation suivante de l'intervalle de confiance sur $\mu_0 p_0$:

$$\left[(17.4\% - 2.0\%) \times (1865 - 65) \; ; \; (17.4\% + 2\%)(1865 + 65) \right]$$
$$= [\, 277, \; 375 \,]$$
$$= [\, 326 \pm 99 \,].$$

Cet intervalle sur p_{OP} est donc la réalisation d'un intervalle de confiance de niveau 90% sur p_{OP} (nous n'avons pas écrit explicitement la formule de ce dernier, mais il serait facile de le faire).

(5) Nous l'avons déjà vu, mais selon le cas :

- 2nd cas : à ajuster la trésorerie (les provisions) ou à avoir une idée du montant moyen attendu de la prime *
- 1er et 3ème cas : à étudier, de manière pessimiste ou optimiste, la rentabilité de l'affaire.

(6) Question supplémentaire

(corrigée page suivante)

[S'il ne s'agissait que de bien estimer p_{OP}, ne pourrait-on pas procéder plus directement et plus efficacement ?]

* Evidemment, pas à fixer les primes, puisque celles-ci sont fortement individuelles et dépendent p.ex. des antécédents de l'assuré !

L'idée, c'est d'intégrer ceux qui n'ont pas eu d'accident responsable au cours de l'année directement dans les frais (frais nuls) au lieu de procéder en deux temps.

Soient ainsi $Z_1, \dots Z_{1472} \in \mathbb{R}^+$ les frais (éventuellement nuls : c'est le cas de $1472 - 256 = 1216$ d'entre eux) occasionnés par les assurés. Ce sont les réalisations de $Z_1 \dots Z_{1472}$ iid selon une certaine loi d'espérance $\mu_0 p_0$ (par définitions de μ_0 et p_0).

L'estimateur naturel est $\bar{Z}_{1472}$ et l'intervalle de confiance (disons à 90% : pour pouvoir comparer avec l'intervalle précédemment obtenu) construit sur lui pour $\mu_0 p_0$ est :

$$\left[\bar{Z}_{1472} \pm z_{95\%} \sqrt{\hat{\sigma}^2_{Z,1472}/1472} \right]$$

Il suffit de calculer les réalisations ; à cet effet, il faut déterminer $\bar{z}_{1472}$ et $s_{z,1472}$, moyenne et écart-type observés.

On le fait par ingénierie inverse :

- $$\bar{z}_{1472} = \frac{1}{1472} \sum_{i=1}^{1472} z_i = \frac{1}{1472} \sum_{i=1}^{256} y_i = \frac{256}{1472} \bar{y}_{256} = \frac{256}{1472} \times 1865 \simeq 324 \text{ €}$$

  (car les 1216 z_i sont nuls !)

- $$s^2_{z,1472} = \frac{1472}{1471} \left(\frac{1}{1472} \sum_{i=1}^{1472} z_i^2 - \left(\frac{1}{1472} \sum_{i=1}^{1472} z_i \right)^2 \right)$$

  ($s^2_{z,1472}$: variance des données (version débiaisée))

  $$= \frac{1}{1472} \sum_{i=1}^{256} y_i^2$$

  ($\bar{z}_{1472} = 324$ €)

  or, $$s^2_{y,256} = \frac{256}{255} \left(\frac{1}{256} \sum_{i=1}^{256} y_i^2 - (\bar{y}_{256})^2 \right)$$

  d'où $$\sum_{i=1}^{256} y_i^2 = \left((\bar{y}_{256})^2 + \frac{255}{256} s^2_{y,256} \right) \times 256$$

  $$\frac{1}{1472} \sum_{i=1}^{256} y_i^2 = \frac{256}{1472} \left(1865^2 + \frac{255}{256} \times 524^2 \right) \simeq 808^2$$

et $s^2_{\bar{z}_{1472}} = \frac{1472}{1471}\left(888^2 - 324^2\right) \cong 740^2$

D'où la réalisation : $\left[\bar{z}_{1472} \pm 1.65 \times \frac{s_{\bar{z}_{1472}}}{\sqrt{1472}}\right]$

(1.65 : quantile à 95%)

$= \left[324 \pm 1.65 \frac{740}{\sqrt{1472}}\right] \simeq \left[324 \pm 32\right]$

ce qui est beaucoup plus précis que l'intervalle [326 ± 99] précédemment obtenu !

Exercice 3. Cf. exercice I de l'examen principal 2007

Attention ! La rédaction ci-dessous est trop laconique, j'attends de vous davantage de détails. Cf. l'exercice 1 pour une rédaction modèle.

(1) Voir un exercice de la partie 2 ; bref rappel des notations :

Existence d'une commande : $x_1, \dots x_{1000} \in \{0;1\}$

réalisations de : $X_1, \dots X_{1000}$ iid $\sim Be(p_0)$

Montant des commandes : $y_1, \dots y_{170} \in \mathbb{R}_+$

réalisations de : $Y_1, \dots Y_{170}$ iid selon une certaine loi d'espérance μ_0

Les paramètres d'intérêt sont p_0 et μ_0 : taux de commande et montant moyen des commandes si on généralisait l'offre à l'ensemble de la population (= les 50 000 clients du fichier).

(2) Dans cet exercice, on prendra un point de vue pessimiste et on cherchera à exhiber des minorants sur p_0 et μ_0, afin de déterminer une valeur planchers de rentabilité.

L'intervalle de confiance asymptotique que l'on retiendra sur p_0 est donc donné par :

$$\left[\bar{X}_{1000} - z_{95\%} \sqrt{\frac{\bar{X}_{1000}(1-\bar{X}_{1000})}{1000}} \; ; \; 100\% \right]$$

Sa réalisation sur nos données ($\bar{x}_{1000} = 17.0\%$) est :

$$\left[0.17 - 1.65 \sqrt{\frac{0.17\,(1-0.17)}{1000}} \; ; \; 1 \right] = [15.0\%; \; 100\%]$$

(3) Nous n'avons pas encore vu les tests statistiques ; mais d'ores et déjà,

l'estimation plancher de 15.0% (à comparer aux 13% habituels) nous souffle que le taux de commande p_0 est significativement plus grand que le taux habituel.

(4) Ici également, et selon la même démarche qu'en (2), on recourt à un intervalle de confiance unilatère.
L'intervalle de confiance asymptotique que l'on retiendra sur μ_0, de niveau 95%, est :

$$\left[\bar{Y}_{170} - z_{95\%} \sqrt{\frac{\hat{\sigma}^2_{Y,170}}{170}} \; ; \; +\infty \right[$$

et sa réalisation sur les données ($\bar{y}_{170} = 73$, $s_{y,170} = 8$) vaut

$$\left[73 - 1.65 \times \frac{8}{\sqrt{170}} \; , \; +\infty \right[\; = \; \left[72.0, \; +\infty \right[$$

(5) Il faut d'abord voir que le chiffre d'affaires en question est donné par $CA = 50\,000\, p_0\, \mu_0$. En effet, on attend $50\,000\, p_0$ commandes, chacune d'un montant moyen de μ_0.

On déduit un intervalle de confiance sur cette quantité CA à partir de ceux sur p_0 et μ_0 :

$$\left[50\,000 \times 15.0\% \times 72.0 \; , \; +\infty \right[$$
$$= \left[540\,000, \; +\infty \right[.$$

Par méthode de Bonferroni, les valeurs proposées ci-dessus correspondent à la réalisation d'un intervalle de confiance de niveau 90% seulement sur CA. (Nous n'avons pas donné la formule théorique explicite de ce dernier, mais il serait facile de le faire.)

(6) Attention ! Cette question n'est pas réellement une question de statistiques, mais plutôt une question de bon sens : il s'agit de voir comment passer d'un chiffre d'affaires à une marge brute (= profit réalisé sur les ventes, marge avant impôts).

Auparavant (sans offre promotionnelle) : 100 € de commandes donnent 40 € de marge

Ancien chiffre d'affaires : 50 000 × 13% × 70 €
= 455 000 €

soit une marge de 455 000 × 40% = 182 000 €
(on retrouve bien le chiffre de l'énoncé)

Avec l'offre : 100 € de commandes ne rapportent que 95 € (prix facturé après remise) et coûtent en réalité 60 € ; soit une marge de 35 €.
Le nouveau taux de marge pour le montant avant remise est donc de 35%.

Chiffres d'affaires minimum attendu avec la promotion : 540 000 €

soit une marge attendue de 540 000 × 35%
= 189 000 €

Conclusion : Meilleure rentabilité de la nouvelle offre (conclusion tirée d'une méthode donnant le vrai résultat avec proba. ≥ 90%)
Remarque : et encore, on a fait ici une analyse précautionneuse et pessimiste (on a exhibé des minorants).

↳ Il faut lancer la nouvelle offre sur tout le fichier clients !

Exercice 4. Cf. § Modélisation et première estimation de l'examen de rattrapage 2008.

Attention! La rédaction ci-dessous est trop laconique, j'attends de vous davantage de détails. Cf. l'exercice 1 pour une rédaction modèle.

(1) Voir un exercice de la partie 2; bref rappel des notations :
On ne considère que la série de données $y_1, \dots y_{172} \in \mathbb{R}^+$ correspond aux montants d'achats déclarés (prévus ou effectués).
On les a modélisés comme la réalisation de $Y_1, \dots Y_{172}$ iid selon une certaine loi d'espérance μ_Y.
μ_Y était le montant moyen des achats déclarés par l'ensemble des clients de Vélizy II.

(2) Cette figure est obtenue, dans le menu "Analyse", par un sous-élément du menu "Statistiques descriptives", p. ex. "Effectifs" (en précisant les deux séries "Salaire" et "Achats" dans la boîte de dialogue correspondante). Note : on lit $\bar{y}_{170} = 376.22$ et $s_{y,170} = 259.82$

(3) On n'a pas de raison de ne pas prendre un intervalle bilatère.
On retient l'intervalle suivant, asymptotique de niveau 95% :

$$\left[\bar{Y}_{170} \pm z_{97.5\%} \sqrt{\hat{\sigma}^2_{Y,170}/170} \right]$$

Sa réalisation vaut $\left[\bar{y}_{170} \pm 1.96 \frac{s_{y,170}}{\sqrt{170}} \right]$

$$= \left[376.22 \pm 1.96 \times \frac{259.82}{\sqrt{170}} \right] = [376.22 \pm 39.05]$$

et on écrira plutôt : $[376 \pm 40]$

Exercice 5. Cf. exercice II de l'examen de rattrapage 2007

* On a déjà vu la modélisation dans un exercice de la partie 2; on en rappelle brièvement les éléments qui nous intéressent.

On dispose des données $x_1, \dots x_{1837} \in \mathbb{R}_+$ fournies par les sondés, et on les a modélisées comme la réalisation de $X_1, \dots X_{1837}$ iid selon une certaine loi d'espérance μ_0. Cette dernière forme notre paramètre d'intérêt : le montant mensuel moyen (sur l'ensemble des foyers français) consacré aux achats hors produits de nécessité.

* On retient la formule suivante, pour l'intervalle de confiance (asymptotique) de niveau 95 % sur μ_0 :

$$\left[\bar{X}_{1837} \pm z_{97.5\%} \sqrt{\hat{\sigma}^2_{X_1,1837} / 1837} \right]$$

Or, on dispose des statistiques d'échantillon

$$\bar{x}_{1837} = 598 \quad \text{et} \quad s_{x_1,1837} = 254,$$

de sorte que la valeur réalisée pour l'intervalle de confiance est :

$$\left[\bar{x}_{1837} \pm 1.96 \frac{s_{x_1,1837}}{\sqrt{1837}} \right] = \left[598 \pm 1.96 \times \frac{254}{\sqrt{1837}} \right]$$

$$= [598 \pm 12].$$

(Là encore, on préfère écrire ± 12 plutôt que $\pm 11.61 \dots$)

Exercice 6. Cf. exercice II, examen principal 2007

On considère donc les données

$z_1, \dots z_{10} \in \mathbb{R}$ suivantes :

z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
-0.1	0.4	-1.0	-0.2	0.3	0.1	-0.1	-0.2	0.2	-0.7

Les patients étant soumis indépendamment au traitement (et notamment: dormant chacun seul dans sa chambre), on peut modéliser $z_1, \dots z_{10}$ comme la réalisation de $Z_1 \dots Z_{10}$ iid selon une certaine loi. Les paramètres de cette loi (espérance Δ, variance σ^2) seraient ceux que l'on obtiendrait si l'on soumettait toute la population française à ce protocole.

En fait, le test de normalité (Shapiro - Wilk) accepte l'hypothèse que $Z_1 \dots Z_{10}$ iid $\sim \mathcal{N}(\Delta, \sigma^2)$.

Ainsi, on a l'intervalle de confiance (non-asymptotique) à 95% suivant (on emploie la formule pour le petits échantillons issus de lois normales) :

$$\left[\overline{Z}_{10} \pm t_{9,\,97.5\%} \sqrt{\hat{\sigma}^2_{Z,10}/10} \right].$$

utilisation de la loi de Student

l'énoncé donne une variance

On dispose des statistiques d'échantillon $\overline{z}_{10} = -0.13$ et $s^2_{z,10} = 0.19$; soit la réalisation

$$\left[\overline{z}_{10} \pm t_{9,\,97.5\%} \sqrt{s^2_{z,10}/10} \right] = \left[-0.13 \pm 2.262 \times \sqrt{\frac{0.19}{10}} \right]$$

$$= [-0.13 \pm 0.32]$$

$$= [-0.45\,;\ 0.19]$$

Puisque $0 \in [-0{,}45 ; 0{,}19]$, on ne sait vraiment pas (malgré $\bar{Z}_{10} = -0.13 < 0$) quel médicament est le plus efficace, on ne peut rien dire du signe de Δ.

Note : on aurait aussi pu proposer l'intervalle

$$]-\infty, \bar{Z}_{10} + t_{9,95\%} \sqrt{S^2_{\bar{Z}_{10}}/10}\,] =]-\infty, 0.123]$$

qui est la réalisation d'un intervalle de confiance unilatère (non asymptotique, fondé sur la loi de Student) de niveau 95%. Là non plus, on ne peut rien conclure quant au signe de Δ.

Conclusion ? Du point de vue des statistiques, on ne peut pas en tirer.

MAIS · vu $\bar{Z}_{10} < 0$,

· vu les boîts à moustaches (meilleure médiane, dispersion plus faible), on soupçonne que l'ancien médicament Morpheus est plus efficace que le nouveau !

↳ Il faudrait mener une étude sur un échantillon plus grand que $n = 10$ personnes pour le prouver !

Sixième Partie

Interlude : deux quizz sur l'estimation

Premier énoncé (sujet posé en 2009)

Seuls accessoires autorisés : calculatrice et tables des lois

Quizz 2 – Estimation ponctuelle et par intervalles – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Question de cours

On se place dans le modèle où $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi de Bernoulli.

1. Donner l'expression d'un estimateur naturel de la variance commune des X_j.
2. Traduire mathématiquement le fait qu'il soit consistant (pas de démonstration demandée de ce fait).
3. Rappeler son espérance et indiquer comment le rendre sans biais.

Lecture de tables

Donnez la valeur u telle que $\mathbb{P}\big\{|N| \geq u\big\} = 16\,\%$, où $N \sim \mathcal{N}(0,1)$; ici, on précisera également la notation pour u.

Encadrez $\mathbb{P}\{T \leq 1.9\}$ lorsque $T \sim \mathcal{T}_5$.

Statistiques et boulangeries

Une troupe d'étudiants a pour mission de mener une étude statistique sur le prix du sandwich jambon-beurre à Paris. Ils se rendent dans 97 boulangeries prises au hasard (en choisissant des coordonnées géographiques au hasard et en déterminant la boulangerie la plus proche de chacune d'elles). Ils obtiennent les mesures suivantes : prix moyen de 4.15 euros (avec écart-type constaté de 72 centimes) et longueur moyenne de 19.6 centimètres (avec écart-type constaté de 3.8 centimètres).

Modéliser très rapidement la situation rencontrée, en ne considérant qu'un seul des deux paramètres d'intérêt possibles (*au choix*), puis donner un intervalle de confiance sur ce dernier. **Attention !** Il sera fait grand cas de la qualité de la rédaction.

Premier corrigé (sujet posé en 2009)

Seuls accessoires autorisés : calculatrice et tables des lois

Quizz 2 – Estimation ponctuelle et par intervalles – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Gilles Stoltz

Question de cours

On se place dans le modèle où $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi de Bernoulli.

1. Donner l'expression d'un estimateur naturel de la variance commune des X_j.
2. Traduire mathématiquement le fait qu'il soit consistant (pas de démonstration demandée de ce fait).
3. Rappeler son espérance et indiquer comment le rendre sans biais.

Attention ! On se plaçait ici dans le cas où $X_1 \ldots X_n$ iid $\sim$ Ber(p_0).

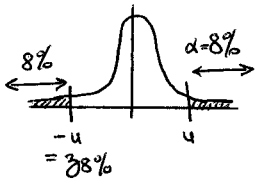
1. C'est $\bar{X}_n(1-\bar{X}_n)$, il correspond à la version biaisée de l'estimateur de la variance.
2. $\bar{X}_n(1-\bar{X}_n) \xrightarrow{\mathbb{P}} p_0(1-p_0)$
3. $\mathbb{E}[\bar{X}_n(1-\bar{X}_n)] = \frac{n-1}{n} p_0(1-p_0)$ d'où $\frac{n}{n-1}\bar{X}_n(1-\bar{X}_n)$ est sans biais.

Note : on a $\frac{n}{n-1}\bar{X}_n(1-\bar{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$

Lecture de tables

Donnez la valeur u telle que $\mathbb{P}\{|N| \geq u\} = 16\,\%$, où $N \sim \mathcal{N}(0,1)$; ici, on précisera également la notation pour u.

$u = z_{1-\alpha} = z_{92\%}$ et l'on lit sur la table des quantiles de la $\mathcal{N}(0,1)$: $z_{92\%} = 1.4051$



Note : $u = -z_{8\%}$

Encadrez $\mathbb{P}\{T \leq 1.9\}$ lorsque $T \sim \mathcal{T}_5$.

Par lecture de la table de la loi de Student :

$\mathbb{P}\{T \leq 1.476\} \leqslant \mathbb{P}\{T \leq 1.9\} \leqslant \mathbb{P}\{T \leq 2.015\}$

$= 1 - \mathbb{P}\{T \geqslant 1.476\} = 90\%$ $\quad = 1 - \mathbb{P}\{T \geqslant 2.015\} = 95\%$

Statistiques et boulangeries

Une troupe d'étudiants a pour mission de mener une étude statistique sur le prix du sandwich jambon-beurre à Paris. Ils se rendent dans 97 boulangeries prises au hasard (en choisissant des coordonnées géographiques au hasard et en déterminant la boulangerie la plus proche de chacune d'elles). Ils obtiennent les mesures suivantes : prix moyen de 4.15 euros (avec écart-type constaté de 72 centimes) et longueur moyenne de 19.6 centimètres (avec écart-type constaté de 3.8 centimètres).

Modéliser très rapidement la situation rencontrée, en ne considérant qu'un seul des deux paramètres d'intérêt possibles (*au choix*), puis donner un intervalle de confiance sur ce dernier. **Attention !** Il sera fait grand cas de la qualité de la rédaction.

Je m'intéresse aux prix uniquement.

Modélisation brève :

Les données de prix $x_1, \dots x_{97} \in [0,20]$ (disons) peuvent, via le choix au hasard des boulangeries, être modélisées comme la réalisation de variables aléatoires $X_1, \dots X_{97}$ iid selon une certaine loi sur $[0,20]$, dont on note μ_0 l'espérance (inconnue).

μ_0 forme le paramètre d'intérêt : c'est le prix moyen sur toutes les boulangeries de Paris pour le jambon-beurre.

Note : on dispose des statistiques d'échantillon $\bar{x}_{97} = 4.15$ et $s_{x,97} = 0.72$

Intervalle de confiance sur μ_0 :

1er cas : les étudiants veulent juste avoir une idée précise des prix.

L'intervalle théorique à 95% est $\hat{I}_{97} = \left[\bar{X}_{97} \pm z_{97.5\%} \sqrt{\frac{\hat{\sigma}^2_{97}}{97}} \right]$

Il admet pour réalisation :

$$\left[\bar{x}_{97} \pm z_{97.5\%} \frac{s_{x,97}}{\sqrt{97}} \right] = \left[4.15 \pm 1.96 \frac{0.72}{\sqrt{97}} \right]$$

$$\simeq [4.15 \pm 0.15]$$

2ème cas : s'ils ont en tête de montrer que les sandwichs sont chers à Paris, ils considéreront l'intervalle à 95%

$$\hat{J}_{97} = \left[\bar{X}_{97} - z_{95\%} \sqrt{\frac{\hat{\sigma}^2_{97}}{97}} ; +\infty \right[$$

qui admet pour réalisation :

$$\left[\bar{x}_{97} - z_{95\%} \frac{s_{x,97}}{\sqrt{97}}, +\infty \right[$$

$$= \left[4.15 - 1.645 \frac{0.72}{\sqrt{97}}, +\infty \right[= [4.03, +\infty[$$

Note : Ici on retient deux chiffres après la virgule parce que l'énoncé donnait toutes les estimées avec cette précision.

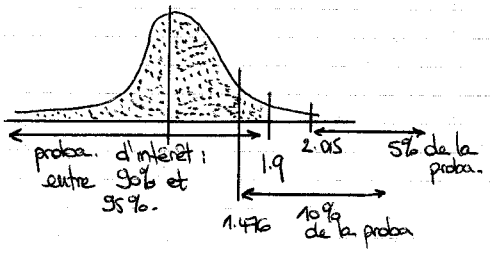
Commentaires sur la correction et erreurs typiques.

Question de cours.

- On se plaçait dans un modèle de Bernoulli, il fallait en tenir compte.
- A la question 1., j'ai souvent vu proposé, à tort, $\frac{1}{m}\bar{X}_n(1-\bar{X}_n)$
- A la question 3., il s'agissait évidemment de faire le lien avec la question (avec l'estimateur naturel $\bar{X}_n(1-\bar{X}_n)$ exhibé).

Lecture de table.

- Faits un dessin ! C'est plus parlant qu'un long calcul, et permet de voir comment passer de z_β (β-quantile) aux quantités $z_{1-\alpha}$ ($(1-\alpha)$-quantils) donnés par les tables.
- Idem pour la loi de Student. Note : environ la moitié d'entre vous a proposé l'encadrement [5%, 10%], qui provient d'une mauvaise représentation de ce que donne la table : faites un dessin et confrontez-le à celui de la table, p. ex. :



- Lecture de quantils : attention, c'est ($\alpha = 0.008$) $z_{99.2\%}$ qui vaut 2.4089 ; et non $z_{92\%}$, qui, lui, vaut 1.4051

Statistiques & boulangeries.

Partie modélisation

- Modéliser de manière concise mais complète ; en particulier, il faut arriver à $X_1, \dots X_{97}$ iid selon une certaine loi dont on précise et interprète le paramètre μ_0 (= moyenne sur TOUTES les boulangeries de Paris, pas juste les sondées)
- μ_0 est inconnu, il est tout à fait FAUX (même si je l'ai encore souvent vu) que μ_0 vale $\bar{x}_{97}$.
- Notations : $\hat{\sigma}_{97}$ = estimateur de l'écart-type (construit sur les X_j)
  $s_{x,97}$ = estimée de l'écart-type (sur les x_j)

Calcul de l'intervalle

- Ne pas écrire la formule générale $\left[\bar{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right]$
  mais la spécifier : prenant un niveau de 95% (soit $\alpha = 5\%$)
  et disposant de $n = 97$ données, nous considérons
  $$\hat{I}_{97} = \left[\bar{X}_{97} \pm z_{97.5\%} \sqrt{\hat{\sigma}_{97}^2 / 97} \right].$$
- Attention au moment de calculer la réalisation, la réalisation de $\sqrt{\hat{\sigma}_{97}^2}$ est $s_{x,97} = 0.72$ ici et non $\sqrt{0.72}$
  ↳ C'est l'erreur typique #1
- Ne pas mélanger des X_j, $\bar{X}_{97}$, $\hat{\sigma}_{97}^2$, etc. avec des x_j, $\bar{x}_{97}$, $s_{x,97}$ etc.
  Procéder en deux temps :
  * la formule théorique
  * la réalisation obtenue (sur les données)

  Attention lors de la rédaction :
  * il y a bien une probabilité 95% que $\mu_0 \in \hat{I}_{97}$

* μ_0 appartient ou pas à la réalisation (probabilité d'appartenance 0 ou 1, donc) de cet intervalle ; mais on ignore quelle assertion entre ces deux est vraie.

↳ Cela conduit à l'erreur typique #2.

- Enfin, on rappelle que l'intervalle $\left[\bar{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right]$ n'est destiné qu'à des variables aléatoires de Bernoulli et ne s'applique pas ici.

↳ C'est l'erreur typique #3.

Pour ne plus <u>jamais</u> revoir ces erreurs, je reproduis une TD copie et d'autres commettant les erreurs.

EXEMPLE DE BONNE COPIE

Population visée : Ensemble des boulangeries de Paris.

Données :

Taille : $x_1, x_2, \ldots, x_{97}$: Prix du sandwich jambon-beurre dans la boulangerie i

Etendue : $\forall i \in [\![1, 97]\!]$, $x_{97} \in [0; 50]$ Le prix > 50 € peu probable pour un sandwich

$\bar{x}_{97} = 4,15$ euros. et $s_{x,97} = 0.72$ €

Modélisation :

$x_1, \ldots, x_{97}$ réalisation de $X_1, \ldots, X_{97}$ indépendantes (car les boulangeries sont choisies au hasard) et identiquement distribuées selon une certaine loi d'espérance μ_0 et de variance σ_0^2.

Paramètre d'intérêt et interprétation :

Le paramètre d'intérêt est μ_0. μ_0 est ici le prix moyen réel du sandwich Jambon-beurre à Paris.

Intervalles de confiance :

$\hat{I}_{97} = \left[\bar{X}_{97} \pm \mathfrak{z}_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_{97}^2}{97}} \right]$ intervalle de confiance au niveau au moins égal à $1-\alpha$ dont la réalisation est pour $\alpha = 5\%$.

$$\left[4,15 \pm \mathfrak{z}_{97,5\%} \sqrt{\frac{0,72^2}{97}} \right] = \left[4,15 \pm 1,96 \sqrt{\frac{0,72^2}{97}} \right].$$

écrire $\left[\bar{x}_{97} \pm \mathfrak{z}_{97,5\%} \frac{s_{x,97}}{\sqrt{97}} \right]$

$= [4,01 ; 4,29]$ ✓

ERREUR #1

Seri
Population sandwich jambon beurre de Paris

$\overline{x} = 4,15$ €

$\overline{s} = 0,72$ €

modélisation trop brève

X VA modélisant le prix des sandwichs parisiens suivant une loi d'espérance μ et de variance σ^2

$$\left[\overline{X}_{97} \pm z_{97,5\%} \sqrt{\frac{\hat{\sigma}^2_{97}}{n_{97}}} \right]$$

La réalisation de cet intervalle de confiance est à 95%

$$\left[4,15 \pm 1,96 \sqrt{\frac{0,72}{97}} \right] \longleftrightarrow 0.72/\sqrt{97}$$

$$[\, 3,98 \ , \ 4,31 \,]$$

ERREUR #2

On considère la population des boulangeries parisiennes.
Pour l'étude, on considère un échantillon aléatoire de
de 97 boulangeries.
On obtient les données $x_1, \dots, x_{97}$
avec $\forall i = 1 \dots 97$ x_i est le prix du sandwich
les x_i modélisent $X_1, \dots, X_{97}$, variables aléatoires i.i.d
qui suivent une certaine loi, d'espérance μ_0, notre
paramètre d'intérêt, qui est le véritable prix moyen
du sandwich en considérant la totalité des boulangeries
parisiennes, c'est lui qu'on cherche à estimer.
On obtient les données suivantes :

$\bar{x}_{97} = 4{,}15$ et $s_{97} = 0{,}72$ (estimée de l'écart type)

On veut déterminer un intervalle de confiance
asymptotique de niveau 95% pour μ_0 :

on a théoriquement : $\mu_0 \in \left[\bar{X}_{97} - z_{97\%} \sqrt{\frac{\hat{\sigma}^2_{\mu_0}}{97}} ; \bar{X}_{97} + z_{97\%} \sqrt{\frac{\hat{\sigma}^2_{\mu_0}}{97}} \right]$
avec proba 95%

en remplaçant par les données :

$\mu_0 \in \left[\bar{x}_{97} \pm z_{97\%} \sqrt{\frac{s^2_{97}}{97}} \right]$ ← c'est une réalisation de l'IC ; on ne sait pas si μ_0 y appartient

$\mu_0 \in \left[4{,}15 \pm 1{,}96 \sqrt{\frac{(0{,}72)^2}{97}} \right]$

$\mu_0 \in [4{,}15 \pm 14\%]$

0.14

ERREUR #2

population boulangeries de Paris

Données $x_1, \dots, x_{97} \in \mathbb{R}^+$: prix d'un jambon beurre

$\overline{x_{97}} = 4{,}15$

Modélisation $X_1, \dots, X_{97}$ iid suivent une certaine loi d'espérance μ_0 et de variance σ_0^2

iid : les boulangeries sont choisies au hasard

μ_0 est le paramètre d'intérêt : il correspond au prix moyen d'une baguette.

Un intervalle de confiance au niveau 95 % de μ_0 nous est donné par la formule théorique

$$\left[\overline{X}_{97} \pm z_{97,5} \times \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right].$$

Sa réalisation est donc la suivante :

$$\left[4{,}15 \pm 1{,}96 \times \frac{0{,}72}{\sqrt{97}} \right]$$ OK

μ_0 a 95 % de chance d'appartenir à cet ensemble.

NON si c'est $[4.15 \pm \dots]$

OUI si c'est $[\overline{X}_{97} \pm \dots]$

ERREUR #3

Population : Les ~~baguettes~~ jambon-beurre des boulangeries parisiennes

données : $x_1, \dots, x_{97} \in [0\ 10]$

$\bar{x}_{97} = 4{,}15$

Modélisation : $X_1, \dots, X_{97}$ suivent une loi inconnue, i.i.d, avec comme paramètre d'intérêt p_0 (le prix)

p_0 est le prix d'un sandwich jambon beurre moyen sur l'ensemble des boulangeries parisiennes.

Intervalle de confiance

NON : $\hat{\sigma}^2_{97}$ (il ne s'agit pas de bio de Bernoulli ici)

$\left[\bar{X}_n \pm z_{97,5} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right]$ est un intervalle de confiance de degré de confiance 95% dont la réalisation avec les données présentes est

$\left[4{,}15 \pm 1{,}96 \sqrt{\frac{4{,}15(1-4{,}15)}{97}} \right]$

Vous êtes embêté parce que le numérateur est <0 !

Second énoncé (sujet posé en 2008)

Seuls accessoires autorisés : Calculatrice et tables des lois

Quizz 2 – Eléments de statistique mathématique

Question de cours

Rappelez la formule de l'estimateur sans biais de la variance (pour un n-échantillon $X_1, \ldots, X_n$ i.i.d. selon une loi de variance σ_0^2). Traduisez mathématiquement son caractère sans biais et son caractère consistant.

Lecture de tables

Dans la table de la loi de Student, lisez ou encadrez le mieux possible, selon les cas, les valeurs suivantes :

Quel est u tel que $\mathbb{P}\{T \leq u\} = 99\,\%$, lorsque $T \sim \mathcal{T}_6$? On donnera sa notation et sa valeur :

Encadrez $\mathbb{P}\{T \geq 2\}$ lorsque $T \sim \mathcal{T}_4$:

Un intervalle de confiance

On a interrogé au hasard 100 étudiants à la cantine. Parmi eux, 70 sont en faveur de la construction d'une église et 30, pour une piscine. Quel est ici le paramètre statistique d'intérêt ? (Lui donner un nom et l'interpréter par une petite phrase.) Estimez-le par intervalle. Note : $\sqrt{0.21} \approx 0.46$.

Second corrigé (sujet posé en 2008)

Etudiant: Gilles Stoltz

Quizz 2 – Eléments de statistique mathématique

Question de cours

Rappelez la formule de l'estimateur sans biais de la variance (pour un n-échantillon $X_1, \ldots, X_n$ i.i.d. selon une loi de variance σ_0^2). Traduisez mathématiquement son caractère sans biais et son caractère consistant.

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X}_n)^2$$

sans biais : $\mathbb{E}\left[\hat{\sigma}_n^2\right] = \sigma_0^2$

consistant : $\hat{\sigma}_n^2 \xrightarrow[n \to +\infty]{\mathbb{P}} \sigma_0^2$

Lecture de tables

Dans la table de la loi de Student, lisez ou encadrez le mieux possible, selon les cas, les valeurs suivantes :

Quel est u tel que $\mathbb{P}\{T \leq u\} = 99\,\%$, lorsque $T \sim \mathcal{T}_6$? On donnera sa notation et sa valeur :

$$u = t_{6,99\%} = 3.143$$

Encadrez $\mathbb{P}\{T \geq 2\}$ lorsque $T \sim \mathcal{T}_4$:

$$\mathbb{P}\{T \geq 2\} \in \left[\mathbb{P}\{T \geq 2.132\} ; \mathbb{P}\{T \geq 1.533\} \right] = [5\%, 10\%]$$

Un intervalle de confiance

On a interrogé au hasard 100 étudiants à la cantine. Parmi eux, 70 sont en faveur de la construction d'une église et 30, pour une piscine. Quel est ici le paramètre statistique d'intérêt ? (Lui donner un nom et l'interpréter par une petite phrase.) Estimez-le par intervalle. Note : $\sqrt{0.21} \approx 0.46$.

La modélisation est similaire à celle du second exercice du quizz #1 de 2008, on la rappelle brièvement. La population visée est l'ensemble des étudiants en cours de scolarité à HEC. Les données sont $x_1, \ldots x_{100} \in \{0,1\}$ (1 pour l'église, 0 pour la piscine) ; elles sont la réalisation de $X_1, \ldots X_{100}$ iid $\sim \mathrm{Ber}(p_0)$, où p_0 est la vraie proportion des étudiants en faveur de l'église (on la connaîtrait si on organisait des élections). Sur les données : $\overline{x}_{100} = 70.0\%$, c'est notre estimée de p_0. On veut indiquer sa précision.

L'intervalle de confiance à 95% sur p_0 est donné par la formule théorique $\hat{I}_n = \left[\overline{X}_n \pm z_{97.5\%} \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \right]$ et sa réalisation sur nos données est :

$$\left[\overline{x}_{100} \pm z_{97.5\%} \sqrt{\frac{\overline{x}_{100}(1-\overline{x}_{100})}{100}} \right] = \left[0.7 \pm 1.96 \sqrt{\frac{0.7\,(1-0.7)}{100}} \right]$$

Conclusion nette en faveur de l'église... $\simeq \left[0.7 \pm 2 \frac{\sqrt{0.21}}{10}\right] = [0.7 \pm 0.092] = [70.0\% \pm 9.2\%]$

Remarques à propos du corrigé du Quizz *2

Question de cours :

- on divise bien par $n-1$ et pas n...
- la notation consacrée pour l'estimateur de σ_0^2 est $\hat{\sigma}_n^2$ (et rien d'autre !)
- attention à bien distinguer les x_i des X_i : un estimateur est fonction des X_i mais une estimée est fonction des x_i
- la consistance correspond à une convergence en probabilité $\xrightarrow{\mathbb{P}}$ et non à une convergence en loi $\xrightarrow{\mathcal{L}}$
- je ne peux évidemment pas accepter les énoncés vagues comme " $\lim \hat{\sigma}_n^2 = \sigma_0^2$ " (en quel sens ?)

Lecture de tables :

- le petit nom de u_γ c'était $t_{q\,99\%}$; et pas $3_{99\%}$ (quantile d'une loi $\mathcal{N}(0,1)$)...
- beaucoup, beaucoup trop d'autre vous ont dit " $\mathbb{P}\{T \geq 2\} \in [90\%, 95\%]$ " : intuitivement, ça ne fait pas un peu beaucoup ? Une loi de Student n'est pas si loin d'une loi normale, or, $\mathbb{P}\{N \geq 2\} \simeq 5\%$ si $N \sim \mathcal{N}(0,1)$...
- l'encadrement parfois proposé $[0\%, 10\%]$ était insuffisant, je voulais $[5\%, 10\%]$.

Un intervalle de confiance :

- la modélisation semble mieux acquise, à l'interprétation du paramètre d'intérêt pas près

- certains recourent à la formule d'IC $[\bar{X}_n \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_n^2/n}]$ ce n'est pas impossible, mais il faut alors se souvenir que $\hat{\sigma}_n^2 = \frac{n}{n-1}\bar{X}_n(1-\bar{X}_n)$; or, tous ceux qui ont utilisé cette formule sont restés cois devant le calcul de l'estimée de la variance
- la loi de Student n'est appropriée que pour le cas d'observations gaussiennes !
- la bonne formule était $\left[\bar{X}_n \pm z_{97.5\%}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$ (où $n=100$). Il était imprécis, mais pas faux, de majorer $\sqrt{\bar{X}_n(1-\bar{X}_n)}$ par $1/2$.
- n'oubliez pas la présence du quantile $z_{97.5\%} = 1.96$!
- un calcul très exact à la calculatrice donne $[61.02\%, 78.98\%]$, de grâce, écrivez alors $[70.0\% \pm 9.0\%]$!!! Le $\pm 0.02\%$ de différence, on s'en fiche vu la taille de l'intervalle ...
- certains en sont restés à $[\bar{X}_n \pm \dots] = [\bar{X}_n \pm 0.09]$ et ne m'ont pas écrit d'intervalle explicite ... Bizarre ! À la fin des fins je veux un intervalle totalement numérique -
- Rappel: $\mathbb{P}\{\mu_0 \in [61\%, 79\%]\}$, ou des choses du même acabit, n'a pas de sens ; μ_0 y est ou pas ; en revanche, $\mathbb{P}\{\mu_0 \in [\bar{X}_n \pm \dots]\}$ a un sens, à comparer donc à $\mathbb{P}\{\mu_0 \in [\bar{x}_n \pm \dots]\}$, qui n'en a pas. $[\bar{X}_n \pm \dots]$ est un IC à 95% et vous vous en donnez la réalisation
- PS: selon la manière de coder, l'intervalle $[21\%, 39\%]$ était bien sûr aussi admissible.

Septième Partie

Introduction aux tests : tests de comparaison d'une moyenne à une valeur de référence

Version rédigée du cours

Résumé : Le chapitre précédent a montré comment, lorsque l'on estime, on peut préciser l'incertitude (la marge d'erreur) de l'estimée proposée. Cette quantification mettait en jeu des intervalles de confiance, dont la demi-longueur typique était proportionnelle à $s_n/\sqrt{n}$, l'estimée s_n de l'écart-type divisée par $\sqrt{n}$, où n est la taille d'échantillon. Le facteur de proportionnalité était donné par un quantile approprié (qui dépendait de la forme des lois sous-jacentes et du niveau de confiance attribué à l'intervalle de confiance, typiquement, 95 %).

Objectif : Pour déterminer si une valeur de référence proposée est compatible ou pas avec les données, on peut par exemple regarder si elle appartient ou non à un intervalle de confiance de niveau 95 % construit sur les données. Cela étant, cette méthode ne permet que de dire s'il semble y avoir compatibilité ou non, elle ne quantifie pas cette compatibilité. Or vous savez que les mathématiciens aiment quantifier... Le bon outil est formé par les tests. Nous verrons tout d'abord la méthodologie des tests, puis, dans un paragraphe destiné à ceux qui veulent approfondir la question, nous étudierons les liens entre tests et intervalles de confiance.

Motivation

Ce chapitre va vous aider à comprendre la mise en garde suivante :

> Il ne faut pas utiliser les statistiques comme les ivrognes utilisent les réverbères :
> pour s'appuyer et non s'éclairer.
> Lord Thorneycroft (homme politique britannique, 1909–1994)

En effet, nous allons voir que les tests statistiques peuvent éventuellement mettre en lumière certains aspects, éclairer partiellement une décision, mais que leur mise en œuvre et son résultat doivent toujours être supervisés par un homme ou une femme, à qui revient la responsabilité de prendre une décision. Plus précisément, nous verrons la force des préjugés et intentions politiques lors du choix des hypothèses à tester. Les statistiques ne doivent pas être source de technocratie : le statisticien n'a pas vocation à se substituer au dirigeant. Réciproquement, le dirigeant ne doit pas demander au statisticien de prendre une décision à sa place.

Plan de ce chapitre

Ce chapitre pourra vous sembler dense et d'accès difficile : c'est souvent le sentiment éprouvé face à une nouvelle théorie mathématique, d'autant plus qu'ici, les questions d'interprétation et de choix politiques ou subjectifs compliquent le propos. Le plan de ce cours est le suivant :

1. deux exemples introductifs, par lesquels on veut introduire le vocabulaire et les concepts fondamentaux liés aux tests statistiques ;

2. une présentation générale, et assez romancée, de la démarche liée à un test statistique ;
3. l'implémentation de cette démarche dans le cadre des tests de comparaison d'une moyenne à une valeur de référence (dans différents cadres mathématiques, selon que l'on teste une moyenne générale ou une proportion et selon que l'échantillon est de grande taille ou non).

1. Méthodologie des tests : deux exemples

1.1. A bas les tricheurs !

EXEMPLE 7.1 (Détection de tricheurs). Pour les introduire à la notion de hasard, un enseignant demande à ses étudiants de lancer 200 fois une pièce pour le lendemain et de noter les résultats. Il passe dans les rangs pour vérifier que les élèves ont effectué leurs devoirs sérieusement... et sermonne tous ceux qui, parmi leurs 200 lancers, n'ont pas eu 6 fois pile ou 6 fois face de suite. Pourquoi et comment ?

En fait, on peut montrer, par le calcul ou par simulations, qu'il y a toutes les chances, ici en l'occurrence une probabilité d'au moins 97 %, que sur 200 lancers d'une pièce équilibrée, le même côté sorte au moins 6 fois de suite. Or, les tricheurs, ceux qui ne lancent pas honnêtement leur pièce, ignorent souvent ce fait et écrivent des 0 ou 1 selon un certain schéma, qui n'est pas du hasard pur. En particulier, jamais, jamais, ils n'osent écrire 6 fois pile ou 6 fois face de suite : ils pensent, à tort, que c'est trop improbable et qu'ils vont se faire prendre s'ils le font.

Si l'élève a été honnête (c'est notre hypothèse H_0 de départ), alors avec probabilité 97 % il a eu 6 fois le même côté à suivre. Il a peut-être été malchanceux et fait partie des 3 % d'honnêtes qui n'ont pas eu la chance de voir apparaître cette séquence typique dans leurs lancers. En tout état de cause, au final, le professeur félicite les 97 % d'honnêtes chanceux et les malhonnêtes qui connaissaient le truc ; et sermonne les 3 % d'honnêtes malchanceux et les malhonnêtes qui ne connaissaient pas le truc.

S'il y a peu d'élèves, et si, comme c'est probable, aucun ne connaissait à l'avance le truc, c'est, somme toute, une manière de procéder qui ne crée que peu d'injustice. (Note : l'injustice la plus insoutenable est quand même la sanction d'élèves honnêtes !)

REMARQUE 7.1 (A retenir !). C'est exactement le principe des tests : on formule une certaine hypothèse de départ, on modélise les données, et on regarde quel est leur comportement typique sous l'hypothèse. Si les valeurs observées ne rentrent pas dans les clous de ce comportement typique, alors on part du principe que c'est l'hypothèse de départ qui est fausse. Evidemment, il y a une petite probabilité que ce ne soit pas le cas et qu'on l'ait rejetée à tort ; inversement, elle peut être fausse alors que le comportement observé est malgré tout acceptable ; mais comme toujours en statistique, on accepte un petit pourcentage d'erreur. Ici, on le voit, il y a deux types d'erreurs possibles, rejeter à tort ou conserver à tort. Ces deux erreurs n'ont par ailleurs pas le même poids : l'erreur la plus grave est de rejeter à tort l'hypothèse.

1.2. La version moderne de la danse de la pluie.

EXEMPLE 7.2 (Les faiseurs de pluie). En Beauce, il pleut habituellement 600 millimètres par an (mm/an). Des scientifiques à l'allure un peu mafieuse prétendent avoir

trouvé une technique pour augmenter localement les précipitations et ainsi faire des économies d'eau, en épandant un produit chimique révolutionnaire par avion sur les nuages. Des tests sur les années 1995–2002 ont donné les résultats suivants :

Année	95	96	97	98	99	00	01	02
Pluviométrie (mm/an)	606	592	639	598	614	607	616	586

Cette technique est-elle efficace ?

On dispose des données $x_1 = 606$, $x_2 = 592$, ..., $x_8 = 586$. La pluviométrie annuelle étant causée par de multiples facteurs et étant la somme de 365 pluviométries journalières, on soupçonne que la loi de sa distribution est gaussienne (avec ou sans épandage). Un météorologue nous confirmant par ailleurs que les conditions climatiques à un jour donné ne dépendent que de celles des deux ou trois derniers jours (c'est l'effet papillon), nous en déduisons que les observations peuvent être considérées comme indépendantes. Etant réalisées dans les mêmes conditions d'épandage, elles sont donc en outre identiquement distribuées. On peut donc modéliser les valeurs observées comme la réalisation de $X_1, \ldots, X_8$, variables aléatoires indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0 cependant inconnus.

Les données conduisent aux estimées respectives $\overline{x}_8 = 607.25$ pour la moyenne et $s_{x,8} = 16.5$ pour l'écart-type. On sent d'ores et déjà que le gain par rapport à la valeur de référence de $\mu_{\text{ref}} = 600$ ne semble pas significatif. On va quantifier cette impression.

On pose comme hypothèse de départ $H_0 : \mu_0 = 600$; c'est une hypothèse de prudence, que l'on ne veut lâcher que si vraiment les données suggèrent qu'elle n'est pas valable. On ne va quand même pas suivre ces scientifiques mafieux sur un coup de tête ! L'hypothèse alternative, vers laquelle on se tournerait si H_0 se révélait en contradiction avec les données et leur modélisation, est celle que suggèrent nos interlocuteurs, à savoir $H_1 : \mu_0 > 600$.

Or, si H_0 est vraie, avec les notations du chapitre précédent, on a

$$T_8 = \sqrt{8}\,\frac{\overline{X}_8 - 600}{\sqrt{\widehat{\sigma}_8^2}} \;\sim\; \mathcal{T}_7 \,.$$

T_8 va être notre statistique de test, elle en a en effet toutes les qualités : c'est une variable aléatoire, que l'on peut calculer entièrement à partir des observations, et dont on connaît la loi sous H_0.

On veut déterminer un intervalle de valeurs typiques pour T_8 (et on verra ensuite si la valeur calculée pour T_8 sur les observations est dans cet intervalle ou pas, ce qui conduira à la conservation ou au rejet de H_0). La loi de $\overline{X}_8$ est centrée autour de μ_0 : si $H_1 : \mu_0 > 600$ est vraie, alors $\overline{X}_8$ et donc T_8 tendront à prendre des valeurs plus grandes que celles qu'elles ne prendraient respectivement si $H_0 : \mu_0 = 600$ était vraie. On donc affirmer que les valeurs de T_8 en-dessous d'un certain seuil sont typiques de H_0, et que celles qui sont au-dessus indiquent une préférence pour H_1. On fixe ce seuil de telle sorte que 95 % du temps, si H_0 est vraie, alors T_8 est en-dessous de ce seuil : ainsi, il doit être pris égal au quantile $t_{7,95\,\%}$, qui vaut 1.895 comme on peut le lire sur la table de la loi de Student. Cela correspond à un niveau d'erreur (risque) de 5 % sous H_0. L'intervalle $]1.895, +\infty[$ obtenu est appelé l'intervalle de rejet. Si la valeur calculée sur les observations pour la statistique de test tombe dedans, alors on rejette H_0.

Sur les données, on a la réalisation suivante de T_8 :

$$\sqrt{8}\,\frac{\overline{x}_8 - 600}{s_{x,8}} = \sqrt{8}\,\frac{607.25 - 600}{16.5} = 1.24\ .$$

Cette valeur n'étant pas dans l'intervalle de rejet, on conserve H_0, et on en déduit que pour le moment, on n'a aucune raison de croire en une augmentation significative de la pluviométrie. On pourra se reporter à la figure 33 pour une illustration graphique de ce raisonnement.

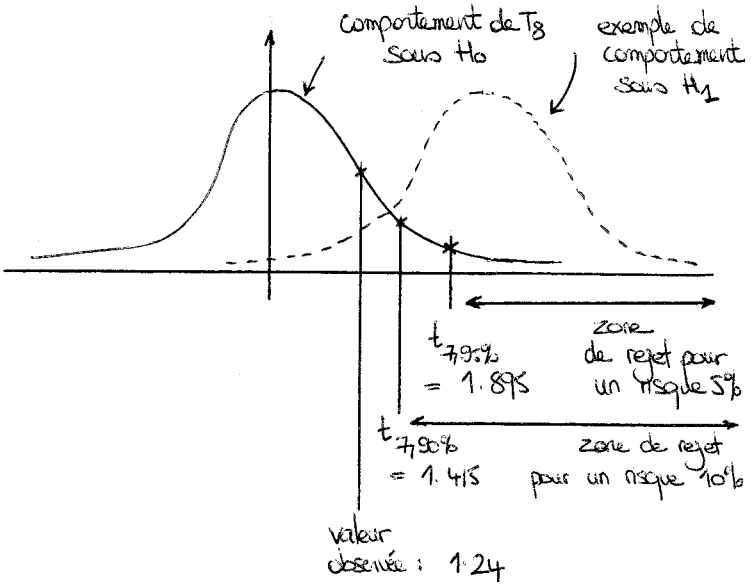


FIGURE 33. Zones de rejet aux risques 5 % et 10 % et valeur réalisée de la statistique de test T_8 pour les hypothèses $H_0 : \mu_0 = 600$ contre $H_1 : \mu_0 > 600$.

Puisque $t_{7,90\,\%} = 1.415$, la conclusion aurait été la même avec un intervalle de valeurs typiques contenant seulement 90 % de la probabilité totale (avec un test de niveau d'erreur de 10 %). Voilà où commence la quantification : on dira alors que la P-valeur du test est supérieure ou égale à 10 %. (On définit cette dernière de manière plus formelle dans la suite de ce cours : on l'interprète comme le degré de crédibilité de H_0.)

Conclusion stratégique : Le plus important maintenant est de tirer une décision pratique de cet enseignement statistique. Ici, on fera remarquer aux scientifiques mafieux que les données et arguments qu'ils procurent ne sont pas assez décisifs au vu du coût de leur méthode et qu'en conséquence, on décline pour le moment leur offre. On pourra cependant les prier de nous recontacter dans quelques temps, lorsque leurs données et/ou leur méthode seront plus probantes (voir la remarque qui suit).

REMARQUE 7.2 (Toute vérité ne dure qu'un temps...). En effet, ici, il serait tout à fait possible que μ_0 soit égale à 608 par exemple et que H_0 soit fausse malgré tout. Dans ce cas, en continuant les relevés pendant encore quelques années [10], on pourra rejeter H_0 et

10. Ceux qui veulent exercer leur esprit mathématique noteront que, par loi des grands nombres, lorsque n est grand,

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_{\text{ref}}}{\sqrt{\widehat{\sigma}_n^2}} \sim \sqrt{n}\,\frac{\Delta}{\sigma_0}$$

passer à H_1. Retenez que plus l'écart Δ (par ailleurs inconnu) entre la vraie valeur μ_0 et la valeur de référence μ_{ref} testée (ici, $\mu_{\text{ref}} = 600$) est petit, plus il faut d'observations pour rejeter H_0.

EXERCICE 7.1. Cet exercice montre comment les scientifiques mafieux pourraient embobiner quelqu'un qui n'a pas eu la chance d'avoir votre excellente formation en statistique. Ainsi, vos interlocuteurs pourraient rétorquer que le test de

$$H_0 : \quad \mu_0 = 615 \qquad \text{vs.} \qquad H_1 : \mu_0 < 615$$

conserve H_0 et qu'il est donc clair que leur technique est performante. Prouvez ce fait et commentez-le (indiquez les conséquences éventuelles à en tirer, ou pas).

CORRECTION 7.1. On considère cette fois-ci la statistique de test

$$T_8' = \sqrt{8}\,\frac{\overline{X}_8 - 615}{\sqrt{\widehat{\sigma}_8^2}}\,;$$

sous H_0, on a par définition des lois de Student que $T_8' \sim \mathcal{T}_7$. Comme, quelle que soit la valeur de μ_0, la moyenne empirique $\overline{X}_8$ est proche de μ_0, on a que sous $H_1 : \mu_0 < 615$, la statistique de test T_8' a tendance à prendre des valeurs plus petites que sous H_0 (voir la figure 34). On rejette donc H_0 lorsque la valeur réalisée de T_8' passe en-dessous d'un certain

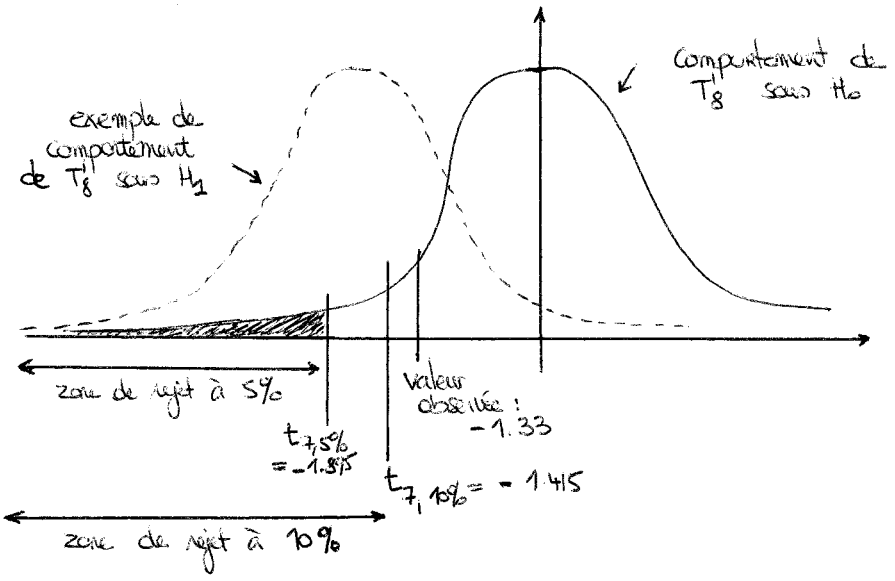


FIGURE 34. Zones de rejet aux risques 5 % (grisée) et 10 % et valeur réalisée de la statistique de test T_8' pour les hypothèses $H_0 : \mu_0 = 615$ contre $H_1 : \mu_0 < 615$.

seuil, à savoir $t_{7,5\%} = -1.895$ pour un niveau d'erreur de 5 % seulement et $t_{7,10\%} = -1.415$ pour un niveau d'erreur de 10 %. Or, cette valeur réalisée vaut sur les données

$$\sqrt{8}\,\frac{\overline{x}_8 - 615}{s_{x,8}} = \sqrt{8}\,\frac{607.25 - 615}{16.5} = -1.33$$

et est au-dessus de ces seuils. Dans les deux cas, on conserve alors $H_0 : \mu_0 = 615$.

et dépasse donc tout seuil fini au bout d'un certain rang, et ce, d'autant plus vite que Δ est grand.

Commentons cette conclusion statistique. On retrouve là encore, comme l'indiquait une remarque précédente, le fait qu'on conserve H_0 peut-être ou sans doute faute de données suffisantes. En conséquence de quoi, notre méthodologie générale de test sur la moyenne nous a conduit à conserver [11] à la fois $H_0 : \mu_0 = 600$ et $H_0 : \mu_0 = 615$. En somme, chacun campe sur ses préjugés et les données ne contredisent suffisamment aucune de ces deux hypothèses, de sorte qu'on peut chacune les tenir pour vraies. Cela étant, cela indique surtout que l'on se trouve dans une zone grise d'incertitude et que l'on ne peut rien conclure fermement. Il s'agit d'une question de confiance ou de défiance envers les scientifiques mafieux, mais les statistiques ne tranchent pas cette question.

REMARQUE 7.3 (A propos de la forme de l'intervalle de rejet). On a mené ici (tant dans le développement que dans la correction de l'exercice) des tests dits unilatères : on ne rejetait, selon le couple H_0 et H_1 retenu, qu'au-dessus (premier test) ou en-dessous (second test) d'une valeur. On verra dans la suite que lorsque H_0 et H_1 sont respectivement de la forme $\mu_0 = \mu_{\text{ref}}$ et $\mu_0 \neq \mu_{\text{ref}}$, on aura affaire à des tests dits bilatères, avec des zones de rejets généralement symétriques, correspondant à un rejet au-dessous d'une première valeur et au-dessus d'une seconde valeur.

11. C'est fondamentalement dû au faible écartement entre les deux valeurs de référence 600 et 615, comparé au nombre, lui aussi faible, de données disponibles.

2. Méthodologie des tests : théorie générale

Le cadre est le cadre désormais habituel, celui de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbb{P}_{\theta_0}$, dont on sait seulement qu'elle appartient à un ensemble de lois possibles $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$. On a construit cette modélisation au vu des données $x_1, \ldots, x_n$.

La construction et la réalisation d'un test se fait selon les étapes suivantes, que nous allons désormais étudier chacune en détails (sauf l'étape préliminaire de modélisation, étudiée dans la partie 2).

Méthodologie des tests statistiques

1. Etape préliminaire : modélisation du problème.
2. Détermination des hypothèses à tester H_0 et H_1.
3. Choix d'une statistique de test T_n, dont on connaît la loi sous H_0.
4. Etude du comportement de T_n sous H_1 et déduction de la forme de la zone de rejet.
5. Calcul de cette zone R pour un niveau fixé puis confrontation aux données ; et/ou calcul de la P-valeur du test sur les données.
6. Conclusion statistique : conservation ou rejet de l'hypothèse de départ H_0 et commentaire éventuel sur la P-valeur.
7. Conclusion stratégique : décision que l'on va prendre une fois éclairé par le résultat statistique.

J'attire votre attention sur le point 7. de cette méthodologie. Il s'agit là de votre valeur ajoutée par rapport au traitement mathématique et du seul intérêt de mettre en œuvre un test statistique : pour prendre une décision. Il faut donc toujours indiquer cette dernière.

2.1. Déterminer les hypothèses à tester (première partie théorique). On définit deux sous-ensembles disjoints de Θ, que l'on note Θ_0 et Θ_1, et l'on formule les hypothèses

$$H_0 : \quad \theta_0 \in \Theta_0 \qquad \text{et} \qquad H_1 : \quad \theta_0 \in \Theta_1 \ .$$

Exemple 7.3. Par exemple, H_0 peut être "la loi est d'espérance égale à 1", soit $H_0 : \mu_0 = 1$; ou "la loi a une espérance positive", soit $H_0 : \mu_0 \geqslant 0$; ou encore, "la variance de la loi est plus petite que 10", soit $H_0 : \sigma_0 \leqslant 10$.

Remarque 7.4 (Disjonction mais pas partition). H_0 et H_1 doivent être disjointes, mais ne sont pas nécessairement la négation l'une de l'autre, comme on l'a vu dans l'exemple des faiseurs de pluie, où l'on ne confrontait pas $\mu_0 \leqslant 600$ contre $\mu_0 > 600$, ni $\mu_0 = 600$ contre $\mu_0 \neq 600$, mais bien $\mu_0 = 600$ contre $\mu_0 > 600$. Le contexte, les connaissances a priori, etc., permettent de déterminer l'ensemble des cas intéressants [12] à considérer.

12. En ce sens, la statistique est une sous-discipline un peu plus expérimentale que les mathématiques pures, au sens où l'intuition permet de ne pas faire une disjonction exhaustive de cas. Un peu comme dans la blague suivante : « Quelle est la différence entre un physicien et mathématicien ? Lorsque l'on lance un

Principe des tests : Un test confronte le modèle postulé par H_0 à des observations ; si la confrontation se passe mal, i.e., si elle indique que les données semblent contredire H_0, alors on passe au modèle indiqué par H_1. (Le paragraphe 2.6 ci-dessous indique les recettes pour déterminer en pratique qui est H_0 et qui est H_1.)

2.2. Choisir une statistique de test.

DÉFINITION 7.1 (Statistique de test). *Une statistique de test T_n est une variable aléatoire ne dépendant que du couple (Θ_0, Θ_1) et des observations $X_1, \ldots, X_n$. Les statistiques de test intéressantes sont celles dont on connaît la loi sous H_0.*

REMARQUE 7.5 (Liens avec les estimateurs ?). Une statistique de test est définie presque comme un estimateur, à part qu'elle peut dépendre de la formulation des hypothèses H_0 et H_1. Ainsi, dans l'exemple des faiseurs de pluie, les statistiques de test dépendaient des valeurs à tester, qui étaient successivement $\mu_{\text{ref}} = 600$ et $\mu_{\text{ref}} = 615$. De plus, la charge psychologique est différente, un estimateur se devant d'avoir une loi concentrant les valeurs probables autour de la quantité d'intérêt, tandis que la statistique de test peut avoir n'importe quelle loi, tant qu'on connaît son comportement sous H_0, et que son comportement sous H_1 est sensiblement différent de celui sous H_0.

C'est vraiment le point crucial de la construction du test. Vous vous demandez évidemment comment trouver la bonne statistique de test : ce sera l'objet de la suite de ce chapitre, et des deux suivants ! Nous ferons œuvre de botanique statistique et nous dresserons la liste, pour différentes situations concrètes, des statistiques de test à considérer et des intuitions qui les soutiennent.

2.3. Détermination de la forme de la zone de rejet. La plupart du temps, les zones de rejet R sont de l'une des formes suivantes :

$$]r, +\infty[\,, \qquad]-\infty, r[\,, \quad \text{ou} \quad]-\infty, r[\,\cup\,]r', +\infty[\,.$$

Pour voir laquelle convient, on étudie le comportement de T_n sous H_1. Si T_n tend à prendre des valeurs plus grandes (respectivement, plus petites) quand la vraie loi est l'une de celles de Θ_1, alors on prend une zone de rejet R de la forme $]r, +\infty[$ (respectivement, $]-\infty, r[$). Si la balance peut pencher des deux côtés, alors, à défaut, on prend une zone de rejet symétrique $R =]-\infty, r[\,\cup\,]r', +\infty[$ (où r et r' peuvent être différents, mais sont souvent en pratique l'opposé l'un de l'autre).

2.4. Calcul de la zone de rejet pour une erreur de première espèce fixée.

DÉFINITION 7.2 (Erreur de première espèce). *On appelle erreur de première espèce α d'un test la probabilité de rejeter à tort H_0. Mathématiquement, si l'on appelle R la zone de rejet, alors*

$$\alpha = \sup_{\theta_0 \in \Theta_0} \mathbb{P}\Big\{ T_n \in R \;\Big|\; X_1, \ldots, X_n \text{ i.i.d. selon } \mathbb{P}_{\theta_0} \Big\},$$

où le conditionnement rappelle simplement que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi $\mathbb{P}_{\theta_0}$, avec $\theta_0 \in \Theta_0$.

objet par la fenêtre, le premier répond qu'il tombe, et le second considère trois cas, selon qu'il tombe, reste suspendu sans bouger, ou s'élève. »

Le principe est que l'on prend R suffisamment petit pour que α soit inférieure ou égale à une certaine valeur seuil. Généralement, on se fixe ce seuil à 5 %. Dans tous les cas, c'est un paramètre d'utilisateur : c'est votre responsabilité [13] de le choisir. 5 % est le niveau standard (en biologie ou médecine), mais s'il s'agit de décisions sensibles et que vraiment, on craint de rejeter H_0 à tort, on peut prendre un seuil plus faible, de 1 %.

EXEMPLE 7.4. Rassurez-vous, la plupart du temps, pour la simplicité du propos et des questions d'examen, il sera facile de calculer ce supremum et d'en déduire la zone de rejet. Ainsi, pour les faiseurs de pluie, dans le premier test où $H_0 : \mu_0 = 600$, les lois correspondant à H_0 étaient toutes les lois $\mathcal{N}(600, \sigma_0^2)$, avec $\sigma_0^2 > 0$ quelconque. Sous toutes ces lois, la statistique T_8 suivait une loi de Student $\mathcal{T}_7$, de sorte que l'erreur de première espèce, vu la forme de la zone de rejet $R =]r, +\infty[$, vaut

$$\alpha = \sup_{\sigma_0^2 > 0} \mathbb{P}\left\{ T_8 > r \;\middle|\; X_1, \dots, X_n \text{ i.i.d. selon } \mathcal{N}(600, \sigma_0^2) \right\} = \mathbb{P}\{U > r\}$$

où U est une variable aléatoire distribuée selon la loi $\mathcal{T}_7$. On avait fixé r tel que $\alpha = 5\,\%$, en prenant $r = t_{7,95\%}$.

Si H_0 avait été $\mu_0 \leqslant 600$, l'ensemble des lois possibles aurait été toutes les lois $\mathcal{N}(\mu_0, \sigma_0^2)$ où $\mu_0 \leqslant 600$ et $\sigma_0^2 > 0$ est quelconque. Dans ce cas, il est intuitif que le pire cas dans le calcul de α est celui où $\mu = 600$; dans tous les autres cas, lorsque $\mu_0 < 600$, les valeurs de T_8 tendent à être plus petites que lorsque $\mu_0 = 600$, et donc la probabilité que T_8 dépasse le seuil r est plus faible que dans le cas $\mu_0 = 600$.

REMARQUE 7.6 (Un test inutile mais d'erreur de première espèce faible). A méditer : le test qui accepte inconditionnellement H_0 a une erreur de première espèce $\alpha = 0$, qui est donc plus petite que toute valeur seuil choisie par l'utilisateur. Mais ce test est inutile : qui ne risque rien n'a rien. A force de se focaliser sur la possibilité de rejeter l'hypothèse H_0 à tort, on la conserve tout le temps et on n'apprend jamais rien.

Ce mauvais test est à rapprocher des intervalles de confiance égaux à l'ensemble des valeurs possibles ($\mathbb{R}$ ou Θ)... Ils ont toujours raison, mais à quel prix !

2.5. Alternative : Calcul de la P-valeur. Au lieu de se fixer un seuil d'erreur et de conclure de manière binaire à la conservation ou au rejet de H_0, il est de loin préférable de quantifier l'attachement à H_0. Cela formera en un certain sens un degré de crédibilité de H_0. (C'est notamment ainsi que procède SPSS.) L'indice quantitatif utilisé est le suivant.

DÉFINITION 7.3 (P-valeur). *Etant donné un test statistique et des données, la* P*-valeur* p *est l'erreur maximale* α *telle que le test considéré accepterait encore la valeur réalisée de la statistique de test sur les données. Elle peut être interprétée comme un indice de crédibilité de* H_0*. Une faible* P*-valeur conduit au rejet de* H_0*.*

EXEMPLE 7.5. Calculons la P-valeur associée au (premier) test et aux données du paragraphe correspondant aux faiseurs de pluie. On y a vu que les tests d'erreurs de première espèce $\alpha = 5\,\%$ et $\alpha = 10\,\%$ conservaient encore l'hypothèse $H_0 : \mu = 600$ au vu des données. Ainsi, on sait que $p \geqslant 10\,\%$. En fait, le test qui aurait comme seuil de

13. Cela étant, le paragraphe suivant sur la P-valeur vous montrera comment les mathématiciens évitent de devoir choisir ce paramètre d'utilisateur et proposent au passage une quantification des choses par ce qu'on appellera la P-valeur.

rejet r la valeur observée pour la statistique 1.24 conserverait encore H_0, mais toute valeur inférieure ou égale conduirait au rejet de H_0. La P-valeur p est donc donnée par l'équation

$$t_{7,1-p} = 1.24 \qquad \text{soit} \qquad p = 12.7\,\% \ .$$

Dit autrement, p est la masse de probabilité associée à cette zone de rejet maximale. (Note : la résolution de l'équation $t_{7,1-p} = 1.24$ se fait en utilisant un logiciel ; les tables que je vous ai fournies ne permettent que d'encadrer p entre 10 % et 20 %.)

Une fois qu'on a la P-valeur, on peut la comparer au seuil dont on a parlé au paragraphe précédent et qui vaut typiquement 5 % (sauf lorsque rejeter à tort H_0 est très coûteux ou aurait des conséquences dramatiques, auquel cas un seuil de 1 % peut être préféré). Si la P-valeur est en-dessous de ce seuil, on rejette H_0, et sinon, on la conserve. Notez que lorsque le seuil est de 5 %, les décisions prises selon que $p = 4.9\,\%$ ou 5.1 % sont en principe différentes, mais en pratique... Dans tous les cas, grâce à la quantification apportée par la P-valeur, vous savez au moins si le rejet ou la conservation vont se faire largement ou sur le fil.

Mon conseil : si ce n'est pas vous le chef, donnez juste la P-valeur, puisque c'est une quantité objective. Ce sera au chef de l'interpréter et de prendre la décision subjective en fonction de cette quantification objective. Si c'est vous le chef, refusez qu'un statisticien prenne une décision à votre place en vous disant simplement si son test conserve ou rejette H_0 : réclamez-lui l'indication de la P-valeur. On résume cette discussion par la grille méthodologique de la figure 35.

Remarque 7.7 (Interprétation de la P-valeur comme un indice de crédibilité de H_0). On ne va pas se le cacher, cette notion de P-valeur est la notion-clé du cours sur les tests, et c'est elle aussi qui a suscité, suscite, et suscitera les incompréhensions de générations d'étudiants...

Imaginez un test avec une erreur de première espèce $\alpha \leqslant 0.1\,\%$. Ce test aura une tendance très forte à conserver H_0 ; qu'il conserve H_0 ne signifie absolument rien, et il est rare qu'il rejette H_0 tant il est conservateur. C'est un peu comme le fait de rester mariés toute sa vie au Moyen-Age : à moins de pouvoir payer des experts de droit canon pour faire reconnaître la nullité du mariage, il n'y avait aucune chance de se séparer. Le fait de se séparer prouvait alors certes qu'on ne s'aimait plus, mais celui de rester ensemble ne prouvait rien quant à lui. Pensons maintenant aux couples actuels : on peut penser qu'une bonne part de ceux qui ne s'aiment plus divorcent (mais d'autres restent ensemble par peur, facilité, ou à cause du reste de conventions). Ce dernier cas de figure correspond aux tests avec une erreur de première espèce plus grande, de l'ordre de 5 % ou 10 %. Il arrive plus souvent que l'on rejette H_0 (évidemment, c'est parfois à tort). Au final, quand on indique la P-valeur p, il faut se représenter le test d'erreur de première espèce $\alpha = p$ correspondant : ce test accepte H_0, par définition de la P-valeur. Si $\alpha = p$ est très faible, H_0 n'est acceptée que parce que le divorce d'avec elle n'est presque pas permis, et donc, autant la rejeter ; si $\alpha = p$ a une valeur plus raisonnable, c'est que H_0 ne semble pas si critiquable que cela et mieux vaut la garder.

On retiendra que la P-valeur p forme en ce sens un degré de crédibilité de H_0 face à H_1. On rejette d'autant plus fermement H_0 que p est petite et on tend d'autant plus à conserver H_0 que p est grande.

Remarque 7.8 (Seconde interprétation). Ceux d'entre vous qui ont bien assimilé ce qui précède seront d'accord avec cette assertion, qui éclaire elle aussi l'interprétation de

Conclusion statistique en fonction de la P*-valeur*

Rappel : la P-valeur correspond au degré de crédibilité de l'hypothèse H_0 face à H_1.

Premier cas : vous êtes le statisticien et c'est votre supérieur qui doit prendre la décision finale.
- Ce n'est donc pas à vous d'endosser la responsabilité de la décision.
- Refusez que votre supérieur se réfugie derrière les statistiques : ne lui dites pas que votre test conserve ou rejette H_0.
- Précisez-lui plutôt la P-valeur que vous avez calculée et laissez-le prendre ensuite la décision en son âme et conscience.

Second cas : vous êtes le supérieur hiérarchique d'un statisticien et vous êtes responsable d'une décision à prendre.
- C'est à vous que l'on demandera des comptes, aussi, demandez à votre subordonné d'éclairer le mieux possible votre décision.
- En conséquence, refusez que ce dernier la prenne à votre place en vous disant simplement si son test conserve ou accepte H_0 et réclamez-lui l'indication d'une P-valeur p.
- Fixez-vous un seuil d'interprétation : 1 % si les conséquences d'un rejet à tort de H_0 seraient dramatiques, 5 % sinon dans les cas plus standards.
- Ensuite,
  + Si p est bien plus petit que ce seuil, rejetez H_0 fermement ;
  + Si p est bien plus grand que ce seuil, campez fermement sur H_0 ;
  + Si p est autour de ce seuil, pas de chance, un abîme de doute se dessine devant vous et il va falloir soit se décider tout de suite et prendre un risque (avec la perspective d'assumer une erreur éventuelle), ou commander une étude complémentaire et reporter la décision.

FIGURE 35. Grille méthodologique d'utilisation stratégique de la P-valeur.

la P-valeur comme un indice de crédibilité de H_0 : la P-valeur la probabilité, calculée en supposant que l'hypothèse de départ H_0 est vraie, d'obtenir, si l'on répétait l'expérience, une valeur pour la statistique de test aussi ou plus contradictoire envers H_0 que la valeur initialement observée.

2.6. Déterminer les hypothèses à tester (seconde partie pratique). Vous n'êtes pas sans avoir remarqué, dans ce qui précède, la dissymétrie entre H_0 et H_1 :
- H_0 tend à être conservée, on ne la rejette que quand il y a de bonnes raisons de le faire, quand il y a désaccord grave avec les données ;
- en particulier, quand on construit un test, on détermine la taille de la zone de rejet en fonction de l'erreur de première espèce α (qui est la probabilité de rejeter H_0 à tort) ; on n'utilise H_1 que pour choisir la forme de cette zone de rejet ;

– ce n'est qu'ensuite, et pas[14] dans ce cours, qu'on s'intéresse réellement à H_1 en calculant l'erreur de deuxième espèce β, i.e., la probabilité de conserver H_0 à tort (alors que H_1 est vraie),

$$\beta = \sup_{\theta_1 \in \Theta_1} \mathbb{P}\Big\{ T_n \notin R \;\Big|\; X_1, \dots, X_n \text{ i.i.d. selon } \mathbb{P}_{\theta_1} \Big\}.$$

Voici, par conséquent, des pistes pour formuler sur un problème concret les hypothèses H_0 et H_1. On prend pour H_0 :

– une hypothèse solidement établie jusqu'à présent (une théorie scientifique communément admise pour l'instant en physique, par exemple, ou un taux habituellement constaté de réponse à une offre) ; on ne veut évidemment la rejeter que si l'on a d'excellentes raisons de le faire, sinon on aura l'air ridicule !
– une hypothèse de prudence, lorsque choisir H_1 à tort est moins grave ou moins coûteux que de rester sur H_0 à tort. Ainsi, on part par exemple du fait H_0 que la commercialisation d'un nouveau produit n'est pas rentable et on espère pouvoir rejeter l'hypothèse et passer à l'alternative H_1 qu'elle est rentable. C'est mieux dans ce sens que dans l'autre, toujours à cause de la tendance qu'a un test à conserver H_0 : on risque uniquement de ne pas faire davantage de profits (en conservant H_0 à tort), ce dont personne ne s'apercevra, alors qu'on écarte avec bonne probabilité (majorée par l'erreur de première espèce) le cas de pertes tangibles (le cas où on rejetterait H_0 à tort).
– une hypothèse à laquelle on est subjectivement attaché (H_0 dénote alors un parti pris : le H_0 des associations de consommateurs n'est souvent pas le même que celui des groupements d'industriels !) ; dans ce cas, le choix de H_0 est hautement politique, puisqu'on a tendance à la conserver ; il faudra bien faire attention à l'absence éventuelle de conclusion si H_0 est conservée ; prenez du recul...

Pour H_1 on prend souvent la négation de H_0, mais pas toujours. Si l'on a un a priori, on peut l'exploiter en restreignant H_1 et en excluant des cas à la fois de H_0 et de H_1. Pensez par exemple au test des faiseurs de pluie, qui considérait $H_0 : \mu = 600$ contre $H_1 : \mu_0 > 600$ et laissant de côté les cas où $\mu_0 < 600$.

Principe à retenir : Les tests ayant tendance à conserver H_0, on n'apprend vraiment quelque chose que lorsque l'on rejette H_0. C'est un progrès négatif. Ainsi, si l'on veut prouver un certain fait, on prend pour H_0 sa négation et pour H_1 ce fait, et l'on espère qu'un test bien construit rejettera H_0.

Un test dit en effet : « ce n'est pas impossible » ou « c'est impossible ». H_0 est vraie jusqu'à preuve du contraire, et cette preuve du contraire peut être apportée par un test (et des données).

EXEMPLE 7.6 (Comment procède l'industrie pharmaceutique pour savoir qu'un nouveau médicament est efficace ?). Il y a deux méthodes. La première est de le tester contre un placebo : on prend comme H_0 le fait que le nouveau médicament ne soit pas meilleur que placebo. Cela étant, ce n'est pas bien difficile pour un laboratoire pharmaceutique de concevoir une substance capable de battre l'effet thérapeutique de gélules vides !

14. Ouf ?

La seconde, qui semble déjà plus raisonnable, est de tester le nouveau médicament contre un médicament déjà établi. Que prendre pour H_0 ? En grossissant le trait, les laboratoires ont tendance à choisir "le nouveau médicament est au moins aussi efficace que celui qui est établi", puisque cela les arrange, les tests ayant tendance à conserver H_0. Une autorité indépendante, surtout si le nouveau médicament est plus cher, requerrait quant à elle la mise en œuvre de tests partant d'une hypothèse H_0 du type "le nouveau médicament n'apporte aucune amélioration par rapport à ceux qui existent", à charge pour les laboratoires de prouver que l'apport du nouveau médicament est tangible.

Dans la pratique, l'autorisation de mise sur le marché est décidée de manière intrinsèque (par rapport à l'efficacité contre placebo) et non de manière globale et relative (par rapport aux médicaments déjà existants). On peut le déplorer ; les industriels jouant quant à eux la carte de l'égalité de traitement de toutes les molécules, nouvellement arrivées ou déjà en position sur le marché.

Pour le côté historique de la chose, je suis profondément persuadé que beaucoup d'avancées de la médecine sont dues en partie aux statistiques. Les protocoles de tests (au moins ceux contre placebo) sont, grâce aux remarques des mathématiciens, très balisés et permettent une interprétation sûre des résultats. C'est en particulier parce que les mathématiciens ont fait remarquer qu'il fallait que les échantillons aient une taille n suffisamment grande et que les répartitions entre groupe placebo et groupe traité soient faites au hasard.

3. Tests de comparaison d'une moyenne à une valeur de référence

On détaille ici la méthodologie générale dans différents cas particuliers. Distinguez bien la vraie valeur inconnue, μ_0 ou p_0, de celle à laquelle on la compare : μ_{ref} ou p_{ref}.

3.1. Cas d'une loi générale et d'une taille d'échantillon grande. Voici, dans ce cas, la statistique de test et son comportement sous différentes alternatives.

PRINCIPE 7.1. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*
Données : $x_1, \dots, x_n \in \mathbb{R}$
Modélisation associée : observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi, admettant un moment d'ordre deux et d'espérance notée μ_0
Hypothèse H_0 : $\mu_0 = \mu_{ref}$
Statistique de test :

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_{ref}}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \rightharpoonup \mathcal{N}(0,1)$
Comportement sous H_1 : lorsque $\mu_0 > \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

La preuve de ce principe découle de faits déjà vus précédemment. La convergence sous H_0 est énoncée par le théorème 5.1. Le comportement sous H_1 découle du fait que $\overline{X}_n$ étant un estimateur consistant de μ_0, il s'en rapproche (et de même $\widehat{\sigma}_n^2$ pour la variance commune σ_0^2 des $X_1, \dots, X_n$), de sorte que le comportement de T_n sous H_1 tend à être celui de $\sqrt{n}(\mu_0 - \mu_{ref})/\sigma_0$.

Le test proposé est asymptotique (quand on déterminera les régions de rejet, elles seront fondées sur les quantiles de la loi normale et ne seront qu'approximativement du niveau requis) ; on négligera ce fait en pratique dès lors que la taille d'échantillon n sera suffisamment grande.

EXERCICE 7.2 (L'évolution du pouvoir d'achat, cf. question 2 exercice II de l'examen de rattrapage 2007). On veut quantifier l'évolution du pouvoir d'achat. En 2004, le montant moyen des achats hors produits de nécessité (par exemple, voyages, abonnement internet, téléphonie mobile, spectables ; par opposition à l'alimentation, au logement, à la voiture, au chauffage, à l'eau, à la téléphonie fixe) était de 637 euros par mois et par foyer selon les données collectées par l'INSEE auprès de plusieurs millions d'habitants lors du recensement partiel. Le gouvernement commande un sondage téléphonique auprès d'environ 2 000 foyers afin de déterminer ce montant en 2008 ; seuls 1 837 foyers arrivent le calculer et à l'indiquer à l'opérateur. Le montant moyen qu'ils déclarent est de 598 euros (avec un écart-type dans les montants de 254 euros). On suppose une inflation de 2 % par an. Le montant consacré aux achats hors produits de nécessité a-t-il significativement diminué, comme tant le déplorent, ou le montant moyen observé de 598 euros est-il compatible avec les données de 2004 ?

CORRECTION 7.2. On commence par modéliser les données $x_1, \ldots, x_{1837}$; vu la manière de les recueillir, elles sont la réalisation d'observations $X_1, \ldots, X_{1837}$ indépendantes et identiquement distribuées selon une certaine loi, admettant un moment d'ordre deux (et même bornée, en fait). On note μ_0 et σ_0 son espérance et son écart-type. Ils sont inconnus mais on a cependant les estimées respectives $\bar{x}_{1837} = 598$ et $s_{1837} = 254$. Notre paramètre d'intérêt est ici μ_0, c'est le montant mensuel moyen (sur l'ensemble des foyers français) consacré aux achats hors produits de nécessité.

Cette modélisation préliminaire étant effectuée, on peut passer au test d'hypothèses proprement dit. Attention, à cause de l'inflation, la valeur de référence est le montant de 2004 convertis en euros courants (de 2007), soit $\mu_{\text{ref}} = 637 \times 1.02^3 \approx 676$. Notez que ceci est une valeur de référence puisqu'elle est calculée sur une très grosse part de la population (plus de 10 % des Français, le recensement étant désormais partiel mais annuel) : c'est une estimation tellement précise que pour une fois, on peut dire qu'elle est égale à la vraie valeur sous-jacente. La question (et le discours ambiant sur le pouvoir d'achat en berne) suggèrent ici un test unilatère de $H_0 : \mu_0 = 676$ contre $H_1 : \mu_0 < 676$.

La statistique de test T_{1837} est fournie par le principe énoncé ci-dessus,

$$T_{1837} = \sqrt{1837}\, \frac{\overline{X}_{1837} - 676}{\sqrt{\widehat{\sigma}^2_{1837}}} \; ;$$

vu la taille de l'échantillon, sous H_0, la loi de T_{1837} est proche de celle de la loi normale $\mathcal{N}(0,1)$.

Sous H_1, puisque $\overline{X}_{1837}$ aura tendance à être proche de $\mu_0 < 676$, la statistique T_{1837} tend à prendre des valeurs plus petites que sous H_0. La zone de rejet est donc de la forme $]-\infty, r[$. On illustre tous les calculs qui suivent par la figure 36.

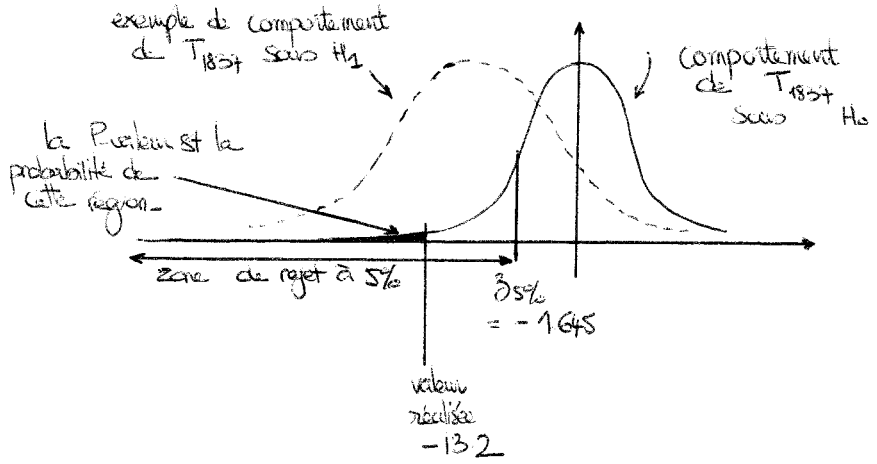


FIGURE 36. Illustration du raisonnement et des calculs associés à l'exercice 7.2.

Construisons dans un premier temps un test d'erreur de première espèce approximativement égale à 5 %. On cherche r tel que, sous H_0,

$$\alpha = \mathbb{P}\{T_{1837} < r\} \approx \mathbb{P}\{N < r\} = 5\,\% \; ,$$

où N suit une loi normale standard. On voit que $r = z_{5\,\%} = -1.645$, le quantile à 5 % de la loi normale standard, convient. Or, le calcul de la valeur réalisée de la statistique de test

donne

$$\sqrt{1837}\,\frac{\overline{x}_{1837}-66}{s_{x,1837}} = \sqrt{1837}\,\frac{598-676}{254} = -13.2\ .$$

On rejette H_0 sans hésitation et on en conclut que le pouvoir d'achat a reculé (froide constatation statistique qu'il s'agirait, selon le contexte, d'exploiter stratégiquement).

J'espère que ce sentiment de rejet "sans hésitation" pique votre curiosité et que vous voulez le quantifier par l'indication de la P-valeur. Cette dernière vaut (approximativement, à cause de la convergence en loi et du caractère asymptotique du test)

$$\mathbb{P}\{N < -13.2\} \leqslant 10^{-39} \qquad \text{où} \quad N \sim \mathcal{N}(0,1)\ .$$

Cette valeur, extrêmement faible, montre à quel point nous sommes sûrs qu'il fallait rejeter H_0. Seuls les tests avec une erreur de première espèce plus petite que ce seuil ridiculement faible auraient conservé H_0 (autant dire que ces tests sont quasiment équivalents au test qui conserve tout le temps H_0, contre vents, marées, et jugement dernier).

Pour vous entraîner : montrez que même si on avait obtenu $\overline{x}_{1837} = 660$ (avec toujours $s_{1837} = 254$), on rejetterait encore H_0 dans ce cas, et sans hésitation là non plus. En effet, la P-valeur serait inférieure à la valeur très faible 4 ‰.

Remarque 7.9 (Variations statistiquement significatives, ou pas). La presse, et en particulier, la presse télévisée, vous annonce souvent des variations (par exemple, concernant le nombre de tués sur la route). Notez bien que rien n'étant constant dans notre bas monde, variations il y a toujours. La vraie et bonne question est de détecter quand ces variations sont significatives, et c'est ce que nous venons de faire sur l'exercice du pouvoir d'achat. La quantification est ici importante pour distinguer l'aléa ordinaire (lié à la constitution de l'échantillon) des évolutions de fond. Cela, les médias ne le font pas du tout, et trop peu de citoyens s'élèvent contre ce fait.

3.2. Taille d'échantillon faible : cas de données gaussiennes. Le cas gaussien décrit ci-dessous est utilisé essentiellement lorsque la taille d'échantillon n est petite, *id est*, $n \leqslant 30$. Lorsqu'elle est grande, il n'est plus nécessaire que les observations soient distribuées selon une loi normale et l'on peut appliquer le test (asymptotique) décrit ci-dessus.

Principe 7.2. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*
Données : $x_1, \ldots, x_n \in \mathbb{R}$
Modélisation associée : observations $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2
Hypothèse H_0 : $\mu_0 = \mu_{ref}$
Statistique de test :

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_{ref}}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \sim \mathcal{T}_{n-1}$
Comportement sous H_1 : lorsque $\mu_0 > \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

Nous avons déjà traité ce cas plus haut, au paragraphe 1.2 (les faiseurs de pluie), mais nous y revenons, pour l'illustrer davantage.

EXERCICE 7.3 (Vous trouvez votre évier trop bas ?). Avez-vous déjà fait la vaisselle dans un évier situé dans une cité U construite dans les années 70 ? Si oui, vous avez remarqué qu'il était très bas. Imaginez que vous soyez un constructeur de cuisines et que vous commenciez à entendre des plaintes [15] sur la hauteur de vos éviers. Vous les construisiez jusqu'à présent pour une taille moyenne de 160 centimètres (vos données vous disaient que la taille moyenne des femmes était justement de 160 centimètres, et vos analyses marketing montrent que c'est elles les prescriptrices d'achat dans les couples en ce qui concerne la cuisine). Vous conduisez donc une étude et mesurez toutes les femmes qui se présentent à votre magasin (en échange d'un petit cadeau). 27 femmes se prêtent au jeu, cela vous donne les hauteurs $x_1, \ldots, x_{27}$. La hauteur moyenne relevée est $\overline{x}_{27} = 164.3$ centimètres et l'écart-type sur les données est de $s_{x,27} = 13.2$ centimètres. Faut-il faire remonter cette information au syndicat des constructeurs de cuisines afin qu'il diligente une étude plus approfondie ? Information dont il vous faudra tenir compte : vous voulez vous présenter à la présidence du syndicat l'an prochain.

CORRECTION 7.3. On a vu dans la partie 2 qu'on pouvait modéliser la taille comme la réalisation d'une variable gaussienne $\mathcal{N}(\mu_0, \sigma_0^2)$ où μ_0 et σ_0^2 sont inconnues. Ici, on a donc affaire à 27 observations $X_1, \ldots, X_{27}$ indépendantes et identiquement distribuées selon cette loi. Le paramètre d'intérêt μ_0 est ici la taille moyenne de l'ensemble des femmes qui fréquentent le magasin en question (la fin de la correction discute le lien avec la taille moyenne des Françaises).

Cette modélisation étant effectuée, on peut passer au test proprement dit. D'abord, le choix des hypothèses : on a envie de prendre pour H_0 le fait que la population féminine n'a pas changé de taille, $H_0 : \mu_0 = 160$, parce qu'on craint de perdre toute crédibilité et de se ridiculiser en alertant le syndicat à tort. (En effet, on préfère de loin que la taille ait changé et ne pas le signaler que de déclencher une fausse alarme.) L'hypothèse alternative est quant à elle $H_1 : \mu_0 > 160$, au vu des retours que l'on a des clients (et du discours ambiant qui annonce que la population française vit mieux et est donc plus grande).

La statistique de test

$$T_{27} = \sqrt{27}\ \frac{\overline{X}_{27} - 160}{\sqrt{\widehat{\sigma}_{27}^2}}$$

est fournie par le principe 7.2 énoncé ci-dessus. Elle suit, sous H_0, la loi $\mathcal{T}_{26}$.

La zone de rejet est de la forme $]r, +\infty[$, pour un certain seuil r à déterminer ; en effet, si μ_0 est effectivement plus grand que $\mu_{\text{ref}} = 160$, alors $\overline{X}_{27}$ et donc T_{27} auront tendance à prendre des valeurs plus grandes que sous H_0. On illustre tous les calculs qui suivent par la figure 37.

Si l'on veut faire un test d'erreur de première espèce $\alpha = 5\,\%$, alors il suffit de prendre $r = t_{26,95\,\%} = 1.706$. Le calcul de la valeur réalisée de T_{27} donne

$$\sqrt{27}\ \frac{\overline{x}_{27} - 160}{\sqrt{s_{27}^2}} = \sqrt{27}\ \frac{164.3 - 160}{13.2} = 1.693\ .$$

15. On retrouvera par exemple la même situation pour les industries textiles (refonte à prévoir ou pas du tableau des mesures corporelles) ; ces dernières ont effectivement lancé il y a deux ans une campagne de nouvelles mesures.

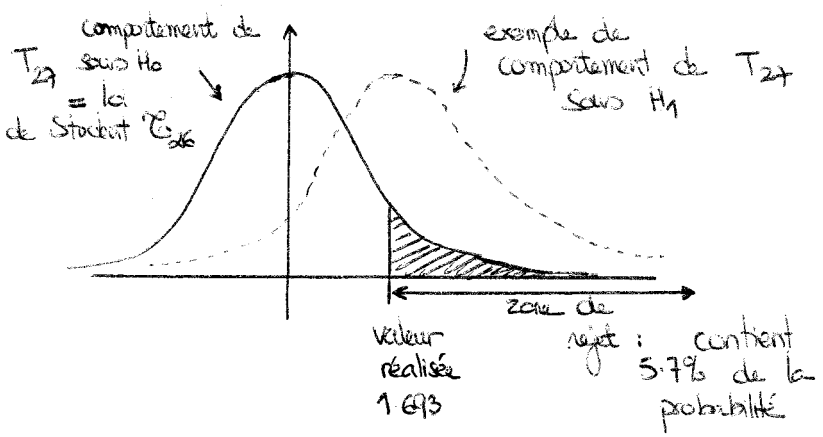


FIGURE 37. Illustration du raisonnement et des calculs associés à l'exercice 7.3.

Comme cette valeur est plus petite que le seuil r, on conserve H_0. De peu ! J'espère que la conservation sur le fil de H_0 vous semble ridicule : c'est bien le problème des tests d'erreur de première espèce fixée. Regardons ce qu'il advient de H_0 avec la P-valeur. Ici, la P-valeur est

$$\mathbb{P}\{T > 1.693\} = 5.7\% \qquad \text{où} \quad T \sim \mathcal{T}_{26} \, .$$

(Avec les tables statistiques seulement, et sans ordinateur, on aurait pu dire simplement que la P-valeur est comprise entre 5 % et 10 %.) On rappelle que c'est le degré d'attachement à H_0.

Le test conduit de manière objective à cette valeur ; l'interprétation et la décision humaines commencent maintenant. On peut se dire que cette P-valeur est suffisamment faible pour que l'on ait de bonnes raisons de rejeter H_0 (un test d'erreur de première espèce non pas à 5 % mais à 6 % le ferait !) et d'alerter le syndicat. On a bien quantifié ici l'attachement à H_0 : il est modéré. Remarquez que cela est sans doute dû à la taille réduite de l'échantillon : seulement 27 femmes et l'on a déjà des doutes... Ceux-ci seront sans doute levés par une étude plus grande – sauf si notre échantillon était biaisé (la taille moyenne dépend de la zone géographique, elle est sans doute plus élevée à Neuilly que dans la Creuse).

Conclusion stratégique. Ici, les statistiques nous ont mis la puce à l'oreille mais ne permettent pas, au vu du faible nombre de données, de trancher définitivement la question en l'état. Ainsi,

- si l'on est d'un tempérament optimiste et que l'on pense que lever ce lièvre servira de manière importante la campagne politique, alors il faudra se lancer ;
- si l'on est d'un naturel pessimiste ou si l'on préfère perdre dignement que prendre le risque de se tromper, alors on se taira ;
- enfin, si l'on a du temps devant soi, on contactera différents collègues, futurs soutiens de campagne, pour effectuer une étude avec un échantillon plus grand.

REMARQUE 7.10 (A ne surtout jamais faire !). Voici une très mauvaise idée qui pourrait vous traverser la tête : on accepte H_0 sur le fil avec une erreur $\alpha = 5\,\%$ fixée à l'avance ? Qu'à cela ne tienne, on incorpore de nouvelles femmes dans l'échantillon jusqu'à tant que la P-valeur descende en-dessous de 5 % ! Attention, dans ce cas, l'échantillon n'est

plus aléatoire, sa taille étant fixée en fonction de la P-valeur. Inclure ou pas un nouveau membre dépend des membres précédents. La théorie mathématique des temps d'arrêt et des martingales montre que les choses se passent alors radicalement différemment. En fait, avec cette méthode, on pourrait toujours tout rejeter, en étant suffisamment patient et en incluant suffisamment de monde. Retenez bien que pour réaliser une expérience statistique, on détermine la taille d'échantillon à l'avance et on inclut des gens au hasard, sinon le caractère indépendant et identiquement distribué n'est pas garanti et toutes les belles méthodes de ce cours s'écroulent !

3.3. Cas d'une fréquence, taille d'échantillon grande. Lorsque les observations sont distribuées selon une loi de Bernoulli $\mathcal{B}(p_0)$, leur variance $p_0(1-p_0)$ est liée à leur espérance p_0 ; on peut donc améliorer le principe précédent en se passant de l'estimation de la variance, puisque l'on connaît cette dernière sous une hypothèse de départ de la forme $H_0 : p_0 = p_{\text{ref}}$, c'est en effet $p_{\text{ref}}(1-p_{\text{ref}})$.

Notez bien la différence cruciale avec la construction des intervalles de confiance : ici, on peut et doit utiliser la valeur p_{ref}, qui est connue, afin de construire la statistique de test.

PRINCIPE 7.3. *Test de comparaison d'une proportion de population p_0 à une valeur de référence p_{ref}*
Données : $x_1, \dots, x_n \in \{0,1\}$
Modélisation associée : observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0
Hypothèse H_0 : $p_0 = p_{\text{ref}}$
Statistique de test :

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1-p_{\text{ref}})}}$$

Comportement sous H_0 : $T_n \rightharpoonup \mathcal{N}(0,1)$
Comportement sous H_1 : lorsque $p_0 > p_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $p_0 < p_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

La preuve découle directement du théorème de la limite centrale (pour le comportement sous H_0 ; aucun besoin de recourir ici au lemme de Slutzky) et de la loi des grands nombres (pour le comportement sous H_1).

EXERCICE 7.4 (Contrôle qualité au service clientèle). Pour économiser un peu d'argent sur le dos de ses clients, un opérateur Internet charge son service clientèle d'accéder favorablement à 25 % des requêtes de ses clients ; pas plus que 25 % à cause du coût de la démarche et parce que cela encouragerait trop de clients à se plaindre (pour l'instant, ils se pensent souvent battus d'avance face aux lourdeurs administratives de l'entreprise) ; mais pas moins non plus, pour ne pas souffrir d'une réputation trop atroce auprès des associations de consommateurs. Le chef du service, vu cet objectif strict, décide que tous les membres du service devront produire un taux de satisfaction des requêtes d'environ 25 %, celui-ci étant mesuré en fin de chaque semaine. Il n'est pas question que des vieilles

rombières frustrées aient un taux individuel outrageusement plus petit, ou que des employés grands seigneurs, syndicalistes communistes à leurs heures perdues, exaucent tout le monde ou presque. Le chef aime l'uniformité et déteste les têtes qui dépassent. Les statistiques sont ses amies : comment va-t-il construire un test permettant de contrôler le travail de ses subordonnés ? Ceux-ci sont au nombre de 20, et il ne voudrait pas non plus passer son temps à vérifier a posteriori leur travail, il voudrait détecter seulement ceux dont il devra éplucher le travail en détails. Chaque employé ayant pour objectif de répondre à 150 courriers par semaine, que va-t-il advenir d'Alban, qui vient de rentrer de congés, la tête encore toute joyeuse, et qui a accédé à 48 requêtes sur les 149 traitées (taux d'acceptation de 32.2 %) ? Et de Ghislaine, qui, à deux mois de la retraite, en a assez, et s'est vengée sur les clients en rejetant 135 demandes sur les 153 qu'elle a vu passer (taux d'acceptation de 11.8 %) ?

CORRECTION 7.4. On note $x_1, \dots, x_n$ le résultat du traitement des courriers par un employé dans une semaine, où n est proche de 150. On code $x_j = 1$ si l'employé accède à la demande du j–ème courrier qu'il a pris au hasard dans le gros sac de courriers de réclamations ; et $x_j = 0$ en cas de refus. On peut modéliser ces données comme la réalisation des variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0, où p_0 est le taux d'acceptation de l'employé cette semaine-là, en fonction de son tempérament habituel et de son humeur particulière du moment. Notez que ce taux est même inconnu à l'employé lui-même : il est difficile de lire l'esprit de quelqu'un, même le sien. En revanche, on voit les réalisations de ce taux à l'épreuve des courriers traités. Le caractère indépendant et identiquement distribué des observations provient du choix au hasard d'une lettre dans un gros sac.

La modélisation étant effectuée, on peut passer au test d'hypothèses, où $p_{\text{ref}} = 0.25$. Ici, il s'agit pour le chef de service de construire un test bilatère de $H_0 : p_0 = 25\,\%$ contre $H_1 : p_0 \neq 25\,\%$. En effet, il veut éviter tant les déviations significatives à la hausse qu'à la baisse.

La statistique de test est fournie par le principe 7.3 ci-dessus,

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - 0.25}{\sqrt{0.25 \times 0.75}}\ ;$$

sous H_0, elle suit approximativement la loi normale $\mathcal{N}(0,1)$. (Ici, exceptionnellement, on garde l'indice n et on le remplace plus tard par les valeurs prises, $n = 149$ ou $n = 153$.) Sous H_1, la statistique T_n tendra à prendre des valeurs plus petites ou plus grandes que sous H_0, selon que $p_0 < 0.25$ ou $p_0 > 0.25$, les deux cas pouvant se produire. On prend donc une zone de rejet symétrique, de la forme $]-\infty, -r[\ \cup\ ]r, +\infty[$, pour un certain seuil r à déterminer.

Construisons dans un premier temps un test d'erreur de première espèce approximativement égale à $\alpha = 2\,\%$, disons. (Le chef ne veut en effet pas trop souvent avoir à vérifier le travail d'un employé, il est déjà débordé, de plus, dans cette entreprise, les syndicats sont puissants et ne sont rendus muets que par des arguments de poids.) On cherche r tel que, sous H_0,

$$\mathbb{P}\{T_n < -r \ \text{ ou } \ T_n > r\} \approx \mathbb{P}\{N < -r \ \text{ ou } \ N > r\} = 2\,\%$$

où N suit une loi normale standard. On voit que $r = z_{99\,\%} = 2.326$, le quantile à 99 % de la loi normale standard, convient. (C'est bien celui à 99 % qu'il faut prendre ici, et non pas celui à 98 %. Cela provient du caractère symétrique de la zone de rejet, voir la figure 38.)

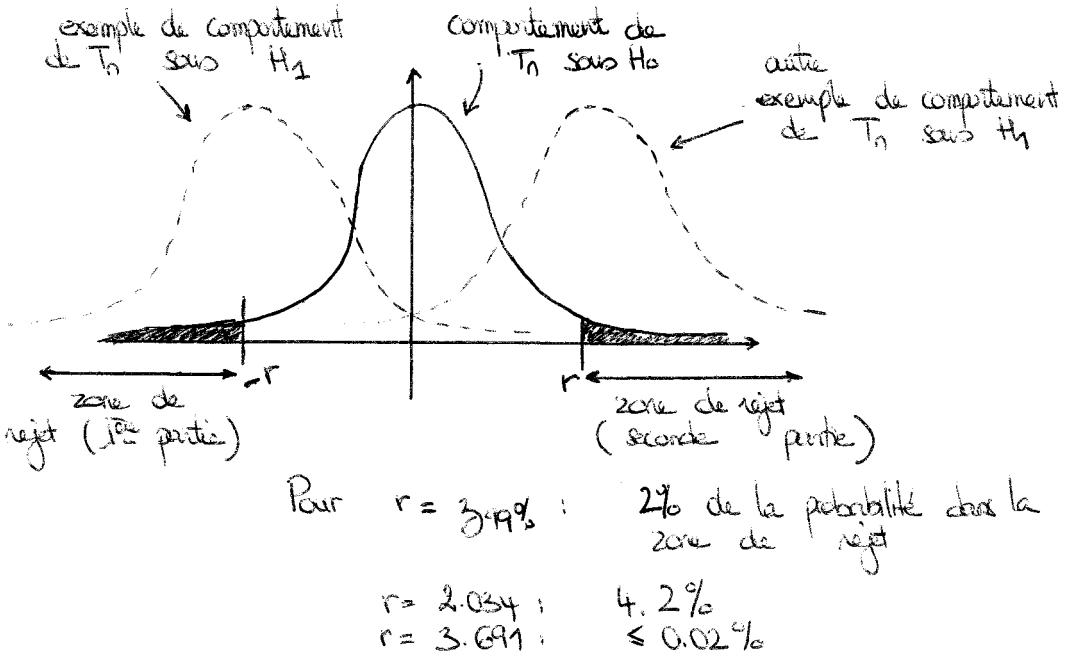


FIGURE 38. Illustration du raisonnement et des calculs associés à l'exercice 7.4.

Or, le calcul de la valeur réalisée de la statistique de test T_{149} donne pour Alban

$$\sqrt{149}\,\frac{\overline{x}_{149}-0.25}{\sqrt{0.25\times 0.75}} = \sqrt{149}\,\frac{(48/149)-0.25}{\sqrt{0.25\times 0.75}} = 2.034\ ,$$

tandis que pour Ghislaine, il s'agit de la valeur réalisée de T_{153} :

$$\sqrt{153}\,\frac{\overline{x}_{153}-0.25}{\sqrt{0.25\times 0.75}} = \sqrt{153}\,\frac{(18/149)-0.25}{\sqrt{0.25\times 0.75}} = -3.691\ .$$

Au niveau fixé de 2 %, le test conserve H_0 dans le cas d'Alban mais rejette H_0 dans le cas de Ghislaine.

Ici encore, on peut quantifier mieux les choses par la P-valeur (il est même préférable, comme toujours de le faire). Celle associée aux données d'Alban vaut (approximativement, à cause de la convergence en loi et du caractère asymptotique du test)

$$\mathbb{P}\{N<-2.034 \text{ ou } N>2.034\} = 2\times\mathbb{P}\{N>2.034\} = 4.2\,\% \qquad \text{où} \quad N\sim\mathcal{N}(0,1)\ .$$

Quant à celle de Ghislaine, elle vaut

$$\mathbb{P}\{N<-3.691 \text{ ou } N>3.691\} \approx 0.02,\,\% \qquad \text{où} \quad N\sim\mathcal{N}(0,1)\ .$$

Conclusion stratégique. Le chef de service, au vu de ces P-valeurs, pourra éventuellement garder Alban à l'œil mais ne le convoquera pas immédiatement (H_0 a pour lui un degré de crédibilité situé entre 4 % et 5 %). Ghislaine, quant à elle, est bonne pour un entretien de recadrage afin qu'elle réalise qu'une humiliante mise au placard l'attend si elle ne fait pas plus attention à ses performances ; en effet, la faiblesse de sa P-valeur indique sans hésitation que quelque chose a mal tourné cette semaine et qu'elle s'est significativement écartée de son objectif de 25 %.

REMARQUE 7.11 (Calcul de la P-valeur pour un test bilatère). Attention, ici, le test étant bilatère, le calcul de la P-valeur est plus délicat. Il faut bien tenir compte du caractère

symétrique de la zone de rejet. Il vient alors, comme on l'a détaillé pour Alban, un facteur 2 (eu égard à la symétrie de la loi normale). Ceci est une remarque fondamentale, qu'il faut absolument que vous assimiliez bien. Nous verrons au chapitre suivant que SPSS réalise des tests presqu'exclusivement bilatères et vous donne toujours cette P-valeur pour une zone de rejet symétrique, comme ici (`bilatérale`, écrit-il).

REMARQUE 7.12. Le test proposé ci-dessus ne suffit pas à vérifier que les employés travaillent bien ; il faudrait aussi étudier la répartition des 0 et des 1 au cours de la semaine, pour voir s'il y a des phénomènes de rattrapage en fin de semaine pour arriver à une bonne fréquence moyenne, ou pire, si, de manière séquentielle et uniforme, sans aucune étude préalable de leur contenu, trois dossiers étaient rejetés, puis un accepté, etc., ce qui ferait parvenir à une proportion d'exactement 25 %.

Le fait, surtout répété dans le temps, d'avoir tout le temps exactement 25 % est aussi suspect que les grandes déviations vers le haut ou le bas ! La modélisation aléatoire commande en effet des fluctuations autour de 25 %, de l'ordre de $\pm 1/\sqrt{n} \approx \pm 8\,\%$ ici (cf. le cours sur les intervalles de confiance). Il est improbable que la fréquence observée soit régulièrement très proche de 25 % (à moins de 1 %, disons).

Un cours de probabilités avancées permettrait de mieux quantifier cela. Nous nous contenterons pour notre part de ces remarques intuitives.

Compléments pour étudiants avancés

4. Liens entre intervalles de confiance et tests

La construction des tests et celle des intervalles de confiance vous semblent proches ? Vous souvenez-vous que dans la partie 1, nous faisions des tests de comparaison à une valeur de référence en regardant si celle-ci était dans l'intervalle de confiance exhibé ?

Il y a effectivement des liens forts entre tests et intervalles de confiance ; ils sont même essentiellement équivalents lorsque H_0 est réduite au test d'une valeur de référence (mais pas lorsque H_0 est plus complexe).

Cela étant, je ne vous donne qu'un aperçu ci-dessous de cette équivalence (qu'on appelle "dualité entre tests et intervalles de confiance") et de ses limites, essentiellement en l'illustrant sur des exemples simples mettant en jeu des lois de Bernoulli.

4.1. Intervalle de confiance $\rightarrow$ test. En gros, si vous avez un intervalle de confiance à 95 %, alors le test qui déciderait de conserver H_0 si et seulement si la valeur de référence est dans l'intervalle de confiance serait un test d'erreur de première espèce $\alpha = 5\,\%$. Cependant, si je vous donnais des valeurs observées, vous ne pourriez que me dire si vous conservez H_0 ou pas, mais vous ne pourriez pas me donner ce que j'aime pourtant tant, une P-valeur.

Ici, l'erreur de première espèce est fixée, et c'est bien contraignant.

4.2. Test $\rightarrow$ intervalle de confiance en théorie. Inversement, si on a une modélisation indexée par un paramètre θ_0 et que l'on a un test d'erreur de première espèce fixée 5 %, alors l'ensemble des valeurs θ_{ref} telles que le test conserve $H_0 : \theta_0 = \theta_{\text{ref}}$ forme un intervalle de confiance à 95 %. L'ennui de cette technique, c'est que ce n'est pas une manière très naturelle de procéder.

Voyons ce que cela donne sur un schéma de Bernoulli. On dispose d'observations $X_1, \ldots, X_n$ i.i.d. selon une loi de Bernoulli de paramètre p_0. Un premier intervalle de confiance asymptotique (bilatère) à 95 % est donné (voir la partie 5) par

$$\widehat{I}_n = \left[\overline{X}_n \pm z_{97.5\,\%} \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} \right] .$$

Le test classique bilatère (voir le principe 7.3) accepte les hypothèses $H_0 : p_0 = p_{\text{ref}}$ (contre $H_1 : p_0 \neq p_{\text{ref}}$) pour les p_{ref} tels que

$$\left| \frac{\sqrt{n}}{\sqrt{p_{\text{ref}}(1-p_{\text{ref}})}} (\overline{X}_n - p_{\text{ref}}) \right| \leqslant z_{97.5\,\%}$$

ce qui conduit à l'intervalle de confiance

$$\widehat{J}_n = \left\{ p : \left| \frac{\sqrt{n}}{\sqrt{p(1-p)}} (\overline{X}_n - p) \right| \leqslant z_{97.5\,\%} \right\} .$$

Il est facile de voir que $\widehat{J}_n$ est bien un intervalle, mais pas d'en donner une formule explicite ! On peut voir que la différence entre $\widehat{I}_n$ et $\widehat{J}_n$ s'amenuise à mesure que n grandit.

Exercices

Je vous propose dans cette partie sept exercices, assez simples et portant tous uniquement sur la comparaison d'une moyenne ou d'une proportion à une valeur de référence. (De nombreux autres exercices sont disponibles dans les quizz.) Nous verrons des situations de tests plus complexes aux cours suivants.

Quatres exercices issus du cours

Effectuez les exercices 7.1 à 7.4, dont l'énoncé et la correction détaillée se trouvent dans la version rédigée de cette partie.

Trois exercices issus des annales

EXERCICE 7.5. Traitez la question 1 de l'exercice III de l'examen principal de 2008 (répartition hommes–femmes dans les salles de sport).

EXERCICE 7.6. Répondez à la question 3 de l'exercice I de l'examen principal de 2007 (augmentation ou non d'un taux de commande).

EXERCICE 7.7. Considérez l'exercice I de l'examen de rattrapage de 2007 (détermination du fait qu'une pièce est biaisée ou équilibrée).

Exercice 1.

Question 1 de l'exercice III de l'examen principal de 2008

> Les données disponibles sont que sur 269 sondés à la sortie des salles, 105 étaient des femmes.

Modélisation : La population visée est l'ensemble des Français pratiquant le sport en salle. On s'intéresse à la proportion p_0 de femmes dans cette population et on veut la comparer à la proportion $p_{ref} = 51.4\%$ générale de femmes parmi l'ensemble des Français : a-t-on $p_0 = p_{ref}$?

On dispose des données $x_1, \dots x_{269} \in \{0,1\}$, où l'on code $x_j = 1$ si le j-ème sondé est une femme. Vu l'interrogation effectuée à un moment au hasard, on peut le modéliser comme la réalisation de $X_1, \dots X_{269}$ iid $\sim Ber(p_0)$, où $p_0 \in [0,1]$ est le paramètre d'intérêt (déjà décrit plus haut).

On observe la statistique $\bar{x}_{269} = 39.0\%$.

Test d'hypothèses / Cas d'un statisticien sans préjugés.

Un statisticien sans préjugé partirait du fait que les proportions de femmes sont identiques et prendrait pour hypothèse alternative une hypothèse bilatère (il n'a là non plus pas de préjugé sur la composition ressentie par les usagers d'une salle). Cela correspond à $H_0 : p_0 = p_{ref}$ contre $H_1 : p_0 \neq p_{ref}$.

La statistique de test est $T_{269} = \sqrt{269}\,\dfrac{\bar{X}_{269} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}$; elle suit approximativement la loi $\mathcal{N}(0,1)$ sous H_0.

Sous H_1, $\bar{X}_{269}$ est proche de p_0 mais peut donc être plus petite ou plus grande que p_{ref}, de sorte que T_{269} peut être plus grande ou plus petite que sous H_0. On prend par conséquent une zone de rejet bilatère, de la forme $]-\infty, -r[\cup]r, +\infty[$.

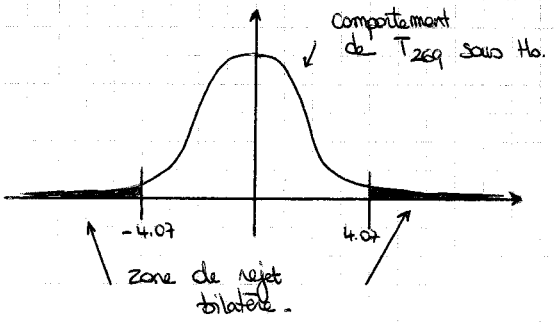
La valeur réalisée de T_{269} est

$$\sqrt{269}\,\frac{\bar{x}_{269} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}} = \sqrt{269}\left(\frac{0.390 - 0.514}{\sqrt{0.514\,(1-0.514)}}\right) = -4.07$$

soit (via les tables) une P-valeur de

$$\mathbb{P}\{N < -4.07 \text{ ou } N > 4.07\} = 2\,\mathbb{P}\{N > 4.07\} \simeq 0.005\,\%$$

Attention, ici la zone de rejet est bilatère, il ne faut pas oublier d'en tenir compte :



Conclusion statistique : on rejette avec force H_0 !

Conclusion stratégique : elle dépend du but recherché ; p. ex., ici, pour faire la publicité de l'appareil PowerPlate aux femmes, il faudrait leur donner des tracts dans des lieux qu'elles fréquentent davantage ; quant aux affiches placées dans les salles de sport, on retiendra qu'il est surtout nécessaire qu'elles s'adressent aux hommes.

Test d'hypothèses / Cas d'un statisticien adepte des salles de sport.

Lui a déjà ressenti l'absence des femmes en salles et testerait plutôt

$$H_0 : p_0 = p_{ref} \quad \text{vs} \quad H_1 : p_0 < p_{ref}$$

Sa zone de rejet serait unilatère ($]-\infty, r[$) et sa P-valeur vaudrait $\mathbb{P}\{N < -4.07\} \simeq 0.0025\,\%$. Les conclusions sont inchangées.

Exercice 2. Question 3 de l'exercice I de l'examen principal 2007

Modélisation (rappel) : La population visée est constituée par l'ensemble des clients du fichier (50 000 foyers) et on s'intéresse au taux de réponse p_0 qu'on aurait si on généralisait une nouvelle offre. On dispose des réactions $x_1, \dots x_{1000} \in \{0,1\}$ des membres d'un échantillon à cette offre : on note $x_j = 1$ lorsque le j-ième membre a effectué une commande (et $x_j = 0$ sinon). Sur l'échantillon, on dispose de l'estimée $\bar{x}_{1000} = 17.0\%$. On a proposé la modélisation des données comme la réalisation de $X_1, \dots X_{1000}$ iid $\sim Ber(p_0)$, via le tirage au hasard dans le fichier clients.

Test d'hypothèses : Changer quelque chose en place pour un nouveau venu est un moment crucial : il ne faut le faire qu'avec prudence et en étant sûr de son coup. C'est pourquoi on prend $H_0 : p_0 = 13\%$ (pas d'augmentation du taux de commande) contre $H_1 : p_0 > 13\%$. Notre souci est de rejeter H_0, ce qui indiquerait que p_0 est significativement plus grand que $p_{ref} = 13\%$.

La statistique de test est $T_{1000} = \sqrt{1000}\left(\frac{\bar{X}_{1000} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right)$, elle suit approximativement une loi $\mathcal{N}(0,1)$ sous H_0.

Sous H_1, $\bar{X}_{1000}$ donc T_{1000} tendent à prendre des valeurs plus grandes. La zone de rejet est donc de la forme $]r, +\infty[$.

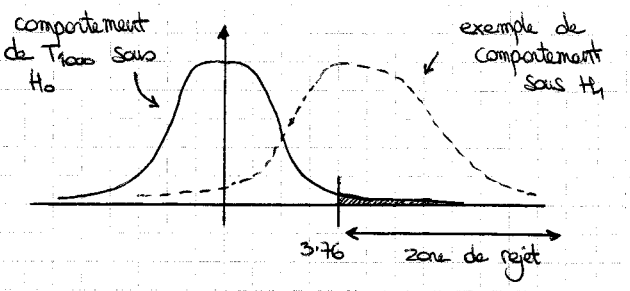
La valeur réalisée de T_{1000} est

$$\sqrt{1000}\left(\frac{\bar{x}_{1000} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right) = \sqrt{1000}\left(\frac{0.17 - 0.13}{\sqrt{0.13 \times (1-0.13)}}\right) = 3.76$$

soit une P-valeur de $\mathbb{P}\{N > 3.76\} \leq 0.01\%$

(où $N \sim \mathcal{N}(0,1)$)

On a le dessin suivant :



Conclusion Statistique : rejet clair de H_0.

Conclusion Stratégique : certes, le taux de réponse a augmenté mais il faut voir maintenant ce qu'il se passe au niveau de la marge par commande (des fois qu'une augmentation du taux de réponse soit compensée par une baisse de la marge par commande !).

Exercice 3. Exercice I de l'examen de rattrapage 2007

Note : on a affaire ici à un problème plutôt probabiliste (via l'interprétation du paramètre p_0 et l'absence de population et échantillon, remplacées par une expérience répétée).

(1) On dispose des résultats de lancers $x_1, \dots x_{1000} \in \{0,1\}$ (disons : $x_j = 1$ si pile au j-ème lancer et $x_j = 0$ pour face), que l'on modélise comme la réalisation de $X_1, \dots X_{1000}$ iid $\sim Ber(p_0)$, où $p_0 \in [0,1]$ est le paramètre d'équilibrage de la pièce.

On va se placer dans la peau de quelqu'un sans préjugé et qui pense que toutes les pièces du monde sont équilibrées ; auquel cas on choisit les hypothèses $H_0 : p_0 = 1/2$ contre $H_1 : p_0 \neq 1/2$ (soit, $p_{ref} = 1/2$).

On considère la statistique de test

$$T_{1000} = \sqrt{1000}\left(\frac{\bar{X}_{1000} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right) = 2\sqrt{1000}\left(\bar{X}_{1000} - 1/2\right)$$

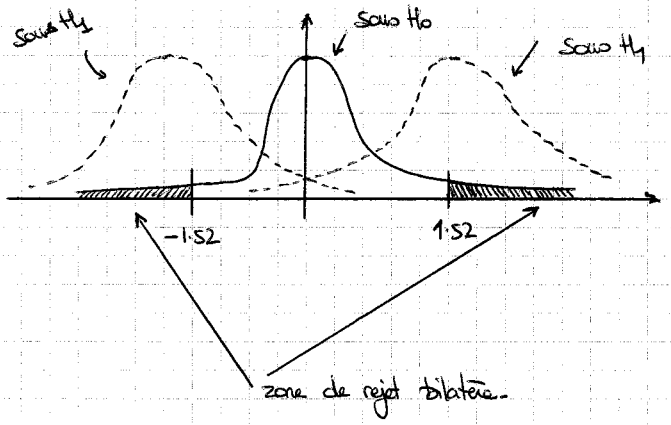
Sous H_0, T_{1000} suit approximativement une loi $\mathcal{N}(0,1)$.
Sous H_1, $\bar{X}_{1000}$ étant proche de p_0, T_{1000} peut prendre des valeurs plus petites ou plus grandes que sous H_0.
On prend donc une zone de rejet bilatère, de la forme

$$]-\infty, -r[\cup]r, +\infty[$$

Comme $\bar{x}_{1000} = 47.6\% = 0.476$, la valeur réalisée de T_{1000} est

$$2\sqrt{1000}\left(\bar{x}_{1000} - 1/2\right) = 2\sqrt{1000}\left(0.476 - 0.50\right) = -1.52$$

soit une P-valeur de :



$p = \mathbb{P}\{N \leq -1.52 \text{ ou } N \geq 1.52\} = 2\,\mathbb{P}\{N \leq -1.52\}$
$= 2(1 - 0.9357) = 6.4\%$ selon les tables.

Conclusion statistique : p légèrement plus grand que 5%, on conserve H_0 (sur le fil)

Conclusion stratégique : si on veut en avoir le coeur net, il faut procéder p. ex. à une nouvelle série de lancers, en plus grand nombre.

(2) On reprend les calculs précédents en se fixant un niveau $\alpha = 5\%$, soit la région de rejet $]-\infty, z_{2.5\%}[\cup]z_{97.5\%}, +\infty[$
$=]-\infty, -1.96[\cup]1.96, +\infty[$

On rejette H_0 lorsque la valeur réalisée est dans cette zone, soit lorsque $|2\sqrt{1000}\,(\bar{x}_{1000} - 1/2)| > 1.96$

i.e.,
$$\left| \frac{x_1 + \cdots x_{1000}}{1000} - 1/2 \right| > 1.96/2\sqrt{1000}$$

soit encore $x_1 + \cdots x_{1000} > \left(1/2 + \frac{1.96}{2\sqrt{1000}}\right) \times 1000$, ie, $x_1 + \cdots x_{1000} \geq 531$

ou $x_1 + \cdots x_{1000} < \left(1/2 - \frac{1.96}{2\sqrt{1000}}\right) \times 1000$, ie, $x_1 + \cdots x_{1000} \leq 469$.

Huitième Partie

Interlude : deux quizz sur les tests de comparaison à une valeur de référence

Premier énoncé (sujet posé en 2009)

Calculatrices et tables uniquement & Tournez la page !

Quizz 3 – Estimation par intervalles et tests de conformité à une valeur de référence – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Question de cours

Dans quelle brasserie travaillait William Gosset, *alias* Student ?

Intervalle de confiance

Une étude menée fin 2008 sur 298 logements parisiens choisis au hasard dans l'annuaire assure que le prix du loyer au mètre carré est de 18.4 euros, avec un écart-type mesuré de 3.2 euros.

1. Modélisez *brièvement* la situation.
2. Entourez un commanditaire : Collectif Jeudi Noir, Observatoire des loyers parisiens, Confédération nationale des propriétaires-bailleurs ; et donnez-lui la réalisation d'un intervalle de confiance sur le prix moyen du loyer au mètre carré à Paris (justifiez sa forme).
3. Question subsidiaire : l'intervalle proposé est-il en contradiction avec le fait que les studios de 25 m^2 sont souvent présentés dans les annonces avec un loyer de 750 euros (soit 30 euros du mètre carré) ?

Tournez la page !

Calculatrices et tables uniquement & Tournez la page !

Test d'hypothèses

Le Figaro rapporte l'information suivante.

> Au congrès des notaires de France qui a eu lieu fin 2008 et a rassemblé plusieurs milliers de participants (soit presque tous les notaires de France), différentes consultations avaient eu lieu, et notamment une quant à l'optimisme des représentants de cette noble profession face à la reprise, ou non, du marché immobilier. 39.0 % s'étaient déclarés optimistes quant à une reprise rapide, au plus tard début 2010.
>
> Courant juillet 2009, une étude téléphonique a été menée par la chambre des notaires : 750 notaires tirés au hasard dans le fichier de la profession ont été contactés et 685 d'entre eux ont accepté de répondre à la question suivante : « Pensez-vous toujours que le marché immobilier reprendra au plus tard début 2010 ou qu'il a déjà repris ? » Parmi eux, 294 ont répondu par l'affirmative.

Que doivent en conclure ses lecteurs ? (On commencera par modéliser *brièvement* la situation, puis on mènera un test d'hypothèses.)

Premier corrigé (sujet posé en 2009)

Notes : Elles sont très bonnes, je suis en moyenne très satisfait de votre travail : BRAVO !

	A	B	C	D	E	Absents (f)	Excusés
94 étudiants :	28	32	18	6	2	5	3

médiane et mode à B

2/3 environ des étudiants ont A ou B

Correction :

1) Question de cours : C'était évidemment une blague, je ne vous demandais pas de retenir ce micro-détail historique, mais suis fort aise de constater que 80% d'entre vous y sont parvenus ! Comme quoi, on retient plus facilement certains détails que certains théorèmes....
Au fait, la réponse était : Guinness (avec deux n).

2) Intervalle de confiance :

Modélisation : La population est l'ensemble des logements locatifs de Paris. On dispose de 298 données de prix de location au m^2, notées $x_1, \dots x_{298} \in \mathbb{R}^+$. Vu le recueil au hasard des données, on peut les modéliser comme étant la réalisation de $X_1, \dots X_{298}$ iid selon une certaine loi, d'espérance (inconnue) notée μ_0. Cette dernière forme le paramètre d'intérêt : le prix moyen au m^2 sur l'ensemble des logements locatifs de Paris. On dispose de l'estimée $\bar{x}_{298} = 18.4$ €. Par ailleurs, l'écart-type

d'échantillon est $s_{298} = 3.2$ €.

Réalisations d'intervalles selon le commanditaire retenu : sur μ_0

- Observatoire des loyers parisiens : il veut simplement une image précise de la réalité, il est neutre. Il voudra donc un intervalle bilatère : une estimée du prix moyen, assortie d'une marge d'erreur.

  L'intervalle théorique à 95% est $\left[\bar{X}_{298} \pm z_{97.5\%} \sqrt{\frac{\hat{\sigma}^2_{298}}{298}} \right]$ et sa réalisation sur les données vaut

  $$\left[\bar{x}_{298} \pm z_{97.5\%} \frac{s_{298}}{\sqrt{298}} \right] = \left[18.4 \pm 1.96 \frac{3.2}{\sqrt{298}} \right]$$
  $$= [18.4 \pm 0.4]$$

  (Note : un chiffre après la virgule suffit, c'est la précision avec laquelle les estimés avaient été donnés.)

- Le collectif Jeudi Noir dénonce les loyers trop chers, il veut une minoration du prix moyen au m² pour montrer que les prix sont déraisonnables : intervalle théorique à 95% :

  $$\hat{I}_{298} = \left[\bar{X}_{298} - z_{95\%} \sqrt{\hat{\sigma}^2_{298}/298} \, , +\infty \right[$$

  de valeur réalisée (après calculs) : $[18.1, +\infty[$.

- Les propriétaires - bailleurs voulant montrer que le prix sont tout à fait raisonnables : $\hat{J}_{298} = \left[0, \bar{X}_{298} + z_{95\%} \sqrt{\hat{\sigma}^2_{298}/298} \right]$

  de valeur réalisée (après calculs) : $[0, 18.7]$.

Question subsidiaire : Ici, on s'intéresse au prix moyen μ_0 de l'ensemble des logements (baux récents et baux plus anciens, parc privé ou HLM, petits ou grandes surfaces). Les 30 €/m² seraient plutôt à

Comparer à μ_1 : le prix moyen au m^2 des studios du parc privé à la relocation. Trois facteurs expliquent un prix plus important pour cette sous-catégorie de logements :

1) les petites surfaces sont plus chères à Paris, à cause de la loi de l'offre (tendue) et de la demande (très forte : nombreux célibataires à Paris). La pression est moins forte pour les grandes surfaces.

2) le parc HLM est moins cher que le parc privé, il tire μ_0 vers le bas ; or les annonces concernent uniquement le parc privé.

3) en ce moment (ou récemment : 2003 - 2008), le marché a été très tendu et les prix, surtout à la relocation, ont fort augmenté. Or, le marché n'a pas toujours été tendu (voir crise immobilière des années 90). Ainsi, les nouveaux baux sont plus chers en moyenne que les anciens.

En conclusion : il n'y a pas de contradiction. μ_0 est simplement une valeur moyenne sur une population fortement hétérogène, de sorte que diverses sous-populations ont un comportement moyen fort différent de μ_0.

Commentaires :

- Il fallait penser à bien justifier la forme (unilatère ou bilatère) des intervalles de confiance en fonction du commanditaire.
- Pour le reste (distinguer variable aléatoire et réalisations, erreurs de rédactions, notations mal choisies, etc.) : cf. corrigé très détaillé des erreurs récurrentes dans le quizz #2.

3) Test :

Modélisation : La population considérée est l'ensemble des notaires de France. On en a tiré un échantillon au hasard, parvenant ainsi à 685 réponses exploitables $x_1, \dots x_{685} \in \{0,1\}$, avec comme code que $x_j = 1$ si le j-ième sondé table sur une reprise. Vu ce tirage au hasard, on peut modéliser ces données comme la réalisation de $X_1 \dots X_{685}$ iid selon une loi de Bernoulli de paramètre p_0. p_0 est le paramètre d'intérêt, c'est la vraie proportion des notaires tablant sur une reprise en juillet 2009 ; p_0 est inconnu mais est estimé par $\bar{x}_{685} = \frac{294}{685} \simeq 42.9\%$.

Test : Il s'agit de voir si l'optimisme des notaires a significativement évolué. En 2008, on avait eu la chance d'interroger tous les notaires de France au presque (il ne s'agissait donc pas d'un sondage en 2008, mais d'une étude exhaustive de la population) ; on avait obtenu une proportion $p_{ref} = 39.0\%$ d'optimistes tablant sur une reprise. Cela forme notre valeur de référence. On teste alors l'hypothèse $H_0 : p_0 = p_{ref}$ (hypothèse dont on voudrait prouver qu'elle ne tient pas et qui correspond au fait que l'optimisme n'a pas changé) contre une hypothèse alternative.

1er cas : Cf. discours ambiant de reprise : on pense à $H_1 : p_0 > p_{ref}$ (augmentation de l'optimisme)

La statistique de test est

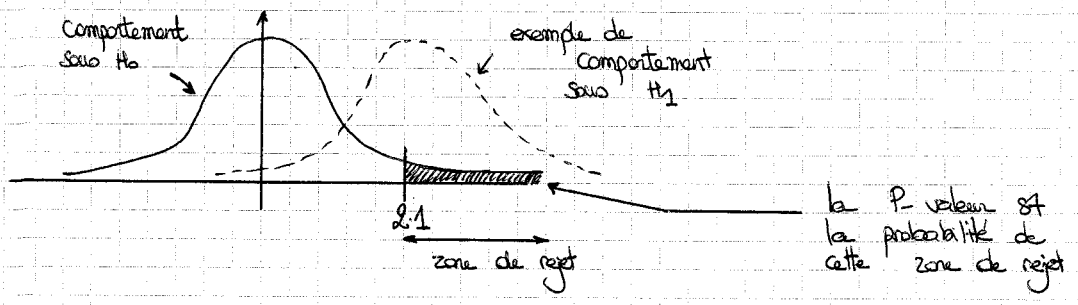
$$T_{685} = \sqrt{685}\left(\frac{\bar{X}_{685} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right)$$

Sous H_0, on a $T_{685} \overset{(d)}{\simeq} \mathcal{N}(0,1)$

Sous H_1, $\bar{X}_{685}$ est donc T_{685} tendant à prendre des valeurs plus grandes : la zone de rejet est de la forme $]r, +\infty[$.

Or, la valeur réalisée de T_{obs} est :

$$\sqrt{685}\left(\frac{\bar{x}_{obs} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right) = \sqrt{685}\left(\frac{0.429 - 0.39}{\sqrt{0.39(1-0.39)}}\right) \simeq 2.1$$



Calcul de la P-valeur : $p = \mathbb{P}\{N \geq 2.1\} = 1 - \mathbb{P}\{N \leq 2.1\}$

(où $N \sim \mathcal{N}(0,1)$) $= 1 - 0.9821 \simeq 1.8\%$

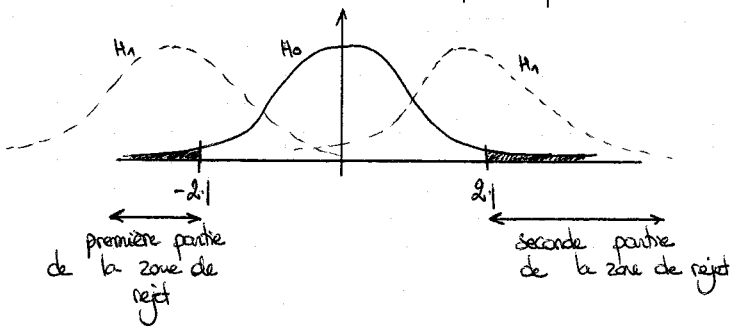
Conclusion statistique : P-valeur faible, bien inférieure à 5%, on rejette H_0 sans hésitation. Ainsi, on peut dire que le taux d'optimisme a significativement augmenté parmi les notaires.

Conclusion stratégique : (Celle d'un lecteur du Figaro) Les notaires voient passer les actes, ils sont donc souvent aux premières loges pour voir les crises ou les reprises (avec cependant un petit délai de 3 mois, cf. le temps de passer du compromis de vente à l'acte authentique). Cependant, la question reposant ici sur un ressenti : il ne s'agit en aucun cas de chiffres objectifs mais de la quantification de sentiments subjectifs. Alors certes, il existe peut-être des signes de reprise ; ou alors (c'est la méthode Coué) les notaires se font simplement le relais d'une communication institutionnelle qui annonce « Tout va bien, Mme la marquise ». Après tout, ils ont leurs études notariales à faire tourner !

En bref : Ce sondage ne dit pas qu'il faut acheter ou que les prix vont augmenter, il dit juste que les notaires reprennent espoir et que sans doute (c'est là la vraie conclusion stratégique) il faut continuer à

surveiller l'évolution du marché immobilier comme le lait sur le feu. P. ex., en continuant à lire le Figaro.

2nd cas : Cas d'un lecteur sans préjugés, il prendra une alternative bilatère $H_1 : p_0 \neq p_{ref}$; sous H_1, T_{obs} prendra alors des valeurs au plus grandes ou plus petits que sous H_0, conduisant à une zone de rejet de la forme $]-\infty, -r[\cup]r, +\infty[$, et à une P-valeur de $2 \times 1.8\% = 3.6\%$, ainsi qu'indiqué sur le dessin suivant.



Commentaires et relevé des erreurs fréquentes :

- Les hypothèses H_0 et H_1 n'étaient pas H_0 : le marché ne reprendra pas contre H_1 : le marché reprendra, mais H_0 : aucune évolution dans l'optimisme des notaires contre H_1 : augmentation de leur optimisme, ce que, par ailleurs, il faut traduire mathématiquement ; avec nos notations, c'était $H_0 : p_0 = p_{ref}$ contre $H_1 : p_0 > p_{ref}$.
- Attention, il faut poser H_1 non pas au vu des données (de la valeur de $\bar{x}_{obs}$) mais au vu du contexte ; si l'on n'a aucun argument a priori pour un test unilatère (ici, c'était le discours ambiant), alors c'est que l'on est sans préjugés et l'on doit recourir à un test bilatère.
- Attention, on n'avait ici que 685 et non 750 données $x_1, x_2, \ldots$

exploitables ! J'ai cependant compté juste à ceux qui avaient mené le test avec $x_1, \dots x_{750}$ (et $\bar{x}_{750} = \frac{294}{750} = 39.2\%$) ; leurs P-valeurs unilatère comme bilatère étaient grands et ils campaient avec force sur H_0.

- ATTENTION à la statistique de test T_{GSS} : sous H_0, on connaît p_0, c'est p_{ref}, inutile de l'estimer par $\bar{X}_{GSS}$. La statistique n'était donc pas $\sqrt{GSS}\left(\frac{\bar{X}_{GSS} - p_{ref}}{\sqrt{\bar{X}_{GSS}(1-\bar{X}_{GSS})}}\right)$ mais $T_{GSS} = \sqrt{GSS}\left(\frac{\bar{X}_{GSS} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right)$

- Enfin j'ai très rarement vu une conclusion stratégique pertinente et/ou qui apporte une valeur ajoutée par rapport à la conclusion statistique (qui était une l'augmentation de l'optimisme des natures) et/ou qui prenne du recul face à l'étude. Bref, vos salaires futurs (mirobolants), par qui seront-ils justifiés sinon par une valeur ajoutée dans le traitement statistique ? Sinon, autant embaucher quelqu'un qui sort d'un IUT de stats ...

Second énoncé (sujet posé en 2008)

Dans les deux problèmes qui suivent, on mettra en œuvre la méthodologie des tests pour répondre à la question posée. On commencera notamment par modéliser, avec concision, la situation statistique.

Une nouvelle campagne marketing

Une grande enseigne de VPC envoie une nouvelle promotion à 1 000 clients choisis au hasard dans son grand fichier de clients. Habituellement, le taux de réponse à ses courriers commerciaux est de 13 %. Ici, cette campagne, créée et pensée par un petit jeune sorti d'HEC, a vu 153 commandes réalisées, soit un taux de réponse de 15.3 %. La différence entre les deux taux est-elle significative, i.e., le directeur général doit-il féliciter sa brillante nouvelle recrue ?

Evolution du salaire moyen dans la fonction publique

Les fonctionnaires se plaignent d'une perte de pouvoir d'achat collective dans les dix dernières années, tandis que le gouvernement rétorque qu'individuellement chaque fonctionnaire gagne plus. L'Etat ne disposant pas de service de ressources humaines, les syndicats réalisent alors un sondage téléphonique pour en avoir le cœur net. Ils obtiennent, sur 345 fonctionnaires sondés, un salaire moyen mensuel net de 2 193 euros en 2008, avec un écart-type mesuré de 573 euros sur l'échantillon. Le dernier grand recensement de 2004 donnait un salaire moyen (en euros constants, on tient donc compte de l'inflation dans le chiffre suivant) de 2 245 euros. Ces chiffres vont-ils compter lors des négociations avec le ministre de la fonction publique ?

Question subsidiaire, à ne traiter que si vous avez le temps
Cet exercice est représentatif de la vraie vie : la méthodologie appliquée par les syndicats est celle décrite plus haut, comparer les salaires moyens de la population de l'ensemble des fonctionnaires, celle du gouvernement actuel est de suivre l'évolution moyenne de pouvoir d'achat d'un fonctionnaire donné. Les deux méthodes parviennent en pratique à des conclusions opposées. Pouvez-vous expliquer pourquoi et indiquer celle qui vous semble la plus juste ?

Second corrigé (sujet posé en 2008)

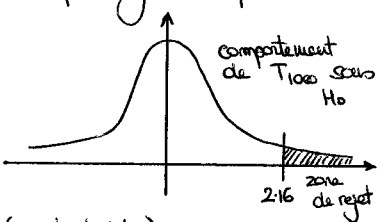
Une nouvelle campagne marketing

Une grande enseigne de VPC envoie une nouvelle promotion à 1 000 clients choisis au hasard dans son grand fichier de clients. Habituellement, le taux de réponse à ses courriers commerciaux est de 13 %. Ici, cette campagne, créée et pensée par un petit jeune sorti d'HEC, a vu 153 commandes réalisées, soit un taux de réponse de 15.3 %. La différence entre les deux taux est-elle significative, i.e., le directeur général doit-il féliciter sa brillante nouvelle recrue ?

La population visée est l'ensemble des clients du grand fichier.
On dispose des données $x_1, \dots x_{1000} \in \{0,1\}$ ($x_j = 1$ si le j-ème client de l'échantillon a effectué une commande), que, via le tirage au hasard, on modélise comme la réalisation de $X_1, \dots X_{1000}$, variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathrm{Ber}(p)$, de paramètre $p_0 \in [0,1]$ inconnu.
p_0 est le paramètre d'intérêt, c'est le taux de commande si on généralisait la promotion à l'ensemble des clients. Il est inconnu mais on dispose de l'estimée $\bar{x}_{1000} = 15.3\%$.

On veut savoir si cette proportion observée sur l'échantillon est significativement plus grande que le taux habituel $p_{\text{ref}} = 13\%$. Le DG étant un homme prudent et avare en compliments, il choisit de tester $H_0 : p_0 = 13\%$ contre $H_1 : p_0 > 13\%$.
La statistique de test est $T_{1000} = \sqrt{1000}\left(\bar{X}_{1000} - 0.13\right) / \sqrt{0.13\,(1-0.13)}$
et sous H_0, T_{1000} suit approximativement une loi $\mathcal{N}(0,1)$.
Sous H_1, $\bar{X}_{1000}$ donc T_{1000} tendant à prendre des valeurs plus grandes que sous H_0.
La zone de rejet est donc de la forme $]r, +\infty[$.
Sur les données on observe la réalisation

$$\sqrt{1000}\left(\frac{\bar{x}_{1000} - 0.13}{\sqrt{0.13 \times (1-0.13)}}\right) = \sqrt{1000}\left(\frac{0.153 - 0.13}{\sqrt{0.13 \times 0.87}}\right) = 2.16$$



soit une P-valeur de $\mathbb{P}\{N > 2.16\} \simeq 1.58\%$ (via les tables) où $N \sim \mathcal{N}(0,1)$
Cette P-valeur très faible permet de rejeter H_0 sans hésitation (conclusion statistique). Ici, la conclusion stratégique est sans appel vu son prix et vu la certitude statistique : le DG doit féliciter votre camarade ! (... et éventuellement lui donner une prime ?)

Quizz 3 – Eléments de statistique mathématique

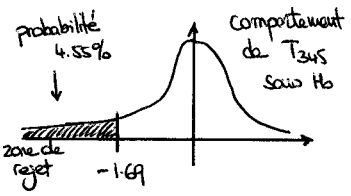
Evolution du salaire moyen dans la fonction publique

Les fonctionnaires se plaignent d'une perte de pouvoir d'achat collective dans les dix dernières années, tandis que le gouvernement rétorque qu'individuellement chaque fonctionnaire gagne plus. L'Etat ne disposant pas de service de ressources humaines, les syndicats réalisent alors un sondage téléphonique pour en avoir le cœur net. Ils obtiennent, sur 345 fonctionnaires sondés, un salaire moyen mensuel net de 2 193 euros en 2008, avec un écart-type mesuré de 573 euros sur l'échantillon. Le dernier grand recensement de 2004 donnait un salaire moyen (en euros constants, on tient donc compte de l'inflation dans le chiffre suivant) de 2 245 euros. Ces chiffres vont-ils compter lors des négociations avec le ministre de la fonction publique ?

La population visée est l'ensemble des fonctionnaires français, dont on veut déterminer le salaire moyen μ_0. On dispose, en 2008, des données $x_1, \dots x_{345} \in \mathbb{R}_+$, représentant chacune le salaire mensuel d'un sondé. Sur l'échantillon, on a $\bar{x}_{345} = 2193$ et $s_{x,345} = 573$. On modélise, via l'interrogation au hasard, ces données comme la réalisation des variables aléatoires $X_1, \dots X_{345}$ indépendantes et identiquement distribuées selon une certaine loi, d'espérance précisément le paramètre μ_0 recherché.

On veut savoir si ce salaire moyen de 2008 peut être dit significativement différent du salaire de 2004, à savoir $\mu_{ref} = 2245$ (exprimé en tenant compte de l'inflation). Ici, le choix des hypothèses est très politique : les syndicats vont prendre H_0 : $\mu_0 = \mu_{ref}$ et espèrent montrer que H_0 doit être rejetée avec force. On la teste contre H_1 : $\mu_0 < \mu_{ref}$ (c'est le ressenti des syndicalistes). On réinterprète ces hypothèses comme H_0 : le pouvoir d'achat est constant vs. H_1 : le pouvoir d'achat a baissé. On considère la statistique de test $T_{345} = \sqrt{345}\,\dfrac{\bar{X}_{345} - 2245}{\sqrt{\hat{\sigma}^2_{345}}}$ qui suit approximativement, sous H_0, la loi $\mathcal{N}(0,1)$.

Sous H_1, $\bar{X}_{345}$ et donc T_{345} tendent à prendre des valeurs plus petites. La zone de rejet est donc de la forme $]-\infty, r[$. Or, sur les données, on observe la réalisation $\sqrt{345}\left(\dfrac{\bar{x}_{345} - 2245}{s_{x,345}}\right) = \sqrt{345}\left(\dfrac{2193 - 2245}{573}\right) = -1.69$



soit une P-valeur de $\mathbb{P}\{N < -1.69\} = 4.55\%$ (en consultant les tables).

Quizz 3 – Eléments de statistique mathématique

Conclusion Statistique : ce serait le rejet (sur le fil, vu la proximité de la P-valeur à 5%) de H_0.

Conclusions Stratégiques : les syndicats souligneront qu'avec seulement 345 sondés, on a déjà tellement de doute qu'on rejette l'hypothèse de pouvoir d'achat constant, et que ce serait sans doute encore pire avec plus de sondés ! Ces chiffres comptèrent donc, et ils les diffuseront dans la presse !

Question subsidiaire, à ne traiter que si vous avez le temps
Cet exercice est représentatif de la vraie vie : la méthodologie appliquée par les syndicats est celle décrite plus haut, comparer les salaires moyens de la population de l'ensemble des fonctionnaires, celle du gouvernement actuel est de suivre l'évolution moyenne de pouvoir d'achat d'un fonctionnaire donné. Les deux méthodes parviennent en pratique à des conclusions opposées. Pouvez-vous expliquer pourquoi et indiquer celle qui vous semble la plus juste ?

→ Le gouvernement regarde la carrière des gens déjà en place : selon des données qu'il diffusait à l'époque, 83% d'entre eux avaient gagné en pouvoir d'achat ces dernières années. Certes, mais pas à cause des augmentations générales du point d'indice, qui sont censés compenser l'inflation, mais à cause de l'avancement individuel à l'ancienneté. En somme, on gagne plus parce qu'on est plus vieux. Les 17% qui ont perdu en pouvoir d'achat sont ceux qui sont au dernier échelon de leur grille de rémunération. C'était cela, "l'avancement d'un fonctionnaire donné".

→ Les syndicats regardent l'ensemble des salaires, y compris les entrants. Et disent que la moyenne des salaires augmente moins vite que l'inflation.

→ Exemple concret : Un prof. de 30 ans en 2008 avec 4 ans d'ancienneté a bien sûr plus de pouvoir d'achat que lors de son recrutement en 2004 (raisonnement du gouvernement) mais moins que ce qu'aurait une prof. de 30 ans avec 4 ans d'ancienneté en 2004.

→ Je pense, vous vous en doutez bien, que le raisonnement des syndicats est plus juste, car il tient compte des entrants et mesure la perte d'attractivité des carrières de la fonction publique.

Remarques détaillées & erreurs fréquentes.

Note : Lisez bien ma correction, elle présente une rédaction-type. J'ai l'impression que certains d'entre vous ne méditent pas assez mes corrigés et font et refont indéfiniment les mêmes erreurs.

Une nouvelle campagne marketing : [Cet exercice avait été fait en cours au remplacement de 17% par 15.3% près.

- Ce n'était pas un test de comparaison de proportions entre deux populations ; on avait une et une seule population (les clients du fichier) et une valeur de référence claire.
- Ici, il valait mieux faire 1 test unilatère $H_0 : p_0 = 0.13$ vs $H_1 : p_0 > 0.13$.

  On prend pour H_0 l'hypothèse de prudence ; si on arrive à la rejeter, alors on aura appris quelque chose. Il n'est pas bon de prendre $H_0 : p_0 \geq 0.13$; vous ne savez pas construire le test dans ce cas : vous ne savez que traiter, en gros, le cas $H_0 : p_0 = p_{ref}$ (mais pas $\geq$ ou $\leq$).
- Si on faisait un test bilatère ($H_1 : p_0 \neq 0.13$) il fallait alors trouver 3.16 % de P-valeur (et rejeter H_0).
- p_0 n'est pas le taux de réponses positives de la campagne d'essai : ce taux vaut 15.3%. p_0 serait le taux de réponses si on généralisait la promo à tout le fichier clients. De toute façon, si p_0 valait 15.3% effectivement, pourquoi s'ennuierait-on à le tester contre 13% ?!?

  ↳ Rappel : je veux que vous me décriviez par une phrase qui est le paramètre d'intérêt et son sens.

- Attention au calcul de la statistique de test : on n'a pas besoin d'estimer la variance, on la connaît sous H_0, c'est $p_{ref}(1-p_{ref}) = 0.13 \times (1-0.13)$. Ceux qui se sont trompés ont trouvé 2.02 au lieu de 2.16 pour la valeur observée de la statistique de test.

- Remarques mineures ou qui concernent mais de monde :
  - Bernoulli (et non pas Bernouilli, rien à voir avec des nouilles)
  - Que veut dire le symbole $\hookrightarrow$?
    Je ne connais que $\sim$ (suit la loi)
    $\longrightarrow$ (converge en loi)
    et $\overset{(d)}{\simeq}$ (suit approximativement la loi)
  - Certains ont inversé ou confondu p_0 et p_{ref} ; p_{ref} est donné par l'énoncé, p_0 est inconnu et à tester.
  - "H_0 : il y a différence significative" ne veut rien dire ; il y a telle différence si $H_0 : p_0 = p_{ref}$ est rejeté, c'est tout.
  - Ici, comme le soulignait l'énoncé, il s'agissait bien de tests et pas d'intervalles de confiance.
  - En particulier, il ne faut pas oublier de préciser ni H_0 ni H_1 !

Evolution du salaire moyen dans la fonction publique.

(Très proche lui aussi d'un autre exercice fait en cours...)

- Si on prenait $H_0 : \mu_0 \leq \mu_{ref}$ vs. $H_1 : \mu_0 > \mu_{ref}$ (ie, on prend pour H_0 le ressenti des syndicats), alors on trouve une P-valeur de 95.45 % (refaits les calculs pour le voir).
Une P-valeur aussi élevée oblige à conserver H_0, mais cela n'a pas le même poids stratégique que le raisonnement du corrigé : ici, on montre juste qu'on n'a pas tort de camper sur ses préjugés.

- Quand $n \geq 30$, $\mathcal{C}_n \overset{(d)}{\simeq} \mathcal{N}(0,1)$, on peut les identifier.

- Pour $H_0 : \mu_0 = \mu_{ref}$ vs. $H_1 : \mu_0 \neq \mu_{ref}$ il fallait trouver une P-valeur de 9.10% (et le rejet de H_0 n'était plus clair du tout ! On sentait juste qu'il serait bon d'avoir plus de données pour conclure...)

- On n'a pas que l'écart-type σ² des salaires sur la population vaut 573, on a juste une estimée de ce paramètre : $s_{7,345} = 573$.

- Enfin, le plus important : il faut toujours conclure par une phrase donnant la décision stratégique ou de management à prendre ! (Ici, s'appuyer sur les données, forcer le gouvernement à les regarder, les diffuser à la presse.)

Neuvième Partie

Compléments sur les tests, et notamment, tests à partir d'échantillons indépendants ou appariés

Version rédigée du cours

Résumé : Le cours précédent vous a appris, à partir d'un problème concret, à formuler des hypothèses (une de départ et une alternative) et à mettre en œuvre un test (i.e., une statistique de test et une forme de zone de rejet) afin de déterminer s'il faut camper sur l'hypothèse de départ ou si les données la contredisent suffisamment gravement pour qu'il faille passer à l'hypothèse alternative. Nous avons introduit une quantité-clé pour réaliser ce choix final : la P-valeur. C'est le niveau d'erreur maximal dans la construction du test qui conduirait encore au non-rejet de H_0. La P-valeur est un indice de crédibilité de H_0 : une faible valeur (plus petite qu'un seuil de 5 % ou 1 %) est le signe qu'il faut préférer l'hypothèse alternative à l'hypothèse de départ.

Nous avions illustré la méthodologie des tests en déclinant différentes situations de tests de comparaison d'une valeur moyenne à une valeur de référence (selon que les lois sous-jacentes étaient normales, de Bernoulli, ou sans forme connue).

Objectif : La vie statistique est bien plus vaste que cela, et un océan de tests s'ouvre à nous. Nous approfondirons essentiellement les tests de comparaison de moyennes de deux populations. Par ailleurs, SPSS, grand absent du chapitre précédent, fait son retour sur scène : à vous les statistiques cliquables ! Nous verrons comment décrypter ses sorties, qui ont l'avantage, une fois qu'on sait les lire, de calculer pour nous et sans erreur les P-valeurs.

1. Avant de commencer, une citation et un retour sur le sens profond des tests

Comme à mon habitude, je soumets une citation [16] et son commentaire à votre sagacité :

> La statistique est un bikini. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital.
> Arthur Koestler (écrivain, journaliste et essayiste hongrois, 1905–1983)

En effet, ainsi que nous l'avons vu au cours précédent et continuerons à le voir dans ce cours-ci et les suivants, si parfois, les tests d'hypothèses mettent en évidence certains phénomènes dans d'autres cas, ils ne peuvent se prononcer. En gros, le statisticien se pose une question, et le test lui répond « ce n'est pas impossible » ou « c'est impossible ». Mais dans le premier cas, on n'est pas sûr d'avoir mis le doigt sur la vérité : on n'a avancé qu'une assertion qui ne contredisait pas les observations. Dans le second cas, en revanche, la connaissance fait un progrès : on sait que telle assertion ne peut être tenue pour vraie. C'est un progrès négatif.

16. Variante personnelle pour être sûr de plaire à tout le monde et promouvoir l'égalité hommes–femmes : « La statistique est un *shorty* moulant. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital. »

2. Retour SPSS sur le test de comparaison d'une seule moyenne à une valeur de référence

Dans le chapitre précédent, nous n'avions pas donné de nom aux différents tests de comparaison d'une moyenne à une valeur de référence. Le test qui part d'une modélisation des données comme réalisations de variables aléatoires indépendantes et identiquement distribuées selon une loi normale (voir le principe 7.2) s'appelle le T-test, parce qu'il met en jeu les lois de Student, que l'on note $\mathcal{T}_k$.

Dans les logiciels de statistique, c'est lui que l'on met en œuvre, même dans le cas des grands échantillons ($n \geqslant 30$, voir le principe 7.1), que l'on ait vérifié au préalable ou pas que les observations étaient bien distribuées selon une loi normale. Cela n'est pas choquant dès lors [17] que l'on se souvient que lorsque la taille d'échantillon n est grande ($n \geqslant 30$), les quantiles de Student sont diablement proches des quantiles correspondants de la loi normale. On peut donc appliquer le T-test comme test universel de comparaison d'une moyenne à une valeur de référence dès lors que l'on a suffisamment d'observations.

Il faut simplement éviter d'appliquer ce T-test lorsque l'on a affaire à un test de comparaison d'une proportion à une valeur de référence (voir le principe 7.3), car pour ce dernier, il est inutile d'estimer la variance, elle est connue sous H_0, et la statistique de test est différente.

La figure 39 réalise un T-test bilatère [18] de comparaison du salaire moyen des infirmières américaines (jeu de données de la partie 2) à la valeur de référence de $\mu_{\text{ref}} = 20$ dollars, i.e., $H_0 : \mu_0 = 20$ contre $H_1 : \mu_0 \neq 20$.

Test-t

Statistiques sur échantillon unique

	N	Moyenne	Ecart-type	Erreur standard moyenne
Salaire horaire	2911	20,0159	4,00309	,07419

Test sur échantillon unique

	Valeur du test = 20					
					Intervalle de confiance 95% de la différence	
	t	ddl	Sig. (bilatérale)	Différence moyenne	Inférieure	Supérieure
Salaire horaire	,214	2910	,831	,01586	-,1296	,1613

FIGURE 39. T-test appliqué aux données de salaire horaire des infirmières américaines : comparaison à la valeur de référence de 20 dollars.

LA MINUTE SPPS 9.1. La figure 39 a été obtenue avec `Analyse / Comparer les moyennes / Test-T pour échantillon unique`. Essayons de comprendre l'ensemble des nombres qui y sont calculés. On note $x_1, \ldots, x_n$ les données disponibles ; n désigne donc le nombre de données valides (non manquantes). Le tableau du haut de la figure 39 correspond alors à :

17. Ou que l'on se souvient que les quantiles de Student majorent ceux de la loi normale, de sorte que l'on obtient bien des tests de niveau désiré ici, car ils sont juste (un peu) plus précautionneux.

18. Cf. la précision entre parenthèses de `bilatérale` en version française et `2-tailed` en version anglaise au niveau de la case donnant la P-valeur

N	Moyenne	Ecart-type	Err. std. moy.
n	$\overline{x}_n$	$s_{x,n}$	$s_{x,n}/\sqrt{n}$

tandis qu'on lit dans celui du bas (la case `ddl` donne le nombre de degrés de liberté de la loi de Student, i.e., le terme $n-1$ fait référence à la loi $\mathcal{T}_{n-1}$) :

Réalisation de T_n	ddl	Sig.	Diff. moy.	IC de niveau $1-\alpha$
$\sqrt{n}\left(\frac{\overline{x}_n - \mu_{ref}}{s_{x,n}}\right)$	$n-1$	P-valeur	$\overline{x}_n - \mu_{ref}$	$\overline{x}_n - \mu_{ref} \pm t_{n-1,1-\alpha/2}\, s_{x,n}/\sqrt{n}$

En particulier, `Erreur standard moyenne` est donc, au facteur multiplicatif près donné par le quantile $t_{n-1,1-\alpha/2}$ (ici, il s'agit de $t_{2910,97.5\,\%} \approx z_{97.5\,\%} = 1.96$), la demi-longueur de l'intervalle de confiance sur la moyenne. Notez que SPSS propose la réalisation d'un intervalle de confiance (ici à 95 %) non pas sur le salaire moyen μ_0 mais sur la différence $\mu_0 - \mu_{ref} = \mu_0 - 20$. Ici, il s'agit de $[-0.1296, 0.1613]$. On en déduit alors que la réalisation d'un intervalle de confiance sur μ_0 est donnée par l'intervalle $[20 - 0.1296, 20 + 0.1613] = [19.8704, 20.1613]$, qu'évidemment on arrondit à $[19.87, 20.17]$ (au moins [19]).

Exploitation de la figure 39 et conclusion : On voit que 2 911 données sont disponibles et que la loi de la statistique de test T_{2911} est la loi de Student $\mathcal{T}_{2910}$. La valeur réalisée pour cette dernière étant de 0.214, on obtient une P-valeur de 83.1 %, soit une conservation très claire de H_0 face à H_1 bilatère.

Remarque 9.1 (Passage d'une P-valeur de test bilatère à celle d'un test unilatère). On peut, sans faire de calcul mais en menant simplement un raisonnement sur une figure, déterminer les P-valeurs du test de $H_0 : \mu_0 = 20$ contre les alternatives $H_1 : \mu_0 > 20$ ou $H_1 : \mu_0 < 20$. Elles valent respectivement environ 41.5 % et 58.5 % (cf. figure 40).

Transition et plan : considération de deux séries de données

On a expliqué en détails comment comparer la moyenne d'une population à une valeur de référence sortie d'un chapeau. Désormais, on va s'intéresser à la comparaison des moyennes associées à deux séries de données $x_1, \ldots, x_n$ d'une part, et $y_1, \ldots, y_m$ d'autre part. On distinguera deux cas :

- ces données sont issues d'échantillons indépendants ; et alors en général, $n \neq m$; dans ce cas, on aura affaire à deux sous-cas, selon que l'on s'intéresse à
  - + une proportion
  - + ou à une "vraie moyenne" ;
- ces données ne sont pas indépendantes pour la simple et bonne raison qu'elles se correspondent deux à deux : x_t et y_t sont liées en un certain sens, pour tout indice $t = 1, \ldots, n$. (Dans ce cas, $n = m$ nécessairement.) On parle alors de données appariées.

19. Ici, vu le résultat du test d'adéquation à la valeur $\mu_{ref} = 20$, qui conserve avec force H_0 (la P-valeur est de 83.1 %), il serait légitime de proposer l'intervalle, très lisible, $[20 \pm 0.20] = [19.80, 20.20]$.

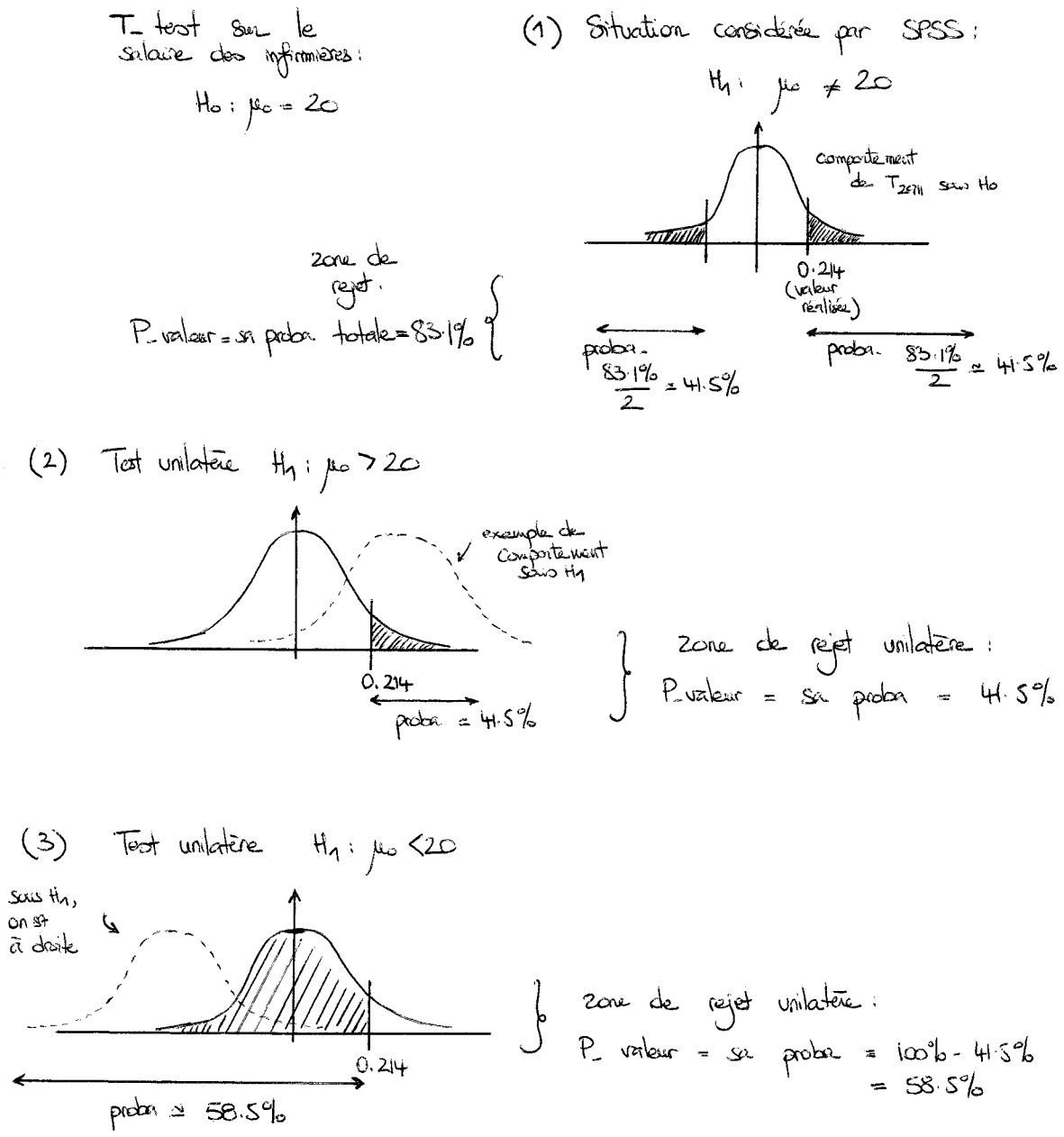


FIGURE 40. Principe de passage de la P-valeur pour un test bilatère, telle que calculée par SPSS, aux P-valeurs pour tests unilatères.

3. Première étude : cas des échantillons appariés

Plus précisément, on parle de données appariées lorsque les deux échantillons considérés sont formés des mêmes individus, pour lesquels on a fait deux mesures d'une même quantité, généralement avec écart temporel et/ou après l'occurrence d'un événement. On dispose donc de deux séries de données, de même taille $x_1, \dots, x_n$ et $y_1, \dots, y_n$, où x_j et y_j ont toutes deux été mesurées sur le j–ième membre de l'échantillon.

EXEMPLE 9.1 (Crèmes hydratantes). On teste une nouvelle formule de crème hydratante contre une formule éprouvée. On utilise pour cela les mêmes sujets (tirés au hasard dans la population), voir la figure 41 ; on peut par exemple tester l'ancienne formule sur la main gauche et la nouvelle, sur celle de droite. On mesure alors l'hydratation deux heures après l'application. Ainsi, on évite les biais causés par la tendance naturelle de chacun à

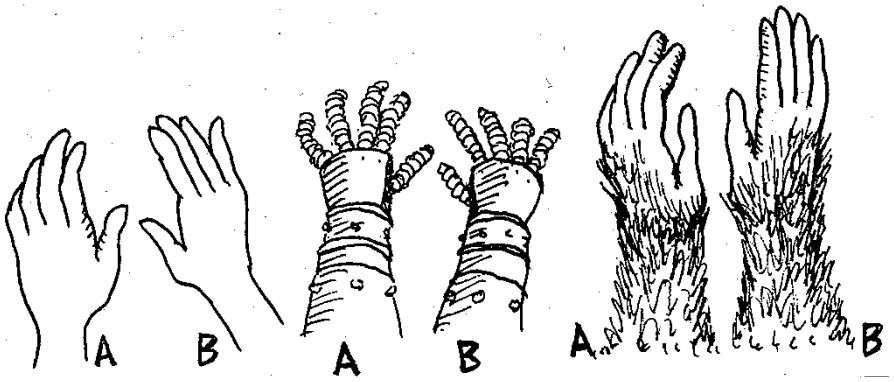


FIGURE 41. Pour choisir entre deux crèmes hydratantes, il vaut mieux recourir à un test sur observations appariées, une crème pour chaque main des sujets.

avoir une peau sèche ou grasse, et qui seraient apparus si l'on avait choisi deux groupes de sujets différents (surtout si ceux-ci étaient de petite taille, comme c'est souvent le cas dans les tests cliniques, pour des raisons de coût). Apparier les observations permet de réduire la variabilité propre aux sujets et de n'observer que celle induite par la différence entre les crèmes.

EXEMPLE 9.2 (Double correction). La technique la plus juste de comparer la manière de noter de deux examinateurs est la double correction, lorsque chacun corrige l'ensemble des copies. Il est particulièrement important de procéder ainsi (plutôt que par division du paquet de copies en deux sous-paquets) lorsqu'il y a peu de copies et qu'elles sont très disparates. C'est moins grave sur un gros concours ; dans ce dernier cas, on impose souvent par avance un objectif commun (une moyenne et un écart-type) aux correcteurs.

LA MINUTE SPPS 9.2. Si l'on regarde les données dans un logiciel, le cas des données appariées correspond à la comparaison de toutes les valeurs de deux colonnes.

Ici, la modélisation est qu'on a d'une part $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi, d'espérance μ_X, et d'autre part, $Y_1, \dots, Y_n$ indépendantes et identiquement distribuées selon une loi éventuellement différente, de moyenne μ_Y. Cela est généralement garanti par la manière de créer l'échantillon (par exemple, par tirage au sort des sujets parmi la population).

Mais ici, on ne pourra pas supposer que les X_j sont indépendantes des Y_j ; la tendance naturelle du sujet j (par exemple à avoir des mains habituellement hydratées ou sèches) influe à la fois sur X_j et sur Y_j. C'est pourquoi on s'intéressera plutôt aux différences $Z_j = X_j - Y_j$. Pour les mêmes raisons que précédemment, on a bien que $Z_1, \ldots, Z_n$ sont indépendantes et identiquement distribuées, selon une certaine loi d'espérance Δ. Avec les notations précédentes, il vient, par linéarité de l'espérance, que $\Delta = \mu_X - \mu_Y$.

Il s'agit de voir si les deux moyennes μ_X et μ_Y sont significativement différentes, ou pas, i.e., de tester $H_0 : \Delta = 0$ (contre l'alternative bilatère $H_1 : \Delta \neq 0$ ou une alternative unilatère, $H_1 : \Delta > 0$ ou $H_1 : \Delta < 0$, selon le contexte).

Il suffit alors d'appliquer directement les techniques du cours précédent (principes 7.1 ou 7.2) aux données z_j associées aux Z_j. Pour y trouver le bon principe à appliquer parmi ces deux principes, il faudra regarder la taille n de l'échantillon et ce que l'on peut dire de la loi commune des Z_j (si elle est normale ou pas). Il n'y a pas donc besoin de détailler davantage ce cas-ci.

La section consacrée aux exercices comportera une illustration de la mise en œuvre de cette méthodologie sur un exercice d'annales comparant les performances de deux somnifères.

4. Seconde étude : échantillons indépendants, le cas des proportions

Dans ce cas, on dispose par exemple de données $x_1, \ldots, x_n \in \{0, 1\}$, mesurant un certain résultat binaire et recueillies sur un premier échantillon tiré d'une première population, et de données $y_1, \ldots, y_m \in \{0, 1\}$, mesurant le même résultat mais sur un second échantillon tiré de manière indépendante d'une seconde population. La question est alors de savoir si les vraies proportions moyennes (inconnues) sur chacune des deux populations sont égales.

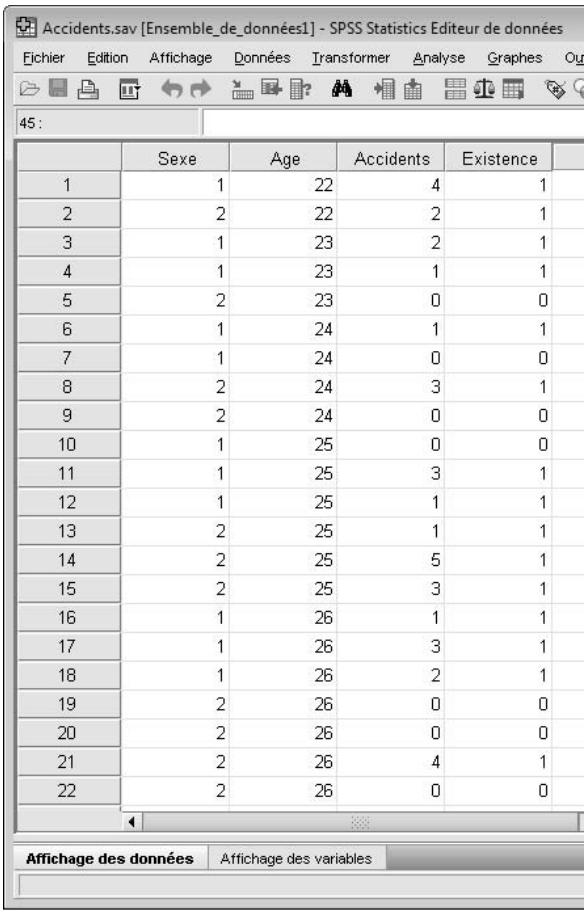
La minute SPPS 9.3. Concrètement, cela correspond cette fois-ci, sous SPSS, au cas de données que l'on lit dans une colonne, mais qu'il faut séparer en deux groupes selon la valeur donnée par une autre colonne. Dit autrement, on compare les valeurs, pour une colonne fixée, des données d'un premier jeu de lignes à celles d'un second jeu de lignes.

Exemple 9.3. Considérons les données de la figure 42. Elles reportent les accidents au cours des cinq dernières années d'assurés tirés au hasard dans le fichier d'une mutuelle. Plus précisément, chaque ligne indique le sexe, l'âge et le nombre d'accidents de l'assuré considéré. Pour votre commodité, j'ai calculé une quatrième colonne indiquant simplement s'il y a eu au moins un accident responsable ou non. Les données qui nous intéressent (les x_j et les y_k) sont précisément celles de cette quatrième colonne. La première colonne distingue les deux populations d'intérêt : les hommes assurés et les femmes assurées.

De manière générale, on suppose que l'on dispose de données, par exemple des réponses à la même question effectuées :

- sur la même population mais lors de deux sondages distincts suffisamment séparés dans le temps et après occurrence d'un événement, comme un changement de politique marketing ;
- sur deux sondages conduits simultanément mais de manière indépendante sur deux populations distinctes.

On note ces deux séries de données $x_1, \ldots, x_n \in \{0, 1\}$ et $y_1, \ldots, y_m \in \{0, 1\}$; ici, les tailles n et m n'ont aucune raison d'être égales.



	Sexe	Age	Accidents	Existence
1	1	22	4	1
2	2	22	2	1
3	1	23	2	1
4	1	23	1	1
5	2	23	0	0
6	1	24	1	1
7	1	24	0	0
8	2	24	3	1
9	2	24	0	0
10	1	25	0	0
11	1	25	3	1
12	1	25	1	1
13	2	25	1	1
14	2	25	5	1
15	2	25	3	1
16	1	26	1	1
17	1	26	3	1
18	1	26	2	1
19	2	26	0	0
20	2	26	0	0
21	2	26	4	1
22	2	26	0	0

FIGURE 42. Début du fichier de données sur les accidents routiers.

On modélise ces données comme les réalisations respectives d'une première série de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_X et celles d'une seconde série $Y_1, \ldots, Y_m$ de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_Y. On veut tester si $H_0 : p_X = p_Y$ contre différentes alternatives possibles, selon les cas, $H_1 : p_X \neq p_Y$, ou $H_1 : p_X < p_Y$, ou encore $H_1 : p_X > p_Y$. L'interprétation de ce test est la suivante : H_0 correspond respectivement aux cas où

- l'événement n'a eu aucune influence sur la population ;
- les deux populations pourtant distinctes ont le même sentiment ou comportement moyen.

Nous essayons de construire maintenant une statistique de test naturelle. En pratique, le bon réflexe est de comparer $\overline{x}_n$ et $\overline{y}_m$, i.e., de s'intéresser, du point de vue théorique, à la quantité $\overline{X}_n - \overline{Y}_m$. Il s'agit de voir si cette statistique prend des valeurs significativement différentes de 0. Pour quantifier ce caractère significatif, il faut tenir compte de la variance de la statistique, qui mesure les écarts naturels à 0. Or, par indépendance, en notant $\sigma^2(Z)$ la variance d'une variable aléatoire Z, il vient :

$$\sigma^2(\overline{X}_n - \overline{Y}_m) = \sigma^2(\overline{X}_n) + \sigma^2(\overline{Y}_m) = \frac{1}{n}\sigma^2(X_1) + \frac{1}{m}\sigma^2(Y_1) = \frac{1}{n}p_X(1-p_X) + \frac{1}{m}p_Y(1-p_Y)$$

Sous H_0, les éléments des deux séries de variables aléatoires ont toutes même variance $p_X(1-p_X) = p_Y(1-p_Y)$. On estime alors cette dernière de manière groupée, en estimant au préalable le paramètre commun $p_X = p_Y$ par

$$\widehat{p}_{n+m} = \frac{X_1 + \ldots + X_n + Y_1 + \ldots + Y_m}{n+m} = \frac{n\overline{X}_n + m\overline{Y}_m}{n+m} \ .$$

On pondère enfin $\overline{X}_n - \overline{Y}_m$ par un estimateur de sa variance pour obtenir la statistique de test :

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{p}_{n+m}(1-\widehat{p}_{n+m})\,(1/n + 1/m)}} \ .$$

Sous H_0 : $p_X = p_Y$, par une version généralisée du théorème de la limite centrale et par application du lemme de Slutzky (détails omis), il vient alors la convergence en loi $T_{n,m} \rightharpoonup \mathcal{N}(0,1)$ lorsque $n \to \infty$ et $m \to \infty$. Sous H_1, selon la forme de cette dernière, la différence $\overline{X}_n - \overline{Y}_m$ et donc $T_{n,m}$ elle-même tendent à prendre des valeurs plus grandes et/ou plus petites que 0.

Cela conduit au principe de test 9.1. Comme il repose sur une convergence en loi fondée sur le théorème de la limite centrale, il nécessite en pratique que les tailles n et m soient suffisamment grandes (toutes deux plus grandes que 30).

PRINCIPE 9.1. *Test de comparaison de proportions p_X et p_Y associées à deux séries de données indépendantes*
Données : $x_1, \ldots, x_n \in \{0,1\}$ et $y_1, \ldots, y_m \in \{0,1\}$
Modélisation associée : deux séries d'observations $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$, indépendantes et chacune formée de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli, avec les paramètres respectifs p_X et p_Y
Hypothèse H_0 : $p_X = p_Y$
Statistique de test :

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{p}_{n+m}(1-\widehat{p}_{n+m})\,(1/n + 1/m)}}$$

où l'estimateur groupé $\widehat{p}_{n+m}$ de la proportion commune sous H_0 est défini par

$$\widehat{p}_{n+m} = \frac{n\overline{X}_n + m\overline{Y}_m}{n+m}$$

Comportement sous H_0 : $T_{n,m} \rightharpoonup \mathcal{N}(0,1)$ lorsque n et m tendent tous deux vers $+\infty$
Comportement sous H_1 : lorsque $p_X > p_Y$, la statistique $T_{n,m}$ tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $p_X < p_Y$, la statistique $T_{n,m}$ tend à prendre des valeurs plus petites que sous H_0.

FIGURE 43. Principe du test de comparaison des proportions de deux populations.

Exercice 9.1 (Video killed the star, continued). Reprenons le cas de la radio généraliste étudiant le taux d'audience du segment étudiant : il s'agit de l'exercice 5.7 (exercice de synthèse de la partie 5). A la dernière question, il était demandé de déterminer si l'audience sur ce segment avait augmenté de manière significative entre deux sondages espacés de six mois. Nous avions raisonné à l'époque en termes d'intervalles de confiance mais avons désormais tous les outils pour mener à bien un test statistique et quantifier la crédibilité d'une (absence d')augmentation.

Correction 9.1. Commençons par rappeler les données sous la forme très commode d'un tableau 2×2 :

Etudiants	Auditeurs	Non-auditeurs	Total
Sondage 1	30	72	102
Sondage 2	338	662	1 000
Total	368	734	1 102

On rappelle les notations employées lors de la résolution de l'exercice dans la partie 5. Les données du premier sondage sont notées $z_1, \dots, z_{102}$ et celles du second, $y_1, \dots, y_{1000}$, et on peut les modéliser respectivement comme la réalisation de $Z_1, \dots, Z_{102}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_Z et comme celle de $Y_1, \dots, Y_{1000}$, indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_Y. Les paramètres p_Z et p_Y sont le taux d'audience du segment étudiant lors, respectivement, du premier et du second sondages (le second ayant lieu, on le rappelle, six mois après le premier, à l'occasion d'un rajeunissement de la grille des programmes effectué afin de plaire aux jeunes adultes).

On dispose des statistiques d'échantillon $\overline{z}_{102} = 30/102 = 29.4\,\%$ et $\overline{y}_{1000} = 338/1000 = 33.8\,\%$. La question est de savoir si à partir de ces estimées, on peut effectivement conclure sans trop de risques d'erreur que $p_Y > p_Z$ (ce sera H_1, ce qu'on voudrait prouver) ou s'il faut s'en tenir à l'hypothèse de prudence $p_Z = p_Y$ (ce sera H_0). Remarquez que l'on ne pense même pas ici au cas où le taux d'audience aurait diminué : on a vraiment tout fait pour le booster. C'est pour cela que le test retenu est unilatère.

La statistique de test est, conformément au principe précédent,

$$T_{102,1000} = \frac{\overline{Z}_{102} - \overline{Y}_{1000}}{\sqrt{\widehat{p}_{1102}(1 - \widehat{p}_{1102})\,(1/102 + 1/1000)}}$$

où l'estimateur groupé $\widehat{p}_{1102}$ de la proportion commune sous H_0 est défini par

$$\widehat{p}_{1102} = \frac{102\,\overline{X}_{102} + 1000\,\overline{Y}_{1000}}{1\,102}\,.$$

Sous H_0, vu que $n = 102$ et $m = 1000$ sont grands ici, $T_{102,1000}$ suit approximativement une loi normale. Par ailleurs, comme la différence $\overline{Z}_{102} - \overline{Y}_{1000}$ est toujours proche de $p_Z - p_Y$, elle a tendance à prendre des valeurs plus petites sous H_1. Des valeurs réalisées petites sont donc signe de H_1, de sorte que la forme de la zone de rejet est $]-\infty, r[$.

On calcule les valeurs réalisées. Celle pour l'estimateur groupé $\widehat{p}_{1102}$ du paramètre est

$$\overline{p}_{1102} = 368/1102 = 33.4\,\%\,,$$

de sorte que la valeur réalisée pour la statistique de test est égale à

$$\frac{\overline{z}_{102} - \overline{y}_{1000}}{\sqrt{\overline{p}_{1102}(1-\overline{p}_{1102})\,(1/102+1/1000)}} = \frac{0.294 - 0.338}{\sqrt{0.334\,(1-0.334)\,(1/102+1/1000)}} = -0.895\ ,$$

soit une P-valeur égale à

$$\mathbb{P}\{Z \leqslant -0.895\} = 18.5\,\% \qquad \text{où} \quad Z \sim \mathcal{N}(0,1)\ .$$

(Voir la figure 44 pour une illustration de l'ensemble du raisonnement précédent.)

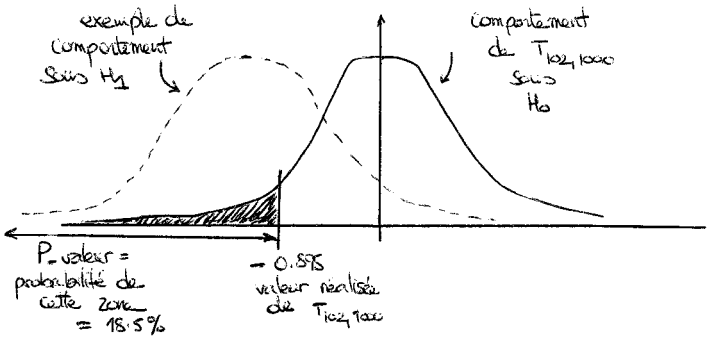


FIGURE 44. Calcul de la P-valeur associée au test de l'exercice 9.1.

Conclusion statistique : On rappelle que la P-valeur forme le degré de crédibilité de H_0 ; elle est ici supérieure aux seuils habituels (de 5 % par exemple). D'un point de vue d'une conclusion statistique, on n'a donc aucune raison de ne pas conserver H_0 ; on ne peut donc affirmer aucune différence significative entre les deux taux d'audience, tout du moins, au vu des données disponibles [20].

Conclusion stratégique : Le directeur de la radio, même s'il a une opinion subjective très positive sur le directeur des programmes, n'a pas encore d'arguments suffisants pour lui décerner une prime exceptionnelle à rendre jaloux tous les autres salariés.

LA MINUTE SPPS 9.4. Le test présenté ci-dessus, dans sa version bilatère, est un cas particulier du test du χ^2 d'indépendance, que nous verrons dans la partie 10. Lorsque les conditions asymptotiques ne sont pas remplies, on utilise un autre test, exact celui-ci mais plus difficile à calculer pour un être humain (mais pas pour SPSS), le test exact de Fisher. SPSS procure les résultats de tous ces tests, avec des commentaires sur la validité asymptotique ou non du premier. Pour lancer le test d'égalité des proportions sur le diagramme 2×2, il faut utiliser `Analyse / Statistiques descriptives / Tableaux croisés` et cliquer sur `Chi-deux` dans l'onglet `Statistiques`.

EXEMPLE 9.4 (Accidents de voiture selon le sexe). Pour les données de la figure 42, on obtient par exemple les résultats reproduits à la figure 45. On note en particulier que SPSS procure un tableau 2×2 de résumé des données, similaire à celui que nous avions également écrit pour l'exercice sur la radio. La P-valeur des deux tests discutés plus haut (et d'autres...) est indiquée dans les trois colonnes les plus à droite du second tableau de

20. Le problème est vraisemblablement que le premier sondage, effectué sur la population générale, n'a pas conduit à un sous-échantillon de population étudiante de taille suffisante ! S'il y avait eu la même proportion de 29.4 % d'étudiants auditeurs mais avec $n = 500$ étudiants interrogés lors du premier sondage, alors la P-valeur aurait été de 4.3 % et on aurait, d'un point de vue d'une conclusion statistique, rejeté H_0.

Tableau croisé Existence d'un accident au moins * Sexe de l'assuré

Effectif

		Sexe de l'assuré		
		Male	Female	Total
Existence d'un accident au moins	Pas d'accident	46	76	122
	Au moins un accident	204	174	378
	Total	250	250	500

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	9,758[a]	1	,002		
Correction[b] pour la continuité	9,118	1	,003		
Rapport de vraisemblance	9,837	1	,002		
Test exact de Fisher				,002	,001
Association linéaire par linéaire	9,738	1	,002		
Nombre d'observations valides	500				

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 61,00.

b. Calculé uniquement pour un tableau 2x2

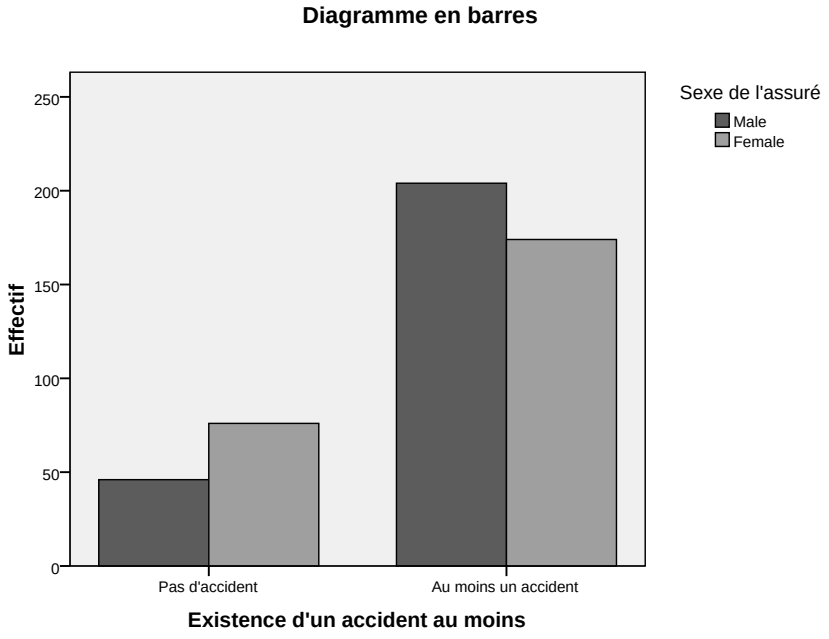


FIGURE 45. Existence d'accidents de voiture au cours des cinq dernières années selon le sexe des assurés.

la figure 45. Toutes ces P-valeurs étant très petites (même pour des tests bilatères), de l'ordre de 0.3 %, on en conclut sans aucun doute possible que les taux d'accident pour les hommes et pour les femmes sont différents (taux plus faible pour les femmes)... et donc que les femmes conduisent mieux ! Cela ne veut cependant pas dire que les hommes coûtent nécessairement plus cher à leurs assurances : il faudrait tenir compte ici du nombre effectif d'accidents et de leur coût respectif. (Note : dans la section consacrée aux exercices, nous recalculerons à la main la P-valeur du test de comparaison de proportions, que l'on lit ici dans la première ligne du second tableau.)

5. Troisième étude : échantillons indépendants, le cas général

Le cas général est plus complexe à traiter. Je vous demande uniquement de savoir lire les rendus SPSS. La théorie sous-jacente est esquissée dans les compléments mathématiques facultatifs de cette partie.

On suppose disposer de données $x_1, \dots, x_n \in \mathbb{R}$, mesurant une certaine quantité et recueillies sur un premier échantillon tiré d'une première population, et de données $y_1, \dots, y_m \in \mathbb{R}$, mesurant la même quantité mais obtenues sur un second échantillon tiré indépendamment d'une seconde population. La question est alors de savoir si les valeurs moyennes de la quantité sur les deux populations sont égales.

On modélise les données $x_1, \dots, x_n$ et $y_1, \dots, y_m$ comme la réalisation de $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$, deux séries de variables aléatoires indépendantes et identiquement distribuées, et indépendantes entre elles. On note μ_X et σ_X d'une part et μ_Y et σ_Y d'autre part l'espérance et l'écart-type communs de ces variables aléatoires. On s'intéresse à la comparaison de μ_X et μ_Y, qui sont les moyennes (évidemment inconnues) d'une certaine quantité sur chacune des deux populations en question.

La statistique naturelle est ici encore fondée sur $\overline{X}_n - \overline{Y}_m$ et comme précédemment, il s'agit cependant de la renormaliser (centrage inutile, mais standardisation nécessaire) pour déterminer à partir de quel seuil les déviations de 0 cessent d'être naturelles et liées à l'aléa pour devenir signe que $\mu_X \neq \mu_Y$. C'est là que les difficultés commencent, car pour effectuer efficacement cette standardisation, il faudrait savoir si σ_X et σ_Y sont égales ou pas. C'est pourquoi on procède à un pré-test d'égalité des variances (test de Levene : hypothèse de départ $H_0' : \sigma_X = \sigma_Y$ contre alternative $H_1' : \sigma_X \neq \sigma_Y$), afin de déterminer à quelle statistique de test recourir ensuite (elles sont différentes selon que les variances sont égales ou pas).

Application SPSS et lecture de tableaux. SPSS effectue uniquement des tests bilatères. Il part de l'hypothèse $H_0 : \mu_X = \mu_Y$, et considère une hypothèse alternative symétrique, $H_1 : \mu_X \neq \mu_Y$ (c'est ce que ferait un observateur extérieur sans préjugés). Cela correspond à une zone de rejet symétrique de la forme $]-\infty, -r[\cup]r, +\infty[$. Lors de l'analyse, SPSS produit un tableau du type suivant :

		Test d'égalité des variances	Tests d'égalité des moyennes (sur 5 à 7 colonnes)
Quantités testées	...	P-valeur de cette égalité	Deux tests différents pour l'égalité des moyennes (un par ligne)

Il faut premièrement déterminer si les variances peuvent être dites égales ou non : c'est l'objet des deux premières colonnes. En fonction du résultat, on lira le T-test de la première ligne (variances égales) ou de la seconde ligne (variances différentes).

LA MINUTE SPPS 9.5. La réalisation du tableau précédent est fournie par `Analyse / Comparer les moyennes / Test-T pour échantillons indépendants`.

EXEMPLE 9.5 (Consommation d'alcool à HEC). A la figure 46, on reproduit les résultats d'une enquête quantitative sur la consommation d'alcool à HEC, effectuée en cours

par une collègue. On notera que certaines réponses paraissent assez fantaisistes... et l'on pourra s'étonner que les étudiants pensent spontanément à ne pas se contenter d'indiquer des nombres entiers de verres bus.

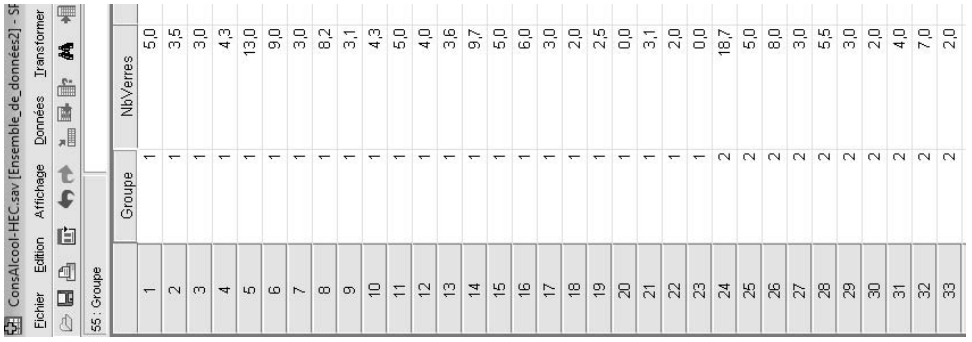


FIGURE 46. Enquête effectuée un vendredi matin d'octobre 2007 : nombre de verres bus au POW de la veille ; résultats donnés selon le groupe d'appartenance (8h, code 1, ou 10h, code 2).

On applique le cadre vu précédemment [21]. SPSS propose le tableau de la figure 47, que nous allons lire ensemble : le pré-test d'égalité des variances associe ici une P-valeur de

Statistiques de groupe

	Groupe	N	Moyenne	Ecart-type	Erreur standard moyenne
Verres bus	8h	23	4,448	3,0598	,6380
	10h	31	7,735	10,1658	1,8258

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Verres bus	Hypothèse de variances égales	5,005	,030	-1,498	52	,140	-3,2877	2,1944	-7,6911	1,1158
	Hypothèse de variances inégales			-1,700	37,021	,098	-3,2877	1,9341	-7,2064	,6311

FIGURE 47. T-test de comparaison des moyennes appliqué aux données de consommation d'alcool : test de la dépendance ou de l'indépendance de la consommation en fonction du groupe d'appartenance (8h ou 10h).

3 % à l'hypothèse H_0' d'égalité des variances, on la rejette donc et on lit la seconde ligne du tableau. Le T-test de comparaison des moyennes associé procure alors quant à lui la P-valeur 9.8 % à l'hypothèse $H_0 : \mu_X = \mu_Y$.

Conclusion statistique : on conserve H_0 et (au moins pour l'instant) on ne peut conclure à une consommation différenciée selon les groupes. (... Bon, en fait, en éliminant les deux ou trois données aberrantes, on pourrait rejeter H_0. Je vous laisse traiter ce cas en exercice.)

Conclusion stratégique : on s'abstiendra de penser que le groupe de 10h est plus fun que celui de 8h (... en l'absence de traitement des données aberrantes).

21. Note culturelle : La modélisation en deux séries de variables aléatoires indépendantes convient toujours ici, mais pas pour les raisons habituelles de tirage de petits échantillons dans chaque population. En effet on interroge pour une fois *tous* les membres de la population mais l'aléa est causé par le fait qu'on ne boit pas exactement la même quantité d'alcool chaque jeudi soir. On compare alors les tendances habituelles μ_X et μ_Y de chacun des deux groupes.

LA MINUTE SPPS 9.6. Regardons quelques autres colonnes du second tableau de la figure 47 : `Différence moyenne` est la valeur réalisée $\overline{x}_n - \overline{y}_m$, ici, $\overline{x}_{23} - \overline{y}_{31}$. `Différence écart-type` est la valeur réalisée des différentes quantités par lesquelles on a renormalisé la différence des moyennes (construites à partir d'estimées de la variance, groupées ou non). Enfin, les deux dernières colonnes proposent des réalisations d'intervalles de confiance sur la différence $\mu_X - \mu_Y$. Les valeurs reportées dans la colonne `ddl` ne peuvent quant à elles être comprises qu'après lecture des compléments de cours facultatifs.

6. Complément (non facultatif) : Tests de normalité

On part de données $x_1, \ldots, x_n$ dont on suppose qu'on peut les modéliser comme la réalisation de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbb{P}_{\theta_0}$. Il s'agit alors de tester que cette loi commune $\mathbb{P}_{\theta_0}$ est une loi normale, i.e., on pose H_0 : il existe $\mu_0 \in \mathbb{R}$ et $\sigma_0 > 0$ tels que $\mathbb{P}_{\theta_0} = \mathcal{N}(\mu_0, \sigma_0^2)$. On teste H_0 contre sa négation H_1 : $\mathbb{P}_{\theta_0}$ n'est pas une loi normale.

Il existe essentiellement deux tests possibles, celui de Shapiro-Wilk et celui de Kolmogorov-Smirnov sous correction de Lilliefors. Ce dernier est aisé à décrire et est présenté dans ses détails mathématiques en annexe facultative. Il est plus délicat de donner la formule du premier, mais il est plus puissant. Je ne vous demande que de savoir mettre en œuvre ces tests sous SPSS et de savoir en lire les tableaux de résultats : voir la figure 48.

Tests de normalité

	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
	Statistique	ddl	Signification	Statistique	ddl	Signification
Salaire horaire	,021	2911	,004	,997	2911	,000

a. Correction de signification de Lilliefors

Tests de normalité

	Type d'infirmière	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
		Statistique	ddl	Signification	Statistique	ddl	Signification
Salaire horaire	Hopital	,019	1945	,113	,999	1945	,457
	Privé	,017	966	,200 [*]	,999	966	,717

a. Correction de signification de Lilliefors

*. Il s'agit d'une borne inférieure de la signification réelle.

FIGURE 48. Tests de normalité appliqués aux salaires des infirmières considérés tous ensemble (tableau du haut) ou séparés selon l'établissement d'exercice (tableau du bas).

LA MINUTE SPPS 9.7. Les tableaux de la figure 48 ont été obtenus par `Analyse / Statistiques descriptives / Explorer`, puis onglet `Diagrammes`, dans lequel on clique sur `Graphes de répartition gaussiens avec tests`.

Conclusion statistique : au vu des P-valeurs respectives (0.4 % et une valeur plus faible que 0.1 %) lues dans les colonnes `Signification`, on peut affirmer que les données de salaire mises toutes ensemble ne peuvent pas être modélisées comme la réalisation de variables aléatoires indépendantes et identiquement distribuées selon une loi normale ; mais qu'en revanche, une fois que l'on coupe ces données en deux séries correspondant chacune à un type d'établissement d'exercice, alors la modélisation normale tient (cf. les

P-valeurs de 11.3 %, et 45.7 % d'une part, pour les salaires d'hôpitaux, et 20.0 % et 71.7 % d'autre part, pour ceux du privé).

Conclusion stratégique : il n'y en a pas, on ne met en œuvre ce genre de tests que pour appliquer une cuisine statistique ultérieure requérant des hypothèses de normalité (par exemple, fondée sur des lois de Student).

Compléments pour étudiants avancés

7. Echantillons indépendants, cas général : compléments mathématiques

Il faut trouver les standardisations adéquates à appliquer à la différence $\overline{X}_n - \overline{Y}_m$. A cet effet, on calcule au préalable la variance de cette différence :

$$\sigma^2(\overline{X}_n - \overline{Y}_m) = \frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2 \ .$$

Il s'agit maintenant de l'estimer et deux cas de figure se présentent, selon que les variances de population σ_X^2 et σ_Y^2 sont égales ou non (selon qu'elles ont été déclarées égales ou non par le pré-test d'égalité des variances).

7.1. Cas de variances de population σ_X^2 et σ_Y^2 différentes. On estime séparément σ_X^2 et σ_Y^2 par leurs estimateurs canoniques,

$$\widehat{\sigma}_{X,n}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 \qquad \text{et} \qquad \widehat{\sigma}_{Y,m}^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2 \ .$$

On considère alors la statistique de test

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\sigma}_{X,n}^2 / n + \widehat{\sigma}_{Y,m}^2 / m}} \ .$$

Sa loi sous H_0 et son comportement sous H_1 sont détaillés dans le principe suivant.

Principe 9.2 (Test d'Aspin-Welch). *On part d'une situation modélisée par le fait qu'une première série d'observations $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus) μ_X et σ_X, tandis qu'une seconde série d'observations $Y_1, \ldots, Y_m$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus) μ_Y et σ_Y. De plus, les deux séries sont indépendantes. On se demande si $\mu_X = \mu_Y$. Le test est fondé sur le résultat suivant : sous $H_0 : \mu_X = \mu_Y$,*

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\sigma}_{X,n}^2 / n + \widehat{\sigma}_{Y,m}^2 / m}} \overset{(d)}{\approx} \mathcal{T}_{\lfloor \nu \rfloor}$$

où $\lfloor \nu \rfloor$ est la partie entière (inférieure) de

$$\nu = \frac{\left(\dfrac{\widehat{\sigma}_{X,n}^2}{n} + \dfrac{\widehat{\sigma}_{Y,m}^2}{m}\right)^2}{\dfrac{\widehat{\sigma}_{X,n}^4}{n^2(n-1)} + \dfrac{\widehat{\sigma}_{Y,m}^4}{m^2(m-1)}} \ ;$$

en particulier, la distribution de $T_{n,m}$ est centrée autour de 0. Lorsque $\mu_X > \mu_Y$, la statistique $T_{n,m}$ tend à prendre des valeurs plus grandes que 0. Lorsque $\mu_X < \mu_Y$, elle tend à prendre des valeurs plus petites que 0.

L'assertion

$$T_{n,m} \overset{(d)}{\approx} \mathcal{T}_{\lfloor \nu \rfloor}$$

signifie que la loi de $T_{n,m}$ est très proche de la loi $\mathcal{T}_{\lfloor \nu \rfloor}$ – une assertion que nous ne quantifierons pas plus précisément. En pratique, on fait comme s'il y avait égalité des distributions lors de la mise en œuvre du test.

Note : en toute rigueur, il faut donc vérifier le caractère normal de la modélisation de chaque série de données pour appliquer ce test (voir la version rédigée du cours : test de Shapiro-Wilk par exemple). Cela étant, en pratique, lorsque n, m (et donc ν) sont grands, l'hypothèse de normalité n'est plus essentielle et la loi suivie par $T_{n,m}$ sous H_0 est approximativement une loi normale standard. C'est un résultat que vous utiliserez en cours de marketing, à ceci près que les P-valeurs, etc., seront toujours calculées par SPSS en utilisant le principe 9.2 et ses lois de Student.

7.2. Cas de variances égales. On suppose ici qu'on peut dire que $\sigma_X = \sigma_Y = \sigma$ (de valeur inconnue). Généralement, cette affirmation est soutenue soit par un pré-test d'égalité des variances (test de Levene : hypothèse de départ H_0' : $\sigma_X = \sigma_Y$ contre alternative H_1' : $\sigma_X \neq \sigma_Y$), soit par le contexte. Ainsi, deux séries de mesures physico-chimiques réalisées sur des objets tirés de deux populations différentes, mais avec le même appareil de mesure et le même opérateur, pourraient être modélisées de la sorte (je vous rappelle que les erreurs de mesure sont typiquement gaussiennes, de variance dépendant de l'appareil et de l'opérateur).

On effectue dans ce cas une estimation groupée de σ, par

$$\widehat{\sigma}^2_{n+m} = \frac{(n-1)\widehat{\sigma}^2_{X,n} + (m-1)\widehat{\sigma}^2_{Y,m}}{n+m-2} \ ,$$

et on considère la statistique de test

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\sigma}^2_{n+m}\,(1/n + 1/m)}} \ .$$

Sa loi sous H_0 et son comportement sous H_1 sont détaillés dans le principe suivant.

PRINCIPE 9.3. *On part d'une situation modélisée par le fait qu'une première série d'observations* $X_1, \ldots, X_n$ *sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus)* μ_X *et* σ*, tandis qu'une seconde série d'observations* $Y_1, \ldots, Y_m$ *sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus)* μ_Y *et* σ*. De plus, les deux séries sont indépendantes. On se demande si* $\mu_X = \mu_Y$*. Le test est fondé sur le résultat suivant : sous* $H_0 : \mu_X = \mu_Y$*,*

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\sigma}^2_{n+m}\,(1/n + 1/m)}} \ \sim \ \mathcal{T}_{n+m-2}$$

où l'estimateur de la variance $\widehat{\sigma}^2_{n+m}$ *(dit estimateur groupé) est défini par*

$$\widehat{\sigma}^2_{n+m} = \frac{(n-1)\widehat{\sigma}^2_{X,n} + (m-1)\widehat{\sigma}^2_{Y,m}}{n+m-2} \ .$$

En particulier, la distribution de $T_{n,m}$ *est centrée autour de* 0*. Lorsque* $\mu_X > \mu_Y$*, la statistique* $T_{n,m}$ *tend à prendre des valeurs plus grandes que* 0*. Lorsque* $\mu_X < \mu_Y$*, elle tend à prendre des valeurs plus petites que* 0*.*

REMARQUE 9.2. Le principe présenté ci-dessus est donc fort similaire au principe 9.2, au remplacement près, dans la définition de la statistique de test, des estimateurs séparés $\widehat{\sigma}^2_{X,n}$ et $\widehat{\sigma}^2_{Y,m}$ dans le dénominateur par la seule version groupée $\widehat{\sigma}^2_{n+m}$. C'est là que l'on exploite l'hypothèse d'égalité des variances. De plus on a ici une loi exacte (et non plus approchée) pour $T_{n,m}$.

REMARQUE 9.3 (A propos de l'hypothèse de normalité). Ici encore, comme pour le test d'Aspin-Welch, on note que lorsque n et m sont grands, le caractère gaussien des observations n'est plus requis, et que $T_{n,m}$ suit approximativement, sous H_0, une loi $\mathcal{N}(0,1)$.

L'intérêt du test du principe 9.3 par rapport à celui du principe 9.2, lorsqu'effectivement les variances sont égales, est que son erreur de deuxième espèce (la probabilité de rejeter à tort H_1 lorsqu'elle est vraie) est plus faible, à erreur de première espèce α fixée (probabilité de rejeter H_0 à tort lorsqu'elle est vraie).

Se pose cependant le problème d'enchaîner deux tests (le pré-test sur l'égalité des variances, puis le T-Test lui-même). Il faut tenir compte du fait que l'erreur de première espèce du test qui combine ces deux sous-tests est la somme des erreurs de première espèce des sous-tests. Cela complique singulièrement le calcul de toute P-valeur, mais rassurez-vous, cela n'étouffera personne dans le cadre du cours de marketing. On vous recommandera même cette procédure en deux temps... !

7.3. Lectures de tableaux de résultats SPSS, suite. A propos du second tableau de la figure 47, vous devriez être en mesure de comprendre maintenant la colonne `ddl`, qui, selon le test considéré, donne la valeur du nombre de degrés de liberté de la loi de Student à considérer, à savoir, ν ou $n+m-2$.

Notez également qu'il est facile d'adapter les résultats des principes 9.3 et 9.2 pour obtenir des intervalles de confiance sur la différence $\mu_X - \mu_Y$. C'est ainsi que SPSS détermine les réalisations des intervalles de confiance proposées dans les deux dernières colonnes.

8. Tests d'ajustement à une loi ou une famille de lois

8.1. Préliminaires : Test d'ajustement à une loi. On part de valeurs observées $x_1, \ldots, x_n$ dont on suppose qu'on les a déjà modélisées comme la réalisation des variables aléatoires $X_1, \ldots, X_n$, indépendantes et identiquement distribuées selon une loi commune notée $\mathbb{P}_{\theta_0}$. On pense ici à une loi de référence $\mathcal{L}_{\text{ref}}$, par exemple la loi normale standard $\mathcal{N}(0,1)$, et on veut tester si $\mathbb{P}_{\theta_0}$ est cette loi $\mathcal{L}_{\text{ref}}$:

$$H_0 : \mathbb{P}_{\theta_0} = \mathcal{L}_{\text{ref}} \qquad \text{contre} \qquad H_1 : \mathbb{P}_{\theta_0} \neq \mathcal{L}_{\text{ref}} \ .$$

C'est ce que l'on appelle un test d'ajustement des observations à une loi donnée.

On a une méthode quasi-universelle pour tester cela, une méthode qui marche quelque soit la loi à densité $\mathcal{L}_{\text{ref}}$. On l'appelle le test de Kolmogorov-Smirnov, et elle compare l'écart entre la fonction de répartition associée à $\mathcal{L}_{\text{ref}}$ et la fonction de réparation $\widehat{F}_n$ associée aux données, appelée fonction de répartition empirique.

Cette dernière est formellement définie comme

$$\widehat{F}_n : x \in \mathbb{R} \longmapsto \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{\{X_j \leqslant x\}} \ ;$$

on saute de $1/n$ à chaque observation, voir l'illustration de la figure 49. On rappelle que la fonction de répartition F_{ref} de $\mathcal{L}_{\text{ref}}$ est quant à elle définie par

$$F_{\text{ref}} : x \in \mathbb{R} \mapsto \mathbb{P}\{L \leqslant x\} \qquad \text{où} \quad L \sim \mathcal{L}_{\text{ref}} \ .$$

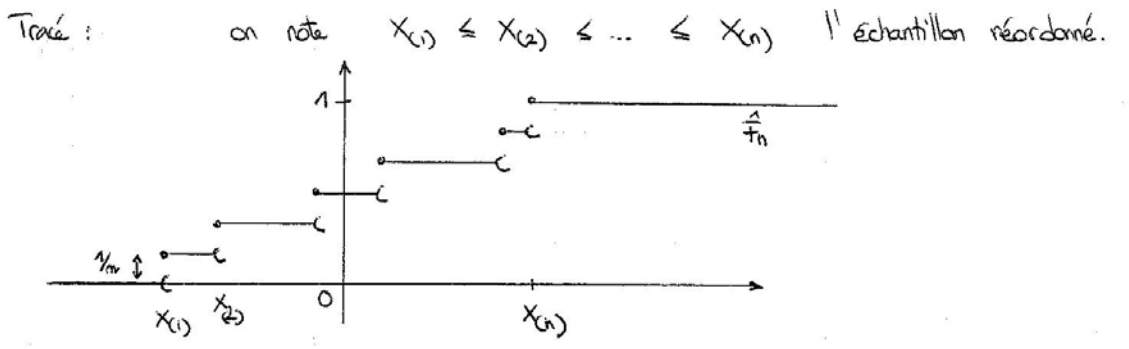


FIGURE 49. Allure typique d'une fonction de répartition empirique.

Le résultat fondamental est que, lorsque $\mathbb{P}_{\theta_0} = \mathcal{L}_{\text{ref}}$, alors $\widehat{F}_n$ est très proche de F_{ref}, c'est le théorème de Glivenko-Cantelli.

THÉORÈME 9.1 (Glivenko-Cantelli). *Sous* $H_0 : \mathbb{P}_{\theta_0} = \mathcal{L}_{\text{ref}}$, *on a*

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_{\text{ref}}(x) \right| \xrightarrow{\mathbb{P}} 0 \ .$$

Un autre résultat essentiel est que la loi de la statistique D_n sous H_0 ne dépend pas de $\mathcal{L}_{\text{ref}}$, tant que cette dernière est à densité. Que $\mathcal{L}_{\text{ref}}$ soit une loi normale standard, une loi exponentielle, une loi uniforme sur $[0, 1]$, etc., peu importe, la loi associée de la statistique D_n sous H_0 ne change pas. On la note $\mathcal{K}_n$ et on l'appelle loi de la statistique de Kolmogorov-Smirnov. Les statisticiens ont tabulé cette loi $\mathcal{K}_n$ (calculé une table de quantiles). On en déduit le principe de test suivant.

PRINCIPE 9.4 (Test de Kolmogorov-Smirnov). *On part d'une situation modélisée par le fait que les observations* $X_1, \ldots, X_n$ *sont indépendantes et identiquement distribuées selon une certaine loi, notée* $\mathbb{P}_{\theta_0}$, *à densité*[22]. *On*[23] *se demande si cette loi est la loi* $\mathcal{L}_{\text{ref}}$. *Le test est fondé sur le résultat suivant : sous* $H_0 : \mathbb{P}_{\theta_0} = \mathcal{L}_{\text{ref}}$,

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_{\text{ref}}(x) \right| \ \sim \ \mathcal{K}_n \ .$$

Lorsque $\mathbb{P}_{\theta_0} \neq \mathcal{L}_{\text{ref}}$, *la statistique* D_n *tend à prendre des valeurs plus grandes que* 0.

De ce principe, vous devez voir comment mettre en œuvre le test correspondant (si je vous donnais une table des quantiles de $\mathcal{K}_n$ – vous pouvez la trouver à la fin des livres que je vous ai mis en référence dans le syllabus de cours).

22. Mais à part ça, totalement inconnue, même sa forme est inconnue.

23. Je vous rappelle qu'on a évidemment une loi candidate en tête, par exemple la loi normale standard $\mathcal{N}(0, 1)$; c'est elle qu'on note $\mathcal{L}_{\text{ref}}$.

8.2. Test de normalité dit de Kolmogorov-Smirnov sous correction de Lilliefors. On part du même cadre que celui des tests de normalité dans la version rédigée du cours, soit de valeurs observées $x_1, \ldots, x_n$ dont on suppose qu'on les a déjà modélisées comme la réalisation de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi commune notée $\mathbb{P}_{\theta_0}$. On veut tester ici si $\mathbb{P}_{\theta_0}$ est une loi normale :

H_0 : Il existe μ et σ^2 tels que $\mathbb{P}_{\theta_0} = \mathcal{N}(\mu, \sigma^2)$

contre $H_1 : \mathbb{P}_{\theta_0}$ n'est pas une loi normale.

On utilise la même statistique que précédemment, à ceci près qu'on remplace $\mathcal{L}_{\text{ref}}$ par la loi $\mathcal{N}(\overline{X}_n, \widehat{\sigma}_n^2)$: on définit

$$D'_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_{\mathcal{N}\left(\overline{x}_n, \widehat{\sigma}_n^2\right)}(x) \right| .$$

La loi de D'_n n'est plus la loi $\mathcal{K}_n$, à cause de l'étape d'estimation des paramètres. Mais là encore, on peut montrer que la loi de D'_n ne dépend pas des vrais paramètres μ_0 et σ_0, ce qui était loin d'être évident. A nouveau, une tabulation de cette loi (appelée loi de la statistique de Kolmogorov-Smirnov sous la correction de Lilliefors) est possible et on peut en déduire un test, dont le principe est tout à fait similaire au principe 9.4, au remplacement près de D_n par D'_n.

On pourra se reporter à la figure 48 et aux commentaires qui la suivent pour la mise en œuvre du test décrit ci-dessus.

La minute SPPS 9.8. Une illustration de la nécessité de savoir ce que SPSS calcule précisément : avec `Analyse / Tests non paramétriques / K-S à 1 échantillon`, il calcule D'_n mais, pour le calcul de la P-valeur, néglige la correction de Lilliefors et utilise la loi $\mathcal{K}_n$. Cela lui permet de proposer en menu le test de différentes distributions, poissonnienne, normale, exponentielle, uniforme ; mais le test réalisé a en général une P-valeur trop optimiste (trop grande). Vous pouvez faire l'essai sur le fichier de données des salaires des infirmières : un premier test de normalité effectué selon cette mauvaise procédure avec le menu `K-S à 1 échantillon` et un second suivant la minute SPSS 9.7.

8.3. Test d'homogénéité de deux populations. On dispose ici de deux séries de données issues d'échantillons indépendants, i.e., tirés indépendamment dans deux populations différentes. On se pose la question de savoir si le comportement des deux populations est le même, i.e., s'il peut être modélisé par la même loi. C'est un test dit d'homogénéité. Remarquez bien que tester l'égalité des comportements moyens (des espérances) des deux lois, ce que nous nous sommes contentés de faire dans la version rédigée du cours, est un objectif bien moins ambitieux. Ici, l'ensemble de la loi, et pas seulement son espérance, nous intéresse.

On modélise cette situation comme auparavant par la considération de deux séries indépendantes $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$, chacune formée de variables aléatoires indépendantes et identiquement distribuées. On note leurs fonctions de répartition empiriques respectives $\widehat{F}_n$ et $\widehat{G}_m$. On se doute que le test reposera sur la comparaison de ces deux fonctions : après tout, si les deux populations suivaient la même loi de fonction de répartition F, alors $\widehat{F}_n$ serait proche de F et de même, $\widehat{G}_m$ serait également proche de F. Par transitivité, $\widehat{F}_n$ et $\widehat{G}_m$ seraient proches. Ici encore, on dispose d'une loi universelle $\mathcal{K}_{n,m}$ qui ne dépend pas de la loi commune, tant que celle-ci est à densité ; elle est également tabulée.

PRINCIPE 9.5 (Test d'homogénéité de Kolmogorov-Smirnov). *On part d'une situation modélisée par le fait qu'une première série d'observations* $X_1, \ldots, X_n$ *sont indépendantes et identiquement distribuées selon une certaine loi* $\mathbb{P}_X$ *(inconnue), tandis qu'une seconde série d'observations* $Y_1, \ldots, Y_m$ *sont indépendantes et identiquement distribuées selon une loi* $\mathbb{P}_Y$ *(également inconnue). De plus, les deux séries sont indépendantes. On se demande si les deux lois sont égales,* $\mathbb{P}_X = \mathbb{P}_Y$*. Le test est fondé sur le résultat suivant : sous* $H_0 : \mathbb{P}_X = \mathbb{P}_Y$*,*

$$D_{n,m} = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - \widehat{G}_m(x) \right| \ \sim \ \mathcal{K}_{n,m} \ .$$

Lorsque $\mathbb{P}_X \neq \mathbb{P}_Y$*, la statistique* D_n *tend à prendre des valeurs plus grandes que* 0*.*

Test de Kolmogorov-Smirnov à deux échantillons

Fréquences

	Groupe	N
Verres bus	8h	23
	10h	31
	Total	54

Test[a]

		Verres bus
Différences les plus extrêmes	Absolue	,202
	Positive	,052
	Négative	-,202
	Z de Kolmogorov-Smirnov	,734
	Signification asymptotique (bilatérale)	,654

a. Critère de regroupement : Groupe

FIGURE 50. Test d'homogénéité entre les consommations d'alcool des deux groupes d'étudiants.

LA MINUTE SPPS 9.9. On met en œuvre ce test avec la séquence `Analyse / Tests non paramétriques / 2 échantillons indépendants`, puis en cochant la case correspondant à `Z de Kolmogorov-Smirnov`. On obtient une figure similaire à la figure 50 sur les données de consommation d'alcool (application du test légitime pour peu que l'on pense avoir affaire à une échelle continue, ce qui est en accord avec les réponses fournies par les étudiants). On voit que les deux lois ne peuvent pas pour l'instant être considérées comme différentes (P-valeur de 65.4 %), au moins en l'absence de traitement des données aberrantes. Précédemment, on avait simplement montré que dans ce cas, leurs moyennes ne pouvaient pas en l'état être considérées comme différentes.

Exercices

Je ne vous propose cette fois-ci que cinq exercices, mais comme vous le verrez, leur correction est, notamment à cause de la rédaction, relativement longue.

Trois exercices issus des annales

Exercice 9.2 (Exercice de synthèse). L'exercice III de l'examen principal 2008 forme un exercice de synthèse assez complet : effectuez-le.

Exercice 9.3 (Comparaison de proportions). Répondez aux questions du paragraphe "Etude de la concurrence d'Internet" de l'examen de rattrapage 2008.

Exercice 9.4 (Données appariées). L'exercice 2 de l'examen principal 2007 propose un cas de données appariées.

Deux exercices issus du cours

Effectuez l'exercice 9.1 de ce cours (le corrigé se trouve dans la version rédigée du cours).

Je vous propose également l'exercice suivant, qui revient sur un tableau de résultats exhibé dans la version rédigée du cours.

Exercice 9.5 (Accidents selon le sexe : à propos de la figure 45). Modélisez les données associées à la figure 45, en justifiant notamment qu'elles entrent dans le cadre du test appliqué. Posez les hypothèses de ce dernier puis refaites à la main les calculs afin de retrouver la P-valeur de la première ligne du tableau de résultats fourni par SPSS.

Par ailleurs, lors du second TP, nous nous entraînerons à lire des tableaux de résultats SPSS, et notamment, à y débusquer les P-valeurs.

Exercice 1. Cf. exercice III de l'examen principal de 2008.

(1) La population visée est celle fréquentant les salles de sport et l'on s'intéresse à la proportion p_0 de femmes à l'intérieur de cette population.
(En particulier, il ne s'agit pas de la proportion de femmes qui fréquentent les salles, mais de la proportion de femmes parmi celles et ceux qui fréquentent les salles de sport.)

On dispose des données $f_1, \dots f_{269}$ où $f_j = 1$ si le j-ème enquêté est une femme, $f_j = 0$ si c'est un homme.
Vu l'interrogation à des heures choisies au hasard à la sortie de salles elles-mêmes choisies au hasard, on peut modéliser ces données comme la réalisation des variables aléatoires $F_1, \dots F_{269}$ iid $\sim Ber(p_0)$, où l'on rappelle que le paramètre d'intérêt p_0 est la proportion de femmes à l'intérieur de la population fréquentant les salles de sport.

Cette sous-population de la population française peut ne pas être représentative de cette dernière si par exemple les proportions de femmes p_0 et $p_{ref} = 51.4\%$ sont significativement différentes.
On est ainsi conduit au test de proportion simple

$$H_0 : p_0 = p_{ref} \quad \text{contre} \quad H_1 : p_0 \neq p_{ref}$$

(choix effectué par un observateur sans préjugés ; quelqu'un ayant déjà assidûment fréquenté les salles de sport pourrait penser à $H_1 : p_0 < p_{ref}$).

La statistique de test est $T_{269} = \sqrt{269}\, \dfrac{\bar{F}_{269} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}$, sous H_0, $T_{269} \overset{(d)}{\approx} \mathcal{N}(0,1)$,

tandis que sous H_1, elle peut prendre des valeurs plus grandes ou

plus petites. La forme de la zone de rejet est donc $]-\infty, -r[\cup]r, +\infty[$.

La valeur réalisée de T_{269} est $\sqrt{269}\left(\frac{\bar{f}_{269} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\right)$

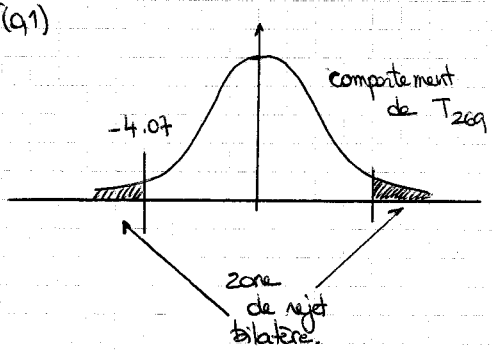
où $\bar{f}_{269} = \frac{105}{269} \approx 39.0\%$

$$= \sqrt{269}\left(\frac{0.390 - 0.514}{\sqrt{0.514\,(1-0.514)}}\right) \approx -4.07$$

On calcule la P-valeur : sur le graphique ci-dessous, la P-valeur vaut

$$\underline{\underline{2}} \times \mathbb{P}\{N \leq -4.07\} \leq 2(1 - 0.999968) \simeq 6 \cdot 10^{-5}$$

où $N \sim \mathcal{N}(0,1)$ $\simeq 0.006\,\%$



Conclusion statistique : la P-valeur est très faible, on rejette sans hésiter H_0 et l'on conclut avec une quasi-certitude que $p_0 \neq p_{ref}$ (en fait, $p_0 < p_{ref}$).

Conclusion stratégique : il faudra toujours garder en tête, d'un point de vue marketing p.ex., que l'on n'a pas affaire à une population homogène au reste des Français et qu'il faudra sans doute créer des campagnes spécifiques.

(2) Il s'agit cette fois de faire un test de comparaison de proportions. On considère toujours comme population l'ensemble des personnes fréquentant les salles de sport et on la divise en deux sous-populations, les hommes et les femmes. On note respectivement $p_♂$ et $p_♀$ les proportions de ces derniers préférant l'image de droite.

On dispose des données $x_1, \dots x_{164}$ d'une part et $y_1, \dots y_{105}$ d'autre part, et, avec les mêmes arguments de tirage au hasard que précédemment, on peut les modéliser comme étant la réalisation de $X_1, \dots X_{164}$ iid $\sim Ber(p_♂)$ et $Y_1, \dots Y_{105}$ iid $\sim Ber(p_♀)$ respectivement. (On rappelle qu'on a utilisé le code 1 pour une préférence envers l'image de droite.) De plus, les X_j sont indépendants des Y_k.

Les données indiquent $\bar{x}_{164} = 75/164 = 45.7\%$ et $\bar{y}_{105} = \frac{54}{105} = 51.4\%$ et on voudrait savoir si elles indiquent des goûts significativement différents. On pose $H_0 : p_♂ = p_♀$ contre $H_1 : p_♂ \neq p_♀$ (cas d'un observateur sans préjugés qui n'a pas de grandes théories sur le fait que les hommes préfèrent les images de femmes, et les femmes, celle d'hommes).

La statistique de test est

$$T_{164,105} = \frac{\bar{X}_{164} - \bar{Y}_{105}}{\sqrt{\hat{p}_{269}(1-\hat{p}_{269})(1/164 + 1/105)}}$$

où $\hat{p}_{269} = \frac{1}{164+105}(X_1 + \dots X_{164} + Y_1 + \dots Y_{105})$.

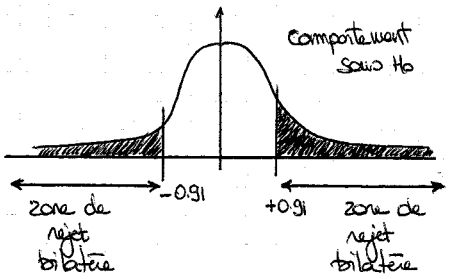
La valeur réalisée de $\hat{p}_{269}$ est $\bar{p}_{269} = \frac{129}{269} = 48.0\%$, de sorte que la valeur réalisée de $T_{164,105}$ est :

$$\frac{\bar{x}_{164} - \bar{y}_{105}}{\sqrt{\bar{p}_{269}(1-\bar{p}_{269})(1/164 + 1/105)}} = \frac{0.457 - 0.514}{\sqrt{0.480(1-0.480)(1/164 + 1/105)}} \simeq -0.91$$

Or, sous H_0, $T_{164,105} \overset{(d)}{\underset{n}{\sim}} \mathcal{N}(0,1)$

tandis que sous H_1, $T_{164,105}$ peut prendre des valeurs au plus grands ou plus petits. La forme de la zone de rejet est donc $]-\infty,-r[\cup]r,+\infty[$.

On calcule la P-valeur à l'aide d'un dessin :



Elle vaut : $\mathbb{P}\{N \leq -0.91\} + \mathbb{P}\{N \geq 0.91\}$ où $N \sim \mathcal{N}(0,1)$

$= 2 \times \mathbb{P}\{N \geq +0.91\}$

$= 2\left(1 - \mathbb{P}\{N \leq 0.91\}\right) = 2(1 - 0.8186)$ par la table de fonction de répartition

$\simeq 36.3\%$

Conclusion statistique : P-valeur élevée, on conserve donc H_0 et on déduit que les données présentées ne permettent pas d'affirmer une différence de goûts significative.

Conclusion stratégique : le DG doit renvoyer la directrice du marketing dans ses cordes et suivre plutôt l'avis des services financiers.

(3) L'achat du gérant de salle :

Il a recueilli les données de prix $x_1, \dots x_{54}$ (Etats-Unis) et $y_1, \dots y_{48}$ (Europe). Il a pris soin de le prendre au hasard de ses recherches sur Internet : on le modélise comme la réalisation de $X_1 \dots X_{54}$ iid selon une certaine loi d'espérance μ_{USA} et de $Y_1, \dots Y_{48}$ iid selon une certaine loi d'espérance μ_{UE}.

Dans le tableau, on lit les estimés respectives $7\,099.6416 \simeq 7100$ € et $6976.8485 \simeq 6977$ € pour μ_{USA} et μ_{UE}.

La question est de savoir si ces données indiquent une différence significative entre μ_{USA} et μ_{UE}, on effectue donc le test

$$H_0 : \mu_{USA} = \mu_{UE} \quad \text{contre} \quad H_1 : \mu_{USA} \neq \mu_{UE}$$

(SPSS procède à un test bilatère).

Lisons la P-valeur dans le tableau : les variances ne pouvant être prises égales (P-valeur de 0.5% à leur test d'égalité), on lit la seconde ligne du second tableau et on note une P-valeur faible égale à 1.4%. On rejette donc H_0 (conclusion statistique).

Conclusion stratégique : Vu les estimés, on a prouvé que c'était moins cher en moyenne en Europe et le gérant devrait donc concentrer ses recherches de prix les plus bas sur cette zone.

Exercice 2. Cf. § Etude de la concurrence d'Internet dans l'examen de rattrapage 2008.

(1) En 2007, on avait recueilli les données $x_1, \dots x_{193}$ ($x_j = 1$ si le j-ème sondé avait fait ou allait faire un complément sur Internet), avec une proportion d'échantillon $\bar{x}_{193} = {}^{35}/_{193} \simeq 18.1\%$

En 2008, il s'agit des données $y_1, \dots y_{172}$ (mêmes codes), avec $\bar{y}_{172} = {}^{41}/_{172} = 23.8\%$

Dans les deux cas, la population étudiée est celle fréquentant Vélizy II, mais on la prend à deux temps distincts et bien séparés.
Vu les sondages menés de manière aléatoire, on peut modéliser ces données comme respectivement la réalisation de $X_1, \dots X_{193}$ iid $\sim$ Ber(p_{2007}) et $Y_1, \dots Y_{172}$ iid $\sim$ Ber(p_{2008}), les X_j étant de plus indépendants des Y_k.

Les paramètres d'intérêt sont p_{2007} et p_{2008} : les proportions, en 2007 et 2008, sur l'ensemble des clients de Vélizy II (plusieurs milliers), qui allaient recourir ou avaient recouru à Internet.

p_{2007} et p_{2008} sont inconnues ; des estimées en sont cependant disponibles : respectivement $\bar{x}_{193} \simeq 18.1\%$ et $\bar{y}_{172} \simeq 23.8\%$.

(2) On va faire un test de comparaison de proportions, et voir si les estimées sont révélatrices de proportions significativement différentes ou non.

On pose $H_0 : p_{2007} = p_{2008}$ contre $H_1 : p_{2007} < p_{2008}$.
Le caractère unilatère est retenu ici à cause de l'impression qu'Internet prend une place de plus en plus grande dans la vie.

La statistique de test est

$$T_{193,172} = \frac{\bar{X}_{193} - \bar{Y}_{172}}{\sqrt{\hat{P}_{365}(1-\hat{P}_{365})\left(1/193 + 1/172\right)}}$$

où l'estimateur groupé est donné par

$$\hat{P}_{365} = \frac{1}{365}\left(X_1 + \dots X_{193} + Y_1 + \dots Y_{172}\right)$$

Sous H_0, $T_{193,172} \overset{(d)}{\simeq} \mathcal{N}(0,1)$

Sous H_1, $\bar{X}_{193} - \bar{Y}_{172}$ et donc $T_{193,172}$ tendent à prendre des valeurs plus petits.

On rejette donc H_0 lorsque $T_{193,172}$ descend en-dessous d'un certain seuil, sort une zone de rejet de la forme $]-\infty, r[$.

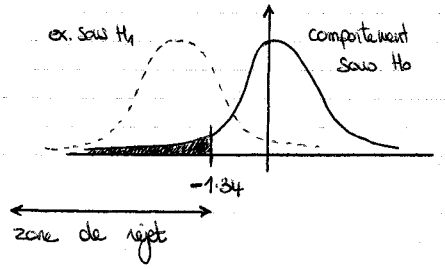
La valeur réalisée de $\hat{P}_{365}$ est $\bar{P}_{365} = \frac{35+41}{193+172} = \frac{76}{365} = 20.8\%$

et celle de $T_{193,172}$ est alors

$$\frac{\bar{x}_{193} - \bar{y}_{172}}{\sqrt{\bar{P}_{365}(1-\bar{P}_{365})\left(1/193 + 1/172\right)}} = \frac{0.181 - 0.238}{\sqrt{0.208\,(1-0.208)\left(1/193 + 1/172\right)}}$$

$$\simeq -1.34$$

On calcule la P-valeur :



Elle vaut $\mathbb{P}\{N \leq -1.34\}$ où $N \sim \mathcal{N}(0,1)$

Par symétrie et manipulations habituelles,

$$\mathbb{P}\{N \leq -1.34\} = 1 - \mathbb{P}\{N \leq 1.34\}$$
$$= 1 - 0.9099 \simeq 9\%$$

Conclusion statistique : on conserve H_0, avec toutefois, un attachement modéré (P-valeur de 9%).

Conclusion stratégique :

- si le syndicat est d'un naturel prudent et rebâcle à la dépense, il conclura de cette étude que pour l'instant, il vaut mieux privilégier un statu quo;
- si le syndicat comporte en son sein des adhérents énergiques ou visionnaires, ceux-ci mettront en avant que déjà en 2008, la P-value est faible, c'est signe d'une évolution des temps, qu'il vaut mieux l'anticiper, etc.

(3) En gros, les principes ont été respectés.
Cependant :

- sur la forme de l'enquête, il y a peut-être un biais dû au fait que l'on n'a visé que le WE (et pas la semaine);
- sur le fond, il aurait été aussi intéressant d'accéder à une autre population visée par le nouveau site : ceux qui ne vont jamais physiquement à Velizy II pour l'instant mais seraient prêts à y aller 2 minutes à un centre de récupération d'achats effectués sur Internet. (On peut les toucher avec un sondage en ligne.)

Exercice 3. Cf. exercice II de l'examen de rattrapage 2007.

Il traite de données appariées.

(1) La figure 1 représente la dispersion des durées de sommeil.
On voit que celles obtenues avec Morpheus sont plus grandes en moyenne (cf. médiane plus grande, cf. traits centraux des boîtes).
De plus, elles sont moins dispersées autour de cette meilleure tendance centrale (boîte de droite plus ramassée que celle de gauche).

Pour ces deux raisons (meilleure efficacité en moyenne, meilleure précision ou garantie sur le résultat), on tendrait intuitivement à pencher pour Morpheus (l'ancien médicament).

Intermède: MODÉLISATION

Les données $x_1, \dots x_{10}$ et $y_1, \dots y_{10}$ sont modélisées comme la réalisation des variables aléatoires $X_1, \dots X_{10}$ et $Y_1, \dots Y_{10}$; les X_j sont iid selon une certaine loi d'espérance μ_X (durée moyenne de sommeil sur patients traités avec DodoPlus) et les Y_k, iid selon une certaine loi d'espérance μ_Y (durée moyenne de sommeil sur patients traités avec Morpheus).

Les caractères iid sont garantis par le choix au hasard des cobayes.*
Note : μ_X et μ_Y sont inconnus, car la population concernée (patients ou futurs patients prenant des somnifères) est très grande.

*Certains petits malins avaient précisé dans leur copie: "et à condition que chacun dorme seul dans son lit."

---

(2) La dépendance entre X_j et Y_j provient de la tendance naturelle du j-ème sujet à dormir beaucoup ou pas: on parle de données appariées.

Note : Contrairement à ce que j'avais lu à l'époque, cette dépendance ne provient pas des restes d'effets du somnifère du jour J au jour J+1, parce que ces derniers disparaissent évidemment en 10h - 12h !

(3) Cf. § modélisation : C'est à cause du choix au hasard des cobayes. Les Z_j sont iid selon une certaine loi admettant une espérance notée Δ et valant, avec les notations précédentes et par linéarité de l'espérance, $\Delta = \mu_X - \mu_Y$.

(4) On effectue le test de Shapiro-Wilk sur les données $z_j = x_j - y_j$.
Cela teste : H_0 : la loi commune des Z_j est une loi normale.

On lit la P-valeur : 31.57 % et on conserve donc H_0.

Conclusion pour la suite : on pensera désormais que $Z_1, \dots Z_{10}$ sont iid $\sim \mathcal{N}(\Delta, \sigma^2)$.

(5) Le laboratoire claime que le nouveau médicament DodoPlus est au moins aussi efficace que Morpheus. Comme ce labo est de mauvaise foi (cf. question (1)) et un peu manipulateur, il pense à :

$$H_0 : \Delta = 0 \quad \text{contre} \quad H_1 : \Delta < 0 \qquad \text{ie } \Delta_{ref} = 0$$
$$(\text{ou : } H_0 : \Delta \geqslant 0)$$

Il joue évidemment sur le fait qu'un test a tendance à conserver H_0.

(6) On recourt à un test de Student, vu que l'on a un échantillon de petite taille mais dont on a pu supposer qu'il était issu d'une loi normale.

La statistique de test est $T_{10} = \sqrt{10}\left(\dfrac{\overline{Z}_{10} - \Delta_{ref}}{\sqrt{\hat{\sigma}_Z^2}}\right)$

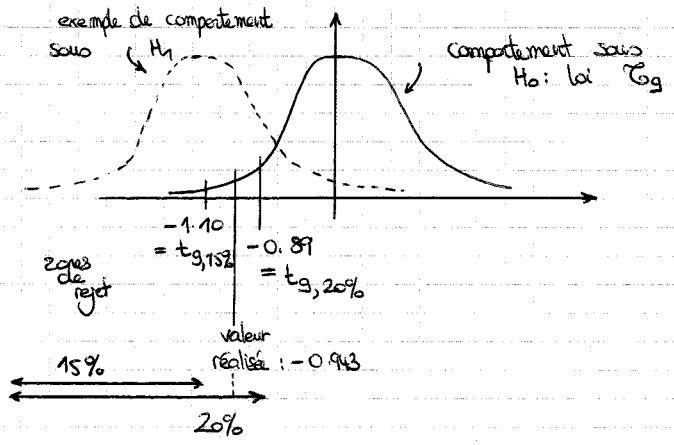
où $\hat{\sigma}_Z^2$ est l'estimateur canonique de

la variance construit sur $Z_1, \dots Z_{10}$.

Sous H_0, $T_{10} \sim \mathcal{T}_9$ tandis que sous H_1, T_{10} tend à prendre des valeurs plus petites. La zone de rejet est donc de la forme $]-\infty, r[$.

La valeur réalisée de T_{10} est $\sqrt{10}\left(\frac{\bar{z}_{10} - 0}{\sqrt{s^2_{z_{10}}}}\right) = \sqrt{10}\left(\frac{-0.13}{\sqrt{0.19}}\right) \simeq -0.943$

On peut alors encadrer la P-valeur avec la table des quantiles de la loi de Student :



La P-valeur $\mathbb{P}[T \leq -0.943]$ où $T \sim \mathcal{T}_9$ appartient à l'intervalle $[15\%, 20\%]$ (un calcul informatique donne $\simeq 18.5\%$).

Conclusion statistique : on conserve H_0 (au vu des 10 seules données dont on dispose).

La question suivante étudie quant à elle les conclusions stratégiques à tirer ou non de cette conclusion statistique.

(7) Il vaut mieux ne tirer (pour les patients) aucune conclusion stratégique au vu de cette étude : 10 données, c'est peu, les hypothèses favorisent le labo, etc. En bref, on a affaire ici à un argument marketing qui essaie de se fonder mathématiquement, mais on est en train de se faire manipuler !

Exercice 4. → Accidents selon le sexe.

On étudie les assurés d'une compagnie donnée et, visiblement, on a tiré au hasard et indépendamment, 250 hommes et 250 femmes dans ce fichier. (Des nombres aussi ronds indiquent bien des échantillonnages séparés.)

Cela a procuré les données $x_1, \dots x_{250}$ pour les hommes et $y_1, \dots y_{250}$ pour les femmes. On note $x_j = 1$ si le j-ème sondé homme (respectivement, $y_j = 1$ si la j-ème sondée femme) a eu au moins un accident responsable au cours des cinq dernières années. Avec : $\bar{x}_{250} = \frac{204}{250} = 81.6\%$ et $\bar{y}_{250} = \frac{174}{250} \simeq 69.6\%$

On les modélise, via leur tirage au hasard, comme la réalisation de $X_1 \dots X_{250}$ iid $\sim \mathrm{Ber}(p_\male)$ et $Y_1, \dots Y_{250}$ iid $\sim \mathrm{Ber}(p_\female)$, où les paramètres d'intérêt (inconnus) $p_\male$ et $p_\female$ désignent respectivement les taux, sur l'ensemble des assurés, des accidents responsables selon le sexe.

On effectue le test bilatère de $H_0 : p_\male = p_\female$ contre $H_1 : p_\male \neq p_\female$.
(Cas d'un statisticien sans préjugés.)

On recourt à la statistique de test

$$T_{250,250} = \frac{\bar{X}_{250} - \bar{Y}_{250}}{\sqrt{\hat{p}_{500}(1-\hat{p}_{500})\left(\frac{1}{250}+\frac{1}{250}\right)}}$$

où $\hat{p}_{500} = \frac{1}{500}\left(X_1 + \dots X_{250} + Y_1 + \dots Y_{250}\right)$

Sous H_0, $T_{250,250} \overset{(d)}{\simeq} \mathcal{N}(0,1)$ tandis que sous H_1, $T_{250,250}$ peut prendre des valeurs au plus grandes ou plus petites.
La zone de rejet est donc de la forme $]-\infty, -r[\cup]r, +\infty[$.

Les valeurs réalisées de $\hat{p}_{500}$ et $T_{250,250}$ sont respectivement

$$\bar{p}_{500} = \frac{1}{500}\left(204 + 174\right) \simeq 75.6\%$$

et $$\frac{\bar{x}_{250} - \bar{y}_{250}}{\sqrt{\bar{p}_{500}(1-\bar{p}_{500})(1/250 + 1/250)}} = \frac{0.816 - 0.696}{\sqrt{0.756(1-0.756)(1/250+1/250)}}$$

$$\simeq 3.12$$

Note : Pour des raisons que nous expliquerons dans la suite de ce polycopié, cette valeur est en fait la racine carrée de la valeur réalisée 9.758 de la statistique du Khi-deux (χ^2) de Pearson.

La P-valeur est donnée, selon le dessin habituel, par

$$\mathbb{P}\{N \leq -3.12 \text{ ou } N \geq 3.12\} = 2 \times \mathbb{P}\{N \geq 3.12\} = 2\left(1 - \mathbb{P}\{N \leq 3.12\}\right)$$

$$\simeq 2(1 - 0.999) \quad \text{selon la table}$$

$$= 0.2\ \%$$

On retrouve bien la P-valeur bilatère de 0.2% du tableau.

Conclusion statistique : rejet très clair de H_0, soit, au vu de $\bar{x}_{250}$ et $\bar{y}_{250}$: affirmation très claire du fait que les femmes conduisent mieux que les hommes.

Conclusion stratégique : créer une compagnie d'assurances dédiée aux femmes, ou, à tout le moins, leur faire payer à conditions (expérience, type de voiture, zone géographique) égales, moins cher qu'aux hommes.
... Et faire une grande campagne de pub à ce sujet !

Dixième Partie

Tests du χ^2

Version rédigée du cours

Résumé : La partie précédente s'est achevée sur les tests de normalité (tests de Shapiro-Wilk et de Kolmogorov-Smirnov avec correction de Lilliefors), i.e., des tests d'ajustement à la famille des lois normales.

Objectif : Nous traitons ici des problèmes d'ajustement pour des lois chargeant un nombre fini de points, typiquement, l'ensemble des valeurs qu'une variable qualitative (ordinale ou nominale) peut prendre, ou qu'un tel couple de variables peut prendre. Nous étudions tout d'abord l'ajustement à une loi de référence, puis apprendrons à tester l'indépendance entre deux variables qualitatives.

1. Motivation : non pas manipuler mais détecter les manipulations

Certains enseignants de statistique aiment commencer leur cours en citant :

> *Do not trust any statistics you did not fake yourself.*
> Winston Churchill (homme politique britannique, 1874–1965)

Cette citation souligne certes le côté parfois politique des statistiques institutionnelles, tant dans le recueil et la présentation des données (changement des critères de définition du chômage par exemple) que dans leur traitement (choix des hypothèses d'un test à mener par exemple).

Mais surtout, elle indique qu'il est tentant d'embellir, voire de truquer ou d'inventer, des données. Alors, certes, je pourrais vous apprendre à manipuler des données. Mais ce serait mal. Par conséquent, comme dans *Harry Potter*, vous aurez au contraire droit dans les pages qui suivent à un cours de défense contre les forces du mal : un cours qui vous permette de détecter certaines manipulations des données. Ainsi, nous verrons comment des statisticiens [24] ont pu détecter des fraudes aux élections présidentielles iraniennes (voir la figure 51).

24. Attention, l'article que je vous propose indique une démarche intéressante mais la mise en œuvre de cette dernière est plutôt mauvaise et a été critiquée par de nombreux statisticiens, qui rappelaient que dans le cas d'espèce il ne fallait pas raisonner en termes d'intervalles de confiance sur des proportions mais plutôt avec un test du χ^2 d'adéquation à une loi uniforme. Nous reverrons cela plus tard dans ce cours.

The Washington Post

The Devil Is in the Digits

By Bernd Beber and Alexandra Scacco
Saturday, June 20, 2009 12:02 AM

Since the declaration of Mahmoud Ahmadinejad's landslide victory in Iran's presidential election, accusations of fraud have swelled. Against expectations from pollsters and pundits alike, Ahmadinejad did surprisingly well in urban areas, including Tehran -- where he is thought to be highly unpopular -- and even Tabriz, the capital city of opposition candidate Mir Hussein Mousavi's native East Azarbaijan province.

Others have pointed to the surprisingly poor performance of Mehdi Karroubi, another reform candidate, and particularly in his home province of Lorestan, where conservative candidates fared poorly in 2005, but where Ahmadinejad allegedly captured 71 percent of the vote. Eyebrows have been raised further by the relative consistency in Ahmadinejad's vote share across Iran's provinces, in spite of wide provincial variation in past elections.

These pieces of the story point in the direction of fraud, to be sure. They have led experts to speculate that the election results released by Iran's Ministry of the Interior had been altered behind closed doors. But we don't have to rely on suggestive evidence alone. We can use statistics more systematically to show that this is likely what happened. Here's how.

We'll concentrate on vote counts -- the number of votes received by different candidates in different provinces -- and in particular the last and second-to-last digits of these numbers. For example, if a candidate received 14,579 votes in a province (Mr. Karroubi's actual vote count in Isfahan), we'll focus on digits 7 and 9.

This may seem strange, because these digits usually don't change who wins. In fact, last digits in a fair election don't tell us anything about the candidates, the make-up of the electorate or the context of the election. They are random noise in the sense that a fair vote count is as likely to end in 1 as it is to end in 2, 3, 4, or any other numeral. But that's exactly why they can serve as a litmus test for election fraud. For example, an election in which a majority of provincial vote counts ended in 5 would surely raise red flags.

Why would fraudulent numbers look any different? The reason is that humans are bad at making up numbers. Cognitive psychologists have found that study participants in lab experiments asked to write sequences of random digits will tend to select some digits more frequently than others.

So what can we make of Iran's election results? We used the results released by the Ministry of the Interior and published on the web site of Press TV, a news channel funded by Iran's government. The ministry provided data for 29 provinces, and we examined the number of votes each of the four main candidates -- Ahmadinejad, Mousavi, Karroubi and Mohsen Rezai -- is reported to have received in each of the provinces -- a total of 116 numbers.

The numbers look suspicious. We find too many 7s and not enough 5s in the last digit. We expect each digit (0, 1, 2, and so on) to appear at the end of 10 percent of the vote counts. But in Iran's provincial results, the digit 7 appears 17 percent of the time, and only 4 percent of the results end in the number 5. Two such departures from the average -- a spike of 17 percent or more in one digit and a drop to 4 percent or less in another -- are extremely unlikely. Fewer than four in a hundred non-fraudulent elections would produce such numbers.

As a point of comparison, we can analyze the state-by-state vote counts for John McCain and Barack Obama in last year's U.S. presidential election. The frequencies of last digits in these election returns never rise above 14 percent or fall below 6 percent, a pattern we would expect to see in seventy out of a hundred fair elections.

But that's not all. Psychologists have also found that humans have trouble generating non-adjacent digits (such as 64 or 17, as opposed to 23) as frequently as one would expect in a sequence of random numbers. To check for deviations of this type, we examined the pairs of last and second-to-last digits in Iran's vote counts. On average, if the results had not been manipulated, 70 percent of these pairs should consist of distinct, non-adjacent digits.

Not so in the data from Iran: Only 62 percent of the pairs contain non-adjacent digits. This may not sound so different from 70 percent, but the probability that a fair election would produce a difference this large is less than 4.2 percent. And while our first test -- variation in last-digit frequencies -- suggests that Rezai's vote counts are the most irregular, the lack of non-adjacent digits is most striking in the results reported for Ahmadinejad.

Each of these two tests provides strong evidence that the numbers released by Iran's Ministry of the Interior were manipulated. But taken together, they leave very little room for reasonable doubt. The probability that a fair election would produce both too few non-adjacent digits and the suspicious deviations in last-digit frequencies described earlier is less than .005. In other words, a bet that the numbers are clean is a one in two-hundred long shot.

Bernd Beber and Alexandra Scacco, Ph.D. candidates in political science at Columbia University, will be assistant professors in New York University's Wilf Family Department of Politics this fall.

FIGURE 51. Un article relatant comment le truquage des élections présidentielles par le gouvernement iranien a été détecté. On lira notamment la phrase : "Each of these two tests provides strong evidence that the numbers released by Iran's Ministry of the Interior were manipulated."

2. Test du χ^2 d'ajustement simple

2.1. Un exemple pour introduire les notations.

EXERCICE 10.1 (Agence immobilière[25]). On dit souvent que c'est au printemps qu'il se vend et s'achète le plus de logements, en prévision des déménagements d'été. Le gérant d'une agence immobilière, fort de ses certitudes, explique donc à ses employés qu'il faut absolument recruter un stagiaire pour les aider dans leurs tâches au printemps (mois d'avril, mai et juin). Pour cela, il s'apprête à leur demander de faire le tour des formations BTS. Il voudrait au préalable les convaincre de la nécessité de cette démarche. Y arrivera-t-il au vu de l'historique, mois par mois, des ventes de l'an dernier ?

Mois	J	F	M	A	M	J	J	A	S	O	N	D
Ventes	1	3	4	6	6	5	3	1	2	1	2	2

Il a donc réalisé 36 ventes. Elles se distribuent selon les saisons comme suit : 8 ventes en hiver, 17 au printemps, 6 en été, 5 en automne. On se demande si la répartition est uniforme suivant la saison, ou s'il y a un pic statistiquement significatif de mouvements immobiliers au printemps. Si la répartition avait été uniforme, on aurait attendu environ 9 ventes à chaque saison (pas exactement 9, bien sûr, mais 9 plus ou moins un certain nombre de ventes à quantifier). On résume ceci dans le tableau suivant.

Saisons	hiver	printemps	été	automne
Ventes attendues	9	9	9	9
Ventes réalisées	8	17	6	5
Fréquences attendues	25%	25%	25%	25%
Fréquences réalisées	22.2%	47.2%	16.7%	13.9%

2.2. Notations et principes généraux. On part de données qualitatives (ordinales ou nominales) $x_1, \ldots, x_n$. Sans perte de généralité, on note $\{1, 2, \ldots, k\}$ l'ensemble des valeurs qu'elles peuvent prendre ; on peut agir de la sorte, quitte à ré-étiqueter les catégories.

On suppose qu'on peut modéliser les données comme la réalisation des variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p} = (p_1, \ldots, p_k)$ sur $\{1, 2, \ldots, k\}$, inconnue.

On pense, pour des raisons subjectives et qui dépendent du contexte, à une certaine loi de référence $\mathbf{p}^{\text{ref}} = (p_1^{\text{ref}}, \ldots, p_k^{\text{ref}})$ et on veut tester

$$H_0 : \mathbf{p} = \mathbf{p}^{\text{ref}} \qquad \text{contre} \qquad H_1 : \mathbf{p} \neq \mathbf{p}^{\text{ref}} .$$

(H_1 se ré-écrit comme : il existe j tel que $p_j \neq p_j^{\text{ref}}$.)

L'intuition du test est la suivante. Vous vous souvenez qu'en général, pour construire un test, on commence par étudier le comportement d'une certaine statistique de test sous

25. Cf. troisième exercice de l'examen principal de 2007

H_0, pour voir comment elle évolue et quelles sont ses valeurs typiques et atypiques. Or, si H_0 est vraie, on s'attend par loi des grands nombres à ce que pour tout $j = 1, \ldots, k$, la proportion observée $\widehat{p}_{j,n}$ de la catégorie j dans les données soit proche de p_j :

$$\widehat{p}_{j,n} = \frac{N_{j,n}}{n} \xrightarrow{\mathbb{P}} p_j \qquad \text{où} \qquad N_{j,n} = \mathrm{Card}\big\{t : X_t = j\big\}$$

compte le nombre de variables aléatoires X_t prenant la valeur j.

Ainsi, une mesure naturelle de la validité de H_0 consiste à regarder la proximité ou la distance de chacune de ces $\widehat{p}_{j,n}$ aux p_j ; plus précisément, un gros bout de théorie mathématique admise (et qui fournit pas mal d'intuitions, hélas, nous n'avons pas le temps de la voir...) amène à considérer la statistique de test suivante :

$$D_n(\mathbf{p}^{\mathrm{ref}}) \overset{\text{not.}}{=} n \sum_{j=1}^{k} \frac{\big(\widehat{p}_{j,n} - p_j^{\mathrm{ref}}\big)^2}{p_j^{\mathrm{ref}}} = \sum_{j=1}^{k} \frac{\big(N_{j,n} - n\, p_j^{\mathrm{ref}}\big)^2}{n\, p_j^{\mathrm{ref}}} \ .$$

Un calcul très simple montre l'égalité des deux définitions ; la première mesure l'écart entre les fréquences attendues et celles réalisées, la seconde, entre les effectifs attendus et ceux réalisés. (Retenez la forme qui vous parle le plus.)

Pour la suite, afin d'étudier le comportement de $D_n(\mathbf{p}^{\mathrm{ref}})$ sous H_0 et H_1, nous aurons besoin de la définition suivante (qui avait déjà été présentée dans la partie 5, mais dans les compléments facultatifs...).

Définition 10.1. *Si $Z_1, \ldots, Z_k$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi $\mathcal{N}(0,1)$, alors on appelle la loi de $Z_1^2 + \ldots + Z_k^2$ la loi du χ^2 à k degrés de liberté et on la note χ^2_k. Les quantiles de ces lois sont fournis dans les tables à la fin du présent polycopié.*

Or, la théorie mathématique assure que sous H_0, la statistique $D_n(\mathbf{p}^{\mathrm{ref}})$ converge en loi vers une loi du χ^2 à $k-1$ degrés[26] de liberté, notée χ^2_{k-1}.

Sous H_1, les écarts entre les $\widehat{p}_{j,n}$ et les p_j tendent à être grands, et le facteur multiplicatif n dans la première définition assure alors que $D_n(\mathbf{p}^{\mathrm{ref}}) \xrightarrow{\mathbb{P}} +\infty$ lorsque $n \to \infty$. Tout cela conduit au principe 10.1.

Ainsi, la zone de rejet associée au test est toujours unilatère ; elle vaut, pour une erreur de première espèce α fixée : $\big]\, c_{k-1,1-\alpha}, +\infty\big[$, où $c_{k-1,1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi χ^2_{k-1}.

Remarque 10.1 (Conditions pratiques d'approximation asymptotique). Notez bien que le test, reposant sur une convergence en loi, est un test asymptotique. En pratique, on obtient de bons résultats dès lors que l'échantillon est suffisamment grand, $n \geqslant 30$, et que toutes les valeurs possibles j conduisent à des effectifs attendus $n\, p_j^{\mathrm{ref}} \geqslant 5$. Si la seconde condition n'est pas satisfaite, il faut alors procéder à un regroupement de classes (l'agrégation de deux classes j et j' en une seule classe). Nous en verrons des exemples concrets. Attention, cette remarque vaut également pour le test du χ^2 d'indépendance, étudié dans la seconde partie de ce cours.

26. Pourquoi $k-1$ et pas k ? C'est facile à retenir : une probabilité sur un ensemble à k éléments est définie par $k-1$ de ses valeurs, la dernière se déduisant des précédentes : on a effectivement $k-1$ degrés de liberté, la terminologie est éclairante.

PRINCIPE 10.1. *Test d'ajustement simple à une loi de référence* $\mathbf{p}^{\text{ref}}$
Données : $x_1, \dots, x_n \in \{1, \dots, k\}$
Modélisation associée : observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p}$ sur $\{1, \dots, k\}$
Hypothèse H_0 **:** $\mathbf{p} = \mathbf{p}^{\text{ref}}$
Statistique de test :

$$D_n(\mathbf{p}^{\text{ref}}) \overset{\text{not.}}{=} n \sum_{j=1}^{k} \frac{(\widehat{p}_{j,n} - p_j^{\text{ref}})^2}{p_j^{\text{ref}}} = \sum_{j=1}^{k} \frac{(N_{j,n} - n\, p_j^{\text{ref}})^2}{n\, p_j^{\text{ref}}}$$

Comportement sous H_0 **:** $D_n(\mathbf{p}^{\text{ref}}) \rightharpoonup \chi^2_{k-1}$ lorsque $n \to \infty$
Comportement sous H_1 **:** $D_n(\mathbf{p}^{\text{ref}})$ tend vers $+\infty$ et prend donc des valeurs beaucoup plus grandes que sous H_0.

FIGURE 52. Principe du test du χ^2 dit d'ajustement simple à une loi de référence.

LA MINUTE SPPS 10.1. Lorsque les conditions asymptotiques ne sont pas remplies, SPSS vous le fait savoir par une note en bas de tableau. Il faut alors fusionner des catégories (ce qui est possible en cliquant bien, par exemple avec `Transformer / Recoder des variables`).

2.3. Retour à l'exemple introductif.

CORRECTION 10.1 (Agence immobilière). La saison de vente d'un logement est notée v_j : ici, on a donc affaire aux données $v_1, \dots, v_{36}$. On les modélise comme étant la réalisation des variables aléatoires $V_1, \dots, V_{36}$ indépendantes et identiquement distribuées selon une certaine loi (inconnue) sur l'ensemble à quatre éléments $\{H, P, E, A\}$ contenant les quatre saisons. En effet, les dates de ventes des logements proposés par cette agence forment un échantillon aléatoire des différents dates de ventes des logements à vendre dans la zone géographique considérée, dès lors que :

- les vendeurs choisissent une agence un peu au hasard,
- il y a de nombreux autres biens sur le marché au même moment, chacun d'eux vivant alors sa vie propre et se vendant d'autant plus vite qu'il est de bonne qualité (ce qui règle le cas du comportement des acheteurs, i.e., du délai de vente).

En revanche, on suppose que les vendeurs mettent leur bien sur le marché au moment qui les arrange, de telle sorte que la promesse de vente, puis l'acte définitif (trois mois après) soient signés à-peu-près au moment qui les arrange : le choix du mois où mettre son bien en vente n'est lui pas du tout déterminé au hasard.

On note $\mathbf{p} = (p_H, p_P, p_E, p_A)$ la loi commune des observations $V_1, \dots, V_{36}$. On se demande si c'est la loi uniforme, soit $H_0 : \mathbf{p} = \mathbf{p}^{\text{ref}}$, où $\mathbf{p}^{\text{ref}} = (1/4, \dots, 1/4)$.

Ici, on a $k = 4$ classes ; les conditions asymptotiques décrites à la remarque 10.1 sont vérifiées (taille d'échantillon de 36 et effectif attendu minimal de $36/4 = 9$) ; et la valeur réalisée pour la statistique D_{36} du χ^2 (qui approximativement sous H_0 la loi χ^2_3) est

$$\frac{(8-9)^2}{9} + \frac{(17-9)^2}{9} + \frac{(6-9)^2}{9} + \frac{(5-9)^2}{9} = 10\ .$$

Les tables que je vous ai fournies permettent d'encadrer la P-valeur entre 1 % et 2.5 %, en lisant la ligne de la loi χ^2_3 (voir la figure 53 pour une illustration). Avec un logiciel statistique, on verrait qu'elle vaut plus exactement 1.9 %. Dans tous les cas, on rejette assez nettement l'hypothèse de répartition uniforme des ventes selon les saisons (conclusion statistique). La conclusion stratégique est qu'une aide ponctuelle aux mois chargés est plus que bienvenue.

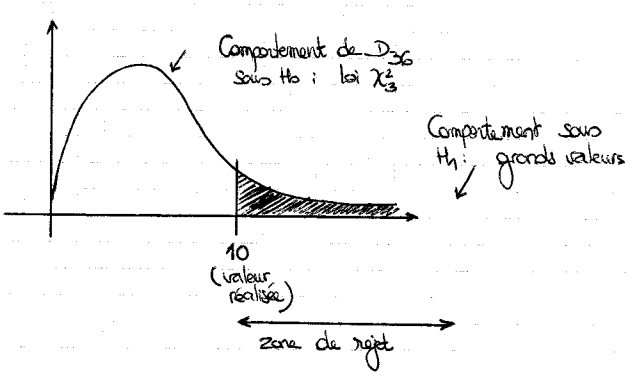


FIGURE 53. Illustration du calcul de la P-valeur dans l'exercice sur les ventes d'une agence immobilière.

LA MINUTE SPPS 10.2. J'ai implémenté l'exemple de l'agence immobilière sous le fichier de données `Immo.sav`, disponible sur le site web du cours. La colonne 2 contient les mois de vente et correspond bien aux données présentées dans l'exemple 10.1. Les colonnes 1 et 3 donnent respectivement le jour et le prix de vente (j'ai rempli ces colonnes au hasard). J'ai déduit les colonnes 4 et 5 (saison et code de la saison) de la colonne 2, via `Transformer / Recoder des variables`. On lance le test par `Analyse / Tests non paramétriques / Khi-deux` et dans la fenêtre qui apparaît, on indique qu'on veut tester l'équiprobabilité des classes sur la variable `CodeSaison`. On obtient alors la figure 54. On retrouve bien toutes les valeurs numériques que nous avions calculées à la main, et notamment, la P-valeur de 1.9 %. La petite note qui suit le second tableau indique que les conditions asymptotiques de la remarque 10.1 sont bien remplies et que le test est donc validement appliqué.

2.4. Autres exemples.

EXEMPLE 10.1 (Détection de tricheries électorales en Iran). Essayons de déterminer la méthodologie employée à la figure 51 : les votes en faveur des 4 candidats en lice dans chacun des 29 provinces ont été fournis par le gouvernement ; ce sont des nombres chacun de l'ordre de plusieurs milliers ou dizaines de milliers. La modélisation consiste à dire que les chiffres des unités de ces grands nombres peuvent être considérés comme la réalisation de variables aléatoires indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p}$ (ils n'ont pas de signification).

Si par ailleurs ces chiffres sont reportés de manière sincère, alors les données auxquelles on a affaire sont formées de 116 chiffres des unités $u_1, \dots, u_{116}$, réalisations de variables aléatoires $U_1, \dots, U_{116}$ indépendantes et identiquement distribuées selon la loi uniforme $\mathbf{p}^{\text{ref}}$ sur $\{0, 1, \dots, 9\}$.

Test du Khi-deux

CodeSaison

	Effectif observé	Effectif théorique	Résidu
Hiver	8	9,0	-1,0
Printemps	17	9,0	8,0
Eté	6	9,0	-3,0
Automne	5	9,0	-4,0
Total	36		

Test

	CodeSaison
Khi-deux	10,000[a]
ddl	3
Signification asymptotique	,019

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 9,0.

FIGURE 54. Résultat du test du χ^2 d'ajustement à une loi uniforme sur les ventes selon la saison.

Sinon, lorsque les chiffres sont manipulés, il faut savoir que les êtres humains ont tendance à mal inventer des nombres aléatoires et sous-représentent certains chiffres (le 5, le 0) et sur-représentent certains autres (le 7). Cela donne lieu à la loi commune $\mathbf{p}$ pour les U_j (qui dépend du manipulateur).

Les auteurs ont ensuite testé $H_0 : \mathbf{p} = \mathbf{p}^{\text{ref}}$ contre $H_1 : \mathbf{p} \neq \mathbf{p}^{\text{ref}}$. En fait, ils ne l'ont pas fait avec un test du χ^2 et cela leur a été vertement reproché par la communauté scientifique. S'ils l'avaient fait, leur conclusion serait restée la même mais aurait eu la force d'une conclusion scientifiquement prouvée : les chiffres de cette élection ont été manipulés ! (L'absence des données précises m'empêche de vous présenter les détails de ce calcul.)

EXEMPLE 10.2 (Détection des tricheries comptables). Une méthodologie similaire peut être employée pour auditer la comptabilité d'une entreprise. La loi du premier chiffre des nombres présents dans ces écritures ne suit pas une loi uniforme mais une loi dite de Benford. Cette loi stipule par exemple que le 1 a une probabilité d'apparition de 30.10 %, le 2, une probabilité 17.61 %, ..., le 9, une probabilité 4.58 %. Or, ceux qui trichent [27], là encore, tendent à modifier cette répartition, en sur- ou sous-représentant significativement certains chiffres.

Ces techniques sont réellement utilisées en pratique, au moins aux Etats-Unis, depuis le milieu des années 90. Il est ainsi de notoriété publique qu'à New-York, le parquet a pu ainsi avoir des doutes sur sept entreprises, qui, après vérifications, falsifiaient effectivement leur bilan. Les auditeurs américains utilisent couramment le test du χ^2 d'ajustement à la loi de Benford pour avoir une première idée de la sincérité des écritures comptables.

Tomasz Michalski (qui est peut-être votre enseignant d'économie) et moi-même avons récemment employé ces tests d'ajustement pour vérifier la sincérité des balances commerciales des Etats (voir l'article : "Do countries falsify economic data strategically ? Some evidence that they do" sur mon site).

27. Un exemple célèbre est donné par une secrétaire qui détournait des fonds : elle avait l'autorisation de signature pour les chèques inférieurs à 1 000 dollars, et devinez quoi, le 9 était sur-représenté dans les lignes comptables, à cause de nombreux débits de 900 dollars et quelques...

3. Test du χ^2 d'indépendance entre deux variables qualitatives

Ce paragraphe présente des résultats absolument essentiels ; ils forment en réalité un cas particulier des tests du χ^2 d'ajustement à une famille de lois présentés, eux, dans les compléments facultatifs.

3.1. Deux exemples pour nous motiver. On va considérer des couples de variables qualitatives.

Exercice 10.2 (Merci HEC Sondages). Cette (défunte ?) association avait organisé, début 2007, un sondage pour les élections présidentielles. J'en ai récupéré les résultats sur leur site et les reproduis en figure 55. Cela étant, ceux qui ont traité les données n'ont indiqué que les proportions observées, et pas du tout les effectifs des différents échantillons. Ceci est mal, très mal (comment calculer des intervalles de confiance ou réaliser des tests sans indications de taille d'échantillons ?), et les auteurs de ces tableaux récapitulatifs de sondage mériteraient de repasser l'examen de statistiques.

Aussi, j'ai essentiellement supposé qu'un vrai travail de bénédictin avait été effectué en interrogeant environ 100 étudiants de première, deuxième et troisième année, et que le tableau obtenu avant calcul des pourcentages était :

		1A	2A	3A	Total
(1)	Royal	9	8	6	23
(2)	Sarkozy	38	36	76	150
(3)	Autre	14	22	0	36
(4)	Indécis	29	26	14	69
(5)	NSPP	3	5	0	8
Total		93	97	96	286

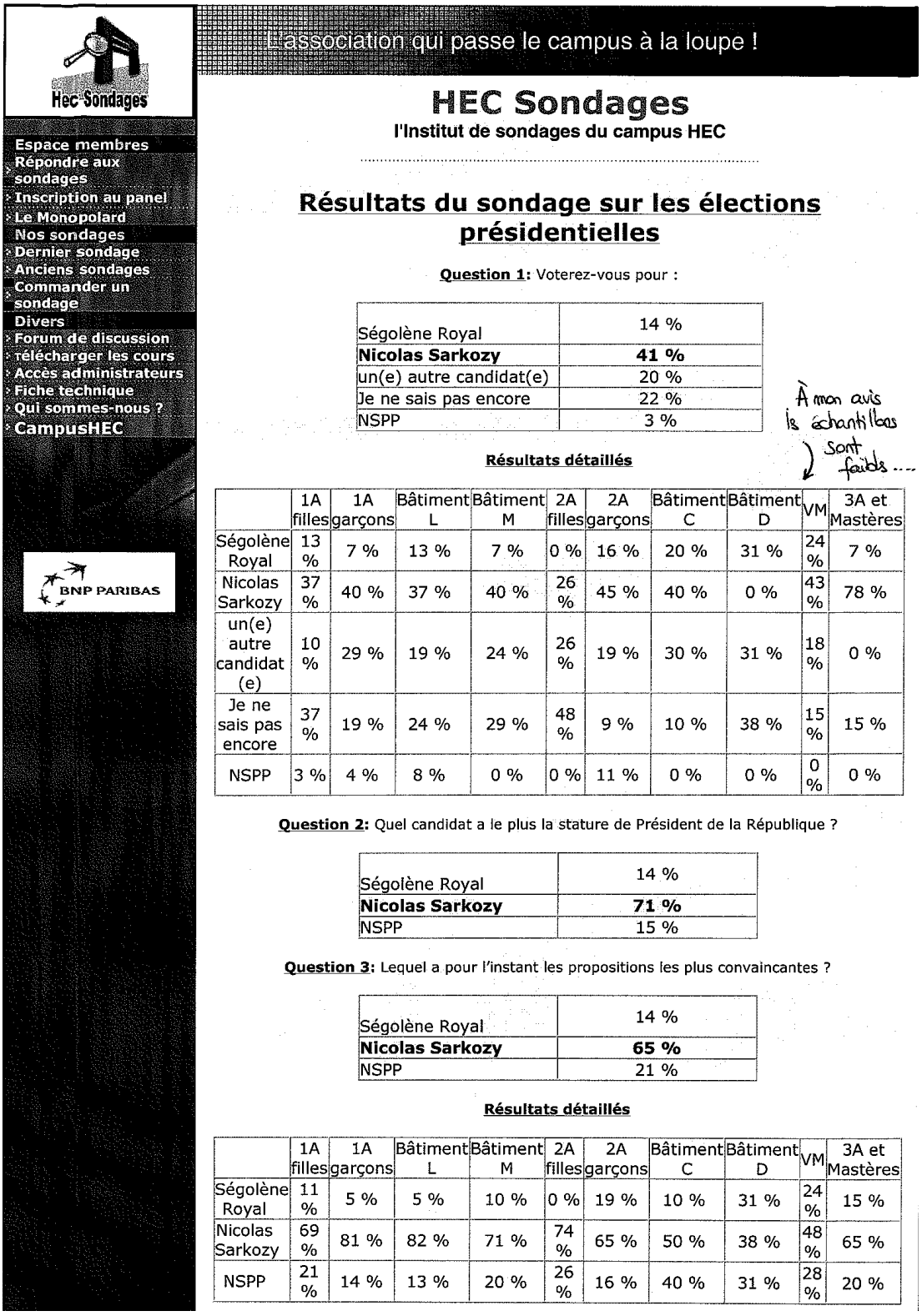
Question : Les opinions politiques dépendent-elles de l'année de scolarité ?

Exercice 10.3 (Prof Grincheux contre Prof Gentil). En 2007 et 2008, une collègue et moi-même utilisions un test pour vérifier [28] que nous notions bien de la même manière dans nos groupes, qu'aucun n'était ni plus sévère ni plus laxiste que l'autre. Je ne retrouve plus nos notes et ai la flemme de les demander à Béatrice Poivre, alors considérons les notes fictives suivantes :

Notes	A	B	C	D	E	F	Total
Prof Grincheux	14	15	26	18	17	5	95
Prof Gentil	21	18	24	19	15	2	99
Total	35	33	50	37	32	7	194

Quel est votre verdict : les réputations sont-elles méritées ou usurpées ?

28. Remarquez bien que l'administration d'HEC effectuait également ce genre de vérifications, lorsqu'il y avait des réclamations ; ce ne devrait plus être le cas désormais à cause de la récente uniformisation des échelles de notations, expliquée page iv.



Hec Sondages

L'association qui passe le campus à la loupe !

Espace membres
Répondre aux sondages
Inscription au panel
Le Monopolard
Nos sondages
Dernier sondage
Anciens sondages
Commander un sondage
Divers
Forum de discussion
télécharger les cours
Accès administrateurs
Fiche technique
Qui sommes-nous ?
CampusHEC

BNP PARIBAS

HEC Sondages

l'Institut de sondages du campus HEC

Résultats du sondage sur les élections présidentielles

Question 1: Voterez-vous pour :

Ségolène Royal	14 %
Nicolas Sarkozy	**41 %**
un(e) autre candidat(e)	20 %
Je ne sais pas encore	22 %
NSPP	3 %

À mon avis les échantillons sont faibles ...

Résultats détaillés

	1A filles	1A garçons	Bâtiment L	Bâtiment M	2A filles	2A garçons	Bâtiment C	Bâtiment D	VM	3A et Mastères
Ségolène Royal	13 %	7 %	13 %	7 %	0 %	16 %	20 %	31 %	24 %	7 %
Nicolas Sarkozy	37 %	40 %	37 %	40 %	26 %	45 %	40 %	0 %	43 %	78 %
un(e) autre candidat (e)	10 %	29 %	19 %	24 %	26 %	19 %	30 %	31 %	18 %	0 %
Je ne sais pas encore	37 %	19 %	24 %	29 %	48 %	9 %	10 %	38 %	15 %	15 %
NSPP	3 %	4 %	8 %	0 %	0 %	11 %	0 %	0 %	0 %	0 %

Question 2: Quel candidat a le plus la stature de Président de la République ?

Ségolène Royal	14 %
Nicolas Sarkozy	**71 %**
NSPP	15 %

Question 3: Lequel a pour l'instant les propositions les plus convaincantes ?

Ségolène Royal	14 %
Nicolas Sarkozy	**65 %**
NSPP	21 %

Résultats détaillés

	1A filles	1A garçons	Bâtiment L	Bâtiment M	2A filles	2A garçons	Bâtiment C	Bâtiment D	VM	3A et Mastères
Ségolène Royal	11 %	5 %	5 %	10 %	0 %	19 %	10 %	31 %	24 %	15 %
Nicolas Sarkozy	69 %	81 %	82 %	71 %	74 %	65 %	50 %	38 %	48 %	65 %
NSPP	21 %	14 %	13 %	20 %	26 %	16 %	40 %	31 %	28 %	20 %

FIGURE 55. Données de sondage telles que présentées sur le site de HEC Sondages (sondage effectué début 2007).

3.2. Principe. Les exemples précédents montrent que l'on part d'une situation avec des couples de données $(x_1, y_1), \ldots, (x_n, y_n)$. Supposons qu'on puisse les modéliser comme

la réalisation de couples de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$, ces couples étant indépendants et identiquement distribués ; ce fait sera évidemment à justifier en préliminaire de tout traitement de données.

On note respectivement $\mathcal{X}$ et $\mathcal{Y}$ les ensembles finis de modalités que peuvent prendre les x_j et les y_j. La loi commune des couples (X_j, Y_j) est alors une certaine loi $\mathbf{p}$ sur l'ensemble fini $\mathcal{X} \times \mathcal{Y}$. La question est de savoir si les X_j sont indépendantes des Y_j, i.e., si $\mathbf{p}$ est une loi-produit (si $\mathbf{p}$ est égale au produit de ses marginales).

A cet effet, on note $r = \operatorname{Card} \mathcal{X}$ et $s = \operatorname{Card} \mathcal{Y}$, et sans perte de généralité, on effectue les identifications $\mathcal{X} = \{1, \ldots, r\}$ et $\mathcal{Y} = \{1, \ldots, s\}$. Enfin, on définit les marginales [29] $\mathbf{p}_{\mathcal{X}}$ et $\mathbf{p}_{\mathcal{Y}}$ de $\mathbf{p}$ et on note

$$\mathbf{p}(x, y) \ , \qquad \mathbf{p}_{\mathcal{X}}(x) \ , \qquad \mathbf{p}_{\mathcal{Y}}(y)$$

les probabilités associées à des éléments $x \in \mathcal{X}$ et $y \in \mathcal{Y}$.

Les hypothèses à tester sont les suivantes. H_0 est l'hypothèse que les couples sont formés de variables indépendantes, i.e., que les X_j sont indépendantes des Y_j. L'hypothèse alternative H_1 est simplement la négation de H_0. H_0 se traduit mathématiquement par le fait que pour tous $x \in \mathcal{X}$ et $y \in \mathcal{Y}$,

$$\mathbf{p}(x, y) = \mathbf{p}_{\mathcal{X}}(x)\, \mathbf{p}_{\mathcal{Y}}(y) \ .$$

On note de la manière suivante, pour tout $x \in \mathcal{X}$ et $y \in \mathcal{Y}$, les effectifs observés :

$$N_{x,y} = \operatorname{Card}\{j : X_j = x \text{ et } Y_j = y\} \ , \quad N_{x,\cdot} = \operatorname{Card}\{j : X_j = x\} \ , \quad N_{\cdot,y} = \operatorname{Card}\{j : Y_j = y\} \ .$$

(On ne note plus explicitement les dépendances en n, pour alléger l'écriture.) Par loi des grands nombres, sous H_0 comme sous H_1, on a alors les estimations suivantes des marginales $\mathbf{p}_{\mathcal{X}}$ et $\mathbf{p}_{\mathcal{Y}}$:

$$\widehat{\mathbf{p}}_{\mathcal{X}} = \left(\frac{N_{1,\cdot}}{n}, \ldots, \frac{N_{r,\cdot}}{n} \right) \qquad \text{et} \qquad \widehat{\mathbf{p}}_{\mathcal{Y}} = \left(\frac{N_{\cdot,1}}{n}, \ldots, \frac{N_{\cdot,s}}{n} \right) \ .$$

Encore par loi des grands nombres, les valeurs observées

$$\frac{N_{x,y}}{n} \qquad \text{et} \qquad \widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y)$$

sont toujours (i.e., sous H_0 comme sous H_1) respectivement proches des probabilités $\mathbf{p}(x, y)$ et $\mathbf{p}_{\mathcal{X}}(x)\, \mathbf{p}_{\mathcal{Y}}(y)$.

Une manière de vérifier si H_0 est vraie est donc de voir si les $N_{x,y}/n$ sont proches ou non des $\widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y)$. C'est ce que fera notre test, puisqu'il considère la statistique

$$D_n^{\text{indep}} \overset{\text{not.}}{=} n \sum_{x=1}^{r} \sum_{y=1}^{s} \frac{\left(N_{x,y}/n - \widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y) \right)^2}{\widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y)} = \sum_{x=1}^{r} \sum_{y=1}^{s} \frac{\left(N_{x,y} - n\, \widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y) \right)^2}{n\, \widehat{\mathbf{p}}_{\mathcal{X}}(x)\, \widehat{\mathbf{p}}_{\mathcal{Y}}(y)} \ .$$

On l'utilise selon le principe suivant.

29. Je vous renvoie à vos cours de prépa pour la définition d'une marginale : c'est la loi induite sur une composante.

PRINCIPE 10.2. *Test d'indépendance de couples de données*
Données : couples $(x_1, y_1), \ldots, (x_n, y_n)$ prenant leurs valeurs dans un ensemble-produit $\{1, \ldots, r\} \times \{1, \ldots, s\}$
Modélisation associée : couples d'observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendants et identiquement distribués selon une certaine loi $\mathbf{p}$ sur l'ensemble-produit $\{1, \ldots, r\} \times \{1, \ldots, s\}$
Hypothèse H_0 : les X_j sont indépendantes des Y_j, i.e., $\mathbf{p}$ est une loi égale au produit de ses marginales
Statistique de test :

$$D_n^{\text{indep}} \overset{\text{not.}}{=} \sum_{x=1}^{r} \sum_{y=1}^{s} \frac{\left(N_{x,y} - n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)\right)^2}{n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)}$$

Comportement sous H_0 : $D_n^{\text{indep}} \rightharpoonup \chi^2_{(r-1)(s-1)}$
Comportement sous H_1 : lorsqu'il n'y a pas indépendance, i.e., que $\mathbf{p}$ n'est pas égale au produit de ses marginales, D_n^{indep} tend vers $+\infty$ et prend donc des valeurs beaucoup plus grandes que sous H_0.

Ici encore, la zone de rejet associée au test est toujours unilatère ; elle vaut, pour une erreur de première espèce α fixée : $\left]\, c_{(r-1)(s-1),1-\alpha},\, +\infty \right[$, où $c_{(r-1)(s-1),1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi $\chi^2_{(r-1)(s-1)}$.

3.3. Application du principe à l'exemple de HEC Sondages. On rappelle que la méthode est, en gros, de comparer le produit des lignes et des colonnes (renormalisé par le nombre total d'observations) aux observations effectivement obtenues dans chaque case. Cela conduit au tableau du haut de la figure 56 : il s'agit de comparer l'effectif observé 9 à celui attendu, $23 \times 93/286 = 7.5$, etc. (pour chacune des 15 cases). Voici maintenant les détails...

CORRECTION 10.2. La modélisation est la suivante. La population est l'ensemble des étudiants de la grande école. En ce qui concerne les données, pour chaque sondé j, on a noté son opinion politique x_j ainsi que son année de scolarité y_j. Ici, j varie entre 1 et 286. Les étendues respectives de ces deux variables sont $\mathcal{X} = \{1, 2, 3, 4, 5\}$ pour les x_j et $\mathcal{Y} = \{1, 2, 3\}$ pour les y_j, selon la table de correspondance indiquée à gauche du tableau de l'énoncé de l'exercice 10.2.

Le choix des sondés ayant été, espérons-le, effectué au hasard, on peut modéliser ces données comme les réalisations de couples de variables aléatoires $(X_1, Y_1), \ldots, (X_{286}, Y_{286})$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p}$ sur l'ensemble-produit $\mathcal{X} \times \mathcal{Y} = \{1, 2, 3\} \times \{1, 2, 3, 4, 5\}$.

Se demander si les opinions dépendent de l'année de scolarité revient à se demander si H_0 : la loi $\mathbf{p}$ est une loi-produit, i.e., si on peut l'écrire comme produit de ses marginales $\mathbf{p}_{\mathcal{X}}$ et $\mathbf{p}_{\mathcal{Y}}$. On estime tout d'abord la loi-produit associée ; elle est obtenue par produit des estimées des marginales

$$\left(\frac{23}{286}, \frac{150}{286}, \frac{36}{286}, \frac{69}{286}, \frac{8}{286}\right) \qquad \text{et} \qquad \left(\frac{93}{286}, \frac{97}{286}, \frac{96}{286}\right) ;$$

ainsi, par exemple, on a un effectif attendu de

$$286 \times \frac{150}{286} \times \frac{93}{286} = 48.8$$

étudiants de première année votant pour Sarkozy en cas d'indépendance ; c'est la réalisation de $286\,\widehat{p}_{\mathcal{X}}(2)\,\widehat{p}_{\mathcal{Y}}(1)$, à comparer à la valeur réalisée de $N_{2,1}$, qui est 38.

Il faut en fait faire cela pour chacune des quinze cases du tableau et on peut faire appel à SPSS à cet effet, afin d'obtenir la figure 56 (où l'on retrouve bien par exemple les valeurs 48.8 et 38 dans la première case de la deuxième ligne).

La minute SPPS 10.3. La figure 56 a été obtenue par `Analyse / Statistiques descriptives / Tableaux croisés`. Puis, dans la fenêtre qui apparaît, on clique sur `Statistiques` pour sélectionner `Chi-deux`, puis on clique sur `Cellules` pour cocher dans le groupe `Effectifs` les valeurs `Observé` et `Attendu`.

Un problème apparaît cependant : les conditions de la remarque 10.1 ne sont pas remplies, et SPSS le signale dans sa note. On lit en effet dans la table des valeurs attendues (effectifs théoriques) plus petites que 5, il faut donc procéder à un regroupement.

On regroupe par exemple ceux qui ne se prononcent pas avec les indécis et on refait le calcul. On obtient cette fois la figure 57. L'asymptotique étant désormais respectée, on peut en tirer des conclusions. On calcule combien vaut la réalisation de la statistique D_{286}^{indep} sur les données. A cet effet, on calcule la somme de douze termes :

$$\frac{(9-7.5)^2}{7.5} + \frac{(38-48.8)^2}{48.8} + \ldots + \frac{(14-25.8)^2}{25.8} \approx 49.157\ .$$

La valeur réalisée de D_{286}^{indep} est supérieure à 49. Or sa loi sous H_0 est approximativement une loi du χ^2 avec $(r-1)(s-1) = 6$ degrés de liberté (on a $r = 3$ mais $s = 4$ à cause du regroupement) et la forme de la zone de rejet est $\left]\, c_{6,1-\alpha},\, +\infty \right[$.

On remarquera que SPSS reprend la valeur réalisée de D_{286}^{indep}, le nombre de degrés de liberté de sa loi limite, ainsi que la P-valeur associée, dans le second tableau de la figure 57.

La P-valeur est très faible, comme l'indique SPSS, quasi-nulle ; un calcul avec un logiciel de statistiques plus efficace montre que

$$\mathbb{P}\{X \geqslant 49\} \leqslant 10^{-8} \qquad \text{où} \quad X \sim \chi^2_6\ .$$

C'est un rejet très clair de l'hypothèse d'indépendance du projet de vote en fonction de l'année (première partie d'une conclusion statistique).

Juste pour avoir le cœur totalement net et pouvoir avancer une explication, regardons s'il y a indépendance entre le vote et le fait d'être 1A ou 2A (i.e., on teste l'homogénéité entre les votes des 1A et 2A). Une démarche similaire à la précédente amène aux résultats obtenus à la figure 58.

La minute SPPS 10.4. La seule différence se trouve du point de vue de l'utilisation de SPSS. On neutralise ici les 3A par un filtrage de données, réalisé avec `Données / Sélectionner des observations`.

Cette fois-ci, la P-valeur est de 61 %, on peut donc dire qu'il y a homogénéité des votes entre 1A et 2A (seconde partie d'une conclusion statistique).

C'est donc l'année de césure qui modifie en profondeur le sentiment politique. Or, si l'on regarde le tableau de données, on voit que ce qui distingue les 3A des 1A et 2A est

Tableau croisé Vote * Année

			Année			
			1A	2A	3A	Total
Vote	Royal	Effectif	9	8	6	23
		Effectif théorique	7,5	7,8	7,7	23,0
	Sarkozy	Effectif	38	36	76	150
		Effectif théorique	48,8	50,9	50,3	150,0
	Autre candidat	Effectif	14	22	0	36
		Effectif théorique	11,7	12,2	12,1	36,0
	Indécis	Effectif	29	26	14	69
		Effectif théorique	22,4	23,4	23,2	69,0
	NSPP	Effectif	3	5	0	8
		Effectif théorique	2,6	2,7	2,7	8,0
	Total	Effectif	93	97	96	286
		Effectif théorique	93,0	97,0	96,0	286,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	51,383[a]	8	,000
...	...	...	...
...	...	...	...
Nombre d'observations valides	286		

a. 3 cellules (20,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 2,60.

FIGURE 56. Opinions politiques et année de scolarité : résultat du test du χ^2 d'indépendance avant regroupement de classes. Note : on a remplacé par ... les valeurs de deux lignes à ne pas regarder, car elles ne sont pas décrites dans ce cours.

Tableau croisé Vote (après regroupement) * Année

			Année			
			1A	2A	3A	Total
Vote (après regroupement)	Royal	Effectif	9	8	6	23
		Effectif théorique	7,5	7,8	7,7	23,0
	Sarkozy	Effectif	38	36	76	150
		Effectif théorique	48,8	50,9	50,3	150,0
	Autre	Effectif	14	22	0	36
		Effectif théorique	11,7	12,2	12,1	36,0
	Indécis ou NSPP	Effectif	32	31	14	77
		Effectif théorique	25,0	26,1	25,8	77,0
	Total	Effectif	93	97	96	286
		Effectif théorique	93,0	97,0	96,0	286,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	49,157[a]	6	,000
Nombre d'observations valides	286		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5.
L'effectif théorique minimum est de 7,48.

FIGURE 57. Opinions politiques et année de scolarité : résultat du test du χ^2 d'indépendance après regroupement de classes.

Tableau croisé Vote (après regroupement) * Année

			Année		
			1A	2A	Total
Vote (après regroupement)	Royal	Effectif	9	8	17
		Effectif théorique	8,3	8,7	17,0
	Sarkozy	Effectif	38	36	74
		Effectif théorique	36,2	37,8	74,0
	Autre	Effectif	14	22	36
		Effectif théorique	17,6	18,4	36,0
	Indécis ou NSPP	Effectif	32	31	63
		Effectif théorique	30,8	32,2	63,0
	Total	Effectif	93	97	190
		Effectif théorique	93,0	97,0	190,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	1,823[a]	3	,610
Nombre d'observations valides	190		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 8,32.

FIGURE 58. Résultat du test du χ^2 d'homogénéité entre les votes des étudiants de première et deuxième années.

la cristallisation [30] des votes des indécis et de ceux en faveur d'autres candidats sur Sarkozy.

Conclusion stratégique ? Ma foi, cela dépend de votre bord politique : si vous êtes de droite, pourquoi ne pas militer pour des stages en entreprise le plus tôt possible, afin que les étudiants commencent à délaisser plus précocement leur tendance naturelle à voter à gauche pour celle des cadres supérieurs, qui est de voter à droite. Si vous êtes de gauche, il s'agira d'informer vos camarades sur ce changement de mentalités insidieux qui risquera de se provoquer en eux lors de leur stage long : dites-leur de ne pas oublier de lire *Libération* et recadrez régulièrement leur univers de gros salaires en leur rappelant la misère du monde.

REMARQUE 10.2. Il faut noter que toute cette étude ne vaut que parce qu'on a supposé un nombre raisonnable de sondés par année. La conclusion serait sans doute différente si les proportions annoncées par HEC Sondages ne portaient que sur le sondage de 10 étudiants 3A... Il faut, encore une fois, absolument préciser les tailles d'échantillons et donner des tables de contingence avec des effectifs (et non des proportions), comme celle que nous avons été constraints d'inventer à l'exemple 10.2 à partir des résultats proposés en pourcentages.

3.4. Application du principe à l'exemple du test d'indépendance (d'homogénéité) des notations.

CORRECTION 10.3. Ici, l'indépendance entre les notations est équivalente à l'homogénéité des notations : peut-on dire que les deux séries de notes proviennent de la même distribution, indépendamment du professeur qui a noté ?

30. Cette agrégation des voix autour de Sarkozy aurait d'ailleurs pu être traitée par un simple test unilatère de proportions.

Pour chaque étudiant, on note (x_j, y_j) le couple formé par son enseignant $x_j \in \mathcal{X}$, où $\mathcal{X} = \{1, 2\}$ (avec la convention que 1 est le grincheux et 2 le gentil) et sa note x_j, qui est dans l'ensemble $\mathcal{Y} = \{A, B, C, D, E, F\}$, identifié à $\mathcal{Y} = \{1, 2, 3, 4, 5, 6\}$.

Chaque enseignant a une certaine manière de noter et on suppose que les copies disponibles cette année-là sont représentatives des copies habituellement écrites par les étudiants. En négligeant l'effet subjectif de notation (qui est de noter plus sèchement après une bonne copie et plus gentiment après une mauvaise copie), il vient que les notes Y_j sont indépendantes et de loi ne dépendant que de l'enseignant. Quitte à considérer l'affectation dans un groupe de cours ou l'autre comme aléatoire (i.e., quitte à regarder les x_j comme réalisations d'une procédure d'affectation aléatoire dont l'administration d'HEC a le secret), nous sommes donc parvenus à la modélisation des données comme la réalisation de couples de variables aléatoires (X_1, Y_1), (X_2, Y_2), ..., (X_{194}, Y_{194}), indépendants et identiquement distribués.

On se demande maintenant si H_0 : la note Y_j est indépendante de l'enseignant X_j. Le même principe que précédemment s'applique :

1. estimation de la loi jointe ;
2. calcul de la table des effectifs attendus et observés et de la valeur réalisée de D_{194}^{indep} à partir de cette table ;
3. détermination de la P-valeur, conclusions statistique et stratégique.

Ici encore, on remarque après le premier traitement (voir les deux premiers tableaux de la figure 59) qu'il faut effectuer un regroupement ; on regroupe les notes E et F sous la même bannière, notée G.

On détaille un peu les calculs pour ce regroupement. Les résultats sont présentés dans les deux tableaux du bas de la figure 59. L'effectif attendu pour le gentil professeur et la note C est déterminé comme la réalisation de $194\,\widehat{p}_{\mathcal{X}}(2)\,\widehat{p}_{\mathcal{Y}}(3)$, soit

$$194 \times \frac{99}{194} \times \frac{50}{194} \approx 25.5$$

et on doit la comparer à la valeur réalisée de $N_{2,3}$, soit 24. On fait de même pour les dix cases de la table et on calcule ensuite la valeur réalisée de D_{194}^{indep}, par sommation sur dix éléments :

$$\frac{(14 - 17.1)^2}{17.1} + \ldots + \frac{(17 - 19.9)^2}{19.9} \approx 2.339 \ .$$

La loi de D_{194}^{indep} sous H_0 étant ici approximativement une loi du χ^2 à 4 degrés de liberté ($r = 2$ et $s = 5$ après regroupement, d'où $(r-1)(s-1) = 4$), cela conduit à la P-valeur (lue dans la sortie SPSS, comme l'ensemble des nombres précédents)

$$\mathbb{P}\{X > 2.339\} = 67.4\,\% \qquad \text{où} \quad X \sim \chi^2_4 \ .$$

Conclusion statistique : On conserve donc l'hypothèse H_0 d'indépendance, sans aucune hésitation, et on en déduit que les réputations sont usurpées, ou qu'à tout le moins, les données dont nous disposons, ne nous permettent pas d'affirmer que ces réputations sont méritées.

Conclusion stratégique : Elle dépend de l'acteur !

– Le professeur catalogué comme grincheux devrait mettre en œuvre le test précédent devant ses étudiants afin de les convaincre, qu'à l'instar de Calimero, la vie et son entourage sont injustes avec lui... et lancer une campagne de communication pour diffuser la nouvelle !

Avant regroupement

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	3,109[a]	5	,683
Nombre d'observations valides	194		

a. 2 cellules (16,7%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 3,43.

Tableau croisé Professeur * Note

			Note						
			A	B	C	D	E	F	Total
Professeur	Grincheux	Effectif	14	15	26	18	17	5	95
		Effectif théorique	17,1	16,2	24,5	18,1	15,7	3,4	95,0
	Gentil	Effectif	21	18	24	19	15	2	99
		Effectif théorique	17,9	16,8	25,5	18,9	16,3	3,6	99,0
	Total	Effectif	35	33	50	37	32	7	194
		Effectif théorique	35,0	33,0	50,0	37,0	32,0	7,0	194,0

Après regroupement

Tableau croisé Professeur * Note (après regroupement)

			Note (après regroupement)					
			A	B	C	D	G	Total
Professeur	Grincheux	Effectif	14	15	26	18	22	95
		Effectif théorique	17,1	16,2	24,5	18,1	19,1	95,0
	Gentil	Effectif	21	18	24	19	17	99
		Effectif théorique	17,9	16,8	25,5	18,9	19,9	99,0
	Total	Effectif	35	33	50	37	39	194
		Effectif théorique	35,0	33,0	50,0	37,0	39,0	194,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,339[a]	4	,674
Nombre d'observations valides	194		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 16,16.

FIGURE 59. Résultats de tests du χ^2 d'homogénéité entre les notes des deux groupes d'étudiants (avant et après regroupement des notes-lettres E et F dans une catégorie G).

- L'administration, si elle reçoit des plaintes d'étudiants, les renverra dans leurs cordes et les cataloguera comme mauvais coucheurs.
- Les étudiants ont intérêt à se taire face à ces données, ils ne pourraient que se décrédibiliser ; il leur faudra patienter et trouver une nouvelle situation où mettre en difficulté, et avec raison cette fois-ci, l'administration et/ou un professeur.

REMARQUE 10.3 (Au risque de me répéter !). Cet exemple explique que les tests d'homogénéité entre deux distributions sont des cas particuliers des tests d'indépendance ; il suffit de considérer la donnée précisant l'index du groupe comme la réalisation d'une variable aléatoire, même fictive, et de tester l'indépendance de l'autre élément des couples de données par rapport à cette variable fictive. Certains (mauvais) livres, mais pas cet excellent polycopié, présentent séparément un test du χ^2 d'indépendance et un test du χ^2 d'homogénéité, alors qu'il s'agit, en théorie et en pratique, du même test !

4. Rappel : cas de deux classes ou de deux fois deux classes

Les tests du χ^2 sont tellement universels que leur considération risque de vous faire oublier les tests les plus simples. Ce paragraphe veut vous les rappeler.

4.1. Test d'ajustement lorsqu'il n'y a que deux classes ($k = 2$).

EXEMPLE 10.3. On lance 1 000 fois un dé et on obtient 212 fois un 6. Le dé est-il biaisé ?

Ici, on dispose de $k = 2$ classes (lancers égaux à 6 vs. autres valeurs) mais on ne va évidemment pas faire une table du χ^2 pour voir si les nombres d'occurrences observés $(212, 788)$ sont bien compatibles ou pas avec les nombres d'occurrences attendues $(166.7, 833.3)$. On fait un test de comparaison[31] à une proportion de référence, voir le principe 7.3. Le test d'ajustement simple du χ^2 n'est à utiliser que lorsqu'il y a strictement plus de deux classes.

REMARQUE 10.4. Les plus attentifs d'entre vous noteront que la statistique de test D_n considérée par le test du χ^2 dans ce cas est égale au carré de la statistique T_n du principe 7.3. Le carré de la loi limite de T_n (le carré d'une loi $\mathcal{N}(0,1)$) est, sans surprise, la loi asymptotique de D_n, une loi χ^2_1. En fait, la vraie différence ici entre implémenter un test du χ^2 et un test de comparaison d'une proportion à une valeur de référence est que ce dernier peut être unilatère, alors que celui du χ^2 est toujours bilatère.

Lorsque l'on dispose de $k \geqslant 3$ classes, alors un test d'ajustement du χ^2 s'impose.

4.2. Test d'indépendance lorsqu'il n'y a que deux fois deux classes.

De même, lorsque $r = s = 2$, i.e., que la table de contingence contient quatre cases, deux lignes et deux colonnes, on se reportera de préférence au principe de comparaison de deux proportions, le principe 9.1. Sous SPSS, ce cas n'est cependant accessible que par la manipulation habituelle, à savoir l'usage de `Tableaux croisés`.

REMARQUE 10.5. Ce test et un test du χ^2 sont égaux en cas d'hypothèse alternative bilatère, mais on peut énoncer une remarque similaire à la précédente quant à la possibilité d'un test unilatère, uniquement avec le principe 9.1.

Si la table contient au moins trois lignes ou trois colonnes, i.e., que $r \geqslant 3$ ou $s \geqslant 3$, on recourra au test du χ^2 d'indépendance.

4.3. Recette !

Dans les deux cas $k = 2$ ou $r = s = 2$, on a $k-1 = 1$ et $(r-1)(s-1) = 1$, de sorte que la loi limite est une loi χ^2_1. La recette est alors la suivante : si vous voyez une telle loi χ^2_1 apparaître, c'est qu'il ne faut pas considérer un test du χ^2, mais un test plus simple !

31. Au fait, quelle P-valeur trouvez-vous ?

Compléments pour étudiants avancés

5. Utilisation du test du χ^2 pour tester l'adéquation à des lois continues

Dans la version rédigée de cette partie, on a énoncé les résultats pour des lois chargeant un nombre fini de points, correspondant à des variables qualitatives. Cependant, si l'on a affaire à une variable quantitative, on peut évidemment la découper en catégories ordinales et ensuite appliquer le test du χ^2. Il faudra bien prendre garde à considérer comme loi de référence sur les catégories la loi induite par la loi continue de référence.

EXEMPLE 10.4 (Tout à fait théorique : test d'ajustement à une loi normale $\mathcal{N}(0,1)$ via un test du χ^2). On part de données obtenues comme la réalisation de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ et on veut tester que leur loi commune est la loi normale standard $\mathcal{N}(0,1)$. A la partie 9, on a indiqué des méthodes efficaces (tests de Shapiro-Wilk et de Kolmogorov-Smirnov sous correction de Lilliefors) à cet effet.

Ici, uniquement pour le bien de l'illustration mathématique, nous indiquons comment on pourrait faire avec le test du χ^2. Par exemple, on considérerait quatre classes $]-\infty, -1[$, $[-1, 0[$, $[0, 1[$ et $[1, +\infty[$, et on testerait que la répartition des X_j s'effectue entre ces quatre classes selon les proportions indiquées par la masse de probabilité que la loi normale induit sur chacune de ces classes ; par exemple, celle affectée à la première classe serait

$$p_1^{\text{ref}} = \mathbb{P}\{N < -1\} = 16\,\% \qquad \text{où} \quad N \sim \mathcal{N}(0,1)\ .$$

Lorsque les classes sont bien choisies, ce test peut se révéler assez efficace.

LA MINUTE SPPS 10.5. A cause de cette possibilité de tester l'ajustement à n'importe quelle loi, même continue, le test du χ^2 est dit non-paramétrique. C'est pourquoi, dans SPSS, il se trouve dans le menu `Analyse / Tests non paramétriques`.

6. Test du χ^2 d'ajustement à une famille de lois

6.1. Principe et exemple. On considère une famille de lois $\mathcal{F} = \{\mathbf{p}^\theta,\ \theta \in \Theta\}$ sur l'ensemble des modalités $\{1, \ldots, k\}$. On part toujours de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi $\mathbf{p}$ et on veut tester que $\mathbf{p} \in \mathcal{F}$. Pour ce faire, on compare la répartition observée $\widehat{\mathbf{p}}_n$ à celle proposée par un élément de $\mathcal{F}$ bien choisi, $\mathbf{p}^{\widehat{\theta}_n}$, et ce, au travers de la statistique de test

$$D_n\left(\mathbf{p}^{\widehat{\theta}_n}\right) \overset{\text{not.}}{=} n \sum_{j=1}^{k} \frac{\left(\widehat{p}_{j,n} - p_j^{\widehat{\theta}_n}\right)^2}{p_j^{\widehat{\theta}_n}} = \sum_{j=1}^{k} \frac{\left(N_{j,n} - n\, p_j^{\widehat{\theta}_n}\right)^2}{n\, p_j^{\widehat{\theta}_n}}\ .$$

Notez bien la similarité avec le test d'ajustement simple : ici, on prend simplement pour loi de référence une loi de la famille de lois de référence, celle indexée par $\widehat{\theta}_n$, où l'estimateur $\widehat{\theta}_n$ du paramètre θ doit être bon en un sens à préciser pour que la procédure fonctionne.

Avant d'indiquer comment on calcule un tel bon estimateur, on revient sur l'exemple de l'agence immobilière.

EXEMPLE 10.5 (Agence immobilière, suite). Tester que deux fois plus de promesses de vente sont signées pendant les mois de printemps que les autres mois revient à tester l'ajustement de la loi commune $\mathbf{p}$ des V_j à la loi $\mathbf{p}^{\text{ref}} = (1/5, 2/5, 1/5, 1/5)$. Si l'on veut tester que deux fois plus de promesses de vente sont signées au printemps qu'en hiver ou en été (sans contrainte par rapport aux ventes d'automne), cela revient en revanche à tester que $\mathbf{p}$ appartient à l'ensemble

$$\mathcal{F} = \big\{(\theta, 2\theta, \theta, 1-4\theta),\ \theta \in [0, 1/4]\big\} .$$

Il faut utiliser toute l'information disponible pour estimer le meilleur θ. La loi des grands nombres assure que $N_{1,n}/n \to \theta$ en probabilité par exemple, mais c'est vrai pour d'autres proportions empiriques. Le mieux est de prendre les trois premières à la fois (la quatrième se déduit d'elles), pour conserver le maximum d'information disponible. Par loi des grands nombres, à nouveau, on a

$$\frac{N_{1,n}+N_{2,n}+N_{3,n}}{n} \xrightarrow{\mathbb{P}} 4\theta \qquad \text{d'où} \quad \widehat{\theta}_n = \frac{N_{1,n}+N_{2,n}+N_{3,n}}{4n} .$$

Un raisonnement mathématique plus rigoureux permet de retrouver ce résultat. On y fait une brève allusion.

> ELÉMENTS CULTURELS. La vraisemblance d'un n–échantillon $X_1, \dots, X_n$ est la probabilité qu'il avait de se produire. Ici, avec les notations précédentes (et en codant 1 pour l'hiver, 2 pour le printemps, 3 pour l'été et 4 pour l'automne), elle vaut donc
>
> $$V_n(\theta) = \theta^{N_{1,n}}\,(2\theta)^{N_{2,n}}\,\theta^{N_{3,n}}(1-4\theta)^{N_{4,n}} .$$
>
> Un bon estimateur $\widehat{\theta}_n$ est la valeur qui maximise la vraisemblance (la terminologie est éclairante), ou de manière équivalente, le logarithme de la vraisemblance. On utilise que $N_{4,n} = n - (N_{1,n} + N_{2,n} + N_{3,n})$ et on dérive selon θ la fonction $\log V_n$; on annule la dérivée pour trouver les extrema,
>
> $$\frac{N_{1,n}}{\theta} + \frac{N_{2,n}}{\theta} + \frac{N_{3,n}}{\theta} - 4\,\frac{n-(N_{1,n}+N_{2,n}+N_{3,n})}{1-4\theta} = 0 ,$$
>
> soit, comme annoncé intuitivement,
>
> $$\widehat{\theta}_n = \frac{N_{1,n}+N_{2,n}+N_{3,n}}{4n} .$$
>
> Cette valeur est bien un maximum, comme on le vérifie en notant que la dérivée seconde est strictement négative en ce point.

On utilise alors le principe suivant. Il ne diffère du principe 10.1 que dans le nombre de degrés de la loi limite du χ^2 : on tient compte de l'estimation du paramètre θ en retirant, par rapport à sa valeur précédente $k-1$, autant de degrés de liberté que le nombre d de paramètres décrivant les éléments de la famille $\mathcal{F}$, parvenant ainsi à une loi du χ^2 à seulement $k-d-1$ degrés de liberté.

PRINCIPE 10.3. *On part d'une situation modélisée par des observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p}$ sur $\{1, \dots, k\}$. On se demande si $\mathbf{p}$ appartient à une certaine famille $\mathcal{F} = \big\{\mathbf{p}^\theta,\ \theta \in \Theta\big\}$ de lois de référence. On note d le nombre minimal de paramètres décrivant les éléments de*

la famille. Le test est fondé sur le résultat suivant : sous $H_0 : \mathbf{p} \in \mathcal{F}$*, si* $\widehat{\theta}_n$ *est un « bon » estimateur de* θ*, alors*

$$D_n\left(\mathbf{p}^{\widehat{\theta}_n}\right) \rightharpoonup \chi^2_{k-d-1} .$$

Lorsque $\mathbf{p} \notin \mathcal{F}$*, la statistique de test tend vers* $+\infty$ *et prend donc des valeurs beaucoup plus grandes que sous* H_0*.*

Dans l'exemple de l'agence immobilière, on a l'estimée $31/(4 \times 36) = 21.5\,\%$ pour θ et on teste donc l'ajustement à la loi $(0.215, 0.43, 0.215, 0.14)$, sous laquelle les effectifs attendus sont $(7.74, 15.48, 7.74, 5.04)$, à comparer aux effectifs observés $(8, 17, 6, 5)$. La taille d'échantillon étant supérieure à 30 et toutes les classes vérifiant la condition de la remarque 10.1 (avoir des effectifs attendus supérieurs à 5), on peut appliquer validement le principe 10.3. La valeur réalisée de la statistique de test D_{36} est

$$\frac{(8-7.74)^2}{7.74} + \frac{(17-15.48)^2}{15.48} + \frac{(6-7.74)^2}{7.74} + \frac{(5-5.04)^2}{5.04} \approx 0.55 .$$

Le nombre de degrés de la loi du χ^2 est $k - d - 1 = 2$, car ici, $k = 4$ et $d = 1$ (un seul paramètre) ; soit une P-valeur de $76\,\%$ (lorsque l'on lit la table de la loi χ^2_2 que je vous ai fournie, on a simplement l'encadrement de la P-valeur entre $10\,\%$ et $90\,\%$).

On conserve donc H_0 sans hésitation, l'ajustement semble tout à fait raisonnable (conclusion statistique). La conclusion stratégique serait par exemple qu'il faut deux fois plus d'employés au printemps qu'en hiver ou l'été : c'est au printemps, et non en été, qu'il faut embaucher des stagiaires !

6.2. Autre exemple.

Exercice 10.4 (Nombre d'appels téléphoniques reçus). Un étudiant suspicieux, mais patient, note chaque jour, pendant 1 000 jours, soit presque quatre ans, le nombre d'appels téléphoniques qu'il reçoit les jours de semaine. En effet, son enseignant lui avait dit à son arrivée à HEC (voir la partie 2) que la loi de ce nombre d'appels était poissonnienne. Quatre ans se sont écoulés, il est presque l'heure de quitter HEC pour le monde du travail et l'étudiant veut vérifier si ce qu'on lui avait enseigné était bien vrai. Il regroupe ses données $x_1, \ldots, x_{1000}$ dans le tableau suivant. Peut-il les exploiter pour se venger de cet enseignant, qui lui avait mis un F ?

Nombre d'appels	0	1	2	3	4	5	6	7	8	9	10	11	$\geqslant 12$
Nombre de jours	14	70	155	185	205	150	115	65	30	5	1	5	0

Correction 10.4. On modélise les données comme la réalisation des variables aléatoires $X_1, \ldots, X_{1000}$, indépendantes et identiquement distribuées selon une loi commune $\mathbf{p}$ sur les entiers naturels $\mathbb{N}$ (même si, en pratique, le nombre d'appels par jour est borné). Ici, on teste l'adéquation à la famille des lois de Poisson $\mathcal{F} = \left\{\mathcal{P}(\lambda),\ \lambda \in \mathbb{R}_+^*\right\}$, aussi note-t-on plutôt λ que θ le paramètre. Un bon estimateur du vrai paramètre λ_0 (si le modèle est adéquat) est la moyenne empirique, $\widehat{\lambda}_{1000} = \overline{X}_{1000}$, comme on l'a vu dans les compléments

de la partie 2 ; on pourrait lui aussi l'obtenir comme estimateur du maximum de vraisemblance du paramètre dans un modèle poissonnien. Ici, l'estimée correspondante [32] vaut $\overline{x}_{1000} = 4$ (tout pile !), on va donc tester l'adéquation à la loi de Poisson de paramètre 4.

Sur 1 000 jours, on attend $1\,000 \times \mathbb{P}\{P = j\}$ jours avec j appels, où P suit une loi de Poisson de paramètre 4. En substituant les valeurs, on obtient le tableau suivant.

Appels	0	1	2	3	4	5	6	7	8	9	10	11	$\geqslant 12$
Observés	14	70	155	185	205	150	115	65	30	5	1	5	0
Attendus	18.3	73.3	146.5	195.4	195.4	156.3	104.2	59.5	29.8	13.2	5.3	1.9	0.9

Nous ne sommes pas dans les conditions de la remarque 10.1 : nous lisons des effectifs attendus plus petits que 5 dans les deux dernières colonnes, et c'est pourquoi nous allons procéder au regroupement des trois dernières colonnes :

Nombre d'appels	0	1	2	3	4	5	6	7	8	9	$\geqslant 10$
Jours observés	14	70	155	185	205	150	115	65	30	5	6
Jours attendus	18.3	73.3	146.5	195.4	195.4	156.3	104.2	59.5	29.8	13.2	8.1

La valeur réalisée de la statistique du χ^2 vaut ainsi

$$\frac{(14-18.3)^2}{18.3} + \frac{(70-73.3)^2}{73.3} + \ldots + \frac{(5-13.2)^2}{13.2} + \frac{(6-8.1)^2}{8.1} \approx 10.2 \ .$$

Cette valeur est à comparer aux quantiles de la loi χ^2_φ : en effet, après regroupement, il n'y a plus que $k = 11$ classes, et il faut enlever deux degrés de liberté ($d = 1$ pour l'estimation de λ_0 et un autre parce qu'on en enlève toujours un). On a ici une P-valeur de 33.2 % (lorsque l'on lit la table de la loi χ^2_2 que je vous ai fournie, on a simplement l'encadrement de la P-valeur entre 10 % et 90 %).

On conserve donc H_0 sans hésitation, l'ajustement semble tout à fait raisonnable (conclusion statistique). La conclusion stratégique est qu'il vaut mieux que l'étudiant parte dignement en ruminant son F plutôt que de faire un scandale... qui prouverait qu'il a bien mérité son F, puisque les données ne contredisent absolument pas l'hypothèse de répartition poissonnienne !

6.3. Liens avec les tests d'indépendance. Le test d'indépendance est en fait un test d'adéquation de la loi commune $\mathbf{p}$ des couples d'observations à la famille des lois-produits

$$\mathcal{F} = \left\{ \mathbf{q}_\mathcal{X} \otimes \mathbf{q}_\mathcal{Y}, \quad \mathbf{q}_\mathcal{X} \text{ loi sur } \mathcal{X}, \ \mathbf{q}_\mathcal{Y} \text{ loi sur } \mathcal{Y} \right\} \ .$$

Ici, on a défini, pour $\mathbf{q}_\mathcal{X}$ et $\mathbf{q}_\mathcal{Y}$ deux lois données respectivement sur $\mathcal{X}$ et $\mathcal{Y}$, la loi-produit $\mathbf{q} = \mathbf{q}_\mathcal{X} \otimes \mathbf{q}_\mathcal{Y}$ par la probabilité qu'elle met sur tout couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

$$\mathbf{q}(x, y) = \mathbf{q}_\mathcal{X}(x)\, \mathbf{q}_\mathcal{Y}(y) \ .$$

32. On aurait pu calculer la variance d'échantillon, qui forme une autre estimée du paramètre inconnu λ_0 ; on aurait obtenu $s^2_{1000} = 3.74$; d'ailleurs, la proximité de $\overline{x}_{1000}$ et s^2_{1000} est un argument en faveur de la loi de Poisson.

Chaque élément de la famille $\mathcal{F}$ est décrit par $d = (r-1)+(s-1)$ paramètres ($r-1$ pour définir la première marginale et $s-1$ pour la seconde).

Par ailleurs, l'estimation efficace des paramètres correspond ici à l'estimation des marginales par la méthode des moments (les fréquences empiriques). On retrouve ainsi la statistique proposée par le principe du test d'indépendance, le principe 10.2, fourni dans la version rédigée de cette partie.

Enfin, le comportement sous H_0 de cette statistique de test est la convergence en loi vers une loi du χ^2 à $k-d-1$ degrés de liberté, où

- $k = rs$ est le cardinal de $\mathcal{X} \times \mathcal{Y}$ (nombre de modalités possibles pour le couple),
- $d = (r-1)+(s-1)$ est le nombre de paramètres décrivant $\mathcal{F}$.

Cela donne au final une loi limite du χ^2 avec $k-d-1 = rs-r-s+1 = (r-1)(s-1)$ degrés de liberté, ainsi qu'annoncé.

Exercices

Exercices issus du cours

Relisez et méditez les exercices 10.1 à 10.3, dont le corrigé se trouve dans la version rédigée du cours ; l'exercice 10.4 (dans les compléments de cours) est quant à lui totalement facultatif.

Un exercice présentant une autre détection de données un peu manipulées

Exercice 10.5 (Théorie génétique). Pour tester ses lois de brassage génétique, Mendel a croisé des plants de pois lisses et jaunes (tous issus d'une longue lignée de pois tous lisses et jaunes) avec des plants pois ridés et verts. A la première génération, tous les plants de pois sont lisses et jaunes. A la deuxième, on obtient 315 plants de poids lisses et jaunes, 108 lisses et verts, 101 jaunes et ridés, 32 verts et ridés. Le caractère ridé étant récessif par rapport au caractère lisse, de même pour la couleur verte par rapport à la jaune, ses lois prévoyaient (faites appel à vos souvenirs de lycée...) une répartition $(9/16, 3/16, 3/16, 1/16)$. Les résultats obtenus sont-ils en accord avec la théorie ? Montrez que la P-valeur sur les données est de 92.5 %. Commentez ce nombre : n'est-il pas trop beau [33] pour être vrai ?

Exercice issu des annales

Exercice 10.6 (Optimisme et âge). Effectuez la question 3 de l'exercice II de l'examen de rattrapage 2007.

Note : l'exercice 10.1 de la version rédigée du cours, qui portait sur une agence immobilière, était lui aussi issu des annales.

33. La plupart des données fournies par Mendel sont de cet acabit ; on soupçonne qu'il a un peu arrangé ses résultats, et notamment, qu'il n'a réalisé ses expériences que sur quelques dizaines de plants et pas quelques centaines...

Exercice : Théorie génétique.

On dispose des données $x_1, \dots x_{556} \in \{LJ, LV, RJ, RV\}$.
À cause du "hasard génétique", on peut les modéliser comme la réalisation des variables aléatoires $X_1, \dots X_{556}$ iid selon une certaine loi $p = (p_{LJ}, p_{LV}, p_{RJ}, p_{RV})$.

Il s'agit de tester l'adéquation à la loi de référence (prévue par la théorie génétique) $p^{ref} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$.

Soit : $H_0 : p = p^{ref}$ contre $H_1 : p \neq p^{ref}$.

À cet effet, on compare les effectifs observés à ceux attendus :

	LJ	LV	RJ	RV
Effectifs observés	315	108	101	32
attendus	312.75	104.25	104.25	34.75
	$= 556 \times \frac{9}{16}$	$= 556 \times \frac{3}{16}$		$= 556 \times \frac{1}{16}$

Note : les conditions asymptotiques sont remplies.

La valeur réalisée de la statistique de test $D_{556}\left((9/16, 3/16, 3/16, 1/16)\right)$ est :

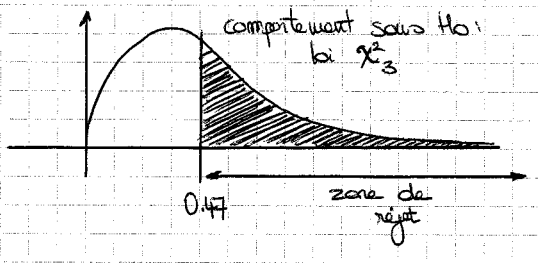
$$\frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} \approx 0.47$$

Or, sous H_0, $D_{556} \overset{(d)}{\simeq} \chi^2_{4-1} = \chi^2_3$

d'où la P-valeur : $p = \mathbb{P}\{X > 0.47\}$ où $X \sim \chi^2_3$

Avec les tables que je vous ai fournies, vous pouvez lire

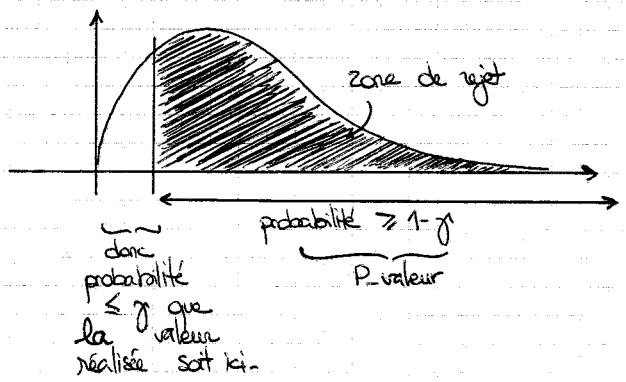
$p = \mathbb{P}\{X > 0.47\} \in [90\%, 95\%]$



Avec un logiciel statistique plus évolué : $p = 92.5\%$

Conclusion statistique : Confirmation éclatante de la théorie développée par Mendel.

Conclusion stratégique : Confirmation trop éclatante ?
Sur toutes ses expériences, Mendel propose de grands P-valeurs, proches de 100%. Or cela est peu probable : même si Ho est vraie, il n'y a qu'une probabilité γ que la P-valeur soit supérieure à $1-\gamma$! Cf le schéma :



Note : ce n'est pas une P-valeur grande qui est suspecte, mais le fait que trop systématiquement, Mendel présentait de grandes P-valeurs.

Exercice : Cf. question 3 de l'exercice II du rattrapage 2007

On commence par modéliser le problème.

On dispose des données (x_j, y_j), $j = 1, \dots 1999$ où :

- $x_j \in \{1,2,3\}$ désigne la catégorie d'âge dans laquelle se trouve le j-ème sondé
- $y_j \in \{1,2,3\}$ désigne son opinion.

On utilise les tables de correspondance :

$x_j =$		$y_j =$	
1 :	20-40	1 :	Optimiste
2 :	40-60	2 :	Pas optimiste
3 :	60 et +	3 :	Sans opinion

Vu l'interrogation au hasard des sondés, on peut modéliser les données comme la réalisation des variables aléatoires (X_j, Y_j) avec $j = 1, \dots 1999$ indépendantes et identiquement distribuées selon une certaine loi $\mathbb{P}$ sur $\{1,2,3\} \times \{1,2,3\} = \{(1,1), (1,2), \dots (3,3)\}$.

On veut tester si H_0 : les X_j sont indépendants des Y_j
ie, si H_0 : $\mathbb{P}$ est une loi - produit.

Pour cela, on compare les effectifs observés :

	Opt.	Pas opt.	Sans op.	Total
20-40	237	392	13	642
40-60	326	298	32	656
60 et +	362	258	81	701
Total	925	948	126	1999

aux effectifs théoriques estimés :

	Opt.	Pas opt.	Sans op.	Total
20-40	297.0	304.5	40.5	642
40-60	303.6	311.1	41.3	656
60 et +	324.4	332.4	44.2	701
Total	925	948	126	1999

Note: les conditions asymptotiques sont vérifiées (toutes les valeurs de cette table sont ≥ 5)

ex: 332.4 a été obtenu comme : $1999 \times \frac{948}{1999} \times \frac{701}{1999}$

ex: 297.0 a été obtenu comme : $1999 \times \frac{925}{1999} \times \frac{642}{1999}$

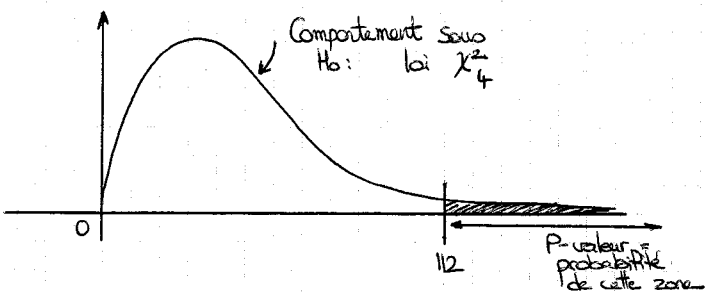
La valeur réalisée de la statistique de test $D^{indép}_{1999}$ est donnée par la somme de neuf termes :

$$\frac{(237 - 297.0)^2}{297.0} + \frac{(392 - 304.5)^2}{304.5} + \cdots + \frac{(81 - 44.2)^2}{44.2} \approx 112$$

Or sous H_0, $D^{indép}_{1999} \overset{(d)}{\approx} \chi^2_{(3-1)(3-1)} = \chi^2_4$ et sous H_1, $D^{indép}_{1999}$ prend de grandes valeurs

soit une zone de rejet de la forme $]r, +\infty[$

et la P-valeur :



$p = \mathbb{P}\{X \geq 112\}$ où $X \sim \chi^2_4$

$\leq \mathbb{P}\{X \geq 18.47\} = 0.1\%$ en lisant les tables fournies

La P-valeur étant très faible, on rejette H_0 et on en conclut (conclusion statistique) que l'optimisme dépend de l'âge, les jeunes semblant plus pessimistes (chômage, prix des logements, etc.) que leurs aînés.

Conclusion stratégique ? S'il s'agit d'un sondage destiné au gouvernement, ce dernier pourra envisager des mesures catégorielles pour rassérénes les jeunes, surtout si c'est un gouvernement de droite à qui les aînés sont toujours acquis !

Onzième Partie

Interlude : quizz sur l'ensemble des parties sur les tests

Enoncé (sujet posé en 2009)

Quizz 4 – Tests d'hypothèses plus complexes – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Cadeaux et comportements d'achat

On considère deux foires aux vins, ayant lieu simultanément dans deux supermarchés similaires d'une même grande agglomération. Dans le premier supermarché, on offre un coffret ouvre-bouteille dès que les achats dépassent 100 euros ; dans le second, aucun cadeau n'était prévu. Sur 130 clients ayant acheté du vin dans le premier, 26 l'ont fait pour plus de 100 euros, tandis que dans le second, ils étaient 15 sur 96.

On résoudra cet exercice selon les questions suivantes et pour une fois, et pour cette fois seulement, on ne modélisera pas le problème (histoire de gagner du temps, tant à la composition, pour vous, qu'à la correction, pour moi).

- Indiquez le nom du test à appliquer :
- Posez les hypothèses de test H_0 et H_1 (avec des mots) :
- Calculez la valeur réalisée de la statistique de test (formule et application numérique) :
- Faites le schéma où on lit les comportements sous H_0 et H_1, la zone de rejet et la P-valeur :
- Indiquez (sans les détails du calcul) la valeur numérique de la P-valeur :
- Conclusion statistique (une phrase) :
- Conclusion stratégique (une phrase) :

Lecture de table (I)

On demande à des hommes quelle est leur marque de chemise préférée entre Nodus, Kenzo et Coton Doux. Les résultats sont reportés dans les tableaux ci-dessous.

Marque de chemise préférée

	Effectif observé	Effectif théorique	Résidu
Nodus	19	14,3	4,7
Kenzo	11	14,3	-3,3
Coton Doux	13	14,3	-1,3
Total	43		

Test

	Marque de chemise préférée
Khi-deux	2,419[a]
ddl	2
Signification asymptotique	,298

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 14,3.

Indiquez les hypothèses testées (avec des mots) :

Formez une conclusion statistique (pas de conclusion stratégique demandée pour une fois) :

Lecture de table (II)

On demande à des consommateurs tests s'ils aiment le goût sucré, s'ils y sont indifférents ou s'ils ne l'aiment pas. On leur demande également de choisir une boisson préférée entre deux boissons au cola. Les résultats sont reportés dans les tableaux ci-dessous.

Tableau croisé boisson préférée * goût pour le sucre

			goût pour le sucre			
			D'accord	Je ne suis pas sûr	Pas d'accord	Total
boisson préférée	rola-cola	Effectif	8	9	7	24
		Effectif théorique	8,4	7,8	7,8	24,0
	koka-cola	Effectif	6	4	6	16
		Effectif théorique	5,6	5,2	5,2	16,0
	Total	Effectif	14	13	13	40
		Effectif théorique	14,0	13,0	13,0	40,0

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	,714[a]	2	,700
Rapport de vraisemblance	,726	2	,695
Association linéaire par linéaire	,024	1	,877
Nombre d'observations valides	40		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 5,20.

Indiquez les hypothèses testées (avec des mots) :

Formez une conclusion statistique (pas de conclusion stratégique demandée pour une fois) :

Corrigé (sujet posé en 2009)

Quizz 4 – Tests d'hypothèses plus complexes – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Gilles Stoltz

Cadeaux et comportements d'achat

[Première version : TEST UNILATÈRE –

On considère deux foires aux vins, ayant lieu simultanément dans deux supermarchés similaires d'une même grande agglomération. Dans le premier supermarché, on offre un coffret ouvre-bouteille dès que les achats dépassent 100 euros ; dans le second, aucun cadeau n'était prévu. Sur 130 clients ayant acheté du vin dans le premier, 26 l'ont fait pour plus de 100 euros, tandis que dans le second, ils étaient 15 sur 96.

On résoudra cet exercice selon les questions suivantes et pour une fois, et pour cette fois seulement, on ne modélisera pas le problème (histoire de gagner du temps, tant à la composition, pour vous, qu'à la correction, pour moi).

– Indiquez le nom du test à appliquer : Test de comparaison de proportions de population p_X et p_Y associées à deux séries de données indépendants.

– Posez les hypothèses de test H_0 et H_1 (avec des mots) :

(cf. préjugé que le consommateur est faible et influençable...)

H_0 : le cadeau est sans effet sur les montants d'achat, i.e., $p_X = p_Y$ (où p_X, p_Y sont les taux d'achat dépassant 100 € dans le premier et le second magasins) vs H_1 : le cadeau incite à dépenser plus, i.e., $p_X > p_Y$

– Calculez la valeur réalisée de la statistique de test (formule et application numérique) :

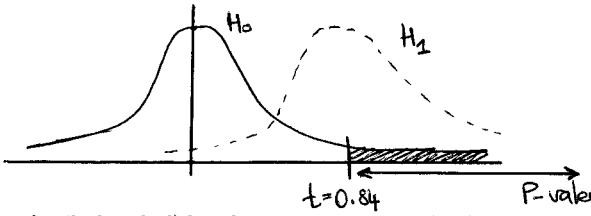
$T_{130,96}$ admet pour réalisation

$$t = \frac{\bar{x}_{130} - \bar{y}_{96}}{\sqrt{\bar{p}_{226}(1-\bar{p}_{226})(1/130 + 1/96)}}$$

(où $\bar{p}_{226} = \frac{26+15}{130+96} \simeq 18.1\%$ est l'estimée de la proportion commune)

$$= \frac{26/130 - 15/96}{\sqrt{0.181\,(1-0.181)\,(1/130+1/96)}} \times 0.84$$

– Faites le schéma où on lit les comportements sous H_0 et H_1, la zone de rejet et la P-valeur :



– Indiquez (sans les détails du calcul) la valeur numérique de la P-valeur : où $N \sim \mathcal{N}(0,1)$

Soit : $\simeq 20.0\%$

– Conclusion statistique (une phrase) :

On conserve H_0 (assez nettement, la P-valeur est grande alors que les échantillons étaient assez grands).

– Conclusion stratégique (une phrase) :

Le cadeau actuel n'augmentant pas le montant des achats :

- on le supprime (c'est de l'argent investi pour rien)
- et/ou on trouve un meilleur cadeau ou une meilleure idée promo.

A Bravo !

Quizz 4 – Tests d'hypothèses plus complexes – 2009

Prénom, nom et indication du groupe théorique (8h ou 10h) :

Une de vos camarades....

Seconde version : TEST BILATÈRE.

Cadeaux et comportements d'achat

On considère deux foires aux vins, ayant lieu simultanément dans deux supermarchés similaires d'une même grande agglomération. Dans le premier supermarché, on offre un coffret ouvre-bouteille dès que les achats dépassent 100 euros ; dans le second, aucun cadeau n'était prévu. Sur 130 clients ayant acheté du vin dans le premier, 26 l'ont fait pour plus de 100 euros, tandis que dans le second, ils étaient 15 sur 96.

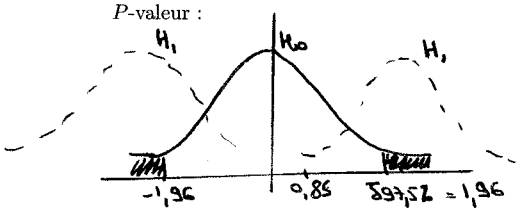
On résoudra cet exercice selon les questions suivantes et pour une fois, et pour cette fois seulement, on ne modélisera pas le problème (histoire de gagner du temps, tant à la composition, pour vous, qu'à la correction, pour moi).

- Indiquez le nom du test à appliquer :
  Test de comparaison de deux proportions sur des échantillons indépendants
- Posez les hypothèses de test H_0 et H_1 (avec des mots) :
  H_0 : Avec ouvre-bouteille et sans, la proportion de gens achetant pour plus de 100€ de vin est la même.
  H_1 : Les proportions sont différentes dans les deux supermarchés
- Calculez la valeur réalisée de la statistique de test (formule et application numérique) :

$$\frac{\bar{X}_{130} - \bar{Y}_{96}}{\sqrt{\hat{P}_{226}(1-\hat{P}_{226})\left(\frac{1}{130}+\frac{1}{96}\right)}} \quad \text{avec} \quad \hat{P}_{226} = \frac{130\bar{X}_{130} + 96\bar{Y}_{96}}{130+96} = 0{,}18$$

$$\text{Valeur réalisée} = \frac{0{,}2 - 0{,}16}{\sqrt{0{,}18(1-0{,}18)\left(\frac{1}{130}+\frac{1}{96}\right)}} \simeq 0{,}85$$

- Faites le schéma où on lit les comportements sous H_0 et H_1, la zone de rejet et la P-valeur :



$$\text{P-valeur} = 2\mathbb{P}(N \geq 0{,}85) = 2\left[1 - \mathbb{P}(N \leq 0{,}85)\right] = 2\left[1 - 0{,}8023\right] = 0{,}40$$

- Indiquez (sans les détails du calcul) la valeur numérique de la P-valeur : 0,40
- Conclusion statistique (une phrase) :
  On conserve forcément H_0
- Conclusion stratégique (une phrase) :
  Il est inutile de s'évertuer à offrir un coffret ouvre-bouteille pour plus de 100€ d'achats de vin, cela n'augmente pas les achats.

Quizz 4 – Tests d'hypothèses plus complexes – 2009

Lecture de table (I)

On demande à des hommes quelle est leur marque de chemise préférée entre Nodus, Kenzo et Coton Doux. Les résultats sont reportés dans les tableaux ci-dessous.

test du χ^2 d'ajustement simple (à la bi $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ici)

Marque de chemise préférée

	Effectif observé	Effectif théorique	Résidu
Nodus	19	14,3	4,7
Kenzo	11	14,3	-3,3
Coton Doux	13	14,3	-1,3
Total	43		

Test

	Marque de chemise préférée
Khi-deux	2,419[a]
ddl	2
Signification asymptotique	,298

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 14,3.

Indiquez les hypothèses testées (avec des mots) :

H_0 : les préférences envers chacune des trois marques sont équitablement distribuées dans la population masculine vs. H_1 : ce n'est pas le cas ie, en moyenne, une des trois marques est préférée aux deux autres.

Formez une conclusion statistique (pas de conclusion stratégique demandée pour une fois) :

P-valeur lue : 29.8%

soit conservation de H_0.

Lecture de table (II)

On demande à des consommateurs tests s'ils aiment le goût sucré, s'ils y sont indifférents ou s'ils ne l'aiment pas. On leur demande également de choisir une boisson préférée entre deux boissons au cola. Les résultats sont reportés dans les tableaux ci-dessous.

test du χ^2 d'indépendance.

Tableau croisé boisson préférée * goût pour le sucre

			goût pour le sucre			
			D'accord	Je ne suis pas sûr	Pas d'accord	Total
boisson préférée	rota-cola	Effectif	8	9	7	24
		Effectif théorique	8,4	7,8	7,8	24,0
	koka-cola	Effectif	6	4	6	16
		Effectif théorique	5,6	5,2	5,2	16,0
	Total	Effectif	14	13	13	40
		Effectif théorique	14,0	13,0	13,0	40,0

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	,714[a]	2	,700
~~Rapport de vraisemblance~~	~~,726~~	~~2~~	~~,695~~
~~Association linéaire par linéaire~~	~~,024~~	~~1~~	~~,877~~
~~Nombre d'observations valides~~	~~40~~		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 5,20.

← on n'a vu ensemble que cette ligne.

Indiquez les hypothèses testées (avec des mots) :

H_0 : indépendance de la boisson préférée et du goût pour le sucre vs. H_1 : le goût pour le sucre détermine au moins en partie la boisson préférée

Formez une conclusion statistique (pas de conclusion stratégique demandée pour une fois) :

P-valeur lue : 70.0%

soit conservation telle de H_0.

Distribution des notes.

Sans surprise, elle est moins bonne que d'habitude (médiane à C):

A	B	C	D	E	Absents, etc.
15	22	22	18	8	9

J'espère que vous vous êtes déjà remis sérieusement au travail !

Remarques et commentaires.

Cadeaux & comportements d'achat.

* Test à appliquer : Cela pourrait éventuellement être un test du χ^2 d'indépendance, mais pas un T-test (suppose des lois normales et non de Bernoulli !); les données correspondaient ici à des échantillons indépendants (et non, appariés).

* Hypothèses : → attention ! il ne s'agissait pas de voir si les proportions sur les deux échantillons étaient différents (elles le sont : $^{26}/_{130} \neq {}^{15}/_{96}$) mais de voir si elles étaient révélatrices de taux moyens (d'achats pour une valeur $\geq 100€$) différents selon qu'un cadeau est offert ou non; il fallait donc prendre garde à la formulation et écrire " H_0 : le cadeau est sans effet ... " (= fait général de psychologie du consommateur) plutôt que " H_0 : le cadeau a été sans effet ... " (= mesci, on sait lire les données réalisées ...).

→ j'ai souvent vu H_0 et H_1 inversées; H_0 est toujours

l'hypothèse d'égalité, par construction du test.

* Statistique de test : → Il ne s'agissait pas de recourir au principe de comparaison d'une proportion à une valeur de référence et de fixer $p_{ref} = 15\%$ p. ex. On n'a aucune raison ici de connaître parfaitement une valeur de référence ; on fait tout à fait autre chose, on compare des taux d'achats significatifs avec et sans un cadeau promo.

→ attention comme toujours de bien distinguer estimateurs et estimés/valeurs réalisées ; je suis tellement las de voir cette distinction toujours non faite que je ne l'ai même plus corrigée.

* Conclusion stratégique : → Il faut, là aussi, regarder vers le futur et vers les actions à entreprendre (ou pas)....

Lecture de table (I).

→ Il ne s'agissait pas d'un test d'indépendance (de qui contre qui ?!), mais d'un test d'ajustement à une loi uniforme. Ce dernier testant l'équirépartition des préférences dans la population masculine.

→ Ici, comme pour l'autre lecture de table et plus généralement, comme pour tout test du χ^2, H_1 est l'exacte négation de H_0. H_1 n'est en particulier pas formulée au vu des données. (Ni ici, ni dans aucun autre test d'ailleurs : H_1 est à formuler a priori, selon le contexte mais sans les données.)

Douzième Partie

Cas de révision pour la modélisation, l'estimation et les tests

Enoncé du cas « Votre Santé »

Source : Jacques Obadia et les autres professeurs du département SIAD d'HEC Paris
... et pas mon cerveau fertile, pour une fois ...

La société Votre Santé

[Note : très ancien exercice, rédigé en francs F !]

La société *Votre Santé* est une entreprise de vente par correspondance de produits de beauté dits « *naturels* ». Elle gère un fichier de 350 000 clients et propose chaque mois une offre promotionnelle accompagnée d'un cadeau. Le taux de réponse à cette offre est généralement de 15%, la marge moyenne par réponse de 340F. Mlle C. Claire, nouvellement en charge de ce fichier, a retenu comme cadeau un abonnement gratuit de six mois, au mensuel « *Votre beauté Madame* ». Elle pense que cela pourrait augmenter le taux de réponse à la prochaine offre ; toutefois cette proposition ne serait rentable que si le taux de réponse dépassait les 17,5% (avec la même marge moyenne évidemment). Elle envisage de tester la réalité de ces hypothèses sur un échantillon de clientes. La précision voulue pour son estimation est de l'ordre de 2%.

Questions

1. Quelle taille d'échantillon doit-elle choisir afin d'atteindre la précision voulue (avec un degré de confiance de 0,95) ?
2. Les résultats d'un sondage sur un échantillon de 1225 clientes vous sont donnés en annexe 1. Donner une estimation par intervalle au degré de confiance 0,95 du pourcentage p_0 de réponses positives attendu à l'offre.
3. Mlle C. Claire se propose de procéder au test d'hypothèse suivant :

$$H_0 : p_0 = 17{,}5\%$$
$$H_1 : p_0 > 17{,}5\%$$

Expliquer pourquoi elle envisage ce test.
Calculer la P-valeur. Qu'en concluez-vous ?

4. Mlle C. Claire pense que les nouveaux clients (inscrits depuis moins de 6 mois) ont un taux de réponse inférieur aux anciens. Confirmer ou infirmer cette hypothèse.
5. Il s'agit dans cette question de déterminer un intervalle de confiance au degré de confiance 0,95 de la marge de la campagne promotionnelle.
   a) Peut-on considérer que la marge moyenne attendue de cette campagne sera la même que pour les campagnes précédentes. On posera cette alternative sous forme de test.
   b) En déduire une estimation par intervalle de la marge totale attendue.
   c) Comment aurait-on fait sans le test effectué en a) ?

Annexe 1 Résultats du sondage

Taille de l'échantillon : 1225 individus

	Total	*Anciens Clients*
Nombre d'individus	1225	850
Nombre de réponses	258	193

Résultats sur la marge

Marge totale	Marge Moyenne	Ecart-type de la marge
85140 F	330 F	165 F

Corrigé du cas « Votre Santé »

Cas « Votre Santé »

0. [A ne pas oublier :] Modélisation :

- Population : les 350 000 clients du fichier
- Echantillon : taille à déterminer, cf. question 1/ ; on la note n

→ On aura alors des données $x_1, \dots x_n \in \{0,1\}$ (x_j : j-ème client achète = 1 ou non = 0) réalisations de $X_1 \dots X_n$ iid $\sim Ber(p_0)$, où p_0 s'interprète comme le taux de réponses positives qu'on aurait si on généralisait l'offre à l'ensemble des clients

1. → Méthode pessimiste (majoration de la variance) : l'IC a priori, de niveau 95%, est $\left[\bar{X}_n \pm z_{97.5\%}\sqrt{\frac{1/4}{n}}\right] \simeq \left[\bar{X}_n \pm 1/\sqrt{n}\right]$ vu que $z_{97.5\%} = 1.96 \simeq 2$. Il suffit de prendre n tel que $1/\sqrt{n} \leq 2\%$, $n = 2500$ convient.

→ Méthode plus optimiste : l'IC a posteriori, de niveau 95%, sera $\left[\bar{X}_n \pm z_{97.5\%}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$ avec sans doute $\bar{X}_n(1-\bar{X}_n)$ proche de son ancienne valeur $15\%(1-15\%)$. La condition sur n est alors $1.96\sqrt{15\%(1-15\%)}/\sqrt{n} \leq 2\%$, elle est vérifiée pour $n \geq 1225$.

2. On obtient (avec la méthode optimiste) les données $x_1 \dots x_{1225}$ avec $\bar{x}_{1225} = \frac{258}{1225} \simeq 21.1\%$

On se doute que p_0 est proche de 21% mais on veut quantifier cela par un IC.

L'IC après sondage (bilatère et de niveau 95%) est $\left[\bar{X}_{1225} \pm z_{97.5\%}\sqrt{\frac{\bar{X}_{1225}(1-\bar{X}_{1225})}{1225}}\right]$

Sa valeur réalisée est $\left[\bar{x}_{1225} \pm 1.96\sqrt{\frac{\bar{x}_{1225}(1-\bar{x}_{1225})}{1225}}\right] = \left[21.1\% \pm 1.96\sqrt{\frac{0.211(1-0.211)}{1225}}\right]$

$\simeq [21.1\% \pm 2.3\%]$

$\subseteq [21.0\% \pm 2.5\%] = [18.5\%, 23.5\%]$

La précision est légèrement moins bonne que 2%... Cf. la méthode optimiste ne garantit pas la précision avec certitude...

On pourrait vouloir plutôt un IC unilatère pour montrer que p_0 est bien supérieur à 17.5% ; au niveau 95%, la formule théorique est :

$$\left[\bar{X}_{1225} - 3{,}95\% \sqrt{\frac{\bar{X}_{1225}(1-\bar{X}_{1225})}{1225}},\ 100\% \right]$$ avec pour valeur réalisée

$$\left[\bar{x}_{1225} - 1.65 \sqrt{\frac{\bar{x}_{1225}(1-\bar{x}_{1225})}{1225}},\ 100\% \right]$$

$$= \dots = [19.1\%,\ 100\%].$$

Conclusion stratégique : [Continuer l'étude pour voir si la marge est la même.]

3. Si elle rejette son H_0 pour H_1, alors elle saura que p_0 est bien $> 17.5\%$ et que la nouvelle promo est à étudier

Conclusion statistique

(On n'apprend réellement d'un test que lorsque l'on rejette H_0.)

Vu la question 2., on se doute qu'on va rejeter H_0. On veut quantifier cela par la P-valeur.

$p_{ref} = 17.5\%$; on utilise la statistique de test

$$T_{1225} = \sqrt{1225}\left(\frac{\bar{X}_{1225} - 0.175}{\sqrt{0.175\,(1-0.175)}} \right)$$

Sous H_0 : $p_0 = 17.5\%$, on a $T_{1225} \overset{(d)}{\simeq} \mathcal{N}(0,1)$

Sous H_1 : $p_0 > 17.5\%$, T_{1225} tend à prendre des valeurs plus grandes, la zone de rejet est donc de la forme $]r_1, +\infty[$.

La valeur réalisée de T_{1225} est : $\sqrt{1225}\left(\bar{x}_{1225} - 0.175\right) \Big/ \sqrt{0.175(1-0.175)}$

$$= \sqrt{1225}\left(\frac{0.211 - 0.175}{\sqrt{0.175(1-0.175)}} \right) \simeq 3.28$$

La P-valeur est $\mathbb{P}\{N > 3.28\}$ où $N \sim \mathcal{N}(0,1)$

$$\leq \mathbb{P}\{N > 3.20\}$$

$$= 1 - 0.999313 \leq 0.07\%,$$ d'après les tables.

Elle est très, très faible, on rejette H_0 sans hésiter (conclusion statistique). On continue donc l'étude, avec la marge. (Conclusion stratégique) ☺

4. Le plus simple est d'appliquer un test de comparaison de proportions entre deux populations (nouveaux vs. anciens clients).

Données d'échantillon : Les anciens clients sont $y_1, \dots y_{850}$, on lit $\bar{y}_{850} = \frac{193}{850} = 22.7\%$
Pour les nouveaux, on a $z_1, \dots z_{375}$, avec 65 (= 258 - 193) réponses, soit $\bar{z}_{375} = 65/375 = 17.3\%$.

Pour être tout à fait clair, on reformule le tableau :

	Anciens	Nouveaux	Total
Ont répondu	193	65	258
N'ont pas rép.	657	310	967
Total	850	375	1225

On modélise, via le tirage au hasard, les données comme réalisations de $Y_1, \dots Y_{850}$ iid $\sim Ber(p_{ancien})$ et $Z_1 \dots Z_{375}$ iid $\sim Ber(p_{nouveau})$ où p_{ancien} et $p_{nouveau}$ seraient les taux de réponse respectifs sur les 2 populations si on généralisait la preuve.

On veut tester H_0 : $p_{ancien} = p_{nouveau}$ vs. H_1 : $p_{anciens} > p_{nouveaux}$ (c'est ce qu'affirme Mlle Claire)

La statistique de test est

$$T_{850,375} = \frac{\bar{Y}_{850} - \bar{Z}_{375}}{\sqrt{\hat{p}_{1225}(1-\hat{p}_{1225})\left(1/850 + 1/375\right)}}$$

où $\hat{p} = \frac{Y_1 + \dots Y_{850} + Z_1 + \dots Z_{375}}{1225}$

Sous H_0, $T_{850,375} \overset{(d)}{\simeq} \mathcal{N}(0,1)$
et sous H_1, $T_{850,375}$ tend à prendre des valeurs plus grandes.

(3

La zone de rejet est donc de la forme $]r_1, +\infty[$.

La valeur réalisée de $T_{850,375}$ est $\dfrac{\bar{y}_{850} - \bar{y}_{375}}{\sqrt{\bar{z}_{1225}(1-\bar{z}_{1225})\left(\frac{1}{850}+\frac{1}{375}\right)}} = \dfrac{0.227 - 0.173}{\sqrt{0.211\,(1-0.211)\left(\frac{1}{850}+\frac{1}{375}\right)}}$

(Avec les notations précédents, la valeur réalisée de $\hat{p}_{1225}$ est $\bar{z}_{1225}$.)

$\simeq 2.13$

La P_valeur est donc $P\{N > 2.13\} = 1 - 0.9834 \leq 1.7\%$

(où $N \sim \mathcal{N}(0,1)$)

On rejette clairement H_0 et on en conclut que les anciens clients sont plus réceptifs (conclusion statistique).

Il faudra sans doute créer une campagne juste pour les nouveaux clients. Le taux de 17.5% n'est pas garanti pour eux (estimée à 17.3% donc IC sur Pnouveau qui contiendrait des valeurs plus petites que 17.5%).

↑ Conclusion stratégique.

Note: On pourrait faire un χ^2 d'indépendance. Cela revient, comme il est écrit dans le poly, à faire une comparaison de proportions bilatère. On trouverait une stat. de test D^{indep}_{1225} avec valeur réalisée $\simeq 4.53$, une loi limite χ^2_1; et une P_valeur égale à 3.4%. (le double de celle du test unilatère).

Je vous laisse faire le calcul détaillé. Il faut arriver à

Valeur réalisée de D^{indep}_{1225}: $\dfrac{(193-179)^2}{179} + \dfrac{(657-671)^2}{671} + \dfrac{(65-79)^2}{79} + \dfrac{(36-296)^2}{296}$

$= 4.53$

soit la P valeur $P\{X > 4.53\} = 3.4\%$ où $X \sim \chi^2_1$

(sur les tables vous liez juste $\in [2.5\%, 5\%]$)

④

5. a. Pour chacune des 258 commandes, on note m_j la marge réalisée : $m_1, \dots m_{258}$ avec $\bar{m}_{258} = 330$, $s_{m,258} = 165$ que l'on modélise par $M_1, \dots M_{258}$ iid selon une certaine loi admettant une espérance μ_0.

μ_0 est la marge moyenne par commande que l'on obtiendrait si on généralisait la promo.

C'est notre paramètre d'intérêt ici. On note $\mu_{ref} = 340$ et on teste :

$$H_0 : \mu_0 = 340 \quad \text{vs} \quad H_1 : \underbrace{\mu_0 < 340}_{\text{alternative pessimiste}}.$$

Note : vu la forme de la promo, la marge n'a aucune raison d'avoir changé ! Ce qui justifie le test.

On considère la statistique de test

$$T_{258} = \frac{\bar{M}_{258} - 340}{\sqrt{\hat{\sigma}^2_{258}/258}}$$

Sous H_0, $T_{258} \overset{(d)}{\simeq} \mathcal{N}(0,1)$

Sous H_1, T_{258} tend à prendre des valeurs plus petits. La zone de rejet est de la forme $]-\infty, r[$ et

T_{258} admet pour valeur réalisée : $\sqrt{258}\left(\frac{\bar{m}_{258} - \mu_{ref}}{s_{m,258}}\right) = \sqrt{258}\left(\frac{330 - 340}{165}\right) = -0.97$

La P-valeur est $P\{N \leq -0.97\} = P\{N \geq 0.97\}$

$= 1 - 0.8340$

$= 16.60\%$

On conserve H_0. On pensera donc dans la suite que μ_0 ne diffère pas significativement de 340 F.

5. b. Cf question 2, on a l'IC $[18.5\%, 23.5\%]$ pour p_0.

Marge attendue : $350\,000 \times \{18.5\%, 23.5\%\} \times 340$

(réalisation d'un IC sur la marge attendue) — # clients — marge moyenne retenue

$= \{ 2\,201\,500 \text{ F} ; 2\,796\,500 \text{ F} \}$ soit $[2{,}2 \text{ MF} ; 2{,}8 \text{ MF}]$

pour arrondir un peu les choses.

5.c. S'il n'y avait pas eu la question 5.a., on aurait calculé un IC sur μ_0 et utilisé la méthode de Bonferroni, comme on en cours, pour déterminer un IC sur $\mu_{0,p_0} \times 350\,000$:

réalisation d'un IC à 95 % sur μ_0 : $\left[330 \pm 1.96 \frac{165}{\sqrt{258}} \right] = [309.8, 350.2]$

réalisation d'un IC à 90% sur μ_{0,p_0} : $[309.8 \times 18.5\%, 350.2 \times 23.5\%]$

$= [57.3 ; 82.3]$

Soit la réalisation d'un IC sur marge totale, à 90 % :

$[350\,000 \times 57.3, 350\,000 \times 82.3]$

$= [2{,}0 \text{ MF} ; 2{,}9 \text{ MF}]$.

↑ un peu moins bon que précédemment (plus large) mais on s'y attendait !

Treizième Partie

Régression linéaire simple

Version rédigée du cours

Résumé : Les chapitres précédents ont présenté les notions d'intervalle de confiance et de test, et en ont donné divers exemples.

Objectif : Nous étudions ici un modèle statistique d'usage fréquent, voire incontournable, à cause de son efficacité : la régression linéaire (simple). Il permet d'expliquer ou de prédire une variable quantitative comme fonction affine d'une autre variable quantitative. (La partie suivante présentera le cas plus général de l'explication ou de la prédiction en fonction de plusieurs variables quantitatives.) En particulier, nous mettrons en œuvre intervalles de confiance et tests sur les paramètres de ce modèle, notamment pour quantifier l'influence de la variable explicative sur la variable à expliquer.

1. Présentation du modèle et de ses objectifs

1.1. Couples de données : variable à expliquer et variable explicative. On considère ici des données se présentant comme des couples de variables $(x_1, y_1), \ldots, (x_n, y_n)$. On imagine que la valeur du premier élément x_j du j–ième couple de données a une influence, en un certain sens, sur celle du second élément y_j et on veut trouver une expression générale de la liaison ainsi formée (i.e., valant pour tous les couples à la fois et même pour de nouveaux couples à venir). On dit en français[34] des y_j qu'elles forment la variable à expliquer et des x_j qu'elles sont la variable explicative.

Remarque 13.1 (Explication statistique vs. explication causale). On ne parle ici évidemment que d'explication statistique : le fait que les variations d'une variable soient liées aux variations d'une autre variable. Cela n'implique pas en général que la variable explicative soit la source ontologique ou la cause profonde de la variation de la variable à expliquer ! On s'en convaincra avec l'exercice 13.4, qui étudie l'existence d'une relation entre maladie mentale et nombre de téléviseurs.

Exemple 13.1 (Prix d'un appartement). A environnement (quartier ou ville) donné, une idée généralement partagée est que la surface d'un appartement détermine assez largement son prix ; le prix de présentation, sans nul doute, mais moins le prix de vente cependant, à cause de la multitude d'autres facteurs objectifs (étage et présence d'un ascenseur, orientation, parking, gardien, année de construction, etc.) et subjectifs (charme) à prendre en compte. On considère la coupure de journal donnée par la figure 60 (datant du début des années 2000, les prix font rêver !). On dispose donc ici de 28 couples (x_j, y_j) et on veut relier le prix y_j d'un appartement (variable à expliquer) en fonction de sa surface x_j (variable explicative). On représente graphiquement ces couples à la figure 61. Au moins pour les surfaces inférieures à 100 mètres carrés, la relation entre prix et surface semble raisonnablement linéaire.

34. La terminologie anglo-saxonne, reprise par SPSS, est celle de variable indépendante et de variable dépendante.

1. CENSIER, bas de R. Mouffetard, pied-à-terre, 28 m², tt confort. Visite vendredi, samedi, dim. 130 000 € à discuter. Facilités	2. CONTRESCARPE, imm. Ancien, pierre de taille, beau duplex caractère, 50 m², poutres, refait neuf, 280 000 €
3. R. St-Simon, en pleine verdure, calme, plein soleil, Superbe appt 4p., 106 m², cuis. aménagée, s. de bains moderne, chff. cent. Parfait état. Px 650 000 à discuter. Agence s'abstenir. Direct. Propriétaire.	4. RAPP 7P., 196 m² standing, 9 fenêtres plein soleil, 800 000 €
5. R. St André-des-Arts, beau liv + chbre, imm. XVIIIe siècle, 55 m², 268 000 €	6. 5e PRES QUAIS, 7 pces, 190 m² caractère, standing, 790 000 €
7. GOBELINS, Beau 5p., 110m², gd cft, soleil, 500 000 €	8. GOBELINS, et. élevé, calme, asc., 2 pièces, 60 m², 320 000 €
9. CENSIER, très grand studio + entrée 48 m², tt cft, ensoleillé, calme, bel imm., 250 000 €	10. PANTHÉON, 7e étage, ascenseur, grand studio 35 m² + terrasse. Vue. 250 000 €
11. RUE MADAME, 3P. + Serv., 86 m², 350 000 €	12. RUE DE SEINE, 3P., tt cft, 65 m², calme, soleil, 300 000 €
13. PANTHEON, bel imm., verdure, magnifique studio 32 m², caractère, 155 000 €	14. SEVRES BAB, 1er ét., 2P., gde cuis., bns, 52 m², état neuf, 245 000 €
15. MONTPARNASSE, Part. vend atelier d'artiste 40 m², duplex, vue imprenable, tout confort, Prix 200 000 €	16. RUE D'ASSAS, imm. gd standing, bel appart 260 m², triple récept. + 5 ch., tt cft (travaux) 2 park., 2 ch. Serv., Prix 1 500 000 € à déb.
17. BD St-GERMAIN, 4P., 70 m², à amén., 4e ét., 325 000 €	18. ÎLE St-LOUIS, Lux. appt., 117 m², en duplex, gde récept., gde chambre, 2 sdb, Terras., parf. et., décor tr. bon goût, 950 000 €
19. JUSSIEU, Charme, gd 3 pces, 90 m², 378 000 €	20. QUARTIER LATIN, 30m² à aménager, prix 78 000 €
21. MONTPARNASSE, Imm. p.d.t., 4-5 P., 105 m², bon état, 375 000 €	22. RUE MAZARINE, 4e ét., sans ascens., 52 m² à rénover. Prix total 200 000 €
23. CENSIER, Bel imm., 4P. 80 m², tt cft, petits travaux, 270 000 €	24. ASSAS LUXEMBOURG, 3P. 60 m² s/arbres, imm. caractère, 295 000 €
25. SUR JARDINS OBSERVATOIRE, 140 m², grand charme, 990 000 €	26. RUE DE SAVOIE, 4e ét., Studio 20 m², dche, 85 000 €. crédit possible
27. PRES LUXEMBOURG, Bel imm., pierre de taille, Appartement 100 m², salon, sal. à manger, 2 chbres, office, cuis., bains, chf. cent., asc., prix : 495 000 €	28. Mo GOBELINS, studio, cuis., s. de bains, 28m², calme. Prix 85.000 €

FIGURE 60. Liste de 28 appartements à vendre.

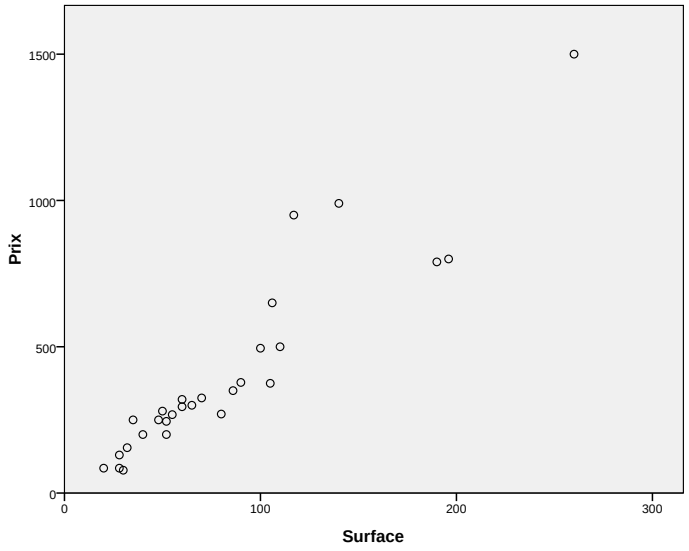


FIGURE 61. Représentation graphique des données de prix et surfaces d'appartements.

1.2. Notion de modèle (sur l'exemple des prix des appartements). On veut écrire les données à expliquer y_j sous la forme suivante :

$$y_j = f(x_j) + e_j \ , \qquad \text{pour tout } j = 1, \ldots, n \, .$$

Cela dit que la variable à expliquer y_j (les prix) est la somme de deux facteurs :

- un facteur dit modélisé ou expliqué $f(x_j)$, parce qu'il ne dépend que de la variable explicative (les surfaces) ;
- un autre facteur e_j dit résiduel, qui englobe tous les paramètres objectifs et subjectifs autres que la surface (ceux précédemment décrits dans l'exemple : garage, charme, étage, etc.).

Dans ce qui suit, on s'intéressera uniquement aux relations f affines, c'est-à-dire du type, pour a et b deux réels (les mêmes pour tout l'échantillon) :

$$y_j = a + bx_j + e_j \ .$$

REMARQUE 13.2 (Autres types de dépendance). Quitte à considérer les $\ln x_j$ ou les x_j^2 en lieu et place des x_j, on peut évidemment aussi s'intéresser à des relations comme

$$y_j = a_0' + b_0' x_j^2 + e_j' \qquad \text{ou} \qquad y_j = a_0'' + b_0'' \ln x_j + e_j'' \ .$$

C'est sur les figures de répartition dans le plan, comme la figure 61, que l'on devine le bon type de relation. Cela demandez du nez et de l'expérience. Souvent, par une transformation Φ appropriée (et non linéaire) des x_j, on arrive à une relation qui semble linéaire entre les y_j et les $\Phi(x_j)$; on s'intéresse alors au modèle linéaire liant les y_j aux $x_j' = \Phi(x_j)$. C'est pourquoi, sans perte de généralité, on se restreint dans la suite aux relations affines, paramétrées par deux réels a et b.

REMARQUE 13.3 (Avertissement !). Un modèle ne conduit presque jamais à un ajustement parfait : les résidus e_j ne sont pas nuls en général, même lorsque a et b sont choisis le mieux possible. C'est parce que les x_j n'épuisent pas toute la complexité des y_j. Un modèle est d'autant meilleur que les e_j sont faibles.

EXEMPLE 13.2 (Un peu d'histoire : pourquoi parle-t-on de régression ?). Sir Galton (homme de science britannique, 1822–1911) étudiait la taille des fils y_j (variable à expliquer) en fonction de la taille des pères x_j (variable explicative). Il a noté un retour vers un comportement moyen : les pères grands donnaient naissance à des fils plus petits, et les pères petits donnaient naissance à des fils plus grands. "Regression" signifie en anglais « retour » (vers la moyenne, ici). D'où, vous commencez à vous y habituer, la mauvaise traduction franglaise « régression », désormais synonyme de relation en statistique.

Mathématiquement, on s'attend à une relation du type $y_j = m_0 + b(y_j - m_0) + e_j$, avec m_0 la taille moyenne de la population. (C'est bien une relation de forme affine $y_j = a + bx_j + e_j$.) Ici, on voudrait déterminer b et notamment, tester que l'encadrement $0 < b < 1$ est significatif, ce qui est la condition mathématique correspondant à un retour vers la moyenne.

Profitons-en, justement, pour décrire les objectifs et motivations situés derrière le concept de régression linéaire.

1.3. Objectifs et motivations. Dans cette partie sur la régression linéaire simple, nous aurons en tête deux objectifs.

Objectif 1 : Existence d'une relation et nature éventuelle. Sur des données, il est facile de calculer les valeurs numériques des meilleurs coefficients a et b, par la méthode dite des moindres carrés, que l'on décrit plus loin. Une question plus ardue est de savoir si la valeur de b est significativement différente de 0, i.e., si la considération d'une influence linéaire des x_j sur les y_j est profitable ou non pour la compréhension des données ; et dans ce cas, à quel point c'est profitable. Là encore, on voit qu'on est confronté à des problèmes de quantifications.

Les questions que l'on se pose sont ainsi les suivantes :

- Y a-t-il réellement une relation linéaire entre la variable à expliquer et la variable explicative, entre le prix d'un appartement et sa surface (voir exemple 13.1), entre le prix d'un forfait de ski et la taille du domaine skiable correspondant (voir exercice 13.3), entre le taux de malades mentaux dans la population anglaise et le nombre de téléviseurs (voir exercice 13.4) ?
- De plus, dans les cas où une telle relation existe : en quelle proportion la variable à expliquer peut-elle être reconstruite à partir de la variable explicative ?
- Et il faudra par ailleurs également interpréter cette relation linéaire : que nous enseigne, d'un point de vue économique et stratégique, la relation linéaire calculée ?

Objectif 2 : Prédiction de nouvelles valeurs et/ou détection de valeurs atypiques. Une fois une relation exhibée et validée, on peut envisager deux autres lignes d'étude.

- La détection de couples atypiques dans les données à disposition : par exemple, dans le cas des forfaits de ski, quelles sont les stations qui sur-facturent honteusement l'accès à leur domaine ? Ou, dans le cas des appartements en vente : quelles sont les bonnes affaires et, a contrario, qui sont les vendeurs trop gourmands ? On essaie de voir, dans chaque cas, les occurrences y_j de la variable à expliquer qui sont beaucoup plus petites ou beaucoup plus grandes que les valeurs attendues au vu de l'occurrence correspondante de la variable explicative x_j.
- La prédiction de nouvelles valeurs : étant donné une nouvelle occurrence x_{n+1} de la variable explicative, à quelles valeurs y_{n+1} dois-je raisonnablement m'attendre ? Cela aide par exemple une station de ski qui renouvelle son offre à fixer ses tarifs ; ou cela permet à une agence de conseiller des prix de vente à de nouveaux clients.

2. Analyse descriptive : choix de la meilleure régression linéaire

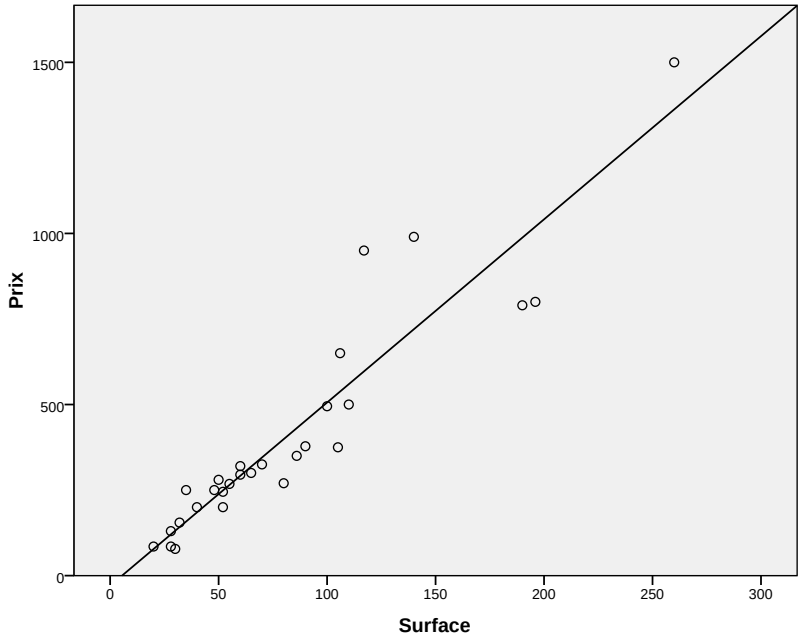
2.1. Calcul des coefficients de la relation linéaire : la méthode des moindres carrés. On considère la fonction

$$F_n : (\alpha,\ \beta) \in \mathbb{R}^2 \longmapsto \sum_{j=1}^{n} \left(y_j - (\alpha + \beta x_j)\right)^2$$

quantifiant pour chaque couple $(\alpha,\ \beta) \in \mathbb{R}^2$ la qualité de l'ajustement linéaire qui lui est associé ; cette quantification est en termes de la somme des carrés des résidus associés. On rappelle que plus les résidus sont petits, meilleur est l'ajustement. Ainsi, on choisit pour couple (a, b) au vu des données un antécédent du minimum de F_n, ce que l'on note par :

$$(a, b) \in \operatorname*{argmin}_{(\alpha,\,\beta)\in\mathbb{R}^2} F_n(\alpha, \beta) = \operatorname*{argmin}_{(\alpha,\,\beta)\in\mathbb{R}^2} \sum_{j=1}^{n} \left(y_j - (\alpha + \beta x_j)\right)^2 .$$

Les logiciels de statistique, comme SPSS, calculent pour nous un tel couple (a, b) — voir la figure 62 — et tracent également la droite de régression, qui est l'ensemble des couples $(x,\ a + bx)$.



Régression

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-29,466	...		...	...
	Surface	5,353	...	...	...	...

a. Variable dépendante : Prix

FIGURE 62. Droite de régression pour les données de prix d'appartements, d'équation $y' = -29.466 + 5.353\,x'$ et correspondant à une modélisation des prix y_j comme fonction de la surface $-29.466 + 5.353\,x_j$ plus des termes résiduels.

Cette droite passe toujours par le point $(\overline{x}_n, \overline{y}_n)$; c'est un fait qui découle des éléments culturels indiqués ci-dessous. Elle passe également par les points notés $\widehat{y}_j = a + bx_j$ pour $j = 1, \dots, n$. Les $\widehat{y}_j$ sont les valeurs dites modélisées ou reconstruites par notre modèle. On rappelle qu'à cause de la contrainte de modèle linéaire, cette reconstruction ne peut être parfaite et que les écarts résiduels

$$e_j = y_j - \widehat{y}_j = y_j - (a + bx_j) \ , \qquad j = 1, \dots, n,$$

quantifient la qualité du modèle linéaire.

La minute SPPS 13.1. On obtient la figure 62 de la manière suivante. Premièrement, on trace avec le `Générateur de diagrammes` la représentation dans le plan déjà considérée à la figure 61, puis on double-clique dessus, on sélectionne les points et on clique sur la bouton `Ajouter une courbe d'ajustement`. Ensuite, on fait calculer le tableau d'analyse de la régression (nous en présentons une version partiellement masquée) : `Analyse / Régresssion / Linéaire`. On y lit les valeurs de a et b, soit, respectivement, -29.466 et 5.353.

Eléments culturels. Revenons sur le calcul de (a, b). Tout d'abord, un minimum global de F_n existe, puisque F_n tend vers $+\infty$ lorsque α et/ou β tendent vers un infini. On obtient l'antécédent de ce minimum par étude de fonction : par calcul des points critiques de F_n. Ce sont les points (α, β) tels que

$$\frac{\partial F_n}{\partial \alpha}(\alpha, \beta) = \frac{\partial F_n}{\partial \beta}(\alpha, \beta) = 0 \ .$$

En résolvant le système de deux équations à deux inconnues, α et β, ainsi obtenu (détails omis !), on obtient un unique couple solution, dont on donne maintenant la valeur explicite, modulo l'introduction de quelques notations.

Définition 13.1. *On définit les quantités empiriques suivantes,*

$$\mathrm{Var}\big(x_1^n\big) = \frac{1}{n}\sum_{j=1}^{n}\big(x_j - \overline{x}_n\big)^2 \qquad \textit{et} \qquad \mathrm{Cov}\big(x_1^n, y_1^n\big) = \frac{1}{n}\sum_{j=1}^{n}\big(y_j - \overline{y}_n\big)\big(x_j - \overline{x}_n\big) \ ,$$

appelées respectivement variance et covariance d'échantillon.

Alors les coefficients a et b donnés par la méthode des moindres carrés sont égaux à

$$b = \frac{\mathrm{Cov}\big(x_1^n, y_1^n\big)}{\mathrm{Var}\big(x_1^n\big)} \qquad \text{et} \qquad a = \overline{y}_n - b\overline{x}_n \ .$$

En particulier, la droite de régression est l'ensemble des points

$$\mathcal{D} = \big\{\big(x, a + bx\big),\ x \in \mathbb{R}\big\} = \left\{\left(x, \overline{y}_n + \frac{\mathrm{Cov}\big(x_1^n, y_1^n\big)}{\mathrm{Var}\big(x_1^n\big)}\big(x - \overline{x}_n\big)\right),\ \ x \in \mathbb{R}\right\} \ .$$

Elle passe bien par les points $\big(x_j, \widehat{y}_j\big)$. Un calcul simple montre par ailleurs que

$$\sum_{j=1}^{n} e_j = \sum_{j=1}^{n}\big(y_j - (a + bx_j)\big) = 0 \ ,$$

c'est-à-dire que certaines des observations y_j se situent au-dessus de la droite de régression tandis que d'autres se situent en-dessous, et que leurs écarts à la droite

se compensent. C'est bien un fait que l'on observe à la figure 62. On note en particulier que l'égalité à 0 ci-dessus montre que la moyenne des valeurs reconstruites est la moyenne des valeurs originelles :

$$\frac{1}{n}\sum_{j=1}^{n}\widehat{y}_j = \overline{y}_n\,.$$

Note : Les formules précisées ci-dessus ne sont évidemment pas à connaître par cœur... et à voir leur tête, on comprend mieux l'intérêt d'un logiciel statistique réalisant les calculs à notre place !

2.2. Appréciation de la qualité de l'ajustement linéaire : le coefficient de détermination r^2. On prouvera dans la partie suivante que, dans un cadre même plus général, la méthode des moindres carrés conduit à la décomposition suivante. (Cela découlera d'un théorème de Pythagore associé à des projections orthogonales.)

THÉORÈME 13.1. *La somme des carrés totale Σ_T est égale à la somme des carrés retrouvée par la régression Σ_E plus la somme des carrés résiduelle Σ_R,*

$$\underbrace{\sum_{j=1}^{n}(y_j-\overline{y}_n)^2}_{\overset{not.}{=}\Sigma_T} = \underbrace{\sum_{j=1}^{n}(\widehat{y}_j-\overline{y}_n)^2}_{\overset{not.}{=}\Sigma_E} + \underbrace{\sum_{j=1}^{n}(y_j-\widehat{y}_j)^2}_{\overset{not.}{=}\Sigma_R}\,.$$

Interprétation des différents termes :

- Σ_T mesure la variabilité des données à expliquer y_j autour de leur moyenne $\overline{y}_n$; à un coefficient $1/(n-1)$, elle est égale à l'estimée débiaisée de la variance.
- Le terme Σ_E mesure la variabilité des reconstructions du modèle $\widehat{y}_j$ autour de leur moyenne, dont les éléments culturels ci-dessus ont montré qu'elle valait également $\overline{y}_n$; la quantité Σ_E est en un sens la variabilité intrinsèque du modèle.
- Le dernier terme, Σ_R, est égal à la somme des carrés des résidus et il mesure donc quant à lui la taille totale de ces résidus (leur variabilité autour de leur moyenne nulle).

DÉFINITION 13.2. *Le coefficient de détermination r^2 est la fraction de la variabilité totale retrouvée par la régression,*

$$r^2 = \frac{\Sigma_E}{\Sigma_T}\,.$$

On a donc $0 \leqslant r^2 \leqslant 1$. Le cas limite $r^2 = 1$ exprime une adéquation linéaire parfaite : les écarts résiduels e_j sont tous nuls, le modèle linéaire semble parfaitement déterministe, au vu des données recueillies. Au contraire, une faible valeur de r^2 indique une faible liaison linéaire entre les x_j et les y_j ; attention ! cela ne signifie pas qu'il n'existe pas de liaison significative entre x_j et y_j : lorsque celle-ci existe, et cela est possible, elle n'est simplement pas linéaire (les y_j peuvent par exemple être linéaires en les x_j^2).

LA MINUTE SPPS 13.2. Avec toujours `Analyse / Régression / Linéaire`, mais en cochant la case `Qualité de l'ajustement` dans `Statistiques`, on peut obtenir la figure 63 dans laquelle on lit, respectivement :

- dans le premier tableau, les valeurs réalisées de $\sqrt{r^2}$ et r^2 ;
- dans le second tableau, et dans cet ordre, les valeurs réalisées de Σ_E, Σ_R et Σ_T.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,930[a]	,865	...	...

a. Valeurs prédites : (constantes), Surface

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2527207,505	...	...	...	...[a]
	Résidu	392963,209	...	...		
	Total	2920170,714	...			

a. Valeurs prédites : (constantes), Surface

b. Variable dépendante : Prix

FIGURE 63. Mesures de la qualité de l'ajustement linéaire (certaines quantités non encore étudiées à ce point du cours ont été masquées).

Ici par exemple, la valeur réalisée pour r^2 est 86.5 %. On traduira à l'homme de la rue (de manière un peu abusive[35] mais simple à comprendre) : la (variabilité de la) surface explique 86.5 % (de la variabilité) du prix de l'appartement.

C'est évidemment la valeur de r^2, grande ou faible (selon n), qui déterminera si la relation linéaire proposée est significative du point de vue statistique (si b a une valeur significativement différente de 0). Mais pour détailler ce point, il nous faut enrichir notre point de vue et passer d'une analyse brute des données à un point de vue stochastique : on va considérer les données comme la réalisation de variables aléatoires.

35. Abusive, parce que cette proportion est celle pour les données considérées mais pas nécessairement celle pour l'ensemble de la population ; c'en est simplement une estimée.

3. Enrichissement du point de vue : modèle linéaire gaussien

On propose un cadre simple et agréable pour travailler en théorie : le cadre dit du modèle linéaire gaussien. Ce n'est évidemment pas un ensemble d'hypothèses minimal, mais il est adapté au niveau de ce cours. Plus précisément, en considérant que les valeurs x_j de la variable explicative sont fixées, on suppose désormais que les données $y_1, \dots, y_n$ correspondent aux réalisations des variables aléatoires indépendantes (mais pas identiquement distribuées)

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j \ , \qquad j = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont elles des variables aléatoires indépendantes et identiquement distribuées selon une selon une loi normale $\mathcal{N}(0, \sigma_0^2)$.

On suppose donc qu'il existe une relation linéaire entre une variable à expliquer et une variable explicative, à un terme stochastique près. Ce dernier englobe de nombreux facteurs explicatifs et tient compte du caractère un peu frustre d'une modélisation linéaire déterministe. Par théorème de la limite centrale flou, puisqu'une foule de petits facteurs indépendants expliquent les variations de la variable à expliquer par rapport à l'ajustement linéaire sous-jacent par la variable explicative, ce terme stochastique est effectivement de loi normale. En outre, nous avons besoin que les différents écarts stochastiques ε_j

- soient indépendants, d'une part ; c'est le cas dès que éléments fondant les couples (Y_j, x_j), par exemple, les appartements, les stations de ski, etc., ne sont pas reliés entre eux et n'ont pas d'influence les uns sur les autres ;
- aient tous même variance σ_0^2, d'autre part ; c'est une condition plus restrictive et qu'il faudra vérifier au moins *a posteriori* ; *a priori*, on peut s'aider de la représentation graphique. Ainsi, celle de la figure 61 semble indiquer par exemple que cette hypothèse peut être tenue pour vraie, mais uniquement pour les surfaces plus petites que 120 m^2.

On rappelle que les seules réalisations auxquelles on aura accès sont celles des Y_j, notées y_j. On n'observera pas en revanche les réalisations des ε_j.

Objectifs (mathématiques) : On a trois paramètres, que l'on va chercher à estimer, encadrer et tester : α_0, β_0 et σ_0^2. Ce traitement mathématique nous permettra de répondre au passage aux différentes questions qui motivaient notre étude.

3.1. Estimateurs des coefficients α_0 et β_0.

Avant on avait noté a et b les coefficients calculés par la méthode des moindres carrés au vu des x_j et y_j. Désormais, on va noter $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ les coefficients calculés au vu des Y_j et des x_j, parce que ce sont des estimateurs des coefficients sous-jacents inconnus α_0 et β_0. Les coefficients a et b correspondent alors à des estimées : ce sont les réalisations des estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

Définition 13.3. *On appelle estimateurs des moindres carrés de* (α_0, β_0) *le couple* $(\widehat{\alpha}_n, \widehat{\beta}_n)$ *tel que*

$$(\widehat{\alpha}_n, \widehat{\beta}_n) \in \operatorname*{argmin}_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{j=1}^{n} \big(Y_j - (\alpha + \beta x_j)\big)^2 \ .$$

Les résidus stochastiques reconstruits

$$\widehat{\varepsilon}_j = Y_j - \widehat{Y}_j = Y_j - (\widehat{\alpha}_n + \widehat{\beta}_n x_j) \ , \qquad j = 1, \dots, n,$$

seront utilisés dans la théorie pour mesurer l'adéquation du modèle. Ils ne sont pas égaux aux ε_j, mais presque : aux erreurs d'estimations près de α_0 et β_0 par $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

THÉORÈME 13.2. *Dans le modèle linéaire gaussien, les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ suivent les lois normales*

$$\widehat{\beta}_n \sim \mathcal{N}\left(\beta_0, \frac{\sigma_0^2}{n\ \mathrm{Var}(x_1^n)}\right) \qquad \text{et} \qquad \widehat{\alpha}_n \sim \mathcal{N}\left(\alpha_0, \frac{\sigma_0^2}{n}\left(1 + \frac{(\overline{x}_n)^2}{\mathrm{Var}(x_1^n)}\right)\right) .$$

Ils sont en particulier sans biais.

ELÉMENTS CULTURELS. Un précédent encart d'éléments culturels fournissait les expressions explicites de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$: tous deux sont des combinaisons linéaires des Y_j. Or, ces dernières sont des variables aléatoires indépendantes, chacune de loi normale, puisque les ε_j le sont également. En conséquence, les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ suivent chacun une loi normale, dont il ne reste plus qu'à déterminer les paramètres. Il est facile de calculer leurs espérances. Ainsi, il suffit de voir, pour $\widehat{\beta}_n$, que

$$\begin{aligned}\mathbb{E}\left[\mathrm{Cov}(x_1^n, Y_1^n)\right] &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)\,\mathbb{E}\left[Y_j - \overline{Y}_n\right] \\ &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)\left((\alpha_0 + \beta_0 x_j) - (\alpha_0 + \beta_0 \overline{x}_n)\right) = \beta_0\ \mathrm{Var}(x_1^n) .\end{aligned}$$

Puis, pour $\widehat{\alpha}_n$,

$$\mathbb{E}\left[\widehat{\alpha}_n\right] = \mathbb{E}\left[\overline{Y}_n\right] - \mathbb{E}\left[\widehat{\beta}_n\right]\ \overline{x}_n = (\alpha_0 + \beta_0 \overline{x}_n) - \beta_0 \overline{x}_n = \alpha_0 .$$

Ne reste plus qu'à déterminer leurs variances. Mais pour le coup, c'est un calcul pas forcément aisé ni agréable... que je ne détaille pas, par conséquent.

Par ailleurs, sachez qu'on peut montrer que parmi tous les estimateurs sans biais de α_0 et β_0 qui sont fonctions linéaires des Y_j, les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ sont ceux de variance minimale. (Ce résultat reste vrai même quand les ε_j ne suivent pas une loi normale mais sont juste supposées admettre un moment d'ordre deux.)

3.2. Estimateur de la variance des résidus. Ce qui concerne l'estimation de la variance σ_0^2 des écarts aléatoires au comportement linéaire est à rapprocher des résultats sur l'estimation de la variance énoncés dans les compléments facultatifs de la partie 5.

On commence par noter que par définition du modèle linéaire gaussien et de la loi du χ^2,

$$\frac{1}{\sigma_0^2}\sum_{j=1}^{n}\varepsilon_j^2 \sim \chi_n^2 .$$

Ici, le mieux qu'on puisse faire est de remplacer les ε_j (inconnus et non observés) par les quantités construites à partir des observations $\widehat{\varepsilon}_j$; i.e., on est amené à considérer la somme des carrés des écarts résiduels, qu'on avait notée Σ_R et dont on peut prouver qu'elle suit également une loi du χ^2, mais à $n-2$ degrés [36] de liberté :

$$\frac{\Sigma_R}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{j=1}^{n}(\widehat{\varepsilon}_j)^2 \sim \chi_{n-2}^2 .$$

36. Cette fois-ci, on a dû estimer deux paramètres, α_0 et β_0, alors que dans le cas d'un unique échantillon $X_1, \ldots, X_n$, cf. la partie 5, on n'avait besoin que d'estimer un paramètre : la moyenne μ_0, de sorte que l'estimateur construit suivait une loi χ_{n-1}^2.

On obtient alors en particulier le théorème suivant (qui utilise le fait qu'une loi du χ^2 à k degrés de liberté admet pour espérance k et que par définition et loi des grands nombres, la suite Z_k/k, où $Z_k \sim \chi^2_k$, converge en probabilité vers 1).

THÉORÈME 13.3. *Dans le modèle linéaire gaussien, on dispose de l'estimateur de la variance σ_0^2 donné par*

$$\widehat{\sigma}_n^2 = \frac{1}{n-2}\sum_{j=1}^{n}\left(\widehat{\varepsilon}_j\right)^2 .$$

Il est sans biais et consistant.

Une propriété supplémentaire qui serait utile si on faisait les preuves des différents théorèmes à venir est qu'en outre, $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

4. Vérification de l'existence d'une relation linéaire

L'existence d'une relation linéaire significative correspond au fait que le coefficient β_0 soit significativement différent de 0. Il s'agit donc de tester $H_0 : \beta_0 = 0$. Si le test conserve H_0, alors on en déduit que la relation linéaire n'est pas suffisamment fondée : la modélisation linéaire de la variable à expliquer en fonction de la variable explicative est pauvre et peu informative, il faut soit changer de variable explicative, soit considérer un autre type de liaison (quadratique, logarithmique, etc.).

4.1. Estimation et test sur β_0 : la variable explicative a-t-elle une influence linéaire significative sur la variable à expliquer ? Le théorème 13.2 montre que

$$\sqrt{n \ \mathrm{Var}(x_1^n)} \ \frac{\widehat{\beta}_n - \beta_0}{\sigma_0} \ \sim \ \mathcal{N}(0,1) \ ,$$

mais on ne peut déduire de ce fait aucun test ni intervalle de confiance sur β_0, car le membre de gauche met en jeu le paramètre inconnu σ_0. En estimant ce paramètre, en injectant les résultats du paragraphe 3.2, on aboutit à l'égalité en loi suivante, de laquelle se déduisent facilement intervalle de confiance sur β_0 et test de comparaison de β_0 à la valeur de référence $\beta_{\mathrm{ref}} = 0$:

$$\sqrt{n \ \mathrm{Var}(x_1^n)} \ \frac{\widehat{\beta}_n - \beta_0}{\sqrt{\widehat{\sigma}_n^2}} \ \sim \ \mathcal{T}_{n-2} \ .$$

Corollaire 13.1. *Un intervalle de confiance exact de niveau $1-p$ sur β_0 est donné par exemple par*

$$\widehat{I}_n = \left[\widehat{\beta}_n \pm t_{n-2,1-p/2} \sqrt{\frac{\widehat{\sigma}_n^2}{n \ \mathrm{Var}(x_1^n)}} \right] \ .$$

Principe 13.1. *Test de l'existence d'une relation linéaire (dans le cadre d'un modèle gaussien)*
Données : $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$
Modélisation associée : valeurs $x_1, \ldots, x_n$ fixées et observations stochastiques $Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j$, où les ε_j sont indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$
Hypothèse H_0 : $\beta_0 = 0$ (i.e., absence de liaison linéaire)
Statistique de test :

$$T_n = \sqrt{n \ \mathrm{Var}(x_1^n)} \ \frac{\widehat{\beta}_n}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \sim \mathcal{T}_{n-2}$
Comportement sous H_1 : T_n tend à prendre des valeurs ou plus grandes ou plus petites sous $H_1 : \beta_0 \neq 0$ (i.e., existence d'une liaison linéaire).

Remarque 13.4 (Totalement culturelle : pour faire le lien avec le r^2). Un calcul simple montre que la statistique de test T_n introduite dans le principe précédent est une fonction de r^2 ; que r^2 joue un rôle dans le test d'existence d'une liaison linéaire ne doit pas vous

surprendre au vu des commentaires qui suivent la définition 13.2. En pratique, pour des raisons qui seront plus apparentes dans la partie 14, les logiciels statistiques considèrent souvent le carré de T_n :

$$D_n = (T_n)^2 = \left(\sqrt{n \ \mathrm{Var}(x_1^n)} \ \frac{\widehat{\beta}_n}{\sqrt{\widehat{\sigma}_n^2}} \right)^2 = (n-2) \frac{\sum_{j=1}^n (\widehat{Y}_j - \overline{Y}_n)^2}{\sum_{j=1}^n (Y_j - \widehat{Y}_j)^2}$$
$$= (n-2) \frac{\Sigma_E}{\Sigma_R} = (n-2) \frac{r^2}{1-r^2} = \frac{\Sigma_E}{\Sigma_R / (n-2)} \ .$$

Cette statistique D_n suit une loi dite de Fisher (ici, à 1 et $n-2$ degrés de liberté). On généralisera au cours de la partie 14, dans le cas de la régression multiple, ces tests d'existence ou d'absence de relation linéaire fondés sur r^2.

On retiendra qu'ici on a essentiellement comparé la meilleure approximation $\overline{Y}_n$ des moindres carrés des Y_j sous l'hypothèse $H_0 : \beta_0 = 0$ à celles proposées sous H_1, et qui sont les $\widehat{Y}_j$.

4.2. Estimation et test sur α_0. Ils procèdent évidemment de la même méthodologie qu'employée ci-dessus ; vous pouvez les définir plus précisément en exercice : vu le temps imparti à ce cours, nous nous contenterons pour notre part de lire et interpréter les valeurs calculées par SPSS. On notera dans la suite T'_n la statistique de test associée au test de $H_0 : \alpha_0 = 0$.

5. En pratique : décryptage des sorties SPSS et interprétation économique de la relation proposée

5.1. Décryptage des sorties SPSS. Dans ce paragraphe, nous allons apprendre à lire les sorties de régression SPSS. Nous en avions déjà commenté précédemment quelques valeurs en en masquant d'autres : nous allons maintenant apprendre le sens de (presque) toutes les cases des tableaux de sorties.

LA MINUTE SPPS 13.3. Les tableaux de la figure 64 ont été obtenus par la série de clics `Analyse / Régression / Linéaire`, où l'on a coché les cases `Estimations`, `Intervalles de confiance` et `Qualité de l'ajustement` dans la fenêtre appelée par le bouton des options `Statistiques`.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,930[a]	,865	,860	122,939

a. Valeurs prédites : (constantes), Surface

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2527207,505	1	2527207,505	167,210	,000[a]
	Résidu	392963,209	26	15113,970		
	Total	2920170,714	27			

a. Valeurs prédites : (constantes), Surface

b. Variable dépendante : Prix

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			95,0% % intervalles de confiance pour B	
		A	Erreur standard	Bêta	t	Sig.	Borne inférieure	Limite supérieure
1	(Constante)	-29,466	41,246		-,714	,481	-114,247	55,316
	Surface	5,353	,414	,930	12,931	,000	4,502	6,204

a. Variable dépendante : Prix

FIGURE 64. Tableaux de régression complets.

Le premier tableau donne les valeurs réalisées des statistiques suivantes :

R	R-deux	R-deux ajusté	Erreur standard
$\sqrt{r^2}$	r^2	...	$\sqrt{\widehat{\sigma}^2_n}$

Il s'agit donc, à gauche, de mesures de la qualité de l'ajustement linéaire, et à droite, d'une estimée de l'écart-type σ_0 commun aux aléas ε_j non observés.

Le deuxième tableau donne les valeurs réalisées des quantités suivantes :

Somme carrés	ddl	Moyenne carrés	D	Sig.
Σ_E	1	Σ_E	D_n	P-val. test $H_0 : \beta_0 = 0$
Σ_R	$n-2$	$\Sigma_R/(n-2)$		
Σ_T	$n-1$			

Dans ce tableau, lorsque la P-valeur lue est petite, on rejette l'hypothèse $H_0 : \beta_0 = 0$ d'absence de liaison linéaire et on conclut à l'existence d'une relation linéaire. Si la P-valeur lue est grande, le modèle linéaire ne tient pas (auquel cas il peut exister des relations d'un autre type, non linéaires) ; il ne faudra alors pas lire ni exploiter les résultats du troisième tableau.

Ici, par exemple, on rejette avec force H_0 (cf. P-valeur quasi-nulle) et on conclut à l'existence d'une relation linéaire singificative.

Enfin, le troisième tableau donne les valeurs réalisées de :

Coeff.	Err. standard	...	t	Sig.	IC (bornes inf. et sup.)
$\widehat{\alpha}_n$	$\sqrt{\frac{\widehat{\sigma}_n^2}{n}\left(1+\frac{(\overline{x}_n)^2}{\mathrm{Var}(x_1^n)}\right)}$		T_n'	P-val. test $H_0 : \alpha_0 = 0$	cf. paragraphe 4.2
$\widehat{\beta}_n$	$\sqrt{\frac{\widehat{\sigma}_n^2}{n\ \mathrm{Var}(x_1^n)}}$	...	T_n	P-val. test $H_0 : \beta_0 = 0$	cf. Corollaire 13.1

La première colonne, nous l'avions déjà vu, précise donc les valeurs des estimées des coefficients α_0 et β_0 de la régression. La seconde colonne, `Erreur standard`, donne, aux facteurs quantiles près, la demi-longueur des intervalles de confiance sur le paramètre considéré ; les valeurs données correspondent, comme toujours, à des estimées de l'écart-type correctement [37] renormalisées. La troisième colonne est ignorée. La quatrième colonne calcule les valeurs réalisées des statistiques de test T_n' (pour $H_0 : \alpha_0 = 0$) et T_n (pour $H_0 : \beta_0 = 0$). Toutes deux suivant une loi $\mathcal{T}_{n-2}$, les P-valeurs de la cinquième colonne s'en déduisent ; on rappelle qu'elles correspondent aux tests bilatères, i.e., à $H_1 : \alpha_0 \neq 0$ et $H_1 : \beta_0 \neq 0$. Ici, on lit respectivement les P-valeurs 48.1 % (l'hypothèse $\alpha_0 = 0$ est conservée) et une valeur quasi-nulle (l'hypothèse $H_0 : \beta_0 = 0$ est en revanche clairement rejetée). Enfin, les deux dernières colonnes donnent les réalisations des intervalles de confiance sur α_0 et β_0, au niveau fixé par l'utilisateur lorsqu'il lance la régression. Notez que le premier contient 0 mais pas le second, comme on s'y attendait vu les résultats des tests de nullité.

Note : on avait expliqué que le test de l'existence d'une relation linéaire, i.e., de $H_0 : \beta_0 = 0$, pouvait être effectué avec la statistique de test D_n (deuxième tableau) ou avec T_n (troisième tableau) ; sans surprise, on lit donc la même P-valeur (quasi-nulle) pour ce test dans les deux tableaux.

37. On remarquera qu'ici, la renormalisation est singulièrement plus complexe que celle par $1/\sqrt{n}$ dans le cas des intervalles de confiance sur la moyenne !

REMARQUE 13.5 (**Attention ! Erreur fréquente !**). La valeur des estimées ne renseigne en rien sur le fait que les paramètres α_0 et β_0 soient significativement différents de 0 ou non ; ici, l'estimée -29.466 pour α_0 montre que ce dernier n'est pas significativement différent de 0 tandis que l'estimée 5.353 pour β_0 prouve que ce coefficient est, lui, significativement différent de 0. C'est une question d'échelle... les valeurs obtenues n'ont pas de sens absolu, il faut vraiment regarder le résultat du test pour conclure.

5.2. Interprétation et validation économiques du modèle exhibé. Ici, il s'agit surtout de valider et d'interpréter le modèle, d'un point de vue économique. La validation statistique a été vue plus haut : on a vu que l'on pouvait retenir, avec force, l'existence d'une relation linéaire et que l'estimée de celle-ci était donnée par

$$\begin{aligned}\text{prix (en milliers d'euros)} = -29.466 + 5.353\times \text{ surface (en m}^2\text{)} \\ + \text{ aléa d'écart-type } 122.939\end{aligned}$$

L'interprétation est une augmentation de prix moyen de 5 350 euros par m^2 supplémentaire ; cela forme presque le prix moyen au m^2 : presque, à cause du facteur constant qui vaut à-peu-près $-30\,000$ euros.

Ce dernier est fort difficilement interprétable d'un point de vue économique. Si encore il était positif, on aurait pu dire qu'il mesurait le prix et le coût des parties communes de l'immeuble. Ici, on rêverait qu'il ne soit pas là : cela tombe bien, on rappelle que dans la batterie de résultats statistiques présentés ci-dessus, on avait vu qu'on pouvait retenir l'hypothèse $\alpha_0 = 0$. On relance donc la régression en demandant à SPSS de forcer $\alpha_0 = 0$.

LA MINUTE SPPS 13.4. On lance une telle régression en forçant le coefficient constant à être nul en décochant `Inclure terme constant`) dans la fenêtre appelée par `Options` lors de la définition de la régression simple.

On voit alors qu'on propose la modélisation

$$\text{prix (en milliers d'euros)} = 5.109\times \text{ surface (en m}^2\text{)} + \text{ aléa d'écart-type } 121.819$$

qui, elle, est facilement interprétable économiquement : le prix au m^2 est de 5 109 euros.

On retiendra de ce paragraphe qu'après l'analyse et la validation statistiques, il ne faut pas oublier l'interprétation et la validation économiques du modèle : c'est le côté explicatif du modèle. Ci-dessous, on fait rapidement allusion à la détection des valeurs atypiques (à la hausse ou à la baisse) et à une utilisation du côté prédictif d'un tel modèle.

6. Prédiction en un nouveau point ; détection des valeurs atypiques

On suppose toujours que les données $y_1, \ldots, y_n$ peuvent être modélisées selon un modèle linéaire gaussien

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j \ , \qquad j = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$. On a indiqué aux paragraphes précédents comment ajuster le modèle aux données, i.e., comment estimer α_0, β_0, et σ_0^2, et comment déterminer s'il existe ou pas une relation linéaire. On passe maintenant à l'exploitation du modèle.

6.1. Prédiction en un nouveau point. Soit un nouveau point x, déterminé par le dispositif expérimental ou choisi par l'utilisateur. On veut dire le plus de choses possibles sur l'observation y_x qui lui sera associée. A cet effet, on note $Y_x = \alpha_0 + \beta_0 x + \varepsilon_x$ la variable aléatoire dont elle devrait être la réalisation (où $\varepsilon_x \sim \mathcal{N}(0, \sigma_0^2)$ est indépendante des variables aléatoires précédentes) ; son espérance est notée $\mu_x = \alpha_0 + \beta_0 x$.

On peut vouloir, dans le cadre de la prévision,

- donner un intervalle de confiance sur μ_x, l'espérance de Y_x,
- ou même, donner un intervalle de prévision, dans lequel Y_x sera avec grande probabilité.

L'annexe des compléments mathématiques facultatifs détaille comment procéder.

Nous nous contenterons ici des tracés SPSS de ces intervalles ; évidemment, les intervalles de prévision sont plus gros que les intervalles de confiance, parce qu'ils tiennent également compte de la variabilité de Y_x par rapport à son espérance μ_x, tandis que les derniers se contentent d'encadrer μ_x.

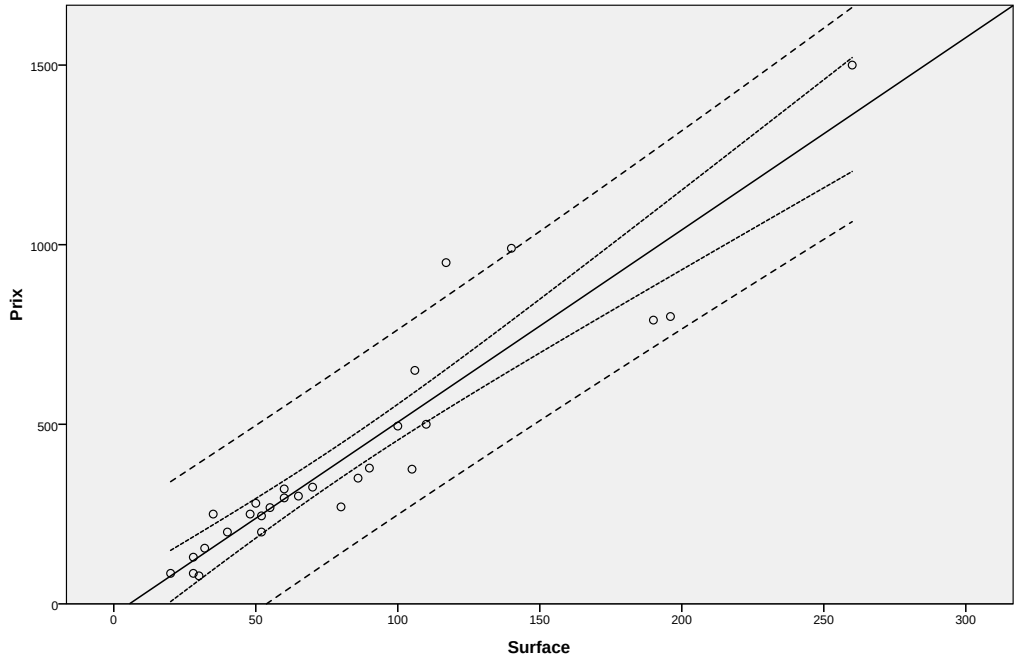


FIGURE 65. Réalisations d'intervalles de confiance (pointillés rapprochés) et de prédiction (pointillés plus éloignés), tous deux de niveau 95 %.

La minute SPPS 13.5. La figure 65 a été obtenue d'une manière similaire à celle de la Minute SPSS 13.1, en sélectionnant simplement, en outre, les boutons radios `Intervalles de confiance : Moyenne` ou `Individuelle`, selon que l'on veut le tracé des intervalles de confiance ou des intervalles de prévision.

Exploitation de la figure 65 : on peut prévoir que le prix d'un nouvel appartement de 100 m^2 à la vente sera compris entre 250 et 750 milliers d'euros ; ce n'est pas une prévision d'une très grande précision... Cette dernière est en revanche bien meilleure en ce qui concerne le prix moyen de l'ensemble des nouveaux appartements de cette superficie.

6.2. Détection et traitement éventuel des valeurs atypiques. On définit une valeur atypique comme un couple (x_j, y_j) tel que y_j n'appartient pas à l'intervalle de prévision calculé sur x_j au vu de l'ensemble des couples de données. Cela correspond à des valeurs obtenues trop grandes ou trop petites par rapport aux attentes.

Exemple 13.3. Sur la figure 65, il y a deux observations atypiques, qui correspondent à des appartements beaucoup plus chers que prévus. L'acheteur devra être prudent face à eux : sont-ce de mauvaises affaires ou leur prix est-il justifié par des éléments tangibles extraordinaires ?

Parfois, des considérations extra-statistiques expliquent pourquoi ces données atypiques méritent un traitement à part ; on peut alors relancer la régression. Les résultats (la précision des prédictions) en sont généralement fort améliorés.

Remarque 13.6 (Ne trichez pas !). Il est ici bien important de justifier la non-considération de ces données atypiques pour l'établissement d'un modèle statistique général ; il faudra évidemment ne pas les oublier et malgré tout les traiter lors de décisions stratégiques ! On ne peut pas totalement les faire disparaître d'un coup de baguette magique, ce serait malhonnête (voir figure 66).

FIGURE 66. Tricher, c'est mal. On peut effacer des données pour le traitement statistique, à condition d'avoir une justification économique de cet omission statistique et de ne pas oublier totalement le cas de ces données atypiques lors de la décision stratégique.

Compléments pour étudiants avancés

7. Estimation et prédiction en un nouveau point x (détails mathématiques)

7.1. Intervalle de confiance sur l'espérance μ_x. Certes, nous avons vu dans la version rédigée de cette partie des intervalles de confiance sur α_0 et β_0, et s'ils sont simultanément vrais (méthode de Bonferroni), alors on peut en déduire un intervalle de confiance sur $\mu_x = \alpha_0 + \beta_0 x$. L'intervalle qu'on obtiendrait ainsi (exercice : explicitez-le !) a pour défaut que sa demi-longueur croît rapidement (linéairement) avec x. On voudrait une formule moins sensible à la valeur de x.

On va procéder de la même manière qu'au paragraphe 3.1 de la version rédigée du cours. Les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ sont combinaisons linéaires des observations gaussiennes indépendantes Y_j ; l'estimateur

$$\widehat{\mu}_{x,n} = \widehat{\alpha}_n + \widehat{\beta}_n\, x$$

est lui aussi une telle combinaison linéaire et suit par conséquent une loi normale (dont il suffit de déterminer espérance et variance). L'espérance est bien μ_x, vu le caractère sans biais de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$. Des calculs similaires à ceux proposés dans les éléments culturels de ce paragraphe 3.1 permettent de déterminer la variance, puis, moyennant une studentisation, conduisent au résultat suivant.

THÉORÈME 13.4. *Dans le modèle linéaire gaussien, l'estimateur* $\widehat{\mu}_{x,n} = \widehat{\alpha}_n + \widehat{\beta}_n\, x$ *suit une loi normale,*

$$\widehat{\mu}_{x,n} \;\sim\; \mathcal{N}\left(\mu_x,\; \frac{\sigma_0^2}{n}\, h_{x,n}\right) \qquad \text{où} \quad h_{x,n} = 1 + \frac{1}{\mathrm{Var}(x_1^n)}\,(x - \overline{x}_n)^2$$

est appelé le levier en x.

Par conséquent,

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2\, h_{x,n}}}\;(\widehat{\mu}_{x,n} - \mu_x) \;\sim\; \mathcal{T}_{n-2}\ .$$

On déduit de la dernière assertion des intervalles de confiance et des tests sur μ_x, selon la cuisine habituelle. Par exemple, un intervalle de confiance exact de niveau $1-p$ sur μ_x est

$$\left[\widehat{\alpha}_n + \widehat{\beta}_n\, x \;\pm\; t_{n-2,1-p/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}\, h_{x,n}}\right]\ .$$

7.2. Intervalle de prévision sur Y_x. On cherche maintenant à donner des indications sur la valeur non plus de μ_x, mais de $Y_x = \mu_x + \varepsilon_x$. Il suffit de tenir compte de l'ajout de la perturbation aléatoire ε_x, qui est indépendante des observations $Y_1, \ldots, Y_n$ (et des aléas correspondants, $\varepsilon_1, \ldots, \varepsilon_n$).

On cherche un intervalle $\widehat{I}_{x,n}$ tel que

$$\mathbb{P}\{Y_x \in \widehat{I}_{x,n}\} \geqslant 1-\alpha$$

pour un niveau $1-\alpha$ fixé par l'utilisateur. $\widehat{I}_{x,n}$ est appelé un intervalle de prévision, plutôt qu'un intervalle de confiance. En effet, les intervalles de confiance portent sur des quantités déterministes ; quand on veut encadrer une quantité aléatoire qui sera observée plus tard, on parle, comme ici, d'intervalle de prévision.

On a de bons espoirs d'aboutir : on a un intervalle de confiance sur μ_x et on a une estimation $\widehat{\sigma}_n^2$ de la variance σ_0^2 de $\varepsilon_x \sim \mathcal{N}(0,\, \sigma_0^2)$. Des calculs simples, utilisant le résultat du théorème 13.4, montrent alors le résultat suivant.

THÉORÈME 13.5. *Dans le modèle linéaire gaussien, la différence entre l'observation Y_x et la prédiction de son espérance $\widehat{\mu}_{x,n}$ suit une loi normale,*

$$Y_x - \widehat{\mu}_{x,n} \;\sim\; \mathcal{N}\left(0,\; \sigma_0^2\left(1+\frac{h_{x,n}}{n}\right)\right) \;.$$

Une studentisation donne alors

$$\frac{1}{\sqrt{\widehat{\sigma}_n^2\,(1+h_{x,n}/n)}}\,(Y_x-\widehat{\mu}_{x,n}) \;\sim\; \mathcal{T}_{n-2}$$

et par conséquent, un intervalle de prévision au niveau $1-\alpha$ est

$$\widehat{I}_{x,n} = \left[\widehat{\mu}_{x,n} \;\pm\; t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma}_n^2\left(1+\frac{h_{x,n}}{n}\right)}\right] \;.$$

Dans ce théorème, il n'est facile de prouver que la première assertion (elle résulte d'une simple addition de lois normales indépendantes : il suffit de sommer leurs espérances et leurs variances). Le reste est admis.

Exercices

Les exercices qui suivent nécessitent des fichiers de données, disponibles, comme à l'accoutumée, sur le site web du cours.

Deux exercices issus des annales

EXERCICE 13.1. Répondez aux questions du paragraphe "Construction d'un modèle prédictif/explicatif" de l'examen de rattrapage 2008.

EXERCICE 13.2. Retrouvez les résultats du rapport de régression du dernier exercice de l'examen de rattrapage 2007, en étudiant sous SPSS le fichier `Orange.sav`. La première colonne est l'étiquette de l'oranger considéré, la deuxième donne l'âge de l'arbre en jours, et la troisième, sa circonférence mesurée (en millimètres). Répondez ensuite aux questions posées lors de l'examen.

Trois exercices supplémentaires

EXERCICE 13.3 (Pensons un peu à la semaine ski...). Le fichier `ForfaitsSki.sav` précise, pour 42 stations de ski alpines, le prix du forfait d'accès aux pistes à la semaine (en euros) et la distance totale (en km) des pistes. Les données datent de 2008. Y a-t-il une relation linéaire entre les deux ? Effectuez l'analyse de régression linéaire en cinq points. Note : il faut charger et étudier les données sous SPSS ; on constatera notamment qu'il vaudrait mieux faire l'analyse sans la station Serre Chevalier.

EXERCICE 13.4 (Regarder la télé rend-il fou ?). Chargez le fichier de données `TV.sav` sous SPSS. Il reporte chaque année (première colonne) le nombre de téléviseurs en service (en milliers, deuxième colonne) et le taux de malades mentaux (nombre pour mille habitants, troisième colonne). L'étude a été réalisée en Grande-Bretagne.

1. Quelle est d'après vous la variable à expliquer et la variable explicative ?
2. Effectuez la régression sous SPSS et étudiez son résultat.
3. Faut-il en conclure que regarder la télé rend fou, et si non, pourquoi ?
4. On pourra se convaincre qu'il faut répondre non à la question précédente en étudiant la régression du taux de malades mentaux par l'indice de l'année. Que se passe-t-il donc ?

EXERCICE 13.5 (Facultatif). Méditez la figure 67 ; que fait-on, quels sont les résultats ? Notez en particulier que la régression (simple ou multiple) sera très utilisée dans vos cours de finance !

Table IV

Predictability Regressions

$$\sigma_{t,T} = a + b\hat{\sigma}_t^T + \varepsilon_{t,T}$$

where $\sigma_{t,T}$, the volatility over the remaining life of the option contract, is regressed against the volatility forecast $\hat{\sigma}_t^T$. This includes the implied standard deviation (ISD) from option prices, a moving average (MA) with a moving window of 20 trading days, and the GARCH time-series model, which is the conditional volatility for the next day based on parameters in Table II. Periods end on February 28, 1992, and start in January 1985 (DM), July 1986 (JY), and March 1985 (SF). Regressions use daily observations, and standard errors are corrected for the induced overlap and heteroskedasticity using the Hansen-White (HW) procedure. Asymptotic HW standard errors in parentheses.

		Slopes On			
Currency	a	ISD	MA(20)	GARCH	R^2
DM	0.323*	0.547*†			0.1564
	(0.115)	(0.138)			
	0.602*		0.190†		0.0540
	(0.084)		(0.099)		
	0.366			0.478*†	0.0499
	(0.191)			(0.227)	
	0.303*	0.669*	−0.099		0.1632
	(0.112)	(0.165)	(0.101)		
	0.401*	0.622*		−0.173	0.1599
	(0.152)	(0.167)		(0.201)	
JY	0.327*	0.496*†			0.0965
	(0.118)	(0.181)			
	0.563*		0.134†		0.0223
	(0.074)		(0.102)		
	−0.063			1.017*	0.0495
	(0.323)			(0.458)	
	0.322*	0.578*	−0.073		0.1004
	(0.117)	(0.204)	(0.111)		
	0.042	0.421*		0.474	0.1051
	(0.289)	(0.177)		(0.399)	
SF	0.392*	0.520*†			0.1454
	(0.149)	(0.175)			
	0.658*		0.182†		0.0542
	(0.087)		(0.099)		
	0.250			0.650*	0.0581
	(0.267)			(0.305)	
	0.370*	0.647*	−0.097		0.1521
	(0.146)	(0.187)	(0.090)		
	0.526*	0.609*		−0.240	0.1490
	(0.210)	(0.201)		(0.262)	

* Significantly different from zero at the 5 percent level.
† Significantly different from unity at the 5 percent level.

FIGURE 67. Un extrait d'un article de finance (envoyé par Christophe Pérignon).

Exercice 1 : Cf. § Construction d'un modèle prédictif / explicatif de l'examen de rattrapage 2008

1) On cherche évidemment à modéliser le montant des achats effectués y_j en fonction des revenus du foyer x_j.

2) Cela correspond à la régression linéaire fournie par la 2nde série de tableaux.

3) [On lit une valeur réalisée de 79,7% pour le coefficient de détermination r^2, ce qui est élevé.] ↔ [Par ailleurs, le test d'existence d'une relation linéaire rejette avec force l'hypothèse que $H_0 : \beta_0 = 0$ (cf. P-valeur quasi-nulle), il existe donc une relation linéaire entre les y_j et les x_j.]

On propose la relation :

$$\text{montant des achats (en euros)} = -85.014\text{€} + 0.333 \times \text{revenus mensuels en euros} + \text{aléa}$$

aléa : d'écart-type estimé 117.158 €

Validation éco : le −85.014 peut sembler étrange, en fait, il est nécessaire statistiquement (le test de $H_0 : \alpha_0 = 0$ admet une P-valeur également quasi-nulle, on ne peut donc pas retenir $\alpha_0 = 0$). Cela étant, ce −85.014 est "compensé" par le revenu minimal qu'est le RMI (autour de 400€) : $-85 + 0.333 \times 400 > 0$!

Exercice 2: Cf. Exercice III de l'examen de rattrapage 2007

1) Les tableaux proposent :

$$\underset{\text{(en mm)}}{\text{Circonférence}} = 17.4\ \text{mm} + 0.107 \times \underset{\text{(en jours)}}{\text{âge}} + \text{aléa}$$

↑ d'écart-type estimé 23.738

2) Oui, il existe une liaison linéaire forte : la P-valeur du test qui prend pour hypothèse $H_0 : \beta_0 = 0$ (pas de relation linéaire) est quasi-nulle, on en déduit l'existence d'une relation linéaire.
En fait, le modèle linéaire semble très adapté ici : le r^2 a une valeur réalisée de 83.5%, ce qui est élevé.

3) * Que penser du 17.4 mm, l'ordonnée à l'origine ? Un arbre de quelques jours, c'est une tige de 2 ou 3 mm de circonférence, pas plus ...
En fait, le test de $H_0 : \alpha_0 = 0$ a une P-valeur de 5.2%, cas borderline, mais on est tenté pour des raisons d'interprétation de conserver H_0 (sur le fil) et de relancer la régression avec ce forçage de α_0 à 0.
On obtient ainsi, comme on pourra le vérifier sous SPSS :

$$\underset{\text{(en mm)}}{\text{Circonférence}} = 0.122 \times \underset{\text{(en jours)}}{\text{âge}} + \text{aléa}$$

↕ d'écart-type estimé 24.785

* On peut vérifier sous SPSS (intervalle de prévision) qu'il n'y a pas de valeurs atypiques.

NOTE : Cet exo cadre avec ce que l'on vous racontait en SVT, les troncs d'arbres et leurs anneaux successifs, d'aires de plus en plus grands mais de section constante !

Exercice 3: (Prix de forfaits de ski)

Les données (ainsi que la régression obtenue avec Serre-Chevalier : droite de régression et intervalles de prévision à 95%) et les sorties de régression sont reproduits ci-après.

1) Avec Serre-Chevalier :

(1) il existe bien une liaison linéaire entre les prix y_i des forfaits et la surface x_i des domaines (cf. P-valeur quasi-nulle pour le test de $H_0: \beta_0 = 0$)

(2) qualité de l'ajustement bonne : valeur réalisée de r^2 égale à 57.1%

(3) relation estimée :

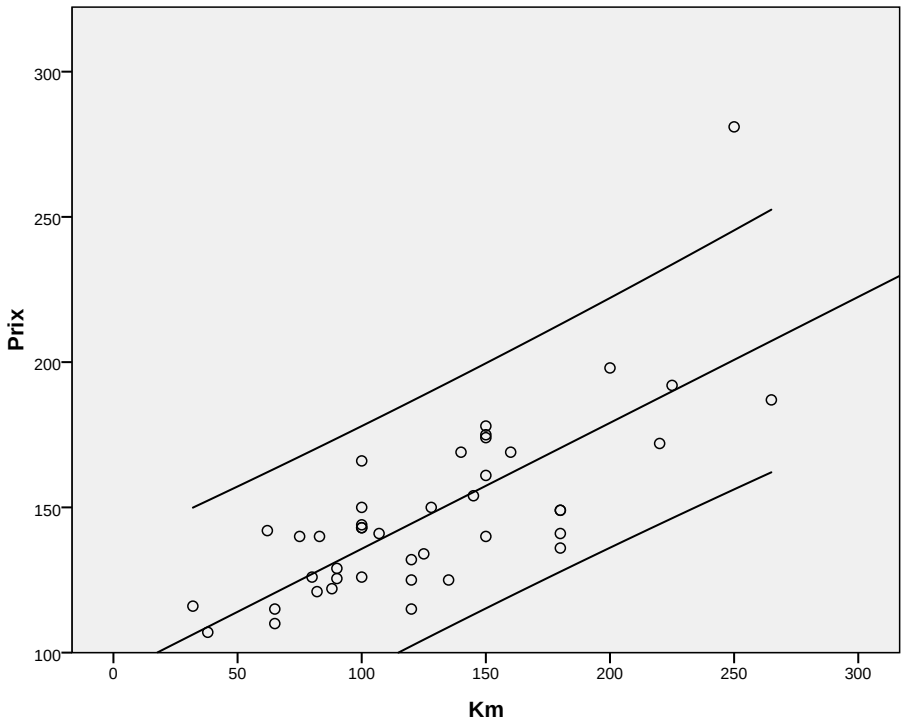
$$\underset{\text{semaine (en €)}}{\text{prix forfait}} = 92.366\,€ + 0.434 \times \underset{\text{(en kms)}}{\text{pistes}} + \underset{\substack{\uparrow \\ \text{d'écart-type} \\ \text{estimé } 20.594}}{\text{aléa}}$$

(4) il y a bien une ordonnée à l'origine (cf. P-valeur quasi-nulle pour le test de $H_0: \alpha_0 = 0$), on l'interprète comme le coût moyen d'accès aux pistes (remonte-pente station → pistes). Ensuite, chaque km supplémentaire de pistes coûte en moyenne 43 cts aux usagers.

(5) Cela dit, Serre-Chevalier est une donnée atypique, il serait plus sûr de refaire l'étude sans elle.

↳ Qui y est déjà allé et aurait une explication extra-statistique au fait de ne pas considérer cette station (ou en tout cas, de la considérer à part) pour l'établissement d'un modèle ?

2) Sans Serre Chevalier : (3) prix = 102.354 + 0.338 × pistes + aléa (éc.-type : 15.539)
// Points (1), (2), (4) faciles à adapter // (5) plus de données atypiques (refaire le graphe pour le vérifier, cf. les intervalles de prévision changent un peu : sont plus petits).



Régression (avec Serre Chevalier)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,756[a]	,571	,560	20,594

a. Valeurs prédites : (constantes), Km

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	22596,541	1	22596,541	53,280	,000[a]
	Résidu	16964,418	40	424,110		
	Total	39560,958	41			

a. Valeurs prédites : (constantes), Km

b. Variable dépendante : Prix

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	92,366	8,247		11,200	,000
	Km	,434	,059	,756	7,299	,000

a. Variable dépendante : Prix

Régression (sans Serre Chevalier)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,749[a]	,560	,549	15,539

a. Valeurs prédites : (constantes), Km

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	12001,429	1	12001,429	49,707	,000[a]
	Résidu	9416,376	39	241,446		
	Total	21417,805	40			

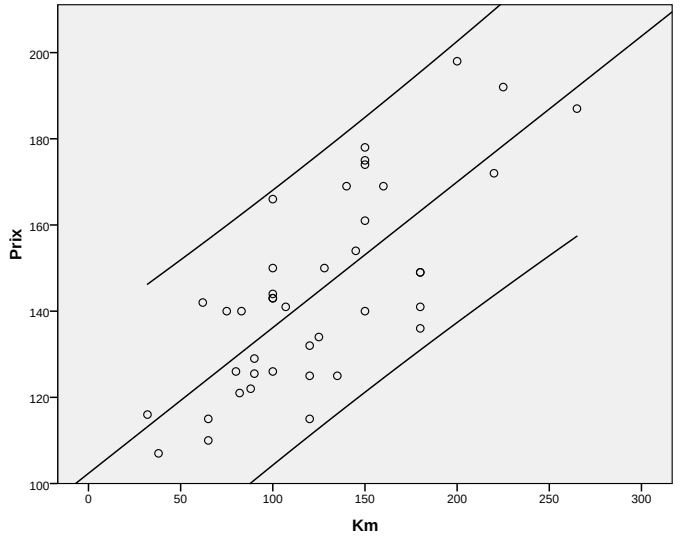
a. Valeurs prédites : (constantes), Km

b. Variable dépendante : Prix

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	102,354	6,474		15,810	,000
	Km	,338	,048	,749	7,050	,000

a. Variable dépendante : Prix



Exercice 4 :

1) Faisons mine de croire un instant que regarder la télé rend idiot (ou fou) : on veut expliquer la proportion y_j de malades mentaux en fonction du nombre x_j de téléviseurs. (Attention, il s'agit d'une explication au sens statistique : une modélisation.)

2) Etude de la régression des y_j par les x_j :

(1) le second tableau (P-valeur quasi-nulle) indique l'existence forte d'une relation linéaire ;

(2) la valeur réalisée de r^2 (98.4%) est très élevée, signe d'un excellent ajustement des données de l'échantillon au modèle linéaire.

(3) on propose la relation estimée :

$$\text{proportion de malades mentaux (en ‰)} = 4.552 + 0.222 \times \text{nombre de télés (en milliers)} + \text{aléa}$$

(aléa : d'écart-type estimé 0.728)

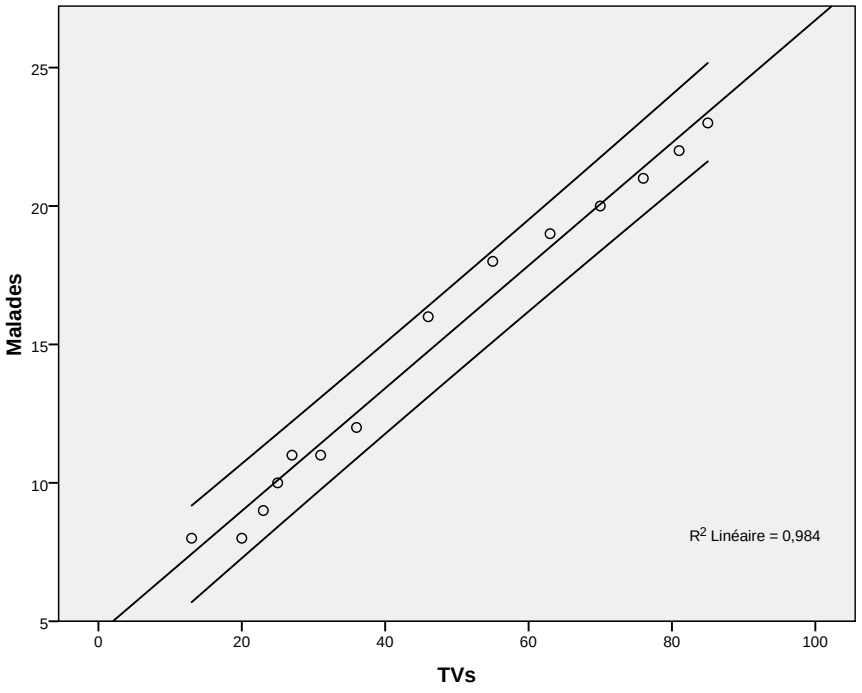
(4) → interprétation : 4.552 = taux de folie résiduel (avant TV) ? (sujette à caution !)

$\uparrow$ estimée de α_0, P-valeur quasi-nulle pour le test de $H_0 : \alpha_0 = 0$, on sait que $\alpha_0 \neq 0$.

→ validation & conclusion : dangereuses. On note une bonne explication statistique mais pas nécessairement causale

3) ↳ en fait, la proportion de malades mentaux et le nombre de télés croissent sans doute tous deux linéairement en fonction d'une troisième variable : le Progrès (ou : le temps qui passe).

C'est cette variable mesurant le temps qui passe qui explique les choses : nos comportements ont changé entre 1970 et 1983, la télé s'est généralisée et on prend mieux soin des



Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,992[a]	,984	,983	,728

a. Valeurs prédites : (constantes), TVs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	393,361	1	393,361	742,976	,000[a]
	Résidu	6,353	12	,529		
	Total	399,714	13			

a. Valeurs prédites : (constantes), TVs

b. Variable dépendante : Malades

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	4,552	,425		10,707	,000
	TVs	,222	,008	,992	27,258	,000

a. Variable dépendante : Malades

malades mentaux, c'est tout.

4) On peut se persuader de la justesse de la réponse précédente en effectuant les régression du taux de malades ou du nombre de tél's par rapport aux années z_j :

y_j malades / années z_j r^2 réalisé de 96.4% } deux valeurs elles aussi très grandes.
x_j TV / années z_j : 97.2% }

Ainsi, { les y_j sont quasiment des fonctions linéaires des z_j / les x_j sont quasiment des fonctions linéaires des z_j }

et par transitivité, les y_j sont quasiment des fonctions linéaires des x_j

On rappelle que les z_j mesurent ici simplement le temps qui passe et l'évolution des comportements.

MORALE DE CET EXERCICE :

La régression linéaire permet de modéliser et prédire, d'expliquer de manière statistique mais pas causale.

Exercice 5:

Le centre des tableau rapporte des estimés du parametre β_0 de liaison linéaires, pour differents jeux de données. Ce β_0 est noté b dans le titre de la page.

Par ailleurs, des tests d'hypothèses sont menés, sur H_0: $\beta_0 = 0$ et sur H'_0: $\beta_0 = 1$. Lorsque les P-valeurs correspondantes sont $\leq 5\%$, et que H_0 ou H'_0 sont rejetées, ces faits sont indexés respectivement par les nots * et †.

La premiere colonne rapporte des estimés de α_0 (a ici).
La derniere colonne donne les valeurs réalisées de r^2, toutes plutôt faibles (inférieures à 15%).

Quatorzième Partie

Régression linéaire multiple

Version rédigée du cours

Résumé : Le chapitre précédent a introduit le modèle statistique de la régression linéaire simple, qui permet de rendre compte une variable quantitative (variable dite à expliquer) comme fonction affine d'une seule autre variable quantitative (variable dite explicative), à quoi s'ajoute un terme résiduel correspondant aux influences d'autres facteurs explicatifs possibles non considérés.

Objectif : Nous étudions ici le modèle de la régression linéaire multiple où l'on dispose de plusieurs variables explicatives pour rendre compte (toujours de manière affine) de la variable à expliquer. Les questions que nous nous poserons seront globalement les mêmes : existence d'une relation linéaire significative, interprétation de cette relation, détection de valeurs atypiques, prédiction. Nous ajouterons à cela une nouvelle ligne d'étude : parfois, il y a tellement de variables explicatives que toutes peuvent ne pas être directement utiles à considérer, auquel cas il s'agit d'en sélectionner un sous-ensemble permettant de prédire suffisamment efficacement la variable à expliquer tout en conservant une relation facilement interprétable.

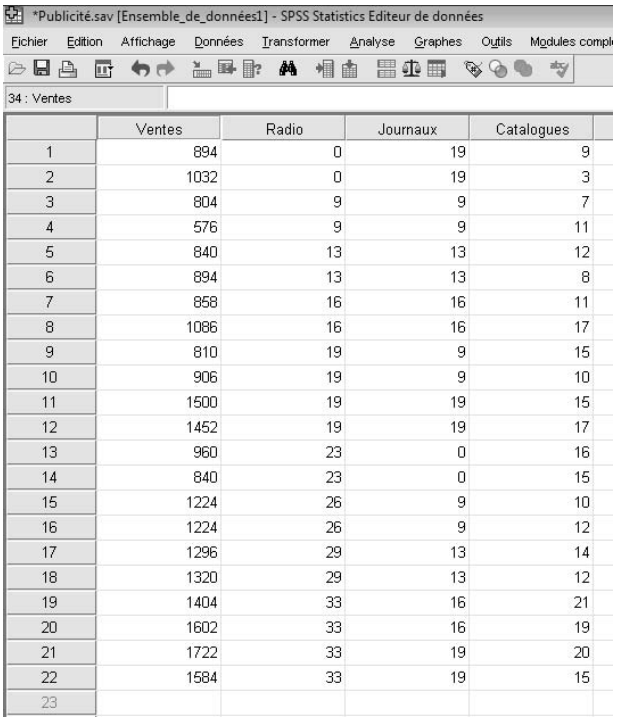
1. Le modèle linéaire multiple

1.1. Un exemple pour commencer.

EXEMPLE 14.1. On considère la figure 68, qui reporte les données $y_1, y_2, \ldots, y_{22}$ à expliquer (des volumes de ventes). A cet effet, elle propose plusieurs variables explicatives : les montants investis respectivement dans des opérations de publicité à la radio, notés $x_{1,1}, x_{1,2}, \ldots, x_{1,22}$, ou dans la presse écrite, $x_{2,1}, x_{2,2}, \ldots, x_{2,22}$, ainsi que ceux destinés à la distribution de catalogues des produits, $x_{3,1}, x_{3,2}, \ldots, x_{3,22}$. Dans chacune de 22 villes françaises de tailles similaires, on a fait varier les trois montants de campagnes publicitaires et on en a mesuré l'impact sur les ventes. On espère ainsi construire un modèle explicatif et prédictif des ventes en fonction de ces budgets publicitaires : déterminer dans quels domaines publicitaires il est crucial d'investir et si cet investissement est rentable ou non.

1.2. Le modèle linéaire gaussien multiple. Décrivons cette situation dans un cadre plus général. Pour chaque situation, indexée par $t = 1, \ldots, n$, on dispose d'une donnée y_t à expliquer en fonction d'un nombre fini k de données explicatives $x_{1,t}, x_{2,t}, \ldots, x_{k,t}$. Pour les mêmes raisons et intuitions que dans la partie 13, on modélise les données à expliquer y_t comme la réalisation des variables aléatoires

$$\begin{aligned} Y_t &= \alpha_0 + \beta_{1,0}\, x_{1,t} + \beta_{2,0}\, x_{2,t} + \ldots + \beta_{k,0}\, x_{k,t} + \varepsilon_t \\ &= \alpha_0 + \sum_{j=1}^{k} \beta_{j,0}\, x_{j,t} + \varepsilon_t \ , \qquad \text{pour } t = 1, \ldots, n, \end{aligned}$$



	Ventes	Radio	Journaux	Catalogues
1	894	0	19	9
2	1032	0	19	3
3	804	9	9	7
4	576	9	9	11
5	840	13	13	12
6	894	13	13	8
7	858	16	16	11
8	1086	16	16	17
9	810	19	9	15
10	906	19	9	10
11	1500	19	19	15
12	1452	19	19	17
13	960	23	0	16
14	840	23	0	15
15	1224	26	9	10
16	1224	26	9	12
17	1296	29	13	14
18	1320	29	13	12
19	1404	33	16	21
20	1602	33	16	19
21	1722	33	19	20
22	1584	33	19	15
23				

FIGURE 68. Le jeu de données d'exemple considéré dans ce cours : montant des ventes (en milliers d'euros) de divers produits d'une enseigne, montant des campagnes de publicité à la radio (en milliers d'euros) et dans la presse écrite (en milliers d'euros), coût (en centaines d'euros) de diffusion de catalogues des produits disponibles.

où les $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$. Les réalisations y_j des Y_j sont observées, les valeurs des $x_{k,j}$ sont observées voire choisies par le statisticien ; en revanche, les coefficients α_0 et $\beta_{1,0}, \ldots, \beta_{k,0}$ de la relation linéaire, de même que les (réalisations des) erreurs aléatoires ε_j, ne le sont pas.

Objectifs mathématiques. On cherche à estimer, encadrer et tester les valeurs des coefficients α_0 et $\beta_{1,0}, \ldots, \beta_{k,0}$, de même que celle de la variance commune σ_0^2 des aléas.

Objectifs stratégiques. Reformulons les objectifs précédents, ainsi que nous l'avions fait dans la partie 13, afin de motiver l'étude mathématique :

- vérifier l'existence d'une relation linéaire significative (i.e., qui contribue significativement à l'explication) ; cela correspond à tester l'hypothèse H_0 que tous les coefficients $\beta_{j,0}$ sont nuls contre l'hypothèse H_1 qu'au moins un de ces coefficients est non nul ;
- déterminer, parmi les coefficients $\beta_{j,0}$ ceux qui sont significativement non nuls et oublier les autres ; cela correspondra à tester différentes hypothèses individuelles sur les coefficients, $H_{0,\ell} : \beta_{\ell,0} = 0$;
- dans le modèle simplifié ainsi obtenu, quantifier la qualité de la reconstruction linéaire, lorsqu'elle est significative ; le critère mathématique sera encore appelé le coefficient de détermination r^2 ;

- interpréter les (estimées des) coefficients α_0 et $\beta_{j,0}$ en termes stratégiques et économiques ;
- construire des intervalles de prévision pour de nouvelles valeurs et détecter les valeurs atypiques parmi les valeurs existantes (celles n'appartenant pas à leur propre intervalle de prévision) ; étudier de plus près ces valeurs atypiques : sont-ce de bonnes affaires ou des escroqueries ?

En fait, le point crucial et délicat de cette étude est la construction d'un modèle simplifié : on cherche à établir un compromis entre considérer un nombre suffisamment grand de variables explicatives, pour bien rendre compte du phénomène à expliquer, sans toutefois que ce nombre ne soit trop grand, afin de conserver une relation facilement interprétable. Il s'agit donc d'isoler un bon (en un certain sens) sous-ensemble des variables explicatives.

2. Un zeste d'algèbre linéaire

2.1. Formulation matricielle compacte. On note de manière compacte le modèle linéaire gaussien sous la forme matricielle suivante :

$$\underline{Y} = \mathbf{X}\,\underline{\beta}_0 + \underline{\varepsilon} \qquad \text{où}$$

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{k,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \cdots & x_{k,n} \end{bmatrix}, \quad \underline{\beta}_0 = \begin{bmatrix} \alpha_0 \\ \beta_{1,0} \\ \vdots \\ \beta_{k,0} \end{bmatrix}, \quad \text{et} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

La matrice $\mathbf{X}$ est connue : c'est la matrice des variables explicatives ; $\underline{Y}$ est observé : c'est le vecteur des données correspondant à la variable à expliquer ; mais $\underline{\beta}_0$ est inconnu (il est même à estimer et à tester) : c'est le vecteur des coefficients de la relation linéaire. De son côté, le vecteur $\underline{\varepsilon}$ des résidus n'est pas observé.

On effectue une hypothèse supplémentaire : que $\mathbf{X}$ soit injective, afin que le vecteur $\underline{\beta}_0$ soit déterminé de manière unique (en effet, si $\mathbf{X}$ n'est pas injective, des valeurs pourtant différentes de $\underline{\beta}_0$ peuvent conduire à la même explication linéaire $\mathbf{X}\,\underline{\beta}_0$).

> Eléments culturels (Quelques mots de plus sur l'injectivité). L'injectivité de $\mathbf{X}$ peut être retraduite par le fait que les colonnes de cette matrice sont non liées ; cela implique en particulier que $k+1 \leqslant n$. Ici, on se place en outre dans le cas $k+1 < n$, sans quoi il existerait une solution parfaitement déterministe au système d'équations donnant le modèle : en effet, $\mathbf{X}$ serait inversible dans le cas où $k+1 = n$ et on pourrait reconstruire parfaitement la réalisation $\underline{y}$ de $\underline{Y}$ à partir de $\mathbf{X}$, sans tenir compte d'aucun aléa. Or, ici, on se place moralement dans une situation où l'on dispose de peu de variables explicatives face à une grande masse de données à expliquer (en fait, on pense à $k+1 \ll n$, ce qui signifie que $k+1$ est très petit face à n), de sorte qu'un aléa est nécessaire pour englober les nombreuses variables explicatives que l'on ne considère pas. En résumé, on se place donc implicitement dans une situation où $k+1 < n$ (et même $k+1 \ll n$) et où la matrice $\mathbf{X}$ est injective, c'est-à-dire ici de rang plein $k+1$ (ses $k+1$ vecteurs-colonnes de taille n forment un système libre).

Dans la suite, nous allons avoir besoin du résultat fondamental suivant : lorsque $\mathbf{X}$ est injective, $\mathbf{X}^T\mathbf{X}$ est une matrice $(k+1)\times(k+1)$ inversible.

> En effet (et cette démonstration est donnée à titre culturel et pour réveiller vos plus beaux souvenirs mathématiques...), $\mathbf{X}^T\mathbf{X}$ étant une matrice carrée, elle est inversible si et seulement si elle est injective, ce qui se traduit par le fait que pour tout vecteur $\underline{v} \in \mathbb{R}^{k+1}$, on ait $\mathbf{X}^T\mathbf{X}\,\underline{v} = \underline{0}$ si et seulement si $\underline{v} = \underline{0}$. Soit donc un vecteur $\underline{v}$ tel que $\mathbf{X}^T\mathbf{X}\,\underline{v} = \underline{0}$. En particulier,
>
> $$\|\mathbf{X}\underline{v}\|_2^2 = \underline{v}^T\,\mathbf{X}^T\mathbf{X}\,\underline{v} = 0$$
>
> où $\|\cdot\|_2$ désigne la norme euclidienne. Ainsi, $\mathbf{X}\underline{v} = \underline{0}$ et par injectivité de $\mathbf{X}$ (notre hypothèse), il vient finalement $\underline{v} = \underline{0}$, ce qui conclut la preuve de l'inversibilité de $\mathbf{X}^T\mathbf{X}$.

2.2. Estimation du vecteur $\underline{\beta}_0$ des coefficients par la méthode des moindres carrés. Comme dans la partie précédente, on considère l'estimateur $\widehat{\underline{\beta}}_n$ des moindres carrés. Avec les notations précédentes (norme[38] euclidienne $\|\cdot\|_2$), ce dernier est défini comme antécédent du minimum d'un écart quadratique :

$$\widehat{\underline{\beta}}_n \in \underset{\underline{b}\in\mathbb{R}^{k+1}}{\operatorname{argmin}} \left\| \underline{Y} - \mathbf{X}\,\underline{b} \right\|_2 .$$

PROPOSITION 14.1. *Sous l'hypothèse d'injectivité de* $\mathbf{X}$*, on a l'expression explicite suivante :*

$$\widehat{\underline{\beta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\underline{Y}.$$

ELÉMENTS CULTURELS. Voici la preuve de la proposition précédente. Vos souvenirs d'algèbre linéaire indiquent que la solution du problème de minimisation définissant l'estimateur est tout vecteur $\widehat{\underline{\beta}}_n$ tel que $\mathbf{X}\,\widehat{\underline{\beta}}_n$ est la projection orthogonale $\Pi_V\,\underline{Y}$ de $\underline{Y}$ sur $V = \operatorname{Im}\mathbf{X}$, l'image de $\mathbf{X}$. En particulier, le produit scalaire entre n'importe quel élément de $\operatorname{Im}\mathbf{X}$ et $\underline{Y} - \Pi_V\,\underline{Y} = \underline{Y} - \mathbf{X}\,\widehat{\underline{\beta}}_n$ est nul, ce que l'on peut ré-écrire, de manière matricielle par

$$\mathbf{X}^T\left(\underline{Y} - \mathbf{X}\,\widehat{\underline{\beta}}_n\right) = \underline{0}\ .$$

Cette égalité se retraduit par

$$\mathbf{X}^T\,\underline{Y} = \mathbf{X}^T\mathbf{X}\,\widehat{\underline{\beta}}_n\ , \qquad \text{soit encore} \quad \widehat{\underline{\beta}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\underline{Y}\ ,$$

où l'on a utilisé un fait prouvé précédemment, à savoir que $\mathbf{X}^T\mathbf{X}$ est inversible.

On conclut ce paragraphe culturel par une indication de propriété théorique de l'estimateur exhibé : il est sans biais.

PROPOSITION 14.2. *L'estimateur* $\widehat{\underline{\beta}}_n$ *est sans biais :* $\mathbb{E}\left[\widehat{\underline{\beta}}_n\right] = \underline{\beta}_0$.

En effet, par linéarité,

$$\mathbb{E}\left[\widehat{\underline{\beta}}_n\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbb{E}[\underline{Y}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\underline{\beta}_0 = \underline{\beta}_0\ ,$$

où l'on a également utilisé que par hypothèse du modèle, $\mathbb{E}[\underline{\varepsilon}] = \underline{0}$.

REMARQUE 14.1. On notera

$$\widehat{\underline{\beta}}_n = \begin{bmatrix} \widehat{\alpha}_n \\ \widehat{\beta}_{1,n} \\ \vdots \\ \widehat{\beta}_{k,n} \end{bmatrix}$$

les composantes de l'estimateur vectoriel $\widehat{\underline{\beta}}_n$ du vecteur $\underline{\beta}_0$.

38. On la définit comme : $\|\underline{x}\|_2 = \sqrt{x_1^2 + \ldots + x_n^2}$ pour tout $\underline{x} = (x_1, \ldots, x_n)$

2.3. Reconstruction et prédiction. Pour un nouveau vecteur de données explicatives $\underline{x}' = (x'_1, \ldots, x'_k)^T$, on utilise alors, pour prédire l'observation $y_{\underline{x}'}$, réalisation de la variable aléatoire $Y_{\underline{x}'}$, l'estimateur ponctuel suivant, construit à partir des données précédentes indexées de 1 à n :

$$\underline{\widehat{\beta}}_n \cdot \underline{x}' = \widehat{\alpha}_n + \widehat{\beta}_{1,n}\, x'_1 + \ldots + \widehat{\beta}_{k,n}\, x'_k \ .$$

Il existe évidemment des techniques pour obtenir des intervalles de confiance sur l'espérance $\mu_{\underline{x}'}$ de $Y_{\underline{x}'}$, d'une part, et des intervalles de prévision sur $Y_{\underline{x}'}$, d'autre part. (Nous ne les détaillons pas.)

On note, pour la suite de cette partie,

$$\widehat{Y}_t = \underline{\widehat{\beta}}_n \cdot \underline{x}_t \qquad \text{pour } t = 1, \ldots, n,$$

la reconstruction de y_t à partir des estimateurs des coefficients, où l'on a utilisé le vecteur $\underline{x}_t = (1, x_{1,t}, \ldots, x_{k,t})$, qui est la t–ème ligne de $\mathbf{X}$.

3. Lecture et interprétation de sorties SPSS

Maintenant que le décor est posé, on peut faire appel à notre vieil ami, SPSS ; nous allons décrire dans le reste de ce cours le contenu des tableaux de régression qu'il nous propose (voir la figure 69).

Régression sur le modèle complet

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,916[a]	,839	,813	138,034

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1793129,948	3	597709,983	31,370	,000[a]
	Résidu	342959,506	18	19053,306		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			intervalles de confiance à 95% pour B	
		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Limite supérieure
1	(Constante)	238,458	112,242		2,124	,048	2,646	474,270
	Radio	23,850	4,524	,749	5,272	,000	14,346	33,354
	Journaux	32,629	5,369	,585	6,078	,000	21,350	43,908
	Gratuits	-,619	10,228	-,009	-,060	,952	-22,107	20,870

a. Variable dépendante : Ventes

FIGURE 69. Tableaux SPSS de résultats de régression linéaire multiple sur les données de l'exemple 14.1.

Ce qu'on peut quand même dire immédiatement, c'est que la première colonne du troisième tableau, celle nommée B (ou A, selon les versions de SPSS), donne le vecteur des estimées correspondant à l'estimateur $\underline{\widehat{\beta}}_n$. (Ceci est à comparer au tableau similaire exhibé au cours précédent.)

EXEMPLE 14.2. Ici, j'espère que vous êtes surpris par l'estimée négative pour le coefficient devant la variable explicative du montant investi dans la publicité dans les journaux gratuits... En fait, on verra que cette estimée n'est pas significativement différente de 0 (ouf !) ; sinon, cela aurait voulu dire que la publicité peut être contre-productive, ce qui aurait été quand même assez étonnant...

L'essentiel du reste de ce cours sera structuré autour des trois tableaux présentés ci-dessus ; ainsi, nous étudierons :

- d'une part, les deux premiers tableaux : le tableau de résumé de la qualité de l'ajustement linéaire et celui s'intéressant à la validité globale du modèle linéaire multiple ;
- d'autre part, le troisième tableau, qui étudie la significativité individuelle de chaque variable explicative face aux autres variables.

3.1. Premier et deuxième tableaux : existence d'un ajustement linéaire significatif et qualité éventuelle de cet ajustement. Il s'agit d'étendre les définitions vues dans la partie précédente : coefficient de détermination r^2, estimateur de la variance σ_0^2 des résidus, ainsi que d'introduire un test de l'existence d'une relation linéaire significative, qui correspond à ce qu'au moins un des coefficients $\beta_{j,0}$ soit significativement différent de 0. C'est pourquoi on testera

$$H_0 : \beta_{1,0} = \beta_{2,0} = \ldots = \beta_{k,0} = 0$$

en espérant que le test rejette H_0, ce qui correspondra à l'existence d'une telle relation.

3.1.1. *Coefficient de détermination r^2 et estimateur de la variance des résidus.* On introduit, comme pour le modèle linéaire simple, les reconstructions des observations Y_t, à partir desquelles on estime les résidus ε_t :

$$\widehat{Y}_t = \underline{\widehat{\beta}}_n \cdot \underline{x}_t \qquad \text{et} \qquad \widehat{\varepsilon}_t = Y_t - \widehat{Y}_t \ , \qquad \text{pour } t = 1, \ldots, n,$$

où l'on a noté $\underline{x}_t = (1, x_{1,t}, \ldots, x_{k,t})$ la t–ème ligne de $\mathbf{X}$ (la t–ème série de variables explicatives). Les vecteurs colonnes de longueur n correspondants seront notés $\mathbf{X}\,\underline{\widehat{\beta}}_n = \underline{\widehat{Y}}$ et $\underline{\widehat{\varepsilon}}$.

Premier fait : la moyenne des reconstructions $\widehat{Y}_t$ est égale à la moyenne des observations Y_t :

$$\frac{1}{n}\sum_{t=1}^{n} \widehat{Y}_t = \frac{1}{n}\sum_{t=1}^{n} Y_t \overset{\text{not.}}{=} \overline{Y}_n \ .$$

En effet, par construction, le vecteur $\underline{\widehat{\varepsilon}} = \underline{Y} - \underline{\widehat{Y}}$ est orthogonal à l'image de $\mathbf{X}$, qui contient en particulier le vecteur $\underline{1} = (1, \ldots, 1)$.

On définit alors les sommes des carrés totale Σ_T, expliquée Σ_E et résiduelle Σ_R de la même manière que dans le cas de la régression simple (et en y mettant les mêmes interprétations, voir page 317) :

$$\Sigma_T \overset{\text{not.}}{=} \sum_{t=1}^{n} (Y_t - \overline{Y}_n)^2 \ , \qquad \Sigma_E \overset{\text{not.}}{=} \sum_{t=1}^{n} (\widehat{Y}_t - \overline{Y}_n)^2 \ , \qquad \text{et} \qquad \Sigma_R \overset{\text{not.}}{=} \sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2 \ .$$

Là encore, les notations matricielles permettent une ré-écriture plus compacte ; notant $\overline{\underline{Y}}$ le vecteur colonne $\overline{Y}_n\,\underline{1}$ (i.e., celui dont les n composantes sont égales à $\overline{Y}_n$), il vient :

$$\Sigma_T = \left\| \underline{Y} - \overline{\underline{Y}} \right\|_2^2 \ , \qquad \Sigma_E = \left\| \underline{\widehat{Y}} - \overline{\underline{Y}} \right\|_2^2 \ , \qquad \text{et} \qquad \Sigma_R \overset{\text{not.}}{=} \left\| \underline{Y} - \underline{\widehat{Y}} \right\|_2^2 = \left\| \underline{\widehat{\varepsilon}} \right\|_2^2 \ .$$

Par une égalité de Pythagore, au vu de la projection orthogonale effectuée [39], il vient la décomposition

$$\Sigma_T = \Sigma_E + \Sigma_R \ .$$

On conserve la définition du coefficient de détermination : $r^2 = \Sigma_E / \Sigma_T$. On généralise également le résultat du Théorème 13.3 de la manière suivante [40].

39. $\underline{\widehat{Y}}$ est la projection orthogonale de $\underline{Y}$ sur l'image de $\mathbf{X}$, qui contient le vecteur $\underline{1}$ et donc également $\overline{\underline{Y}}$

40. Désormais, on perd $k+1$ degrés de liberté dans la loi de Σ_R/σ_0^2, qui est donc une loi $\chi^2_{n-(k+1)}$, et ceci, pour une raison similaire à celle vue dans le cours précédent : on a en effet besoin d'estimer $k+1$ paramètres (les $k+1$ coefficients de la relation linéaire). La renormalisation par $1/\big(n-(k+1)\big)$ pour l'estimateur de la variance s'en déduit.

THÉORÈME 14.1. *Dans le modèle linéaire gaussien multiple, on dispose de l'estimateur de la variance σ_0^2 donné par*

$$\widehat{\sigma}_n^2 = \frac{1}{n-(k+1)} \sum_{t=1}^{n} (\widehat{\varepsilon}_t)^2 = \frac{\Sigma_R}{n-(k+1)} \ .$$

Il est sans biais et consistant.

On note que l'estimateur de la variance ne prend pas seulement en compte Σ_R mais également le nombre k de variables explicatives considérées. Cela n'est pas le cas du coefficient de détermination r^2, puisque

$$r^2 = 1 - \frac{\Sigma_R}{\Sigma_T} \ ;$$

comme il est plus juste de rapporter la qualité de l'ajustement au nombre de variables utilisées, on cherche une version, que l'on dira corrigée ou ajustée, de r^2 et que l'on notera r^2_{ajust}. Les contraintes que l'on s'impose sont

- qu'elle soit d'autant plus grande que $\widehat{\sigma}_n^2$ soit petit (i.e., qu'elle soit une fonction décroissante de $\widehat{\sigma}_n^2$) ;
- qu'elle soit nulle dans le cas de $k = 0$ variable explicative (r^2 étant également nul dans ce cas : il n'y que des résidus et rien n'est expliqué) ;
- qu'elle ait une expression simple.

La solution retenue par la littérature et la pratique statistiques est

$$r^2_{\text{ajust}} = 1 - \frac{n-1}{n-(k+1)}(1-r^2) = 1 - \frac{(n-1)\,\widehat{\sigma}_n^2}{\Sigma_T} \ .$$

3.1.2. *Lecture du premier tableau.* Nous avons maintenant tous les outils pour lire le premier tableau de la figure 69 : il est donné par la réalisation du tableau suivant.

R	R-deux	R-deux ajusté	Erreur standard
$\sqrt{r^2}$	r^2	r^2_{ajust}	$\sqrt{\widehat{\sigma}_n^2}$

Exploitation (en général). La qualité de l'ajustement sur l'échantillon donné par un modèle fixé est évaluée par la valeur réalisée de r^2. En revanche, pour comparer les performances de différents modèles, on recourt aux estimées de l'écart-type des résidus et du r^2_{ajust}, car elles seules [41] tiennent compte du nombre de variables (entre deux modèles, on choisit celui de plus fort r^2_{ajust} réalisé ou, de manière équivalente, celui de plus faible $\sqrt{\widehat{\sigma}_n^2}$ réalisé).

Exploitation (sur notre exemple). Nous ne nous intéressons pour l'instant qu'à un seul modèle, le modèle complet. On notera simplement que la valeur réalisée de son coefficient de détermination r^2 est importante, elle vaut en effet 83.9 % (valeur ajustée : 81.3 %).

41. On rappelle que r^2 augmente mécaniquement lorsque le nombre de variables explicatives augmente ; ce n'est pas le cas de sa version ajustée.

3.1.3. *Test de la validité globale d'un modèle linéaire.* On veut tester ici si le modèle linéaire apporte une contribution significative à l'explication statistique. C'est le cas si au moins une variable est significative dans la régression linéaire, i.e., si dans le test d'hypothèses

$$H_0 : \beta_{1,0} = \ldots = \beta_{k,0} = 0 \qquad \text{vs.} \qquad H_1 : \text{il existe } j \text{ tel que } \beta_{j,0} \neq 0\,,$$

on peut rejeter H_0.

On avait vu dans la partie précédente que l'on pouvait considérer à cet effet un test fondé sur une statistique notée D_n et définie à partir du coefficient de détermination r^2 (ou, de manière équivalente, à partir du rapport entre la somme des carrés expliquée Σ_E et la somme des carrés résiduelle Σ_R).

On généralise sa définition en tenant compte, là aussi, du nombre de variables explicatives :

$$D_n = \frac{\dfrac{1}{k}\displaystyle\sum_{t=1}^{n}(\widehat{Y}_t - \overline{Y}_n)^2}{\dfrac{1}{n-(k+1)}\displaystyle\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2} = \frac{\Sigma_E \,/\, k}{\Sigma_R \,/\, (n-(k+1))} = \frac{n-(k+1)}{k}\,\frac{r^2}{1-r^2}\,.$$

(En fait, D_n est donc une fonction croissante de r^2.) Répétons-le : on divise par k le numérateur pour compenser la forcément meilleure explication du phénomène par k variables plutôt que par une seule.

Sous H_0, la statistique F suit une loi dite de Fisher $\mathcal{F}(k, n-(k+1))$, et sous H_1, puisqu'alors il existe une relation linéaire, elle tend à prendre des valeurs grandes : Σ_E est plus grande et Σ_R est plus petite. On en déduit un principe de test fondé sur D_n (et utilisant une zone de rejet unilatère), que l'on ne détaille cependant pas davantage.

3.1.4. *Lecture du deuxième tableau.* Nous avons maintenant tous les outils pour lire le deuxième tableau de la figure 69 : il est donné par la réalisation du tableau suivant.

Somme carrés	ddl	Moyenne carrés	D	Sig.
Σ_E	k	$\Sigma_E \,/\, k$	D_n	P-val. test $H_0 : \forall\, j,\ \beta_{j,0} = 0$
Σ_R	$n-(k+1)$	$\Sigma_R \,/\, (n-(k+1))$		
Σ_T	$n-1$			

Exploitation (en général). On regarde essentiellement la P-valeur située dans la dernière colonne. Si elle est plus petite que 5 %, et dans ce cas seulement, on déduit qu'il existe une relation linéaire entre les variables à expliquer et les variables explicatives. Si la P-valeur est plus grande que 5 %, on partira du principe qu'il n'y a pas de relation *linéaire* significative ; cela ne veut cependant pas dire qu'il n'y a pas de relation du tout entre les variables à expliquer et les variables explicatives, cette dernière pouvant être non

linéaire [42].

Exploitation (sur notre exemple). Ici, tout va bien, la P-valeur est quasi-nulle, on en déduit avec force l'existence d'une relation linéaire et on peut passer à la lecture du troisième tableau.

3.2. Troisième tableau : relation proposée, test des significativités individuelles des variables. Ici, il s'agit de tester qu'une variable explicative donnée, disons, la j–ème, apporte une contribution significative à l'explication linéaire *face aux autres variables* ; c'est-à-dire que l'on teste

$$H_0 : \beta_{j,0} = 0 \qquad \text{vs.} \qquad H_1 : \beta_{j,0} \neq 0 ;$$

lorsque H_0 est conservée, la variable j n'apporte pas de telle contribution significative.

REMARQUE 14.2 (Attention ! Le test est face aux autres variables.). Dans le test précédent, même lorsque H_0 est conservée, la j–ème variable peut parfois malgré tout apporter une explication linéaire significative à la variable à expliquer, lorsqu'elle est considérée isolément ou en présence de moins de variables explicatives ; c'est simplement qu'en présence de toutes les autres variables explicatives, elle n'est plus utile. C'est généralement parce qu'elle est redondante avec une ou plusieurs autre(s) variable(s) explicative(s).

On commence là encore par un peu de théorie : il s'agit de quantifier quand l'estimée du coefficient $\beta_{j,0}$ est significativement différente de 0. A cet effet, il nous faut déterminer la loi des estimateurs en jeu. Chaque coefficient de

$$\underline{\widehat{\beta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underline{Y}$$

étant combinaison linéaire d'observations normales indépendantes, il suit une loi normale. Le vecteur $\underline{\widehat{\beta}}_n$ forme même ce qu'on appelle un vecteur gaussien. Les règles de manipulation des vecteurs gaussiens permettent d'arriver facilement aux résultats décrits ci-après, que nous admettrons. On part de

$$\widehat{\alpha}_n \sim \mathcal{N}\left(\alpha_0, \sigma_0^2 (\mathbf{X}^T\mathbf{X})^{-1}_{1,1}\right) \qquad \text{et} \qquad \widehat{\beta}_{p,n} \sim \mathcal{N}\left(\beta_{p,0}, \sigma_0^2 (\mathbf{X}^T\mathbf{X})^{-1}_{p+1,p+1}\right)$$

(pour tout $p = 1, \ldots, k$), puis, par standardisation, on parvient au théorème suivant.

THÉORÈME 14.2. *Dans le modèle linéaire gaussien multiple,*

$$T'_n = \frac{\widehat{\alpha}_n - \alpha_0}{\sqrt{\widehat{\sigma}_n^2 (\mathbf{X}^T\mathbf{X})^{-1}_{1,1}}} \sim \mathcal{T}_{n-(k+1)} \qquad \textit{et} \qquad T_{p,n} = \frac{\widehat{\beta}_{p,n} - \beta_{p,0}}{\sqrt{\widehat{\sigma}_n^2 (\mathbf{X}^T\mathbf{X})^{-1}_{p+1,p+1}}} \sim \mathcal{T}_{n-(k+1)}$$

pour tout $p = 1, \ldots, k$.

On en déduit des intervalles de confiance (précisés dans le tableau 3) et des tests à une valeur de référence. Le troisième tableau de la figure 69 reporte justement les valeurs réalisées de ces résultats (avec pour les tests, la comparaison à la valeur de référence 0, i.e., le fait que le coefficient correspondant, α_0 ou $\beta_{j,0}$, est nul). On dit qu'il présente une

42. Ainsi, si pour des raisons externes (par exemple, économiques), on est certain qu'il existe une telle relation non linéaire, on effectuera au préalable une transformation non linéaire des variables explicatives (choisie en fonction des tracés dans $\mathbb{R}^{k+1}$ des variables à expliquer en fonctions des explicatives) et on retentera une régression linéaire à partir de ces explicatives transformées.

Coeff.	Err. standard	[Ignorée]	t	Sig.	IC(α) – bornes inf. et sup.
$\widehat{\alpha}_n$	$\sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{1,1}}$		T'_n	P-val. $H_0 : \alpha_0 = 0$	$\left[\widehat{\alpha}_n \pm t_{n-(k+1),1-\alpha/2} \sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{1,1}}\right]$
$\widehat{\beta}_{1,n}$	$\sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{2,2}}$		$T_{1,n}$	P-val. $H_0 : \beta_{1,0} = 0$	$\left[\widehat{\beta}_{1,n} \pm t_{n-(k+1),1-\alpha/2} \sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{2,2}}\right]$
...	...		...	...	
...	...		...	...	
$\widehat{\beta}_{k,n}$	$\sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{k+1,k+1}}$		$T_{k,n}$	P-val. $H_0 : \beta_{k,0} = 0$	$\left[\widehat{\beta}_{k,n} \pm t_{n-(k+1),1-\alpha/2} \sqrt{\widehat{\sigma}_n^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{k+1,k+1}}\right]$

TABLE 3. Décryptage mathématique du troisième tableau de la figure 69 : estimées des coefficients, tests de significativité individuelle face aux autres variables explicatives, intervalles de confiance.

étude des validités marginales.

Exploitation (en général). On regarde la colonne des P-valeurs. Les variables à qui correspond une P-valeur plus grande que 5 % ne sont pas individuellement significatives face aux autres variables aléatoires, les autres le sont. Attention, une variable indiquée comme non individuellement significative face aux autres peut quand même contribuer à l'explication :

- soit sa contribution à l'explication serait en fait non linéaire et une transformation non linéaire préalable s'impose ;
- soit, ainsi qu'on l'a déjà évoqué plus haut, elle est fortement corrélée à une ou plusieurs autres variables et lorsque ces dernières sont présentes, cette variable est alors inutile ; réciproquement, il est alors souvent le cas qu'une de ces variables à qui elle est corrélée soit également, à cause d'elle, marquée comme non individuellement significative face aux autres variables.

Dans tous les cas, du point de vue statistique, on ne peut laisser en l'état un modèle dans lequel toutes les variables ne sont pas individuellement significatives. Il faut supprimer au moins une variable non significative (cas de corrélations) et/ou effectuer une transformation non linéaire sur une autre (cas où l'on sait pour des raisons extra-statistiques que la variable est quand même individuellement significative) et relancer la régression.

A noter : ainsi, l'ajout ou la suppression d'une variable explicative peut changer, dans un sens comme dans l'autre, le caractère individuellement significatif des variables déjà ou encore présentes.

Exploitation (sur notre exemple). La variable explicative `Gratuits` (qui donne le montant des budgets consacrés à la diffusion de catalogues gratuits) n'est pas individuellement significative ; ici, on a d'autant plus envie de la supprimer de la liste des variables explicatives que l'estimée de son coefficient est négative, ce qui est quand même difficilement interprétable économiquement ! Il faudrait relancer la régression avec uniquement les variables explicatives `Radio` et `Journaux`. (On le fait plus loin.)

REMARQUE 14.3 (Une erreur trop souvent commise !). La valeur absolue de l'estimée du coefficient de régression correspondant à une variable explicative donnée n'est absolument pas une indication de son caractère individuellement significatif ! En particulier, que sa valeur soit petite peut simplement être le fait de l'échelle des mesures (cf. les facteurs multiplicatifs selon que les variables explicatives sont données dans telle ou telle unité, ici, en milliers d'euros, mais elles auraient pu l'être en euros ou centaines d'euros). Pour voir si un coefficient est significatif ou non du point de vue de la régression linéaire multiple, il n'y a pas le choix, il faut conduire un test. Ce dernier compare en fait la valeur du coefficient estimé à la valeur reportée dans la colonne `Erreur standard`, ce qui permet de quantifier les déviations naturelles autour de 0 de celles qui indiquent que le coefficient est significatif (non nul).

Dans le même genre d'idées et pour les mêmes raisons, on se gardera de hiérarchiser l'importance des variables explicatives en fonction des estimées des coefficients.

3.3. Un autre exemple. Je vous propose de jeter un œil très rapide à l'exemple suivant, issu de l'examen principal de l'année 2009–10 et qui sera étudié bien plus en

détails (notamment en ce qui concerne les interprétations économiques et stratégiques à formuler) lors de la correction du dit examen.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,897[a]	,804	,794	17,249

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	112433,232	5	22486,646	75,578	,000[a]
	Résidu	27372,656	92	297,529		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	42,736	8,603		4,968	,000
	Altitude station	-,002	,006	-,020	-,332	,741
	Nombre de remontées	-,004	,060	-,004	-,064	,949
	Altitude pistes	,023	,005	,323	4,354	,000
	Nombre de pistes	,418	,091	,364	4,586	,000
	Nombre de lits	,001	,000	,386	6,020	,000

a. Variable dépendante : Prix forfait semaine

FIGURE 70. Tableaux SPSS de résultats de régression linéaire multiple sur un nouveau jeu de données.

On lit tout d'abord, grâce aux notes de bas de tableau, qu'il s'agit d'établir une relation linéaire entre le prix du forfait semaine (variable à expliquer) et les variables explicatives suivantes, décrivant chaque station : sa capacité de logement (le nombre de lits), son altitude, le nombre de remontées, l'altitude et le nombre de ses pistes.

L'explication linéaire par ces variables est significative : on lit, dans le second tableau, une P-valeur quasi-nulle pour le test de validité globale. Cela est confirmé par la valeur élevée (80.4 %) du coefficient de détermination r^2, lu dans le premier tableau. On continue donc l'étude et on s'intéresse au troisième tableau, qui présente l'étude des validités marginales. Or, deux variables, l'altitude de la station et le nombre de remontées, sont déclarées non individuellement significatives : la P-valeur des tests de nullité de leur coefficient est élevée, bien plus grande que 5 % (respectivement, 74.1 % et 94.9 %).

Le modèle contient donc au moins une variable non significative et ne peut être considéré en l'état comme bon pour l'analyse ; il faut le simplifier. A cet effet, il faut relancer l'analyse en supprimant une des deux variables non significatives, de préférence la moins significative en l'état : le nombre de remontées. Il faudra refaire l'étude dans le modèle à

quatre variables ainsi obtenu ; il se pourrait que toutes les variables y soient alors significatives [43] ou pas [44].

En fait, le caractère non significatif de ces deux variables est heureux, car elles sont associées à des estimées de coefficients négatives et qui auraient été difficilement interprétables : on pense plutôt que le forfait est d'autant plus cher que la station est haute et/ou qu'il y a de remontées...

REMARQUE 14.4 (Rappel : le caractère significatif se lit uniquement sur les P-valeurs...).
... et non sur les valeurs des coefficients. Ainsi, le coefficient estimé 0.001 devant le nombre de lits est-il significativement non nul tandis que ceux pourtant plus grands correspondant à l'altitude de la station ou au nombre de remontées ne le sont pas. C'est souvent une question d'échelles de valeurs et d'unités : chaque variable est associée à sa propre échelle (le nombre de remontées est de l'ordre de 10, l'altitude est mesurée en mètres et est donc de l'ordre de deux milliers, etc.) ; celle-ci transparaît dans la colonnes `Erreur standard`, qui mesure la précision de l'estimation des coefficients.

43. Par exemple, si les caractères non significatifs dans le modèle complet provenaient uniquement de la redondance entre les variables d'altitude de la station et de nombre de remontées entre elles

44. S'il y avait redondance au moins partielle avec une autre variable, par exemple, l'altitude des pistes.

4. Comparaison et choix de modèles linéaires ; procédures de sélection de variables explicatives

Ce n'est pas une tâche aisée que de comparer des modèles ni d'en choisir un. Lorsque k variables explicatives sont disponibles, $2^k - 1$ modèles avec au moins une variable explicative sont disponibles, ce qui peut faire beaucoup ! On cherche donc des méthodes automatiques de constructions de modèles satisfaisants (qui ne seront cependant pas nécessairement des modèles optimaux).

On a également des contraintes d'interprétabilité : on ne veut généralement pas considérer trop de variables explicatives mais un petit nombre (en pratique, au plus une variable explicative pour six données à expliquer). Il faut donc effectuer également un compromis entre un nombre à la fois suffisant de variables explicatives (pour avoir une bonne valeur réalisée de r^2) et pas trop grand (afin que la relation exhibée reste interprétable facilement).

Nous présenterons les différentes méthodes de comparaison et choix dans l'ordre suivant :

1. comment choisir entre plusieurs modèles possédant le même nombre $k \geqslant 1$ de variables explicatives,
2. avec comme cas particulier du cas précédent, la détermination de la meilleure variable explicative ;
3. comment comparer deux modèles reposant sur des nombres différents de variables explicatives ;
4. une première méthode automatique : la méthode de sélection "backward" ;
5. une seconde méthode automatique : la méthode "forward".

Toutes ces techniques reposent essentiellement sur des considérations statistiques, mais avec la réserve que ces dernières peuvent toujours être nuancées par des considérations extra-statistiques, notamment économiques.

4.1. Choix à nombre $k \geqslant 1$ de variables explicatives fixées. A nombre k de variables fixées, on retient le modèle avec le meilleur r^2 (i.e., avec la plus petite estimée de l'écart-type des résidus).

Ainsi, dans l'exemple 14.1 (figure 68), si on veut considérer deux variables, il faut calculer les régressions sur les trois couples de variables explicatives possibles et ne conserver que celui conduisant au plus grand r^2. Le résultat est reporté aux figures 71 et 72. Il s'agit en l'occurrence du couple de variables `Radio` et `Journaux`.

Remarque 14.5 (Considérations extra-statistiques). Cela étant, entre deux modèles avec des valeurs réalisées du coefficient r^2 proches, on pourrait retenir le modèle avec la valeur légèrement plus faible si, du point de vue économique, les variables considérées nous semblent plus raisonnables.

4.2. Détermination de la meilleure variable explicative prise isolément. Dans le cas $k = 1$, il est facile d'obtenir une réponse rapidement sous SPSS sans calculer toutes les régressions simples possibles. La figure 73 illustre cela.

La minute SPPS 14.1. La figure 73 a été obtenue par `Analyse / Correlation / Bivariée`.

Régression Ventes / Radio & Journaux

Variables

Modèle	Variables introduites	Variables supprimées	Méthode
1	Radio, Journaux[a]	.	Entrée

a. Toutes variables requises saisies.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,916[a]	,839	,823	134,366

a. Valeurs prédites : (constantes), Radio, Journaux

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1793060,221	2	896530,111	49,658	,000[a]
	Résidu	343029,233	19	18054,170		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Radio, Journaux

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	235,168	95,577		2,461	,024
	Journaux	32,571	5,140	,584	6,337	,000
	Radio	23,646	2,935	,742	8,058	,000

a. Variable dépendante : Ventes

Régression Ventes / Radio & Gratuits

Variables introduites/supprimées

Modèle	Variables introduites	Variables supprimées	Méthode
1	Gratuits, Radio[a]	.	Entrée

a. Toutes variables requises saisies.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,714[a]	,510	,458	234,720

a. Valeurs prédites : (constantes), Gratuits, Radio

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1089310,991	2	544655,496	9,886	,001[a]
	Résidu	1046778,463	19	55093,603		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits, Radio

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	607,838	160,463		3,788	,001
	Radio	19,067	7,575	,598	2,517	,021
	Gratuits	10,609	17,106	,147	,620	,543

a. Variable dépendante : Ventes

FIGURE 71. Deux régressions avec deux couples de variables explicatives, sur les données de l'exemple 14.1.

Régression Ventes / Journaux & Gratuits

Variables introduites/supprimées

Modèle	Variables introduites	Variables supprimées	Méthode
1	Gratuits, Journaux[a]	.	Entrée

a. Toutes variables requises saisies.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,769[a]	,592	,549	214,298

a. Valeurs prédites : (constantes), Gratuits, Journaux

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1263536,381	2	631768,191	13,757	,000[a]
	Résidu	872553,073	19	45923,846		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits, Journaux

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	250,869	174,218		1,440	,166
	Journaux	27,706	8,208	,496	3,376	,003
	Gratuits	39,586	10,582	,550	3,741	,001

a. Variable dépendante : Ventes

FIGURE 72. Une troisième régression avec un autre couple de variables explicatives, sur les données de l'exemple 14.1.

Elle représente le tableau des corrélations empiriques. On ne va pas définir ces dernières de manière intrinsèque, faute de temps, mais on se contentera de noter (cf. les régressions simples reportées sous le tableau croisé des corrélations) que ces dernières sont égales chacune à $\pm\sqrt{r^2}$, où r^2 est la valeur du coefficient de détermination associé à la régression simple correspondante. (Le signe est en fait celui de l'estimée du coefficient de régression $\widehat{\beta}_{p,n}$ associé.) La P-valeur du test de validité de la régression simple correspondante est également reportée dans le tableau croisé des corrélations.

Comme le choix de la meilleure variable explicative est celui maximisant le r^2, ou, de manière équivalente, maximisant la valeur absolue de la corrélation, on le lit directement sur le tableau croisé des corrélations, sans avoir besoin [45] d'effectuer toutes les régression simples.

Ici, c'est la variable `Radio` qui procure la meilleure explication linéaire.

4.3. Comparaison entre deux modèles donnés, fondés sur des nombres différents de variables explicatives. Nous avons vu précédemment que dans ce cas, sauf considérations extra-statistiques (économiques par exemple), on comparait, de manière équivalente, les valeurs réalisées des r^2_{ajust} ou les estimées des écarts-types des résidus. Le modèle avec la plus grande valeur réalisée de r^2_{ajust}, i.e., la plus petite estimée de l'écart-type, l'emporte.

45. Nous ne les avons toutes calculées et reportées ici que dans le but, espéré pédagogique, de comparer les résultats synthétiques du tableau croisé à ceux des régressions simples.

Corrélations

Corrélations

		Ventes	Radio	Journaux	Gratuits
Ventes	Corrélation de Pearson	1	,707**	,539*	,589**
	Sig. (bilatérale)		,000	,010	,004
	N	22	22	22	22
Radio	Corrélation de Pearson	,707**	1	-,060	,737**
	Sig. (bilatérale)	,000		,791	,000
	N	22	22	22	22
Journaux	Corrélation de Pearson	,539**	-,060	1	,078
	Sig. (bilatérale)	,010	,791		,731
	N	22	22	22	22
Gratuits	Corrélation de Pearson	,589**	,737**	,078	1
	Sig. (bilatérale)	,004	,000	,731	
	N	22	22	22	22

Ventes / Radio

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,707[a]	,500	,475	231,081

a. Valeurs prédites : (constantes), Radio

Ventes / Radio[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1068122,070	1	1068122,070	20,003	,000[a]
	Résidu	1067967,384	20	53398,369		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Radio

b. Variable dépendante : Ventes

Ventes / Journaux

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,539[a]	,291	,255	275,247

a. Valeurs prédites : (constantes), Journaux

Ventes / Journaux[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	620874,436	1	620874,436	8,195	,010[a]
	Résidu	1515215,019	20	75760,751		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Journaux

b. Variable dépendante : Ventes

Ventes / Gratuits

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,589[a]	,347	,314	264,181

a. Valeurs prédites : (constantes), Gratuits

Ventes / Gratuits[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	740256,945	1	740256,945	10,607	,004[a]
	Résidu	1395832,509	20	69791,625		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits

b. Variable dépendante : Ventes

FIGURE 73. Tableau croisé des corrélations sur les données de l'exemple 14.1 (et, pour comparaison, régressions simples possibles des `Ventes` sur les variables explicatives possibles).

4.4. Première procédure automatique : méthode "backward". On rappelle qu'*a priori*, il y a $2^k - 1$ modèles possibles : on cherche donc une procédure automatique de sélection (qui n'ait pas besoin d'être supervisée par un être humain).

Une première manière simple de procéder est considérer toutes les variables initialement, d'effectuer la régression, puis d'éliminer la moins significative lorsqu'une des variables est effectivement non individuellement significative ; dans ce cas, on élimine celle pour qui le test de significativité a la plus grande P-valeur, ou, de manière équivalente, dont la valeur observée pour la statistique de Student est la plus petite en valeur absolue.

On continue ainsi la procédure, jusqu'à tomber sur un modèle dans lequel toutes les variables sont individuellement significatives.

EXEMPLE 14.3. Dans notre exemple, en lisant la régression complète de la figure 69, on se rend compte que la variable `Gratuits` est la seule non significative ; on l'enlève et on relance une analyse de régression linéaire. Dans le modèle à deux variables alors obtenu, toutes les variables sont significatives (voir la figure 71), de sorte que la procédure s'arrête ici et recommande donc finalement le modèle construit sur `Radio` et `Journaux`.

4.5. Seconde procédure automatique : méthode "forward". Elle ajoute des variables séquentiellement : elle considère d'abord la meilleure variable explicative pour la régression simple (cf. ci-dessus) ; puis, une fois qu'on a déjà k variables, elle ajoute la variable qui est telle que le modèle à $k + 1$ variables a le meilleur r^2 parmi l'ensemble des modèles qu'on aurait pu proposer et dont toutes les variables sont significatives. On s'arrête lorsque l'on ne peut plus ajouter de variables qui soient significatives dans le modèle étendu.

EXEMPLE 14.4. Dans notre exemple, nous avons déjà vu que la meilleure variable explicative est `Radio`. On regarde ensuite les couples `Radio` et `Journaux` d'une part, `Radio` et `Gratuits` d'autre part (voir la figure 71). Le second modèle n'est pas admissible (variable `Gratuits` non significative) mais le premier modèle l'est, c'est celui-ci qui forme le modèle retenu à la seconde itération. Or, il n'est dès lors plus possible d'ajouter de variable qui reste significative dans le modèle étendu (cf. `Gratuits` à la figure 69). La procédure s'arrête ici et recommande donc finalement le modèle construit sur `Radio` et `Journaux`.

REMARQUE 14.6. Note : les procédures "backward" et "forward" ne conduisent pas toujours à la même recommendation ! Pire, il existe en fait une troisième procédure : la procédure "stepwise". C'est un mélange entre "backward" et "forward" ; elle tient compte du fait que dans "forward", l'ajout d'une variable peut rendre non significatives des variables précédemment ajoutées (et dont le statut n'est plus jamais remis en cause une fois qu'elles sont entrées dans le modèle), et que dans backward, on peut enlever des variables qui auraient été significatives dans le modèle simplifié finalement obtenu (et dont le statut n'est jamais remis en cause non plus une fois qu'elles ont été supprimées du modèle). Nous ne la détaillons[46] pas.

46. Pas cette année en tout cas. Peut-être sera-ce le cas pour vos camarades plus chanceux de l'année prochaine ?

5. Variables explicatives qualitatives

5.1. Considération d'une unique variable qualitative, à valeurs binaires. Lorsque l'une, et l'une seulement des variables explicatives, disons la k–ème, prend uniquement les valeurs 0 et 1 (selon l'absence ou la présence d'une modalité), le coefficient correspondant $\beta_{k,0}$ est à interpréter comme un terme correctif additif à apporter en moyenne sur la valeur de la variable à expliquer lorsque la modalité considérée est présente.

Exemple 14.5 (Forfait de ski en fonction du massif). Pour modéliser le prix du forfait semaine, on aurait pu considérer une variable $x_{k,t}$ indiquant si oui (valeur 1) ou non (valeur 0) la station est alpine. L'estimée de $\beta_{k,0}$ indiquerait le tarif moyen supplémentaire éventuel des stations alpines par rapport aux stations pyrénéennes, vosgiennes, du Massif Central, etc. ; je dis bien éventuel car on aurait alors pu tester si ce coefficient est significativement non nul ou pas.

5.2. Autres cas. Les cas où
- il y a au moins une variable qualitative avec plus de trois valeurs possibles,
- il y a au moins deux variables qualitatives, qu'elles prennent des valeurs binaires ou non,
- l'on veut modéliser un impact plus fin qu'un terme correctif additif (par exemple, un changement de coefficient de pente selon la valeur de la variable qualitative),

sont plus compliqués et ne sont pas à notre programme cette année. ... Peut-être l'an prochain ?

Compléments pour étudiants avancés

Pour une fois... il n'y en a pas !

Exercices

Je vous propose trois exercices (les deux premiers étant assez longs). Les énoncés sont disponibles ci-dessous ; tous mettent en jeu des fichiers de données disponibles sur le site web du cours. Ma correction proposera les réponses aux questions posées, de même que [47] l'ensemble des sorties SPSS.

Deux exercices de synthèse

EXERCICE 14.1 (Source : Jacques Obadia, professeur honoraire HEC Paris). Cet exercice est à faire avec le fichier de données `Pompes.sav` disponible sur le site web du cours.

La direction marketing d'un distributeur d'essence souhaite établir un modèle expliquant les ventes de ses stations services situées dans les grands centres urbains. Le tableau de données précise, pour 45 stations de ce type, les informations suivantes :

- les ventes de la station (exprimées en milliers de litres),
- le nombre de pompes de la station,
- le nombre de concurrents dans la zone desservie par la station,
- le trafic quotidien (exprimé en milliers de voitures).

Jetez un œil aux données et effectuez sous SPSS les différentes régressions de la variable à expliquer (les volumes de ventes) en fonction d'une, deux ou trois variables explicatives. Répondez ensuite aux questions suivantes :

- Expliquez pourquoi le modèle à trois variables ne peut être retenu. Quelle variable faut-il éliminer de ce modèle pour obtenir un modèle à deux variables ?
- Le modèle proposé à la question précédente est-il valide statistiquement ?
- Commentez et/ou interprétez la relation proposée par ce modèle ; notamment : que pensez-vous de la validité économique du modèle, i.e., permet-il de mieux comprendre la réalité ?
- Pourquoi le modèle à deux variables défini à la première question et étudié à la deuxième est-il préférable aux modèles à une seule variable ?

EXERCICE 14.2 (Source : Gilles Mauffrey, professeur HEC Paris). Cet exercice est à faire avec le fichier de données `Elec.sav` disponible sur le site web du cours.

Une entreprise souhaite évaluer l'importance relative de l'influence de ses vendeurs et de ses prix sur ses ventes. A cet effet, elle a réparti ses clients en un certain nombre de zones géographiques. Pour chacune d'entre elles, les variables suivantes ont été mesurées :

- les ventes (nombre de ventes, en centaines),

47. Et ce, pour la facilité de lecture du corrigé dans les transports en commun ou tous ces moments où l'on est seul et où l'on a rien de mieux à faire que réviser ses cours, par opposition aux temps utiles où l'on est assis devant son ordinateur mais surfe sur FaceBook.

- le nombre de vendeurs,
- la moyenne des prix facturés par l'entreprise (en euros, par produit),
- la moyenne des prix facturés par la concurrence (en euros, par produit),
- l'indice des prix à la consommation (base 100).

L'indice 100 des prix est l'indice de la zone de plus bas prix (la région Auvergne).

Jetez un œil aux données et effectuez sous SPSS les différentes régressions de la variable à expliquer (le nombre de ventes) en fonction d'une, deux, trois ou quatre variables explicatives. Répondez ensuite aux questions suivantes :

- Quelle est la meilleure variable explicative prise isolément ?
- Quel est le meilleur couple de variables explicatives ?
- Que pensez-vous des résultats de la régression complète (avec les quatre variables explicatives) ou de celles avec trois variables explicatives ?
- Comment expliquez-vous que certaines variables significatives dans un modèle à deux variables ne le soient plus dans un modèle à trois ou quatre variables ?
- Combien de variables explicatives est-il nécessaire de considérer ici, d'après vous : une, deux, trois, quatre ? Lesquelles ? Quel est le modèle statistique que vous recommanderiez finalement ? Par quelle(s) procédure(s) automatique(s) l'obtient-on également ?
- Le modèle ainsi obtenu vous semble-t-il raisonnable du point de vue de la modélisation économique ? Quelle relation proposez-vous alors, finalement ? Enfin, donnez une idée de l'exploitation stratégique possible du modèle ainsi construit.

Un exercice issu des annales

EXERCICE 14.3 (Salaires en sortie d'école de commerce). Répondez aux questions de l'exercice IV de l'examen principal 2008. Je vous ai mis les données utilisées dans le fichier `BS.sav` disponible sur le site web du cours. Question subsidiaire : quel modèle recommandent les procédures de sélection séquentielle "backward" et "forward" ?

Correction 14.1 (Pompes à essence). On commence, pour ceux qui ont la flemme d'ouvrir SPSS, par préciser les données :

Ventes	Pompes	Concurrents	Trafic	Ventes	Pompes	Concurrents	Trafic
203	4	4	13	249	13	15	19
262	18	21	18	242	9	9	23
247	16	19	10	220	6	7	11
239	11	12	15	268	21	24	16
241	10	11	19	225	9	11	8
217	4	4	12	240	10	11	18
224	8	10	9	234	9	10	16
249	15	17	14	226	9	11	8
242	10	11	21	243	13	15	14
268	20	23	18	222	4	4	17
233	10	12	12	230	7	7	17
250	16	19	12	230	8	9	14
236	10	11	14	223	3	2	20
231	9	10	12	255	14	15	22
254	17	20	14	245	11	12	21
238	12	14	11	239	7	7	25
242	11	12	17	239	10	11	18
241	9	9	22	257	14	15	24
260	16	18	22	224	7	8	12
282	21	23	28	247	12	13	20
258	17	19	17	232	9	10	13
265	20	23	16	235	8	8	19
249	14	16	17				

De même, on propose aux pages suivantes les 7 régressions possibles des ventes en fonction de 1, 2 ou 3 explicatives parmi le nombre de pompes (Pompes), le nombre de concurrents (Concurrents) et le trafic quotidien (Trafic).

Ventes / Pompes

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,934[a]	,872	,869	5,621

a. Valeurs prédites : (constantes), Pompes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	9257,688	1	9257,688	293,003	,000[a]
	Résidu	1358,624	43	31,596		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Pompes

b. Variable dépendante : Ventes

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	205,802	2,234		92,135	,000
	Pompes	3,121	,182	,934	17,117	,000

a. Variable dépendante : Ventes

Ventes / Concurrents

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,899[a]	,808	,804	6,879

a. Valeurs prédites : (constantes), Concurrents

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	8581,634	1	8581,634	181,361	,000[a]
	Résidu	2034,677	43	47,318		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Concurrents

b. Variable dépendante : Ventes

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	209,116	2,597		80,529	,000
	Concurrents	2,528	,188	,899	13,467	,000

a. Variable dépendante : Ventes

Ventes / Trafic

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,520[a]	,270	,253	13,423

a. Valeurs prédites : (constantes), Trafic

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2868,934	1	2868,934	15,923	,000[a]
	Résidu	7747,377	43	180,172		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Trafic

b. Variable dépendante : Ventes

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	212,757	7,414		28,696	,000
	Trafic	1,737	,435	,520	3,990	,000

a. Variable dépendante : Ventes

Ventes / Pompes & Concurrents

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,976[a]	,952	,950	3,477

a. Valeurs prédites : (constantes), Concurrents, Pompes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	10108,449	2	5054,225	417,983	,000[a]
	Résidu	507,862	42	12,092		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Concurrents, Pompes

b. Variable dépendante : Ventes

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	200,360	1,527		131,250	,000
	Pompes	12,115	1,078	3,625	11,237	,000
	Concurrents	-7,607	,907	-2,706	-8,388	,000

a. Variable dépendante : Ventes

Ventes / Pompes & Trafic

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,990[a]	,981	,980	2,188

a. Valeurs prédites : (constantes), Trafic, Pompes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	10415,294	2	5207,647	1088,070	,000[a]
	Résidu	201,017	42	4,786		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Trafic, Pompes

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	189,995	1,337		142,055	,000
	Pompes	2,883	,073	,863	39,708	,000
	Trafic	1,129	,073	,338	15,552	,000

a. Variable dépendante : Ventes

Ventes / Trafic & Concurrents

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,990[a]	,980	,979	2,274

a. Valeurs prédites : (constantes), Trafic, Concurrents

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	10399,060	2	5199,530	1005,196	,000[a]
	Résidu	217,251	42	5,173		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Trafic, Concurrents

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	188,081	1,413		133,113	,000
	Concurrents	2,385	,063	,848	38,154	,000
	Trafic	1,393	,074	,417	18,744	,000

a. Variable dépendante : Ventes

Ventes / Pompes & Trafic & Concurrents

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,991[a]	,981	,980	2,213

a. Valeurs prédites : (constantes), Trafic, Concurrents, Pompes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	10415,574	3	3471,858	709,119	,000[a]
	Résidu	200,737	41	4,896		
	Total	10616,311	44			

a. Valeurs prédites : (constantes), Trafic, Concurrents, Pompes

b. Variable dépendante : Ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	189,767	1,653		114,804	,000
	Pompes	2,551	1,389	,763	1,837	,074
	Concurrents	,276	1,150	,098	,239	,812
	Trafic	1,159	,146	,347	7,920	,000

a. Variable dépendante : Ventes

Exercice 1.

* Rejet du modèle à 3 variables.

Son r^2 est bon (même excellent : 98.1 %), sa validité globale est assurée (cf. second tableau : P-valeur quasi-nulle)
mais il y a un problème avec le tableau des validités marginales : la variable Concurrents n'est clairement pas individuellement significative face aux autres variables (cf. P-valeur de 81.2 %) ; de plus, on peut également avoir des doutes sur la significativité de la variable Pompes face aux autres variables même si cette fois-ci, les choses se jouent plus sur le fil (P-valeur de 7.4 %, supérieure au seuil habituel de 5 %).

* Passage à un modèle à 2 variables.

On applique une procédure "backward" et de ces deux variables non significatives, on préfère retirer celle qui semble la moins significative (= celle avec la plus grande P-valeur), soit Concurrents

↳ On propose donc le modèle Ventes / Pompes & Trafic.

* Validité statistique de ce modèle à 2 variables.

Ce modèle est valide globalement (P-valeur quasi-nulle dans son second tableau de régression) et les deux variables qu'il contient sont chacune individuellement significatives (cf. troisième tableau, dit des validités marginales : les P-valeurs de ses deuxième et troisième lignes sont quasi-nulles).
On note par ailleurs que la valeur réalisée du r^2 (ici, 98.1 %) reste

très élevée.

* Interprétation du modèle (des coefficients).

→ La relation estimée est la suivante :

$$\underset{\text{(en milliers de litres)}}{\text{Ventes}} = 189.995 + 2.883 \times \underset{\text{(nombre de pompes)}}{\text{Pompes}} + 1.129 \times \text{Trafic (en milliers de voitures)} + \text{aléa d'écart-type estimé } 2.188$$

→ Cela étant, vu les valeurs typiques pour Pompes (de l'ordre de 10) et pour Trafic (de l'ordre de 20), on voit donc que les variations des ventes, bien que significatives selon le nombre de pompes et la quantité de trafic, restent modestes face au débit de base (autour de 200 milliers de litres pour une station avec 2 pompes et un trafic faible).

↳ En gros, une station-service même mal équipée et mal située, c'est le jackpot ?

* Modèle à 2 variables contre les modèles à 1 variable.

Du point de vue statistique, la comparaison équitable se fait par comparaison des valeurs réalisées des r^2_{ajust} (ou, de manière équivalente, des estimés des écarts-types).

Meilleur modèle à 1 variable : Pompes avec réalisations suivantes : r^2_{ajust} : 86.9% ($\sqrt{\hat{\sigma}^2_{MS}}$: 5.624)

VS

Meilleur modèle à 2 variables : Pompes & Trafic, r^2_{ajust} : 98.0% ($\sqrt{\hat{\sigma}^2_{MS}}$: 2.188)

(On rappelle que "meilleur" correspond à la réalisation de r^2_{ajust} la plus grande, i.e., à l'estimée de l'écart-type la plus petite.)

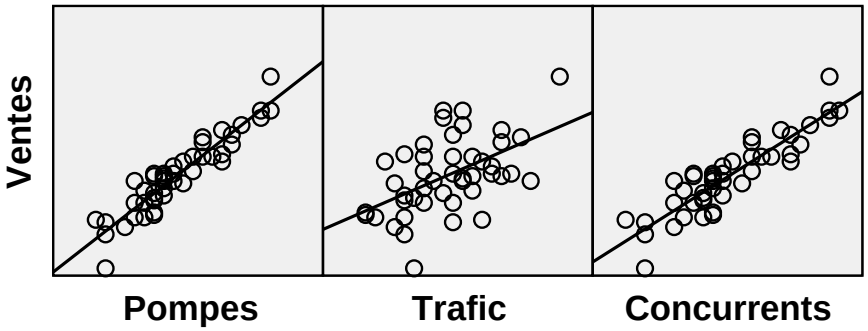
↳ Le modèle à 2 variables Pompes & Trafic est donc préférable.

* Note : Cela peut vous surprendre au vu de la régression complète, mais ce n'est pas Trafic mais Pompes la meilleure variable explicative.

* Remarque subsidiaire.

Le modèle Ventes / Pompes & Trafic a été obtenu par sélection "backward".

On vérifie facilement que c'est également le modèle retenu par la sélection "forward".

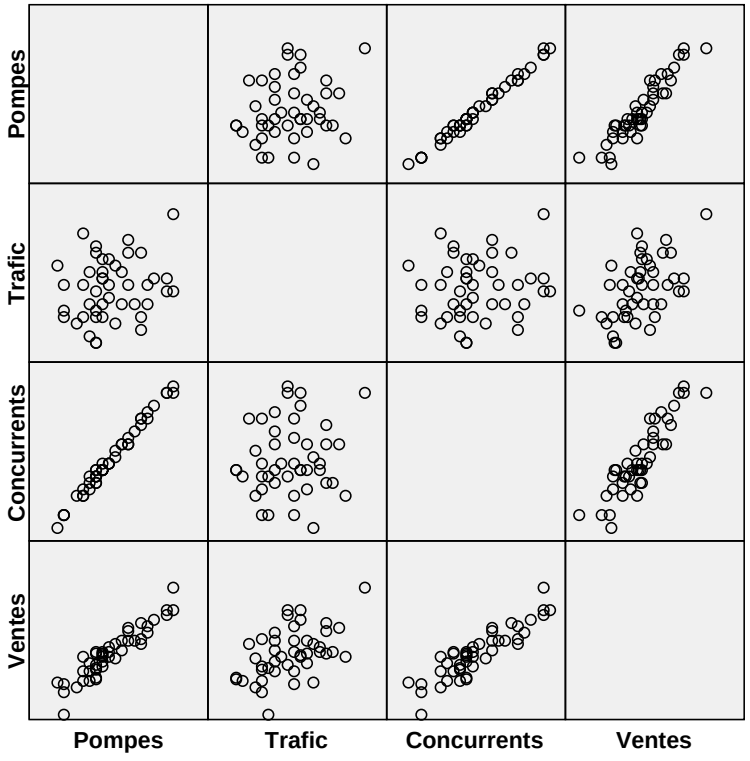


Corrélations

Corrélations

		Ventes	Pompes	Concurrents	Trafic
Ventes	Corrélation de Pearson	1	,934**	,899**	,520**
	Sig. (bilatérale)		,000	,000	,000
	N	45	45	45	45
Pompes	Corrélation de Pearson	,934**	1	,995**	,211
	Sig. (bilatérale)	,000		,000	,164
	N	45	45	45	45
Concurrents	Corrélation de Pearson	,899**	,995**	1	,121
	Sig. (bilatérale)	,000	,000		,427
	N	45	45	45	45
Trafic	Corrélation de Pearson	,520**	,211	,121	1
	Sig. (bilatérale)	,000	,164	,427	
	N	45	45	45	45

**. La corrélation est significative au niveau 0.01 (bilatéral).



CORRECTION 14.2 (Ventes en fonction du nombre de vendeurs et des prix). On commence, pour ceux qui ont la flemme d'ouvrir SPSS, par préciser les données :

Ventes	Vendeurs	PrixEntreprise	PrixConcurrence	IndicePrix
50	5	30	25	100.0
120	7	30	26	102.5
140	11	33	28	106.5
135	16	34	30	110.4
163	16	33	31	114.0
233	16	36	34	119.1
241	21	40	37	134.7
255	27	45	42	160.1
286	26	50	48	174.9
330	30	53	54	183.0
389	33	58	58	194.2
425	36	60	61	209.3
445	38	71	72	235.6
472	37	80	81	268.8
501	37	90	93	293.4
510	38	92	92	299.3
490	36	92	90	303.1
505	37	94	94	310.3

De même, on propose aux pages suivantes les 15 régressions possibles des ventes en fonction de 1, 2, 3 ou 4 variables explicatives parmi le nombre de vendeurs (Vendeurs), les prix de l'entreprise concernée (PrixEntreprise), ceux de la concurrence (PrixConcurrence) et l'indice des prix (IndicePrix).

Ventes / Vendeurs

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,969[a]	,939	,935	39,670

a. Valeurs prédites : (constantes), Nombre de vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	384514,013	1	384514,013	244,332	,000[a]
	Résidu	25179,765	16	1573,735		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Nombre de vendeurs

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-21,634	23,544		-,919	,372
	Nombre de vendeurs	13,018	,833	,969	15,631	,000

a. Variable dépendante : Nombre de ventes

Ventes / PrixEntreprise

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,954[a]	,910	,905	47,904

a. Valeurs prédites : (constantes), Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	372976,916	1	372976,916	162,531	,000[a]
	Résidu	36716,861	16	2294,804		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-35,263	29,785		-1,184	,254
	Prix entreprise	6,195	,486	,954	12,749	,000

a. Variable dépendante : Nombre de ventes

Ventes / PrixConcurrence

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,966[a]	,933	,929	41,501

a. Valeurs prédites : (constantes), Prix de la concurrence

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	382135,883	1	382135,883	221,867	,000[a]
	Résidu	27557,895	16	1722,368		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix de la concurrence

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-6,783	23,783		-,285	,779
	Prix de la concurrence	5,835	,392	,966	14,895	,000

a. Variable dépendante : Nombre de ventes

Ventes / IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,962[a]	,926	,921	43,525

a. Valeurs prédites : (constantes), Indice des prix

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	379382,724	1	379382,724	200,261	,000[a]
	Résidu	30311,053	16	1894,441		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-49,024	27,767		-1,766	,097
	Indice des prix	1,922	,136	,962	14,151	,000

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & PrixEntreprise

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,990[a]	,980	,977	23,545

a. Valeurs prédites : (constantes), Prix entreprise, Nombre de vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	401378,056	2	200689,028	362,005	,000[a]
	Résidu	8315,721	15	554,381		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix entreprise, Nombre de vendeurs

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-47,480	14,739		-3,222	,006
	Nombre de vendeurs	7,726	1,079	,575	7,158	,000
	Prix entreprise	2,876	,522	,443	5,515	,000

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & PrixConcurrence

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,991[a]	,982	,979	22,258

a. Valeurs prédites : (constantes), Prix de la concurrence, Nombre de vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	402262,635	2	201131,317	405,990	,000[a]
	Résidu	7431,143	15	495,410		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix de la concurrence, Nombre de vendeurs

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-30,728	13,297		-2,311	,035
	Nombre de vendeurs	7,034	1,104	,523	6,374	,000
	Prix de la concurrence	2,970	,496	,492	5,985	,000

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,989[a]	,978	,976	24,245

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	400876,703	2	200438,351	340,995	,000[a]
	Résidu	8817,075	15	587,805		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-51,655	15,473		-3,338	,004
	Nombre de vendeurs	7,271	1,202	,541	6,047	,000
	Indice des prix	,943	,179	,472	5,276	,000

a. Variable dépendante : Nombre de ventes

Ventes / PrixEntreprise & PrixConcurrence

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,979[a]	,958	,952	34,056

a. Valeurs prédites : (constantes), Prix de la concurrence, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	392296,763	2	196148,382	169,122	,000[a]
	Résidu	17397,015	15	1159,801		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix de la concurrence, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	80,894	35,473		2,280	,038
	Prix entreprise	-16,464	5,562	-2,536	-2,960	,010
	Prix de la concurrence	21,128	5,177	3,497	4,081	,001

a. Variable dépendante : Nombre de ventes

Ventes / PrixEntreprise & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,968[a]	,937	,929	41,440

a. Valeurs prédites : (constantes), Indice des prix, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	383934,960	2	191967,480	111,787	,000[a]
	Résidu	25758,818	15	1717,255		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-67,250	28,709		-2,342	,033
	Prix entreprise	-11,291	6,935	-1,739	-1,628	,124
	Indice des prix	5,390	2,134	2,698	2,526	,023

a. Variable dépendante : Nombre de ventes

Ventes / PrixConcurrence & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,966[a]	,933	,924	42,743

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	382288,695	2	191144,348	104,622	,000[a]
	Résidu	27405,082	15	1827,005		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	6,387	51,710		,124	,903
	Prix de la concurrence	7,567	6,000	1,252	1,261	,227
	Indice des prix	-,574	1,983	-,287	-,289	,776

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & PrixEntreprise & PrixConcurrence

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,992[a]	,983	,980	22,204

a. Valeurs prédites : (constantes), Prix de la concurrence, Nombre de vendeurs, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	402791,578	3	134263,859	272,333	,000[a]
	Résidu	6902,200	14	493,014		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Prix de la concurrence, Nombre de vendeurs, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-3,457	29,481		-,117	,908
	Nombre de vendeurs	6,226	1,349	,463	4,614	,000
	Prix entreprise	-4,604	4,445	-,709	-1,036	,318
	Prix de la concurrence	7,576	4,474	1,254	1,693	,113

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & PrixEntreprise & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,990[a]	,980	,976	24,180

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	401508,418	3	133836,139	228,909	,000[a]
	Résidu	8185,360	14	584,669		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-43,523	17,301		-2,516	,025
	Nombre de vendeurs	8,200	1,496	,610	5,482	,000
	Prix entreprise	5,246	5,047	,808	1,039	,316
	Indice des prix	-,793	1,680	-,397	-,472	,644

a. Variable dépendante : Nombre de ventes

Ventes / PrixEntreprise & PrixConcurrence & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,982[a]	,965	,957	32,197

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	395180,379	3	131726,793	127,067	,000[a]
	Résidu	14513,399	14	1036,671		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	43,482	40,347		1,078	,299
	Prix entreprise	-22,438	6,363	-3,456	-3,526	,003
	Prix de la concurrence	17,578	5,337	2,909	3,294	,005
	Indice des prix	3,015	1,808	1,509	1,668	,118

a. Variable dépendante : Nombre de ventes

Ventes / PrixEntreprise & PrixConcurrence & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,982[a]	,965	,957	32,197

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	395180,379	3	131726,793	127,067	,000[a]
	Résidu	14513,399	14	1036,671		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Prix de la concurrence, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	43,482	40,347		1,078	,299
	Prix entreprise	-22,438	6,363	-3,456	-3,526	,003
	Prix de la concurrence	17,578	5,337	2,909	3,294	,005
	Indice des prix	3,015	1,808	1,509	1,668	,118

a. Variable dépendante : Nombre de ventes

Ventes / Vendeurs & PrixEntreprise & PrixConcurrence & IndicePrix

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,992[a]	,983	,978	22,918

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs, Prix de la concurrence, Prix entreprise

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	402865,954	4	100716,488	191,762	,000[a]
	Résidu	6827,824	13	525,217		
	Total	409693,778	17			

a. Valeurs prédites : (constantes), Indice des prix, Nombre de vendeurs, Prix de la concurrence, Prix entreprise

b. Variable dépendante : Nombre de ventes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	-1,218	31,005		-,039	,969
	Nombre de vendeurs	6,611	1,728	,492	3,825	,002
	Prix entreprise	-2,680	6,869	-,413	-,390	,703
	Prix de la concurrence	7,446	4,631	1,232	1,608	,132
	Indice des prix	-,601	1,597	-,301	-,376	,713

a. Variable dépendante : Nombre de ventes

Exercice 2.

* Meilleure variable explicative prise isolément :

Il suffit de déterminer parmi les 4 modèle suivants celui qui a le meilleur (plus grand) r^2 réalisé :

Vents / Vendeurs :	93.9%
Vents / Prix Entreprise :	91.0%
Vents / Prix Concurrence :	93.3%
Vents / Indice Prix :	92.6%

Ces 4 régressions simples sont tracés plus bas, en bas de la page des illustrations.

↳ C'est donc le nombre de vendeurs. Note : Vents / Vendeurs est un modèle statistiquement valide.

* Meilleur couple de variables explicatives :

On applique le même principe et parmi les 6 modèls de régression à deux variables, on retient celui qui a le plus grand r^2 réalisé : il s'agit de

Vents / Vendeurs et Prix Concurrence.

Notes : - C'est bien un modèle statistiquement valide (validité globale du modèle et validité partielle : toutes les variables sont linéairement significatives.

- Comparons les valeurs réalisées des r^2 ajustés du meilleur modèle à 1 variable (→ 93.5%) et du meilleur modèle à 2 variable (→ 97.9%) : il y a augmentation du r^2 ajusté, cela signifie que l'apport de la nouvelle variable Prix Concurrence est significatif. (De manière équivalente, on note une forte diminution de l'estimée de l'écart-type des aléas.)

* Modèles de régression à 3 ou 4 variables explicatives :

Tous ces modèles ont un sens global (cf. P-valeur du test reporté dans les tableaux d'ANOVA) mais dans chacun d'entre eux, il y a

toujours au moins une variable non individuellement significative face aux autres (cf. toujours au moins une des P-valeurs des tests du tableau Coefficients qui est $> 0.05 = 5\%$).

Aucun de ces modèles ne peut donc être validé par l'analyse statistique.

* Comment expliquer une perte de significativité individuelle lors de l'ajout d'une nouvelle variable (ou, de manière équivalente, un retour du caractère individuellement significatif d'une variable lorsque l'on en retire une autre du modèle) :

On rappelle que ce caractère de significativité individuelle correspond au fait qu'une variable donnée aide (ou pas) à améliorer l'ajustement linéaire déjà réalisé par les autres variables.

Si, comme c'est le cas ici, des variables sont très fortement corrélées* (cf. tableau de corrélations : toutes les corrélations sont significatives), alors lorsque l'on en considère déjà une, l'autre n'apporte rien de significatif.

La situation du modèle à 4 variable est intéressante à ce sujet : il contient trois variables non individuellement significatives par rapport aux autres (Prix Entreprise, Prix Concurrence, Indice Prix). Or chacune de ces trois apporte quelque chose au modèle, comme on peut le voir en notant que dans les modèles

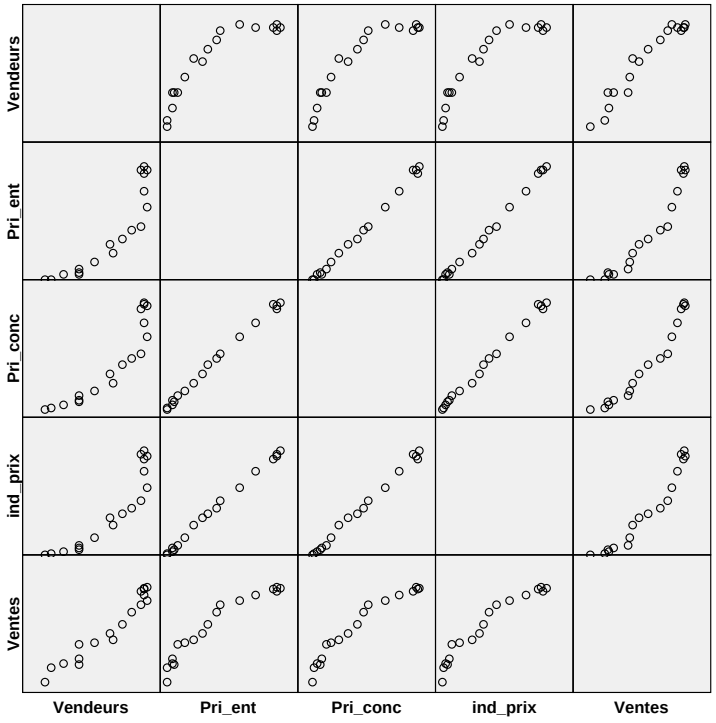
Ventes / Vendeurs & Prix Entreprise
Ventes / Vendeurs & Prix Concurrence
Ventes / Vendeurs & Indice Prix

ces trois variables sont chaque fois individuellement significatives.

* Cela semble surtout le cas de Prix Entreprise et Prix Concurrence, voir les figures page suivante.

* Recommandation finale :

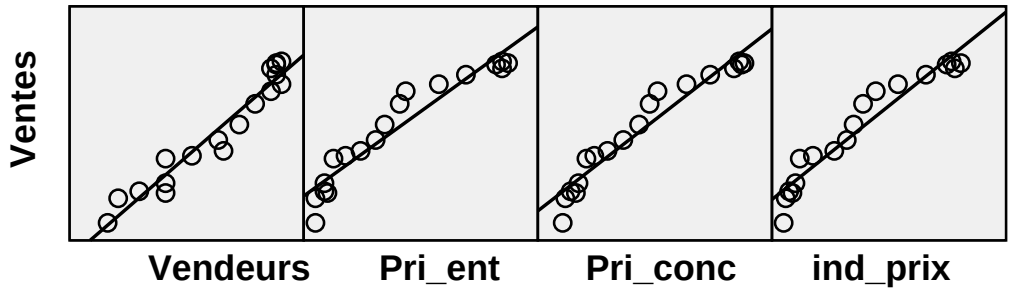
On a vu : - qu'aucun modèle à 3 ou 4 variables explicatives n'était valide statistiquement (existence de variables



Corrélations

		Ventes	Vendeurs	Pri_ent	Pri_conc	ind_prix
Ventes	Corrélation de Pearson	1	,969**	,954**	,966**	,962**
	Sig. (bilatérale)		,000	,000	,000	,000
	N	18	18	18	18	18
Vendeurs	Corrélation de Pearson	,969**	1	,889**	,906**	,906**
	Sig. (bilatérale)	,000		,000	,000	,000
	N	18	18	18	18	18
Pri_ent	Corrélation de Pearson	,954**	,889**	1	,998**	,998**
	Sig. (bilatérale)	,000	,000		,000	,000
	N	18	18	18	18	18
Pri_conc	Corrélation de Pearson	,966**	,906**	,998**	1	,998**
	Sig. (bilatérale)	,000	,000	,000		,000
	N	18	18	18	18	18
ind_prix	Corrélation de Pearson	,962**	,906**	,998**	,998**	1
	Sig. (bilatérale)	,000	,000	,000	,000	
	N	18	18	18	18	18

**. La corrélation est significative au niveau 0.01 (bilatéral).



non individuellement significatives)

- que parmi les modèls à 1 ou 2 variables explicatives, le meilleur était

Ventes / Vendeurs & Prix Concurrence

(cf. meilleur réalisation de r^2 ajusté).

C'est, du point de vue statistique, le modèle qu'on a envie de retenir.

Note: Cette recommandation statistique finale apparait également :

→ comme le résultat de la procédure de sélection backward :

1) retrait de IndicePrix (cf. plus grande P-valeur pour le tst de non significativité individuelle dans le modèle à 4 variables, cette P-valeur de 71.3% étant par ailleurs supérieure à 5%)
2) retrait de PrixEntreprise du modèle à 3 variables restant (cf. plus grande P-valeur, cette P-valeur de 31.8% étant supérieure à 5%)
3) arrivée à un modèle à 2 variables (Vendeurs, Prix Concurrence) où cette fois-ci, chacune est individuellement significative
   ↳ arrêt de la procédure Backward

→ comme le résultat de la procédure de sélection forward

1) choix de Vendeurs, la variable la plus significative (ici, dans les modèls à 1 variable, toutes les P-valeurs de significativité sont quasi-nulles : on lit 0.000 dans les tableaux; cependant, en regardant pour qui la valeur réalisée de la statistique de tst, lue dans la colonne "t", est la plus grande, on en déduit la plus petite P-valeur, et à elle correspond la variable la plus significative)
2) ajout de Prix Concurrence (même remarque que ci-dessus : comparer les valeurs dans la colonne "t")

3) Après, plus aucun ajout de variable étant individuellement significative dans le modèle étendu ainsi obtenu.

↳ arrêt de la procédure Forward.

* Validation économique, relation finale, commentaires et exploitation stratégique :

Validation économique : Les réponses précédentes montrent que le modèle Ventes / Vendeurs & PrixConcurrence est le meilleur modèle pour reconstruire les données d'échantillon.

On s'intéresse maintenant à son pouvoir explicatif / prédictif : éclaire-t-il la situation, d'un point de vue économique ?

Les variables Vendeurs et PrixConcurrence sont associées à des (estimées des) coefficients de régression positifs, ce qui est effectivement fort interprétable :

- si le nombre de vendeurs est augmenté, les ventes augmentent
- si le prix de la concurrence augmentent, les ventes de l'entreprise étudiée augmentent.

Ce modèle est en particulier beaucoup plus interprétable que le suivant : Ventes / Vendeurs & PrixEntreprise, dans lequel, lorsque l'entreprise augmente ses prix, les ventes augmentent aussi ! En fait, là encore, cela est sans doute dû uniquement aux grandes corrélations linéaires entre les variables explicatives de prix.

↳ Le modèle est sans doute plus explicatif que prédictif.

Relation finale :

$$\underset{\text{(nombre de ventes, en centaines)}}{\text{Ventes}} = -30.728 + 7.034 \times \underset{\text{(nombre de...)}}{\text{Vendeurs}} + 2.970 \times \underset{\text{(moyenne par produit, en euros)}}{\text{PrixConcurrence}} + \text{aléa, d'écart-type estimé } 22.258$$

Remarque : l'ordonnée à l'origine −30.728 est, elle, difficilement interprétable.

Exploitation stratégique :

On pourrait utiliser ce modèle pour voir comment varient, quand on ajoute ou enlève un vendeur, les ventes et donc la marge réalisée, et comparer cela au coût salarial d'un vendeur, afin de déterminer s'il faut embaucher ou licencier le cas échéant.

Cependant, la qualité de cette exploitation stratégique est à nuancer en fonction de ce que l'on a déjà souligné plus haut : le modèle est sans doute plus explicatif que prédictif.

CORRECTION 14.3 (Exercice des annales, 2008). Sa correction peut être trouvée aux pages suivantes.

Exercice 3. Cf. Exercice IV de l'examen principal de 2008.

* Réponses aux questions posées lors de l'examen.

1) Ce modèle est globalement valide (cf. P-valeur quasi-nulle dans le tableau d'ANOVA) mais le tableau des coefficients indique qu'il y a un problème avec les validités marginales : la variable des frais de scolarité (Fees) n'est pas significative face au taux de sélection (SelectionRate), cf. P-valeur lue de 92.3% pour l'hypothèse que son coefficient dans la régression est nul.
On peut donc valider statistiquement ce modèle : il faut enlever la variable Fees.

2) Pour comparer les 4 modèles, on compare leurs valeurs réalisées de r^2_{ajust} (le meilleur modèle est celui avec la plus grande valeur réalisée de r^2_{ajust}).

↳ On retient : Salary / SelectionRate (r^2_{ajust} : 90.1%)

Note : Salary / SelectionRate et Salary / SelectionRate & Fees ont tous deux un r^2 réalisé de 90.5%, mais on avait vu que le r^2 n'est pas un indicateur juste pour comparer des modèles utilisant des nombres différents de variables explicatives.

Relation proposée :

$$\text{Salaire moyen de sortie (en k\$)} = 83.215 - 0.708 \times \text{Taux de sélection} + \text{Aléa d'écart-type estimé } 2.825$$

Note : Le coefficient de pente est négatif et c'est bien naturel : plus une école est sélective, meilleure elle est !

3) On regarde le modèle Salary / PropPhDFaculty :
Le test de validité globale (cf. tableau ANOVA) conserve l'hypothèse d'absence de relation linéaire, avec une P-valeur de 14.8%.

L'excellence du corps professoral n'a donc aucune influence linéaire sur le salaire en sortie d'école.
En fait, il pourrait avoir une influence non-linéaire mais même en traçant le graphique des salaires en fonction de PropPhDFaculty, on ne détecte rien de très visible.

À noter cependant :

- ce taux a une influence, puisque c'est un des critères principaux pour le ranking des écoles, mais cette influence est difficile à percevoir (noyée au milieu de nombreuses influences, et finalement, on peut l'agréger à toutes les explications mises dans le terme stochastique) ;
- en fait, 100% des nouveaux recrutés depuis 5/10 ans ont un PhD ce n'est plus ce diplôme qui détermine fondamentalement l'excellence du corps professoral. Le recrutement universitaire est désormais organisé comme un marché, un candidat a une valeur de marché, et il faudrait plutôt voir quelles écoles attirent les meilleurs candidats.

4) [Désolé, je n'avais pu résister à vous faire découvrir cette conclusion démagogique !]

On a retenu le modèle Salary / SelectionRate : ce sont donc les étudiants (et le mode de recrutement par nombreux classes + publicité pour avoir de nombreux candidats) qui font la réputation d'une école. En fait, c'est le "tampon" gagné à 20 ans par l'admission dans une école qui conditionne

beaucoup de choses (mais pas tout : ne vous reposez pas sur vos lauriers !).

* Procédures automatiques.

Cf. deux sorties SPSS supplémentaires page suivante.

→ Forward :
- on part de SelectionRate (meilleure explicative) ;
- or, dans aucun des modèles Salary / SelectionRate & Fees ni Salary / SelectionRate & PropPhDFac, la variable ajoutée n'est individuellement significative ;
- on s'arrête donc à la première étape et l'on recommande Salary / SelectionRate.

→ Backward :
- on part du modèle à 3 variables : on enlève Fees, qui a la plus grande P-valeur (74.4% > 75%) et est donc la moins significative
- dans Salary / SelectionRate & PropPhDFac, PropPhDFac n'est pas significative (P-valeur de 77.4%), on l'enlève
- on arrive à Salary / SelectionRate, qui est acceptable.

En fait, les cas avec de belles et grandes valeurs de r^2 ont été inventés par des profs !

* Remarque : Dans la vraie vie, il est difficile d'obtenir de bons r^2 (plus grands que 40%). Ici, on obtient des valeurs fabuleuses (> 90%) à cause de la variable SelectionRate, que j'ai inventée. Les variables Fees et PropPhDFac du "Financial Times" apportent une explication linéaire beaucoup plus limitée (r^2 de 1.8% et 7.6% en régr. simple).

Régression complète

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,952[a]	,906	,894	2,924

a. Valeurs prédites : (constantes), SelectionRate, Fees, PropPhDFaculty

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2049,026	3	683,009	79,865	,000[a]
	Résidu	213,802	25	8,552		
	Total	2262,828	28			

a. Valeurs prédites : (constantes), SelectionRate, Fees, PropPhDFaculty

b. Variable dépendante : Salary

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	85,571	5,920		14,456	,000
	PropPhDFaculty	-,023	,054	-,034	-,426	,674
	Fees	-,031	,094	-,025	-,330	,744
	SelectionRate	-,713	,048	-,958	-14,804	,000

a. Variable dépendante : Salary

Régression sur deux variables, après suppression de la moins significative

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,951[a]	,905	,898	2,874

a. Valeurs prédites : (constantes), SelectionRate, PropPhDFaculty

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2048,096	2	1024,048	123,993	,000[a]
	Résidu	214,732	26	8,259		
	Total	2262,828	28			

a. Valeurs prédites : (constantes), SelectionRate, PropPhDFaculty

b. Variable dépendante : Salary

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	84,371	4,589		18,386	,000
	PropPhDFaculty	-,012	,043	-,018	-,291	,774
	SelectionRate	-,712	,047	-,957	-15,074	,000

a. Variable dépendante : Salary

Quinzième Partie

Annales d'examen (tous les énoncés, la plupart des corrigés)

Examen principal, session 2009–10 : énoncé uniquement

Examen 2009 du cours "Eléments de statistique mathématique"

Gilles Stoltz

Les exercices qui suivent sont indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix.

Le sujet est long, il en sera tenu compte dans la notation.

Il est demandé de numéroter soigneusement les réponses et de rédiger de manière complète et précise, mais également la plus concise possible.

Durée : 2 heures – **Tous documents autorisés, calculatrice autorisée**

Table de la loi normale : fournie dans le sujet

Exercice I : Quizz sur la théorie mathématique

Répondez, sur votre copie, par un mot ou un nombre aux questions suivantes. Il est inutile de justifier votre réponse (dans cet exercice uniquement).

(1) $x_1, \ldots, x_n$ désignent (a) les valeurs observées ou (b) sont des variables aléatoires.

(2) Si l'on prend $X_1, \ldots, X_{100}$ indépendantes et identiquement distribuées selon une loi $\mathcal{T}_3$, alors il est plus probable que $\overline{X}_{100}$ soit plus proche de (a) la valeur 0 ou (b) la valeur 3.

(3) Dans un modèle où $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathcal{B}(p_0)$, l'estimateur $\overline{X}_n(1-\overline{X}_n)$ est un estimateur sans biais de la variance $p_0\,(1-p_0)$: vrai ou faux ?

(4) Que vaut $\mathbb{P}\{N \leq 2.12\}$ lorsque $N \sim \mathcal{N}(0,1)$?

(5) Le quantile d'ordre α d'une loi de fonction de répartition F bijective est le nombre q_α tel que (a) $F(q_\alpha) = 1-\alpha$, ou (b) est tel qu'avec probabilité $100\,\%\,\alpha$, une nouvelle réalisation de la loi est plus petite que lui, ou (c) n'est ni l'un ni l'autre.

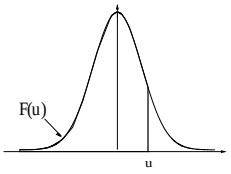
(6) On rejette H_0 lorsque la P-valeur est grande (plus grande que $5\,\%$ par exemple) : vrai ou faux ?

(7) Une modélisation de régression linéaire est d'autant meilleure que son coefficient de détermination r^2 est grand : vrai ou faux ?

Table de la loi normale

Loi normale : fonction de répartition

Pour une valeur $u \geqslant 0$, la table ci-dessous renvoie la valeur $F(u)$ de la fonction de répartition F de la loi normale centrée réduite au point u.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table pour les grandes valeurs de u :

u	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
u	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
u	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
u	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1

Exercice II : Valeur d'un stock

On considère une entreprise produisant des matériels high-tech. Elle dispose d'un stock important de pièces détachées, qui fait l'objet d'un inventaire permanent assuré par un système central d'information, à partir de bordereaux d'entrée (livraisons des fournisseurs) et de bons de sortie émis par la production.

Un auditeur responsable du contrôle de la comptabilité de l'entreprise a décidé de vérifier la valeur réelle du stock de pièces détachées. La diversité des articles constitutifs du stock l'a conduit à distinguer deux catégories :

- les articles de petite valeur (inférieure à 10 euros), pour lesquels il existe de nombreuses références : 1532 ;
- les articles de coût unitaire important (supérieur à 10 euros), pour lesquels il n'existe que 180 références.

Le contrôle d'une référence consiste en deux opérations :

- la détermination du bon prix unitaire de l'article correspondant (i.e., vérifier l'absence d'erreur de saisie de ce prix dans le système d'information) ;
- le recomptage des unités disponibles ; cette dernière opération prenant un certain temps.

A partir de ces deux informations, on peut alors déterminer la valeur réelle du stock existant pour cette référence.

Le temps mis par les responsables du stock pour recompter les références est le suivant : ils sont capables d'en recompter environ 50 par heure, étant 5 employés entraînés. Au vu des valeurs comptables déclarées (reproduites ci-dessous en Table 1) et du temps disponible, l'auditeur décide :

- de contrôler en totalité les références correspondant à des articles de coût unitaire important ;
- d'effectuer des contrôles aléatoires sur un nombre de références à déterminer pour les références correspondant à des articles de petite valeur.

Pour dimensionner le contrôle aléatoire, il effectue un coup de sonde sur 50 références choisies au hasard dans le catalogue et détermine la moyenne, sur ces 50 références, de leurs valeurs comptable (avant contrôle) et réelle (après contrôle). Les résultats en sont reproduits à la Table 2.

Coût unitaire	Nombre de références	Valeur totale en stock déclarée
Inférieur à 10 euros	1532	3 366 495
Supérieur à 10 euros	180	2 625 380

Table 1: Valeurs comptables déclarées.

Variable	Moyenne	Ecart-type
Valeurs comptables	2 315.83	777.35
Valeurs réelles	2 304.10	753.74
Ecarts	−11.73	110.32

Table 2: Résultats du coup de sonde sur 50 références choisies au hasard.

Les questions qui suivent visent à comprendre la démarche de l'auditeur, puis à déterminer le nombre de références supplémentaires à vérifier afin d'atteindre un objectif fixé.

(1) Expliquer la démarche de l'auditeur : pourquoi contrôle-t-il toutes les références d'une catégorie mais utilise-t-il une méthode aléatoire pour contrôler celles de l'autre catégorie ?

On étudie dans un premier temps l'information apportée par la *valeur réelle* des 50 références contrôlées sur les 1532.

(2) Modéliser la situation rencontrée (préciser notamment la population visée, le paramètre d'intérêt, etc.).

(3) Déduire du coup de sonde la réalisation d'un intervalle de confiance à 95 % sur la valeur réelle totale du stock d'articles de petite valeur ; indiquer sa précision.

(4) On suppose que la précision recherchée était égale à ±1 % de la valeur comptable actuelle totale de ces mêmes articles : constater que la précision obtenue en (3) n'est pas suffisante et déterminer la taille d'échantillon qui permettrait d'obtenir la précision souhaitée. Que pensez-vous de cette taille ?

On cherche donc une procédure plus précise et plus intelligente. A cet effet, on étudie désormais l'information apportée par les *écarts* entre valeurs comptable et réelle des 50 références contrôlées.

(5) Modéliser cette nouvelle situation.

(6) Calculer la réalisation d'un intervalle de confiance à 95 % sur la différence entre les valeurs comptable et réelle du stock.

(7) En déduire un intervalle de confiance sur la valeur réelle du stock ; indiquer sa précision.

(8) Déterminer la taille d'échantillon qui permettrait d'obtenir la précision souhaitée à la question (4). Que pensez-vous de cette taille cette fois-ci ?

(9) Prendre un instant de réflexion : qu'est-ce qui fait que, profondément, cette seconde démarche est plus efficace ?

Exercice III : Patch anti-tabac

Un laboratoire pharmaceutique envisage de lancer sur le marché une nouvelle formule de patch anti-tabac ; il ne veut le faire effectivement que s'il peut garantir formellement une efficacité strictement supérieure à 60 % d'arrêt au moins temporaire (plus de deux mois) du tabagisme. Des essais ont été réalisés sur un panel de 100 fumeurs choisis au hasard ; on en reporte ci-dessous les résultats.

Sexe	Nombre de sondés	Nombre d'arrêts temporaires
F	41	28
H	59	36
Tous	100	64

(1) Modéliser la situation décrite (sans distinguer pour l'instant entre hommes et femmes).

(2) Formuler les hypothèses du test du laboratoire. (On pensera avoir affaire à un laboratoire scientifiquement honnête.)

(3) Mettre en œuvre la démarche de test (de bout en bout) et conclure.

(4) Peut-on faire une différence sur l'efficacité du médicament selon le sexe ?

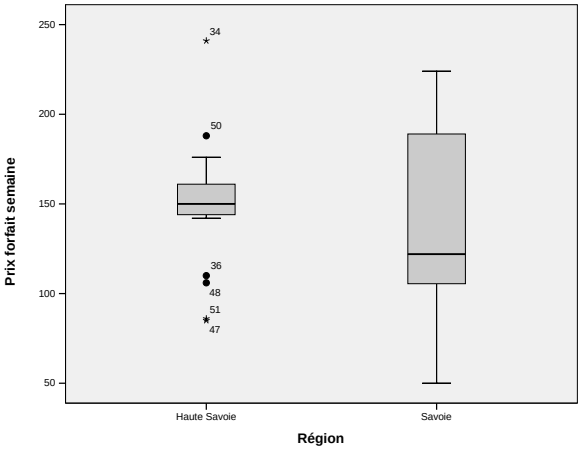
Exercice IV : Modélisation du prix du forfait de ski

On considère les données de la figure reproduite page suivante. On dispose, pour 98 stations de ski françaises :

- de leur nom ;
- du prix du forfait semaine ;
- de l'altitude de la station village ;
- du nombre de remontées ;
- de l'altitude du sommet des pistes ;
- du nombre de pistes ;
- du nombre de lits (hôtels et appartements en location saisonnière) ;
- de leur région d'appartenance, cette dernière étant codée par un nombre entre 1 et 8 (exemple : la Haute-Savoie est codée par 4, la Savoie par 7).

Comparaison de tendances centrales

On dit généralement que skier en Haute-Savoie est plus élitiste et plus cher qu'en Savoie.



(1) Comment appelle-t-on la représentation précédente ? Que suggère-t-elle ?

(2) Au vu des sorties SPSS fournies en annexe, que pensez-vous de l'affirmation énoncée plus haut et de ce que vous a suggéré la représentation graphique ? Vous indiquerez quelle sortie vous exploitez, quelle est la procédure qu'elle implémente et quel est son résultat.

On essaie maintenant de voir si on peut expliquer malgré tout les répartitions différentes observées dans la représentation graphique. On va construire à cet effet un modèle des prix en fonction des caractéristiques de chacune des stations.

Ski2.sav [Ensemble_de_données1] - SPSS Statistics Éditeur de données

Fichier Edition Affichage Données Transformer Analyse Graphes Outils Modules complémentaires Fenêtre Aide

1 : Station | Alpe d'Huez

	Station	Forfait	AltitudeStation	Remontées	AltitudePistes	Pistes	Lits	Région	Va
1	Alpe d'Huez	197	1860	85	3330	108	32000	3	
2	Alpe du Grand Serre	107	1400	20	2200	34	3500	3	
3	Areches Beaufort	110	1050	15	2300	30	5500	7	
4	Auris	104	1600	15	2175	19	4500	3	
5	Auron	120	1600	27	2450	38	10480	1	
6	Aussois	90	1500	11	2750	21	3000	7	
7	Autrans	78	1050	15	1710	18	8000	3	
8	Avoriaz	168	1800	39	2460	44	16200	4	
9	Bessans	70	1710	4	2200	4	1580	7	
10	Beuil	130	1430	27	2100	59	2500	1	
11	Bonneval sur arc	110	1800	10	3050	18	1567	7	
12	Chamonix	241	1035	49	3840	69	56908	4	
13	Chamrousse	136	1650	26	2255	36	15000	3	
14	Chatel	153	1200	39	2200	40	18000	4	
15	Combloux	110	900	80	1850	35	9000	4	
16	Courchevel	192	1100	67	3200	102	32000	7	
17	Crest Voland Corennoz	102	1150	17	1950	26	6500	7	
18	Eaux Bonnes Courette	100	1400	23	2400	30	7800	6	
19	Flaine	163	1600	377	2500	132	9500	4	
20	Flumet	50	1000	40	2070	20	6025	7	
21	Font Romeu	140	1800	23	2250	40	18000	6	
22	Gérardmer	80	680	20	1150	20	6400	8	
23	Isola 2000	122	1900	24	2610	47	8000	1	
24	La Bonhomme	85	830	11	1235	13	5000	8	
25	La Bresse	126	900	29	1350	39	4900	8	
26	La Clusaz	153	1100	57	2600	76	19500	4	
27	La Foux Val d'Allos	154	1800	20	2600	32	12000	1	
28	La Mongie	138	1800	52	2550	70	12000	6	
29	La Norma	115	1350	18	2750	25	3500	7	
30	La Plagne	223	1250	110	3250	123	46000	7	
31	La Rosière	145	1110	18	2600	33	7800	7	
32	La Tania	192	1350	67	3200	102	3500	7	
33	La Toussuire	122	1800	54	2600	89	8000	7	
34	Lans en Vercors	78	1020	16	1807	19	3500	3	
35	Le Collet d'Allevard	99	1450	13	2100	18	5012	3	

Construction d'un modèle linéaire

On exploitera là encore les sorties SPSS fournies en annexe.

Une seule variable explicative

(1) Parmi les modèles linéaires avec une variable explicative : quels sont ceux qui sont valides statistiquement ?

(2) Parmi ces derniers, quel est le meilleur modèle (i.e., quelle est la meilleure variable explicative prise isolément) ?

(3) Pourquoi et comment cette meilleure variable explicative influe-t-elle sur le prix du forfait ? (On demande une explication économique.)

Couples de variables explicatives

(4) Parmi les modèles linéaires formés sur des couples de variables explicatives : quels sont ceux qui sont valides statistiquement ?

(5) Parmi ces derniers, quel est le meilleur modèle (i.e., quel est le meilleur couple de variables explicatives) ?

(6) Pourquoi et comment ce meilleur couple de variables explicatives influe-t-il sur le prix du forfait ? (On demande une explication économique.)

Régression complète

(7) Que pensez-vous de la régression complète (i.e., sur toutes les variables) ?

(8) Si vous en pensez du bien : écrivez la relation qu'elle propose.
Si vous en pensez du mal : que faut-il lui faire ?

Méthodes 1 et 2

(9) Comment appelle-t-on respectivement les méthodes de sélection de variables notées 1 et 2 ? Expliquez leur cheminement.

(10) Quel(s) modèle(s) recommandent-elles respectivement, finalement ?

(11) Justifiez que ce(s) modèle(s) est (sont) valide(s) ; écrivez la (les) relation(s) proposée(s).

(12) Effectuez une interprétation économique de la (des) relation(s) ainsi écrite(s).

Conclusions

(13) Comment expliquer alors les répartitions de prix différentes entre la Savoie et la Haute-Savoie ?

(14) Conclusion stratégique : si vous lisez cette question, c'est que vos vacances n'ont jamais été aussi proches. Alors, où irez-vous skier ? J'irai pour ma part à Valloire...

Test-t

Statistiques de groupe

	Région	N	Moyenne	Ecart-type	Erreur standard moyenne
Prix forfait semaine	Haute Savoie	19	148,21	35,672	8,184
	Savoie	32	138,44	47,842	8,457

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
										Intervalle de confiance 95% de la différence
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Prix forfait semaine	Hypothèse de variances égales	5,343	,025	,771	49	,444	9,773	12,676	-15,700	35,246
	Hypothèse de variances inégales			,830	46,309	,411	9,773	11,769	-13,912	33,458

Régressions simples

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,252[a]	,064	,054	36,927

a. Valeurs prédites : (constantes), Altitude station

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	8900,781	1	8900,781	6,527	,012[a]
	Résidu	130905,107	96	1363,595		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Altitude station

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	97,197	14,374		6,762	,000
	Altitude station	,027	,011	,252	2,555	,012

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,472[a]	,223	,215	33,646

a. Valeurs prédites : (constantes), Nombre de remontées

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	31126,888	1	31126,888	27,495	,000[a]
	Résidu	108678,999	96	1132,073		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de remontées

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	116,229	4,623		25,140	,000
	Nombre de remontées	,434	,083	,472	5,244	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,722[a]	,521	,516	26,419

a. Valeurs prédites : (constantes), Altitude pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	72800,246	1	72800,246	104,302	,000[a]
	Résidu	67005,642	96	697,975		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Altitude pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	11,698	12,141		,963	,338
	Altitude pistes	,051	,005	,722	10,213	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,760[a]	,578	,574	24,787

a. Valeurs prédites : (constantes), Nombre de pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	80825,821	1	80825,821	131,558	,000[a]
	Résidu	58980,067	96	614,376		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	88,459	4,596		19,248	,000
	Nombre de pistes	,873	,076	,760	11,470	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,780[a]	,609	,605	23,863

a. Valeurs prédites : (constantes), Nombre de lits

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	85140,465	1	85140,465	149,518	,000[a]
	Résidu	54665,423	96	569,431		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	97,278	3,766		25,829	,000
	Nombre de lits	,003	,000	,780	12,228	,000

a. Variable dépendante : Prix forfait semaine

Régressions sur couples de variables explicatives

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,514[a]	,264	,248	32,911

a. Valeurs prédites : (constantes), Nombre de remontées, Altitude station

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	36907,008	2	18453,504	17,037	,000[a]
	Résidu	102898,880	95	1083,146		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de remontées, Altitude station

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	88,242	12,932		6,824	,000
	Altitude station	,022	,009	,204	2,310	,023
	Nombre de remontées	,414	,081	,450	5,085	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,744[a]	,553	,543	25,651

a. Valeurs prédites : (constantes), Altitude pistes, Altitude station

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	77297,770	2	38648,885	58,739	,000[a]
	Résidu	62508,118	95	657,980		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Altitude pistes, Altitude station

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	21,911	12,419		1,764	,081
	Altitude station	-,023	,009	-,216	-2,614	,010
	Altitude pistes	,059	,006	,842	10,196	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,771[a]	,594	,585	24,446

a. Valeurs prédites : (constantes), Nombre de pistes, Altitude station

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	83034,622	2	41517,311	69,474	,000[a]
	Résidu	56771,266	95	597,592		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de pistes, Altitude station

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	71,786	9,785		7,336	,000
	Altitude station	,014	,007	,128	1,923	,058
	Nombre de pistes	,849	,076	,739	11,138	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,803[a]	,645	,638	22,855

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	90182,938	2	45091,469	86,325	,000[a]
	Résidu	49622,949	95	522,347		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	71,194	9,137		7,791	,000
	Altitude station	,020	,007	,191	3,107	,002
	Nombre de lits	,003	,000	,765	12,474	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,774[a]	,599	,591	24,279

a. Valeurs prédites : (constantes), Altitude pistes, Nombre de remontées

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	83806,061	2	41903,031	71,086	,000[a]
	Résidu	55999,827	95	589,472		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Altitude pistes, Nombre de remontées

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	15,271	11,188		1,365	,176
	Nombre de remontées	,269	,062	,292	4,321	,000
	Altitude pistes	,045	,005	,640	9,453	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,764[a]	,584	,575	24,737

a. Valeurs prédites : (constantes), Nombre de pistes, Nombre de remontées

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	81674,262	2	40837,131	66,737	,000[a]
	Résidu	58131,626	95	611,912		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de pistes, Nombre de remontées

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	87,860	4,615		19,039	,000
	Nombre de remontées	-,099	,084	-,108	-1,178	,242
	Nombre de pistes	,960	,106	,836	9,089	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,809[a]	,654	,647	22,558

a. Valeurs prédites : (constantes), Nombre de lits, Nombre de remontées

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	91463,222	2	45731,611	89,869	,000[a]
	Résidu	48342,665	95	508,870		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Nombre de remontées

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	92,968	3,764		24,696	,000
	Nombre de remontées	,209	,059	,227	3,525	,001
	Nombre de lits	,003	,000	,701	10,889	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,847[a]	,717	,711	20,411

a. Valeurs prédites : (constantes), Nombre de pistes, Altitude pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	100229,959	2	50114,979	120,298	,000[a]
	Résidu	39575,929	95	416,589		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de pistes, Altitude pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	28,252	9,599		2,943	,004
	Altitude pistes	,031	,005	,441	6,825	,000
	Nombre de pistes	,602	,074	,524	8,114	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,853[a]	,727	,722	20,027

a. Valeurs prédites : (constantes), Nombre de lits, Altitude pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	101701,335	2	50850,668	126,778	,000[a]
	Résidu	38104,553	95	401,101		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Altitude pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	38,218	9,720		3,932	,000
	Altitude pistes	,029	,005	,415	6,426	,000
	Nombre de lits	,002	,000	,548	8,488	,000

a. Variable dépendante : Prix forfait semaine

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,863[a]	,744	,739	19,392

a. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	104080,884	2	52040,442	138,386	,000[a]
	Résidu	35725,004	95	376,053		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	83,013	3,662		22,670	,000
	Nombre de pistes	,526	,074	,458	7,097	,000
	Nombre de lits	,002	,000	,508	7,864	,000

a. Variable dépendante : Prix forfait semaine

Régression complète

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,897[a]	,804	,794	17,249

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	112433,232	5	22486,646	75,578	,000[a]
	Résidu	27372,656	92	297,529		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

b. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	42,736	8,603		4,968	,000
	Altitude station	-,002	,006	-,020	-,332	,741
	Nombre de remontées	-,004	,060	-,004	-,064	,949
	Altitude pistes	,023	,005	,323	4,354	,000
	Nombre de pistes	,418	,091	,364	4,586	,000
	Nombre de lits	,001	,000	,386	6,020	,000

a. Variable dépendante : Prix forfait semaine

Méthode 1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,897[a]	,804	,794	17,249
2	,897[b]	,804	,796	17,156
3	,897[c]	,804	,798	17,075

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

b. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Altitude pistes, Nombre de pistes

c. Valeurs prédites : (constantes), Nombre de lits, Altitude pistes, Nombre de pistes

ANOVA[d]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	112433,232	5	22486,646	75,578	,000[a]
	Résidu	27372,656	92	297,529		
	Total	139805,888	97			
2	Régression	112432,003	4	28108,001	95,494	,000[b]
	Résidu	27373,885	93	294,343		
	Total	139805,888	97			
3	Régression	112398,240	3	37466,080	128,497	,000[c]
	Résidu	27407,648	94	291,571		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Nombre de remontées, Altitude pistes, Nombre de pistes

b. Valeurs prédites : (constantes), Nombre de lits, Altitude station, Altitude pistes, Nombre de pistes

c. Valeurs prédites : (constantes), Nombre de lits, Altitude pistes, Nombre de pistes

d. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	42,736	8,603		4,968	,000
	Altitude station	-,002	,006	-,020	-,332	,741
	Nombre de remontées	-,004	,060	-,004	-,064	,949
	Altitude pistes	,023	,005	,323	4,354	,000
	Nombre de pistes	,418	,091	,364	4,586	,000
	Nombre de lits	,001	,000	,386	6,020	,000
2	(Constante)	42,703	8,542		4,999	,000
	Altitude station	-,002	,006	-,020	-,339	,736
	Altitude pistes	,023	,005	,324	4,427	,000
	Nombre de pistes	,414	,069	,360	6,004	,000
	Nombre de lits	,001	,000	,386	6,060	,000
3	(Constante)	42,096	8,312		5,065	,000
	Altitude pistes	,022	,004	,309	5,341	,000
	Nombre de pistes	,415	,069	,361	6,057	,000
	Nombre de lits	,001	,000	,393	6,460	,000

a. Variable dépendante : Prix forfait semaine

Variables exclues[c]

Modèle						Statistiques de colinéarité
		Bêta dans	t	Sig.	Corrélation partielle	Tolérance
2	Nombre de remontées	-,004[a]	-,064	,949	-,007	,502
3	Nombre de remontées	-,006[b]	-,087	,930	-,009	,504
	Altitude station	-,020[b]	-,339	,736	-,035	,612

a. Valeurs prédites dans le modèle : (constantes), Nombre de lits, Altitude station, Altitude pistes, Nombre de pistes

b. Valeurs prédites dans le modèle : (constantes), Nombre de lits, Altitude pistes, Nombre de pistes

c. Variable dépendante : Prix forfait semaine

Méthode 2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,780[a]	,609	,605	23,863
2	,863[b]	,744	,739	19,392
3	,897[c]	,804	,798	17,075

a. Valeurs prédites : (constantes), Nombre de lits

b. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes

c. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes, Altitude pistes

ANOVA[d]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	85140,465	1	85140,465	149,518	,000[a]
	Résidu	54665,423	96	569,431		
	Total	139805,888	97			
2	Régression	104080,884	2	52040,442	138,386	,000[b]
	Résidu	35725,004	95	376,053		
	Total	139805,888	97			
3	Régression	112398,240	3	37466,080	128,497	,000[c]
	Résidu	27407,648	94	291,571		
	Total	139805,888	97			

a. Valeurs prédites : (constantes), Nombre de lits

b. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes

c. Valeurs prédites : (constantes), Nombre de lits, Nombre de pistes, Altitude pistes

d. Variable dépendante : Prix forfait semaine

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	97,278	3,766		25,829	,000
	Nombre de lits	,003	,000	,780	12,228	,000
2	(Constante)	83,013	3,662		22,670	,000
	Nombre de lits	,002	,000	,508	7,864	,000
	Nombre de pistes	,526	,074	,458	7,097	,000
3	(Constante)	42,096	8,312		5,065	,000
	Nombre de lits	,001	,000	,393	6,460	,000
	Nombre de pistes	,415	,069	,361	6,057	,000
	Altitude pistes	,022	,004	,309	5,341	,000

a. Variable dépendante : Prix forfait semaine

Variables exclues[d]

Modèle						Statistiques de colinéarité
		Bêta dans	t	Sig.	Corrélation partielle	Tolérance
1	Altitude station	,191[a]	3,107	,002	,304	,993
	Nombre de remontées	,227[a]	3,525	,001	,340	,878
	Altitude pistes	,415[a]	6,426	,000	,550	,688
	Nombre de pistes	,458[a]	7,097	,000	,589	,646
2	Altitude station	,138[b]	2,707	,008	,269	,971
	Nombre de remontées	-,046[b]	-,632	,529	-,065	,512
	Altitude pistes	,309[b]	5,341	,000	,483	,624
3	Altitude station	-,020[c]	-,339	,736	-,035	,612
	Nombre de remontées	-,006[c]	-,087	,930	-,009	,504

a. Valeurs prédites dans le modèle : (constantes), Nombre de lits

b. Valeurs prédites dans le modèle : (constantes), Nombre de lits, Nombre de pistes

c. Valeurs prédites dans le modèle : (constantes), Nombre de lits, Nombre de pistes, Altitude pistes

d. Variable dépendante : Prix forfait semaine

Examen de rattrapage, session 2009–10 : énoncé uniquement

Examen de rattrapage 2009–10 des cours de Statistiques

Xavier Boute, Gilles Mauffrey et Gilles Stoltz

Les exercices qui suivent sont indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix.

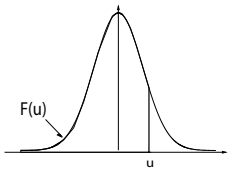
Il est demandé de numéroter soigneusement les réponses et de rédiger de manière complète et précise, mais également la plus concise possible.

Durée : 2 heures – **Tous documents autorisés, calculatrice autorisée**

Table de la loi normale : fournie dans le sujet

Table de la loi normale

Pour une valeur $u \geqslant 0$, la table ci-dessous renvoie la valeur $F(u)$ de la fonction de répartition F de la loi normale centrée réduite au point u.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table pour les grandes valeurs de u :

u	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
u	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
u	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
u	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1

Exercice I : Estimation (7 points)

La société Canf vend en ligne des produits de grande consommation ; elle utilise les services d'un prestataire pour ses expéditions en France métropolitaine. Or, le service client a reçu récemment un certain nombre de mails se plaignant de délais non respectés. Son responsable voudrait avoir une idée précise du taux de satisfaction des clients vis-à-vis de la livraison. Il a réussi à dégager un budget de 6 000 euros pour mener une enquête téléphonique et le sous-traitant retenu pour la réaliser facture 3 euros l'obtention de chaque réponse exploitable. (En effet, il faut évidemment effectuer bien plus qu'un appel en moyenne pour avoir une réponse exploitable, cela prend donc un peu de temps, donc de l'argent.)

1. Modéliser le problème : quelle est la population étudiée, quel est le paramètre étudié, quelles sont les variables disponibles ?

2. Quelle taille d'échantillon serait-elle nécessaire pour pouvoir garantir *a priori* une précision de 2 % avec un degré de confiance à 95 % ?

3. Le sondage réalisé a donné les résultats suivants :

Satisfaits	1 780
Non satisfaits	220
Total	2000

   (a) Expliquer la taille de l'échantillon.

   (b) Donner une estimation par intervalle du paramètre d'intérêt.

4. La hiérarchie du responsable du service client pourrait-elle lui reprocher d'avoir gaspillé de l'argent ? Autrement dit, notre responsable n'aurait-il pas pu faire des économies ?
   Indiquer comment il aurait pu procéder pour mener (ou faire mener) son enquête de manière plus économique et chiffrer l'ordre de grandeur des économies qui auraient ainsi été réalisées.

Exercice II : Tests d'hypothèses (7 points)

Depuis six mois, les ventes hebdomadaires de Kidem, le produit phare de la maison FLANY, ne cessent de baisser ; ainsi, la moyenne des quantités vendues par point de vente a dégringolé de 5 000 unités à 4 500 seulement. La chef de produit a fait étudier un nouvel emballage plus jeune et dynamique pour Kidem ; ce dernier a reçu un très bon accueil auprès d'un panel de consommateurs. Elle voudrait maintenant vérifier que cette nouvelle présentation ramènera la rentabilité de Kidem à son niveau antérieur, même si elle est plus coûteuse que le conditionnement précédent (la marge unitaire brute du produit va ainsi légèrement baisser et passer de $0,52$ euros à $0,50$ euros).

1. Montrer que pour que la marge totale brute revienne au niveau voulu, les ventes hebdomadaires moyennes doivent être d'au moins 5 200 unités.
2. La chef de produit va présenter le nouvel emballage dans 500 points de ventes et noter les ventes réalisées pendant une semaine.
   (a) Modéliser le problème : quelle est la population étudiée, quel est le paramètre d'intérêt μ_0, quelles seront les variables disponibles ?
   (b) L'objectif de la chef de produit, qui se souvient de son cours de statistique, est de mettre en œuvre un test d'hypothèses. Or, elle hésite entre les deux jeux d'hypothèses suivants :

   $$\text{(a)} \begin{cases} H_0 : \mu_0 \geqslant 5\,200 \\ H_1 : \mu_0 < 5\,200 \end{cases} \quad \text{vs.} \quad \text{(b)} \begin{cases} H_0 : \mu_0 \leqslant 5\,200 \\ H_1 : \mu_0 > 5\,200 \end{cases}$$

   Lequel de ces deux jeux vous semble-t-il le mieux adapté à la situation ?
3. Après l'expérimentation d'une semaine, les résultats suivants sont parvenus :

Nombre de points de vente	500
Ventes hebdomadaires : moyenne	5 290
Ventes hebdomadaires : écart-type	889,53

   Quelle décision doit prendre la chef de produit ? (On prendra, par exemple, un risque de première espèce de 5 %.)

Exercice III : Régression linéaire (6 points)

On a relevé, sur un échantillon de 50 pharmacies et pendant une période d'un mois, le nombre de prescriptions exécutées comportant le médicament X–Form du laboratoire Antal. On se propose d'expliquer les variations de ces prescriptions en fonction des variables suivantes :

- le nombre de médecins dans le quartier de la pharmacie ;
- le nombre moyen de visites par médecin et par an du visiteur médical du laboratoire Antal associé à la zone géographique ;
- le nombre moyen d'échantillons gratuits distribués dans le mois par le visiteur médical sur la zone ;
- la note du visiteur médical au test de QI effectué au moment de son recrutement.

Dans les questions qui suivent, on prendra un risque de première espèce de 5 %. Les tableaux de résumé de toutes les régressions dont il est question dans l'énoncé sont disponibles aux pages suivantes.

1. A propos des régressions linéaires à une variable explicative :
   (a) analyser chacune de ces régressions d'un point de vue statistique et économique ;
   (b) quelle vous paraît être la meilleure variable explicative prise individuellement ?
2. La régression complète (sur les quatre variables explicatives) est-elle valable statistiquement ? Economiquement ?
3. A propos de la régression sur trois variables explicatives :
   (a) justifier le choix des trois variables considérées ;
   (b) le modèle proposé est-il satisfaisant ?
   (c) pouvez-vous expliquer d'un point de vue statistique pourquoi la variable précisant le nombre d'échantillons gratuits est ici significative alors qu'elle ne l'était pas lors de la régression à une seule variable ?
4. Quel serait le nombre de prescriptions prévu pour une pharmacie située dans une zone où il y 10 médecins, où le nombre moyen de visites est de 3,5 par médecin et par an, où 4 échantillons gratuits sont distribués en moyenne chaque mois et où le visiteur a un QI mesuré de 110 ?

Régressions sur une variable explicative

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,179[a]	,032	,012	2,336

a. Valeurs prédites : (constantes), Echantillons distribués

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	8,629	1	8,629	1,582	,215[a]
	Résidu	261,871	48	5,456		
	Total	270,500	49			

a. Valeurs prédites : (constantes), Echantillons distribués

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	52,777	2,190		24,097	,000
	Echantillons distribués	,320	,254	,179	1,258	,215

a. Variable dépendante : Nombre de prescriptions

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,468[a]	,219	,202	2,098

a. Valeurs prédites : (constantes), Nombre de visites

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	59,175	1	59,175	13,441	,001[a]
	Résidu	211,325	48	4,403		
	Total	270,500	49			

a. Valeurs prédites : (constantes), Nombre de visites

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	52,537	,861		61,017	,000
	Nombre de visites	1,095	,299	,468	3,666	,001

a. Variable dépendante : Nombre de prescriptions

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,058[a]	,003	-,017	2,370

a. Valeurs prédites : (constantes), Nombre de médecins

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	,911	1	,911	,162	,689[a]
	Résidu	269,589	48	5,616		
	Total	270,500	49			

a. Valeurs prédites : (constantes), Nombre de médecins

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	54,760	1,867		29,329	,000
	Nombre de médecins	,079	,197	,058	,403	,689

a. Variable dépendante : Nombre de prescriptions

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,312[a]	,098	,079	2,255

a. Valeurs prédites : (constantes), QI du visiteur

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	26,395	1	26,395	5,190	,027[a]
	Résidu	244,105	48	5,086		
	Total	270,500	49			

a. Valeurs prédites : (constantes), QI du visiteur

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	49,822	2,513		19,829	,000
	QI du visiteur	,059	,026	,312	2,278	,027

a. Variable dépendante : Nombre de prescriptions

Régression sur les quatre variables explicatives

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,637[a]	,405	,353	1,891

a. Valeurs prédites : (constantes), QI du visiteur, Nombre de visites, Nombre de médecins, Echantillons distribués

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	109,670	4	27,418	7,671	,000[a]
	Résidu	160,830	45	3,574		
	Total	270,500	49			

a. Valeurs prédites : (constantes), QI du visiteur, Nombre de visites, Nombre de médecins, Echantillons distribués

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	48,361	2,980		16,231	,000
	Echantillons distribués	-,667	,291	-,373	-2,292	,027
	Nombre de visites	1,788	,386	,764	4,635	,000
	Nombre de médecins	,190	,162	,139	1,174	,247
	QI du visiteur	,064	,022	,342	2,919	,005

a. Variable dépendante : Nombre de prescriptions

Régression sur trois variables explicatives

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,622[a]	,387	,347	1,898

a. Valeurs prédites : (constantes), QI du visiteur, Nombre de visites, Echantillons distribués

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	104,747	3	34,916	9,690	,000[a]
	Résidu	165,753	46	3,603		
	Total	270,500	49			

a. Valeurs prédites : (constantes), QI du visiteur, Nombre de visites, Echantillons distribués

b. Variable dépendante : Nombre de prescriptions

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
1	(Constante)	49,915	2,681		18,620	,000
	Echantillons distribués	-,649	,292	-,363	-2,222	,031
	Nombre de visites	1,702	,380	,727	4,476	,000
	QI du visiteur	,067	,022	,358	3,068	,004

a. Variable dépendante : Nombre de prescriptions

Examen principal, session 2008–09 : énoncé

Examen 2008 du cours "Eléments de statistique mathématique"

Groupes de Gilles Stoltz

Les exercices qui suivent sont indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix. Cependant, ils sont présentés dans un ordre qui me semble croissant de difficulté.

Il est demandé de numéroter soigneusement les réponses et de rédiger manière complète et précise, mais concise.

Ne soyez pas troublés par la longueur du sujet d'examen ; il en sera tenu compte lors de la correction. Je ne m'attends pas à ce que vous ayez le temps de répondre à toutes les questions.

Durée : 2 heures – **Tous documents autorisés, calculatrice autorisée**

Table de la loi normale : disponible à la fin du sujet

Exercice I : Quizz sur la théorie mathématique

Répondez, sur votre copie, par un mot ou un nombre aux questions suivantes. Il est inutile de justifier votre réponse (dans cet exercice uniquement).

(1) $x_1, \ldots, x_n$ désignent (a) les valeurs observées ou (b) sont des variables aléatoires.

(2) Si l'on prend $X_1, \ldots, X_{100}$ indépendantes et identiquement distribuées selon une loi $\mathcal{T}_3$, alors il est plus probable que $\overline{X}_{100}$ soit plus proche de (a) la valeur 0 ou (b) la valeur 3.

(3) Dans un modèle où $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi de Bernoulli $\mathcal{B}(p_0)$, l'estimateur $\overline{X}_n(1-\overline{X}_n)$ est un estimateur sans biais de la variance $p_0\,(1-p_0)$: vrai ou faux ?

(4) Que vaut $\mathbb{P}\{N \leq 2.12\}$ lorsque $N \sim \mathcal{N}(0,1)$?

(5) Le quantile d'ordre α d'une loi de fonction de répartition F bijective est le nombre q_α tel que (a) $F(q_\alpha) = 1-\alpha$, ou (b) est tel qu'avec probabilité $100\,\%\,\alpha$, une nouvelle réalisation de la loi est plus petite que lui, ou (c) n'est ni l'un ni l'autre.

(6) On rejette H_0 lorsque la P-valeur est grande (plus grande que $5\,\%$ par exemple) : vrai ou faux ?

(7) Une modélisation de régression linéaire est d'autant meilleure que son coefficient de détermination r^2 est grand : vrai ou faux ?

Exercice II : Création d'une compagnie d'assurance

Un ancien d'HEC voudrait exploiter un créneau délaissé et créer une compagnie d'assurance réservée aux étudiants, avec des tarifs ajustés pour eux, à l'image, par exemple, de la MAIF qui est dédiée aux enseignants. Pour fixer ses tarifs, il commence par une étude préliminaire sur les risques et sinistres rencontrés par les étudiants (pour voir s'ils lui coûteront cher ou pas). Il fait ainsi interroger au téléphone des étudiants de toute la France, pour leur demander s'ils ont eu ou non un accident lors de l'année écoulée, et si oui, dans le cas où ils en étaient responsables, quel est le montant total des sinistres encourus pour eux et pour les autres voitures impliquées (en somme, le total à charge pour l'assurance).

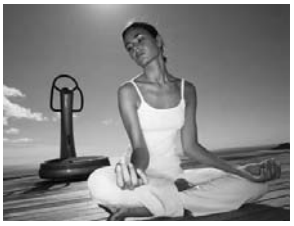
Sur les sondés, on ne s'intéresse qu'aux 1 472 qui sont assurés : 256 d'entre eux rapportent un accident pour lequel ils sont responsables, avec une moyenne de frais encourus de 1 865 euros (et un écart-type mesuré de 524 euros).

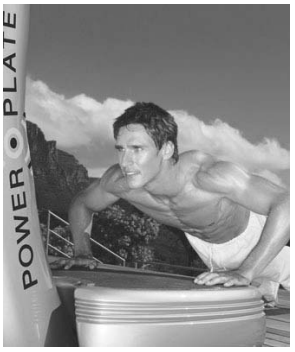
1. Modéliser le problème : décrire la population ciblée, l'échantillon retenu, les données recueillies, et comment on peut les modéliser mathématiquement ; indiquer les *deux* paramètres d'intérêt, avec pour chacun d'entre eux, une interprétation (ce qu'il représente).
2. Préciser, pour chacun des deux paramètres, un estimateur, et l'estimée correspondante.
3. Construire un intervalle de confiance à 95 % sur chacun des deux paramètres. On pourra les prendre bilatères ou unilatères, mais on justifiera soigneusement la forme retenue (par la précision de l'objectif poursuivi, d'après vous, par l'entrepreneur).
4. En déduire finalement un intervalle de confiance sur le montant total à prévoir pour les sinistres à charge par assuré étudiant. Quel est le niveau de cet intervalle de confiance ?
5. A quoi va servir, d'après vous, l'estimation de la question précédente ?

Exercice III : Campagne marketing pour un appareil de musculation

Campagne publicitaire en France

Un fabricant d'appareil de musculation hésite entre deux campagnes publicitaires. Elles sont représentées ci-dessous ; on les appellera dans ce qui suit "image de gauche" et "image de droite".





Elle engage des enquêteurs pour les poster à la sortie de différentes salles de sport. Ils fournissent les résultats suivants :

Image préférée	Gauche	Droite	Total
Hommes	89	75	164
Femmes	51	54	105
Total	140	129	269

Dans la réponse à chacune des deux questions qui suivent, on commencera par modéliser le problème et formuler soigneusement les deux hypothèses du test à effectuer.

1. Sachant que dans la population française, il y a 51.4 % de femmes, peut-on dire que la population qui fréquente les salles de sport n'est pas représentative de la population française ? Justifiez votre affirmation par la mise en œuvre d'un test.
2. A quel point peut-on affirmer que les goûts des hommes et des femmes pratiquant un sport en salle sont différents ? Est-ce au point qu'il faut songer à faire deux campagnes publicitaires distinctes, une destinée aux hommes et une autre aux femmes ? C'est ce que veut faire la directrice du marketing, mais les services financiers renâclent. Aidez le directeur général à y voir plus clair : effectuez un test d'hypothèses.

L'achat raisonné d'un gérant de salle

Un gérant de salle de sport à qui des dizaines de clients, alléchés par la campagne, ont demandé de commander l'appareil, veut s'en tirer le mieux possible. Il se demande si par hasard, comme pour les iPhone, les prix ne seraient pas moins chers aux Etats-Unis

qu'en Europe. Une recherche rapide sur Internet et par téléphone lui permet de trouver 54 prix de vente aux Etats-Unis et 48 en Europe. Il convertit les prix américains en euros (au taux du jour) et obtient comme résultat de la comparaison sous SPSS le tableau suivant :

Group Statistics

	Country	N	Mean	Std. Deviation	Std. Error Mean
Price	US	54	7099,6416	288,60563	39,27425
	EU	48	6976,8485	205,57792	29,67262

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Price	Equal variances assumed	8,245	,005	2,447	100	,016	122,79316	50,18824	23,22112	222,36521
	Equal variances not assumed			2,495	95,636	,014	122,79316	49,22328	25,08099	220,50534

Où doit-il effectuer sa commande, sachant que les frais de livraison sont facturés le même prix par le transporteur aussi éloigné soit le lieu de livraison du lieu d'expédition ? Vous indiquerez les hypothèses testées et expliquerez comment lire la sortie SPSS ci-dessus.

Exercice IV : Salaire d'embauche en sortie de business school

On étudie les données de la page suivante, qui indiquent, pour différentes écoles de commerce, le salaire moyen annuel trois ans après la sortie (en milliers de dollars), le pourcentage du corps professoral permanent détenant un doctorat, les droits de scolarité (en milliers d'euros), et le taux de sélection à l'entrée (rapport entre le nombre de dossiers présentés et le nombre de dossiers acceptés). Ces données sont tirées (sauf la dernière colonne) du classement Financial Times 2008.

On cherche à expliquer le salaire en fonction des caractéristiques des écoles.

1. On considère le modèle linéaire avec pour variables explicatives le montant des frais de scolarité et le taux de sélection à l'entrée. Que pensez-vous de ce modèle : est-ce un bon modèle, explique-t-il bien la réalité et avec un minimum de variables ?
2. Quel modèle, parmi les quatre qui sont proposés, retiendriez-vous, et pourquoi ? Indiquez la relation qu'il propose entre la variable à expliquer et la/les variable(s) explicative(s).
3. En particulier, que pensez-vous de l'influence de l'excellence du corps professoral (mesuré en termes de proportion de docteurs) sur le salaire à la sortie ?
4. Concluez : quelle est la variable qui vous garantit un bon salaire futur, i.e., qui fait la réputation d'une école ?

School	Salary	Prop. PhD Fac.	Fees	Sel. rate
HEC Paris	75	92	16	20
Mannheim Business School	68	79	2	21
ESCP-EAP European School of Management	66	87	16	30
Stockholm School of Economics	66	98	0	25
Grenoble Graduate School of Business	64	71	11	35
Essec Business School	61	91	14	23
Cems	60	80	0	30
Copenhagen Business School	57	89	0	40
Warsaw School of Economics	54	78	0	39
EM Lyon	53	89	15	38
Edhec Business School	53	82	17	35
Maastricht University	53	86	1	40
Rotterdam School of Management	52	100	2	41
Solvay Business School	52	98	1	49
Audencia Nantes	51	78	14	43
ESC Toulouse	51	73	12	48
Vienna University of Economics and Business	51	91	1	47
Esade Business School	50	78	20	44
Nyenrode Business Universiteit	50	51	23	49
Vlerick Leuven Gent Management School	49	98	8	50
ESC Rouen School of Management	48	64	15	51
Reims Management School	46	67	16	53
IAG-Louvain School of Management	46	100	1	54
Universiteit Antwerpen Management School	45	100	5	52
ESC Lille	43	60	16	59
Euromed Marseille Ecole de Management	42	80	15	57
ESC Tours-Poitiers (ESCEM)	41	65	14	58
IAE Aix-en-Provence School of Management	41	91	3	59
ESC Clermont	38	66	13	62

Salary / SelectionRate & Fees

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,951[a]	,905	,898	2,878

a. Predictors: (Constant), SelectionRate, Fees

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	83,215	2,061		40,378	,000
	Fees	-,007	,075	-,006	-,097	,923
	SelectionRate	-,707	,045	-,950	-15,563	,000

a. Dependent Variable: Salaray

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2047,477	2	1023,739	123,599	,000[a]
	Residual	215,350	26	8,283		
	Total	2262,828	28			

a. Predictors: (Constant), SelectionRate, Fees

b. Dependent Variable: Salaray

Salary / Fees

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,135[a]	,018	-,018	9,071

a. Predictors: (Constant), Fees

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41,280	1	41,280	,502	,485[a]
	Residual	2221,547	27	82,280		
	Total	2262,828	28			

a. Predictors: (Constant), Fees

b. Dependent Variable: Salaray

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	54,162	2,752		19,679	,000
	Fees	-,165	,233	-,135	-,708	,485

a. Dependent Variable: Salaray

Salary / SelectionRate

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,951[a]	,905	,901	2,825

a. Predictors: (Constant), SelectionRate

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2047,399	1	2047,399	256,603	,000[a]
	Residual	215,429	27	7,979		
	Total	2262,828	28			

a. Predictors: (Constant), SelectionRate

b. Dependent Variable: Salaray

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	83,173	1,978		42,047	,000
	SelectionRate	-,708	,044	-,951	-16,019	,000

a. Dependent Variable: Salaray

Salary / PropPhDFaculty

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,275[a]	,076	,042	8,801

a. Predictors: (Constant), PropPhDFaculty

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	171,487	1	171,487	2,214	,148[a]
	Residual	2091,340	27	77,457		
	Total	2262,828	28			

a. Predictors: (Constant), PropPhDFaculty

b. Dependent Variable: Salaray

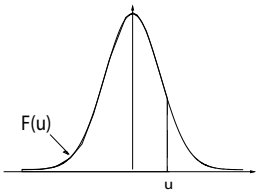
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	37,451	10,325		3,627	,001
	PropPhDFaculty	,185	,124	,275	1,488	,148

a. Dependent Variable: Salaray

Table de la loi normale

Pour une valeur $u \geqslant 0$, la table ci-dessous renvoie la valeur $F(u)$ de la fonction de répartition F de la loi normale centrée réduite au point u.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table pour les grandes valeurs de u :

u	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
u	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
u	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
u	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1

Examen principal, session 2008–09 : corrigé

Exercice I

Cet exercice a été repris dans l'examen principal de la session 2009–10.

Exercice II

La question 1 est traitée par l'exercice 2.2, page 41. Les questions suivantes sont résolues par l'exercice 5.8, page 117.

Exercice III

La question 1 de la partie "Campagne publicitaire en France" est abordée dans l'exercice 7.5, page 183. Le reste des questions est traité par l'exercice 9.2, page 239.

Exercice IV

Il correspond à l'exercice 14.3, page 372.

Examen de rattrapage, session 2008–09 : énoncé

Examen (de rattrapage) 2008 du cours "Eléments de statistique mathématique"

Groupes de Gilles Stoltz

Les blocs de questions qui suivent sont largement indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix. Cependant, ils sont présentés dans un ordre qui me semble croissant de difficulté.

Il est demandé de numéroter soigneusement les réponses et de rédiger manière complète et précise, mais concise.

Cet examen de rattrapage ne s'adressant qu'à des étudiants recalés à la première session, et à aucun absent, il est plus simple qu'un examen ordinaire. Je vous rappelle cependant que dans une telle situation, vu le règlement intérieur de la grande école, la note ne peut dépasser E.

Durée : 2 heures – **Tous documents autorisés, calculatrice autorisée**

Table de la loi normale : disponible à la fin du sujet

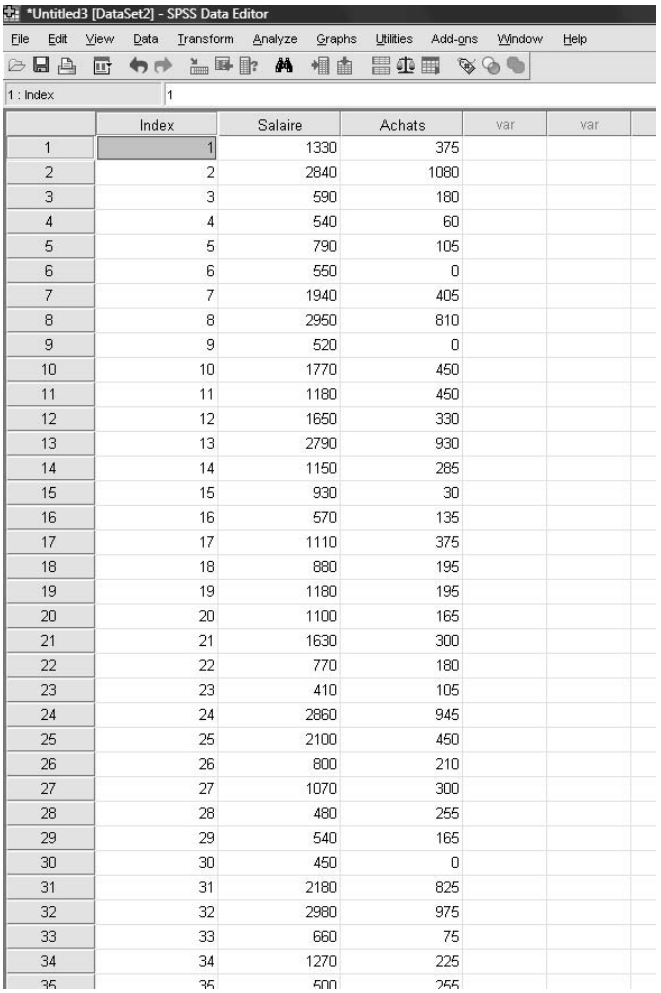
L'étude dont vous êtes le héros

Vous êtes stagiaire dans une entreprise de conseil. Le syndicat des commerçants du centre commercial régional Vélizy 2 a commandé à cette dernière une étude sur, notamment, le montant moyen des achats de leurs clients, afin d'ajuster leurs offres. Un exemple entre de nombreux autres est qu'en fonction de ce montant, il vaut peut-être mieux que les parfumeurs proposent des coffrets cadeaux à 40 euros plutôt qu'à 60 euros. Une question secondaire de l'étude est de mesurer la concurrence induite par les achats sur Internet : le centre se demande s'il doit ajouter une boutique virtuelle sur son site Internet, regroupant l'ensemble des produits vendus par les enseignes qui ne disposent pas déjà d'un telle boutique au niveau national (c'est le cas de la Fnac et d'Ikéa).

Votre méthodologie est de vous rapprocher le plus possible de votre terrain et c'est plein de bonne volonté que vous écrivez un questionnaire et vous rendez deux jours de suite, un samedi et un dimanche de décembre, au dit centre commercial (à quelques minutes à peine en voiture depuis HEC, ce qui vous permet de faire une bonne nuit). Vous vous êtes fixé l'objectif de 200 questionnaires remplis. Après dépouillement, il s'avère que seuls 172 d'entre eux sont exploitables (les autres ayant été remplis à moitié seulement et/ou avec une écriture illisible).

Vous aviez demandé dans la première partie questionnaire le montant moyen mensuel net des salaires et allocations par adulte du foyer, ainsi que le budget (éventuellement prévisionnel) consacré cette année aux cadeaux de Noël. Vous avez bravement rentré ces données sous `SPSS`, et cela se présente comme indiqué à la figure 1.

Dans la seconde partie du questionnaire, vous avez simplement demandé aux sondés s'ils effectuent cette année des compléments d'achats sur Internet en plus de leurs courses à Vélizy. (Il se trouve que le syndicat vous avait mis à disposition les résultats d'un sondage similaire effectué en 2007.)



	Index	Salaire	Achats	var	var
1	1	1330	375		
2	2	2840	1080		
3	3	590	180		
4	4	540	60		
5	5	790	105		
6	6	550	0		
7	7	1940	405		
8	8	2950	810		
9	9	520	0		
10	10	1770	450		
11	11	1180	450		
12	12	1650	330		
13	13	2790	930		
14	14	1150	285		
15	15	930	30		
16	16	570	135		
17	17	1110	375		
18	18	880	195		
19	19	1180	195		
20	20	1100	165		
21	21	1630	300		
22	22	770	180		
23	23	410	105		
24	24	2860	945		
25	25	2100	450		
26	26	800	210		
27	27	1070	300		
28	28	480	255		
29	29	540	165		
30	30	450	0		
31	31	2180	825		
32	32	2980	975		
33	33	660	75		
34	34	1270	225		
35	35	500	255		

FIGURE 1 – Vos données, telles qu'elles se présentent une fois rentrées sous SPSS.

Modélisation et première estimation

1. Etape préliminaire : modélisez le problème mathématiquement. En particulier, indiquez par quoi sont formées les observations, qui est la population ciblée, et quel(s) est (sont) le(s) paramètre(s) d'intérêt.
   Vous répondrez à cette question à la fois
   – par des formules mathématiques
   – *et* par des phrases d'explication ou de commentaires.
2. On veut un peu mieux connaître les données. Comment obtient-on la figure suivante sous SPSS ?

Statistics

		Salaire	Achats
N	Valid	172	172
	Missing	0	0
Mean		1386,22	376,22
Median		1180,00	300,00
Std. Deviation		697,503	259,820
Minimum		400	0
Maximum		2980	1080

3. Calculez alors un intervalle de confiance à 95 % sur le montant moyen du budget consacré aux achats de Noël.

Etude de la concurrence d'Internet

Cette année, sur les 172 sondés, 41 ont fait ou vont faire un complément de courses sur Internet. Ils étaient 35 sur 193 dans le sondage de l'an dernier.

1. Modélisez le problème mathématiquement.
2. La proportion de clients effectuant un tel complément a-t-elle significativement augmenté entre 2007 et 2008 ?
3. Que pensez-vous de la méthodologie que vous avez employée : prête-t-elle quelque part le flanc à la critique ? Si oui, indiquez comment il aurait fallu procéder. Si non, rappelez les principes généraux de conduite de sondage et expliquez en quoi ils sont ici respectés.

Construction d'un modèle prédictif/explicatif

1. On veut expliquer une variable en fonction de l'autre : qui en fonction de qui, d'après vous ?
2. On pense à un modèle linéaire : on demande à SPSS d'effectuer une régression linéaire. Des résultats obtenus sont indiqués dans les tables 1 et 2 : quel tableau retenez-vous ?
3. Ecrivez le modèle proposé par ce tableau. Est-ce un modèle satisfaisant ? Formulez tout autre commentaire pertinent.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,893[a]	,798	,797	314,519

a. Predictors: (Constant), Achats

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6,638E7	1	6,638E7	670,999	,000[a]
	Residual	1,682E7	170	98921,906		
	Total	8,319E7	171			

a. Predictors: (Constant), Achats

b. Dependent Variable: Salaire

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	484,069	42,285		11,448	,000
	Achats	2,398	,093	,893	25,904	,000

a. Dependent Variable: Salaire

TABLE 1 – Première régression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,893[a]	,798	,797	117,158

a. Predictors: (Constant), Salaire

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9210165,794	1	9210165,794	670,999	,000[a]
	Residual	2333427,811	170	13726,046		
	Total	1,154E7	171			

a. Predictors: (Constant), Salaire

b. Dependent Variable: Achats

Coefficients[a]

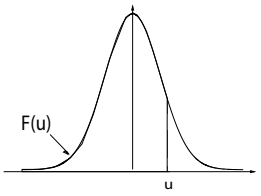
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-85,014	19,921		-4,268	,000
	Salaire	,333	,013	,893	25,904	,000

a. Dependent Variable: Achats

TABLE 2 – Seconde régression.

Table de la loi normale

Pour une valeur $u \geqslant 0$, la table ci-dessous renvoie la valeur $F(u)$ de la fonction de répartition F de la loi normale centrée réduite au point u.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Table pour les grandes valeurs de u :

u	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
u	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
u	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
u	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1

Examen de rattrapage, session 2008–09 : corrigé

Modélisation et première estimation

La question 1 est résolue par l'exercice 2.3, page 41. Les questions suivantes sont traitées par l'exercice 5.10, page 118.

Etude de la concurrence d'Internet

La question 1 est traitée par l'exercice 2.3, page 41. Les questions suivantes sont l'objet de l'exercice 9.3, page 239.

Construction d'un modèle prédictif/explicatif

Les questions posées ici sont résolues par l'exercice 13.1, page 333.

Examen principal, session 2007–08 : énoncé

Examen 2007 du cours "Eléments de statistique mathématique"

Groupes de Gilles Stoltz

Les trois exercices qui suivent sont indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix. Cependant, ils sont présentés dans un ordre qui me semble croissant de difficulté.

Dans l'ensemble du présent sujet, pour répondre à une question, on pourra admettre les résultats des questions précédentes (notamment à l'exercice II). Il est demandé de numéroter soigneusement les réponses.

Ne soyez pas troublés par la longueur du sujet d'examen ; il en sera tenu compte lors de la correction. Je ne m'attends pas à ce que vous ayez le temps de répondre à toutes les questions.

Durée : 2 heures – **Documents et calculatrice autorisés**

Exercice I

Une grande enseigne de la vente par correspondance a identifié dans son fichier de clients 50 000 foyers équipés soit d'un ordinateur, soit d'une console de jeu. Tous les trimestres depuis plusieurs années, un catalogue spécifique leur est envoyé, présentant des produits multimédia au même prix que les grandes enseignes comme la Fnac, Surcouf ou Micromania. La stratégie marketing avait reposé jusqu'à présent sur une politique de vente « satisfait ou remboursé », ainsi que dans l'envoi, avec le catalogue, d'un CD contenant des versions de démonstration de certains logiciels. Les résultats moyens (moyenne sur plusieurs trimestres) étaient pour l'instant les suivants :

- taux de commande moyen par trimestre : 13 %,
- chiffre d'affaire moyen par commande : 70 euros,
- soit une marge brute moyenne par trimestre de 182 000 euros (la marge brute par commande étant en moyenne de 40 % du chiffre d'affaires de la commande).

Pour le prochain trimestre, le nouveau directeur marketing, ancien d'HEC, suggère une politique de prix plus agressive, avec 5 % de remise sur tous les produits. Avant de prendre une décision définitive, il lance un test sur un échantillon de 1 000 clients. Le tableau suivant en donne les résultats. La question est bien évidemment de déterminer si cette promotion est rentable.

Taille de l'échantillon	1 000
Nombre de commandes	170
Montant moyen des commandes (*avant remise*)	73
Ecart-type du montant des commandes (*avant remise*)	8

1. Précisez la population, les variables et paramètres en jeu pour l'étude du taux d'achat et du chiffre d'affaire par client ; on formalisera donc un certain modèle statistique.
2. Donnez un intervalle de confiance à 95 % du taux de commande avec l'offre promotionnelle.
3. Peut-on dire que le taux de commande a augmenté par rapport aux trimestres précédents ? (Effectuez un test unilatère à préciser.)
4. Donnez un intervalle de confiance à 95 % du chiffre d'affaire moyen par commande (*avant remise*).
5. Déduisez des questions 2. et 4. un intervalle de confiance sur le chiffre d'affaire total que la société obtiendrait si elle envoyait la promotion à l'ensemble des clients concernés. De quel niveau est cet intervalle de confiance ?
6. Concluez : la promotion semble-t-elle rentable ?

Données : voici une table de quantiles z_β. On rappelle que z_β est le réel positif tel que $\mathbb{P}\{N \leqslant z_\beta\} = \beta$ lorsque N suit la loi $\mathcal{N}(0,1)$.

β	0.80	0.95	0.975	0.99	0.999
z_β	0.84	1.65	1.96	2.33	3.09

Exercice II

Avertissement : En 2007, nous utilisions encore le logiciel `R`, remplacé depuis par `SPSS`. Vous pouvez quand même faire cet exercice, il vous suffit de faire preuve d'un peu d'adaptabilité à la question 4.

Un laboratoire pharmaceutique veut tester l'efficacité de sa nouvelle formule de somnifère, DodoPlus, contre [1] celui le plus couramment prescrit et qui jouit de la meilleure réputation, Morpheus. Il recrute dix volontaires, et au jour J (respectivement, J+7), leur administre une pilule de DodoPlus (respectivement, Morpheus). Les longueurs des nuits (en heures), $x_1, \dots, x_{10}$ au jour J, puis $y_1, \dots, y_{10}$ au jour J+7, sont résumées dans le tableau suivant.

J	8.4	9.8	7.5	8.2	9.4	9.2	7.9	8.3	8.1	7.5
J+7	8.5	9.4	8.5	8.4	9.1	9.1	8.0	8.5	7.9	8.2

Le laboratoire axe sa compagne de communication autour des éléments suivants : DodoPlus est aussi efficace que Morpheus alors qu'il est bien moins cher.

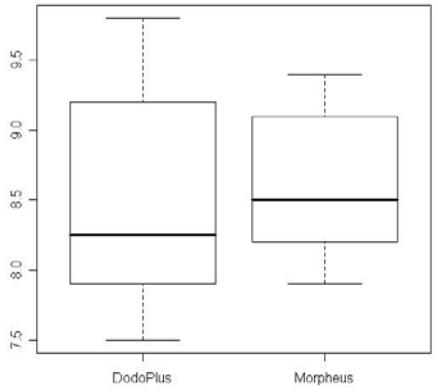


FIGURE 1 – Boîtes à moustache des données

1. Les données sont représentées à la figure 1. Quel est votre sentiment : votre intuition vous suggère-t-elle qu'un médicament est plus efficace qu'un autre ?
2. On modélise les données comme la réalisation des variables aléatoires $X_1, \dots, X_{10}$ et $Y_1, \dots, Y_{10}$. Expliquez pourquoi pour chaque j, l'observation X_j ne peut être indépendante de Y_j.
3. On considère les différences $Z_j = X_j - Y_j$. Indiquez pourquoi on peut dire, en revanche, que les $Z_1, \dots, Z_{10}$ sont indépendantes et identiquement distribuées.

1. Malheureusement, dans la vraie vie pharmaceutique, les nouveaux médicaments sont le plus souvent testés contre placebo, et non pas contre les formules déjà existantes...

4. On a envie de dire que les Z_j suivent une loi normale, de paramètres Δ et σ^2 inconnus. Quel est le nom du test pouvant vérifier cela ? Sous `R`, sa mise en œuvre donne :

```
> print(D)
     J  J7
1  8.4 8.5
2  9.8 9.4
3  7.5 8.5
4  8.2 8.4
5  9.4 9.1
6  9.2 9.1
7  7.9 8.0
8  8.3 8.5
9  8.1 7.9
10 7.5 8.2
> shapiro.test(D$J - D$J7)

        Shapiro-Wilk normality test

data:  D$J - D$J7
W = 0.9148, p-value = 0.3157
```

   Que conclure ?

5. Comment formuler, avec les paramètres introduits, un test qui soutienne l'affirmation du laboratoire ? Précisez bien l'hypothèse testée, ainsi que l'hypothèse alternative.

6. Achevez la construction du test et mettez-le en œuvre. Quelle est la conclusion du laboratoire au vu des données ?
   On donne la table des réalisations z_j des Z_j :

z_j	-0.1	0.4	-1.0	-0.2	0.3	0.1	-0.1	-0.2	0.2	-0.7

   La moyenne $\overline{z}_{10}$ des observations z_j vaut -0.13, leur variance (version débiaisée) est égale à $s^2_{z,10} = 0.19$.

   Voici enfin une table de quantiles $t_{k,\beta}$. On rappelle que $t_{k,\beta}$ est le réel positif tel que $\mathbb{P}\{T \leqslant t_{k,\beta}\} = \beta$ lorsque T suit la loi $\mathcal{T}_k$.

β	0.15	0.20	0.50	0.95
$k = 9$	-1.10	-0.89	0.00	1.83

7. Quelle est votre réaction ou sentiment face à cette conclusion ? Seriez-vous convaincu par l'affirmation suivante du laboratoire :

   > Des tests scientifiques montrent que DodoPlus est aussi efficace que les meilleurs somnifères, et il est bien moins cher, contribuant ainsi à la réduction des dépenses de santé.

Exercice III

Un agent immobilier se demande s'il lui faut, à certains mois de l'année, solliciter auprès des écoles des environs l'envoi de stagiaires (mal rémunérés) pour l'aider. L'opinion qu'il s'est forgée par l'expérience est en effet que la fréquence des ventes d'appartements (plus exactement, des signatures de promesses de vente) n'est pas uniforme au cours de l'année.

Ses tables donnent, mois par mois sur l'année écoulée, le nombre de mandats qui se sont conclus par une signature :

A	M	J	J	A	S	O	N	D	J	F	M
6	6	5	3	1	2	1	2	2	1	3	4

Ces données confirment-elles son sentiment empirique ?

Données : voici une table de quantiles $\chi^2_{k,\beta}$. On rappelle que $\chi^2_{k,\beta}$ est le réel positif tel que $\mathbb{P}\{X \leqslant \chi^2_{k,\beta}\} = \beta$ lorsque X suit la loi χ^2_k.

β	0.40	0.60	0.80	0.95	0.98	0.99
$k=2$	1.02	1.83	3.22	5.99	7.82	9.21
$k=3$	1.87	2.95	4.64	7.81	9.84	11.34
$k=4$	2.75	4.04	5.99	9.49	11.67	13.28

Examen principal, session 2007–08 : corrigé

Exercice I

La question 1 est traitée par l'exercice 2.4, page 41. Les questions 2, 4, 5 et 6 sont résolues par l'exercice 5.9, page 117, tandis que la question 3 est abordée par l'exercice 7.6, page 183.

Exercice II

Il correspond à l'exercice 9.4, page 239. (Il est à noter qu'un angle d'attaque différent de cet exercice, par des intervalles de confiance, a été proposé par l'exercice 5.12, page 118.)

Exercice III

Cet exercice est désormais un exemple du cours (il ne l'était bien sûr pas en 2007 !) ; vous trouverez sa correction page 259 (correction numérotée 10.1).

Examen de rattrapage, session 2007–08 : énoncé

Examen (de rattrapage) 2007 du cours "Eléments de statistique mathématique"

Groupes de Gilles Stoltz

Les trois exercices qui suivent sont indépendants les uns des autres et peuvent donc être abordés dans un ordre laissé au libre choix. Cependant, ils sont présentés dans un ordre qui me semble croissant de difficulté.

Dans l'ensemble du présent sujet (et notamment à l'exercice II), pour répondre à une question, on pourra admettre les résultats des questions précédentes. Il est demandé de numéroter soigneusement les réponses.

Durée : 2 heures – **Tous documents autorisés, calculatrice autorisée**

Exercice I

1. On lance 1 000 fois une pièce et on obtient 476 piles : la pièce est-elle biaisée ?
2. Quels sont les nombres limites (minimum et maximum) de piles à obtenir lors d'une expérience consistant à lancer 1 000 fois une pièce afin de rejeter l'hypothèse qu'elle est équilibrée ?

Exercice II

On veut quantifier l'évolution du pouvoir d'achat et son impact sur le moral des Français.

En 2004, le montant moyen des achats hors produits de nécessité (par exemple, voyages, abonnement internet, téléphonie mobile, spectables, et contrairement à l'alimentation, logement, voiture, chauffage, eau, téléphonie fixe) était de 637 euros par mois et par foyer selon les données collectées par l'INSEE auprès de plusieurs millions d'habitants lors du recensement partiel.

Le gouvernement commande un sondage téléphonique auprès de 2 000 foyers environ (1 999 seront interrogés) afin de déterminer ce montant en 2007 ; seuls 1 837 de ces foyers arrivent le calculer et à l'indiquer à l'opérateur. Le montant moyen qu'ils déclarent est de 598 euros (avec un écart-type mesuré dans les montants de 254 euros).

Par ailleurs, il est demandé aux sondés s'ils sont optimistes ou non quant à leur capacité d'achat pour les années à venir. Les résultats (absolus) sont présentés par catégories d'âge.

Age	Optimistes	Pas optimistes	Pas d'opinion
20 – 40	237	392	13
40 – 60	326	298	32
60 et +	362	258	81

1. Donner un intervalle de confiance à 95 % du budget actuellement consacré aux prestations hors produits de nécessité.
2. En supposant une inflation de 2 % par an, peut-on dire que le montant consacré aux achats hors produits de nécessité a diminué ? Faire un test d'hypothèses en précisant bien les hypohtèses mises en jeu, la statistique de test et la zone de rejet.
3. Peut-on dire que l'optimisme des Français est indépendant de l'âge ?

Exercice III

On étudie la circonférence d'orangers en fonction de leur âge. On obtient les résultats de mesures suivants.

ID	age	circumference
1	118	30
2	484	58
3	664	87
4	1004	115
5	1231	120
6	1372	142
7	1582	145
8	118	33
9	484	69
10	664	111
11	1004	156
12	1231	172
13	1372	203
14	1582	203
15	118	30
16	484	51
17	664	75
18	1004	108
19	1231	115
20	1372	139
21	1582	140
22	118	32
23	484	62
24	664	112
25	1004	167
26	1231	179
27	1372	209
28	1582	214
29	118	30
30	484	49
31	664	81
32	1004	125
33	1231	142
34	1372	174
35	1582	177

On effectue ensuite sous SPSS la régression linéaire de la circonférence en fonction de l'âge ; on obtient le résumé de régression reproduit à la page suivante.

1. Quelle est la relation proposée entre circonférence et âge ?
2. Y a-t-il une dépendance forte entre la circonférence et l'âge ?
3. Commentez le nombre 17.400 du résumé de régression (dernier tableau, première case),

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,914[a]	,835	,830	23,738

a. Predictors: (Constant), Age

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	93771,541	1	93771,541	166,416	,000[a]
	Residual	18594,744	33	563,477		
	Total	112366,286	34			

a. Predictors: (Constant), Age

b. Dependent Variable: Circumference

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	17,400	8,623		2,018	,052
	Age	,107	,008	,914	12,900	,000

a. Dependent Variable: Circumference

Examen de rattrapage, session 2007–08 : corrigé

Exercice I

Il correspond à l'exercice 7.7, page 183.

Exercice II

La question 1 a été traitée par l'exercice 5.11, page 118. L'exercice 7.2, page 172, a quant à lui abordé la question 2. Enfin, la question 3 est résolue par l'exercice 10.6, page 279. (Il est à noter que la modélisation des différentes données introduites n'était pas explicitement demandée ici mais a été effectuée dans l'exercice 2.5, page 41.)

Exercice III

Il correspond à l'exercice 13.2, page 333.

Seizième Partie

Fiches de synthèse

Rappel des contenus étudiés partie après partie

Dans ce polycopié, nous avons étudié les points suivants, dans cet ordre :

1. Panorama de la démarche statistique (par construction, les sondages se trompent parfois, par exemple, lors de l'élection municipale 2007 dans le cinquième arrondissement de Paris) ; rappels de probabilités du cours de classes préparatoires.
2. Recueil, représentation et modélisation de données : de la collecte des données au seuil du traitement mathématique ; ce dernier permet de dire des choses sur des paramètres de population sans pour autant interroger tous les éléments de cette dernière : il suffit d'en considérer un échantillon aléatoire ! En pratique, il est difficile de constituer un tel échantillon (de le tirer sans biais dans la population d'intérêt, cf. les sondages téléphoniques).
3. Notions fondamentales de statistique : estimation (estimateurs, utilisés en théorie, vs. estimées, réalisations des estimateurs sur les données) et quantiles des lois usuelles.
4. Les intervalles de confiance, ou comment briller en entreprise non seulement en ne donnant pas trop de chiffres après la virgule, mais aussi et surtout, en procurant une estimée assortie d'une marge d'erreur ; cf. Georges Elgozy : « Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales. »
5. Les tests d'hypothèses, qui permettent de quantifier l'attachement à une assertion de départ (choisie subjectivement) face à une assertion alternative ; on a vu l'utilisation potentiellement politique et polémique des tests (l'exemple des faiseurs de pluie) : on n'apprend essentiellement des tests que lorsqu'ils rejettent l'hypothèse de départ. La quantité-clé associée à un test est sa P-valeur.
6. Les tests d'hypothèses pour comparer deux séries de données, indépendantes ou appariées : ils permettent par exemple de comparer les effets de deux médicaments ou de deux crèmes hydratantes (données appariées), les taux d'accidents des hommes et des femmes, les consommations d'alcool aux POW des deux groupes d'étudiants, etc. (données indépendantes).
7. Les tests du χ^2, qui permettent d'une part (tests d'ajustement simple) de détecter de nombreuses tricheries (électorales en Iran, fraudes fiscales, données embellies par le généticien Mendel) ; et d'autre part (tests d'indépendance), de regarder si les opinions politiques dépendent de l'année de scolarité à HEC ou si deux enseignants notent de la même façon ou non.
8. La régression linéaire simple, qui est un premier modèle d'explication d'une variable quantitative (prix d'un appartement ou prix d'un forfait de ski) comme fonction affine d'une autre variable (la surface de l'appartement ou la taille du domaine skiable) à quoi s'ajoute un aléa de modélisation.

9. La régression linéaire multiple, qui part du même principe mais considère cette fois-ci plusieurs variables quantitatives et non plus une seule, ... ce qui rend les choses mathématiquement plus compliquées à écrire et à se représenter, mais dont le résultat est facilement lisible dans les sorties SPSS.

Panorama de la démarche statistique (cf. partie 1)

A retenir pour le reste de votre carrière

Le champ de la statistique est large, c'est même une discipline fondamentalement citoyenne qui permet de voir les chiffres avec du recul et de ne pas se laisser berner par un discours de communication (publicitaire, médiatique ou gouvernementale, quelle que soit l'orientation politique du gouvernement). La bonne pratique de la statistique montre que les chiffres ne parlent pas d'eux-mêmes et qu'il s'agit de quantifier des impressions. Bien plus, la statistique est une école de la modestie puisqu'à l'occasion, et par construction même, elle se trompe.

A retenir pour la suite du cours et pour l'examen

Le premier cours était un cours d'introduction et d'échauffement, faisant le lien avec les (bons ou mauvais) souvenirs de votre vie avant le concours. Il n'est pas nécessaire de retenir les constructions d'intervalles de confiance qui y avaient été exhibées car elles ont été revues plus en détails dans les cours ultérieurs. Il faut même définitivement oublier l'utilisation de l'inégalité de Chebychev-Markov, dont on a montré qu'elle conduisait à des intervalles trop grossiers.

En revanche, il faut connaître la version "universitaire" de la loi des grands nombres et du théorème de la limite centrale (et effacer les énoncés équivalents mais moins adaptés vus en classes préparatoires) :

THÉORÈME 16.1 (Loi des grands nombres). *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes et identiquement distribuées admettant une espérance* μ. *Alors la moyenne empirique converge vers l'espérance,*

$$\overline{X}_n \overset{\text{not.}}{=} \frac{1}{n}(X_1 + \ldots + X_n) \xrightarrow{\mathbb{P}} \mu \ .$$

THÉORÈME 16.2 (Théorème de la limite centrale). *Soient* $X_1, \ldots, X_n$ *des variables aléatoires indépendantes et identiquement distribuées, de loi commune admettant un moment d'ordre deux, d'espérance et de variance communes notées* μ *et* σ^2. *Alors, on a la convergence en loi*

$$\frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu) \rightharpoonup \mathcal{N}(0,\,1) \ .$$

On rappelle la définition de la convergence en loi (vers la loi normale standard) : pour tout intervalle $[a,\, b]$, on a

$$\mathbb{P}\left\{a \leqslant \frac{\sqrt{n}}{\sigma}\,(\overline{X}_n - \mu) \leqslant b\right\} \longrightarrow \mathbb{P}\{a \leqslant Z \leqslant b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx \ .$$

Recueil, représentation et modélisation de données (cf. partie 2)

In God we trust, all others bring data.
Edwards Deming (universitaire américain, consultant pour l'industrie, 1900–93)

A retenir pour le reste de votre carrière

Avant de vous lancer sur un marché, il faut identifier une population cible puis la cerner, à coups de sondages (aléatoires). Vous définissez un questionnaire puis en reportez les réponses dans une grande matrice, en codant les données qualitatives selon une certaine table de correspondance. La matrice ainsi obtenue constituera le matériau brut de votre étude statistique et sera la source de votre crédibilité : *In God we trust, all others bring data.* Et surtout, ne faites pas comme cet ancien élève qui est venu me voir, sans données, sans étude préalable, et qui voulait que je l'aide à modéliser le marché auquel il s'intéressait : la statistique se fonde sur des données, ce n'est pas de la voyance !

A retenir pour la suite du cours et pour l'examen

Il suffit de retenir de cette partie quelques grandes idées et quelques mots de vocabulaire. Premièrement, il faut être conscient des différents types de données possibles.

Données			
Qualitatives		Quantitatives	
Nominales	Ordinales	Discrètes	Continues
Marque de voiture	Rang d'un classement	Nombre d'enfants	Salaire
Sexe	Niveau d'éducation	Nombre de ventes	Taille
Statut conjugal	Degré de satisfaction		

TABLE 4. Résumé de la discussion sur les différents types de données. Note : les données qualitatives doivent être codées par des entiers (1, 2, 3, etc.) lors de leur traitement sous un logiciel statistique.

Ensuite, il faut connaître le principe de la chaîne de recueil, description et traitement des données, décrite à la figure 74 ; les séances suivantes du cours portent sur les statistiques inférentielles mentionnées dans la dernière case. Nous avons en revanche vu dans la partie 2 comment recueillir, décrire et modéliser des données.

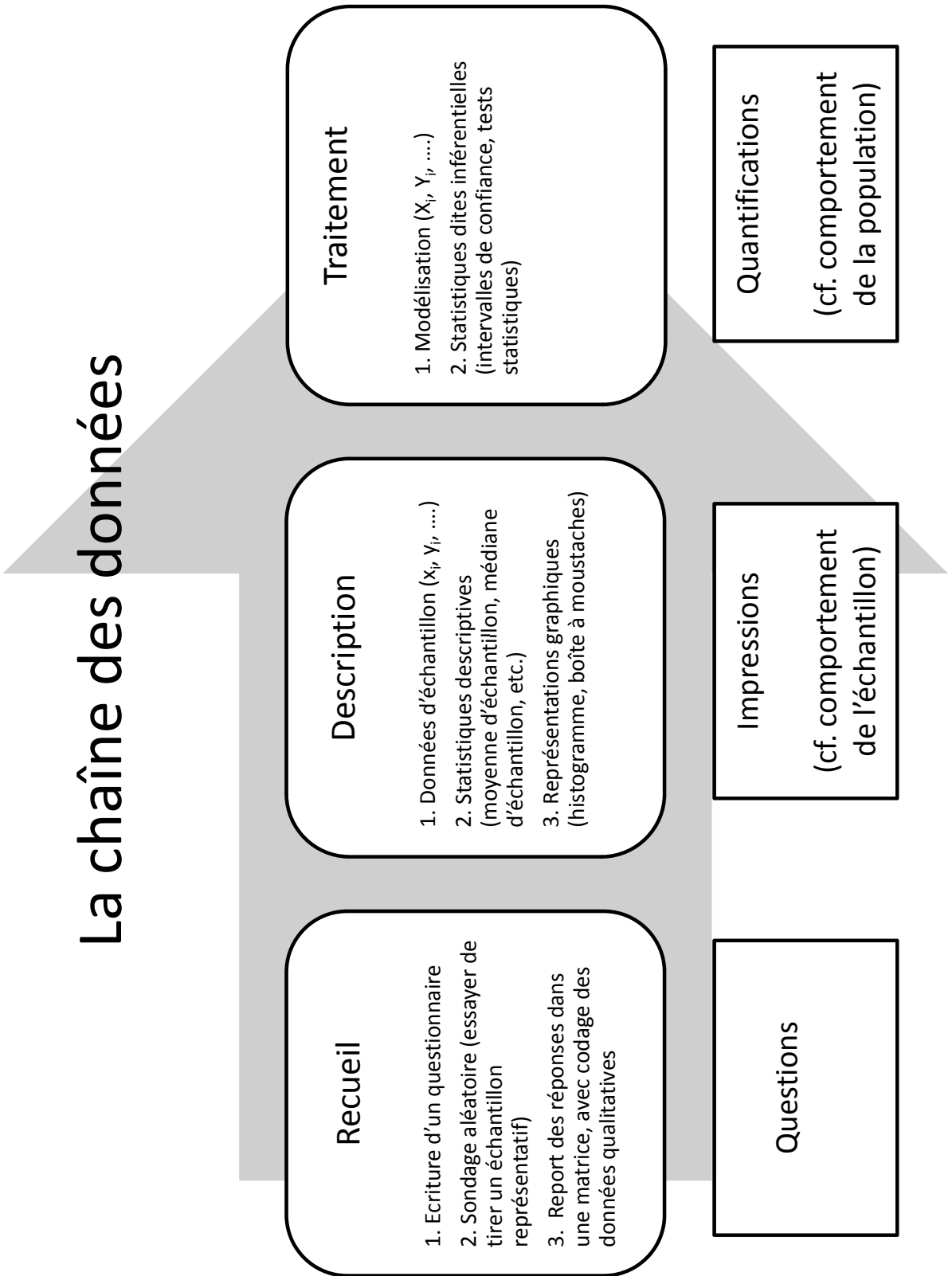


FIGURE 74. La chaîne de recueil, description et traitement des données.

Pour la représentation graphique des données, une méthode nouvelle est apparue en la personne des boîtes à moustaches, dont il faut connaître le principe de construction.

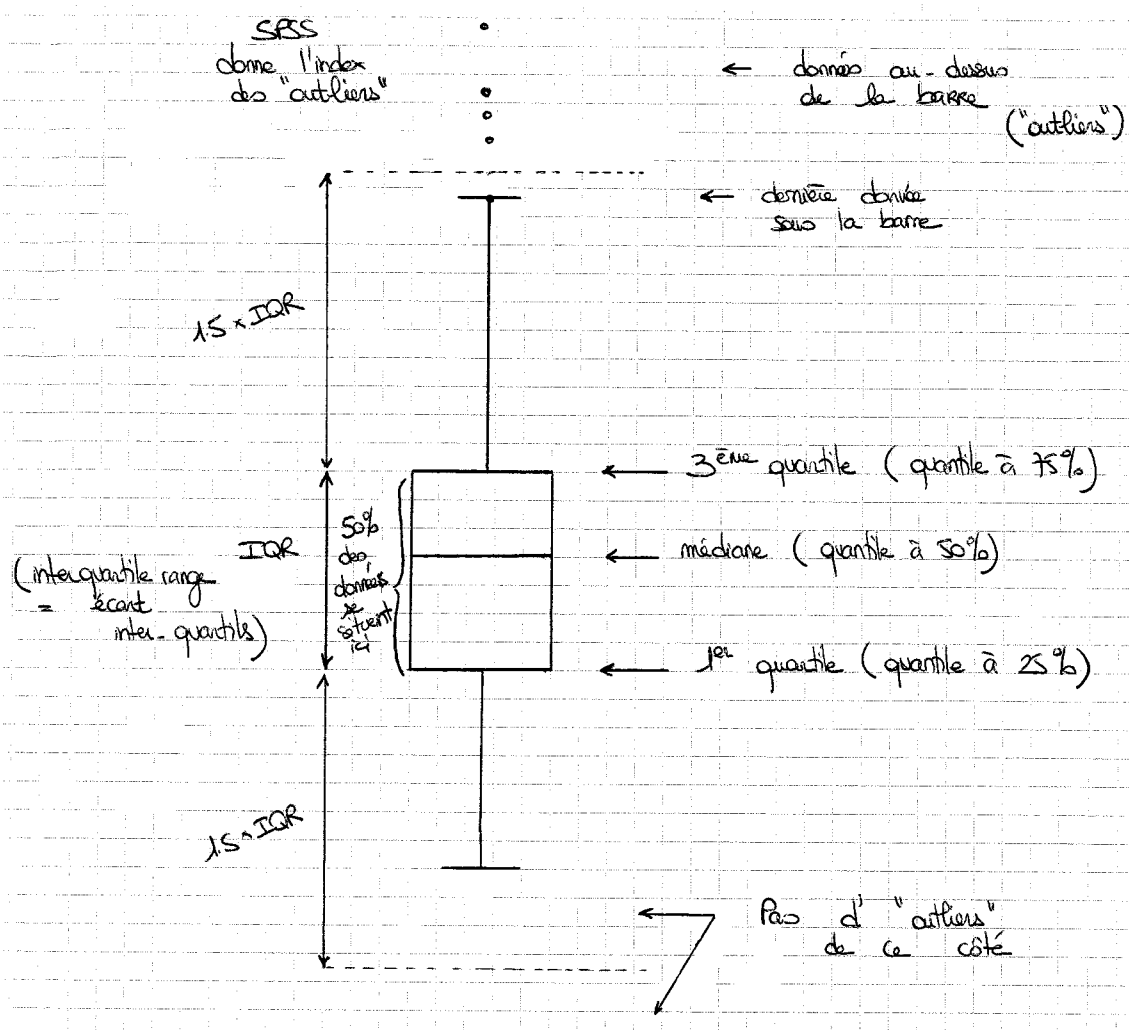


FIGURE 75. Principe de construction des boîtes à moustaches.

Pour le recueil et la modélisation des données, on retiendra les faits suivants :

- lors du recueil des données, il est notamment bon d'interroger des individus au hasard, dans une grande population : cela conduit à une modélisation des données $x_1, \ldots, x_n$ comme la réalisation de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi commune, qui est celle qui gouverne la population ; on parle de tirage d'un échantillon au hasard ;
- les deux lois particulières que nous verrons le plus souvent pour cette loi de population sont la loi normale (de paramètres μ_0 et σ_0^2) et la loi de Bernoulli (de paramètre p_0) : la première modélise des situations où un grand nombre de paramètres influent sur la valeur d'une donnée, la seconde est utilisée pour les sondages ;
- lorsque l'on ne sait pas la forme de la loi de la population, on va juste s'intéresser à son espérance μ_0 ou à son écart-type σ_0.

Il faut en particulier bien distinguer la population, qui nous intéresse, de l'échantillon qu'on en tire ; on observe diverses quantités sur ce dernier et à partir d'elles, on veut mieux connaître les quantités correspondantes de la population.

En pratique, sur un exercice, la modélisation est une démarche en six temps, rappelée dans la figure suivante.

La démarche de modélisation

1. Définir la population effectivement visée.
2. Rappeler la table de correspondance (comment on code les données qualitatives proposées par l'énoncé, lorsqu'il y en a) ;
3. Préciser les données (leur donner un nom, rappeler leur nombre, définir leur étendue, i.e., l'ensemble des valeurs pouvant être prises).
4. Décrire les données (rappeler les statistiques d'échantillon disponibles).
5. Modéliser les données, au vu de la méthode de recueil utilisée : souvent, sous la forme de réalisations de variables aléatoires indépendantes et identiquement distribuées, selon une loi dont on précisera si on en connaît la forme (normale, Bernoulli) ou pas.
6. Préciser les paramètres d'intérêt de cette loi et surtout, les interpréter en termes de comportement de la population (en une phrase, donner le sens et le but de la connaissance de ces paramètres). Ces paramètres sont inconnus ; il s'agit souvent de la moyenne de certaines quantités sur l'ensemble de la population.

FIGURE 76. Démarche de modélisation à reprendre avant de traiter tout exercice.

Estimation ponctuelle et quantiles des lois usuelles (cf. partie 4)

Une fois n'est pas coutume, ce qu'il s'agit de retenir pour la suite de votre carrière d'une part, et pour la suite du cours d'autre part, forment deux ensembles disjoints !

A retenir pour le reste de votre carrière

Pas grand-chose dans la version rédigée du cours, ... qui a préparé uniquement le terrain pour les parties suivantes. En revanche, dans les compléments facultatifs, ceux qui auront lu le paragraphe sur les différentes manières d'estimer la tendance centrale ont dû être convaincus qu'il faut refuser la dictature de la moyenne, car cette dernière est trop sensible aux données extrêmes (dites également atypiques, ou "outliers"). Il faut penser notamment à la médiane et à la moyenne calculée sur les 95 % d'observations les plus centrales ; mais il existe de nombreux autres estimateurs possibles, dits robustes. On n'a que l'embarras du choix, et c'est bien le problème : il n'y a pas d'unique meilleure estimée...

A retenir pour la suite du cours et pour l'examen

Estimation. On se place dans le cadre d'un modèle $X_1, \ldots, X_n$ de variables aléatoires indépendantes et identiquement distribuées selon une certaine loi $\mathbb{P}_{\theta_0}$, appartenant à la famille $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$.

Définition 16.1 (Estimateur). *Un estimateur est toute variable aléatoire construite uniquement à partir des observations* $X_1, \ldots, X_n$. *En particulier, il ne doit pas dépendre de quantités inconnues, telles que* θ_0 *ou* $\mathbb{P}_{\theta_0}$.

Définition 16.2 (Estimée). *Une estimée est la réalisation d'un estimateur sur les données* $x_1, \ldots, x_n$. *Autrement dit, l'estimée est la valeur que l'on peut calculer en remplaçant les* X_j *par les* x_j *dans la définition de l'estimateur correspondant.*

Définition 16.3 (Estimateur sans biais). *Un estimateur* $\widehat{g}_n$ *de* $g(\theta_0)$ *est dit sans biais lorsque, quel que soit le vrai paramètre* θ_0,

$$\mathbb{E}[\widehat{g}_n] = g(\theta_0)\ ;$$

il est dit biaisé sinon.

Définition 16.4 (Estimation consistante). *Une suite* $(\widehat{g}_n)$ *d'estimateurs de* $g(\theta_0)$ *est dite consistante lorsque*

$$\widehat{g}_n \xrightarrow{\mathbb{P}} g(\theta_0)\ .$$

Exemple 16.1 (Cas général). Dans tout modèle, l'estimateur

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$$

est un estimateur sans biais de la variance σ_0^2. Il est également consistant.

EXEMPLE 16.2 (Cas des sondages). Dans un modèle de Bernoulli, on utilise souvent l'estimateur (biaisé) de la variance

$$\overline{X}_n(1-\overline{X}_n) = \frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{X}_n)^2 \ .$$

Il est consistant.

Quantiles d'une loi. Il faut :
- connaître la notion de quantile, et savoir en placer sur un graphique de densité (voir les nombreux graphiques présentés dans la version rédigée du cours et dans les exercices) ;
- utiliser les notations usuelles z_β pour la loi normale, $t_{k,\beta}$ pour les lois de Student et $c_{k,\beta}$ (ou $\chi^2_{k,\beta}$) pour les lois du χ^2 ;
- savoir lire des tables de quantiles ou de fonctions de répartition de lois (nous avons vu ensemble comment procéder lors des exercices).

Estimation par intervalles (cf. partie 5)

Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.
Georges Elgozy (économiste français, 1909–1989)

A retenir pour le reste de votre carrière

La morale de ce cours est simple : il ne faut jamais plus donner un nombre (une estimée) sans précision. En particulier, il faut se méfier comme de la peste des nombres avec beaucoup de chiffres après la virgule ! Remettez donc sèchement en place vos futurs collègues lorsqu'ils tomberont dans ce travers : « Certes, vous me proposez cette estimation, mais avec quelle précision je vous prie ? » Mieux, clouez-leur le bec en lâchant dans un soupir la citation d'Elgozy : « Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales. » Attention ! Votre boss risque de vous prendre pour un génie des maths : et il aura raison, car si vous procédez ainsi, c'est que vous avez bien mieux compris votre cours de statistiques que vos collègues...

A retenir pour la suite du cours et pour l'examen

Dans ce qui suit, on se place à nouveau dans le cadre de variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbb{P}_{\theta_0}$. On s'intéresse à l'estimation d'une certaine quantité $g(\theta_0)$ fonction de cette loi.

Définition 16.5. *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est la donnée d'un couple d'estimateurs $\widehat{g}_n \leqslant \widehat{g}'_n$, à partir de qui on construit*

$$\widehat{I}_n = \left[\widehat{g}_n,\ \widehat{g}'_n\right] .$$

Définition 16.6 (Niveau). *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est dit de niveau au moins égal à $1-\alpha$, où $\alpha \in [0,1]$, si, quelle que soit la valeur de θ_0,*

$$\mathbb{P}\left\{g(\theta_0) \in \widehat{I}_n\right\} \geqslant 1-\alpha .$$

Définition 16.7 (Niveau asymptotique). *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est dit de niveau asymptotique au moins égal à $1-\alpha$, où $\alpha \in [0,1]$, si quel que soit θ_0, pour tout $\varepsilon > 0$, il existe un rang N tel que pour tout $n \geqslant N$,*

$$\mathbb{P}\left\{g(\theta_0) \in \widehat{I}_n\right\} \geqslant 1-\alpha-\varepsilon .$$

Note : c'est le cas en particulier lorsque l'on peut garantir

$$\lim_{n\to\infty} \mathbb{P}\left\{g(\theta_0) \in \widehat{I}_n\right\} = 1-\alpha .$$

La définition 16.7 est cependant plus générale, au sens où il n'y est pas nécessaire qu'une limite existe.

Vue d'ensemble des résultats. La taille des intervalles de confiance mesure la précision de l'estimation. Elle est fonction de l'écart-type et de la taille d'échantillon : plus précisémenent, elle est typiquement proportionnelle à $s_n/\sqrt{n}$, où s_n est une estimée de l'écart-type. C'est pour cela que l'écart-type est une statistique si importante et qu'on dit qu'il mesure la dispersion des données.

La figure 77 résume mal ceci : elle ne s'intéresse qu'à la précision en fonction de la taille d'échantillon mais omet la notion d'écart-type. En revanche, elle a bien raison de souligner que la qualité de l'estimation est indépendante de la taille de la population.

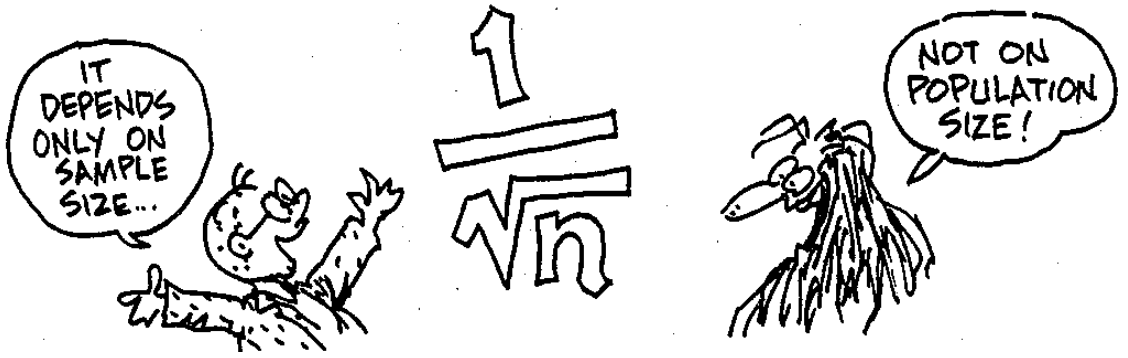


FIGURE 77. La qualité de l'estimation ne dépend que de la taille de l'échantillon, pas de celle de la population dans laquelle il a été tiré.

Rappel. Dans l'application des formules ci-dessous, on veillera à arrondir les résultats obtenus, de sorte à ne pas présenter trop de chiffres significatifs.

Estimation d'une moyenne pour un échantillon de taille n grande ($n \geqslant 30$). On a les intervalles de confiance suivants.

COROLLAIRE 16.1. *Soient des variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi admettant un moment d'ordre deux et d'espérance notée μ_0. On note z_β le β−quantile de la loi $\mathcal{N}(0,1)$. Alors les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ pour la moyenne μ_0 :*

$$\left[\overline{X}_n - z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ \overline{X}_n + z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \overset{\text{not.}}{=} \left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] ;$$

$$\left]-\infty,\ \overline{X}_n + z_{1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] ;$$

et

$$\left[\overline{X}_n - z_{1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}},\ +\infty\right[.$$

Estimation d'une proportion pour un échantillon de taille n grande ($n \geqslant 30$). On a les intervalles de confiance suivants.

COROLLAIRE 16.2. *Soient des variables aléatoires $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi $\mathcal{B}(p_0)$ de Bernoulli de paramètre p_0. On note*

z_β le β–quantile de la loi $\mathcal{N}(0,1)$. Alors les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ pour la proportion p_0 :

$$\left[\overline{X}_n \pm z_{1-\alpha/2}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right] \ ;$$

$$\left[0, \quad \overline{X}_n + z_{1-\alpha}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right] \ ;$$

et

$$\left[\overline{X}_n - z_{1-\alpha}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \quad 1\right] \ .$$

Estimation d'une moyenne pour un échantillon de taille n petite ($n \leqslant 30$). Le principe sous-jacent est le suivant.

Définition–Théorème 16.1. *Soit $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2. Alors la loi de*

$$\frac{\overline{X}_n - \mu_0}{\sqrt{\widehat{\sigma}_n^2/n}} = \sqrt{\frac{n}{\widehat{\sigma}_n^2}}\,(\overline{X}_n - \mu_0)$$

est indépendante de μ_0 et σ_0^2 ; on l'appelle la loi de Student à $n-1$ degrés de liberté, et on la note $\mathcal{T}_{n-1}$.

On a comme conséquence de ce principe les intervalles suivants.

Corollaire 16.3. *On note $t_{n-1,\beta}$ le β–quantile de la loi $\mathcal{T}_{n-1}$. Alors, sous les hypothèses de la définition–théorème précédente, les intervalles suivants sont des intervalles de confiance de la moyenne μ_0, non asymptotiques et exactement de niveau $1-\alpha$,*

$$\left[\overline{X}_n - t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}, \ \overline{X}_n + t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \overset{\text{not.}}{=} \left[\overline{X}_n \pm t_{n-1,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \ ;$$

$$\left]-\infty, \ \overline{X}_n + t_{n-1,1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}\right] \ ;$$

et

$$\left[\overline{X}_n - t_{n-1,1-\alpha}\sqrt{\frac{\widehat{\sigma}_n^2}{n}}, \ +\infty\right[\ .$$

Planification d'expériences. Pour garantir une précision ε fixée sur l'estimation d'un paramètre, valant à un niveau $1-\alpha$:

– s'il s'agit d'une proportion p_0, on prend

$$n \geqslant \left(\frac{z_{1-\alpha/2}}{2\varepsilon}\right)^2 \ ;$$

cependant, une fois l'expérience réalisée, on pourra recourir aux intervalles de confiance plus précis du corollaire 16.1 ;

– dans le cas d'une moyenne μ_0, on estime tout d'abord par s_t l'écart-type de la population σ_0 sur une premier échantillon de taille t, puis on tire un second échantillon

de taille n telle que

$$z_{1-\alpha/2}\sqrt{\frac{s_t^2}{n}} \leqslant \varepsilon \ , \qquad \text{soit} \quad n \geqslant s_t^2 \left(\frac{z_{1-\alpha/2}}{\varepsilon}\right)^2 \ .$$

Intervalles simultanés : méthode de Bonferroni. Si $\widehat{I}$ est un intervalle de confiance de niveau $1-\alpha$ pour un paramètre μ_0 et $\widehat{J}$ est un intervalle de confiance de niveau $1-\beta$ pour un paramètre μ_0', alors l'événement donné par les appartenances simultanées $\mu_0 \in \widehat{I}$ et $\mu_0' \in \widehat{J}$ survient avec probabilité au moins $1-(\alpha+\beta)$.

On en déduit des intervalles de confiance de niveau $1-(\alpha+\beta)$ sur toutes les quantités issues de μ_0 et μ_0', comme on l'a vu dans les exercices. On peut proposer la formulation générale suivante. Soit g une fonction de deux variables : on estime $g(\mu_0, \mu_0')$ par le sous-ensemble

$$\left[\min_{\widehat{I}\times\widehat{J}} g, \quad \max_{\widehat{I}\times\widehat{J}} g\right] .$$

Ce qui précède vaut en toute généralité, qu'il y ait indépendance ou non entre les observations ayant servi à construire $\widehat{I}$ et celles pour $\widehat{J}$. C'est la méthode de Bonferroni.

Remarques de conclusion et d'interprétation. Proposer un intervalle de confiance de niveau $1-\alpha$, c'est donc faire un pari, que l'on a une probabilité $1-\alpha$ au moins de gagner (dit autrement, que l'on a un risque au plus α de perdre !). Cependant, on n'observe jamais le résultat du pari : on ne sait jamais si la réalisation de l'intervalle proposé contient effectivement ou non le paramètre d'intérêt.

Construire un intervalle de confiance nécessite un arbitrage entre trois quantités : le niveau $1-\alpha$, la taille d'échantillon n et la précision ε. La donnée de deux de ces quantités fixe la troisième. Par exemple,

- en planification, on fixe α et ε et l'on détermine alors n ;
- en exploitation de résultats, on dispose de n et on se fixe α, après quoi on peut indiquer la précision ε que l'on garantit.

Tests de comparaison d'une moyenne à une valeur de référence (cf. partie 7)

Il ne faut pas utiliser les statistiques comme les ivrognes utilisent les réverbères : pour s'appuyer et non s'éclairer.
Lord Thorneycroft (homme politique britannique, 1909–1994)

A retenir pour le reste de votre carrière

La citation précédente montre que si les tests statistiques peuvent éventuellement mettre en lumière certains aspects et éclairer partiellement une décision (notamment lorsqu'ils rejettent l'hypothèse de départ, moins lorsqu'ils la conservent), leur mise en œuvre et son résultat doivent toujours être supervisés par un homme ou une femme, à qui revient la responsabilité de prendre une décision. En effet, les statistiques ne doivent pas être source de technocratie : le statisticien n'a pas vocation à se substituer au dirigeant. Réciproquement, le dirigeant ne doit pas demander au statisticien de prendre une décision à sa place.

Nous avons également souligné la force des préjugés et intentions politiques lors du choix des hypothèses à tester, puisque l'hypothèse de départ H_0 tend à être conservée.

A retenir pour la suite du cours et pour l'examen

Principes généraux sur les tests. Si les tests d'hypothèses peuvent mettre en évidence certains phénomènes, dans d'autres cas, ils ne peuvent se prononcer. En gros, le statisticien se pose une question (qui est H_0), et le test lui répond « ce n'est pas impossible » (conservation de H_0) ou « c'est impossible » (rejet de H_0 et passage à H_1).

Mais dans le premier cas, on n'est pas sûr d'avoir mis le doigt sur la vérité : on n'a avancé qu'une assertion qui ne contredisait pas les observations. Dans le second cas, en revanche, la connaissance fait un progrès : on sait que telle assertion ne peut être tenue pour vraie. C'est un progrès négatif.

La méthodologie des tests est rappelée à la page suivante.

Méthodologie des tests statistiques

1. Etape préliminaire : modélisation du problème.
2. Détermination des hypothèses à tester H_0 et H_1 (selon le contexte, et selon votre intuition et bon sens).
3. Choix d'une statistique de test T_n, dont on connaît la loi sous H_0 (voir les principes rappelés dans les pages suivantes).
4. Etude du comportement de T_n sous H_1 (*idem*) et déduction de la forme de la zone de rejet R.
5. Calcul de cette zone R pour un niveau fixé puis confrontation aux données ; et/ou calcul de la P-valeur du test sur les données.
6. Conclusion statistique : conservation ou rejet de l'hypothèse de départ H_0 et commentaire éventuel sur la P-valeur.
7. Conclusion stratégique : décision (en termes de management) ou action à entreprendre une fois éclairé par le résultat statistique de l'étape précédente.

Principe des tests : Un test confronte le modèle postulé par H_0 à des observations ; si la confrontation se passe mal, i.e., si elle indique que les données semblent contredire H_0, alors on passe au modèle indiqué par H_1.

DÉFINITION 16.8 (Erreur de première espèce ; P-valeur). *L'erreur de première espèce α d'un test est la probabilité de rejeter à tort H_0 lorsqu'elle est vraie.*

Etant donné un test statistique et des données, la P-valeur p est l'erreur maximale α telle que le test considéré accepterait encore la valeur réalisée de la statistique de test sur les données. Elle peut être interprétée comme un indice de crédibilité de H_0. Une faible P-valeur conduit au rejet de H_0.

En pratique, on calcule la P-valeur en déterminant la taille (en probabilité) de la zone de rejet qui s'arrêterait exactement à la valeur observée de la statistique de test.

Choix des hypothèses : H_0 est une hypothèse communément admise ou correspondant à un comportement prudent. On contrôle la probabilité de rejeter à tort H_0 (c'est α, qui est faible par construction), mais pas celle de conserver à tort H_0 (lorsque c'est H_1 qui est vraie). H_0 est donc également appelée l'hypothèse conservative, parce qu'un test a tendance à la conserver, sauf lorsque les données la contredisent gravement.

Conclusion statistique : Voir le tableau d'interprétation en fonction de la P-valeur page suivante.

Conclusion statistique en fonction de la P*-valeur*

Rappel : la P-valeur correspond au degré de crédibilité de l'hypothèse H_0 face à H_1.

Premier cas : vous êtes le statisticien et c'est votre supérieur qui doit prendre la décision finale.
- Ce n'est donc pas à vous d'endosser la responsabilité de la décision.
- Refusez que votre supérieur se réfugie derrière les statistiques : ne lui dites pas que votre test conserve ou rejette H_0.
- Précisez-lui plutôt la P-valeur que vous avez calculée et laissez-le prendre ensuite la décision en son âme et conscience.

Second cas : vous êtes le supérieur hiérarchique d'un statisticien et vous êtes responsable d'une décision à prendre.
- C'est à vous que l'on demandera des comptes, aussi, demandez à votre subordonné d'éclairer le mieux possible votre décision.
- En conséquence, refusez que ce dernier la prenne à votre place en vous disant simplement si son test conserve ou accepte H_0 et réclamez-lui l'indication d'une P-valeur p.
- Fixez-vous un seuil d'interprétation : 1 % si les conséquences d'un rejet à tort de H_0 seraient dramatiques, 5 % sinon dans les cas plus standards.
- Ensuite,
  + Si p est bien plus petit que ce seuil, rejetez H_0 fermement ;
  + Si p est bien plus grand que ce seuil, campez fermement sur H_0 ;
  + Si p est autour de ce seuil, pas de chance, un abîme de doute se dessine devant vous et il va falloir soit se décider tout de suite et prendre un risque (avec la perspective d'assumer une erreur éventuelle), ou commander une étude complémentaire et reporter la décision.

Principes particuliers : tests de comparaison à une valeur de référence.

Cas d'une loi générale et d'une taille d'échantillon grande.

PRINCIPE 16.1. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*
Données : $x_1, \ldots, x_n \in \mathbb{R}$
Modélisation associée : observations $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi, admettant un moment d'ordre deux et d'espérance notée μ_0
Hypothèse H_0 **:** $\mu_0 = \mu_{ref}$
Statistique de test :

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_{ref}}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 **:** $T_n \rightharpoonup \mathcal{N}(0,1)$
Comportement sous H_1 **:** lorsque $\mu_0 > \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

Cas d'une loi normale et d'une taille d'échantillon petite.

PRINCIPE 16.2. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*
Données : $x_1, \ldots, x_n \in \mathbb{R}$
Modélisation associée : observations $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2
Hypothèse H_0 **:** $\mu_0 = \mu_{ref}$
Statistique de test :

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu_{ref}}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 **:** $T_n \sim \mathcal{T}_{n-1}$
Comportement sous H_1 **:** lorsque $\mu_0 > \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{ref}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

Cas d'une fréquence et d'une taille d'échantillon grande.

Ici, il n'est pas nécessaire d'estimer la variance : elle est connue sous H_0.

PRINCIPE 16.3. *Test de comparaison d'une proportion de population* p_0 *à une valeur de référence* p_{ref}
Données : $x_1, \dots, x_n \in \{0, 1\}$
Modélisation associée : observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0
Hypothèse H_0 **:** $p_0 = p_{ref}$
Statistique de test :

$$T_n = \sqrt{n}\, \frac{\overline{X}_n - p_{ref}}{\sqrt{p_{ref}(1 - p_{ref})}}$$

Comportement sous H_0 **:** $T_n \rightharpoonup \mathcal{N}(0, 1)$
Comportement sous H_1 **:** lorsque $p_0 > p_{ref}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $p_0 < p_{ref}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0.

Compléments sur les tests (cf. partie 9)

La statistique est un bikini. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital.
Arthur Koestler (écrivain, journaliste et essayiste hongrois, 1905–1983)

A retenir pour le reste de votre carrière

Nous avons vu dans ce chapitre de nombreux tests, il en existe en fait des dizaines d'autres, un pour chaque type de situation à laquelle on peut penser. Trouver le bon test requiert une connaissance profonde et vaste de la statistique, ainsi qu'un peu d'expérience. Ce n'est donc pas nécessairement à votre portée. Il faut donc retenir essentiellement deux choses de ce cours :

- dans les situations simples où il s'agit de comparer des résultats (proportions ou moyennes d'achat par exemple) de deux campagnes menées indépendamment, il ne suffit pas de comparer les valeurs lues sur les deux échantillons, il faut également déterminer si elles sont significativement différentes, ce que l'on fait en appliquant des T-tests de comparaisons de proportions ou de moyennes ;
- dans les situations où l'on veut tester des choses plus compliquées, on peut penser à faire appel à des statisticiens professionnels afin qu'ils apportent leur éclairage : souvent, ils auront un test idoine sous la main.

A retenir pour la suite du cours et pour l'examen

Traitement des données appariées. Les données appariées correspondent, sous SPSS, à la comparaison des valeurs de deux colonnes, par exemple lorsque l'on applique deux traitements différents simultanément ou successivement aux mêmes cobayes. On dispose deux séries de données de même longueur, $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$. Le meilleur moyen d'éliminer la variabilité intrinsèque à chaque sujet est de considérer les différences $z_j = x_j - y_j$ et d'effectuer un test sur leur moyenne, comme on l'a vu dans la partie 7. On teste ici usuellement que la moyenne des différences est nulle ($H_0 : \Delta = 0$, où Δ désigne l'espérance commune des variables aléatoires $Z_1, \ldots, Z_n$ modélisant les z_j).

Comparaison de deux populations. On se place ici dans un cadre où une première série de données $x_1, \ldots, x_n$ a été obtenue par échantillonnage sur une première population, et une seconde série de données $y_1, \ldots, y_m$, de longueur en général différente, a été obtenue quant à elle sur une seconde population. Concrètement, sous SPSS, cela va correspondre à comparer les valeurs d'une colonne fixée entre deux sous-ensembles de lignes du tableau de données.

Cas des proportions. Dans ce cas, les données sont dans $\{0, 1\}$. Lorsque n et m sont grands, on peut appliquer le principe énoncé à la page suivante.

PRINCIPE 16.4. *Test de comparaison de proportions p_X et p_Y associées à deux séries de données indépendantes*

Données : $x_1, \dots, x_n \in \{0,1\}$ et $y_1, \dots, y_m \in \{0,1\}$

Modélisation associée : deux séries indépendantes d'observations $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$, chacune formée de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli, avec les paramètres respectifs p_X et p_Y

Hypothèse H_0 : $p_X = p_Y$

Statistique de test :

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{p}_{n+m}(1-\widehat{p}_{n+m})\,(1/n + 1/m)}}$$

où l'estimateur groupé $\widehat{p}_{n+m}$ de la proportion commune sous H_0 est défini par

$$\widehat{p}_{n+m} = \frac{n\overline{X}_n + m\overline{Y}_m}{n+m}$$

Comportement sous H_0 : $T_{n,m} \rightharpoonup \mathcal{N}(0,1)$ lorsque n et m tendent tous deux vers $+\infty$

Comportement sous H_1 : lorsque $p_X > p_Y$, la statistique $T_{n,m}$ tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $p_X < p_Y$, la statistique $T_{n,m}$ tend à prendre des valeurs plus petites que sous H_0.

Cas des moyennes de variables quantitatives. Dans le cas général, il s'agit de déterminer au préalable si les variances des deux populations sont égales ou pas ; on recourt à un pré-test d'égalité des variances. Selon le résultat du test (lu dans les deux premières colonnes, la P-valeur correspond à l'indice de crédibilité de l'hypothèse d'égalité des variances), selon qu'il conserve ou pas l'hypothèse d'égalité des variances, on lit la première ou la seconde ligne des colonnes présentant les T-tests d'égalité des moyennes. Ici, je ne vous demande que de savoir lire les sorties SPSS correspondantes ; un exemple avait été fourni :

Statistiques de groupe

	Groupe	N	Moyenne	Ecart-type	Erreur standard moyenne
Verres bus	8h	23	4,448	3,0598	,6380
	10h	31	7,735	10,1658	1,8258

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes						
									Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Différence écart-type	Inférieure	Supérieure
Verres bus	Hypothèse de variances égales	5,005	,030	-1,498	52	,140	-3,2877	2,1944	-7,6911	1,1158
	Hypothèse de variances inégales			-1,700	37,021	,098	-3,2877	1,9341	-7,2064	,6311

Culture et confiture. Pour le reste, soyez simplement conscient qu'il existe d'autres tests que sur la moyenne ou la comparaison de moyennes : des tests sur la variance ou la comparaison de variances, bien sûr, mais aussi des tests d'ajustement à une loi ou à une famille de lois (ils répondent par exemple à la question : "Peut-on dire que les valeurs observées sont la réalisation d'un échantillon gaussien ?"), des tests de comparaison des lois de deux échantillons, etc.

Il faut également savoir être en mesure de lire les sorties SPSS correspondantes, ou tout du moins, de voir où y est calculée la P-valeur : dans la colonne `Signification`.

Tests du χ^2 (cf. partie 10)

Do not trust any statistics you did not fake yourself.
Winston Churchill (homme politique britannique, 1874–1965)

A retenir pour le reste de votre carrière

Les tests du χ^2 permettent souvent de détecter des écarts à des comportements attendus ; entre autres, ils permettent de découvrir des manipulations de données, comme des fraudes dans les écritures comptables.

Ils forment également un moyen simple de tester l'indépendance d'une variable qualitative par rapport à une autre (niveau de satisfaction en fonction de la catégorie d'âge par exemple).

A retenir pour la suite du cours et pour l'examen

Test du χ^2 d'ajustement simple. On part de données qualitatives (ordinales ou nominales) $x_1, \dots, x_n$ modélisées comme la réalisation de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribués selon une certaine loi $\mathbf{p} = (p_1, \dots, p_k)$ sur $\{1, 2, \dots, k\}$.

On pense, pour des raisons subjectives, à une loi de référence $\mathbf{p}^{\text{ref}} = (p_1^{\text{ref}}, \dots, p_k^{\text{ref}})$ et on veut tester

$$H_0 : \mathbf{p} = \mathbf{p}^{\text{ref}} \qquad \text{contre} \qquad H_1 : \mathbf{p} \neq \mathbf{p}^{\text{ref}} \ .$$

(H_1 se ré-écrit comme : il existe j tel que $p_j \neq p_j^{\text{ref}}$.)

On considère

$$\widehat{p}_{j,n} = \frac{N_{j,n}}{n} \qquad \text{où} \qquad N_{j,n} = \text{Card}\big\{t : X_t = j\big\}$$

puis la statistique de test

$$D_n(\mathbf{p}^{\text{ref}}) \overset{\text{not.}}{=} n \sum_{j=1}^{k} \frac{\big(\widehat{p}_{j,n} - p_j^{\text{ref}}\big)^2}{p_j^{\text{ref}}} = \sum_{j=1}^{k} \frac{\big(N_{j,n} - n\, p_j^{\text{ref}}\big)^2}{n\, p_j^{\text{ref}}} \ .$$

Le principe du test est alors décrit à la page suivante.

PRINCIPE 16.5. *Test d'ajustement simple à une loi de référence* $\mathbf{p}^{\text{ref}}$
Données : $x_1, \ldots, x_n \in \{1, \ldots, k\}$
Modélisation associée : observations $X_1, \ldots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p}$ sur $\{1, \ldots, k\}$
Hypothèse H_0 **:** $\mathbf{p} = \mathbf{p}^{\text{ref}}$
Statistique de test :

$$D_n(\mathbf{p}^{\text{ref}}) \overset{\text{not.}}{=} n \sum_{j=1}^{k} \frac{\left(\widehat{p}_{j,n} - p_j^{\text{ref}}\right)^2}{p_j^{\text{ref}}} = \sum_{j=1}^{k} \frac{\left(N_{j,n} - n\, p_j^{\text{ref}}\right)^2}{n\, p_j^{\text{ref}}}$$

Comportement sous H_0 **:** $D_n(\mathbf{p}^{\text{ref}}) \rightharpoonup \chi^2_{k-1}$ lorsque $n \to \infty$
Comportement sous H_1 **:** $D_n(\mathbf{p}^{\text{ref}})$ tend vers $+\infty$ et prend donc des valeurs beaucoup plus grandes que sous H_0.

Ici, la zone de rejet est toujours unilatère, et vaut donc $\left]\, c_{k-1,1-\alpha}, +\infty\right[$, pour une erreur de première espèce approximativement égale à α, où $c_{k-1,1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi χ^2_{k-1}.

REMARQUE 16.1 (Conditions pratiques d'approximation asymptotique). Le test est asymptotique et en pratique il faut vérifier que $n \geqslant 30$ et que les effectifs attendus pour toutes les modalités j vérifient $n\, p_j^{\text{ref}} \geqslant 5$. Si la seconde condition n'est pas satisfaite, il faut alors procéder à un regroupement de classes.

LA MINUTE SPPS 16.1. Il faut savoir lire les sorties SPSS et notamment, y retrouver la P-valeur du test, de même que décoder la petite note sous le tableau SPSS qui indique si les conditions asymptotiques sont remplies ou pas. Un exemple de tableau à lire est :

Test du Khi-deux

CodeSaison

	Effectif observé	Effectif théorique	Résidu
Hiver	8	9,0	-1,0
Printemps	17	9,0	8,0
Eté	6	9,0	-3,0
Automne	5	9,0	-4,0
Total	36		

Test

	CodeSaison
Khi-deux	10,000[a]
ddl	3
Signification asymptotique	,019

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 9,0.

Test du χ^2 d'indépendance entre deux variables qualitatives. On part de couples de données $(x_1, y_1), \ldots, (x_n, y_n)$ modélisés commes la réalisation des couples de variables aléatoires (X_1, Y_1), $\ldots$, (X_n, Y_n), indépendants et identiquement distribués selon une certaine loi $\mathbf{p}$ sur l'ensemble-produit $\{1, \ldots, r\} \times \{1, \ldots, s\}$.

On teste l'hypothèse H_0 d'indépendance entre les X_j et les Y_j, i.e., le fait que $\mathbf{p}$ soit une loi-produit (égale aux produits de ses marginales).

On note, pour tous x et y,

$$N_{x,y} = \text{Card}\{j : X_j = x \text{ et } Y_j = y\}\ , \quad N_{x,\cdot} = \text{Card}\{j : X_j = x\}\ , \quad N_{\cdot,y} = \text{Card}\{j : Y_j = y\}\ ,$$

puis on considère les estimateurs des marginales

$$\widehat{\mathbf{p}}_{\mathcal{X}} = \left(\frac{N_{1,\cdot}}{n}, \ldots, \frac{N_{r,\cdot}}{n}\right) \qquad \text{et} \qquad \widehat{\mathbf{p}}_{\mathcal{Y}} = \left(\frac{N_{\cdot,1}}{n}, \ldots, \frac{N_{\cdot,s}}{n}\right)\ .$$

On introduit la statistique de test

$$D_n^{\text{indep}} \overset{\text{not.}}{=} \sum_{x=1}^{r}\sum_{y=1}^{s} \frac{\left(N_{x,y} - n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)\right)^2}{n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)}$$

et on l'utilise selon le principe suivant.

PRINCIPE 16.6. *Test d'indépendance de couples de données*
Données : couples $(x_1, y_1), \ldots, (x_n, y_n)$ prenant leurs valeurs dans un ensemble-produit $\{1, \ldots, r\} \times \{1, \ldots, s\}$
Modélisation associée : couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendants et identiquement distribués selon une certaine loi $\mathbf{p}$ sur l'ensemble-produit $\{1, \ldots, r\} \times \{1, \ldots, s\}$
Hypothèse H_0 : les X_j sont indépendantes des Y_j, i.e., $\mathbf{p}$ est une loi égale au produit de ses marginales
Statistique de test :

$$D_n^{\text{indep}} \overset{\text{not.}}{=} \sum_{x=1}^{r}\sum_{y=1}^{s} \frac{\left(N_{x,y} - n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)\right)^2}{n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)}$$

Comportement sous H_0 : $D_n^{\text{indep}} \rightharpoonup \chi^2_{(r-1)(s-1)}$
Comportement sous H_1 : lorsqu'il n'y a pas indépendance, i.e., que $\mathbf{p}$ n'est pas une loi-produit, D_n^{indep} tend vers $+\infty$ et prend donc des valeurs beaucoup plus grandes que sous H_0.

Ici encore, la zone de rejet est toujours unilatère, et vaut donc $\left]\, c_{(r-1)(s-1),1-\alpha},\, +\infty\right[$, pour une erreur de première espèce approximativement égale à α.

REMARQUE 16.2 (Conditions pratiques d'approximation asymptotique). Le test est asymptotique et en pratique il faut vérifier que $n \geqslant 30$ et que pour tous les couples (x, y), on a des effectifs attendus, égaux aux réalisations de $n\,\widehat{\mathbf{p}}_{\mathcal{X}}(x)\,\widehat{\mathbf{p}}_{\mathcal{Y}}(y)$, qui soient tous plus grands que 5. Si la seconde condition n'est pas satisfaite, il faut alors procéder à un regroupement de classes.

LA MINUTE SPPS 16.2. Il faut savoir lire les sorties SPSS et notamment, y retrouver la P-valeur du test, de même que décoder la petite note sous le tableau SPSS qui indique si les conditions asymptotiques sont remplies ou pas. Un exemple de tableau à lire est :

Tableau croisé Vote (après regroupement) * Année

			Année			
			1A	2A	3A	Total
Vote (après regroupement)	Royal	Effectif	9	8	6	23
		Effectif théorique	7,5	7,8	7,7	23,0
	Sarkozy	Effectif	38	36	76	150
		Effectif théorique	48,8	50,9	50,3	150,0
	Autre	Effectif	14	22	0	36
		Effectif théorique	11,7	12,2	12,1	36,0
	Indécis ou NSPP	Effectif	32	31	14	77
		Effectif théorique	25,0	26,1	25,8	77,0
	Total	Effectif	93	97	96	286
		Effectif théorique	93,0	97,0	96,0	286,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	49,157[a]	6	,000
Nombre d'observations valides	286		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 7,48.

Remarques : Cas de $k = 2$ classes ou de $r = s = 2$ deux classes. Les tests du χ^2 sont tellement universels que leur considération risque de vous faire oublier les tests les plus simples (et les plus efficaces). Ce paragraphe veut vous les rappeler.

Test d'ajustement lorsqu'il n'y a que deux classes. Dans le test d'ajustement simple, lorsqu'il n'y a que $k = 2$ classes, on fait plutôt un test de comparaison à une proportion de référence (voir le principe page 501). Le test du χ^2 n'est à utiliser que lorsqu'il y a strictement plus de deux classes.

Test d'indépendance lorsqu'il n'y a que deux fois deux classes. De même, lorsque $r = s = 2$, i.e., que la table de contingence contient quatre cases, deux lignes et deux colonnes, on recourra plutôt au test de comparaison de proportions de deux populations (voir le principe page 504). Le test du χ^2 d'indépendance n'est à utiliser que si la table de contingence contient au moins trois lignes ou trois colonnes.

Recette mnémotechnique. Dans les deux cas précédemment cités, si l'on essayait d'appliquer la méthodologie des tests du χ^2, la loi limite serait une loi χ^2_1, carré d'une loi normale. Tomber sur une loi χ^2_1 signifie ainsi qu'il existe un moyen plus simple de mener le test.

Régression linéaire simple (cf. partie 13)

A retenir pour le reste de votre carrière

Le modèle de la régression linéaire est le modèle le simple pour établir une liaison entre deux variables quantitatives. Vous l'aviez déjà vu au lycée ou en classes préparatoires, nous en avons donné un nouvel éclairage, celui de la statistique inférentielle, qui permet de déterminer si le modèle linéaire établi (par méthode des moindres carrés) contribue significativement, d'un point de vue statistique, à l'explication d'une variable par une autre.

Des formules compliquées permettent, une fois le modèle établi, de prévoir de nouvelles observations (leurs valeurs moyennes ou leurs valeurs individuelles) et/ou de détecter les valeurs atypiques, beaucoup plus grandes ou beaucoup plus petites que les valeurs attendues.

Enfin, la régression étant une vraie science en soi, il faut savoir que si les hypothèses mises sur le modèle (dit linéaire gaussien) sont fortes, il existe cependant des techniques avancées (l'analyse des résidus) pour vérifier qu'elles tiennent ; il ne faut jamais oublier cette validation *a posteriori*, même si vraisemblablement, vous sous-traiterez la tâche (encore faut-il penser à son existence !).

A retenir pour la suite du cours et pour l'examen

Rappel du cadre et des objectifs. On parle de modèle linéaire gaussien lorsque des couples de données $(x_1, y_1), \dots, (x_n, y_n)$ sont disponibles et que les variables à expliquer y_j peuvent être modélisées de la manière suivante, en fonction des variables explicatives x_j : les y_j sont les réalisations des variables aléatoires

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j \ , \qquad j = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0,\, \sigma_0^2)$.

Seules les x_j et les y_j sont observées et à partir d'elles, on veut estimer, encadrer et tester les trois paramètres du modèle que sont α_0, β_0 et σ_0.

Définition 16.9. *On appelle estimateurs des moindres carrés de* $(\alpha_0,\, \beta_0)$ *le couple* $(\widehat{\alpha}_n,\, \widehat{\beta}_n)$ *antécédent du minimum d'une certaine fonction :*

$$(\widehat{\alpha}_n,\, \widehat{\beta}_n) \in \operatorname*{argmin}_{(\alpha,\, \beta) \in \mathbb{R}^2} \ \sum_{j=1}^{n} \left(Y_j - (\alpha + \beta x_j)\right)^2 \ .$$

On dispose par ailleurs d'expressions explicites pour $\widehat{\alpha}_n$ et $\widehat{\beta}_n$, ce qui permet de calculer leurs valeurs réalisées. A partir de ces estimateurs, on construit les prédictions du modèle $\widehat{Y}_j$ et les résidus associés :

$$\widehat{Y}_j = \widehat{\alpha}_n + \widehat{\beta}_n x_j \qquad \text{et} \qquad \widehat{\varepsilon}_j = Y_j - \widehat{Y}_j \ , \qquad j = 1, \dots, n \ .$$

Evidemment, on a les valeurs réalisées des Y_j, on n'a donc pas besoin de les prédire ou de les estimer, mais ces quantités $\widehat{Y}_j$ jouent un rôle important dans l'analyse théorique.

PROPOSITION 16.1. *L'estimateur de la variance σ_0^2 donné par*

$$\widehat{\sigma}_n^2 = \frac{1}{n-2}\sum_{j=1}^{n}(\widehat{\varepsilon}_j)^2 = \frac{\Sigma_R}{n-2}$$

est sans biais.

Question 1 : Existence d'une relation linéaire significative. L'existence d'une relation linéaire significative correspond au fait que le coefficient β_0 soit significativement différent de 0. Il s'agit donc de tester $H_0 : \beta_0 = 0$. Si le test conserve H_0, alors on en déduit que la relation linéaire n'est pas suffisamment fondée : la modélisation linéaire de la variable à expliquer en fonction de la variable explicative est pauvre et peu informative, il faut soit changer de variable explicative, soit considérer un autre type de liaison (quadratique, logarithmique, etc.).

PRINCIPE 16.7. *Test de l'existence d'une relation linéaire (dans le cadre d'un modèle gaussien)*
Données : $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$
Modélisation associée : valeurs $x_1, \ldots, x_n$ fixées et observations stochastiques $Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j$, où les ε_j sont indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$
Hypothèse H_0 : $\beta_0 = 0$ (i.e., absence de liaison linéaire)
Statistique de test :

$$T_n = \sqrt{n\ \mathrm{Var}(x_1^n)}\ \frac{\widehat{\beta}_n}{\sqrt{\widehat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \sim \mathcal{T}_{n-2}$
Comportement sous H_1 : T_n tend à prendre des valeurs ou plus grandes ou plus petites sous $H_1 : \beta_0 \neq 0$ (i.e., existence d'une liaison linéaire).

Note : on définit $D_n = (T_n)^2$ et on peut effectuer un test d'absence de liaison linéaire également avec D_n.

Question 2 : Proportion de reconstruction.

THÉORÈME 16.3. *La somme des carrés totale Σ_T est égale à la somme des carrés expliquée par la régression Σ_E plus la somme des carrés résiduelle Σ_R,*

$$\underbrace{\sum_{j=1}^{n}(Y_j - \overline{Y}_n)^2}_{\overset{\text{not.}}{=}\Sigma_T} = \underbrace{\sum_{j=1}^{n}(\widehat{Y}_j - \overline{Y}_n)^2}_{\overset{\text{not.}}{=}\Sigma_E} + \underbrace{\sum_{j=1}^{n}(Y_j - \widehat{Y}_j)^2}_{\overset{\text{not.}}{=}\Sigma_R}.$$

DÉFINITION 16.10. *Le coefficient de détermination r^2 est la fraction de la variabilité totale expliquée par la régression,*

$$r^2 = \frac{\Sigma_E}{\Sigma_T}.$$

Plus r^2 *est grand, meilleur est le modèle linéaire.*

Question 3 : Interprétation de la relation linéaire. Avant de l'effectuer, il faut également déterminer si α_0 est significativement différent de 0 ou non. On écrit ensuite la relation linéaire proposée et on donne, autant que faire se peut, une interprétation économique de chaque coefficient (pente et ordonnée à l'origine). Evidemment, ce paragraphe n'est pas du ressort des mathématiques mais de celui du bon sens.

Question 4 : Détection des valeurs atypiques. En demandant à SPSS de tracer les intervalles de prévision à 95 %, on peut voir des couples (x_j, y_j) tels que y_j n'appartient pas à l'intervalle de prévision construit sur x_j. Ces valeurs y_j sont donc beaucoup plus grandes ou beaucoup plus petites que les valeurs qui auraient été attendues. C'est un fait qu'il s'agit d'interpréter économiquement ; cela peut correspondre à de bonnes affaires ou au contraire, des tentatives d'escroquerie.

En pratique : lecture de sorties SPSS. De ce cours, il faut surtout retenir comment lire les sorties SPSS, comme celle-ci :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,930[a]	,865	,860	122,939

a. Valeurs prédites : (constantes), Surface

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	2527207,505	1	2527207,505	167,210	,000[a]
	Résidu	392963,209	26	15113,970		
	Total	2920170,714	27			

a. Valeurs prédites : (constantes), Surface

b. Variable dépendante : Prix

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			95,0% % intervalles de confiance pour B	
		A	Erreur standard	Bêta	t	Sig.	Borne inférieure	Limite supérieure
1	(Constante)	-29,466	41,246		-,714	,481	-114,247	55,316
	Surface	5,353	,414	,930	12,931	,000	4,502	6,204

a. Variable dépendante : Prix

Le premier tableau donne les valeurs réalisées des statistiques suivantes :

R	R-deux	R-deux ajusté	Erreur standard
$\sqrt{r^2}$	r^2	...	$\sqrt{\widehat{\sigma}_n^2}$

Le deuxième tableau donne les valeurs réalisées de :

Somme carrés	ddl	Moyenne carrés	D	Sig.
Σ_E	1	Σ_E	D_n	P-val. test $H_0 : \beta_0 = 0$
Σ_R	$n-2$	$\Sigma_R/(n-2)$		
Σ_T	$n-1$			

Dans ce tableau, lorsque la P-valeur lue est petite, on rejette l'hypothèse $H_0 : \beta_0 = 0$ d'absence de liaison linéaire et on conclut à l'existence d'une relation linéaire. Si la P-valeur lue est grande, le modèle linéaire ne tient pas (auquel cas il peut exister des relations d'autres types, non linéaires) ; il ne faudra alors pas lire ni exploiter les résultats du troisième tableau.

Enfin, le troisième tableau donne les valeurs réalisées de :

Coeff.	Err. standard	...	t	Sig.	IC (bornes inf. et sup.)
$\widehat{\alpha}_n$	$\sqrt{\frac{\widehat{\sigma}_n^2}{n}\left(1+\frac{(\overline{x}_n)^2}{\mathrm{Var}(x_1^n)}\right)}$		T'_n	P-val. test $H_0 : \alpha_0 = 0$	IC sur α_0
$\widehat{\beta}_n$	$\sqrt{\frac{\widehat{\sigma}_n^2}{n\ \mathrm{Var}(x_1^n)}}$	...	T_n	P-val. test $H_0 : \beta_0 = 0$	IC sur β_0

Les deux tests mis en œuvre ci-dessus utilisent des statistiques suivant sous H_0 la loi de Student. Dans la seconde ligne, on teste $H_0 : \beta_0 = 0$ contre $H_1 : \beta_0 \neq 0$ et dans la première ligne, il s'agit de $H_0 : \alpha_0 = 0$ contre $H_1 : \alpha_0 \neq 0$.

En pratique : comment exploiter les sorties SPSS ainsi lues. On procède ainsi.

1. On commence par regarder si le coefficient de pente β_0 est significatif ou non (ou on regarde la P-valeur du test avec D_n ou de celui avec T_n) ; s'il l'est, alors il existe une liaison linéaire significative et on continue le traitement des données ; et sinon, il faut chercher un autre type de relation (et en tout cas, cesser de perdre son temps sur cette sortie SPSS-là !).

2. On lit alors la valeur réalisée du coefficient de détermination r^2 et on regarde s'il est grand (autour de 40 % ou plus), modéré (autour de 25 %) ou faible (inférieur à 10 %, disons). C'est un indicateur de la qualité de l'ajustement linéaire.

3. Ensuite, on peut passer à l'écriture de la relation. Soit, par exemple, la relation suivante [48]

$$\begin{aligned}\text{prix (en milliers d'euros)} = 24.546 + 4.735\times\ &\text{surface (en } m^2\text{)} \\ &+ \text{aléa d'écart-type } 67.939\end{aligned}$$

4. Il ne faut pas oublier l'étape de validation selon le contexte ; il s'agit souvent d'une validation économique, qui consiste à trouver une interprétation aux coefficients. Ainsi, la relation précédente indique une augmentation moyenne du prix d'un appartement de 4 735 euros par m^2 supplémentaire, à quoi s'ajoute un coût fixe pour les parties communes de 25 000 euros environ. Lorsque l'interprétation de l'ordonnée à l'origine (l'estimée de α_0) pose problème, il faut notamment regarder le test de $H_0 : \alpha_0 = 0$ pour savoir si on peut la prendre nulle ou non ; à défaut, il existe peut-être un phénomène de compensation, ou la relation linéaire peut également n'être vraie que dans un certain intervalle (on se référera aux exercices pour des exemples concrets de telles situations).

5. Enfin, en traçant les intervalles de prévision sur le graphique de dispersion des données, on peut se rendre compte qu'il existe des données atypiques, que l'on aurait envie d'enlever de l'étude pour la rendre plus précise ; il faut justifier cette omission statistique par des arguments économiques ou, en tout cas, extra-statistiques. Ces données, qui sont beaucoup plus grandes ou beaucoup plus petites que les valeurs qui auraient été attendues, peuvent correspondre à de bonnes affaires ou au contraire, à des tentatives d'escroquerie.

48. Qui n'est pas celle que l'on peut lire dans la sortie SPSS reproduite ci-dessus, mais en est inspirée !

Régression linéaire multiple (cf. partie 14)

A retenir pour le reste de votre carrière

Le modèle de la régression linéaire multiple permet d'expliquer des liaisons (linéaires) entre une variable quantitative à expliquer et différentes variables explicatives ; dans ce cours, nous n'avons vu que le cas de variables explicatives quantitatives, mais en fait, on peut également tenir compte des effets de variables qualitatives. Dans le même genre d'idées, il est également possible de construire des modèles de régression non linéaires, selon l'intuition économique que l'on a sur le comportement des données. C'est un sujet fort complexe et fort délicat que de trouver une bonne modélisation : en pratique, la régression linéaire multiple forme souvent une bonne première approche mais il vous faudra recourir à des statisticiens professionnels si cette dernière ne vous donne pas satisfaction (i.e., si son indice de qualité principal, le coefficient de détermination r^2 est faible). Seul cet expert extérieur saura vous guider dans la jungle des différents modèles généralisés possibles.

A retenir pour la suite du cours et pour l'examen

Le modèle gaussien multiple. On dispose de n données y_t, où $t = 1, \ldots, n$, à expliquer, chacune en fonction d'un nombre fini k de données explicatives, dont les valeurs correspondant à y_t sont notées $x_{1,t}, x_{2,t}, \ldots, x_{k,t}$. Pour les mêmes raisons et intuitions que dans la partie 13, on modélise les données à expliquer y_t comme la réalisation des variables aléatoires

$$\begin{aligned} Y_t &= \alpha_0 + \beta_{1,0}\, x_{1,t} + \beta_{2,0}\, x_{2,t} + \ldots + \beta_{k,0}\, x_{k,t} + \varepsilon_t \\ &= \alpha_0 + \sum_{j=1}^{k} \beta_{j,0}\, x_{j,t} + \varepsilon_t \ , \qquad \text{pour } t = 1, \ldots, n, \end{aligned}$$

où les $\varepsilon_1, \ldots, \varepsilon_n$ (les résidus) sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$.

Les réalisations y_j des Y_j sont observées, les valeurs des $x_{k,j}$ sont observées voire choisies par le statisticien ; en revanche, les coefficients α_0 et $\beta_{1,0}$, ..., $\beta_{k,0}$ de la relation linéaire, de même que les (réalisations des) résidus ε_j, ne le sont pas. On effectue par ailleurs l'hypothèse que $\mathbf{X}$ soit injective.

Etude d'une proposition de modèle linéaire donné. En pratique, on fait face à des sorties SPSS comme :

Régression sur le modèle complet

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,916[a]	,839	,813	138,034

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	1793129,948	3	597709,983	31,370	,000[a]
	Résidu	342959,506	18	19053,306		
	Total	2136089,455	21			

a. Valeurs prédites : (constantes), Gratuits, Journaux, Radio

b. Variable dépendante : Ventes

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés			intervalles de confiance à 95% pour B	
Modèle		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Limite supérieure
1	(Constante)	238,458	112,242		2,124	,048	2,646	474,270
	Radio	23,850	4,524	,749	5,272	,000	14,346	33,354
	Journaux	32,629	5,369	,585	6,078	,000	21,350	43,908
	Gratuits	-,619	10,228	-,009	-,060	,952	-22,107	20,870

a. Variable dépendante : Ventes

Le premier tableau présente des indicateurs résumant la qualité de l'ajustement linéaire, mesurée par

- le coefficient de détermination r^2 (la proportion de la variabilité de la variable à expliquer pouvant être reconstruite à partir des variables explicatives) ;
- la version ajustée du précédent, r^2_{ajust}, obtenue après correction prenant en compte la nécessaire meilleure reconstruction lorsque le nombre de variables explicatives augmente ;
- l'estimée de l'écart-type σ_0 des résidus ε_t.

R	R-deux	R-deux ajusté	Erreur standard
$\sqrt{r^2}$	r^2	r^2_{ajust}	Estimée de σ_0

Le second tableau s'intéresse à la validité globale du modèle (test de Fisher) : le modèle linéaire proposé contribue-t-il significativement à l'explication statistique de la variable à expliquer ? Ce que l'on peut reformuler mathématiquement comme le test de $H_0 : \beta_{1,0} = \beta_{2,0} = \ldots = \beta_{k,0} = 0$. On regardera surtout si la P-valeur fournie par la dernière colonne est plus grande ou plus petite que 5 % (la validité du modèle, i.e., son caractère explicatif significatif n'étant retenu que si cette P-valeur est plus petite que ce seuil de 5 %).

Somme carrés	ddl	Moyenne carrés	D	Sig.
...	...	...	...	P-val. test $H_0 : \forall j,\ \beta_{j,0} = 0$
...	...	...		
...	...			

Enfin, le troisième tableau propose les estimées des coefficients et étudie la validité marginale du modèle, i.e., le fait que toutes les variables explicatives considérées sont individuellement significatives ou non face aux autres variables ; ce qui correspond aux tests des hypothèses $H_0 : \alpha_0 = 0$ et $H_0 : \beta_{j,0} = 0$ pour $j = 1, \ldots, k$. On lit essentiellement :

- la première colonne, qui fournit les estimées des différents coefficients ;
- la colonne `Sig.`, qui précise quels sont les coefficients significativement non nuls dans le modèle considéré (la P-valeur de la quatrième colonne doit être inférieure à 5 % afin que la variable correspondante puisse être dite significative face aux autres variables).

Coeff.	Err. std.	...	t	Sig.	Int. confiance
Estimée de α_0	...	...	...	P-val. $H_0 : \alpha_0 = 0$	IC sur α_0
Estimée de $\beta_{1,0}$	...	...	...	P-val. $H_0 : \beta_{1,0} = 0$	IC sur $\beta_{1,0}$
...	...	...	...	...	...
Estimée de $\beta_{k,0}$	...	...	...	P-val. $H_0 : \beta_{k,0} = 0$	IC sur $\beta_{k,0}$

Interprétation des sorties SPSS. En pratique, lorsque l'on a affaire à une sortie de régression multiple, on procède ainsi.

1. On commence par regarder la validité globale du modèle, i.e., le résultat du test de Fisher mené dans le deuxième tableau. Seulement dans le cas où la P-valeur lue est plus petite que 5 %, on continue l'étude.

2. On étudie ensuite la validité marginale : si toutes les variables sont significatives ou pas, ce que l'on lit dans le troisième tableau. Seulement dans le cas où toutes les P-valeurs lues (pour les variables explicatives [49] uniquement) sont plus petites que 5 %, on continue l'étude.

3. On lit seulement alors la valeur réalisée du coefficient de détermination ajusté r^2_{ajust} et on regarde s'il est grand (autour de 40 % ou plus), modéré (autour de 25 %) ou faible (inférieur à 10 %, disons). Il indique la qualité de l'ajustement linéaire, rapporté au nombre de variables explicatives utilisées.

4. Enfin, si les coefficients sont tous significatifs, on peut passer à l'écriture de la relation linéaire proposée (cf. format présenté à la fiche de synthèse précédente).

5. Il ne faut pas oublier l'étape de validation selon le contexte (il s'agit souvent d'une validation économique) ; il s'agit également d'interpréter les paramètres α_0 et $\beta_{p,0}$ du modèle.

49. La P-valeur du coefficient noté `(Constante)` par SPSS et correspondant à l'ordonnée à l'origine α_0 peut être grande, cela ne pose pas de problème.

6. On ne considérera pas ici l'étude des données atypiques, elle est en effet plus délicate que dans le cas de la régression simple. Mais SPSS sait le faire sur demande.

Comparaison et choix de modèles. Premièrement, on ne considère ici que des modèles statistiquement valides au sens où les étapes 1. et 2. du processus d'interprétation ci-dessous ont été remplies. Il faut mélanger différents critères : des critères statistiques assez automatiques, à nuancer cependant le cas échéant avec des considérations économiques. L'objectif est d'obtenir un compromis entre un nombre de variables explicatives à la fois suffisamment petit (afin de proposer une relation facilement interprétable) et grand (afin de garantir une bonne qualité d'ajustement).

- Pour comparer deux modèles, on recourra de manière équivalente à la comparaison de leurs valeurs réalisées de r^2_{ajust} (la plus grande l'emporte) ou de celle de leurs estimées de l'écart-type des résidus (la plus petite l'emporte).
- Il existe deux méthodes de sélection automatique, la méthode "backward" qui enlève des variables une à une tant qu'il reste des variables non individuellement significatives, et la méthode "forward", qui en ajoute tant qu'elle peut ajouter des variables restant significatives dans le modèle étendu obtenu.

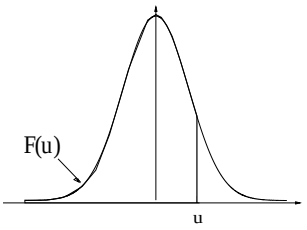
Note : on rappelle que le coefficient r^2 augmente mécaniquement avec l'inclusion de nouvelles variables et qu'il ne forme donc pas un critère juste pour comparer deux modèles ne reposant pas sur le même nombre de variables explicatives.

Dix-septième Partie

Tables des lois statistiques

Loi normale : fonction de répartition

Pour une valeur $u \geqslant 0$, la table ci-dessous renvoie la valeur $F(u)$ de la fonction de répartition F de la loi normale centrée réduite au point u.



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

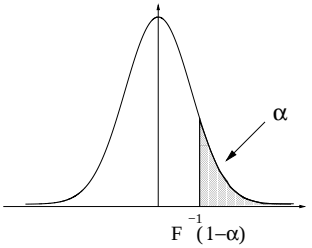
Table pour les grandes valeurs de u :

u	3.0	3.1	3.2	3.3	3.4
$F(u)$	0.99865	0.999032	0.999313	0.999517	0.999663
u	3.5	3.6	3.7	3.8	3.9
$F(u)$	0.999767	0.999841	0.999892	0.999928	0.999952
u	4.0	4.1	4.2	4.3	4.4
$F(u)$	0.999968	0.999979	0.999987	0.999991	0.999995
u	4.5	4.6	4.7	4.8	4.9
$F(u)$	0.999997	0.999998	0.999999	0.999999	1

FIGURE 78. Table de la fonction de répartition de la loi normale standard

Loi normale : quantiles

Pour une valeur $\alpha \in]0; 0.5[$, la table ci-dessous renvoie la valeur $F^{-1}(1-\alpha)$ de la fonction quantile F^{-1} de la loi normale centrée réduite au point $1-\alpha$.

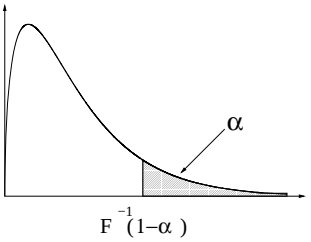


α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.00	∞	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656
0.01	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0749
0.02	2.0537	2.0335	2.0141	1.9954	1.9774	1.9600	1.9431	1.9268	1.9110	1.8957
0.03	1.8808	1.8663	1.8522	1.8384	1.8250	1.8119	1.7991	1.7866	1.7744	1.7624
0.04	1.7507	1.7392	1.7279	1.7169	1.7060	1.6954	1.6849	1.6747	1.6646	1.6546
0.05	1.6449	1.6352	1.6258	1.6164	1.6072	1.5982	1.5893	1.5805	1.5718	1.5632
0.06	1.5548	1.5464	1.5382	1.5301	1.5220	1.5141	1.5063	1.4985	1.4909	1.4833
0.07	1.4758	1.4684	1.4611	1.4538	1.4466	1.4395	1.4325	1.4255	1.4187	1.4118
0.08	1.4051	1.3984	1.3917	1.3852	1.3787	1.3722	1.3658	1.3595	1.3532	1.3469
0.09	1.3408	1.3346	1.3285	1.3225	1.3165	1.3106	1.3047	1.2988	1.2930	1.2873
0.10	1.2816	1.2759	1.2702	1.2646	1.2591	1.2536	1.2481	1.2426	1.2372	1.2319
0.11	1.2265	1.2212	1.2160	1.2107	1.2055	1.2004	1.1952	1.1901	1.1850	1.1800
0.12	1.1750	1.1700	1.1650	1.1601	1.1552	1.1503	1.1455	1.1407	1.1359	1.1311
0.13	1.1264	1.1217	1.1170	1.1123	1.1077	1.1031	1.0985	1.0939	1.0893	1.0848
0.14	1.0803	1.0758	1.0714	1.0669	1.0625	1.0581	1.0537	1.0494	1.0450	1.0407
0.15	1.0364	1.0322	1.0279	1.0237	1.0194	1.0152	1.0110	1.0069	1.0027	0.9986
0.16	0.9945	0.9904	0.9863	0.9822	0.9782	0.9741	0.9701	0.9661	0.9621	0.9581
0.17	0.9542	0.9502	0.9463	0.9424	0.9385	0.9346	0.9307	0.9269	0.9230	0.9192
0.18	0.9154	0.9116	0.9078	0.9040	0.9002	0.8965	0.8927	0.8890	0.8853	0.8816
0.19	0.8779	0.8742	0.8705	0.8669	0.8633	0.8596	0.8560	0.8524	0.8488	0.8452
0.20	0.8416	0.8381	0.8345	0.8310	0.8274	0.8239	0.8204	0.8169	0.8134	0.8099
0.21	0.8064	0.8030	0.7995	0.7961	0.7926	0.7892	0.7858	0.7824	0.7790	0.7756
0.22	0.7722	0.7688	0.7655	0.7621	0.7588	0.7554	0.7521	0.7488	0.7454	0.7421
0.23	0.7388	0.7356	0.7323	0.7290	0.7257	0.7225	0.7192	0.7160	0.7128	0.7095
0.24	0.7063	0.7031	0.6999	0.6967	0.6935	0.6903	0.6871	0.6840	0.6808	0.6776
0.25	0.6745	0.6713	0.6682	0.6651	0.6620	0.6588	0.6557	0.6526	0.6495	0.6464
0.26	0.6433	0.6403	0.6372	0.6341	0.6311	0.6280	0.6250	0.6219	0.6189	0.6158
0.27	0.6128	0.6098	0.6068	0.6038	0.6008	0.5978	0.5948	0.5918	0.5888	0.5858
0.28	0.5828	0.5799	0.5769	0.5740	0.5710	0.5681	0.5651	0.5622	0.5592	0.5563
0.29	0.5534	0.5505	0.5476	0.5446	0.5417	0.5388	0.5359	0.5330	0.5302	0.5273
0.30	0.5244	0.5215	0.5187	0.5158	0.5129	0.5101	0.5072	0.5044	0.5015	0.4987
0.31	0.4959	0.4930	0.4902	0.4874	0.4845	0.4817	0.4789	0.4761	0.4733	0.4705
0.32	0.4677	0.4649	0.4621	0.4593	0.4565	0.4538	0.4510	0.4482	0.4454	0.4427
0.33	0.4399	0.4372	0.4344	0.4316	0.4289	0.4261	0.4234	0.4207	0.4179	0.4152
0.34	0.4125	0.4097	0.4070	0.4043	0.4016	0.3989	0.3961	0.3934	0.3907	0.3880
0.35	0.3853	0.3826	0.3799	0.3772	0.3745	0.3719	0.3692	0.3665	0.3638	0.3611
0.36	0.3585	0.3558	0.3531	0.3505	0.3478	0.3451	0.3425	0.3398	0.3372	0.3345
0.37	0.3319	0.3292	0.3266	0.3239	0.3213	0.3186	0.3160	0.3134	0.3107	0.3081
0.38	0.3055	0.3029	0.3002	0.2976	0.2950	0.2924	0.2898	0.2871	0.2845	0.2819
0.39	0.2793	0.2767	0.2741	0.2715	0.2689	0.2663	0.2637	0.2611	0.2585	0.2559
0.40	0.2533	0.2508	0.2482	0.2456	0.2430	0.2404	0.2378	0.2353	0.2327	0.2301
0.41	0.2275	0.2250	0.2224	0.2198	0.2173	0.2147	0.2121	0.2096	0.2070	0.2045
0.42	0.2019	0.1993	0.1968	0.1942	0.1917	0.1891	0.1866	0.1840	0.1815	0.1789
0.43	0.1764	0.1738	0.1713	0.1687	0.1662	0.1637	0.1611	0.1586	0.1560	0.1535
0.44	0.1510	0.1484	0.1459	0.1434	0.1408	0.1383	0.1358	0.1332	0.1307	0.1282
0.45	0.1257	0.1231	0.1206	0.1181	0.1156	0.1130	0.1105	0.1080	0.1055	0.1030
0.46	0.1004	0.0979	0.0954	0.0929	0.0904	0.0878	0.0853	0.0828	0.0803	0.0778
0.47	0.0753	0.0728	0.0702	0.0677	0.0652	0.0627	0.0602	0.0577	0.0552	0.0527
0.48	0.0502	0.0476	0.0451	0.0426	0.0401	0.0376	0.0351	0.0326	0.0301	0.0276
0.49	0.0251	0.0226	0.0201	0.0175	0.0150	0.0125	0.0100	0.0075	0.0050	0.0025

FIGURE 79. Table des quantiles de la loi normale standard

Loi du χ^2 : quantiles

Pour un degré de liberté n entre 1 et 30 et pour certaine valeur de α, la table ci-dessous renvoie la valeur $F^{-1}(1-\alpha)$ de la fonction quantile F^{-1} de la loi du χ^2 à n degrés de liberté au point $1-\alpha$.



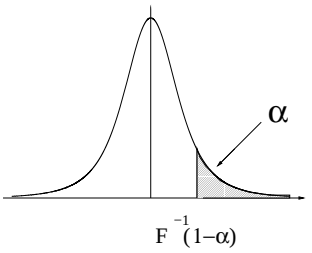
n \ α	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.001
1	0.0002	0.001	0.0039	0.0158	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	0.87	1.24	1.64	2.2	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.12
9	2.09	2.7	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	31.26
12	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	8.55	22.31	25.	27.49	30.58	37.7
16	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32.	39.25
17	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	42.31
19	7.63	8.91	10.12	11.65	27.2	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.31
21	8.9	10.28	11.59	13.24	29.62	32.67	35.48	38.93	46.8
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.2	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.4	13.85	15.66	33.2	36.42	39.36	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.2	13.84	15.38	17.29	35.56	38.89	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.3
30	14.95	16.79	18.49	20.6	40.26	43.77	46.98	50.89	59.7

Lorsque le nombre de degrés de liberté n est supérieur à 30, on peut utiliser l'approximation suivante : $\sqrt{2\chi^2(p)} - \sqrt{2p-1}$ suit approximativement la loi normale centrée réduite.

FIGURE 80. Table des quantiles de la loi du χ^2 (en fonction du nombre de degrés de liberté)

Loi de Student : quantiles

Pour un certain degré de liberté n et pour certaine valeur de α, la table ci-dessous renvoie la valeur $F^{-1}(1-\alpha)$ de la fonction quantile F^{-1} de la loi de Student à n degrés de liberté au point $1-\alpha$.



n \ α	0.45	0.3	0.2	0.1	0.05	0.025	0.01	0.001
1	0.158	0.727	1.376	3.078	6.314	12.706	31.821	318.309
2	0.142	0.617	1.061	1.886	2.920	4.303	6.965	22.327
3	0.137	0.584	0.978	1.638	2.353	3.182	4.541	10.215
4	0.134	0.569	0.941	1.533	2.132	2.776	3.747	7.173
5	0.132	0.559	0.920	1.476	2.015	2.571	3.365	5.893
6	0.131	0.553	0.906	1.440	1.943	2.447	3.143	5.208
7	0.130	0.549	0.896	1.415	1.895	2.365	2.998	4.785
8	0.130	0.546	0.889	1.397	1.860	2.306	2.896	4.501
9	0.129	0.543	0.883	1.383	1.833	2.262	2.821	4.297
10	0.129	0.542	0.879	1.372	1.812	2.228	2.764	4.144
11	0.129	0.540	0.876	1.363	1.796	2.201	2.718	4.025
12	0.128	0.539	0.873	1.356	1.782	2.179	2.681	3.930
13	0.128	0.538	0.870	1.350	1.771	2.160	2.650	3.852
14	0.128	0.537	0.868	1.345	1.761	2.145	2.624	3.787
15	0.128	0.536	0.866	1.341	1.753	2.131	2.602	3.733
16	0.128	0.535	0.865	1.337	1.746	2.120	2.583	3.686
17	0.128	0.534	0.863	1.333	1.740	2.110	2.567	3.646
18	0.127	0.534	0.862	1.330	1.734	2.101	2.552	3.610
19	0.127	0.533	0.861	1.328	1.729	2.093	2.539	3.579
20	0.127	0.533	0.860	1.325	1.725	2.086	2.528	3.552
21	0.127	0.532	0.859	1.323	1.721	2.080	2.518	3.527
22	0.127	0.532	0.858	1.321	1.717	2.074	2.508	3.505
23	0.127	0.532	0.858	1.319	1.714	2.069	2.500	3.485
24	0.127	0.531	0.857	1.318	1.711	2.064	2.492	3.467
25	0.127	0.531	0.856	1.316	1.708	2.060	2.485	3.450
26	0.127	0.531	0.856	1.315	1.706	2.056	2.479	3.435
27	0.127	0.531	0.855	1.314	1.703	2.052	2.473	3.421
28	0.127	0.530	0.855	1.313	1.701	2.048	2.467	3.408
29	0.127	0.530	0.854	1.311	1.699	2.045	2.462	3.396
30	0.127	0.530	0.854	1.310	1.697	2.042	2.457	3.385
40	0.126	0.529	0.851	1.303	1.684	2.021	2.423	3.307
80	0.126	0.526	0.846	1.292	1.664	1.990	2.374	3.195
120	0.126	0.526	0.845	1.289	1.658	1.980	2.358	3.160
∞	0.126	0.524	0.842	1.282	1.645	1.96	2.327	3.091

FIGURE 81. Table des quantiles de la loi de Student (en fonction du nombre de degrés de liberté)