



Apprendre à oser®

HEC Formation Fondamentale L3

ELÉMENTS DE STATISTIQUE POUR CITOYENS D'AUJOURD'HUI ET MANAGERS DE DEMAIN

Vincent Eli
Jean-Michel Gautier
Benjamin Petiau
Dinah Rosenberg
Gilles Stoltz

Année universitaire 2016–2017

Table des matières

Mise en bouche : La statistique dans la presse!	5
Partie I. Version rédigée du cours	9
Chapitre 1. Panorama de la démarche statistique	11
1 Des données à une modélisation	11
2 Rappels de cours	13
3 Tests d'hypothèses	19
Chapitre 2. Recueil, représentation et modélisation de données	21
1 Les différents types de données	21
2 Le recueil des données et son but	24
3 Modélisation mathématique : les lois classiques et le cas général	32
4 Résumé	39
Chapitre 3. Introduction aux tests d'hypothèses	41
1 Méthodologie des tests : trois exemples	41
2 Méthodologie des tests : théorie générale	47
3 Rappels sur les quantiles d'une loi	56
Chapitre 4. Tests de comparaison à une valeur de référence	59
1 Cas d'une loi générale et d'une taille d'échantillon grande	59
2 Cas d'une proportion, taille d'échantillon grande	64
3 Taille d'échantillon faible : cas de données gaussiennes	70
Chapitre 5. Tests à partir d'échantillons indépendants ou appariés	75
1 Retour SPSS sur le test de comparaison d'une seule moyenne à une valeur de référence	75
2 Première étude : cas des échantillons appariés	79
3 Seconde étude : échantillons indépendants, le cas des proportions	81
4 Troisième étude : échantillons indépendants, le cas général	89

Chapitre 6. Tests du χ^2	95
1 Test du χ^2 d'ajustement simple	95
2 Test du χ^2 d'indépendance entre deux variables qualitatives	103
3 Rappel : cas de deux classes ou de deux fois deux classes	112
Chapitre 7. Estimation par intervalles	113
1 Rappels mathématiques (lecture en diagonale possible)	113
2 Interprétation et exploitation des formules d'intervalles de confiance	117
3 Nouveauté — Intervalles de confiance simultanés	124
4 Nouveauté — Planification de sondages	126
Chapitre 8. Régression linéaire simple	131
1 Présentation du modèle et de ses objectifs	131
2 Analyse descriptive : choix de la meilleure régression linéaire	135
3 Enrichissement du point de vue : modèle linéaire gaussien	139
4 Vérification de l'existence d'une relation linéaire	142
5 Décryptage des sorties SPSS et interprétation économique	144
6 Intermède : justification de la (longue) liste des formules	147
7 Prévission en un nouveau point; détection des valeurs atypiques	149
Chapitre 9. Régression linéaire multiple	153
1 Le modèle linéaire multiple	153
2 Un zeste d'algèbre linéaire	155
3 Lecture et interprétation de sorties SPSS	158
4 Comparaison et choix de modèles linéaires; procédures de sélection de variables explicatives	166
5 Variables explicatives qualitatives	171
6 Résumé des compétences nécessaires	171
Partie II. Exercices issus de quiz et d'examens	173
Chapitre 10. Extraits de quiz	175
Chapitre 11. Extraits d'examens	233
Partie III. Fiches de synthèse	299
Partie IV. Tables des lois statistiques	321

Mise en bouche : La statistique dans la presse !

20 SPORT

MERCREDI 16 NOVEMBRE 2011

ANALYSE De plus en plus de clubs de L1 se servent de ces données pour rationaliser la vie de leur club

LES STATS ONT LA COTE

ANTOINE MAES

Il y a la fiction. Où Brad Pitt révolutionne le base-ball en incarnant Billy Beane, coach des A's d'Oakland en 2002, dont le recours massif aux statistiques lui permet d'emmener son équipe vers le record de victoires consécutives en MLB. *Le Stratège* sort aujourd'hui en salle. Et il y a la réalité : en 2011, alors que les championnats anglais de foot de première et deuxième division se sont massivement tournés vers les chiffres – à l'instar de ce qui se fait dans les sports américains depuis longtemps –, pour mieux comprendre les performances de leurs joueurs, la Ligue 1 a un train de retard. « Mais c'est en voie de démocratisation », indique Olivier Broca, directeur commercial d'Amisco, société niçoise qui fournit en données des clubs comme Lille, Lyon, Rennes, Toulouse...

Des économies en vue

Le but ? Mieux analyser le rendement individuel de ses joueurs, affiner les critères de recrutement, et fournir des études des adversaires plus « scientifiques ». « On ne devine pas tout, et heureusement, remarque David Collet, directeur commercial d'Opta, entreprise concurrente d'Amisco. Pour les clubs, c'est un outil supplémentaire, en aucun cas une garantie de victoire. » Une nuance bien comprise en Angleterre, mais moins en France. « En France, on oppose une logique de moyennes humaines, mais aussi de



L'entraîneur lillois Rudi Garcia (à droite) est un des adeptes en France de l'utilisation des statistiques.

culture », poursuit David Collet. « C'est aussi une question de génération, même si quelqu'un comme Aimé Jacquet, en 1998, a été l'un des premiers à s'en servir », remarque Olivier Broca. Aujourd'hui, des coachs comme Frédéric Antonetti, Alain Casanova ou Rudi Garcia sont de fervents défenseurs de la méthode statistique.

C'est que leur utilisation massive peut aussi être un bon moyen de faire des économies. Exemple le plus frappant dans le recrutement. Là où il fallait tout un maillage de scouts sur place, Opta et Amisco offrent un service qui réduit considérablement les coûts. Les statistiques envahissent maintenant des champs comme la formation ou la pré-

paration physique. « Au bout de quelques années, avec toutes ces données, on fabrique un patrimoine qui s'apparente à l'ADN d'un club. Les entraîneurs peuvent changer, on a toujours cet outil à disposition. On n'a pas encore bien compris à quel point ça allait révolutionner le football », promet David Collet. ■

Lire aussi en page cinéma

« Ce n'est pas le GPS qui décide qui sera sur le terrain »

PASCAL VALENTINI

Préparateur physique du Racing-Métro.

A quoi sert le GPS que les joueurs portent à l'entraînement au Racing ?

L'idée est de rationaliser les entraînements et de savoir ce que les joueurs y font exactement. Tout ce qu'on faisait au pifomètre est désormais rationalisé grâce au GPS. On affine par joueur et par groupe de joueurs.

Les statistiques issues de ces GPS ont-elles changé les entraînements ?

Oui, le GPS nous autorise des entraînements moins physiques... On adapte le nombre de joueurs (5-5), petites surfaces, grandes surfaces, rugby à 7...

Existe-t-il un avant et un après-GPS ?

L'outil devient indispensable au fil du temps. Mais ce n'est pas le GPS qui décide qui sera sur le terrain le match suivant. Ce n'est qu'un outil, et il en existe d'autres. Les données transmises par le GPS sont analysées et elles doivent l'être de manière intelligente.

Est-ce grâce au GPS que le Racing a de bons résultats ?

C'est la réflexion contraire. C'est la volonté globale de s'améliorer qui nous a fait venir au GPS. Mais c'est un vrai coût. Donc il faut que ce genre de démarche corresponde à un moment de l'évolution du club. On ne met pas une Rolls-Royce dans les mains d'un gamin de 16 ans. ■

RECUEILLI PAR J. L. M.

■ DRÔLES DE CHIFFRES

100 En productivité absolue par minute, sur une base 100, le Montpelliérain Souleymane Camara est le meilleur joueur du championnat de France en termes de buts et de tentatives.

88,1 Sur une base 100, le Toulousain Antoine Devaux est le deuxième meilleur joueur de L1 à la récupération, derrière le Parisien Blaise Matuidi.

6^e En croisant toutes les données (buts et tirs, distribution, récupération...), Valenciennes est 6^e. Le podium (PSG, Montpellier, Lille) est aussi celui du championnat.

(Source Opta.)

A Valenciennes, la méfiance règne

Si la Ligue 1 commence doucement à se mettre aux statistiques, il reste encore quelques irréductibles. Exemple à Valenciennes où l'entraîneur Daniel Sanchez n'a pas le goût des chiffres. « Ça peut être trompeur. Par exemple, quand on parle de passe ratée, qu'elle soit à 10 m ou très proche entre deux joueurs, ça n'a rien à voir », lâche le technicien arrivé cet été dans le Nord. Même si ces satanées statistiques lui ont récemment donné un bon coup de main en championnat. « Lors de notre match à Rennes, on avait noté que quand cette équipe menait 1 à 0, elle se faisait quasiment rejoindre à chaque fois. Avant le match j'ai dit aux joueurs : "Ne vous affolez pas si on prend un but" et on a égalisé. » ■

F. L.

Le métier de statisticien sera le plus «sexy» de la décennie

Mis à jour le 18/10/2010 à 11:28 |



Hal Varian a créé le Google Price Index qui se base sur les prix des produits achetés en ligne. Crédits photo : IPON-BONESS/SIPA/SIPA

La Toile et les moteurs de recherche sont d'extraordinaires pourvoyeurs d'informations et de statistiques. Le tout est de savoir s'en servir.

Les Bourses et spécialement Wall Street sont rythmées par la publication des indicateurs économiques de fréquence mensuelle. Ce rituel va-t-il disparaître lorsque ces mêmes chiffres seront renouvelés quotidiennement ? Ainsi Google a-t-il développé un indice des prix, le [Google Price Index](#)¹, dont l'intitulé est calqué sur le Consumer Price Index américain. Sauf que le GPI sera établi chaque jour, alors que le CPI est calculé par le Bureau of Labor Statistics mensuellement et connu avec quinze jours de retard.

Le GPI reste encore dans sa phase expérimentale, a expliqué la semaine dernière Hal Varian, le chef économiste de Google, à la conférence de la National Association of Business Economists à Denver. La méthode repose sur la collecte des prix des produits achetés en ligne. Certes la gamme en est moins large que celle «du panier de la ménagère» dans le monde réel. Les dépenses liées au logement y sont moins bien représentées. Elles ne couvrent que 18 % de l'indice Google au lieu de 40 % pour le Bureau du travail américain. Avec un sens consommé de la publicité, Hal Varian a déclaré malgré tout que le GPI exprimait «une tendance déflationniste très nette» depuis Noël dernier. Un avis particulièrement sensible à un moment où le spectre de la déflation tétanise les esprits.

Spécialiste mondialement reconnu pour ses travaux sur les réseaux, longtemps professeur à l'université Berkeley en Californie, Hal Varian est entré en 2002 chez Google comme chef économiste. Il est devenu l'invité vedette des congrès où il ne se lasse pas de répéter «que le job sexy (sic) des dix prochaines années sera celui de statisticien». Avec son parc d'un million de serveurs, un budget de recherche et d'investissement de 3,1 milliards de dollars (en 2009), le moteur de recherche le plus puissant de la planète a pour mission «d'organiser l'information à l'échelle mondiale et de la rendre universellement accessible». Ces termes figurent dans sa note de présentation.

La révolution que prophétise Hal Varian est simple et radicale : les statisticiens vont désormais pouvoir décrire l'économie en temps réel au lieu de la regarder dans le rétroviseur. «Prévoir le présent», tel est son mantra. «Par exemple, s'il y a une rumeur de licenciements, les gens vont chercher à s'informer (via Internet) sur la situation du marché du travail et sur l'adresse de l'agence locale pour l'emploi. Il n'est pas étonnant que le taux de chômage soit corrélé avec ces demandes de renseignement des particuliers», expliquait-il le 4 août dernier devant le Congrès mondial des statistiques de Vancouver. De même est-il possible de connaître à la seconde la diffusion d'une épidémie de grippe en observant le nombre de recherches d'information sur la maladie répertoriées sur la Toile. Du moins existe-t-il une corrélation entre les deux phénomènes.

Les experts ont de tout temps cherché à établir de telles corrélations, la base de leur métier. Le seul changement est que les nouvelles technologies de l'information offrent des possibilités insoupçonnées. Par leur rapidité, leur exhaustivité et leur facilité d'accès.

Les instituts de statistiques nationaux exerçaient une prééminence de fait. L'explication en était principalement le coût élevé de la collecte des «données», mot trompeur, car elles étaient tout sauf gratuites. L'Insee doit ainsi mobiliser 150 enquêteurs pour relever les étiquettes à travers toute la France. La «fabrication» de l'indice des prix à la consommation lui revient à environ 10 millions d'euros par an. «Nous devons respecter énormément de contraintes, qui relèvent d'une réglementation internationale très stricte. Il faut comparer les prix de produits à qualité strictement constante. Ce n'est pas facile et ce n'est pas, semble-t-il, ce que fait l'indice Google ! En même temps nous utilisons nous aussi Internet pour collecter l'information ainsi que les sources professionnelles comme la base Galileo pour les billets d'avion», explique Jean-Louis Lhéritier, responsable de la division «prix et conditions de vie des ménages» à l'Insee. De même des travaux sont en cours pour utiliser les tickets de caisse des magasins, ce qui simplifierait le travail.

Les nouvelles technologies constituent «un défi pour la statistique publique qui va devoir se repositionner», reconnaît Paul Schreyer, directeur des comptes nationaux à l'OCDE. Grâce à elles, d'énormes progrès ont déjà été obtenus dans la connaissance des dépenses de santé, entre autres. La puissance des fichiers administratifs permet également d'obtenir une information économique et sociale à un niveau local de plus en plus fin, note-t-on à l'OCDE. Pour sa part, Eurostat, qui fédère les offices nationaux de l'Union européenne, a lancé un groupe de réflexion «pour mieux utiliser les nouvelles sources qui traînent sur la Toile». Sans oublier les problèmes de déontologie que cela implique.

«Les organismes publics occupaient un monopole naturel pour des raisons de coût et de technicité. Tout cela vole en éclats», observe Jacky Fayolle, administrateur de l'Insee, reconverti dans le secteur privé. Les statisticiens ont toujours reflété la société dans laquelle ils officient, du totalitarisme soviétique à l'État néolibéral. Une nouvelle ère s'offre à eux. Ces hérauts sauront se repérer et nous guider dans le maelström de l'information qui nous accable. Ils seront forcément «sexy».



Publié le 9 septembre 2011 à 19h54 - Mis à jour le 11 septembre 2011 à 15h37

France

Par Hélène Favier

11-Septembre : la «curieuse» collaboration des étudiants d'HEC



Les deux logos - celui de la junior entreprise de HEC et celui de ReOpen911 - trônent côte-à-côte. © CAPTURE D'ECRAN SITE REOPEN911

La junior entreprise de la célèbre école a réalisé un sondage pour l'association ReOpen911.

D'un côté la junior entreprise d'HEC, la très réputée «École des hautes études commerciales» de Paris. De l'autre, l'association polémique ReOpen911 (pour la réouverture de l'enquête sur le 11-Septembre), considérée comme la représentante des «conspirationnistes» en France. A priori, rien ne rapproche la prestigieuse école et l'association qui remet en cause la version officielle des attentats du 11-Septembre. Si ce n'est que ReOpen911 publie, vendredi, une étude sur le 11-Septembre «co-brandée» avec HEC.

ReOpen911 présente «un sondage exclusif HEC»

Selon ce sondage réalisé par les étudiants d'HEC en juin dernier, 58% des Français «expriment des doutes sur la version officielle, contre seulement 31% qui n'ont pas de doutes» sur ce qui s'est passé

le 11-Septembre 2001. «Un tiers (34%) envisagent également que les autorités américaines aient pu être impliquées dans la réalisation des attentats». Un sondage qui satisfait donc pleinement ReOpen911, dont la thèse est que les attentats du 11-Septembre ont surtout «servi de prétexte à une dérive ultra-sécuritaire, liberticide et anti-démocratique de nos sociétés sous couvert de guerre au terrorisme».

Joint par Europe1.fr, la junior entreprise d'HEC se défend d'avoir apporté un quelconque crédit à la thèse «conspirationniste», lors de cette collaboration. «Avec cette étude, nous avons procédé comme d'habitude, nous avons élaboré des questions, nous les avons fait valider à notre client et nous avons interrogé 500 personnes par téléphone», explique David Le Louarn, chargé de projet sur ce sondage.

«On refuse très peu d'études, tant qu'elles sont à notre portée et qu'elles ne sont pas en contradiction avec notre code de déontologie», ajoute l'étudiant pour qui ce sondage et son client rentraient bien dans les règles déontologiques que s'est fixée la junior entreprise. «Cette étude n'est pas moins intéressante qu'une autre. Si un adversaire de ReOpen911 nous avait commandé un sondage, nous l'aurions fait de la même manière», renchérit-il, avant d'insister : «nous ne refusons aucun client pour ses idées».

HEC sert à «crédibiliser» ReOpen911 ?

Reste que du côté de ReOpen911, on se satisfait de ce coup d'éclairage. «Comme nous sommes diabolisés, nous essayons de tout faire pour nous crédibiliser : nous avons recours à des experts, des scientifiques et donc à HEC pour ce sondage», indique Arnaud

Imprimé avec joliprint

<http://www.europe1.fr/France/11-Septembre-la-curieuse-collaboration-des-etudiants-d-HEC-710565/index.html>

Page 1



Publié le 9 septembre 2011 à 19h54 - Mis à jour le 11 septembre 2011 à 15h37

11-Septembre : la «curieuse» collaboration des étudiants d'HEC

Bérard, membre du Conseil d'administration de ReOpen911. «Si nous avions fait le sondage nous-mêmes, on nous aurait accusés d'avoir truqué les résultats», ajoute-t-il, content que son logo soit désormais présenté côte-à-côte de celui de la célèbre HEC.

Interrogé sur le crédit qu'il apporte à une association comme ReOpen911 avec ce «sondage exclusif HEC» : l'étudiant de l'école, David Le Louarn, répond «c'est plutôt flatteur, cela veut dire que nous sommes crédibles».

Contactée par Europe1.fr, la direction d'HEC n'a pas encore donné suite à nos demandes de précisions.

Première partie

Version rédigée du cours

Panorama de la démarche statistique

Le but de ce chapitre est d'illustrer quelques concepts et outils fondamentaux de la statistique 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 types de situations différentes conduisant à une modélisation similaire,

- la répétition d'une expérience (par exemple, on lance une pièce plusieurs fois pour déterminer si elle est équilibrée); c'est le genre de situation que l'on rencontre dans les sciences expérimentales... et que nous ne croiserons donc pas dans ce cours!
- 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'une étude marketing ou lors d'un sondage électoral); c'est le genre de situation que nous traiterons, en revanche.

Où se situe l'aléa ? — Packaging A ou packaging B

On demande à des consommateurs s'ils préfèrent, pour un produit de grande consommation qu'on veut relancer¹, le packaging A ou le packaging B. On interroge n personnes dans un panel de consommateurs et on inscrit les résultats dans un tableau. On code la préférence pour A par 0 et celle pour B par 1.

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

On note $x_1, x_2, x_3, \dots$ les résultats successifs. Par exemple, le troisième consommateur interrogé préfère le packaging B : $x_3 = 1$.

1. Ou alors, il peut s'agir de la question de renommer ou pas un produit ou une enseigne (par exemple, "Camaïeu Homme" vers "Jules", cela est arrivé en l'an 2000; ou AGF vers Allianz en 2009; ou GDF Suez vers Engie en 2015). 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.

Le mathématicien se place aux instants avant les interrogations d'un consommateur donné et considère celles-ci comme des expériences aléatoires : il note $X_1, X_2, \dots$ les variables aléatoires correspondantes. Elles représentent la réponse du consommateur avant que² celui-ci ne soit choisi et interrogé. 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).

Si on tire un petit nombre de consommateurs dans un grand panel, c'est comme si on les tirait avec remise. Cela conduit à une modélisation des X_t en variables aléatoires 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 si le packaging A ou B est préférable revient à se demander si $p_0 < 1/2$ ou $p_0 > 1/2$.

Mais que représente p_0 , au fait ? Quelle est son interprétation ? On a déjà souligné que ce n'était pas la probabilité qu'un consommateur préfère B : chaque consommateur répond de manière déterministe, selon une préférence gravée dans son cœur ; l'aléa n'est pas dans la bouche ou dans la tête du consommateur. Le paramètre p_0 est plutôt la proportion de consommateurs dans le grand panel en faveur de B ; si ce panel est vraiment très grand, les enquêteurs n'ont pas le temps d'interroger tout le monde et ne peuvent accéder à la vraie valeur de p_0 – d'où la nécessité de méthodes statistiques pour malgré tout, pouvoir dire des choses avec un degré de confiance raisonnable sur p_0 .

A retenir pour la suite : le point de départ de toutes nos études statistiques sera une modélisation. Ici, il s'agit du 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 de réalisations $x_1, x_2, \dots, x_n$ issues des variables aléatoires $X_1, X_2, \dots, X_n$ indépendantes et identiquement distribuées selon la loi (inconnue) $\mathcal{B}(p_0)$. On fera ensuite usage de ces réalisations obtenues pour répondre à des questions posées sur p_0 .

2. Notez bien que l'aléa ne se situe pas dans ce que va répondre un consommateur donné mais uniquement dans le choix au hasard de ce consommateur dans un grand panel ! Nous ne ferons pas nôtre le proverbe : « Souvent consommateur varie. »

2. Rappels de cours

Le mathématicien développe des méthodes théoriques mettant en jeu les X_t et étudie par exemple comment les utiliser au mieux pour construire des 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. Vous avez vu cela (peut-être de manière rapide) en classes préparatoires.

2.1. Deux outils bien connus pour commencer

La loi des grands nombres 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 $X_1, \dots, X_n$ des variables aléatoires indépendantes et identiquement distribuées admettant une espérance μ . Alors la moyenne empirique converge vers l'espérance,*

$$\bar{X}_n \stackrel{\text{not.}}{=} \frac{1}{n}(X_1 + \dots + X_n) \xrightarrow{\mathbb{P}} \mu .$$

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, $\bar{X}_n$ doit être proche de μ à partir d'un certain rang. Ce sens est, tout à fait précisément, que pour tout $\varepsilon > 0$,

$$\mathbb{P}\left\{|\bar{X}_n - \mu| > \varepsilon\right\} \longrightarrow 0 .$$

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.

Pour simplifier les notations et lorsqu'il sera clair au vu du contexte de quelles variables aléatoires on parle, on écrira le plus souvent σ^2 au lieu de $\sigma^2(X)$.

Remarque 1.2. 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 $x_1, \dots, x_n$. Les lettres grecques seront quant à elles employées lors de l'énoncé du modèle statistique sous-jacent aux $X_1, \dots, X_n$. La figure 1.1 essaie d'imprimer cette distinction en vous.

Théorème 1.2 (Théorème limite central). *Soient des variables aléatoires $X_1, \dots, X_n$ 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*

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

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 ,

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



FIGURE 1.1 – Les lettres grecques sont pour l'énoncé du modèle théorique et les lettres romaines, pour le traitement des données concrètes.

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

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

Remarque 1.3. Distinguez bien les symboles $\rightarrow$ (convergence “déterministe”), $\xrightarrow{\mathbb{P}}$ (convergence en probabilité) et $\xrightarrow{\text{loi}}$ (convergence en loi).

Remarque 1.4. *A priori*, grâce à mon action de lobbying au sein de la commission de refonte des programmes de mathématiques de classes préparatoires, vous ne devriez plus énoncer le théorème limite central comme la convergence en loi des

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

(ou par toute autre formulation similaire) vers la loi normale standard. Ce qui hélas était le cas auparavant : or, X_n^* est une bien mystérieuse quantité, alors que la formulation du théorème ci-dessus met en jeu l'estimateur $\bar{X}_n$ de μ .

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 limite central indique donc que $\bar{X}_n$ suit (approximativement) une loi normale de moyenne μ et de variance σ^2/n ,

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

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

2.2. Intervalles de confiance asymptotiques

Repartons de l'exemple du packaging et supposons qu'en 100 consommateurs sondés, on ait eu 58 votes en faveur de A et 42 en faveur de B, soit une moyenne observée

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

On pense donc (avec raison) que le vrai paramètre p_0 se situe sans doute autour de la valeur $0.42 = 42\%$ 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 définitivement, le packaging A est préféré par l'ensemble des panélistes, on ne peut en revanche être formel dans le second cas.

Pour quantifier ces incertitudes, on fait donc appel à la théorie, en passant, dans un premier temps, par les X_i . 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 $\sigma^2 = p_0(1 - p_0)$. Le théorème limite central implique que

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

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

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

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 $u = 1.96$. La valeur 95% pour le niveau de confiance 95% est celle généralement considérée (c'est une convention). Comme

$$\begin{aligned} \frac{\sqrt{n}}{\sqrt{p_0(1 - p_0)}} |\bar{X}_n - p_0| &\leq 1.96 \\ \iff p_0 \in I_n &\stackrel{\text{not.}}{=} \left[\bar{X}_n - \frac{1.96 \sqrt{p_0(1 - p_0)}}{\sqrt{n}}, \bar{X}_n + \frac{1.96 \sqrt{p_0(1 - p_0)}}{\sqrt{n}} \right] , \end{aligned}$$

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 $\hat{I}_n$ le contenant et ne dépendant plus de p_0 ; on dispose de deux méthodes à cet effet.

Méthode 1 (à oublier) : par majoration, en utilisant que $1.96 \sqrt{x(1 - x)} \leq 1.96/2 \leq 1$, qui découle de l'inégalité $x(1 - x) \leq 1/4$, il vient :

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

Méthode 2 (ce sera la méthode efficace) : la combinaison du lemme dit de Slutsky³ et du théorème limite central assure que

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

3. Au programme uniquement des classes EC/S ; son énoncé sera donné ou rappelé dans un chapitre ultérieur, n'ayez crainte.

d'où l'on tire, par des calculs similaires à ceux qui précèdent, l'intervalle de confiance asymptotique de niveau 95%

$$\hat{I}_n = \left[\bar{X}_n - 1.96\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + 1.96\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right].$$

Remarque 1.5. Une convention utile pour la suite : on mettra un petit chapeau $\hat{\cdot}$ à toutes les quantités qui ne dépendent que des observations. Notez bien la différence entre I_n , qui est un intervalle aléatoire ne nous servant que du point de vue des calculs théoriques, et $\hat{I}_n$, qu'on peut déterminer entièrement en injectant les valeurs observées (les x_t) dans sa formule et qui est donc un intervalle de confiance.

On conclut donc de cette étude théorique qu'avec probabilité asymptotiquement au moins 95 %, le vrai paramètre p_0 se situe dans l'intervalle $\hat{I}_n$, construit uniquement sur les observations.

Notez que l'aléa (le fait qu'on ne soit sûr qu'à 95 % et pas à 100 % de cet intervalle théorique) est causé par les événements atypiques, comme par exemple, obtenir au moins 90 fois la réponse B sur 100 consommateurs interrogés, alors que les préférences pour A et B seraient par ailleurs égales dans la population de panélistes. 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 recourir à une valeur de u' telle que $\mathbb{P}\{|Z| \leq u'\} = 99\%$ et de reprendre tous les calculs précédents en remplaçant u par u' . 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 1.2 plus loin.

On veut repasser maintenant aux données et appliquer notre formule théorique (celle pour $\hat{I}_n$). Les valeurs observées $x_1, \dots, x_{100}$, de moyenne $\bar{x}_{100} = 0.42$, conduisent à l'intervalle suivant pour la valeur p_0 :

$$\left[\bar{x}_{100} - 1.96\sqrt{\frac{\bar{x}_{100}(1-\bar{x}_{100})}{100}}, \bar{x}_{100} + 1.96\sqrt{\frac{\bar{x}_{100}(1-\bar{x}_{100})}{100}} \right] = [0.42 \pm 0.097] = [42\% \pm 9.7\%],$$

que l'on arrondit à $[42\% \pm 10\%]$. En d'autres mots, on peut (seulement) dire avec une grande confiance que le vrai paramètre p_0 représentant le taux de préférence de B sur le panel est de $42\% \pm 10\%$. On ne peut donc pas encore exclure avec une bonne certitude statistique le cas où c'est le packaging B qui serait préféré.

Remarque 1.6. Attention, s'il y a effectivement une probabilité (asymptotiquement) de 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 $[32\%, 52\%]$: en effet, p_0 appartient ou pas à un tel intervalle déterministe, il n'y a pas d'autre issue. On pourrait déterminer ce fait en interrogeant tous les membres du panel (mais cela prendrait du temps). En pratique, on ne le fait pas et on espère avoir eu de la chance lors de la réalisation de l'intervalle de confiance.

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. Evidemment, à parier, on parierait plutôt sur le fait de ne pas avoir eu le gros lot.

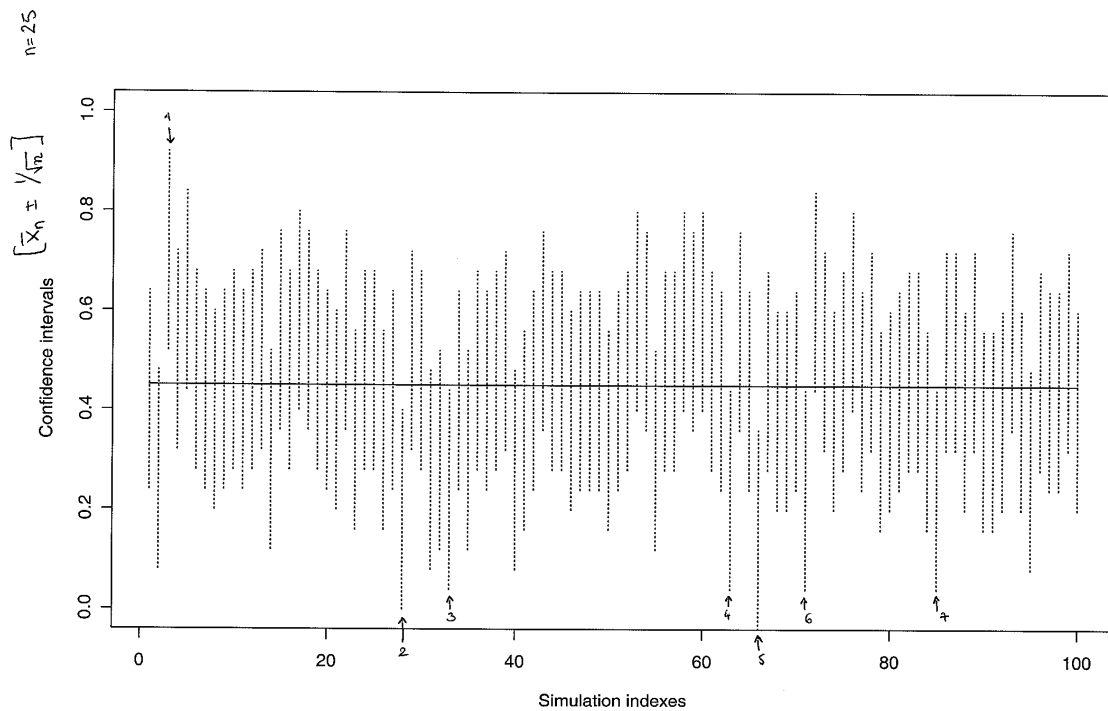
Ici, l'indication de la probabilité *a priori* de 95 % dans le cas des intervalles de confiance nous rassure quant à la validité de ce que l'on obtient après l'expérience sans toutefois que ce sentiment rassurant puisse être une certitude mathématique ; on a juste une indication de bonne confiance statistique.

Remarque 1.7 (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 ? Au moins $1/\sqrt{1\,000}$, soit 3 %, et même mieux lorsque les opinions sont tranchées (que $\bar{x}_n$ ou p_0 sont éloignés de 50%). Mais dans tous les cas, que penser de ces unes de journaux télévisés annonçant, en 2007 (mais plus en 2012!), 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 !

2.3. Deux dessins pour conclure sur les intervalles de confiance

La figure 1.2 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 $p_0 = 0.45$ et on l'a estimé par intervalle comme si on ne le connaissait pas : c'est une simulation typique sous Scilab que vous avez sûrement effectuée l'an dernier.

La figure 1.3 vous montre le genre d'énoncés (pas nécessairement romantiques, on est d'accord) que peut être amené à effectuer un statisticien.



7 intervalles sur les 100 (indiqués par une flèche) ne contiennent pas le vrai paramètre

FIGURE 1.2 – 100 répétitions de la même expérience : on tire informatiquement $X_1, \dots, X_{25}$ indépendantes et identiquement distribuées selon la loi de Bernoulli de paramètre $p_0 = 0.45$ et on construit à chaque fois $\hat{I}_{25}$; on s'est trompé 7 fois sur les 100.



FIGURE 1.3 – La statistique donne des intervalles de confiance. C'est utile, mais pas nécessairement romantique.

3. Tests d'hypothèses

Nous jetons un œil à ce qui sera la première notion nouvelle de ce cours : les tests d'hypothèses statistiques. Ce paragraphe aurait pu admettre pour sous-titre : « Pourquoi deux résultats contradictoires peuvent être simultanément vrais ! » En effet, 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 non rentable d'une promotion commerciale pour nous, mais dans d'autres domaines, 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.

On continue de considérer l'exemple du choix entre deux packagings, qui a été modélisé comme la réalisation de variables aléatoires $X_1, \dots, X_{100}$ indépendantes et identiquement distribuées, selon une loi de Bernoulli de paramètre noté p_0 . On rappelle que p_0 est la proportion (inconnue) de l'ensemble des clients préférant le packaging B au A. Supposons que B soit le packaging actuel ; en changer pour le nouveau packaging, A, engendrerait des coûts supplémentaires.

Un directeur du marketing prudent partira de $H_0 : p_0 \geq 1/2$ (qui correspond à H_0 : “les clients préfèrent l'ancien packaging”), à cause du coût au changement. L'hypothèse alternative serait $H_1 : p_0 < 1/2$. Or, l'intervalle de confiance asymptotique à 95 % précédemment obtenu était $[42 \% \pm 10 \%]$, qui contient des valeurs supérieures à $1/2$. La conclusion est que les données, exploitées pour l'instant⁴ 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 packaging. 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 $\bar{x}_{100} = 42 \%$.

Si un revanche, on avait eu affaire à un directeur marketing ayant proposé ce changement de dénomination et essayant de la faire passer coûte que coûte à sa hiérarchie sous des arguments faussement scientifiques, on aurait pu partir de $H'_0 : p_0 \leq 1/2$ (qui correspond à H'_0 : “les clients préfèrent le nouveau packaging”) contre l'hypothèse alternative $H'_1 : p_0 > 1/2$. Là encore, le test aurait conservé H'_0 , ce qui signifie que les données n'indiquent pas que le nouveau packaging n'est pas préféré ; mais cela n'est pas du tout équivalent à dire que les données indiquent formellement que le nouveau packaging est préféré !

Remarque 1.8. 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

4. Nous verrons comment effectuer une exploitation directe sans détour par les intervalles de confiance.

qu'on conserve l'hypothèse de départ, on n'a pas vraiment progressé : 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 statistique, une hypothèse est vraie tant qu'elle n'a pas été gravement contredite ! C'était la première minute citoyenne...

Ainsi, les tests 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).

En fait, on verra que les tests statistiques, c'est parfois (quand on conserve H_0) savoir bien conclure par des doubles négations et ne pas trop se mouiller...

Recueil, représentation et modélisation de données

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 recueillies¹ 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 un fichier de données, on a regroupé les salaires horaires d'infirmières américaines. Les données se présentent comme indiqué à la figure 2.1 : 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 «individu» (par groupe de valeurs observées pour la même personne) ; tandis que chaque colonne donne les valeurs prises sur les individus pour une quantité mesurée donnée (par exemple, la quatrième colonne liste les salaires). On appelle ces quantités les «variables».

On remarque que la matrice ne comporte que des nombres.

La minute SPSS 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

1. 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 !

hourlywagedata.sav [Ensemble_de_données4] - SPSS Statistics Editeur de données

Fichier Edition Affichage Données Transformer Analyse Graphes Outils

30 :

	position	agerange	yrsscale	hourwage	va
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		.

Affichage des données Affichage des variables

FIGURE 2.1 – 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).

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 2.1.

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?

On peut noter les données de la ligne i par le quadruplet (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, comme c'est le cas par exemple à la ligne 22. (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 SPSS 2.2. Pour compter les valeurs manquantes, on utilise Analyse / Statistiques descriptives / Effectifs.

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 2.1 – Table de correspondance pour les données de la figure 2.1.

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'occurrence, le salaire horaire). Ici, il est difficile de dire si la mesure effectuée est discrète ou continue : on saute 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 2.1 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 2.2.

La minute SPSS 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 2.2 – 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 2.2.

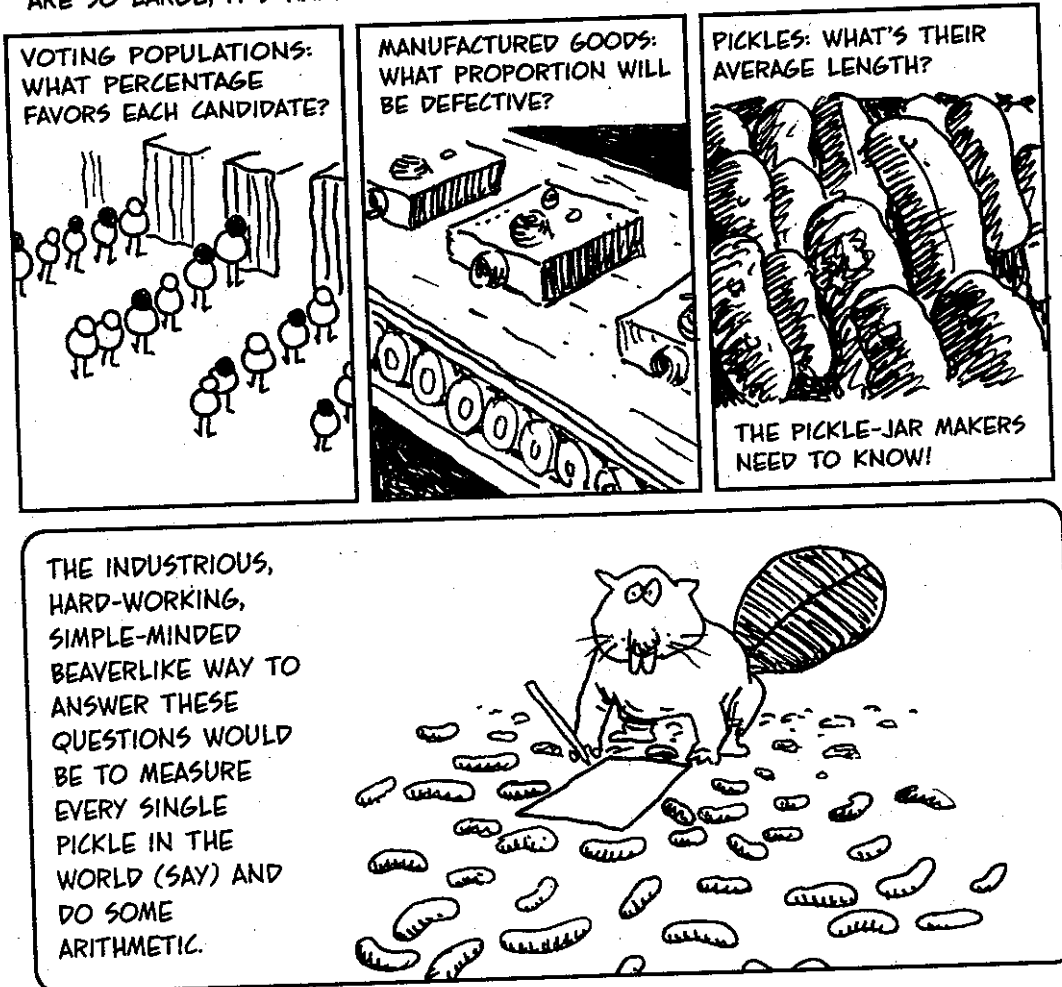
On constitue donc un échantillon, composé dans l'exemple introductif de 3 000 infirmières, on en interroge ses membres et on reporte les réponses dans le tableau de données. Evidemment, telle ou telle réponse est inexploitable, et c'est ce qui donne les 2911 groupes de données complètes de la figure 2.1.

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

THE PROBLEM WITH THE WORLD IS THAT THE COLLECTIONS OF STUFF IN IT ARE SO LARGE, IT'S HARD TO GET THE INFORMATION WE WANT:



BUT WE AREN'T BEAVERS—WE'RE STATISTICIANS! WE'RE LOOKING FOR THE EASY WAY OUT...

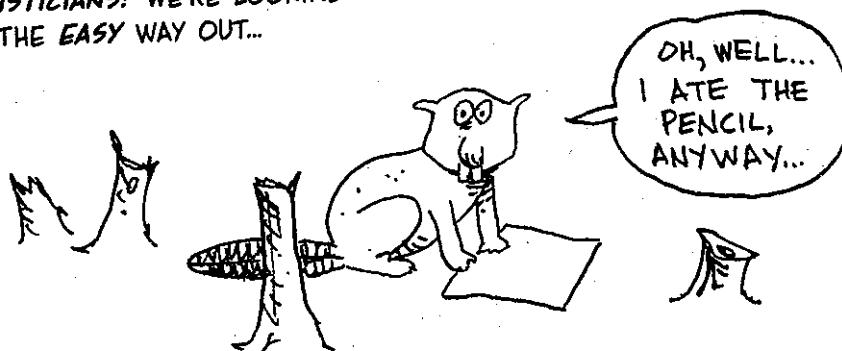


FIGURE 2.2 – Vous n'êtes pas des petites bêtes industrielles : en conséquence de quoi, vous ferez désormais des sondages plutôt que des décomptes exhaustifs sur des populations immenses.

NOT TO PROLONG THE MYSTERY, THE WAY TO GET STATISTICALLY DEPENDABLE RESULTS IS TO CHOOSE THE SAMPLE AT **random**.

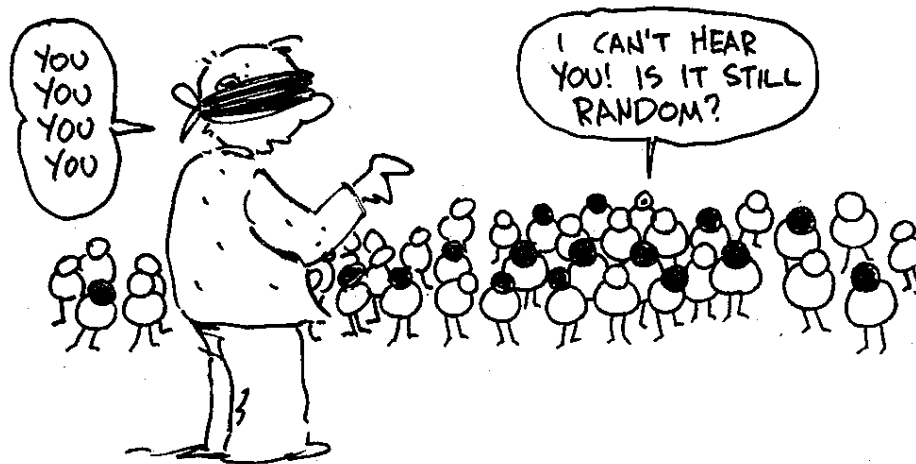


FIGURE 2.3 – La bonne manière de collecter des données : interroger au hasard !

Exemple 2.1. Les données de salaire $h_1, \dots, h_{2911}$ des infirmières peuvent être considérées comme la réalisation d'un échantillon $H_1, \dots, H_{2911}$ 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 $h_1, \dots, h_{2911}$ à ceux en majuscules, $H_1, \dots, H_{2911}$.

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 sans considération² de leur lieu de travail, 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 le paramètre 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.).

Cela étant, toutes ces hypothèses et conclusions valent dans un monde idéal. En pratique, il existe ce qu'on appelle des biais, qui favorisent la représentation de telle ou telle sous-catégorie de

2. En revanche, si l'on tirait d'abord uniformément au hasard le lieu de travail, puis une infirmière uniformément au hasard dans ce lieu de travail, il pourrait y avoir un biais, les infirmières issus de lieux petits en effectifs seraient sur-représentées.

la population dans l'échantillon. Ces biais peuvent venir soit du sondeur (biais géographiques et temporels, biais lié au médium d'interrogation, utilisation d'un mauvais aléa, voire... manipulation lors de la constitution de l'échantillon), soit des sondés (biais de motivation ou de sincérité).

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 2.4).

Dans le même genre d'idées, on peut penser aux biais terribles liés aux sondages par téléphone dans les années 1940 et 1950 (ou comment ne sonder que les riches) ; ou à ceux liés aux sondages par Internet de nos jours.

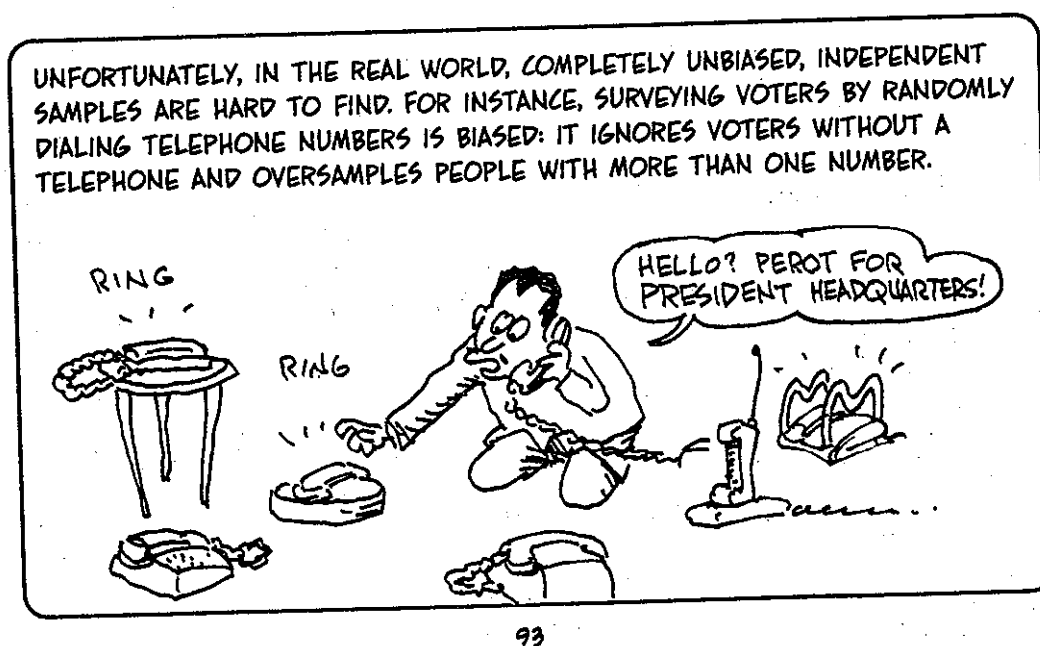


FIGURE 2.4 – Il est difficile d'obtenir un échantillon aléatoire bien représentatif lors d'un sondage téléphonique aléatoire : on n'a pas accès à l'opinion de ceux qui n'ont pas de téléphone et l'opinion de ceux qui ont plusieurs lignes compte davantage !

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 semaine de quatre jours (et demi),

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 (toutes les deux / trois heures ?) et n'interroger qu'un nombre de clients raisonnable : en proportion de la fréquentation (sans quoi les clients des heures creuses seraient sur-représentés).

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 facilité et d'endogamie). C'est celui qui vous menace le plus et je le constate via FaceBook : souvent, vos camarades 3A font tourner leurs questionnaires (écrits sur SurveyMonkey ou GoogleForms) par cette plate-forme et comptent sur le bouche-à-oreille pour leur diffusion et la collecte de réponse. Ils pensent détecter les biais en incluant des questions sur l'âge, le sexe, etc., mais c'est insuffisant. En effet, leur réseau et les réseaux des amis de leur réseau sont constitués de gens qui leur ressemblent et partagent donc, dans une certaine mesure, les mêmes goûts et intérêts. Pas vraiment un échantillon représentatif de la population française (mais parfois, assez représentatif des CSP+ de l'Ouest parisien ou francilien).

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 :

2123717 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).

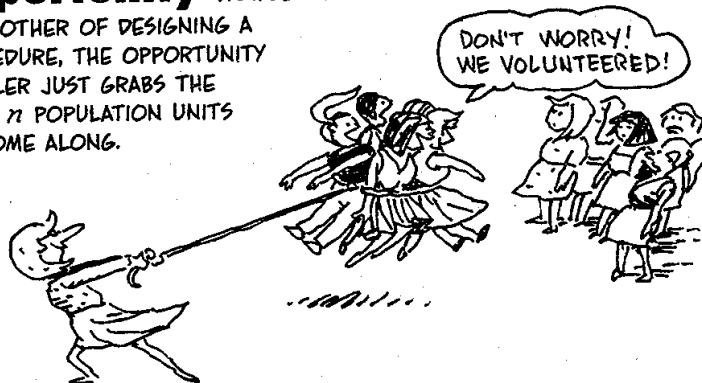
- 1) Il y a un fort biais (de motivation) : l'échantillon n'est nullement représentatif de l'ensemble des Français.
- 2) La portée statistique est quasi-nulle : tous les gens motivés et concernés (= la population visée, en fait) ont voté et on a dépuillé : il n'y a plus d'incertitude, on connaît ici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

FIGURE 2.5 – Un extrait de quizz posé à l'automne 2009.

Exemple 2.5 (Biais de motivation). Les figures 2.5 et 2.6 illustrent 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 ou un peu bizarres, qui ont souvent davantage envie de parler avec vous). 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. A cet égard, la votation citoyenne sur le statut de la poste, étudiée à la figure 2.5 (c'est un extrait de quizz), n'est nullement représentative de l'opinion générale de la population française. Enfin, un autre exemple fameux d'étude sans doute entachée d'un biais de motivation est constitué par une étude de 2001 portant sur des étudiants américains en "Spring Break" à Cancún au Mexique ; je vous laisse deviner sur quoi elle portait...

A COMMONLY USED METHOD IS ESPECIALLY PRONE TO BIAS: IT'S CALLED AN **opportunity** SAMPLE. AVOIDING ALL THE BOTHER OF DESIGNING A PROCEDURE, THE OPPORTUNITY SAMPLER JUST GRABS THE FIRST n POPULATION UNITS TO COME ALONG.



A CLASSIC EXAMPLE IS SHERE HITE'S BOOK, *WOMEN AND LOVE*. 100,000 QUESTIONNAIRES WENT TO WOMEN'S ORGANIZATIONS (AN OPPORTUNITY SAMPLE), ONLY 4.5% WERE FILLED OUT AND RETURNED (RESPONSE BIAS). SO HER "RESULTS" WERE BASED ON A SAMPLE OF WOMEN WHO WERE HIGHLY MOTIVATED TO ANSWER THE SURVEY'S QUESTIONS, FOR WHATEVER REASON.



FIGURE 2.6 – 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.

Exemple 2.6 (Biais de manipulation). La figure 2.7 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...

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.



FIGURE 2.7 – Un dessin tiré du *Canard enchaîné* (le contexte étant la sélection de figurants lors des déplacements présidentiels souvent petits et encartés UMP).

Ce que l'on peut quand même dire dès maintenant, c'est qu'en pratique, pour réaliser «physiquement» une enquête en ligne, on peut recourir aux sites web SurveyMonkey s'il y a peu de questions, et créer un formulaire Google Forms sinon. Dans les deux cas, on récupère les réponses sous la forme de fichiers texte ou de tableurs, et on les traite *via* un logiciel statistique (*modulo* un codage numérique des données qualitatives si besoin). Réaliser une telle enquête pourra arriver rapidement lors de votre scolarité, ... par exemple pour l'épreuve professionnelle de la campagne JÉ : voir un exercice de quiz plus loin dans ce polycopié !

Remarque 2.5 (Une autre méthode de sondage). Pour parer aux biais cités précédemment, on recourt souvent à une méthode de sondage par quotas, qui 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.

2.3. Intermède : lisons la prose des Anciens

... Parce qu'aussi surprenant que cela puisse sembler, ce cours de statistiques vous servira dans votre vie professionnelle future ! ... Exercice : Prenez connaissance de la lettre reproduite à la figure 2.8 et proposez une réponse aimable et informative. Voilà la réponse que j'ai écrite à l'époque ; la quantification des marges d'incertitudes procède de techniques que nous avons vues rapidement au chapitre 1 et reverrons plus en détails au chapitre 7.

Bonsoir XXX,

Ca me paraît OK.

Il y a un autre biais (qu'on a vu en cours) : le biais de motivation. Ceux qui ont à se plaindre rempliront plus volontiers votre questionnaire que ceux qui sont satisfaits !

Cependant, ce que vous appelez biais numérique, c'est ce qu'on a appelé l'aléa d'échantillonnage. Votre marge d'incertitude finale sur le taux sera au plus de $\pm 1/\sqrt{n} = \pm 10\%$ ici. Mais ce sera sans doute mieux, parce que les intervalles de confiance a posteriori prennent en compte le taux réalisé (et on gagne quand ce taux est loin de 50%).

Monsieur,

Je vous contacte dans le cadre du stage que j'effectue actuellement en boutique chez Orange. J'aurais besoin de votre avis et de votre aide sur un questionnaire que je compte rédiger et administrer.

En effet, je travaille à améliorer l'Excellence Commerciale, sorte de classement des magasins Orange basé sur la satisfaction client de la boutique.

J'ai défini la satisfaction client par trois facteurs (variables), à savoir, la qualité de la réponse apportée au problème du client, la qualité du contact et de l'accueil du client, et la confiance qu'a le client dans la prise en compte de son avis. C'est cette troisième variable que j'aimerais évaluer, car je pense que tous les sondages et autres bornes électroniques récoltant les feedback que nous avons mis en place ne permettent pas de la mesurer. Cette variable se situe à un niveau supérieur, puisque le sondé doit être interrogé sur le sondage même auquel il vient de répondre.

A ce titre, je compte élaborer un questionnaire de cinq questions. J'ai déjà administré un questionnaire-test à une vingtaine de personnes, puis je l'ai modifié et recentré sur la problématique. Les questions sont les suivantes :

- Avez-vous remarqué la borne qui permet de donner votre avis ? (Question qui vise à mesurer la visibilité de la borne et de voir quelle importance le client y accorde)
- Appréciez-vous la démarche qui consiste à mettre une borne pour recueillir les avis des clients ?
- Pensez-vous que votre avis aura un impact inexistant, faible ou fort sur l'amélioration des prestations en boutique ? (Question maîtresse)
- Répondrez-vous aussi au sondage qui vous sera éventuellement envoyé par mail si l'on vous explique que l'objectif de ce sondage est différent de celui de la borne ?
- Recommanderiez-vous notre boutique (question qui vise à créer des sous-groupes pour mener éventuellement une étude comparative). Si non, pourquoi ?

J'ai déjà décelé quelques biais pour l'étude :

- Le biais temporel : il faut faire attention à distribuer les questionnaires à des moments différents afin de bien couvrir toutes les conditions possibles de vente en boutique (beaucoup ou peu de monde dans la boutique, etc.), ainsi que tout le panel de clientèle (le client du midi, businessman, n'est pas celui du soir, etc.).
- Le biais numérique. J'estime pouvoir récolter une centaine de réponses environ. Question : quelle sera alors ma marge d'erreur ?

Auriez-vous des remarques à faire par rapport à ce questionnaire ? Des améliorations à apporter ? Et sur quelle marge d'erreur puis-je m'appuyer ?

Merci beaucoup pour votre aide,

XXX, ex-élève de L3

FIGURE 2.8 – La copie d'un mail qui a été envoyé par un de vos prédécesseurs pendant son stage de fin de L3 (fin mai 2011).

Là encore, je vous renvoie au cours : fiches de synthèse du chapitre 7 et de nombreux exercices sur le sujet...

Et la réponse de l'étudiant à la réponse :

Ok, super ! Merci pour votre réponse rapide.

J'avais pensé au biais de motivation mais j'avais oublié de le marquer :D.

Bonnes vacances, et à septembre si on a encore l'heur d'avoir un cours de stats en 2A...

... Au fait, chers étudiants actuels, il n'y a plus de cours de statistiques en deuxième année, ceci est votre premier et dernier cours formel de statistiques (mais les cours de marketing vous introduiront à quelques méthodes statistiques supplémentaires). Et l'expérience montre que ce cours est utile dès le premier stage, et également pour les missions JE !

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 2.2.

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 : les observations $X_1, \dots, X_n$ 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 générale»

Sauf cas particuliers (décrits dans les paragraphes suivants), on recourra donc généralement à cette seconde modélisation, vague mais générique :

$X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une certaine loi, d'espérance μ_0 (et de variance σ_0^2).

3.2. La loi de Bernoulli

Elle modélise le comportement d'une 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.

On considère à cet effet la démarche générale de modélisation suivante.

La démarche de modélisation

1. Définir la population effectivement visée, et préciser l'échantillon qui en a été tiré.
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); elles se rapportent bien sûr à l'échantillon.
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.

Attention! Pour la description des données comme pour l'interprétation du paramètre d'intérêt, on choisira avec soin le temps et le mode du verbe (présent, passé composé, futur simple, conditionnel), selon le contexte et les intentions de l'étude menée.

Correction 2.1. Voici la solution de l'exercice, au vu de la démarche générale proposée :

1. Population visée : les étudiants HEC en cours de scolarité et vivant sur le campus ; échantillon : on a sélectionné 94 répondants au hasard dans cette population.
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 (d'échantillon) : $x_1, \dots, x_{94}$, appartenant à l'ensemble $\{0, 1\}$.
4. Description des données : on observe une moyenne sur l'échantillon de

$$\bar{x}_{94} = \frac{1}{94} \sum_{i=1}^{94} x_i = \frac{51}{94} \approx 54.3 \% .$$

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.

3.3. La loi normale

Elle est la loi d'observations quantitatives qui résultent de la combinaison de nombreux effets ; c'est le théorème limite central qui explique pourquoi elle est si fréquente. Attention : on n'y recourra que lorsque la taille d'échantillon sera petite (et sinon, on se contentera de la «loi générale»).

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 2.7 (Retour à l'exemple des salaires des infirmières). La figure 2.9 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) : graphiquement, il ressort l'impression que la loi commune des salaires $H_1, \dots, H_{2911}$ 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.

La minute SPSS 2.4. Pour obtenir l'histogramme de la figure 2.9, 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

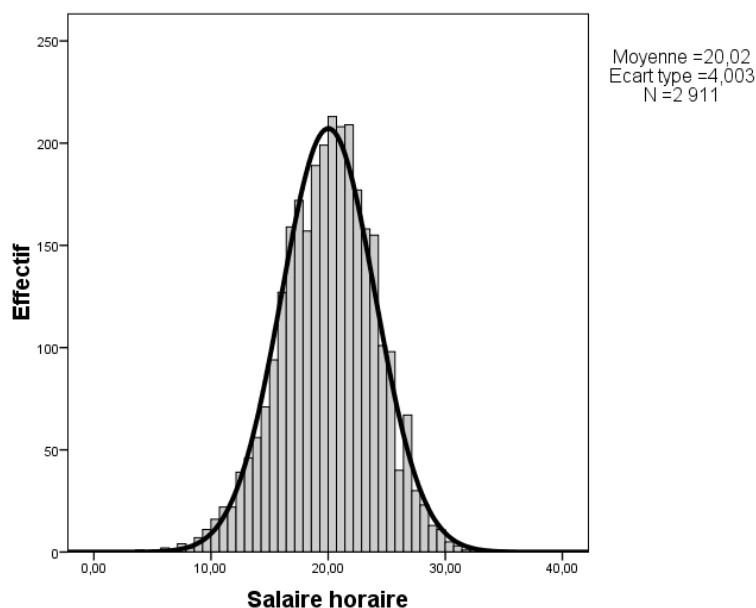


FIGURE 2.9 – Histogramme des salaires horaires déclarés par l'échantillon d'infirmières sondées.

telle somme de petits phénomènes aléatoires conduit à une loi normale : c'est ce que dit le théorème limite central.

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 des membres d'une population, puisque la taille dépend des gènes, mais aussi de l'environnement, et notamment, de l'alimentation ;
- le rendement d'un champ, qui dépend de l'exposition de la parcelle, de la pluviométrie, de la qualité de la terre.

3.4. La loi log-normale (exclusivement pour la culture et parce qu'on va parler d'argent)

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 $x_1, x_2, \dots$, il suffit de voir si les transformées $\ln x_1, \ln x_2, \dots$ 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 limite central, à appliquer dans une exponentielle : les écarts de salaires entre deux professions à qualifications différentes sont généralement

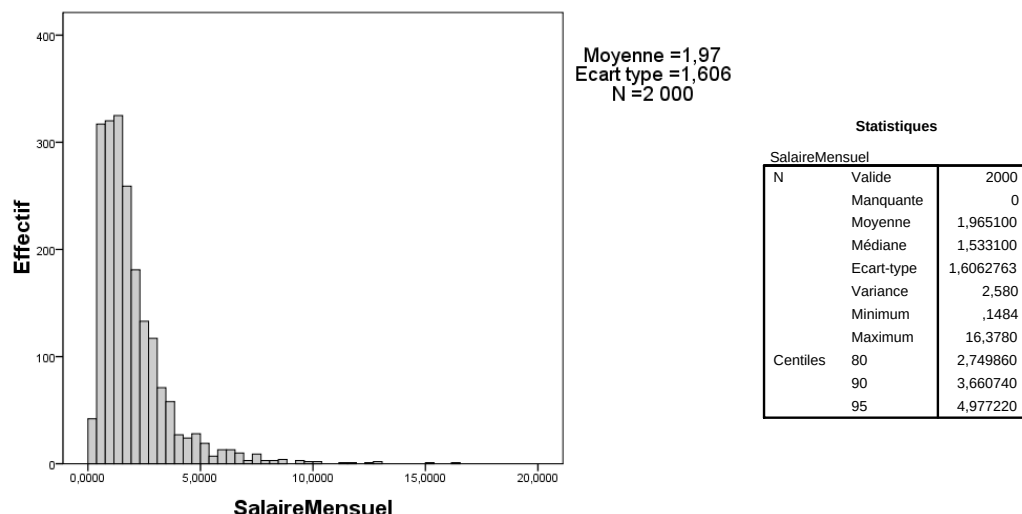


FIGURE 2.10 – Statistiques descriptives des données simulées de salaires mensuels nets (en milliers d'euros).

fixés en facteurs multiplicatifs, de la forme $1 + a \approx e^a$, où a est petit entre deux professions proches (par exemple, d'ouvrier à contremaître, ou de professeur certifié à professeur agrégé).

La minute SPSS 2.5. Sous SPSS, c'est avec Analyse / Statistiques descriptives / Effectifs que l'on calcule la moyenne et la médiane, le minimum et le maximum, ainsi que quelques quantiles³ (les déciles à 80 % et 90 %, puis le quantile à 95 % par exemple). On peut également faire afficher un histogramme des données. On obtient quelque chose de similaire à la figure 2.10.

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 médiane 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 : le salaire du Français moyen plutôt que le salaire moyen des Français. 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 de la fin des années 2000 ; 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 (à 80 % et à 90 %) sont trop élevés sur ces données simulées. Ainsi, d'après l'INSEE (voir le tableau de la figure 2.11), 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 % en 2015 est donc autour de 3 680 euros nets actuels environ, si l'on tient compte d'une évolution moyenne de ces hauts salaires de 2 % 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

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 %.

3 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 2.11 – 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).

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 2.12. On retiendra notamment que le ressenti des 50 % les plus centraux de la population est décrit par la boîte centrale.

Remarque 2.10. Attention ! Les moustaches ne correspondent pas aux déciles à 10% et à 90%, contrairement à ce que l'on a (mal) pu vous enseigner. Ils sont donnés par 1,5 fois l'écart-interquartile reporté à partir du premier ou du troisième quartile, arrêté à la donnée immédiatement au-dessus ou en-dessous.

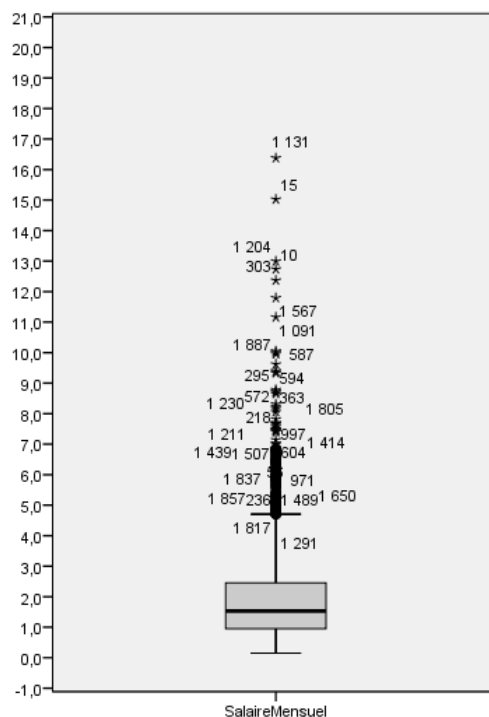
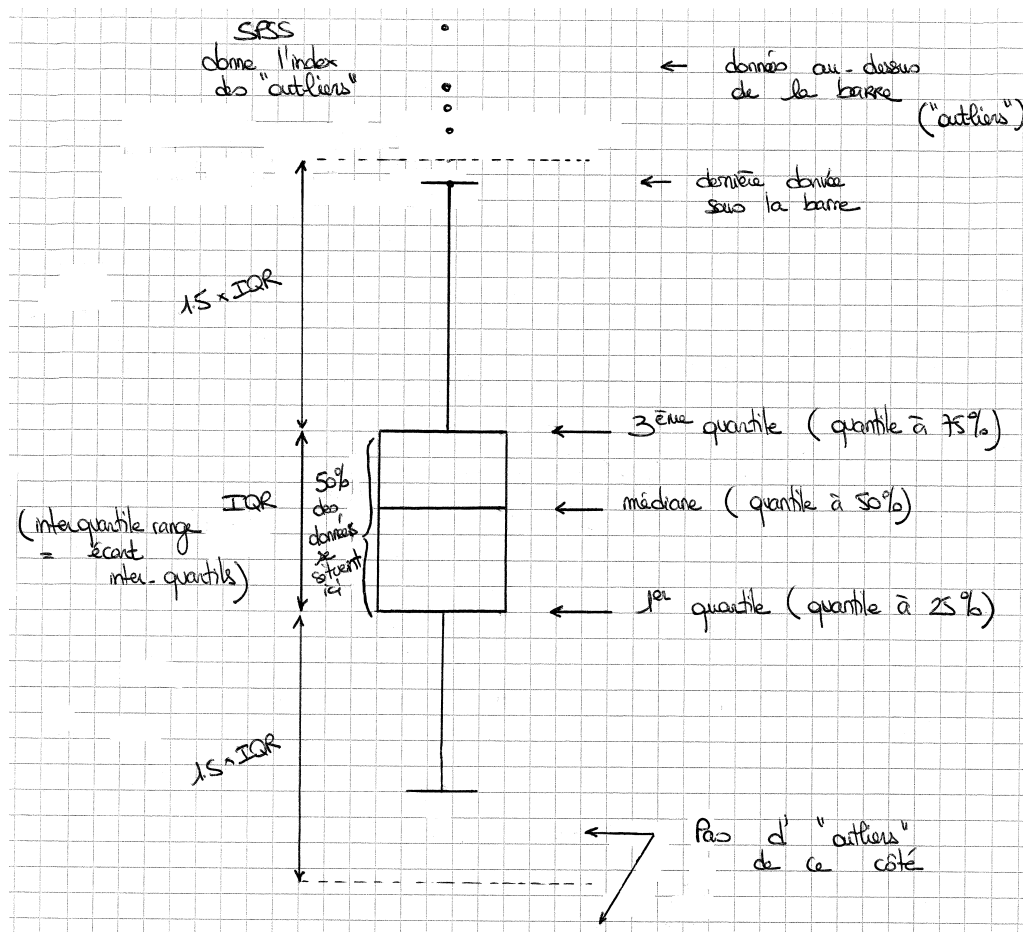


FIGURE 2.12 – Boîtes à moustaches : principe et application.

4. Résumé

Il faut retenir de ce chapitre le principe de la chaîne de recueil, description et traitement des données ; les chapitres suivants porteront sur les statistiques inférentielles mentionnées dans la dernière case.

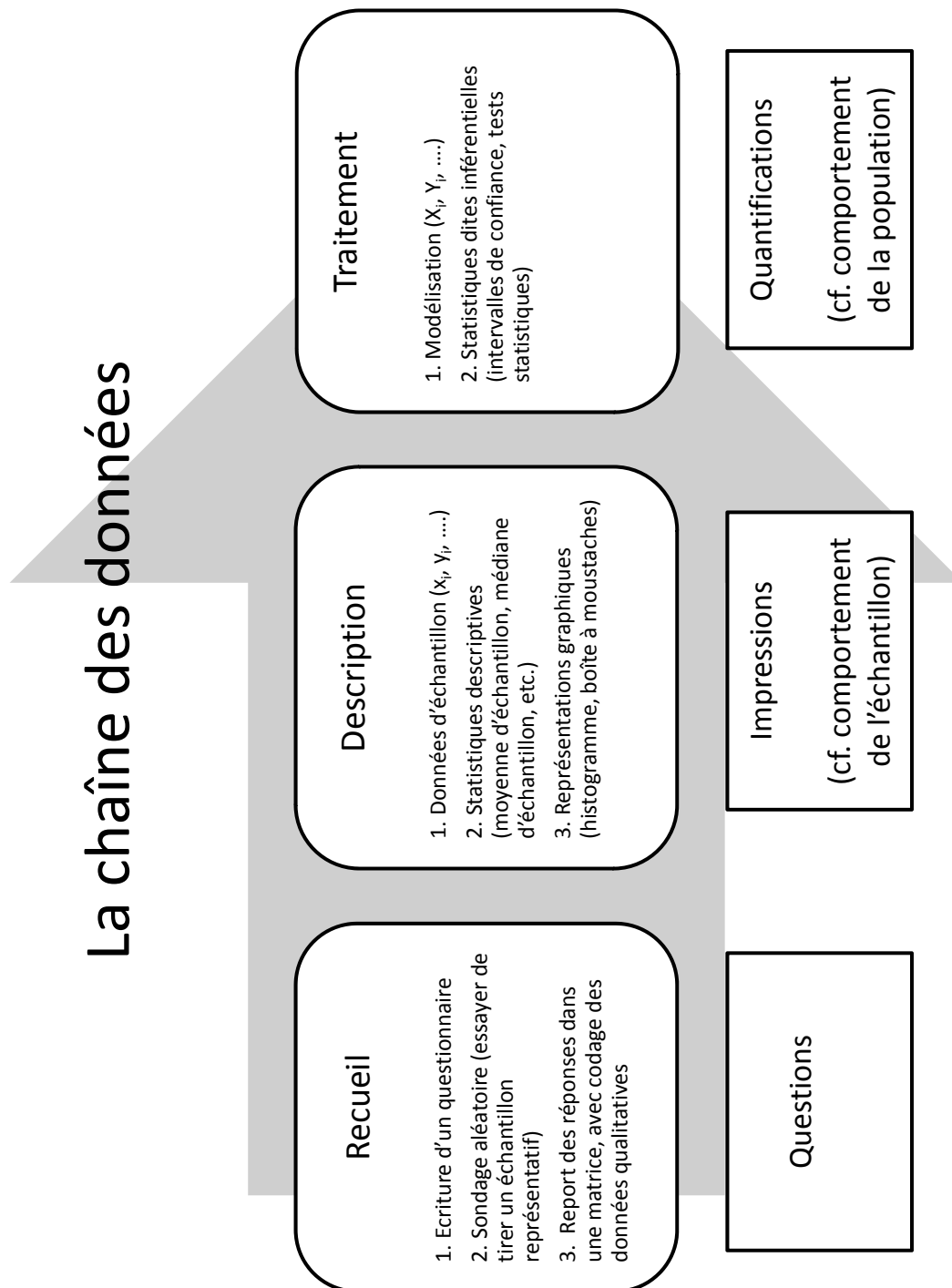


FIGURE 2.13 – La chaîne de recueil, description et traitement des données.

Introduction aux tests d'hypothèses

Objectifs. 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. trois 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.

En revanche, la mise en œuvre 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) est laissée pour le chapitre suivant.

1. Méthodologie des tests : trois exemples

1.1. «The lady tasting tea»

Ce premier exemple est cité par Pere Grima dans *La Certitude absolue et autres illusions* :

À la fin des années 1920, à Cambridge, en Angleterre, un groupe de professeurs, leurs épouses et leurs invités prennent le thé dehors, profitant d'un après-midi agréable. Après avoir bu une première gorgée, une dame, une tasse à la main, déclare avoir noté que la

saveur du thé est différente selon que le thé est servi avant ou après le lait. Poliment, bien sûr, une personne fait quelques remarques sur la difficulté d'admettre ce fait et cela donne lieu à une discussion au cours de laquelle on brandit toutes sortes d'arguments tirés du monde de la physique et la chimie : la composition du produit résultant est la même si l'on verse d'abord le thé ou le lait, les particules dissoutes sont à la fin toutes les mêmes, le degré de température ne peut pas influencer le goût, etc. Non, il est impossible de distinguer une tasse d'une autre... ou peut-être a-t-on manqué quelque chose ?

Une des personnes présentes, un homme d'environ 40 ans nommé Ronald Aylmer Fisher, propose de dissiper tous les doutes possibles au moyen d'une procédure «révolutionnaire» : effectuer une épreuve de dégustation. Bien sûr, on ne peut pas procéder avec une seule tasse de chaque type, car la probabilité de tirer l'une d'entre elles au hasard est juste de $1/2$ et l'on ne saurait pas si cela est dû au hasard, ou si l'on est réellement capable de distinguer un mélange de l'autre. Mais si vous disposez de 4 tasses de chaque type, la probabilité due au hasard est seulement de 1 sur 70 (il y a 70 façons distinctes de choisir 4 objets parmi 8). Dans ces conditions, on peut donc affirmer avec certitude être en mesure de distinguer une préparation de l'autre avec une faible probabilité d'erreur, probabilité connue, de surcroît.

A l'époque, Fisher était déjà un professeur célèbre qui publiera en 1935 un texte de référence ayant marqué un avant et un après dans les stratégies de recueil de données par l'expérimentation. Son livre, intitulé *The Design of Experiments*, présente dans son deuxième chapitre certains des concepts clefs utilisant cette situation comme fil conducteur.

Le raisonnement précédant la dégustation de thé

Dès le départ, nous supposons que la dégustatrice de thé ne peut pas distinguer les deux préparations. C'est ce qui nous semble normal et nous croirons en ses capacités uniquement si les données recueillies dans le cadre d'une expérience bien conçue et contrôlée vont à l'encontre de l'hypothèse initiale. «Aller à l'encontre» signifie que nos résultats sont peu probables dans le cas où l'on ne sait vraiment pas faire la distinction et nous définissons ainsi ce que nous entendons par «peu probable» : ce qui survient dans moins de 5% des cas, dans moins de 1% des cas ou toute autre valeur.

Si nous sommes disposés à nous fier à l'expérimentation seulement si ses résultats peuvent se produire par hasard (c'est-à-dire, sans aucun mérite de la part de la dame) dans moins de 5% des cas, on ne pourra [peut-être] pas se contenter de faire goûter 3 tasses de chaque type puisqu'il existe 20 façons possibles de choisir 3 objets parmi 6 et qu'il n'y a qu'une manière correcte de le faire. Donc, la possibilité d'y arriver par le seul hasard est de 1 sur 20, soit 5%. [...] Si nous avons 4 tasses de chaque type, le nombre de sélections possibles de 4 tasses sera de 70, et comme il n'y a qu'un seul ensemble de 4 tasses dans lequel on a versé d'abord le thé, puis le lait, la probabilité de le trouver par hasard est de 1 sur 70, soit 1,4%. Si l'on se trompe d'une tasse sur les 4 tasses choisies, on ne pourra pas considérer raisonnablement que l'on sait les distinguer, puisque la probabilité que cela se produise par hasard est d'environ 23%. Mais nous ne devons pas tout miser sur le seul raisonnement mathématique. Il faut également être attentif aux détails lors de la réalisation de l'expérience et ne pas donner d'indices à la dégustatrice... Fisher a donc insisté expressément pour que les tasses soient présentées dans un ordre aléatoire :

«Notre expérience consistera à préparer huit autres tasses de thé, quatre selon un type de préparation et quatre selon l'autre type, et à les apporter ordre aléatoire) à la dégustatrice pour qu'elle nous donne son opinion. On lui expliquera préalablement en quoi consistera le test : on lui apportera huit tasses de thé, quatre de chaque type, dans un ordre aléatoire

(décidé par des dés, la roulette, des cartes, etc. ou tout simplement par des numéros tirés au hasard) et sa tâche consistera à séparer les deux séries de 4 tasses en les classant, si elle le peut, selon qu'a été versé en premier le thé ou le lait.»

Qu'est-il arrivé ensuite ? Fisher ne raconte pas dans son livre quel a été le résultat de l'expérience, mais parmi les personnes présentes se trouvait le professeur Hugh Smith, qui a raconté cette histoire à David Salsburg, auteur d'un excellent livre sur l'explosion des statistiques au XX^e siècle, *The Lady Tasting Tea*. L'ouvrage commence par raconter cette histoire, qui, bien sûr, donne son titre au livre. Le professeur Smith nous raconte que, finalement, la dame a bien identifié chacune des tasses.

Remarque 3.1 (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 et on passe à l'hypothèse alternative.

Ici, l'hypothèse de départ, notée H_0 , est que la dégustatrice ne peut pas distinguer les tasses selon leur mode de préparation, alors que l'hypothèse alternative H_1 est qu'elle le peut. On suppose que H_0 soit vraie, on effectue la dégustation, et si l'événement de faible probabilité sous H_0 se produit, à savoir, la parfaite classification des 8 tasses, alors on en déduit non pas que la dégustatrice a eu de la chance, mais qu'elle possède un don : on rejette H_0 et on passe à H_1 .

1.2. A bas les 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 3.2 (A retenir, suite !). 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.

1.3. Commercialisation d'un 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 : 64 d'entre eux ont connu un arrêt temporaire.

- Modélisez la situation et précisez notamment le paramètre d'intérêt.
- Effectuez un test adéquat sur ce paramètre (on pensera avoir affaire à un laboratoire scientifiquement honnête).
- Concluez : que doit faire l'équipe marketing du laboratoire ?

Modélisation. La population visée est l'ensemble des fumeurs prêts à s'arrêter de fumer (suffisamment motivés pour s'acheter des patchs!). On en prend au hasard 100 personnes, sur qui on teste le patch : c'est notre échantillon. On note $x_j = 1$ si la j -ème personne a connu un arrêt au moins provisoire de sa consommation, $x_j = 0$ sinon. On dispose ainsi de données $x_1, \dots, x_{100} \in \{0, 1\}$. On peut supposer qu'elles sont la réalisation de variables aléatoires $X_1, \dots, X_{100}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0 . Ce paramètre p_0 est le paramètre d'intérêt : c'est le taux, parmi les futurs clients si le produit était commercialisé, d'arrêt au moins temporaire de consommation. Il est inconnu, mais on peut l'estimer par $\bar{x}_{100} = 64\%$. Cependant, notre volonté est surtout de savoir s'il est supérieur à 60% ou non. (Est-ce que la proportion d'échantillon de 64% est significativement supérieure à 60%, sachant que l'échantillon est de taille 100 ?)

Choix des hypothèses à tester. On pose comme hypothèse de départ $H_0 : p_0 \leq 60\%$; 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 ternir la réputation de notre laboratoire pour une publicité un peu trop alléchante ! L'hypothèse alternative, vers laquelle on se tournerait si H_0 se révélait en contradiction avec les données et leur modélisation, est celle qu'espère le service marketing, à savoir $H_1 : p_0 > 60\%$.

Nous expliquerons ci-dessous comment, au vu de H_1 , il suffit de considérer le cas limite $p_0 = 60\%$ sous H_0 .

Choix d'une statistique de test. Sous l'hypothèses qu'on a exactement $p_0 = 60\%$, le théorème limite central assure que

$$T_n = \frac{\sqrt{n}}{\sqrt{0.6 \times (1 - 0.6)}} (\bar{X}_n - 0.6) \rightarrow \mathcal{N}(0, 1) .$$

Nous allons retenir T_n comme statistique de test.

Comme $\bar{X}_n$ est typiquement proche de p_0 (par loi des grands nombres), on en déduit que sous H_1 , la moyenne $\bar{X}_n$ et donc la statistique de test T_n tendent à prendre des valeurs plus grandes que sous H_0 . Que T_n soit au-dessus d'un certain seuil nous mettra la puce à l'oreille en faveur de H_1 .

D'autre part, que se passe-t-il sous $H_0 : p_0 \leq 60\%$? Intuitivement, les valeurs prises par T_n sont d'autant plus petites que p_0 est petit. Le cas limite est donc le cas où $p_0 = 60\%$, auquel cas T_n suit à peu près une loi normale standard. (T_n est "plus petite" qu'une loi normale lorsque $p_0 < 60\%$.)

Comportement sur les données. De ce qui précède, on peut donc affirmer que les valeurs de T_{100} 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_{100} est en-dessous de ce seuil, et donc, seul 5 % du temps, T_{100} est-elle au-dessus. Ainsi, ce seuil doit être pris égal au quantile¹ noté $z_{95\%}$, qui vaut 1.645 comme on peut le lire sur la table de la loi normale située à la fin de ce polycopié de cours. Cela correspond à un niveau d'erreur (risque) de 5 % sous H_0 . L'intervalle $]1.645, +\infty[$ 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_{100} :

$$\frac{\sqrt{100}}{\sqrt{0.6(1-0.6)}} (\bar{x}_{100} - 0.6) = \frac{\sqrt{100}}{\sqrt{0.6(1-0.6)}} (0.64 - 0.6) = 0.82 .$$

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 pouvoir certifier un taux d'efficacité supérieur à 60% : la prudence commande de penser que l'efficacité est inférieure à 60%, malgré le taux d'échantillon obtenu de 64%. On pourra se reporter à la figure 3.1 pour une illustration graphique de ce raisonnement.

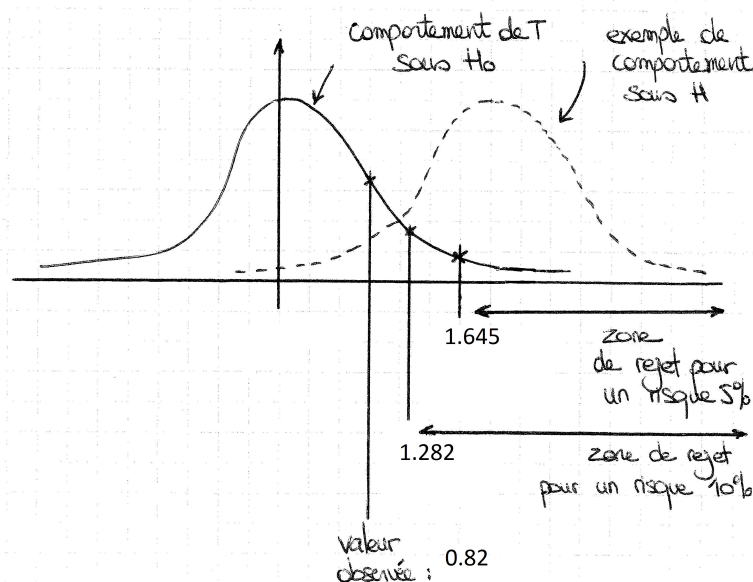


FIGURE 3.1 – Zones de rejet aux risques 5 % et 10 % et valeur réalisée de la statistique de test T_{100} pour les hypothèses $H_0 : p_0 = 60\%$ (cas limite) contre $H_1 : p_0 > 60\%$.

Puisque $z_{90\%} = 1.282$, 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 définira 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. Certes, le laboratoire ne peut communiquer honnêtement sur un taux

1. Pour des rappels sur la notion de quantiles et les notations afférentes, voir les annexes de ce chapitre.

garanti de 60%.

- On peut (paresseusement) proposer que le service R&D améliore le produit avant de refaire une étude clinique dont on espère qu'elle conclura à de meilleurs résultats.
- On peut aussi refaire une étude sur un échantillon de taille plus grande (exercice : montrez par le calcul qu'un taux de 64% de réussite sur un échantillon de 410 individus aurait, lui, permis de rejeter H_0).
- On peut enfin communiquer sur un taux garanti plus faible, par exemple 55%, histoire d'avoir un taux strictement supérieur à 50% au titre de l'impact psychologique. (Exercice : montrez ce fait par le calcul.)

A retenir : la conservation de H_0 ne prouve rien. Si l'on effectuait le test de $H'_0 : p_0 \geq 65\%$ contre $H'_1 : p_0 < 65\%$, on conserverait là encore H'_0 . (Exercice : écrire les détails du raisonnement.) Pourtant, cette hypothèse $H'_0 : p_0 \geq 65\%$ et l'hypothèse précédente $H_0 : p_0 \leq 60\%$ sont parfaitement contradictoires du point de vue de la logique.

La subtilité est que nous n'avons pas validé ni prouvé ces hypothèses, nous avons juste souligné que les données ne contredisaient gravement aucune d'entre elles. Bien sûr, l'une des deux est fausse, mais nous ne savons pas (encore) laquelle. On retrouve là encore, comme l'indiquait une remarque précédente, le fait qu'on conserve les hypothèses de départ soit faute de données suffisantes (c'est le cas ici), soit parce qu'elles sont vraies.

Conserver H_0 ne veut donc en général rien dire de plus que le fait que l'on se trouve dans une zone grise d'incertitude et que l'on ne peut rien conclure fermement. En revanche, rejeter H_0 , c'est établir que les données contredisent gravement l'hypothèse de départ, et là, on a fait un progrès (négatif). Il y a dissymétrie entre le traitement des deux hypothèses.

Remarque 3.3 (A propos de la forme de l'intervalle de rejet). On a mené ici un test dit unilatère : on ne rejetait H_0 qu'au-dessus d'une certaine valeur. On verra dans la suite que lorsque H_0 et H_1 sont respectivement de la forme $p_0 = p_{ref}$ et $p_0 \neq p_{ref}$, 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.

2. Méthodologie des tests : théorie générale

Nous tirons maintenant des principes généraux des trois exemples précédents.

Le cadre est le cadre désormais habituel, celui de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi inconnue $\mathbb{P}_{\theta_0}$, d'espérance et de variance (inconnues) respectivement notées μ_0 et σ_0^2 . On a construit cette modélisation au vu des données $x_1, \dots, x_n$.

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 le chapitre 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 l'ensemble des lois possibles, que l'on note Θ_0 et Θ_1 , et l'on formule les hypothèses

$$H_0 : \theta_0 \in \Theta_0 \quad \text{et} \quad H_1 : \theta_0 \in \Theta_1 .$$

Exemple 3.1. 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 $H_0 : \mu_0 \geq 0$; ou encore, “la variance de la loi est plus petite que 10”, soit $H_0 : \sigma_0^2 \leq 10$.

Remarque 3.4 (Disjonction mais pas partition). H_0 et H_1 doivent être disjointes, mais ne sont pas nécessairement la négation l'une de l'autre. Le contexte, les connaissances *a priori*, etc., permettent de déterminer l'ensemble des cas intéressants² à considérer. Par exemple, dans le cadre de l'application

2. 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 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. »

d'une promotion, on pourra se demander si H_0 : le taux de commande reste le même ou si H_1 : il augmente. On n'envisagera vraisemblablement pas le cas où la promotion ferait baisser les commandes !

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 3.1 (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 $X_1, \dots, X_n$. Les statistiques de test intéressantes sont celles dont on connaît la loi sous H_0 .*

Remarque 3.5 (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 du patch anti-tabac, la statistique de test dépendait de la valeur à tester, qui était égale à $p_{\text{ref}} = 60\% = 0.6$. 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 des chapitres 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 :

$$]r, +\infty[, \quad]-\infty, r[, \quad \text{ou} \quad]-\infty, r[\cup]r', +\infty[.$$

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 3.2 (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*

$$\alpha = \sup_{\theta_0 \in \Theta_0} \mathbb{P} \left\{ T_n \in R \mid X_1, \dots, X_n \text{ i.i.d. selon } \mathbb{P}_{\theta_0} \right\},$$

où le conditionnement rappelle simplement que les variables aléatoires $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi $\mathbb{P}_{\theta_0}$, avec $\theta_0 \in \Theta_0$.

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 responsabilité³ 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 3.2. 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 : il suffit de deviner quel est le cas limite. Ainsi, pour le patch anti-tabac, dans $H_0 : p_0 \leq 60\%$ contre $H_1 : p_0 > 60\%$, le cas limite sous H_0 , et donc le seul cas à considérer, était $H_0 : p_0 = 60\%$.

Remarque 3.6 (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 3.3 (P-valeur). *Etant donné un test statistique et des données, la P-valeur est l'erreur maximale q 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 3.3. Calculons la P-valeur associée au (premier) test et aux données du paragraphe correspondant au patch anti-tabac. On y a vu que les tests d'erreurs de première espèce $\alpha = 5\%$ et $\alpha = 10\%$ conservaient encore l'hypothèse $H_0 : p_0 \leq 60\%$ au vu des données. Ainsi, on sait que la P-valeur est supérieure à 10 %. En fait, le test qui aurait comme seuil de rejet r la valeur observée pour la statistique 0.82 conserverait encore H_0 , mais toute valeur inférieure ou égale conduirait au rejet de H_0 . La P-valeur q est donc donnée par l'équation

$$z_{1-q} = 0.82 \quad \text{soit} \quad q = 20.6\%$$

en utilisant les tables de la loi normale. Dit autrement, q est la masse de probabilité associée à cette zone de rejet maximale.

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

3. 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.

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 q .
- 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 q est bien plus petit que ce seuil, rejetez H_0 fermement ;
 - + Si q est bien plus grand que ce seuil, campez fermement sur H_0 ;
 - + Si q 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 3.2 – Grille méthodologique d'utilisation stratégique de la P-valeur.

que $q = 4.9\%$ 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 3.2.

Remarque 3.7 (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 $\alpha \leq 0.1\%$. 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 q , il faut se représenter le test d'erreur de première espèce $\alpha = q$ correspondant : ce test accepte H_0 , par définition de la P-valeur. Si $\alpha = q$ 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 $\alpha = q$ 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 q forme en ce sens un degré de crédibilité de H_0 face à H_1 . On rejette d'autant plus fermement H_0 que q est petite et on tend d'autant plus à conserver H_0 que q est grande.

Remarque 3.8 (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 la P-valeur comme un indice de crédibilité de H_0 : la P-valeur est 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 ⁴ 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),

$$\beta = \sup_{\theta_1 \in \Theta_1} \mathbb{P} \left\{ T_n \notin R \mid X_1, \dots, X_n \text{ i.i.d. selon } \mathbb{P}_{\theta_1} \right\}.$$

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

4. Ouf ?

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...
- la négation de ce que l'on voudrait prouver, et H_1 est alors ce que l'on veut prouver. On exploite là le fait que les tests ayant tendance à conserver H_0 , on n'apprend vraiment quelque chose que lorsque ils rejettent H_0 . Ils procèdent par progrès négatifs.

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 .

Principe à retenir. Un test dit : « ce n'est pas impossible⁵ » lorsque H_0 est conservée ou « c'est impossible » lorsque H_0 est rejetée. L'hypothèse H_0 est vraie jusqu'à preuve du contraire, et cette preuve du contraire peut être apportée par un test (et des données).

Exemple 3.4 (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 !

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

5. Dans ce cas, la conclusion du statisticien se doit d'être prudente. Elle utilisera souvent une double négation et sera de la forme : « On ne peut pas dire que [H_0 ne soit pas vraie] ».

échantillons aient une taille n suffisamment grande et que les répartitions entre groupe placebo et groupe traité soient faites au hasard.

2.7. Résumés de la méthodologie

Comme à mon habitude, je sou mets une citation⁶ 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 et continuerons à le voir, 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.

Retrouvez cette dualité dans la coupure de presse reproduite aux pages suivantes, qui est l'un des (nombreux) articles revenant sur la controverse autour des études de nocivité des OGM proposées par Gilles-Eric Séralini en septembre 2012.

6. Variante personnelle pour être sûr de plaire à tout le monde et promouvoir l'égalité femmes-hommes : « La statistique est un *shorty* moulant. Ce qu'elle révèle est suggestif, ce qu'elle cache est vital. »

Étude de Séralini sur les OGM : pourquoi sa méthodologie est statistiquement bonne

Modifié le 09-10-2012 à 11h00 21 réactions | 5085 lu

 Temps de lecture : 3 minutes



Par **Paul Deheuvels**
Membre de l'Académie des Sciences

Source :
Le Plus / Le Nouvel Observateur

LE PLUS. L'étude choc sur la nocivité des OGM, à l'initiative du professeur Séralini et de son équipe, suscite des réactions houleuses : les détracteurs sont en effet nombreux à remettre en cause la validité méthodologique de ces recherches. Paul Deheuvels, membre de l'Académie des Sciences et statisticien renommé, nous explique pourquoi l'étude de Séralini est fiable.

Édité par Gaëlle-Marie Zimmermann Auteur parrainé par Guillaume Malaurie

Recommander

151

Envoyer

Tweeter

47

4

Share

0

RÉAGIR



Maïs et OGM (CLOSON DENIS/SOPIX/SIPA)

L'objet de cette tribune est d'apporter mon soutien le plus ferme à G.E Séralini, professeur à l'Université de Caen. J'ai pris connaissance du fait que le professeur Séralini avait réalisé des études comparatives ayant eu, récemment, un considérable retentissement.

À ma connaissance, ces enquêtes ont été conçues et analysées en conformité aux meilleures normes professionnelles en vigueur, et devraient être appréciées comme telles.

La durée de l'expérience : un gage de fiabilité

L'un des points les plus intéressants et innovants de ces études est qu'elles ont été menées sur des durées inhabituellement longues. C'est un fait notoire, connu des expérimentateurs, que les analyses du même genre réalisées au cours de périodes plus courtes ne sont pas aptes à détecter des effets toxiques à long terme.

Le cas particulier du développement de tumeurs relève de cette catégorie, et il n'est donc pas surprenant que les études existantes, ayant été développées sur des durées plus courtes que celle des travaux de G.E. Séralini, n'aient pas révélé, jusqu'ici, de différences significatives entre les effets des traitements.

Je considère donc que le professeur Seralini a fourni des indices très solides, propres à établir l'existence de certains effets toxiques inattendus de la part de produits considérés auparavant comme sûrs.

Bien sûr, il se trouvera toujours des personnes pour contester ces résultats, du fait même qu'ils sont susceptibles d'avoir d'importantes conséquences industrielles.

Leur point de vue ne pourra être soutenu qu'à la lumière de résultats d'expérience nouveaux, puisqu'il n'y a pas de doute sur le sérieux et la fiabilité des conclusions actuelles du Professeur Seralini.

Des reproches statistiques infondés : explications concrètes

On reproche, par exemple, à G.E. Seralini d'avoir réalisé son étude avec des blocs de 10 animaux cobayes, alors que l'usage le plus courant est d'en utiliser 50.

Cette critique est infondée. En effet, dans une étude statistique, on peut chercher soit à démontrer l'innocuité d'un produit, soit à mettre en évidence sa nocivité.

Prenons l'exemple d'un produit qu'on soupçonne être un poison. Si on l'administre à 5 patients et s'il y a 5 morts, il n'y a pas besoin d'aller plus loin pour décider qu'il est toxique. Si, par contre, les 5 patients survivent, cela ne veut pas dire pour autant que le produit soit inoffensif, car les 5 patients ne sont pas nécessairement représentatifs de toute la population.

C'est pour cette raison que les études commanditées par les entreprises qui souhaitent qu'on admette l'innocuité de leurs produits imposent que les échantillons utilisés soient d'une importance minimale.

Dans le cas présent, il est naturel d'imposer un minimum de 50 cobayes (ici, des rats) pour les analyses comparatives. C'est tout à fait différent si on recherche des effets toxiques. Dans ce cas, le fait de déceler des différences significatives sur des petits groupes (10 dans le cas de l'étude de G.E.Seralini) renforce la conclusion, au lieu de l'affaiblir.

Deux types d'erreurs dans les études statistiques

Je conclus en insistant sur le principe suivant. Dans une étude statistique, si on décèle des effets toxiques, il y a toujours un risque de se tromper, mais l'erreur, s'il y a, peut facilement être dissipée par des analyses ultérieures. En revanche, si des études ne décèlent pas d'effets, on commet généralement une faute en en déduisant que ces effets n'existent pas. Car il y a deux types d'erreurs:

1. Déceler une toxicité alors que celle-ci n'existe pas;
2. Admettre l'innocuité alors que la toxicité existe.

La plupart du temps, on contrôle le risque (1) (le standard est de fixer les règles de décision de sorte que sa probabilité soit fixée à 5%). En revanche, le risque (2) dépend de la taille de l'étude, et est souvent très important (40% voire 60% ou plus).

C'est pour limiter le risque (2) qu'on cherche à imposer des études faisant intervenir beaucoup d'expériences individuelles (ou de cobayes). Par contre, lorsqu'on décèle la toxicité, il n'y a que 5% de probabilité d'erreur, quel que soit le nombre de cobayes utilisés.

3. Rappels sur les quantiles d'une loi

Le minimum à savoir consiste en ce qui suit : pour $\beta \in]0, 1[$, on appelle z_β le quantile d'ordre β , défini de manière unique par

$$\mathbb{P}\{Z \leq z_\beta\} = \beta \quad \text{où} \quad Z \sim \mathcal{N}(0, 1).$$

On lit la valeur de ces quantiles dans la table de la loi normale nommée «Loi normale : quantiles» à la fin de ce polycopié.

Ne lisez ce qui suit que si vous voulez en savoir un peu plus sur le sujet.

3.1. Définition théorique

Nous n'aurons presque toujours qu'à déterminer les quantiles de lois admettant une densité f strictement positive sur $\mathbb{R}$ ou $\mathbb{R}_+$: loi normale, loi de Student, loi du χ^2 . Les définitions suivantes ne considèrent donc que ce cadre. On énoncera simplement une remarque faisant allusion à la définition des quantiles empiriques (sur des données).

Définition 3.4 (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 $x \in \mathbb{R}$,*

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

Définition–Théorème 3.1 (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 $\mathbb{R} \rightarrow]0, 1[$. 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,

$$\mathbb{P}\{X \leq F^{-1}(\beta)\} = \beta \quad \text{et} \quad \mathbb{P}\{X \geq F^{-1}(\beta)\} = 1 - \beta.$$

On illustre cela à la figure 3.3. En particulier, la médiane est le quantile à 50 %.

Remarque 3.9. 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 $\mathbb{R}_+ \rightarrow [0, 1[$; 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

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

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

3.2. 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 .

Définition et calculs de quantiles.

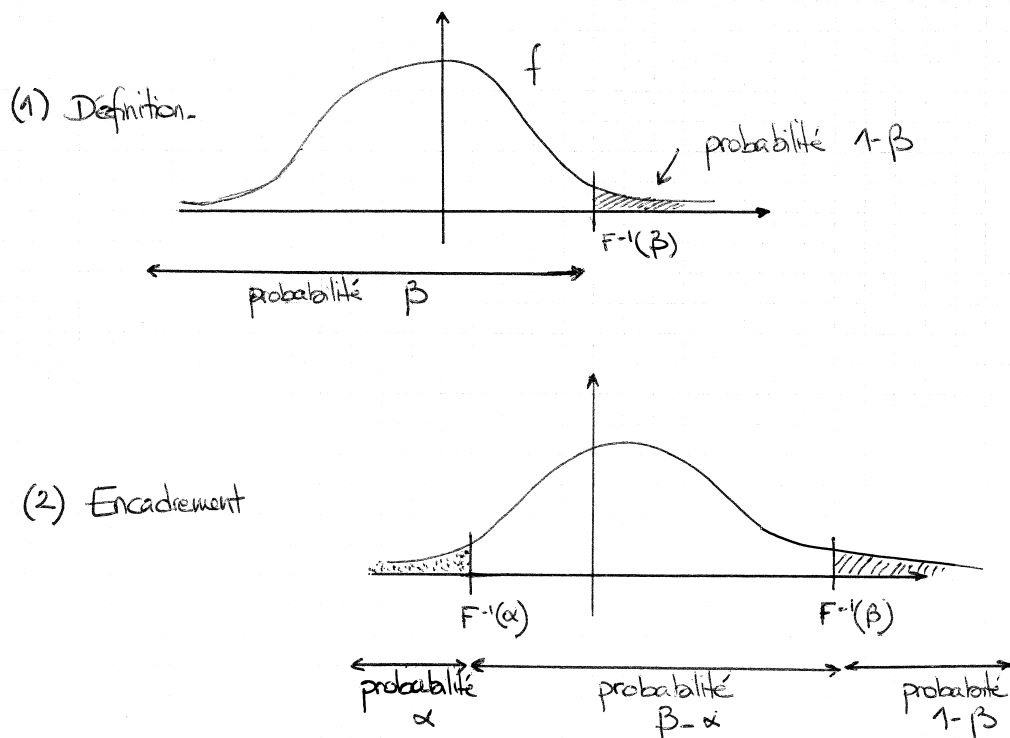


FIGURE 3.3 – Illustration graphique de la définition des quantiles, pour une loi chargeant tout $\mathbb{R}$: la loi normale.

- La loi dite de Student $\mathcal{T}_k$ 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_k^2 , selon son nombre de degrés de liberté k : on note ses quantiles $\chi_{k,\beta}^2$ ou $c_{k,\beta}$. (On verra que la loi du χ^2 ne charge que $\mathbb{R}_+$.)

Remarque 3.10 (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 ,

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

comme on peut s'en persuader sur un dessin.

3.3. 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_k^2 ;
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, cela viendra au cours des exercices.

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 $x_1, \dots, x_n$

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.

Tests de comparaison à une valeur de référence

Objectifs. Nous allons décliner les principes dégagés au chapitre précédent afin de déterminer à quel point une valeur de référence proposée, μ_{ref} ou p_{ref} , est compatible ou pas avec les données recueillies, qui estiment le paramètre de population μ_0 ou p_0 .

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. Ce sera le seul point à retenir même si, pour la culture, nous fournissons des explications (non exigibles) du fonctionnement de ce principe, afin qu'il ne soit pas qu'une vague recette. On note

$$\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

l'estimateur empirique (débiaisé) de la variance associé aux données $X_1, \dots, X_n$.

Principe 4.1. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*

Données : $x_1, \dots, x_n \in \mathbb{R}$ où $n \geq 30$

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 : $\mu_0 = \mu_{\text{ref}}$

Statistique de test :

$$T_n = \sqrt{n} \frac{\bar{X}_n - \mu_{\text{ref}}}{\sqrt{\hat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \rightarrow \mathcal{N}(0, 1)$

Comportement sous H_1 : lorsque $\mu_0 > \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_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.

1.1. Preuve de ce principe (lecture facultative ; rappels de classes préparatoires)

On se place dans un modèle où la vraie loi sous-jacente admet un moment d'ordre deux (c'est presque toujours le cas en pratique).

Rappels sur la variance empirique. Un estimateur sans biais de la variance σ_0^2 est défini par

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

Cet estimateur est également convergent :

$$\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma_0^2.$$

Lemmes. La stabilité de la convergence en probabilité par application de fonctions continues est au programme des filières EC/E et EC/S. En revanche, seuls les étudiant-e-s issu-e-s des classes EC/S ont vu le lemme de Slutsky.

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

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

En particulier,

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

Lemme 4.2 (de Slutsky). *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 :*

$$Y_n \rightharpoonup Y \quad \text{et} \quad Z_n \xrightarrow{\mathbb{P}} z;$$

alors on a les convergences en loi

$$Y_n + Z_n \rightharpoonup Y + z \quad \text{et} \quad Y_n Z_n \rightharpoonup zY.$$

Ingrédient final : théorème limite central. Il assure que :

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

Le lemme de Slutsky et la convergence de l'estimateur de la variance (associée à la stabilité de la convergence en probabilité) montrent que cette convergence est encore vraie lorsque la variance (inconnue) est estimée :

$$Z_n = \sqrt{\frac{\sigma_0^2}{\hat{\sigma}_n^2}} \xrightarrow{\mathbb{P}} 1$$

puis

$$\sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0) \rightharpoonup \mathcal{N}(0, 1).$$

Conclusion 1 : comportement de T_n sous H_0 . Puisque par définition $\mu_0 = \mu_{\text{ref}}$ sous H_0 , le résultat précédent explique la première partie du principe de test.

Conclusion 2 : comportement de T_n sous H_1 . Il découle du fait que $\bar{X}_n$ étant un estimateur convergent de μ_0 , il s'en rapproche (et de même $\hat{\sigma}_n^2$ 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 $\sqrt{n}(\mu_0 - \mu_{\text{ref}})/\sigma_0$.

1.2. Application de ce principe (lecture obligatoire)

Nous considérons l'exercice suivant, rédigé en 2008, portant sur le salaire moyen des fonctionnaires.

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 ?

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 2193 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 2245 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, vu 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.

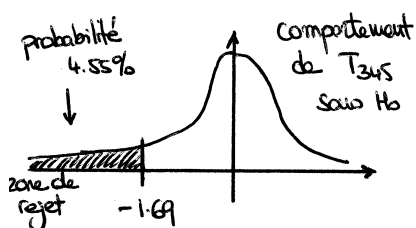
On veut savoir si ce salaire moyen de 2008 peut être dit significativement différent du salaire de 2004, à savoir $\mu_{\text{ref}} = 2245$ (exprimé en tenant compte de l'inflation). Ici, le choix des hypothèses est très politique : les syndicats vont prendre $H_0: \mu_0 = \mu_{\text{ref}}$ et espèrent montrer que H_0 doit être rejetée avec force. On la teste contre $H_1: \mu_0 < \mu_{\text{ref}}$ (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 $T_{345} = \sqrt{345} \frac{\bar{X}_{345} - 2245}{\sqrt{\hat{\sigma}_{345}^2}}$ qui suit approximativement, sous H_0 , la loi $\mathcal{U}(0,1)$.

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 $]-\infty, -c]$. Or, sur les données, on observe

$$\text{la réalisation } \sqrt{345} \left(\frac{\bar{x}_{345} - 2245}{s_{x,345}} \right) = \sqrt{345} \left(\frac{2193 - 2245}{573} \right) = -1.69$$

soit une P-valeur de

$$P\{N < -1.69\} = 4.55\% \quad (\text{on consulte les tables}).$$



Conclusion Statistique: ce serait le rejet (sur la fil, ou 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 doutes qu'on rejette l'hypothèse de pouvoir d'achat constant, et que ce serait sans doute encore pire avec plus de sondés ! Les chiffres

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 ?

compteront donc, et ils les diffuseront dans la presse !

→ 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. C'est, 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 un 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.

2. Cas d'une proportion, 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 $p_{\text{ref}}(1 - p_{\text{ref}})$.

En pratique, on se ramène toujours à une hypothèse de départ de la forme $H_0 : p_0 = p_{\text{ref}}$ par identification d'un cas limite.

Principe 4.2. *Test de comparaison d'une proportion de population p_0 à une valeur de référence p_{ref}*

Données : $x_1, \dots, x_n \in \{0, 1\}$ où $n \geq 30$

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 :

$$T_n = \sqrt{n} \frac{\bar{X}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}}$$

Comportement sous H_0 : $T_n \rightarrow \mathcal{N}(0, 1)$

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 .

2.1. Preuve de ce principe

La preuve découle directement du théorème limite central (pour le comportement sous H_0 ; aucun besoin de recourir ici au lemme de Slutsky) et de la loi des grands nombres (pour le comportement sous H_1).

2.2. Application de ce principe

Nous considérons l'exercice suivant, rédigé en 2013, portant sur des retards de trains.

On s'intéresse ici aux retards des trains, à l'instar de l'un de vos enseignants discourant sur France 5 dans l'émission *La quotidienne* du mercredi 23 octobre 2013; et en particulier, de ceux des RER, et encore plus précisément, au cas de notre RER C chéri. La SNCF annonce que plus de 90 % de ses trains de banlieue sont à l'heure, en moyenne sur l'année écoulée (toutes lignes et tous créneaux horaires confondus). Un usager l'utilisant quotidiennement veut tester ce fait sur son trajet à lui (Paris Saint-Michel-Notre-Dame / Jouy-en-Josas, départ à 6h33, arrivée sans retard à 7h39).

- Imaginez un état d'esprit pour ce voyageur du quotidien et déduisez-en le jeu d'hypothèses à tester. (*Avec des mots; plusieurs réponses sont possibles.*)

Il collecte donc des données pendant trois mois, ce qui nous fait 62 jours ouvrés : son train est arrivé en retard 11 fois et à l'heure 51 fois.

- Traitez ces données en effectuant un test d'hypothèses; on commencera par proposer une modélisation.

3. Retards de train (5 points)

On s'intéresse ici aux retards des trains, à l'instar de l'un de vos enseignants discourant sur France 5 dans l'émission *La quotidienne* du mercredi 23 octobre 2013; et en particulier, de ceux des RER, et encore plus précisément, au cas de notre RER C chéri.



La SNCF annonce que plus de 90 % de ses trains de banlieue sont à l'heure, en moyenne sur l'année écoulée (toutes lignes et tous créneaux horaires confondus). Un usager l'utilisant quotidiennement veut tester ce fait sur son trajet à lui (Paris Saint-Michel-Notre-Dame / Jouy-en-Josas, départ à 6h33, arrivée sans retard à 7h39).

- Imaginez un état d'esprit pour ce voyageur du quotidien et déduisez-en le jeu d'hypothèses à tester. (*Plusieurs réponses sont possibles.*)

Trois réponses étaient possibles (voir corrigé); en voici une...
Cela fait des années que ce voyageur emprunte cet itinéraire à cette heure - et tout va bien, merci, il est à l'heure pour faire cours. Il a un préjugé plutôt bienveillant et ne voudrait renier son choix de vie que si les données sont très convaincantes:
 H_0 : Pas plus de retards sur cet itinéraire à cette heure que sur l'ensemble des trains à l'ensemble des heures
Il collecte donc des données pendant trois mois, ce qui nous fait 62 jours ouvrés : son train est arrivé en retard 11 fois et à l'heure 51 fois. Nous allons traiter ces données, en commençant par une modélisation beaucoup plus brève qu'à l'ordinaire (afin de gagner du temps, dans ce quizz qui est long). vs. H_1 : Davantage de retards.

- Indiquez et interprétez le paramètre d'intérêt :

On s'intéresse au taux p_0 de trains sur cet itinéraire et à cette heure qui sont en retard (taux moyen sur l'année); et on veut le comparer au taux $p_{ref} = 10\%$ de trains en retard sur l'ensemble du réseau et à l'ensemble des heures, soit:

$H_0: p_0 \leq p_{ref}$ vs. $H_1: p_0 > p_{ref}$, avec comme cas limite: $H_0: p_0 = p_{ref}$

On peut, d'ailleurs, le modéliser comme la réalisation de variables aléatoires $X_1, \dots, X_{62}$ iid $\sim \text{Ber}(p_0)$ car l'état du trafic un jour donné ne dépend pas (ou très peu) de l'état des autres jours.

Quizz 2 - Tests d'hypothèses - 2013

- Indiquez les données disponibles, en précisant le codage éventuel et les statistiques d'échantillon :

On dispose des relevés $x_1, \dots, x_{62} \in \{0, 1\}$, où $x_j = 1$ si le train du j-ème jour de l'étude était en retard, $x_j = 0$ sinon.

Estimée de p_0 : $\bar{x}_{62} = \frac{11}{62} \approx 17.7\%$

Cette valeur est-elle significativement supérieure à 10%?

- Précisez la statistique de test (formule théorique) et son comportement sous les différentes hypothèses retenues :

$$T_{62} = \sqrt{62} \left(\frac{\bar{x}_{62} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}} \right)$$

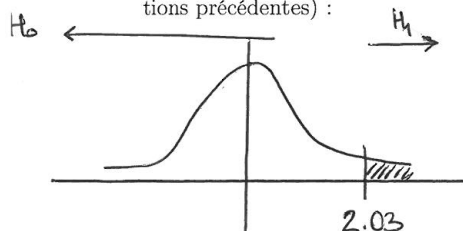
Sous H_0 , $T_{62} \stackrel{(d)}{=} \mathcal{N}(0, 1)$

Sous H_1 , T_{62} tend à prendre des valeurs plus grandes.

- Calculez sa valeur réalisée (sa valeur numérique) :

$$= \sqrt{62} \left(\frac{\bar{x}_{62} - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}} \right) = \sqrt{62} \left(\frac{0.177 - 0.1}{\sqrt{0.1(1-0.1)}} \right) = 2.03$$

- Calculez la P-valeur (par exemple en vous aidant d'un schéma récapitulant les réponses aux deux questions précédentes) :



$$\begin{aligned} P\text{-valeur} &= P(\mathcal{N}(0, 1) > 2.03) \\ &= 1 - 0.9788 \\ &\approx 2.1\% \end{aligned}$$

cf. tables de loi

- Conclusion statistique (dont une phrase non technique compréhensible par tous) :

P-valeur < 5% (assez nettement) $\rightarrow$ On rejette $H_0 \rightarrow$ Les données montrent que ce bout de trajet connaît un taux de retards supérieur à la moyenne de 10%.

- Conclusion stratégique (que va faire le voyageur, maintenant qu'il est éclairé par cette étude statistique ?) :

Les 10% de retards moyens sont calculés sur l'ensemble des trajets et des heures : avec un peu de chance, le taux de retards étudié ici sur le trajet St Michel - Joux n'est dû qu'à l'heure. Le voyageur pourrait étudier ce qu'il est de son trajet favori un peu plus tôt encore ou plus tard. Si la même mauvaise performance en ressort, peut-être y a-t-il un problème (p. ex. technique) récurrent sur la ligne ? Il créera une association d'usagers pour tirer cela au clair.

Commentaires sur l'exercice 3.

Autres jeux d'hypothèses possibles :

Règle : H_0 est le préjugé, dont on ne se départit que si les données le contredisent gravement.

* Etat d'esprit neutre (le voyageur débarque de province et n'a aucun préjugé) :

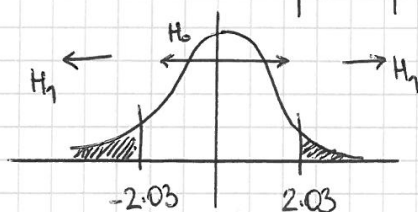
$H_0 : p_0 = p_{ref}$
 $\hookrightarrow$ Même taux de retards qu'au global

vs. $H_1 : p_0 \neq p_{ref}$
 $\hookrightarrow$ Taux différent, à la hausse ou à la baisse

Comportement de T_{02}

• sous H_0 : $\stackrel{(d)}{=} d(0,1)$

• sous H_1 : valeurs plus grandes ou plus petits $\rightarrow$ zone de rejet bilatère $]-\infty, -r[\cup]r, +\infty[$



$$P\text{-valeur} = 2 \times 2.1\% = 4.2\%$$

$< 5\%$ $\hookrightarrow$ on rejette H_0 , et, on regardant les données, on voit que la différence se fait à la hausse. Même conclusion stratégique.

* Etat d'esprit énévée / pessimiste (le voyageur a pour préjugé que rien ne va jamais pour lui) :

$H_0 : p_0 \geq p_{ref}$
 $\hookrightarrow$ Plus de retards qu'au global

vs. $H_1 : p_0 < p_{ref}$
 $\hookrightarrow$ Moins de retards

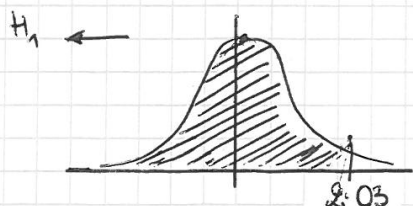
Cas limite :

$H_0 : p_0 = p_{ref}$ vs. $H_1 : p_0 < p_{ref}$

Comportement de T_{02} :

• sous H_0 : $\stackrel{(d)}{=} d(0,1)$

• sous H_1 : valeurs plus petits $\rightarrow$ zone de rejet $]-\infty, r[$



$$P\text{-valeur} = 97.9\% \gg 5\%$$

$\rightarrow$ Le voyageur conserve son hyp. H_0 ,

Il n'a aucune raison de s'en départir (mais cela ne prouve pas a priori que H_0 soit vraie : elle n'est pas remise en cause par les données, simplement, lorsque l'on pose le test ainsi).

Conclusion Stratégique : Il conserve son préjugé et ne change donc rien à sa routine, si tant est qu'elle le satisfait.

Erreurs et manquements souvent constatés :

- * L'erreur la plus fréquente a été de penser que l'étude portait sur l'ensemble des trains de banlieue (ou, tant qu'on y est, sur l'ensemble des trains de France ou du monde, pour grossir encore le trait !) : non, on n'étudie que ce qui se passe sur un trajet donné et à une heure donnée. On ne cherche donc pas à confirmer ou infirmer le 10% de retards au global, on cherche à voir comme notre taux de retard "local" se compare à ce taux global de référence, $p_{ref} = 10\%$.
- * La justification de hypothèses était parfois faible.
- * Il y a souvent eu des problèmes de cohérence : si on note 1 les retards et 0 les trains à l'heure, alors p_0 , p_{ref} et $\bar{x}_{G2}$ sont tous trois des taux de retard. On peut bien sûr coder 1 pour un train à l'heure, 0 pour un retard, auquel cas p_0 , p_{ref} et $\bar{x}_{G2}$ sont des taux de trains à l'heure ; et T_{G2} admet pour valeur réalisée -2.03 , etc. Bien sûr, cela ne change pas les P-valeurs en fonction de préjugés !
 ↳ Or il y a souvent eu des mélanges (ex: p_0 taux de retard vs. $\bar{x}_{G2}$ taux de trains à l'heure)

Pot-pourri et florilège de états d'esprit de voyageurs.

Merci à tous ceux qui m'ont fait rire tout en proposant à la suite une solution pertinente !
 Je cite : → Ceux-là ont bien compris ce qui me fait plaisir.

Malik Felag : « La personne présentée sur la vidéo a manifestement l'air mécontente, elle lève les sourcils dans un signe d'agacement. [...] »

Ayméric Capmarty : « Ce voyageur n'a pas l'air content du tout ! Mais comme il est gentil, magnanime et compréhensif envers le service public, il va supposer que les trains n'ont pas trop de retards pour ensuite montrer que c'est faux (peut-être). »

Rémi Thivolle : « Le voyageur qui, fort de son savoir statistique, semble avoir un esprit de contradiction assez développé envers les chiffres officiels de tout genre, adoptera comme hypothèse : [...] »

Vianney de Dreuille : « Ce voyageur est exaspéré par les retards permanents de la SNCF. Mais ce voyageur étant très prudent (prof de stats), il choisit la hypothèse : [...] »

Tristan Chabert : « Imaginons que ce voyageur est un professeur dans une grande école de Jury et qu'il doit donc arriver sans faute pour son cours de Sh. [...] »

De bonnes contributions également d'Olivier Attal et Robert Le Pecl-

À tous : BRAVO et MERCI.

Mon seul regret : Mesdames, pourquoi n'avez-vous pas osé ?

3. Taille d'échantillon faible : cas de données gaussiennes

Les résultats précédents sont asymptotiques et reposent sur le théorème limite central. On n'en croit les résultats en pratique que lorsque la taille d'échantillon n est suffisamment grande¹, disons $n \geq 30$. 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 chapitre 2 ; mais en particulier, cette solution ne vaut pas dans le cas d'un modèle de Bernoulli, où il s'agit d'estimer une proportion ou une fréquence : nous ne traiterons pas dans notre cours du cas du test sur une proportion dans le cas d'une taille d'échantillon faible.

Cette solution repose sur la loi dite de Student², que l'on introduit maintenant.

3.1. Loi de Student

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 $\bar{X}_n \sim \mathcal{N}(\mu_0, \sigma_0^2/n)$ et donc

$$\frac{\bar{X}_n - \mu_0}{\sqrt{\sigma_0^2/n}} = \sqrt{\frac{n}{\sigma_0^2}} (\bar{X}_n - \mu_0) \sim \mathcal{N}(0, 1).$$

(A gauche, il s'agit de la même variable aléatoire que celle considérée au théorème limite central.) 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 $\hat{\sigma}_n^2$ conduit à la loi de Student.

Définition–Théorème 4.1. *Soit $X_1, \dots, X_n$ des variables aléatoires indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2 . Alors la loi de*

$$\frac{\bar{X}_n - \mu_0}{\sqrt{\hat{\sigma}_n^2/n}} = \sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0)$$

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 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)

$$X'_1 = \frac{X_1 - \mu_0}{\sigma_0}, \dots, X'_n = \frac{X_n - \mu_0}{\sigma_0} ;$$

elles sont indépendantes et identiquement distribuées selon la loi $\mathcal{N}(0, 1)$. On note

$$\bar{X}'_n \quad \text{et} \quad \hat{\sigma}'_n{}^2$$

les estimateurs de la moyenne et de la variance empiriques construits sur les X'_j et on note, par un calcul immédiat, que

$$\sqrt{\frac{n}{\hat{\sigma}'_n{}^2}} (\bar{X}_n - \mu_0) = \sqrt{\frac{n}{\hat{\sigma}'_n{}^2}} \bar{X}'_n ,$$

ce qui conclut la preuve. □

1. Et pour les tests sur une proportion, on aime également à rajouter les contraintes $np_{\text{ref}} \geq 5$ et $n(1 - p_{\text{ref}}) \geq 5$, que vous avez vues en classe de terminale et que nous reverrons également dans le chapitre 6 consacré aux tests du χ^2 .

2. 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.

On peut voir (par exemple sur les tables) que, à $\beta \geq 0.5$ fixé, $t_{k,\beta}$ est une suite décroissante avec k , de limite z_β . On retrouve cela sur la figure 4.1, 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 confondue avec la loi $\mathcal{N}(0, 1)$ lorsque $k \geq 30$). On dit que la loi de Student admet des queues plus épaisses que la queue de la loi normale.

Loi de Student

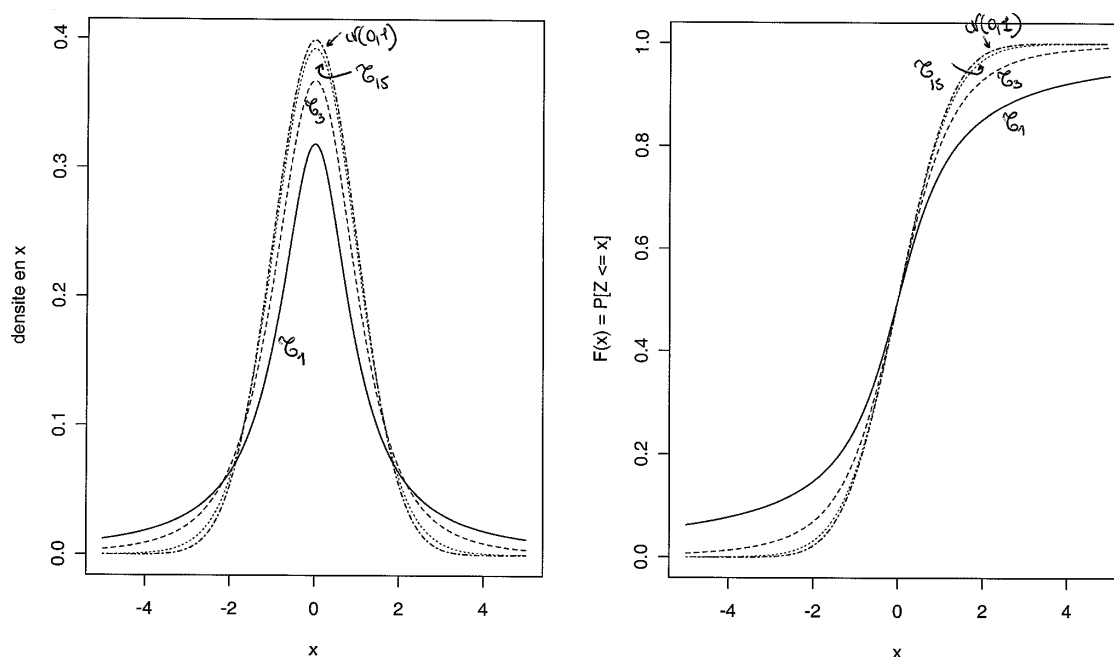


FIGURE 4.1 – Densités (à gauche) et fonctions de répartitions (à droite) des lois de Student, pour différents degrés de liberté, et pour leur loi limite, la loi normale standard $\mathcal{N}(0, 1)$.

3.2. Enoncé du principe de test

Principe 4.3. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*

Données : $x_1, \dots, x_n \in \mathbb{R}$ où $n \leq 30$

Modélisation associée : observations $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2

Hypothèse H_0 : $\mu_0 = \mu_{\text{ref}}$

Statistique de test :

$$T_n = \sqrt{n} \frac{\bar{X}_n - \mu_{\text{ref}}}{\sqrt{\hat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \sim \mathcal{T}_{n-1}$

Comportement sous H_1 : lorsque $\mu_0 > \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0 .

Le test proposé est non asymptotique, 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 gaussiennes... En revanche, la distribution des valeurs observées est souvent raisonnablement proche d'une loi normale, de sorte que le test ci-dessus est d'un comportement théorique raisonnablement proche de celui attendu pour des données exactement gaussiennes.

En fait, la remarque sur la convergence décroissante des quantiles ci-dessus montre qu'une utilisation des lois de Student est plus robuste que le recours à une loi normale standard : les P-valeurs sont rendues plus grandes par le recours à une loi de Student que par celui à la loi normale.

3.3. Application de ce principe (hélas très sexiste)

Exercice 4.1 (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 plaintes³ 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 $x_1, \dots, x_{27}$. La hauteur moyenne relevée est $\bar{x}_{27} = 164.3$ 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 4.1. On a vu dans le chapitre 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 $X_1, \dots, X_{27}$ 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

$$T_{27} = \sqrt{27} \frac{\bar{X}_{27} - 160}{\sqrt{\hat{\sigma}_{27}^2}}$$

est fournie par le principe 4.3 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 $\bar{X}_{27}$ et donc T_{27} auront tendance à prendre des valeurs

3. 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 récemment une campagne de nouvelles mesures.

plus grandes que sous H_0 . On illustre tous les calculs qui suivent par la figure 4.2.

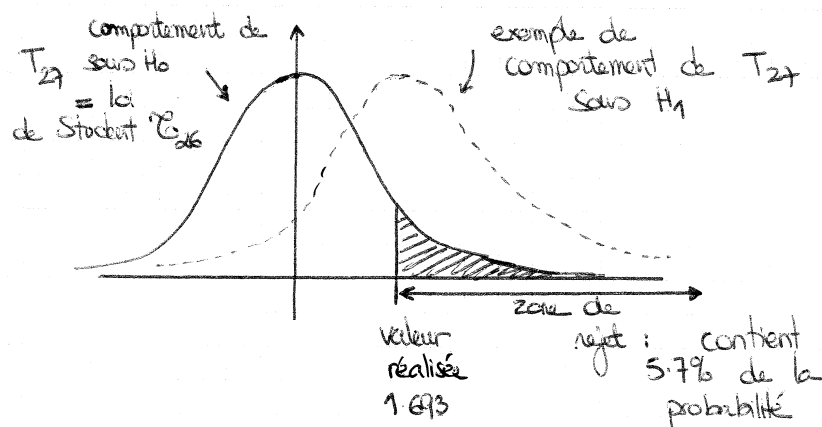


FIGURE 4.2 – Illustration du raisonnement et des calculs associés à l'exercice 4.1.

Si l'on veut faire un test d'erreur de première espèce $\alpha = 5\%$, alors il suffit de prendre le seuil $r = t_{26,95\%} = 1.706$. Le calcul de la valeur réalisée de T_{27} donne

$$\sqrt{27} \frac{\bar{x}_{27} - 160}{\sqrt{s_{27}^2}} = \sqrt{27} \frac{164.3 - 160}{13.2} = 1.693.$$

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

$$\mathbb{P}\{T > 1.693\} = 5.7\% \quad \text{où } T \sim T_{26}.$$

(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 4.1 (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 !

Tests à partir d'échantillons indépendants ou appariés

Résumé. Les deux chapitres précédents vous ont 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 avons illustré la méthodologie des tests en déclinant différentes situations de tests de comparaison d'une valeur moyenne d'échantillon à une valeur de référence.

Objectifs de ce chapitre. Nous allons nous intéresser à la comparaison de deux échantillons, issus de deux populations différentes : les moyennes (de population) peuvent-elles être dites égales, au vu des moyennes d'échantillons observées ?

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. 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 4.3) 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 ($n \geq 30$, voir le principe 4.1), 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 lors¹ que l'on se souvient que lorsque la taille d'échantillon n est grande ($n \geq 30$), les quantiles de Student sont diablement proches des quantiles correspondants de la loi normale. On peut donc appliquer le T-test comme test

1. 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 P-valeurs plus grandes avec des lois de Student : les tests associés sont simplement (un peu) plus précautionneux.

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 4.2), 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 5.1 réalise un T-test bilatère² de comparaison du salaire moyen des infirmières américaines (jeu de données du chapitre 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 5.1 – T-test appliqué aux données de salaire horaire des infirmières américaines : comparaison à la valeur de référence de 20 dollars.

La minute SPSS 5.1. La figure 5.1 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 $x_1, \dots, x_n$ les données disponibles; n désigne donc le nombre de données valides (non manquantes). Le tableau du haut de la figure 5.1 correspond alors à :

N	Moyenne	Ecart-type	Err. std. moy.
n	$\bar{x}_n$	$s_{x,n}$	$s_{x,n}/\sqrt{n}$

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 $1 - \alpha$
$\sqrt{n} \left(\frac{\bar{x}_n - \mu_{\text{ref}}}{s_{x,n}} \right)$	$n - 1$	P-valeur	$\bar{x}_n - \mu_{\text{ref}}$	$\bar{x}_n - \mu_{\text{ref}} \pm t_{n-1, 1-\alpha/2} s_{x,n}/\sqrt{n}$

Exploitation de la figure 5.1 et conclusion. On voit que 2911 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 : ces données ne permettent pas d'établir une différence de salaire entre les deux types d'infirmières.

2. 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.

Remarque 5.1 (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 5.2).

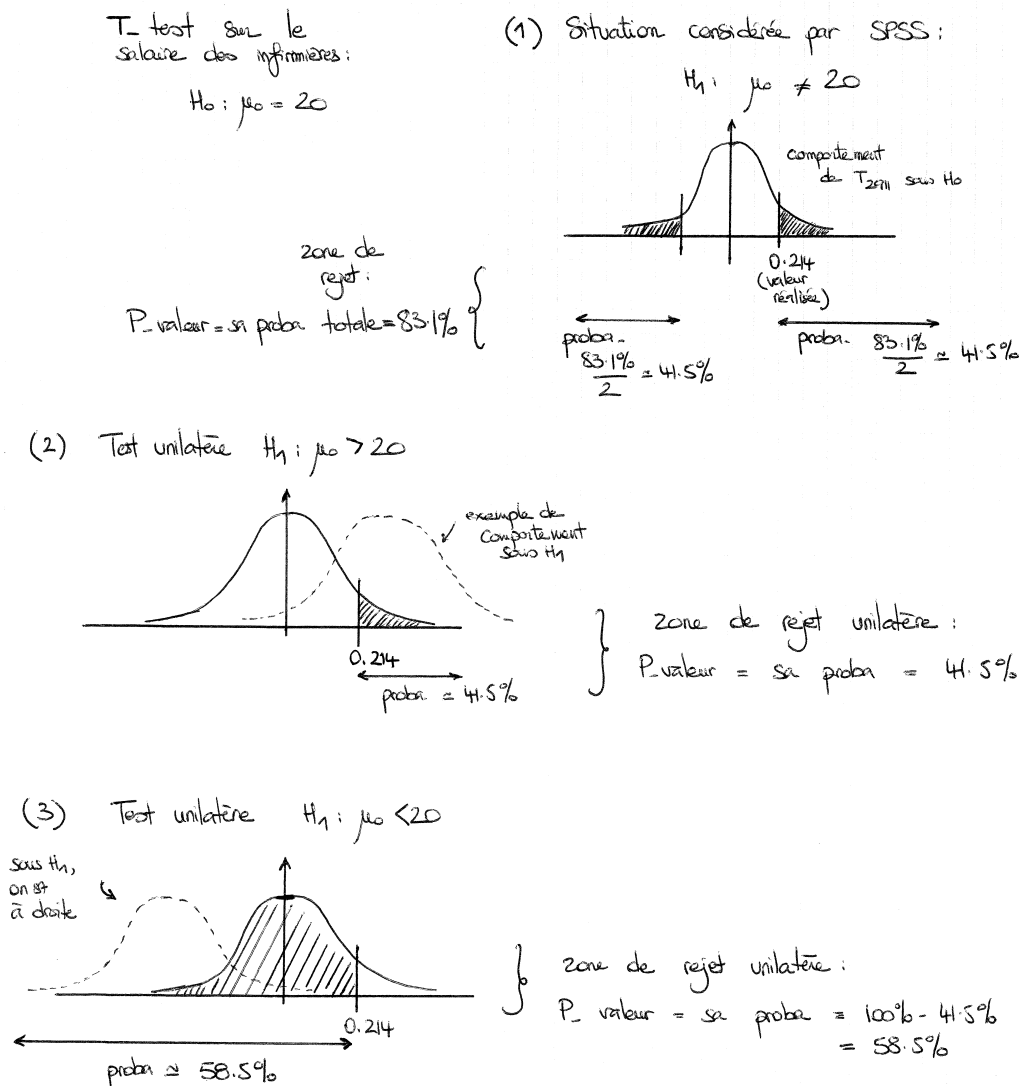


FIGURE 5.2 – Principe de passage de la P-valeur pour un test bilatère, telle que calculée par SPSS, aux P-valeurs pour tests unilatères.

Liens avec les intervalles de confiance. Au risque d'anticiper sur le cours sur les intervalles de confiance, notez qu'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 moins³).

3. 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 %),

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 $x_1, \dots, x_n$ d'une part, et $y_1, \dots, y_m$ 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 $t = 1, \dots, n$. (Dans ce cas, $n = m$ nécessairement.) On parle alors de données appariées.

il serait légitime de proposer l'intervalle, très lisible, $[20 \pm 0.20] = [19.80, 20.20]$.

2. 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 -ème membre de l'échantillon.

Exemple 5.1 (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 5.3 ; 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 avoir une peau sèche ou grasse, et qui seraient apparus si l'on

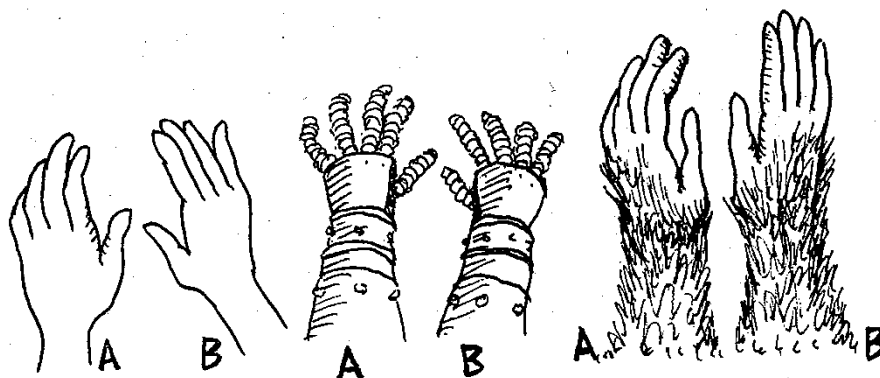


FIGURE 5.3 – Pour choisir entre deux crèmes hydratantes, il vaut mieux recourir à un test sur observations appariées, une crème pour chaque main des sujets.

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 5.2 (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 SPSS 5.2. 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 $Z_1, \dots, Z_n$ 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).

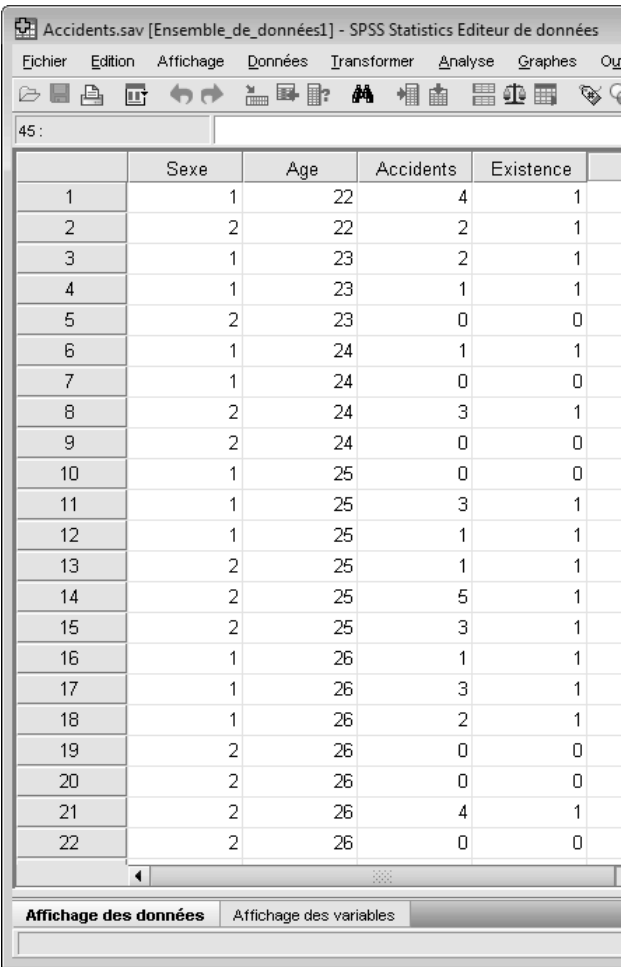
Il suffit alors d'appliquer directement les techniques du cours précédent (principes 4.1 ou 4.3) aux données z_j associées aux Z_j . Pour y trouver le bon principe à appliquer parmi ces deux principes, il faudra regarder la taille n de l'échantillon et ce que l'on peut dire de la loi commune des Z_j (si elle est normale ou pas). Il n'y a pas donc besoin de détailler davantage ce cas-ci.

3. Seconde étude : échantillons indépendants, le cas des proportions

Dans ce cas, on dispose par exemple de données $x_1, \dots, x_n \in \{0, 1\}$, mesurant un certain résultat binaire et recueillies sur un premier échantillon tiré d'une première population, et de données $y_1, \dots, y_m \in \{0, 1\}$, 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 SPSS 5.3. 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 5.3. Considérons les données de la figure 5.4. 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.



	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 5.4 – Début du fichier de données sur les accidents routiers.

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 $x_1, \dots, x_n \in \{0, 1\}$ et $y_1, \dots, y_m \in \{0, 1\}$; ici, les tailles n et m n'ont aucune raison d'être égales.

3.1. Obtention du principe de test (lecture facultative)

On modélise ces données comme les réalisations respectives d'une première série de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_X et celles d'une seconde série $Y_1, \dots, Y_m$ 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 $\bar{x}_n$ et $\bar{y}_m$, i.e., de s'intéresser, du point de vue théorique, à la quantité $\bar{X}_n - \bar{Y}_m$. 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 :

$$\sigma^2(\bar{X}_n - \bar{Y}_m) = \sigma^2(\bar{X}_n) + \sigma^2(\bar{Y}_m) = \frac{1}{n}\sigma^2(X_1) + \frac{1}{m}\sigma^2(Y_1) = \frac{1}{n}p_X(1-p_X) + \frac{1}{m}p_Y(1-p_Y)$$

Sous H_0 , les observations des deux séries 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

$$\hat{p}_{n+m} = \frac{X_1 + \dots + X_n + Y_1 + \dots + Y_m}{n+m} = \frac{n\bar{X}_n + m\bar{Y}_m}{n+m}.$$

On pondère enfin $\bar{X}_n - \bar{Y}_m$ par un estimateur de sa variance pour obtenir la statistique de test :

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{p}_{n+m}(1-\hat{p}_{n+m})(1/n + 1/m)}}.$$

Sous $H_0 : p_X = p_Y$, par une version généralisée du théorème limite central et par application du lemme de Slutsky (détails omis), il vient alors la convergence en loi $T_{n,m} \rightarrow \mathcal{N}(0, 1)$ lorsque $n \rightarrow \infty$ et $m \rightarrow \infty$. Sous H_1 , selon la forme de cette dernière, la différence $\bar{X}_n - \bar{Y}_m$ et donc $T_{n,m}$ elle-même tendent à prendre des valeurs plus grandes et/ou plus petites que 0.

3.2. Enoncé du principe de test

Cela conduit au principe de test suivant. Comme il repose sur une convergence en loi fondée sur le théorème limite central, il nécessite en pratique que les tailles n et m soient suffisamment grandes (toutes deux plus grandes que 30).

Principe 5.1. *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 $y_1, \dots, y_m \in \{0, 1\}$ où $n \geq 30$ et $m \geq 30$

Modélisation associée : deux séries d'observations $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$, 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 :

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{p}_{n+m}(1 - \hat{p}_{n+m})(1/n + 1/m)}}$$

où l'estimateur groupé $\hat{p}_{n+m}$ de la proportion commune sous H_0 est défini par

$$\hat{p}_{n+m} = \frac{n\bar{X}_n + m\bar{Y}_m}{n + m}$$

Comportement sous H_0 : $T_{n,m} \rightarrow \mathcal{N}(0, 1)$ 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 5.5 – Principe du test de comparaison des proportions de deux populations.

3.3. Applications du principe

Exercice 5.1 (Video killed the star). Prenons le cas d'une radio généraliste étudiant le taux d'audience du segment étudiant. Il est demandé de déterminer si l'audience sur ce segment a augmenté de manière significative entre deux sondages espacés de six mois, le second ayant lieu après un rajeunissement de la grille des programmes effectué afin de plaire aux jeunes adultes.

Etudiants	Auditeurs	Non-auditeurs	Total
Sondage 1	30	72	102
Sondage 2	338	662	1 000
Total	368	734	1 102

On pourrait raisonner 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 5.1. Les données du premier sondage sont notées $z_1, \dots, z_{102}$ et celles du second sondage, $y_1, \dots, y_{1000}$, 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.

On dispose des statistiques d'échantillon $\bar{z}_{102} = 30/102 = 29.4\%$ et $\bar{y}_{1000} = 338/1000 = 33.8\%$. 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,

$$T_{102,1000} = \frac{\bar{Z}_{102} - \bar{Y}_{1000}}{\sqrt{\hat{p}_{1102}(1 - \hat{p}_{1102})(1/102 + 1/1000)}}$$

où l'estimateur groupé $\hat{p}_{1102}$ de la proportion commune sous H_0 est défini par

$$\hat{p}_{1102} = \frac{102 \bar{X}_{102} + 1000 \bar{Y}_{1000}}{1102}.$$

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 $\bar{Z}_{102} - \bar{Y}_{1000}$ 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 $\hat{p}_{1102}$ du paramètre est

$$\bar{p}_{1102} = 368/1102 = 33.4\%,$$

de sorte que la valeur réalisée pour la statistique de test est égale à

$$\frac{\bar{z}_{102} - \bar{y}_{1000}}{\sqrt{\bar{p}_{1102}(1 - \bar{p}_{1102})(1/102 + 1/1000)}} = \frac{0.294 - 0.338}{\sqrt{0.334(1 - 0.334)(1/102 + 1/1000)}} = -0.895,$$

soit une P-valeur égale à

$$\mathbb{P}\{Z \leq -0.895\} = 18.5\% \quad \text{où } Z \sim \mathcal{N}(0, 1).$$

(Voir la figure 5.6 pour une illustration de l'ensemble du raisonnement précédent.)

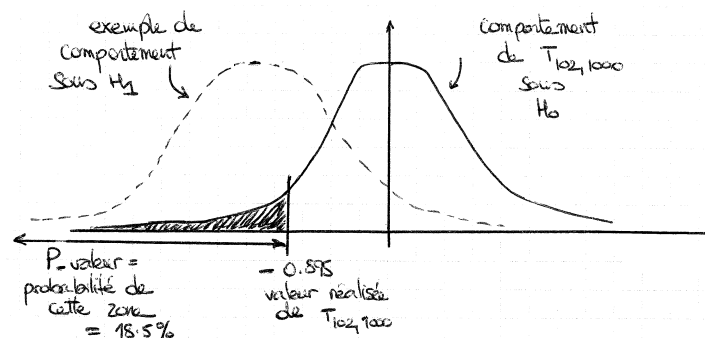


FIGURE 5.6 – Calcul de la P-valeur associée au test de l'exercice 5.1.

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 disponibles⁴.

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 SPSS 5.4. Le test présenté ci-dessus, dans sa version bilatère, est un cas particulier du test du χ^2 d'indépendance, que nous verrons dans le chapitre 6. 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 5.4 (Accidents de voiture selon le sexe). Pour les données de la figure 5.4, on obtient par exemple les résultats reproduits à la figure 5.7. On note en particulier que SPSS procure un tableau 2×2 de résumé des données, similaire à celui que nous avons 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 la figure 5.7. 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.

Exercice 5.2 (Accidents de voiture selon le sexe). Effectuer à la main le calcul menant à la P-valeur du test de comparaison de proportions, que l'on lit ici dans la première ligne du second tableau de la figure 5.7. (La correction manuscrite est détaillée aux pages suivantes.)

4. 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		Total
		Male	Female	
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 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

Diagramme en barres

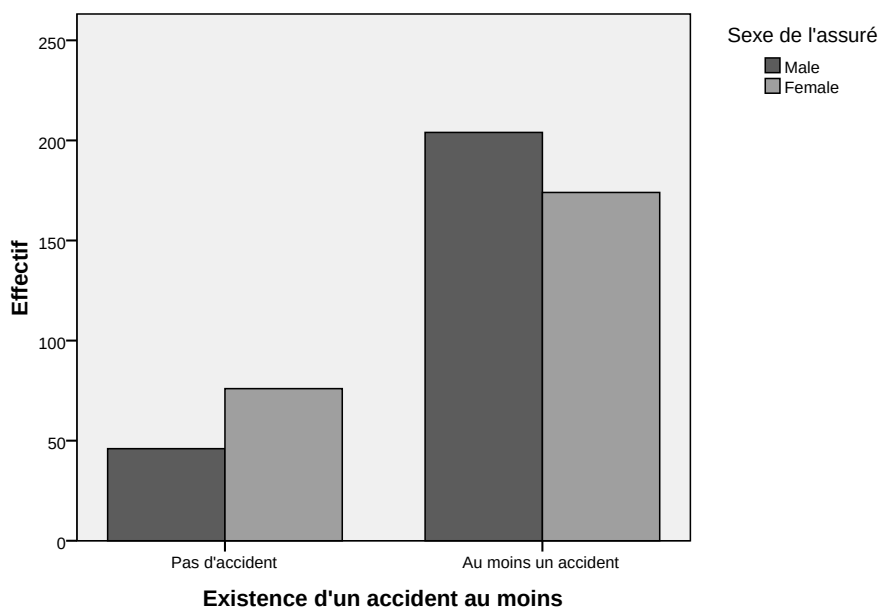


FIGURE 5.7 – Existence d'accidents de voiture au cours des cinq dernières années selon le sexe des assurés.

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 $\bar{y}_{250} = \frac{174}{250} = 69.6\%$

On les modélise, vu leur tirage au hasard, comme la réalisation de $X_1, \dots, X_{250} \text{ iid } \sim \text{Ber}(p_\sigma)$ et $Y_1, \dots, Y_{250} \text{ iid } \sim \text{Ber}(p_\varphi)$, où les paramètres d'intérêt (inconnus) p_σ et p_φ désignent respectivement les taux, sur l'ensemble des assurés, des accidents responsables selon le sexe.

On effectue le test bilatère de $H_0: p_\sigma = p_\varphi$ contre $H_1: p_\sigma \neq p_\varphi$. (Cas d'un statisticien sans préjugés.)

On recourt à la statistique de test

$$T_{250,250} = \frac{\bar{x}_{250} - \bar{y}_{250}}{\sqrt{\hat{p}_{500}(1-\hat{p}_{500})(\frac{1}{250} + \frac{1}{250})}}$$

où $\hat{p}_{500} = \frac{1}{500} (X_1 + \dots + X_{250} + Y_1 + \dots + Y_{250})$

Sous H_0 , $T_{250,250} \stackrel{(d)}{\approx} \mathcal{U}(0,1)$ tandis que sous H_1 , $T_{250,250}$ peut prendre des valeurs ou plus grandes ou plus petites.

La zone de rejet est donc de la forme $]-\infty, -r[\cup]r, +\infty[$.

Les valeurs réalisés de $\hat{p}_{500}$ et $T_{250,250}$ sont respectivement

$$\bar{p}_{500} = \frac{1}{500} (204 + 174) = 75.6\%$$

$$\text{et } \frac{\bar{z}_{250} - \bar{y}_{250}}{\sqrt{\bar{p}_{500}(1 - \bar{p}_{500})(\frac{1}{250} + \frac{1}{250})}} = \frac{0.816 - 0.696}{\sqrt{0.756(1 - 0.756)(\frac{1}{250} + \frac{1}{250})}} \approx 3.12$$

Note : Pour des raisons que nous expliquerons dans la suite de ce polycopié, valeur est en fait la racine carrée de la valeur réelle 9.758 de la statistique du Khi-deux (χ^2) de Pearson.

La P-valeur est donnée, selon le dessin habituel, par

$$\begin{aligned} P\{N \leq -3.12 \text{ ou } N \geq 3.12\} &= 2 \times P\{N \geq 3.12\} = 2(1 - P\{N \leq 3.12\}) \\ &\approx 2(1 - 0.999) \quad \text{selon la table} \\ &= 0.2\% \end{aligned}$$

On retrouve bien la P-valeur bilatérale de 0.2% du tableau.

Conclusion statistique : rejet très clair de H_0 , soit, au vu de $\bar{z}_{250}$ 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!

Mise à jour

La correction de cet exercice avait été rédigée en 2008-2009, avant qu'au printemps 2012 un arrêt de la Cour européenne de Justice ne juge discriminatoire les différences de tarifications selon le sexe (les seules différences de tarifications peuvent venir des antécédents d'accidents ou de leur absence).

4. 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 à la fin de ce chapitre.

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 $\bar{X}_n - \bar{Y}_m$ 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).

4.1. Lecture de tableaux SPSS

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)

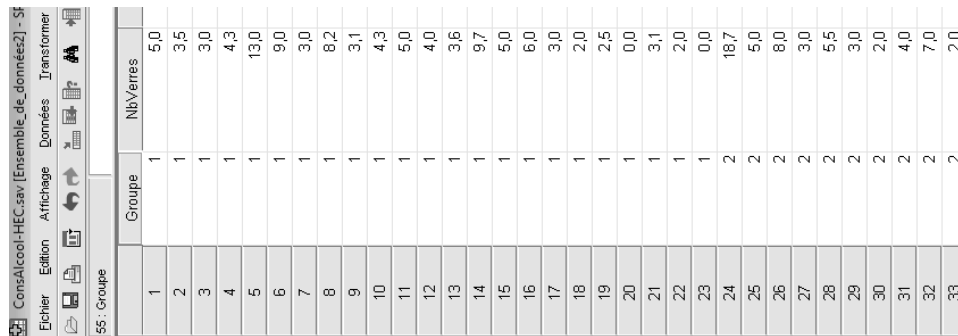
Il faut premièrement déterminer si les variances peuvent être dites égales ou non : c'est l'objet des deux premières colonnes. En fonction du résultat, on lira le T-test de la première ligne (variances égales) ou de la seconde ligne (variances différentes).

La minute SPSS 5.5. La réalisation du tableau précédent est fournie par Analyse / Comparer les moyennes / Test-T pour échantillons indépendants.

4.2. Application de la lecture de tableaux SPSS

Exemple 5.5 (Consommation d'alcool à HEC). A la figure 5.8, on reproduit les résultats d'une enquête quantitative sur la consommation d'alcool à HEC, effectuée en cours par Veronika Czellar. 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.



Group	N	NbVerres
1	1	5,0
2	1	3,5
3	1	3,0
4	1	4,3
5	1	13,0
6	1	9,0
7	1	3,0
8	1	8,2
9	1	3,1
10	1	4,3
11	1	5,0
12	1	4,0
13	1	3,6
14	1	9,7
15	1	5,0
16	1	6,0
17	1	3,0
18	1	2,0
19	1	2,5
20	1	0,0
21	1	3,1
22	1	2,0
23	1	0,0
24	2	18,7
25	2	5,0
26	2	8,0
27	2	3,0
28	2	5,5
29	2	3,0
30	2	2,0
31	2	4,0
32	2	7,0
33	2	2,0

FIGURE 5.8 – 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 précédemment⁵. SPSS propose le tableau de la figure 5.9, que nous allons lire ensemble : le pré-test d'égalité des variances associe ici une P-valeur de 3 % à l'hypothèse H'_0

Statistiques de groupe

Groupes	N	Moyenne	Ecart-type	Erreur standard moyenne
Verres bus	8h	23	4,448	3,0598
	10h	31	7,735	10,1658

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 5.9 – 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).

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).

La minute SPSS 5.6. Regardons quelques autres colonnes du second tableau de la figure 5.9 : Différence moyenne est la valeur réalisée $\bar{x}_n - \bar{y}_m$, ici, $\bar{x}_{23} - \bar{y}_{31}$. 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

5. 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.

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 qui suivent.

4.3. Compléments mathématiques (lecture très facultative)

Il faut donc trouver les standardisations adéquates à appliquer à la différence $\bar{X}_n - \bar{Y}_m$. A cet effet, on calcule au préalable la variance de cette différence :

$$\sigma^2(\bar{X}_n - \bar{Y}_m) = \frac{1}{n}\sigma_X^2 + \frac{1}{m}\sigma_Y^2.$$

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).

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,

$$\hat{\sigma}_{X,n}^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \quad \text{et} \quad \hat{\sigma}_{Y,m}^2 = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2.$$

On considère alors la statistique de test

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{\sigma}_{X,n}^2/n + \hat{\sigma}_{Y,m}^2/m}}.$$

Sa loi sous H_0 et son comportement sous H_1 sont détaillés dans le principe suivant.

Principe 5.2 (Test d'Aspin-Welch). *On part d'une situation modélisée par le fait qu'une première série d'observations $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus) μ_X et σ_X , tandis qu'une seconde série d'observations $Y_1, \dots, Y_m$ 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 $H_0 : \mu_X = \mu_Y$,*

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{\sigma}_{X,n}^2/n + \hat{\sigma}_{Y,m}^2/m}} \stackrel{(d)}{\approx} \mathcal{T}_{\lfloor v \rfloor}$$

où $\lfloor v \rfloor$ est la partie entière (inférieure) de

$$v = \frac{\left(\frac{\hat{\sigma}_{X,n}^2}{n} + \frac{\hat{\sigma}_{Y,m}^2}{m} \right)^2}{\frac{\hat{\sigma}_{X,n}^4}{n^2(n-1)} + \frac{\hat{\sigma}_{Y,m}^4}{m^2(m-1)}};$$

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

$$T_{n,m} \stackrel{(d)}{\approx} \mathcal{T}_{[v]}$$

signifie que la loi de $T_{n,m}$ est très proche de la loi $\mathcal{T}_{[v]}$ – 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. Cela étant, en pratique, lorsque n , m (et donc v) 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 5.2 et ses lois de Student.

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

$$\hat{\sigma}_{n+m}^2 = \frac{(n-1)\hat{\sigma}_{X,n}^2 + (m-1)\hat{\sigma}_{Y,m}^2}{n+m-2},$$

et on considère la statistique de test

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{\sigma}_{n+m}^2 (1/n + 1/m)}}.$$

Sa loi sous H_0 et son comportement sous H_1 sont détaillés dans le principe suivant.

Principe 5.3. *On part d'une situation modélisée par le fait qu'une première série d'observations $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres (inconnus) μ_X et σ , tandis qu'une seconde série d'observations $Y_1, \dots, Y_m$ 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 $H_0 : \mu_X = \mu_Y$,*

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{\sigma}_{n+m}^2 (1/n + 1/m)}} \sim \mathcal{T}_{n+m-2}$$

où l'estimateur de la variance $\hat{\sigma}_{n+m}^2$ (dit estimateur groupé) est défini par

$$\hat{\sigma}_{n+m}^2 = \frac{(n-1)\hat{\sigma}_{X,n}^2 + (m-1)\hat{\sigma}_{Y,m}^2}{n+m-2}.$$

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 5.2. Le principe présenté ci-dessus est donc fort similaire au principe 5.2, au remplacement près, dans la définition de la statistique de test, des estimateurs séparés $\hat{\sigma}_{X,n}^2$ et $\hat{\sigma}_{Y,m}^2$ dans le dénominateur par la seule version groupée $\hat{\sigma}_{n+m}^2$. 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 5.3 (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 5.3 par rapport à celui du principe 5.2, 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...!

Lectures de tableaux de résultats SPSS, suite. A propos du second tableau de la figure 5.9, 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, v ou $n + m - 2$.

Notez également qu'il est facile d'adapter les résultats des principes 5.3 et 5.2 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.

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 (ordinaire ou nominale) peut prendre, ou qu'un tel couple de variables peut prendre.

Nous étudions

- le test de l'ajustement à une loi de référence, utile par exemple pour vérifier que des objectifs de qualité sont atteints ou détecter des manipulations;
- le test de l'indépendance entre deux variables qualitatives, utile par exemple en marketing pour savoir s'il faut segmenter une population.

1. Test du χ^2 d'ajustement simple

Il sert à vérifier que des données qualitatives sont distribuées selon une certaine loi de référence.

1.1. Éléments mathématiques menant au principe de test

On part de données qualitatives (ordinales ou nominales) $x_1, \dots, x_n$. Sans perte de généralité, on note $\{1, 2, \dots, k\}$ 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 de variables aléatoires $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une certaine loi $\mathbf{p} = (p_1, \dots, p_k)$ sur $\{1, 2, \dots, k\}$, inconnue.

On pense, pour des raisons subjectives et qui dépendent du contexte, à une certaine loi de référence $\mathbf{p}^{\text{ref}} = (p_1^{\text{ref}}, \dots, p_k^{\text{ref}})$ et on veut tester

$$H_0 : \mathbf{p} = \mathbf{p}^{\text{ref}} \quad \text{contre} \quad H_1 : \mathbf{p} \neq \mathbf{p}^{\text{ref}} .$$

(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 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 $j = 1, \dots, k$, la proportion observée $\hat{p}_{j,n}$ de la catégorie j dans les données soit proche de p_j :

$$\hat{p}_{j,n} = \frac{N_{j,n}}{n} \xrightarrow{\mathbb{P}} p_j \quad \text{où} \quad N_{j,n} = \text{Card}\{t : X_t = j\}$$

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 $\hat{p}_{j,n}$ 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 :

$$D_n(\mathbf{p}^{\text{ref}}) \stackrel{\text{not.}}{=} n \sum_{j=1}^k \frac{(\hat{p}_{j,n} - p_j^{\text{ref}})^2}{p_j^{\text{ref}}} = \sum_{j=1}^k \frac{(N_{j,n} - n p_j^{\text{ref}})^2}{n p_j^{\text{ref}}}.$$

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 $D_n(\mathbf{p}^{\text{ref}})$ sous H_0 et H_1 , nous aurons besoin de la définition suivante.

Définition 6.1. Si $Z_1, \dots, Z_k$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi $\mathcal{N}(0, 1)$, alors on appelle la loi de $Z_1^2 + \dots + Z_k^2$ la loi du χ^2 à k degrés de liberté et on la note χ_k^2 .

Les quantiles de ces lois sont fournis dans les tables à la fin du présent polycopié. On reproduit à la figure 6.1 des tracés des fonctions densité et de répartition de lois du χ^2 , de différents degrés de liberté.

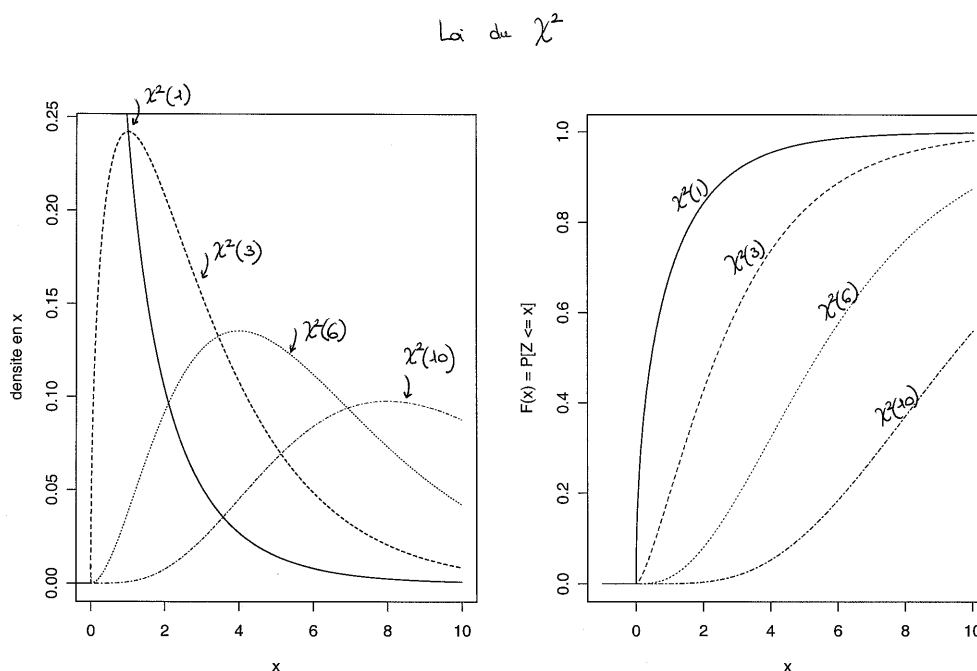


FIGURE 6.1 – Densités (à gauche) et fonctions de répartitions (à droite) des lois du χ^2 , pour différents degrés de liberté.

Or, la théorie mathématique assure que sous H_0 , la statistique $D_n(\mathbf{p}^{\text{ref}})$ converge en loi vers une loi du χ^2 à $k - 1$ degrés¹ de liberté, notée χ_{k-1}^2 .

1. Pourquoi $k - 1$ et pas k ? C'est facile à retenir : une probabilité sur un ensemble à k éléments est définie par $k - 1$

Sous H_1 , les écarts entre les $\hat{p}_{j,n}$ et les p_j tendent à être grands, et le facteur multiplicatif n dans la première définition assure alors que $D_n(\mathbf{p}^{\text{ref}}) \xrightarrow{\mathbb{P}} +\infty$ lorsque $n \rightarrow \infty$. Tout cela conduit au principe de test énoncé ci-dessous.

1.2. Enoncé du principe de test

Principe 6.1. *Test du χ^2 d'ajustement 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 :

$$D_n(\mathbf{p}^{\text{ref}}) \stackrel{\text{not.}}{=} n \sum_{j=1}^k \frac{(\hat{p}_{j,n} - p_j^{\text{ref}})^2}{p_j^{\text{ref}}} = \sum_{j=1}^k \frac{(N_{j,n} - n p_j^{\text{ref}})^2}{n p_j^{\text{ref}}}$$

Comportement sous H_0 : $D_n(\mathbf{p}^{\text{ref}}) \rightarrow \chi_{k-1}^2$ lorsque $n \rightarrow \infty$

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 .

Ainsi, la zone de rejet associée au test est toujours unilatère ; elle vaut, pour une erreur de première espèce α fixée : $] c_{k-1, 1-\alpha}, +\infty[$, où $c_{k-1, 1-\alpha}$ désigne le quantile d'ordre $1 - \alpha$ de la loi χ_{k-1}^2 .

Remarque 6.1 (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, $n \geq 30$, et que toutes les valeurs possibles j conduisent à des effectifs attendus $n p_j^{\text{ref}} \geq 5$. Si la seconde condition n'est pas vérifiée, il faut alors procéder à un regroupement de classes (l'agrégation de deux classes j et j' en une seule classe) ou à la suppression de certaines classes. 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 chapitre.

La minute SPSS 6.1. 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), ou supprimer des catégories (en effectuant une sélection des observations).

de ses valeurs, la dernière se déduisant des précédentes : on a effectivement $k - 1$ degrés de liberté, la terminologie est éclairante.

1.3. Application 1 : 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)

Certes, il semble y avoir débat sur la paternité véritable de cette phrase (d'aucuns disent qu'on l'attribue à tort à Churchill, voire, que ce serait une phrase de propagande nazie). Mais, qui que soit son auteur, cette citation souligne d'une part 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).

Aussi et surtout, elle indique d'autre part 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 exemples 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 allons voir comment des statisticiens ont pu (croire) détecter des fraudes aux élections présidentielles iraniennes (voir l'article relaté² à la figure 6.2).

Les données sont les suivantes : 29 régions, 4 candidats, soit 116 totaux de voix obtenues, desquels on extrait 116 chiffres des unités. Un extrait des données est reproduit à la figure 6.3.

On voit que les totaux de votes en faveur des 4 candidats en lice dans chacun des 29 provinces (nombres fournis par le gouvernement) 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 devraient pouvoir être considérés comme la réalisation de variables aléatoires indépendantes et identiquement distribuées selon une certaine loi p (ils n'ont pas de signification parce que ce sont des chiffres des unités, c'est-à-dire, le dernier chiffre significatif; cela ne serait bien sûr pas vrai pour le premier chiffre significatif!).

Ainsi, si ces totaux de voix ont été reportés de manière sincère, alors les données de chiffres des unités 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 p^{ref} sur $\{0, 1, \dots, 9\}$.

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 p pour les U_j (qui dépend du manipulateur et qui est inconnue; la seule chose que l'on sache est que cela n'est pas p^{ref}).

Les auteurs ont ensuite testé $H_0 : p = p^{\text{ref}}$ (le report des résultats a été sincère) contre $H_1 : p \neq p^{\text{ref}}$ (les résultats ont été manipulés). 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 (et c'est la seule manière juste et honnête de le faire), leur conclusion aurait été moins tranchée : on voit, à la figure 6.4, que la P-valeur est de 7.7%, on ne peut donc pas exclure un report sincère des résultats par les autorités iraniennes.

2. Attention, l'article que je vous propose ici 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.

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 6.2 – 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.”

Case	Province	Candidat	Voix	Chiffre
20	Kurdistan	Ahmadinejad	315689	9
21	Markazi	Ahmadinejad	572988	8
22	Mazandaran	Ahmadinejad	1289257	7
23	Qazvin	Ahmadinejad	498061	1
24	Qom	Ahmadinejad	422457	7
25	Semnan	Ahmadinejad	295177	7
26	Sistan and Baluchistan	Ahmadinejad	450269	9
27	Tehran	Ahmadinejad	3371523	3
28	Yazd	Ahmadinejad	337178	8
29	Zanjan	Ahmadinejad	444480	0
30	Ardabil	Karroubi	2319	9
31	Azarbaijan, East	Karroubi	7246	6
32	Azarbaijan, West	Karroubi	21609	9
33	Bushehr	Karroubi	3563	3
34	Chahar Mahaal and Bakhtiari	Karroubi	4127	7
35	Fars	Karroubi	16277	7
36	Gilan	Karroubi	7183	3
37	Golestan	Karroubi	10097	7
38	Hamadan	Karroubi	12032	2
39	Hormozgan	Karroubi	5126	6
40	Ilam	Karroubi	7471	1

Affichage des données
 Affichage des variables

FIGURE 6.3 – Les données associées à l'article reproduit à la figure 6.2.

Refaisons les calculs ensemble en appliquant le principe vu ci-dessus. On commence par remarquer que les conditions de validité asymptotique sont remplies (taille d'échantillon de $116 \geq 30$ et effectifs attendus tous égaux à 11.6 donc supérieurs à 5). La loi de référence est $p^{\text{ref}} = (1/10, \dots, 1/10)$. La valeur réalisée de la statistique de test est une somme de 10 termes,

$$\frac{(9 - 11.6)^2}{11.6} + \frac{(11 - 11.6)^2}{11.6} + \dots + \frac{(13 - 11.6)^2}{11.6} = 15.552.$$

La loi limite est une loi χ^2_9 . Les tables que je vous ai fournies permettent d'encadrer la P-valeur entre 5 % et 10 %, en lisant la ligne de la loi χ^2_9 (voir la figure 6.5 pour une illustration). Avec notre logiciel statistique, on a vu qu'elle valait plus exactement 7.7 %.

La minute SPSS 6.2. Le traitement SPSS de l'exemple est le suivant. J'avais déduit la colonne 4 (chiffres des unités) de la colonne 3, via Transformer / Recoder des variables. J'ai lancé le test par Analyse / Tests non paramétriques / Khi-deux et dans la fenêtre qui apparaît, j'ai indiqué qu'on voulait tester l'équiprobabilité des classes sur la variable Chiffre. J'ai alors obtenu la figure 6.4. On retrouve bien toutes les valeurs numériques que nous avons calculées à la main, et notamment, la P-valeur de 7.7 %. La petite note qui suit le second tableau indique que les conditions asymptotiques de la remarque 6.1 sont bien remplies et que le test est donc validement appliqué.

Test du Khi-deux

Chiffre			
	Effectif observé	Effectif théorique	Résidu
0	9	11,6	-2,6
1	11	11,6	-,6
2	8	11,6	-3,6
3	9	11,6	-2,6
4	10	11,6	-1,6
5	5	11,6	-6,6
6	14	11,6	2,4
7	20	11,6	8,4
8	17	11,6	5,4
9	13	11,6	1,4
Total	116		

Test	
	Chiffre
Khi-deux	15,552 ^a
ddl	9
Signification asymptotique	,077

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 11,6.

FIGURE 6.4 – Sortie du test du χ^2 d'ajustement à une loi uniforme pour les unités des résultats électoraux iraniens.

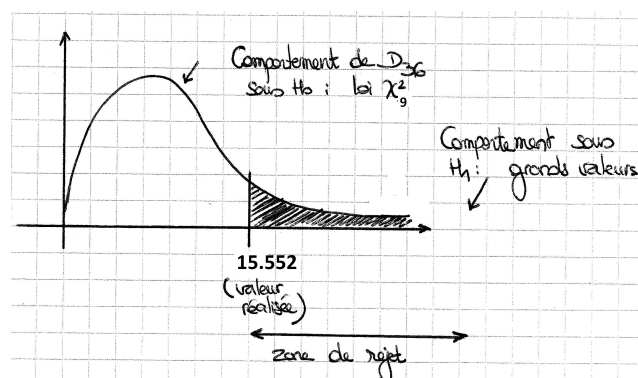


FIGURE 6.5 – Illustration du calcul de la P-valeur dans l'exemple des données d'élections en Iran.

1.4. Autres exemples d'application

Exemple 6.1 (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 trichent³, 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 might" sur mon site). Ces résultats ont été simplifiés pour donner lieu à un exercice de quiz, reproduit page 203.

Outre la détection de tricheries et de manipulations, on peut citer deux autres grandes applications des tests du χ^2 d'ajustement à une loi :

- la vérification de la représentativité d'un échantillon, cf. l'exercice «Epreuve pré-professionnelle de la campagne JE» page 194 ;
- la vérification que des objectifs commerciaux sont atteints, cf. l'exercice «Temps d'attente téléphonique pour joindre un service clients» page 196.

3. 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...

2. Test du χ^2 d'indépendance entre deux variables qualitatives

Les résultats présentés ici seront surtout importants en marketing : ils permettent de répondre à la question de savoir s'il est pertinent, ou pas, de segmenter ses clients (c'est-à-dire de distinguer des sous-populations au sein d'une population cible selon leur comportement en tant que consommateurs).

2.1. Deux exemples pour nous motiver

On va considérer des couples de variables qualitatives.

Exercice 6.1 (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 6.6. 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 6.2 (Prof Grincheux contre Prof Gentil). En 2007 et 2008, une collègue enseignante et moi-même utilisons un test pour vérifier⁴ que nous notions bien de la même manière dans nos groupes, qu'aucun de nous deux n'était ni plus sévère ni plus laxiste que l'autre. Je ne retrouve plus nos notes et ai eu la flemme de les demander au bureau de promotion, 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 ?

4. 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 l'uniformisation des échelles de notations décidée au début des années 2010.

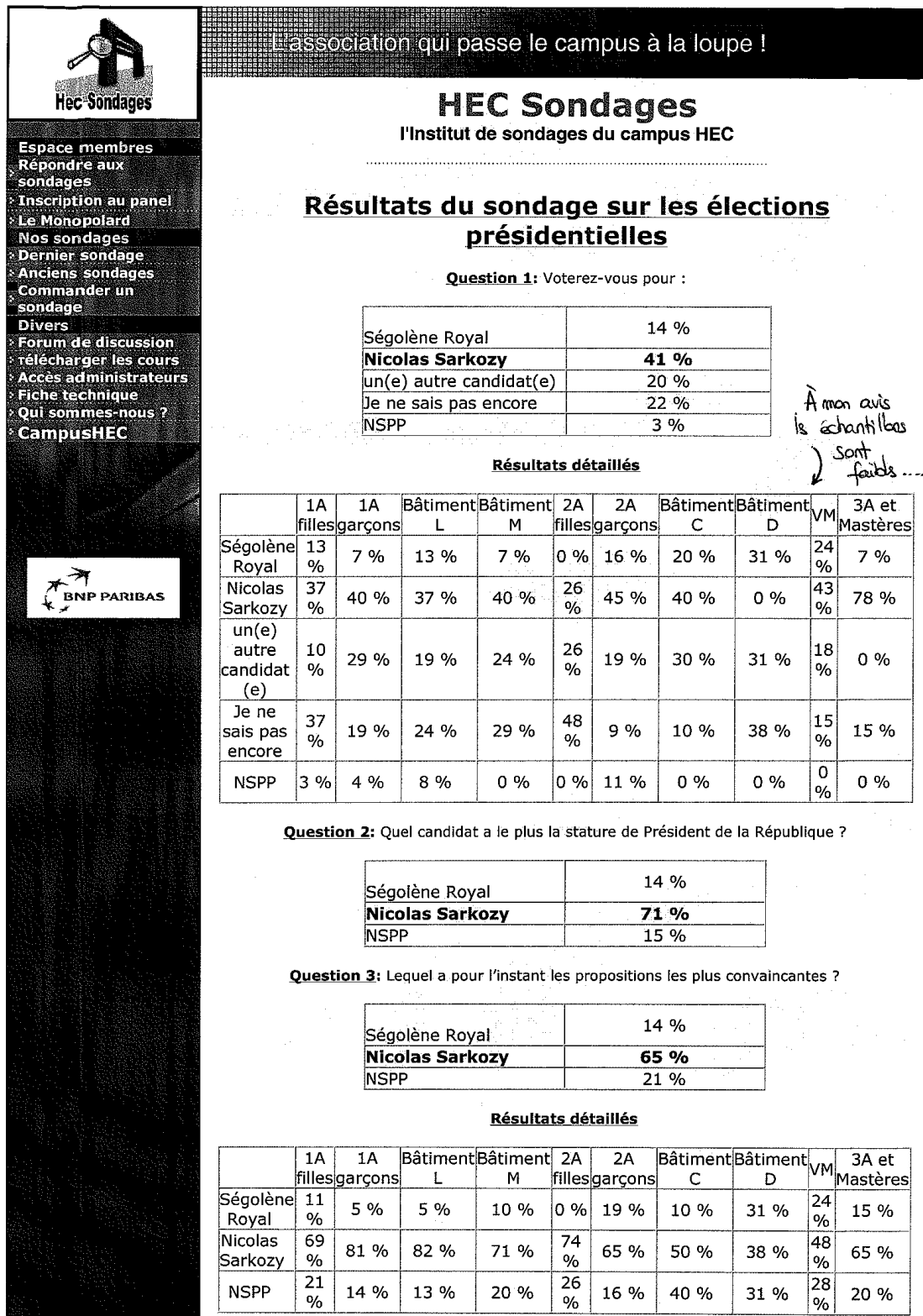


FIGURE 6.6 – Données de sondage telles que présentées sur le site de HEC Sondages (sondage effectué début 2007).

2.2. Éléments mathématiques menant au principe de test

En étudiant les exemples précédents, on remarque que l'on part d'une situation avec des couples de données $(x_1, y_1), \dots, (x_n, y_n)$. Supposons qu'on puisse les modéliser comme la réalisation de couples de variables aléatoires $(X_1, Y_1), \dots, (X_n, Y_n)$, 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 $r = \text{Card } \mathcal{X}$ et $s = \text{Card } \mathcal{Y}$, et sans perte de généralité, on effectue les identifications $\mathcal{X} = \{1, \dots, r\}$ et $\mathcal{Y} = \{1, \dots, s\}$. Enfin, on définit les marginales⁵ $\mathbf{p}_{\mathcal{X}}$ et $\mathbf{p}_{\mathcal{Y}}$ de $\mathbf{p}$ et on note

$$\mathbf{p}(x, y), \quad \mathbf{p}_{\mathcal{X}}(x), \quad \mathbf{p}_{\mathcal{Y}}(y)$$

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 :

$$N_{x,y} = \text{Card}\{j : X_j = x \text{ et } Y_j = y\}, \quad N_{x,\cdot} = \text{Card}\{j : X_j = x\}, \quad N_{\cdot,y} = \text{Card}\{j : Y_j = y\} .$$

(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 $\mathbf{p}_{\mathcal{Y}}$:

$$\hat{\mathbf{p}}_{\mathcal{X}} = \left(\frac{N_{1,\cdot}}{n}, \dots, \frac{N_{r,\cdot}}{n} \right) \quad \text{et} \quad \hat{\mathbf{p}}_{\mathcal{Y}} = \left(\frac{N_{\cdot,1}}{n}, \dots, \frac{N_{\cdot,s}}{n} \right) .$$

Encore par loi des grands nombres, les valeurs observées

$$\frac{N_{x,y}}{n} \quad \text{et} \quad \hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y)$$

sont toujours (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 $\hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y)$. C'est ce que fera notre test, puisqu'il considère la statistique

$$D_n^{\text{indep not.}} \equiv n \sum_{x=1}^r \sum_{y=1}^s \frac{\left(N_{x,y}/n - \hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y) \right)^2}{\hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y)} = \sum_{x=1}^r \sum_{y=1}^s \frac{\left(N_{x,y} - n \hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y) \right)^2}{n \hat{\mathbf{p}}_{\mathcal{X}}(x) \hat{\mathbf{p}}_{\mathcal{Y}}(y)} .$$

2.3. Enoncé du principe de test

Attention ! Notez que vous n'avez pas le choix ici des hypothèses H_0 et H_1 ; la première se réfère toujours à l'indépendance et la seconde, à l'absence d'indépendance. Il n'y a pas à tergiverser sur leur choix (une fois n'est pas coutume).

5. Je vous renvoie à vos cours de classes préparatoires pour la définition d'une marginale : c'est la loi induite sur une composante.

Principe 6.2. Test d'indépendance de couples de données

Données : couples $(x_1, y_1), \dots, (x_n, y_n)$ prenant leurs valeurs dans un ensemble-produit $\{1, \dots, r\} \times \{1, \dots, s\}$

Modélisation associée : couples d'observations $(X_1, Y_1), \dots, (X_n, Y_n)$ indépendants et identiquement distribués selon une certaine loi $\mathbf{p}$ sur l'ensemble-produit $\{1, \dots, r\} \times \{1, \dots, s\}$

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 :

$$D_n^{\text{indép}} \stackrel{\text{not.}}{=} \sum_{x=1}^r \sum_{y=1}^s \frac{(N_{x,y} - n \hat{p}_X(x) \hat{p}_Y(y))^2}{n \hat{p}_X(x) \hat{p}_Y(y)}$$

Comportement sous H_0 : $D_n^{\text{indép}} \xrightarrow{\text{d}} \chi_{(r-1)(s-1)}^2$

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, $D_n^{\text{indép}}$ 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 : $] c_{(r-1)(s-1), 1-\alpha}, +\infty[$, où $c_{(r-1)(s-1), 1-\alpha}$ désigne le quantile d'ordre $1 - \alpha$ de la loi $\chi_{(r-1)(s-1)}^2$.

2.4. 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 (re-normalisé par le nombre total d'observations) aux observations effectivement obtenues dans chaque case. Cela conduit au tableau du haut de la figure 6.7 : 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 6.1. 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 6.1.

Le choix des sondés ayant été, espérons-le, effectué au hasard, on peut modéliser ces données comme les réalisations de couples de variables aléatoires $(X_1, Y_1), \dots, (X_{286}, Y_{286})$ 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 $\mathbf{p}_X$ et $\mathbf{p}_Y$. On estime tout d'abord la loi-produit associée; elle est obtenue par produit des estimées des marginales

$$\left(\frac{23}{286}, \frac{150}{286}, \frac{36}{286}, \frac{69}{286}, \frac{8}{286} \right) \quad \text{et} \quad \left(\frac{93}{286}, \frac{97}{286}, \frac{96}{286} \right);$$

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 $286 \hat{p}_X(2) \hat{p}_Y(1)$, 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 6.7 (où l'on retrouve bien par exemple les valeurs 48.8 et 38 dans la première case de la deuxième ligne).

La minute SPSS 6.3. La figure 6.7 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 6.1 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 6.8. L'asymptotique étant désormais respectée, on peut en tirer des conclusions. On calcule combien vaut la réalisation de la statistique $D_{286}^{\text{indép}}$ sur les données. A cet effet, on calcule la somme de douze termes :

$$\frac{(9 - 7.5)^2}{7.5} + \frac{(38 - 48.8)^2}{48.8} + \dots + \frac{(14 - 25.8)^2}{25.8} \approx 49.157.$$

La valeur réalisée de $D_{286}^{\text{indép}}$ 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 $] c_{6,1-\alpha}, +\infty[$.

On remarquera que SPSS reprend la valeur réalisée de $D_{286}^{\text{indép}}$, le nombre de degrés de liberté de sa loi limite, ainsi que la P-valeur associée, dans le second tableau de la figure 6.8.

La P-valeur est très faible, comme l'indique SPSS, quasi-nulle; un calcul avec un logiciel de statistiques plus efficace montre que

$$\mathbb{P}\{X \geq 49\} \leq 10^{-8} \quad \text{où } X \sim \chi_6^2.$$

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).

Dans le simple but d'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 6.9.

La minute SPSS 6.4. 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 ne peut donc pas rejeter une homogénéité des votes entre 1A et 2A (seconde partie d'une conclusion statistique); et même, par un glissement sémantique valide uniquement dans le cas des tests du χ^2 (qui sont «puissants»), on peut affirmer l'homogénéité des votes entre 1A et 2A.

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 la cristallisation⁶ des votes

6. Cette agrégation des voix autour de Sarkozy aurait d'ailleurs pu être traitée par un simple test unilatère de proportions (à la condition expresse qu'on l'ait soupçonnée à l'avance).

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 6.7 – 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			Total
			1A	2A	3A	
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 6.8 – 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 6.9 – Résultat du test du χ^2 d'homogénéité entre les votes des étudiants de première et deuxième années.

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 (version «Hémisphère droit»), 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 (tendance «Contre-courant»), 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 6.2. 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 contraints d'inventer à l'exemple 6.1 à partir des résultats proposés en pourcentages.

2.5. Application du principe à l'exemple du test d'indépendance (d'homogénéité) des notations

Correction 6.2. 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 ?

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 y_j , qui est dans l'ensemble formé des lettres $\{A, B, C, D, E, F\}$, 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 $(X_1, Y_1), (X_2, Y_2), \dots, (X_{194}, Y_{194})$, 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 6.10) 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 6.10. L'effectif attendu pour le gentil professeur et la note C est déterminé comme la réalisation de $194 \hat{p}_X(2) \hat{p}_Y(3)$, 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 :

$$\frac{(14 - 17.1)^2}{17.1} + \dots + \frac{(17 - 19.9)^2}{19.9} \approx 2.339 .$$

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)

$$\mathbb{P}\{X > 2.339\} = 67.4 \% \quad \text{où } X \sim \chi_4^2 .$$

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 méritées⁷.

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 !
- L'administration, si elle reçoit des plaintes d'étudiants, les renverra dans leurs cordes et les cataloguera comme mauvais coucheurs.

7. On rappelle que les tests du χ^2 sont dits « puissants » : pour eux, et pour eux seuls, on peut faire le glissement sémantique consistant à dire que H_0 est vraie alors qu'elle n'est simplement pas rejetée.

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						Total
			A	B	C	D	E	F	
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)					Total
			A	B	C	D	G	
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 6.10 – 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).

- Les étudiants ont intérêt à se taire face à ces données, ils ne pourraient que se décrédibiliser⁸ ; 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 6.3 (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 !

8. Rappelez-moi en cours de vous narrer l'histoire de l'étudiant se plaignant des affectations d'électifs en fonction de la place dans l'alphabet et comment, après m'avoir consulté, l'administration l'a renvoyé dans ses cordes (en le cataloguant comme ch...).

3. 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.

3.1. Test d'ajustement lorsqu'il n'y a que deux classes ($k = 2$)

Exemple 6.2. 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 comparaison⁹ à une proportion de référence, voir le principe 4.2. Le test d'ajustement simple du χ^2 n'est à utiliser que lorsqu'il y a strictement plus de deux classes.

Remarque 6.4. 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 4.2. 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 $k \geq 3$ classes, alors un test d'ajustement du χ^2 s'impose.

3.2. Test d'indépendance lorsqu'il n'y a que deux fois deux classes

De même, lorsque $r = s = 2$ (mais plus si $r \geq 3$ ou $s \geq 3$), i.e., que la table de contingence contient quatre cases, deux lignes et deux colonnes, on devrait se reporter de préférence au principe de comparaison de deux proportions, le principe 5.1.

9. Au fait, quelle P-valeur trouvez-vous ?

Estimation par intervalles

A l'aide des tests d'hypothèses du chapitre 4 nous pouvons dire si telle ou telle valeur de référence est compatible avec les observations, et à quel degré. Une autre manière de procéder est de proposer un intervalle de valeurs compatibles avec les données : un intervalle de confiance.

Vous avez déjà vu pour la plupart cette notion en classes préparatoires. Ce chapitre peut donc être vu comme un intermède agréable où l'on retrouve des notions connues après la série de chapitres sur les tests d'hypothèses. Nous allons cependant voir deux nouveautés essentielles :

- la planification de sondages ;
- le traitement d'intervalles de confiance simultanés.

1. Rappels mathématiques (lecture en diagonale possible)

Je commence ce chapitre à dessein par des formules très techniques, *a priori* vues en grande partie en classes préparatoires. Vous pouvez sauter la lecture de cette partie.

En revanche, l'application de ces formules et la rédaction à utiliser, ainsi que des commentaires méthodologiques, sont procurés plus loin dans ce chapitre : eux seront importants à méditer, ce qui n'est pas du tout le cas des éléments de mathématiques rappelés ci-dessous (enfin, certains éléments peuvent être nouveaux, notamment à ceux qui ne sont pas issus des filières EC/S ; mais peu nous chaut des mathématiques, dans ce cours, ce qui compte, ce sont les mises en œuvre et les interprétations).

1.1. Le minimum de vocabulaire pour commencer

Comme dans les chapitres précédents, on suppose qu'on a déjà modélisé les valeurs observées $x_1, \dots, x_n$ comme étant la réalisation de $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon une loi commune $\mathbb{P}_{\theta_0}$ inconnue et on s'intéresse à l'estimation d'une quantité $g(\theta_0)$ dérivée de cette loi (généralement son espérance μ_0).

Définition 7.1. *Un intervalle de confiance $\widehat{I}_n$ pour $g(\theta_0)$ est la donnée d'un couple d'estimateurs $\widehat{g}_n \leq \widehat{g}'_n$, à partir de qui on construit*

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

Définition 7.2 (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 ,*

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

Définition 7.3 (Niveau asymptotique). *Un intervalle de confiance $\hat{I}_n$ 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 , il existe une suite (α_n) telle que $\alpha_n \rightarrow \alpha$ et*

$$\mathbb{P}\{g(\theta_0) \in \hat{I}_n\} \geq 1 - \alpha_n .$$

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

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

La définition 7.3 est cependant plus générale, au sens où il n'y est pas nécessaire qu'une limite existe.

Remarque 7.1. Il est difficile de dire ce qu'est un « bon » intervalle de confiance. 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 7.1, 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 rappeler comment, justement, les exploiter intelligemment.



FIGURE 7.1 – 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 !

1.2. Intervalles de confiance asymptotiques sur la moyenne

On a rappelé à la page 60 la convergence fondamentale suivante :

$$\sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0) \rightharpoonup \mathcal{N}(0, 1) .$$

C'est une adaptation du théorème limite central avec substitution de la variance inconnue par un estimateur convergent (en l'occurrence, l'estimateur empirique débiaisé de la variance). On en déduit les intervalles de confiance suivants.

Corollaire 7.1. *On rappelle qu'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 :*

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

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

$$\text{et} \quad \left[\bar{X}_n - z_{1-\alpha} \sqrt{\frac{\hat{\sigma}_n^2}{n}}, +\infty \right[.$$

On propose la démonstration à partir de (1.2) uniquement pour la culture.

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

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

de sorte que par (1.2),

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

où Z suit une loi normale $\mathcal{N}(0, 1)$. L'égalité de la dernière probabilité à $1 - \alpha$ peut être vue ainsi : la loi normale est symétrique par rapport à 0, ainsi,

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

puis, par inclusion d'événements,

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

(Faites un dessin pour mieux suivre le raisonnement ci-dessus ; il sera similaire à la figure 3.3.) $\square$

Remarque 7.2 (De l'importance d'estimer la variance...). D'aucuns pourraient vouloir tirer de la convergence en loi donnée par le théorème limite central que, par exemple, l'intervalle

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

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 7.1.

1.3. Cas particulier : estimation par intervalle asymptotique d'une proportion

On se place ici dans le cas d'un modèle de Bernoulli, dans lequel on sait que $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées selon une loi $\mathcal{B}(p_0)$ appartenant à la famille de lois $\{\mathcal{B}(p), p \in [0, 1]\}$. Le vrai paramètre p_0 est inconnu et on cherche à l'estimer par un intervalle.

Or, les éléments culturels ci-dessus, et notamment la preuve de (1.2) à la page 60, montrent que la seule propriété que l'on utilise de l'estimateur de la variance est son caractère convergent. Ainsi, dans le cas d'une proportion, on tend à remplacer l'estimateur $\hat{\sigma}_n^2$ de la variance $p_0(1 - p_0)$ par $\bar{X}_n(1 - \bar{X}_n)$, qui a une expression plus simple à calculer et plus agréable à l'œil... et qui est toujours convergent, par stabilité de la convergence en probabilité par passage aux fonctions continues (voir page 60). Il vient ici

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

et les mêmes techniques que dans la preuve du corollaire 7.1 permettent de tirer des intervalles de confiance asymptotiques à partir de cette convergence en loi.

Corollaire 7.2. 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 :

$$\begin{aligned} & \left[\bar{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right] ; \\ & \left[0, \bar{X}_n + z_{1-\alpha} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \right] ; \\ \text{et} \quad & \left[\bar{X}_n - z_{1-\alpha} \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, 1 \right] . \end{aligned}$$

1.4. Intervalles de confiance (exacts) sur la moyenne pour taille d'échantillon petite

Les résultats précédents sont asymptotiques et reposent sur le théorème limite central. On n'en croit les résultats en pratique que lorsque la taille d'échantillon n est suffisamment grande, disons $n \geq 30$. Lorsque ce n'est pas le cas, on a une solution, mais uniquement dans le cas du modèle gaussien, fondé sur la loi de Student (voir page 70).

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 , on a par définition

$$\sqrt{\frac{n}{\hat{\sigma}_n^2}} (\bar{X}_n - \mu_0) \sim \mathcal{T}_{n-1},$$

d'où l'on tire les intervalles de confiance suivants. (La preuve de ces résultats est totalement similaire à ceux du corollaire 7.1, au remplacement des lois normales et de leurs quantiles par ceux de la loi de Student.)

Corollaire 7.3. On note $t_{n-1,\beta}$ le β -quantile de la loi $\mathcal{T}_{n-1}$. Alors, sous les hypothèses de normalité ci-dessus, les intervalles suivants sont des intervalles de confiance de la moyenne μ_0 , non asymptotiques et exactement de niveau $1 - \alpha$,

$$\begin{aligned} & \left[\bar{X}_n - t_{n-1,1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}}, \bar{X}_n + t_{n-1,1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] \stackrel{\text{not.}}{=} \left[\bar{X}_n \pm t_{n-1,1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] ; \\ & \left[-\infty, \bar{X}_n + t_{n-1,1-\alpha} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] ; \\ \text{et} \quad & \left[\bar{X}_n - t_{n-1,1-\alpha} \sqrt{\frac{\hat{\sigma}_n^2}{n}}, +\infty \right] . \end{aligned}$$

Remarque 7.3. On rappelle qu'on peut voir (par exemple sur les tables) que, à $\beta \geq 0.5$ 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. C'est une question de robustesse. (On retrouve ces considérations sur la figure 4.1.) Par précaution, SPSS utilise les quantiles des lois de Student quelle que soit la valeur de la taille d'échantillon n .

2. Interprétation et exploitation des formules d'intervalles de confiance

La partie précédente correspond (au moins en partie) à ce que vous avez vu en classes préparatoires à propos des intervalles de confiance : des formules théoriques. Vous manquez de mise en pratique et en particulier, ne vous êtes jamais interrogé sur la forme d'intervalle de confiance à adopter (majorant, minorant, encadrant) et devez apprendre la rédaction idoine.

2.1. Application des formules : résolution d'un exercice-type

Exercice 7.1 (Collectif Jeudi Noir; ou autres associations de lobbyistes du logement). 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. Sélectionnez 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 m² sont souvent présentés dans les annonces avec un loyer de 750 euros (soit 30 euros du mètre carré)?

Pour la résolution de cet exercice, détaillée ci-dessous, il «suffit» d'appliquer les bonnes formules, en l'occurrence ici, celles du corollaire 7.1. Ici, 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. Attention! on prendra également garde au fait 3. que les énoncés procurent la plupart du temps l'estimée de l'écart-type et non celle de la variance.

- 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 7.1 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. Que veut faire le commanditaire? Il s'agit de cerner ses intentions.
- Pour 3., cela signifie que souvent les énoncés donnent la valeur réalisée de $\sqrt{\hat{\sigma}_n^2}$ et qu'il faut prendre garde à ne pas passer une fois de trop à la racine carrée...

Voici une correction de cet exercice :

Modélisation : La population est l'ensemble des logements locatifs de Paris. On dispose de 298 données de prix de location au m², 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 m² sur l'ensemble des logements locatifs de Paris. On dispose de l'estimée $\bar{x}_{298} = 18.4 \text{ €}$. Par ailleurs, l'écart-type

d'échantillon est $s_{298} = 3.2 \text{ €}$.

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 estimation du prix moyen, assortie d'une marge d'erreur.

L'intervalle théorique à 95% est $\left[\bar{x}_{298} \pm z_{97.5\%} \sqrt{\frac{\hat{\sigma}_{298}^2}{298}} \right]$ 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 auraient été donnés.)

- Le collectif Jeudi Noir dénonce les loyers trop chers, il veut une minoration du prix moyen au m^2 pour montrer que les prix sont déraisonnables : intervalle théorique à 95% :

$$\hat{I}_{298} = \left[\bar{x}_{298} - z_{95\%} \sqrt{\frac{\hat{\sigma}_{298}^2}{298}}, +\infty \right]$$

de valeur réalisée (après calculs) : $[18.1, +\infty[$.

- Les propriétaires - bailleurs veulent montrer que les prix sont tout à fait raisonnables :

$$\hat{J}_{298} = \left[0, \bar{x}_{298} + z_{95\%} \sqrt{\frac{\hat{\sigma}_{298}^2}{298}} \right]$$

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 grands surfaces). Les 30 €/m^2 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 petits 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 biens 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.

2.2. Point sur le vocabulaire

Confiance versus risque. Attention ! Le vocabulaire consacré est optimiste : on parle de niveau de confiance (par exemple, à 95 %) et pas de niveau de risque (à 5 % dans ce cas). Dans le monde de l'entreprise, il est résolument meilleur de dire «L'étude a démontré, avec un niveau de confiance de 95 %, que ...» plutôt que de se focaliser sur les échecs possibles et de mettre en garde son auditoire en disant «L'étude a démontré, avec un risque de 5 %, que ...» Evidemment, du point de vue mathématique, les deux formulations sont identiques, mais l'impact psychologique n'est pas le même ! Nous n'utiliserons que la première formulation (et là encore, c'est un changement par rapport à certaines des mauvaises habitudes que l'on a pu vous enseigner en classes préparatoires).

Confiance versus probabilité. N'oubliez jamais la remarque 1.6 formulée dans le premier chapitre : il y a une probabilité 0 ou 1 que le vrai prix au mètre carré μ_0 (tel que mesuré des années plus tard, lorsque l'observatoire des loyers parisiens créé dans le cadre de la loi ALUR sera en régime stationnaire de fonctionnement) soit dans chacun des intervalles de confiance proposés, par exemple $[0, 18.7]$. En revanche, il y a une probabilité (environ¹) $0.95 = 95\%$ que μ_0 soit dans l'intervalle théorique

$$\left[0, \bar{X}_{298} + z_{95\%} \sqrt{\frac{\hat{\sigma}_{298}^2}{298}} \right].$$

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

Par conséquent, on écrira : «l'étude montre qu'avec confiance 95%, le prix au mètre carré de la location mensuelle est inférieur à 18.7 euros» (et non pas «avec probabilité»).

2.3. A retenir pour votre vie en entreprise

Méditez 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)

Désormais, vous ne présenterez plus vos résultats comme avant, et saurez qu'il est souvent ridicule de présenter des pourcentages avec de nombreuses décimales, que cela marque un manque de recul profond et vous décrédibiliserait² 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.

2.4. Sondages et vie politique

Finissons sur quelques considérations citoyennes.

Remarque 7.4 (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

1. La probabilité est 95 % à la limite quand $n \rightarrow \infty$; ici, n est suffisamment grand, $n = 298$, de sorte que l'on est proche de 95 %.

2. 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 !

entendu qu'elle était autour de 3 % : retrouvons³ 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 7.2 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.

Cependant, les sondeurs fournissent désormais aux journalistes des grilles de lecture des résultats des sondages, comme illustré à la figure 7.2 et même encore plus précisément, à la figure 7.3 (notez le vocabulaire incorrect «chance» au lieu de «confiance»!).

Remarque 7.5 (Chirac, Le Pen et Jospin ; ou de l'usage pas assez systématique des intervalles de confiance). Les sondeurs précisent donc des marges d'erreur... 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 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 (nous allons voir qu'elles sont 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 systématiquement le droit à une information complète avec marges d'erreur ? Les journalistes de la presse écrite la donnent de plus en plus, mais ils ne le font pas encore systématiquement ; quant à ceux des journaux télévisuels, ils jugent que cela embrouillerait le Français moyen...

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

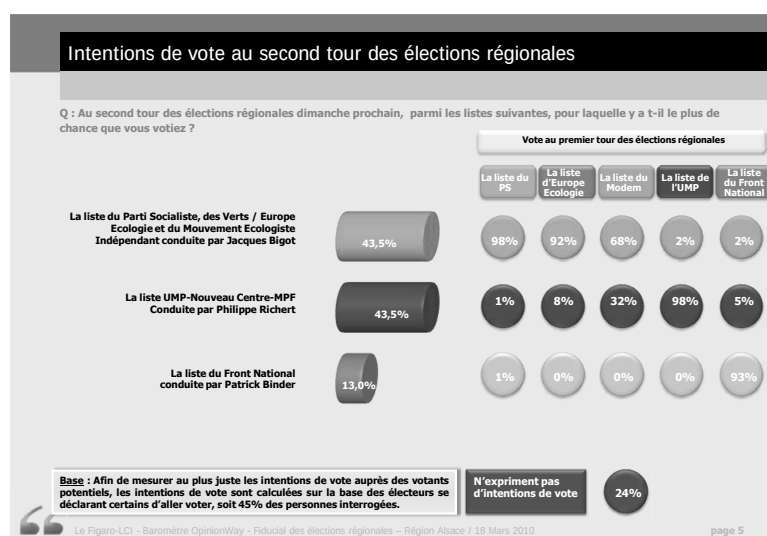
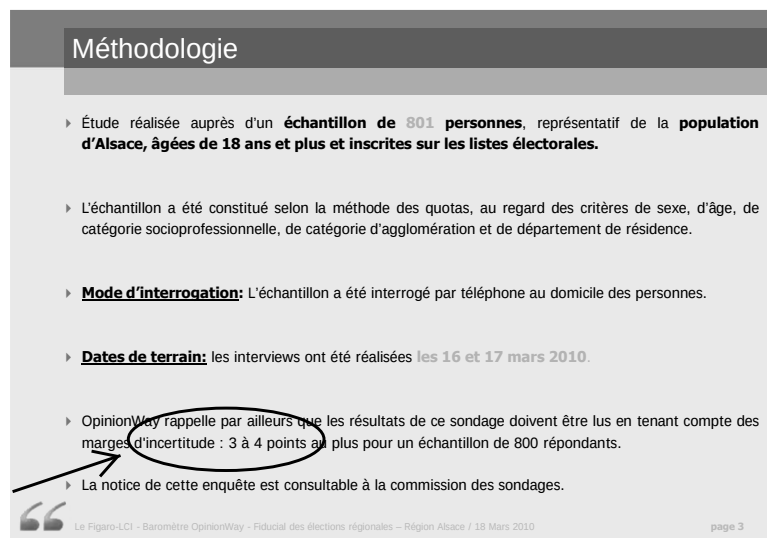


FIGURE 7.2 – Un exemple de résultat de sondage fourni par un institut avec indication de la marge d'erreur (d'incertitude) possible.

Méthodologie

Ce document présente les résultats d'une étude réalisée par l'Ifop. Elle respecte fidèlement les principes scientifiques et déontologiques de l'enquête par sondage. Les enseignements qu'elle indique reflètent un état de l'opinion à l'instant de sa réalisation et non pas une prédiction.

Aucune publication totale ou partielle ne peut être faite sans l'accord exprès de l'Ifop.

Retrouvez les sondages et analyses de l'Ifop sur :



iPhone

iPad

www.ifop.com

@ifopinion

Ifop Opinion

Etude réalisée par l'Ifop pour :	La Manif Pour Tous
Echantillon	Echantillon de 1005 personnes, représentatif de la population française âgée de 18 ans et plus. La représentativité de l'échantillon a été assurée par la méthode des quotas (sexe, âge, profession de la personne interrogée) après stratification par région et catégorie d'agglomération.
Mode de recueil	Les interviews ont eu lieu par questionnaire auto-administré en ligne (CAWI - Computer Assisted Web Interviewing).
Dates de terrain	Du 11 au 13 décembre au 2012

PRÉCISION RELATIVE AUX MARGES D'ERREUR

La théorie statistique permet de mesurer l'incertitude à attacher à chaque résultat d'une enquête. Cette incertitude s'exprime par un intervalle de confiance situé de part et d'autre de la valeur observée et dans lequel la vraie valeur a une probabilité déterminée de se trouver. Cette incertitude, communément appelée « marge d'erreur », varie en fonction de la taille de l'échantillon et du pourcentage observé comme le montre le tableau ci-dessous :

INTERVALLE DE CONFIANCE A 95% DE CHANCE										
Et si l'effectif est...	Si le pourcentage trouvé est...									
	5 ou 95%	10 ou 90%	20 ou 80%	30 ou 70%	40 ou 60%	50%				
50	6,2	8,5	11,3	13,0	13,9	14,1				
100	4,4	6,0	8,0	9,2	9,8	10,0				
200	3,1	4,2	5,7	6,5	6,9	7,1				
250	2,8	3,8	5,1	5,8	6,2	6,3				
300	2,5	3,5	4,6	5,3	5,7	5,8				
350	2,3	3,2	4,3	4,9	5,2	5,3				
400	2,2	3,0	4,0	4,6	4,9	5,0				
450	2,1	2,8	3,8	4,5	4,8	4,9				
500	2,1	2,7	3,6	4,3	4,6	4,7				
600	1,8	2,4	3,3	3,7	4,0	4,1				
700	1,6	2,3	3,0	3,5	3,7	3,8				
800	1,5	2,1	2,8	3,2	3,5	3,5				
900	1,4	2,0	2,6	3,0	3,2	3,3				
1000	1,4	1,8	2,5	2,8	3,0	3,1				
2000	1,0	1,3	1,8	2,1	2,2	2,2				
4000	0,7	0,9	1,3	1,5	1,6	1,6				
6000	0,6	0,8	1,1	1,3	1,4	1,4				
10000	0,4	0,6	0,8	0,9	0,9	1,0				

Exemple de lecture du tableau : dans le cas d'un échantillon de 1000 personnes, si le pourcentage mesuré est de 10%, la marge d'erreur est égale à 1,8. Le vrai pourcentage est donc compris entre 8,2% et 11,8%.

FIGURE 7.3 – Un exemple de grille de lecture de résultats de sondage fourni par un institut : tableau résumant les marges d'erreur (d'incertitude).

3. Nouveauté — Intervalles de confiance simultanés

A retenir : L'exploitation simultanée d'un intervalle de confiance de niveau $1 - \alpha$ et d'un intervalle de confiance $1 - \beta$ vaut avec confiance seulement $1 - (\alpha + \beta)$. Il y a une correction de niveau à faire pour la considération de plusieurs intervalles de confiance (les risques d'erreur s'ajoutent). On l'appelle la correction de Bonferroni.

Exemples d'application. Supposons que nous ayons deux intervalles de confiance pour deux paramètres, disons $[30\%, 45\%]$ pour p_0 et $[20\%, 25\%]$ pour q_0 , tous deux de niveau 95%. Pour chaque quantité qui pourrait nous intéresser ($p_0 - q_0$, ou $(p_0 + q_0)/2$, ou encore p_0/q_0), on regarde les intervalles obtenus en exploitant les deux intervalles de départ et on applique les règles habituelles de manipulation des inégalités pour identifier les bornes de l'intervalle à calculer. Par exemple :

- La quantité $p_0 - q_0$ est estimée par l'intervalle $[30\% - 25\%, 45\% - 20\%] = [5\%, 25\%]$. On «croise» ici les bornes, à cause du signe $-$ devant q_0 .
- La quantité $(p_0 + q_0)/2$ est estimée par l'intervalle $[(30\% + 20\%)/2, (45\% + 25\%)/2] = [25\%, 35\%]$. Pas besoin de «croiser» ici, on calcule séparément la moyenne des bornes inférieures et supérieures.
- La quantité p_0/q_0 est estimée par l'intervalle $[30\%/25\%, 45\%/20\%] = [1.2, 2.25]$. Ici, on «croise» à nouveau.

Tout cela ne sont que des exemples, il faut raisonner à chaque fois dans le cas d'espèce de l'exercice pour voir s'il y a lieu de croiser ou pas les bornes, ou même de faire des manipulations plus compliquées.

Détails mathématiques de l'obtention de la correction. Ici, il s'agit d'estimer par intervalle deux paramètres (ou plus) à la fois, disons μ_0 et μ'_0 . On propose deux intervalles $\hat{I}$ et $\hat{J}$ tels que, par exemple,

$$\mathbb{P}\{\mu_0 \in \hat{I}\} = 1 - \alpha \quad \text{et} \quad \mathbb{P}\{\mu'_0 \in \hat{J}\} = 1 - \beta$$

et on se demande ce que l'on peut garantir sur

$$\mathbb{P}\{\mu_0 \in \hat{I} \text{ et } \mu'_0 \in \hat{J}\}.$$

L'intérêt est alors qu'on a un intervalle de confiance sur toute quantité de la forme $g(\mu_0, \mu'_0)$, avec g continue : il est donné par

$$\left[\min_{\hat{I} \times \hat{J}} g, \max_{\hat{I} \times \hat{J}} g \right],$$

car en effet,

$$\mathbb{P}\left\{g(\mu_0, \mu'_0) \in \left[\min_{\hat{I} \times \hat{J}} g, \max_{\hat{I} \times \hat{J}} g \right]\right\} \geq \mathbb{P}\{\mu_0 \in \hat{I} \text{ et } \mu'_0 \in \hat{J}\}.$$

La plupart du temps, $g(x, y) = xy$, ou $g(x, y) = x + y$, ou encore $g(x, y) = x - y$.

A priori, 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. Elle est à oublier ! (Nous y reviendrons.)

3.1. Méthode de Bonferroni

En général, on a utilisé les mêmes individus (le même échantillon de personnes ou d'objets) pour construire les deux intervalles $\hat{I}$ et $\hat{J}$: ils ne sont donc pas du tout indépendants et on ne peut pas

écrire

$$\mathbb{P}\{\mu_0 \in \hat{I} \text{ et } \mu'_0 \in \hat{J}\}$$

comme un produit d'événements. On procède alors ainsi (ce qui suit s'appelle la méthode de Bonferroni) :

$$\mathbb{P}\{\mu_0 \in \hat{I} \text{ et } \mu'_0 \in \hat{J}\} = 1 - \mathbb{P}\{\mu_0 \notin \hat{I} \text{ ou } \mu'_0 \notin \hat{J}\} \geq 1 - (\mathbb{P}\{\mu_0 \notin \hat{I}\} + \mathbb{P}\{\mu'_0 \notin \hat{J}\}) = 1 - (\alpha + \beta).$$

Remarque 7.6. En pratique, on prendra souvent des niveaux $\alpha = \beta = 2.5\%$ pour $\hat{I}$ et $\hat{J}$, 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.

A retenir (bis) : L'exploitation simultanée d'un intervalle de confiance de niveau $1 - \alpha$ et d'un intervalle de confiance $1 - \beta$ vaut avec confiance seulement $1 - (\alpha + \beta)$. Il y a une correction de niveau à faire pour la considération de plusieurs intervalles de confiance (les risques d'erreur s'ajoutant).

3.2. Cas à oublier : indépendance

Quand bien même $\hat{I}$ et $\hat{J}$ seraient construits sur des échantillons indépendants (pas sur les mêmes individus ou objets) et que l'on pourrait écrire, par indépendance d'événements

$$\mathbb{P}\{\mu_0 \in \hat{I} \text{ et } \mu'_0 \in \hat{J}\} = \mathbb{P}\{\mu_0 \in \hat{I}\} \mathbb{P}\{\mu'_0 \in \hat{J}\} = (1 - \alpha)(1 - \beta) = 1 - (\alpha + \beta) + \alpha\beta,$$

on voit que le niveau garanti sous cette hypothèse d'indépendance serait certes un peu supérieur au niveau pouvant être garanti d'une manière générale par la méthode de Bonferroni, mais d'un facteur $\alpha\beta$ absolument négligeable. En effet, lorsque $\alpha = \beta = 5\%$ par exemple, il vient $\alpha\beta = 0.25\%$!

4. Nouveauté — 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 l'écart-type ; lorsque l'intervalle de confiance était bilatère, sa demi-largeur indiquait la précision ε de l'estimation.

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 ?

Parfois seulement on ne se pose pas cette question et on recueille autant d'observations qu'on le peut, mais dans de nombreux cas, on veut réfléchir à la taille d'échantillon à l'avance, notamment pour organiser le travail de collecte. Cette situation est illustrée par la figure 7.4.

4.1. Planification : cas de l'estimation d'une proportion

On reprend les notations du paragraphe 1.3. Dans ce cas, on a vu que les intervalles de confiance (bilatères) sont de la forme

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

La précision dépend donc des observations à venir via $\bar{X}_n$! On va majorer cette précision de la même manière qu'à la remarque 7.4 et on choisira n tel que ce majorant soit plus petit que ε . Toujours en utilisant que $x(1-x) \leq 1/4$, il vient

$$\hat{I}_n \subset \hat{J}_n \stackrel{\text{not.}}{=} \left[\bar{X}_n \pm z_{1-\alpha/2} \frac{1}{2\sqrt{n}} \right] .$$

$\hat{J}_n$ est également un intervalle de confiance, asymptotiquement de niveau $1 - \alpha$; il est plus large que $\hat{I}_n$, mais sa précision (sa demi-largeur) est indépendante des observations. Il suffit de prendre, à α et ε fixés, n suffisamment grand pour que

$$z_{1-\alpha/2} \frac{1}{2\sqrt{n}} \leq \varepsilon , \quad \text{soit} \quad n \geq \left(\frac{z_{1-\alpha/2}}{2\varepsilon} \right)^2 .$$

En particulier, lorsque $\alpha = 5\%$, il s'agit de prendre $n \geq 1/\varepsilon^2$.

Remarque 7.7 (Planification vs. exploitation d'une expérience). Evidemment, après avoir conduit l'expérience, on peut délaisser l'intervalle $\hat{J}_n$ pour $\hat{I}_n$: la précision sera alors d'autant meilleure que $\bar{X}_n$ est éloigné de $1/2$. On retiendra pour la suite que pour planifier, on considère des intervalles du type de $\hat{J}_n$, mais que si l'on dispose des données, on se tourne vers ceux du type de $\hat{I}_n$. L'exercice suivant illustre cela.

Exercice 7.2 (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 commande 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 moyenne qu'on aurait eue si on avait pu interroger tous les habitants un par un ?

OUR METHOD IS TO TAKE A **SAMPLE**... A RELATIVELY SMALL SUBSET OF THE TOTAL POPULATION, THE WAY POLLSTERS DO AT ELECTION TIME.



AN OBVIOUS QUESTION IS: HOW BIG A SAMPLE DO WE HAVE TO TAKE TO GET MEANINGFUL RESULTS?



AND THE ANSWER, WHICH YOU SHOULD INSCRIBE IN YOUR BRAIN FOREVERMORE, WILL TURN OUT TO BE: IF n IS THE NUMBER OF ITEMS IN THE SAMPLE, THEN EVERYTHING IS GOVERNED BY

$$\frac{1}{\sqrt{n}}.$$

GOVERNED BY $\frac{1}{\sqrt{n}}$? DIDN'T EVEN KNOW IT WAS ON THE BALLOT!



FIGURE 7.4 – Planification de sondages et mauvais jeux de mots associés.

Correction 7.1. La précision est de $\varepsilon = 0.02$ et si l'on conserve le niveau asymptotique habituel de 95 %, il s'agit de récupérer

$$n \geq \frac{1}{\varepsilon^2} = \frac{1}{0.02^2} = 2500$$

réponses, selon les calculs précédents. Evidemment, cela obligera à interroger autour de $n' = 3\,000$ habitants (foyers), à cause de ceux qui refusent de répondre. 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 $x_1, \dots, x_{2500} \in \{0, 1\}$, telles que

$$\bar{x}_{2500} = \frac{355}{2\,500} = 14.2\% .$$

On les modélise comme la réalisation de $X_1, \dots, X_{2500}$ 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

$$\left[\bar{x}_{2500} \pm 1.96 \sqrt{\frac{\bar{x}_{2500}(1 - \bar{x}_{2500})}{2500}} \right]$$

et sa réalisation sur les données vaut

$$\left[14.2\% \pm 1.96 \sqrt{\frac{\bar{x}_{2500}(1 - \bar{x}_{2500})}{2500}} \right] = \left[14.2\% \pm 1.96 \sqrt{\frac{0.142(1 - 0.142)}{2500}} \right] = [14.2\% \pm 1.4\%] .$$

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 7.3 (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 7.2 (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 $x_1, \dots, x_{1000}$ ici), on distingue alors deux cas :

- si la moyenne d'échantillon $\bar{x}_{1000}$ 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 7.4) et donc ne contiendra pas le seuil de 10 % ; on conclura à la rentabilité ;

- sinon, on a $\bar{x}_{1000} \leq 13\%$ et en particulier, $x \mapsto x(1-x)$ étant croissante sur $[0, 1/2]$, la précision est majorée par

$$1.96 \sqrt{\frac{\bar{x}_{1000}(1 - \bar{x}_{1000})}{1000}} \leq 1.96 \sqrt{\frac{0.13(1 - 0.13)}{1000}} \approx 1.1\% .$$

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.

4.2. 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 $x(1-x) \leq 1/4$, et on pouvait majorer par avance la précision indépendamment des observations,

$$z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \leq z_{1-\alpha/2} \frac{1}{2\sqrt{n}} .$$

Dans le cas général, il s'agirait de majorer, à l'avance, avant de conduire l'expérience,

$$z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} ,$$

or, on n'a aucune idée de la valeur qui va être réalisée pour $\hat{\sigma}_n^2$! 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,

$$z_{1-\alpha/2} \sqrt{\frac{s_t^2}{n}} \leq \varepsilon , \quad \text{soit} \quad n \geq s_t^2 \left(\frac{z_{1-\alpha/2}}{\varepsilon} \right)^2 .$$

n donne alors la taille du nouvel échantillon de données à recueillir, notées $x_{t+1}, \dots, x_{t+n}$. 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 7.4. 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'affaire. 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 dégagant un espace plus calme dans le marché et en embauchant sept enquêteurs pendant une demi-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 7.3. 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 :

$$\left[\bar{x}_{30} \pm z_{97.5\%} \sqrt{\frac{s_{x,30}^2}{30}} \right] = \left[47.32 \pm 1.96 \frac{12.05}{\sqrt{30}} \right] = [47.32 \pm 4.31] \approx [43.0, 51.7] .$$

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 $n \geq (1.96 \times 12.05)^2$, 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

$$\left[\bar{x}_{558} \pm z_{97.5\%} \sqrt{\frac{s_{x,558}^2}{558}} \right] = \left[48.51 \pm 1.96 \frac{11.37}{\sqrt{558}} \right] = [48.51 \pm 0.94] \approx [47.5, 49.5] .$$

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...).

Régression linéaire simple

Les chapitres précédents ont présenté les notions d'intervalle de confiance et de test, et en ont donné divers exemples. 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 prévoir une variable quantitative comme fonction affine d'une autre variable quantitative. (Le chapitre suivant présentera le cas plus général de l'explication ou de la prévision 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 $(x_1, y_1), \dots, (x_n, y_n)$. 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 français¹ des y_j qu'elles forment la variable à expliquer et des x_j qu'elles sont la variable explicative.

Remarque 8.1 (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 page 228, qui étudie l'existence d'une relation entre maladie mentale et nombre de téléviseurs (de même qu'avec l'article qui suit son énoncé, intitulé : «Le chocolat engendre-t-il des tueurs en série ?»).

Exemple 8.1 (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 8.1 (datant du début des années 2000, les prix font rêver !). On dispose donc ici de 28 couples

1. 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 m ² , tt confort. Visite vendredi, samedi, dim. 130 000 € à discuter. Facilités	2. CONTRESCARPE, imm. Ancien, pierre de taille, beau duplex caractère, 50 m ² , poutres, refait neuf, 280 000 €
3. R. St-Simon, en pleine verdure, calme, plein soleil, Superbe appt 4p., 106 m ² , cuis. aménagée, s. de bains moderne, chff. cent. Parfait état. Px 650 000 à discuter. Agence s'abstenir. Direct. Propriétaire.	4. RAPP 7P., 196 m ² standing, 9 fenêtres plein soleil, 800 000 €
5. R. St André-des-Arts, beau liv + chbre, imm. XVIIIe siècle, 55 m ² , 268 000 €	6. 5 ^e PRES QUAIS, 7 pces, 190 m ² caractère, standing, 790 000 €
7. GOBELINS, Beau 5p., 110m ² , gd cft, soleil, 500 000 €	8. GOBELINS, et. élevé, calme, asc., 2 pièces, 60 m ² , 320 000 €
9. CENSIER, très grand studio + entrée 48 m ² , tt cft, ensoleillé, calme, bel imm., 250 000 €	10. PANTHÉON, 7 ^e étage, ascenseur, grand studio 35 m ² + terrasse. Vue. 250 000 €
11. RUE MADAME, 3P. + Serv., 86 m ² , 350 000 €	12. RUE DE SEINE, 3P., tt cft, 65 m ² , calme, soleil, 300 000 €
13. PANTHEON, bel imm., verdure, magnifique studio 32 m ² , caractère, 155 000 €	14. SEVRES BAB, 1 ^{er} ét., 2P., gde cuis., bns, 52 m ² , état neuf, 245 000 €
15. MONTPARNASSE, Part. vend atelier d'artiste 40 m ² , duplex, vue imprenable, tout confort, Prix 200 000 €	16. RUE D'ASSAS, imm. gd standing, bel appart 260 m ² , triple récept. + 5 ch., tt cft (travaux) 2 park., 2 ch. Serv., Prix 1 500 000 € à déb.
17. BD St-GERMAIN, 4P., 70 m ² , à amén., 4 ^e ét., 325 000 €	18. ÎLE St-LOUIS, Lux. appt., 117 m ² , en duplex, gde récept., gde chambre, 2 sdb, Terras., parf. et., décor tr. bon goût, 950 000 €
19. JUSSIEU, Charme, gd 3 pces, 90 m ² , 378 000 €	20. QUARTIER LATIN, 30m ² à aménager, prix 78 000 €
21. MONTPARNASSE, Imm. p.d.t., 4-5 P., 105 m ² , bon état, 375 000 €	22. RUE MAZARINE, 4 ^e ét., sans ascens., 52 m ² à rénover. Prix total 200 000 €
23. CENSIER, Bel imm., 4P. 80 m ² , tt cft, petits travaux, 270 000 €	24. ASSAS LUXEMBOURG, 3P. 60 m ² s/arbres, imm. caractère, 295 000 €
25. SUR JARDINS OBSERVATOIRE, 140 m ² , grand charme, 990 000 €	26. RUE DE SAVOIE, 4 ^e ét., Studio 20 m ² , dche, 85 000 €. crédit possible
27. PRES LUXEMBOURG, Bel imm., pierre de taille, Appartement 100 m ² , salon, sal. à manger, 2 chbres, office, cuis., bains, chf. cent., asc., prix : 495 000 €	28. Mo GOBELINS, studio, cuis., s. de bains, 28m ² , calme. Prix 85.000 €

FIGURE 8.1 – Liste de 28 appartements à vendre.

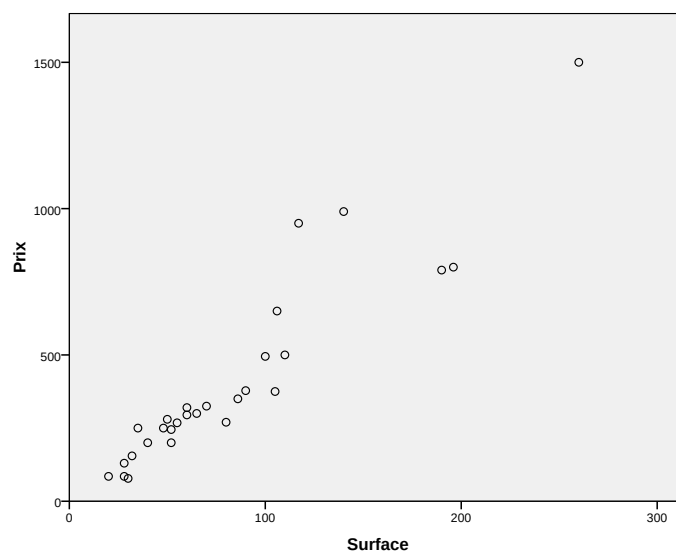


FIGURE 8.2 – Représentation graphique des données de prix et surfaces d'appartements.

(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 8.2. Au moins pour les surfaces inférieures à 100 mètres carrés, la relation entre prix et surface semble raisonnablement linéaire.

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 :

$$y_j = f(x_j) + e_j, \quad \text{pour tout } j = 1, \dots, n.$$

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 8.2 (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

$$y_j = a'_0 + b'_0 x_j^2 + e'_j \quad \text{ou} \quad y_j = a''_0 + b''_0 \ln x_j + e''_j.$$

C'est sur les figures de répartition dans le plan, comme la figure 8.2, que l'on devine le bon type de relation. Cela demande 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 8.3 (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 8.2 (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 française « régression », désormais synonyme de relation en statistique.

Mathématiquement, on s'attend à une relation du type $y_j = m_0 + b(x_j - m_0) + e_j$, 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. Cet exemple est traité plus en détails comme exercice de quizz : voir page 221.

Profitions-en, justement, pour décrire les objectifs et motivations situés derrière le concept de régression linéaire.

1.3. Objectifs et motivations

Dans ce chapitre 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 8.1), entre le prix d'un forfait de ski et la taille du domaine skiable correspondant (voir exercice page 226), entre le taux de maladies mentales dans la population anglaise et le nombre de téléviseurs (voir exercice 228) ?
- 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 : Préviation 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 préviation 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 : par moindres carrés

On considère la fonction

$$F_n : (\alpha, \beta) \in \mathbb{R}^2 \mapsto \sum_{j=1}^n (y_j - (\alpha + \beta x_j))^2$$

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 petits², 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 :

$$(a, b) \in \operatorname{argmin}_{(\alpha, \beta) \in \mathbb{R}^2} F_n(\alpha, \beta) = \operatorname{argmin}_{(\alpha, \beta) \in \mathbb{R}^2} \sum_{j=1}^n (y_j - (\alpha + \beta x_j))^2 .$$

Les logiciels de statistique, comme SPSS, calculent pour nous un tel couple (a, b) — voir la figure 8.3 — et tracent également la droite de régression, qui est l'ensemble des couples $(x, a + bx)$.

Cette droite passe toujours par le point $(\bar{x}_n, \bar{y}_n)$; c'est un fait qui découle des éléments culturels indiqués ci-dessous. Elle passe également par les points notés $\hat{y}_j = a + bx_j$ pour $j = 1, \dots, n$. Les $\hat{y}_j$ 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

$$e_j = y_j - \hat{y}_j = y_j - (a + bx_j) , \quad j = 1, \dots, n,$$

quantifient la qualité du modèle linéaire.

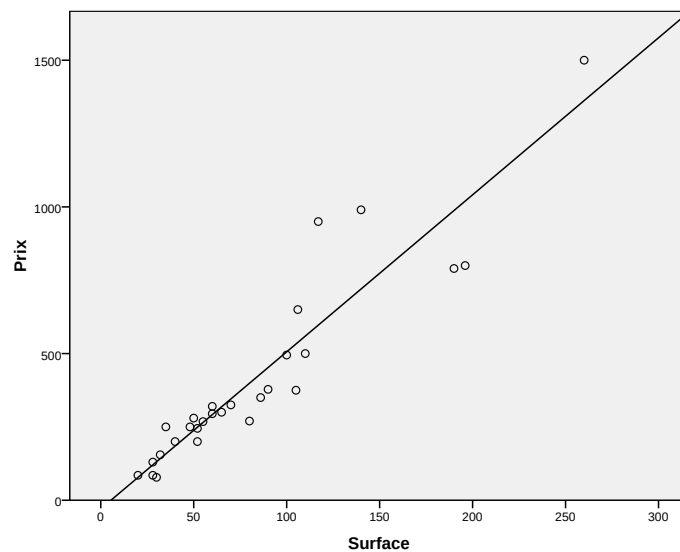
La minute SPSS 8.1. On obtient la figure 8.3 de la manière suivante. Premièrement, on trace avec le Générateur de diagrammes la représentation dans le plan déjà considérée à la figure 8.2, puis on double-clique dessus, on sélectionne les points et on clique sur le bouton Ajouter une courbe d'ajustement. Ensuite, on fait calculer le tableau d'analyse de la régression (nous en présentons une version partiellement masquée) : Analyse / Régression / Linéaire. On y lit les valeurs de a et b , soit, respectivement, -29.466 et 5.353 .

ÉLÉMENTS 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.

2. On parle donc de moindres carrés. Au fait, pourquoi ce critère, et pas un autre, comme les moindres valeurs absolues par exemple ? Tout simplement parce que ce critère donne lieu à des formules explicites, que l'on a donc pu utiliser bien avant l'avènement de l'informatique. De nos jours, on peut calculer numériquement les solutions à des problèmes n'admettant pas solution générale explicite, comme les moindres valeurs absolues ; mais l'Histoire joue un grand poids ici, dans la détermination du critère retenu !



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 8.3 – Droite de régression pour les 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.

Définition 8.1. On définit les quantités empiriques suivantes,

$$\text{Var}(x_1^n) = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n)^2 \quad \text{et} \quad \text{Cov}(x_1^n, y_1^n) = \frac{1}{n} \sum_{j=1}^n (y_j - \bar{y}_n)(x_j - \bar{x}_n),$$

appelées respectivement variance et covariance d'échantillon.

Alors les coefficients a et b donnés par la méthode des moindres carrés sont égaux à

$$b = \frac{\text{Cov}(x_1^n, y_1^n)}{\text{Var}(x_1^n)} \quad \text{et} \quad a = \bar{y}_n - b\bar{x}_n.$$

En particulier, la droite de régression est l'ensemble des points

$$\mathcal{D} = \{(x, a + bx), x \in \mathbb{R}\} = \left\{ \left(x, \bar{y}_n + \frac{\text{Cov}(x_1^n, y_1^n)}{\text{Var}(x_1^n)}(x - \bar{x}_n) \right), x \in \mathbb{R} \right\}.$$

Elle passe bien par les points $(x_j, \hat{y}_j)$. Un calcul simple montre par ailleurs que

$$\sum_{j=1}^n e_j = \sum_{j=1}^n (y_j - (a + bx_j)) = 0,$$

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 8.3. On note en particulier que l'égalité à 0 ci-dessus montre que la

moyenne des valeurs reconstruites est la moyenne des valeurs originelles :

$$\frac{1}{n} \sum_{j=1}^n \hat{y}_j = \bar{y}_n.$$

Note : on comprend bien 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 le chapitre suivant 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 8.1. *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 ,*

$$\underbrace{\sum_{j=1}^n (y_j - \bar{y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_T} = \underbrace{\sum_{j=1}^n (\hat{y}_j - \bar{y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_E} + \underbrace{\sum_{j=1}^n (y_j - \hat{y}_j)^2}_{\stackrel{\text{not.}}{=} \Sigma_R}.$$

Interprétation des différents termes :

- Σ_T mesure la variabilité des données à expliquer y_j autour de leur moyenne $\bar{y}_n$; 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 $\hat{y}_j$ autour de leur moyenne, dont les éléments culturels ci-dessus ont montré qu'elle valait également $\bar{y}_n$; 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 8.2. *Le coefficient de détermination r^2 est la fraction de la variabilité totale retrouvée par la régression,*

$$r^2 = \frac{\Sigma_E}{\Sigma_T}.$$

Ainsi $0 \leq r^2 \leq 1$. 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. 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 SPSS 8.2. Avec toujours Analyse / Régression / Linéaire, mais en cochant la case Qualité de l'ajustement dans Statistiques, on peut obtenir la figure 8.4 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 .

Ici par exemple, la valeur réalisée pour r^2 est 86.5 %. On traduira à l'homme de la rue (de manière un peu abusive³ mais simple à comprendre) : la (variabilité de la) surface explique 86.5 % (de la variabilité) du prix de l'appartement.

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	...	...	...	...
Résidu	392963,209	...	...	...	...
Total	2920170,714	...	...	...	...

a. Valeurs prédites : (constantes), Surface

b. Variable dépendante : Prix

FIGURE 8.4 – Mesures de la qualité de l'ajustement linéaire (certaines quantités non encore étudiées à ce point du cours ont été masquées).

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.

3. 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)

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j, \quad j = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont elles des variables aléatoires indépendantes et identiquement distribuées 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 frustré d'une modélisation linéaire déterministe. Par théorème limite central 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 8.2 semble indiquer par exemple que cette hypothèse peut être tenue pour vraie, mais uniquement pour les surfaces plus petites que 120 m².

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

On avait auparavant 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 $\hat{\alpha}_n$ et $\hat{\beta}_n$ 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 $\hat{\alpha}_n$ et $\hat{\beta}_n$.

Définition 8.3. On appelle estimateurs des moindres carrés de (α_0, β_0) le couple $(\hat{\alpha}_n, \hat{\beta}_n)$ tel que

$$(\hat{\alpha}_n, \hat{\beta}_n) \in \underset{(\alpha, \beta) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{j=1}^n (Y_j - (\alpha + \beta x_j))^2.$$

Les résidus stochastiques reconstruits

$$\hat{\varepsilon}_j = Y_j - \hat{Y}_j = Y_j - (\hat{\alpha}_n + \hat{\beta}_n x_j), \quad j = 1, \dots, n,$$

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 $\hat{\alpha}_n$ et $\hat{\beta}_n$.

Théorème 8.2. *Dans le modèle linéaire gaussien, les estimateurs $\hat{\alpha}_n$ et $\hat{\beta}_n$ suivent les lois normales*

$$\hat{\beta}_n \sim \mathcal{N}\left(\beta_0, \frac{\sigma_0^2}{n \operatorname{Var}(x_1^n)}\right) \quad \text{et} \quad \hat{\alpha}_n \sim \mathcal{N}\left(\alpha_0, \frac{\sigma_0^2}{n} \left(1 + \frac{(\bar{x}_n)^2}{\operatorname{Var}(x_1^n)}\right)\right).$$

Ils sont en particulier sans biais.

ÉLÉMENTS CULTURELS. Un précédent encart d'éléments culturels fournissait les expressions explicites de $\hat{\alpha}_n$ et $\hat{\beta}_n$: 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 $\hat{\alpha}_n$ et $\hat{\beta}_n$ 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 $\hat{\beta}_n$, que

$$\begin{aligned} \mathbb{E}[\operatorname{Cov}(x_1^n, Y_1^n)] &= \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n) \mathbb{E}[Y_j - \bar{Y}_n] \\ &= \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x}_n) ((\alpha_0 + \beta_0 x_j) - (\alpha_0 + \beta_0 \bar{x}_n)) = \beta_0 \operatorname{Var}(x_1^n). \end{aligned}$$

Puis, pour $\hat{\alpha}_n$,

$$\mathbb{E}[\hat{\alpha}_n] = \mathbb{E}[\bar{Y}_n] - \mathbb{E}[\hat{\beta}_n] \bar{x}_n = (\alpha_0 + \beta_0 \bar{x}_n) - \beta_0 \bar{x}_n = \alpha_0.$$

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 $\hat{\alpha}_n$ et $\hat{\beta}_n$ 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

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 $\hat{\varepsilon}_j$; 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 $n - 2$ degrés⁴ de liberté :

$$\frac{\Sigma_R}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{j=1}^n (\hat{\varepsilon}_j)^2 \sim \chi_{n-2}^2.$$

4. Cf. on a dû estimer deux paramètres, α_0 et β_0 , ce qui fait perdre 2 degrés de liberté ; c'est très «mathémagique» comme explication, j'en suis bien conscient...

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 $Z_k \sim \chi_k^2$, converge en probabilité vers 1).

Théorème 8.3. *Dans le modèle linéaire gaussien, on dispose de l'estimateur de la variance σ_0^2 donné par*

$$\hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{j=1}^n (\hat{\varepsilon}_j)^2 .$$

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, $\hat{\sigma}_n^2$ est indépendant de $\hat{\alpha}_n$ et $\hat{\beta}_n$.

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 8.2 montre que

$$\sqrt{n \operatorname{Var}(x_1^n)} \frac{\hat{\beta}_n - \beta_0}{\sigma_0} \sim \mathcal{N}(0, 1),$$

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 $\beta_{\text{ref}} = 0$:

$$\sqrt{n \operatorname{Var}(x_1^n)} \frac{\hat{\beta}_n - \beta_0}{\sqrt{\hat{\sigma}_n^2}} \sim \mathcal{T}_{n-2}.$$

Corollaire 8.1. *Un intervalle de confiance exact de niveau $1 - p$ sur β_0 est donné par exemple par*

$$\hat{I}_n = \left[\hat{\beta}_n \pm t_{n-2, 1-p/2} \sqrt{\frac{\hat{\sigma}_n^2}{n \operatorname{Var}(x_1^n)}} \right].$$

Principe 8.1. *Test de l'existence d'une relation linéaire (dans le cadre d'un modèle gaussien)*

Données : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$

Modélisation associée : valeurs $x_1, \dots, x_n$ 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 :

$$T_n = \sqrt{n \operatorname{Var}(x_1^n)} \frac{\hat{\beta}_n}{\sqrt{\hat{\sigma}_n^2}}$$

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 8.4 (Totalelement 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 8.2. En pratique, pour des raisons qui seront plus apparentes dans le chapitre 9, les logiciels statistiques considèrent souvent le carré de T_n :

$$\begin{aligned} D_n = (T_n)^2 &= \left(\sqrt{n \operatorname{Var}(x_1^n)} \frac{\hat{\beta}_n}{\sqrt{\hat{\sigma}_n^2}} \right)^2 = (n-2) \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y}_n)^2}{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2} \\ &= (n-2) \frac{\Sigma_E}{\Sigma_R} = (n-2) \frac{r^2}{1-r^2} = \frac{\Sigma_E}{\Sigma_R / (n-2)} . \end{aligned}$$

Cette statistique D_n suit une loi dite de Fisher (ici, à 1 et $n-2$ degrés de liberté). On généralisera au cours du chapitre 9, 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 $\bar{Y}_n$ des moindres carrés des Y_j sous l'hypothèse $H_0 : \beta_0 = 0$ à celles proposées sous H_1 , et qui sont les $\hat{Y}_j$.

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 avons 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 SPSS 8.3. Les tableaux de la figure 8.5 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 8.5 – 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	...	$\sqrt{\hat{\sigma}_n^2}$

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 significative.

Enfin, le troisième tableau donne les valeurs réalisées de :

Coeff.	Err. standard	...	t	Sig.	IC (bornes inf. et sup.)
$\hat{\alpha}_n$	$\sqrt{\frac{\hat{\sigma}_n^2}{n} \left(1 + \frac{(\bar{x}_n)^2}{\text{Var}(x_1^n)} \right)}$		T'_n	P-val. test $H_0 : \alpha_0 = 0$	cf. paragraphe 4.2
$\hat{\beta}_n$	$\sqrt{\frac{\hat{\sigma}_n^2}{n \text{Var}(x_1^n)}}$	...	T_n	P-val. test $H_0 : \beta_0 = 0$	cf. Corollaire 8.1

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 correctement⁵ 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.

Remarque 8.5 (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

5. 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!

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

$$\text{prix (en milliers d'euros)} = -29.466 + 5.353 \times \text{surface (en m}^2\text{)} + \text{aléa d'écart-type 122.939}$$

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 SPSS 8.4. 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. Intermède : justification de la (longue) liste des formules

Pourquoi nous vous avons-vous fait subir la longue litanie des formules absconses des pages précédentes, pourquoi n'apprenons-nous pas simplement à lire les tableaux fournis par SPSS (comme ceux de la figure 8.5) sans re-définir les quantités qui y sont présentes ? Tout simplement parce que nous ne pouvons pas préjuger du logiciel qui sera à votre disposition en entreprise ! Les figures 8.6 et 8.7 montrent par exemple les résumés de régression linéaire obtenus respectivement avec le logiciel R et Excel. Le lecteur attentif reconnaîtra tous les chiffres des tableaux de la figure 8.5 (et quelques autres relatifs notamment à l'étude des résidus) mais constatera également qu'ils sont ordonnés différemment ! Cependant, les noms («R-squared» pour r^2 , «somme des carrés», etc.) permettent de retrouver qui est qui.

R Console

```
> D = read.csv("Appartements.csv", sep=";", dec=",")
> summary(lm(prix ~ surface, data = D))

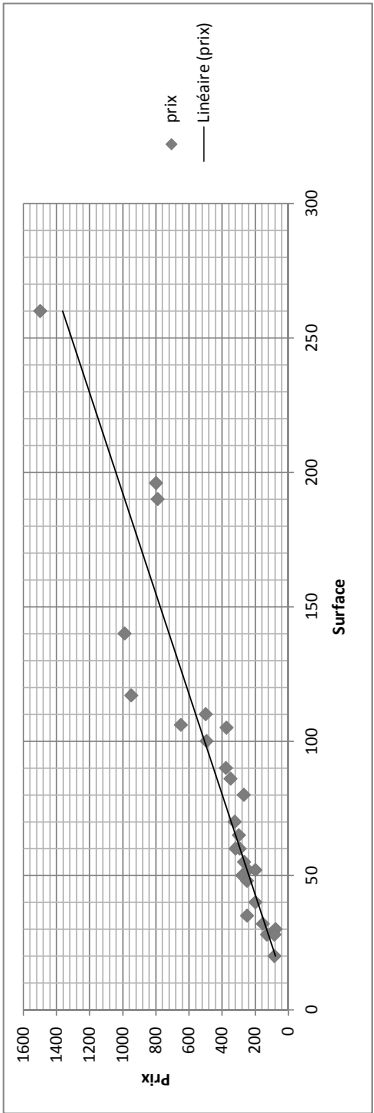
Call:
lm(formula = prix ~ surface, data = D)

Residuals:
    Min       1Q   Median       3Q      Max
-219.7543  -54.6924   -0.4285   23.9544  353.1456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -29.466     41.245  -0.714    0.481
surface         5.353       0.414  12.931 7.86e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 122.9 on 26 degrees of freedom
Multiple R-squared:  0.8654,    Adjusted R-squared:  0.8603
F-statistic: 167.2 on 1 and 26 DF,  p-value: 7.862e-13
```

FIGURE 8.6 – Sortie de régression linéaire obtenue sous R pour le jeu de données de prix d'appartements.



RAPPORT DÉTAILLÉ

Statistiques de la régression	
Coefficient de détermination multiple	0.930285669
Coefficient de détermination R^2	0.865431426
Coefficient de détermination R^2	0.860255712
Erreur-type	122.938855
Observations	28

ANALYSE DE VARIANCE

	Degré de liberté	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Régression	1	2527207.505	2527207.505	167.2100431	7.86227E-13
Résidus	26	392963.2091	15113.96958		
Total	27	2920170.714			

	Coefficients	Erreur-type	Statistique t	Probabilité	Limite inférieure	Limite supérieure	pour seuil de confiance = 95%
Constante	-29.46578921	41.24550534	-0.714400005	0.481345163	-114.2471388	55.3155604	
Variable X 1	5.3531636	0.413979978	12.93097224	7.86227E-13	4.502215577	6.204111623	

FIGURE 8.7 – Sortie de régression linéaire obtenue sous Excel pour le jeu de données de prix d'appar-
tements.

7. Préviation en un nouveau point ; détection des valeurs atypiques

On suppose toujours que les données $y_1, \dots, y_n$ peuvent être modélisées selon un modèle linéaire gaussien

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j, \quad j = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale $\mathcal{N}(0, \sigma_0^2)$. 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.

7.1. Préviation 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 préviation,

- donner un intervalle de confiance sur μ_x , l'espérance de Y_x ,
- ou même, donner un intervalle de préviation, dans lequel Y_x sera avec grande probabilité.

La fin de cette partie (de lecture facultative) détaille comment procéder. Nous nous contenterons ici des tracés SPSS de ces intervalles; évidemment, les intervalles de préviation 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 (de l'ordre de ± 2 fois l'écart-type), tandis que les derniers se contentent d'encadrer μ_x .

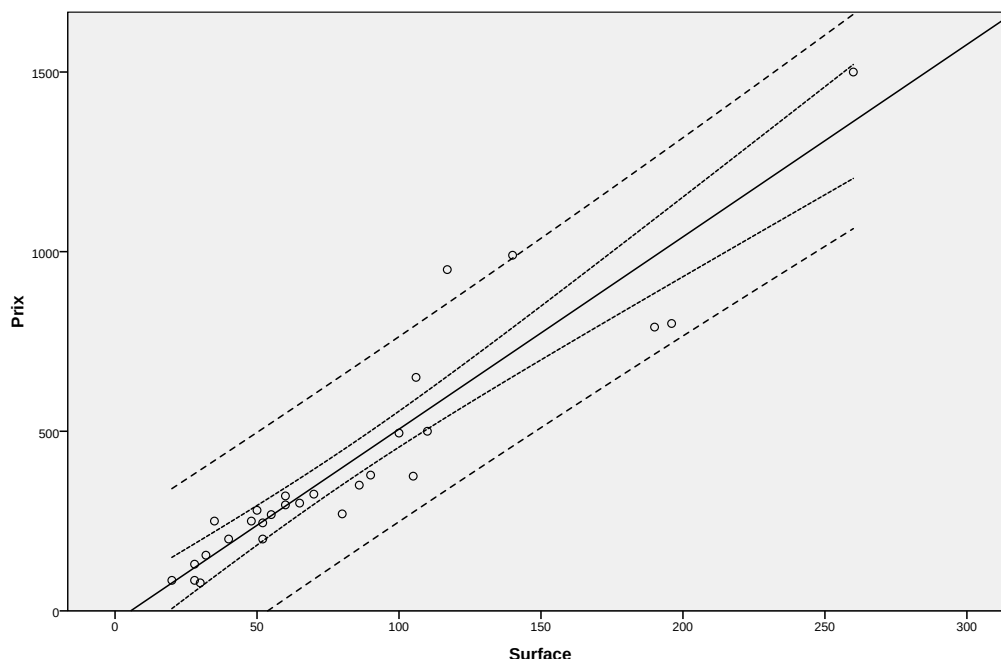


FIGURE 8.8 – Réalisations d'intervalles de confiance (pointillés rapprochés) et de préviation (pointillés plus éloignés), tous deux de niveau 95 %.

La minute SPSS 8.5. La figure 8.8 a été obtenue d'une manière similaire à celle de la Minute SPSS 8.1, 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 8.8 : on peut prévoir que le prix d'un nouvel appartement de 100 m² 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.

7.2. 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 8.3. Sur la figure 8.8, 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 prévisions) en sont généralement fort améliorés.

Remarque 8.6 (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 8.9).

7.3. Détails mathématiques (lecture facultative)

Intervalle de confiance sur l'espérance μ_x Certes, nous avons vu précédemment dans ce chapitre 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 ci-dessus. Les estimateurs $\hat{\alpha}_n$ et $\hat{\beta}_n$ sont combinaisons linéaires des observations gaussiennes indépendantes Y_j ; l'estimateur

$$\hat{\mu}_{x,n} = \hat{\alpha}_n + \hat{\beta}_n x$$

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 $\hat{\alpha}_n$ et $\hat{\beta}_n$. 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.

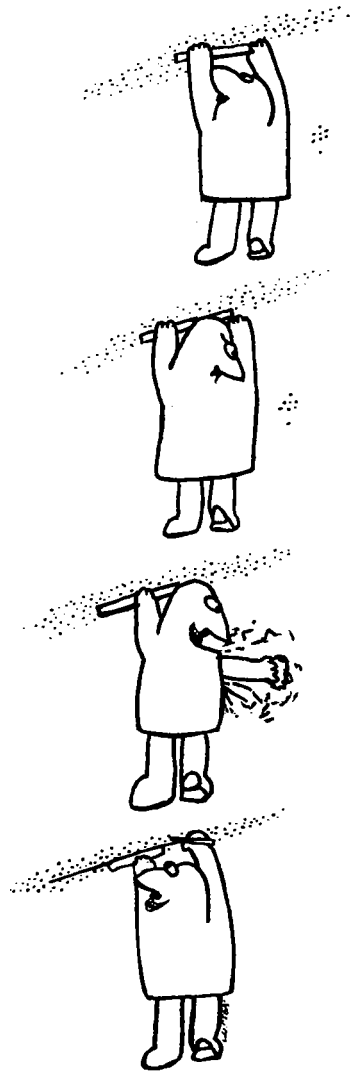


FIGURE 8.9 – Tricher, c'est mal. On peut effacer des données pour le traitement statistique, à condition d'avoir une justification économique de cette omission statistique et de ne pas oublier totalement le cas de ces données atypiques lors de la décision stratégique.

Théorème 8.4. Dans le modèle linéaire gaussien, l'estimateur $\hat{\mu}_{x,n} = \hat{\alpha}_n + \hat{\beta}_n x$ suit une loi normale,

$$\hat{\mu}_{x,n} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_0^2}{n} h_{x,n}\right) \quad \text{où} \quad h_{x,n} = 1 + \frac{1}{\text{Var}(x_1^n)} (x - \bar{x}_n)^2$$

est appelé le levier en x .

Par conséquent,

$$\sqrt{\frac{n}{\hat{\sigma}_n^2 h_{x,n}}} (\hat{\mu}_{x,n} - \mu_x) \sim \mathcal{T}_{n-2}.$$

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

$$\left[\hat{\alpha}_n + \hat{\beta}_n x \pm t_{n-2, 1-p/2} \sqrt{\frac{\hat{\sigma}_n^2}{n} h_{x,n}} \right].$$

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 $Y_1, \dots, Y_n$ (et des aléas correspondants, $\varepsilon_1, \dots, \varepsilon_n$).

On cherche un intervalle $\hat{I}_{x,n}$ tel que

$$\mathbb{P}\{Y_x \in \hat{I}_{x,n}\} \geq 1 - \alpha$$

pour un niveau $1 - \alpha$ fixé par l'utilisateur. $\hat{I}_{x,n}$ 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 $\hat{\sigma}_n^2$ de la variance σ_0^2 de $\varepsilon_x \sim \mathcal{N}(0, \sigma_0^2)$. Des calculs simples, utilisant le résultat du théorème 8.4, montrent alors le résultat suivant.

Théorème 8.5. Dans le modèle linéaire gaussien, la différence entre l'observation Y_x et la prévision de son espérance $\hat{\mu}_{x,n}$ suit une loi normale,

$$Y_x - \hat{\mu}_{x,n} \sim \mathcal{N}\left(0, \sigma_0^2 \left(1 + \frac{h_{x,n}}{n}\right)\right).$$

Une studentisation donne alors

$$\frac{1}{\sqrt{\hat{\sigma}_n^2 (1 + h_{x,n}/n)}} (Y_x - \hat{\mu}_{x,n}) \sim \mathcal{T}_{n-2}$$

et par conséquent, un intervalle de prévision au niveau $1 - \alpha$ est

$$\hat{I}_{x,n} = \left[\hat{\mu}_{x,n} \pm t_{n-2, 1-\alpha/2} \sqrt{\hat{\sigma}_n^2 \left(1 + \frac{h_{x,n}}{n}\right)} \right].$$

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.

La seconde assertion indique ce que nous écrivions précédemment : la variabilité de Y_x par rapport à son espérance μ_x est de l'ordre de ± 2 fois l'écart-type estimé. En effet, $t_{n-2, 97.5\%}$ est proche de 2 et on néglige le terme de levier $h_{x,n}/n$.

Régression linéaire multiple

Le chapitre précédent a introduit le modèle statistique de la régression linéaire simple, qui permet de rendre compte d'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.

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, prévision. 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 prévoir suffisamment efficacement la variable à expliquer tout en conservant une relation facilement interprétable.

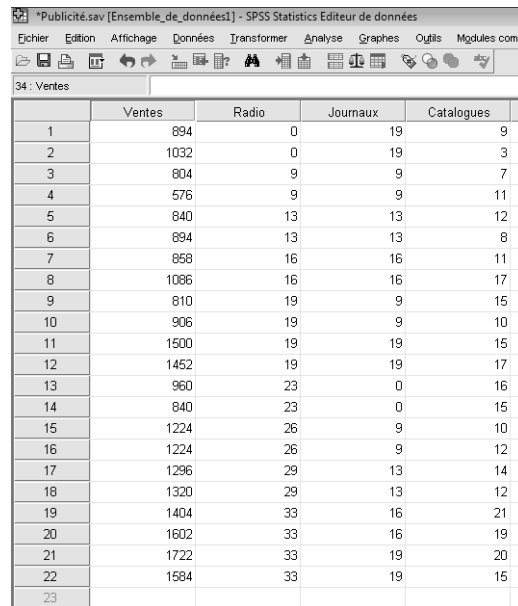
1. Le modèle linéaire multiple

Exemple 9.1. On considère la figure 9.1, qui reporte les données $y_1, y_2, \dots, y_{22}$ 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 $x_{1,1}, x_{1,2}, \dots, x_{1,22}$, ou dans la presse écrite, $x_{2,1}, x_{2,2}, \dots, x_{2,22}$, ainsi que ceux destinés à la distribution de catalogues des produits, $x_{3,1}, x_{3,2}, \dots, x_{3,22}$. 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.

1.1. Le modèle linéaire gaussien multiple

Pour chaque ville (chaque «individu» statistique), indexée par $t = 1, \dots, n$, on dispose d'une donnée y_t à expliquer en fonction d'un nombre fini k de données explicatives $x_{1,t}, x_{2,t}, \dots, x_{k,t}$. Pour les mêmes raisons et intuitions que dans le chapitre 8, on modélise les données à expliquer y_t comme la réalisation des variables aléatoires

$$\begin{aligned} Y_t &= \alpha_0 + \beta_{1,0} x_{1,t} + \beta_{2,0} x_{2,t} + \dots + \beta_{k,0} x_{k,t} + \varepsilon_t \\ &= \alpha_0 + \sum_{j=1}^k \beta_{j,0} x_{j,t} + \varepsilon_t, \quad \text{pour } t = 1, \dots, n, \end{aligned}$$



	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 9.1 – Le jeu de données considéré : montant des ventes (en milliers d'euros) de divers produits, 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 $\varepsilon_1, \dots, \varepsilon_n$ 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 $\beta_{1,0}, \dots, \beta_{k,0}$ 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 $\beta_{1,0}, \dots, \beta_{k,0}$, 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 au chapitre 8, 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 ; c'est 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 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 ?

Le point crucial 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, pour avoir une relation facilement interprétable. Il s'agit donc d'isoler un bon sous-ensemble des variables explicatives.

2. Un zeste d'algèbre linéaire

L'ensemble de ce paragraphe (technique) est destiné à votre culture et épanouissement mathématiques et peut être sauté en première lecture.

Formulation matricielle compacte. On note de manière compacte le modèle linéaire gaussien sous la forme matricielle suivante :

$$\underline{Y} = \mathbf{X} \underline{\beta}_0 + \underline{\varepsilon} \quad \text{où}$$

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{k,1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & \cdots & x_{k,n} \end{bmatrix}, \quad \underline{\beta}_0 = \begin{bmatrix} \alpha_0 \\ \beta_{1,0} \\ \vdots \\ \beta_{k,0} \end{bmatrix}, \quad \text{et} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

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 $\mathbf{X} \underline{\beta}_0$).

ÉLÉMENTS 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 $k+1 \leq n$. 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 $\mathbf{X}^T \mathbf{X}$.

Estimation du vecteur des coefficients par la méthode des moindres carrés. Comme dans le chapitre précédent, on considère l'estimateur $\hat{\underline{\beta}}_n$ des moindres carrés. Avec les notations précédentes (norme¹ euclidienne $\|\cdot\|_2$), ce dernier est défini comme antécédent du minimum d'un écart quadratique :

$$\hat{\underline{\beta}}_n \in \operatorname{argmin}_{\underline{b} \in \mathbb{R}^{k+1}} \|\underline{Y} - \underline{X} \underline{b}\|_2.$$

Proposition 9.1. *Sous l'hypothèse d'injectivité de $\underline{X}$, on a l'expression explicite suivante :*

$$\hat{\underline{\beta}}_n = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}.$$

ÉLÉMENTS 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 $\hat{\underline{\beta}}_n$ tel que $\underline{X} \hat{\underline{\beta}}_n$ est la projection orthogonale $\Pi_V \underline{Y}$ de $\underline{Y}$ sur $V = \operatorname{Im} \underline{X}$, l'image de $\underline{X}$. En particulier, le produit scalaire entre n'importe quel élément de $\operatorname{Im} \underline{X}$ et $\underline{Y} - \Pi_V \underline{Y} = \underline{Y} - \underline{X} \hat{\underline{\beta}}_n$ est nul, ce que l'on peut ré-écrire, de manière matricielle par

$$\underline{X}^T (\underline{Y} - \underline{X} \hat{\underline{\beta}}_n) = \underline{0}.$$

Cette égalité se traduit par

$$\underline{X}^T \underline{Y} = \underline{X}^T \underline{X} \hat{\underline{\beta}}_n, \quad \text{soit encore} \quad \hat{\underline{\beta}}_n = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y},$$

où l'on a utilisé un fait prouvé précédemment, à savoir que $\underline{X}^T \underline{X}$ est inversible.

On conclut ce paragraphe culturel par une indication de propriété théorique de l'estimateur exhibé : il est sans biais.

Proposition 9.2. *L'estimateur $\hat{\underline{\beta}}_n$ est sans biais : $\mathbb{E}[\hat{\underline{\beta}}_n] = \underline{\beta}_0$.*

En effet, par linéarité,

$$\mathbb{E}[\hat{\underline{\beta}}_n] = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \mathbb{E}[\underline{Y}] = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{X} \underline{\beta}_0 = \underline{\beta}_0,$$

où l'on a également utilisé que par hypothèse du modèle, $\mathbb{E}[\underline{\varepsilon}] = \underline{0}$.

Remarque 9.1. On notera

$$\hat{\underline{\beta}}_n = \begin{bmatrix} \hat{\alpha}_n \\ \hat{\beta}_{1,n} \\ \vdots \\ \hat{\beta}_{k,n} \end{bmatrix}$$

les composantes de l'estimateur vectoriel $\hat{\underline{\beta}}_n$ du vecteur $\underline{\beta}_0$.

Reconstruction et prévision. Pour un nouveau vecteur de données explicatives $\underline{x}' = (x'_1, \dots, x'_k)^T$, on utilise alors, pour prévoir 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 :

$$\hat{\underline{\beta}}_n \cdot \underline{x}' = \hat{\alpha}_n + \hat{\beta}_{1,n} x'_1 + \dots + \hat{\beta}_{k,n} x'_k.$$

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.)

1. On la définit comme : $\|\underline{x}\|_2 = \sqrt{x_1^2 + \dots + x_n^2}$ pour tout $\underline{x} = (x_1, \dots, x_n)$

On note, pour la suite de ce chapitre,

$$\hat{Y}_t = \hat{\underline{\beta}}_n \cdot \underline{x}_t \quad \text{pour } t = 1, \dots, n,$$

la reconstruction de y_t à partir des estimateurs des coefficients et du vecteur $\underline{x}_t = (1, x_{1,t}, \dots, x_{k,t})$, 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 9.2).

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 9.2 – Tableaux SPSS de résultats de régression linéaire multiple sur les données de l'exemple 9.1.

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 $\hat{\beta}_n$. (Ceci est à comparer au tableau similaire exhibé au cours précédent.)

Exemple 9.2. 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 le chapitre précédent : 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

$$H_0 : \beta_{1,0} = \beta_{2,0} = \dots = \beta_{k,0} = 0$$

en espérant que le test rejette H_0 , ce qui correspondra à l'existence d'une telle relation.

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 :

$$\hat{Y}_t = \hat{\beta}_n \cdot \underline{x}_t \quad \text{et} \quad \hat{\varepsilon}_t = Y_t - \hat{Y}_t, \quad \text{pour } t = 1, \dots, n,$$

où l'on a noté $\underline{x}_t = (1, x_{1,t}, \dots, x_{k,t})$ la t -ème ligne de $\mathbf{X}$ (la t -ème série de variables explicatives). Les vecteurs colonnes de longueur n correspondants seront notés $\mathbf{X} \hat{\beta}_n = \hat{\mathbf{Y}}$ et $\hat{\varepsilon}$.

Premier fait : la moyenne des reconstructions $\hat{Y}_t$ est égale à la moyenne des observations Y_t :

$$\frac{1}{n} \sum_{t=1}^n \hat{Y}_t = \frac{1}{n} \sum_{t=1}^n Y_t \stackrel{\text{not.}}{=} \bar{Y}_n.$$

En effet, par construction, le vecteur $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ est orthogonal à l'image de $\mathbf{X}$, qui contient en particulier le vecteur $\underline{1} = (1, \dots, 1)$.

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 137) :

$$\Sigma_T \stackrel{\text{not.}}{=} \sum_{t=1}^n (Y_t - \bar{Y}_n)^2, \quad \Sigma_E \stackrel{\text{not.}}{=} \sum_{t=1}^n (\hat{Y}_t - \bar{Y}_n)^2, \quad \text{et} \quad \Sigma_R \stackrel{\text{not.}}{=} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2.$$

Là encore, les notations matricielles permettent une ré-écriture plus compacte ; notant $\bar{\mathbf{Y}}$ le vecteur colonne $\bar{Y}_n \underline{1}$ (i.e., celui dont les n composantes sont égales à $\bar{Y}_n$), il vient :

$$\Sigma_T = \|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2, \quad \Sigma_E = \|\hat{\mathbf{Y}} - \bar{\mathbf{Y}}\|_2^2, \quad \text{et} \quad \Sigma_R \stackrel{\text{not.}}{=} \|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2 = \|\hat{\varepsilon}\|_2^2.$$

Par une égalité de Pythagore, au vu de la projection orthogonale effectuée², 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 8.3 de la manière suivante³.

2. $\hat{\mathbf{Y}}$ est la projection orthogonale de $\mathbf{Y}$ sur l'image de $\mathbf{X}$, qui contient le vecteur $\underline{1}$ et donc également $\bar{\mathbf{Y}}$

3. Désormais, on perd $k+1$ degrés de liberté dans la loi de Σ_R / σ_0^2 , qui est donc une loi $\chi_{n-(k+1)}^2$, 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 $1/(n-(k+1))$ pour l'estimateur de la variance s'en déduit.

Théorème 9.1. Dans le modèle linéaire gaussien multiple, on dispose de l'estimateur de la variance σ_0^2 donné par

$$\hat{\sigma}_n^2 = \frac{1}{n - (k + 1)} \sum_{t=1}^n (\hat{\varepsilon}_t)^2 = \frac{\Sigma_R}{n - (k + 1)} .$$

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_{ajust}^2 . Les contraintes que l'on s'impose sont

- qu'elle soit d'autant plus grande que $\hat{\sigma}_n^2$ soit petit (i.e., qu'elle soit une fonction décroissante de $\hat{\sigma}_n^2$);
- 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;
- ... et enfin, que $r_{\text{ajust}}^2 \approx r^2$.

La solution retenue par la littérature et la pratique statistiques est

$$r_{\text{ajust}}^2 = 1 - \frac{n - 1}{n - (k + 1)} (1 - r^2) = 1 - \frac{(n - 1) \hat{\sigma}_n^2}{\Sigma_T} .$$

Lecture du premier tableau. Nous avons maintenant tous les outils pour lire le premier tableau de la figure 9.2 : il est donné par la réalisation du tableau suivant.

R	R-deux	R-deux ajusté	Erreur standard
$\sqrt{r^2}$	r^2	r_{ajust}^2	$\sqrt{\hat{\sigma}_n^2}$

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_{ajust}^2 , car elles seules⁴ tiennent compte du nombre de variables (entre deux modèles, on choisit celui de plus fort r_{ajust}^2 réalisé ou, de manière équivalente, celui de plus faible $\sqrt{\hat{\sigma}_n^2}$ 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 %).

4. On rappelle que r^2 augmente mécaniquement lorsque le nombre de variables explicatives augmente; ce n'est pas le cas de sa version ajustée.

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 on peut rejeter H_0 dans le test d'hypothèses

$$H_0 : \beta_{1,0} = \dots = \beta_{k,0} = 0 \quad \text{vs.} \quad H_1 : \text{il existe } j \text{ tel que } \beta_{j,0} \neq 0.$$

On avait vu dans le chapitre précédent 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 :

$$D_n = \frac{\frac{1}{k} \sum_{t=1}^n (\hat{Y}_t - \bar{Y}_n)^2}{\frac{1}{n - (k + 1)} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2} = \frac{\Sigma_E / k}{\Sigma_R / (n - (k + 1))} = \frac{n - (k + 1)}{k} \frac{r^2}{1 - r^2}.$$

(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.

Lecture du deuxième tableau. Nous avons maintenant tous les outils pour lire le deuxième tableau de la figure 9.2 : il est donné par la réalisation du tableau suivant.

Somme carrés	ddl	Moyenne carrés	D	Sig.
Σ_E	k	Σ_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 linéaire⁵.

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.

5. 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.

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

$$H_0 : \beta_{j,0} = 0 \quad \text{vs.} \quad H_1 : \beta_{j,0} \neq 0;$$

lorsque H_0 est conservée, la variable j n'apporte pas de telle contribution significative.

Remarque 9.2 (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 soit redondante avec une ou plusieurs autre(s) variable(s) explicative(s), soit parce qu'elle est intrinsèquement inutile.

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

$$\hat{\underline{\beta}}_n = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underline{Y}$$

étant combinaison linéaire d'observations normales indépendantes, il suit une loi normale. Le vecteur $\hat{\underline{\beta}}_n$ 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

$$\hat{\alpha}_n \sim \mathcal{N}(\alpha_0, \sigma_0^2 (\mathbf{X}^T \mathbf{X})_{1,1}^{-1}) \quad \text{et} \quad \hat{\beta}_{p,n} \sim \mathcal{N}(\beta_{p,0}, \sigma_0^2 (\mathbf{X}^T \mathbf{X})_{p+1,p+1}^{-1})$$

(pour tout $p = 1, \dots, k$), puis, par standardisation, on parvient au théorème suivant.

Théorème 9.2. Dans le modèle linéaire gaussien multiple,

$$T'_n = \frac{\hat{\alpha}_n - \alpha_0}{\sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{1,1}^{-1}}} \sim \mathcal{T}_{n-(k+1)} \quad \text{et} \quad T_{p,n} = \frac{\hat{\beta}_{p,n} - \beta_{p,0}}{\sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{p+1,p+1}^{-1}}} \sim \mathcal{T}_{n-(k+1)}$$

pour tout $p = 1, \dots, k$.

On en déduit des intervalles de confiance (précisés dans le tableau 9.1) et des tests à une valeur de référence. Le troisième tableau de la figure 9.2 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 étude des validités marginales (ou des significativités individuelles).

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.

Coeff.	Err. standard	[Ignorée]	t	Sig.	IC(α) – bornes inf. et sup.
$\hat{\alpha}_n$	$\sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{1,1}^{-1}}$		T'_n	P-val. $H_0 : \alpha_0 = 0$	$\left[\hat{\alpha}_n \pm t_{n-(k+1), 1-\alpha/2} \sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{1,1}^{-1}} \right]$
$\hat{\beta}_{1,n}$	$\sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{2,2}^{-1}}$		$T_{1,n}$	P-val. $H_0 : \beta_{1,0} = 0$	$\left[\hat{\beta}_{1,n} \pm t_{n-(k+1), 1-\alpha/2} \sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{2,2}^{-1}} \right]$
...	...		...	...	
...	...		...	...	
$\hat{\beta}_{k,n}$	$\sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{k+1,k+1}^{-1}}$		$T_{k,n}$	P-val. $H_0 : \beta_{k,0} = 0$	$\left[\hat{\beta}_{k,n} \pm t_{n-(k+1), 1-\alpha/2} \sqrt{\hat{\sigma}_n^2 (\mathbf{X}^T \mathbf{X})_{k+1,k+1}^{-1}} \right]$

TABLE 9.1 – Décryptage mathématique du troisième tableau de la figure 9.2 : estimées des coefficients, tests de significativité individuelle face aux autres variables explicatives, intervalles de confiance.

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 9.3 (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 (mais on serait davantage en droit de le faire au vu de leurs P-valeurs de significativités individuelles).

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 pages 244 et suivantes (notamment en ce qui concerne les interprétations économiques et stratégiques à formuler).

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

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 $k \geq 1$ 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 $k \geq 1$ de variables explicatives fixées

A nombre k de variables fixées, on retient, parmi les modèles dans lesquels toutes les variables explicatives sont individuellement significatives, celui avec le meilleur r^2 (i.e., avec la plus petite estimée de l'écart-type des résidus).

Ainsi, dans l'exemple 9.1 (figure 9.1), 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 9.4 et 9.5. Il s'agit en l'occurrence du couple de variables Radio et Journaux.

Remarque 9.5 (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 9.6 illustre cela.

La minute SPSS 9.1. La figure 9.6 a été obtenue par Analyse / Correlation / Bivariée.

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

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 9.4 – Deux régressions avec deux couples de variables explicatives, sur les données de l'exemple 9.1.

Régression Ventes / Journaux & Gratuits

Variables introduites/supprimées

Modèle	Variables introduites	Variables supprimées	Méthode
1	Gratuits, ^a Journaux	.	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	t	Sig.
		A	Erreur standard	Bêta		
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 9.5 – Une troisième régression avec un autre couple de variables explicatives, sur les données de l'exemple 9.1.

coefficient de détermination associé à la régression simple correspondante. (Le signe est en fait celui de l'estimée du coefficient de régression $\hat{\beta}_{p,n}$ 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 besoin⁸ 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.

8. 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 9.6 – Tableau croisé des corrélations sur les données de l'exemple 9.1 (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 9.3. Dans notre exemple, en lisant la régression complète de la figure 9.2, 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 9.4), 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 9.4. 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 9.4). 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 9.2). La procédure s'arrête ici et recommande donc finalement le modèle construit sur *Radio* et *Journaux*.

Remarque 9.6. Note : les procédures “backward” et “forward” ne conduisent pas toujours à la même recommandation ! 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 détaillons⁹ pas.

9. 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 9.5 (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 explicative qualitative avec plus de trois valeurs possibles,
- 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),
- la variable à expliquer est qualitative (régression logistique),

sont plus compliqués et ne sont pas à notre programme cette année. ... Peut-être l'an prochain ?

6. Résumé des compétences nécessaires

Ce chapitre de cours est un peu aride et ne peut être bien compris qu'en résolvant de nombreux exercices d'examen, comme nous le ferons ensemble. Campagne BDA ou pas, les séances de cours sur la régression linéaire sont donc vitales. L'expérience montre que les listeux BDA dévalident le cours de statistique plus souvent qu'à leur tour à cause de leur absence à ces séances cruciales. C'est pourquoi nous proposons la «check-list» suivante de compétences.

Le minimum vital pour obtenir E au problème de régression à l'examen

Cette liste est cumulative ; l'absence d'une des compétences suivantes suffit à dévalider le problème de régression, donc l'examen, donc tout le cours de statistique !

1. Savoir valider statistiquement un modèle (validité globale, validités marginales)
2. Savoir écrire le modèle linéaire proposé, sous une forme intelligible (en donnant le nom des variables), exploitable (en précisant les unités), et statistiquement correcte (ne pas oublier l'existence d'un terme de résidu, dont la taille typique est caractérisée par son écart-type ; attention à l'écriture du terme de résidu)
3. Savoir valider économiquement un modèle (*a minima* : se demander si les signes des coefficients sont logiques ou pas)
4. Savoir prévoir une nouvelle observation et/ou déterminer si telle observation du jeu de données est conforme ou non au modèle (en exploitant le terme de résidu, et le fait qu'il soit typiquement compris entre ± 2 fois son écart-type)

Deuxième partie

Exercices issus de quiz et d'examens

Corrigés des exercices

Disponibles sur le site web du cours

<http://www.hec.fr/stoltz>

Principe des quiz

Composés d'un ou plusieurs exercices courts

Temps de composition de 30 à 40 minutes

Sans documents

Index des exercices par thème

Recueil de données

- Une alléchante mission JE page 177

Tests de comparaison à une valeur de référence

- Révolte populaire page 178
- Téléphones Banana page 179
- Fermeture de MegaUpload page 180
- Enquête de santé publique : prévalence de l'asthme page 182

Tests à partir d'échantillons indépendants ou appariés

- Concours de musique : vidéo versus son page 184
- Cadeaux et comportements d'achat page 186
- Comparaison de prix : Franprix versus Monoprix page 187
- Comparaison de salaires : hommes versus femmes page 187
- Prix du forfait de ski en Savoie et Haute-Savoie voir question (2) page 244

Tests du χ^2 d'ajustement

- Offres promotionnelles pour aller au cinéma page 193
- Epreuve pré-professionnelle de la campagne JE page 194
- Temps d'attente téléphonique pour joindre un service clients page 196
- Théorie du hasard génétique de Mendel page 202
- COP21 — Mise en évidence d'un réchauffement climatique ? page 198
- Report de votes page 200
- Détection de tricheries comptables grâce à la loi de Benford (partie 1) page 203

Tests du χ^2 d'indépendance

- Détection de tricheries comptables grâce à la loi de Benford (partie 2) page 206
- Naufrage du Titanic page 207
- Sociologie des habitudes de lecture page 208
- Couleurs de cheveux en Ecosse page 210
- Etude sociologique des configurations familiales page 211
- La clientèle des palaces parisiens page 212
- Publicité et sexisme page 214

Intervalles de confiance

- Création d'une compagnie d'assurance page 215
- Pré-indemnisation par une assurance page 216
- Décompte de la population du royaume selon Laplace page 218
- Les sondeurs se contredisent-ils ? (Lecture du *Canard enchaîné*) page 220

Régression linéaire simple

- L'exemple historique de Galton : taille des fils et taille des pères page 221
- Décomptes de manifestants en France, 1984–2014 page 224
- Forfait de ski en fonction de la taille du domaine page 226
- Regarder la télé rend-il fou ? page 228
- Croissance des orangers page 231

Une alléchante mission JE

Comment auriez-vous effectué la mission JE suivante? On supposera le questionnaire fourni; la question porte sur sa bonne administration. Expliquez comment il aurait fallu procéder, devinez comment les étudiants ont au contraire procédé.

La JE recrute: 135€ pour administrer 30 questionnaires à la sortie de magasins spécialisés dans la chasse et pêche.



Boîte de réception x

	FLEISCHMANN, Marie Bonjour à tous, Dans le cadre d'une mission pour un cabinet de conseil en str...	7 janv. (Il y a 6 jours) ☆
	FLEISCHMANN, Marie Nous précisons que les frais de transports seront remboursés.	7 janv. (Il y a 6 jours) ☆
	BRACQUART, Florent Bonjour à tous, Nous recherchons encore des personnes susceptibles d'administ...	11 janv. (Il y a 2 jours) ☆
	BRACQUART, Florent <florent.bracquart@hec.edu> À Marie, GE, GE, GE, GE Recrutement terminé, les postulants sélectionnés seront contactés dans la journée afin de commencer la mission dès demain. Merci à tous, 	11 janv. (Il y a 2 jours) ☆

Révolte populaire

Un gouvernement songe à une profonde réforme fiscale mais craint une agitation populaire à quelques mois des élections ; les renseignements généraux estiment que tant que 15 % de la population n'est pas prête à se mobiliser activement (dans la rue ou sur les réseaux sociaux) contre la réforme cette dernière peut passer. Un institut de sondage interroge 1 000 Français, parmi lesquels 980 s'expriment : 131 sont prêts à agir activement mais les 849 autres, tout en ne soutenant pas nécessairement la réforme, affirment qu'ils se contenteront au mieux de pester dans leur coin. Un conseiller du Premier ministre affirme que la proportion observée étant de $131/980 \approx 13.4\%$, ce qui est bien inférieur à 15 %, le gouvernement doit se lancer. Qu'en pensez-vous ? Nous allons effectuer un test d'hypothèses pour répondre à cette question.

- Indiquez la population visée :

- Indiquez les données disponibles :

- Précisez et interprétez le paramètre d'intérêt :

- Posez les hypothèses à tester (en justifiant leur choix) :

- Calculez la valeur réalisée de la statistique de test (formule et application numérique) :

- Faites le schéma où on lit la zone de rejet et la P -valeur :

- Indiquez la valeur numérique de la P -valeur :
- Conclusion statistique (une phrase) :

- Conclusion stratégique (une phrase) :

Les téléphones Banana (contrôle de qualité et campagne marketing associée)

Laisser tomber son smartphone peut être un drame : on perd ses contacts, tout lien au monde, et en plus, cet accident n'est pas couvert par la garantie et il faudra déboursier plusieurs centaines d'euros pour un smartphone de remplacement. Bref, voilà un créneau tout trouvé pour Banana, un concurrent d'Apple et d'Orange : cette entreprise a construit un smartphone révolutionnaire qui peut encaisser au moins 260 chocs en moyenne, selon l'argumentaire commercial développé. Sceptique, une association civique et féminine d'utilisateurs de téléphonie mobile, «Téléphonées», mène un test sur 60 appareils de Banana, dont les résultats sont reproduits ci-dessous.

Echantillon	$n = 60$ téléphones
Nombre moyen de chocs avant panne	$\bar{x} = 255.1$
Ecart-type des nombres de chocs	$s = 24.4$

Que faut-il en conclure ? On répondra selon le canevas suivant.

- Indiquez la population visée :
- Indiquez les données disponibles :
- Précisez et interprétez le paramètre d'intérêt :
- Posez les hypothèses à tester (en justifiant leur choix) :
- Calculez la valeur réalisée de la statistique de test (formule et application numérique) :
- Faites le schéma où on lit la zone de rejet et la P -valeur :
- Indiquez la valeur numérique de la P -valeur :
- Conclusion statistique (une phrase) :
- Conclusion stratégique (une phrase) :
- Allons plus loin (*selon le temps restant*) : pensez-vous qu'un client donné sera satisfait d'une garantie en moyenne, telle que développée par l'argumentaire commercial ?

Fermeture de MegaUpload

Depuis la fermeture, le 19 janvier 2012, du site de téléchargement direct et gratuit MegaUpload ordonnée par le gouvernement américain par l'intermédiaire du FBI, les sites légaux de télévision en *replay* (télévision de rattrapage) et de vente de VOD (*video on demand*) se félicitent d'une importante hausse de leur trafic. S'il est vrai que cette hausse est flagrante dans les jours qui ont suivi la fermeture de MegaUpload, plus d'un mois après, cette tendance n'est pas si évidente. En étudiant la fréquentation du site **m6replay.fr** (en pourcentage d'utilisateurs de l'Internet) sur les deux années qui précèdent la fermeture de MegaUpload, on constate qu'elle s'élève en moyenne avant le 19 janvier à 0.021 %. Une représentation graphique des fréquentations mesurées à des instants réguliers (environ une mesure tous les jours) mais aléatoirement espacés (pas toutes les mesures aux mêmes heures) après la fermeture est reproduite en annexe. Elle est issue du site de mesures de trafic Internet **alexa.com**. On ne s'intéressera qu'à 40 mesures, celles réalisées entre le 20 janvier 2012 et le 29 février 2012, dont on résume les caractéristiques ci-dessous.

Echantillon	$n = 40$ mesures de trafic
Moyenne	0.023 %
Ecart-type	0.005 %

Que faut-il en conclure ? On répondra selon le canevas suivant. (Pour gagner du temps, la modélisation sera ici effectuée de manière très rapide, une fois n'est pas coutume.)

- Indiquez les données disponibles et le paramètre d'intérêt :

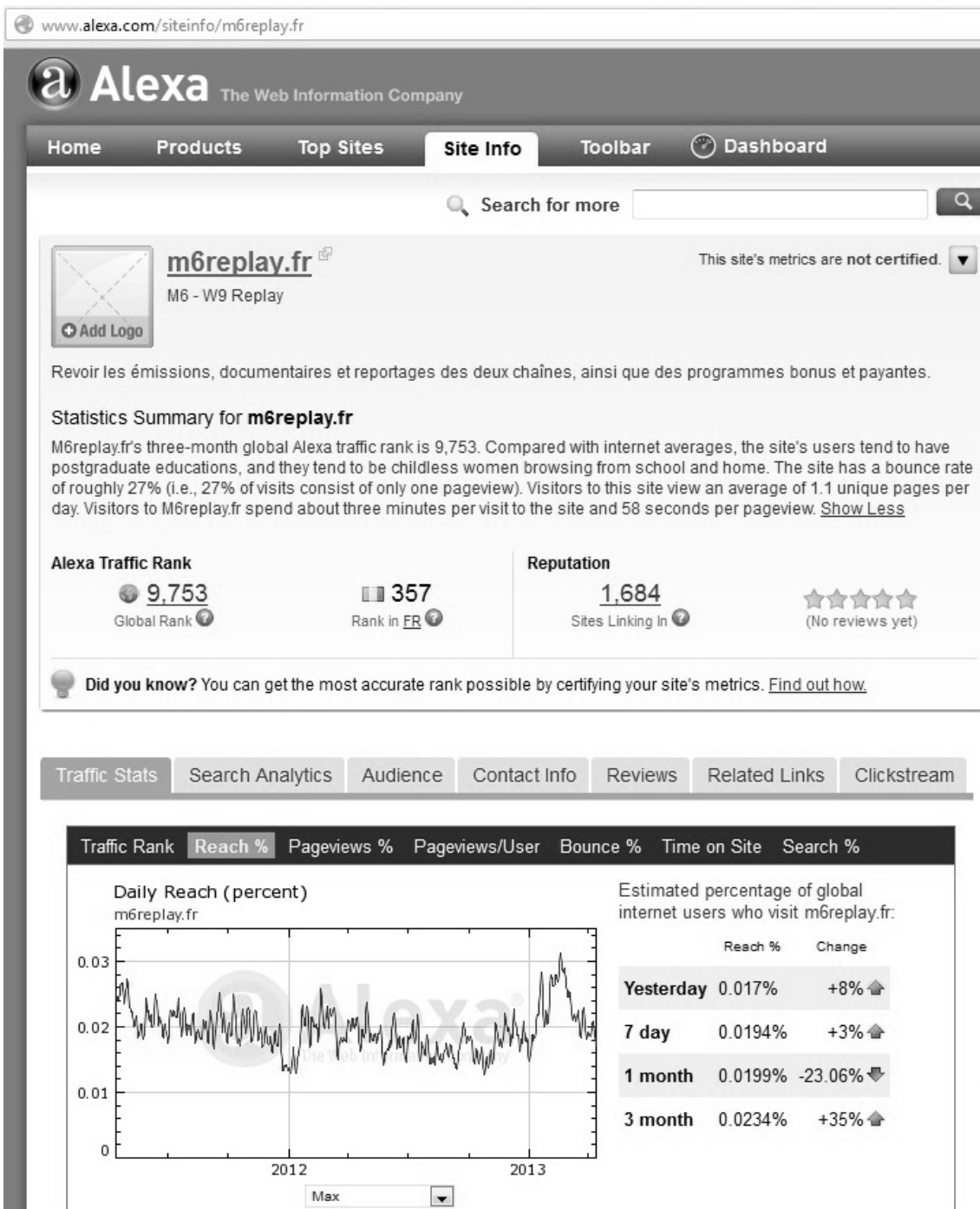
- Posez les hypothèses à tester (en justifiant leur choix) :

- Calculez la valeur réalisée de la statistique de test (formule et application numérique) :

- Faites le schéma où on lit la zone de rejet et la P -valeur :

- Indiquez la valeur numérique de la P -valeur :
- Conclusion statistique (une phrase) :

- Conclusion stratégique (une phrase ; choisissez au préalable un commanditaire de l'étude) :



Enquête de santé publique : prévalence de l'asthme

Cet exercice est inspiré de faits réels. Au niveau national, la prévalence de l'asthme chez les enfants en CM2 est de 9.7% selon une étude nationale effectuée par l'Institut de veille sanitaire (INVS) : c'est le taux des enfants ayant été traités au cours des 12 derniers mois. Dans une ville moyenne de montagne, encaissée, et à proximité de laquelle un incinérateur a été ouvert, l'air est de mauvaise qualité une grande partie de l'année. Les parents se plaignent que leurs enfants ont fréquemment des affections respiratoires. Pour en avoir le cœur net, une fédération de parents d'élèves effectue une enquête auprès des enfants de CM2 de la ville; après relances, elle obtient exactement 300 réponses concernant l'asthme : 37 enfants déclarent avoir suivi un traitement dans les 12 derniers mois, soit 12.3% des enfants.

Que doit faire cette association de parents? On effectuera d'abord un traitement statistique des données, puis on procurera une conclusion stratégique.

- Indiquez et interprétez le paramètre d'intérêt (en précisant la population au passage), ainsi que la valeur de référence :

- Indiquez les données disponibles et les statistiques d'échantillon :

- Formulez et justifiez le jeu d'hypothèses à tester (d'abord avec des mots, puis avec des équations).

- Précisez la statistique de test (formule théorique) et son comportement sous les différentes hypothèses retenues :

Recherche sur les déterminants des scores aux concours de musique

Cet exercice est inspiré d'un article de recherche de Chia Jung Tsay, "Sight over sound in the judgment of music performance".

Le thème de cette recherche est de comprendre les déterminants qui affectent le succès à des concours de musique classique de très haut niveau. Il est connu qu'il est très difficile de prévoir le gagnant parmi les trois finalistes de tels concours car ils sont tous trois excellents. Une première étude cherche à savoir si en écoutant la bande son des prestations des candidats il est possible de retrouver le gagnant (plus efficacement qu'en devinant aléatoirement !). Dans une seconde étude la même question est posée : peut-on deviner le gagnant mieux qu'au hasard, mais en regardant une vidéo sans le son ? De façon paradoxale, et que ce soit pour des novices ou des professionnels, il semble que la vidéo sans le son permette mieux de déceler le gagnant à un tel concours que le son.

Première étude : utilité de la bande son (6 points)

Dans une première étude on va proposer à 105 personnes de tenter de deviner le gagnant d'un concours parmi les trois finalistes en écoutant la bande son de leur prestation au concours. On cherche à savoir si la bande son leur permet de faire mieux (ou moins bien) que le hasard pour deviner le gagnant. On effectue pour cela un test d'hypothèses.

- Modélisez la situation autant que possible (puisque les données ne sont pas encore disponibles). Indiquez notamment la population étudiée, l'échantillon retenu, et interprétez le paramètre d'intérêt (*1 point*).

- Formulez et justifiez le jeu d'hypothèses à tester, d'abord avec des mots et ensuite seulement avec des équations (*1.5 points*).

- Précisez la statistique de test (formule théorique) et son comportement sous les différentes hypothèses retenues (*1 point*).

- L'expérience est menée et parmi ces 105 personnes, 29 arrivent à retrouver le gagnant en écoutant la musique. Calculez la valeur réalisée de la statistique de test (sa valeur numérique, *0.5 point*).

- Calculez la P -valeur du test face à ces données, en traçant au préalable un schéma récapitulant les réponses aux deux questions précédentes (*1 point*).

- Proposez des conclusions pour cette étude (sort de l'hypothèse H_0 , puis une phrase non technique compréhensible par tou-te-s et commençant par «Ces données...» ; *1 point*).

Seconde étude : utilité de la vidéo (2 points)

Dans une deuxième étude on propose à 105 autres personnes de tenter de deviner le gagnant d'un concours parmi les trois finalistes en regardant la vidéo sans le son de leur prestation au concours (en l'occurrence, 54 d'entre elles). Là encore, on cherche à savoir si la video muette leur permet de faire mieux (ou moins bien) que le hasard pour deviner le gagnant. On effectue pour cela un test d'hypothèses.

- Expliquez comment adapter les calculs précédents pour former une conclusion statistique ; précisez notamment la P -valeur obtenue (*2 points*).

Conclusion générale (2 points)

- Les deux études ci-dessus permettent-elles à elles seules d'affirmer que le public est mieux apte à détecter le gagnant d'un tel concours à l'aide de la video que du son ? Pourquoi ? (*1 point*)

- Si vous avez formulé une réponse négative à la question précédente — Comment pourrait-on répondre alors à la question posée, quel test devrait-on effectuer ? Rédigez très concisément sa mise en œuvre en vous focalisant sur l'obtention d'une P -valeur (*1 point*).

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) :

Exercice 1 — Comparaison de prix : Franprix versus Monoprix

Cet exercice porte sur des données réelles, recueillies par Benjamin Petiau (enseignant au premier semestre) au cours de l'automne 2015, dans deux magasins de proximité de Versailles, une enseigne Franprix et un magasin Monoprix. Il s'agit de relevés des prix de produits de consommation courante. Ces relevés de prix correspondent aux deux premières variables du jeu de données, **Franprix** et **Monoprix** (voir capture d'écran des données page suivante). Pour chaque produit comparé, il a ensuite calculé la différence de prix ainsi que le logarithme naturel du rapport des prix. Ces variables sont **Différence**, définie comme **Franprix** – **Monoprix** ; et **LnRapport**, définie par $\ln(\text{Franprix}/\text{Monoprix})$.

1. De quel type de données (simples, appariées, indépendantes) s'agit-il ?
2. Imaginez la question que Benjamin Petiau avait en tête avant d'effectuer le recueil des données et déduisez-en le jeu d'hypothèses testées (justifiez le choix des hypothèses).

Diverses sorties du logiciel **SPSS** sont procurées aux pages suivantes : diagrammes de dispersion des données, statistiques descriptives sur les variables considérées, mise en œuvre de tests statistiques.

3. Calculez la valeur des variables **Différence** et **LnRapport** pour le premier produit.
4. Modélisez les données relatives à la variable **Différence** : population objet de l'étude, paramètre d'intérêt ; échantillon, données disponibles, résumés statistiques.
5. Effectuez le test d'hypothèses et procurez une P -valeur : rappelez la statistique de test en jeu, son comportement sous H_0 , son comportement sous H_1 ; calculez sa valeur réalisée, tracez un schéma récapitulatif, et calculez-y la P -valeur.
6. Formez une conclusion statistique lisible et compréhensible par le grand public (une phrase commençant par «Ces données...») ; puis formez une conclusion stratégique.

La question qui suit s'appuie sur les T -tests menés sous **SPSS** et reproduits dans les pages suivantes. (Vous pourrez vérifier en partie votre résultat de la question 5. à partir de cette sortie.)

7. Pourquoi peut-il sembler plus naturel de considérer la variable **LnRapport** plutôt que la variable **Différence** ? Et quelle est la valeur de référence associée, au vu du jeu d'hypothèses retenu ?
8. Déduisez la P -valeur procurée par le test mené sur **LnRapport** à partir de celle calculée par **SPSS** (en clair : ne faites pas des calculs similaires à ceux de la question 5).
9. Formez une conclusion statistique lisible et quantifiant le cas échéant de manière naturelle les phénomènes mis en évidence.

Exercice 2 — Comparaison de salaires : hommes versus femmes

Les données ci-dessous sont extraites d'une récente «enquête emploi» de l'INSEE (pour ceux qui sont en licence d'économie à Paris I : c'est le jeu de données que nous utilisons pour les travaux pratiques sous SAS). Le jeu de données reporte pour chaque sondé son âge, son sexe, la région dans laquelle il habite, sa catégorie socio-professionnelle, son niveau de diplôme, etc., et son salaire mensuel net moyen.

On se demande s'il y a une discrimination salariale entre hommes et femmes. On adopte le point de vue d'une association féministe, qui croit savoir d'expérience qu'une telle discrimination existe. Plus précisément, elle s'intéresse à un phénomène de plafond de verre. C'est pourquoi nous restreignons l'étude à la catégorie socio-professionnelle 37 (cadres administratifs et commerciaux d'entreprises) et à la région parisienne.

1. Formulez un jeu d'hypothèses statistiques (avec des mots, sans équation) en justifiant son choix.
2. Etudiez la sortie **SPSS** procurée : expliquez comment la lire.
3. Quelle conclusion statistique indique-t-elle ? Précisez la P -valeur lue pour votre jeu d'hypothèses (à l'aide d'un schéma). Ecrivez également une phrase compréhensible par le grand public et commençant par «Ces données...».
4. A-t-on ou non prouvé l'existence d'une discrimination salariale ?

supermarche.sav [Jeu_de_données1] - IBM SPSS Statistics Editeur de données

	V1	Franprix	Monoprix	Différence	LnRapport
1	Vodka Polialov (1L)	17,63	16,74	,89	,05
2	Chocolat Lindt 70%	2,73	2,47		
3	Coca Zéro	3,13	2,99	,14	,05
4	Papier toilette LOTUS	2,96	2,97	-,01	,00
5	Nettoyant VU Lunettes	3,55	2,83	,72	,23
6	Paic Citron	4,92	3,85	1,07	,25
7	Dés d'épaule Fleury Michon	1,50	1,30	,20	,14
8	Pot Crème Vanille La Laitière	2,21	1,92	,29	,14
9	Fromage Blanc Câlin	2,05	1,85	,20	,10
10	Pot Activia Pêche	1,86	1,61	,25	,14
11	Jus Tropicana	3,03	2,86	,17	,06
12	Taille fine Chocolat	1,96	1,62	,34	,19
13	SojaSun Noisette	1,74	1,74	,00	,00
14	Gervita	2,00	1,82	,18	,09
15	Chocolat Danette	1,62	1,40	,22	,15
16	Steack haché Charal	2,25	2,62	-,37	-,15
17	Camembert Président	1,77	1,61	,16	,09
18	Crème Fleurette	2,32	1,85	,47	,23
19	Carte d'or Vanille	7,16	7,70	-,54	-,07
20	Fleury Michon (*2)	1,89	1,91	-,02	-,01
21	Krispolls Sans sucre ajouté	1,75	1,55	,20	,12
22	Sucre en Poudre Dady	1,40	1,16	,24	,19

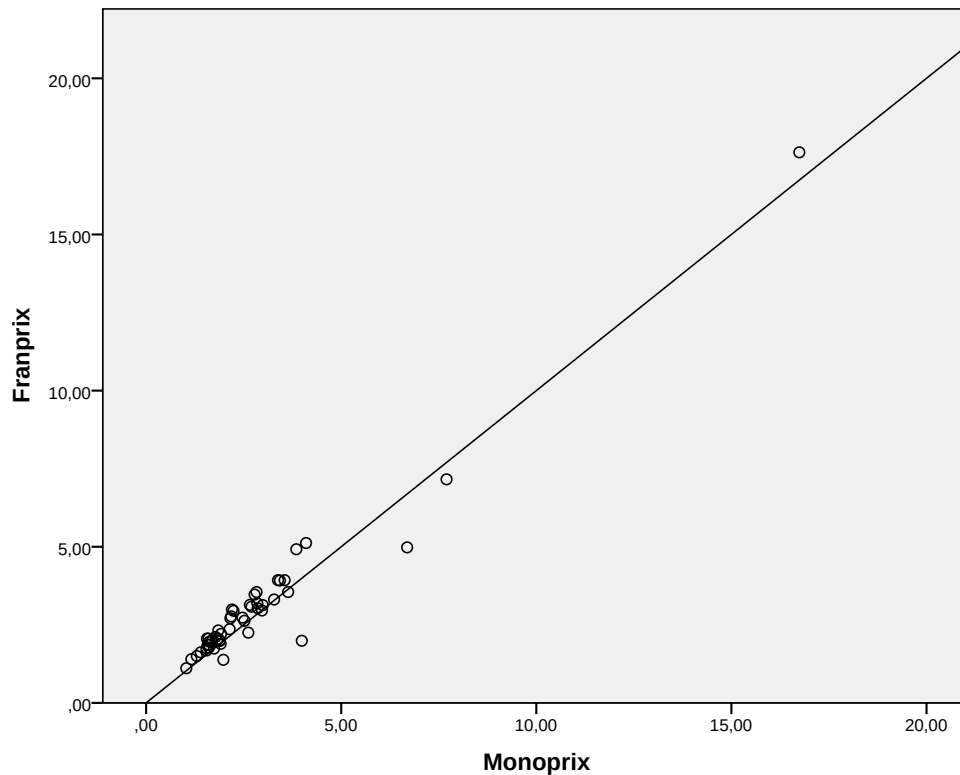


Diagramme de dispersion des données

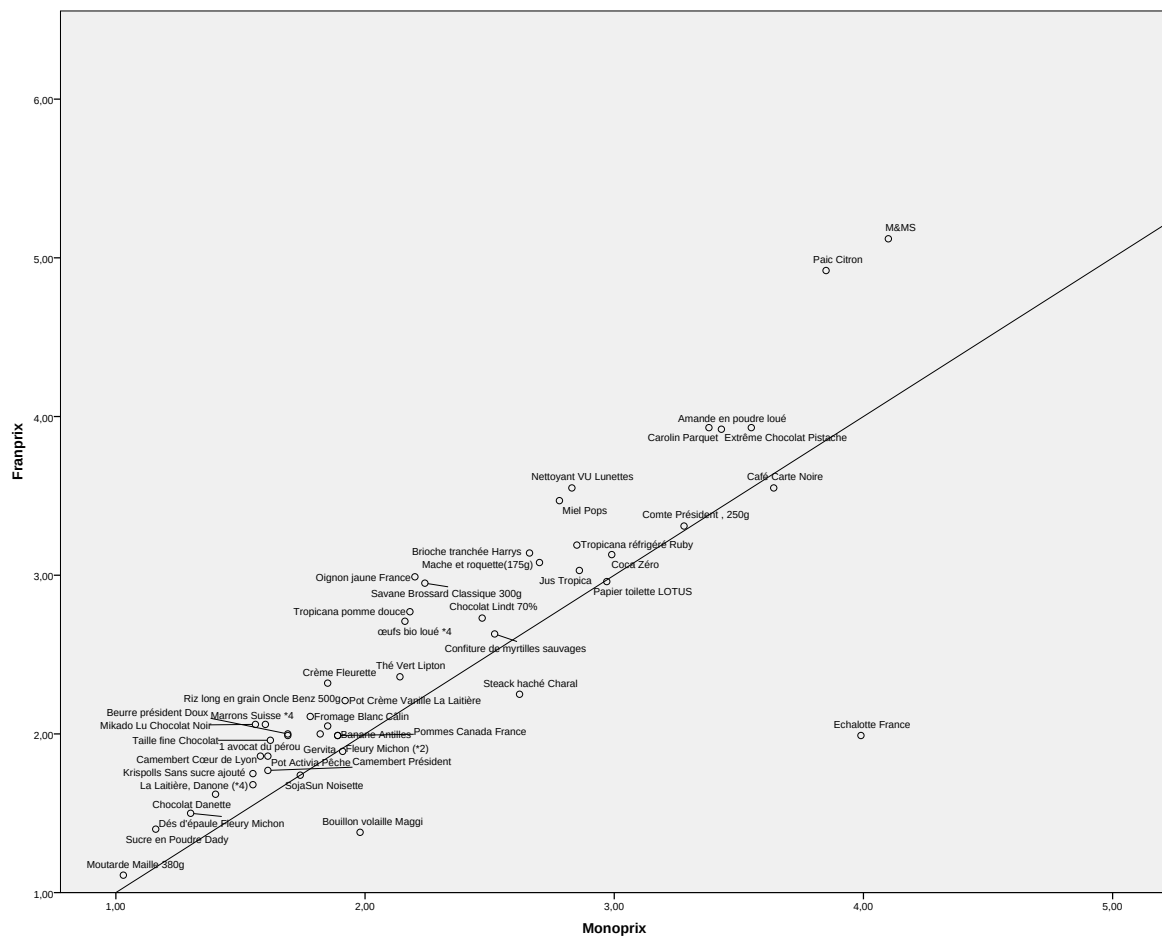


Diagramme de dispersion des données (zoom sur les plus petits prix)

Statistiques

		Franprix	Monoprix	Différence	LnRapport
N	Valide	51	51	51	51
	Manquant	0	0	0	0
Moyenne		2,9741	2,7667	,2075	,0901
Médiane		2,3200	2,1600	,2500	,1214
Ecart type		2,37978	2,34611	,53437	,17093
Minimum		1,11	1,03	-2,00	-,70
Maximum		17,63	16,74	1,07	,31

Test sur échantillon unique

	Valeur de test = 0					
	t	ddl	Sig. (bilatéral)	Différence moyenne	Intervalle de confiance de la différence à 95 %	
					Inférieur	Supérieur
Différence	2,772	50	,008	,20745	,0572	,3577
LnRapport	3,765	50	,000	,09012	,0420	,1382

Prénom, nom :

Barème : chaque question vaut 0.5 point ; à la correction, je cocherai le carré ☐ lorsque les points sont acquis, je le laisserai vide sinon. Il y a 20 carrés au total, soit 10 points.

Exercice 1 — 7 points

- ☐ Données simples, appariées, ou indépendantes ?
- ☐ Hypothèses retenues (en mots) :

- ☐ Justification rapide :

- ☐ Valeurs de **Différence** et **LnRapport** pour le deuxième produit :
- ☐ Population objet de l'étude et paramètre d'intérêt :

- ☐ Echantillon, données disponibles, résumés statistiques :

- ☐ Formule de la statistique de test, valeur réalisée :

- ☐ Comportements sous H_0 et sous H_1 :

- ☐ Schéma récapitulatif, P -valeur :

☐ Conclusion statistique : «Ces données...»

☐ Conclusion stratégique :

☐ Pourquoi LnRapport et valeur de référence associée :

☐ *P*-valeur lue pour LnRapport :

Test sur échantillon unique

	Valeur de test = 0					
	t	ddl	Sig. (bilatéral)	Différence moyenne	Intervalle de confiance de la différence à 95 %	
					Inférieur	Supérieur
Différence	2,772	50	,008	,20745	,0572	,3577
LnRapport	3,765	50	,000	,09012	,0420	,1382

☐ Conclusion statistique quantifiant le cas échéant les phénomènes mis en évidence :

Exercice 2 — 3 points

- ☐ Hypothèses retenues (en mots) :
- ☐ Justification rapide :
- ☐ Comment lire la sortie suivante (quelles cases lire, qu'y lit-on) :

Statistiques de groupe

	Sexe	N	Moyenne	Ecart type	Moyenne erreur standard
Salaire net mensuel	Hommes	179	3431,46	3895,437	291,159
	Femmes	147	2434,90	1282,947	105,816

Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes						Intervalle de confiance de la différence à 95 %	
		F	Sig.	t	ddl	Sig. (bilatéral)	Différence moyenne	Différence erreur standard		Inférieur	Supérieur
Salaire net mensuel	Hypothèse de variances égales	8,195	,004	2,972	324	,003	996,560	335,371		336,781	1656,339
	Hypothèse de variances inégales			3,217	223,375	,001	996,560	309,791		386,074	1607,047

- ☐ P -valeur pour votre jeu d'hypothèses (avec schéma) :
- ☐ Conclusion statistique : «Ces données...»
- ☐ A-t-on prouvé l'existence d'une discrimination salariale ?

Offres promotionnelles pour aller au cinéma

Cet exercice est inspiré de faits réels : l'offre accompagnant dont il est question ici existe réellement pour une grande chaîne de cinémas et dans certains quartiers. L'entreprise Hugécé dispose de quatre grands complexes cinématographiques à Paris : au Forum des Halles (30 salles), à Montparnasse (10+3 salles), sur les Champs-Élysées (6+7 salles), et à Bercy (25 salles). Ils sont fréquentés de manière inégale : ils représentent chacun en moyenne 43%, 20%, 13% et 24% de sa clientèle du week-end. Pour damer le pion à la concurrence, Hugécé a lancé il y a plusieurs années déjà une carte d'abonnement. Elle veut maintenant fidéliser également les non-détenteurs d'une telle carte et attirer éventuellement une nouvelle clientèle ; elle voudrait également viser à équilibrer la fréquentation de ses cinémas. Elle a ainsi l'idée de permettre à ses abonnés d'acheter une place à tarif réduit pour un accompagnant. Cette offre est cependant limitée en nombre : seules 2 000 telles places d'accompagnants sont disponibles chaque week-end. La première semaine, cette offre est proposée dans l'ensemble de ses complexes et il n'y a pas de quota par complexe. Le succès est immédiat et les résultats sont reproduits dans le tableau suivant. Que doit faire l'équipe marketing pour la deuxième semaine d'expérimentation ?

Cinéma

	Effectif observé	N théorique	Résidus
Forum des Halles	957	860,0	97,0
Montparnasse	438	400,0	38,0
Champs-Élysées	158	260,0	-102,0
Bercy	447	480,0	-33,0
Total	2000		

Tests statistiques

	Cinéma
Khi-deux	56,835 ^a
ddl	3
Sig. asymptotique	,000

a. 0 cellules (0,0%) ont des fréquences théoriques inférieures à 5.
La fréquence théorique minimum d'une cellule est 260,0.

- ☐ Nom du test mis en œuvre et jeu d'hypothèses testées (en mots)
- ☐ Le test est-il validement appliqué ? Précisez la P -valeur obtenue.
- ☐ Conclusion statistique (non technique, pouvant être lue par tout collaborateur de l'entreprise)
- ☐ Conclusion stratégique : suggestion d'action pour la deuxième semaine d'expérimentation

Epreuve pré-professionnelle de la campagne JE

Cet exercice est tiré de faits 100% réels, ayant eu lieu en novembre 2012. Un groupe d'étudiants de la liste JE de Nicolas Hubert (liste finalement élue!) interroge des étudiants sur le campus afin de réaliser une enquête sur la culture à HEC, thème de l'épreuve pré-professionnelle de la campagne JE de novembre 2012. Les 200 étudiants sondés se répartissent de la manière suivante, selon le programme et le sexe :

	Hommes	Femmes	Total
Grande Ecole	78	73	151
MS / MSc	17	12	29
MBA	10	7	17
Doctorat	1	2	3
Total	106	94	200

Sophie Garnichat, de la Direction de la communication, indique que les 2 983 étudiants en cours de scolarité se répartissent quant à eux de la manière suivante dans ces catégories :

	Hommes	Femmes	Total
Grande Ecole	39.4%	31.3%	70.7%
MS / MSc	9.4%	7.0%	16.4%
MBA	6.5%	3.6%	10.1%
Doctorat	1.5%	1.3%	2.8%
Total	56.8%	43.2%	100%

A partir des trois couples de tableaux reproduits aux deux pages suivantes, déterminer si l'échantillon de sondés est représentatif de l'ensemble des étudiants en cours de scolarité à HEC. On rédigera la réponse sous la forme d'un texte expliquant notamment quels tests statistiques sont mis en œuvre, avec quelles hypothèses, et quelles conclusions on peut tirer (ou pas) de chaque couple de tableaux.

Catégories (asexuées)

	Effectif observé	Effectif théorique	Résidu
Grande Ecole	151	141,3	9,7
MS-MSc	29	32,9	-3,9
MBA	17	20,2	-3,2
Doctorat	3	5,6	-2,6
Total	200		

Test

	Catégories (asexuées)
Khi-deux	2,841 ^a
ddl	3
Signification asymptotique	,417

a. 0 cellule (,0%) a des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 5,6.

Catégories regroupées

	Effectif observé	Effectif théorique	Résidu
Grande Ecole / Homme	78	78,7	-,7
Grande Ecole / Femme	73	62,6	10,4
MS-MSc / Homme	17	18,8	-1,8
MS-MSc / Femme	12	14,1	-2,1
MBA ou Doctorat / Homme	11	16,0	-5,0
MBA ou Doctorat / Femme	9	9,9	-,9
Total	200		

Test

	Catégories regroupées
Khi-deux	3,852 ^a
ddl	5
Signification asymptotique	,571

a. 0 cellule (,0%) a des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 9,9.

Catégorie

	Effectif observé	Effectif théorique	Résidu
Grande Ecole / Homme	78	78,7	-,7
Grande Ecole / Femme	73	62,6	10,4
MS-MSc / Homme	17	18,8	-1,8
MS-MSc / Femme	12	14,1	-2,1
MBA / Homme	10	13,1	-3,1
MBA / Femme	7	7,2	-,2
Doctorat / Homme	1	2,9	-1,9
Doctorat / Femme	2	2,7	-,7
Total	200		

Test

	Catégorie
Khi-deux	4,368 ^a
ddl	7
Signification asymptotique	,737

a. 2 cellules (25,0%) ont des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 2,7.

Question subsidiaire : Si (et seulement si) il vous reste du temps, expliquez comment les “effectifs théoriques” ont été obtenus dans les trois couples de tableaux.

Temps d'attente téléphonique pour joindre un service clients

Un opérateur Internet cherche à attirer des clients en leur promettant que son service clients est facile à joindre. A cet effet, et avant de communiquer sur cet argument, il effectue un audit interne et cherche à voir si 50 % des clients attendent moins de 2 minutes et 90 % des clients attendent moins de 5 minutes, comme il le souhaite. (En tout cas, ce sont ces objectifs chiffrés que la direction générale a demandé au service clientèle de réaliser.) Le cas limite d'attente est donc que 50 % des clients attendent 2 minutes ou moins, 40 % des clients attendent entre 2 et 5 minutes, et 10 % attendent plus de 5 minutes.

Ses deux centres d'appels sont testés, ce qui conduit à la considération des deux jeux de données reproduits en annexe. Etudiez, selon la grille d'analyse ci-dessous, les résultats obtenus par un certain traitement statistique.

Premier jeu de données

Répondez de manière concise aux questions suivantes.

- Précisez le nom du test mis en œuvre.

- Indiquez les hypothèses testées (avec des mots, d'abord, puis en termes mathématiques plus précis).

- Le test est-il validement appliqué?

- Quelle est la P -valeur obtenue?

- Formez une conclusion statistique, puis une conclusion stratégique.

Second jeu de données

Les hypothèses testées étant les mêmes, répondez uniquement aux questions suivantes.

- Quelle est la P -valeur obtenue?

- Formez une conclusion statistique, puis une conclusion stratégique.

Premier jeu de données

Temps d'attente

	Effectif observé	Effectif théorique	Résidu
<= 2 min	58	55,5	2,5
> 2 min & <= 5 min	44	44,4	-,4
> 5 min	9	11,1	-2,1
Total	111		

Test

	Temps d'attente
Khi-deux	,514 ^a
ddl	2
Signification asymptotique	,774

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 11,1.

Second jeu de données

Temps d'attente

	Effectif observé	Effectif théorique	Résidu
<= 2 min	58	52,0	6,0
> 2 min & <= 5 min	44	41,6	2,4
> 5 min	2	10,4	-8,4
Total	104		

Test

	Temps d'attente
Khi-deux	7,615 ^a
ddl	2
Signification asymptotique	,022

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 10,4.

COP21 — Mise en évidence d'un changement climatique ?

A l'occasion de la COP21 tenue en 2015 à Paris, voici un exercice sur le réchauffement climatique. Vos enseignants se sont intéressés à l'évolution de la température à Paris et ont cherché à comparer les températures de la seconde moitié du 20ème siècle avec celles du passé de long terme, notamment de la période pré-industrielle (avant 1850). Ils se fondent sur les données de l'article "Les températures mensuelles en région parisienne de 1676 à 2008" de Daniel Rousseau (Conseil supérieur de la météorologie) paru dans la revue *Météorologie*, numéro 67 daté de novembre 2009.

Nous analysons les 49 relevés de température effectués à la station météo du parc du Montsouris à Paris de 1960 à 2008, en nous concentrant sur les températures moyennes du mois d'août de chaque année de cette période. Ces températures sont-elles distribuées de la même manière que les températures moyennes des mois d'août du passé de très long terme (période pré-industrielle) ? Dans ce passé de très long terme, les températures moyennes en août suivaient peu ou prou une loi normale de moyenne $18,5^{\circ}\text{C}$ avec un écart-type de $1,5^{\circ}\text{C}$.

Nous décidons donc de grouper les températures moyennes du mois d'août en quatre catégories, avec pour seuils $18,5^{\circ}\text{C}$ et $18,5^{\circ}\text{C} \pm 1,5^{\circ}\text{C}$. Pour chacune ces catégories, nous précisons la fréquence attendue sous hypothèse d'une loi normale de moyenne $18,5^{\circ}\text{C}$ et d'écart-type de $1,5^{\circ}\text{C}$:

- Catégorie 1 : $]-\infty ; 17^{\circ}\text{C}[$, fréquence attendue : 15,9%
- Catégorie 2 : $[17^{\circ}\text{C} ; 18,5^{\circ}\text{C}[$, fréquence attendue : 34,1%
- Catégorie 3 : $[18,5^{\circ}\text{C} ; 20^{\circ}\text{C}[$, fréquence attendue : 34,1%
- Catégorie 4 : $[20^{\circ}\text{C} ; +\infty[$, fréquence attendue : 15,9%

Voici le résultat de l'analyse sous SPSS.

Températures -- groupement #1			
	Effectif observé	N théorique	Résidus
... < 17	2	7,8	-5,8
17 <= ... < 18.5	15	16,7	-1,7
18.5 <= ... < 20	17	16,7	,3
20 <= ...	15	7,8	7,2
Total	49		

Tests statistiques	
	Températures -- groupement #1
Khi-deux	11,155 ^a
ddl	3
Sig. asymptotique	,011

a.0 cellules (0,0%) ont des fréquences théoriques inférieures à 5.
La fréquence théorique minimum d'une cellule est 7,8.

- Quel est le test mis en oeuvre ici ?
- Enoncer en mots les hypothèses testées.
- Ce test est-il appliqué de façon valide ?
- Quelle est la P -valeur obtenue ?

- Formuler une conclusion statistique ; puis écrire une phrase dans un style grand public qui expliquerait ce que nous avons mis ou pas en évidence.

Suite à la précédente étude, nous décidons d'aller plus loin et consultons un modèle du GIEC. Celui-ci prévoit pour Paris et pour la période 1960-2008, une hausse moyenne de 0,7°C de la température moyenne des mois d'août par rapport à la distribution du passé de long terme (pré-industriel). Nous souhaitons donc tester ce modèle à partir des données.

- Quelles sont les catégories utilisées désormais, avec quelles fréquences attendues ?

- L'analyse sous SPSS est reproduite ci-dessous. Formuler une conclusion statistique à partir d'elle ; puis écrire une phrase dans un style grand public qui expliquerait ce que nous avons mis ou pas en évidence. (Il n'est pas besoin ici de rappeler le nom du test, les hypothèses testées, etc. Seule la conclusion nous intéresse.)

Températures -- groupement #2

	Effectif observé	N théorique	Résidus
Catégorie 1	9	7,8	1,2
Catégorie 2	18	16,7	1,3
Catégorie 3	15	16,7	-1,7
Catégorie 4	7	7,8	-,8
Total	49		

Tests statistiques

	Températures -- groupement #2
Khi-deux	,542 ^a
ddl	3
Sig. asymptotique	,909

a. 0 cellules (0,0%) ont des fréquences théoriques inférieures à 5.
La fréquence théorique minimum d'une cellule est 7,8.

Report de votes

On étudie les reports de votes entre les deux tours des élections cantonales 2015 dans les deux communes formant le canton de Bagneux : Bourg-la-Reine et Bagneux. Au premier tour, étaient en lice des binômes de candidats représentant les partis suivants : PCF, PS, UMP, FN; et un binôme DVD (divers droite, dissidents de l'UDI). Ils ont obtenu des scores très différents selon les villes, mais au niveau global, ce sont le PCF et l'UMP qui se sont qualifiés pour le second tour. L'objet de cet exercice est d'étudier la dynamique de report des votes. On sait par des experts politiques que dans le contexte présent, les électeurs du FN se reportent à 75% sur l'UMP, parti d'opposition et parti de droite, et à 25% vont s'abstenir ou voter blanc. On veut voir si le report des électeurs du PS sur le PCF et des DVD sur l'UMP est total. Pour raisonner à effectifs constants, on considère l'abstention comme un pseudo-parti.

Cas de Bagneux (2 points). Il y a 21 197 inscrits sur les listes électorales. Les scores au premier tour ont été, en pourcentage des inscrits :

- abstention, vote blanc ou nul — 60.20%
- PCF — 20.08%
- PS — 5.50%
- DVD — 2.76%
- UMP — 5.14%
- FN — 6.32%

On effectue alors le test suivant.

Vote à Bagneux			
	Effectif observé	Effectif théorique	Résidu
Abstention	13087	13095,5	-8,5
PCF	5486	5422,2	63,8
UMP	2624	2679,3	-55,3
Total	21197		

Test	
	Vote
Khi-deux	1,898 ^a
ddl	2
Signification asymptotique	,387

a. 0 cellules (.0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 2679,3.

— Comment l'effectif théorique de 2679.3 a-t-il été obtenu pour l'UMP ?

— Précisez le nom du test vu en cours et mis en œuvre ici.

— Indiquez les hypothèses testées (avec des mots puis de manière mathématique).

- Le test est-il validement appliqué?
- Quelle est la P -valeur obtenue?
- Formez une conclusion statistique (une phrase sur le statut de H_0 et une phrase *non technique* compréhensible par tou-te-s et commençant par «Ces données...»).

Cas de Bourg-la-Reine (1 point). Il y a 21 197 inscrits sur les listes électorales. Les scores au premier tour ont été, en pourcentage des inscrits :

- abstention, vote blanc ou nul — 49.80%
- PCF — 6.27%
- PS — 14.20%
- DVD — 5.81%
- UMP — 18.08%
- FN — 5.84%

On effectue alors deux tests, l'un selon le même modèle que précédemment avec reports 1/4-3/4 entre abstention et UMP pour le FN, reports totaux pour le PS et les DVD ; et l'autre, en modifiant le report du PS pour 90% vers le PCF et 10% pour l'UMP (pour modéliser un effet gauche caviar de réticence au vote communiste).

Vote à Bourg-la-Reine / Report intégral

	Effectif observé	Effectif théorique	Résidu
Abstention	7098	7033,9	64,1
PCF	2520	2808,9	-288,9
UMP	4104	3879,2	224,8
Total	13722		

Test

	Vote
Khi-deux	43,323 ^a
ddl	2
Signification asymptotique	,000

a. 0 cellules (.0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 2808,9.

Vote à Bourg-la-Reine / Report partiel

	Effectif observé	Effectif théorique	Résidu
Abstention	7098	7033,9	64,1
PCF	2520	2614,0	-94,0
UMP	4104	4074,1	29,9
Total	13722		

Test

	Vote
Khi-deux	4,187 ^a
ddl	2
Signification asymptotique	,123

a. 0 cellules (.0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 2614,0.

Analyser de manière concise ces deux sorties de tests : que peut-on en conclure ?

Validation expérimentale par Mendel de sa théorie du hasard génétique

Pour tester ses lois du hasard génétique, Mendel a croisé des plants de pois lisses et jaunes avec des plants de pois ridés et verts, le caractère ridé étant selon lui récessif par rapport au caractère lisse, de même pour la couleur verte par rapport à la jaune. Il n'a gardé de la première génération engendrée que les plants de pois lisses et jaunes. A la deuxième génération, obtenue par croisement de plants de la première génération, il a noté 315 plants de pois lisses et jaunes, 108 lisses et verts, 101 jaunes et ridés, 32 verts et ridés. Or ses lois prévoyaient une répartition (9/16, 3/16, 3/16, 1/16) de ces différents caractères. Vient-il de les confirmer ? On étudie les tableaux suivants.

Phénotype			
	Effectif observé	Effectif théorique	Résidu
Lisses et jaunes	315	312,8	2,3
Lisses et verts	108	104,3	3,8
Ridés et jaunes	101	104,3	-3,3
Ridés et verts	32	34,8	-2,8
Total	556		

Test	
	Phénotype
Khi-deux	,470 ^a
ddl	3
Signification asymptotique	,925

a. 0 cellules (.0%) ont des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 34,8.

- Précisez le nom du test mis en œuvre.
- Indiquez les hypothèses testées (avec des mots).
- Le test est-il validement appliqué ?
- Quelle est la P -valeur obtenue ?
- Formez une conclusion statistique (dont une phrase non technique compréhensible par tou-te-s).
- Formez une conclusion stratégique (qu'aurait dû faire Mendel si la statistique avait existé à son époque et s'il avait disposé du résultat de la question précédente?).

Détection de tricheries comptables grâce à la loi de Benford (partie 1)

La balance des paiements (*balance of payments*, BoP) est un document statistique élaboré sous forme comptable, élément de la comptabilité nationale recensant les flux de biens, de services, de revenus, de transferts de capitaux, et les flux financiers que les résidents d'un pays dans leur ensemble (particuliers, entreprises ou État) entretiennent avec ceux du reste du monde.

On étudie les balances des paiements transmises trimestriellement au FMI par différents états dans une période allant de 1989 à 2007 ; on considère pour chaque trimestre et chaque état 13 entrées de cette balance (soldes des échanges de biens, des investissements directs, etc.). Une théorie mathématique bien établie indique que la loi des premiers chiffres significatifs¹ de ces entrées doit suivre une loi dite de Benford... en tout cas si les états n'ont pas manipulé les données !

Cette loi indique que, parmi les valeurs possibles, qui sont celles entre 1 et 9, le 1 doit survenir environ 30.1 % du temps, le 2, environ 17.6 %, etc., jusqu'au 9, qui ne doit apparaître que 4.6 % du temps.

On groupe les états en trois catégories, selon leur type de régime monétaire : taux flottant, taux fixe, taux mixte (ni tout à fait flottant, ni tout à fait fixe). Un modèle économique simple montre que les états en régime monétaire à taux fixe ont des incitations plus fortes à manipuler les données et à embellir le descriptif de leur situation que ceux à taux flottant. C'est pourquoi on veut effectuer un test d'authenticité des données comptables présentées : vérifier qu'elles n'ont pas été manipulées.

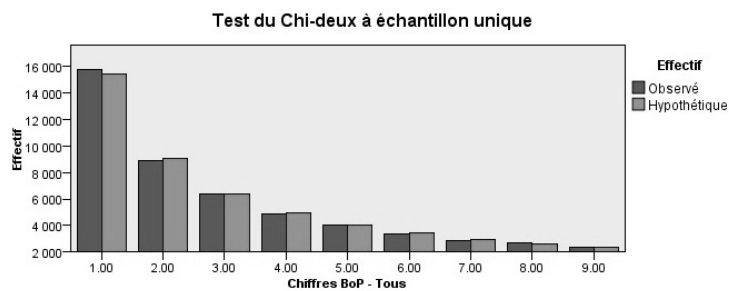
— Donnez le nom du test mis en œuvre aux pages suivantes, rappelez les hypothèses testées (avec des mots).

— Indiquez les conditions d'application valide du test et constatez qu'elles sont réalisées (comment le sait-on?).

— Pour chacune des trois situations étudiées (tous les états, uniquement ceux en taux fixe, uniquement ceux en taux flottant), précisez la P-valeur obtenue, formez une conclusion statistique et, lorsque vous le pouvez, une conclusion économique et/ou stratégique (pensez que vous êtes le FMI).

Reportez les réponses en bas de chaque mise en œuvre en précisant de quelle situation il s'agit parmi les trois listées à la question précédente ; ne les écrivez pas ici.

1. Par exemple, le premier chiffre significatif de 1 235 416 est 1.



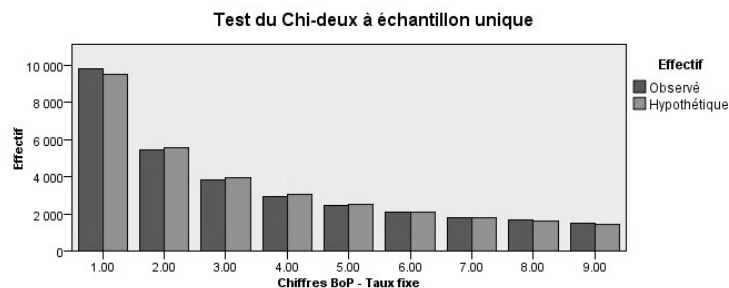
N Total	51,281
Statistique de test	13.681
Degrés de liberté	8
Sig. asymptotique (test bilatéral)	.090

1. Il y a 0 cellules (0%) avec des valeurs attendues inférieures à 5. La valeur minimum attendue est 2,358.926.

Situation étudiée : Tous les Etats — Taux fixe — Taux flottant

P-valeur :

Conclusions statistique et éventuellement stratégique :



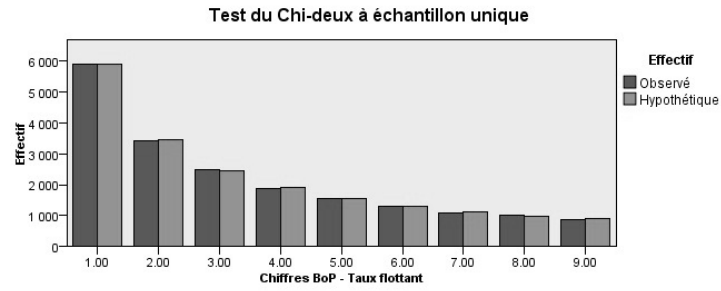
N Total	31,736
Statistique de test	20.655
Degrés de liberté	8
Sig. asymptotique (test bilatéral)	.008

1. Il y a 0 cellules (0%) avec des valeurs attendues inférieures à 5. La valeur minimum attendue est 1,459.856.

Situation étudiée : Tous les Etats — Taux fixe — Taux flottant

P-valeur :

Conclusions statistique et éventuellement stratégique :



N Total	19,545
Statistique de test	4,766
Degrés de liberté	8
Sig. asymptotique (test bilatéral)	.782

1. Il y a 0 cellules (0%) avec des valeurs attendues inférieures à 5. La valeur minimum attendue est 699,070.

Situation étudiée : Tous les Etats — Taux fixe — Taux flottant

P-valeur :

Conclusions statistique et éventuellement stratégique :

Détection de tricheries comptables grâce à la loi de Benford (partie 2)

On considère les mêmes données que dans la partie 1 de cet exercice et on ajoute un indice d'efficacité gouvernementale, pouvant prendre des valeurs entre 0 % et 100 % ; plus les valeurs sont grandes, meilleure est l'efficacité. On effectue alors les calculs suivants.

Tableau croisé : Mesure de l'efficacité gouvernementale * Type de régime monétaire

			Type de régime monétaire		Total
			Taux flottant	Taux fixe	
Mesure de l'efficacité gouvernementale	0 -- 25	Effectif	538	3961	4499
		Effectif théorique	1929,4	2569,6	4499,0
	25,1 -- 50	Effectif	3124	6358	9482
		Effectif théorique	4066,3	5415,7	9482,0
	50,1 -- 75	Effectif	3358	8077	11435
		Effectif théorique	4903,9	6531,1	11435,0
	75,1 -- 100	Effectif	11670	6496	18166
		Effectif théorique	7790,4	10375,6	18166,0
Total	Effectif	18690	24892	43582	
	Effectif théorique	18690,0	24892,0	43582,0	

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	6374,954 ^a	3	,000
Rapport de vraisemblance	6683,830	3	,000
Association linéaire par linéaire	5172,880	1	,000
Nombre d'observations valides	43582		

a. 0 cellules (.0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 1929,38.

Etudiez les résultats obtenus, selon la grille d'analyse ci-dessous.

— Précisez le nom du test mis en œuvre.

— Indiquez les hypothèses testées (avec des mots).

— Le test est-il validement appliqué ?

— Quelle est la P -valeur obtenue ? Formez une conclusion statistique.

— [Seulement s'il vous reste du temps] En termes économiques, que vient-on de mettre en évidence ? Expliquez les interactions possibles de ce résultat avec les conclusions tirées des études d'authenticité des données en fonction du régime monétaire.

Titanic

Le naufrage du Titanic a été un grand drame humain. Mais certains s'en seraient-ils mieux sortis que d'autres ? Considérez les données suivantes (ce sont les vraies données!) :

Tableau croisé Classe * Survécu

			Survécu		Total
			Non	Oui	
Classe	Membres d'équipage	Effectif	673	212	885
		Effectif théorique	599,1	285,9	885,0
	Première classe	Effectif	122	203	325
		Effectif théorique	220,0	105,0	325,0
	Deuxième classe	Effectif	167	118	285
		Effectif théorique	192,9	92,1	285,0
	Troisième classe	Effectif	528	178	706
		Effectif théorique	477,9	228,1	706,0
	Total	Effectif	1490	711	2201
		Effectif théorique	1490,0	711,0	2201,0

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	190,401 ^a	3	,000
Rapport de vraisemblance	180,901	3	,000
Association linéaire par linéaire	,000	1	,991
Nombre d'observations valides	2201		

a. 0 cellule (,0%) a un effectif théorique inférieur à 5. L'effectif théorique minimum est de 92,06.

Etudiez les résultats obtenus par le traitement statistique, selon la grille d'analyse ci-dessous.

— Précisez le nom du test mis en œuvre.

— Indiquez les hypothèses testées (avec des mots).

— Le test est-il validement appliqué ?

— Quelle est la P -valeur obtenue ? Encadrez-la dans un des tableaux ci-dessus. Formez une conclusion statistique.

— Que vient-on de démontrer ? Ecrivez une phrase grand public décrivant un sentiment ou formant un jugement sur la nature humaine.

Sociologie des habitudes de lecture

Cet exercice s'appuie sur les tableaux reproduits à la page suivante. Ils reposent sur l'enquête permanente de l'INSEE sur les modes de vie des ménages et s'intéressent plus particulièrement aux nombres de livres lus par les Français au cours de l'année écoulée ; ils les catégorisent à cet effet selon la CSP (catégorie socio-professionnelle). Nous allons exploiter ces tableaux selon le canevas suivant.

- Combien de sondés a-t-on interrogés ? Parmi eux, combien d'agriculteurs ?
- Précisez le nom des tests mis en œuvre.
- Indiquez pourquoi il ne faut exploiter que le second tableau et précisez comment il a été obtenu à partir du premier tableau. Pourra-t-on dire quelque chose du nombre de livres lus par les agriculteurs ?
- Indiquez les hypothèses testées (avec des mots).
- Quelle est la P -valeur obtenue ?
Entourez également l'endroit où vous lisez sa valeur.
- Formez une conclusion statistique (dont une phrase non technique compréhensible par tou-te-s).
- Supposez que vous soyez une éditrice : formez une conclusion stratégique (quelle action mener?).

Tableau croisé PCS * Lecture

			Lecture			Total
			Un livre par mois ou plus	Moins d'un livre par mois	Aucun livre dans l'année écoulée	
PCS	Agriculteurs	Effectif	3	5	10	18
		Effectif théorique	2,8	7,6	7,6	18,0
	Artisans, commerçants, chefs d'entreprise	Effectif	6	18	25	49
		Effectif théorique	7,7	20,7	20,6	49,0
	Cadres et professions intellectuelles supérieures	Effectif	36	70	23	129
		Effectif théorique	20,3	54,5	54,3	129,0
	Professions intermédiaires	Effectif	35	102	58	195
		Effectif théorique	30,6	82,3	82,1	195,0
	Employés	Effectif	37	117	94	248
		Effectif théorique	39,0	104,7	104,4	248,0
	Ouvriers (y compris ouvriers agricoles)	Effectif	9	56	131	196
		Effectif théorique	30,8	82,7	82,5	196,0
	Retraités	Effectif	76	162	221	459
		Effectif théorique	72,1	193,7	193,1	459,0
	Autres inactifs	Effectif	34	104	70	208
		Effectif théorique	32,7	87,8	87,5	208,0
Total	Effectif	236	634	632	1502	
	Effectif théorique	236,0	634,0	632,0	1502,0	

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	121,562 ^a	14	,000
Rapport de vraisemblance	127,370	14	,000
Association linéaire par linéaire	10,555	1	,001
Nombre d'observations valides	1502		

a. 1 cellules (4,2%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 2,83.

Tableau croisé PCS * Lecture

			Lecture			Total
			Un livre par mois ou plus	Moins d'un livre par mois	Aucun livre dans l'année écoulée	
PCS	Artisans, commerçants, chefs d'entreprise	Effectif	6	18	25	49
		Effectif théorique	7,7	20,8	20,5	49,0
	Cadres et professions intellectuelles supérieures	Effectif	36	70	23	129
		Effectif théorique	20,3	54,7	54,1	129,0
	Professions intermédiaires	Effectif	35	102	58	195
		Effectif théorique	30,6	82,7	81,7	195,0
	Employés	Effectif	37	117	94	248
		Effectif théorique	38,9	105,1	103,9	248,0
	Ouvriers (y compris ouvriers agricoles)	Effectif	9	56	131	196
		Effectif théorique	30,8	83,1	82,2	196,0
	Retraités	Effectif	76	162	221	459
		Effectif théorique	72,1	194,5	192,4	459,0
	Autres inactifs	Effectif	34	104	70	208
		Effectif théorique	32,7	88,2	87,2	208,0
Total	Effectif	233	629	622	1484	
	Effectif théorique	233,0	629,0	622,0	1484,0	

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	120,019 ^a	12	,000
Rapport de vraisemblance	125,622	12	,000
Association linéaire par linéaire	13,239	1	,000
Nombre d'observations valides	1484		

a. 0 cellules (0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 7,69.

Couleurs de cheveux en Ecosse

Les données suivantes ont été recueillies par un des pères fondateurs de la statistique, Fisher, dans un district écossais.

Tableau croisé Sexe * Couleur de cheveux

			Couleur de cheveux					Total
			Blond	Roux	Châtain	Brun	Noir de jais	
Sexe	Hommes	Effectif	592	119	849	504	36	2100
		Effectif théorique	614,4	116,8	825,3	516,5	27,0	2100,0
	Femmes	Effectif	544	97	677	451	14	1783
		Effectif théorique	521,6	99,2	700,7	438,5	23,0	1783,0
Total		Effectif	1136	216	1526	955	50	3883
		Effectif théorique	1136,0	216,0	1526,0	955,0	50,0	3883,0

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	10,467 ^a	4	,033
Rapport de vraisemblance	10,756	4	,029
Association linéaire par linéaire	1,722	1	,189
Nombre d'observations valides	3883		

a. 0 cellules (,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 22,96.

Répondez de manière concise aux questions suivantes.

- Précisez le nom du test mis en œuvre.
- Indiquez les hypothèses testées (uniquement avec des mots).
- Le test est-il validement appliqué ?
- Quelle est la P -valeur obtenue ?
- Formez une conclusion statistique (mais pas, pour une fois, de conclusion stratégique, à moins que vraiment, vous n'ayez une idée géniale de telle conclusion...).
- Si (et seulement si) il vous reste du temps, expliquez à côté du tableau comment les “effectifs théoriques” ont été obtenus pour les hommes blonds.

Etude sociologique des configurations familiales

L'enquête «Histoire de vie» a été réalisée en 2003 par l'INSEE sur 8 403 sondés, que l'on interroge en particulier sur leurs contextes familiaux, loisirs, sentiments d'identité, etc. Les tableaux suivants s'intéressent aux données recueillies suivantes : la catégorie socio-professionnelle [CSP] de celui que les sondés considèrent comme leur père (lorsqu'un tel père existe) et le fait pour les sondés d'avoir ou non été élevés par leur mère seule jusqu'à 18 ans (les autres contextes familiaux étant par exemple d'avoir été élevés par leurs deux parents, leur père seul, un parent et un beau-parent, etc.).

			Elevé par		Total
			mère seule	autre	
CSP père	Agriculteur	Effectif	22	990	1012
		Effectif théorique	56,9	955,1	1012,0
	Indépendant	Effectif	50	801	851
		Effectif théorique	47,8	803,2	851,0
	Cadre	Effectif	60	572	632
		Effectif théorique	35,5	596,5	632,0
	Intermédiaire	Effectif	57	800	857
		Effectif théorique	48,2	808,8	857,0
	Employé	Effectif	50	690	740
		Effectif théorique	41,6	698,4	740,0
	Ouvrier	Effectif	161	2861	3022
		Effectif théorique	169,9	2852,1	3022,0
	Total	Effectif	400	6714	7114
		Effectif théorique	400,0	6714,0	7114,0

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	44,629 ^a	5	,000
Rapport de vraisemblance	48,111	5	,000
Association linéaire par linéaire	3,243	1	,072
Nombre d'observations valides	7114		

a. 0 cellules (.0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 35,54.

- Précisez le nom du test vu en cours et mis en œuvre ici.
 - Indiquez les hypothèses testées (avec des mots).
 - Le test est-il validement appliqué?
 - Quelle est la P -valeur obtenue? Entourez-la également dans les tableaux ci-dessus.
 - Formez une conclusion statistique (une phrase sur le statut de H_0 et une phrase *non technique* compréhensible par tou-te-s et commençant par «Ces données...»).
- On ne demande pas ici de conclusion stratégique mais il faudra le cas échéant effectuer un retour précis sur les données.*

La clientèle des palaces parisiens

Cet exercice est inspiré de données réelles d'une étude de l'Office du tourisme et des congrès de Paris, 2010. La France et Paris sont des destinations touristiques de premier plan. En 2010, une étude a été effectuée auprès de trois établissements d'hôtellerie de luxe afin de mieux connaître leur clientèle et de pouvoir la comparer, le cas échéant : en l'occurrence, les palaces Plaza Athénée, Ritz et Fouquet's Barrière. Les données recueillies et leur traitement sont précisés dans les tableaux ci-dessous.

Tableau croisé Origine * Hôtel

			Hôtel			Total
			Plaza Athénée	Ritz	Fouquet's Barrière	
Origine	Amérique du Nord	Effectif	20	23	5	48
		Effectif théorique	19,2	14,6	14,2	48,0
	Amérique du Sud	Effectif	15	6	9	30
		Effectif théorique	12,0	9,1	8,9	30,0
	Grande-Bretagne	Effectif	20	7	10	37
		Effectif théorique	14,8	11,3	11,0	37,0
	Europe (dont France)	Effectif	20	15	11	46
		Effectif théorique	18,4	14,0	13,6	46,0
	Moyen-Orient	Effectif	10	11	23	44
		Effectif théorique	17,6	13,4	13,0	44,0
	Russie	Effectif	8	6	11	25
		Effectif théorique	10,0	7,6	7,4	25,0
	Asie	Effectif	4	6	3	13
		Effectif théorique	5,2	4,0	3,9	13,0
	Total	Effectif	97	74	72	243
		Effectif théorique	97,0	74,0	72,0	243,0

Tests du khi-deux

	Valeur	ddl	Sig. approx. (bilatérale)
khi-deux de Pearson	32,234 ^a	12	,001
Rapport de vraisemblance	32,646	12	,001
Association linéaire par linéaire	9,370	1	,002
N d'observations valides	243		

a. 2 cellules (9,5%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 3,85.

Tableau croisé Origine * Hôtel

			Hôtel			Total
			Plaza Athénée	Ritz	Fouquet's Barrière	
Origine	Amérique du Nord	Effectif	20	23	5	48
		Effectif théorique	19,4	14,2	14,4	48,0
		% dans Origine	41,7%	47,9%	10,4%	100,0%
	Amérique du Sud	Effectif	15	6	9	30
		Effectif théorique	12,1	8,9	9,0	30,0
		% dans Origine	50,0%	20,0%	30,0%	100,0%
	Grande-Bretagne	Effectif	20	7	10	37
		Effectif théorique	15,0	10,9	11,1	37,0
		% dans Origine	54,1%	18,9%	27,0%	100,0%
	Europe (dont France)	Effectif	20	15	11	46
		Effectif théorique	18,6	13,6	13,8	46,0
		% dans Origine	43,5%	32,6%	23,9%	100,0%
	Moyen-Orient	Effectif	10	11	23	44
		Effectif théorique	17,8	13,0	13,2	44,0
		% dans Origine	22,7%	25,0%	52,3%	100,0%
	Russie	Effectif	8	6	11	25
		Effectif théorique	10,1	7,4	7,5	25,0
		% dans Origine	32,0%	24,0%	44,0%	100,0%
	Total	Effectif	93	68	69	230
		Effectif théorique	93,0	68,0	69,0	230,0
		% dans Origine	40,4%	29,6%	30,0%	100,0%

Tests du khi-deux

	Valeur	ddl	Sig. approx. (bilatérale)
khi-deux de Pearson	30,604 ^a	10	,001
Rapport de vraisemblance	31,152	10	,001
Association linéaire par linéaire	11,218	1	,001
N d'observations valides	230		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 7,39.

- ☐ Nom du test mis en œuvre et jeu d'hypothèses testées (en mots)

- ☐ Qu'est-ce qui distingue les deux applications du test ? Lesquelles sont valides, lesquelles ne le sont pas ?

- ☐ Précisez la (les) P -valeur(s) obtenue(s).

- ☐ Conclusion statistique (non technique, pouvant être lue par tout destinataire du rapport de l'office de tourisme)

- ☐ Refaisons les calculs : comment obtient-on l'effectif attendu de touristes d'Amérique du Nord pour le Ritz ?

Publicité et sexisme

Cet exercice est inspiré d'un article de recherche de Leslie Zebrowitz McArthur et Beth Resko, "The portrayal of men and women in american television commercials".

Les auteurs s'inscrivent dans un projet général consistant à évaluer la prévalence des stéréotypes sexistes. Ils s'intéressent en particulier aux publicités à la télévision. Ils recueillent ainsi sur les trois plus grandes chaînes américaines des publicités à différentes heures du jour. Ils commencent par coder le personnage principal de la publicité selon que c'est un homme, une femme (seules ces deux catégories de publicités seront prises en compte dans l'étude) ; ou autre chose (un animal par exemple). L'objet de l'article est de déterminer si les publicités diffèrent selon divers critères selon que le personnage principal est masculin ou féminin :

- Les lieux choisis sont-ils différemment répartis (entre maison et bureau, campagne et ville, etc.) ?
- Les arguments utilisés sont-ils différents (scientifiques, liés au plaisir, à une menace, etc.) ?
- Le personnage est-il défini par lui-même ou dans son rôle par rapport à d'autres (époux, parents, enfant, employé, etc.) ?

Dans cet exercice on s'intéresse au produit vanté par la publicité. Les produits sont répartis en quatre catégories : produits de soins corporels, produits ménagers, alimentation, autres. Le résultat des observations est résumé dans les tableaux joints.

Tableau croisé Type * Sexe

			Sexe		Total
			Femmes	Hommes	
Type	Soins corporels	Effectif	36	21	57
		Effectif théorique	32,5	24,5	57,0
	Produits ménagers	Effectif	33	20	53
		Effectif théorique	30,2	22,8	53,0
	Alimentation	Effectif	16	29	45
		Effectif théorique	25,6	19,4	45,0
	Autres catégories	Effectif	34	20	54
		Effectif théorique	30,7	23,3	54,0
Total		Effectif	119	90	209
		Effectif théorique	119,0	90,0	209,0

Tests du khi-deux

	Valeur	ddl	Sig. approx. (bilatérale)
khi-deux de Pearson	10,703 ^a	3	,013
Rapport de vraisemblance	10,660	3	,014
Association linéaire par linéaire	,657	1	,417
N d'observations valides	209		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 19,38.

- Quel est le test mis en œuvre ici ?
- Énoncer en mots les hypothèses testées.
- Ce test est-il appliqué de façon valide ?
- Quelle est la P -valeur obtenue ?
- Formuler une conclusion statistique ; puis écrire une phrase dans un style grand public qui expliquerait ce que les auteurs ont mis ou pas en évidence.

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 ?

☐ ☐ Conclusion : montant de la pré-indemnisation (détails du calcul et résultat)

☐ Niveau de confiance associé au résultat de la question précédente

Méthode 2

☐ Population visée, paramètre d'intérêt, résumés des données

☐ Justification de la forme de l'intervalle de confiance à calculer sur ce paramètre

☐ Valeur de cet intervalle (détails du calcul, résultat, et indication d'un niveau de confiance)

☐ Conclusion : montant estimé de la perte totale et donc de la pré-indemnisation (détails du calcul, résultat, et indication d'un niveau de confiance)

Conclusion générale. Comparez les deux méthodes quant à leur méthodologie : quelles sont les hypothèses implicitement considérées lors de la mise en œuvre de chaque méthode? C'est le moment de présenter la modélisation associée à la méthode 1 et son principe de fonctionnement.

☐ Réponse :

Décompte de la population du royaume selon Laplace

Nous allons nous intéresser au décompte de la population tel qu'on l'effectuait au 18ème siècle. Un mémoire de Laplace à destination de l'Académie des Sciences, *Sur les naissances, les mariages et les morts à Paris depuis 1771 jusqu'en 1784, et dans toute l'étendue de la France, pendant les années 1781 et 1782*, revient sur l'utilisation des registres de naissances tenus dans les paroisses pour l'estimation de la population totale. On n'avait alors aucun décompte exhaustif de cette dernière mais les dits registres de naissance étaient, eux, soigneusement tenus depuis que François Ier l'avait ordonné, en août 1539.

On pouvait donc connaître le nombre total de naissances d'une année donnée en France et on voulait, par multiplication, en déduire la taille de la population totale. Le facteur de multiplication M_0 était appelé "facteur multiplicatif des naissances".

1. De nos jours, on s'intéresse plutôt à l'inverse de cette quantité, c'est-à-dire au rapport $1/M_0 = p_0 \in [0, 1]$ des naissances sur la taille de la population. Comment appelle-t-on p_0 ? (0.5 point)

L'objet du mémoire de Laplace était de déterminer comment procéder pour estimer précisément ce facteur M_0 . L'objectif était qu'il fournisse avec bonne confiance une estimée de la taille de la population à plus ou moins 500 000 habitants. A cet effet, Laplace, après calculs, a proposé de s'intéresser aux naissances issues d'un échantillon de 1 000 000 d'habitants, constitué à partir de paroisses représentant la diversité du royaume. Constituer cet échantillon n'aurait pas été une tâche facile mais c'est le cadre de pensée que Laplace se fixe pour son mémoire à l'Académie.

2. Supposons formé un tel échantillon de 1 000 000 d'habitants, et supposons qu'on ait reporté 38 500 naissances pour eux tous. Effectuer la modélisation correspondant à ces données. (2 points)

Les calculs du mémoire de Laplace se fondaient sur une version du théorème limite central pour les lois de Bernoulli, que l'on a appelé par la suite le théorème de Moivre–Laplace; ils donnent lieu aux formules que nous avons vues en cours (et qu'il ne s'agit pas de re-démontrer, mais simplement d'appliquer ici).

3. Calculer une estimation par intervalle de p_0 . On justifiera la forme de l'intervalle retenu. (2.5 points)

4. En déduire une estimation par intervalle du facteur multiplicatif des naissances M_0 , puis une estimation par intervalle de la population française totale en 1781; commenter le résultat obtenu. On consultera le tableau reporté par Laplace dans son mémoire. (1.5 points)

5. [Cette question, contrairement aux précédentes, s'intéresse à une situation fictive.] Imaginons que Turgot, qui était alors en charge des finances du royaume pour Louis XVI, ait demandé par ailleurs un mémoire estimant la consommation moyenne de sel par habitant et par an (résultat : 5.2 ± 0.4 kilogrammes, au niveau de confiance standard). On peut alors estimer la quantité totale de sel consommée dans le royaume chaque année.

– Déterminer un intervalle de confiance, de niveau de confiance à préciser, sur cette quantité totale. (1.5 points)

– Pourquoi Turgot s'intéresse-t-il à cette quantité? (question bonus)

ET LES MORTS, ETC.

45

Population du Royaume, l'île de Corse comprise, suivant l'ordre des généralités.
pendant l'année 1781.

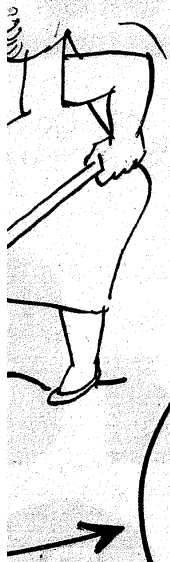
NUMÉROS qui consistent l'ordre des généralités et provinces.	DÉNOMINATION des généralités du Royaume, l'île de Corse comprise, distinguées en pays d'Élections et en pays d'États, la ville de Paris étant distinguée de la généralité comme capitale du Royaume.	NAISSANCES.	MARIAGES.	PROFENSIONS en religion.	MORTS			EXCÉDENT des naissances sur les morts.
					dans la société civile.	en religion.	Total des morts.	
	PARIS (ville).....	20 232	4 970	87	20 037	123	20 160	+ 52
	<i>Généralités en pays d'Élections.</i>							
1	Paris.....	44 451	10 210	53	42 994	87	43 081	+ 1 370
2	Orléans.....	26 294	6 641	25	28 870	58	28 928	+ 1 634
3	Tours.....	49 334	12 593	59	53 243	95	53 338	+ 4 004
4	Poitiers.....	27 377	7 523	21	27 468	38	27 506	+ 29
5	Bourges.....	20 440	4 920	29	20 897	25	20 922	+ 452
6	Limoges.....	26 181	7 433	30	22 840	24	22 864	+ 3 317
7	La Rochelle.....	17 027	4 612	22	21 211	22	21 233	+ 4 206
8	Bordeaux.....	54 802	14 924	48	44 732	65	44 797	+ 10 005
9	Auch.....	34 527	8 469	27	27 037	24	27 061	+ 7 466
10	Montauban.....	21 569	5 296	13	19 971	24	19 995	+ 1 574
11	Grenoble.....	27 338	6 250	31	20 848	39	20 887	+ 6 451
12	Lyon.....	24 624	5 823	30	19 983	59	20 042	+ 4 582
13	Riom.....	27 761	6 815	44	18 693	58	18 751	+ 9 010
14	Moulins.....	25 067	6 996	36	23 168	27	23 195	+ 1 872
15	Châlons.....	30 925	7 238	23	29 965	12	29 977	+ 948
16	Le Clermontois.....	1 459	317	"	1 212	"	1 212	+ 247
17	Soissons.....	16 580	3 889	15	16 099	28	16 127	+ 147
18	Amiens.....	10 598	5 044	14	20 761	40	20 801	+ 203
19	Rouen.....	27 801	7 765	51	27 297	87	27 384	+ 417
20	Caen.....	24 719	6 067	47	22 495	62	22 557	+ 2 162
21	Alençon.....	18 799	4 954	33	19 117	26	19 143	+ 344
	<i>Généralités en pays d'États.</i>							
22	Rennes.....	91 330	22 920	100	88 537	171	88 708	+ 2 622
23	Perpignan.....	7 514	1 727	1	7 030	6	7 036	+ 458
24	Montpellier.....	71 099	15 849	78	51 824	93	51 917	+ 19 182
25	Aix.....	27 846	5 698	33	21 961	68	22 029	+ 5 817
26	Dijon.....	42 488	10 216	72	41 148	98	41 246	+ 1 242
27	Besançon.....	27 614	6 110	31	21 760	54	21 814	+ 5 800
28	Strasbourg.....	25 312	5 613	31	19 668	50	19 718	+ 6 194
29	Metz.....	13 129	2 597	25	11 948	55	12 003	+ 1 176
30	Nancy.....	32 052	6 647	84	28 277	89	28 366	+ 3 686
31	Valenciennes.....	10 798	2 506	43	7 694	51	7 745	+ 3 053
32	Lille.....	28 898	6 886	147	26 435	189	26 624	+ 1 774
33	Île de Corse.....	4 921	985	18	3 940	21	3 961	+ 960
	Résultats du Royaume, l'île de Corse comprise.....	970 406	236 503	1 400	879 170	968	881 138	+ 89 268

Observations. — Dans la colonne de l'excédent des naissances sur les morts, le signe + indique que le nombre des naissances surpasse celui des morts, et le signe — indique que le nombre des morts surpasse celui des naissances.

Les généralités d'Orléans, de Tours, de Poitiers, de Bourges, de La Rochelle, de Soissons, d'Amiens et d'Alençon ont été atteintes d'épidémies et de maladies qui y ont occasionné une mortalité considérable, puisque le nombre des décès surpasse celui des naissances; mais cependant le résultat de toutes les généralités présente un Tableau satisfaisant, puisque le nombre total des naissances surpasse celui des morts de 89,268.

Les sondeurs se contredisent-ils ?

Commentez brièvement cet extrait du *Canard enchaîné* (date : août 2010).



ards

minimares

- **Noté**, ce titre du « Monde » (22/8) : « Niches fiscales : M. Sarkozy annonce 10 milliards d'euros de hausses d'impôts ». Même avec les niches, il faut appeler un chat un chat.
- **Ce titre**, dans « Les Echos » (20/8) : « Voté il y a trois ans, le paquet fiscal est resté un talon d'Achille pour la majorité ». Un talon menacé de rupture.
- **Le « JDD »** (22/8) consacre un quart de page à expliquer la hausse de 2 points de la cote de Sarko, selon l'Ifop. Manque de bol, un sondage Viavoix pour « Libé » (23/8), réalisé aux mêmes dates, le donne en baisse de 1 point. Il en faudrait, des tiers de page, pour expliquer les contradictions entre sondeurs !
- **Cours de fiscalité** donné par le secrétaire général de l'UMP, Xavier Bertrand (« Le Parisien », 21/8) : « Baisser la réduction d'impôts de certains, ce n'est pas augmenter les impôts pour tous. » Ni toucher
- **Ségolène Royal** a rope 1, 21/8) les p Sarko sur la sécurité fie d'« agitation frère cace ». Seulement :
- **La même Ségolène** Figaro », 18/8) : « J sur moi et mon père les hommes, donc voir, donc moi fille forcément castrati que « je » parle de
- **De Vincent Peillon** péen PS (« Le Par « On a laissé se con velle oligarchie, mé fonctionnaires pass nets ministériels, liés litique, faisant des fo dans le privé. » Pei parler aux anciens cabinets Fabius ou aujourd'hui salariés c nault.
- **En marge des jour** écologistes, Denis

mbe !

omme. C'est t à l'intéressé,

L'exemple historique de Galton : taille des fils et taille des pères

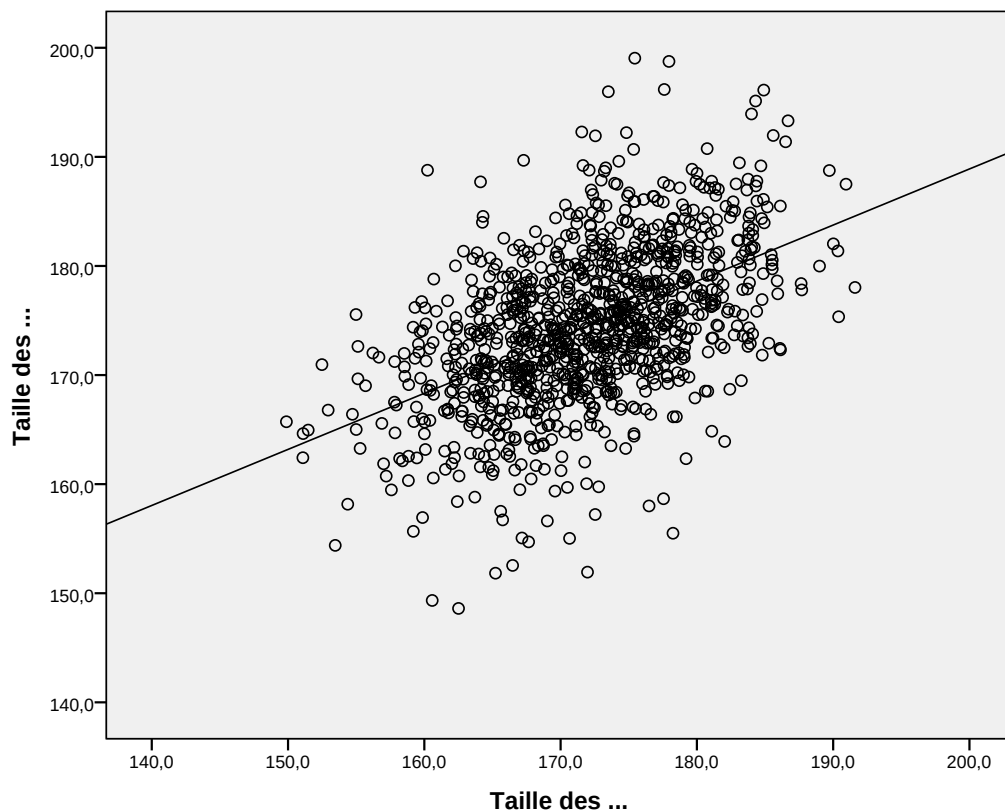
Sir Galton (homme de science britannique, 1822–1911) étudiait la taille des fils en fonction de la taille des pères. Il a noté un retour vers un comportement moyen : les pères plus grands que la moyenne donnaient naissance à des fils plus petits qu'eux, et les pères plus petits que la moyenne donnaient naissance à des fils plus grands qu'eux.

Voici une représentation graphique et quelques statistiques descriptives correspondant aux données qu'il a recueillies.

Faisons connaissance avec les données...

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Taille des pères	1078	149,9	191,6	171,925	6,9720
Taille des fils	1078	148,6	199,0	174,458	7,1493
N valide (listwise)	1078				



Il a ensuite effectué une analyse de régression linéaire, reproduite à la page suivante.

Sortie de régression #1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,501 ^a	,251	,251	6,1889

a. Valeurs prédites : (constantes), Taille des pères

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	13835,971	1	13835,971	361,235	,000 ^a
	Résidu	41212,829	1076	38,302		
	Total	55048,800	1077			

a. Valeurs prédites : (constantes), Taille des pères

b. Variable dépendante : Taille des fils

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	95,0% % intervalles de confiance pour B	
		A	Erreur standard	Bêta			Borne inférieure	Limite supérieure
1	(Constante)	86,072	4,654		18,493	,000	76,940	95,204
	Taille des pères	,514	,027	,501	19,006	,000	,461	,567

a. Variable dépendante : Taille des fils

Sortie de régression #2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,501 ^a	,251	,251	6,0353

a. Valeurs prédites : (constantes), Taille des fils

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	13157,942	1	13157,942	361,235	,000 ^a
	Résidu	39193,204	1076	36,425		
	Total	52351,146	1077			

a. Valeurs prédites : (constantes), Taille des fils

b. Variable dépendante : Taille des pères

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	95,0% % intervalles de confiance pour B	
		A	Erreur standard	Bêta			Borne inférieure	Limite supérieure
1	(Constante)	86,633	4,491		19,289	,000	77,820	95,446
	Taille des fils	,489	,026	,501	19,006	,000	,438	,539

a. Variable dépendante : Taille des pères

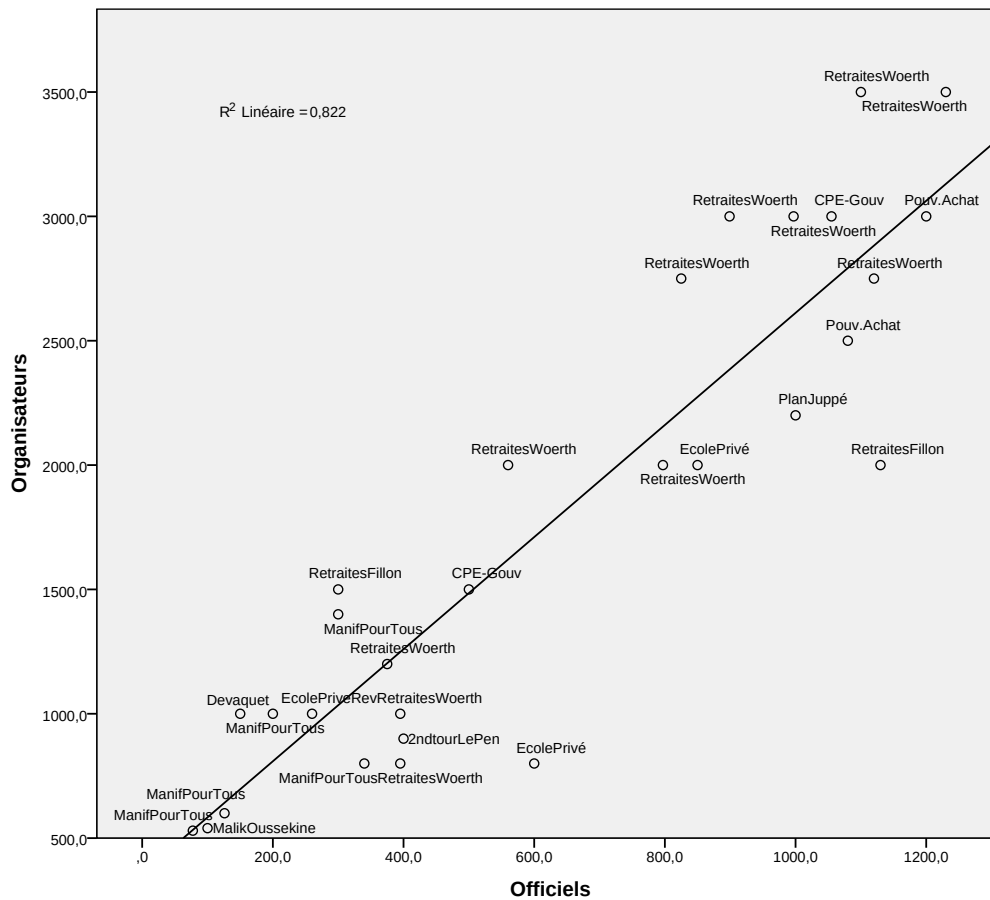
Répondez très concisément aux questions suivantes, en numérotant vos réponses.

1. Quelle est la variable à expliquer, quelle est la variable explicative? Compléter en conséquence les étiquettes des axes du graphique de l'énoncé.
2. Par conséquent, quelle sortie **SPSS** parmi les deux reproduites faut-il lire?
3. Y a-t-il une relation linéaire significative entre la variable à expliquer et la variable explicative?
4. Préciser une mesure de la qualité du modèle linéaire étudié et la commenter.
5. Ecrire la relation linéaire proposée par la sortie **SPSS** retenue.
6. Transformer un peu la présentation de cette relation afin de faire apparaître l'assertion de Galton, selon laquelle on observe un retour à la moyenne.
7. Interpréter les coefficients de la relation ainsi obtenue; Galton avait-il raison, i.e., observe-t-on un retour à la moyenne statistiquement significatif?

Décomptes de manifestants en France, 1984–2014

Le tableau ci-dessous compare, pour différentes manifestations de la période 1984 – 2014, les décomptes officiels (ceux effectués par les autorités) et les décomptes fournis par les organisateurs des manifestations. Que peut-on conclure du traitement de régression linéaire simple reproduit à la page suivante ?

Date	Cause	Org.	Officiels
4 mars 1984	EcolePrivé	800.0	600.0
24 juin 1984	EcolePrivé	2 000.0	850.0
4 décembre 1986	Devaquet	1 000	200
10 décembre 1986	MalikOussekiné	600	126
16 janvier 1994	EcolePriveRev	1 000	260
12 décembre 1995	PlanJuppé	2 200	1 000
1 mai 2002	2ndtourLePen	900	400
13 mai 2003	RetraitesFillon	2 000	1 130
3 juin 2003	RetraitesFillon	1 500	300
18 mars 2006	CPE-Gouv	1 500	500
28 mars 2006	CPE-Gouv	3 000	1 055
28 mai 2009	Pouv.Achat	2 500	1 080
19 mars 2009	Pouv.Achat	3 000	1 200
23 mars 2010	RetraitesWoerth	800	395
27 mai 2010	RetraitesWoerth	1 000	395
24 juin 2010	RetraitesWoerth	2 000	797
7 septembre 2010	RetraitesWoerth	2 750	1 120
23 septembre 2010	RetraitesWoerth	3 000	997
2 octobre 2010	RetraitesWoerth	3 000	899
12 octobre 2010	RetraitesWoerth	3 500	1 230
16 octobre 2010	RetraitesWoerth	2 750	825
19 octobre 2010	RetraitesWoerth	3 500	1 100
28 octobre 2010	RetraitesWoerth	2 000	560
6 novembre 2010	RetraitesWoerth	1 200	375
13 janvier 2013	ManifPourTous	800	340
24 mars 2013	ManifPourTous	1 400	300
26 mai 2013	ManifPourTous	1 000	150
2 février 2014	ManifPourTous	540	100
5 octobre 2014	ManifPourTous	530	78



Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,906 ^a	,822	,815	413,4445

a. Prédicteurs : (Constante), Officiels

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	21243842,21	1	21243842,21	124,279	,000 ^b
	Résidus	4615281,926	27	170936,368		
	Total	25859124,14	28			

a. Variable dépendante : Organisateurs

b. Prédicteurs : (Constante), Officiels

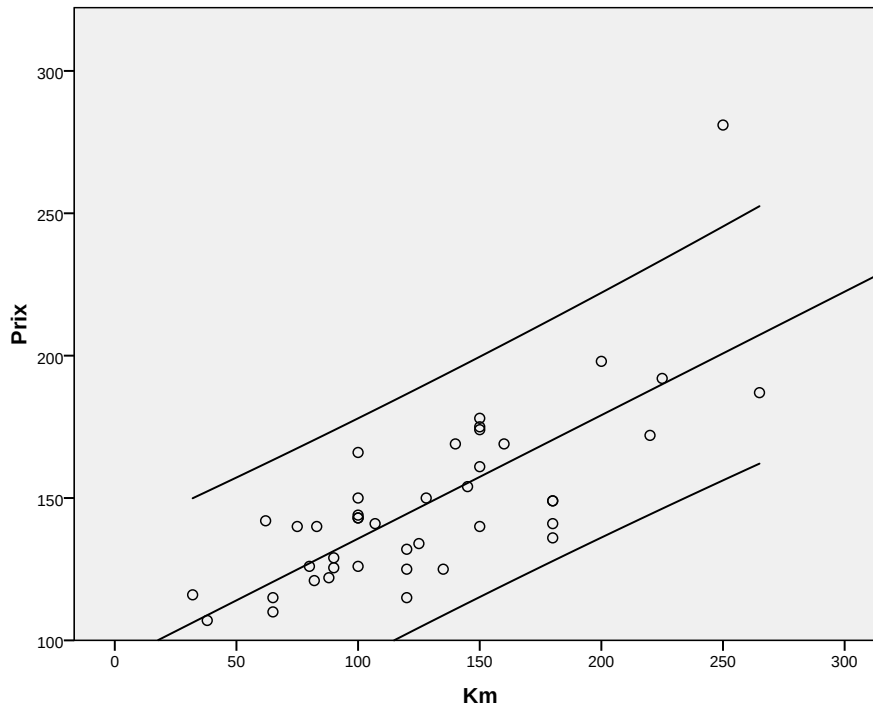
Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	358,197	149,261		2,400	,024
	Officiels	2,254	,202	,906	11,148	,000

a. Variable dépendante : Organisateurs

Forfait de ski en fonction de la taille du domaine

Etudions les données qui consistent, pour 42 stations de ski alpines, en 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.



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			

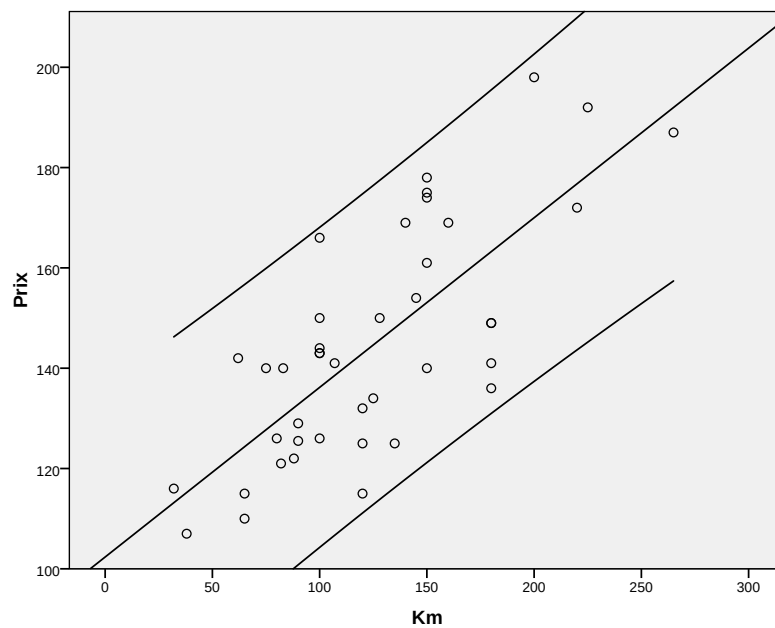
a. Valeurs prédites : (constantes), Km

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

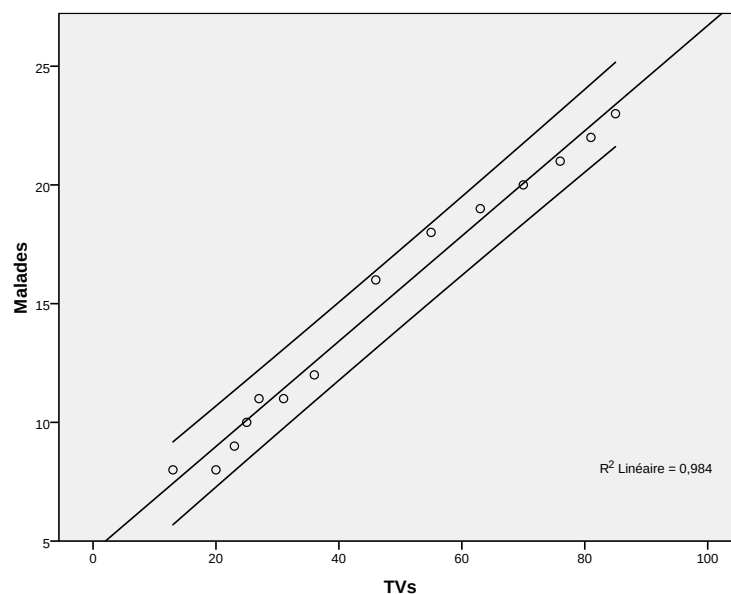


Regarder la télé rend-il fou ?

Les données considérées ici consistent en des triplets : index de l'année, taux d'équipement en téléviseurs de la population (en %), taux de malades mentaux (nombre pour mille habitants). L'étude a été réalisée en Grande-Bretagne.

1. Quelle est d'après vous la variable à expliquer et la variable explicative ?
2. Etudiez le résultat de la régression sous SPSS.
3. Faut-il en conclure que regarder la télé rend fou, et si non, pourquoi ?
4. *Question subsidiaire* : En exploitant le fichier **TV.sav** disponible sur le site web du cours, 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 ?

Un autre exemple montrant qu'une corrélation linéaire n'est pas nécessairement signe de causalité est procuré par un article tiré (d'un blog) du *Monde*, reproduit à la fin de l'énoncé : «Le chocolat engendre-t-il des tueurs en série?»



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
1	(Constante)	4,552	,425		10,707
	TVs	,222	,008	,992	27,258

a. Variable dépendante : Malades

M Blogs

21 novembre 2012, par Pierre Barthélémy

Le chocolat engendre-t-il des tueurs en série ?



Le 10 octobre dernier est parue dans le fort sérieux *New England Journal of Medicine* une étude qui a fait les délices de la presse généraliste. On y apprenait qu'il existait un lien de corrélation extrêmement significatif entre la consommation de chocolat par un pays et le nombre de prix Nobel que ledit pays décrochait. L'information a eu tant plus d'écho qu'elle a été publiée pendant la semaine où les Nobel 2012 étaient décernés. Tout en expliquant qu'une corrélation ne signifie pas forcément un lien de cause à effet, l'auteur de cette étude, Franz Messerli, fait tout pour en dénicher un ! Pour ce cardiologue, tout est dans les flavanols, des molécules antioxydantes présentes dans le cacao, dont plusieurs études ont montré qu'elles améliorent les fonctions cognitives. Tout se tient : les pays où l'on mange beaucoup de chocolat font des habitants plus intelligents et ont donc plus de prix Nobel. La possibilité que le lien de cause à effet soit inversé – dans les pays les plus intelligents et donc les plus comblés en prix Nobel, on sait les vertus bienfaisantes du chocolat sur la santé et on en mange davantage – est, selon Franz Messerli, concevable mais improbable.

Ce genre d'étude fait les choux gras de la presse et plaît beaucoup au public, notamment parce que le mécanisme présenté est à la fois astucieux et simple à comprendre. Il n'empêche : le glissement d'un simple lien de corrélation statistique à un lien de cause à effet est un exercice périlleux. La raison principale tient au fait que si X et Y sont statistiquement corrélés, cela ne signifie pas forcément que l'un provoque l'autre : X et Y peuvent parfaitement être deux conséquences de la même cause. L'exemple le plus fréquemment cité est le suivant : dans une ville, plus les gens achètent de cornets de glace, plus il y a de noyades. Le lien semble "évident" : les baigneurs gavés de glace à la vanille ou à la fraise font des malaises en se baignant et en meurent. C'est aller un peu vite. En réalité, si l'on compte plus de glaces achetées et plus de noyades, c'est avant tout parce qu'il fait plus chaud. C'est parce qu'ils recherchent la fraîcheur que les gens consomment plus de glaces et se baignent davantage que d'ordinaire... Une cause unique produit deux courbes semblables mais sans relation l'une avec l'autre. Enfin, il se peut également qu'une corrélation existe mais qu'elle soit le fait d'une pure coïncidence. Ainsi, on a pu montrer qu'il y avait une corrélation entre la répartition mondiale d'un arbre, le gommier rouge, et celles des langues tonales.

Cette étude tendant à trouver un lien de cause à effet entre consommation de chocolat et prix Nobel a fait des émules critiques, en la personne de James Winters et Sean Roberts, deux jeunes chercheurs britanniques spécialisés en dans la science du langage et la cognition. Publiée sur leur site Internet, replicatedtypo, leur réponse à Franz Messerli est plutôt amusante. Les deux hommes ont repris la méthodologie du cardiologue et trouvé, entre autres, une magnifique corrélation entre la consommation de chocolat d'un pays et... le nombre de tueurs en série que ce pays engendre. Le côté obscur des flavanols vous apparaît subitement... Tout à coup, vous ne regardez plus votre tablette de chocolat ou votre boîte de cacao de la même façon. Va-t-il falloir les interdire ? Faisaient-ils partie du menu de Jack l'Eventreur ? Ou s'agit-il plutôt de la preuve par l'absurde que la méthodologie et les conclusions tirées dans l'étude de Franz Messerli ne sont pas des plus rigoureuses ?

Pour James Winters et Sean Roberts, Franz Messerli a probablement tort lorsqu'il écrit qu'"il est difficile d'identifier un dénominateur commun plausible qui pourrait sous-tendre à la fois la consommation de chocolat et le nombre de lauréats du Nobel". Quand on regarde la liste des nations les plus récompensées (par rapport à leur population), on se rend compte qu'il s'agit essentiellement de pays occidentaux, lesquels ont les moyens de se payer du chocolat, l'ont intégré à leur culture alimentaire depuis des siècles et ont souvent développé une industrie chocolatière. Et c'est le même groupe de pays qui, historiquement, a le plus investi dans la science...

Même si les chercheurs, de par leur formation, sont habitués à se méfier du passage hâtif de la corrélation au lien de cause à effet, il arrive parfois qu'ils se laissent prendre au piège. J'ai ainsi le souvenir d'une étude parue dans *Nature* en 1999 montrant que les petits enfants s'endormant avec la lumière allumée étaient souvent plus myopes que les autres. D'où une campagne médiatique pour prévenir des méfaits des veilleuses. N'ayant pu confirmer ce résultat, d'autres scientifiques se sont attaqués autrement au problème et se sont rendus compte, quelques mois plus tard, que c'était dans les familles où les parents étaient myopes que l'on laissait plus volontiers la lumière dans la chambre des enfants. La myopie étant essentiellement héréditaire, la corrélation s'expliquait d'une tout autre manière...

Pierre Barthélémy (@PasseurSciences sur Twitter)

TABLE 11.2 Examples of Spurious Relationships

Observed Spurious Relationship*	Reason for the Relationship (the Third Variable)
Amount of ice cream sold and deaths by drownings (Moore, 1993)	Season: Ice cream sales and drownings tend to be high during the warm months of the year.
Size of left hand and size of right hand	Genetics: The size of both hands is due to genetic makeup.
Height of sons and height of daughters (Davis, 1985)	Genetics: Heights of sons and daughters are both due to their parents' genetic makeup.
Ministers' salaries and price of vodka	Area (i.e., urban or rural): In urban areas, prices and salaries tend to be higher.
Shoe size and reading performance for elementary school children	Age: Older children have larger shoe sizes and read better.
Number of doctors in region and number of people dying from disease	Population density: In highly dense areas, there are more doctors and more people die.
Number of police officers and number of crimes (Glass & Hopkins, 1996)	Population density: In highly dense areas, there are more police officers and more crimes.
Number of homicides and number of churches	Population density: In highly dense areas, there are more homicides and more churches.
Number of storks sighted and the population of Oldenburg, Germany, over a six-year period (Box, Hunter, & Hunter, 1978)	Time: Both variables were increasing over time.
Number of public libraries and the amount of drug use	Time: Both were increasing during the 1970s.
Teachers' salaries and the price of liquor (Moore and McCabe, 1993)	Time: Both tend to increase over time.
Tea drinking and lung cancer	Smoking: Tea drinkers have a lower risk only because they smoke less.

*All but one of the spurious relationships in the first column shows a positive relationship. That is, as one of the variables increases, the other variable also increases. The one negative relationship is the relationship between tea drinking and lung cancer.

Croissance des orangers

ID	age (jrs)	circonf. (mm)
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 étudie la circonférence d'orangers (en millimètres) en fonction de leur âge (en jours). On obtient les résultats de mesures ci-dessus et la régression linéaire ci-dessous.

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^a

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	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17,400	8,623		2,018	,052
	Age	,107	,008	,914	12,900	,000

a. Dependent Variable: Circumference

Corrigés des extraits d'examens

Disponibles sur le site web du cours

<http://www.hec.fr/stoltz>

Principe des examens

Composés de deux problèmes indépendants :

- l'un portant sur les intervalles de confiance et tests d'hypothèses,
- l'autre sur la régression linéaire simple et multiple.

La validation suppose une réussite minimale de chacun des deux problèmes, c'est-à-dire

- pour le premier, le calcul correct d'au moins un intervalle de confiance et d'au moins une P-valeur ;
- pour l'autre, la lecture et l'écriture correcte d'au moins une sortie de régression (y compris les unités et le terme de résidu), ainsi que la mise en œuvre des concepts de validités statistique et économique, de même que la prévision en un nouveau point et/ou la détection des valeurs atypiques du modèle.

Seul document autorisé : une feuille A4 de résumé (recto-verso) ; rien de plus !

Index des problèmes

Intervalles de confiance et tests d'hypothèses

- | | |
|---|----------|
| — Création d'un club d'achat sur Internet | page 234 |
| — Valeur d'un stock | page 236 |
| — Audit de la Cour des comptes | page 238 |
| — Nouvelle politique de prix pour une entreprise de VPC | page 243 |

Régression linéaire simple et multiple

- | | |
|---|----------|
| — Modélisation du prix du forfait de ski | page 244 |
| — Prévision des résultats des élections présidentielles américaines | page 259 |
| — Emission de polluants par une voiture («scandale Volkswagen») | page 276 |
| — Référendums écologiques en Suisse | page 288 |

Création d'un club d'achat sur Internet

Un entrepreneur a l'idée suivante : créer un club sur Internet, auquel l'adhésion coûtera seulement 5 euros par an et qui permettra d'avoir accès tout au long de l'année à des offres exclusives, négociées avec de grands partenaires. Il lui faut dans un premier temps construire son fichier de clients ; nous étudierons dans un second temps deux exemples représentatifs de l'exploitation qui peut être faite de ce fichier.

Création du fichier clients

L'entrepreneur crée et finance la structure (hébergement du site Internet, investissement informatique, etc.) à partir de ses fonds propres. Il cherche en revanche à construire le fichier de clientèle de manière financièrement neutre, *id est*, uniquement avec les adhésions enregistrées. Or, la location d'une adresse mail dans un fichier clients d'une autre société coûte 40 centimes. Trois tels fichiers sont considérés et sont issus respectivement d'une société de vente par correspondance (VPC), d'un fournisseur d'accès à Internet (FAI) et d'un libraire en ligne.

- (1) Quel est le taux d'adhésion minimum devant être observé sur un fichier clients extérieur afin qu'il contribue de manière financièrement neutre à la création du fichier clients propre ?

L'entrepreneur exploite là encore ses fonds propres pour tenter un coup de sonde sur chacun des trois fichiers clients. Les résultats sont reportés au tableau suivant.

Fichier	Taille totale	Nombre de sondés	Nombre d'adhésions
VPC	200 000	2 000	150
FAI	300 000	2 000	210
Libraire	150 000	2 000	240

On veut déterminer parmi ces trois fichiers, ceux qui peuvent être exploités par l'entrepreneur avec une garantie forte de neutralité financière.

- (2) Ecrire le modèle associé au sondage effectué à partir du fichier VPC (définir la population étudiée, préciser le paramètre d'intérêt, etc.).
- (3) Indiquer si ce fichier doit être exploité ou non : on effectuera à cet effet un test, dont on précisera la P-valeur obtenue sur les données.
- (4) Refaire, de manière très concise, les calculs pour les deux autres fichiers et conclure sur leur sort (toujours *via* l'indication de P-valeurs).

On veut également avoir une idée de la taille attendue du fichier clients propre constitué de la sorte.

- (5) Préciser pour chacun des fichiers retenus un intervalle de confiance à 95 % sur le nombre d'adhésions qu'il rapportera.
- (6) Calculer alors un intervalle de confiance sur le nombre total d'adhésions attendues et lui attribuer un niveau de confiance.

Exploitation du fichier clients

Elle peut être de deux types :

- un démarchage classique, qui consiste à demander à un partenaire d'effectuer un geste commercial (réduction tarifaire par exemple) tout en lui faisant payer l'accès au fichier clients du club, en mettant en avant que l'opération sera rentable vu la qualité de ce fichier clients, composé de membres motivés ;

- la création de buzz ou de partenariats prestigieux, qui ont pour but secondaire le développement de l'image du club ; dans ce cas il s'agit de créer des objets en série limitée, uniquement disponibles pour les membres du club et introuvables dans les autres enseignes (traditionnelles ou sur Internet) ; ainsi, une gamme de chemises M&H signées par le couturier Laurent de Saint-Yves.

On étudie tout d'abord un cas de démarchage classique. On veut proposer à un partenaire de consentir une réduction tarifaire importante, de sorte que son taux de marge moyen descend à 5 % sur le chiffre d'affaires ainsi réalisé. Afin de déterminer le prix qu'on lui demandera pour utiliser le fichier, on effectue un sondage sur un panel de 1 000 membres. On note pour chacun d'entre eux le montant qu'il est prêt à dépenser vu cette offre exceptionnelle (on note 0 euro lorsque le membre se déclare non intéressé par l'offre). On obtient un montant moyen de 43.50 euros (écart-type mesuré : 12.70 euros).

- (7) Calculer un intervalle de confiance sur la marge attendue par client pour le partenaire.
- (8) En déduire un prix de location des adresses du fichier. (On raisonnera comme un entrepreneur et on expliquera la fixation du prix à partir des éléments quantitatifs obtenus à la question précédente.)

On étudie ensuite le cas d'un partenariat de prestige, portant sur des chaussures pour femmes. Afin de le négocier au mieux avec une marque de chaussures de luxe, on cherche à savoir, à ± 4 euros près, combien chaque cliente serait prête à dépenser en moyenne. (En effet, pour être crédible lors des discussions, on veut pouvoir exhiber des données : *In God we trust, all others bring data.*)

- (9) On interroge 50 clientes, qui déclarent un montant prévisionnel d'achats de 194.50 euros (écart-type : 75.90 euros). Cet échantillon est-il suffisant ? Combien de clientes supplémentaires doit-on interroger, le cas échéant, pour obtenir la précision désirée ?

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
Ecart	−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 références disponibles au total.

- (2) Modéliser la situation rencontrée (préciser notamment la population visée, le paramètre d'intérêt, etc.).
- (3) Dédire 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 $\pm 1\%$ 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 ?

Audit de la Cour des comptes

Cet exercice repose sur le référé 71261 de la Cour des comptes, daté du 27 novembre 2014. Il porte sur les subventions pour travaux divers d'intérêt local : la fameuse «réserve parlementaire», sur laquelle les parlementaires peuvent faire affecter des aides aux collectivités locales de leur circonscription d'origine, et la moins connue «réserve ministérielle», d'un montant total bien plus faible. Nous étudierons certaines parties de ce rapport, au besoin en inventant des données plausibles (ce sera alors de la politique-fiction ou de la comptabilité-fiction), à défaut de publication des données brutes par la Cour. Les pages 1 et 6 du rapport sont reproduites pour information à la page suivante. Il n'est pas nécessaire de les lire en détails, le sujet en reproduira les phrases importantes. Par exemple, la Cour indique sa méthodologie dans les premières lignes de son rapport :

La Cour a procédé au contrôle de la gestion des subventions pour travaux divers d'intérêt local, en application de l'article L. 111-3 du code des juridictions financières.

Pour ce faire, elle a, d'une part, analysé les données financières de l'ensemble des décisions d'attribution intervenues de 2006 à 2012, soit plus de 60 000 subventions d'un montant total dépassant le milliard d'euros. D'autre part, elle a interrogé vingt-cinq préfetures et s'est rendue sur place dans huit autres, ce qui lui a permis de réaliser un examen approfondi d'un échantillon représentatif de 550 dossiers de subventions attribuées en 2011 et en 2012. En revanche, elle s'est abstenue d'examiner les modalités selon lesquelles les subventions sont réparties entre les parlementaires.

La Cour qualifie d'autorité son échantillon de représentatif, mais sans justifier plus avant ce fait. Nous verrons comment elle aurait pu s'en assurer. La Cour effectue ensuite son analyse en revenant sur divers éléments : conformité de la dépense à l'esprit de cette réserve (mise en œuvre d'une solidarité nationale) et conformité de la procédure d'attribution suivie ; efficacité de la dépense (c'est-à-dire, étude des coûts de gestion) ; et équité des affectations (quelles sont les collectivités locales bénéficiaires?).

Les questions qui suivent ne sont pas nécessairement posées dans l'ordre logique de leur traitement par la Cour, mais plus prosaïquement, par ordre d'importance et de difficulté dans le cadre de notre cours de statistique.

Partie 1 : Questions fondamentales et éliminatoires (3 points)

Dans la période 2006–2012 considérée, sur les 60 000 subventions attribuées, 62,5% l'ont été par des parlementaires issus de la majorité de la chambre (Sénat ou Assemblée nationale) considérée, et 37,5% l'ont été par ceux de l'opposition. Parmi l'échantillon de 550 dossiers considérés, 362 ont été attribués par des parlementaires de la majorité et 188 par ceux de l'opposition.

1. L'échantillon considéré est-il représentatif au sens des attributions majorité / opposition ? On commencera par modéliser la situation considérée. (1.5 points ; une réponse gravement incorrecte est éliminatoire)

La Cour étudie également la conformité réglementaire des subventions attribuées. A propos de celles de l'échantillon, elle déclare :

Les vérifications de la Cour sur 550 dossiers ont montré que, parmi eux, plus de 40 % ne comportaient pas toutes les pièces justificatives requises, portaient sur des dépenses inéligibles ou auraient dû appeler une instruction plus approfondie des services de l'État sur leur conformité.

2. Que peut-elle en déduire sur l'ensemble des dossiers de la période 2006–2012 considérée ? On calculera un certain intervalle de confiance, de forme à justifier, et on en donnera le résultat sous la forme d'une phrase qui aurait pu s'insérer naturellement dans le rapport, à la suite du paragraphe reproduit ci-dessus. On portera en particulier une grande attention au style de rédaction. (1.5 points ; une réponse gravement incorrecte est éliminatoire)

Cour des comptes



Le 27 NOV. 2014

Le Premier président

à

Monsieur Michel Sapin
Ministre des finances et des comptes publics

Monsieur Bernard Cazeneuve
Ministre de l'intérieur

Réf. : n° 71261

Objet : Les subventions pour travaux divers d'intérêt local

La Cour a procédé au contrôle de la gestion des subventions pour travaux divers d'intérêt local, en application de l'article L. 111-3 du code des juridictions financières.

Pour ce faire, elle a, d'une part, analysé les données financières de l'ensemble des décisions d'attribution intervenues de 2006 à 2012, soit plus de 60 000 subventions d'un montant total dépassant le milliard d'euros. D'autre part, elle a interrogé vingt-cinq préfectures et s'est rendue sur place dans huit autres, ce qui lui a permis de réaliser un examen approfondi d'un échantillon représentatif de 550 dossiers de subventions attribuées en 2011 et en 2012. En revanche, elle s'est abstenue d'examiner les modalités selon lesquelles les subventions sont réparties entre les parlementaires.

À l'issue de la contradiction avec les administrations et collectivités concernées, la Cour m'a demandé, en application de l'article R. 143-1 du code précité, d'appeler votre attention sur les principales observations résultant de ses travaux.

Le contrôle de la Cour montre que, au-delà d'un effort de transparence financière récent qu'il faut saluer, les nombreuses difficultés et le coût de gestion de cette procédure particulière justifieraient une réforme profonde de celle-ci. En outre, une réflexion est certainement nécessaire pour revoir les contours et la forme que devraient prendre les concours de l'État aux collectivités territoriales en vue du financement de leurs investissements, à l'heure où une nouvelle étape de la décentralisation rend nécessaire une redéfinition de ses formes d'intervention dans les territoires. En 2013, plus de 10 000 subventions ont été attribuées. Si le nombre de subventions est quasiment constant depuis trois ans¹, leur montant global a diminué et s'est établi en 2013 à 108 M€².

Cour des comptes – Référé n° 71261

6/9

présentés sous un libellé suffisamment générique afin de ne pas compromettre l'octroi de la subvention malgré l'absence de réalisation conforme au dessin initial.

b. L'absence de coordination de différents dispositifs d'aides de l'État

Puisieurs dispositifs de subventionnement de l'État peuvent bénéficier, sans coordination, à une même catégorie d'opérations, en particulier en matière d'équipements scolaires, sportifs ou socioculturels. On peut citer notamment le fonds national d'aménagement et de développement du territoire (FNADT), la dotation d'équipement aux territoires ruraux ou des dispositifs de soutien spécifiques, par exemple, pour la modernisation des bibliothèques¹¹ ou l'élaboration des documents d'urbanisme.

Une gestion efficiente devrait conduire à regrouper au sein de ces dispositifs les aides ayant le même objet.

c. Un caractère incitatif souvent difficile à démontrer

Un nombre important de subventions ne contribuent que de manière très marginale au financement des projets. La part de la subvention attribuée par rapport au coût moyen des projets s'est établie à 6,6 % en moyenne en 2012, en diminution par rapport à 2011 (7,5 %)¹². En 2012, près de 7 % des subventions contribuaient pour moins de 2 % au financement des opérations¹³ et leur nombre était en augmentation, passant de 318 en 2006 à 671 en 2012. Dans certains départements, la dispersion des subventions est très grande et certaines représentent moins d'un millième du coût du projet. Bien que, pour la réserve ministérielle, un montant minimal de subvention de 5 000 € soit appliqué, les règles en vigueur autorisent à subventionner des biens de faible valeur unitaire, conduisant en pratique à attribuer des subventions d'un montant inférieur à leur coût de gestion.

2. Une gestion coûteuse

Comparée à d'autres, cette procédure a un coût de gestion élevé au regard des montants unitaires attribués. Près de 85 équivalents temps plein travaillés (ETPT)¹⁴ du ministère de l'intérieur, selon une estimation basse, sont mobilisés pour la gestion de ces subventions. Le coût moyen de mise en œuvre, pour les seules « rémunérations et charges sociales », peut être estimé à 385 € par subvention versée. Des subventions d'un montant inférieur à ce coût sont attribuées chaque année. En moyenne, 10 M€ attribués au titre des travaux divers d'intérêt local requièrent 5,8 ETPT pour leur gestion. Ce coût est plus de 2,5 fois supérieur à celui de la gestion de la dotation d'équipement des territoires ruraux.

Dans ces conditions, le maintien en l'état de la procédure des subventions pour travaux divers d'intérêt local, redondante avec des dispositifs mieux encadrés poursuivant des objectifs similaires, sinon identiques, n'apparaît pas compatible avec l'objectif d'un bon emploi des deniers publics.

Le soutien à des collectivités financièrement déstabilisées par des circonstances exceptionnelles ou anormales, auquel devraient contribuer, selon l'information donnée au Parlement dans les documents budgétaires, les subventions pour travaux divers d'intérêt local est largement perdu de vue. Ces subventions sont, en pratique, une procédure d'aide à l'investissement des collectivités territoriales venant s'ajouter à des dispositifs mieux définis et encadrés.

¹¹ 99 dossiers ont été subventionnés au titre des travaux divers d'intérêt local en 2012, alors que 80 M€ de crédits sont prévus pour aider ces équipements au sein du programme 122 – Concours spécifiques et administration.

¹² Pour la dotation d'équipement des territoires ruraux, l'objectif de performance retenu par l'administration en 2012 était qu'une proportion de 65 % des crédits attribués représente entre 20 % et 30 % de la dépense subventionnable.

¹³ Cette proportion s'accroît puisque la part des subventions inférieures à 2 % s'élève en 2013 à plus de 7,5 %.

¹⁴ Y compris les effets de la cellule ministérielle.

Partie 2 : Conformité selon majorité et opposition (2 points)

Par ailleurs, le rapporteur de la Cour calcule pour lui-même (mais sans les reporter dans le rapport) les statistiques suivantes : sur les 40%, soit 220, dossiers non conformes dans l'échantillon dont il était question ci-dessus, 141 ont été portés par des parlementaires de la majorité et 79 par l'opposition.

3. Peut-on dire que l'une des deux postures (majorité et opposition) soit associée à des comportements plus vertueux? On parle de « posture » parce qu'évidemment au cours du temps et des législatures, droite et gauche changent de posture dans une chambre. On se rappellera par ailleurs que les hypothèses d'un test statistique ne doivent jamais être fixées au vu des données. (2 points)

Partie 3 : Vérification plus poussée de la représentativité (2 points)

La Cour a voulu vérifier que les 550 dossiers représentatifs (notre échantillon) l'étaient bien, non pas uniquement au sens des attributions majorité / opposition, mais également selon deux autres critères : les montants des subventions accordées et les affectations sectorielles. A ce dernier sujet, la Cour écrit :

2. L'affectation sectorielle

Les travaux de voirie ou d'aménagement urbain représentaient en 2012 plus du tiers des subventions attribuées, les dépenses d'équipement des services des collectivités et les investissements scolaires ou périscolaires environ 10 %, les unes comme les autres. Ces dépenses n'ont rien d'exceptionnel mais sont au contraire habituelles pour les collectivités territoriales.

Des subventions sont affectées en proportion importante ou de manière répétée à des dépenses qui pourraient répondre à un objectif de politique publique comme l'acquisition de défibrillateurs ou la restauration du patrimoine religieux (5,6 % des subventions attribuées en 2012). Elles peuvent être attribuées pendant plusieurs exercices successifs, comme la Cour l'a constaté, pour la restauration d'églises par certaines communes.

Nous supposons que sur l'ensemble des dossiers de 2012, au nombre d'environ 10 000, la répartition selon les montants et les affectations est la suivante. Ces répartitions de 2012 formeront notre répartition de référence (grâce à une hypothèse de stationnarité des dites répartitions entre les années).

Affectations sectorielles		Montants	
Voirie / aménagement urbain	35.1%	Inférieur à 500 euros	12.6%
Equipement des services	9.2%	Entre 501 et 2 500 euros	21.4%
Investissements [péri-]scolaires	11.3%	Entre 2 501 et 10 000 euros	25.6%
Politiques publiques	5.6%	Entre 10 001 et 20 000 euros	21.9%
Autres dépenses (les bonnes dépenses !)	38.8%	Supérieur à 20 001 euros	18.5%

Nous calculons les traitements statistiques reproduits en haut de la page suivante.

4. Comment s'appellent les procédures statistiques mises en œuvre? L'ont-elles validement été? Quels en sont les résultats? (1.5 points)

Un traitement supplémentaire aurait été de vérifier la bonne répartition géographique. Elle devrait être assurée puisque la Cour a pris soin d'interroger 33 préfectures. Cependant, si on voulait la vérifier (*id est*, vérifier que le nombre de dossiers attribués par région administrative métropolitaine hors Corse existant en 2012 est bien proportionnel à sa population) :

5. Pourrait-on utiliser le découpage de la page suivante, ou serait-il trop fin et la procédure statistique associée souffrirait-elle d'un problème de validité? (0.5 point)

Secteur

	Effectif observé	N théorique	Résidus
Voirie / aménagement urbain	187	193,0	-6,0
Equipement des services	54	50,6	3,4
Investissements [péri-] scolaires	67	62,2	4,9
Politiques publiques	28	30,8	-2,8
Autres (et bonnes) dépenses	214	213,4	,6
Total	550		

Tests statistiques

	Secteur
Khi-deux	1,053 ^a
ddl	4
Sig. asymptotique	,902

a. 0 cellules (0,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 30,8.

Montant

	Effectif observé	N théorique	Résidus
<= 500	60	69,3	-9,3
500 < ... <= 2500	130	117,7	12,3
2500 < ... <= 10000	145	140,8	4,2
10000 < ... <= 20000	116	120,5	-4,5
> 20000	99	101,8	-2,8
Total	550		

Tests statistiques

	Montant
Khi-deux	2,897 ^a
ddl	4
Sig. asymptotique	,575

a. 0 cellules (0,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 69,3.

1	Ile-de-France	11 852 851	18.9%
2	Champagne-Ardenne	1 336 053	2.1%
3	Picardie	1 918 155	3.1%
4	Haute-Normandie	1 839 393	2.9%
5	Centre-Val de Loire	2 556 835	4.1%
6	Basse-Normandie	1 475 684	2.4%
7	Bourgogne	1 642 734	2.6%
8	Nord-Pas-de-Calais	4 042 015	6.4%
9	Lorraine	2 350 657	3.7%
10	Alsace	1 852 325	3.0%
11	Franche-Comté	1 173 440	1.9%
12	Pays de la Loire	3 601 113	5.7%
13	Bretagne	3 217 767	5.1%
14	Poitou-Charentes	1 777 773	2.8%
15	Aquitaine	3 254 233	5.2%
16	Midi-Pyrénées	2 903 420	4.6%
17	Limousin	741 720	1.2%
18	Rhône-Alpes	6 283 541	10.0%
19	Auvergne	1 350 682	2.2%
20	Languedoc-Roussillon	2 670 460	4.3%
21	Provence-Alpes-Côte d'Azur	4 916 690	7.8%
	Total	62 757 541	100.0%

Partie 4 : Estimation des coûts de gestion (3 points)

Nous nous intéressons enfin au paragraphe suivant du rapport :

2. Une gestion coûteuse

Comparée à d'autres, cette procédure a un coût de gestion élevé au regard des montants unitaires attribués. Près de 85 équivalents temps plein travaillés (ETPT)¹⁴ du ministère de l'intérieur, selon une estimation basse, sont mobilisés pour la gestion de ces subventions. Le coût moyen de mise en œuvre, pour les seules « rémunérations et charges sociales », peut être estimé à 385 € par subvention versée. Des subventions d'un montant inférieur à ce coût sont attribuées chaque année. En moyenne, 10 M€ attribués au titre des travaux divers d'intérêt local requièrent 5,8 ETPT pour leur gestion. Ce coût est plus de 2,5 fois supérieur à celui de la gestion de la dotation d'équipement des territoires ruraux.

Nous allons imaginer comment la Cour est parvenue à ces chiffres. Peu de fonctionnaires du Ministère de l'Intérieur, qui est gestionnaire de la réserve parlementaire, passent l'intégralité de leur temps de travail annuel à cette gestion. Mais beaucoup y passent une certaine proportion de leur temps. La Cour fait interroger les 314 fonctionnaires du bureau dédié aux collectivités territoriales, et leur demande à chacun quelle proportion du temps de travail de l'année 2012 ils ont passé à la gestion de cette réserve.

6. Chaque fonctionnaire répond bien entendu de manière assez approximative. Expliquer, si possible grâce à une modélisation stochastique (avec des variables aléatoires), que ces données vont bien permettre malgré tout d'estimer notre paramètre d'intérêt, qui est le temps total consacré par ces fonctionnaires à cette tâche de gestion. Cette question est difficile et peut être sautée en première lecture. *(1 point)*
7. Les données recueillies sont que la proportion moyenne de temps de travail annuel déclarée par ces fonctionnaires vaut 31.5% (avec un écart-type de 46.3% dans les données). Pouvez-vous retrouver le chiffre de 85 ETPT avancé ci-dessus par la Cour des comptes ? *(1.5 points)*
8. Par ailleurs, le coût salarial moyen mensuel de ces fonctionnaires est de 3 243 euros (salaires et cotisations, tant salariales que patronales). Ce chiffre est certain, car la Cour a eu accès à l'ensemble des fiches de paie des fonctionnaires considérés. En déduire l'estimation avancée du coût de gestion d'un dossier de subvention ; cette estimation est-elle une estimation haute ou une estimation basse ? On pourra supposer que les 60 000 dossiers de la période 2006–2012 se sont répartis de manière homogène selon les années. *(0.5 point)*

Nouvelle politique de prix pour une entreprise de VPC

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 (<i>avant remise</i>)	73
Ecart-type du montant des commandes (<i>avant remise</i>)	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 ?

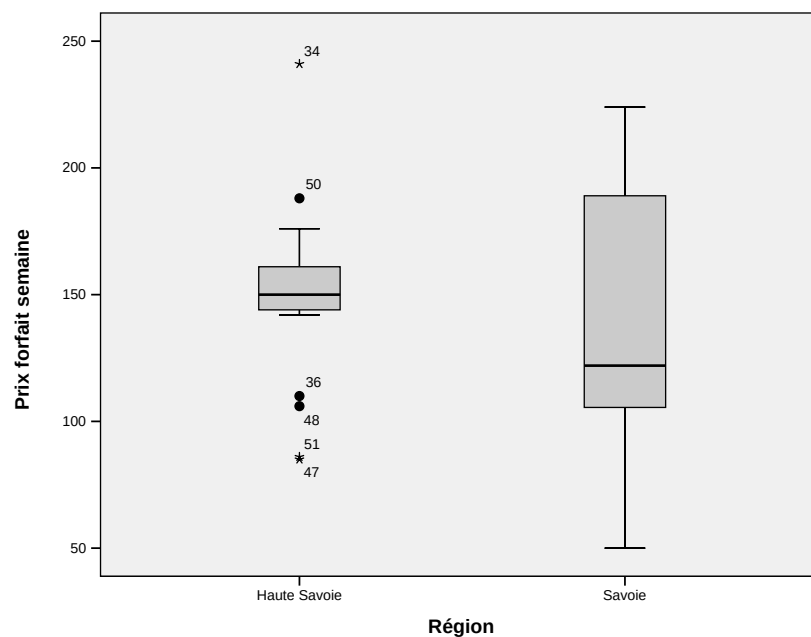
Modélisation du prix du forfait de ski

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.

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	V6	
1	Alpe d'Huez	197	1860	85	3330	108	32000	3		
2	Alpe du Grand Serre	107	1400	20	2200	34	3600	3		
3	Arcaches 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	Aussais	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 Coremoz	102	1150	17	1950	26	6500	7		
18	Eaux Bonnes Courrette	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	660	20	1150	20	6400	8		
23	Isola 2000	122	1800	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'Allervard	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	5,343	,025	,771	49	,444	9,773	12,676	-15,700	35,246	
Hypothèse de variances égales										
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

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

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

Prévision des résultats des élections présidentielles américaines

L'exercice qui suit est tiré de l'article «Prediction for the 2012 United States presidential election using multiple regression model», écrit par trois chercheurs de l'Université de Delhi et publié, en septembre 2012, dans le journal académique *The Journal of Prediction Markets*. Autant mettre notre lecteur en garde : ce problème d'examen se veut aussi une mise au pilori de cet article assez représentatif de ce que peut être un mauvais article de recherche, fondé sur un jeu de données certes valable mais appliquant une méthodologie statistique à tout le moins incorrecte, voire manipulatrice. Bref.

On cherche dans un premier temps à expliquer et dans un second temps à prévoir les résultats des élections présidentielles américaines, entre 1948 et 2012. Plus précisément, on s'intéresse

- à la variable **Vote**, la proportion des voix obtenue par le parti au pouvoir aux élections ayant lieu à la fin du mandat présidentiel (par rapport au total des voix obtenues par les deux partis ultra-majoritaires, républicain et démocrate),

que l'on veut modéliser en fonction de deux types de variables explicatives :

- des variables économiques (taux de chômage, inflation, croissance, etc.) ;
- des variables politiques (présence ou absence de scandales pendant le terme écoulé, résultats des élections de mi-mandat, cote au dernier sondage de popularité).

Par exemple, pour l'élection de 2008, au bout des deux mandats de George W. Bush, l'observation est formée par le score du candidat républicain à l'élection (en l'occurrence, John McCain, qui a eu 46.6 % des suffrages).

Les auteurs choisissent cette mesure plutôt que celle des scores d'un parti donné et fixé parce qu'ils ont en tête une hypothèse (plus ou moins explicitée) d'inertie politique : les électeurs les plus importants, ceux qui ne sont pas acquis à vie à un parti (les «swing voters»), changent leur parti de prédilection lorsqu'ils sont mécontents de celui au pouvoir, et à défaut, reconduisent ce dernier.

Description des variables explicatives

Les auteurs considèrent de nombreuses variables économiques. Les plus naturelles sont le taux de chômage, le taux de croissance, et l'inflation, trois mesures de la bonne santé économique d'un pays. A celles-ci, ils ajoutent de nombreuses autres valeurs, ne semblant être limités que par leur imagination dans cet inventaire à la Prévert.

On retrouve donc dans le jeu de données les variables

- **UnemplRate**, le taux de chômage annuel moyen lors du mandat écoulé (en %) ;
- **IntRate**, le taux d'intérêt annuel moyen servi par le Trésor lors du mandat écoulé (en %) ;
- **Infl**, l'inflation annuelle moyenne lors des trois premières années et demie du mandat écoulé (en %) ;
- **HealthBudg**, la part du budget fédéral, en proportion du PIB, consacrée aux prestations sociales liées à la santé (en %) ;
- **GrowthRate**, la croissance mesurée lors des trois premiers trimestres de l'année électorale (en %) ;
- **BudgSurDef**, l'excès ou le déficit de recettes par rapport aux dépenses du budget prévisionnel, rapportés à ce dernier (en %) ;
- **PublDebt**, la dette fédérale, mesurée en proportion du PIB (en %) ;
- **GoldPrices**, le prix moyen de l'or pendant le mandat écoulé (en dollars) ;
- **OilPrices**, le prix moyen du baril de pétrole brut pendant le mandat écoulé (en dollars) ;
- **ExchRate**, le taux de change moyen pendant le mandat entre la livre sterling et le dollar.

Les variables politiques sont quant à elles :

- **Scandal**, une variable prenant trois valeurs 0, 1 ou 2, selon qu'au cours du mandat écoulé, il n'y a eu aucun scandale politique, au moins un scandale n'ayant pas soulevé la question d'un *impeachment*, ou un scandale retentissant ayant mené ou ayant failli mener à un *impeachment* ;
- **RatingJune**, une mesure de popularité de l'action présidentielle effectuée au moins de juin de l'année électorale (précédant donc l'élection de quelques mois) ;
- **MidTerm**, une variable prenant deux valeurs +1 et -1 selon que le parti du président a conservé ou non la majorité à la chambre des représentants lors des élections de mi-mandat (qui voient un renouvellement total de cette chambre).

Faisons connaissance avec les données

Les données, telles qu'elles se présentent sous SPSS, sont reproduites en annexe (page 3). Les questions qui suivent visent à faire mieux connaissance avec elles ; répondez-y fort brièvement.

- (1) Beaucoup de variables économiques sont considérées. Pensez-vous qu'il y en ait trop ? Aurait-on pu en trouver d'autres ?
- (2) En 2000, George W. Bush a succédé à Bill Clinton. Que pensez-vous de la valeur de la variable **Vote** en 2000 ?
- (3) Combien de scandales retentissants compte-t-on (et question subsidiaire : qui sont-ils) ?
- (4) Choisissez une variable explicative non discutée dans les questions précédentes et effectuez un commentaire ou une observation pertinente (pas nécessairement une critique!).

Etude de variables explicatives prises isolément

- (5) Parmi l'ensemble des 13 modèles de régression linéaire simples reproduits en annexe (pages 4–10), lesquels sont valides statistiquement ? (On indiquera comment on les détermine et on ne dressera la liste que de ces derniers.)
- (6) Parmi ces modèles valides statistiquement, lesquels sont valides économiquement ?
- (7) Ecrire la relation linéaire proposée par le meilleur des modèles valides à la fois statistiquement et économiquement.

Construction d'un modèle explicatif

On lit ici les sorties SPSS intitulées «Méthode automatique # x », où $x \in \{1, 2\}$ (pages 11 et suivantes). On se souviendra qu'il est généralement admis qu'il ne faut pas considérer plus d'une variable explicative pour 5 ou 6 observations.

- (8) Que fait chacune des méthodes, quel est son nom, son principe ? Certaines de ces méthodes conduisent-elles à des modèles contenant trop de variables explicatives par rapport à ce qui est généralement admis ?

On lit maintenant la sortie SPSS intitulée «Modèle final» (page 16).

- (9) Justifier brièvement que ce modèle est valide statistiquement et écrire la relation linéaire proposée.
- (10) Effectuer la validation économique de la relation précédente.
Quantifier également l'impact électoral d'un scandale.
- (11) La méthode #1 proposait le même modèle que celui étudié ici, mais avec des coefficients légèrement différents : pourquoi ? (On pensera à regarder quelles variables explicatives souffrent de valeurs manquantes ; quel est le modèle exploitant le plus de données ?)

Etude du caractère prédictif du modèle construit

Nous avons construit un modèle explicatif. Nous testons maintenant son caractère prédictif de manière rétrospective. En septembre 2012, la mesure de popularité de Barack Obama était de 46.4 %, le taux de croissance sur les trois premiers trimestres était de 1.62 %, et il n'y avait pas eu de scandale pendant le mandat 2009–2012.

- (12) En déduire une prévision du score de Barack Obama lors des élections de novembre 2012.
La comparer ensuite au score réalisé, 51.9 % : l'écart entre les deux valeurs est-il compatible avec le modèle ?
Le modèle construit à la question (10) prévoyait-il de manière nette le vainqueur de l'élection ?

Critique de l'article

Nous avons identifié à la question (7) les variables qui semblaient importantes. Regarder le tableau intitulé «Table A : Analysis of Influence of Economic Variables» (reproduit page 17). Selon les auteurs de l'article sur lequel ce problème est fondé, il est censé justifier le fait que la seule variable économique influente soit le taux de croissance.

- (13) Etes-vous convaincu de ce fait, à la seule vue de ce tableau ?

*Presidential-Elections.sav [Ensemble_de_donnees1] - IBM SPSS Statistics Editeur de données															
Fichier Edition Affichage Données Transformer Analyse Graphes Utilitaires Fenêtre Aide															
47 : Scandal															
	Year	Vote	UnempRate	IntRate	Inf	HealthBudg	GrowthRate	BudgSurDef	PublDebt	GoldPrices	OilPrices	ExchRate	Scandal	RatingJune	MidTerm
1	1944	53,774	.	.	,000	1,228	4,279	-22,700	91,490	33,850	.	4,032	.	.	.
2	1948	52,370	3,80	.	,000	3,679	3,579	4,600	93,580	34,710	2,770	4,032	1	39,5	-1
3	1952	44,595	3,00	.	2,362	3,070	,691	-,400	72,255	34,600	2,770	2,793	1	31,5	1
4	1956	57,764	4,10	2,730	1,935	3,567	-1,451	,900	62,272	34,990	2,940	2,793	0	72,0	-1
5	1960	49,913	5,50	3,210	1,967	4,692	,377	,100	54,291	35,270	2,910	2,809	1	59,0	-1
6	1964	61,344	5,20	3,500	1,260	4,777	5,109	-,900	46,916	35,100	3,000	2,793	0	74,0	1
7	1968	49,596	3,60	5,660	3,139	5,924	5,043	-2,900	38,133	39,310	3,180	2,392	1	41,0	1
8	1972	61,789	5,60	4,440	4,815	7,739	5,914	-2,000	35,145	58,420	3,600	2,500	0	57,5	-1
9	1976	48,948	7,70	5,050	7,630	9,854	3,751	-4,200	34,485	124,740	13,100	1,805	2	45,0	-1
10	1980	44,697	7,10	13,350	7,831	9,842	-3,597	-2,700	42,277	615,000	37,420	2,326	1	33,6	1
11	1984	59,170	7,50	10,230	5,259	9,771	5,440	-4,800	50,896	361,000	28,750	1,337	0	54,0	-1
12	1988	53,902	5,50	7,570	2,906	9,419	2,178	-3,100	61,941	437,000	14,870	1,783	1	50,0	-1
13	1992	46,545	7,50	3,520	3,280	11,600	2,662	-4,700	70,736	343,820	19,250	1,767	1	37,3	-1
14	1996	54,736	5,40	5,300	2,062	11,594	3,121	-1,400	70,299	387,810	20,460	1,563	1	55,0	-1
15	2000	50,265	4,00	6,240	1,605	10,543	1,219	2,400	54,835	279,110	27,390	1,515	2	57,5	-1
16	2004	51,233	5,50	1,350	2,325	11,889	2,690	-3,500	61,420	409,720	37,660	1,832	0	48,5	1
17	2008	46,600	5,80	1,920	3,052	13,000	,220	-3,200	71,221	871,960	91,480	1,852	1	29,0	-1
18															
19															
40															

Visible : 15 variables s

Régressions simples sur variables économiques

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,013 ^a	,000	-,071	5,787473

a. Valeurs prédites : (constantes), UnemplRate

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	,078	1	,078	,002	,962 ^a
Résidu	468,928	14	33,495		
Total	469,006	15			

a. Valeurs prédites : (constantes), UnemplRate

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	52,358	5,697		9,191	,000
	UnemplRate	-,049	1,016	-,013	-,048	,962

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,106 ^a	,011	-,071	5,805074

a. Valeurs prédites : (constantes), IntRate

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	4,620	1	4,620	,137	,718 ^a
Résidu	404,387	12	33,699		
Total	409,006	13			

a. Valeurs prédites : (constantes), IntRate

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	53,567	3,020		17,736	,000
	IntRate	-,181	,490	-,106	-,370	,718

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,189 ^a	,036	-,033	5,683766

a. Valeurs prédites : (constantes), Infl

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	16,733	1	16,733	,518	,484 ^a
	Résidu	452,273	14	32,305		
	Total	469,006	15			

a. Valeurs prédites : (constantes), Infl

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	53,658	2,599		20,644	,000
	Infl	-,487	,677	-,189	-,720	,484

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,210 ^a	,044	-,024	5,659246

a. Valeurs prédites : (constantes), HealthBudg

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	20,627	1	20,627	,644	,436 ^a
	Résidu	448,379	14	32,027		
	Total	469,006	15			

a. Valeurs prédites : (constantes), HealthBudg

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	54,924	3,803		14,444	,000
	HealthBudg	-,346	,431	-,210	-,803	,436

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,561 ^a	,315	,266	4,789723

a. Valeurs prédites : (constantes), GrowthRate

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	147,826	1	147,826	6,444	,024 ^a
	Résidu	321,180	14	22,941		
	Total	469,006	15			

a. Valeurs prédites : (constantes), GrowthRate

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	49,334	1,617		30,515	,000
	GrowthRate	1,194	,470	,561	2,538	,024

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,094 ^a	,009	-,062	5,762076

a. Valeurs prédites : (constantes), BudgSurDef

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	4,185	1	4,185	,126	,728 ^a
	Résidu	464,821	14	33,202		
	Total	469,006	15			

a. Valeurs prédites : (constantes), BudgSurDef

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	52,416	1,705		30,737	,000
	BudgSurDef	,201	,566	,094	,355	,728

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,221 ^a	,049	-,019	5,644191

a. Valeurs prédites : (constantes), PublDebt

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	23,009	1	23,009	,722	,410 ^a
	Résidu	445,997	14	31,857		
	Total	469,006	15			

a. Valeurs prédites : (constantes), PublDebt

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	56,504	5,380		10,503	,000
	PublDebt	-,077	,090	-,221	-,850	,410

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,354 ^a	,125	,063	5,413964

a. Valeurs prédites : (constantes), GoldPrices

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	58,652	1	58,652	2,001	,179 ^a
	Résidu	410,354	14	29,311		
	Total	469,006	15			

a. Valeurs prédites : (constantes), GoldPrices

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	54,108	1,965		27,530	,000
	GoldPrices	-,008	,006	-,354	-1,415	,179

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,352 ^a	,124	,061	5,417092

a. Valeurs prédites : (constantes), OilPrices

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	58,178	1	58,178	1,983	,181 ^a
Résidu	410,828	14	29,345		
Total	469,006	15			

a. Valeurs prédites : (constantes), OilPrices

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	53,758	1,799		29,890	,000
	OilPrices	-,086	,061	-,352	-1,408	,181

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,063 ^a	,004	-,067	5,776603

a. Valeurs prédites : (constantes), ExchRate

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	1,838	1	1,838	,055	,818 ^a
Résidu	467,168	14	33,369		
Total	469,006	15			

a. Valeurs prédites : (constantes), ExchRate

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	50,965	5,012		10,168	,000
	ExchRate	,502	2,140	,063	,235	,818

a. Variable dépendante : Vote

Régressions simples sur variables politiques

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,652 ^a	,425	,384	4,389789

a. Valeurs prédites : (constantes), Scandal

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	199,223	1	199,223	10,338	,006 ^a
Résidu	269,784	14	19,270		
Total	469,006	15			

a. Valeurs prédites : (constantes), Scandal

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	56,612	1,783		31,744	,000
	Scandal	-5,563	1,730	-,652	-3,215	,006

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,807 ^a	,651	,626	3,417675

a. Valeurs prédites : (constantes), RatingJune

ANOVA^b

Modèle	Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1 Régression	305,479	1	305,479	26,153	,000 ^a
Résidu	163,527	14	11,681		
Total	469,006	15			

a. Valeurs prédites : (constantes), RatingJune

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	35,636	3,329		10,704	,000
	RatingJune	,336	,066	,807	5,114	,000

a. Variable dépendante : Vote

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,224 ^a	,050	-,018	5,640901

a. Valeurs prédites : (constantes), MidTerm

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	23,529	1	23,529	,739	,404 ^a
	Résidu	445,477	14	31,820		
	Total	469,006	15			

a. Valeurs prédites : (constantes), MidTerm

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	51,601	1,521		33,920	,000
	MidTerm	-1,308	1,521	-,224	-,860	,404

a. Variable dépendante : Vote

Méthode automatique #1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,798 ^a	,637	,607	3,518565
2	,885 ^b	,783	,743	2,841167
3	,939 ^c	,882	,847	2,193011

a. Valeurs prédites : (constantes), RatingJune

b. Valeurs prédites : (constantes), RatingJune, GrowthRate

c. Valeurs prédites : (constantes), RatingJune, GrowthRate, Scandal

ANOVA^d

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	260,443	1	260,443	21,037	,001 ^a
	Résidu	148,564	12	12,380		
	Total	409,006	13			
2	Régression	320,212	2	160,106	19,834	,000 ^b
	Résidu	88,795	11	8,072		
	Total	409,006	13			
3	Régression	360,913	3	120,304	25,015	,000 ^c
	Résidu	48,093	10	4,809		
	Total	409,006	13			

a. Valeurs prédites : (constantes), RatingJune

b. Valeurs prédites : (constantes), RatingJune, GrowthRate

c. Valeurs prédites : (constantes), RatingJune, GrowthRate, Scandal

d. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	35,313	3,886		9,087	,000
	RatingJune	,339	,074	,798	4,587	,001
2	(Constante)	35,358	3,138		11,268	,000
	RatingJune	,302	,061	,710	4,928	,000
	GrowthRate	,795	,292	,392	2,721	,020
3	(Constante)	40,793	3,059		13,336	,000
	RatingJune	,244	,051	,573	4,737	,001
	GrowthRate	,695	,228	,343	3,045	,012
	Scandal	-2,824	,971	-,352	-2,909	,016

a. Variable dépendante : Vote

Variables exclues^d

Modèle		Bêta dans	t	Sig.	Corrélation partielle	Statistiques de colinéarité
						Tolérance
1	UnemplRate	,140 ^a	,712	,491	,210	,822
	IntRate	,086 ^a	,466	,650	,139	,945
	Infl	,168 ^a	,827	,426	,242	,751
	HealthBudg	,148 ^a	,608	,556	,180	,541
	GrowthRate	,392 ^a	2,721	,020	,634	,950
	BudgSurDef	-,474 ^a	-2,591	,025	-,616	,614
	PublDebt	-,111 ^a	-,619	,549	-,183	,992
	GoldPrices	,107 ^a	,427	,678	,128	,512
	OilPrices	,067 ^a	,284	,781	,085	,588
	ExchRate	-,126 ^a	-,641	,535	-,190	,827
	Scandal	-,408 ^a	-2,575	,026	-,613	,822
	MidTerm	-,038 ^a	-,206	,840	-,062	,993
2	UnemplRate	,085 ^b	,524	,612	,163	,807
	IntRate	,146 ^b	,998	,342	,301	,925
	Infl	,154 ^b	,946	,367	,287	,750
	HealthBudg	,061 ^b	,304	,767	,096	,526
	BudgSurDef	-,280 ^b	-1,268	,233	-,372	,384
	PublDebt	,006 ^b	,037	,971	,012	,901
	GoldPrices	,356 ^b	1,862	,092	,507	,441
	OilPrices	,215 ^b	1,145	,279	,341	,546
	ExchRate	,007 ^b	,039	,970	,012	,746
	Scandal	-,352 ^b	-2,909	,016	-,677	,803
	MidTerm	-,043 ^b	-,291	,777	-,092	,993
3	UnemplRate	,049 ^c	,386	,708	,128	,799
	IntRate	,164 ^c	1,548	,156	,459	,923
	Infl	,158 ^c	1,310	,223	,400	,750
	HealthBudg	,108 ^c	,700	,501	,227	,520
	BudgSurDef	-,003 ^c	-,014	,989	-,005	,270
	PublDebt	-,045 ^c	-,370	,720	-,122	,881
	GoldPrices	,239 ^c	1,484	,172	,443	,404
	OilPrices	,119 ^c	,770	,461	,249	,515
	ExchRate	-,112 ^c	-,841	,422	-,270	,680
	MidTerm	-,169 ^c	-1,565	,152	-,463	,879

a. Valeurs prédites dans le modèle : (constantes), RatingJune

b. Valeurs prédites dans le modèle : (constantes), RatingJune, GrowthRate

c. Valeurs prédites dans le modèle : (constantes), RatingJune, GrowthRate, Scandal

d. Variable dépendante : Vote

Méthode automatique #2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,999 ^a	,998	,977	,844432
2	,999 ^b	,997	,983	,736403
3	,996 ^c	,991	,962	1,099114
4	,993 ^d	,986	,955	1,186902

a. Valeurs prédites : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef, PublDebt

b. Valeurs prédites : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

c. Valeurs prédites : (constantes), MidTerm, GrowthRate, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

d. Valeurs prédites : (constantes), MidTerm, GrowthRate, HealthBudg, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

ANOVA^e

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	408,293	12	34,024	47,716	,113 ^a
	Résidu	,713	1	,713		
	Total	409,006	13			
2	Régression	407,922	11	37,084	68,384	,014 ^b
	Résidu	1,085	2	,542		
	Total	409,006	13			
3	Régression	405,382	10	40,538	33,557	,007 ^c
	Résidu	3,624	3	1,208		
	Total	409,006	13			
4	Régression	403,371	9	44,819	31,815	,002 ^d
	Résidu	5,635	4	1,409		
	Total	409,006	13			

a. Valeurs prédites : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef, PublDebt

b. Valeurs prédites : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

c. Valeurs prédites : (constantes), MidTerm, GrowthRate, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

d. Valeurs prédites : (constantes), MidTerm, GrowthRate, HealthBudg, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

e. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	-,070	10,817		-,007	,996
	UnemplRate	-2,126	,493	-,502	-4,310	,145
	IntRate	,659	,149	,386	4,431	,141
	Infl	,921	,712	,347	1,293	,419
	HealthBudg	1,025	,375	,553	2,734	,223
	GrowthRate	1,027	,364	,506	2,817	,217
	BudgSurDef	-1,016	,626	-,388	-1,622	,352
	PublDebt	,077	,107	,180	,722	,602
	OilPrices	,057	,020	,242	2,811	,218
	ExchRate	6,899	1,851	,629	3,727	,167
	Scandal	-1,496	,774	-,186	-1,932	,304
	RatingJune	,493	,077	1,159	6,413	,098
	MidTerm	-1,798	,575	-,301	-3,130	,197
2	(Constante)	3,565	8,349		,427	,711
	UnemplRate	-2,088	,428	-,493	-4,882	,039
	IntRate	,676	,128	,396	5,279	,034
	Infl	,434	,201	,164	2,164	,163
	HealthBudg	1,163	,281	,628	4,136	,054
	GrowthRate	,785	,126	,387	6,246	,025
	BudgSurDef	-1,313	,412	-,501	-3,190	,086
	OilPrices	,055	,018	,233	3,137	,088
	ExchRate	6,857	1,613	,625	4,250	,051
	Scandal	-1,594	,665	-,199	-2,398	,139
	RatingJune	,507	,065	1,191	7,799	,016
	MidTerm	-2,117	,320	-,354	-6,613	,022
3	(Constante)	-,638	12,119		-,053	,961
	UnemplRate	-1,851	,617	-,437	-3,000	,058
	IntRate	,829	,159	,486	5,206	,014
	HealthBudg	1,231	,417	,664	2,951	,060
	GrowthRate	,815	,186	,402	4,374	,022
	BudgSurDef	-1,457	,606	-,557	-2,404	,096
	OilPrices	,057	,026	,242	2,186	,117
	ExchRate	7,904	2,297	,720	3,440	,041
	Scandal	-1,241	,962	-,155	-1,290	,287
	RatingJune	,508	,097	1,193	5,235	,014
	MidTerm	-2,204	,474	-,368	-4,649	,019
4	(Constante)	-12,321	8,698		-1,417	,230

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
4	UnemplRate	-2,321	,538	-,548	-4,319	,012
	IntRate	,922	,153	,540	6,005	,004
	HealthBudg	1,480	,399	,799	3,706	,021
	GrowthRate	,765	,197	,377	3,886	,018
	BudgSurDef	-2,065	,412	-,789	-5,012	,007
	OilPrices	,073	,025	,310	2,970	,041
	ExchRate	9,726	1,957	,886	4,970	,008
	RatingJune	,612	,058	1,439	10,608	,000
	MidTerm	-2,330	,501	-,389	-4,650	,010

a. Variable dépendante : Vote

Variables exclues^e

Modèle		Bêta dans	t	Sig.	Corrélation partielle	Statistiques de colinéarité
						Tolérance
1	GoldPrices	-18,425 ^a	.	.	-1,000	5,135E-6
2	GoldPrices	-7,445 ^b	-,417	,749	-,385	7,073E-6
	PublDebt	,180 ^b	,722	,602	,585	,028
3	GoldPrices	-,487 ^c	-2,193	,160	-,840	,026
	PublDebt	-,125 ^c	-1,343	,311	-,689	,269
	Infl	,164 ^c	2,164	,163	,837	,231
4	GoldPrices	-,560 ^d	-2,306	,104	-,800	,028
	PublDebt	-,078 ^d	-,664	,554	-,358	,293
	Infl	,119 ^d	1,011	,386	,504	,246
	Scandal	-,155 ^d	-1,290	,287	-,597	,205

a. Valeurs prédites dans le modèle : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef, PublDebt

b. Valeurs prédites dans le modèle : (constantes), MidTerm, GrowthRate, Infl, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

c. Valeurs prédites dans le modèle : (constantes), MidTerm, GrowthRate, HealthBudg, Scandal, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

d. Valeurs prédites dans le modèle : (constantes), MidTerm, GrowthRate, HealthBudg, IntRate, UnemplRate, RatingJune, OilPrices, ExchRate, BudgSurDef

e. Variable dépendante : Vote

Modèle final

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,939 ^a	,882	,853	2,147258

a. Valeurs prédites : (constantes), RatingJune, GrowthRate, Scandal

ANOVA^b

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	413,677	3	137,892	29,907	,000 ^a
	Résidu	55,329	12	4,611		
	Total	469,006	15			

a. Valeurs prédites : (constantes), RatingJune, GrowthRate, Scandal

b. Variable dépendante : Vote

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	40,599	2,718		14,936	,000
	GrowthRate	,750	,219	,353	3,426	,005
	Scandal	-2,743	,946	-,321	-2,901	,013
	RatingJune	,245	,046	,588	5,298	,000

a. Variable dépendante : Vote

Economic Factors: The following table lists the regression results for various models containing economic variables as independent variables and *VOTE* as dependent variable:

Table A: Analysis of Influence of Economic Variables

Model	Year	R ² (%)	p-Statistic	
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH + \beta_3 INFLATION$ + $\beta_4 UNEMPLOYMENT_RATE$	1948-2008	33.775	GROWTH* INFLATION UNEMPLOYMENT_RATE	0.046 0.551 0.756
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH$ + $\beta_3 CHANGE_UNEMPLOYMENT_RATE$	1952-2008	31.948	GROWTH* CHANGE_UNEMPLOYMENT_RATE	0.036 0.903
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH$ + $\beta_3 HEALTHCARE$ + $\beta_4 DEBT$	1948-2008	37.503	GROWTH* HEALTHCARE DEBT	0.040 0.359 0.581
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH$ + $\beta_3 INTEREST_RATE$	1956-2008	30.368	GROWTH* INTEREST_RATE	0.054 0.992
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH + \beta_3 GOLD$ + $\beta_4 OIL$	1952-2008	43.928	GROWTH* GOLD OIL	0.070 0.549 0.175
<i>VOTE</i> = $\beta_1 + \beta_2 GROWTH + \beta_3 DEFICIT$ + $\beta_4 EXCHANGE_RATE$	1948-2008	35.882	GROWTH* DEFICIT EXCHANGE_RATE	0.025 0.450 0.848

*Denotes significant p-value at 5 % level of significance

Emission de polluants par une voiture

Nous avons évidemment eu l'idée de cet exercice en contemplant le scandale Volkswagen : nous nous sommes demandés si une étude statistique simple n'aurait pas permis de le détecter. Bien sûr, la première étape d'une telle démarche est de collecter des données. Celles utilisées dans cet exercice sont extraites de la page

<https://www.data.gouv.fr/fr/datasets/emissions-de-co2-et-de-polluants-des-vehicules-commercialises-en-france/>

du site gouvernemental de données ouvertes [data.gouv.fr](https://www.data.gouv.fr/); en réalité, nous avons extrait 836 individus statistiques sur les 55 044 que contenait le fichier. Ce sujet d'examen retrace la manière dont nous avons abordé ce jeu de données et les conclusions que nous avons pu en tirer (ou pas).

Pour chacun de ces individus statistiques, qui sont des modèles de voitures, il est reporté les variables suivantes :

- `lib_mrq_utac`, la marque;
- `lib_mod_doss`, le modèle; voir précisions en note de pied de page¹;
- `Diesel`, une variable binaire valant 1 si la voiture consomme du diesel et 0 sinon;
- `puiss_admin_98`, la puissance administrative (nombre sans unité – le sujet expliquera plus bas comment elle est calculée);
- `puiss_max`, la puissance maximale (en kW);
- `sixrapports`, une variable binaire valant 1 si le nombre de rapports est supérieur ou égal à six, 0 sinon; le fait d'avoir six rapports au moins est censé réduire la consommation de carburant, voir précisions en note de pied de page².
- `conso_urb`, la consommation urbaine de carburant (en l/100km),
- `conso_exurb`, la consommation mixte de carburant (en l/100km),
- `conso_mixte`, la consommation extra-urbaine de carburant (en l/100km),
- `masse`, la masse (en kg),
- `co2`, l'émission de CO₂ (en g/km).

On peut penser que l'émission de CO₂ est proportionnelle (strictement?) à la consommation de carburant, et cette dernière est influencée par de nombreuses variables reportées ci-dessus. Toutefois, la consommation de carburant est une variable difficile à mesurer (et on peut tricher sur sa mesure), contrairement à ce qui se passe pour ces autres variables. Cette idée est au cœur de ce sujet. Le site [data.gouv.fr](https://www.data.gouv.fr/) précise que les données étudiées ici ont été fournies par l'UTAC. Enquête faite pour vous, sur

https://fr.wikipedia.org/wiki/Union_technique_de_l'automobile,_du_motocycle_et_du_cycle

cet organisme est l'Union technique de l'automobile, du motorcycle et du cycle et c'est une entreprise privée française, appartenant en quasi-totalité au Comité des constructeurs français d'automobiles (CCFA). Les données ne sont donc que collectées, mais pas validées, par le gouvernement, qui ne vérifie pas les mesures.

Faisons connaissance avec les données (1 point)

Il s'agit de commenter la matrice des diagrammes de dispersion reportée page 4.

1. On précisera à chaque réponse quelles cases de la matrice sont utilisées.
 - (a) Que pensez-vous de la modélisation linéaire de `co2` en fonction de chacune des autres variables, pour quelles variables cette modélisation semble-t-elle bonne voire excellente, ou au contraire, pauvre? Justifier la réponse. (0.5 point)
 - (b) Identifiez deux sous-groupes de variables fortement linéairement corrélées entre elles. (0.5 point)

1. Il peut y avoir plusieurs sous-modèles d'un modèle, distincts dans leurs caractéristiques techniques telles que mesurées par les autres variables, mais nous n'avons pas reporté la variable de sous-modèle permettant de faire ces distinctions.

2. Extrait du site <http://vendre.autobiz.fr/> — Les rapports sont liés à la boîte de vitesses; cette dernière est aujourd'hui typiquement proposée en 6 rapports sur la plupart des voitures à vocation routière, au lieu des 5 couramment utilisés depuis deux décennies. Alors que le premier rapport doit être assez court (grande démultiplication) pour faciliter le lancement du véhicule sans risque de calage, le dernier rapport (faible démultiplication) a longtemps été calculé pour atteindre la vitesse maximale. La nécessité de réduire la consommation a introduit une nouvelle règle : le dernier rapport doit abaisser le régime de rotation du moteur afin de l'utiliser dans une zone où il consomme moins. Afin d'éviter soit l'allongement du premier rapport, soit des écarts trop élevés entre chacun d'entre eux, un sixième rapport est ajouté, uniquement utilisé lorsque la voiture a atteint sa vitesse de croisière.

Régressions linéaires simples (2.5 points)

2. Retrouver quantitativement les résultats de la question 1.(a) à partir des sorties de régressions simples des pages 5–8 ; préciser en préambule quel indicateur statistique il faut regarder. (0.5 point)
3. Valider statistiquement, puis écrire, interpréter et valider économiquement les deux modèles suivants : `co2 / Diesel` et `co2 / masse`. (2 points ; attention, les erreurs majeures à cette question sont éliminatoires.)

Réduction du nombre de variables (1.5 points)

Nous préparons ici le terrain d'une analyse de régression multiple, en tâchant de dresser la liste des variables explicatives à y considérer. *Dans tout le problème, on considérera que des r^2 valant plus de 99.5% sont en fait égaux à 100% (seuls des problèmes d'arrondis dans les données empêchent vraisemblablement l'égalité à 100%).*

4. En exploitant les sorties de régression de la page 9, expliquer combien de variables parmi `conso_urb`, `conso_exurb` et `conso_mixte` semblent réellement utiles. Confronter aux résultats de la question 1.(b). Expliquer en particulier comment la variable `conso_mixte` semble avoir été obtenue : par mesure ou par déduction ? (0.5 point)
5. En exploitant la sortie de régression de la page 10, retrouver de manière quantitative et commenter le reste des résultats de la question 1.(b). (0.5 point)

En réalité, la page

https://fr.wikipedia.org/wiki/Cheval_fiscal

explique que la puissance administrative (dite aussi puissance fiscale) est calculée à partir de l'émission de CO2 et de la puissance maximale. Nous ne la retiendrons donc pas pour la suite. En découvrant cela, notre curiosité a été titillée. Est-on sûr que les émissions de CO2 ne sont pas calculées et sont réellement mesurées ? Après recherche, nous constatons que le site

<http://www.ecoscore.be/fr/comment-calculer-les-emissions-de-co2-en-fonction-de-la-quantite-de-carburant-consomme>

donne une formule exacte, dépendant du type de carburant, déduisant ces émissions à partir de la consommation.

6. Retrouvez ce fait sur les sorties de régression reproduites page 11 : expliquez comment ces sorties de régression ont été obtenues et ce qu'elles montrent. Il n'est pas nécessaire d'écrire précisément les relations obtenues, il s'agit de commenter leur sens et intérêt. (0.5 point)

Conclusion intermédiaire : nous éliminons les variables de consommation et la variable de puissance administrative pour le reste de l'étude !

Régression linéaire multiple (2.5 points)

Il s'agit d'expliquer la variable `co2` en fonction des variables restantes.

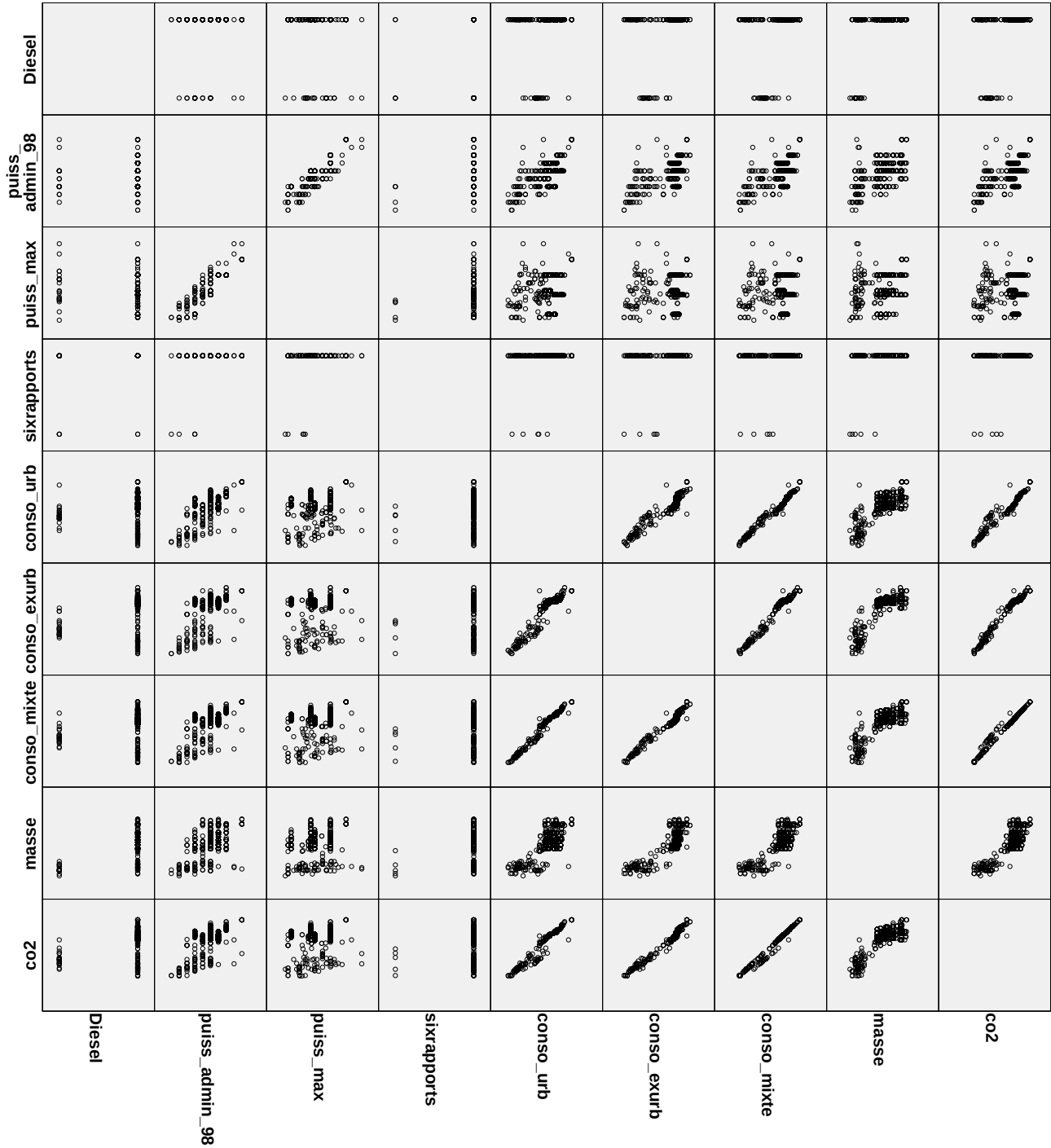
7. Rappeler le nom et le principe de la méthode conduisant aux suites de régressions linéaires multiples reportée à la page 12. (0.5 point)
8. Valider statistiquement le modèle obtenu, puis écrire, interpréter et valider économiquement la relation qu'il propose. (2 points ; attention, les erreurs majeures à cette question sont éliminatoires.)

Scandale Volkswagen (2.5 points)

9. Doit-on (ou peut-on) mettre en doute l'émission de CO2 reportée pour la Volkswagen Golf ? On répondra à cette question en appliquant le modèle de la question 8. On consultera l'extrait des données reproduit en page 3. (2.5 points)

Case	lib_mrj	lib_mod_doss	Diesel	puiss_admin_	puiss_max	sixrapports	conso_urb	conso_exurb	conso_mixte	masse	co2
1	VOLKSWAGEN	GOLF	1	4	66	0	4.6	3.3	3.8	1280	98
2	VOLKSWAGEN	TOURAN	1	8	103	1	6.5	4.6	5.3	1560	139
3	VOLKSWAGEN	PASSAT	1	5	77	1	5.2	3.8	4.3	1499	114
4	VOLKSWAGEN	SCIROCCO	1	7	103	1	6.1	4.1	4.8	1379	127
5	ALFA-ROMEO	BRERA	0	12	147	1	11.8	6.0	8.1	1505	189
6	ALFA-ROMEO	BRERA	1	9	120	1	7.1	4.4	5.4	1555	142
7	ALFA-ROMEO	GIULIETTA	0	9	125	1	7.6	4.6	5.7	1365	131
8	ALFA-ROMEO	GIULIETTA	0	9	120	1	7.8	4.6	5.8	1365	134
9	BMW	SERIE 1	0	7	100	1	7.0	4.5	5.4	1375	125
10	BMW	SERIE 1	1	7	105	1	5.1	3.6	4.1	1395	109
11	BMW	SERIE 1	1	10	135	1	5.9	4.1	4.8	1505	126
12	BMW	SERIE 1	1	12	160	1	6.0	4.2	4.9	1470	128
13	BMW	SERIE 1	0	13	160	1	8.6	5.4	6.6	1425	154
14	CHEVROLET	CAPTIVA	1	11	135	1	8.5	6.0	6.9	2033	183
15	CITROEN	C4	1	5	84	1	4.4	3.4	3.8	1280	97
16	CITROEN	C4 AIRCROSS	1	8	110	1	6.8	4.9	5.6	1570	146
17	CITROEN	C4 PICASSO	0	9	115	1	8.3	4.9	6.1	1371	142
18	FIAT	BRAVO	1	6	77	1	6.3	4.1	4.9	1395	129

FIGURE 1 – Un extrait des données considérées.



Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,375 ^a	,140	,139	21,817

a. Prédicteurs : (Constante), Diesel

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	64873,969	1	64873,969	136,294	,000 ^b
Résidus	396972,806	834	475,987		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), Diesel

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	140,455	4,651		30,196	,000
	Diesel	55,032	4,714	,375	11,674	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,584 ^a	,341	,341	19,099

a. Prédicteurs : (Constante), puiss_admin_98

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	157637,719	1	157637,719	432,169	,000 ^b
Résidus	304209,056	834	364,759		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), puiss_admin_98

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	95,512	4,785		19,960	,000
	puiss_admin_98	10,693	,514	,584	20,789	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,193 ^a	,037	,036	23,090

a. Prédicteurs : (Constante), puiss_max

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	17213,957	1	17213,957	32,288	,000 ^b
Résidus	444632,818	834	533,133		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), puiss_max

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	163,947	5,356		30,613	,000
	puiss_max	,281	,050	,193	5,682	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,194 ^a	,037	,036	23,088

a. Prédicteurs : (Constante), sixrapports

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	17295,681	1	17295,681	32,448	,000 ^b
Résidus	444551,094	834	533,035		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), sixrapports

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	135,400	10,325		13,114	,000
	sixrapports	58,991	10,356	,194	5,696	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,962 ^a	,925	,925	6,450

a. Prédicteurs : (Constante), conso_urb

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	427148,891	1	427148,891	10266,971	,000 ^b
	Résidus	34697,884	834	41,604		
	Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), conso_urb

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	18,444	1,747		10,556	,000
	conso_urb	19,161	,189	,962	101,326	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,974 ^a	,949	,949	5,328

a. Prédicteurs : (Constante), conso_exurb

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	438168,047	1	438168,047	15432,930	,000 ^b
	Résidus	23678,728	834	28,392		
	Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), conso_exurb

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-18,858	1,724		-10,941	,000
	conso_exurb	33,256	,268	,974	124,229	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,992 ^a	,984	,984	3,005

a. Prédicteurs : (Constante), conso_mixte

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	454314,133	1	454314,133	50300,805	,000 ^b
Résidus	7532,643	834	9,032		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), conso_mixte

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-8,116	,907		-8,945	,000
	conso_mixte	27,380	,122	,992	224,278	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,767 ^a	,588	,588	15,102

a. Prédicteurs : (Constante), masse

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	271634,123	1	271634,123	1190,998	,000 ^b
Résidus	190212,652	834	228,073		
Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), masse

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	71,406	3,592		19,881	,000
	masse	,054	,002	,767	34,511	,000

a. Variable dépendante : co2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,904 ^a	,817	,817	,5056

a. Prédicteurs : (Constante), conso_exurb

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	950,282	1	950,282	3717,706	,000 ^b
Résidus	213,179	834	,256		
Total	1163,461	835			

a. Variable dépendante : conso_urb

b. Prédicteurs : (Constante), conso_exurb

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-,750	,164		-4,588	,000
	conso_exurb	1,549	,025	,904	60,973	,000

a. Variable dépendante : conso_urb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,999 ^a	,998	,998	,0333

a. Prédicteurs : (Constante), conso_exurb, conso_urb

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	605,116	2	302,558	273182,362	,000 ^b
Résidus	,923	833	,001		
Total	606,039	835			

a. Variable dépendante : conso_mixte

b. Prédicteurs : (Constante), conso_exurb, conso_urb

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	,039	,011		3,564	,000
	conso_urb	,421	,002	,584	184,762	,000
	conso_exurb	,544	,004	,440	139,377	,000

a. Variable dépendante : conso_mixte

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,886 ^a	,785	,785	,596

a. Prédicteurs : (Constante), puiss_max

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1082,827	1	1082,827	3052,528	,000 ^b
	Résidus	295,846	834	,355		
	Total	1378,673	835			

a. Variable dépendante : puiss_admin_98

b. Prédicteurs : (Constante), puiss_max

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	1,667	,138		12,067	,000
	puiss_max	,071	,001	,886	55,250	,000

a. Variable dépendante : puiss_admin_98

Régression simple dans le cas Diesel = 1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
	Diesel = 1 (sélectionné)			
1	1,000 ^a	,999	,999	,592

a. Prédicteurs : (Constante), conso_mixte

ANOVA^{a,b}

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	390700,609	1	390700,609	1114160,468	,000 ^c
Résidus	284,743	812	,351		
Total	390985,351	813			

a. Variable dépendante : co2

b. Sélection exclusive des observations pour lesquelles Diesel = 1

c. Prédicteurs : (Constante), conso_mixte

Coefficients^{a,b}

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-1,243	,188		-6,627	,000
	conso_mixte	26,516	,025	1,000	1055,538	,000

a. Variable dépendante : co2

b. Sélection exclusive des observations pour lesquelles Diesel = 1

Régression simple dans le cas Diesel = 0

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
	Diesel = 0 (sélectionné)			
1	,998 ^a	,997	,996	1,023

a. Prédicteurs : (Constante), conso_mixte

ANOVA^{a,b}

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	5966,528	1	5966,528	5702,231	,000 ^c
Résidus	20,927	20	1,046		
Total	5987,455	21			

a. Variable dépendante : co2

b. Sélection exclusive des observations pour lesquelles Diesel = 0

c. Prédicteurs : (Constante), conso_mixte

Coefficients^{a,b}

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	-4,186	1,928		-2,171	,042
	conso_mixte	23,890	,316	,998	75,513	,000

a. Variable dépendante : co2

b. Sélection exclusive des observations pour lesquelles Diesel = 0

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,772 ^a	,596	,594	14,981
2	,771 ^b	,595	,593	14,997

a. Prédicteurs : (Constante), Diesel, puiss_max, sixrapports, masse

b. Prédicteurs : (Constante), Diesel, puiss_max, masse

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	275348,700	4	68837,175	306,725	,000 ^b
	Résidus	186498,075	831	224,426		
	Total	461846,775	835			
2	Régression	274732,227	3	91577,409	407,197	,000 ^c
	Résidus	187114,548	832	224,897		
	Total	461846,775	835			

a. Variable dépendante : co2

b. Prédicteurs : (Constante), Diesel, puiss_max, sixrapports, masse

c. Prédicteurs : (Constante), Diesel, puiss_max, masse

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	49,881	7,451		6,694	,000
	puiss_max	,072	,033	,049	2,175	,030
	sixrapports	11,724	7,074	,038	1,657	,098
	masse	,051	,002	,725	29,300	,000
	Diesel	9,156	3,667	,062	2,497	,013
2	(Constante)	59,069	4,984		11,851	,000
	puiss_max	,079	,033	,054	2,406	,016
	masse	,051	,002	,727	29,383	,000
	Diesel	10,589	3,568	,072	2,968	,003

a. Variable dépendante : co2

Variables exclues^a

Modèle	Bêta In	t	Sig.	Corrélation partielle	Statistiques de colinéarité
					Tolérance
2 sixrapports	,038 ^b	1.657	,098	,057	,902

a. Variable dépendante : co2

b. Prédicteurs dans le modèle : (Constante), Diesel, puiss_max, masse

Référendums écologiques en Suisse

Cet exercice est tiré de l'article «What does it take to sell environmental policy? An empirical analysis of referendum data», publié en 2006 dans le journal *Environmental and Resource Economics* et co-écrit par Daniel Halbherr (du département de marketing d'HEC Paris et qui sera peut-être votre enseignant pour le cours de marketing en M1).

Il repose sur des données réelles de référendums suisses portant sur des questions environnementales. Pour les 45 tels référendums ayant eu lieu entre 1977 et 2003 et pour chacun des 26 cantons suisses, on veut expliquer la proportion de votes en faveur de la mesure environnementale (variable : **Vote oui** ; unités : %) en fonction de diverses variables explicatives dont on dresse la liste ci-dessous. On dispose ainsi de $45 \times 26 = 1\,170$ séries de données.

Les variables explicatives considérées sont les suivantes :

- **Impact conso** est une variable binaire indiquant si la mesure soumise au vote restreindra la consommation (code : 1 ; p.ex., en empêchant de conduire un véhicule certains jours, en empêchant la construction d'une autoroute, etc.) ou au contraire (code : 0), sera neutre sur la consommation (p.ex., remplacement de l'énergie nucléaire par des énergies renouvelables) ou l'augmentera (p.ex., en favorisant les transports publics).
- **Impact taxes** est une variable binaire indiquant si la mesure soumise au vote aura pour effet d'augmenter les contributions fiscales directes ou indirectes des citoyens ou usagers (code : 1 ; p.ex., augmentation des tarifs de péage autoroutier, création d'une éco-taxe, etc.) ou au contraire (code : 0), n'entraînera aucune hausse pour les charges publiques, voire les diminuera (p.ex., non-construction d'une autoroute initialement planifiée).
- **Soutien** est une variable binaire indiquant si la mesure soumise au vote bénéficie (code : 1) ou non (code : 0) du soutien de l'organisation patronale suisse majoritaire, nommée Economiesuisse (en un mot!).
- **Préfér. environ.** (unités : %) est une variable quantitative reportant la part de la population en âge de voter considérant les problèmes environnementaux comme importants (source : enquête annuelle réalisée par l'institut de recherche GfS).
- **Taux croissance** (unités : %) désigne le taux de croissance du produit intérieur brut cantonal dans l'année précédent le scrutin.
- **Taux chômage** (unités : %) désigne le taux de chômage cantonal dans l'année précédent le scrutin.
- **Densité** (unités : nombre d'habitants par km²) est une mesure de la densité de population, effectuée en 1997 (source : gouvernement suisse, <http://www.admin.ch>).
- **Contre-prop.** est une variable binaire indiquant si une contre-proposition plus modérée est simultanément proposée au vote référendaire (codes : 1 si c'est le cas et 0 sinon). C'est une particularité du système référendaire suisse.

Un extrait des données, telles qu'elles se présentent sous SPSS, est reproduit à la page suivante. On notera que les deux premières colonnes indiquent respectivement le canton suisse (ZH : Zürich dans l'extrait) et les référendums concernés (ces derniers, par indication de leur date, même s'il peut y avoir plusieurs référendums le même jour). Par ailleurs, des statistiques descriptives pour les différentes variables (explicatives et à expliquer) sont également procurées plus loin dans le sujet.

Partie 1 : Faisons connaissance avec les données (1 point)

Il faut répondre aux questions qui suivent en quelques minutes, et de manière très brève. On utilisera les statistiques descriptives fournies et l'extrait des données.

* Suisse.sav [Jeu de données] - IBM SPSS Statistics Editeur de données												
Fichier Édition Affichage Données Transformer Analyse Marketing direct Graphiques Utilitaires Fenêtre Aide												
3 :												
	Canton	Date	VoteOui	Conso	Taxes	Soutien	PreEnv	TauxCrois	TauxChomage	Densite	ContreProp	
1	ZH	25-Sep-1977	46,54	0	0	0	74	2,02	,40	683	0	
2	ZH	26-Feb-1978	43,38	0	0	0	74	2,02	,40	683	0	
3	ZH	28-May-1978	40,95	1	0	0	74	2,02	,40	683	0	
4	ZH	18-Feb-1979	45,24	0	0	0	73	2,02	,27	683	0	
5	ZH	20-May-1979	78,16	0	0	0	73	2,02	,27	683	1	
6	ZH	27-Feb-1983	51,71	0	0	1	71	-27	,30	683	0	
7	ZH	26-Feb-1984	58,37	0	1	0	71	1,28	,70	683	0	
8	ZH	23-Sep-1984	44,81	0	1	0	71	1,28	,70	683	0	
9	ZH	23-Sep-1984	41,64	0	0	0	71	1,28	,70	683	0	
10	ZH	07-Dec-1986	43,13	0	0	0	73	2,67	,80	683	0	
11	ZH	06-Dec-1987	58,51	0	0	0	78	4,76	,60	683	0	
12	ZH	06-Dec-1987	66,84	0	1	1	78	4,76	,60	683	0	
13	ZH	12-Jun-1988	47,76	0	0	0	74	3,33	,50	683	0	
14	ZH	01-Apr-1990	31,83	1	0	0	70	3,78	,40	683	0	
15	ZH	01-Apr-1990	36,56	1	0	0	70	3,78	,40	683	0	
16	ZH	01-Apr-1990	32,94	1	0	0	70	3,78	,40	683	0	
17	ZH	01-Apr-1990	37,53	1	0	0	70	3,78	,40	683	0	
18	ZH	23-Sep-1990	70,30	0	0	0	70	3,78	,40	683	1	
19	ZH	23-Sep-1990	52,68	0	0	0	70	3,78	,40	683	0	
20	ZH	23-Sep-1990	45,01	0	0	0	70	3,78	,40	683	0	
21	ZH	03-Mar-1991	43,65	0	0	0	61	2,30	,30	683	0	
22	ZH	03-Mar-1991	43,08	0	0	0	61	2,30	,30	683	0	
23	ZH	17-May-1992	69,92	0	0	0	50	-1,31	,70	683	1	
24	ZH	27-Sep-1992	62,56	0	1	1	50	-1,31	,70	683	0	
25	ZH	07-Mar-1993	61,64	0	1	1	47	-2,26	2,10	683	0	
26	ZH	06-Jun-1993	46,19	0	0	0	47	-2,26	2,10	683	0	
27	ZH	20-Feb-1994	76,17	0	1	1	47	-4,15	4,10	683	0	
28	ZH	20-Feb-1994	79,89	0	0	1	47	-4,15	4,10	683	0	
29	ZH	20-Feb-1994	75,10	0	0	1	47	-4,15	4,10	683	0	
30	ZH	20-Feb-1994	55,49	0	0	0	47	-4,15	4,10	683	0	
31	ZH	12-Mar-1995	56,62	0	0	1	31	1,80	4,50	683	0	
32	ZH	09-Jun-1996	83,45	0	0	1	20	,97	4,00	683	0	
33	ZH	07-Jun-1998	37,69	1	0	1	19	1,87	5,20	683	0	
34	ZH	27-Sep-1998	27,35	0	0	0	19	1,87	5,20	683	0	
35	ZH	27-Sep-1998	66,94	0	0	1	19	1,87	5,20	683	0	
36	ZH	29-Nov-1998	65,51	0	1	1	19	1,87	5,20	683	0	

Statistiques descriptives

	N	Minimum	Maximum	Somme	Moyenne	Ecart type
Vote oui	1170	9,97	99,00	52620,41	44,9747	17,14809
Impact conso	1170	0	1	234	,20	,400
Impact taxes	1170	0	1	260	,22	,416
Soutien	1170	0	1	312	,27	,442
Préfér. environ.	1170	14	78	58422	49,93	23,208
Taux croissance	1170	-7,97	14,82	1302,71	1,1134	2,97745
Taux chômage	1170	,00	7,80	2114,28	1,8071	1,79176
Densité	1170	26	5218	517635	442,42	994,224
Contre-prop.	1170	0	1	182	,16	,363
N valide (liste)	1170					

1. Combien de mesures référendaires avaient-elles un impact sur la consommation? Sur les contributions directes ou indirectes? Combien bénéficiaient-elles du soutien d'Economiesuisse? Combien ont été accompagnées de contre-propositions? (*0.5 point*)
2. Comment semble varier **Préfér. environ.** au cours du temps? Donnez des ordres de grandeur typiques pour **Taux croissance**, **Taux chômage** et **Densité**, pas uniquement pour Zürich. (*0.5 point*)

Partie 2 : Régressions linéaires simples / questions éliminatoires (3 points)

Les deux questions qui suivent sont fondamentales, et des réponses gravement incorrectes sont éliminatoires.

3. Parmi l'ensemble des régressions linéaires simples reproduites à la fin du sujet : lesquelles sont valides statistiquement, lesquelles sont valides économiquement ou politiquement? Regroupez ces dernières selon leur qualité respective. On écrira une réponse synthétique. (*1 point ; une réponse gravement incorrecte est éliminatoire*)
4. Ecrivez la relation proposée par les deux meilleurs modèles à la fois valides statistiquement et économiquement / politiquement. Interprétez les coefficients de ces deux relations. (*2 points ; une réponse gravement incorrecte est éliminatoire*)

Partie 3 : Régression linéaires multiples (4.5 points)

5. Que pensez-vous de la régression linéaire complète, celle sur toutes les variables explicatives? Que faut-il en faire, le cas échéant? (*0.5 point*)
6. Rappelez le nom et le procédé des méthodes intitulées «Méthode 1» et «Méthode 2». Recommandent-elles le même modèle? Et si oui, ce modèle est-il valide statistiquement, puis économiquement et politiquement? (*1 point*)
7. Classez l'ensemble des huit variables explicatives par ordre décroissant d'influence. On pourra raisonner selon les valeurs des coefficients de pente, associées au résultat de la question 2, d'une part, mais aussi et surtout selon une autre méthode plus objective, d'autre part. Cette question est difficile. (*1 point*)
8. Pourquoi a-t-on envie de considérer les deux autres modèles indiqués en fin de sujet? Parmi l'ensemble des modèles de régression linéaire multiple, expliquez celui qui a votre préférence. (*0.5 point*)
9. Ecrivez la relation qu'il propose et interprétez soigneusement tous ses coefficients. (*1.5 points*)

Partie 4 : Typicité ou atypicité des observations (1.5 points)

10. Le résultat du 27 février 1983 à Zürich est-il typique du modèle construit à la question précédente? (*1.5 points*)

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,460 ^a	,212	,211	15,22832

a. Prédicteurs : (Constante), Impact conso

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	72891,460	1	72891,460	314,321	,000 ^b
Résidus	270861,098	1168	231,902		
Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Impact conso

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	48,921	,498		98,284	,000
	Impact conso	-19,733	1,113	-,460	-17,729	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,199 ^a	,039	,039	16,81357

a. Prédicteurs : (Constante), Impact taxes

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	13563,550	1	13563,550	47,979	,000 ^b
Résidus	330189,008	1168	282,696		
Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Impact taxes

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	43,155	,557		77,426	,000
	Impact taxes	8,190	1,182	,199	6,927	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,456 ^a	,208	,208	15,26528

a. Prédicteurs : (Constante), Soutien

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	71574,969	1	71574,969	307,151	,000 ^b
	Résidus	272177,589	1168	233,029		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Soutien

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	40,258	,521		77,249	,000
	Soutien	17,687	1,009	,456	17,526	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,169 ^a	,028	,028	16,90996

a. Prédicteurs : (Constante), Préf. environ.

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	9766,888	1	9766,888	34,156	,000 ^b
	Résidus	333985,670	1168	285,947		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Préf. environ.

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	38,756	1,173		33,030	,000
	Préf. environ.	,125	,021	,169	5,844	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,153 ^a	,024	,023	16,95237

a. Prédicteurs : (Constante), Taux croissance

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	8089,288	1	8089,288	28,148	,000 ^b
Résidus	335663,271	1168	287,383		
Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Taux croissance

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	45,958	,529		86,853	,000
	Taux croissance	-,883	,167	-,153	-5,305	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,141 ^a	,020	,019	16,98336

a. Prédicteurs : (Constante), Taux chômage

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	6861,209	1	6861,209	23,788	,000 ^b
Résidus	336891,349	1168	288,434		
Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Taux chômage

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	42,531	,705		60,299	,000
	Taux chômage	1,352	,277	,141	4,877	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,179 ^a	,032	,031	16,87717

a. Prédicteurs : (Constante), Densité

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	11060,821	1	11060,821	38,832	,000 ^b
	Résidus	332691,737	1168	284,839		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Densité

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	43,606	,540		80,738	,000
	Densité	,003	,000	,179	6,232	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,130 ^a	,017	,016	17,01072

a. Prédicteurs : (Constante), Contre-prop.

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	5774,539	1	5774,539	19,956	,000 ^b
	Résidus	337978,019	1168	289,365		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Contre-prop.

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	44,021	,541		81,343	,000
	Contre-prop.	6,130	1,372	,130	4,467	,000

a. Variable dépendante : Vote oui

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,743 ^a	,552	,549	11,51110

a. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Taux croissance, Soutien, Préfér. environ.

ANOVA^a

Modèle	Somme des carrés	ddl	Carré moyen	F	Sig.
1 Régression	189913,763	8	23739,220	179,157	,000 ^b
Résidus	153838,795	1161	132,505		
Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Taux croissance, Soutien, Préfér. environ.

Coefficients^a

Modèle	Coefficients non standardisés		Coefficients standardisés	t	Sig.
	B	Ecart standard	Bêta		
1 (Constante)	20,287	1,490		13,620	,000
Impact conso	-13,046	,912	-,304	-14,310	,000
Impact taxes	1,046	,877	,025	1,192	,233
Soutien	20,564	,933	,531	22,047	,000
Préfér. environ.	,317	,020	,429	16,136	,000
Taux croissance	-,056	,126	-,010	-,439	,660
Taux chômage	1,086	,268	,114	4,060	,000
Densité	,003	,000	,162	8,071	,000
Contre-prop.	16,882	1,057	,357	15,967	,000

a. Variable dépendante : Vote oui

Méthode 1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,743 ^a	,552	,549	11,51110
2	,743 ^b	,552	,550	11,50710
3	,743 ^c	,552	,550	11,50964

a. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Taux croissance, Soutien, Préfér. environ.

b. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Soutien, Préfér. environ.

c. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Soutien, Préfér. environ.

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	189913,763	8	23739,220	179,157	,000 ^b
	Résidus	153838,795	1161	132,505		
	Total	343752,558	1169			
2	Régression	189888,171	7	27126,882	204,865	,000 ^c
	Résidus	153864,387	1162	132,413		
	Total	343752,558	1169			
3	Régression	189687,859	6	31614,643	238,652	,000 ^d
	Résidus	154064,699	1163	132,472		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Taux croissance, Soutien, Préfér. environ.

c. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Soutien, Préfér. environ.

d. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Soutien, Préfér. environ.

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	20,287	1,490		13,620	,000
	Impact conso	-13,046	,912	-,304	-14,310	,000
	Impact taxes	1,046	,877	,025	1,192	,233
	Soutien	20,564	,933	,531	22,047	,000
	Préfér. environ.	,317	,020	,429	16,136	,000
	Taux croissance	-,056	,126	-,010	-,439	,660
	Taux chômage	1,086	,268	,114	4,060	,000
	Densité	,003	,000	,162	8,071	,000
	Contre-prop.	16,882	1,057	,357	15,967	,000
2	(Constante)	20,228	1,483		13,641	,000
	Impact conso	-13,108	,900	-,306	-14,561	,000
	Impact taxes	1,075	,874	,026	1,230	,219
	Soutien	20,599	,929	,531	22,172	,000
	Préfér. environ.	,316	,020	,427	16,189	,000
	Taux chômage	1,104	,265	,115	4,173	,000
	Densité	,003	,000	,162	8,063	,000
	Contre-prop.	16,970	1,038	,359	16,348	,000
3	(Constante)	20,667	1,440		14,357	,000
	Impact conso	-13,384	,872	-,312	-15,350	,000
	Soutien	20,847	,907	,538	22,984	,000
	Préfér. environ.	,313	,019	,423	16,153	,000
	Taux chômage	1,071	,263	,112	4,070	,000
	Densité	,003	,000	,162	8,089	,000
	Contre-prop.	16,876	1,035	,357	16,299	,000

a. Variable dépendante : Vote oui

Variables exclues^a

Modèle		Bêta In	t	Sig.	Corrélation partielle	Statistiques de colinéarité
						Tolérance
2	Taux croissance	-,010 ^b	-,439	,660	-,013	,801
3	Taux croissance	-,012 ^c	-,532	,595	-,016	,805
	Impact taxes	,026 ^c	1,230	,219	,036	,857

a. Variable dépendante : Vote oui

b. Prédicteurs dans le modèle : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Impact taxes, Soutien, Préfér. environ.

c. Prédicteurs dans le modèle : (Constante), Contre-prop., Densité, Impact conso, Taux chômage, Soutien, Préfér. environ.

Méthode 2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,460 ^a	,212	,211	15,22832
2	,598 ^b	,357	,356	13,75802
3	,645 ^c	,416	,415	13,12007
4	,716 ^d	,513	,512	11,98423
5	,739 ^e	,545	,543	11,58632
6	,743 ^f	,552	,550	11,50964

a. Prédicteurs : (Constante), Impact conso

b. Prédicteurs : (Constante), Impact conso, Soutien

c. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ.

d. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop.

e. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop., Densité

f. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop., Densité, Taux chômage

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	72891,460	1	72891,460	314,321	,000 ^b
	Résidus	270861,098	1168	231,902		
	Total	343752,558	1169			
2	Régression	122859,059	2	61429,530	324,538	,000 ^c
	Résidus	220893,499	1167	189,283		
	Total	343752,558	1169			
3	Régression	143041,769	3	47680,590	276,993	,000 ^d
	Résidus	200710,789	1166	172,136		
	Total	343752,558	1169			
4	Régression	176433,127	4	44108,282	307,114	,000 ^e
	Résidus	167319,431	1165	143,622		
	Total	343752,558	1169			
5	Régression	187493,948	5	37498,790	279,336	,000 ^f
	Résidus	156258,610	1164	134,243		
	Total	343752,558	1169			
6	Régression	189687,859	6	31614,643	238,652	,000 ^g
	Résidus	154064,699	1163	132,472		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Impact conso

c. Prédicteurs : (Constante), Impact conso, Soutien

d. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ.

e. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop.

f. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop., Densité

g. Prédicteurs : (Constante), Impact conso, Soutien, Préfér. environ. , Contre-prop., Densité, Taux chômage

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	48,921	,498		98,284	,000
	Impact conso	-19,733	1,113	-,460	-17,729	,000
2	(Constante)	44,334	,531		83,497	,000
	Impact conso	-16,814	1,021	-,392	-16,460	,000
	Soutien	15,012	,924	,387	16,248	,000
3	(Constante)	34,280	1,058		32,414	,000
	Impact conso	-15,782	,979	-,368	-16,125	,000
	Soutien	17,363	,907	,448	19,133	,000
	Préfér. environ.	,185	,017	,250	10,828	,000
4	(Constante)	25,922	1,111		23,339	,000
	Impact conso	-13,694	,904	-,320	-15,141	,000
	Soutien	22,127	,886	,571	24,979	,000
	Préfér. environ.	,268	,016	,362	16,217	,000
	Contre-prop.	16,266	1,067	,344	15,248	,000
5	(Constante)	24,554	1,084		22,643	,000
	Impact conso	-13,694	,874	-,320	-15,661	,000
	Soutien	22,127	,856	,571	25,836	,000
	Préfér. environ.	,268	,016	,362	16,774	,000
	Contre-prop.	16,266	1,031	,344	15,771	,000
	Densité	,003	,000	,179	9,077	,000
6	(Constante)	20,667	1,440		14,357	,000
	Impact conso	-13,384	,872	-,312	-15,350	,000
	Soutien	20,847	,907	,538	22,984	,000
	Préfér. environ.	,313	,019	,423	16,153	,000
	Contre-prop.	16,876	1,035	,357	16,299	,000
	Densité	,003	,000	,162	8,089	,000
	Taux chômage	1,071	,263	,112	4,070	,000

a. Variable dépendante : Vote oui

Autre modèle #1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,739 ^a	,545	,543	11,58632

a. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Préfér. environ. , Soutien

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	187493,948	5	37498,790	279,336	,000 ^b
	Résidus	156258,610	1164	134,243		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Contre-prop., Densité, Impact conso, Préfér. environ. , Soutien

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	24,554	1,084		22,643	,000
	Impact conso	-13,694	,874	-,320	-15,661	,000
	Soutien	22,127	,856	,571	25,836	,000
	Préfér. environ.	,268	,016	,362	16,774	,000
	Densité	,003	,000	,179	9,077	,000
	Contre-prop.	16,266	1,031	,344	15,771	,000

a. Variable dépendante : Vote oui

Autre modèle #2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,716 ^a	,513	,512	11,98423

a. Prédicteurs : (Constante), Contre-prop., Impact conso, Préfér. environ. , Soutien

ANOVA^a

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	176433,127	4	44108,282	307,114	,000 ^b
	Résidus	167319,431	1165	143,622		
	Total	343752,558	1169			

a. Variable dépendante : Vote oui

b. Prédicteurs : (Constante), Contre-prop., Impact conso, Préfér. environ. , Soutien

Coefficients^a

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Ecart standard	Bêta		
1	(Constante)	25,922	1,111		23,339	,000
	Impact conso	-13,694	,904	-,320	-15,141	,000
	Soutien	22,127	,886	,571	24,979	,000
	Préfér. environ.	,268	,016	,362	16,217	,000
	Contre-prop.	16,266	1,067	,344	15,248	,000

a. Variable dépendante : Vote oui

Troisième partie

Fiches de synthèse

Fiches de synthèse

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 bernier 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.

Chapitre 2 : Recueil, représentation et modélisation de données

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 !

Types de données possibles. Ils sont résumés dans ce tableau. Les données qualitatives doivent être codées par des entiers (1, 2, 3, etc.) lors de leur traitement sous un logiciel statistique.

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		

Modéliser les données. En pratique, sur un exercice, la modélisation est une démarche en six temps :

La démarche de modélisation

1. Définir la population effectivement visée, et préciser l'échantillon qui en a été tiré.
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); elles se rapportent bien sûr à l'échantillon.
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.

Attention! Pour la description des données comme pour l'interprétation du paramètre d'intérêt, on choisira avec soin le temps et le mode du verbe (présent, passé composé, futur simple, conditionnel), selon le contexte et les intentions de l'étude menée.

Les deux types de lois que nous rencontrerons le plus souvent pour la loi de population des points 5. et 6. sont :

- la loi de Bernoulli, de paramètre p_0 , où p_0 représente la part de la population telle que [...];
- une «loi générale» dont on s'intéressera juste à la moyenne μ_0 , qui représente une moyenne de population, et éventuellement à son écart-type σ_0 .

Représentation des données : boîte à moustaches. Pour la représentation graphique des données, nous avons revu les boîtes à moustaches, dont il faut connaître le principe de construction (en particulier : les conventions utilisées ici et partout dans le monde différent sans doute de celles que vous aviez vues dans le secondaire en France).

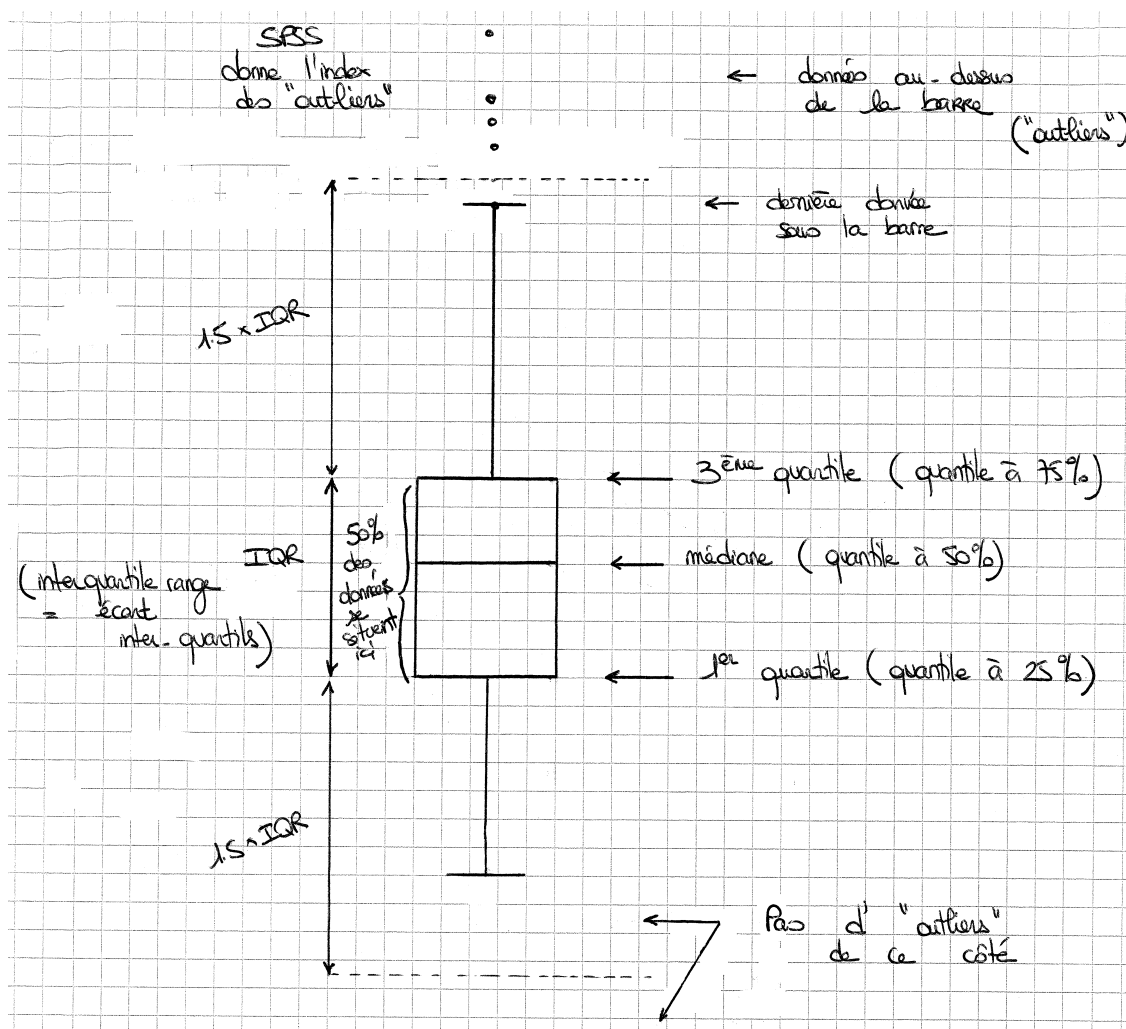


FIGURE S.1 – Principe de construction des boîtes à moustaches.

Chaîne des données. Considérez en résumé la chaîne de recueil, description et traitement des données; le reste des chapitres du cours portent sur les statistiques inférentielles mentionnées dans la dernière case. Nous avons en revanche vu comment recueillir des données (au hasard, par tirage aléatoire dans la population visée) et comment les décrire et les modéliser (voir ci-dessus).

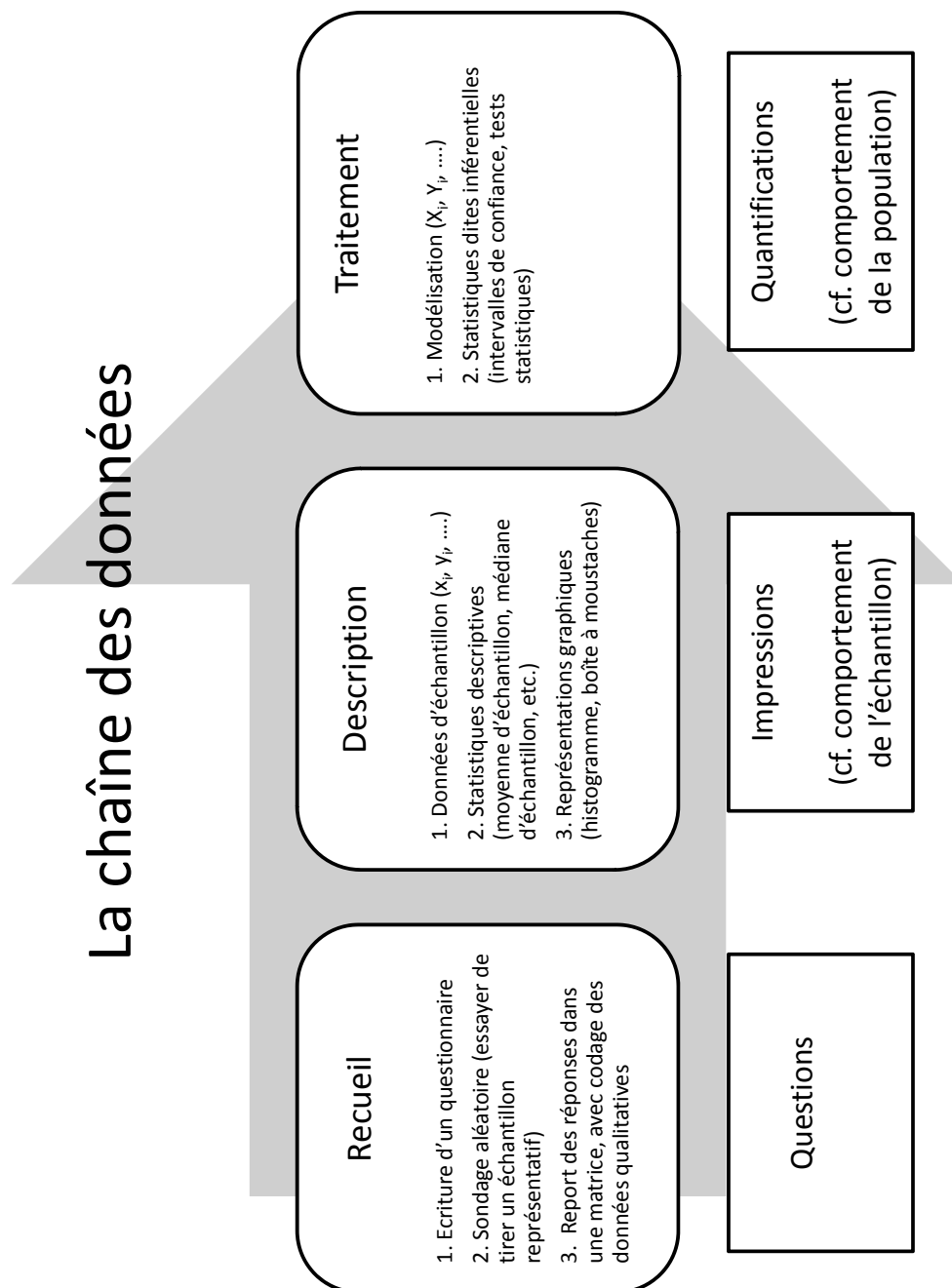


FIGURE S.2 – La chaîne de recueil, description et traitement des données.

Chapitre 3 : Introduction aux tests d'hypothèses

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). Attention ! 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. Il y a une réelle dissymétrie entre les deux hypothèses H_0 et H_1 , et le choix de H_0 en particulier est très politique puisqu'un test tend à camper sur H_0 .

Critères de choix pour H_0 et H_1 .

- Méthode ou faits communément admis, préjugé, avis d'une autorité supérieure : H_0
- Prudence : H_0
- Action risquée : H_1
- Ce qu'on veut prouver : H_1 (et alors, la négation de ce qu'on veut prouver : H_0)

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.

Notion de 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 . 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.

Décision en fonction de la P-valeur p :

- si $p \gg 5\%$, conservation (non-rejet) de H_0 ;
- si $p \ll 5\%$, rejet de H_0 ;
- si $p \approx 5\%$, zone de doute (que conclure ?).

Chapitre 4 : Tests de comparaison à une valeur de référence

On note

$$\bar{X}_n \stackrel{\text{not.}}{=} \frac{1}{n}(X_1 + \dots + X_n) \quad \text{et} \quad \hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

la moyenne empirique et la variance empirique des observations.

Principe S.1. *Test de comparaison d'une moyenne de population μ_0 à une valeur de référence μ_{ref}*

Données : $x_1, \dots, x_n \in \mathbb{R}$ où $n \geq 30$

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 : $\mu_0 = \mu_{\text{ref}}$

Statistique de test :

$$T_n = \sqrt{n} \frac{\bar{X}_n - \mu_{\text{ref}}}{\sqrt{\hat{\sigma}_n^2}}$$

Comportement sous H_0 : $T_n \rightarrow \mathcal{N}(0, 1)$

Comportement sous H_1 : lorsque $\mu_0 > \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus grandes que sous H_0 ; lorsque $\mu_0 < \mu_{\text{ref}}$, la statistique T_n tend à prendre des valeurs plus petites que sous H_0 .

Principe S.2. *Test de comparaison d'une proportion de population p_0 à une valeur de référence p_{ref}*

Données : $x_1, \dots, x_n \in \{0, 1\}$ où $n \geq 30$

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 :

$$T_n = \sqrt{n} \frac{\bar{X}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1 - p_{\text{ref}})}}$$

Comportement sous H_0 : $T_n \rightarrow \mathcal{N}(0, 1)$

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 .

Chapitre 5 : Tests à partir d'échantillons indépendants ou appariés

Données appariées. Les données appariées correspondent dans un tableau (Excel, SPSS) à la comparaison des valeurs de deux colonnes, recueillies donc pour les mêmes individus (exemple : comportements avant et après l'application d'une promotion).

On dispose ainsi de deux séries de données de même longueur, $x_1, \dots, x_n$ et $y_1, \dots, y_n$. 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 celui du principe S.1. On teste ici usuellement que la moyenne des différences est nulle, $H_0 : \Delta = 0$, où Δ désigne l'espérance commune des variables aléatoires $Z_1, \dots, Z_n$ modélisant les z_j .

Données indépendantes. On se place ici dans un cadre où une première série de données $x_1, \dots, x_n$ a été obtenue par échantillonnage sur une première population, et une seconde série de données $y_1, \dots, y_m$, de longueur en général différente, a été obtenue quant à elle sur une seconde population. Concrètement, sous Excel ou SPSS, cela va correspondre à comparer les valeurs d'une colonne fixée entre deux sous-ensembles de lignes du tableau de données.

Données indépendantes : cas des proportions Dans ce cas, les données sont dans $\{0, 1\}$. On peut appliquer un test d'homogénéité du χ^2 (qui est toujours bilatère) ou recourir au principe suivant (qui permet de faire des tests unilatères).

Principe S.3. *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 $y_1, \dots, y_m \in \{0, 1\}$ où $n \geq 30$ et $m \geq 30$

Modélisation associée : deux séries d'observations $X_1, \dots, X_n$ et $Y_1, \dots, Y_m$, 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 :

$$T_{n,m} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\hat{p}_{n+m}(1 - \hat{p}_{n+m})(1/n + 1/m)}}$$

où l'estimateur groupé $\hat{p}_{n+m}$ de la proportion commune sous H_0 est défini par

$$\hat{p}_{n+m} = \frac{n\bar{X}_n + m\bar{Y}_m}{n + m}$$

Comportement sous H_0 : $T_{n,m} \rightarrow \mathcal{N}(0, 1)$ 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 .

Données indépendantes : cas de variables quantitatives générales. 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
	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	
Verres bus	Hypothèse de variances égales	5,005	,030	-1,498	52	,140	-3,2877	2,1944	-7,6911
	Hypothèse de variances inégales			-1,700	37,021	,098	-3,2877	1,9341	-7,2064
									1,1158
									,6311

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, on rejette l'hypothèse H_0 d'égalité des variances (P-valeur 3%) et on lit donc une P-valeur d'égalité des moyennes de populations égale à 9.8%, qui nous impose de conserver l'hypothèse H_0 d'égalité des moyennes.

Chapitre 6 : Tests du χ^2

Il s'agit simplement de savoir lire les sorties SPSS et notamment, y retrouver les P-valeurs des tests, de même que décoder la petite note qui indique si les conditions asymptotiques de validité sont remplies ou pas.

Test du χ^2 d'ajustement à une loi. Une fois n'est pas coutume, il n'y a pas le choix des hypothèses :

- H_0 : Conformité du comportement de l'échantillon à la loi postulée
- H_1 : Non-conformité

Les applications de ce test sont :

- la détection de tricheries et de manipulations ;
- la vérification de la représentativité d'un échantillon ;
- la vérification que des objectifs commerciaux sont atteints.

Il faut savoir lire ce genre de sorties SPSS :

Test du Khi-deux

Chiffre			
	Effectif observé	Effectif théorique	Résidu
0	9	11,6	-2,6
1	11	11,6	-,6
2	8	11,6	-3,6
3	9	11,6	-2,6
4	10	11,6	-1,6
5	5	11,6	-6,6
6	14	11,6	2,4
7	20	11,6	8,4
8	17	11,6	5,4
9	13	11,6	1,4
Total	116		

Test	
	Chiffre
Khi-deux	15,552 ^a
ddl	9
Signification asymptotique	,077

a. 0 cellules (,0%) ont des fréquences théoriques inférieures à 5. La fréquence théorique minimum d'une cellule est 11,6.

Et notamment y retrouver :

- les conditions d'application valide du test (taille d'échantillon $n \geq 30$ et tous les effectifs attendus ≥ 5 , cf. note de bas de tableau) ;
- la P-valeur calculée, ici, 7.7%.

Si condition de validité du test n'est pas vérifiée, il faut alors procéder à un regroupement de certaines classes ou à la suppression de certaines classes.

Test du χ^2 d'indépendance (aussi dit d'homogénéité). Une fois n'est pas coutume, il n'y a pas le choix des hypothèses :

- H_0 : Indépendance des deux variables considérées
- H_1 : Dépendance

Cet test est notamment utilisé pour déterminer l'utilité de la segmentation d'une population de clients en fonction de comportements de consommation éventuellement significativement différents.

Il faut savoir lire ce genre de sorties SPSS :

Tableau croisé Vote (après regroupement) * Année

			Année			Total
			1A	2A	3A	
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.

Et notamment y retrouver :

- les conditions d'application valide du test (taille d'échantillon $n \geq 30$ et tous les effectifs attendus ≥ 5 , cf. note de bas de tableau);
- la P-valeur calculée, ici, ...% (lire uniquement la ligne «Khi-deux de Pearson», nous n'avons pas vu les autres tests en cours).

Si condition de validité du test n'est pas vérifiée, il faut alors procéder à un regroupement de certaines classes ou à la suppression de certaines classes.

Chapitre 7 : Estimation par intervalles

La morale de ce chapitre est simple : il ne faut jamais plus donner un nombre (une estimation) sans précision ni indication de niveau de confiance. 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 et quel niveau de confiance, 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. »

Formules théoriques On rappelle qu'on note z_β le β -quantile de la loi $\mathcal{N}(0, 1)$. Les valeurs les plus utiles sont

- $z_{97.5\%} = 1.96$ pour les intervalles bilatères (symétriques) ;
- $z_{95\%} = 1.645$ pour les intervalles unilatères (majorants ou minorants).

Cas de lois générales. Les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1 - \alpha$ pour la moyenne μ_0 :

$$\begin{aligned} & \left[\bar{X}_n - z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}}, \bar{X}_n + z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] \stackrel{\text{not.}}{=} \left[\bar{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] ; \\ & \left[-\infty, \bar{X}_n + z_{1-\alpha} \sqrt{\frac{\hat{\sigma}_n^2}{n}} \right] ; \\ \text{et} \quad & \left[\bar{X}_n - z_{1-\alpha} \sqrt{\frac{\hat{\sigma}_n^2}{n}}, +\infty \right] . \end{aligned}$$

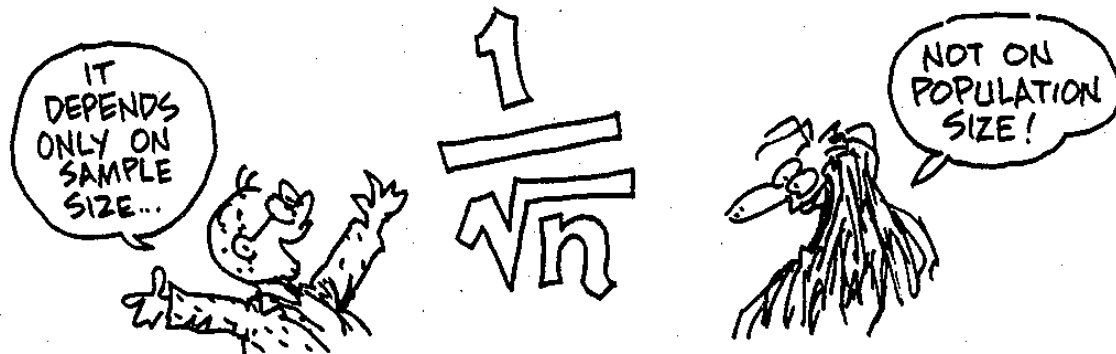
On rappelle que ces formules mettent en jeu l'écart-type empirique :

$$\sqrt{\hat{\sigma}_n^2} = \hat{\sigma}_n .$$

Cas de proportions. Les intervalles suivants sont des intervalles de confiance asymptotiques de niveau $1 - \alpha$ pour la proportion p_0 :

$$\begin{aligned} & \left[\bar{X}_n \pm z_{1-\alpha/2} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right] ; \\ & \left[0, \bar{X}_n + z_{1-\alpha} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right] ; \\ \text{et} \quad & \left[\bar{X}_n - z_{1-\alpha} \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}, 1 \right] . \end{aligned}$$

Remarque. La qualité de l'estimation par intervalles ne dépend donc pas de la taille de la population, mais uniquement de celle de l'échantillon !



Application pratique des formules.

- Ecrire «avec un niveau de confiance de 95 %» (et non pas parler de risque, trop pessimiste, ou de probabilité, terme à réserver aux études théoriques sur variables aléatoires et incompatible avec le calcul pratique sur des données du résultat d'un intervalle de confiance).
- Bien justifier la forme de l'intervalle retenu (symétrique car neutre, ou au contraire majorant ou minorant) en fonction des intentions du commanditaire de l'étude.

Nouveauté 1 : Intervalles de confiance simultanés / correction de Bonferroni. L'exploitation simultanée d'un intervalle de confiance de niveau $1 - \alpha$ et d'un intervalle de confiance $1 - \beta$ vaut avec confiance seulement $1 - (\alpha + \beta)$. Il y a une correction de niveau à faire pour la considération de plusieurs intervalles de confiance (les risques d'erreur s'ajoutant). On l'appelle la correction de Bonferroni.

Nouveauté 2 : 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

$$n \geq \left(\frac{z_{1-\alpha/2}}{2\varepsilon} \right)^2, \quad \text{c'est-à-dire} \quad n \geq \frac{1}{\varepsilon^2} \quad \text{lorsque } 1 - \alpha = 95\%;$$

cependant, une fois l'expérience réalisée, on pourra recourir aux intervalles de confiance plus précis énoncés précédemment, du type

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

- 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

$$z_{1-\alpha/2} \sqrt{\frac{s_t^2}{n}} \leq \varepsilon, \quad \text{soit} \quad n \geq s_t^2 \left(\frac{z_{1-\alpha/2}}{\varepsilon} \right)^2.$$

Chapitre 8 : Régression linéaire simple

Le modèle de la régression linéaire est le modèle le plus simple pour établir une liaison entre deux variables quantitatives. Vous l'avez 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; nous n'avons cependant pas eu le temps d'étudier cette validation *a posteriori* en cours.

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

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j, \quad j = 1, \dots, n,$$

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 .

On appelle estimateurs des moindres carrés de (α_0, β_0) le couple $(\hat{\alpha}_n, \hat{\beta}_n)$ antécédent du minimum d'une certaine fonction :

$$(\hat{\alpha}_n, \hat{\beta}_n) \in \underset{(\alpha, \beta) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{j=1}^n (Y_j - (\alpha + \beta x_j))^2.$$

On dispose par ailleurs d'expressions explicites pour $\hat{\alpha}_n$ et $\hat{\beta}_n$, ce qui permet de calculer leurs valeurs réalisées. A partir de ces estimateurs, on construit les prévisions du modèle $\hat{Y}_j$ et les résidus associés :

$$\hat{Y}_j = \hat{\alpha}_n + \hat{\beta}_n x_j \quad \text{et} \quad \hat{\varepsilon}_j = Y_j - \hat{Y}_j, \quad j = 1, \dots, n.$$

Evidemment, on a les valeurs réalisées des Y_j , on n'a donc pas besoin de les prévoir ou de les estimer, mais ces quantités $\hat{Y}_j$ jouent un rôle important dans l'analyse théorique.

Enfin, on dispose d'un estimateur sans biais de la variance σ_0^2 , donné par

$$\hat{\sigma}_n^2 = \frac{1}{n-2} \sum_{j=1}^n (\hat{\varepsilon}_j)^2 = \frac{\Sigma_R}{n-2}.$$

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 S.4. *Test de l'existence d'une relation linéaire (dans le cadre d'un modèle gaussien)*

Données : $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$

Modélisation associée : valeurs $x_1, \dots, x_n$ 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 :

$$T_n = \sqrt{n \operatorname{Var}(x_1^n)} \frac{\hat{\beta}_n}{\sqrt{\hat{\sigma}_n^2}}$$

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).

Question 2 : Proportion de reconstruction

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 ,

$$\underbrace{\sum_{j=1}^n (Y_j - \bar{Y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_T} = \underbrace{\sum_{j=1}^n (\hat{Y}_j - \bar{Y}_n)^2}_{\stackrel{\text{not.}}{=} \Sigma_E} + \underbrace{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}_{\stackrel{\text{not.}}{=} \Sigma_R}.$$

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

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	...	$\sqrt{\hat{\sigma}_n^2}$

Le deuxième tableau¹ 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	$n - 1$			

Enfin, le troisième tableau donne les valeurs réalisées de :

Coeff.	Err. standard	...	t	Sig.	IC (bornes inf. et sup.)
$\hat{\alpha}_n$	$\sqrt{\frac{\hat{\sigma}_n^2}{n} \left(1 + \frac{(\bar{x}_n)^2}{\text{Var}(x_1^n)} \right)}$		T'_n	P-val. test $H_0 : \alpha_0 = 0$ vs. $H_1 : \alpha_0 \neq 0$	IC sur α_0
$\hat{\beta}_n$	$\sqrt{\frac{\hat{\sigma}_n^2}{n \text{Var}(x_1^n)}}$	...	T_n	P-val. test $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$	IC sur β_0

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.

En pratique : comment exploiter les sorties SPSS ainsi lues

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 ici la relation suivante :

$$\begin{aligned} \text{prix (en milliers d'euros)} = & -29.466 + 5.353 \times \text{surface (en m}^2\text{)} \\ & + \text{résidu d'écart-type } 122.939 \end{aligned}$$

(où l'on peut préciser que le terme constant -29.466 n'est pas significativement différent de zéro).

L'interprétation du terme de résidu est qu'il prend typiquement des valeurs comprises entre -2 et $+2$ fois l'écart-type ; c'est une correction individuelle à la modélisation déterministe globale.

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 signes des coefficients (et parfois même, aux valeurs des coefficients).

Ainsi, la relation précédente admet bien un coefficient de pente positif : le prix d'un appartement augmente avec la surface, ce qui semble économiquement logique. La valeur de ce coefficient de pente admet même (modulo le terme d'ordonnée à l'origine négatif) une interprétation : c'est le prix au mètre carré. Lorsque l'interprétation de l'ordonnée à l'origine (l'estimée de α_0) pose problème, comme c'est le cas ici, 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.

Chapitre 9 : Régression linéaire multiple

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 ou de variables qualitatives binaires, 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.

Le modèle gaussien multiple. On dispose de n données y_t , où $t = 1, \dots, n$, à expliquer, chacune en fonction d'un nombre fini k de données explicatives, dont les valeurs correspondant à y_t sont notées $x_{1,t}, x_{2,t}, \dots, x_{k,t}$. Pour les mêmes raisons et intuitions que dans le chapitre précédent, on modélise les données à expliquer y_t comme la réalisation des variables aléatoires

$$\begin{aligned} Y_t &= \alpha_0 + \beta_{1,0} x_{1,t} + \beta_{2,0} x_{2,t} + \dots + \beta_{k,0} x_{k,t} + \varepsilon_t \\ &= \alpha_0 + \sum_{j=1}^k \beta_{j,0} x_{j,t} + \varepsilon_t, \quad \text{pour } t = 1, \dots, n, \end{aligned}$$

où les $\varepsilon_1, \dots, \varepsilon_n$ (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 $\beta_{1,0}, \dots, \beta_{k,0}$ de la relation linéaire, de même que les (réalisations des) résidus ε_j , ne le sont pas.

Tournez la page !

En pratique : lecture de sorties SPSS

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	
		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Limite supérieure
Modèle	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 $H_0 : \beta_{1,0} = \beta_{2,0} = \dots = \beta_{k,0} = 0$ contre l'hypothèse H_1 qu'au moins un de ces coefficients soit non nul. 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 $j = 1, \dots, k$ chaque fois contre $H_1 : \alpha_0 \neq 0$ et $H_0 : \beta_{j,0} \neq 0$. 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 $\beta_{k,0}$

En pratique : comment exploiter les sorties SPSS ainsi lues

1.1. On commence par regarder la validité globale du modèle, i.e., le résultat du test de Fisher mené dans le deuxième tableau. On continue l'étude seulement si la P-valeur lue est plus petite que 5 %.

1.2. 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 explicatives² uniquement) sont plus petites que 5 %, on continue l'étude.

2. On lit seulement alors la valeur réalisée du coefficient de détermination ajusté r_{ajust}^2 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.

3. 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).

4. Il ne faut pas oublier l'étape de validation économique (au sens large, il s'agit parfois d'une validation politique) : il s'agit d'interpréter les paramètres α_0 et $\beta_{p,0}$ du modèle. On entend par cela, *a minima*, un commentaire sur leur signe ; et, lorsque c'est possible, des commentaires supplémentaires sur leurs valeurs (sur leurs ordres de grandeur).

5. Pour l'identification des données atypiques ou la prévision de nouvelles valeurs, on utilisera le fait que typiquement, le terme de résidu est (au moins en première approximation) compris entre ± 2 fois son écart-type.

2. 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.

Comparaison et choix de modèles Premièrement, on ne considère ici que des modèles statistiquement valides au sens où les étapes 1.1. et 1.2. 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.

Le minimum vital pour obtenir E au problème de régression à l'examen

Ce chapitre de cours est un peu aride et ne peut être bien compris qu'en résolvant de nombreux exercices d'examen, comme nous le ferons ensemble. Campagne BDA ou pas, les séances de cours sur la régression linéaire sont donc vitales. L'expérience montre que les listeux BDA dévalident le cours de statistique plus souvent qu'à leur tour à cause de leur absence à ces séances cruciales. C'est pourquoi nous proposons la «check-list» suivante de compétences.

Le minimum vital pour obtenir E au problème de régression à l'examen

Cette liste est cumulative ; l'absence d'une des compétences suivantes suffit à dévalider le problème de régression, donc l'examen, donc tout le cours de statistique !

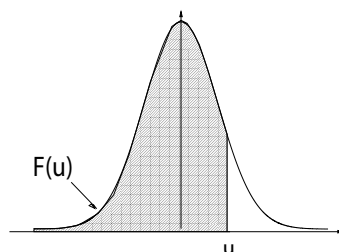
1. Savoir valider statistiquement un modèle (validité globale, validités marginales)
2. Savoir écrire le modèle linéaire proposé, sous une forme intelligible (en donnant le nom des variables), exploitable (en précisant les unités), et statistiquement correcte (ne pas oublier l'existence d'un terme de résidu, dont la taille typique est caractérisée par son écart-type ; attention à l'écriture du terme de résidu)
3. Savoir valider économiquement un modèle (*a minima* : se demander si les signes des coefficients sont logiques ou pas)
4. Savoir prévoir une nouvelle observation et/ou déterminer si telle observation du jeu de données est conforme ou non au modèle (en exploitant le terme de résidu, et le fait qu'il soit typiquement compris entre ± 2 fois son écart-type)

Quatrième partie

Tables des lois statistiques

Loi normale : fonction de répartition

Pour une valeur $u \geq 0$, 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

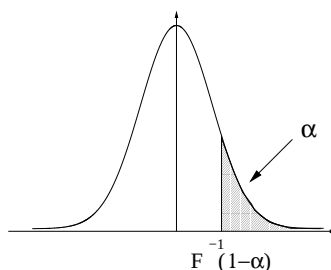
Table pour les grandes valeurs de u :

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 S.3 – 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 $1 - \alpha$.

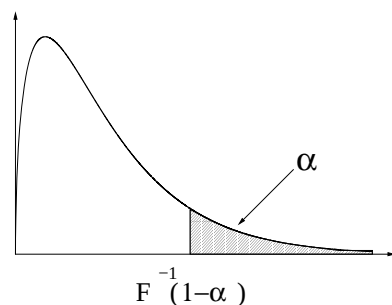


α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.00	∞	3.0902	2.8782	2.7478	2.6521	2.5758	2.5121	2.4573	2.4089	2.3656
0.01	2.3263	2.2904	2.2571	2.2262	2.1973	2.1701	2.1444	2.1201	2.0969	2.0749
0.02	2.0537	2.0335	2.0141	1.9954	1.9774	1.9600	1.9431	1.9268	1.9110	1.8957
0.03	1.8808	1.8663	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 S.4 – 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 $1 - \alpha$.



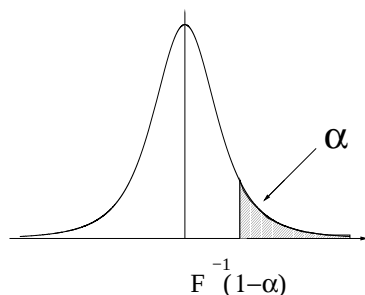
$n \backslash \alpha$	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.001
1	0.0002	0.001	0.0039	0.0158	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21	13.82
3	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.3	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	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 S.5 – 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 $1 - \alpha$.



$n \backslash \alpha$	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 S.6 – Table des quantiles de la loi de Student (en fonction du nombre de degrés de liberté)