HEC PARIS

Apprendre à oser®

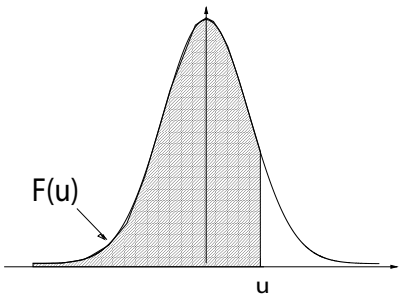
HEC / Grande Ecole L3

STATISTIQUE

Gilles Fortin-Stoltz

(stoltz@hec.fr)

Année universitaire 2023-24



	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

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

1

Vocabulaire et modélisation statistiques

La statistique, contrairement à la voyance, s'appuie sur des données ; pour être crédible, dans la vie comme en entreprise, il vaut mieux s'appuyer sur des assertions étayées et chiffrées que sur des opinions. Ainsi, avec une fine allusion à la devise imprimée sur les billets de dollar américain :

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

1. Vocabulaire statistique

L'exemple fil rouge de ce chapitre est le suivant : comme chaque année en mai, les médias s'intéressent au budget vacances des Français et résument les résultats d'une enquête menée par un institut de sondage sur 1 000 foyers français. A ceux-ci, deux questions ont-été posées. Êtes-vous partis en vacances l'an dernier (réponses possibles : oui ou non) ? Le cas échéant, combien avez-vous dépensé pour ces vacances (on reporte la valeur 0 euro pour ce montant en cas d'absence de vacances). Un extrait du jeu de données ainsi constitué est reproduit ci-contre.

ID	Départ	Budget
1	Oui	2 100
2	Oui	1 350
3	Non	0
4	Oui	1 650
5	Non	0
...	...	...
...	...	...
1 000	Oui	1 800

Individu statistique : c'est l'unité élémentaire de l'étude statistique ; ici, les individus sont des foyers français (au sens de résidents en France).

Population : c'est l'ensemble des individus statistiques ; ici, les environ 30 millions de foyers français.

Echantillon : c'est le sous-ensemble aléatoire d'individus constitué au sein de la population pour l'étude ; les questions n'ont été posées qu'aux individus membres de l'échantillon, ici, 1 000 foyers.

Variables d'intérêt : les quantités mesurées sur chaque individu (effectivement mesurées, pour les individus de l'échantillon, ou qu'on pourrait mesurer, pour les autres individus de la population) ; ici, d'une part, le fait d'être, ou non, parti en vacances, et d'autre part, le montant dépensé.

Dans le tableau reportant les données,

- chaque ligne correspond à un individu statistique, ceux-ci étant d'ailleurs numérotés (cf. colonne ID pour identifiant ou individu) ;
- chaque colonne correspond à une variable d'intérêt : départ en vacances, budget alloué.

Exemple (Les individus statistiques ne sont pas nécessairement des personnes !). Si une étude était conduite par une entreprise sur la fiabilité des machines à café qu'elle place chez différents clients (universités, hôpitaux, etc.), les individus statistiques en question seraient des machines à café.

1.1. Objectifs de l'étude

Paramètres d'intérêt : les objectifs de l'étude statistique, correspondant, à notre niveau, à la moyenne des variables d'intérêt sur l'ensemble de la population.

Ici, nous nous intéressons

— d'une part, à la proportion p_0 de l'ensemble des foyers français partant en vacances ;

— d'autre part, au budget moyen μ_0 consacré par l'ensemble des foyers français aux vacances.

Ces quantités p_0 et μ_0 existent et sont parfaitement déterminées, mais seul un être omniscient, capable de lire les pensées et de connaître l'histoire des $N = 30$ millions de foyers français, serait en mesure d'en donner leur valeur exacte, en effectuant une moyenne de N observations.

Comment fait un être humain plus ordinaire ? Il s'appuie sur les données d'échantillon.

1.2. Données d'échantillon

Le sondage a permis de recueillir deux séries de $n = 1\,000$ observations. On note

— d'une part, $x_1, x_2, \ldots, x_{1\,000}$ les réponses binaires obtenues à la question de l'existence de vacances, avec le codage $x_j = 1$ si le j–ème foyer sondé est parti en vacances, et $x_j = 0$ sinon ;

— d'autre part, $y_1, y_2, \ldots, y_{1\,000}$ les montants déclarés pour le budget vacances, où y_j désigne le montant du j–ème foyer sondé.

Ces deux séries de données peuvent chacun être résumées ainsi.

Cas des proportions. On procure la fréquence empirique sur l'échantillon,

$$\overline{x}_{1\,000} = \frac{1}{1\,000} \sum_{j=1}^{1\,000} x_j \,.$$

Grâce au codage en 0/1, la fréquence empirique $\overline{x}_{1\,000}$ représente bien la proportion de foyers de l'échantillon partis en vacances.

Cas de variables quantitatives générales. Pour les budgets vacances, on reporte d'une part la moyenne d'échantillon $\overline{y}_{1\,000}$, et d'autre part un écart-type d'échantillon $s_{y,1\,000}$:

$$\overline{y}_{1\,000} = \frac{1}{1\,000} \sum_{j=1}^{1\,000} y_j \qquad \text{et} \qquad s_{y,1\,000} = \sqrt{\frac{1}{999} \sum_{j=1}^{1\,000} \big(y_j - \overline{y}_{1\,000}\big)^2} \,.$$

Le fait que l'on divise par 999 et non par 1 000 dans la définition de l'écart-type $s_{y,1\,000}$ est une longue et peu importante histoire, résumée de manière technique, comme certains ont pu le voir en classes préparatoires, par le fait que « l'estimateur de la variance nécessite une normalisation par $n - 1$, et non par n, pour être sans biais ».

Il n'est nul besoin de comprendre en détails la formule d'écart-type d'échantillon ; il suffit de retenir que cet écart-type d'échantillon est une mesure de la dispersion des données $y_1, y_2, \ldots, y_{1\,000}$ autour de leur moyenne $\overline{y}_{1\,000}$.

Exemple. Des valeurs réalistes seraient les suivantes. Sur les 1 000 foyers sondés, 652 sont partis en vacances, soit une proportion $\overline{x}_{1\,000} = 65.2\%$ d'entre eux. En moyenne, les 1 000 foyers sondés ont dépensé $\overline{y}_{1\,000} = 1\,641$ euros pour leurs vacances, avec un écart-type dans les données sous-jacentes de $s_{y,1\,000} = 2\,289$ euros.

2. Modélisation

Là où les choses commencent à être intéressantes (et justifient l'existence de ce cours), c'est qu'il existe des liens entre les moyennes d'échantillons $\overline{x}_n$ et $\overline{y}_n$ et les moyennes de population p_0 et μ_0, qui sont nos paramètres d'intérêt. Supposons que l'échantillon ait été constitué par tirage uniformément au hasard dans la population des 30 millions de foyers français — la dernière partie de ce chapitre reviendra sur la difficulté à assurer cela.

Tirer un échantillon au hasard revient à postuler un modèle d'urne : imaginez une grande urne avec autant de boules que de foyers français, donc environ 30 millions de boules. Sur chaque boule est écrit si le foyer correspondant est ou non parti en vacances l'année écoulée, de même que le montant consacré à ces vacances. Quand on tire 1 000 boules uniformément au hasard, il convient de se demander si on le fait avec ou sans remise. Evidemment, en pratique, un institut de sondage n'interrogera pas deux fois le même foyer : le tirage s'effectue sans remise. Pour autant, on peut approximer ce tirage sans remise par un tirage avec remise (ce qui correspond à approximer une loi hypergéométrique par une loi binomiale), car on ne prélève qu'un petit échantillon au sein d'une grande population. Or, qui dit tirage avec remise dit variables aléatoires indépendantes et identiquement distribuées.

Cas des proportions. Ainsi, dans le cas des proportions :

> Les données $x_1, \dots, x_{1\,000} \in \{0, 1\}$ de départ en vacances sont la réalisation des variables aléatoires indépendantes et identiquement distribuées $X_1, \dots, X_{1\,000}$.

La loi commune des X_j est une loi de Bernoulli (puisque les X_j prennent les valeurs 0 et 1), mais de quel paramètre ? Ce paramètre est la proportion de l'ensemble des boules avec la valeur 1 : c'est exactement la définition de p_0 !

La loi des grands nombres assure que $\overline{X}_n$ convergence en probabilité vers p_0 lorsque $n \to \infty$, ce que l'on peut reformuler de manière plus intuitive par le fait qu'il y a toutes les chances, lorsque n est grand, que la valeur réalisée $\overline{x}_n$ de $\overline{X}_n$ soit proche de p_0.

Pour autant, et contrairement aux formulations trompeuses des journalistes, la part p_0 des foyers français partis en vacances au cours de l'année écoulée n'est pas exactement égale à la proportion $\overline{x}_{1\,000} = 65.2\%$ des foyers de l'échantillon l'ayant fait. On peut tout au plus dire qu'on pense que p_0 est proche de cette valeur $\overline{x}_{1\,000} = 65.2\%$, mais il est important, et même vital, de distinguer les deux concepts p_0 et $\overline{x}_{1\,000}$ dans le cas de ce cours.

Remarque 1.1 (Interprétations probabiliste vs. statistique). Une interprétation probabiliste de p_0 serait que c'est la probabilité qu'un foyer (pris au hasard) soit parti en vacances ; nous préférerons de loin l'interprétation statistique comme la part de l'ensemble des foyers partis en vacances.

Cas de variables quantitatives générales. Le modèle du tirage d'urne avec remise indique ici encore que :

> Les données $y_1, \dots, y_{1\,000}$ de budgets vacances sont la réalisation des variables aléatoires indépendantes et identiquement distribuées $Y_1, \dots, Y_{1\,000}$.

Il est plus difficile ici de concevoir la loi commune des Y_j : on peut la décrire comme une loi discrète correspondant aux montants écrits sur les environ 30 millions de boules de l'urne (une par foyer),

qu'on pourrait appeler la loi de population. Cette loi est complexe, mais nous avons déjà rencontré son espérance : c'est exactement μ_0. On peut donc préciser la modélisation de la manière suivante, et c'est de toute façon tout ce qui nous importe pour appliquer la loi des grandes nombres :

> Les données $y_1, \ldots, y_{1\,000}$ de budgets vacances sont la réalisation des variables aléatoires indépendantes et identiquement distribuées $Y_1, \ldots, Y_{1\,000}$, dont la loi commune est d'espérance $\mathbb{E}[Y_1] = \mu_0$.

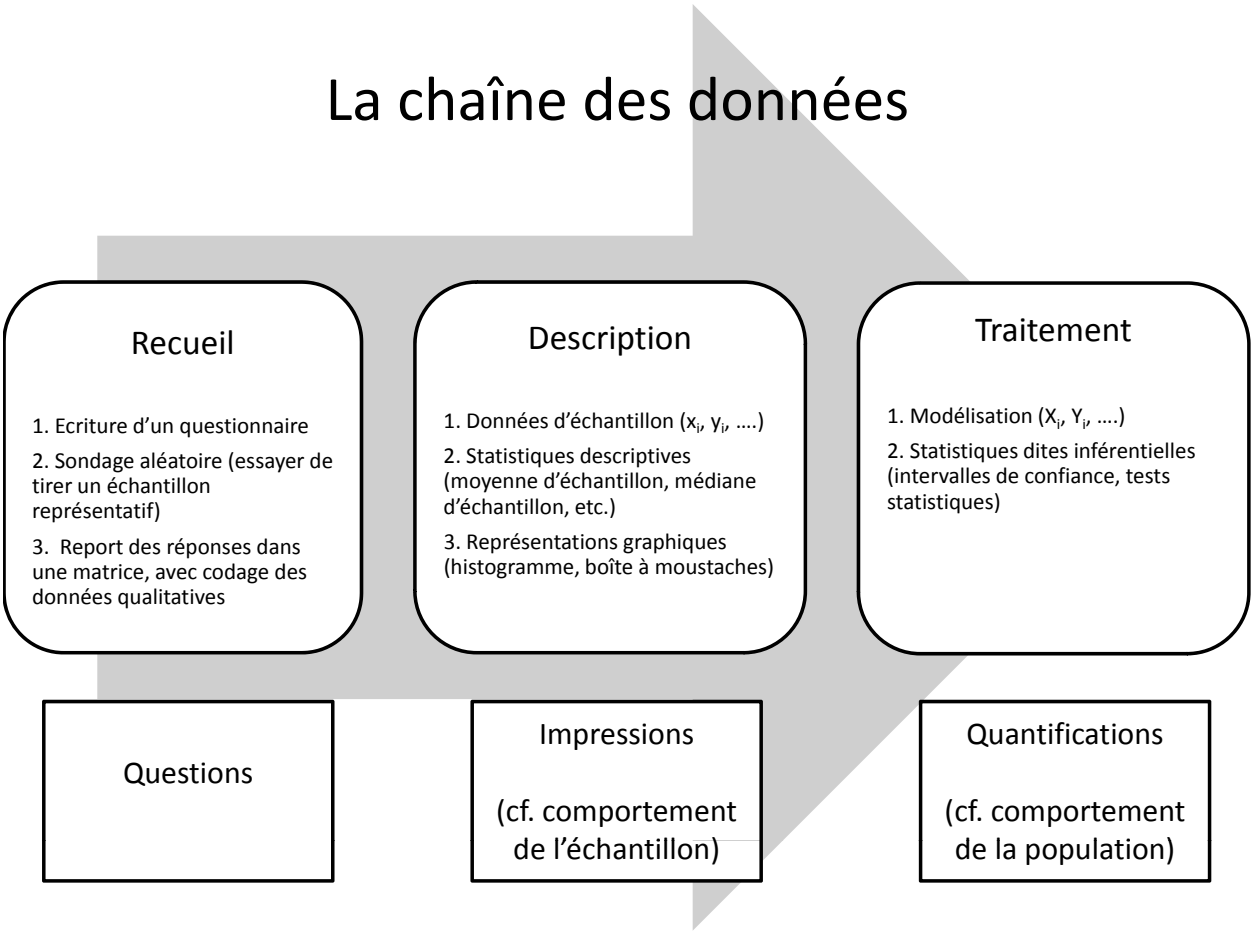
Là encore, par loi des grands nombres, il y a toutes les chances que le « vrai » budget vacances moyen des Français μ_0 soit proche de $\overline{y}_{1\,000} = 1\,641$ euros.

Remarque 1.2 (Intervalles de confiance). Le chapitre suivant aura pour objet de formuler des assertions plus précises que celle-ci, en remplaçant « il y a toutes les chances que » par « avec confiance 95% » et en assortissant l'estimation ponctuelle 1 641 d'une marge d'erreur : on parlera d'intervalles de confiance.

La chaîne de collecte et de traitement des données

On peut résumer l'ensemble de la démarche suivie par l'infographie suivante. Ce cours sera centré sur la troisième et dernière case, concernant au traitement dit inférentiel : quelles généralisations à la population peuvent être effectuées, ou non, à partir des observations effectuées sur un échantillon.

A retenir : Il faut (et comme ce cours le montrera, il suffit de) recourir à un échantillon faute de pouvoir étudier de manière exhaustive une grande population.



3. Biais d'échantillonnage vs. Echantillons « représentatifs »

Les modélisations décrites précédemment reposent sur des tirages uniformément au hasard au sein de la population : c'est en général chose difficile, sauf pour un cas particulier, qui formera notre idéal.

Cas particulier (idéal) : liste exhaustive de la population. On dispose parfois d'une liste exhaustive de la population : c'est le cas si cette dernière correspond à l'ensemble des clients d'une entreprise, auquel cas un fichier clients dresse leur liste. Il est alors facile de tirer uniformément au hasard (via un programme informatique) un petit nombre de clients pour tester une offre sur eux.

C'est également le cas si l'on s'intéresse à un type de commerce particulier : supposons que l'on veuille étudier le prix moyen μ_0 des sandwiches jambon-beurre dans les boulangeries de Paris. Profitons-en pour réviser le vocabulaire : la variable d'intérêt est alors le prix d'un sandwich jambon-beurre, les individus statistiques sont des boulangeries, la population est l'ensemble des boulangeries de Paris, et le paramètre d'intérêt μ_0 est, comme indiqué, le prix moyen μ_0 des sandwiches jambon-beurre dans les boulangeries de Paris. Les Pages Jaunes nous fournissent la liste de toutes ces boulangeries : il suffit d'exporter cette liste, d'y piocher uniformément au hasard (par ordinateur) une cinquantaine de magasins, et de contacter ces derniers (soit par téléphone, soit en se rendant en boutique).

Mais en général, ce n'est pas aussi simple. Il est difficile d'expliquer comment bien procéder, il est en revanche facile de pointer différents biais de constitution possibles dans les échantillons.

Biais de motivation. Lorsque l'on effectue une enquête de satisfaction, cette dernière est un exutoire pour les clients mécontents, qui y participent plus volontiers que les clients contents. La bonne pratique est donc de fixer à l'avance, aléatoirement, une liste de clients devant répondre et de se fixer à cette liste, en relançant autant de fois que nécessaire les non-répondants jusqu'à ce qu'ils répondent.

Il ne faut pas hésiter, pour ce type d'enquête, comme de manière générale, à penser à des incitations à participer — comme des bons d'achat de faible valeur ou, pour les sondages de rue, à une boisson ou un bonbec. Sinon, qui viendra vous parler, dans la rue ? Ceux qui ont le temps et ont envie de parler, alors que ceux qui sont pressés, comme les parents de jeunes enfants, passeraient leur chemin.

Le thème de l'enquête peut aussi influencer le choix de participer ou non au sondage : ne le révélez jamais et demandez plutôt « Voulez-vous répondre à un questionnaire sur des thèmes variés ? »

Biais d'endogamie / Biais lié au medium. Il est difficile, pour vous mais aussi désormais pour les sondeurs professionnels, d'effectuer un échantillonnage en population générale. Côté sondeurs, c'est faute d'un annuaire des téléphones portables — au bon vieux temps existait l'annuaire des lignes fixes, que les sondeurs exploitaient pour choisir des Français au hasard. Un moyen de contournement a été mis en œuvre par l'Etat début 2019 : par génération au hasard de numéros de téléphone (fixe ou portable), voir l'extrait d'article du Parisien. Certes, cela augmentait la probabilité de sélection des détenteurs de plusieurs lignes téléphoniques...

Une pratique courante, mais statistiquement criminelle, consiste à écrire un questionnaire et à le faire tourner sur les réseaux sociaux : ce faisant, on collecte l'opinion de gens qui nous ressemblent. Le sondage quotidien du Figaro n'est pas plus fiable : il renvoie l'opinion de ses lecteurs les plus jeunes, ceux qui consultent son site, il y a donc un double biais d'orientation politique et d'âge. La même question, posée sur le site de Marianne ou de Libération, recevrait sans doute des réponses fort différentes !

Biais temporels / Nécessité d'un plan d'expérience. Le message important dans cette partie, c'est qu'un échantillonnage se prépare : on parle de plan d'expérience. Si vous voulez mener une étude qualitative sur les clients d'un supermarché, il faut interroger des clients en fonction de l'affluence, i.e., en nombre proportionnel aux passages en caisse : peu le jeudi après-midi à 14h, mais beaucoup

Tirage au sort du grand débat : comment l'Etat va-t-il obtenir votre numéro de téléphone ?

Emmanuelle Wargon a indiqué ce jeudi que les Français seraient tirés au sort via leur numéro de téléphone pour participer aux conférences citoyennes.

Le principe est simple : un logiciel va composer de façon aléatoire des numéros de téléphone, commençant par 01, 02, 03, etc., jusqu'à 09. « C'est un outil de génération automatique qui fait ressortir des numéros de téléphone, qu'ils existent ou pas », précise au Parisien l'entourage de la secrétaire d'Etat Emmanuelle Wargon. Cette dernière, en charge du grand débat avec son collègue ministre Sébastien Lecornu, avait avancé cette piste ce jeudi matin sur RTL. Si une personne répond, il lui sera demandé de se présenter et si elle souhaite participer (la charte précise des questions doit encore être totalement arbitrée). [...] Le gouvernement a exclu en revanche de passer par les listes électorales pour tirer au sort les citoyens à appeler. « Je rencontre plein de gens qui me disent qu'ils ne sont inscrits sur ces listes électorales et c'est le cas de quatre ou cinq millions de Français, il n'est pas question de les exclure », explique Pascal Perrineau. Pas question non plus de passer par un annuaire classique, par exemple celui des Pages blanches ou d'un opérateur de téléphone. Cela aurait exclu les Français qui sont inscrits sur liste rouge ou sur la plateforme anti-démarchage Bloctel. Au final, cette génération aléatoire de numéros « est la solution qui permet d'exclure le moins de monde », vante un conseiller gouvernemental.

FIGURE 1.1 – Extrait d'un article du Parisien, daté du 7 février 2019.

le vendredi soir ou le samedi. Il vous faut discuter avec le commanditaire en amont, pour récupérer ces informations et construire votre planning d'échantillonnage en fonction de cela, en recrutant le nombre d'enquêteurs adéquats par jour et tranche horaire.

Là encore, il est interdit de fonctionner à l'opportunité et d'aller enquêter quand cela vous arrange, ou de vous contenter de vos propres observations non préparées — ces dernières semblent vous indiquer par exemple que les salles de cinéma sont toujours pleines, alors même que les exploitants vous diraient que la plupart du temps (mais pas le vendredi soir ou le samedi soir), les salles sont plutôt vides. Par biais d'opportunité, on entend qu'il ne faut pas nécessairement prendre toutes les réponses ou observations que l'on peut obtenir, toutes celles qui se présentent, mais qu'il faut établir au préalable un protocole d'enquête.

Un dernier exemple, pour faire le lien avec le point suivant, consiste en une remarque sur les sondages téléphoniques, du temps de leur apogée dans les années 1990 : qui décrochait alors le téléphone ? « Curieusement », plus souvent les femmes et les retraités que les hommes actifs.

Constitution par quotas / Redressement. Les sondeurs professionnels (et idéalement les administrateurs de la JE) doivent se renseigner au préalable sur la population visée et sa répartition en différentes sous-catégories d'âges, revenus, etc. La constitution de l'échantillon est alors pilotée par l'obtention de répartitions en son sein similaires à celles dans la population, c'est ce que l'on appelle la méthode par quotas.

Au cas où, pour des raisons diverses (notamment des défections dans une enquête au long cours), les quotas ne seraient finalement plus respectés dans l'échantillon, il est alors possible d'affecter des poids différents aux individus de l'échantillon pour sur-représenter ou sous-représenter certains groupes. C'est une cuisine de sondeurs à manier avec précaution...

Nous verrons plus tard, grâce aux tests du χ^2 d'ajustement à une loi, comment s'assurer, rétrospectivement, qu'un échantillon est bien représentatif d'une population, au sein où il est réparti en catégories de manière non significativement différente de la répartition générale de la population en ces catégories.

Autres éléments à considérer. Un dernier mot porte sur la rédaction des questions, dont il faut

qu'elle soit la plus neutre possible et n'induise pas de réponse particulière. Ce n'est pas la même chose de demander « Que pensez-vous de [telle réforme sociétale] ? » et « Pensez-vous que [telle réforme sociétale] soit un sujet prioritaire vu la situation économique actuelle du pays ? »

Contre-exemple gouvernemental

Je vous propose de méditer, en conclusion de ce chapitre, sur l'échec statistique magistral suivant, relaté par

> Benoît Riandey et Isabelle Widmer, L'enseignement des sondages à l'usage du plus grand nombre : quelques réflexions tirées de l'expérience, Statistique et Enseignement, 1(1), 47–63, avril 2010

2.1 Chaluter n'est pas sonder

Au printemps 1994, le premier ministre Edouard Balladur organise une vaste consultation des jeunes âgés de 15 à 25 ans et recueille un million cinq cent trente neuf mille questionnaires. Suite à ce succès quantitatif, en meilleur politique que statisticien, il déclare que « *Un million cinq cent mille réponses volontaires valent mieux qu'une enquête par sondage* ». Au contraire, les spécialistes nient toute valeur scientifique à ce vrac de questionnaires et la Société Française de Statistique (SFdS) demande dans un communiqué à la presse que, pour disposer d'une information scientifiquement fondée, ce questionnaire soit soumis à un échantillon représentatif. Ainsi, la première consultation estimait à 33 % la proportion de jeunes « *pas du tout confiants dans l'avenir* », tandis que quelques mois plus tard SCP Communication estimait cette proportion à seulement 13 %. Le décalage temporel ne rend certes pas directement comparables ces proportions, mais il conforte l'idée que l'inquiétude pour l'avenir aurait été une très forte motivation pour répondre à la consultation organisée par le Premier Ministre. Le succès de cette consultation postale était alors l'expression du mécontentement politique des jeunes, ce qui n'est guère un critère de représentativité.

Exercice : lecture critique de mail #1

Qu'auriez-vous répondu au message suivant, qui m'a été adressé par un de vos prédécesseurs lors d'un stage sur le terrain en fin de L3 ?

Monsieur,

Je vous contacte dans le cadre du stage que j'effectue actuellement en boutique chez XXX. 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 XXX 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 autre 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 ?

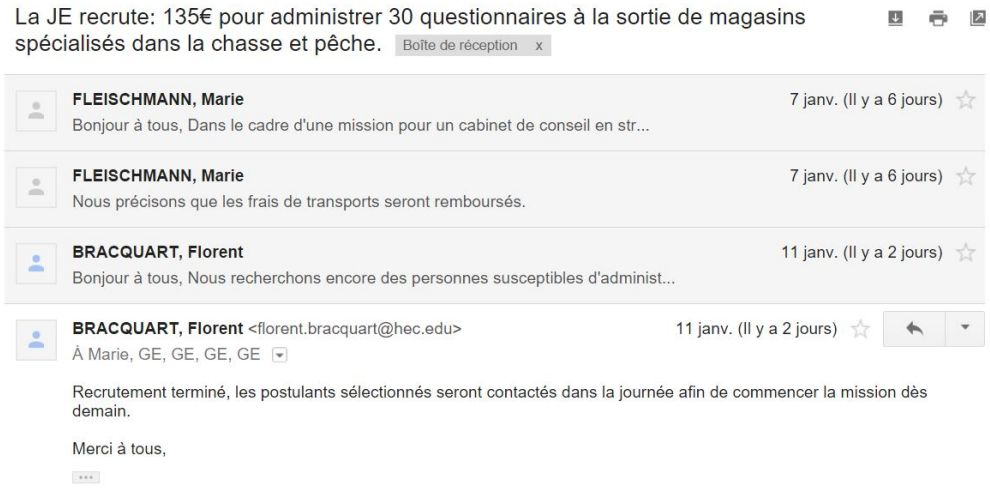
Auriez-vous des remarques à faire par rapport à ce questionnaire ? Des améliorations à apporter ? Et sur quelle marge d'erreur puis-je m'appuyer ?

Merci beaucoup pour votre aide, XXX, ex-élève de L3

Exercice : lecture critique de mail #2

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.



Exercice : modélisation #1

On sélectionne 200 clients au hasard dans notre fichier clients et on leur propose une offre commerciale qu'on voudrait tester avant généralisation éventuelle à tous les clients. Parmi eux, 156 passent commande. Modélisez cette situation : individus statistiques, population visée, paramètre d'intérêt, taille d'échantillon, données disponibles, résumé des données d'échantillon.

Procurez des descriptions contenant un verbe conjugué pour le paramètre d'intérêt et le résumé des données d'échantillon.

Exercice : modélisation #2

Modélisez la situation de la page 27 (exercice TwitterAudit) : individus statistiques, population visée, paramètre d'intérêt, taille d'échantillon, données disponibles, résumé des données d'échantillon.

Exercice : modélisation #3

Modélisez la situation du premier utilisateur de la page 47 (exercice Application podomètre) : individus statistiques, population visée, paramètre d'intérêt, taille d'échantillon, données disponibles, résumé des données d'échantillon.

Comment justifie-t-on ici que l'échantillon est représentatif ?

2

Intervalles de confiance

Dans ce chapitre, nous viserons à résoudre le problème de politique publique suivant.

Fixation du forfait post-stationnement par une ville

En janvier 2018, une réforme du contrôle de stationnement a eu lieu, avec pour effet essentiel de transférer complètement sa mise en œuvre aux communes qui, en contrepartie, peuvent fixer librement le montant dû pour non-paiement du stationnement (appelé FPS, forfait post-stationnement). Une ville veut calculer un montant de FPS incitatif pour encourager les usagers à payer le stationnement. Toutefois, les usagers étant électeurs, elle veut malgré tout laisser ce montant de FPS à un niveau politiquement acceptable et explicable. (Elle ne compte pas le fixer à 1 000 euros par exemple, qui serait un montant extrêmement incitatif !) Pour fonder la décision politique sur des éléments objectifs, elle va ainsi mener une étude statistique.

La ville dispose d'une brigade de dix agents municipaux en charge du contrôle du stationnement et ne veut pas augmenter la taille de cette brigade. La brigade parvient actuellement à un certain taux p_0 (inconnu) de détection des fraudeurs au stationnement. Notons par ailleurs μ_0 le temps moyen de stationnement dans la ville, et supposons-le identique entre fraudeurs et bons payeurs. C'est une sorte de nécessité statistique, puisque nous pouvons facilement estimer le temps de stationnement moyen des bons payeurs (en observant les tickets de parcmètre), mais il est difficile et chronophage de recueillir des observations pour celui des fraudeurs (cela suppose essentiellement d'être à côté d'une voiture stationnant en infraction tout le temps de son stationnement). On peut également penser qu'intuitivement, les fradeurs stationnent pour des temps plutôt plus courts que les usagers payant leur stationnement.

Sachant que l'heure de stationnement coûte 2 euros, le montant du FPS qui rendrait en moyenne indifférents les usagers au fait de payer à l'horodateur ou non serait de $c_0 = 2\mu_0/p_0$. La ville veut donc estimer c_0 et à cet effet, va estimer μ_0 et p_0. A cet effet, la brigade effectue des contrôles à des moments variés d'une semaine typique, en des endroits variés de la ville. Elle contrôle au total 2 159 véhicules, dont 563 ont été en infraction au stationnement. Par ailleurs, le temps de stationnement moyen relevé pour les 1 596 véhicules en règle est de 2h 41 minutes, avec un écart-type de 1h 08 minutes dans ces 1 596 données.

L'idée est d'aller au-delà de l'estimation ponctuelle de $2 \times (2 + 41/60)/(563/2\,159) \approx 20.58$ euros que l'on pourrait immédiater proposer, et de suivre une méthodologie plus sûre.

1. Concept d'intervalle de confiance / Formules pour estimer μ_0

Partons, pour une variable d'intérêt donnée, d'observations $x_1, \ldots, x_n$ issues d'un échantillon représentatif extrait d'une population de moyenne μ_0. Par loi des grands nombres, la moyenne d'échantillon $\overline{x}_n$ est proche de la moyenne de population μ_0, mais nous voulons quantifier cette proximité. Nous allons plutôt proposer des intervalles de valeurs plausibles pour μ_0, au lieu d'indiquer une valeur ponctuelle (une unique valeur).

Trois types de formulations sont possibles à cet effet.

— Les intervalles symétriques (en langage technique : les intervalles bilatères) correspondent à dire qu'avec confiance 95%, la moyenne de population μ_0 se situe dans un intervalle centré en $\overline{x}_n$, i.e., de la forme $[\overline{x}_n \pm \varepsilon_n]$. Le terme ε_n est appelé la marge d'erreur.

— Les estimations basses (en langage technique : les intervalles unilatères minorants) correspondent à dire qu'avec confiance 95%, la moyenne de population μ_0 est supérieure à des corrections à la baisse de $\overline{x}_n$, i.e., se situe dans un intervalle de la forme $[\overline{x}_n - \varepsilon'_n, +\infty[$. Dit autrement, avec confiance 95%, la moyenne de population vaut au moins $\overline{x}_n - \varepsilon'_n$.

— De même, les estimations hautes (en langage technique : les intervalles unilatères majorants) correspondent à dire qu'avec confiance 95%, la moyenne de population μ_0 est inférieure à des corrections à la hausse de $\overline{x}_n$, i.e., se situe dans un intervalle de la forme $]-\infty, \overline{x}_n + \varepsilon'_n]$. Dit autrement, avec confiance 95%, la moyenne de population vaut au plus $\overline{x}_n + \varepsilon'_n$.

Comme les exercices l'illustreront, les intervalles symétriques sont à utiliser dans les cas neutres, tandis que les intervalles unilatères s'imposent dans les cas avec des intentions particulières.

Exemple. La ville devrait-elle estimer de manière neutre le temps de stationnement moyen, ou a-t-elle des intentions particulières ? On pourrait dire qu'elle veut mener une étude la plus neutre possible pour éclairer le choix de ses élus. Mais on pourrait également avancer qu'on préfère fixer le montant du forfait post-stationnement de sorte à être raisonnablement sûr qu'en moyenne, les automobilistes ont intérêt à payer leur stationnement au parcmètre, ce qui conduirait à privilégier une estimation haute sur le montant c_0 du forfait, dont on verra qu'elle requiert une estimation haute sur le temps moyen de stationnement μ_0.

Formules d'intervalles de confiance sur μ_0. Rappelons que nous notons $\overline{x}_n$ et $s_{x,n}$ la moyenne et l'écart-type des données $x_1, \ldots, x_n$. Des raisonnements mathématiques en annexe de ce chapitre, reposant sur l'utilisation conjointe du théorème limite central et du lemme de Slutzky, procurent les trois formules (de réalisations) d'intervalles de confiance suivantes :

— Intervalle symétrique $\quad \left[\overline{x}_n \pm 1.96\, \dfrac{s_{x,n}}{\sqrt{n}}\right]$

— Estimation basse $\quad \overline{x}_n - 1.645\, \dfrac{s_{x,n}}{\sqrt{n}}$

— Estimation haute $\quad \overline{x}_n + 1.645\, \dfrac{s_{x,n}}{\sqrt{n}}$

Exemple. On rappelle que les données pour l'estimation du temps de stationnement moyen μ_0 sont une moyenne d'échantillon $\overline{x}_{1596} = 2\text{h } 41$ minutes, avec un écart-type $s_{x,1596} = 68$ minutes ; il convient de convertir l'écart-type en minutes afin d'obtenir une marge d'erreur en minutes :

$$1.96\, \frac{68}{\sqrt{1\,596}} = 3.33617... \leqslant 4 \text{ minutes.}$$

On arrondit les marges d'erreur à la hausse pour toujours agrandir l'intervalle proposer et ainsi conserver le niveau de confiance 95%. Ici, on arrondit à la minute près, par lisibilité. Conclusion :

> Avec grande confiance, le temps moyen de stationnement μ_0 sur l'ensemble des véhicules se situe entre 2h 37 minutes et 2h 45 minutes.

Remarque 2.1 (Attention à l'interprétation !). Nous ne disons pas du tout que 95% des véhicules stationnent pour une durée comprise entre 2h 37 minutes et 2h 45 minutes ! Nous disons simplement que si un être omniscient pouvait observer μ_0, il y aurait toutes les chances que cette moyenne μ_0 sur un très grand nombre de véhicules soit dans l'intervalle 2h 37 minutes – 2h 45 minutes. Au vu de l'écart-type 1h 08 minutes des observations, on peut plutôt penser que les temps de stationnement typiques s'étalent entre des durées très courtes (quelques minutes) et plusieurs heures. Une autre manière de reformuler ces observations, c'est de vous faire remarquer que la phrase suivante est terriblement incorrecte (puisqu'elle parle de valeurs individuelles) :

> Avec grande confiance, un véhicule stationne entre 2h 37 minutes et 2h 45 minutes dans cette ville.

Au contraire, celle-ci trouve grâce à mes yeux (elle parle bien d'une valeur moyenne) :

> Avec grande confiance, les véhicules stationnent en moyenne entre 2h 37 minutes et 2h 45 minutes dans cette ville.

Les deux mots « en moyenne » (et le passage au pluriel) changent complètement le sens de l'ensemble ! Et c'est pourquoi je serai souvent si pointilleux sur la présence de ces mots, « en moyenne », lorsqu'ils sont nécessaires.

Remarque 2.2 (Confiance vs. probabilité). Nous ne disons pas que μ_0 appartient à l'intervalle 2h 37 minutes – 2h 45 minutes avec grande probabilité, mais avec grande confiance. Ici, il n'y a plus rien d'aléatoire : μ_0 vaut ce qu'il vaut (mais nous ne le connaissons pas) et nous avons confiance que notre procédure d'estimation a bien marché, parce qu'elle avait une probabilité 95% de bien marcher (voir l'annexe). C'est à l'inverse des tirages du Loto : quelques minutes avant le tirage, nous avons une probabilité faible de remporter le gros lot, mais quelques minutes après le tirage, il n'y a plus d'aléa, les choses sont ce qu'elles sont ; si nous ne sommes pas informés du résultat, nous avons malgré tout confiance que nous n'avons pas remporté le gros lot.

Exemple. Voici le calcul de l'estimation haute :

$$\bar{x}_{1596} + 1.645\,\frac{s_{x,1596}}{\sqrt{1\,596}} = 2\text{h } 41 \text{ minutes} + 1.645\,\frac{68}{\sqrt{1\,596}} = 2\text{h } 41 \text{ minutes} + 2.80... \leqslant 2\text{h } 44 \text{ minutes}.$$

Ici encore, on agrandit l'intervalle de confiance au moment des arrondis, et comme il s'agit d'une estimation haute, c'est à la hausse qu'on arrondit. Notre conclusion est :

> Avec grande confiance, le temps moyen de stationnement μ_0 sur l'ensemble des véhicules est inférieur à 2h 44 minutes.

En renonçant à borner inférieurement μ_0, nous donnons une borne supérieure de 2h 44 minutes plus petite que celle de 2h 45 minutes issue de l'intervalle symétrique. C'est fondamentalement lié au facteur 1.645 des intervalles unilatères remplaçant le facteur 1.96 de l'intervalle symétrique. L'annexe explique que cette différence de facteurs correspond à une exclusion soit symétrique (2.5% de chaque côté) soit unilatère (5% d'un côté seulement) de valeurs sous la loi normale standard.

2. Formules d'intervalles de confiance sur p_0

Dans le cas de données binaires, un écart-type empirique se calcule à partir de la proportion d'échantillon $\bar{x}_n$ selon $\sqrt{\bar{x}_n(1-\bar{x}_n)}$. Les formules précédentes peuvent donc se reformuler ainsi, pour l'estimation d'une proportion p_0 d'une variable binaire parmi une population :

— Intervalle symétrique $\left[\bar{x}_n \pm 1.96\,\frac{\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}\right]$

— Estimation basse $\bar{x}_n - 1.645\,\frac{\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}$

— Estimation haute $\bar{x}_n + 1.645\,\frac{\sqrt{\bar{x}_n(1-\bar{x}_n)}}{\sqrt{n}}$

Exemple. Intéressons-nous au taux d'infraction au stationnement p_0, obtenu comme la moyenne des variables binaires (1 pour infraction, 0 pour stationnement en règle) associées aux milliers de véhicules ayant stationné dans un passé proche dans la ville. Ce taux p_0 est évidemment inconnu à la municipalité (seul un être omniscient le connaîtrait), mais qu'à cela ne tienne, la municipalité peut l'estimer à partir des données recueillies : au total 2 159 véhicules ont été contrôlés, dont 563 ont été en infraction au stationnement, soit

$$\bar{x}_{2159} = \frac{563}{2159} = 0.26076\ldots = 26.076\ldots\%\,.$$

Proposer cette estimation ponctuelle est peu précautionneux (en outre, il faudrait l'arrondir) : la ville devrait, dans le cadre d'une démarche neutre, assortir cette estimation d'une marge d'erreur, en l'occurrence :

$$1.96\,\frac{\sqrt{563/2159\,(1-563/2159)}}{2159} = 0.01852\ldots \approx 1.853\%.$$

En réalité, pour bien faire, il faut ici arrondir simultanément $\bar{x}_{2159}$ et la marge d'erreur, et calculer d'une part,

$$\bar{x}_{2159} - 1.96\,\frac{\sqrt{\bar{x}_{2159}(1-\bar{x}_{2159})}}{\sqrt{2159}} = \frac{563}{2159} - 1.96\,\frac{\sqrt{563/2159\,(1-563/2159)}}{2159} = 0.24224\ldots \geqslant 24.2\%\,,$$

et d'autre part,

$$\bar{x}_{2159} + 1.96\,\frac{\sqrt{\bar{x}_{2159}(1-\bar{x}_{2159})}}{\sqrt{2159}} = \frac{563}{2159} + 1.96\,\frac{\sqrt{563/2159\,(1-563/2159)}}{2159} = 0.27928\ldots \leqslant 28.0\%\,.$$

Après quoi, on peut réécrire l'intervalle obtenu $[24.2\%,\, 28.0\%] = [26.1\% \pm 1.9\%]$ sous une forme symétrique.

Notre conclusion est :

> Avec grande confiance, le taux d'infraction p_0 sur l'ensemble des véhicules vaut 26.1% avec une marge d'erreur de 1.9% (ou : est compris entre 24.2% et 28.0%).

On peut même à vrai dire pousser l'arrondi jusque 26% ± 2%.

Exemple. Nous verrons plus loin que pour l'estimation haute sur $c_0 = 2\mu_0/p_0$ que nous appellerons de nos vœux, il conviendra d'avoir une estimation basse sur p_0. C'est celle-ci que nous calculons maintenant : avec grande confiance, le taux d'infraction p_0 est supérieur à

$$\bar{x}_{2159} - 1.645\,\frac{\sqrt{\bar{x}_{2159}(1-\bar{x}_{2159})}}{\sqrt{2159}} = \frac{563}{2159} - 1.645\,\frac{\sqrt{563/2159\,(1-563/2159)}}{2159} = 0.2452250\ldots \geqslant 24.5\%\,.$$

Remarque 2.3 (Application aux sondages d'opinion). Le cas d'application le plus typique des formules d'intervalles de confiance sur p_0 est formé par les sondages d'opinion sur des questions fermées (pour / contre) ou des votes dans le cadre d'un duel (comme le second tour des élections présidentielles). Les sondeurs fournissent à leurs lecteurs journalistes une grille permettant de déduire la marge d'erreur $1.96\sqrt{\bar{x}_n\,(1-\bar{x}_n)/n}$ à partir des valeurs approchées de $\bar{x}_n$ (colonnes) et n (lignes) ; voir la figure 2.1. Le cas le plus typique est donné par la ligne $n = 1\,000$, qui conduit à des marges d'erreur d'au plus 3.1%.

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** **@ifopopinion** **Ifop Opinion**

Etude réalisée par l'Ifop pour :	L···
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

FIGURE 2.1 – Un exemple de grille de lecture de résultats de sondage fourni par un institut : méthodologie et tableau résumant les marges d'erreur. (Le sujet importe peu mais si vous insistez, je vous le révélerai en cours...)

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	4,3	4,6	4,7
500	1,9	2,7	3,6	4,1	4,4	4,5
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%.

3. Intermède : Réflexions sur un miracle et sur les arrondis

Un miracle statistique. La précision de l'estimation, i.e., la marge d'erreur des intervalles symétriques, ne dépend pas de la taille de la population, mais uniquement de la taille d'échantillon (et certes de la dispersion du phénomène étudié, correspondant à un écart-type de population σ_0 et estimé par l'écart-type d'échantillon $s_{x,n}$).

Or, certains citoyens ou certains cadres pensent, à tort, que pour toute étude statistique, il convient d'interroger une certaine fraction de la population (comme 5% ou 10% de la population). Cela rendrait impossible la conduite de sondages sur les plus de 45 millions de Français adultes ! Mais non, ce qui est déterminant, c'est la taille n de l'échantillon, pas la taille N de la population... et bien entendu, le fait de disposer d'un échantillon représentatif, par exemple, tiré uniformément au hasard dans la population.

Ce miracle statistique est permis (comme vous l'avez compris en annexe) par ce miracle de la théorie des probabilités qu'est le théorème limite central. Que serait notre vie sans le théorème limite central ?

Un commentaire utile pour la suite est que puisque la précision varie en fonction de la taille n de l'échantillon comme $1/\sqrt{n}$, il faut par exemple multiplier la taille d'échantillon par 4 pour diviser la marge d'erreur par 2.

Attention aux arrondis. J'aime citer l'aphorisme ironique suivant (et je vous encourage à l'apprendre, et à le citer vous aussi, en réaction à des présentations chiffrées mal menées et indignes de votre rang) :

> Dans toute statistique, l'inexactitude du nombre est compensée par la précision des décimales.
> Alfred Sauvy (économiste, démographe et sociologue français, 1898–1990) ou Georges Elgozy (économiste et statisticien français, 1909–89), les sources hésitent

Ainsi, pensez bien à arrondir vos résultats d'intervalles de confiance, pour obtenir à la fois des nombres lisibles et ne transmettant pas une fausse impression de précision. Prenez garde toutefois à effectuer ces arrondis en augmentant toujours la taille de l'intervalle : ainsi, arrondissez les estimations basses à la baisse, les estimations hautes à la hausse, et les marges d'erreur à la hausse.

4. Intervalles de confiance simultanés (correction de Bonferroni)

Règle à retenir : L'exploitation simultanée des résultats de deux intervalles de confiance 95% (indépendants ou non) ne vaut qu'avec confiance 90%. Autrement dit, on doit sommer les risques d'erreur et noter que le risque d'erreur final est de $5\% + 5\% = 10\%$.

Cette règle est prouvée en appendice. Nous allons surtout nous concentrer sur son application, qui illustrera en particulier ce que l'on entend par « exploitation simultanée » de deux intervalles de confiance.

Exemple. Dans le cas d'une approche neutre, nous avons calculé précédemment les intervalles de confiance symétriques suivants : avec confiance 95%, le taux d'infraction p_0 se situe entre 24.2% et 28.0%, tandis qu'avec confiance 95%, le temps moyen de stationnement se situe entre 2h 37 minutes (soit $2+37/60$ heures) et 2h 45 minutes (soit 2.75 heures). Que peut-on dire du montant de FPS visé, à savoir $c_0 = 2\mu_0/p_0$? On rappelle que les intervalles de confiance indiquent les valeurs plausibles pour les paramètres μ_0 et p_0 ; ici, les valeurs plausibles pour c_0 sont donc :

$$\left\{ \frac{2m}{q} : \ q \in [24.2\%, 28.0\%] \text{ et } m \in [2 + 37/60, 2.75] \right\}$$
$$= \left[2\,\frac{2 + 37/60}{0.28},\ 2\,\frac{2.75}{0.242} \right] = [18.69..., 22.72...] \subset [18, 23] ,$$

où l'on a croisé les bornes d'intervalles de confiance ici, puisque l'objectif $2\mu_0/p_0$ est croissant en μ_0 mais décroissant en p_0. Nous avons arrondi à l'euro près parce que les contraventions sont en euros entiers. Cet intervalle de confiance $[18, 23]$ sur c_0 vaut avec confiance seulement 90%, car il résulte de l'exploitation simultanée de deux intervalles de confiance 95%.

Quelle valeur la ville doit-elle finalement fixer, au vu de cet intervalle de confiance, en gardant en tête que c_0 doit être un montant légèrement supérieur à la valeur d'équilibre exacte rendant les automobilistes indifférents au fait de payer leur stationnement ? Bien entendu, c'est la borne supérieure 23 euros que nous retiendrons (ou certes une valeur politiquement arrondie à 25 euros), ce qui aurait justifié une approche moins neutre : unilatère.

Exemple. On peut également combiner les intervalles unilatères pour p_0, à savoir $\geqslant 24.5\%$, et pour μ_0, à savoir $\leqslant 2 + 44/60$, et obtenir l'intervalle de confiance suivant pour c_0 :

$$\left\{ \frac{2m}{q} : \ q \in [24.5\%, 100\%] \text{ et } m \in [0, 2 + 44/60] \right\} = \left[0,\ 2\,\frac{2 + 44/60}{0.245} \right] = [0, 22.312...] \subset [0, 23] .$$

L'estimation haute pour c_0 avant arrondi entier est, comme attendu, plus basse que celle obtenue par intervalles bilatères, même si ici, du fait des arrondis à l'entier, on obtient le même résultat final.

Remarque 2.4 (Attention aux formes des intervalles que l'on combine)**.** Il n'est en revanche pas possible de combiner les intervalles de confiance $\leqslant 27.7\%$ sur p_0 et $\leqslant 2+44/60$ sur μ_0 pour en déduire un intervalle de confiance sur $c_0 = 2\mu_0/p_0$.

En revanche, on peut combiner les intervalles de confiance $\leqslant 27.7\%$ sur p_0 et $\geqslant 2+38/60$ sur μ_0 pour en déduire une estimation basse de $c_0 = 2\mu_0/p_0$, en l'occurrence, $2\times(2+38/60)/0.277 = 19.013... \geqslant 19$ euros. Là, on obtient une borne meilleure, même après arrondi, que dans le cas bilatère, valant elle aussi avec confiance 90%.

Retenez deux choses des applications précédentes :

1. Il convient de se demander à l'avance quelle genre de conclusion nous intéressera sur le paramètre final (estimation haute, estimation basse, ou les deux).
2. En appliquant les règles usuelles sur les inégalités (selon qu'on applique une fonction croissante ou décroissante), on déduit du vœu précédent la forme des intervalles de confiance à calculer sur les paramètres intermédiaires.

5. Planification de sondages

Objectif : savoir combien sonder d'individus statistiques. Jusqu'à présent, nous calculions des intervalles de confiance à partir de données déjà disponibles. Mais en pratique, vous constaterez, au moment de collecter ces données, qu'une de vos interrogations essentielles sera... de savoir combien en collecter, i.e., combien d'individus statistiques sonder. La réponse n'est pas dans les livres, mais procède de deux éléments : (i) un vœu de précision clairement exprimé, par exemple, par votre manager ; et (ii) un coup de sonde formé par une cinquantaine d'individus.

Nous allons exploiter ici une remarque effectuée dans l'intermède ci-dessus :

Un commentaire utile pour la suite est que puisque la précision varie en fonction de la taille n de l'échantillon comme $1/\sqrt{n}$, il faut par exemple multiplier la taille d'échantillon par 4 pour diviser la marge d'erreur par 2.

Elément 1 : un vœu de précision. Votre manager (ou l'énoncé, dans le cas d'un examen de statistique) annoncera clairement qu'on veut estimer telle moyenne de population μ_0 avec une précision (marge d'erreur) $\pm\varepsilon$ près.

Elément 2 : un coup de sonde. Vous disposez déjà de m données, ou allez recueillir de ce pas m données, avec $m \geqslant 50$.

Méthodologie. Avec l'échantillon final de taille $n \geqslant m$, la précision sera $1.96\, s_{x,n}/\sqrt{n}$, et on veut qu'elle soit plus petite que ε. La difficulté est qu'à la fois n et $s_{x,n}$ sont inconnus. Mais par loi des grands nombres (et c'est là que joue le fait que $n \geqslant m \geqslant 50$), on a $s_{x,m} \approx \sigma_0 \approx s_{x,n}$, où σ_0 désigne l'écart-type de la variable étudiée dans la population. Ainsi, la précision finale, peut être approchée puis contrôlée ainsi :

$$1.96\,\frac{s_{x,n}}{\sqrt{n}} \approx 1.96\,\frac{s_{x,m}}{\sqrt{n}} = \underbrace{1.96\,\frac{s_{x,m}}{\sqrt{m}}}_{\text{marge actuelle}}\ \sqrt{\frac{n}{m}} \overset{\text{but}}{\leqslant} \varepsilon\,,$$

et ce, en augmentant multiplicativement la taille d'échantillon m par un facteur (objectif / marge actuelle)^2 :

$$n \geqslant m\left(\frac{\text{marge actuelle}}{\text{objectif}}\right)^2 = m\left(\frac{1.96\, s_{x,m}/\sqrt{m}}{\varepsilon}\right)^2 = \left(\frac{1.96\, s_{x,m}}{\varepsilon}\right)^2.$$

En général, la marge actuelle issue du coup de sonde aura déjà été calculée, comme l'illustre l'exemple suivant. La formule ci-dessus dit exactement que l'on veut améliorer la marge actuelle d'un certain facteur multiplicatif γ (correspondant au rapport marge actuelle sur objectif ε) et qu'à cet effet, il convient de multiplier la taille actuelle d'échantillon m par un facteur γ^2.

Exemple. Supposons que le maire ait désiré communiquer sur une étude statistique très fiable et qu'il ait souhaité que l'on puisse estimer le temps moyen de stationnement à ± 1 minute près. Combien de tickets de stationnement faudrait-il observer à cet effet ? La marge d'estimation actuelle, avec $m = 1\,596$, est de ± 3.34 minutes (en revenant à la valeur intermédiaire avant arrondi à l'entier telle que calculée page 16). On veut l'améliorer d'un facteur $\gamma = 3.34$, il conviendrait donc de relever

$$n = m\gamma^2 = 1\,596\,(3.34)^2 \approx 17\,805$$

tickets de stationnement (y compris les $m = 1\,596$ déjà disponibles). C'est beaucoup, cela prendrait un temps certain, de sorte qu'éventuellement, le directeur des services pourrait négocier un compromis avec le maire pour un objectif de précision plutôt à ± 2 minutes près, qui nécessiterait seulement 4 452 observations.

Résumé de la méthodologie. Lorsque l'on a aucune information sur les ordres de grandeur de la variable dont on veut estimer la moyenne de population μ_0, on effectue un coup de sonde avec un petit échantillon de taille $m \geqslant 50$, ce qui permet d'obtenir un ordre de grandeur de l'écart-type de la varible, à partir duquel on peut calculer une taille d'échantillon n nécessaire pour obtenir la précision $\pm\varepsilon$ désirée. Si n semble trop grand (trop coûteux en temps ou en argent), il convient de renégocier l'objectif ε avec le décideur.

Remarque 2.5 (Cas des proportions). Tout ce qui a été décrit ci-dessus vaut également pour les proportions, en remplaçant $s_{x,m}$ par $\sqrt{\overline{x}_m\,(1-\overline{x}_m)}$.

Appendice facultatif : Preuves mathématiques

Formules pour μ_0. Nous prouvons ci-dessous que si $X_1, X_2, \dots$ sont des variables aléatoires indépendantes et identiquement distribuées, d'espérance commune μ_0 et d'écart-type commun σ_0, alors

$$\mathbb{P}\left(\mu_0 \in \left[\overline{X}_n \pm 1.96\,\frac{\hat{\sigma}_n}{\sqrt{n}}\right]\right) \longrightarrow 0.95 \qquad \text{et} \qquad \mathbb{P}\left(\mu_0 \geqslant \overline{X}_n - 1.645\,\frac{\hat{\sigma}_n}{\sqrt{n}}\right) \longrightarrow 0.95$$

(et une formule similaire pour l'estimation haute). Les formules

$$\left[\overline{x}_n \pm 1.96\,\frac{s_{x,n}}{\sqrt{n}}\right] \qquad \text{et} \qquad \geqslant \overline{x}_n - 1.645\,\frac{s_{x,n}}{\sqrt{n}}$$

correspondent à la réalisation sur les données de ces intervalles aléatoires.

En effet, le théorème limite central assure la convergence en loi

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

Vous avez prouvé en classes préparatoires (au moins pour certains d'entre vous) la convergence en probabilité

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \xrightarrow{\mathbb{P}} \sigma_0^2\,, \qquad \text{d'où (par continuité)} \qquad \hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2} \xrightarrow{\mathbb{P}} \sigma_0\,.$$

Par lemme de Slutzky, il vient alors la convergence en loi

$$\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_0\right) \rightharpoonup \mathcal{N}(0,1)\,. \tag{2.1}$$

Par conséquent, une des définitions de la convergence en loi assure que

$$\mathbb{P}\left(\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_0\right) \in [-1.96,\, 1.96]\right) \longrightarrow \mathbb{P}(Z \in [-1.96,\, 1.96]) = 0.95\,, \qquad \text{où} \quad Z \sim \mathcal{N}(0,1)\,.$$

En effet, 1.96 a précisément été choisi (après calculs) pour que $\mathbb{P}(Z \in [-1.96, 1.96]) = 0.95$. Dit autrement, $P(Z \geqslant 1.96) = \mathbb{P}(Z \leqslant -1.96) = 0.025$: on exclut symétriquement 2.5% de la masse de probabilité à gauche et à droite. Le raisonnement est conclu pour l'intervalle symétrique en notant l'équivalence

$$\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_0\right) \in [-1.96,\, 1.96] \quad \Longleftrightarrow \quad \mu_0 \in \left[\overline{X}_n \pm 1.96\,\frac{\hat{\sigma}_n}{\sqrt{n}}\right].$$

Quant à l'intervalle unilatère, il suffit d'adapter le raisonnement ci-dessus. Premièrement, 1.645 a précisément été choisi (après calculs) pour que $\mathbb{P}(Z \leqslant 1.645) = 0.95$. La convergence en loi (2.1) entraîne donc, par définition,

$$\mathbb{P}\left(\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_0\right) \leqslant 1.645\right) \longrightarrow \mathbb{P}(Z \leqslant 1.645) = 0.95\,,$$

et le raisonnement est conclu par l'équivalence

$$\frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_0\right) \leqslant 1.645 \quad \Longleftrightarrow \quad \mu_0 \geqslant \overline{X}_n - 1.645\,\frac{\hat{\sigma}_n}{\sqrt{n}}\,.$$

Correction de Bonferroni. Il s'agit de prouver que l'exploitation simultanée des résultats de deux intervalles de confiance à 95% ne vaut qu'avec confiance 90%.

Ceci se réécrit ainsi, en en revenant aux probabilités et à des variables aléatoires : étant donné deux intervalles de confiance aléatoires $\widehat{I}$ et $\widehat{J}$ pour des paramètres μ_0 et μ_0', vérifiant

$$\mathbb{P}(\mu_0 \in \widehat{I}) = 0.95 \quad \text{et} \quad \mathbb{P}(\mu_0' \in \widehat{J}) = 0.95\,,$$

en visant l'estimation d'un objectif combiné $g(\mu_0, \mu_0')$, et en notant

$$g(\widehat{I}, \widehat{J}) = \{g(x, y) : x \in \widehat{I} \text{ et } y \in \widehat{J}\}$$

le résultat de l'exploitation simultanée, montrer que

$$\mathbb{P}\Big(g(\mu_0, \mu_0') \in g(\widehat{I}, \widehat{J})\Big) \geqslant 0.90\,.$$

A cet effet, on note tout d'abord qu'on a l'implication suivante, par définitions :

$$\text{si } \mu_0 \in \widehat{I} \text{ et } \mu_0' \in \widehat{J} \qquad \text{alors} \qquad g(\mu_0, \mu_0') \in g(\widehat{I}, \widehat{J})\,,$$

ce qui entraîne l'inégalité

$$\mathbb{P}\Big(g(\mu_0, \mu_0') \in g(\widehat{I}, \widehat{J})\Big) \geqslant \mathbb{P}(\mu_0 \in \widehat{I} \text{ et } \mu_0' \in \widehat{J})\,.$$

Il serait tentant à ce stade d'utiliser une formule $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$, mais cela n'est sûr de valoir que lorsque A et B sont indépendants. Cela correspondrait ici au fait que $\widehat{I}$ et $\widehat{J}$ auraient été obtenus à partir de deux échantillons eux-mêmes créés indépendamment, ce qui n'est pas le cas en général (et en tout cas, pas le cas du tout dans l'exemple fil rouge de ce chapitre). On utilise plutôt le raisonnement général suivant, qui ne nécessite pas d'hypothèse d'indépendance : en notant $\overline{A}$ et $\overline{B}$ les complémentaires de A et B,

$$\mathbb{P}(A \cap B) = 1 - \underbrace{\mathbb{P}(\overline{A} \cup \overline{B})}_{\leqslant \mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B})} \geqslant 1 - \Big(\mathbb{P}(\overline{A}) + \mathbb{P}(\overline{B})\Big)\,.$$

Ainsi,

$$\mathbb{P}(\mu_0 \in \widehat{I} \text{ et } \mu_0' \in \widehat{J}) \geqslant 1 - \Big(\underbrace{\mathbb{P}(\mu_0 \notin \widehat{I})}_{=0.05} + \underbrace{\mathbb{P}(\mu_0' \notin \widehat{J})}_{=0.05}\Big) = 0.90\,,$$

ce qui conclut la démonstration.

Remarque 2.6 (Indépendance). Dans les cas où $\widehat{I}$ et $\widehat{J}$ seraient malgré tout indépendants, on aurait une minoration plus précise par

$$\mathbb{P}(\mu_0 \in \widehat{I} \text{ et } \mu_0' \in \widehat{J}) = \mathbb{P}(\mu_0 \in \widehat{I})\,\mathbb{P}(\mu_0' \in \widehat{J}) = 0.95 \times 0.95 = 90.25\%\,,$$

soit un gain totalement négligeable de 0.25% dans le niveau de confiance. Oubliez à jamais cette remarque !

Publicité pour un escalier — 8 points

La publicité suivante veut illustrer qu'un escalier est un élément important dans un logement ; assurément, le nombre 89 019 est issu d'un intervalle de confiance, la petite étoile après ce nombre fait d'ailleurs référence à une enquête menée par BVA Group en juillet 2017. J'invente les données suivantes (faute d'avoir pu mettre la main sur l'étude).

Parmi les sondés, nous ne nous intéressons bien évidemment qu'à ceux possédant un escalier, ils sont 534. On leur demande à chacun combien de fois ils le parcourent par jour et combien de temps ils pensent conserver leur escalier avant travaux / changements. Les réponses sont :

— une moyenne de 12.60 montées et descentes par jour (avec un écart-type de 2.41 dans ces données) ;
— une moyenne de 19.34 années sans travaux / changements (avec un écart-type de 4.35 dans ces données).

On considérera qu'une année dure 365.25 jours.

☐ Quelle forme doit-on retenir pour l'estimation par intervalle, en vue de rédiger une belle phrase s'insérant naturellement dans une publicité ? Entourez votre solution :

Symétrique (bilatère) / estimation basse (unilatère minorante) / estimation haute (unilatère majorante)

Procurez une belle phrase, complète et agréable à lire, indiquant à la fois

☐ le premier paramètre d'intérêt considéré et

☐☐ l'estimation par intervalle correspondante (en fournissant les détails du calcul de cette dernière) :

☐☐ Procurez simplement le résultat numérique pour le second paramètre (avec les détails de calcul) :

☐☐ Concluez en complétant la phrase à trous suivante, en commençant par une précision technique qui ne serait pas écrite dans la publicité et qui quantifierait le niveau de confiance :

$\underbrace{\qquad\qquad\qquad\qquad}_{\text{précision technique}}$ et en moyenne, vous allez monter et descendre fois votre escalier.

TwitterAudit — 10 points

Le site TwitterAudit (see `https://www.twitteraudit.com`) définit ainsi sa méthodologie :

> "Each audit takes a sample of up to 5,000 [...] Twitter followers for a user and calculates a score for each follower. This score is based on number of tweets, date of the last tweet, and ratio of followers to friends. We use these scores to determine whether any given user is real or fake. Of course, this scoring method is not perfect but it is a good way to tell if someone with lots of followers is likely to have increased their follower count by inorganic, fraudulent, or dishonest means."

Dans ce qui suit, nous supposerons que TwitterAudit étudie toujours les comptes d'exactement 5,000 followers, et que leur classification en compte réel / compte fallacieux est exempte d'erreur (alors que bien entendu, cette classification ne l'est jamais). Commençons par étudier le compte de Donald Trump :

Capture d'écran de `https://www.twitteraudit.com/realdonaldtrump` (page web consultée le 27 mars 2018 vers 17h)

Analysons d'abord l'image

Nous allons comprendre que cette image est trompeuse.

☐ Définir la population étudiée ici (en particulier, quantifier le nombre d'individus dans la population).

☐ Expliquer pourquoi il semble plus vraisemblable que le score affiché de 73% soit en fait 73.4%.
D'après vous, est-ce un score calculé sur la population ou sur un échantillon de 5,000 comptes ?

Calcul d'un intervalle de confiance

Nous cherchons maintenant à procurer de meilleurs éléments chiffrés dans l'image.

☐ Comment obtient-on le nombre total de vrais followers à partir du paramètre d'intérêt statistique ? Préciser bien entendu ce dernier au cours de votre réponse.

☐ Quelle forme d'intervalle de confiance sur ce paramètre vaut-il mieux retenir ? Justifier soigneusement.

☐ ☐ Effectuer le calcul d'un intervalle de confiance correspondant aux questions précédentes.
(Fournir les détails du calcul, sans quoi aucun point ne sera accordé.)

Meilleur résumé visuel

Nous allons maintenant expliquer à TwitterAudit comment il aurait fallu mieux présenter les résultats.

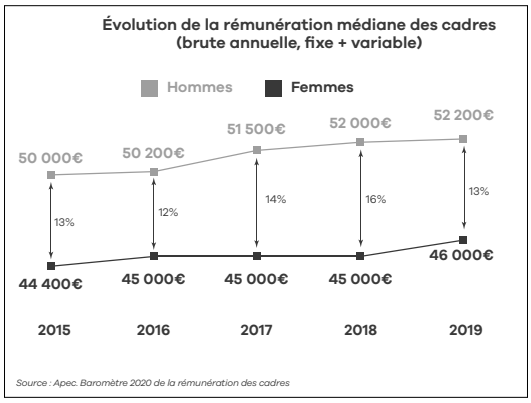
☐ ☐ Le nombre 36,378,957 est en réalité imprécis : expliquer pourquoi.
Proposer un nombre ou un intervalle de nombres de substitution.

Bien entendu, les mêmes critiques et une solution similaire peuvent être considérées pour 13,183,655. Dès lors, suggérer finalement une «boîte» de résumé des données et de leur traitement qui soit à la fois plus précise et plus honnête que la «boîte» TwitterAudit ci-dessus, tout en restant lisible par le grand public.

☐ ☐ A cet effet, remplir le cadre ci-dessous avec des nombres bien choisis.

Ecarts salariaux femmes–hommes chez les cadres

Les données ci-dessous ont été inventées par mes soins et elles grossissent le trait par rapport à la réalité, résumée par une figure issue d'une enquête de l'APEC (association pour l'emploi des cadres), portant sur 16 000 cadres :



Supposons qu'une étude à petite échelle ait été effectuée parmi des cadres et ait fourni les données d'échantillon résumées ci-dessous, pourtant sur des salaires mensuels nets rapportés à un temps plein :

	Sexe	N	Moyenne	Ecart-type	`Standard error'
Salaire mensuel net	Hommes	179	3431,46	3895,437	291,159
	Femmes	147	2434,90	1282,947	105,816

1. En raisonnant par intervalles de confiance, peut-on démontrer à partir de ces données (supposées représentatives) l'existence d'un écart significatif entre les salaires moyens des cadres femmes et hommes ?
2. Est-ce qu'une différence significative entre ces moyennes prouve par elle-même l'existence d'une discrimination salariale ?
3. Combien de sondés faudrait-il considérer dans chaque catégorie pour estimer les salaires mensuels moyens à une erreur de ± 100 euros près ?

Création d'une compagnie d'assurance

Une jeune diplômée d'HEC veut lancer une compagnie d'assurance pensée notamment pour les cadres débordés : 100% en ligne et pouvant être jointe même (et surtout) le dimanche. Avant de creuser davantage son idée, de monter des dossiers pour être incubée, ou de lever des fonds, elle voudrait vérifier rapidement que les clients potentiels ne subissent pas une accidentologie supérieure à la moyenne. En effet, elle voudrait que cette assurance coûte le même prix en moyenne que les autres assurances ; i.e., elle ne joue pas la carte du prix pour se démarquer, mais celle des services.

Elle recourt donc à un institut de sondages (ou à la JE d'une école de commerce, selon son budget) pour effectuer l'enquête suivante : sonder 4 000 personnes majeures vivant en France, représentatives de l'ensemble de la population, leur demander si elles sont assurées pour des risques voiture et si, le cas échéant, elles seraient intéressées par une telle offre d'assurance en ligne. Parmi les 4 000 sondés, il reste 1 472 sondés pour qui les deux réponses sont positives. Les enquêteurs leur demandent alors en outre s'ils ont eu un accident responsable au cours de l'année écoulée, et le cas échéant, le montant moyen des frais encourus (pour eux et les autres véhicules impliqués). En l'occurrence, 256 accidents responsables sont reportés, avec un montant moyen de dommages de 1 865 euros (et un écart-type de 524 euros parmi les montants encourus déclarés).

1. Justifier que le montant moyen des dommages responsables serait de $1\,865 \pm 65$ euros pour les clients potentiels.

2. Combien de clients potentiels ayant été responsables d'un accident faudrait-il interroger de plus pour parvenir à une marge d'erreur $\pm$ 15 euros sur ce montant moyen ?

3. En conséquence, combien de clients potentiels (i.e., assurés et intéressés, mais n'ayant pas nécessairement été responsables d'un accident) faut-il interroger de plus à cet effet ?

Enquêtes sur les allocations sociales non réclamées — 15 points

Supposons que la Cour des comptes veuille étudier l'efficacité budgétaire du système français d'allocations sociales, et qu'elle désire, en particulier, se pencher sur la question des allocations sociales non versées... faute de demande (soit que les bénéficiaires potentiels n'aient pas conscience qu'ils seraient éligibles, soit qu'ils aient renoncé aux démarches parce qu'elles leur semblaient trop complexes).

Il convient de mener une étude par catégorie d'allocation ; prenons l'exemple de la prime d'activité, versée sous conditions aux salariés aux revenus les plus faibles. La Cour, disposant de moyens et de contacts importants, monte une enquête avec l'aide de l'INSEE et de la CAF, où plusieurs milliers de salariés (sélectionnés uniformément au hasard parmi l'ensemble des salariés) sont interrogés et où leur situation est individuellement revue par un agent de la CAF.

Taux de non-recours

Parmi ces milliers de sondés, 861 sont éligibles à la prime d'activité, dont 550 qui l'ont effectivement sollicitée et 311 qui ne la perçoivent pas. Il s'agit de tirer de ces observations un intervalle de confiance sur le taux global de non-recours p_0 à la prime d'activité (sur l'ensemble des salariés éligibles) ; ce taux étant défini comme la part des salariés éligibles à la prime qui ne la perçoivent pas.

☐ Quelle forme d'intervalle convient-il de retenir ici, et pourquoi ?

— Estimation basse
— Estimation haute parce que :
— Symétrique

☐ Ecrivez le calcul à taper sur la calculatrice pour obtenir un intervalle de confiance sur p_0 et copiez-en le résultat sans l'avoir arrondi.

☐ Effectuez un arrondi vers des pourcentages entiers.

☐ Présentez le résultat de la question précédente au sein d'une belle phrase (comme celles vues en cours).

Montants non versés : moyenne

Pour les 311 salariés ne percevant pas la prime, l'agent de la CAF calcule le montant qu'ils auraient pu percevoir chaque trimestre ; il obtient une moyenne de 134 euros, avec un écart-type de 37 euros dans les données.

☐☐☐ Donnez l'intervalle de confiance que l'on peut tirer de ces faits : au sein d'une belle phrase précisant le paramètre d'intérêt μ_0, en présentant les détails du calcul, et en arrondissant les nombres à l'entier.

Montants non versés : total

La CAF verse actuellement la prime d'activité à $N = 4.42$ millions de salariés. Quel est le montant total M des primes d'activité non versées, sur l'ensemble des salariés et par trimestre ?

☐ Exprimer M en fonction des paramètres N, p_0 et μ_0 introduits à la page précédente. On pourra commencer par se demander comment exprimer le nombre total de salariés éligibles en fonction de N et p_0.

☐ Ecrire les calculs qui mèneraient à un intervalle de confiance sur M, sans les effectuer.

☐ Quel serait le niveau de confiance de l'intervalle de la question précédente ?

Planification

Supposons qu'on ait voulu connaître le montant moyen des primes d'activité non sollicitées à ± 2 euros près.

☐☐☐ Combien de salariés éligibles mais n'ayant pas sollicité la prime d'activité aurait-il fallu interroger ?

Formes alternatives

Imaginons un parti politique d'opposition qui voudrait dénoncer l'inefficacité gouvernementale et montrer que de gros montants d'allocations sont non réclamés et donc non versés. On supposera que ce parti est d'une grande méticulosité statistique et ne veut énoncer que des assertions honnêtes.

☐☐ Quelle forme d'intervalle de confiance sur M devrait-il choisir, et partant, quelles formes d'intervalles devrait-il retenir pour μ_0 et p_0 ? Justifiez votre réponse.

3

Tests d'hypothèses : comparaison à une référence

Ce chapitre vous présente une vraie démarche scientifique, qui formule d'abord des hypothèses H_0 et H_1 puis effectue une expérience statistique pour collecter des données, qui soit permettront de rejeter H_0 et passer à H_1 (progrès négatif), soit obligeront à camper sur H_0 (absence de progrès).

1. Introduction : choix des hypothèses

Deux hypothèses, aux rôles tout à fait dissymétriques, sont en jeu.

— L'hypothèse H_0 constitue un point de départ naturel, auquel on est attaché et dont on ne veut se départir que s'il le faut, que si les données contredisent gravement (et pas juste légèrement) cette hypothèse H_0.

— L'hypothèse H_1 est appelée hypothèse alternative, et n'a d'existence que par défaut : c'est la conclusion à laquelle on aboutit lorsque l'on rejette H_0.

Remarque 3.1 (Le doute profite à H_0.). L'analogie la plus éclairante est formée par les procès pénaux, où le doute profite à l'accusé et où il faut monter un dossier solide ("beyond a reasonable doubt" en anglais) pour condamner quelqu'un. Ici, l'accusé est H_0 : on ne la rejette (i.e., on ne la condamne) que si l'on dispose d'éléments à charge suffisamment forts et éloquents, pas juste si l'on pense qu'elle pourrait ne pas être vraie. La charge de la preuve est sur le rejet de H_0.

Le vocabulaire est éclairant. Comme par construction un test est attaché à H_0, on n'accepte jamais H_0 ni ne la valide, au mieux, on la conserve ("fail to reject" en anglais) ; dans ce cas, il y a absence de progrès. En revanche, on peut, en cas de contradiction grave avec les données, rejeter H_0 (et passer à H_1), auquel cas, un progrès négatif est effectué.

Ces principes nous conduisent en pratique aux critères suivants pour choisir nos couples d'hypothèses (selon les exercices, l'un ou l'autre critère s'applique mieux) :

— H_0 : le contraire de ce que l'on veut prouver vs. H_1 : ce que l'on veut prouver.

— H_0 : une situation naturelle vs. H_1 : une situation qui susciterait des réflexions ou des actions.

On voit que dans le choix des hypothèses, ce qui est critique, c'est de déterminer H_1, et un sens, H_0 s'en déduit. La fin de ce chapitre expliquera que pour les cas simples de tests à une valeur de référence (proportion p_{ref} ou moyenne μ_{ref}), l'hypothèse H_0 peut toujours être prise comme l'égalité $p_0 = p_{ref}$ ou $\mu_0 = \mu_{ref}$, tandis qu'il faut justifier si l'on prend pour H_1 une hypothèse bilatère $p_0 \neq p_{ref}$ ou $\mu_0 \neq \mu_{ref}$, ou une hypothèse unilatère avec $<$ ou $>$.

Nous traiterons en cours les deux exemples introductifs suivants. La clé, dans les deux cas, sera de trouver une variable aléatoire T_n, avec les caractéristiques suivantes. D'une part, T_n peut être calculée à partir des données $x_1, \dots, x_n$ de l'échantillon, vues comme réalisations de variables aléatoires $X_1, \dots, X_n$; on appelle T_n une statistique de test. D'autre part, la loi de T_n est parfaitement connue sous H_0 et son comportement sous H_1 est fort différent, ce qui nous permettra de faire la part des choses entre H_0 et H_1.

Exemple historique : "The Lady tasting tea". Peut-on distinguer le goût du thé au lait selon que le lait a été versé avant ou après le thé ? Muriel Bristol prétend qu'elle le peut. Cet exemple, issu d'une situation réelle, a permis à Ronald Fisher (1890–1962) de fonder, avec son ouvrage *The Design of Experiments* (1935) la théorie moderne des tests statistiques. Demandez-vous pour l'heure qui sont H_0 et H_1 parmi les deux assertions « Muriel Bristol a le super-pouvoir de distinguer ces goûts » et « Elle ne l'a pas ».

Autre exemple à méditer : Si j'étais prof de lycée... Je demanderais à mes élèves de lancer 200 fois une pièce en guise de devoirs, et de me noter la suite des pile ou face obtenue. Certains tricheraient tandis que d'autres effectueraient de manière honnête leurs devoirs, et j'aurais une méthode pour détecter les tricheurs.

Demandez-vous pour l'heure qui sont H_0 et H_1 parmi les deux assertions « Un élève donné a fait ses devoirs de manière honnête » et « Il a triché ».

Déterminez également le degré de sévérité associé ici aux faits de rejeter à tort H_0, d'une part, et de conserver à tort H_0, d'autre part. Laquelle de ces deux erreurs veut-on contrôler en priorité, et de quelle erreur est-on obligé de s'accomoder ?

2. Test de comparaison à une proportion de référence p_{ref}

Les exemples que nous traitons dans cette section sont tirés de l'article de recherche

> « Filles/garçons, l'offre de loisirs : asymétrie des sexes, décrochage des filles et renforcement des stéréotypes », E. Maruéjouls et Y. Raibaud, *Ville École Intégration*, vol. 67, pp. 86-91, 2012

qui étudie notamment la participation des garçons et des filles aux activités publiques sportives ou musicales (i.e., organisées par des communes ou subventionnées par elles). Bien évidemment, ce genre d'études statistiques a vocation à inspirer les politiques publiques, afin de les rendre plus inclusives si nécessaire.

Les communes considérées pour recruter les échantillons sont toutes trois situées dans la banlieue de Bordeaux et nous supposons que les comportements des enfants qui y habitent sont représentatifs de ceux de l'ensemble de la France. Pour chaque activité sportive ou musicale, les chercheurs se demandent si on peut dire que les garçons et les filles y participent de manière égale, i.e., si l'activité est neutre du point de vue du genre, ou si au contraire, il existe un biais de genre (au détriment des filles ou au détriment des garçons). On fixe le paramètre d'intérêt comme la proportion p_0 de filles participant à l'activité considérée (dans toute la France ; on rappelle qu'on a supposé l'échantillon représentatif). On considère qu'il y a autant de filles que de garçons en France, i.e, on va comparer p_0 à la proportion de référence $p_{ref} = 50\%$.

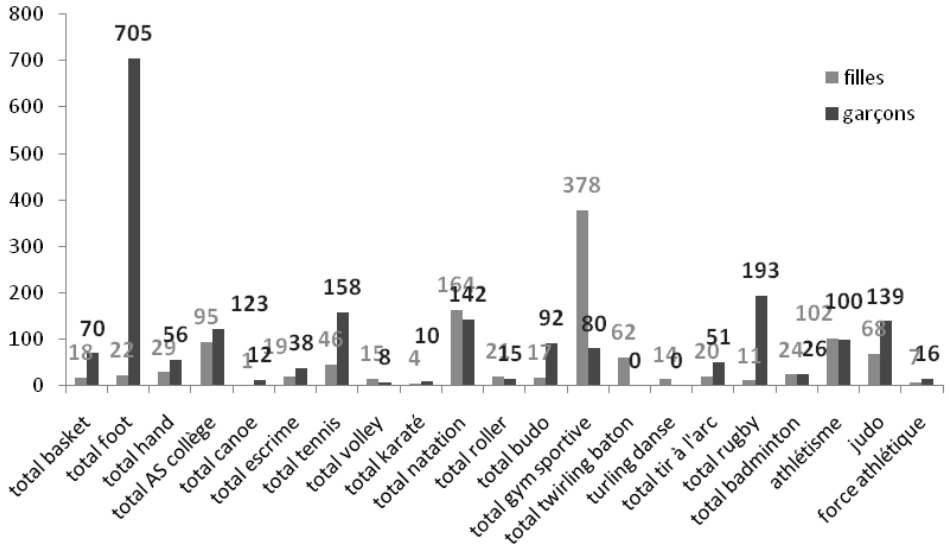


FIGURE 3.1 – Pratiques sportives par sexe dans les trois communes étudiées.

1. Choix des hypothèses. L'approche se doit d'être fondamentalement neutre et identique pour toutes les activités. Or, comme nous le verrons plus tard, nous n'avons pas le droit de fixer les hypothèses au vu des données ; au contraire, il faut les fixer avant même toute collecte de données et concevoir intégralement l'expérience statistique avant de la mener. C'est pourquoi nous allons recourir ici à des hypothèses dites bilatères :
– $H_0 : p_0 = p_{ref}$, i.e., l'activité est neutre du point du vue du genre ;
– H_1 [ce qui susciterait réflexion] : $p_0 \neq p_{ref}$, i.e., l'activité connaît un biais de genre, dans un sens ou dans l'autre.
Une fois ces hypothèses posées, les chercheurs ont recueilli les données résumées dans la Figure 3.1.

2. Conception du test statistique. Supposons que l'on dispose de variables aléatoires $X_1, X_2, \ldots$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0. Le paramètre p_0 est inconnu, mais on veut tester son égalité à p_{ref}.

Or, par loi des grands nombres, $\overline{X}_n$ est proche de p_0. Pour tester si $p_0 = p_{ref}$, il est donc naturel de considérer l'écart $\overline{X}_n - p_{ref}$. On rappelle que l'on veut construire une statistique de test T_n telle que (i) son comportement soit parfaitement déterminé sous H_0 et (ii) son comportement sous H_1 soit très différent du comportement sous H_0.

Pour (i), on se place donc sous $H_0 : p_0 = p_{ref}$. La loi de l'écart $\overline{X}_n - p_{ref}$ dépend beaucoup de p_{ref}, mais la loi de l'écart normalisé

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

est universelle, au moins à la limite. En effet, le théorème limite central [1] indique la convergence en loi $T_n \rightharpoonup \mathcal{N}(0,1)$.

Pour (ii), on se place sous $H_1 : p_0 \neq p_{ref}$ et on distingue deux cas. Le premier cas est que $H_1' : p_0 > p_{ref}$.

1. Ici, il n'est pas besoin de le combiner au lemme de Slutzly, car l'écart-type est parfaitement connu sous l'hypothèse $H_0 : p_0 = p_{ref}$, c'est $\sqrt{p_{ref}(1-p_{ref})}$; il n'y a pas besoin de l'estimer de manière convergente par $\sqrt{\overline{X}_n(1-\overline{X}_n)}$.

Par loi des grands nombres, on a donc

$$\text{sous } H_1' : p_0 > p_{\text{ref}}, \qquad \overline{X}_n - p_{\text{ref}} \xrightarrow{\mathbb{P}} p_0 - p_{\text{ref}} > 0\,, \qquad \text{d'où} \qquad T_n \xrightarrow{\mathbb{P}} +\infty$$

(par multiplication de la limite finie positive $p_0 - p_{\text{ref}} > 0$ par la limite infinie de $\sqrt{n}$). Symétriquement, $T_n \to -\infty$ en probabilité sous $H_1'' : p_0 < p_{\text{ref}}$. Ainsi, dans tous les cas,

$$\text{sous } H_1 : p_0 \neq p_{\text{ref}}, \qquad |T_n| \xrightarrow{\mathbb{P}} +\infty\,.$$

Résumé de (i) et (ii) : sous H_0, la statistique de test T_n tend à prendre des valeurs autour de 0, puisqu'une loi $\mathcal{N}(0,1)$ prend le plus souvent des valeurs entre -2 et $+2$; au contraire, sous H_1, la statistique de test T_n tend à prendre de grandes valeurs positives ou négatives. Nous constatons bien les deux comportements très différents sous H_0 et H_1 que nous appelions de nos vœux.

3. Collecte des données. Nous allons étudier le cas de la natation, pour laquelle les chercheurs ont compté 164 filles and 142 garçons inscrits en club. Ces données correspondent à $n = 306$ données correspondant à la réalisation de $X_1, \ldots, X_{306}$ indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre p_0. Cette modélisation vaut si chaque enfant choisit son sport de manière indépendante des autres enfants, avec des lois de choix possiblement différentes pour les filles et les garçons ; la limite de cette modélisation est bien entendu que les enfants peuvent être influencés par le choix de leurs amis...

La proportion de filles dans l'échantillon est de $164/306 = 0.5359... \approx 53.6\%$, et tester les hypothèses bilatères $H_0 : p_0 = 50\%$ vs. $H_1 : p_0 \neq 50\%$ revient à se poser la question suivante :

> La proportion d'échantillon 53.6% est-elle significativement différente de 50% ?

Vous noterez que tout le sel de la question est dans l'adverbe « significativement » !

4. Application du test : concept de P-valeur. On commence par calculer la valeur réalisée de la statistique de test

$$t_{306} = \sqrt{306}\,\frac{\overline{x}_{306} - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1-p_{\text{ref}})}} = \sqrt{306}\,\frac{164/306 - 0.5}{\sqrt{0.5(1-0.5)}} = 1.2576... \approx 1.26\,.$$

Cette valeur réalisée 1.26 peut être interprétée comme une version normalisée de l'écart $53.6\% - 50\% = 3.6\%$. Il est difficile d'apprécier si l'écart nominal 3.6% est significatif, mais avec un tout petit peu d'expérience, on peut immédiatement apprécier si l'écart normalisé 1.26 est significatif, ou non.

A cet effet, la question que l'on se pose est la suivante :

> Cette valeur 1.26 est-elle une valeur typique sous H_0, i.e., une valeur typique pour une loi $\mathcal{N}(0,1)$, ou au contraire, cette valeur 1.26 est-elle plutôt à rattacher aux comportements de grandes valeurs négatives ou positives établis sous H_1 ?

On raisonne à partir de la Figure 3.2, sur laquelle on a porté le comportement de la statistique de test T_n sous H_0 (en l'occurrence, une loi normale, dont on a représenté la densité de probabilité) et sous H_1 (de grandes valeurs positives ou négatives, symbolisées par des flèches). Les valeurs autour de 0 amèneront à conserver H_0, tandis que les grandes valeurs positives ou négatives nous feront rejeter H_0 au profit de H_1. On hachure la zone au-delà de la valeur $t = 1.26$ et de manière symétrique, celle entre $-\infty$ et -1.26, du fait du caractère bilatère du test, qui ne nous permettait pas de savoir à l'avance dans laquelle des deux zones nous l'écart normalisé se situerait. La probabilité de cette zone hachurée s'appelle la P–valeur. Son interprétation est exactement la suivante :

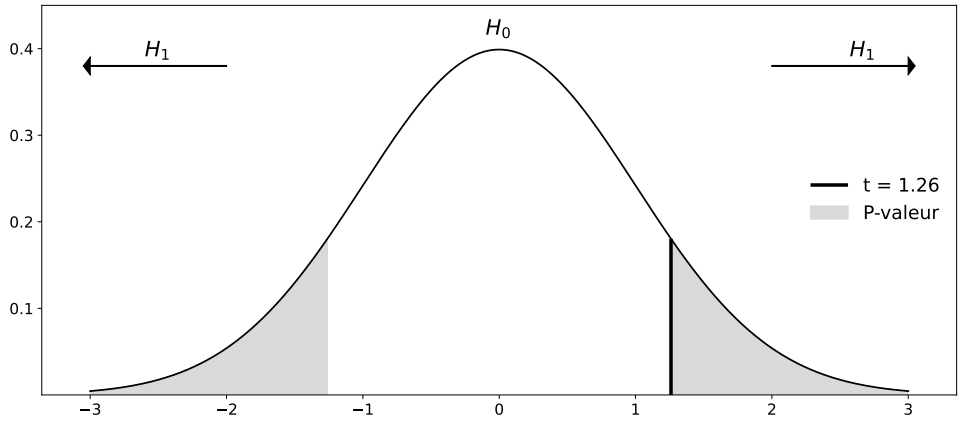


FIGURE 3.2 – Appréciation du caractère typique de l'écart normalisé $t = 1.26$ dans le test bilatère $H_0 : p_0 = p_{ref}$ vs. $H_1 : p_0 \neq p_{ref}$.

> La P–valeur est la probabilité, si l'on répétait l'expérience, d'obtenir une valeur de la statistique de test t au moins aussi contradictoire avec H_0 que celle que l'on vient d'obtenir.

En particulier,

— une P–valeur faible signifie que la valeur actuelle de t est déjà très contradictoire avec H_0, puisqu'il est difficile d'obtenir des valeurs aussi contradictoires ou plus contradictoires qu'elle ;

— une P–valeur forte signifie au contraire qu'il est facile d'obtenir des valeurs pour t plus contradictoires que la valeur actuelle, qui est donc peu contradictoire avec H_0.

Par convention et pour des raisons historiques (liées aux travaux du pionnier Ronald Fisher, 1890–1962), le seuil entre « faible » et « forte » est fixé à 5%, au moins en sciences de gestion. (En physique, le seuil est plus bas, plutôt autour de 0.1%.)

Ainsi, la P–valeur est bien une mesure de typicité : une grande P–valeur montre que la valeur actuellement obtenue pour t est typique, car il est facile d'observer des valeurs plus grandes en valeur absolue et donc davantage signe de H_1 que la valeur actuelle.

Une autre manière de formuler cette « mesure de typicité » est de d'interpréter la P–valeur comme un indice de crédibilité de H_0 au vu des données et face à H_1.

4bis. Application du test : calcul effectif de la P–valeur. Nous avons défini le concept de P–valeur, mais il reste à la calculer de manière numérique sur la Figure 3.2. A cet effet, nous utilisons la table de la loi normale fournie au tout début de ce polycopié. Elle est fondée sur l'illustration de la Figure 3.3.

La table fournit, pour certaines valeurs de u, la valeur $F(u)$ de la fonction de répartition F de la loi $\mathcal{N}(0,1)$ en u. Ces valeurs de u, lorsqu'elles sont comprises entre 0 et 3, sont graduées de 0.01 en 0.01. Par exemple, $u = 1.26$ est décomposée en $u = 1.2 + 0.06$, et la valeur de $F(1.26)$ est à lire en ligne 1.2 et colonne 0.06 ; en l'occurrence, $F(1.26) = 0.8962 = 89.62\%$. La P-valeur correspond ici (comparer les Figures 3.2 et 3.3 et utiliser des arguments de symétrie) à

$$\mathbb{P}(Z \leqslant -1.26) + \mathbb{P}(Z \geqslant 1.26) = 2 \times \mathbb{P}(Z \geqslant 1.26) = 2 \times \big(1 - \underbrace{\mathbb{P}(Z \leqslant 1.26)}_{=F(1.26)}\big) = 2 \times (100\% - 89.62\%) = 20.76\%,$$

où Z désigne une variable aléatoire distribuée selon la loi $\mathcal{N}(0,1)$.

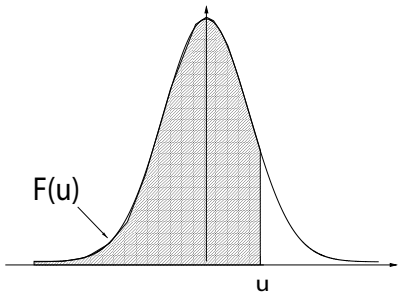


FIGURE 3.3 – Figure d'illustration de la table de la loi normale : pour tout u, on lit dans le tableau sous-jacent la valeur $F(u)$ de la fonction de répartition F de la loi $\mathcal{N}(0,1)$ en u.

Ici, la P–valeur est supérieure à 5%, l'hypothèse H_0 est donc suffisamment crédible au vu des données, nous ne pourrons pas la rejeter, i.e., nous allons la conserver.

5. Conclusion lisible par le grand public. Votre conclusion se doit à la fois de (i) traduire exactement la force du résultat statistique et (ii) d'être lisible.

Pour (i), il s'agit ici de transmettre l'idée qu'on conserve H_0, mais sans la valider ; simplement, on ne s'en départ pas, on ne la rejette pas. Fisher dit à cet égard :

> The hypothesis H_0 "is never proved or established, but is possibly disproved, in the course of experimentation".

Il est par exemple interdit d'écrire ici comme conclusion que « ~~Ces données démontrent que la natation est une activité non genrée~~ », c'est une conclusion bien trop affirmative pour une simple conservation de H_0.

Pour (ii), j'entends par là que les conclusions trop techniques comme « On conserve $H_0 : p_0 = p_{\text{ref}}$ » ont beau être exactes, elles ne me conviennent pas à l'examen. En effet, dans votre vie professionnelle future, vous devrez pouvoir communiquer vos résultats à des groupes de décideurs variés, non nécessairement munis d'une culture technique.

En l'occurrence ici, deux conclusions élégantes possibles sont :

> Ces données ne permettent pas d'exclure que la natation soit une activité non genrée.
>
> Ces données ne mettent en évidence aucun biais significatif en termes de genre concernant la natation.

Dans les deux cas, le message que l'on veut transmettre, c'est que la proportion d'échantillon 53.6% n'est pas significativement différente de la proportion de référence 50%.

Encore une fois, cela ne prouve pas que la natation soit une activité non genrée : simplement, avec cet échantillon de 306 individus, on peut pas rejeter l'idée que la natation soit une activité non genrée. Qui sait si avec des milliers d'individus on ne pourrait pas mettre en évidence un biais de genre ?

5bis. Cas du handball. Reprenons l'ensemble de la méthodologie avec les données du handball, à savoir 29 filles et 56 garçons sur les $n = 29 + 56 = 85$ individus sondés. Cela correspond à une proportion empirique de $29/85 \approx 34.1\%$ de filles, dont on se demande si elle est significativement différente de 50%. On calcule la statistique de test (écart normalisé)

$$t = \sqrt{85}\,\frac{129/85 - 0.5}{\sqrt{0.5(1-0.5)}} = -2.9285... \approx -2.93\,,$$

et elle conduit, selon le même principe que ci-dessus, à la P–valeur

$$2 \times (100\% - 99.83\%) = 0.34\%,$$

très inférieure à 5%. On rejette donc fermement H_0 et on passe à H_1. Ici, on a prouvé, avec une certitude plus que raisonnable, que H_0 est invalide, et on peut donc écrire une conclusion grand public qui soit affirmative. Le message que l'on veut transmettre, c'est que la proportion d'échantillon 34.1% est significativement différente de la proportion de référence 50%, en l'occurrence, qu'elle lui est significativement inférieure :

> Ces données montrent que le handball constitue une activité genrée, en l'occurrence, les garçons y sont surreprésentés.

Remarque 3.2 (Dissymétrie des forces des conclusions). Les tests ne font des progrès que négatifs ; dit autrement, les forces des conclusions sont dissymétriques :

— lorsque l'on conserve H_0, on conclut simplement que H_0 n'est pas impossible ;

— lorsque l'on rejette H_0, on conclut qu'H_0 est impossible (à un niveau de certitude raisonnable).

3. Test de comparaison à une moyenne de référence μ_{ref}

Une fois n'est pas coutume, nous nous intéressons ici à un exemple typique de marketing.

Supposons que nous dirigions une entreprise de e–commerce depuis de longs mois. Nous observons qu'en moyenne, nos clients passent des commandes de $\mu_{\text{ref}} = 28$ euros par mois. Cette valeur de 28 euros est une vérité statistique, obtenue comme moyenne des commandes des dizaines de milliers de clients de notre fichier clients (ou d'absences de commandes : nous notons évidemment 0 euro lorsque le client en question n'a rien commandé).

Nous voulons maintenant booster nos ventes et voulons tester à cet effet une offre commerciale (peu importe en quoi elle consiste pour cet exercice). Comme cette offre est éventuellement difficile à mettre en œuvre à grande échelle, ou un peu coûteuse, etc., nous voulons d'abord tester son impact sur un petit nombre de clients choisis au hasard, disons 500 clients. Grâce à eux, nous allons collecter des montants d'achats $x_1, \ldots, x_{500}$. Ceux-ci nous permettront de discuter notre paramètre d'intérêt μ_0, qui est le montant moyen d'achats qu'effectueraient l'ensemble de nos clients si nous leur offrions cette offre commerciale. Nous souhaitons bien entendu savoir si μ_0 est strictement supérieur à μ_{ref}.

1. Choix des hypothèses. Il s'agit ici de prouver que l'offre commerciale fonctionne ; or ce qu'on veut prouver c'est H_1, donc $H_1 : \mu_0 > \mu_{\text{ref}}$. L'hypothèse H_0 est donnée par les contraires de H_1, mais uniquement les contraires plausibles (il n'est pas obligatoire de couvrir [2] tous les cas). Il est difficile d'imaginer qu'une offre commerciale amène à baisser les ventes, c'est pourquoi H_0 correspondrait plutôt à l'absence d'impact de l'offre, i.e., $H_0 : \mu_0 = \mu_{\text{ref}}$.

Les hypothèses sont dites ici unilatères, parce que des déviations par rapport à la valeur de référence n'ont lieu sous H_1 que dans une direction (ici, par valeurs supérieures), contrairement à l'exemple sur les activités genrées, dont les hypothèses étaient bilatères.

2. En fait, nous expliquerons de toute façon en conclusion de ce chapitre pourquoi tester $H_0 : \mu_0 = \mu_{\text{ref}}$ vs. $H_1 : \mu_0 > \mu_{\text{ref}}$ et $H_0 : \mu_0 \leqslant \mu_{\text{ref}}$ vs. $H_1 : \mu_0 > \mu_{\text{ref}}$ mène à la même P-valeur (par « passage au cas limite »).

2. Conception du test statistique. Supposons que l'on dispose de variables aléatoires $X_1, X_2, \ldots$ indépendantes et identiquement distribuées selon une loi d'espérance μ_0 et admettant un écart-type σ_0. Ici, les hypothèses ne nous renseignent plus sur la valeur de l'écart-type. Nous allons donc nous référer à la page 23 et employer à nouveau la convergence en loi (2.1), qui nous donne au passage l'expression de la statistique de test T_n :

$$\text{sous } H_0 : \mu_0 = \mu_{\text{ref}}, \qquad T_n = \frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_{\text{ref}}\right) \rightharpoonup \mathcal{N}(0,1)\,.$$

Sous $H_1 : \mu_0 > \mu_{\text{ref}}$, la loi des grands nombres assure que $\overline{X}_n - \mu_{\text{ref}} \to \mu_0 - \mu_{\text{ref}} > 0$ en probabilité, d'où :

$$\text{sous } H_1 : \mu_0 > \mu_{\text{ref}}, \qquad T_n = \frac{\sqrt{n}}{\hat{\sigma}_n}\left(\overline{X}_n - \mu_{\text{ref}}\right) \xrightarrow{\mathbb{P}} +\infty\,.$$

Le test étant unilatère ici, les déviations sous H_1 n'ont lieu que dans une direction : les grandes valeurs positives. Nous verrons les incidences en termes de calcul de la P-valeur.

3. Collecte des données. Il se trouve que les 500 clients à qui l'offre commerciale a été proposée effectuent des achats d'un montant moyen (sur le mois considéré) de 29.76 euros (avec un écart-type de 21.45 euros dans ces montants : lié au fait que certains commandent beaucoup et d'autres, rien du tout).

Dit autrement, nous disposons de données $x_1, \ldots, x_{500}$, que l'on peut résumer par $\overline{x}_{500} = 29.76$ euros et $s_{x,500} = 21.45$ euros, et qui sont la réalisation de variables aléatoires $X_1, \ldots, X_{500}$ indépendantes et identiquement distribuées selon une loi d'espérance μ_0. Cette modélisation vaut parce que l'on tire uniformément au hasard un petit nombre de clients dans un fichiers clients avec plusieurs dizaines de milliers d'individus (cf. modèle de tirage dans une urne, avec ou sans remise).

Tester les hypothèses unilatères $H_0 : \mu_0 = 28$ vs. $H_1 : mu_0 \neq 28$ revient à se poser la question suivante :

> La moyenne d'échantillon 29.76 euros est-elle significativement supérieure à 28 euros ?

Là encore, tout l'intérêt de la question est dans l'adverbe « significativement », qui correpond au fait de pouvoir, ou pas, généraliser la relation $\overline{x}_{500} > 28$ observée à l'ensemble du fichier clients, i.e. au fait que $\mu_0 > 28$.

4. Calcul de la P-valeur et sort de H_0. La valeur réalisée de la statistique de test, i.e., la version normalisée de l'écart $29.76 - 28 = 1.76$, vaut

$$t_{500} = \sqrt{500}\,\frac{\overline{x}_{500} - \mu_{\text{ref}}}{s_{x,500}} = \sqrt{500}\,\frac{29.76 - 28}{21.45} = 1.8347... \approx 1.83\,.$$

La P–valeur est calculée en utilisant cette fois-ci (voir Figure 3.4) que les déviations sous H_1 n'ont lieu que vers les grandes valeurs positives :

$$\mathbb{P}(Z \geqslant 1.83) = 1 - \underbrace{\mathbb{P}(Z \leqslant 1.83)}_{=F(1.83)} = 100\% - 96.64\% = 3.36\% < 5\%\,.$$

On rejette donc l'hypothèse H_0 au profit de H_1.

5. Conclusion lisible par le grand public. Nous avons rejeté H_0, et avons donc effectué un progrès négatif :

> Ces données montrent que l'offre commerciale augmente bien le montant moyen des ventes.

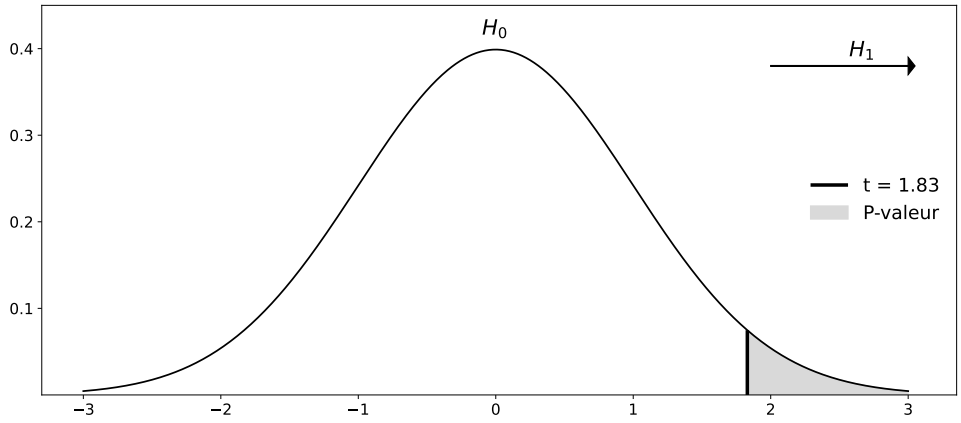


FIGURE 3.4 – Calcul de la P–valeur pour $t = 1.83$ dans le test unilatère $H_0 : \mu_0 = \mu_{ref}$ vs. $H_1 : \mu_0 > \mu_{ref}$.

Remarque 3.3 (Résumé des variations). Dans cet exemple, nous avons considéré deux variations par rapport au cas des possibles biais de genre. D'une part, nous avons testé des moyennes générales, le changement correspondant étant donné par une statistique de test où un écart-type empirique est utilisé au dénominateur. D'autre part, nous avons utilisé des hypothèses unilatères, le changement correspondant étant que dans le calcul de la P–valeur, il n'y a qu'une zone de rejet à considérer (ici, à droite) au lieu de deux zones symétriques par rapport à l'axe des ordonnées dans le cas bilatère.

4. Conclusion : choix des hypothèses

J'aimerais revenir sur le choix des hypothèses, thème qui avait ouvert ce chapitre, et formuler différents commentaires ou remarques.

Pourquoi il est interdit de fixer les hypothèses au vu des données. Dans les exemples traités précédemment, nous avons toujours fixé nos hypothèses et conçu l'expérience avant de collecter les données. C'est la bonne manière de faire. Si nous observions les données avant de choisir les hypothèses, nous ne mènerions sans doute que des tests unilatères, même dans des contextes très neutres où nous n'avions aucune idée à l'avance des déviations attendues. Cela diviserait artificiellement les P–valeurs par 2 (comparer les Figures 3.1 et 3.4).

Il suffit de justifier la forme de H_1. Nous avons vu qu'une bonne statistique de test était une variable aléatoire calculable sur les données, de comportement parfaitement déterminé sous H_0 et exhibant des comportements fort différents sous H_1. A notre niveau, le comportement « parfaitement déterminé sous H_0 » ne sera obtenu que lorsque H_0 est de la forme $\mu_0 = \mu_{ref}$ ou $p_0 = p_{ref}$. Il suffira alors de justifier la forme de H_1, entre

$$\begin{array}{lccc} & H_1 : \mu_0 \neq \mu_{ref}, & H_1 : \mu_0 > \mu_{ref}, & H_1 : \mu_0 < \mu_{ref}; \\ \text{ou} & H_1 : p_0 \neq p_{ref}, & H_1 : p_0 > p_{ref}, & H_1 : p_0 < p_{ref}. \end{array}$$

Passage au cas limite. Dans certains exercices, les hypothèses s'écrivent naturellement sous la forme (disons dans le cas de proportions)

$$H_0 : p_0 \leqslant p_{ref} \quad \text{vs.} \quad H_1 : p_0 > p_{ref},$$

et on peut montrer que la P-valeur obtenue sur des données avec les hypothèses précédentes est la même qu'avec les hypothèses

$$H_0 : p_0 = p_{\text{ref}} \quad \text{vs.} \quad H_1 : p_0 > p_{\text{ref}}.$$

Le cas $p_0 = p_{\text{ref}}$ apparaît comme le cas limite de $p_0 \leqslant p_{\text{ref}}$. C'est la valeur de p_0 sous $H_0 : p_0 \leqslant p_{\text{ref}}$ qui est la plus proche des valeurs considérées sous H_1, c'est donc la valeur intuitivement la plus difficile à tester, et cela explique, toujours intuitivement, pourquoi il suffit de tester ce cas-là.

Cette réduction d'une hypothèse H_0 formulée avec des inégalités larges $\leqslant$ ou $\geqslant$ à l'unique cas d'égalité sera appelée « passage au cas limite ». (Une justification mathématique précise mais totalement facultative sera mise à disposition sous BlackBoard pour les fans de mathématiques ; elle est omise de ce polycopié.)

Cas d'égalité toujours dans H_0. Des arguments similaires à ceux du point précédent expliquent que tester $H_0 : p_0 < p_{\text{ref}}$ vs. $H_0 : p_0 \geqslant p_{\text{ref}}$ revient à tester $H_0 : p_0 \leqslant p_{\text{ref}}$ vs. $H_0 : p_0 \geqslant p_{\text{ref}}$. Autrement dit, on n'échappe pas au cas d'égalité sous H_0, et c'est pourquoi je compterai comme fausses des hypothèses énoncées sous la forme type $H_0 : p_0 < p_{\text{ref}}$ ou $H_0 : p_0 > p_{\text{ref}}$.

De la même manière, tester $H_0 : p_0 \neq p_{\text{ref}}$ contre n'importe quelle hypothèse H_1 (mais les étudiants ont souvent en tête $H_1 : p_0 = p_{\text{ref}}$) revient à tester $H_0 : p_0 \in [0\%, 100\%]$ contre ce H_1. Il ne vous échappe pas qu'il est rigoureusement impossible de rejeter cette hypothèse H_0, et qu'en réalité, la P-valeur de cette hypothèse $H_0 : p_0 \neq p_{\text{ref}}$ sera toujours de 100%. Bref, il ne faut jamais choisir $H_0 : p_0 \neq p_{\text{ref}}$!

5. Défi mathématique facultatif

Cet exercice totalement facultatif (pour lequel je posterai une correction sous BlackBoard) a pour vocation de faire méditer le fait suivant :

> Même lorsque H_0 est vraiment vraie, on n'obtient pas toujours de grandes P–valeurs ; en l'occurrence, une petite fraction du temps on obtient de faibles P–valeurs conduisant au rejet à tort de H_0.

Plus précisément :

1. Montrez que si $T \sim \mathcal{N}(0,1)$, alors $F(T) \sim \mathcal{U}_{[0,1]}$, où F désigne la fonction de répartition de la loi $\mathcal{N}(0,1)$ et où $\mathcal{U}_{[0,1]}$ désigne la loi uniforme sur $[0,1]$.
2. Déduisez-en que lorsque $H_0 : \mu_0 = \mu_{\text{ref}}$ est vraie et qu'on la teste contre une hypothèse H_1 unilatère, alors la P–valeur est uniformément distribuée entre 0% et 100%.

Moralité : Lorsque H0 est vraie, on a par exemple que 95% du temps, la P–valeur est supérieure à 5%, mais également (attention !), que par exemple 25% du temps, cette P–valeur est inférieure à 25%. Même quand on sait par ailleurs que H0 est vraie (parce qu'on l'a prouvée par des méthodes extra-statistiques), on ne peut pas obtenir tout le temps de grandes P–valeurs.

J'expliquerai en cours comment cette observation a été appliquée rétrospectivement aux travaux du généticien Mendel, dont on a pu démontrer qu'il avait embelli ses données en multipliant la taille de ses échantillons par un facteur 10. C'était pour une bonne cause — pour dépasser les réticences de ses contemporains face à la théorie du brassage génétique au hasard — mais c'est statistiquement condamnable.

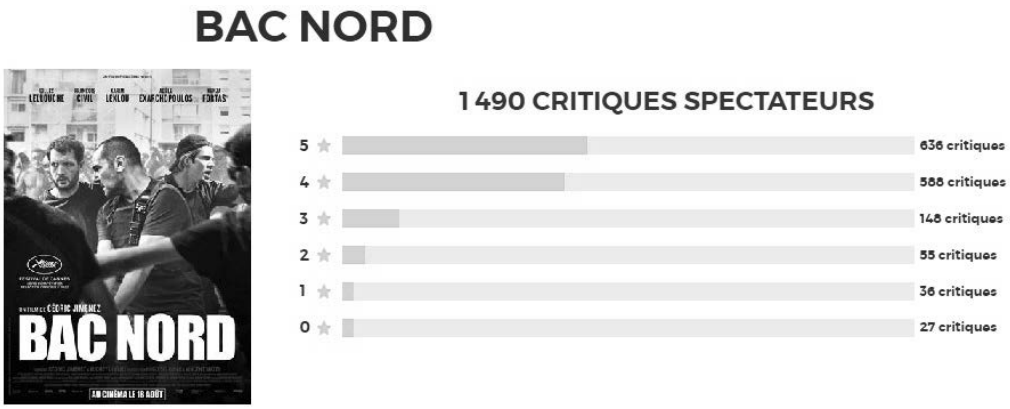
Représentativité d'un échantillon

Un échantillon de 200 personnes est constitué pour une étude sur [ce que vous voulez, par exemple, des aménagements locaux dans une ville moyenne en France]. Il contient 95 femmes et 105 hommes. Une association féministe se plaint que la voix des femmes est encore une fois minorée, car elles représentent 51.5% de la population et que l'échantillon aurait donc dû compter 103 femmes. Que doit en penser le maire, et que doit-il faire ? Fondez votre réponse sur un test d'hypothèses et une P-valeur.

Variantes : Mêmes questions avec 88 femmes et 112 hommes ; puis avec 117 femmes et 83 hommes.

Prix Allociné des critiques amateurs — 10 points + 1 point bonus

Supposons que le bien connu site Allociné (qui, entre autres, recueille les avis des spectateurs sur les films qu'ils ont vus au cinéma) veuille décerner des prix des critiques amateurs. Pour cela, le site ne retient, pour chaque film, que les notes (entre 1 et 5 étoiles) assorties d'un commentaire critique. Par exemple, pour le film *Bac Nord*, seules 1 490 notes décernées sont assorties de critiques spectateurs, sur les 14 542 notes décernées au 20 septembre 2022. Ces notes avec critiques se répartissent ainsi :



Le site se fixe un seuil de 4 sur 5 pour décerner un prix des critiques amateurs, mais hésite entre deux stratégies :

(P) ne jamais décerner un tel prix qu'à bon escient, quand il est vraiment sûr que la note moyenne franchit le seuil de 4 sur 5 (et restera dans cette zone)—c'est la stratégie P, pour prudente ;

(FTBR) ne jamais passer à côté d'un film et décerner le prix dès qu'il est possible que la note moyenne franchisse (un jour) le seuil de 4 sur 5—c'est la stratégie FTBR, qui fait feu de tout bois raisonnable.

Les deux stratégies reposent sur l'hypothèse que les critiques déjà écrites sont représentatives des critiques futures. Le choix de la collation du prix s'effectuera au moment de la 1 500^{e} critique.

Stratégie P et film *Bac Nord*

On se place dans le cadre de la stratégie P.

☐ Enoncez les hypothèses qu'elle considère.

☐ Justifiez brièvement votre réponse.

Bac Nord obtient quelques jours après 10 critiques de plus et arrive au seuil de 1 500 critiques. La moyenne actuelle de ses notes assorties de critiques est de 4.11 sur 5, avec un écart-type dans les notes de 1.09. Convient-il de lui décerner le prix ? Répondez en mettant en œuvre un test d'hypothèses, c'est-à-dire,

☐☐☐ procurez un schéma récapitulant les comportements de la statistique de test adéquate sous H_0 et H_1, calculez la valeur réalisée de cette statistique de test (*procurez les détails du calcul, c'est-à-dire : écrivez le calcul ce que vous tapez sur votre calculatrice*), et indiquez la P-valeur obtenue.

☐ Conclusions : Quel est le sort de vos hypothèses ? Le prix doit-il être décerné, selon la stratégie P ?

Stratégie FTBR et film *BRI Sud*

On se place désormais dans le cadre de la stratégie FTBR.

☐ Enoncez les hypothèses qu'elle considère, sans les justifier.

Le film *BRI Sud* vient d'obtenir 1 500 critiques de spectateurs, pour une note moyenne de 3.95, avec un écart-type dans les notes de 1.43.

☐☐☐ Convient-il de lui décerner le prix, selon la stratégie FTBR ? Justifiez votre réponse.

Question bonus (difficile)

☐ Que peut-on dire, *sans calculs*, sur la P-valeur qui serait obtenue pour *BRI Sud* avec les hypothèses de la stratégie P ? *Il s'agit d'une question bonus, montant le barème à 36 points.*

Application podomètre — 10 points

De récentes recommandations relayées par la presse pour combattre le sédentarisme sont qu'il est sain de marcher 10 000 pas par jour. Des applications podomètre ont vu le jour pour aider leurs utilisateurs à tenir un journal statistique de leurs performances (ou de leur manque de performances) à cet égard.
Supposons que vous veuilliez lancer une telle application, qui ait un petit supplément d'âme : l'affichage d'une pastille de couleur pour indiquer à chaque utilisateur si, en maintenant le rythme de marche actuel,

- nous sommes certains qu'il remplira bien, en moyenne, l'objectif d'au moins 10 000 pas par jour (pastille verte) ;
- nous sommes certains qu'il ne le remplira pas en moyenne (pastille rouge) ;
- nous ne pouvons pas encore nous prononcer avec une certitude raisonnable (pastille orange), en l'absence d'un historique suffisant.

Création du test sous-jacent

☐ Quels sont les individus statistiques considérés ici ? Quel est le paramètre d'intérêt μ_0 visé par l'étude ? Et quelle hypothèse est-elle cruciale pour garantir que nous disposerons d'un échantillon représentatif ?

☐ ☐ Quelles hypothèses faut-il retenir, et pourquoi, au vu de l'objectif de code couleur ternaire ? Justifiez.

Premier jeu de données

Un premier utilisateur emploie l'application pendant 49 jours, au cours desquels il obtient une moyenne de nombres de pas quotidiens de 10 532, avec un écart-type de 3 154 pas dans ces 49 données. Mettez en œuvre un test d'hypothèses idoine

☐ ☐ en traçant un schéma des comportements de la statistique de test sous les hypothèses retenues,
en calculant la valeur numérique de la statistique de test (procurez tous les détails du calcul),
et indiquant la P–valeur associée.

Un logiciel statistique fournit la sortie suivante à partir de ces données.

	N	Moyenne	Ecart type	Erreur standard
Nombre de pas	49	10531,993	3154,049	450,578

Test sur échantillon unique

Valeur de test = 10000

	t	df	Signification p unilatéral	Signification p bilatéral	Différence moyenne	Intervalle de confiance de la différence à 95 % Inférieur	Intervalle de confiance de la différence à 95 % Supérieur
Nombre de pas	1,181	48	,122	,244	532,0	-374,0	1437,9

☐ Expliquez comment y vérifier vos réponses numériques précédentes et fournissez tout commentaire pertinent à cet égard.

☐ Quelle couleur devrait voir cet utilisateur ? Rédigez également une conclusion lisible par le grand public.

Deuxième jeu de données

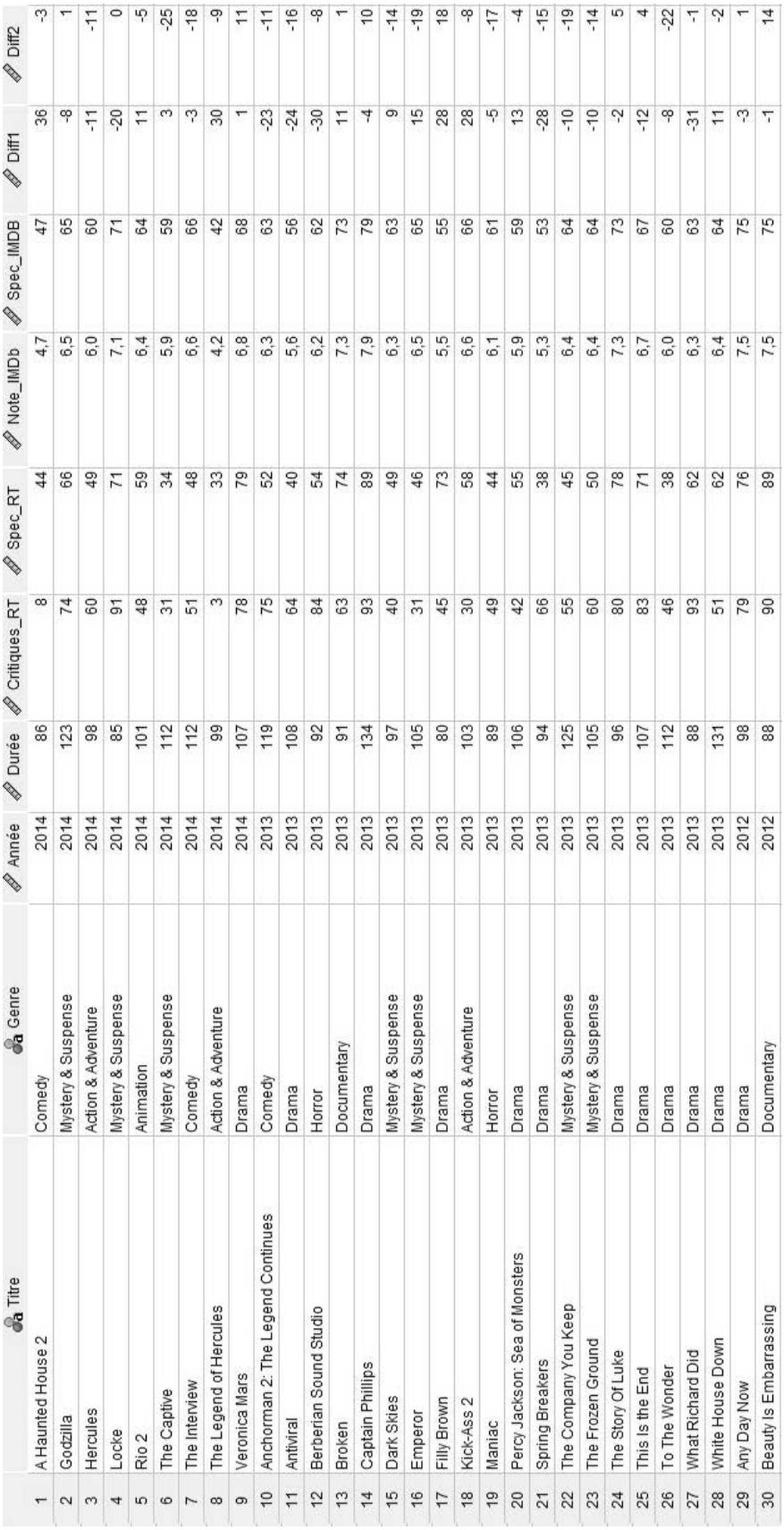
Une deuxième utilisatrice a employé l'application pendant 115 jours et a obtenu une moyenne de 10 452 pas, avec un écart-type de 2 356 steps.

☐ ☐ Calculez la P–valeur associée à ces données et indiquez le code couleur correspondant. (Procurez les détails de vos calculs.)

Troisième jeu de données

Un troisième utilisateur obtient une moyenne de 9 759 steps pendant 62 jours.

☐ Sait-on, au vu de ces informations, quel(s) code(s) couleur(s) il doit voir, ou ne doit pas voir ?



	Titre	Genre	Année	Durée	Critiques_RT	Spec_RT	Note_IMDb	Spec_IMDB	Diff1	Diff2
1	A Haunted House 2	Comedy	2014	86	8	44	4,7	47	36	-3
2	Godzilla	Mystery & Suspense	2014	123	74	66	6,5	65	-8	1
3	Hercules	Action & Adventure	2014	98	60	49	6,0	60	-11	-11
4	Locke	Mystery & Suspense	2014	85	91	71	7,1	71	-20	0
5	Rio 2	Animation	2014	101	48	59	6,4	64	11	-5
6	The Captive	Mystery & Suspense	2014	112	31	34	5,9	59	3	-25
7	The Interview	Comedy	2014	112	51	48	6,6	66	-3	-18
8	The Legend of Hercules	Action & Adventure	2014	99	3	33	4,2	42	30	-9
9	Veronica Mars	Drama	2014	107	78	79	6,8	68	1	11
10	Anchorman 2: The Legend Continues	Comedy	2013	119	75	52	6,3	63	-23	-11
11	Antiviral	Drama	2013	108	64	40	5,6	56	-24	-16
12	Berberian Sound Studio	Horror	2013	92	84	54	6,2	62	-30	-8
13	Broken	Documentary	2013	91	63	74	7,3	73	11	1
14	Captain Phillips	Drama	2013	134	93	89	7,9	79	-4	10
15	Dark Skies	Mystery & Suspense	2013	97	40	49	6,3	63	9	-14
16	Emperor	Mystery & Suspense	2013	105	31	46	6,5	65	15	-19
17	Filly Brown	Drama	2013	80	45	73	5,5	55	28	18
18	Kick-Ass 2	Action & Adventure	2013	103	30	58	6,6	66	28	-8
19	Maniac	Horror	2013	89	49	44	6,1	61	-5	-17
20	Percy Jackson: Sea of Monsters	Drama	2013	106	42	55	5,9	59	13	-4
21	Spring Breakers	Drama	2013	94	66	38	5,3	53	-28	-15
22	The Company You Keep	Mystery & Suspense	2013	125	55	45	6,4	64	-10	-19
23	The Frozen Ground	Mystery & Suspense	2013	105	60	50	6,4	64	-10	-14
24	The Story Of Luke	Drama	2013	96	80	78	7,3	73	-2	5
25	This Is the End	Drama	2013	107	83	71	6,7	67	-12	4
26	To The Wonder	Drama	2013	112	46	38	6,0	60	-8	-22
27	What Richard Did	Drama	2013	88	93	62	6,3	63	-31	-1
28	White House Down	Drama	2013	131	51	62	6,4	64	11	-2
29	Any Day Now	Drama	2012	98	79	76	7,5	75	-3	1
30	Beauty Is Embarrassing	Documentary	2012	88	90	89	7,5	75	-1	14

FIGURE 3.5 – Extrait du jeu de données d'évaluations de films considéré tout au long du Chapitre 4 ; le jeu de données complet comporte 651 observations, et seules les 30 premières sont reportées ici.

4

Test de comparaison de moyennes appariées

Nous allons considérer tout au long de ce chapitre un jeu de données d'évaluations de films, effectuées par des spectateurs et des critiques de cinéma aux Etats-Unis. Les individus statistiques sont des films ; l'échantillon en comporte 651. Un extrait du jeu de données est fourni à la Figure 3.5. Les variables considérées sont décrites dans le tableau suivant. Pour la bonne compréhension du propos, indiquons qu'aux Etats-Unis, deux sites jouent le rôle du site français Allociné, à savoir : Rotten Tomatoes (RT, qui utilise une échelle de notation sur 100) et IMDb (notation sur 10).

`Titre`	Titre du film
`Genre`	Genre cinématographique (drame, comédie, horreur, etc.)
`Année`	Année de sortie en salles
`Durée`	Durée en minutes
`Critiques_RT`	Note moyenne (sur 100) des critiques professionnelles collectées par RT
`Spec_RT`	Note moyenne (sur 100) des critiques spectateurs collectées par RT
`Note_IMDb`	Note moyenne (sur 10) des critiques spectateurs collectées par IMDb
`Spec_IMDb`	= `Note_IMDb` $\times 10$
`Diff1`	= `Spec_RT` − `Critiques_RT`
`Diff2`	= `Spec_RT` − `Spec_IMDb`

Les questions que nous allons nous poser sont les suivantes :
- Critiques et spectateurs de RT notent-ils de manière identique en moyenne ?
- Les spectateurs de RT et d'IMDb notent-ils de manière identique en moyenne ?

On considère donc ici des couples de variables d'intérêt. Le premier couple est formé par `Critiques_RT` et `Spec_RT`. On espère que l'échantillon de 651 films soit représentatif des milliers de films référencés par RT et IMDb, qui constituent la population. Il s'agit donc de comparer ici d'une part les moyennes de population μ_{crit} et μ_{RT} de ces deux variables. D'autre part, le second couple de variables est formé par `Spec_RT` et `Spec_IMDb`, de moyennes de population notées μ_{RT} et μ_{IMDb}.

Déterminons les hypothèses à tester. A cet effet, interrogeons-nous sur le contexte. Nous pouvons imaginer que les critiques ont des critères d'évaluations différents du grand public, ce qui peut à la fois défavoriser les blockbusters, que le grand public peut plébisciter, et favoriser des films d'auteurs qui sembleront obscurs au grand public. Certains autres films seront quant à eux conjointement encensés ou massacrés. Que se passera-t-il en moyenne ? C'est difficile à dire, et c'est pourquoi nous recourrons

à des hypothèses bilatères[1], en prenant $H_1 : \mu_{RT} \neq \mu_{crit}$.

De même, il est naturel de penser que les utilisateurs de deux sites différents représentent malgré tout de manière identique les goûts de l'ensemble de la population américaine, et le fait que leurs avis divergent en moyenne susciterait réflexion. Nous adoptons une approche neutre, car nous ne savons pas à l'avance qui, le cas échéant, noterait mieux en moyenne parmi les spectateurs de RT et d'IMDb, d'où encore des hypothèses bilatères, avec $H_1 : \mu_{RT} \neq \mu_{IMDb}$.

En résumé, nous allons procéder aux tests

$$H_0 : \mu_{RT} = \mu_{crit} \quad \text{vs.} \quad H_1 : \mu_{RT} \neq \mu_{crit} \qquad \text{et} \qquad H_0 : \mu_{RT} = \mu_{IMDb} \quad \text{vs.} \quad H_1 : \mu_{RT} \neq \mu_{IMDb}\,.$$

1. Concept de données appariées

Pour chaque test à mener et pour chaque individu statistique, nous disposons de deux mesures : par exemple, pour le film Godzilla (individu numéro 2), `Spec_RT` vaut $x_2 = 66$ et `Critiques_RT`, $y_2 = 74$.

Nous disposons donc plus globalement de deux séries de données, les évaluations moyennes des spectateurs de RT, notées $x_1, x_2, \ldots, x_{651}$, et les évaluations moyennes des critiques de RT, $y_1, y_2, \ldots, y_{651}$, se correspondant deux à deux. I.e., pour chaque film $j \in \{1, 2, \ldots, 651\}$, les évaluations correspondent au couple (x_j, y_j). C'est pourquoi l'on parle de données appariées[2].

Méthodologie de traitement : introduction. Ici, on a affaire à deux moyennes de populations inconnues, μ_{RT} et μ_{crit}. On ne peut pas se ramener à effectuer un test de comparaison à une moyenne de référence donnée par une moyenne empirique $\overline{x}_{651}$ ou $\overline{y}_{651}$, car d'une part ce serait choisir les hypothèses au vu des données (ce qui est interdit), et d'autre part, on négligerait ainsi une des deux variabilités entre celle relative à l'estimation de μ_{RT} et celle relative à μ_{crit}.

En revanche, on peut s'intéresser plutôt à la différence de population $\Delta_0 = \mu_{RT} - \mu_{crit}$, qui correspond à la moyenne, sur la population de films, des différences `Diff1` $=$ `Spec_RT` $-$ `Critiques_RT`. Les données intéressantes pour raisonner sur cette différence de population Δ_0 sont les différences $z_1, \ldots, z_{651}$ observées sur l'échantillon et définies comme $z_j = x_j - y_j$.

Pour chaque j, les deux mesures x_j et y_j sont très liées, car elles portent sur le même film ; dit autrement, les variables aléatoires X_j et Y_j dont elles sont issues ne sont absolument pas indépendantes. Dans la Figure 3.5, cela revient à dire que les valeurs lues dans les deux colonnes `Spec_RT` et `Critiques_RT` ne sont pas indépendantes.

En revanche, on va supposer que les variables aléatoires $Z_1 = X_1 - Y_1, \quad \ldots, \quad Z_{651} = X_{651} - y_{651}$ sont indépendantes et identiquement distribuées selon une loi d'espérance Δ_0. Cela revient à dire que les écarts d'évaluations entre critiques et spectateurs sont indépendants entre films, et de même loi d'espérance Δ_0. En effet, critiques et spectateurs possèdent des grilles d'évaluations constantes même si possiblement différentes (d'où l'identique distribution des écarts), qu'ils appliquent de manière intrinsèque (et non par comparaison) à des films d'un échantillon choisi au hasard parmi un grand ensemble de films (d'où le caractère indépendant entre films des écarts entre les deux évaluations). Dit autrement, dans la Figure 3.5, en ce qui concerne la colonne `Diff1`, les valeurs lues dans les lignes

1. On rappelle au passage que H_0 consiste en le cas d'égalité, pour tous les tests que nous verrons dans ce cours et au prix éventuel d'un passage au cas limite. Votre mission est de justifier la forme de H_1.

2. En référence à des paires (des couples) ; et non, comme je le lis parfois dans des copies, des données appareillées (parce qu'elles entendraient mal ?), ou pire, des données avariées (pas assez fraîches, pas assez récentes ?).

sont indépendantes entre elles.

Méthodologie de traitement : conclusion. Nous considérons les différences $z_j = x_j - y_j$, issues de variables aléatoires $Z_1, \dots, Z_{651}$ indépendantes et identifiquement distribuées selon une loi d'espérance $\Delta_0 = \mu_{RT} - \mu_{crit}$, et

$$\text{pour tester } H_0 : \mu_{RT} = \mu_{crit} \text{ vs. } H_1 : \mu_{RT} \neq \mu_{crit}, \qquad \text{nous testons } H_0 : \Delta_0 = 0 \text{ vs. } H_1 : \Delta_0 \neq 0.$$

Nous nous sommes ramenés à appliquer un test de comparaison à la valeur de référence $\mu_{ref} = 0$ sur la série de données des différences $z_1, \dots, z_{651}$. A cet effet, nous considérons donc la statistique de test

$$t_{651} = \sqrt{651}\, \frac{\bar{z}_{651}}{s_{z,651}}.$$

2. Mises en œuvre sur l'exemple des films

Nous résumons les données disponibles par les indicateurs statistiques suivants.

		Critiques_RT	Spec_RT	Spec_IMDB	Diff1	Diff2
N		651	651	651	651	651
Moyenne		57,7	62,4	64,9	4,7	-2,6
Médiane		61	65	66	2	0
Ecart type		28,4	20,2	10,8	20,2	12,1
Minimum		1	11	19	-49	-36
Maximum		100	97	90	74	32
Percentiles	25	33	46	59	-10	-12
	50	61	65	66	2	0
	75	83	80	73	17	8

Comparaison de μ_{RT} et μ_{crit} : calculs. Nous nous intéressons à la variable `Diff1`. Les données de différences peuvent être résumées par $\bar{z}_{651} = 4.7$ et $s_{z,651} = 20.2$, d'où

$$t_{651} = \sqrt{651}\, \frac{4.7}{20.2} \approx 5.9.$$

La P-valeur se calcule comme sur la Figure 3.2 page 37 et est donnée par $2\,\mathbb{P}(Z \geqslant 5.9)$; la table de la loi normale montre qu'elle est quasi-nulle. On rejette fermement H_0 et on en déduit $\Delta_0 \neq 0$, en l'occurrence, au vu des valeurs d'échantillon, que $\Delta_0 > 0$. On reformule cette conclusion technique à destination du grand public :

> Ces données montrent que les critiques et les spectateurs du site RT n'évaluent pas les films de la même manière, en moyenne ; en l'occurrence, les spectateurs écrivent des évaluations en moyenne plus élevées.

Nous venons de démontrer un effet (les évaluations moyennes μ_{RT} et μ_{crit} sont différentes), et de l'assortir d'une direction d'effet (en l'occurrence, $\mu_{RT} > \mu_{crit}$). Nous pouvons également procurer une taille d'effet, i.e., une quantification de l'écart $\Delta_0 = \mu_{RT} - \mu_{crit}$. Cela est possible par intervalle de confiance, à partir des données $z_1, \text{ldots}, z_{651}$. Au test bilatère correspond un intervalle de confiance également bilatère : avec grande confiance, les spectateurs de RT mettent en moyenne des évaluations supérieures à celles des critiques d'un facteur additif de

$$\bar{z}_{651} \pm 1.96\, \frac{s_{z,651}}{\sqrt{651}} = 4.7 \pm 1.96\, \frac{20.2}{\sqrt{651}} = 4.7 \pm 1.5517... \approx 4.7 \pm 1.6 = [3.1,\, 6.3].$$

Comparaison de μ_{RT} et μ_{crit} : lecture d'une sortie logiciel. Nous pouvons vérifier nos calculs avec la sortie de logiciel suivante :

		Moyenne	N	Ecart type	Moyenne d'erreur standard
Paire 1	Spec_RT	62,36	651	20,223	,793
	Critiques_RT	57,69	651	28,403	1,113

Test des échantillons appariés

		Différences appariées							Signification	
				Moyenne d'erreur standard	Intervalle de confiance de la différence à 95 %					
		Moyenne	Ecart type		Inférieur	Supérieur	t	df	p unilatéral	p bilatéral
Paire 1	Spec_RT - Critiques_RT	4,674	20,165	,790	3,122	6,226	5,914	650	<,001	<,001

En particulier, nous retrouvons, dans le second tableau, la valeur réalisée $t \approx 5.9$ de la statistique de test, la P–valeur bilatère quasi-nulle, et la taille d'effet $[3.1, 6.3]$ après arrondis.

Pour être exhaustif, voici les formules de toutes les quantités calculées :

	Moyenne	N	Ecart-type	Erreur standard
V1	$\overline{x}_n$	n	$s_{x,n}$	$s_{x,n}/\sqrt{n}$
V2	$\overline{y}_n$	n	$s_{y,n}$	$s_{y,n}/\sqrt{n}$

		Ecart-	Erreur	Int. confiance				Signification	
	Moy.	type	standard	Inférieur	Supérieur	t	df	p unilat.	p bilat.
V1 - V2	$\overline{z}_n$	$s_{z,n}$	$s_{z,n}/\sqrt{n}$	$\overline{z}_n \pm 1.96\,(s_{z,n}/\sqrt{n})$		$\sqrt{n}(\overline{z}_n/s_{z,n})$	$n-1$	$\mathbb{P}(Z > \lvert t\rvert)$	$2\,\mathbb{P}(Z > \lvert t\rvert)$

La seule case qui devrait vous sembler mystérieuse est celle intitulée "df", qui signifie "degrees of freedom" (« degrés de liberté ») : en réalité, ce logiciel statistique fonde ses calculs de P–valeur non pas sur la loi normale, mais sur la loi dite de Student, avec $n-1$ degrés de liberté, ce qui donne lieu à des P–valeurs légèrement plus grandes. Voir l'appendice (totalement facultatif) de ce chapitre pour plus de détails.

Comparaison de μ_{RT} et μ_{IMDb}. Pour μ_{RT} et μ_{IMDb}, on reprend la démarche ci-dessus, en injectant de nouvelles valeurs numériques, correspondant à la colonne `Diff2` du tableau de résumés statistiques ; à savoir $\overline{z}_{651} = -2.6$ et $s_{z,651} = 12.1$. Cela conduit à $t_{651} = \sqrt{651}(-2.6/12.1) \approx 5.4$ et donc à une P–valeur elle aussi quasi-nulle. Nous confirmons ces calculs avec la sortie de logiciel suivante, qui procure également une quantification de la taille d'effet en l'intervalle de confiance $[-3.6, -1.6]$ après arrondis :

		Moyenne	N	Ecart type	Moyenne d'erreur standard
Paire 1	Spec_RT	62,36	651	20,223	,793
	Spec_IMDB	64,93	651	10,847	,425

Test des échantillons appariés

		Différences appariées							Signification	
				Moyenne d'erreur standard	Intervalle de confiance de la différence à 95 %					
		Moyenne	Ecart type		Inférieur	Supérieur	t	df	p unilatéral	p bilatéral
Paire 1	Spec_RT - Spec_IMDB	-2,568	12,132	,475	-3,502	-1,635	-5,402	650	<,001	<,001

Nous pouvons formuler la conclusion grand public suivante :

> Ces données montrent que les spectateurs du site RT et ceux d'IMDb n'évaluent pas les films de la même manière, en moyenne ; en l'occurrence, une fois les évaluations ramenées sur la même échelle sur 100, les spectateurs de RT écrivent des évaluations en moyenne moins élevées : avec grande confiance, de 2.6 points, avec une marge d'erreur de $\pm$ 1 point.

3. Autres exemples de données appariées

Voici quelques exemples supplémentaires de situations de données appariées, pour en approfondir le concept.

Exemple (Prix de produits). Un exemple tout à fait typique est donné par les prix de différents produits ou services (qui forment les individus) dans deux magasins ou chez deux fournisseurs donnés. Bien évidemment, le prix chez un fournisseur donne une idée du prix chez l'autre fournisseur (ces deux prix sont liés), mais l'on suppose que les différences de prix entre produits sont indépendantes. Je propose deux exercices sur ce sujet à la fin de ce chapitre.

Exemple (Double correction). La technique la plus juste pour comparer la manière de noter de deux examinateurs est la double correction, lorsque chacun corrige un même ensemble de copies (échantillon) : pour chaque copie (individu statistique), deux mesures (notations) sont effectuées, très liées.

Exemple (Etudes cliniques). Le rêve, dans des études cliniques, est de pouvoir considérer des données appariées, car cela permet de mieux supprimer la variabilité due aux sujets. En cosmétique, c'est assez facile : pour tester la différence d'hydratation apportée par deux crèmes, on peut en appliquer une sur la main droite et l'autre sur la main gauche. Cela peut valoir aussi pour les somnifères : on teste un premier somnifère pendant une semaine, on attend plusieurs semaines, et on teste le second. Hélas, en général, et notamment pour les médicaments visant à traiter des maladies, on ne peut pas agir ainsi. Si un sujet prend un premier médicament, il commencera à guérir, au moins un peu, et on ne pourra pas tester le second médicament dans les mêmes conditions et comparer les effets des deux médicaments toutes choses égales par ailleurs. Dans ces cas-là, on doit hélas procéder en comparant deux échantillons indépendants (voir chapitre suivant), ce qui crée plus de variabilité puisque les deux groupes de sujets ne sont plus identiques.

Exemple (Salaires hommes–femmes). Dans cet exemple, on considère comme individus statistiques des couples hétérosexuels, sur lesquels on effectue deux mesures, liées : le salaire de l'homme et celui de la femme. Ces mesures sont liées à cause de plusieurs phénomènes : une tendance à l'endogamie matrimoniale, i.e., une tendance à se marier avec un partenaire de son milieu, ce qui implique que si l'un des membres du couple a un salaire élevé, l'autre aussi (corrélation) ; et également, le fait que lorsqu'un des conjoints perçoit un salaire vraiment élevé, l'autre conjoint peut ne pas travailler, par exemple pour s'occuper des enfants, et avoir un salaire nul (anti-corrélation). Bref, il existe des dépendances complexes entre les deux salaires. Toutefois, si l'objet de l'étude est constitué par les différences moyennes de salaires au sein des couples hétérosexuels, et s'il s'agit de voir si les hommes gagnent en moyenne significativement plus que les femmes au sein de ces couples, c'est assez simple : on sonde un certain nombre n de tels couples en les choisissant uniformément au hasard dans la population active, et on fonde l'étude sur les différences de salaires $z_1, \dots, z_n$ observées au sein de chaque couple. Ces différences sont bien, vu la méthode d'échantillonnage, les réalisations de variables aléatoires $Z_1, \dots, Z_n$ indépendantes et identiquement distribuées selon une loi admettant pour espérance $\Delta_0 = \mu_H - \mu_F$, où μ_H et μ_F désignent les salaires moyens des hommes et des femmes en couple hétérosexuel. Ici, on ne compare pas du tout les salaires moyens des femmes et des hommes au sein de la population

active, car ces derniers tiendraient également compte des célibataires et intégreraient les membres des couples dont le conjoint est retraité (alors que notre étude sur données appariées repose sur des couples d'actifs).

Résumé. On fait face à des données appariées lorsque pour les mêmes individus statistiques, on effectue deux mesures

— de la même variable mais dans des conditions différentes (par exemple, prix de produits dans deux magasins ; notes données sur les mêmes copies par deux correcteurs ; salaires en 2017 et 2023 pour les mêmes sondés ; montants de commandes avec et sans promotion sur deux trimestres successifs pour les mêmes clients) ;

— au sein d'une « paire », chaque paire étant considérée comme un individu statistique (salaires au sein d'un couple hétérosexuel ; tailles de duos père–fils).

Visuellement, les données appariées se présentent dans un tableur sous la forme suivante :

ID	Variable 1	Variable 2	Différence
1	16	12	4
2	18	16	2
3	30	29	1
4	29	32	-3
...	...	...	

La colonne ID permet de numéroter les individus statistiques, et pour chacun d'eux, deux mesures sont effectuées (`Variable 1` et `Variable 2`), et on se ramène à considérer les écarts entre ces deux mesures (`Différence`).

4. Appendice facultatif : Loi de Student

Vous aimerez, j'en suis sûr, le contexte dans lequel la loi de Student 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 et que l'on utilise pour calculer les P–valeurs pour les tailles d'échantillons faibles, $n \geqslant 30$, et lorsque l'on a affaire à des données gaussiennes.

Mais comme pour $n \geqslant 30$, la loi de Student à $n-1$ degrés de liberté est très proche de la loi normale standard $\mathcal{N}(0,1)$, calculer les P–valeurs avec la loi de Student (comme les sorties logiciel précédentes) ou avec la loi $\mathcal{N}(0,1)$ ne change presque rien, même si, avec la loi de Student, on obtient des P–valeurs légèrement supérieures. Par vœu d'universalité et de simplicité, les logiciels statistiques utilisent alors la loi de Student en toute occasion, que la taille d'échantillon n soit grande ou petite.

Construction de la loi de Student. Lorsque les observations $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées selon une loi normale de paramètres μ_0 et σ_0^2, alors $\overline{X}_n \sim \mathcal{N}(\mu_0, \sigma_0^2/n)$ et donc

$$\frac{\sqrt{n}}{\sigma_0}\,(\overline{X}_n - \mu_0) \;\sim\; \mathcal{N}(0,1)\,.$$

A gauche, il s'agit de la même variable aléatoire que celle considérée au théorème limite central. Pour obtenir une statistique de test pour la comparaison à une valeur de référence μ_{ref} (dans ce chapitre, $\mu_{\text{ref}} = 0$), il a fallu estimer σ_0. C'est précisément ce qui conduit à la loi de Student. Dans la définition–théorème suivante, on note S_n la quantité qui correspond à la statistique de test T_n, au remplacement de μ_0 par μ_{ref} près, que l'on peut effectuer sous $H_0 : \mu_0 = \mu_{\text{ref}}$.

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

$$S_n = \frac{\sqrt{n}}{\widehat{\sigma}_n} (\overline{X}_n - \mu_0)$$

est indépendante de μ_0 et σ_0^2 ; on l'appelle la loi de Student à $n-1$ degrés de liberté, et on la note $\mathcal{T}_{n-1}$.

Démonstration. Il s'agit de voir pourquoi la loi de la variable aléatoire considérée ne dépend ni de μ_0 ni de σ_0^2. On introduit les variables aléatoires (non observées, mais cela n'a pas d'importance)

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

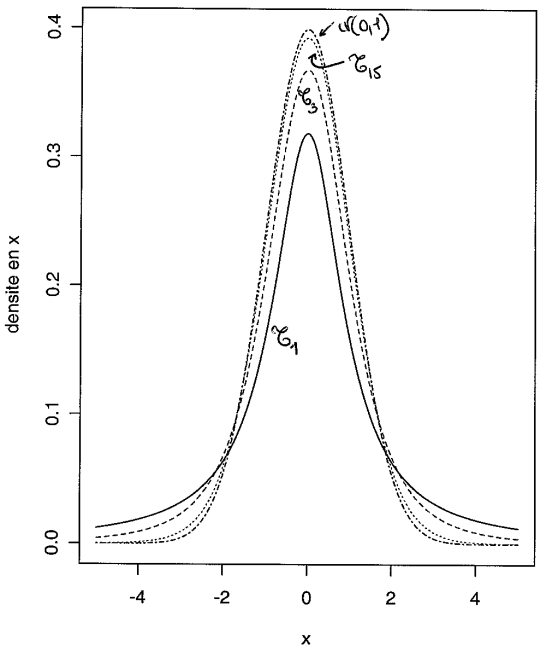
elles sont indépendantes et identiquement distribuées selon la loi $\mathcal{N}(0,1)$. On note $\overline{X'}_n$ et $\widehat{\sigma'}_n^2$ les estimateurs de la moyenne et de la variance empiriques construits sur les X'_j. Par un calcul immédiat, on montre que

$$\frac{\sqrt{n}}{\widehat{\sigma}_n} (\overline{X}_n - \mu_0) = \frac{\sqrt{n}}{\widehat{\sigma}'_n} \overline{X'}_n ,$$

ce qui conclut la preuve. □

Remarque 4.1 (Convergence). Par théorème limite central et lemme de Slutzky, on a bien que $\mathcal{T}_n$ tend vers $\mathcal{N}(0,1)$ lorsque $n \to \infty$.

Densités et queues de distribution. La figure ci-contre présente des tracés des fonctions densité de lois de Student, de différents degrés de liberté, ainsi que leur limite, la loi normale standard. Je n'y ai pas tracé $\mathcal{T}_{29}$, qui est déjà quasiment confondue avec la loi $\mathcal{N}(0,1)$.



On voit que les queues de distribution sont plus épaisses avec la loi de Student ; donc, pour $n \geqslant 30$, les P–valeurs calculées avec une loi de Student $\mathcal{T}_{n-1}$ sont plus grandes, mais à peine plus grandes, que lorsqu'elles sont calculées avec la loi $\mathcal{N}(0,1)$.

Comparaison de prix : Franprix versus Monoprix

Cet exercice porte sur des données réelles, recueillies [1] par Benjamin Petiau (enseignant passé de ce cours) 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`. 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(`Franprix`/`Monoprix`).

		Franprix	Monoprix	Différence	LnRapport
1	Vodka Polialov (1L)	17,63	16,74	,89	,05
2	Chocolat Lindt 70%	2,73	2,47	,26	,10
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 Tropica	3,03	2,86	,17	,06

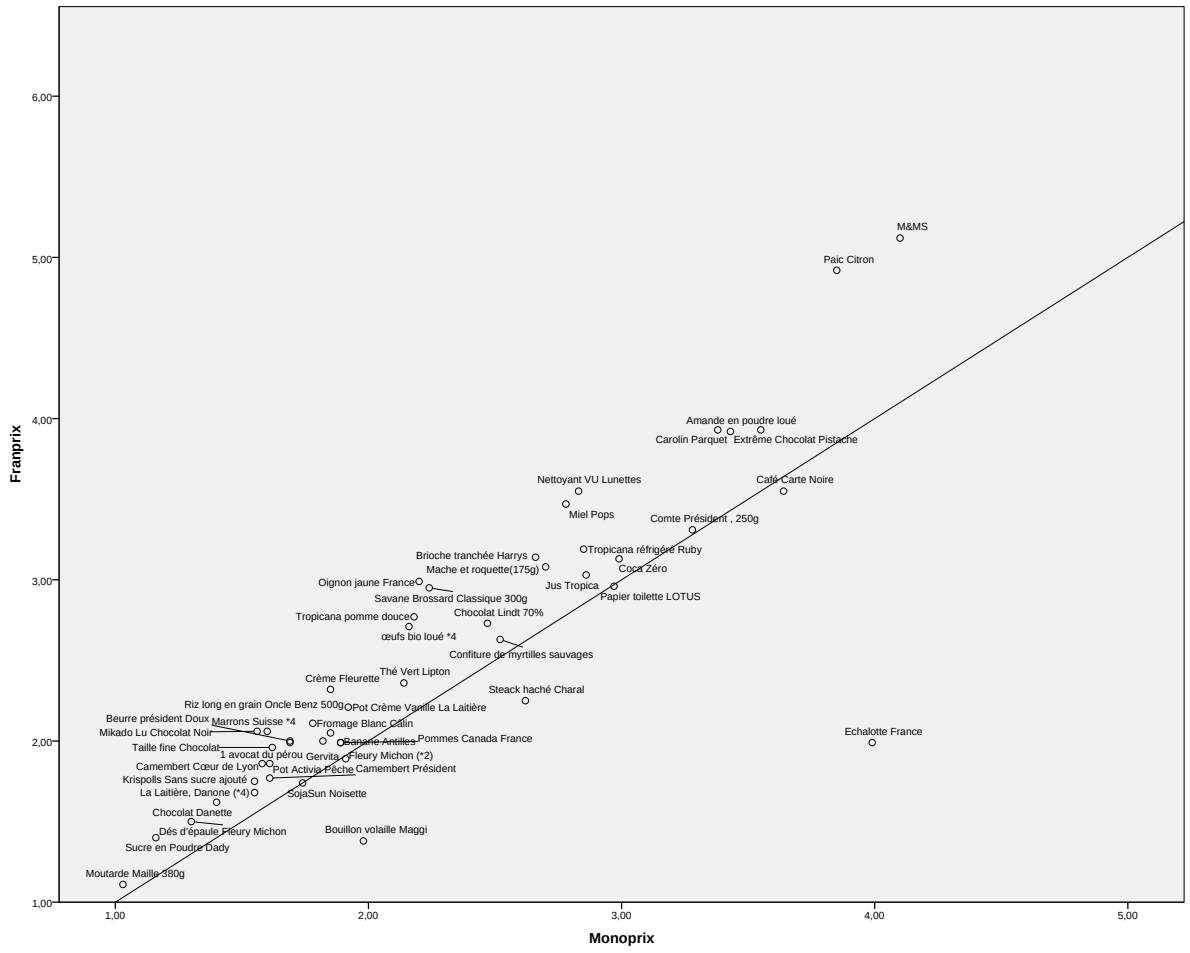


Diagramme de dispersion des données (zoom sur les plus petits prix)

1. Une version moderne de cet exercice pourrrait être construite à partir de `https://www.quiestlemoinscher.leclerc/`, avec au préalable une étude critique de la méthodologie qui y est employée, à savoir, utiliser les prix collectés sur Internet, donc correspondant au « drive », pour les correspondants vs. les prix en magasins pour E.Leclerc. Cette enquête reporte des données d'échantillon brutes, et non des estimations par intervalles ou des P–valeurs.

1. De quel type de données s'agit-il ? Répondre en un adjectif, sans justification :

2. Imaginez la question que Benjamin Petiau avait en tête avant d'effectuer le recueil des données et déduisez-en le(s) paramètre(s) d'intérêt, ainsi que le jeu d'hypothèses testées ? Justifiez vos réponses.

Nous considérons maintenant les sorties logiciel suivantes.

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					
					Intervalle de confiance de la différence à 95 %	
	t	ddl	Sig. (bilatéral)	Différence moyenne	Inférieur	Supérieur
Différence	2,772	50	,008	,20745	,0572	,3577
LnRapport	3,765	50	,000	,09012	,0420	,1382

Nous commençons par nous intéresser à une évaluation « additive » des écarts de prix potentiels.

3. Effectuez à la main le test des hypothèses correspondantes et procurez une P–valeur.

4. Formez une conclusion statistique lisible par le grand public et indiquant, le cas échéant, une direction et une taille d'effet.

5. A la maison : traitez le cas d'une évaluation « multiplicative » des écarts de prix potentiels (i.e., telle enseigne est $x\%$ plus chère que l'autre).

Augmentation des prix suite au passage à l'euro — 9 points

En l'honneur des 20 ans du passage à l'euro, divers médias ont parlé de l'augmentation moyenne des prix à la consommation sur 20 ans (par exemple, JT 13h France 2 le 21 février 2022, La Dépêche du Midi le 1er janvier 2022 citant l'INSEE, etc.). Certains chiffrent cette augmentation à 30%, d'autres à 26%, etc. : nous allons nous demander si l'on peut montrer, avec une étude simple fondée sur un jeu de données petit avec seulement 50 références, que cette augmentation est bien d'au moins 25%.

Nous supposons qu'un statisticien prévoyant a relevé des prix en janvier 2002 sur des produits de grande consommation encore disponibles en janvier 2022, date à laquelle il complète son jeu de données. Un *extrait* de ce dernier est fourni ci-dessous ; les première et troisième colonnes correspondent à des prix relevés sur le terrain, tandis que les deuxième et quatrième colonnes ont été calculées.

Produits	Prix 2002	Prix 2002 + 25%	Prix 2022	Prix 2022 - (Prix 2002 + 25%)
Lait	1,62	2,03	1,99	-0,04
6 œufs	1,95	2,44	2,99	0,55
Beurre	2,45	3,06	2,99	-0,07
Yaourt vanille	1,65	2,06	1,79	-0,27
Crème fraîche	0,89	1,11	2,65	1,54

☐ Quel couple d'hypothèses faut-il considérer au vu du contexte ?

- ◯ H_0 : augmentation de plus de 25% contre H_1 : augmentation de moins de 25%
- ◯ H_0 : augmentation de 25% contre H_1 : augmentation de plus ou de moins de 25%
- ◯ H_0 : augmentation de moins de 25% contre H_1 : augmentation de plus de 25%
- ◯ H_0 : augmentation de plus ou de moins de 25% contre H_1 : augmentation de 25%

☐ Justifiez brièvement ce choix :

Les données peuvent être résumées par les statistiques descriptives ci-dessous.

☐ De quel type de données s'agit-il ? Répondre en un adjectif :

☐ Entourez ci-dessous les statistiques descriptives qui seront utiles pour mener le test, et seulement celles-ci.

	N	Minimum	Maximum	Moyenne	Ecart type
Prix 2002	50	,30	6,99	2,3930	1,49651
Prix 2002 + 25%	50	,375	8,738	2,99125	1,870641
Prix 2022	50	1,35	9,50	3,2645	1,86570
Prix 2022 - (Prix 2002 + 25%)	50	-3,15	3,85	,2733	1,06395

☐ Introduisez un paramètre d'intérêt correspondant aux deux réponses précédentes et reformulez les hypothèses en fonction de lui.

Mettez maintenant en œuvre le test d'hypothèses

☐☐☐ en procurant un schéma récapitulant les comportements de la statistique de test adéquate sous H_0 et H_1, en calculant la valeur réalisée de cette statistique de test (*procurez les détails du calcul, c'est-à-dire : écrivez le calcul ce que vous tapez sur votre calculatrice*), et en indiquant la P-valeur obtenue.

☐ Concluez—Le test d'hypothèses mené ci-dessus a permis de montrer que l'augmentation des prix a été d'au moins 25% sur la période 2002–2022 :

◯ Oui
◯ Non (l'échantillon est trop petit)
◯ Non (l'augmentation est inférieure à 25%)
◯ Autre réponse (préciser)

Plombémie, des parents vers les enfants (10 points)

La plombémie est le taux de plomb dans le sang. Une étude américaine, à savoir

> Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. and Saah, M. : Lead absorption in children of employees in a lead related industry. *American Journal of Epimediology*, 115, pages 549–55, 1982.

a comparé les plombémies (mesurées en mg/dL) d'enfants [« Exposés »] dont un parent a travaillé dans des usines de production de batteries utilisant du plomb à celles d'enfants de leur voisinage [« Contrôle »], dont aucun parent n'est en contact professionnel avec du plomb. Les couples d'enfants ainsi constitués l'ont été en particulier en considération de l'âge, de l'exposition au trafic automobile, et plus généralement, de l'environnement, afin de pouvoir raisonner toutes choses égales par ailleurs. Les données collectées sont reproduites et tracées à la page suivante.

☐ De quel type de données s'agit-il ?

☐ Quel est le paramètre d'intérêt ? (Donner une réponse en lien avec le type de données identifié précédemment.)

☐☐ Choisir un jeu d'hypothèses pertinent au vu du contexte et des objectifs implicites de l'étude : à formuler d'abord en mots, puis en équations. Justifier ce choix.

☐ Dans le tableau ci-contre, entourer les données pertinentes pour réaliser ce test (et uniquement celles-là).

	N	Minimum	Maximum	Moyenne	Ecart type
Exposés	33	10	73	31,85	14,407
Contrôle	33	7	25	15,88	4,540
Différence	33	-9	60	15,97	15,864
N valide (liste)	33				

☐☐☐ Préciser une statistique de test adaptée, ses comportements sous H_0 et sous H_1, et calculer sa valeur réalisée. Tracer un schéma récapitulatif de ces éléments, et calculer une P-valeur.

☐ Former une conclusion statistique compréhensible par le grand public et commençant par « Ces données... ». Ajouter si possible une conclusion stratégique.

ID	Contrôle	Exposés	Différence
1	16	38	22
2	18	23	5
3	18	41	23
4	24	18	-6
5	19	37	18
6	11	36	25
7	10	23	13
8	15	62	47
9	16	31	15
10	18	34	16
11	18	24	6
12	13	14	1
13	19	21	2
14	10	17	7
15	16	16	0
16	16	20	4
17	24	15	-9
18	13	10	-3
19	9	45	36
20	14	39	25
21	21	22	1
22	19	35	16
23	7	49	42
24	18	48	30
25	19	44	25
26	12	35	23
27	11	43	32
28	22	39	17
29	25	34	9
30	16	13	-3
31	13	73	60
32	11	25	14
33	13	27	14

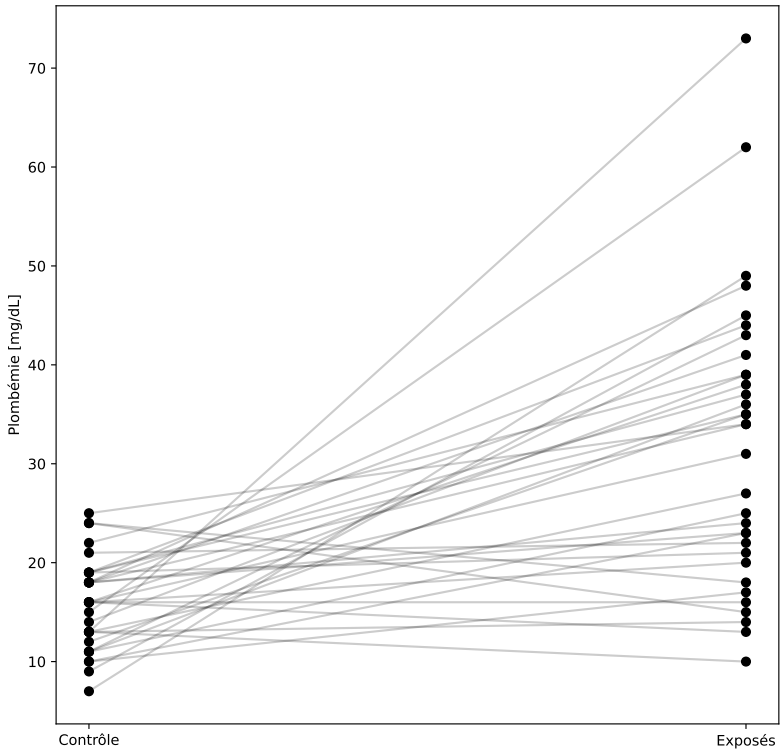


FIGURE 1 – Le jeu de données considéré (à gauche) et une représentation graphique des données (à droite).

☐ Vérifier vos résultats avec la sortie SPSS ci-dessous. Procurer tout commentaire utile.

		Moyenne	N	Ecart type	Moyenne erreur standard
Paire 1	Exposés	31,85	33	14,407	2,508
	Contrôle	15,88	33	4,540	,790

Test des échantillons appariés

		Différences appariées					t	ddl	Sig. (bilatéral)
		Moyenne	Ecart type	Moyenne erreur standard	Intervalle de confiance de la différence à 95 % Inférieur	Supérieur			
Paire 1	Exposés - Contrôle	15,970	15,864	2,762	10,345	21,595	5,783	32	,000

5

Comparaison de moyennes ou proportions indépendantes

Nous allons considérer dans ce chapitre deux échantillons indépendants, i.e., constitués indépendamment au hasard dans deux populations distinctes (ou à tout le moins, représentatifs de deux populations distinctes). Nous étudierons par exemple la répartition filles–garçons [F–G] dans la fréquentation des structures de loisirs encadrés en fonction de l'âge, et disposerons d'un premier échantillon d'enfants scolarisés dans le primaire, et d'un second échantillon de collégiens. Un second exemple portera sur la comparaison des échelles de notations de deux professeurs : un premier échantillon sera constitué avec des étudiants notés par le premier professeur [1] puis comparé à un second échantillon d'étudiants notés, sur le même sujet, par le second professeur [2].

Visuellement, des deux exemples se présentent ainsi dans un tableur : pour chaque individu statistique, on reporte son groupe d'appartenance et l'observation mesurée pour la variable d'intérêt. (Comparez ces exemples à la manière dont se présentent dans un tableau des données appariées, voir page 54.)

ID	Groupe	Sexe
1	Primaire	F
2	Collège	G
3	Primaire	G
4	Primaire	F
5	Collège	G
...	...	...

ID	Prof.	Note
1	1	12
2	2	16
3	1	9
4	1	15
5	2	8
...	...	...

1. Comparaison de proportions indépendantes

Nous allons nous référer de nouveau à l'article mentionné au paragraphe 2 page 34 : il s'intéresse également à la fréquentation en termes de genre des structures de loisirs en fonction de l'âge, parmi les populations des enfants du primaire et des adolescents du collège. Ces structures accueillent les enfants et les adolescents les mercredis après-midis et pendant les vacances scolaires. (Dans ce qui suit, nous avons repris les effectifs décrits dans l'article en les divisant par un facteur 10.)

Les chercheurs n'avaient aucune idée pré-conçue quant à une possible évolution de la répartition filles–garçons entre les activités pour enfants et celles pour adolescents (ni sur sa direction éventuelle), et voulaient tester l'existence même d'une telle évolution qui, si elle avait lieu, devrait susciter réflexion.

Nous prendrons par conséquent pour paramètres d'intérêt les proportions de filles p_0 et d'adolescentes q_0 participant aux activités proposées, et testerons $H_0 : p_0 = q_0$, i.e., la répartition filles–garçons n'évolue pas de l'enfance à l'adolescence, contre $H_1 : p_0 \neq q_0$, i.e., cette répartition évolue, dans un sens (par diminution de la proportion de filles) ou dans l'autre (par augmentation de la proportion de filles).

Remarque 5.1 (deux paramètres inconnus ; aucune valeur de référence). Il s'agit bien ici de comparer deux proportions de filles p_0 q_0 entre elles ; elles sont toutes deux inconnues. Il ne s'agit pas du tout de voir si d'une part, la répartition filles-garçons est paritaire chez les enfants dans les structures de loisirs, et d'autre part, si elle l'est chez les adolescents, ce qui constituerait deux tests de comparaison à la proportion de référence p_{ref}.

Les données collectées dans les trois villes de l'agglomération bordelaise sont que 394 garçons, 364 filles, soit 758 enfants, et 518 adolescents garçons, 293 adolescentes, soit 811 adolescents des deux sexes, participent à ces activités. Nous notons ces données $x_1, \ldots, x_{758}$ pour le premier échantillon, où $x_j = 1$ si le j–ème enfant est une fille et $x_j = 0$ si c'est un garçon, et $y_1, \ldots, y_{811}$ pour le second échantillon, où $y_k = 1$ si la k–ème personne adolescents est de sexe féminim.

Nous les résumons dans le tableau 2×2 suivant—le fait, d'ailleurs, de pouvoir résumer des données dans un tableau 2×2 (appelé "contigency table" en anglais), est le signe que nous avons affaire à une comparaison de proportions indépendantes. Nous supposons par ailleurs que les données collectées sont représentatives de l'ensemble de la fréquentation des structures de loisirs en France.

	Enfants	Ados	Total
Filles	364	293	657
Garçons	394	518	912
Total	758	811	1 569

Nous pouvons donc reformuler la question posée ainsi :

> Les proportions de filles $\overline{x}_{758} = 364/758 \approx 48.0\%$ et d'adolescentes $\overline{y}_{811} = 293/811 \approx 36.1\%$ sont-elles significativement différentes ?

Statistique de test. Il est intuitif de comparer $\overline{x}_{758}$ et $\overline{y}_{811}$ et donc de construire une statistique de test admettant pour numérateur $\overline{x}_{758} - \overline{y}_{811}$:

$$t = \frac{\overline{x}_{758} - \overline{y}_{811}}{?}.$$

Il s'agit alors de pouvoir normaliser cette différence pour que d'une part, sous $H_0 : p_0 = q_0$, elle suive un comportement connu et « universel », au sens où il ne dépend pas de la valeur commune $p_0 = q_0$, et que d'autre part, sous H_1, cette différence normalisée ait un comportement fort différent. L'appendice de lecture facultatif explique qu'en l'occurrence, on estime la valeur commune $p_0 = q_0$ sous H_0 en rassemblant les $758 + 811 = 1\,569$ observations x_j et y_k dans une unique série de données $z_1, \ldots, z_{1569}$. Alors,

$$\overline{z}_{1569} = \frac{x_1 + \ldots + x_{758} + y_1 + \ldots + y_{811}}{758 + 811} = \frac{364 + 293}{758 + 811} = \frac{657}{1\,569}$$

estime la proportion globale de filles parmi les enfants et les adolescentes. Il est à noter que les calculs ci-dessus s'appuient sur le tableau 2×2, qui permet de lire directement les effectifs semi-totaux. A

défaut de proportion de référence p_{ref}, on normalise ici par l'écart-type associé à $\overline{z}_{1569}$:

$$t = \sqrt{\dots}\,\frac{\overline{x}_{758} - \overline{y}_{811}}{\sqrt{\overline{z}_{1569}(1-\overline{z}_{1569})}}\,.$$

Il reste à tenir compte des tailles d'échantillon $n = 758$ et $m = 811$. Lorsqu'il n'y avait qu'une taille d'échantillon n, nous avions un facteur $\sqrt{n}$ au numérateur ou, de manière équivalente, un facteur $\sqrt{1/n}$ au dénominateur. Ici (voir toujours l'appendice facultatif, c'est une question de calcul de variance), nous aurons un facteur $\sqrt{1/n + 1/m}$ au dénominateur, i.e., la statistique de test :

$$t = \frac{\overline{x}_{758} - \overline{y}_{811}}{\sqrt{(1/758 + 1/811)\;\overline{z}_{1569}(1-\overline{z}_{1569})}} = \frac{\dfrac{364}{758} - \dfrac{293}{811}}{\sqrt{\left(\dfrac{1}{758} + \dfrac{1}{811}\right)\dfrac{657}{1\,569}\left(1 - \dfrac{657}{1\,569}\right)}} = 4.77\,.$$

Notre formule générale de la statistique de test, pour des tailles d'échantillon n et m, sera :

$$t_{n,m} = \frac{\overline{x}_n - \overline{y}_m}{\sqrt{(1/n + 1/m)\;\overline{z}_{n+m}(1-\overline{z}_{n+m})}}\,, \tag{5.1}$$

où les données $z_1, \dots, z_{n+m}$ correspondent au regroupement des deux séries de données $x_1, \dots, x_n$ et $y_1, \dots, y_m$.

P–valeur et conclusion. L'appendice facultatif explique que la statistique de test suit approximativement une loi normale sous $H_0 : p_0 = q_0$ et que sous $H_1 : p_0 \neq q_0$, ici bilatère, elle prend de grandes valeurs négatives ou positives. Un schéma comme celui de la Figure 3.2 page 37 s'applique et fournit la P–valeur

$$2 \times \mathbb{P}(Z \geqslant 4.77) = 2(1 - 0.999999) = 0.0002\%\,,$$

qui est quasi-nulle. On rejette très fermement H_0. Cela démontre que $p_0 \neq q_0$ (effet) et en l'occurrence, vu que $\overline{x}_{758} = 364/758 \approx 48.0\% > \overline{y}_{811} = 293/811 \approx 36.1\%$, cela démontre que $p_0 > q_0$ (direction d'effet). Ces conclusions techniques correspondent à la conclusion grand public suivante :

> Ces données montrent de manière très tranchée que la répartition filles–garçons dans les structures de loisirs évolue avec l'âge (effet), en l'occurrence, la part de filles baisse significativement à l'adolescence par rapport à l'âge primaire (direction d'effet).

Remarques : deux erreurs à éviter

Les considérations ci-dessous valent tant pour le test de comparaison de deux proportions indépendantes vu ci-dessus que pour le test de comparaison de deux moyennes indépendantes que nous verrons ci-dessous.

Erreur 1 : négliger la moitié de la variabilité. La fréquence d'échantillon $\overline{y}_{811}$ n'est pas exactement égale à q_0, même si elle en est proche. Il est donc rigoureusement interdit de prendre $p_{\text{ref}} = \overline{y}_{811}$, de poser des hypothèses comme

$$\text{~~$H_0 : p_0 = \overline{y}_{811}$~~} \qquad \text{vs.} \qquad \text{~~$H_1 : p_0 \neq \overline{y}_{811}$~~}$$

et de recourir au test de comparaison à une proportion de référence du Chapitre 3. En effet, les hypothèses ci-dessus (posées au vu des données, ce qui est un crime) ne s'intéressent qu'à la variabilité de l'estimation de p_0 par $\overline{x}_{758}$ et négligent l'autre moitié de la variabilité, à savoir, celle de l'estimation de q_0 par $\overline{y}_{811}$. Il faut bien évidemment traiter les deux variabilités, en les mettant sur le même plan, et c'est exactement ce que fait la statistique de test $t_{n,m}$ définie en (5.1).

Erreur 2 : effectuer deux tests avec $p_{ref} = 50\%$. Il pourrait être tentant ici d'effectuer

— d'une part un test bilatère de comparaison de p_0 à la valeur de référence $p_{ref} = 50\%$, à partir des données $x_1, \dots, x_{758}$, i.e., un test des hypothèses $H_0' : p_0 = 50\%$ vs. $H_1' : p_0 \neq 50\%$; on obtient pour lui une t–valeur de 1.09 puis une P–valeur supérieure à 27.5% ;

— d'autre part, le test de $H_0'' : q_0 = 50\%$ vs. $H_1'' : q_0 \neq 50\%$ à partir des données $y_1, \dots, y_{811}$, pour qui on obtient la t–valeur -7.9 et une P–valeur quasi-nulle.

On conserve donc $H_0' : p_0 = 50\%$ mais on rejette H_0'' et on se tourne vers $H_1'' : q_0 \neq 50\%$ (en l'occurrence, on a montré que $q_0 < 50\%$).

Toutefois, il ne faut pas déduire de cela que l'on a démontré ici aussi que $q_0 < p_0$, car conserver $H_0' : p_0 = 50\%$ ne veut pas dire valider $H_0' : p_0 = 50\%$. La seule certitude raisonnable est que $q_0 < 50\%$, mais pour p_0, on ne sait pas la placer exactement par rapport à 50%. (On ne peut juste pas exclure que p_0 vaille 50%.) Par conséquent, on ne sait pas non plus placer p_0 par rapport à q_0.

2. Comparaison de moyennes indépendantes

Nous nous intéressons désormais à la comparaison de deux moyennes de populations μ_0 et m_0.

Nous imaginons la situation suivante : un gros paquet d'un millier de copies d'étudiants ayant tous composé sur le même sujet doit être corrigé. Bien entendu, un seul professeur ne peut y venir à bout seul et il faut diviser la tâche. Deux professeurs sont volontaires, mais avant de leur confier à chacun la moitié des copies, on veut vérifier que leurs échelles de notations sont identiques en moyenne. I.e., on note μ_0 et m_0 les moyennes qu'ils obtiendraient chacun sur le millier de copies s'ils les corrigeaient chacun toutes : les individus statistiques sont ici les copies, et les moyennes de population μ_0 et m_0 sont évidemment hypothétiques, car il n'arrivera jamais que les deux professeurs corrigent chacun l'ensemble des copies — l'objectif est précisément ici de diviser le travail. Nonobstant cela, il s'agit de tester $H_0 : \mu_0 = m_0$ (les deux professeurs notent de manière identique en moyenne) contre $H_0 : \mu_0 \neq m_0$ (l'un des deux note plus sévèrement que l'autre). Le test est bilatère car aucun contexte spécifique n'est fourni, comme le seraient par exemple des réputations respectives.

		Prof. 1	Prof. 2	Total
Notes	6,5	1	0	1
	7,0	1	0	1
	9,0	3	2	5
	9,5	1	1	2
	10,0	1	1	2
	10,5	1	0	1
	11,0	1	1	2
	11,5	1	1	2
	12,0	2	4	6
	12,5	1	0	1
	13,0	0	3	3
	13,5	1	1	2
	14,0	3	0	3
	14,5	3	3	6
	15,0	3	2	5
	15,5	1	3	4
	16,0	2	0	2
	16,5	3	0	3
	17,0	1	3	4
	17,5	2	3	5
	18,0	1	3	4
	18,5	5	0	5
	19,0	1	2	3
	19,5	0	3	3
	20,0	1	0	1
Total		40	36	76

Le bureau de scolarité sélectionne un petit nombre de copies uniformément au hasard. Ainsi, le premier professeur corrige un échantillon de 40 copies, ce qui aboutit aux notes $x_1, \dots, x_{40}$ tandis que le second professeur, un peu pris par le temps, corrige un échantillon de seulement 36 copies, d'où une seconde série de notes $y_1, \dots, y_{36}$. Les données sont décrites dans leur intégralité ci-contre et résumées dans la sortie logiciel ci-dessous, où l'on lit par exemple : $\overline{x}_{40} = 14.30$ avec $s_{x,40} = 3.56$, et $\overline{y}_{36} = 14.85$ avec $s_{y,36} = 3.19$.

La sortie logiciel met par ailleurs en œuvre le test de comparaison des moyennes.

	Prof	N	Moyenne	Ecart type	Erreur standard
Notes	Prof. 1	40	14,300	3,5641	,5635
	Prof. 2	36	14,847	3,1933	,5322

Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes					
						Signification		Différence	
		F	Sig.	t	df	p unilatéral	p bilatéral	moyenne	Erreur standard
Notes	Hypothèse de variances égales	,257	,614	-,702	74	,242	,485	-,5472	,7797
	Hypothèse de variances inégales			-,706	73,999	,241	,482	-,5472	,7751

La question qui est posée, qui est de tester $H_0 : \mu_0 = m_0$ vs. $H_1 : \mu_0 \neq m_0$, peut être reformulée la suivante :

> Les moyennes d'échantillon $\overline{x}_{40} = 14.30$ pour le premier professeur et $\overline{y}_{36} = 14.85$ pour le second professeur sont-elles significativement différentes ?

Statistique(s) de test. Selon le même principe que pour le test de comparaison de proportions indépdantes, il est intuitif de comparer $\overline{x}_{40}$ et $\overline{y}_{36}$ et donc de construire une statistique de test admettant pour numérateur $\overline{x}_{40} - \overline{y}_{36}$:

$$t = \frac{\overline{x}_{40} - \overline{y}_{36}}{?}.$$

Se pose alors la question de la normalisation. Dans le cas des proportions, les écarts-types de population $\sigma_{x,0}$ et $\sigma_{y,0}$ se déduisaient des proportions de population et étaient nécessairement égaux. Ici, $H_0 : \mu_0 = m_0$ n'entraîne pas nécessairement l'égalité des écarts-types $\sigma_{x,0}$ et $\sigma_{y,0}$. Ceux-ci peuvent être différents ou égaux, et selon la situation, la doctrine statistique traditionnelle propose de diviser par la normalisation

$$\sqrt{s_{x,n}^2/n + s_{y,m}^2/m} \qquad \text{ou} \qquad \sqrt{s_{z,n+m}^2(1/n + 1/m)},$$

où $s_{z,n+m}^2$ désigne l'écart-type de la série de données $z_1, \ldots, z_{n+m}$ correspondent au regroupement des deux séries de données $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$. Comme l'explique plus en détails l'appendice facultatif, deux formules pour la statistique de test $t_{n,m}$ sont donc possibles, et le choix de l'une ou l'autre dépend de l'égalité des écarts-types $\sigma_{x,0}$ et $\sigma_{y,0}$. On teste cette dernière de manière bilatère via les hypothèses $H_0' : \sigma_{x,0} = \sigma_{y,0}$ vs. $H_1' : \sigma_{x,0} \neq \sigma_{y,0}$.

Avec tous ces éléments, nous pouvons décrire le contenu des cases les plus importantes du second tableau de la sortie logiciel ci-dessus :

	Test d'ég. des variances		Test d'égalité des moyennes					
					Signification		Différence	
	F	Sig.	t	df	unilatère	bilatère	moyenne	Erreur standard
Variances égales	...	P–valeur	$\dfrac{\overline{x}_n - \overline{y}_m}{\sqrt{s_{z,n+m}^2(1/n + 1/m)}}$	...	P–valeurs		$\overline{x}_n - \overline{y}_m$	$\sqrt{s_{z,n+m}^2(1/n + 1/m)}$
Variances différentes			$\dfrac{\overline{x}_n - \overline{y}_m}{\sqrt{s_{x,n}^2/n + s_{y,m}^2/m}}$	...	P–valeurs		$\overline{x}_n - \overline{y}_m$	$\sqrt{s_{x,n}^2/n + s_{y,m}^2/m}$

Nous procédons à sa lecture, qui s'effectue en deux temps.

1. Le pré-test d'égalité des variances renvoie une P–valeur de 61.4%, nous conservons donc l'hypothèse $H_0' : \sigma_{x,0} = \sigma_{y,0}$ d'égalité des variances et allons estimer cette variance commune de manière conjointe, à l'aide de la série de données regroupées.
2. I.e., nous lisons ici, pour le test d'égalité des moyennes proprement dit, la première ligne du tableau, et allons exploiter la P–valeur bilatère 48.5%.

Remarque 5.2 (Doxa statistique). Bien entendu, dans d'autres sorties logiciel, la P–valeur du test de comparaison des variances serait $< 5\%$, nous conclurions à des variances $\sigma_{x,0}$ et $\sigma_{y,0}$ différentes, et nous lirions alors la second ligne du tableau en ce qui concerne le test d'égalité des moyennes proprement dit. Pour autant, vous constaterez rapidement que les deux P–valeurs pour les tests d'égalité de moyennes sont toujours très proches, à quelques (dizièmes de) pourcents près. Il me semble peu efficace d'enchaîner deux tests (celui pour l'égalité des variances et celui pour l'égalité des moyennes) et à tout le moins, il faudrait effectuer une correction de Bonferroni et changer le seuil de lecture des P–valeurs. Mais la pratique statistique est différente, c'est celle que j'ai expliquée ci-dessus.

Conclusion pour cet exemple. La P–valeur pour le test de $H_0 : \mu_0 = m_0$ vs. $H_1 : \mu_0 \neq m_0$ vaut 48.5%, nous conservons donc l'hypothèse d'égalité des moyennes, et écrivons :

> Ces données ne permettent pas d'exclure que les deux professeurs notent, en moyenne, de manière identique.
>
> Ces données ne s'opposent donc pas à la division du travail de correction des copies.

Attention, la précision « en moyenne » est cruciale ci-dessus, pour rappeler que nous n'avons fait que comparer deux moyennes. Le chapitre suivant expliquera comment comparer plus finement les manières de noter, via une comparaison des lois de notation des deux professeurs effectuée par un test du χ^2 d'homogénéité.

3. Appendice facultatif : Intuitions pour les statistiques de test

Dans tout cet appendice, nous considérons deux séries de données $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$ issues de variables aléatoires $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$, toutes indépendantes et telles que les $X_1, \ldots, X_n$ ont même loi, d'une part, et les $Y_1, \ldots, Y_m$ également, d'autre part. On rappelle qu'on note $\sigma^2_{x,0}$ et $\sigma^2_{y,0}$ les variances respectives des X_j et Y_k. Leurs espérances seront notées, selon le contexte (données binaires ou données quantitatives générales), p_0 et q_0, ou μ_0 et m_0.

Variance de la différence. Commençons par calculer la variance de la différence $\overline{X}_n - \overline{Y}_m$: par indépendance,

$$\mathrm{Var}(\overline{X}_n - \overline{Y}_m) = \mathrm{Var}(\overline{X}_n) + \mathrm{Var}(\overline{Y}_m) = \frac{\sigma^2_{x,0}}{n} + \frac{\sigma^2_{y,0}}{m}.$$

Sous $H_0 : p_0 = q_0$ ou $H_0 : \mu_0 = m_0$, la variable aléatoire

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\sigma^2_{x,0}/n + \sigma^2_{y,0}/m}} \tag{5.2}$$

est centrée–réduite. Ce n'est pas une statistique de test car elle n'est pas calculable à partir des observations, les variances $\sigma^2_{x,0}$ et $\sigma^2_{y,0}$ étant inconnues.

3.1. Cas des proportions

Nous supposons que $0 < p_0,\, q_0 < 1$. Dans le cas de données binaires $x_j \in \{0,1\}$ et $y_k \in \{0,1\}$, i.e., lorsque les X_j et Y_k suivent des lois de Bernoulli de paramètres respectifs p_0 et q_0, on note que

$$\text{sous } H_0 : p_0 = q_0, \qquad \sigma^2_{x,0} = p_0(1-p_0) = q_0(1-q_0) = \sigma^2_{y,0}\,,$$

et il est naturel d'estimer cette valeur commune par

$$\hat{p}(1-\hat{p})\,, \qquad \text{où} \qquad \hat{p} = \frac{X_1 + \ldots + X_n + Y_1 + \ldots + Y_m}{n+m}\,.$$

Sous H_0, par loi des grands nombres, $\hat{p}$ est un estimateur convergent de $p_0 = q_0$ et par image continue, lorsque $n + m \to \infty$,

$$\sqrt{\hat{p}(1-\hat{p})} \xrightarrow{\mathbb{P}} \sqrt{p_0(1-p_0)}\,.$$

On effectue la substitution dans (5.2) et on considère par conséquent la statistique de test

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

Comportement de $T_{n,m}$ sous H_0. Deux convergences en loi utiles pour l'étude des comportements de $T_{n,m}$, et découlant toutes deux du théorème limite central combiné au lemme de Slutzky, sont que sous $H_0 : p_0 = q_0$,

$$\sqrt{n}\,\frac{\overline{X}_n - p_0}{\sqrt{\hat{p}(1-\hat{p})}} = \underbrace{\sqrt{n}\,\frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)}}}_{\rightharpoonup \mathcal{N}(0,1)}\,\underbrace{\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{p_0(1-p_0)}}}_{\to 1} \rightharpoonup \mathcal{N}(0,1)\,,$$

ce que nous réécrivons de manière intuitive par

$$\frac{\overline{X}_n - p_0}{\sqrt{\hat{p}(1-\hat{p})}} \overset{(d)}{\approx} \mathcal{N}(0,\, 1/n)\,, \qquad \text{et de même,} \qquad \frac{\overline{Y}_m - p_0}{\sqrt{\hat{p}(1-\hat{p})}} \overset{(d)}{\approx} \mathcal{N}(0,\, 1/m)\,,$$

de sorte que, par indépendance de $\overline{X}_n$ et $\overline{Y}_m$,

$$T_{n,m} = \frac{1}{\sqrt{1/n + 1/m}}\left(\frac{\overline{X}_n - p_0}{\sqrt{\hat{p}(1-\hat{p})}} - \frac{\overline{Y}_m - p_0}{\sqrt{\hat{p}(1-\hat{p})}}\right) \overset{(d)}{\approx} \mathcal{N}\left(0,\, \frac{1}{1/n + 1/m}(1/n + 1/m)\right) = \mathcal{N}(0,1)\,.$$

Comportement de $T_{n,m}$ sous H_1. Par loi des grands nombres, quelles que soient les valeurs de p_0 et q_0, on a $\overline{X}_n \to p_0$ et $\overline{Y}_n \to q_0$ en probabilité, et on en déduit en particulier que lorsque $n+m \to \infty$,

$$\mathbb{P}(\hat{p} \in [p_0, q_0]) \longrightarrow 1\,.$$

Par conséquent, lorsque $n + m \to \infty$,

$$\mathbb{P}\left(T_{n,m}\sqrt{1/n + 1/m} \geqslant \frac{p_0 - q_0}{\kappa_0}\right) \longrightarrow 1 \qquad \text{où} \qquad \kappa_0 = \min\{p_0, q_0\}\,(1 - \max\{p_0, q_0\}) > 0\,.$$

Cela démontre en particulier les comportements suivants sous H_1 :

$$\text{si } p_0 > q_0, \quad T_{n,m} \xrightarrow{\mathbb{P}} +\infty \qquad \text{et} \qquad \text{si } p_0 < q_0, \quad T_{n,m} \xrightarrow{\mathbb{P}} -\infty\,.$$

Cela achève de démontrer les comportements utilisés page 65 pour calculer la P–valeur lors de la mise en œuvre du test de comparaison de proportions.

3.2. Cas des moyennes générales

On peut adapter les arguments du cas des proportions ; à y bien regarder, le seul point à revoir concerne les questions d'estimations convergentes. L'hypothèse $H_0 : \mu_0 = m_0$ est muette quant à l'égalité ou à la différence des variances $\sigma^2_{x,0}$ et $\sigma^2_{y,0}$, c'est pourquoi un pré-test d'égalité des variances est mis en place, afin de déterminer celle qui sera moralement la meilleure des deux normalisations possibles, à réinjecter dans l'équation (5.2).

Cas de variances différentes. Plus précisément, si l'on rejette $H_0' : \sigma^2_{x,0} = \sigma^2_{y,0}$ et que l'on passe à $H_1' : \sigma^2_{x,0} \neq \sigma^2_{y,0}$, on procède à une estimation convergente séparée de $\sigma^2_{x,0}$ et $\sigma^2_{y,0}$: lorsque n et m tendent vers $+\infty$,

$$\widehat{\sigma}^2_{x,n} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 \xrightarrow{\mathbb{P}} \sigma^2_{x,0} \qquad \text{et} \qquad \widehat{\sigma}^2_{y,m} = \frac{1}{m-1}\sum_{k=1}^{m}(Y_k - \overline{Y}_n)^2 \xrightarrow{\mathbb{P}} \sigma^2_{y,0}\,,$$

puis on démontre que

$$T_{n,m} = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\widehat{\sigma}^2_{x,n}/n + \widehat{\sigma}^2_{y,m}/m}}$$

converge en loi vers $\mathcal{N}(0,1)$ sous $H_0 : \mu_0 = m_0$, tandis que $T_{n,m} \to +\infty$ et $-\infty$ en probabilité lorsque $\mu_0 > m_0$ et $\mu_0 < m_0$, respectivement.

Cas de variances égales. En fait, ces comportements valent que l'on ait $\sigma^2_{x,0} \neq \sigma^2_{y,0}$ ou $\sigma^2_{x,0} = \sigma^2_{y,0}$. Mais dans ce dernier cas, on préfère recourir à l'estimation groupée

$$\widehat{\sigma}^2_{z,n+m} = \frac{1}{n+m-2}\left(\sum_{j=1}^{n}(X_j - \overline{Z}_{n+m}) + \sum_{k=1}^{m}(Y_k - \overline{Z}_{n+m})\right),$$

$$\text{où} \qquad \overline{Z}_{n+m} = \frac{1}{n+m}\left(\sum_{j=1}^{n} X_j + \sum_{k=1}^{m} Y_k\right),$$

et plutôt considérer

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

pour qui on peut également démontrer une convergence vers $\mathcal{N}(0,1)$ sous H_0 et des limites $\pm\infty$ sous H_1. Vous aurez noté que dans la formule pour $\widehat{\sigma}^2_{z,n+m}$, on divise la somme des sommes par un facteur $n+m-2$ et non $n+m-1$: asymptotiquement, cela n'a pas d'importance, mais lorsque l'on suppose que les X_j et Y_k suivent des lois normales, on peut démontrer que

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

où l'on rappelle que $\mathcal{T}_{n+m-2}$ désigne la loi de Student à $n+m-2$ degrés de liberté (voir page 55).

Auto-évaluation de performance — 9 points

Pour son stage de terrain de 3 semaines, une jeune étudiante de L3 a été affectée comme représentante de commerce stagiaire dans une entreprise. Son chef d'équipe lui a appris le *pitch* produit utilisé depuis des mois et avec succès par des dizaines d'autres représentants de commerce. Toutefois, elle pense avoir une meilleure idée de *pitch*. Comme elle est malgré tout prudente, elle veut tester son idée pendant quelques heures avant de l'utiliser pour de plus longues périodes de temps. Ainsi, au jour 1 de son stage, elle se fond dans le moule et suit scrupuleusement les instructions commerciales. Mais au jour 2, elle recourt à son propre *pitch*. Les résultats sont les suivants : au jour 1, elle a discuté avec 534 prospects, parmi lesquels 64 ont souscrit à son produit ; au jour 2, elle a obtenu 67 souscriptions lors de 526 contacts téléphoniques. Que doit-elle faire ?

☐ Quels sont les paramètres d'intérêt ?

☐☐ Précisez les hypothèses testées ici. Formulez-les en mots agréables puis en équations. Justifiez votre choix.

Mettez en œuvre le test d'hypothèses

☐ en procurant un schéma récapitulant les comportements de la statistique de test adéquate sous H_0 et H_1,
☐☐ en calculant la valeur réalisée de cette statistique de test (écrivez ce que vous tapez sur votre calculatrice),
☐ en indiquant la P-valeur obtenue.

☐ Formez une conclusion statistique, en entourant un élément de chaque colonne ; vos choix doivent suivre la formulation de vos hypothèses et la P-valeur obtenue :

A. Ces données montrent que
B. Ces données suggèrent que
C. Ces données ne permettent pas de montrer que
D. Ces données ne permettent pas d'exclure que

1. le *pitch* de l'entreprise est le plus efficace des deux
2. les deux *pitch*s sont aussi efficaces l'un que l'autre
3. le *pitch* de la stagiaire est le plus efficace des deux
4. le *pitch* de l'entreprise est le moins efficace des deux

☐ Formez une conclusion stratégique : quel *pitch* doit-elle utiliser les jours suivants ?

Stimulation tactile — 8 points

Différentes études, en marketing (Jacob Hornik, "Tacticle stimulation and consumer response", *Journal of Consumer Research*, vol. 19(3), pages 449–458, 1992) et en psychologie (Nicolas Guéguen, "Courtship compliance : the effect of touch on women's behavior", *Social Influence*, vol. 2(2), pages 81–97, 2007) ont montré qu'un léger toucher, par exemple de l'avant-bras, même non consciemment perçu, permet de mettre la personne touchée dans de meilleures dispositions par rapport aux requêtes formulées par la personne qui la touche. En l'occurrence, illustrons cela par exemple de vente de vêtements pour la marque B&G qui a longtemps savamment misé sur ses vendeurs et vendeuses.
Prenons deux boutiques B&G dans des zones de chalandise économiquement comparables. Dans le premier magasin, les vendeurs évitent tout contact physique avec la clientèle tandis qu'ils ont instruction dans le second magasin de toucher légèrement l'avant-bras des clients de sorte que ces derniers ne s'en rendent pas compte. Nous nous intéressons aux taux d'achat p_0 et q_0 respectivement obtenus dans chaque magasin (p_0 sans toucher, et q_0 avec toucher, donc).

Test bilatère de $H_0 : p_0 = q_0$ vs. $H_1 : p_0 \neq q_0$

Une fois ces hypothèses bilatères fixées (la fin du sujet déterminera qui les a fixées ainsi), l'expérience statistique est réalisée et les comportements des 120 premiers clients accueillis sont étudiés dans chaque magasin. Dans le premier magasin (sans toucher), 12 clients effectuent des achats alors qu'ils sont 23 dans le second magasin (avec toucher). Mettez en œuvre le test des hypothèses $H_0 : p_0 = q_0$ vs. $H_1 : p_0 \neq q_0$

☐ en procurant un schéma récapitulant les comportements de la statistique de test adéquate sous H_0 et H_1,
☐☐ en calculant la valeur réalisée de cette statistique (*écrivez le calcul que vous tapez sur votre calculatrice*),
☐ et en concluant à une P−valeur.

☐ Formez une conclusion en choisissant une lettre et un chiffre :

[Début]
A. Les données ne permettent pas d'exclure que
B. Les données suggèrent que
C. Les données montrent que
D. Les données ne permettent pas d'affirmer que

[Fin]
1. les taux d'achat sont différents avec et sans toucher
2. les taux d'achat sont identiques avec et sans toucher
3. le taux d'achat augmente lorsqu'il y a toucher
4. le taux d'achat décroît lorsqu'il y a toucher

Test unilatère de $H_0 : p_0 = q_0$ vs. $H_1 : p_0 < q_0$

Nous testons désormais $H_0 : p_0 = q_0$ vs. $H_1 : p_0 < q_0$, avec les mêmes données ; à cet effet,

☐ procurez un schéma récapitulant les comportements de la statistique de test adéquate sous ces H_0 et H_1, calculez la P–valeur associée.

(Aucune formulation de conclusion grand public n'est requise.)

Test unilatère de $H_0 : p_0 = q_0$ vs. $H_1 : p_0 > q_0$

Nous testons enfin le troisième couple d'hypothèses possible, à savoir $H_0 : p_0 = q_0$ vs. $H_1 : p_0 > q_0$, toujours avec les mêmes données. Là encore,

☐ procurez un schéma récapitulant les comportements de la statistique de test adéquate sous ces H_0 et H_1,

☐ calculez la P–valeur associée.

Formulez des conclusions cohérentes avec les hypothèses testées ici et la P–valeur obtenue, en procurant des couples lettre–numéro (avec le même codage qu'à la page précédente) :

☐ Première conclusion possible Lettre : _ Nombre : _

Seconde conclusion possible Lettre : _ Nombrer : _

Qui choisit quelles hypothèses ?

Considérons une chercheuse en marketing et la directrice des ventes de B&G. Quel couple d'hypothèses devrait choisir chacune parmi ceux-ci ?

(Couple 1) $H_0 : p_0 = q_0$ vs. $H_1 : p_0 \neq q_0$
(Couple 2) $H_0 : p_0 = q_0$ vs. $H_1 : p_0 < q_0$
(Couple 3) $H_0 : p_0 = q_0$ vs. $H_1 : p_0 > q_0$

Reportez simplement le numéro (pour une fois sans justification) :

☐ Chercheuse en marketing : Couple _

Directrice des ventes : Couple _

Consommation d'alcool lors des POWs d'HEC

Cet exercice se fonde sur des données réelles collectées par Veronika Czellar, qui enseignait avec moi à l'automne 2008... données que je compte mettre à jour avec vous à l'automne 2023, où nous avons à nouveau cours de statistique les vendredis matins, après une pause de 15 ans. En 2008, les cours avaient lieu à 8h (!) ou 10h, au lieu de 9h40 et 11h20 comme à l'automne 2023. Nous avions des groupes de 40 étudiants environ — mais vous verrez dans le jeu de données que l'absentéisme était assez prononcé. Nous avons demandé à chaque étudiant présent d'indiquer sa consommation d'alcool durant la soirée et la nuit de la veille.

1. Quels paramètres d'intérêt visions-nous ? Quelles hypothèses avions-nous en tête d'après vous ?

		Groupe		
		8h-9h45	10h-11h45	Total
Nombre de verres	,0	2	2	4
	,6	0	1	1
	1,0	0	1	1
	2,0	2	3	5
	2,5	1	0	1
	3,0	3	5	8
	3,1	2	0	2
	3,5	1	0	1
	3,6	1	0	1
	4,0	1	2	3
	4,3	2	0	2
	5,0	3	4	7
	5,5	0	1	1
	6,0	1	1	2
	7,0	0	2	2
	8,0	0	2	2
	8,2	1	0	1
	9,0	1	0	1
	9,7	1	0	1
	10,0	0	1	1
	11,0	0	1	1
	12,0	0	1	1
	13,0	1	0	1
	14,0	0	1	1
	18,7	0	1	1
	31,0	0	1	1
	51,0	0	1	1
Total		23	31	54

Nous avions obtenu 54 réponses, résumées ci-contre.

2. Voyez-vous certaines valeurs aberrantes ?
3. De quel type de données s'agit-il ?

Les sorties logiciel ci-dessous ont été obtenues en réalisant les tests d'hypothèses adéquats, d'abord sur les données originelles, puis sur ces mêmes données nettoyées de certaines valeurs aberrantes.

4. Entourez les P–valeurs finales que vous y lisez.
5. Ecrivez une conclusion à destination de Xavier Boute (comme s'il faisait partie du grand public, alors que lui aussi enseignait ce cours à l'automne 2008 !).

	Groupe	N	Moyenne	Ecart type	Moyenne d'erreur standard
Nombre de verres	8h-9h45	23	4,448	3,0598	,6380
	10h-11h45	31	7,735	10,1658	1,8258

Test sur données originelles

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes					
						Signification		Différence moyenne	Erreur standard
		F	Sig.	t	df	p unilatéral	p bilatéral		
Nombre de verres	Hypothèse de variances égales	5,005	,030	-1,498	52	,070	,140	-3,2877	2,1944
	Hypothèse de variances inégales			-1,700	37,021	,049	,098	-3,2877	1,9341

	Groupe	N	Moyenne	Ecart type	Moyenne d'erreur standard
Nombre de verres	8h-9h45	23	4,448	3,0598	,6380
	10h-11h45	28	4,968	3,6137	,6829

Test sur données nettoyées en enlevant les valeurs > 15

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes					
						Signification		Différence moyenne	Erreur standard
		F	Sig.	t	df	p unilatéral	p bilatéral		
Nombre de verres	Hypothèse de variances égales	,906	,346	-,547	49	,293	,587	-,5200	,9501
	Hypothèse de variances inégales			-,556	48,942	,290	,580	-,5200	,9346

Ecarts salariaux femmes–hommes chez les cadres, version revisitée

Nous reprenons avec la méthodologie des tests d'hypothèses sur un exercice déjà traité par intervalles de confiance, à partir des mêmes données. Nous obtenons la sortie logiciel suivante :

	SEXE	N	Moyenne	Ecart type	Erreur standard
Salaire net mensuel	Hommes	179	3431,46	3895,437	291,159
	Femmes	147	2434,90	1282,947	105,816

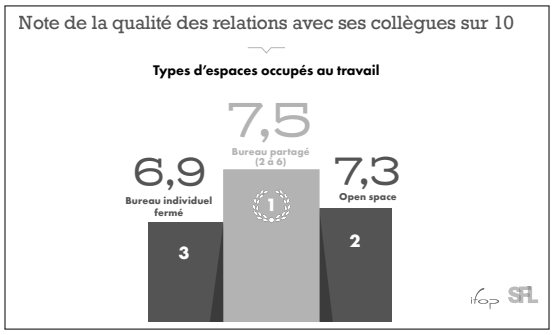
Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes							
						Signification		Différence moyenne	Erreur standard	Intervalle de confiance de la différence à 95 %	
		F	Sig.	t	df	p unilatéral	p bilatéral			Inférieur	Supérieur
Salaire net mensuel	Hypothèse de variances égales	8,195	,004	2,972	324	,002	,003	996,560	335,371	336,781	1656,339
	Hypothèse de variances inégales			3,217	223,375	<,001	,001	996,560	309,791	386,074	1607,047

1. Quelles hypothèses sont testées dans la sortie ci-dessus ? (Attention à bien fournir tous les mots cruciaux.)
2. Entourez dans la sortie la P–valeur finale pour le test mené.
3. Quelles hypothèses auriez-vous voulu mener, et quelle P–valeur y aurait-elle été associée ?
4. Comparez la taille d'effet lue dans la sortie (ou une taille d'effet pouvant être recalculée, en lien avec les hypothèses de la question 3) à celle que nous avions obtenue par intervalles de confiance. Pouvez-vous avancer une explication à cette comparaison ?

Enquête sur le bien-être au travail — 6 points

Dans cet exercice, nous allons effectuer une lecture critique de la sortie graphique ci-dessous, extraite d'une présentation d'une enquête sur le bien-être au travail. A défaut d'avoir pu récupérer les données originelles, j'ai inventé des données cohérentes : que l'on a interrogé 1 619 salariés, dont 901 disposant d'un bureau partagé et 427 travaillant en *open space*. Il leur a notamment été demandé d'indiquer, sur une échelle de 1 à 10, la qualité de leur relation de travail avec leurs collègues. La présentation résume les réponses obtenues ainsi :



Un traitement des données originelles par un logiciel statistique produit la sortie suivante.

	Type	N	Moyenne	Ecart type	Moyenne d'erreur standard
Note	Bureau partagé	901	7,48	1,835	,061
	Open space	427	7,30	2,131	,103

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	df	Sig. (bilatérale)	Différence moyenne	Std. standard	Inférieur	Supérieur
Note	Hypothèse de variances égales	17,612	,000	1,531	1326	,126	,174	,114	-,049	,397
	Hypothèse de variances inégales			1,452	735,141	,147	,174	,120	-,061	,409

☐ Quels sont les paramètres d'intérêt étudiés ici ? (Attention à bien fournir tous les mots cruciaux.)

☐ Quelles hypothèses sont testées dans la sortie ci-dessus ?

☐☐ Entourez dans la sortie la P-valeur finale pour le test mené.
Conservez-vous ou rejetez-vous votre hypothèse H_0 ? Ecrivez une conclusion lisible par le grand public.

☐☐ Commentez l'accord ou le désaccord entre le résultat de votre traitement statistique et la présentation :
La présentation fournit-elle une image juste et lisible de la réalité, ou est-elle trompeuse ?
J'attends une réponse justifiée, *concise*, percutante, et avec un positionnement clair et assumé.

6

Tests du χ^2, d'ajustement et d'indépendance

(Prononciation : χ^2 se lit "khi-deux" ou "khi-carré" ; il s'agit de la lettre grecque χ et non d'un X.)

Les tests du χ^2 généralisent les tests vus pour les proportions :

— Le test du χ^2 d'ajustement à une loi de référence étend le test de comparaison à une proportion de référence p_{ref} au cas où les données, catégorielles, peuvent prendre non plus seulement 2 valeurs, mais 3 valeurs ou plus.

— Le test du χ^2 d'indépendance, appelé aussi test du χ^2 d'homogénéité, étend le test de comparaison de deux proportions indépendantes (qui se fondait sur des données organisées dans un tableau 2×2) au cas où les deux variables, celle mesurant le groupe d'appartenance et la variable d'intérêt, toutes deux catégorielles, peuvent prendre chacune 2 valeurs ou plus — i.e., ce test traite de données organisées dans des tableaux $r \times s$, où $r \geqslant 2$ et $s \geqslant 2$.

1. Test du χ^2 d'ajustement à une loi de référence

Ce test part de données catégorielles : on note $c_0, c_1, \ldots, c_k$ les $k+1 \geqslant 2$ valeurs possiblement prises.

Exemple. Le cas des données binaires considéré jusque-là correspond à $\{c_0, c_1\} = \{0, 1\}$. Le fil rouge de cette partie sera à propos de $\{c_0, c_1, \ldots, c_9\} = \{0, 1, \ldots, 9\}$. Un test de représentativité d'un échantillon en termes d'âge adulte pourrait considérer c_0, c_1, c_2 donnés par les catégories d'âge 18–35 ans, 36–59 ans, et 60+ ans.

On postule une certaine loi de référence $p_{0,\text{ref}}, \ldots, p_{k,\text{ref}}$ pour les valeurs $c_0, \ldots, c_k$. On dispose de données $x_1, x_2, \ldots, x_n$, issues de variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$, dont on se demande si leur loi commune est, ou non, cette loi de référence $p_{0,\text{ref}}, \ldots, p_{k,\text{ref}}$. Ici, pour des raisons mathématiques profondes, on n'a pas le choix des hypothèses, ce sera toujours H_0 la conformité (l'égalité) à la loi de référence vs. H_1 la non-conformité :

H_0 [conformité] : les données sont issues de la loi de référence $p_{0,\text{ref}}, \ldots, p_{k,\text{ref}}$

H_1 [non-conformité] : les données sont issues d'une autre loi.

Exemple. L'exemple fil rouge de cette partie est expliqué très en détails dans l'extrait de l'article de presse "The Devil Is in the Digits" reproduit plus loin. Il s'agit de voir si les résultats des élections présidentielles de 2009 en Iran ont été reportés de manière sincère par les autorités, ou si, au contraire,

Province	Candidat	Score	Unités
Ardabil	Ahmadinejad	325 911	1
Ardabil	Karroubi	2 319	9
Ardabil	Mousavi	302 825	5
Ardabil	Rezaee	6 578	8
Azarbaijan, East	Ahmadinejad	1 131 111	1
Azarbaijan, East	Karroubi	7 246	6
Azarbaijan, East	Mousavi	837 858	8
Azarbaijan, East	Rezaee	16 920	0
Azarbaijan, West	Ahmadinejad	623 946	6
Azarbaijan, West	Karroubi	21 609	9
Azarbaijan, West	Mousavi	656 508	8
Azarbaijan, West	Rezaee	12 199	9
Bushehr	Ahmadinejad	299 357	7
Bushehr	Karroubi	3 563	3
Bushehr	Mousavi	177 268	8
Bushehr	Rezaee	7 607	7

FIGURE 6.1 – Extrait du jeu de données d'élections présidentielles de 2009 en Iran : scores des 4 candidats sur les 4 premières provinces d'Iran (sur 29).

elles les ont altérés (auquel cas, on peut penser que c'était en défaveur notamment du candidat Karroubi). La méthodologie est la suivante : les individus statistiques sont les couples province–candidat, il y en a $29 \times 4 = 116$. Pour chaque individu statistique, on observe un score (un nombre de voix) : en général, de grands nombres, voir l'extrait du jeu de données en haut de cette page. Dans ces nombres, ce qui est porteur de sens, c'est l'ordre de grandeur (10 puissance combien ?) et les premiers chiffres du nombre (appelés les premiers chiffres significatifs). Ce qui en revanche n'est pas porteur de sens et n'est que du « bruit statistique », ce sont les chiffres des unités : que le candidat Ahmadinejad obtienne un score de 325 911 dans la province d'Ardabil (première observation du jeu de données) n'est pas un fait fondamentalement différent d'un score de 325 910 ou 325 912. Ces variations d'unités sont tout à fait aléatoires et plutôt liées à une somme de micro-événements (telle personne était-elle trop malade pour aller voter ?) qu'à une lame de fond (un fort soutien du pays pour tel ou tel candidat). On s'attend donc, et l'article l'explique, à ce que les chiffres des unités des 116 scores ne prennent pas de valeurs particulières, i.e., soient uniformément distribués entre les valeurs possibles $c_0 = 0, \quad c_1 = 1, \quad \ldots, \quad c_9 = 9$. Notre loi de référence est donc ici la loi uniforme

$$(p_{0,\mathrm{ref}}, \ldots, p_{9,\mathrm{ref}}) = (10\%, \ldots, 10\%)\,.$$

Statistique de test. Comme aux chapitres précédents, la question est d'exhiber une statistique de test (i.e., une quantité aléatoire pouvant être calculée à partir des données) dont le comportement aléatoire est parfaitement connu sous H_0 et qui ait des comportements différents sous H_1.

Pour chaque modalité possible c_j, où $j \in \{0, 1, \ldots, k\}$, on désigne par $N_{j,n}$ le nombre d'observations parmi $X_1, \ldots, X_n$ prenant la valeur c_j, et par $\widehat{p}_{j,n}$ la fréquence empirique de cette modalité c_j :

$$N_{j,n} = \mathrm{Card}\big\{t \in \{1, \ldots, n\} : X_t = c_j\big\} \qquad \text{et} \qquad \widehat{p}_{j,n} = \frac{N_{j,n}}{n}\,,$$

où Card désigne le cardinal (le nombre d'éléments) d'un ensemble. Il est naturel de confronter les effectifs observés et les effectifs « théoriques », dits également effectifs « attendus », issus de la loi de référence ; ou, de manière équivalente, les fréquences empiriques et la loi de référence : pour tout $j \in \{0, 1, \ldots, k\}$,

$$N_{j,n} \quad \text{vs.} \quad n\,p_{j,\mathrm{ref}} \qquad \text{ou} \qquad \widehat{p}_{j,n} \quad \text{vs.} \quad p_{j,\mathrm{ref}}\,.$$

The Devil Is in the Digits

The Washington Post

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. [...]

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. [...]

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.

Il s'agit en réalité d'effectuer une comparaison simultanée de toutes ces valeurs. Suite à des travaux initiés par Karl Pearson en 1900 et des raisons mathématiques profondes[1], la comparaison simultanée s'effectue ainsi :

$$T_n = \sum_{j=0}^{k} \frac{(N_{j,n} - n\,p_{j,\text{ref}})^2}{n\,p_{j,\text{ref}}} = \sum_{j=0}^{k} \left(\sqrt{n}\,\frac{\widehat{p}_{j,n} - p_{j,\text{ref}}}{\sqrt{p_{j,\text{ref}}}} \right)^2.$$

Les deux expressions sont égales et ont des mérites différents. Par exemple, il est plus simple de calculer soi-même la valeur réalisée de T_n avec la première expression ; heureusement, comme vous allez le voir, nous allons plutôt confier cette tâche à un logiciel statistique. La seconde expression permet de faire un peu le lien avec le théorème limite central et la statistique de test utilisée pour le test de comparaison à une proportion de référence.

1. A savoir, une version multi-dimensionnelle du théorème limite central dans $\mathbb{R}^{k+1}$ associée au théorème projection orthogonale de Cochran et à une image continue par le carré de la norme euclidienne — des résultats classiquement énoncés et prouvés au cours du second semestre d'un M1 de mathématiques, dans le cadre du cours de statistique mathématique, et donc, tout à fait hors de notre portée, hélas...

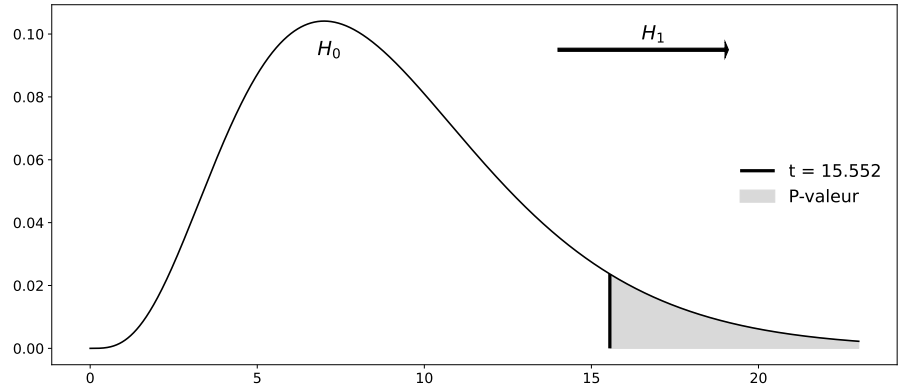


FIGURE 6.2 – Schéma de principe du calcul de la P–valeur pour la valeur réalisée $t = 15.552$ de la statistique de test et la loi du χ^2 à 9 degrés de liberté.

Comportement sous H_1 : Notons $p_{0,\text{vraie}}, \dots, p_{k,\text{vraie}}$ la « vraie » loi des observations $X_1, \dots, X_n$. L'hypothèse H_1 dit exactement que cette vraie loi est différente de la loi de référence $p_{0,\text{ref}}, \dots, p_{k,\text{ref}}$. Par conséquent, il existe au moins une modalité c_j (et en fait, au moins deux) telle(s) que $p_{j,\text{vraie}} \neq p_{j,\text{ref}}$. Par loi des grands nombres,

$$\frac{\widehat{p}_{j,n} - p_{j,\text{ref}}}{\sqrt{p_{j,\text{ref}}}} \xrightarrow{\mathbb{P}} \frac{p_{j,\text{vraie}} - p_{j,\text{ref}}}{\sqrt{p_{j,\text{ref}}}} \neq 0\,, \qquad \text{d'où} \qquad \left(\sqrt{n}\,\frac{\widehat{p}_{j,n} - p_{j,\text{ref}}}{\sqrt{p_{j,\text{ref}}}}\right)^2 \xrightarrow{\mathbb{P}} +\infty\,,$$

d'où $T_n \to +\infty$. Sous H_1, la statistique de test T_n tend à prendre de grandes valeurs positives.

Comportement sous H_0 : $p_{j,\text{vraie}} = p_{j,\text{ref}}$ pour tout j : On admet que pour les raisons citées à la note de bas de page 1, la statistique de test T_n converge en loi vers une loi dite du χ^2 à k degrés de liberté. (Voir l'appendice facultatif pour plus de détails sur la loi du χ^2.)

Ces comportements sont représentés de manière stylisés sur la Figure 6.2.

Incise : conditions de validité. La note de bas de page 1 cite les théorèmes à appliquer pour obtenir la convergence en loi sous H_0. Ce sont des résultats asymptotiques, et des statisticiens ont établi par simulations dans les années 1970 quelles seraient des conditions « pratiques » pour que T_n suive approximativement une loi du χ^2 à k degrés de liberté :

— que la taille d'échantillon n soit supérieure à 30 ;

— que tous les effectifs attendus $n\,p_{j,\text{ref}}$ soient supérieurs à 5.

Lorsque la seconde condition n'est pas vérifiée, on peut procéder soit à des regroupements, soit à des suppressions de classes. Nous illustrerons cela dans les exercices.

Lecture de la P–valeur. Votre mission principale est de pouvoir lire une sortie logiciel comme celle de la Figure 6.3. La colonne des effectifs observés nous y résume les 116 données, en les catégorisant. Les effectifs théoriques valent ici tous $116 \times 10\% = 11.6$, ils sont tous supérieurs à 5, ce qu'indique la note de bas de tableau. En outre, la taille totale d'échantillon $n = 116$ est bien supérieure à 30, le test est validement appliqué.

Nous pourrions alors calculer manuellement la valeur réalisée de la statistique de test :

$$t = \frac{(9-11.6)^2}{11.6} + \frac{(11-11.6)^2}{11.6} + \cdots + \frac{(17-11.6)^2}{11.6} + \frac{(13-11.6)^2}{11.6} = 15.552\,,$$

Chiffre unités

	Effectif observé	N théorique	Résidus
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 statistique

	Chiffre unités
Khi-carré	15,552[a]
df	9
Sig. asymptotique	,077

a. 0 cellules (0,0%) ont des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 11,6.

Modalités	Effectifs observés	Effectifs théoriques	Résidus
c_0	$N_{0,n}$	$n\,p_{0,\text{ref}}$	$N_{0,n} - n\,p_{0,\text{ref}}$
c_1	$N_{1,n}$	$n\,p_{1,\text{ref}}$	$N_{1,n} - n\,p_{1,\text{ref}}$
...	...	...	...
...	...	...	...
...	...	...	...
c_k	$N_{k,n}$	$n\,p_{k,\text{ref}}$	$N_{k,n} - n\,p_{k,\text{ref}}$
Total	n		

Test statistique	
Khi-carré	[valeur de D_n]
df	k
Sig. asymptotique	[P–valeur]

Commentaire sur effectifs théoriques tous $\geqslant 5$, ou non

FIGURE 6.3 – Sortie logiciel [haut] correspondant à la mise en œuvre sur les données de la Figure 6.1 du test du χ^2 d'ajustement à la loi uniforme sur $\{0, 1, \ldots, 9\}$, et explication stylisée [bas] du contenu des tableaux.

mais nous nous contenterons plutôt de la contempler dans la sortie logiciel. A partir de $t = 15.552$ et du nombre de degrés de liberté $k = 9$ de la loi du χ^2 limite ("degrees of freedom" en anglais, expression abrégée en "df" dans la sortie logiciel), nous pouvons déterminer la P–valeur, selon le principe illustré à la Figure 6.2. Une détermination « manuelle » requerrait que je vous fournisse des tables de fonction de répartition de la loi du χ^2 pour de nombreuses valeurs de k, ce qui ne serait pas très écologique ni très pratique. Aussi, là encore, nous laissons faire le logiciel et lisons dans la sortie la P–valeur finale de 7.7%.

Conclusion. Nous conservons, certes sur le fil, l'hypothèse H_0 de conformité à la loi uniforme des chiffres des unités des scores reportés. Avec ces données et par ce seul traitement statistique de test du χ^2, nous ne pourrions pas exclure avec une certitude raisonnable un report sincère des résultats et nous n'aurions donc pas d'arguments suffisamment forts pour faire voter une résolution contre l'Iran en Conseil de sécurité de l'ONU. Il faudrait exhiber d'autres arguments à cet effet.

Remarque 6.1 (Mais l'article, lui, trouvait une P–valeur $< 5\%$!). L'article écrit (cf. fréquences observées de $20/116 = 17.2\%$ pour $c_7 = 7$ et $5/116 = 4.3\%$ pour $c_5 = 5$) :

> 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.

La dernière phrase est une version grand public d'une P–valeur égale à 4% pour eux... mais dont je me demande d'où ils la sortent. J'ai eu vent de cet article car je lisais, sur des blogs de statisticiens, des dénonciations de la méthodologie employée ici par les auteurs. Ils semblent avoir choisi leurs hypothèses au vu des données (en prenant les pires déviations), et ne semblent pas tester simultanément toutes les déviations comme le fait de manière juste et équitable le test du χ^2 d'ajustement ; cela conduit évidemment à minorer la P–valeur.

Utilisations du test du χ^2 d'ajustement à une loi

1. Vérification de l'authenticité de données. Nous avons illustré ci-dessus une première utilisation du test pour détecter des manipulations [2] de données. Un autre exemple est discuté dans

> Tomasz Michalski and Gilles Stoltz, Do countries falsify economic data strategically ? Some evidence that they might, The Review of Economics and Statistics, 95(2):591-616, 2013. Téléchargeable ici : `https://shs.hal.science/halshs-00482106`

et porte sur l'ajustement à la loi de Benford, qui est la loi du premier chiffre significatif des nombres qui nous entourent, et notamment des nombres issus de phénomènes économiques. Cette loi indique qu'environ 30.1% du temps on observe 1 comme premier chiffre significatif, environ 17.6% un 2, etc., jusqu'environ 4.6% du temps un 9.

2. Atteinte d'objectifs commerciaux. Parce que nous sommes à HEC... Une entreprise peut fixer une loi de référence et vouloir vérifier ensuite qu'elle est bien respectée. Voir à ce sujet l'exercice sur les temps d'attente lors d'appels à un service clients.

3. Représentativité d'un échantillon. Il s'agit ici du même principe que par exemple le test de représentativité femmes–hommes vu en exercice page 43, mais étendu au cas de plus de deux catégories : par exemple, au cas de multiples catégories d'âges ou à des couples sexe–programme comme dans l'exercice sur la campagne JE énoncé en fin de ce chapitre. Attention, l'hypothèse H_0 de conformité correspond ici à la représentativité : nous ne pourrons donc jamais prouver celle-ci, nous pourrons au mieux ne pas l'exclure. Cela conduira à des conclusions du type : « On ne peut pas exclure que cet échantillon soit représentatif » ou encore, « Aucun biais significatif n'est présent dans la constitution cet échantillon ».

2. Test du χ^2 d'indépendance / d'homogénéité

L'exemple fil rouge de cette partie sera une question de notation équitable :

		A	B	C	D	E	F	Total
Prof	Grincheux	14	15	26	18	17	5	95
	Gentil	21	18	24	19	15	2	99
Total		35	33	50	37	32	7	194

2. Dans Harry Potter, nous serions donc dans la cadre du cours de Défense contre les forces du Mal. J'ai également plaisir à vous rappeler cette citation faussement attribuée à Winston Churchill par la propagande nazie : "Do not trust statistics that you did not fake yourself".

Les deux professeurs utilisent-ils les mêmes échelles de notation, ou notent-ils de manière différente ?

Ici, nous allons comparer les lois de répartition des notes-lettres, ou dit autrement, les profils de notation des deux enseignants. Nous ne comparerons plus simplement de notes moyennes comme au chapitre précédent (d'autant plus qu'ici, avec des notes-lettres, nous serions bien peine de calculer des moyennes !). Le test que nous allons introduire s'appelle le test du χ^2 d'homogénéité (parce que nous étudions l'égalité des loi de répartitions de notes) ou test du χ^2 d'indépendance (parce que nous étudions l'indépendance de la loi de la note par rapport au professeur).

Nous disposons de couples de données $(x_1, y_1), \ldots, (x_n, y_n)$, où pour le t–ème étudiant parmi les $n = 194$ considérés, x_t désigne son professeur et y_t sa note-lettre. Il y a $r = 2$ modalités possibles pour les x_t, à savoir Grincheux et Gentil, et $s = 6$ modalités possibles pour les y_t, à savoir A–F. C'est pourquoi les données ont été présentées ci-dessus sous la forme d'un tableau 2×6. Nous supposons que ces données sont la réalisation de couples de variables aléatoires indépendantes et identiquement distribués $(X_1, Y_1), \ldots, (X_n, Y_n)$. Les hypothèses testées peuvent être formulées (très) mathématiquement ainsi :

- H_0 [homogénéité / indépendance] : les lois conditionnelles des Y_t sachant $X_t = x$ sont égales, quel que soit x ; ou : la loi commune des (X_t, Y_t) est une loi-produit, i.e., Y_t est indépendante de X_t.
- H_1 [manque d'homogénéité / dépendance] : ces lois conditionnelles varient avec x ; ou : la loi des (X_t, Y_t) n'est pas le produit de ses marginales, i.e., Y_t n'est indépendante de X_t.

Remarque 6.2 (Test « très »bilatère). Dans l'exemple fil rouge, on ne peut pas préciser H_1 : des phénomènes plus complexes que « un professeur note plus sévèrement que l'autre » peuvent survenir ; par exemple, un professeur pourrait utiliser une échelle de notes resserrée et donner surtout des B - C - D, alors que l'autre professeur pourrait exploiter mieux l'ensemble des notes possibles et donner de nombreux A mais aussi de nombreux E et F.

Statistique de test. Ici, nous ne disposons pas de loi de référence, qui, sous H_0 pourrait être la loi conditionelle commune. Qu'à cela ne tienne, nous allons l'estimer par les fréquences empiriques globales, lorsque toutes les données sont mises dans le même sac : pour chaque lettre $L \in \{A, \ldots, F\}$,

$$\widehat{p}_L = \frac{N_{\cdot,L,n}}{n} \qquad \text{où} \qquad N_{\cdot,L,n} = \mathrm{Card}\big\{t \in \{1, \ldots, n\} : Y_t = L\big\}.$$

Exemple. Dans l'exemple fil rouge, cela revient à estimer que la loi conditionnelle commune sous H_0 donne une probabilité d'environ $35/194 \approx 18.04\%$ à la note A.

Nous répercutons alors cette loi conditionnelle commune estimée dans chaque groupe de valeurs selon l'autre variable. Soit

$$N_{i,\cdot,n} = \mathrm{Card}\big\{t \in \{1, \ldots, n\} : X_t = i\big\}$$

le nombre d'étudiants notés par le professeur i (avec par exemple $i = 1$ pour Grincheux et $i = 2$ pour Gentil). Alors, sous l'hypothèses H_0 d'homogénéité, on s'attend à avoir $N_{i,\cdot,n}\, \widehat{p}_L$ étudiants du professeur i obtenant la note L.

Ces quantités $N_{i,\cdot,n}\, \widehat{p}_L$, pour $i \in \{1, 2\}$ et $L \in \{A, \ldots, F\}$ seront nos effectifs attendus.

Exemple. Dans l'exemple fil rouge, cela revient à attendre $99 \times 35/194 \approx 17.9$ notes A données par le professeur Gentil. Nous obtenons également des effectifs attendus $95 \times 32/194 \approx 15.7$ de notes E pour le professeur Grincheux. Plus généralement, nous obtenons les effectifs attendus résumés dans le

tableau suivant :

			A	B	C	D	E	F	Total
Prof	Grincheux	Effectif observé	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 observé	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 observé	35	33	50	37	32	7	194
		Effectif théorique	35,0	33,0	50,0	37,0	32,0	7,0	194,0

Nous confrontons alors conjointement les effectifs observés et attendus, selon les professeurs i et les notes L,

$$N_{i,L,n} = \mathrm{Card}\big\{t \in \{1,\dots,n\} : X_t = i \text{ et } Y_t = L\big\} \qquad \text{vs.} \qquad N_{i,\cdot,n}\,\widehat{p}_L = \frac{N_{i,\cdot,n}\,N_{\cdot,L,n}}{n},$$

d'une manière similaire au cas du test d'ajustement :

$$T_n = \sum_{i=1}^{2} \sum_{L=A}^{F} \frac{\big(N_{i,L,n} - N_{i,\cdot,n}\,\widehat{p}_L\big)^2}{N_{i,\cdot,n}\,\widehat{p}_L}.$$

On peut démontrer que sous l'hypothèse H_0, la statistique de test converge vers une loi du χ^2 à $(r-1)(s-1) = (2-1)(6-1) = 5$ degrés de liberté, tandis que sous H_1, on observe la convergence en probabilité $T_n \to +\infty$. Un schéma de principe similaire à celui de la Figure 6.2 s'applique donc ici.

La valeur réalisée de la statistique de test est ici :

$$t = \frac{(14-17.1)^2}{17.1} + \frac{(15-16.2)^2}{16.2} + \cdots + \frac{(2-3.6)^2}{3.6} \approx 3.109\,.$$

Conditions de validité et lecture de la P–valeur. Les mêmes conditions de validité s'appliquent que pour le test du χ^2 d'ajustement à une loi, à savoir : taille totale d'échantillon $n = 194$ supérieure à 30 (c'est le cas) et effectifs attendus tous supérieurs à 5. Ici, ce n'est pas le cas : on obtient des effectifs attendus de 3.4 et 3.6 pour la note F. Il faut traiter ce problème et nous avons deux solutions possibles :

— regrouper la lettre F avec une autre lettre, le plus logique étant de la regrouper avec le E ;

— supprimer la catégorie F, et ignorer les F donnés par les professeurs.

Les deux corrections sont mises en œuvre dans les sorties logiciel de la page 85.

Dans la première de ces sorties, on retrouve les valeurs $t = 3.109$ et le nombre de degrés de liberté (df : "degrees of freedom") de la loi du χ^2 asymptotiquement suivie par T_n ; mais on ne peut pas exploiter la P–valeur lue, car les conditions d'application sont en défaut, comme le souligne la note de bas de tableau, qui pointe deux cellules problématiques—en l'occurrence, celles pour la note F.

Les deux sorties suivantes correspondent, elles, à des applications valides du test du χ^2 d'homogénéité, et procurent les P–valeurs respectives 67.4% (en cas de fusion des catégories E et F) et 80.1% (en cas de suppression de la catégorie F). Dans les deux cas, nous conservons l'hypothèse H_0.

Des conclusions grand public sont les suivantes (et aucune des deux n'est de nature à justifier les surnoms Grincheux et Gentil accordés) :

> Ces données ne permettent pas d'exclure que les deux professeurs partagent la même échelle de notation.
>
> Ces données ne mettent en évidence aucune différence significative entre les profils de notation des deux enseignants.

Tableau croisé Prof * Note (données originelles)

			Note						
			A	B	C	D	E	F	Total
Prof	Grincheux	Effectif observé	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 observé	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 observé	35	33	50	37	32	7	194
		Effectif théorique	35,0	33,0	50,0	37,0	32,0	7,0	194,0

Test du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	3,109[a]	5	,683
N 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 Prof * Note (fusion E et F)

			Note (fusion E et F)					
			A	B	C	D	E ou F	Total
Prof	Grincheux	Effectif observé	14	15	26	18	22	95
		Effectif théorique	17,1	16,2	24,5	18,1	19,1	95,0
	Gentil	Effectif observé	21	18	24	19	17	99
		Effectif théorique	17,9	16,8	25,5	18,9	19,9	99,0
Total		Effectif observé	35	33	50	37	39	194
		Effectif théorique	35,0	33,0	50,0	37,0	39,0	194,0

Test du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,339[a]	4	,674
N d'observations valides	194		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 16,16.

Tableau croisé Prof * Note (suppression des F)

			Note					
			A	B	C	D	E	Total
Prof	Grincheux	Effectif observé	14	15	26	18	17	90
		Effectif théorique	16,8	15,9	24,1	17,8	15,4	90,0
	Gentil	Effectif observé	21	18	24	19	15	97
		Effectif théorique	18,2	17,1	25,9	19,2	16,6	97,0
Total		Effectif observé	35	33	50	37	32	187
		Effectif théorique	35,0	33,0	50,0	37,0	32,0	187,0

Test du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	1,645[a]	4	,801
N d'observations valides	187		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 15,40.

Utilisations du test du χ^2 d'homogénéité / indépendance

En sciences de gestion : nécessité de segmentation d'une population. En marketing, le test du χ^2 d'homogénéité est utile pour déterminer si les préférences des consommateurs sont homogènes entre différents groupes (femmes–hommes, groupes d'âges, etc.), ou si au contraire, il convient de segmenter le marché et de décliner des produits ou campagnes de publicité différents selon les groupes. Nous illustrerons cela dans l'exercice sur les habitudes de lecture.

De même, on peut tester l'homogénéité de la satisfaction dans les différents groupes constitutifs d'une population, pour mieux comprendre les axes d'amélioration éventuels. Là encore, nous illustrerons cela dans par un exercice.

Auto-défense statistique. Au printemps 2022, les étudiants de filires universitaires m'ont fait remarquer que le test du χ^2 d'homogénéité serait l'outil idéal pour déterminer si les notes-lettres obtenues pour vos filières (GEP ou licences), et comptant pour 23 ECTS dans votre GPA, dépendaient, ou non, de ces filières. Dit autrement, la répartition des notes obtenue dépend-elle du fait d'être parti en GEP ou d'avoir suivi une licence ? J'ai suggéré aux étudiants, pour tester cela, de collecter 100 données uniformément au hasard ainsi : pour chacune des deux promotions précédentes, 25 notes de GEP et 25 notes de licence, pour des étudiants choisis au hasard et pourchassés jusqu'à ce qu'ils répondent. Il faudra sans doute réduire le nombre de catégories de notes en les quatre catégories A, B, C et [D - E - F], voire en trois catégories A, B et [C - D - E - F] afin de respecter la condition que tous les effectifs attendus soient supérieurs à 5.

Variante : second jeu de données de notes

Il se trouve que toutes les mises en œuvre de tests du χ^2 de ce chapitre ont jusque-là conservé H_0. Je voudrais maintenant traiter un exemple où H_0 est rejetée — et où un effet est donc prouvé — pour vous expliquer comment lire les directions d'effet.

Considérons les répartitions de notes suivantes :

		Note [2nd jeu de données]						
		A	B	C	D	E	F	Total
Prof	Grincheux v2	12	11	17	25	21	9	95
	Gentil v2	21	18	24	19	15	2	99
Total		33	29	41	44	36	11	194

En le traitant avec un logiciel statistique, nous obtenons la sortie logiciel de la page 87. Le test du χ^2 d'homogénéité est validement appliqué (taille totale d'échantillon $n = 194 \geqslant 30$ et tous les effectifs attendus $\geqslant 5$, comme indiqué en note de bas de tableau), et il renvoie la P–valeur 4.2%. On rejette donc l'hypothèse d'homogénéité :

> Ces données montrent que les deux enseignants utilisent des échelles de notation différentes, i.e., que les répartitions des notes obtenues sont différentes entre eux.

Lecture des directions d'effet. Un effet a été démontré et a été traduit dans la conclusion grand public ci-dessus. Mais comment lire les directions d'effet ? Dans chaque cellule, il s'agit de comparer les effectifs observés à ceux attendus. Par exemple, pour la note A et le professeur Grincheux, 16.2 notes A étaient attendues sous une hypothèse d'homogénéité des notations et seules 12 sont observées : il y

Tableau croisé Prof * Note [2nd jeu de données]

			Note						
			A	B	C	D	E	F	Total
Prof	Grincheux v2	Effectif observé	12	11	17	25	21	9	95
		Effectif théorique	16,2	14,2	20,1	21,5	17,6	5,4	95,0
	Gentil v2	Effectif observé	21	18	24	19	15	2	99
		Effectif théorique	16,8	14,8	20,9	22,5	18,4	5,6	99,0
Total		Effectif observé	33	29	41	44	36	11	194
		Effectif théorique	33,0	29,0	41,0	44,0	36,0	11,0	194,0

Test du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	11,534[a]	5	,042
N d'observations valides	194		

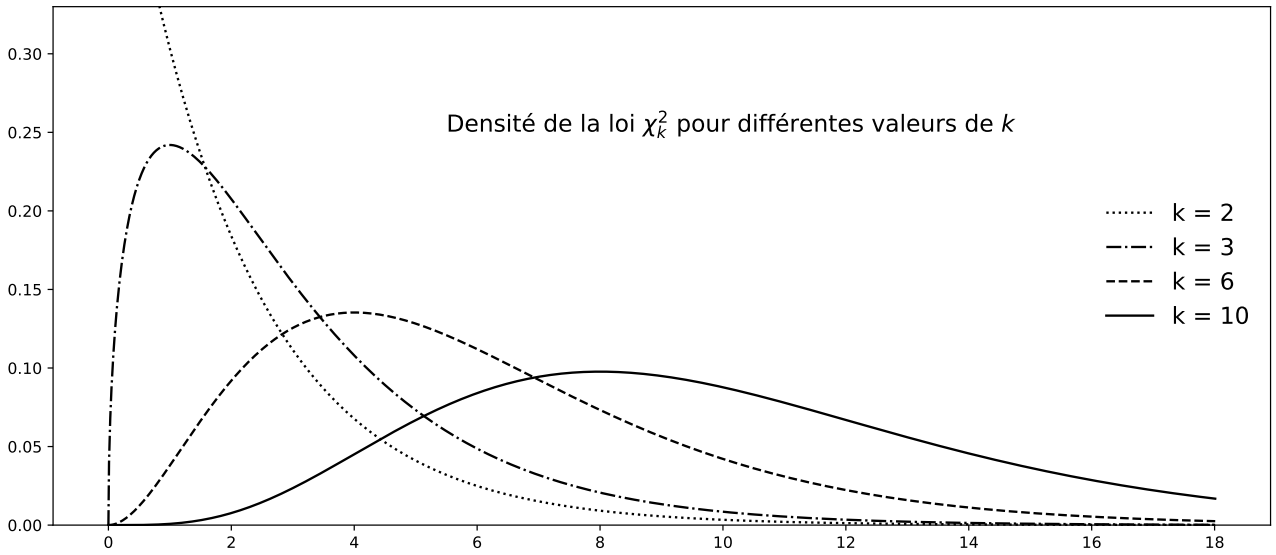
a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 5,39.

a une sous-représentation des A. De même, pour ce professeur, il y a une sous-représentations des B et des C, et une sur-représentation des notes D - E - F : ce professeur est plus sévère dans ses notations (plus sévère que le point de comparaison donné par les répartitions globales de notes, et donc plus sévère que l'autre professeur).

3. Appendice : définition de la loi du χ^2

Définition. *Si* $Z_1, \dots, Z_k$ *sont des variables aléatoires indépendantes et identiquement distribuées selon la la loi normale standard* $\mathcal{N}(0,1)$*, 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^2_k.

Voici le tracé de la fonction de densité de probabilité de quelques lois χ^2_k.



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 l'automne 2012). Les 200 étudiants sondés et l'ensemble des 2 983 étudiants alors en cours de scolarité (répartition de population fournie par la direction d'HEC) se répartissent respectivement de la manière suivante, selon le programme et le sexe :

	Hommes	Femmes	Total
Grande Ecole	78	73	151
Master	17	12	29
MBA	10	7	17
Doctorat	1	2	3
Total	106	94	200

vs.

	Hommes	Femmes	Total
Grande Ecole	39.4%	31.3%	70.7%
Master	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%

Enoncez les hypothèses testées dans la sortie logiciel suivante, expliquez la méthodologie suivie, et indiquez quel est le résultat final. Procurez une conclusion lisible par le grand public et que la liste Hubert aura, espérons-le, pris le soin d'écrire en avant-propos de son rapport.

Test / Première version

	Effectif observé	N théorique	Résidus
Grande Ecole / Homme	78	78,8	-,8
Grande Ecole / Femme	73	62,6	10,4
Master / Homme	17	18,8	-1,8
MAster / Femme	12	14,0	-2,0
MBA / Homme	10	13,0	-3,0
MBA / Femme	7	7,2	-,2
Doctorat / Homme	1	3,0	-2,0
Doctorat / Femme	2	2,6	-,6
Total	200		

Khi-carré	4,364[a]
df	7
Sig. asymptotique	,737

a. 2 cellules ont des effectifs th®oriques inf®rieurs ¨ 5

Test / Seconde version

	Effectif observé	N théorique	Résidus
Grande Ecole / Homme	78	78,8	-,8
Grande Ecole / Femme	73	62,6	10,4
Master / Homme	17	18,8	-1,8
Master / Femme	12	14,0	-2,0
MBA / Homme	10	13,0	-3,0
MBA / Femme	7	7,2	-,2
Doctorat / H+F	3	5,6	-2,6
Total	200		

Khi-carré	4,099[a]
df	6
Sig. asymptotique	,663

a. 0 cellules ont des effectifs th®oriques inf®rieurs ¨ 5

Temps d'attente téléphonique pour joindre un service clients

Un opérateur Internet chercher 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 pour voir si, conformément aux objectifs qu'il s'était fixé, au moins 50% des clients attendent moins de 2 minutes et 90% des clients attendent moins de 5 minutes. Ses deux centres d'appels sont testés par des « clients mystère », qui effectuent chacun 50 appels sur chaque centre, à des jours et heures variés durant deux semaines. Les résultats et leur traitement statistique sont procurés ci-dessous :

Premier centre d'appel

Temps d'attente

	Effectif observé	N théorique	Résidus
< 2 min	58	55,5	2,5
2 - 5 min	44	44,4	-,4
> 5 min	9	11,1	-2,1
Total	111		

Khi-carré	,514[a]
df	2
Sig. asymptotique	,774

a. 0 cellules (0,0%) ont des effectifs th®oriques inf®rieurs ¨ 5

Second centre d'appel

Temps d'attente

	Effectif observé	N théorique	Résidus
< 2 min	58	52,0	6,0
2 - 5 min	44	41,6	2,4
> 5 min	2	10,4	-8,4
Total	104		

Khi-carré	7,615[a]
df	2
Sig. asymptotique	,022

a. 0 cellules (0,0%) ont des effectifs th®oriques inf®rieurs ¨ 5

Indiquez quel test est mis en œuvre, énoncez les hypothèses testées (en mots et de manière plus mathématique), précisez si le test est validement appliqué, et indiquez quel est le résultat final. Procurez des conclusions lisibles par le grand public (et des conclusions stratégiques pour l'opérateur Internet).

Sociologie des habitudes de lecture

Cet exercice s'inspire des résultats de l'enquête permanente de l'INSEE sur les modes de vie des ménages, mais en divisant la taille d'échantillon par un facteur 10. On s'intéresse au nombre de livres lus par les Français au cours de l'année écoulée, et on les catégorise à cet effet par CSP. Les données sont traitées par un logiciel statistique comme indiqué ci-contre.

1. Enoncez les hypothèses testées et expliquez la méthodologie suivie ; en particulier, peut-on dire quelque chose du nombre de livres lus par les agriculteurs ?
2. Indiquez le résultat final et procurez une conclusion lisible par le grand public et suffisamment informative.
3. Supposez que vous soyez une éditrice : quelle(s) action(s) devriez-vous mener ?

Traitement #1

			Lecture			
			Un livre par mois ou plus	Moins d'un livre par mois	Aucun livre dans l'année écoulée	Total
CSP	Agriculteurs	Effectif observé	3	5	10	18
		Effectif attendu	2,8	7,6	7,6	18,0
	Artisans, commerçants, chefs d'entreprise	Effectif observé	6	18	25	49
		Effectif attendu	7,7	20,7	20,6	49,0
	Cadres et professions intellectuelles supérieures	Effectif observé	36	70	23	129
		Effectif attendu	20,3	54,5	54,3	129,0
	Professions intermédiaires	Effectif observé	35	102	58	195
		Effectif attendu	30,6	82,3	82,1	195,0
	Employés	Effectif observé	37	117	94	248
		Effectif attendu	39,0	104,7	104,4	248,0
	Ouvriers	Effectif observé	9	56	131	196
		Effectif attendu	30,8	82,7	82,5	196,0
	Retraités	Effectif observé	76	162	221	459
		Effectif attendu	72,1	193,7	193,1	459,0
	Autres inactifs	Effectif observé	34	104	70	208
		Effectif attendu	32,7	87,8	87,5	208,0
Total		Effectif observé	236	634	632	1502
		Effectif attendu	236,0	634,0	632,0	1502,0

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	121,562[a]	14	<,001
N d'observations valides	1502		

a. 1 cellules (4,2%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 2,83.

Traitement #2

			Lecture			
			Un livre par mois ou plus	Moins d'un livre par mois	Aucun livre dans l'année écoulée	Total
CSP	Artisans, commerçants, chefs d'entreprise	Effectif observé	6	18	25	49
		Effectif attendu	7,7	20,8	20,5	49,0
	Cadres et professions intellectuelles supérieures	Effectif observé	36	70	23	129
		Effectif attendu	20,3	54,7	54,1	129,0
	Professions intermédiaires	Effectif observé	35	102	58	195
		Effectif attendu	30,6	82,7	81,7	195,0
	Employés	Effectif observé	37	117	94	248
		Effectif attendu	38,9	105,1	103,9	248,0
	Ouvriers	Effectif observé	9	56	131	196
		Effectif attendu	30,8	83,1	82,2	196,0
	Retraités	Effectif observé	76	162	221	459
		Effectif attendu	72,1	194,5	192,4	459,0
	Autres inactifs	Effectif observé	34	104	70	208
		Effectif attendu	32,7	88,2	87,2	208,0
Total		Effectif observé	233	629	622	1484
		Effectif attendu	233,0	629,0	622,0	1484,0

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	120,019[a]	12	<,001
N d'observations valides	1484		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 7,69.

Enquête de satisfaction

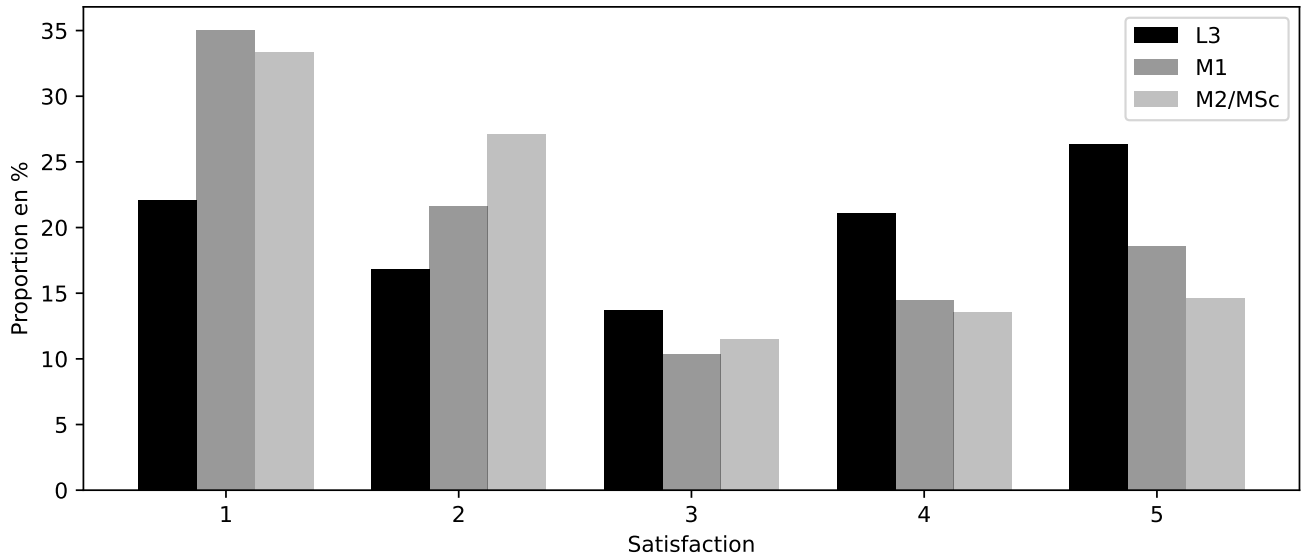
Cet exercice est inspiré par un mémoire écrit vers le début des années 2010 par un étudiant d'HEC dans le cadre de sa majeure de journalisme (avec le CFJ).

Il avait pris pour sujet la satisfaction des étudiants d'HEC par rapport au programme Grande Ecole, et avait notamment catégorisé les retours (qualitatifs et quantitatifs) en fonction de l'année à laquelle une formation de la Grande Ecole avait été intégrée : dès la L3 (étudiants français issus des classes préparatoires, vous, donc) ; en M1 (admis directs internationaux et français) ; en M2/MSc (étudiants venant suivre un master spécialisé, en général après avoir obtenu un diplôme d'ingénieur dans une autre école).

Le tout-HEC s'était évidemment arraché ce mémoire, qui a marqué durablement son époque par les sujets évoqués et les verbatims parfois douloureusement révélateurs. Ce mémoire, en revanche, était indigent sur le plan quantitatif, et je ne peux que le déplorer, alors même qu'HEC offre une formation en sciences de gestion et management axée sur le côté quantitatif...

En particulier, le mémoire résumait ainsi les réponses obtenues auprès d'un échantillon d'étudiants à la question de leur satisfaction globale quant à la formation suivie à HEC, sur une échelle de 1 (satisfaction la plus basse) à 5 (satisfaction la plus haute), par un tableau et une illustration graphique.

			Satisfaction					
			5 (haute)	4	3	2	1 (basse)	Total
Année d'intégration	L3	Effectif observé	21	16	13	20	25	95
		% dans Année d'intégration	22,1%	16,8%	13,7%	21,1%	26,3%	100,0%
	M1	Effectif observé	34	21	10	14	18	97
		% dans Année d'intégration	35,1%	21,6%	10,3%	14,4%	18,6%	100,0%
	M2 / MSc	Effectif observé	32	26	11	13	14	96
		% dans Année d'intégration	33,3%	27,1%	11,5%	13,5%	14,6%	100,0%
Total		Effectif observé	87	63	34	47	57	288
		% dans Année d'intégration	30,2%	21,9%	11,8%	16,3%	19,8%	100,0%



Le mémoire écrivait : « On voit bien que la satisfaction varie selon l'année d'intégration. » Pourquoi cette impression se dégage-t-elle (i.e., quels nombres comparez-vous dans le tableau) ? Et pouvait-il conclure avec les seuls éléments ci-dessus à cette variabilité ?

Après consultation d'un professeur de statistique (appelons-le G. Statz), les responsables de la Grande Ecole étudient plutôt la sortie logiciel suivante.

			Satisfaction					
			5 (haute)	4	3	2	1 (basse)	Total
Année d'intégration	L3	Effectif observé	21	16	13	20	25	95
		Effectif attendu	28,7	20,8	11,2	15,5	18,8	95,0
	M1	Effectif observé	34	21	10	14	18	97
		Effectif attendu	29,3	21,2	11,5	15,8	19,2	97,0
	M2 / MSc	Effectif observé	32	26	11	13	14	96
		Effectif attendu	29,0	21,0	11,3	15,7	19,0	96,0
Total		Effectif observé	87	63	34	47	57	288
		Effectif attendu	87,0	63,0	34,0	47,0	57,0	288,0

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	11,302[a]	8	,185
N d'observations valides	288		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 11,22.

1. Enoncez les hypothèses testées. Le test est-il validement appliqué ?
2. Le traitement statistique ci-dessus prouve-t-il que la satisfaction varie selon l'année d'intégration ? Fournissez une réponse argumentée et suffisamment détaillée, y compris une conclusion de lecture agréable à l'ensemble des collaborateurs de la Grande Ecole.

Consommation d'alcool lors des POWs d'HEC, version 2

Nous reprenons sous un nouvel angle des données déjà traitées dans le cadre des tests de comparaison de deux moyennes indépendante et obtenons la sortie logiciel ci-dessous :

			Nombre de verres			
			0 - 3 (inclus)	3 - 6 (exclus)	6 (inclus) et +	Total
Groupe	8h-9h45	Effectif observé	8	10	5	23
		Effectif attendu	8,8	7,5	6,6	23,0
	10h-11h45	Effectif observé	12	7	10	29
		Effectif attendu	11,2	9,5	8,4	29,0
Total		Effectif observé	20	17	15	52
		Effectif attendu	20,0	17,0	15,0	52,0

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	2,335[a]	2	,311
N d'observations valides	52		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 6,63.

Indiquez quel test est mis en œuvre, et énoncez les hypothèses testées, de telle sorte qu'il soit clair quels objets ou quantités mathématiques sont ici comparés (étant entendu qu'il ne s'agit plus de comparer des moyennes !).

Précisez si le test est validement appliqué, indiquez quel est le résultat technique final, et rédigez une conclusion lisible par le grand public (façon de parler : aucun d'entre nous n'a envie que le grand public se penche sur le douloureux problème de la consommation d'alcool à HEC).

Comment les effectifs de 5 et 6.6 ont-ils été obtenus pour les étudiants de 8h ayant bu plus de 6 verres ?

7

Régression linéaire simple

Dans ce chapitre, nous considérons des couples de données quantitatives $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$.

Exemple. Dans notre exemple fil rouge, représenté à la Figure 7.1, les individus statistiques sont des appartements, la population est l'ensemble des appartements de Versailles, et un échantillon de $n = 42$ appartements a été constitué de manière aléatoire en 2013 par un intervenant de ce cours, Benjamin Petiau. Les deux variables d'intérêt sont, pour chaque appartement j de l'échantillon, sa surface x_t (reportée en m^2) et son prix initial à la vente y_t (reporté en milliers d'euros, kEuros).

On appelle y la variable à expliquer et x, la variable explicative. Il s'agit évidemment d'une explication « statistique » (à quel point peut-on déduire la valeur de y de celle de x ?) et pas d'une explication causale.

Pour des raisons diverses, et au moins en première approche, nous étudions l'intérêt de modéliser linéairement (en réalité, de manière affine) y en fonction de x : nous pensons à des « explications statistiques » et à une réalité sous-jacente respectivement de la forme

$$\widehat{y}_t = a + b\,x_t \qquad \text{et} \qquad y_t = \alpha_0 + \beta\,x_t + e_t\,, \qquad \text{pour} \qquad j \in \{1, \ldots, n\},$$

où a et b sont des coefficients estimés (ordonnée à l'origine et pente), tandis que α_0 et β_0 désignent des coefficients postulés par un modèle stochastique. Ce dernier suppose que

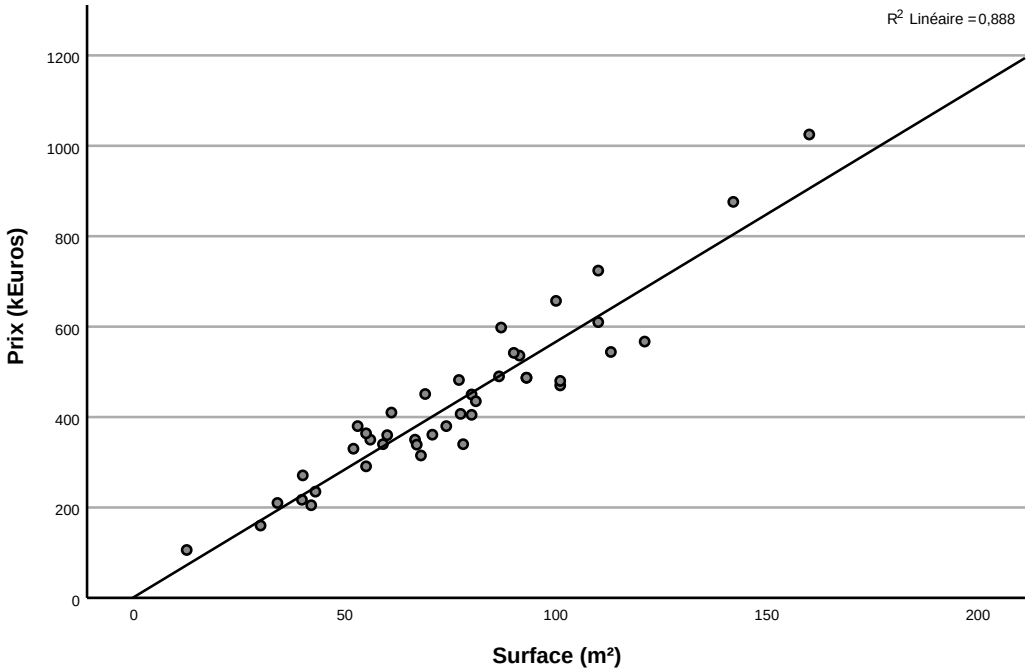
> les observations y_t sont la réalisation de variables aléatoires $Y_t = \alpha_0 + \beta\,x_t + \varepsilon_t$,

où les variables aléatoires ε_t, de réalisations e_t, sont appelés des termes de résidus : des écarts entre la partie modélisée $\alpha_0 + \beta_0\,x_t$, fonction déterministe de la variable explicative, et les valeurs observées Y_t, de réalisations y_t. Les coefficients α_0 et β_0 sont inconnus. Les termes de résidus sont imputables au fait que de nombreux facteurs autres que la variable explicative influent sur la variable à expliquer.

Exemple. Pour les prix des appartements, la surface ne fait pas tout, et des considérations comme l'existence ou l'absence d'un balcon, la sectorisation sur des écoles réputées, la proximité de transports en commun, l'existence d'un parking souterrain, etc., font varier le prix, à la hausse ou à la baisse. C'est l'impact combiné de tous ces facteurs que l'on cache pudiquement dans le terme de résidu.

Pourquoi modéliser ? L'objectif est ici de rendre compte de manière simple d'une relation possiblement complexe, par une relation affine. Cela permet de mieux appréhender le phénomène complexe :

> Tous les modèles sont faux mais certains sont utiles.
> Georges Box (statisticien américain, 1919–2013)



Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,942[a]	,888	,885	60,665

a. Prédicteurs : (Constante), Surface (m²)

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1163593,829	1	1163593,829	316,171	<,001[b]
	Résidus	147210,576	40	3680,264		
	Total	1310804,405	41			

a. Variable dépendante : Prix (kEuros)

b. Prédicteurs : (Constante), Surface (m²)

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	1,861	25,805		,072	,943
	Surface (m²)	5,648	,318	,942	17,781	<,001

a. Variable dépendante : Prix (kEuros)

FIGURE 7.1 – Nuage de points des données (haut) et analyse de régression linéaire produite par un logiciel statistique (bas). Le logiciel écrit « variable dépendante » (traduction littérale de “dependent variable” en anglais) au lieu de variable à expliquer (et par opposition à “independent variable”, qui est la variable explicative).

Une question cruciale, à laquelle nous répondrons, sera évidemment : à quel point rend-on compte, par une relation affine, du phénomène complexe ? Il faut en tout cas retenir que l'on cherche un compromis entre simplicité de l'explication (une relation affine, c'est simple !) et pouvoir explicatif : on veut quand même que les explications statistiques $\widehat{y}_t$, que nous appellerons aussi reconstructions, soient proches des observations y_t.

Pourquoi modéliser de manière affine ? On pense à des modélisations affines notamment parce que l'on trace le nuage de points et qu'on y lit une direction affine. Les exercices illustreront que parfois on y lit une relation logarithmique ou une autre relation connue, auquel cas on effectuera au préalable une pré-transformation sur la variable explicative. Par exemple, on introduira $x'_t = \ln x_t$ et on cherche des relations affines en x'_t, donc de la forme

$$\widehat{y}_t = a + b\, x'_t = \widehat{y}_t = a + b\, \ln x_t\,.$$

1. Statistiques descriptives

Dans cette partie, nous n'avons pas besoin de voir les y_t comme la réalisation de variables aléatoires Y_t : nous nous contentons d'un point de vue de statistiques descriptives, de niveau lycée – classes préparatoires. La partie suivante expliquera comment injecter un point de vue de statistique inférentielle, ce qui est évidemment l'objectif dans ce cours.

1.1. Droite de régression

La droite de régression $y = a + bx$ est obtenue par minimisation au sens des moindres carrés[1] Formellement, les coefficients a et b sont les paramètres minimisant la somme des écarts quadratiques par rapport à l'ensemble des droites possibles, paramétrées de manière générique par des coefficients u et v :

$$E_n(a,b) = \min_{(u,v)\in\mathbb{R}^2} E_n(u,v)\,, \qquad \text{où} \qquad E_n : (u,v) \in \mathbb{R}^2 \longmapsto \sum_{t=1}^{n} \left(y_t - (u + v x_t)\right)^2. \tag{7.1}$$

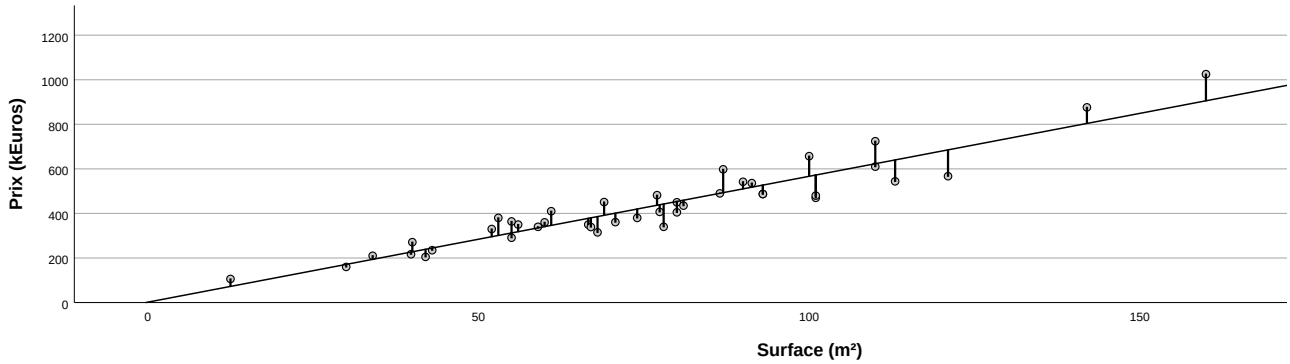
Des expressions explicites existent pour a et b (et sont procurées en appendice, avec une preuve de leur dérivation). Ces expressions sont complexes et difficiles à calculer manuellement, nous nous contenterons de les lire dans les sorties de logiciel. En l'occurrence, les valeurs de a et b sont situées dans le troisième tableau, ici $a \approx 1.861$ et $b \approx 5.648$:

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	1,861				
	Surface (m²)	5,648				

a. Variable dépendante : Prix (kEuros)

1. 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 certes calculer numériquement des solutions à des problèmes n'admettant pas solution générale explicite, comme pour le critère de moindres valeurs absolues. Mais le poids de l'Histoire pousse pour le critère de moindres carrés...

La droite de régression est tracée en haut de la Figure 7.1 et également ci-dessous. Pour chaque abscisse x_t, on rappelle qu'on a noté y_t l'observation correspondante (marquée par un $\circ$ dans le nuage de points) et par $\widehat{y}_t = a + b$ l'explication statistique (la reconstruction) correspondante. Les résidus signés $e_t = y_t - \widehat{y}_t$ sont représentés ci-dessous : ils correspondent à l'écart signé entre l'observation et la droite de régression.



On peut démontrer (en notant que l'on a affaire à une projection orthogonale, voir appendice) que la somme des résidus signés est nulle :

$$\sum_{t=1}^{n} e_t = \sum_{t=1}^{n} (y_t - \widehat{y}_t) = 0\,, \qquad \text{ou encore,} \qquad \overline{y}_n = \frac{1}{n}\sum_{t=1}^{n} \widehat{y}_t\,, \tag{7.2}$$

i.e., les observations y_t et leurs reconstructions $\widehat{y}_t$ ont même moyenne. Les barres au-dessus et en-dessous de la droite de régression se compensent donc dans le graphique. Une conséquence intéressante de ce fait est :

La droite de régression passe par le point moyen $(\overline{x}_n, \overline{y}_n)$ du nuage de points.

1.2. Qualité de l'ajustement : le critère du r^2

Nous introduisons les sommes de carrés Σ_{tot}, Σ_{regr} et Σ_{res}, qui mesurent respectivement une variabilité totale, une variabilité retrouvée par la régression, et une variabilité résiduelle :

$$\underbrace{\sum_{t=1}^{n} (y_t - \overline{y}_n)^2}_{=\Sigma_{tot}} = \underbrace{\sum_{t=1}^{n} (\widehat{y}_t - \overline{y}_n)^2}_{=\Sigma_{regr}} + \underbrace{\sum_{t=1}^{n} (y_t - \widehat{y}_t)^2}_{=\Sigma_{res}}$$

(où l'égalité $\Sigma_{tot} = \Sigma_{regr} + \Sigma_{res}$ procède de la projection orthogonale sous-jacente à la définition de la droite de régression, voir appendice). La somme Σ_{tot} correspond à une mesure de la variabilité des y_t autour de leur moyenne $\overline{y}_n$, on l'interprète comme une variabilité [2] totale. La somme Σ_{regr} se rapporte à une variable retrouvée par la régression, correspondant à la variabilité des $\widehat{y}_t$ autour de leur moyenne également égale à $\overline{y}_n$ selon (7.2). Enfin, Σ_{res}, la somme des résidus au carrés, correspond à la variabilité des résidus autour de leur moyenne, qui est nulle.

On introduit le coefficient de détermination

$$r^2 = \frac{\Sigma_{regr}}{\Sigma_{tot}}\,, \qquad \text{la part de variabilité totale retrouvée par la régression.}$$

2. Une sorte de variance au terme $1/(n-1)$ près, même si, comme nous le verrons ci-dessous, les y_t ne sont pas issues de variables aléatoires Y_t de même loi.

Il est toujours compris entre 0% et 100% et sera notre indicateur de qualité du modèle. Les valeurs du coefficient r^2, de sa racine carrée $\sqrt{r^2}$, ainsi que des sommes de carrés Σ_{tot}, Σ_{regr} et Σ_{res} peuvent être lues dans les deux premiers tableaux de la page 96 :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,942[a]	,888		

a. Prédicteurs : (Constante), Surface (m²)

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1163593,829				
	Résidus	147210,576				
	Total	1310804,405				

a. Variable dépendante : Prix (kEuros)

En l'occurrence, ici $r^2 = 88.8\%$, ce que l'on peut traduire ainsi :

> 88.8% des (variations de) prix sont expliqué(e)s par les (variations de) surfaces.

Remarque 7.1 (Même un r^2 de 20% peut être très bien !). On ne s'attend pas toujours à des r^2 aussi élevés. Le marché immobilier est particulier parce que tous ses acteurs raisonnent fondamentalement en prix au m^2, donc de manière linéaire. Pour des phénomènes plus complexes, notamment économiques, nous pourrons déjà être heureux avec un r^2 de l'ordre de 20%, lorsqu'il est obtenu avec une variable explicative particulièrement simple à appréhender.

Transition : statistiques descriptives → statistique inférentielle. Ainsi s'arrête l'approche de statistiques descriptives et commence la vraie nouveauté par rapport à vos cours précédents : une approche inférentielle, permettant de réaliser des tests d'hypothèses et de fournir des intervalles de confiance, notamment sur les coefficients α_0 et β_0.

2. Approche de statistique inférentielle

Nous allons effectuer une hypothèse relativement forte de modélisation : que les résidus ε_t, i.e., les écarts aux parties modélisées $\alpha_0 + \beta_0 x_t$, soient indépendants et identiquement distribués (i.i.d.) selon une loi normale standard $\mathcal{N}(0, \sigma_0^2)$, où σ_0^2 est une variance inconnue. On parle de modèle linéaire gaussien :

> Les observations y_t sont la réalisation de $Y_t = \alpha_0 + \beta_0 x_t + \varepsilon_t$, où les résidus $\varepsilon_1, \ldots, \varepsilon_n$ sont i.i.d. selon une loi $\mathcal{N}(0, 1)$.

Nous avons donc affaire à trois paramètres d'intérêt : α_0, β_0 et σ_0, qu'il conviendra d'estimer et de tester.

Justification (facultative) de l'hypothèse de modèle linéaire gaussien. Le caractère indépendant provient en général du fait que les individus n'ont pas d'influence les uns sur les autres : ici, le

fait que tel appartement soit plus cher que la moyenne $\alpha_0 + \beta_0 x_t$ ne dépend que de ses caractéristiques propres et n'implique que tel autre appartement sera plus ou moins cher. Le fait que les résidus suivent une loi normale procède de versions généralisées du théorème limite central, qui indiquent qu'une somme de nombreux facteurs indépendants (balcon ou pas, garage ou pas, affectation scolaire, charme, exposition, etc.) conduit à une loi normale. L'hypothèse la plus forte est en réalité l'homoscedasticité : le fait que la variance des résidus ε_t ne dépende pas des variables explicatives x_t. Cela se traduit sur le nuage de points par le fait que les écarts $y - \widehat{y}$ à la droite de régression sont du même ordre de grandeur quelle que soient les ordres de grandeurs des x. C'est une hypothèse que l'on peut vérifier a posteriori, par diverses techniques, appelées « analyse des résidus » et que nous ne traiterons pas dans le cadre de ce cours.

2.1. Quelques éléments de raisonnement mathématique

Je ne veux vous donner ici que quelques éléments de raisonnement. Etudier de manière approfondie le modèle linéaire gaussien et prouver l'ensemble des résultats nécessaires à la lecture des sorties de logiciel statistique prend un semestre entier (par exemple en L3 d'économie ou en M1 de mathématiques).

On note $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ les estimateurs dont les coefficients a et b définis en (7.1) sont les réalisations, et $\widehat{Y}_t = \widehat{\alpha}_n + \widehat{\beta}_n x_t$, qui admettent pour réalisations les $\widehat{y}_t$.

Vu les formules explicites, $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ apparaissent comme des combinaisons linéaires des variables aléatoires indépendantes Y_t, chacun distribuée selon une loi normale (d'espérances certes différentes : $\alpha_0 + \beta_0 x_t$). Par conséquent, $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ suivent des lois normales, et il n'est pas difficile de calculer leurs paramètres : en l'occurrence, leurs espérances respectives sont α_0 et β_0 (i.e., ces deux estimateurs sont sans biais) et leurs écarts-types sont calculables, mais dépendent de σ_0, qui est inconnu.

Ces écarts-types sont de la forme $\sigma_0 S_{\alpha,n}$ et $\sigma_0 S_{\beta,n}$, où les quantités $S_{\alpha,n}$ et $S_{\beta,n}$ sont parfaitement connues et ne dépendent que des variables explicatives $x_1, \dots, x_n$. (Voir les formules pour $S_{\alpha,n}$ et $S_{\beta,n}$ en appendice.) On estime alors σ_0 par

$$\sqrt{\widehat{\sigma}_n^2}, \qquad \text{où} \qquad \widehat{\sigma}_n^2 = \frac{1}{n-2} \sum_{t=1}^{n} (Y_t - \widehat{Y}_t)^2 .$$

On peut montrer que $\widehat{\sigma}_n^2$ est un estimateur sans biais de σ_0^2, car en fait, $(n-2)\widehat{\sigma}_n^2/\sigma_0^2$ suit une loi du χ^2 à $n-2$ degrés de liberté. En outre, par le théorème de projection orthogonale de Cochran, $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

Ces considérations permettent alors de montrer que les deux statistiques de test suivantes suivent des lois de Student à $n-2$ degrés de liberté :

$$\frac{\widehat{\alpha}_n}{\widehat{\sigma}_n S_{\alpha,n}} \sim \mathcal{T}_{n-2} \qquad \text{et} \qquad \frac{\widehat{\beta}_n}{\widehat{\sigma}_n S_{\beta,n}} \sim \mathcal{T}_{n-2} ,$$

ce qui permet d'une part, de construire des tests pour

$$H_0 : \alpha_0 = 0 \ \text{ vs. } \ H_1 : \alpha_0 \neq 0 \qquad \text{et} \qquad H_0 : \beta_0 = 0 \ \text{ vs. } \ H_1 : \beta_0 \neq 0 ,$$

et d'autre part, des intervalles de confiance

$$[\widehat{\alpha}_n \pm t_{n-2,97.5\%}\, \widehat{\sigma}_n S_{\alpha,n}] \qquad \text{et} \qquad [\widehat{\beta}_n \pm t_{n-2,97.5\%}\, \widehat{\sigma}_n S_{\beta,n}]$$

sur α_0 et β_0, où $t_{n-2,97.5\%} \approx 1.96$ désigne le quantile à 97.5% de la loi $\mathcal{T}_{n-2}$.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,942[a]	,888	,885	60,665

a. Prédicteurs : (Constante), Surface (m²)

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1163593,829	1	1163593,829	316,171	<,001[b]
	Résidus	147210,576	40	3680,264		
	Total	1310804,405	41			

a. Variable dépendante : Prix (kEuros)

b. Prédicteurs : (Constante), Surface (m²)

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	1,861	25,805		,072	,943	-50,293	54,015
	Surface (m²)	5,648	,318	,942	17,781	<,001	5,006	6,290

a. Variable dépendante : Prix (kEuros)

Correspondance entre la sortie logiciel (haut) et les formules exhibées (bas) ; les cases avec ✗ sont omises pour l'instant et la plupart d'entre elles seront définies au chapitre prochain

Récapitulatif des modèles

R	R-deux	R-deux ajusté	Erreur standard de l'estimation
$\sqrt{r^2}$	$r^2 = \Sigma_{regr}/\Sigma_{tot}$	✗	$\widehat{\sigma}_n$

ANOVA

Modèle	Somme des carrés	ddl	Carrés moyens	F	Sig.
Régression	Σ_{regr}	1	Σ_{regr}	✗	✗
Résidu	Σ_{res}	$n-2$	$\Sigma_{res}/(n-2)$		
Total	Σ_{tot}	$n-1$			

Coefficients

	Coefficients non standardisés					Intervalle de confiance à 95% pour B
Modèle	B	Err. std.	Beta	t	Sig.	
Constante	$\widehat{\alpha}_n$	$\widehat{\sigma}_n\, S_{\alpha,n}$		$\widehat{\alpha}_n/(\widehat{\sigma}_n\, S_{\alpha,n})$	P–val. $H_0 : \alpha_0 = 0$	$[\widehat{\alpha}_n \pm t_{n-2,97.5\%}\, \widehat{\sigma}_n\, S_{\alpha,n}]$
Var. explic.	$\widehat{\beta}_n$	$\widehat{\sigma}_n\, S_{\beta,n}$	✗	$\widehat{\beta}_n/(\widehat{\sigma}_n\, S_{\beta,n})$	P–val. $H_0 : \beta_0 = 0$	$[\widehat{\beta}_n \pm t_{n-2,97.5\%}\, \widehat{\sigma}_n\, S_{\beta,n}]$

2.2. Intervalles de confiance / Prévision de nouvelles valeurs

Intervalles de confiance. Les sorties logiciel fournissent des intervalles de confiance de niveau 95% sur α_0 et β_0. Par combinaison (et application de la correction de Bonferroni), on peut évidemment en déduire, pour toute valeur x de la variable explicative, un intervalle de confiance sur $\alpha_0 + \beta_0 x$. Mais il y a une formule directe plus efficace, voir appendice.

Dans notre exemple fil rouge, cela correspond à intervalle de confiance sur le prix moyen $\alpha_0 + \beta_0 x$ (de l'ensemble) des appartements de surface x en m^2. Notez bien ici l'utilisation du pluriel.

Les intervalles de confiance portent sur des moyennes effectuées sur un grand nombre d'individus. Au contraire, les intervalles de prévision portent sur des valeurs individuelles.

Intervalles de prévision. Une question naturelle, dans notre exemple fil rouge, est :

> Quelle est la fourchette de prix attendue pour un appartement de surface x en m^2 ?

Notez bien ici l'utilisation du singulier : on s'intéresse au prix d'un appartement donné. Evidemment, la variabilité sera ici bien plus grande que pour les intervalles de confiance : il y a d'une part la variabilité causée par l'estimation (mais comme dans le cas des intervalles de confiance), mais aussi et surtout, d'autre part, la variabilité intrinsèque liée au terme de résidu. C'est cette seconde variabilité qui l'emporte, de loin.

On peut reformuler mathématiquement la question comme :

> Proposer un intervalle aléatoire $\widehat{I}$ tel que pour une nouvelle observation $Y_x = \alpha_0 + \beta_0 x + \varepsilon_x$ associée à une valeur x de la variable explicative, on ait :
>
> $$\mathbb{P}(Y_x \in \widehat{I}) \approx 95\% .$$

L'appendice fournit une formule explicite pour un tel intervalle, appelé intervalle de prévision parce qu'il porte sur une valeur individuelle Y_x, et non sur une moyenne de population comme les intervalles de confiance.

Nous utiliserons la formule approchée suivante :

$$\widehat{I} \approx \underbrace{\widehat{\alpha}_n + \widehat{\beta}_n x}_{=\widehat{Y}_x} \pm 2\widehat{\sigma}_n .$$

Cette approximation est très naturelle : on néglige l'erreur d'estimation de $\alpha_0 + \beta_0 x$ par $\widehat{Y}_x$ et celle de σ_0 par $\widehat{\sigma}_n$, et on tient en revanche grand compte du fait que le résidu ε_x, qui suit une loi $\mathcal{N}(0, \sigma_0^2)$, prend avec probabilité 95% des valeurs entre $-1.96\sigma_0$ et $1.96\sigma_0$, que l'on approche par $-2\widehat{\sigma}_n$ et $2\widehat{\sigma}_n$.

3. Exploitation pratique des sorties logiciel

Les fondements étant posés, nous allons maintenant expliquer quels élements exploiter dans les sorties logiciel pour aboutir à des interprétations et conclusions parlant au cœur du grand public (et de vos futurs employeurs). Nous découpons la démarche de lecture en cinq temps.

Etape 1 : validité statistique du modèle. Par « validité statistique », on entend l'existence d'une relation linéaire significative : la variable explicative contribue-t-elle significativement à l'explication statistique de la variable à expliquer ? Vu la relation postulée, $Y_t = \alpha_0 + \beta_0 x_t + \varepsilon_t$, se demander si les variables explicatives x_t contribuent à l'explication revient à se demander si $\beta_0 \neq 0$. Vu les hypothèses testées (l'égalité est forcément dans H_0) :

Un modèle de régression linéaire est statistiquement valide si la variable explicative contribue significativement à l'explication, i.e., si l'on peut rejeter $H_0 : \beta_0 = 0$ au profit de $H_1 : \beta \neq 0$.

Dans notre exemple fil rouge, on se demande donc si le coefficient de pente estimé 5.648 est significativement différent de 0, et c'est le cas, car la t–valeur est de 17.781 et donc la P–valeur est quasi-nulle (le logiciel écrit < 0.001). Le modèle de régression Prix / Surface est bien valide statistiquement.

Lorsqu'un modèle n'est pas statistiquement valide, on arrête son étude à ce stade : les étapes suivantes sont sans objet.

Etape 2 : qualité du modèle. L'intérêt statistique du modèle étant démontré, nous quantifions maintenant sa qualité, via le critère du r^2. Nous avons déjà expliqué comment procéder : ici,

88.8% des (variations de) prix sont expliqué(e)s par les (variations de) surfaces.

Etape 3 : écriture de la relation. Il s'agit de proposer une relation affine avec terme résiduel, dont on précise l'écart-type estimé $\widehat{\sigma}_n$. Cette relation doit être raisonnablement lisible par le grand public et vos managers, et compréhensible en soi (exit les x et les y) ; par exemple :

$$\begin{aligned}\text{Prix (en kEuros)} \quad &= 1.861 \\ &+ 5.648 \times \text{Surface (en m}^2\text{)} \\ &+ \text{Résidu (d'écart-type estimé 60.665)}\end{aligned}$$

Il ne faut oublier ni de préciser les unités (kEuros et m^2) ni d'inclure le terme résiduel : sans terme résiduel, vous proposeriez une relation déterministe qui donnerait l'impression que 100% de la variable à expliquer est expliqué par la variable explicative, alors que ce n'est pas le cas. C'est une erreur grave, dans le cadre de cours, que d'oublier de reporter un terme de résidu.

Comme nous l'avons indiqué ci-dessus (et le reverrons ci-dessous, à l'étape 5), le terme de résidu est à interpréter comme menant à des écarts typiques, pour des valeurs individuelles, de l'ordre de $\pm 2\widehat{\sigma}_n$ par rapport à la partie modélisée.

Etape 4 : validation économique, voire interprétation des coefficients. Dans cette étape, nous cherchons à voir si le modèle établi et écrit précédemment fait sens. Il convient a minima de vérifier que le bon sens n'est pas contredit, en termes de co-variations : dans la relation affine, le coefficient de pente est ici positif, ce qui signifie que les prix augmentent avec les surfaces. C'est parfaitement logique, et nous validons économiquement le modèle.

On peut alors se poser la question d'interpréter les coefficients 1.861 et 5.648. En réalité, 1.861 n'est pas significativement différent de 0 et n'est donc pas à interpréter (cf. P–valeur de 94.3% pour le test de $H_0 : \alpha_0 = 0$ vs. $H_0 : \alpha_0 \neq 0$, on conserve donc H_0). La valeur 5.648 mesure un prix au m^2 : ici, approximativement et en moyenne, 5 648 euros au mètre carré.

Pour bien faire, on pourrait relancer la régression en forçant l'ordonnée à l'origine à être nulle, pour avoir une valeur plus juste du prix approximatif et moyen au mètre carré, en l'occurrence, 5 669 euros au mètre carré :

Coefficients[a,b]

Modèle		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95% pour B	
		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Borne supérieure
1	Surface (m²)	5,669	,114	,992	49,811	<,001	5,439	5,899

a. Variable dépendante : Prix (kEuros) b. Régression linéaire à l'origine

Les exercices nous permettront de nous entraîner à effectuer des validations économiques et à interpréter des coefficients : ce sont des compétences qui s'acquièrent uniquement par la pratique.

Etape 5 : Prévision de nouvelles valeurs / Détection de valeurs atypiques. Les énoncés d'examens pourront vous demander, à ce stade, de répondre à des questions comme

> Quels sont les prix attendus pour un appartement de 80 m^2 ?

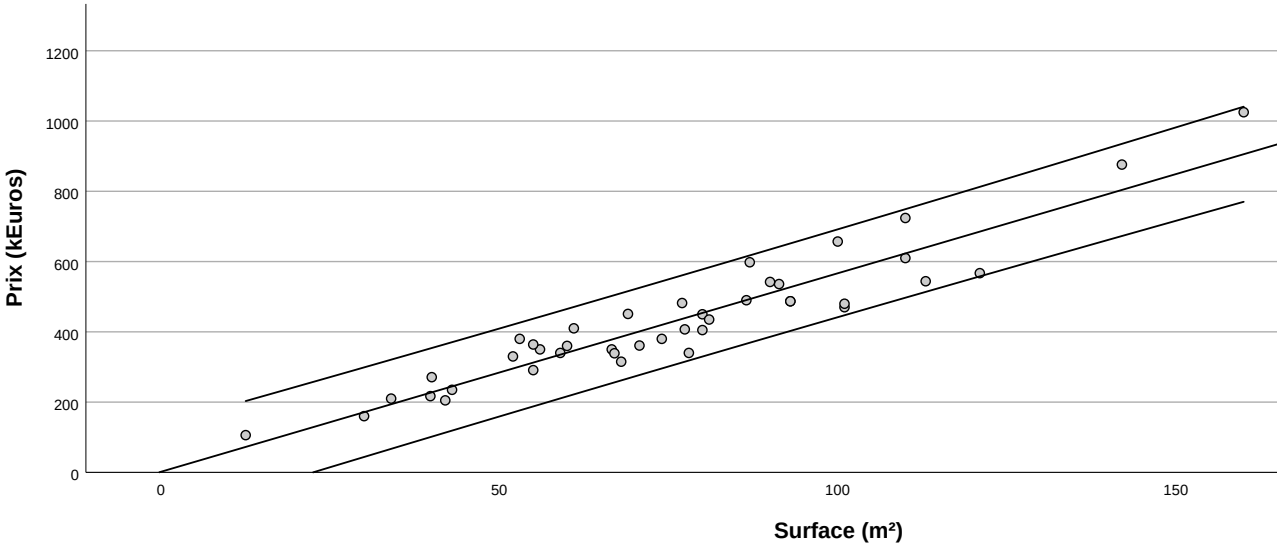
Il s'agit de proposer alors l'intervalle de prévision approché

$$1.861 + 5.648 \times 80 \pm 2 \times 60.665 = 453.701 \pm 121.33 \subset [330; 575]$$

(et non pas juste la valeur moyenne approximative 453.701).

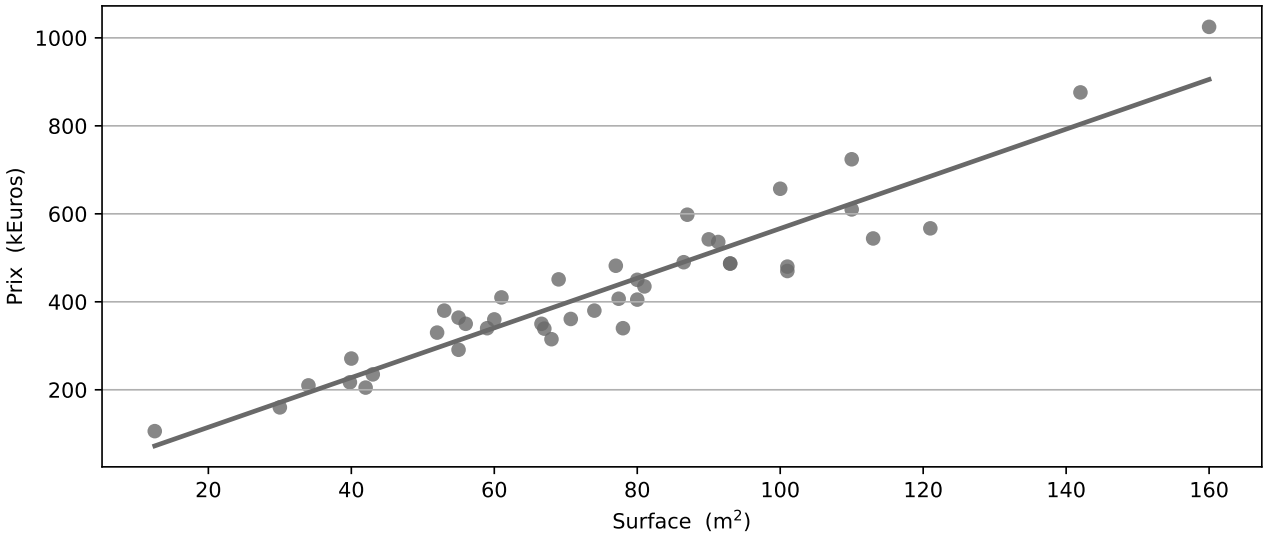
Notez que cela ne vous sera jamais demandé que pour des valeurs de x à l'intérieur de l'intervalle de valeurs sur lequel le modèle a été construit : il convient de ne pas extrapoler la relation. Ici, les valeurs couvertes vont essentiellement jusqu'à 120 m^2, il est imprudent d'utiliser la relation au-delà.

Ce qu'on appelle valeur atypique, c'est une observation y_t qui n'appartiendrait pas à l'intervalle de prévision fondé sur la valeur correspondante x_t de la variable explicative. Dans notre exemple, nous pouvons tracer les intervalles de prévision en fonction de x (en première approche, les bornes sont les parallèles à la droite de régression, décalées de $-2\widehat{\sigma}_n$ et $2\widehat{\sigma}_n$), et constater qu'aucune observation n'est atypique. On dit également que toutes les observations sont compatibles avec le modèle :



4. D'autres logiciels existent !

Dans ce polycopié et à l'examen, j'ai fait le choix du logiciel SPSS (commercialisé par IBM), parce qu'il produit des sorties agréables à l'œil. Son ergonomie et sa capacité à traiter les données sont limités. Les statisticiens professionnels en entreprise ("data scientists" comme on dit de nos jours) lui préfèrent Python (un environnement gratuit et "open source") : le résumé de régression linéaire que Python propose sur notre exemple de prix d'appartement est reproduit à la Figure 7.2. Les nombres que vous y lisez sont évidemment les mêmes, mais la présentation est un peu différente (heureusement, les noms des quantités calculées, bien que donnés en anglais, correspondent aux noms que nous avons introduits).



OLS Regression Results

Dep. Variable:	Price	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	316.2
Date:	Tue, 11 Jul 2023	Prob (F-statistic):	1.35e-20
Time:	13:41:52	Log-Likelihood:	-231.00
No. Observations:	42	AIC:	466.0
Df Residuals:	40	BIC:	469.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.8610	25.805	0.072	0.943	-50.293	54.015
Surface	5.6477	0.318	17.781	0.000	5.006	6.290

Omnibus:	0.605	Durbin-Watson:	2.398
Prob(Omnibus):	0.739	Jarque-Bera (JB):	0.668
Skew:	0.004	Prob(JB):	0.716
Kurtosis:	2.382	Cond. No.	224.

FIGURE 7.2 – Sorties produites par Python, le langage (open source, gratuit) qui est désormais le standard industriel pour le traitement de données. Les sorties précédentes étaient produites, elles, avec SPSS (un environnement commercialisé par IBM).

5. Appendice facultatif : formules diverses

Nous commençons cette appendice de lecture facultative par introduire des notations vectorielles : soient

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \text{et} \qquad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

de même que le sous-espace vectoriel $\mathcal{X}$ de $\mathbb{R}^n$ engendré par $\mathbf{1}$ et $\mathbf{x}$,

$$\mathcal{X} = \mathrm{Vect}(\mathbf{1}, \mathbf{x}) = \left\{ u\mathbf{1} + v\mathbf{x} : (u, v) \in \mathbb{R}^2 \right\}.$$

Droite de régression : coefficients a et b. Le calcul de la droite de régression revient exactement à minimiser la norme euclidienne (au carré) entre $\mathbf{y}$ et $\mathcal{X}$, i.e., à projeter orthogonalement $\mathbf{y}$ sur $\mathcal{X}$. On note $\Pi_{\mathcal{X}}$ l'opérateur de projection, et les coefficients a et b dont nous cherchons l'expression explicite sont tels que

$$\Pi_{\mathcal{X}}\mathbf{y} = \widehat{\mathbf{y}} = a\mathbf{1} + b\mathbf{x},$$

et ils sont caractérisés par le fait que $\mathbf{1}$ et $\mathbf{x}$ sont orthogonaux à $\mathbf{y} - \Pi_{\mathcal{X}}\mathbf{y} = \mathbf{y} - (a\mathbf{1} + b\mathbf{x})$, ce que l'on traduit par :

$$\frac{1}{n}\langle \mathbf{x}, \mathbf{y} - (a\mathbf{1} + b\mathbf{x}) \rangle = \frac{1}{n}\sum_{t=1}^{n} x_t y_t - a\overline{x}_n - b\frac{1}{n}\sum_{t=1}^{n} x_t^2 = 0,$$

$$\text{et} \qquad \frac{1}{n}\langle \mathbf{1}, \mathbf{y} - (a\mathbf{1} + b\mathbf{x}) \rangle = \overline{y}_n - a - b\,\overline{x}_n = 0.$$

L'équation $\overline{y}_n = a - b\,\overline{x}_n$ montre bien, comme annoncé, que la droite de régression passe par le point moyen $(\overline{x}_n, \overline{y}_n)$.

Le système de deux équations à deux inconnues ci-dessus peut également être obtenu en annulant le gradient de la fonction E_n de la page 97 : cela correspond à la condition nécessaire de point critique vérifiée par (a, b). L'approche géométrique initiale est plus « générale », au sens où c'est elle qui s'étend le mieux en dimensions supérieures, quand il y a plus qu'une variable explicative, comme ce sera le cas au chapitre suivant. C'est également elle qui permet de justifier le plus facilement la décomposition $\Sigma_{\text{tot}} = \Sigma_{\text{regr}} + \Sigma_{\text{res}}$.

Dans tous les cas, on résout facilement le système en injectant $a = \overline{y}_n - b\,\overline{x}_n$ dans la première équation :

$$0 = \frac{1}{n}\sum_{t=1}^{n} x_t y_t - a\overline{x}_n + b\frac{1}{n}\sum_{t=1}^{n} x_t^2 = \left(\frac{1}{n}\sum_{t=1}^{n} x_t y_t\right) - \overline{x}_n\,\overline{y}_n + b\,(\overline{x}_n)^2 - b\frac{1}{n}\sum_{t=1}^{n} x_t^2 = \mathrm{Cov}(x_1^n, y_1^n) - b\,\mathrm{Var}(x_1^n)$$

$$\text{où l'on a introduit} \qquad \mathrm{Var}(x_1^n) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x}_n)^2 = \frac{1}{n}\sum_{t=1}^{n} x_t^2 - (\overline{x}_n)^2$$

$$\text{et} \qquad \mathrm{Cov}(x_1^n, y_1^n) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x}_n)(y_t - \overline{y}_n) = \frac{1}{n}\sum_{t=1}^{n} x_t y_t - \overline{x}_n\,\overline{y}_n.$$

Par conséquent,

$$b = \frac{\mathrm{Cov}(x_1^n, y_1^n)}{\mathrm{Var}(x_1^n)}, \qquad \text{réalisation de} \qquad \widehat{\beta}_n = \frac{\mathrm{Cov}(x_1^n, Y_1^n)}{\mathrm{Var}(x_1^n)},$$

et $a = \overline{y}_n - b\,\overline{x}_n$, réalisation de $\widehat{\alpha}_n = \overline{Y}_n - \widehat{\beta}_n\,\overline{x}_n$.

On note pour la suite que les estimateurs $\widehat{\beta}_n$ et $\widehat{\alpha}_n$ sont bien des combinaisons linéaires des variables aléatoires indépendantes gaussiennes $Y_1, \dots, Y_n$. Ces estimateurs $\widehat{\beta}_n$ et $\widehat{\alpha}_n$ suivent donc également des lois gaussiennes.

Décomposition menant au r^2. Par projection orthogonale $\widehat{\mathbf{y}}$ de $\mathbf{y}$ sur $\mathcal{X}$ et du fait que $\overline{y}_n \mathbf{1} \in \mathcal{X}$, on a exactement, en notant $\|\cdot\|$ la norme euclidienne,

$$\Sigma_{\text{tot}} = \|\mathbf{y} - \overline{y}_n \mathbf{1}\|^2 = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 + \|\widehat{\mathbf{y}} - \overline{y}_n \mathbf{1}\|^2 = \Sigma_{\text{res}} + \Sigma_{\text{regr}}.$$

Estimateurs de α_0 et β_0. Comme indiqué plus haut, ils suivent des lois normales. On peut montrer que plus précisément,

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

Ces estimateurs ont des écarts-types respectivement notés $\sigma_0 S_{\beta,n}$ et $\sigma_0 S_{\alpha,n}$, où

$$S_{\beta,n} = \sqrt{\frac{1}{n \operatorname{Var}(x_1^n)}} \qquad \text{et} \qquad S_{\alpha,n} = \sqrt{\frac{1}{n}\left(1 + \frac{(\overline{x}_n)^2}{\operatorname{Var}(x_1^n)}\right)}.$$

Comme indiqué dans le corps du chapitre, $S_{\alpha,n}$ et $S_{\beta,n}$ sont bien des quantités parfaitement connues et ne dépendant que des variables explicatives $x_1, \dots, x_n$.

Il est à vrai dire immédiat de montrer que $\widehat{\beta}_n$ et $\widehat{\alpha}_n$ ont pour espérances β_0 et α_0, i.e., sont sans biais :

$$\begin{aligned}\mathbb{E}\Big[\operatorname{Cov}(x_1^n, Y_1^n)\Big] &= \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x}_n)\,\mathbb{E}\Big[Y_t - \overline{Y}_n\Big] \\ &= \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x}_n)\,((\alpha_0 + \beta_0 x_t) - (\alpha_0 + \beta_0 \overline{x}_n)) = \beta_0 \operatorname{Var}(x_1^n),\end{aligned}$$

puis, $\mathbb{E}[\widehat{\alpha}_n] = \mathbb{E}\Big[\overline{Y}_n\Big] - \mathbb{E}\Big[\widehat{\beta}_n\Big]\,\overline{x}_n = (\alpha_0 + \beta_0 \overline{x}_n) - \beta_0 \overline{x}_n = \alpha_0$. Le calcul des variances est moins agréable à poser, en revanche, mais il est élémentaire.

Intervalle de confiance sur $\alpha_0 + \beta_0 x$. L'estimateur ponctuel $\widehat{\mu}_x = \widehat{\alpha}_n + \widehat{\beta}_n x$ suit une loi normale, car c'est une combinaison linéaire de $\widehat{\alpha}_n$ et de $\widehat{\beta}_n$ et donc des $Y_1, \dots, Y_n$. On peut calculer son écart-type, substituer σ_0 par $\widehat{\sigma}$ et obtenir que $\widehat{\mu}_x$ normalisé par $\widehat{\sigma}_n^2 h_{x,n}/n$ suit une loi de Student à $n-2$ degrés de liberté. Cela donne lieu à l'intervalle de confiance, de niveau 95% :

$$\left[\widehat{\alpha}_n + \widehat{\beta}_n x \;\pm\; t_{n-2,\,97.5\%}\sqrt{\frac{\widehat{\sigma}_n^2}{n} h_{x,n}}\right] \qquad \text{où} \qquad h_{x,n} = 1 + \frac{1}{\operatorname{Var}(x_1^n)}(x - \overline{x}_n)^2.$$

En pratique (sur données réelles), le levier $h_{x,n}$ prend typiquement des valeurs entre 1 et 4 hors extrapolation, i.e., pour des valeurs x dans l'intervalle engendré par les x_t. La quantité $\pm t_{n-2,\,97.5\%}\sqrt{h_{x,n}/n}$ peut donc être approximativement majorée par $\pm 4/\sqrt{n}$.

Intervalle de prévision pour $Y_x = \alpha_0 + \beta_0 x + \varepsilon$. Selon le même principe que pour l'intervalle de confiance, on obtient :

$$\left[\widehat{\alpha}_n + \widehat{\beta}_n x \;\pm\; t_{n-2,1-97.5\%}\sqrt{\widehat{\sigma}_n^2\left(1 + \frac{h_{x,n}}{n}\right)}\right].$$

où le bruit ε ajouté, de variance σ_0^2, a conduit au terme $1+$ dans la racine carrée. Le facteur $1 + h_{x,n}/n$ n'est pas tellement plus grand que 1 et $t_{n-2,97.5\%}$ vaut environ 2, d'où la formule approchée $\pm 2\widehat{\sigma}_n$ pour la demi-largeur.

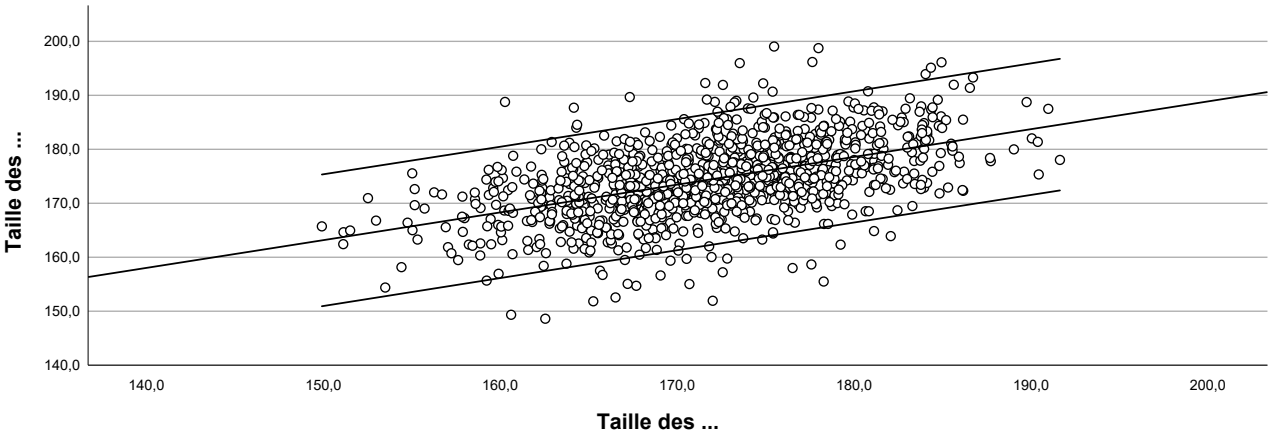
L'exemple historique de Galton : tailles des fils et tailles des pères

Sir Galton (homme de science britannique, 1822–1911) étudiait la taille des fils en fonction de la taille des pères. Il avait noté un retour vers un comportement moyen : les pères grands donnaient naissance à des fils plus petits qu'eux, et les pères petits donnaient naissance à des fils plus grands qu'eux. "Regression" signifie en anglais « retour » (vers un comportement moyen ici, Galton parlait de "regression to the mean")... d'où, par cet exemple historique marquant, la terminologie « régression linéaire », désormais synonyme de « modélisation linéaire », qu'on observe, ou non, un retour vers un comportement moyen dans les données sous-jacentes.

L'objectif de cet exercice est de voir si l'on peut rétrospectivement démontrer que les impressions de Galton étaient fondées, avec des outils postérieurs à son époque.

1. Quelle est la variable à expliquer, quelle est la variable explicative ? Complétez en conséquence les libellés des axes du nuage de points et indiquez quelle sortie de régression il conviendra d'exploiter.
2. Le modèle calculé est-il valide statistiquement, est-il de bonne qualité ?
3. Ecrivez la relation linéaire calculée. Est-il aisé d'en interpréter ses paramètres ?
4. Rappelez par quel point particulier passe la droite de régression et à l'aide du tableau ci-dessus, transformez la relation calculée en une relation équivalente et permettant de lire que Galton avait bien raison.

	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 (liste)	1078				



Sortie #1

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,501[a]	,251	,251	6,0353

a. Prédicteurs : (Constante), Taille des fils

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	13157,942	1	13157,942	361,235	<,001[b]
	Résidu	39193,204	1076	36,425		
	Total	52351,146	1077			

a. Variable dépendante : Taille des pères

b. Prédicteurs : (Constante), Taille des fils

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	86,633	4,491		19,289	<,001	77,820	95,446
	Taille des fils	,489	,026	,501	19,006	<,001	,438	,539

a. Variable dépendante : Taille des pères

Sortie #2

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,501[a]	,251	,251	6,1889

a. Prédicteurs : (Constante), Taille des pères

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	13835,971	1	13835,971	361,235	<,001[b]
	Résidu	41212,829	1076	38,302		
	Total	55048,800	1077			

a. Variable dépendante : Taille des fils

b. Prédicteurs : (Constante), Taille des pères

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	86,072	4,654		18,493	<,001	76,940	95,204
	Taille des pères	,514	,027	,501	19,006	<,001	,461	,567

a. Variable dépendante : Taille des fils

Modélisation de l'espérance de vie

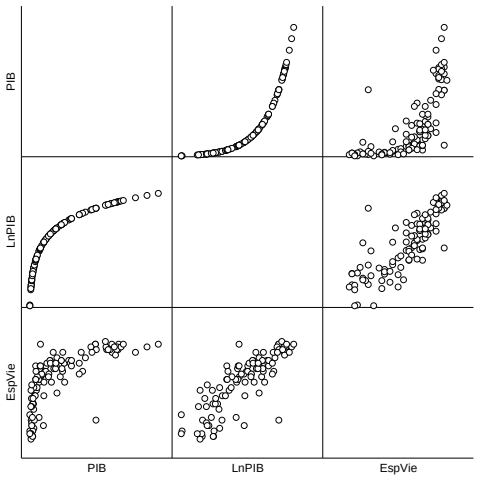
Nous allons modéliser l'espérance de vie des habitants d'un pays (à partir des données réelles, datant de 2012).

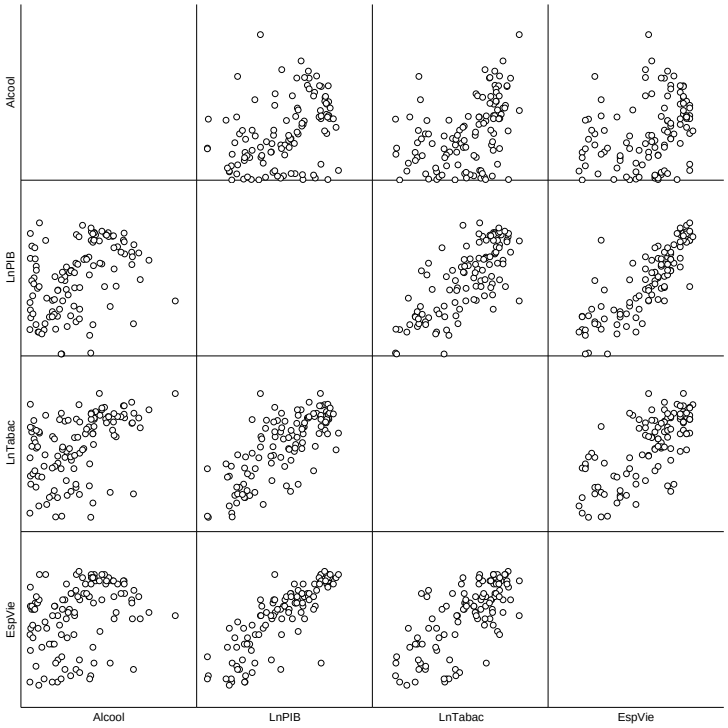
Variable	Définition	Unités
`Pays`	Nom du pays (114 pays considérés au total)	
`EspVie`	Espérance de vie	années
`PIB`	Produit intérieur brut, en millers de dollars, par habitant	K$/hab
`Alcool`	Consommation annuelle moyenne d'alcool (pur) par adulte	L/hab
`Tabac`	Nombre annuel moyen de cigarettes consommées par adulte	/hab
`LnPIB`	$= \ln($`PIB`$)$	
`LnAlcool`	$= \ln($`Alcool`$)$	
`LnTabac`	$= \ln($`Tabac`$)$	

Voici un court extrait du jeu de données.

	Pays	EspVie	PIB	Alcool	Tabac	LnPIB	LnAlcool	LnTabac
31	Egypt	71	6,367	,32	1275	1,851	-1,14	7,15
32	El Salvador	72	7,442	3,99	429	2,007	1,38	6,06
33	Estonia	75	18,274	17,24	1983	2,905	2,85	7,59
34	Ethiopia	54	1,014	4,10	87	,014	1,41	4,47
35	Finland	80	34,402	13,10	1351	3,538	2,57	7,21
36	France	81	34,092	12,48	2058	3,529	2,52	7,63
37	Gambia	60	1,973	3,58	171	,679	1,28	5,14
38	Germany	80	35,930	12,14	1702	3,582	2,50	7,44
39	Ghana	60	1,610	3,11	161	,476	1,13	5,08
40	Greece	80	28,834	11,01	4313	3,362	2,40	8,37
41	Guatemala	69	4,871	7,10	609	1,583	1,96	6,41

Voici enfin quelques nuages de points.





Modèle EspVie / Alcool

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,306[a]	,093	,085	9,896

a. Prédicteurs : (Constante), Alcool

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	1131,310	1	1131,310	11,552	<,001[b]
	de Student	10968,550	112	97,933		
	Total	12099,860	113			

a. Variable dépendante : EspVie

b. Prédicteurs : (Constante), Alcool

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	64,736	1,631		39,687	<,001
	Alcool	,607	,179	,306	3,399	<,001

a. Variable dépendante : EspVie

Modèle EspVie / LnTabac

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,710[a]	,504	,499	7,321

a. Prédicteurs : (Constante), LnTabac

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	6096,721	1	6096,721	113,746	<,001[b]
	de Student	6003,138	112	53,599		
	Total	12099,860	113			

a. Variable dépendante : EspVie

b. Prédicteurs : (Constante), LnTabac

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	20,395	4,636		4,399	<,001
	LnTabac	7,276	,682	,710	10,665	<,001

a. Variable dépendante : EspVie

Modèle EspVie / PIB

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,712[a]	,507	,503	7,296

a. Prédicteurs : (Constante), PIB

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	6137,292	1	6137,292	115,282	<,001[b]
	de Student	5962,568	112	53,237		
	Total	12099,860	113			

a. Variable dépendante : EspVie

b. Prédicteurs : (Constante), PIB

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	61,991	,964		64,277	<,001
	PIB	,522	,049	,712	10,737	<,001

a. Variable dépendante : EspVie

Modèle EspVie / LnPIB

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,846[a]	,716	,714	5,539

a. Prédicteurs : (Constante), LnPIB

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	8664,039	1	8664,039	282,428	<,001[b]
	de Student	3435,821	112	30,677		
	Total	12099,860	113			

a. Variable dépendante : EspVie

b. Prédicteurs : (Constante), LnPIB

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	56,054	,943		59,411	<,001
	LnPIB	6,718	,400	,846	16,806	<,001

a. Variable dépendante : EspVie

1. Quelle est la meilleure variable explicative prise isolément entre `PIB` et `LnPIB` ? Fondez votre réponse d'abord sur des éléments graphiques, puis sur des éléments quantitatifs.
2. Ecrivez la relation obtenue avec la meilleure variable entre `PIB` et `LnPIB`. Interprétez les valeurs numériques des coefficients, en vous aidant du tableau ci-dessous.

	N	Minimum	Maximum	Moyenne	Ecart type
EspVie	114	47	83	69,30	10,348
PIB	114	,396	57,238	13,99882	14,118813
LnPIB	114	-,927	4,047	1,97135	1,303337
N valide (liste)	114				

3. L'espérance de vie en Gambie est-elle compatible avec le modèle écrit à la question précédente ?
4. Parmi les deux autres régressions linéaires simples reportées : lesquelles sont valides statistiquement, lesquelles sont en outre valides économiquement ?

	Station	Massif	Alpes	Pyrenees	PrixForfait	AltitudeBas	AltitudeSommet	Denivele	Pistes	Remontees
184	Thollon les Memises	Alpes du Nord	1	0	123	1000	2000	1000	16	14
185	Tignes	Alpes du Nord	1	0	285	1550	3450	1900	74	39
186	Turini Camp dArgent	Alpes du Sud	1	0	75	1600	1920	320	4	3
187	Val Cenis	Alpes du Nord	1	0	175	1300	2800	1500	57	28
188	Val dAllos - La Foux	Alpes du Sud	1	0	196	1800	2600	800	40	18
189	Val dAllos - Le Seignus	Alpes du Sud	1	0	141	1500	2425	925	24	10
190	Val dAzun	Pyrénées	0	1	48	1350	1600	250	10	2
191	Val dIsere	Alpes du Nord	1	0	285	1850	3456	1606	76	42
192	Val Louron	Pyrénées	0	1	145	1465	2100	635	20	10
193	Val Pelens	Alpes du Sud	1	0	75	1600	1750	150	7	2
194	Val Thorens	Alpes du Nord	1	0	250	2300	3230	930	82	32
195	Valberg	Alpes du Sud	1	0	172	1430	2100	670	56	23
196	Valfrejus	Alpes du Nord	1	0	155	1550	2737	1187	23	10
197	Valloire	Alpes du Nord	1	0	200	1430	2600	1170	89	30
198	Valmeinier	Alpes du Nord	1	0	200	1500	2600	1100	42	14
199	Valmorel	Alpes du Nord	1	0	225	1400	2550	1150	61	34
200	Vars	Alpes du Sud	1	0	200	1650	2750	1100	72	26
201	Vaujany	Alpes du Nord	1	0	172	1650	2800	1150	30	20
202	Ventron	Vosges	0	0	123	870	1150	280	10	7
203	Villard de Lans	Alpes du Nord	1	0	178	1143	2050	907	52	20
204	Villard Reculas	Alpes du Nord	1	0	103	1480	3300	1820	12	8

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
PrixForfait	204	40	285	151,43	60,591
AltitudeBas	204	600	2300	1316,39	292,991
AltitudeSommet	204	970	3600	2159,53	583,323
Dénivelé	204	60	2350	846,54	495,308
Pistes	204	1	235	37,79	34,747
Remontées	204	1	108	18,61	18,008
Alpes	204	0	1	,72	,450
Pyrénées	204	0	1	,15	,355
N valide (liste)	204				

FIGURE 7.3 – Extrait du jeu de données de forfaits de ski considéré tout au long du Chapitre 8 [gauche] et statistiques descriptives associées [droite].

8

Régression linéaire multiple

Le jeu de données considéré ici a été créé par Benjamin Petiau (ex-enseignant dans ce cours) en décembre 2017, à partir du site `https://www.skiinfo.fr/`. Il a extrait, pour 204 stations de ski françaises, le prix du forfait de ski hebdomadaire (variable à expliquer) ainsi que plusieurs variables explicatives décrites ci-dessous. Un extrait du jeu de données et des statistiques descriptives sur les variables sont fournis ci-contre.

Variable	Définition	Unités
`Station`	Nom de la station	
`Massif`	Nom du massif dans lequel se situe la station	
`Alpes`	1 si la station se situe dans les Alpes (du Nord ou du Sud), 0 sinon	
`Pyrénees`	1 si la station se situe dans les Pyrénées, 0 sinon	
`Prix Forfait`	Prix d'un forfait de ski hebdomadaire	euros
`Altitude Bas`	Altitude du bas du domaine (du village-station)	mètres
`Altitude Sommet`	Altitude du sommet de la station	mètres
`Dénivelé`	Différence entre les deux altitudes précédentes	mètres
`Pistes`	Nombre total de pistes (vertes, bleues, rouges, noires)	
`Remontées`	Nombre total de remontées mécaniques (de toutes sortes)	

Objectif : nous allons modéliser la variable `PrixForfait` comme fonction affine de plusieurs variables explicatives ; c'est ce qu'on appelle une régression linéaire multiple (par opposition au chapitre précédente de régression linéaire simple, avec une unique variable explicative).

Nous commençons par postuler une relation fondée uniquement sur $K = 3$ variables explicatives choisies arbitrairement — relation de la forme :

$$\begin{aligned}\text{Prix Forfait (en euros)} \quad &= \alpha_0 \\ &\quad + \beta_0^{(1)} \times \text{Altitude Sommet (en m)} \\ &\quad + \beta_0^{(2)} \times \text{Dénivelé (en m)} \\ &\quad + \beta_0^{(3)} \times \text{Pistes (sans unités)} \\ &\qquad + \text{Résidu gaussien centré (d'écart-type } \sigma_0)\end{aligned}$$

Nous verrons plus tard comment choisir de bons sous-ensembles de variables explicatives.

Méthode pédagogique de ce chapitre : contrairement au chapitre précédent, qui fournissait l'ensemble des formules nécessaires et détaillait les raisonnements mathématiques, nous allons plutôt

nous concentrer sur la lecture et l'exploitation de sorties logiciel, en soulignant les éléments nouveaux introduits par la considération simultanée de plusieurs variables explicatives.

1. Lecture et exploitation d'une sortie de régression linéaire multiple

Nous suivons les 5 mêmes étapes qu'au Chapitre 7 en régression linéaire simple. Les changements principaux portent sur l'étape 1 de validité statistique, et des remarques importantes seront également effectuées à l'étape 4 sur la hiérarchisation des variables lors de l'interprétation des coefficients.

Etape 1 : Validité statistique

Elle se décompose en deux temps : (i) validité statistique globale, pour vérifier que le modèle contribue significativement à l'explication statistique, et (ii) validités statistiques individuelles, pour s'assurer que le modèle n'est pas plus compliqué que nécessaire. Ces deux temps coïncidaient exactement en régression linéaire simple car ils se réduisaient tous deux à tester que l'unique coefficient de pente était nul.

Etape 1(a) : Validité statistique globale. Que le modèle contribue significativement à l'explication statistique, qu'il soit utile, correspond au fait qu'au moins un coefficient de pente $\beta_0^{(j)}$ soit non nul. On teste

$$H_0 : \beta_0^{(1)} = \beta_0^{(2)} = \ldots = \beta_0^{(K)} = 0 \qquad \text{vs.} \qquad H_1 : \text{il existe } j \in \{1, \ldots, K\} \text{ tel que } \beta_0^{(j)} \neq 0$$

Le test (appelé F–test, fondé sur une statistique de test suivant une loi dite de Fisher) est conduit dans le second tableau de la sortie logiciel. Bien évidemment, on souhaite lire une P–valeur $< 5\%$, afin de prouver qu'au moins une variable explicative est utile pour une modélisation affine.

On peut montrer que la F–statistique est une fonction croissante de r^2 et décroissante de n, de sorte que le F–test est équivalent à comparer r^2 à une quantité de l'ordre de $1/\sqrt{n}$. Cela explique pourquoi, lorsque l'on construit des modèles à partir d'observations nombreuses, des modèles avec un r^2 de valeurs nominales faibles sont malgré tout globalement valides.

Etape 1(b) : Validités statistiques individuelles (marginales). Ceci étant effectué, on veut alors vérifier que le modèle n'est pas inutilement compliqué et que chaque variable considérée contribue suffisamment, de manière incrémentale, à l'explication linéaire. Par « de manière incrémentale », on entend non pas sa contribution intrinsèque (dans le cadre d'un modèle de régression simple par exemple), mais sa contribution étant donné qu'on considère déjà les autres variables explicatives. Pour la variable j, cela correspond au test

$$H_0 : \beta_0^{(j)} = 0 \text{ dans le modèle avec les } K-1 \text{ autres variables}$$
$$H_1 : \beta_0^{(j)} \neq 0 \text{ dans le modèle avec les } K-1 \text{ autres variables}$$

La précision « dans le modèle avec les $K-1$ autres variables » est importante : encore une fois, ce n'est pas un test « absolu » d'intérêt de la variable, mais un test « relatif », un test de l'intérêt d'ajouter la variable j dans un modèle où il n'y aurait que les $K-1$ autres variables.

On ne veut retenir que des modèles dont toutes les variables sont individuellement valides ; sinon, il faut d'abord simplifier le modèle. Attention, à cause des raisons possibles de non-validités individuelles, cette simplification doit se faire de manière précautionneuse, par exemple en enlevant une à une les variables non individuellement significatives tant que cela est nécessaire (méthode de sélection "backward"). Il ne faut pas procéder brutalement en supprimant toutes les variables non individuellement significatives. Ceci s'explique à cause des deux raisons (et demie) possibles de non-significativités individuelles :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,885[a]	,783	,780	28,418

a. Prédicteurs : (Constante), Pistes, AltitudeSommet, Dénivelé

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	583744,991	3	194581,664	240,949	<,001[b]
	Résidu	161513,048	200	807,565		
	Total	745258,039	203			

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Pistes, AltitudeSommet, Dénivelé

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	16,433	9,792		1,678	,095
	AltitudeSommet	,039	,007	,376	5,627	<,001
	Dénivelé	,028	,009	,232	3,288	,001
	Pistes	,706	,071	,405	9,874	<,001

a. Variable dépendante : PrixForfait

Sortie logiciel [haut] et explications de contenus [bas], avec des notations similaires à celles du Chapitre 7 ; les cases avec ✗ sont omises du fait qu'elles correspondent à des formules bien plus complexes à écrire qu'en régression linéaire simple

Récapitulatif des modèles

R	R-deux	R-deux ajusté	Erreur standard de l'estimation
$\sqrt{r^2}$	r^2	$r^2_{\text{adj}} = 1 - \frac{n-1}{n-1-K}(1-r^2)$	$\widehat{\sigma}_n$

ANOVA

Modèle	Somme des carrés	ddl	Carrés moyens	F	Sig.
Régression	Σ_{regr}	K	Σ_{regr}/K	$\frac{\Sigma_{\text{regr}}/K}{\Sigma_{\text{res}}/(n-1-K)}$	P–valeur
Résidu	Σ_{res}	$n-1-K$	$\Sigma_{\text{res}}/(n-1-K)$		pour Étape 1(a)
Total	Σ_{tot}	$n-1$			

Coefficients

	Coefficients non standardisés		Beta		
Modèle	B	Err. std.		t	Sig.
Constante	$\widehat{\alpha}_n$	✗		✗	P–val. $H_0 : \alpha_0 = 0$
Var. explicative 1	$\widehat{\beta}_n^{(1)}$	✗	✗	✗	P–val. pour Étape 1(b)
Var. explicative 2	$\widehat{\beta}_n^{(2)}$	✗	✗	✗	P–val. pour Étape 1(b)
...	...	✗	✗	✗	...
Var. explicative K	$\widehat{\beta}_n^{(K)}$	✗	✗	✗	P–val. pour Étape 1(b)

— La variable a effectivement peu ou rien à voir avec le problème. Souvent, la régression linéaire simple correspondante était non statistiquement valide ou à tout le moins, souffrait d'un r^2 faible.

— La variable est très linéairement liée à une autre variable ou à une combinaison linéaire de variables parmi les $K-1$ autres variables (question de redondance, donc). Pour en avoir le cœur net, on peut calculer la régression linéaire de la variable en question sur les $K-1$ autres variables et observer si l'on obtient un r^2 très grand.

— La demie-raison est que la variable explicative considérée a bien une influence, mais non linéaire, sur la variable à expliquer. Cette demie-raison est bien plus technique et est hors de la portée de ce cours.

Exemple. Ici, le modèle Prix Forfait / Altitude Sommet, Dénivelé, Pistes est bien valide globalement et chacune de ses variables explicatives est bien individuellement valide. On peut noter que l'ordonnée à l'origine estimée 16.433 n'est, elle, pas significativement différente de 0 (cf. P–valeur de 9.5% dans le test bilatère avec $H_0 : \alpha_0 = 0$), mais peu importe : cela ne remet pas en cause le sous-ensemble de trois variables considéré.

Autres étapes

Etape 1bis : Validité économique du modèle. Nous intercalerons souvent ici la validation économique du modèle, avant d'aller plus loin. Elle consiste à s'assurer que toutes les co-variations sont logiques, entre chaque variable explicative et la variable à expliquer, i.e., que les coefficients de pente ont tous un signe correspondant à la logique économique.

Ici, on observe trois coefficients estimés positifs, ce qui veut dire que le prix du forfait augmente avec l'altitude du sommet (cf. qualité de la neige), le dénivelé de la station (cf. longueur des pistes), et le nombre des pistes (cf. taille du domaine), trois facteurs en effet influant positivement sur le prix du forfait.

Etape 2 : Qualité du modèle. Le critère du r^2 est défini similairement au cas de la régression linéaire simple et bénéficie de la même interprétation de proportion de variabilité totale retrouvée par le modèle de régression. Ici,

> 78.3% (des variations) des prix des forfaits sont expliqué(e)s par les (variations des) altitudes au sommet, dénivelés et nombres de pistes.

Toutefois, on peut montrer que le r^2 augmente nécessairement avec l'inclusion de nouvelles variables. Il faut donc pondérer l'augmentation observée du r^2 (la qualité du modèle) avec le nombre de variables considéré (la complexité du modèle) et atteindre un bon compromis entre qualité (un r^2 grand) et complexité (pas trop de variables si possible).

La doxa statistique recourt à une version dite ajustée du r^2, définie comme

$$r^2_{\text{adj}} = 1 - \frac{n-1}{n-1-K}(1-r^2)$$

et voulant pénaliser les grandes valeurs de K. La doxa statistique vous encouragerait à choisir, parmi deux modèles statistiquement et économiquement valides, celui avec le plus grand r^2_{adj}. Toutefois, vous constaterez vite dans les sorties logiciel que la correction à la baisse apportée par le r^2_{adj} sur le r^2 est bien trop faible pour pouvoir compenser, en général, l'augmentation de r^2 avec le nombre de variables, de sorte que même comparer les r^2_{adj} conduira à choisir, en général, le modèle le plus complexe.

Nous recourrons dans les problèmes à des comparaisons plus « humaines », appréciant les écarts de r^2

entre deux modèles aux nombres de variables différentes à l'aune du nombre et surtout, de l'identité, des variables présentes dans un modèle et pas dans l'autre.

Etape 3 : Ecriture du modèle. Cette étape est tout à fait similaire à l'étape correspondante en régression linéaire simple, et les deux mêmes commandements s'appliquent : (i) inclure un terme de résidu et (ii) reporter les unités. Ici, on lit la relation estimée

$$\begin{aligned}\text{Prix Forfait (en euros)} &= 16.433 \\ &\quad + 0.039 \times \text{Altitude Sommet (en m)} \\ &\quad + 0.028 \times \text{Dénivelé (en m)} \\ &\quad + 0.706 \times \text{Pistes (sans unités)} \\ &\qquad + \text{Résidu gaussien centré (d'écart-type estimé 28.418)}\end{aligned}$$

Etape 4 : Interprétation des coefficients. Comme en régression linéaire simple, interpréter la valeur numérique de l'ordonnée à l'origine (qui est une constante de recalage du modèle) est en général difficile, sauf peut-être lorsque les variables explicatives peuvent être toutes nulles. Ici, ce n'est pas le cas : l'altitude, le dénivelé, le nombre de pistes ne sont jamais nulles.

En revanche, on peut là aussi interpréter facilement les coefficients de pente, en n'oubliant pas la précision « approximativement et en moyenne » et en rajoutant ici « toutes choses égales par ailleurs », pour indiquer que l'on fait bouger une variable sans toucher aux autres :

— Lorsque l'altitude au sommet augmente de 100 m, le prix du forfait augmente, approximativement et en moyenne, et toutes choses égales par ailleurs, de 3.90 euros.
— Lorsque le dénivelé augmente de 100 m, le prix du forfait augmente, approximativement et en moyenne, et toutes choses égales par ailleurs, de 2.80 euros.
— Lorsque le nombre de pistes augmente de 10, le prix du forfait augmente, approximativement et en moyenne, et toutes choses égales par ailleurs, de 7.06 euros.

Etape 4bis : Considérations sur l'influence des variables. Il ne faut jamais hiérarchiser les variables en fonction des valeurs nominales des coefficients, car ces derniers dépendent fortement des unités ; par exemple, la relation précédente est équivalente à celle-ci, où l'on reporte l'altitude en km :

$$\begin{aligned}\text{Prix Forfait (en euros)} &= 16.433 \\ &\quad + 3.9 \times \text{Altitude Sommet (en km)} \\ &\quad + 0.028 \times \text{Dénivelé (en m)} \\ &\quad + 0.706 \times \text{Pistes (sans unités)} \\ &\qquad + \text{Résidu gaussien centré (d'écart-type estimé 28.418)}\end{aligned}$$

Or l'importance des variables, elle, ne dépend pas des unités ! Il faut bien évidemment apprécier l'importance via une version normalisée des coefficients, donnée par les statistiques de test t (plus t est grand, plus la variable est importante) et les P–valeurs de significativités individuelles (plus une P–valeur est petite, plus la variable est importante).

Et encore une fois, il faut vraiment se retenir de penser qu'un coefficient nominal apparamment petit (comme 0.028 ci-dessus pour le dénivelé) correspond à une variable avec peu d'influence : il peut très bien, et c'est le cas ici, être multiplié par des valeurs très grandes.

Etape 5 : Intervalles de prévision. La dernière étape est similaire à l'étape correspondante en régression simple : les valeurs attendues pour de nouvelles valeurs des variables explicatives sont données par la prévision ponctuelle du modèle en ces valeurs $\pm 2s$, où s est l'écart-type estimé des résidus. On dit également qu'une observation est compatible avec le modèle si elle appartient à son propre intervalle de prévision, et qu'elle est atypique sinon.

2. Comparaison de modèles et sélection des variables explicatives

Comparaison 2 à 2. Nous avons déjà évoqué le critère r^2_{adj} avancé par la doxa statistique, et l'approche plus subjective que je propose d'effectuer un compromis entre qualité, telle que mesurée par le r^2, et simplicité, i.e., avoir un modèle avec peu de variables.

Sélection "backward" : simplificative itérative. Une première méthode pour construire une suite de modèle vous paraîtra naturelle, vu les commentaires effectués autour de la gestion de non-validités individuelles : partir du modèle complet (avec toutes les variables explicatives) et supprimer itérativement, une par une, la variable la moins individuellement significative (i.e., associée à la plus grande P–valeur individuelle) tant que cela est nécessaire. C'est ce que l'on appelle la sélection "backward". Elle est mise en œuvre dans notre exemple ci-contre.

Dans notre exemple, on a une liaison linéaire entre Altitude Sommet, Altitude Bas et Dénivelé (la troisième est la différence des deux premières), ce qui explique que dans le modèle complet, nous n'avons pas considéré Altitude Bas. Nous omettons également pour l'instant les deux variables binaires Alpes et Pyrénées. Le modèle complet admet donc quatre variables explicatives. C'est le modèle #1 dans la sortie de sélection "backward". Il n'est pas statistiquement valide, car s'il est bien globalement valide, un problème de non-significativité individuelle se pose pour la variable Remontées (certes sur le fil, elle est associée à une P–valeur de 5.1%). On la retire et on recalcule la régression : le modèle avec trois variables obtenu, qui est celui que nous avions étudié en première partie de ce chapitre, est bien statistiquement et économiquement valide, lui.

Sélection "forward" : enrichissement itératif. La sélection "forward" consiste à partir de la meilleure variable explicative prise isolément (celle avec le plus grand r^2) et à enrichir itérativement le modèle tant que cela est possible : à chaque étape, on regarde si l'on peut ajouter une variable qui soit individuellement significative dans le modèle ainsi complété (et si plusieurs ajouts sont possibles, on retient celui qui conduit au meilleur r^2). A noter : une fois qu'une variable est incluse, sa présence n'est plus remise en cause.

Ici, le cheminement est en trois étapes, et l'on constate que la meilleure variable explicative prise isolément est le dénivelé. On aboutit au même modèle qu'en sélection "backward", à savoir, celui étudié au début de ce chapitre. En général, il n'y a toutefois pas de raison que les résultats des deux méthodes de sélection coïncident.

Quel est, à mes yeux, le meilleur modèle parmi tous les modèles calculés ? Les modèles en lice sont ceux stastistiquement et économiquement valides : à ce stade, dans ce chapitre, ces modèles sont donc ceux à lire dans la sortie de sélection "forward" et uniquement ceux-là, à savoir :

Prix Forfait / Dénivelé : $r^2 = 64.0\%$
Prix Forfait / Dénivelé, Pistes : $r^2 = 74.9\%$
Prix Forfait / Dénivelé, Pistes, Altitude Sommet : $r^2 = 78.3\%$

Les questions que l'on se pose sont, d'une part : une augmentation de 3.4% du r^2 mérite-t-elle, à nos yeux, d'ajouter la variable Altitude Sommet dans la régression ? Vous lisez le compromis que je pose entre augmentation de la qualité et augmentation de la complexité. La réponse est personnelle et subjective ; je tendrais pour ma part à répondre « Non ». De même, on se demande, d'autre part, si une augmentation du r^2 de 10.9% mérite l'inclusion de la variable Pistes.

Une dernière remarque. La pratique estime qu'il convient de retenir au plus $n/6$ variables explicatives, où n est le nombre d'observations disponibles. Vous verrez que dans les sujets que nous traiterons, nous serons toujours loin de ce seuil.

Sélection Backward

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,887[a]	,787	,783	28,217
2	,885[b]	,783	,780	28,418

a. Prédicteurs : (Constante), Remontées, AltitudeSommet, Dénivelé, Pistes

b. Prédicteurs : (Constante), AltitudeSommet, Dénivelé, Pistes

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	586810,318	4	146702,580	184,249	<,001[b]
	de Student	158447,721	199	796,220		
	Total	745258,039	203			
2	Régression	583744,991	3	194581,664	240,949	<,001[c]
	de Student	161513,048	200	807,565		
	Total	745258,039	203			

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Remontées, AltitudeSommet, Dénivelé, Pistes

c. Prédicteurs : (Constante), AltitudeSommet, Dénivelé, Pistes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	14,927	9,753		1,530	,127
	AltitudeSommet	,041	,007	,393	5,877	<,001
	Dénivelé	,024	,009	,197	2,718	,007
	Pistes	,483	,134	,277	3,605	<,001
	Remontées	,521	,266	,155	1,962	,051
2	(Constante)	16,433	9,792		1,678	,095
	AltitudeSommet	,039	,007	,376	5,627	<,001
	Dénivelé	,028	,009	,232	3,288	,001
	Pistes	,706	,071	,405	9,874	<,001

a. Variable dépendante : PrixForfait

Sélection Forward

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,800[a]	,640	,638	36,441
2	,865[b]	,749	,746	30,509
3	,885[c]	,783	,780	28,418

a. Prédicteurs : (Constante), Dénivelé

b. Prédicteurs : (Constante), Dénivelé, Pistes

c. Prédicteurs : (Constante), Dénivelé, Pistes, AltitudeSommet

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	477015,042	1	477015,042	359,215	<,001[b]
	de Student	268242,997	202	1327,936		
	Total	745258,039	203			
2	Régression	558173,367	2	279086,684	299,845	<,001[c]
	de Student	187084,672	201	930,770		
	Total	745258,039	203			
3	Régression	583744,991	3	194581,664	240,949	<,001[d]
	de Student	161513,048	200	807,565		
	Total	745258,039	203			

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Dénivelé

c. Prédicteurs : (Constante), Dénivelé, Pistes

d. Prédicteurs : (Constante), Dénivelé, Pistes, AltitudeSommet

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	68,581	5,061		13,550	<,001
	Dénivelé	,098	,005	,800	18,953	<,001
2	(Constante)	66,848	4,242		15,760	<,001
	Dénivelé	,068	,005	,555	12,624	<,001
	Pistes	,716	,077	,411	9,338	<,001
3	(Constante)	16,433	9,792		1,678	,095
	Dénivelé	,028	,009	,232	3,288	,001
	Pistes	,706	,071	,405	9,874	<,001
	AltitudeSommet	,039	,007	,376	5,627	<,001

a. Variable dépendante : PrixForfait

Modélisation du forfait de ski, suite : prise en compte de variables binaires

On enrichit la modélisation linéaire du forfait de ski en tenant compte des massifs. A cet effet, on considère que l'impact du massif est additif sur le prix et uniforme au sein d'un massif : autrement dit, on introduit des paramètres s'ajoutant uniformément à tous les prix des stations d'un massif ou groupe de massifs donné.

1. Expliquer au préalable comment l'utilisation des variables `Alpes` et `Pyrénées` permet de tenir compte des trois groupes de massifs.
2. Commenter brièvement d'un point de vue statistique et économique les deux modèles de régression présentés. Conclure ensuite : observe-t-on, toutes choses égales par ailleurs, des écarts de prix significatifs entre les massifs ? Si oui, expliquer quels écarts sont mis en évidence.
3. Est-ce que la réponse à la question précédente contredit la sortie statistique ci-dessous ? Commencer par expliquer ce que cette sortie démontre et ensuite seulement, discuter l'existence ou l'absence d'une contradiction.

Statistiques de groupe

	Alpes	N	Moyenne	Ecart type	Moyenne d'erreur standard
PrixForfait	0	57	130,21	49,410	6,544
	1	147	159,66	62,643	5,167

Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes							
						Signification		Différence moyenne	Erreur standard	Intervalle de confiance de la différence à 95 %	
		F	Sig.	t	df	p unilatéral	p bilatéral			Inférieur	Supérieur
PrixForfait	Hypothèse de variances égales	4,133	,043	-3,184	202	<,001	,002	-29,449	9,248	-47,685	-11,214
	Hypothèse de variances inégales			-3,532	128,425	<,001	<,001	-29,449	8,338	-45,947	-12,951

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,888[a]	,788	,783	28,200

a. Prédicteurs : (Constante), Pyrénées, AltitudeSommet, Pistes, Alpes

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	587007,433	4	146751,858	184,540	<,001[b]
	de Student	158250,606	199	795,229		
	Total	745258,039	203			

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Pyrénées, AltitudeSommet, Pistes, Alpes

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	2,044	7,966		,257	,798
	AltitudeSommet	,065	,005	,626	14,239	<,001
	Pistes	,786	,068	,451	11,548	<,001
	Alpes	-25,629	6,928	-,190	-3,699	<,001
	Pyrénées	-15,565	8,252	-,091	-1,886	,061

a. Variable dépendante : PrixForfait

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,885[a]	,784	,781	28,380

a. Prédicteurs : (Constante), Alpes, Pistes, AltitudeSommet

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	584177,783	3	194725,928	241,775	<,001[b]
	de Student	161080,256	200	805,401		
	Total	745258,039	203			

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Alpes, Pistes, AltitudeSommet

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	-,047	7,939		-,006	,995
	AltitudeSommet	,061	,004	,592	14,697	<,001
	Pistes	,803	,068	,461	11,843	<,001
	Alpes	-16,108	4,776	-,120	-3,373	<,001

a. Variable dépendante : PrixForfait

Modélisation des évaluations de films

On reprend le jeu de données d'évaluations de films présenté dans le chapitre sur les données appariées.

Régressions linéaires simples. On veut vérifier la compatibilités des avis spectateurs sur IMDb et sur Rotten Tomatoes. Pour cela, on considère le modèle linéaire `Note_IMDb` / `Spec_RT`.

1. Le modèle calculé est-il valide statistiquement et économiquement ? Si oui, écrivez la relation qu'il propose et interprétez-en soigneusement la valeur numérique des coefficients.
2. Quel est l'indicateur statistique qui quantifierait la compatibilité globale des avis spectateurs sur IMDb et sur Rotten Tomatoes, et que pensez-vous de sa valeur ?
3. Intéressons-nous maintenant à une compatibilité inviduelle. Prenons le film Hercules : ses notes moyennes spectacteurs sur IMDb et sur Rotten Tomatoes sont-elles compatibles selon le modèle discuté ici ?

Le jeu de données comporte une variable `NbAvis_IMDb` que nous n'avions pas décrite dans le chapitre sur les données appariées et qui pour chaque films, indique le nombre d'avis spectateurs déposés sur IMDb. Des statistiques descriptives sur cette variable sont :

	N	Minimum	Maximum	Moyenne	Ecart type
NbAvis_IMDb	651	180	893008	57532,98	112124,387
N valide (liste)	651				

On se demande si les films les plus vus sont des films populaires. Pour cela, on considère le modèle linéaire `Note_IMDb` / `NbAvis_IMDb`.

4. Le modèle calculé est-il valide statistiquement et économiquement ? Si oui, écrivez la relation qu'il propose et commentez la valeur du coefficient de pente uniquement.
5. Toujours pour le film Hercules : sa note moyenne sur IMDb est-elle compatible avec le nombre d'avis mis selon le modèle de régression discuté dans cette série de questions, à savoir 109 633 ?

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,865[a]	,748	,748	,5450

a. Prédicteurs : (Constante), Spec_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	572,093	1	572,093	1926,312	<,001[b]
	Résidus	192,746	649	,297		
	Total	764,839	650			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Spec_RT

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	3,600	,069		51,954	<,001
	Spec_RT	,046	,001	,865	43,890	<,001

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,331[a]	,110	,108	1,0243

a. Prédicteurs : (Constante), NbAvis_IMDb

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	83,874	1	83,874	79,937	<,001[b]
	Résidus	680,965	649	1,049		
	Total	764,839	650			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), NbAvis_IMDb

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	6,309	,045		139,789	,000
	NbAvis_IMDb	3,204E-6	,000	,331	8,941	<,001

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,030[a]	,001	-,001	1,0851

a. Prédicteurs : (Constante), Année

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	,688	1	,688	,585	,445[b]
	Résidus	764,150	649	1,177		
	Total	764,839	650			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Année

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	12,418	7,748		1,603	,109
	Année	-,003	,004	-,030	-,765	,445

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,268[a]	,072	,071	1,0459

a. Prédicteurs : (Constante), Durée

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	54,959	1	54,959	50,240	<,001[b]
	Résidus	708,864	648	1,094		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Durée

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	4,908	,227		21,605	<,001
	Durée	,015	,002	,268	7,088	<,001

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,765[a]	,585	,585	,6991

a. Prédicteurs : (Constante), Critiques_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	447,644	1	447,644	915,909	<,001[b]
	Résidus	317,195	649	,489		
	Total	764,839	650			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Critiques_RT

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	4,808	,062		77,455	,000
	Critiques_RT	,029	,001	,765	30,264	<,001

a. Variable dépendante : Note_IMDb

On étudie maintenant si le niveau d'exigence des spectacteurs (ou le niveau des films) varie au cours du temps. Pour cela, on considère le modèle linéaire `Note_IMDb` / `Année`.

6. Quelle réponse apporteriez-vous quant à ces variations de niveaux au cours du temps ?

Enfin, on considère également les régressions `Note_IMDb` / `Durée` et `Note_IMDb` / `Critiques_RT`.

7. Parmi les 5 modèles de régression linéaire simple calculés jusque-là, lequel propose-t-il la meilleure modélisation ? Justifiez votre réponse.

Régressions linéaires multiples. On considère maintenant le modèle de régression linéaire multiple complet `Note_IMDb` / `Durée`, `Année`, `NbAvis_IMDb`, `Critiques_RT`, `Spec_RT`.

8. Que pensez-vous de ce modèle ? Justifiez votre réponse. Si vous pensez du bien de ce modèle : écrivez la relation qu'il propose. Si vous en pensez du mal, écrivez la relation proposée par un modèle alternatif proche, dont vous pensez du bien, et dont la détermination sera par ailleurs justifiée.
9. Quel est le nom de la méthode donnant lieu aux sorties de régression intitulées « Suite de régressions linéaires multiples » ? Dans cette suite, expliquez, en justifiant soigneusement votre choix, quel modèle réalise le meilleur compromis entre performance et simplicité.

Nous étudions maintenant l'impact du genre du film (drame, comédie, ou autre catégorie) sur la note moyenne reçue sur IMDb, toutes choses égales par ailleurs, en contrôlant par la durée, le nombre d'avis sur IMDb, et les notes moyennes critiques et spectacteurs sur Rotten Tomatoes. On introduit à cet effet deux variables binaires `GenreDrama` et `GenreComedy` valant respectivement 1 si et seulement si le genre est drame et comédie, et 0 sinon.

10. Que pouvez-vous dire de cet impact, à partir des deux sorties de régression linéaire multiple tenant compte du genre ? Il est demandé une réponse synthétique avec une bonne hauteur de vue.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,900[a]	,810	,808	,4749

a. Prédicteurs : (Constante), Spec_RT, Année, Durée, NbAvis_IMDb, Critiques_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	618,572	5	123,714	548,511	<,001[b]
	de Student	145,251	644	,226		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Spec_RT, Année, Durée, NbAvis_IMDb, Critiques_RT

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	-2,418	3,530		-,685	,494
	Durée	,005	,001	,088	4,712	<,001
	Année	,003	,002	,028	1,601	,110
	NbAvis_IMDb	5,126E-7	,000	,053	2,733	,006
	Critiques_RT	,012	,001	,303	12,472	<,001
	Spec_RT	,033	,001	,622	25,131	<,001

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,899[a]	,809	,808	,4755

a. Prédicteurs : (Constante), Spec_RT, Durée, NbAvis_IMDb, Critiques_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	617,994	4	154,499	683,343	<,001[b]
	de Student	145,829	645	,226		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Spec_RT, Durée, NbAvis_IMDb, Critiques_RT

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	3,230	,118		27,470	<,001
	Durée	,005	,001	,083	4,509	<,001
	NbAvis_IMDb	5,795E-7	,000	,060	3,165	,002
	Critiques_RT	,011	,001	,301	12,388	<,001
	Spec_RT	,033	,001	,621	25,063	<,001

a. Variable dépendante : Note_IMDb

Suite de régressions linéaires multiples

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,865[a]	,748	,748	,5449
2	,892[b]	,796	,796	,4905
3	,898[c]	,806	,805	,4788
4	,899[d]	,809	,808	,4755

a. Prédicteurs : (Constante), Spec_RT

b. Prédicteurs : (Constante), Spec_RT, Critiques_RT

c. Prédicteurs : (Constante), Spec_RT, Critiques_RT, Durée

d. Prédicteurs : (Constante), Spec_RT, Critiques_RT, Durée, NbAvis_IMDb

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	571,392	1	571,392	1924,121	<,001[b]
	de Student	192,432	648	,297		
	Total	763,823	649			
2	Régression	608,144	2	304,072	1263,716	<,001[c]
	de Student	155,679	647	,241		
	Total	763,823	649			
3	Régression	615,729	3	205,243	895,285	<,001[d]
	de Student	148,095	646	,229		
	Total	763,823	649			
4	Régression	617,994	4	154,499	683,343	<,001[e]
	de Student	145,829	645	,226		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), Spec_RT

c. Prédicteurs : (Constante), Spec_RT, Critiques_RT

d. Prédicteurs : (Constante), Spec_RT, Critiques_RT, Durée

e. Prédicteurs : (Constante), Spec_RT, Critiques_RT, Durée, NbAvis_IMDb

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	3,600	,069		51,962	<,001
	Spec_RT	,046	,001	,865	43,865	<,001
2	(Constante)	3,647	,062		58,372	<,001
	Spec_RT	,035	,001	,647	25,898	<,001
	Critiques_RT	,012	,001	,309	12,359	<,001
3	(Constante)	3,108	,112		27,783	<,001
	Spec_RT	,034	,001	,635	25,947	<,001
	Critiques_RT	,011	,001	,300	12,264	<,001
	Durée	,006	,001	,102	5,752	<,001
4	(Constante)	3,230	,118		27,470	<,001
	Spec_RT	,033	,001	,621	25,063	<,001
	Critiques_RT	,011	,001	,301	12,388	<,001
	Durée	,005	,001	,083	4,509	<,001
	NbAvis_IMDb	5,795E-7	,000	,060	3,165	,002

a. Variable dépendante : Note_IMDb

Impact du genre

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,902[a]	,814	,812	,4701

a. Prédicteurs : (Constante), GenreComedy, NbAvis_IMDb, Critiques_RT, GenreDrama, Durée, Spec_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	621,744	6	103,624	468,964	<,001[b]
	de Student	142,080	643	,221		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), GenreComedy, NbAvis_IMDb, Critiques_RT, GenreDrama, Durée, Spec_RT

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	3,361	,121		27,752	<,001
	Durée	,004	,001	,076	4,043	<,001
	NbAvis_IMDb	6,105E-7	,000	,063	3,362	<,001
	Critiques_RT	,011	,001	,289	11,928	<,001
	Spec_RT	,033	,001	,618	25,199	<,001
	GenreDrama	-,047	,041	-,021	-1,147	,252
	GenreComedy	-,245	,060	-,077	-4,115	<,001

a. Variable dépendante : Note_IMDb

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,902[a]	,814	,812	,4702

a. Prédicteurs : (Constante), GenreComedy, NbAvis_IMDb, Critiques_RT, Durée, Spec_RT

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	621,453	5	124,291	562,218	<,001[b]
	de Student	142,370	644	,221		
	Total	763,823	649			

a. Variable dépendante : Note_IMDb

b. Prédicteurs : (Constante), GenreComedy, NbAvis_IMDb, Critiques_RT, Durée, Spec_RT

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		B	Erreur standard	Bêta		
1	(Constante)	3,364	,121		27,775	<,001
	Durée	,004	,001	,072	3,895	<,001
	NbAvis_IMDb	6,225E-7	,000	,064	3,432	<,001
	Critiques_RT	,011	,001	,288	11,904	<,001
	Spec_RT	,033	,001	,617	25,170	<,001
	GenreComedy	-,223	,056	-,070	-3,956	<,001

a. Variable dépendante : Note_IMDb

Discrimination salariale femmes–hommes ?

Nous allons étudier le cas d'une banque des Etats-Unis accusée de discrimination salariale dans les années 1970. L'affaire a duré plusieurs décennies et nous nous concentrerons ici sur une partie des données, ainsi présentées par Ramsey, F.L. et Schafer, D.W. (2013), "The Statistical Sleuth : A Course in Methods of Data Analysis (3rd ed)", Cengage Learning :

> Data on employees from one job category (skilled, entry-level clerical) of a bank that was sued for sex discrimination. The data are on 32 male and 61 female employees, hired between 1965 and 1975.
>
> `Bsal`: Annual salary at time of hire (in USD)
> `Sal77`: Salary as of March 1975 (in USD)
> `Sex`: Sex of employee (1 if woman, 0 if man)
> `Senior`: Seniority (months since first hired)
> `Age`: Age of employee (in months)
> `Educ`: Education (in years)
> `Exper`: Work experience prior to employment with the bank (months)
>
> Reference : Roberts, H.V. (1979), "Harris Trust and Savings Bank : An Analysis of Employee Compensation", Report 7946, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

Un extrait du jeu de données est fourni ci-dessous.

	SexT	Sex	Bsal	Sal77	Senior	Age	Educ	Exper
10	Male	0	6900	10920	75	416	15	132
11	Male	0	6900	10920	89	481	12	175
12	Male	0	5400	12660	91	331	15	18
13	Male	0	6000	12960	66	355	15	64
14	Male	0	6000	12360	86	348	15	25
15	Female	1	5100	8940	95	640	15	165
16	Female	1	4800	8580	98	774	12	381
17	Female	1	5280	8760	98	557	8	190
18	Female	1	5280	8040	88	745	8	90

Prise en main du sujet. Nous allons tout d'abord essayer de mieux comprendre les deux variables de salaire.

1. Sachant que dans la période considérée l'inflation a parfois été très élevée (et souvent à plus de 4%), les résultats des deux modèles de régression linéaire `Sal77` / `Senior` et `Bsal` / `Senior` sont-ils économiquement cohérents et que traduisent-ils ?

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,286[a]	,082	,072	683,707

a. Prédicteurs : (Constante), Senior

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	3784914,704	1	3784914,704	8,097	,005[b]
	de Student	42538375,618	91	467454,677		
	Total	46323290,323	92			

a. Variable dépendante : Bsal

b. Prédicteurs : (Constante), Senior

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95,0% pour B	
Modèle		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Borne supérieure
1	(Constante)	7047,746	576,307		12,229	<,001	5902,984	8192,509
	Senior	-19,779	6,951	-,286	-2,845	,005	-33,587	-5,972

a. Variable dépendante : Bsal

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,126[a]	,016	,005	1785,116

a. Prédicteurs : (Constante), Senior

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	4674672,150	1	4674672,150	1,467	,229[b]
	de Student	289984243,98	91	3186640,044		
	Total	294658916,13	92			

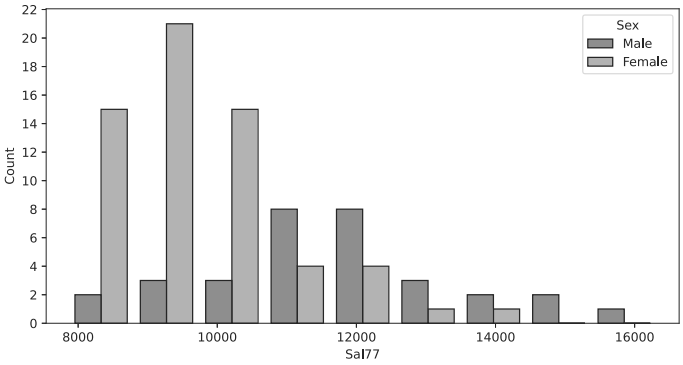
a. Variable dépendante : Sal77

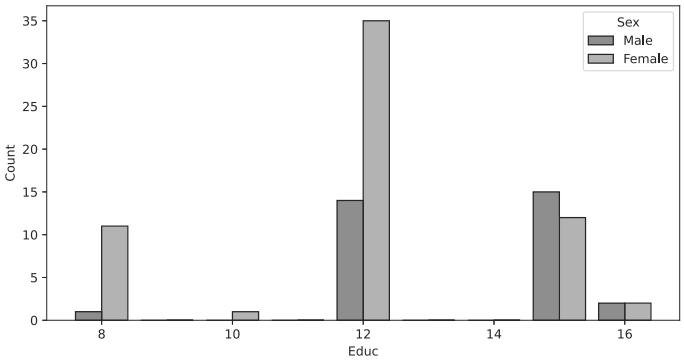
b. Prédicteurs : (Constante), Senior

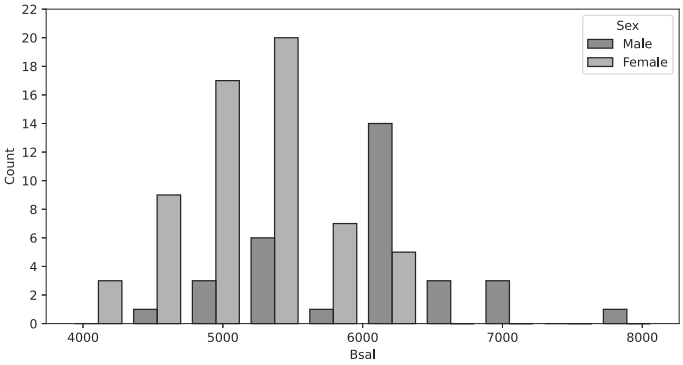
Coefficients[a]

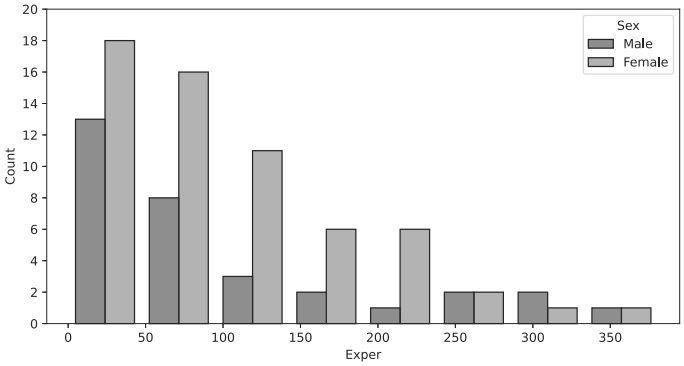
		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95,0% pour B	
Modèle		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Borne supérieure
1	(Constante)	8584,280	1504,702		5,705	<,001	5595,374	11573,185
	Senior	21,981	18,149	,126	1,211	,229	-14,069	58,032

a. Variable dépendante : Sal77









Statistiques de groupe

	Sex	N	Moyenne	Ecart type	Moyenne d'erreur standard
Sal77	Male	32	11681,25	1963,210	347,050
	Female	61	9717,05	1253,001	160,430
Bsal	Male	32	5956,88	690,733	122,106
	Female	61	5138,85	539,871	69,123

Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes							
						Signification		Différence moyenne	Erreur standard	Intervalle de confiance de la différence à 95 %	
		F	Sig.	t	df	p unilatéral	p bilatéral			Inférieur	Supérieur
Sal77	Hypothèse de variances égales	7,259	,008	5,872	91	<,001	<,001	1964,201	334,475	1299,808	2628,594
	Hypothèse de variances inégales			5,137	44,612	<,001	<,001	1964,201	382,337	1193,950	2734,451
Bsal	Hypothèse de variances égales	,344	,559	6,293	91	<,001	<,001	818,023	129,997	559,799	1076,247
	Hypothèse de variances inégales			5,830	51,329	<,001	<,001	818,023	140,313	536,376	1099,669

2. A partir des histogrammes et de la sortie logiciel ci-dessus, expliquer pourquoi une impression de discrimination se dégage, mais pourquoi la prouver requerra un peu d'efforts.
3. Expliquer les liens, similitudes et différences entre la sortie logiciel ci-dessus et les deux sorties de régression ci-contre.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,524[a]	,275	,267	1532,364

a. Prédicteurs : (Constante), Sex

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	80978297,277	1	80978297,277	34,486	<,001[b]
	de Student	213680618,85	91	2348138,669		
	Total	294658916,13	92			

a. Variable dépendante : Sal77

b. Prédicteurs : (Constante), Sex

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	11681,250	270,886		43,122	<,001	11143,168	12219,332
	Sex	-1964,201	334,475	-,524	-5,872	<,001	-2628,594	-1299,808

a. Variable dépendante : Sal77

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,551[a]	,303	,296	595,571

a. Prédicteurs : (Constante), Sex

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	14045183,150	1	14045183,150	39,597	<,001[b]
	de Student	32278107,172	91	354704,474		
	Total	46323290,323	92			

a. Variable dépendante : Bsal

b. Prédicteurs : (Constante), Sex

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	5956,875	105,283		56,580	<,001	5747,743	6166,007
	Sex	-818,023	129,997	-,551	-6,293	<,001	-1076,247	-559,799

a. Variable dépendante : Bsal

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,547[a]	,299	,291	1506,673

a. Prédicteurs : (Constante), Age

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	88083020,259	1	88083020,259	38,802	<,001[b]
	de Student	206575895,87	91	2270064,790		
	Total	294658916,13	92			

a. Variable dépendante : Sal77

b. Prédicteurs : (Constante), Age

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	13703,562	553,968		24,737	<,001	12603,172	14803,952
	Age	-6,979	1,120	-,547	-6,229	<,001	-9,204	-4,753

a. Variable dépendante : Sal77

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,421[a]	,177	,168	1632,190

a. Prédicteurs : (Constante), Educ

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	52230917,682	1	52230917,682	19,606	<,001[b]
	de Student	242427998,45	91	2664043,939		
	Total	294658916,13	92			

a. Variable dépendante : Sal77

b. Prédicteurs : (Constante), Educ

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	6264,513	947,606		6,611	<,001	4382,210	8146,815
	Educ	330,129	74,557	,421	4,428	<,001	182,030	478,228

a. Variable dépendante : Sal77

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,372[a]	,138	,129	1670,316

a. Prédicteurs : (Constante), Exper

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	40773096,398	1	40773096,398	14,614	<,001[b]
	de Student	253885819,73	91	2789954,063		
	Total	294658916,13	92			

a. Variable dépendante : Sal77

b. Prédicteurs : (Constante), Exper

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Borne supérieure
1	(Constante)	11131,681	259,511		42,895	<,001	10616,194	11647,168
	Exper	-7,320	1,915	-,372	-3,823	<,001	-11,123	-3,516

a. Variable dépendante : Sal77

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,422[a]	,178	,169	1631,062

a. Prédicteurs : (Constante), Bsal

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	52565976,477	1	52565976,477	19,759	<,001[b]
	de Student	242092939,65	91	2660361,974		
	Total	294658916,13	92			

a. Variable dépendante : Sal77

b. Prédicteurs : (Constante), Bsal

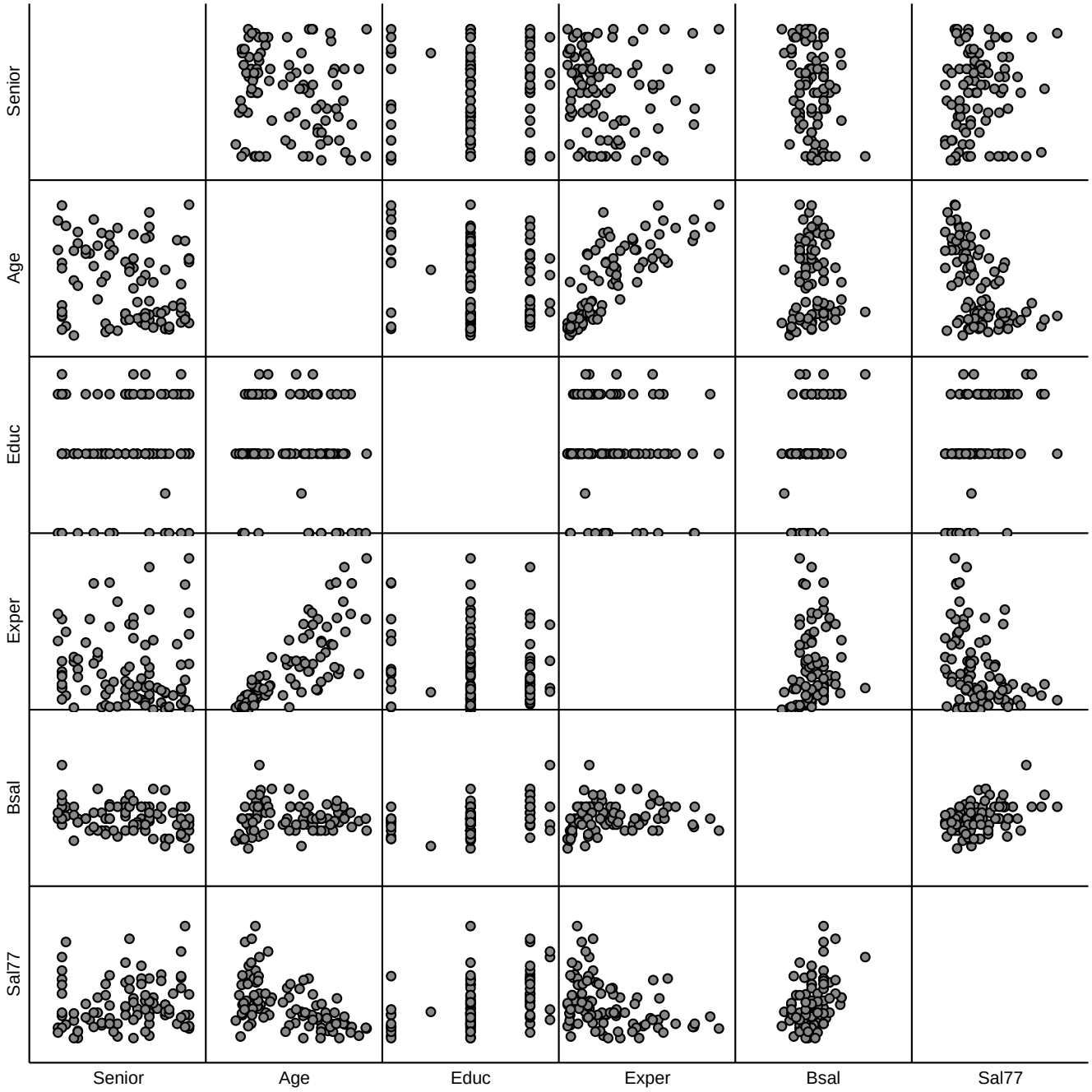
Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés			Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta	t	Sig.	Borne inférieure	Borne supérieure
1	(Constante)	4618,890	1309,925		3,526	<,001	2016,885	7220,895
	Bsal	1,065	,240	,422	4,445	<,001	,589	1,541

a. Variable dépendante : Sal77

Régressions linéaires simples. Nous allons prendre pour variable d'étude principale `Sal77`, à cause des raisons apparues en question 1.

4. Parmi l'ensemble des modèles de régression linéaire simple proposés pour `Sal77`, y compris ceux avec `Senior` et `Sex` comme variable explicative :
   — Quels sont ceux qui sont valides statistiquement ?
   — Quels sont ceux qui sont valides économiquement ? On commencera par rappeler brièvement en quoi consiste la validation économique.
   — Quel est le meilleur modèle, d'un point de vue quantitatif ?
5. Ecrire la relation proposée par le modèle `Sal77` / `Educ`. Interpréter soigneusement les valeurs numériques du coefficient de pente et de l'ordonnée à l'origine.
6. Quels sont les salaires attendus en 1977 pour des employés avec 15 années de scolarité ?


Senior
Age
Educ
Exper
Bsal
Sal77
Senior
Age
Educ
Exper
Bsal
Sal77

Régressions linéaires multiples. Nous considérons les deux sorties logiciel ci-contre (où nous avons supprimé diverses notes de bas de tableau pour augmenter la lisibilité). L'objectif est d'établir un modèle régression multiple de `Sal77` en fonction (de certaines) des variables suivantes : `Senior`, `Age`, `Educ`, `Exper`, et bien entendu, `Sex`.

7. Que pensez-vous du modèle complet ? Indiquez au préalable où vous le lisez, et expliquez, le cas échéant, ce qui a été entrepris au vu de votre jugement sur ce modèle complet.
8. Donnez un nom à la seconde série de tableaux ci-contre. En tenant les validités économiques acquises, quel modèle réalise, d'après vous, le meilleur compromis entre qualité et interprétabilité ?
9. Ecrivez la relation proposée par le modèle `Sal77` / `Age`, `Sex`.
10. Vient-on de mettre en évidence une discrimination salariale selon le sexe ? Quantifier précisément son impact salarial le cas échéant : quel montant moyen de différence de salaire les données disponibles assurent-elles avec grande confiance ? (Quel montant serait considéré par un tribunal ?) Que pensez-vous de ce montant en regard des ordres de grandeurs des salaires annuels ?
11. Le salaire de l'emplyée #15 est-il compatible avec le modèle `Sal77` / `Age`, `Sex` ?

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,706[a]	,499	,470	1303,058
2	,706[b]	,499	,476	1295,771
3	,704[c]	,495	,478	1292,482

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	146936337,20	5	29387267,439	17,307	<,001[b]
	de Student	147722578,93	87	1697960,677		
	Total	294658916,13	92			
2	Régression	146904954,79	4	36726238,697	21,874	<,001[c]
	de Student	147753961,34	88	1679022,288		
	Total	294658916,13	92			
3	Régression	145983656,19	3	48661218,731	29,130	<,001[d]
	de Student	148675259,94	89	1670508,539		
	Total	294658916,13	92			

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	11144,796	1672,827		6,662	<,001	7819,872	14469,720
	Senior	1,846	13,582	,011	,136	,892	-25,148	28,841
	Educ	170,331	63,765	,217	2,671	,009	43,590	297,071
	Exper	-2,022	2,706	-,103	-,747	,457	-7,401	3,357
	Age	-4,043	1,848	-,317	-2,188	,031	-7,717	-,370
	Sex	-1389,772	330,759	-,371	-4,202	<,001	-2047,191	-732,353
2	(Constante)	11314,298	1109,045		10,202	<,001	9110,305	13518,290
	Educ	170,344	63,409	,217	2,686	,009	44,332	296,355
	Exper	-1,981	2,675	-,101	-,741	,461	-7,296	3,334
	Age	-4,089	1,807	-,320	-2,263	,026	-7,680	-,498
	Sex	-1389,975	328,906	-,371	-4,226	<,001	-2043,606	-736,344
3	(Constante)	11597,261	1038,538		11,167	<,001	9533,707	13660,815
	Educ	169,122	63,226	,216	2,675	,009	43,492	294,752
	Age	-5,199	1,008	-,407	-5,160	<,001	-7,201	-3,197
	Sex	-1300,205	304,989	-,347	-4,263	<,001	-1906,213	-694,198

a. Variable dépendante : Sal77

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,547[a]	,299	,291	1506,673
2	,674[b]	,455	,443	1335,946
3	,704[c]	,495	,478	1292,482

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	88083020,259	1	88083020,259	38,802	<,001[b]
	de Student	206575895,87	91	2270064,790		
	Total	294658916,13	92			
2	Régression	134031364,55	2	67015682,276	37,549	<,001[c]
	de Student	160627551,58	90	1784750,573		
	Total	294658916,13	92			
3	Régression	145983656,19	3	48661218,731	29,130	<,001[d]
	de Student	148675259,94	89	1670508,539		
	Total	294658916,13	92			

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	13703,562	553,968		24,737	<,001	12603,172	14803,952
	Age	-6,979	1,120	-,547	-6,229	<,001	-9,204	-4,753
2	(Constante)	14060,601	496,210		28,336	<,001	13074,792	15046,409
	Age	-5,612	1,029	-,440	-5,452	<,001	-7,656	-3,567
	Sex	-1533,022	302,136	-,409	-5,074	<,001	-2133,267	-932,776
3	(Constante)	11597,261	1038,538		11,167	<,001	9533,707	13660,815
	Age	-5,199	1,008	-,407	-5,160	<,001	-7,201	-3,197
	Sex	-1300,205	304,989	-,347	-4,263	<,001	-1906,213	-694,198
	Educ	169,122	63,226	,216	2,675	,009	43,492	294,752

a. Variable dépendante : Sal77

Cas du salaire d'embauche. La plupart des sujets et cours traitant de ce jeu de données s'intéresse plutôt à `Bsal`, malgré les difficultés soulevées à la question 1.

12. La sortie logiciel ci-dessous considérait les variables `Educ`, `Exper`, `Age`, et `Sex` pour modéliser linéairement `Bsal`. Pourquoi pensez-vous qu'elle soit effectivement bien plus sympathique à discuter en cours ?

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,551[a]	,303	,296	595,571
2	,603[b]	,363	,349	572,437
3	,643[c]	,414	,394	552,358

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression	14045183,150	1	14045183,150	39,597	<,001[b]
	de Student	32278107,172	91	354704,474		
	Total	46323290,323	92			
2	Régression	16831743,945	2	8415871,973	25,683	<,001[c]
	de Student	29491546,377	90	327683,849		
	Total	46323290,323	92			
3	Régression	19169458,575	3	6389819,525	20,943	<,001[d]
	de Student	27153831,747	89	305099,233		
	Total	46323290,323	92			

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Intervalle de confiance à 95,0% pour B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(Constante)	5956,875	105,283		56,580	<,001	5747,743	6166,007
	Sex	-818,023	129,997	-,551	-6,293	<,001	-1076,247	-559,799
2	(Constante)	4864,933	387,882		12,542	<,001	4094,339	5635,528
	Sex	-691,808	132,232	-,466	-5,232	<,001	-954,510	-429,107
	Educ	80,698	27,673	,260	2,916	,004	25,721	135,675
3	(Constante)	4204,624	443,832		9,473	<,001	3322,739	5086,509
	Sex	-765,494	130,341	-,515	-5,873	<,001	-1024,478	-506,509
	Educ	92,145	27,021	,296	3,410	<,001	38,456	145,834
	Age	1,192	,431	,236	2,768	,007	,336	2,048

a. Variable dépendante : Bsal

9

Résumé de cours

Chapitre 1. Vocabulaire et modélisation statistiques

Vocabulaire. En termes de vocabulaire, il faut retenir :

- Population = ensemble des invidivus statistiques
- Echantillon = petit sous-ensemble d'individus sélectionnés
- Variable d'intérêt = les quantités mesurées ou pouvant être mesurées sur les individus
- Paramètre d'intérêt = une certaine proportion p_0 ou moyenne de population μ_0 (moyenne de la variable d'intérêt sur tous les individus de la population)
- Statistiques d'échantillon = fréquence d'échantillon $\overline{x}_n$; ou moyenne d'échantillon $\overline{x}_n$ et écart-type d'échantillon $s_{x,n}$

Modélisation. Lorsque le sondage a été effectué par tirage uniformément au hasard, les données $x_1, \dots, x_n$ peuvent être vues comme la réalisation de $X_1, \dots, X_n$ indépendantes et identiquement distribuées selon soit une loi de Bernoulli de paramètre p_0, soit une loi générale d'espérance μ_0. Cette observation permet de relier (via des intervalles de confiance ou des tests d'hypothèse) la moyenne d'échantillon $\overline{x}_n$ au paramètre d'intérêt p_0 ou μ_0.

Echantillons « représentatifs » ou biais possibles. Un cas idéal est de disposer d'une liste exhaustive de la population. A défaut, les sondeurs procèdent par quotas pour garantir des échantillons « représentatifs », i.e., distribués comme la population. Les sondeurs sont également attentifs à éviter différents biais : de motivation, d'endogamie, liés aux canaux (en ligne, téléphone, courrier, etc.) utilisés, et de manière générale, planifient en amont la collecte de données, en variant les espaces et les temps en fonction de la population.

Chapitre 2. Intervalles de confiance

Formes des intervalles de confiance. On corrige la moyenne d'échantillon $\overline{x}_n$ en un intervalle de valeurs plausibles pour p_0 ou μ_0. Cette correction peut s'effectuer

- de manière symétrique (bilatère), dans les cas neutres : on assortit $\overline{x}_n$ d'une marge d'erreur $\pm\dots$;
- d'un seul côté (de manière unilatère), dans les cas où des objectifs le commandent,
  - avec une estimation basse de la forme $\overline{x}_n - \dots$ et à laquelle μ_0 devrait être supérieure,
  - ou avec une estimation haute de la forme $\overline{x}_n + \dots$ et à laquelle μ_0 devrait être inférieure.

Formules pour une proportion p_0. Avec confiance 95%, la proportion de population p_0 est :

— dans l'intervalle symétrique $\left[\overline{x}_n \pm 1.96\,\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}}\right]$;

— supérieure à l'estimation basse $\overline{x}_n - 1.645\,\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}}$;

— inférieure à l'estimation haute $\overline{x}_n + 1.645\,\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}}$.

Par exemple, à partir de $n = 2\,159$ véhicules contrôlés avec 563 en infraction au stationnement, soit une proportion d'échantillon $\overline{x}_{2\,159} = 563/2\,159$, on peut émettre, entre autres, une des deux assertions suivantes sur la « vraie » proportion p_0 de l'ensemble des véhicules stationnés (en différents lieux et temps), voir détails de calculs page 18 :

— Avec grande confiance, la proportion p_0 vaut 26.1% ± 1.9%, i.e., se situe entre 24.2% et 28.0%.
— Avec grande confiance, la proportion p_0 est supérieure à 24.5%.

Formules pour une moyenne μ_0. Avec confiance 95%, la moyenne de population μ_0 est :

— dans l'intervalle symétrique $\left[\overline{x}_n \pm 1.96\,\frac{s_{x,n}}{\sqrt{n}}\right]$;

— supérieure à l'estimation basse $\overline{x}_n - 1.645\,\frac{s_{x,n}}{\sqrt{n}}$;

— inférieure à l'estimation haute $\overline{x}_n + 1.645\,\frac{s_{x,n}}{\sqrt{n}}$.

Par exemple, sur un échantillon de $n = 1\,596$ voitures stationnées, on obtient une moyenne d'échantillon de temps de stationnement de $\overline{x}_{1\,596}$ = 2h 41 min, associée à un écart-type d'échantillon de $s_{x,1\,596} = 68$ minutes. Après calculs et arrondis (voir page 16 pour les détails), on peut énoncer l'une des deux assertions suivantes sur le temps moyen μ_0 de stationnement payé (en différents lieux et temps) par l'ensemble des véhicules en règle :

— Avec grande confiance, le temps moyen μ_0 vaut 2h 41 min ± 4 min, i.e., se situe entre 2h 37 min et 2h 45 min.
— Avec grande confiance, le temps moyen μ_0 est inférieur à 2h 44 min.

Gestion des arrondis. On rappelle qu'il convient d'arrondir les résultats numériques de telle sorte à agrandir les intervalles de confiance en résultant : donc les marges d'erreur à la hausse, les estimations basses à la baisse, et les estimations hautes à la hausse.

Intervalles de confiance simultanés. L'exploitation combinée du résultat de deux intervalles de confiance de niveau 95% ne vaut qu'avec confiance 90%.

Pour cette combinaison, on applique évidemment les règles usuelles de gestion des inégalités. Par exemple, pour obtenir une estimation haute sur $2\mu_0/p_0$, on doit combiner une estimation haute sur μ_0 (notre « inférieure à 2h 44 min ») et une estimation basse sur p_0 (notre « supérieure à 24.5% »).

Planification. Pour déterminer le nombre total n de sondés à interroger pour obtenir un intervalle de confiance bilatère sur μ_0 avec une marge d'erreur $\pm\varepsilon$ imposée par le contexte (l'énoncé à HEC ou votre manager plus tard), on procède en deux temps :

1. on effectue un coup de sonde sur $m \geqslant 30$ ou $m \geqslant 50$ individus ;
2. on calcule alors le nombre d'individus total n (soit $n - m$ individus supplémentaires) à sonder pour obtenir la précision $\pm\varepsilon$ désirée.

Le calcul de n est donné par les observations suivantes : avec le coup de sonde et avec l'échantillon

final, on dispose des précisions (des marges d'erreur) respectives

$$1.96\,\frac{s_{x,m}}{\sqrt{m}} \qquad \text{et} \qquad 1.96\,\frac{s_{x,n}}{\sqrt{n}} \approx 1.96\,\frac{s_{x,m}}{\sqrt{n}} \approx \frac{1}{\gamma}\left(1.96\,\frac{s_{x,m}}{\sqrt{m}}\right),$$

sous l'hypothèse que l'écart-type ne varie pas trop, $s_{x,m} \approx s_{x,n}$, et en exprimant n sous la forme $n = m\gamma^2$. Cela conduit à la règle :

> La taille d'échantillon n est donnée par la taille m du coup de sonde fois le carré γ^2 du facteur d'amélioration γ à apporter, entre la marge du coup de sonde et l'objectif ε : $n = m\gamma^2$.

Par exemple, si la marge avec un coup de sonde de $n = 67$ données est de ± 4.31 euros, et que l'objectif de précision est fixé à ± 1.50 euros, on cherche une amélioration d'un facteur $\gamma = 4.31/1.50$ et on retient un échantillon final de taille $n = m\gamma^2 = 67\,(4.31/1.50)^2 = 553.15...$, arrondie à 554.

Ce qui est écrit ci-dessus pour l'estimation d'une moyenne μ_0 s'adapte au cas d'une proportion p_0.

Chapitre 3. Tests d'hypothèses : comparaison à une référence

Choix des hypothèses. Les critères de choix généraux sont les suivants :

- H_1 est ce que l'on veut prouver, ou ce qui susciterait réflexion ou action, et H_0 est définie en fonction de ce choix de H_1.
- H_0 contient toujours le cas d'égalité, et même, par passage au cas limite, H_0 est réduite au cas d'égalité.
- Ainsi, il faut et il suffit de justifier si H_1 est bilatère (déviations dans les deux sens), ou unilatère (déviations uniquement vers les valeurs supérieures ou inférieures).

Dans ce chapitre où l'on veut comparer la proportion de population p_0 à une proportion de référence p_{ref}, ou la moyenne de population μ_0 à une moyenne de référence μ_{ref}, cela signifie que H_1 est nécessairement de l'une des formes suivantes :

$$\begin{array}{rccc} & H_1 : p_0 \neq p_{ref}, & H_1 : p_0 < p_{ref}, & H_1 : p_0 > p_{ref}, \\ \text{ou} & H_1 : \mu_0 \neq \mu_{ref}, & H_1 : \mu_0 < \mu_{ref}, & H_1 : \mu_0 > \mu_{ref}, \end{array}$$

tandis qu'à la limite, H_0 est $H_0 : p_0 = p_{ref}$ ou $H_0 : \mu_0 = \mu_{ref}$.

Encore une fois, il faut et il suffit de justifier si H_1 est de la forme $\neq$, $<$ ou $>$.

Statistique de test = écart renormalisé. Une statistique de test t_n normalise l'écart entre la fréquence d'échantillon $\overline{x}_n$ et la fréquence de référence p_{ref}, ou entre la moyenne d'échantillon $\overline{x}_n$ et la moyenne de référence μ_{ref}, afin de donner un comportement plus universel à cet écart : selon une loi normale standard sous les hypothèses $H_0 : p_0 = p_{ref}$ et $H_0 : \mu_0 = \mu_{ref}$.

Test contre une proportion de référence p_{ref} : formule. La statistique de test admet la réalisation

$$t_n = \sqrt{n}\,\frac{\overline{x}_n - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}.$$

Son comportement sous H_0 est approximativement issu d'une loi normale standard $\mathcal{N}(0,1)$.

Sous H_1, la statistique de test est issue d'un comportement différent (qui provient du fait que $\overline{x}_n$ est proche de p_0) :

- Sous $H_1 : p_0 > p_{ref}$, la statistique de test t_n tend à prendre des valeurs positives grandes ; on place une flèche $\rightarrow$ (moyen mnémotechnique : de même sens que $>$) sur le dessin de la P–valeur.

— Sous H_1 : $p_0 < p_{ref}$, ce sont de grandes valeurs négatives, représentées par une flèche $\leftarrow$ (de même sens que le signe $<$).
— Sous H_1 : $p_0 \neq p_{ref}$, les deux comportements sont possibles, on place deux flèches $\leftarrow$ et $\rightarrow$.

Test contre une moyenne de référence μ_{ref} : formule. La statistique de test admet la réalisation

$$t_n = \sqrt{n}\,\frac{\overline{x}_n - \mu_{ref}}{s_{x,n}}\,.$$

Son comportement sous H_0 est approximativement issu d'une loi normale standard $\mathcal{N}(0,1)$.

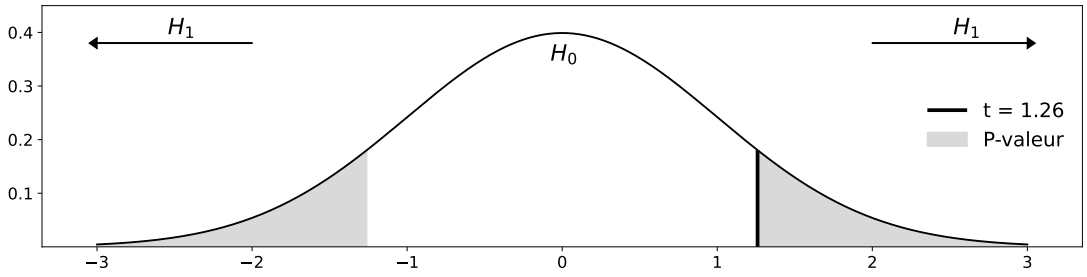
Sous H_1, la statistique de test est issue d'un comportement différent (qui provient du fait que $\overline{x}_n$ est proche de μ_0) :

— Sous H_1 : $\mu_0 > \mu_{ref}$, la statistique de test t_n tend à prendre des valeurs positives grandes et on place une flèche $\rightarrow$ (moyen mnémotechnique : de même sens que $>$) sur le dessin de la P–valeur.
— Sous H_1 : $\mu_0 < \mu_{ref}$, ce sont de grandes valeurs négatives, représentées par une flèche $\leftarrow$ (de même sens que le signe $<$).
— Sous H_1 : $\mu_0 \neq \mu_{ref}$, les deux comportements sont possibles, on place deux flèches $\leftarrow$ et $\rightarrow$.

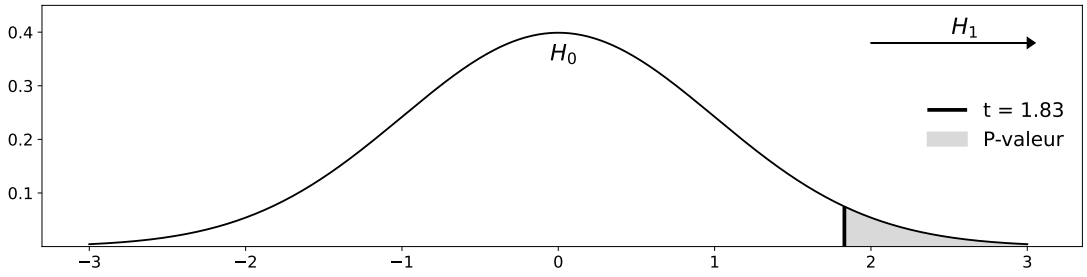
P–valeur. Une fois la valeur numérique de t_n calculée et les comportements sous H_1 déterminés, on peut tracer une figure pour apprécier si la valeur de t_n est typique de H_0 ou est au contraire atypique.

Cette figure permet de lire la P–valeur, qui est un indice de crédibilité de H_0 au des données et face à H_1. La P–valeur est donnée par la probabilité, si l'on répétait l'expérience, d'obtenir une valeur de la statistique de test au moins aussi contradictoire avec H_0 que la valeur t_n actuelle. Si la P–valeur est faible, c'est que la valeur présente est déjà très contradictoire ; si la P–valeur est grande, c'est que la valeur présente est tout à fait compatible avec H_0.

Un premier exemple calcule la P–valeur pour $t = 1.26$ dans le test bilatère $H_0 : p_0 = p_{ref}$ ou $\mu_0 = \mu_{ref}$ vs. $H_1 : p_0 \neq p_{ref}$ ou $\mu_0 \neq \mu_{ref}$.



Un second exemple calcule la P–valeur pour $t = 1.83$ dans le test unilatère $H_0 : p_0 = p_{ref}$ ou $\mu_0 = \mu_{ref}$ vs. $H_1 : p_0 > p_{ref}$ ou $\mu_0 > \mu_{ref}$.



Conclusions : statistique et grand public. Les hypothèses H_0 et H_1 jouent des rôles très dissymétriques : on ne veut se départir de H_0 que s'il le faut vraiment, que si les données sont en contradiction grave avec H_0. On retient un seuil de P–valeur de 5% à cet effet :

— si la P–valeur est $< 5\%$, on rejette H_0 et on passe à H_1 ;
— si la P–valeur est $> 5\%$, on conserve H_0 (ce qui ne signifie pas qu'on valide H_0).

Les tests étant par définition attachés à H_0, ils ne font de progrès que négatifs.

— Lorsque H_0 est rejetée, on peut conclure de manière affirmative, car avec une certitude raisonnable, on a étali un fait :

> « Ces données montrent [ce qui correspond à H_1]. »

— Lorsque H_0 est rejetée, la conclusion doit traduire le fait que tout au plus l'on n'a pas pu rejeter H_0 avec les données disponibles (soit que H_0 soit effectivement vraie, soit que l'on ne dispose de pas suffisamment de données pour l'infirmer), ce qui en réalité ne nous fait pas vraiment avancer et est plutôt une non-conclusion :

> « Ces données ne permettent pas d'exclure [ce qui correspond à H_0]. »

Chapitre 4. Test de comparaison de moyennes appariées

Concept de données appariées. On fait face à des données appariées lorsque pour les mêmes individus statistiques, on effectue deux mesures

— de la même variable mais dans des conditions différentes (par exemple, prix de produits dans deux magasins ; notes données sur les mêmes copies par deux correcteurs ; salaires en 2017 et 2023 pour les mêmes sondés ; montants de commandes avec et sans promotion sur deux trimestres successifs pour les mêmes clients) ;
— au sein d'une « paire », chaque paire étant considérée comme un individu statistique (salaires au sein d'un couple hétérosexuel ; tailles de duos père–fils).

Visuellement, les données appariées se présentent dans un tableur sous la forme suivante :

ID	Variable 1	Variable 2	Différence
1	16	12	4
2	18	16	2
3	30	29	1
4	29	32	-3
...	...	...	

La colonne ID permet de numéroter les individus statistiques, et pour chacun d'eux, deux mesures sont effectuées (`Variable 1` et `Variable 2`), et on se ramène à considérer les écarts entre ces deux mesures (`Différence`).

Traitement par étude des différences. Si n désigne la taille d'échantillon, on dispose donc de couples de données (x_1, y_1), ..., (x_n, y_n), où les mesures x_j et $_j$ effectuées toutes deux sur le même individu j sont très liées.

C'est pourquoi on s'intéresse plutôt à la série des différences $z_j = x_j - y_j$ pour $j = 1, \ldots, n$, que l'on résume par sa moyenne d'échantillon $\bar{z}_n$ et son écart-type d'échantillon $s_{z,n}$.

Test d'existence d'une différence moyenne significative. On s'intéresse à la différence moyenne de population Δ_0 entre les deux variables, et on teste $H_0 : \Delta_0 = 0$ (les deux variables considérées sont égales en moyenne sur la population) vs. une hypothèse H_1 bilatère ($\Delta_0 \neq 0$) ou unilatère ($\Delta_0 < 0$ ou

$\Delta_0 > 0$). C'est-à-dire que l'on se ramène à un test de comparaison à la moyenne de référence $\mu_{\text{ref}} = 0$. En particulier, la statistique de test considérée est

$$t_n = \sqrt{n}\,\frac{\overline{z}_n - 0}{s_{z,n}} = \sqrt{n}\,\frac{\overline{z}_n}{s_{z,n}}.$$

Effet, direction d'effet, taille d'effet. Si le test précédent rejette H_0, alors un effet est démontré : les deux variables sont différentes en moyenne sur la population.

Lorsque le test est bilatère, il convient encore de voir quelle moyenne de population est plus grande que l'autre : c'est la direction d'effet. Elle est donnée par le signe de $\overline{z}_n$.

On peut enfin proposer une taille d'effet, i.e., un intervalle de confiance sur Δ_0. Dans le cas bilatère, il est donné par

$$\left[\overline{z}_n \pm 1.96\,\frac{s_{z,n}}{\sqrt{n}}\right].$$

Chapitre 5. Comparaison de moyennes ou proportions indépendantes

On part de deux séries de données $x_1, \dots, x_n$ et $y_1, \dots, y_m$, de longueurs non nécessairement identiques, obtenues par deux échantillonnages indépendants. On veut déterminer si les fréquences ou moyennes d'échantillon $\overline{x}_n$ et $\overline{y}_m$ sont significativement différentes, c'est-à-dire si au vu de leurs valeurs, on peut conclure que les fréquences de population p_0 et q_0 ou les moyennes de population μ_0 et m_0 correspondantes sont différentes.

On introduit la série regroupée $z_1, \dots, z_{n+m}$ correspondant à la juxtaposition des deux séries de données $x_1, \dots, x_n$ et $y_1, \dots, y_m$.

Comparaison de deux proportions indépendantes p_0 et q_0

Tableau 2×2. Dans ce cas, modulo recodage, les données sont binaires ($x_j \in \{0,1\}$ et $y_k \in \{0,1\}$) et peuvent être résumées dans un tableau 2×2 de la forme suivante, où par exemple, $N_{x,0}$ désigne le nombre de 0 dans l'échantillon des x_j et $N_{z,1}$, le nombre de 1 au total, lorsque l'on fusionne les x_j et les y_k en un unique super-échantillon :

Echantillons	Modalités		Total
	0	1	
Echantillon des x	$N_{x,0}$	$N_{x,1}$	n
Echantillon des y	$N_{y,0}$	$N_{y,1}$	m
Total	$N_{z,0}$	$N_{z,1}$	$n+m$

Statistique de test : formule et comportements. La statistique de test admet la réalisation

$$t_{n,m} = \frac{\overline{x}_n - \overline{y}_m}{\sqrt{(1/n + 1/m)\;\overline{z}_{n+m}(1-\overline{z}_{n+m})}}, \qquad \text{où}$$

— $\overline{x}_n = N_{x,1}/n$ est la proportion de 1 dans la série des x_j,
— $\overline{y}_m = N_{y,1}/m$ est celle dans la série des y_k,

— et $\overline{z}_{n+m} = N_{z,1}/(n+m) = (N_{x,1} + N_{y,1})/(n+m)$ est celle dans la série regroupée correspondant à l'ensemble des x_j et y_k.

Sous $H_0 : p_0 = q_0$, la valeur de $t_{n,m}$ est approximativement issus d'une loi normale standard $\mathcal{N}(0,1)$.

Sous H_1, la statistique de test est issue d'un comportement différent (qui provient du fait que $\overline{x}_n$ et $\overline{y}_m$ sont respectivement proches de p_0 et q_0) :

— Sous $H_1 : p_0 > q_0$, la statistique de test $t_{n,m}$ tend à prendre des valeurs positives grandes et on place une flèche $\rightarrow$ (moyen mnémotechnique : de même sens que $>$) sur le dessin de la P–valeur.
— Sous $H_1 : p_0 < q_0$, ce sont de grandes valeurs négatives, représentées par une flèche $\leftarrow$ (de même sens que le signe $<$).
— Sous $H_1 : p_0 \neq q_0$, les deux comportements sont possibles, on place deux flèches $\leftarrow$ et $\rightarrow$.

P–valeur et conclusion. Le calcul de la P–valeur et la forme des conclusions sont alors les mêmes que dans le cas du Chapitre 3.

Comparaison de deux moyennes indépendantes μ_0 et m_0

Hypothèses testées : objectif et moyen. Nous voulons tester ici (c'est l'objectif) l'égalité de deux moyennes de populations, $H_0 : \mu_0 = m_0$, contre une hypothèse H_1 unilatère ou bilatère.

Pour des raisons techniques, il conviendra de commencer (c'est un moyen) par tester l'égalité des variances des deux populations. Selon le résultat de ce pré-test, on appliquera l'une ou l'autre parmi deux formules possibles pour la statistique de test de $H_0 : \mu_0 = m_0$.

Lecture de sorties logiciel. Pour ce test, il ne vous sera demandé que de savoir lire des sorties logiciel, de la forme suivante (et non pas de le mettre en œuvre vous-même) :

	Prof	N	Moyenne	Ecart type	Erreur standard
Notes	Prof. 1	40	14,300	3,5641	,5635
	Prof. 2	36	14,847	3,1933	,5322

Test des échantillons indépendants

		Test de Levene sur l'égalité des variances		Test t pour égalité des moyennes					
						Signification		Différence	
		F	Sig.	t	df	p unilatéral	p bilatéral	moyenne	Erreur standard
Notes	Hypothèse de variances égales	,257	,614	-,702	74	,242	,485	-,5472	,7797
	Hypothèse de variances inégales			-,706	73,999	,241	,482	-,5472	,7751

Elle admet la structure suivante (voir page 67 pour plus de détails) :

	Test d'ég. des variances		Test d'égalité des moyennes					
					Signification		Différence	
	F	Sig.	t	df	unilatère	bilatère	moyenne	Erreur standard
Variances égales	...	P–valeur	$\dfrac{\overline{x}_n - \overline{y}_m}{\sqrt{s^2_{z,n+m}(1/n + 1/m)}}$	...	P–valeurs		$\overline{x}_n - \overline{y}_m$	$\sqrt{s^2_{z,n+m}(1/n + 1/m)}$
Variances différentes			$\dfrac{\overline{x}_n - \overline{y}_m}{\sqrt{s^2_{x,n}/n + s^2_{y,m}/m}}$	...	P–valeurs		$\overline{x}_n - \overline{y}_m$	$\sqrt{s^2_{x,n}/n + s^2_{y,m}/m}$

On lit, dans cet ordre :

1. La P-valeur du pré-test d'égalité des variances (pour savoir quelle formule utiliser ensuite) ; ici, 61.4%, nous conservons donc l'hypothèse de variances égales et exploiterons la première ligne du reste du tableau.
2. On passe au test d'égalité des moyennes proprement dit : on lit la P-valeur dans la colonne dédiée (et dans la ligne correspond au pré-test sur les variances) ; ici, si $H_1 : \mu_0 \neq m_0$ est bilatère, nous nous intéressons à la P–valeur bilatère 48.5%. Nous conservons $H_0 : \mu_0 = m_0$, et aucun effet n'est démontré :

   > Ces données ne permettent pas d'exclure que les deux professeurs notent, en moyenne, de manière identique.

3. Pour quantifier la taille d'effet éventuelle (pas ci-dessus mais voir par exemple page 75), un intervalle de confiance symétrique sur la différence des moyennes $\mu_0 - m_0$ peut être fourni dans la partie droite du tableau.

Chapitre 6. Tests du χ^2, d'ajustement et d'indépendance

Test du χ^2 d'ajustement à une loi. Il étend le test de comparaison à une proportion de référence aux cas où il y a plus de deux modalités. Une loi de référence $(p_{0,\text{ref}}, \ldots, p_{k,\text{ref}})$ est postulée, des données $x_1, \ldots, x_n \in \{0, 1, \ldots, k\}$ sont recueillies, et les hypothèses testées sont :

H_0 [conformité] : les données sont issues de la loi de référence $p_{0,\text{ref}}, \ldots, p_{k,\text{ref}}$

H_1 [non-conformité] : les données sont issues d'une autre loi.

Pour que le test soit appliqué validement, il faut que la taille d'échantillon n soit plus grande que 30 et que tous les effectifs attendus $np_{j,\text{ref}}$ soient plus grands que 5.

Le test confronte les effectifs observés à ces effectifs attendus, de manière globale, et procure une P–valeur, à lire dans la sortie logiciel. Un exemple est fourni ci-dessous, où l'on lit une P–valeur de 7.7% et où l'on converve donc l'hypothèse H_0 de conformité à la loi de référence postulée :

Chiffre unités

	Effectif observé	N théorique	Résidus
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 statistique

	Chiffre unités
Khi-carré	15,552[a]
df	9
Sig. asymptotique	,077

a. 0 cellules (0,0%) ont des effectifs théoriques inférieurs à 5. L'effectif théorique minimum d'une cellule est 11,6.

Plus généralement, lorsque la P–valeur est $> 5\%$, on ne peut pas exclure que les données suivent la loi de référence, et lorsque la P-valeur est $< 5\%$, on affirme que les données ne suivent pas la loi prescrite.

Test du χ^2 d'indépendance. Le test du χ^2 d'indépendance, appelé aussi test du χ^2 d'homogénéité, étend le test de comparaison de deux proportions indépendantes, qui correspondait à des données organisées dans un tableau 2×2, peuvent prendre chacune 2 valeurs ou plus, i.e., ce test traite de données organisées dans des tableaux $r \times s$, où $r \geqslant 2$ et $s \geqslant 2$.

Les données $(x_1, y_1), \ldots, (x_n, y_n)$ se présentent sous la forme de couples de variables catégorielles, prenant chacune $r \geqslant 2$ et $s \geqslant 2$ modalités possibles. Les hypothèses testées peuvent être formulées intuitivement ainsi :

- H_0 [homogénéité / indépendance] : les valeurs des x_t n'influent pas sur les valeurs des y_t
- H_1 [manque d'homogénéité / dépendance] : les valeurs des y_t sont déterminées au moins en partie par les valeurs des x_t

On calcule des effectifs attendus sous H_0 en reportant les fréquences empiriques globales des modalités prises par les y_j pour chaque modalité des x_t.

Pour que le test soit appliqué validement, il faut que la taille d'échantillon n soit plus grande que 30 et que tous les effectifs attendus ainsi calculés soient plus grands que 5.

Le test confronte les effectifs observés à ces effectifs attendus, de manière globale, et procure une P–valeur, à lire dans la sortie logiciel. Un exemple est fourni ci-dessous, où l'on lit une P–valeur de 4.2% et où l'on rejette donc l'hypothèse H_0 d'indépendance entre la note et le professeur :

Tableau croisé Prof * Note [2nd jeu de données]

			Note						
			A	B	C	D	E	F	Total
Prof	Grincheux v2	Effectif observé	12	11	17	25	21	9	95
		Effectif théorique	16,2	14,2	20,1	21,5	17,6	5,4	95,0
	Gentil v2	Effectif observé	21	18	24	19	15	2	99
		Effectif théorique	16,8	14,8	20,9	22,5	18,4	5,6	99,0
Total		Effectif observé	33	29	41	44	36	11	194
		Effectif théorique	33,0	29,0	41,0	44,0	36,0	11,0	194,0

Test du khi-carré

	Valeur	df	Signification asymptotique (bilatérale)
Khi-deux de Pearson	11,534[a]	5	,042
N d'observations valides	194		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 5,39.

Dit autrement (et à destination du grand public), ces données montrent que les deux enseignants n'utilisent pas les mêmes échelles de notations. En l'occurrence, après confrontation au sein de chaque cellule entre les effectifs observé et attendu, ces données montrent que l'enseignant de la première ligne utilise une échelle de notation plus sévère que l'enseignant de la seconde ligne.

Chapitre 7. Régression linéaire simple

Nous étudions l'explication statistique d'une variable dite variable à expliquer en fonction d'une variable explicative, selon une relation postulée $Y = \alpha_0 + \beta_0 x + \varepsilon$, où l'erreur ε suit une loi normale centrée de variance σ_0^2.

Il s'agit fondamentalement d'estimer et de tester les trois paramètres α_0, β_0 et σ_0. Dans la sortie logiciel simplifiée ci-dessous, seules les cases les plus nécessaires à notre niveau sont remplies :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1		,627		37,115

a. Prédicteurs : (Constante), AltitudeSommet

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression					
	Résidu					
	Total					

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), AltitudeSommet

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	-26,136				,010
	AltitudeSommet	,082				<,001

a. Variable dépendante : PrixForfait

Etape 0 : Identification du modèle. La variable à expliquer ("dependent variable" en anglais) est indiquée sous le tableau : c'est le prix du forfait semaine (en euros). La variable explicative ("independent variable" en anglais) est celle reportée dans le troisième tableau : c'est l'altitude au sommet de la station.

On se réfère au modèle calculé ici comme Prix Forfait / Altitude Sommet.

Etape 1 : Validité statistique du modèle. Il s'agit de vérifier si la variable explicative participe de manière significative à la modélisation affine : c'est le cas si et seulement si $\beta_0 \neq$. La P–valeur du test de $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$ est reportée dans le troisième tableau, colonne Sig., ligne correspondant à la variable explicative.

Ici, on lit < 0.001 pour cette P–valeur et H_0 est donc fermement rejetée. On a démontré que le coefficient de pente β_0 était différent de 0 ou, dit autrement, que le coefficient estimé $b = 0.082$ était significativement différent de 0. Le modèle Prix Forfait / Altitude Sommet est bien statistiquement valide.

Etape 1bis : Validité économique du modèle. Parfois, on demandera en amont de valider économiquement le modèle, i.e., de commenter si les co-variations exhibées pour les variables à expliquer et explicative sont logiques.

Ici, l'estimée $b = 0.082$ pour β_0 indique des co-variations de même sens : le prix du forfait augmente avec l'altitude au sommet. C'est économiquement logique et correspond à des domaines plus intéressants et où la neige est de meilleure qualité. Le modèle Prix Forfait / Altitude Sommet est bien économiquement valide.

Etape 2 : Qualité du modèle. L'indicateur est le r^2 (à lire dans le premier tableau), qui quantifie la proportion de variabilité totale expliquée par la régression linéaire. Plus le r^2 est grand, meilleur est le modèle.

Ici, 62.7% (des variations) des prix des forfaits sont expliqué(e)s par les (variations des) altitudes au sommet.

Etape 3 : Ecriture du modèle. On lit les coefficients estimés d'ordonnée à l'origine a et de pente b dans le troisième tableau, et l'écart-type estimé s des résidus dans le premier tableau. On procure une relation du type « $y = a + bx+$ résidu d'écart-type s » en remplaçant y et x par le nom des variables et incluant bien (i) des indications des unités et (ii) un terme de résidu. L'omission du terme de résidu est une erreur grave, car la relation semblerait purement déterministe et expliquer 100% du phénomène alors même que $r^2 < 100\%$.

Ici,

$$\begin{aligned}\text{Prix Forfait (en euros)} \quad &= \text{-26.136} \\ &\quad + 0.082 \times \text{Altitude Sommet (en m)} \\ &\quad + \text{Résidu (d'écart-type estimé 37.115)}\end{aligned}$$

Etape 4 : Interprétations. Il s'agit a minima de discuter la validité économique, si cela n'a pas déjà été effectué. Le coefficient de pente est en général aisé à interpréter, comme une variation moyenne approximative en fonction de la variable explicative. L'ordonnée à l'origine est souvent plus délicate à interpréter et il faut prendre garde à ne pas commettre d'extrapolation : il n'est pas toujours possible que la variable explicative vaille 0 (c'est le cas ici). Fondamentalement, cette ordonnée à l'origine est un coefficient permettant de recaler la relation.

Ici, pour le coefficient de pente 0.082, on peut dire qu'une augmentation de l'altitude au sommet de 100 m entraîne une augmentation du forfait, approximativement et en moyenne, de 8.2 euros. On ne peut pas interpréter le coefficient d'ordonnée à l'origine -26.136 en soi ; notez que ce coefficient est bien significativement différent de 0, cf. P–valeur dans le troisième tableau sur la ligne « (Constante) ».

Etape 5 : Intervalles de prévision. Une observation associée à une valeur x de la variable à expliquer se situe avec grande confiance dans un intervalle, dit de prévision, $a+bx \pm 2s$. Cet intervalle porte sur une valeur individuelle (une observation), et non sur une moyenne de population comme pour les intevalles de confiance. Un couple de données (x, y) qui est tel que y appartient bien à l'intervalle de prévision construit en x est dit compatible avec le modèle.

Ici, quels sont les prix attendus pour une station d'altitude 3300 m ? Il s'agit de

$$-26.136 + 0.082 \times 3300 \pm 2 \times 37.115 = 244.464 \pm 74.23\,, \qquad \text{arrondi à} \qquad 170 - -319 \quad \text{ou} \quad 245 \pm 75\,.$$

Alpe d'Huez a une altitude au sommet de 3300 m et le prix de son forfait est de 267 euros ; il est bien compatible avec le modèle établi.

Chapitre 8. Régression linéaire multiple

Ce chapitre étend le précédent : désormais, nous considérons simultanément plusieurs variables explicatives, comme dans la sortie logiciel simplifiée ci-dessous, où seules les cases les plus nécessaires à notre niveau sont remplies :

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1		,783	,780	28,418

a. Prédicteurs : (Constante), Pistes, AltitudeSommet, Dénivelé

ANOVA[a]

Modèle		Somme des carrés	ddl	Carré moyen	F	Sig.
1	Régression					<,001[b]
	Résidu					
	Total					

a. Variable dépendante : PrixForfait

b. Prédicteurs : (Constante), Pistes, AltitudeSommet, Dénivelé

Coefficients[a]

		Coefficients non standardisés		Coefficients standardisés		
Modèle		B	Erreur standard	Bêta	t	Sig.
1	(Constante)	16,433				,095
	AltitudeSommet	,039			5,627	<,001
	Dénivelé	,028			3,288	,001
	Pistes	,706			9,874	<,001

a. Variable dépendante : PrixForfait

Il s'agit d'étudier le modèle Prix Forfait / Altitude Sommet, Dénivelé, Pistes. Nous allons focaliser ce résumé de cours sur les différences par rapport à la régression linéaire simple. Les étapes avec les différences les plus notables seront surlignées en gris, les autres, non.

Etape 1 : Validité statistique du modèle. Elle se décompose en deux temps.

1. Validité statistique globale : au moins une variable contribue significativement à l'explication linéaire. Elle est assurée dès que la P–valeur lue dans le second tableau est $< 5\%$, ce qui est le cas ici.
2. Validités statistiques individuelles (marginales) : chaque variable apporte une contribution incrémentale significative au modèle linéaire, étant donné la présence des autres variables explicatives, i.e., le modèle n'est pas inutilement compliqué. C'est le cas si toutes les P–values lues dans les lignes des coefficients de pente du troisième tableau sont $< 5\%$, ce qui est le cas ici (peu importe que l'ordonnée à l'origine, elle, ne soit pas significativement différente de 0).

Etape 1bis : Validité économique du modèle. Il s'agit de justifier si les co-variations de la variable à expliquer et de chaque variable explicative sont logiques, i.e., si les signes des coefficients de pente sont logiques.

Etape 2 : Qualité du modèle. Le r^2 indique toujours la proportion de variabilité totale retrouvée par le modèle de régression ; ici,

> 78.3% (des variations) des prix des forfaits sont expliqué(e)s par les (variations des) altitudes au sommet, dénivelés et nombres de pistes.

La doxa statistique a introduit une (très / trop) légère correction du r^2 en fonction du nombre de variables explicatives, appelée version ajustée r^2_{adj}. Nous comparerons plutôt deux modèles de régression linéaire multiple tous deux valides statistiquement et économiquement en appréciant subjectivement les écarts de r^2 et de nombres (et d'identités) des variables explicatives.

Etape 3 : Ecriture du modèle. Il faut penser aux unités et au terme de résidu : ici,

$$\begin{aligned}\text{Prix Forfait (en euros)} \quad &= 16.433 \\ &\quad + 0.039 \times \text{Altitude Sommet (en m)} \\ &\quad + 0.028 \times \text{Dénivelé (en m)} \\ &\quad + 0.706 \times \text{Pistes (sans unités)} \\ &\qquad + \text{Résidu gaussien centré (d'écart-type estimé } 28.418)\end{aligned}$$

Etape 4 : Interprétation des coefficients de pente. Chaque coefficient de pente est interprété en prenant garde d'écrire « approximativement et en moyenne » et « toutes choses égales par ailleurs » (i.e., en fixant la valeur des autres variables explicatives). Par exemple :

> Lorsque l'altitude au sommet augmente de 100 m, le prix du forfait augmente, approximativement et en moyenne, et toutes choses égales par ailleurs, de 3.90 euros.

Il ne faut surtout pas hiérarchiser les variables en fonction des valeurs nominales de leurs coefficients de pente, car ces derniers ne sont pas intrinsèques et dépendent beaucoup des unités. La hiérarchisation se fait sur les P–valeurs ou les t–valeurs.

Etape 5 : Intervalles de prévision. Ils sont obtenus en assortissant la prévision ponctuelle donnée par le modèle d'une marge $\pm 2s$, où s est l'écart-type des résidus, exactement comme en régression linéaire simple.

Méthodes de sélection. Deux manières automatiques de sélectionner de bons sous-ensembles de variables sont :

- la sélection "backward", qui part d'un modèle complet (avec toutes les variables explicatives possibles) et le simplifie itérativement tant que nécessaire, en enlevant à chaque étape la moins utile des variables individuellement non significatives, s'il en reste ;
- la sélection "forward", qui part de la meilleure variable explicative prise isolément et enrichit itérativement le modèle tant que cela est possible, en ajoutant à chaque étape une variable étant individuellement significative dans le modèle étendu.