HEC
PARIS

The more you know, the more you dare®

HEC / Master in Management

Statistique : Fiches de synthese / Cours condense

Coordinateur du cours :

Gilles Fortin-Stoltz (stoltz@hec.fr)

Equipe enseignante :

Benjamin Petiau (benjamin.petiau@gmail.com)

Dinah Rosenberg (rosenberg@hec.fr)

Dan Xie (dan.xie@hec.edu)

Année universitaire 2022–2023

1

Synthèse — Modélisation

Partie commune. On considère des individus statistiques : leur ensemble est la population, et on en tire un échantillon aléatoire de taille n. Sur chaque individu de l'échantillon, on effectue une mesure de la variable d'intérêt, cela conduit aux données $x_1, \dots, x_n$, qu'on peut résumer par la moyenne d'échantillon $\overline{x}_n$.

Cas d'une proportion. La variable d'intérêt est binaire, $x_j \in \{0, 1\}$. Le paramètre d'intérêt est p_0, la proportion de la population pour qui la variable vaut 1.

Exemple : Un sondage politique sur $n = 1000$ habitants montre que $\overline{x}_{1000} = 53\%$ des sondés pensent qu'une crise est bien gérée. La variable d'intérêt est l'opinion positive, codée par 1, négative, de chaque individu. Le résultat $\overline{x}_{1000} = 53\%$ du sondage nous permettra d'estimer (avec des intervalles de confiance) la proportion p_0 de l'ensemble des habitants (qui peuvent être plusieurs millions) à penser que la crise est bien gérée. Cette proportion de population p_0 forme notre paramètre d'intérêt.

Cas d'une variable quantitative générale. On résume alors les données $x_1, \dots, x_n$, qui sont des nombres réels, par leur moyenne $\overline{x}_n$ et leur écart-type $s_{x,n}$. Le paramètre d'intérêt est μ_0, la moyenne de ce qu'on mesure sur l'ensemble de la population. Cette moyenne de population μ_0 forme notre paramètre d'intérêt.

Exemple : On veut connaître le temps moyen passé par les enfants de 6 ans devant des écrans les jours de semaines. On sélectionne au hasard 150 familles avec un enfant de 6 ans, qu'on observe chacun pendant une journée. Le temps moyen observé $\overline{x}_{150} = 103$ minutes sur l'échantillon estime le temps moyen μ_0 passé par jour devant des écrans par l'ensemble des enfants de 6 ans. On détermine aussi à partir des données un écart-type de $s_{x,150} = 37$ minutes.

2

Synthèse — Intervalles de confiance, résultats élémentaires

1. Forme des intervalles de confiance

Les intervalles de confiance expliquent comment généraliser la proportion ou la moyenne d'échantillon $\overline{x}_n$ à l'ensemble de la population.
Pour bien choisir la forme, il faut penser à la phrase de conclusion qu'on veut écrire : « Avec grande confiance, la proportion p_0 de (ou la moyenne μ_0 sur) l'ensemble de la population est (ou serait, ou sera)... »

Conclusion neutre. Dans ce cas, on dit juste que cette proportion p_0 ou cette moyenne μ_0 sont égales à $\overline{x}_n$ modulo une certaine marge d'erreur, ce qu'on note $\overline{x}_n \pm ...$ (et qu'on prononce : « plus ou moins »).

Estimation basse. On veut ici dire que p_0 ou μ_0 seront supérieures à une certaine valeur (elles vaudront au moins telle valeur). Ainsi, on corrige $\overline{x}_n$ à la baisse pour être sûr avec grande confiance que p_0 ou μ_0 soient supérieures à cette valeur basse.

Estimation haute. Même chose avec pour objectif cette fois que p_0 ou μ_0 soient inférieures à une certaine correction à la hausse de $\overline{x}_n$ (elles vaudront au plus telle valeur).

2. Formules dans le cas d'une proportion

Conclusion neutre. Avec grande confiance, la proportion de population p_0 vaut

$$\overline{x}_n \qquad \text{à une marge d'erreur de} \qquad \pm 1.96\frac{\sqrt{\overline{x}_n(1-\overline{x}_n)}}{\sqrt{n}} \qquad \text{près.}$$

Exemple : Avec $n = 1000$ et $\overline{x}_{1000} = 53\%$, on déduit qu'avec grande confiance, la part p_0 de l'ensemble des habitants trouvant que la crise est bien gérée vaut

$$53\% \qquad \text{avec une marge d'erreur de} \qquad \pm 1.96\frac{\sqrt{0.53(1-0.53)}}{\sqrt{1000}} = \pm 0.03093... = \pm 3.093...\% \approx \pm 3.1\%\,.$$

En clair, on propose l'intervalle $[53\% - 3.1\%,\ 53\% + 3.1\%] = [49.9\%,\ 56.1\%]$ comme intervalle de valeurs plausibles pour p_0.
Note : on arrondit toujours les marges d'erreur à la hausse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).

Estimation basse. Avec grande confiance, la proportion de population p_0 est plus grande que

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

Exemple : Avec $n = 1000$ et $\bar{x}_{1000} = 53\%$, on déduit qu'avec grande confiance, la part p_0 de l'ensemble des habitants trouvant que la crise est bien gérée est au moins de

$$53\% - 1.645\frac{\sqrt{0.53(1-0.53)}}{\sqrt{1000}} = 0.50403... = 50.403...\% \approx 50.4\%\,.$$

Note : on arrondit toujours les estimations basses à la baisse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).

Estimation haute. Avec grande confiance, la proportion de population p_0 est inférieure à

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

Exemple : Avec $n = 1000$ et $\bar{x}_{1000} = 53\%$, on déduit qu'avec grande confiance, la part p_0 de l'ensemble des habitants trouvant que la crise est bien gérée est au plus de

$$53\% + 1.645\frac{\sqrt{0.53(1-0.53)}}{\sqrt{1000}} = 0.55596... = 55.596...\% \approx 55.6\%\,.$$

Note : on arrondit toujours les estimations hautes à la hausse (ainsi, on ne fait qu'agrandir légèrement l'intervalle calculé).

3. Formules dans le cas d'une moyenne générale

Conclusion neutre. Avec grande confiance, la moyenne de population μ_0 vaut

$$\bar{x}_n \qquad \text{à une marge d'erreur de} \qquad \pm 1.96\,\frac{s_{x,n}}{\sqrt{n}} \quad \text{près.}$$

Exemple : Avec $\bar{x}_{150} = 103$ et $s_{x,150} = 37$, obtenus pour un échantillon de taille $n = 150$, on déduit qu'avec grande confiance, le temps moyen μ_0 passé devant les écrans par l'ensemble des enfants de 6 ans est de

$$103 \text{ min} \qquad \text{avec une marge d'erreur de} \qquad \pm 1.96\frac{37}{\sqrt{150}} = \pm 5.9212... \approx \pm 6 \text{ min}\,.$$

En clair, on propose l'intervalle $[103-6,\ 103+6] = [97,\ 109]$ comme intervalle de valeurs plausibles pour μ_0.

Note : on ne dit pas du tout que 95% des enfants de 6 ans regardent quotidiennement des écrans pendant un temps compris entre 97 et 109 minutes ; l'assertion d'intervalle de confiance ne porte que sur le temps moyen μ_0 (apprécié sur l'ensemble des enfants), qui n'est pas connu avec certitude mais dont on pense avec grande confiance qu'il se situe entre 97 et 109 minutes.

Estimation basse. Avec grande confiance, la moyenne de population μ_0 est plus grande que

$$\bar{x}_n - 1.645\,\frac{s_{x,n}}{\sqrt{n}}\,.$$

Exemple : Avec $\bar{x}_{150} = 103$ et $s_{x,150} = 37$, obtenus pour un échantillon de taille $n = 150$, on déduit qu'avec grande confiance, le temps moyen μ_0 passé devant les écrans par l'ensemble des enfants de 6 ans est au moins de

$$103 - 1.645\frac{37}{\sqrt{150}} = 98.0303 \approx 98 \text{ minutes}\,.$$

Estimation haute. Avec grande confiance, la moyenne de population μ_0 est inférieure à

$$\bar{x}_n + 1.645\,\frac{s_{x,n}}{\sqrt{n}}\,.$$

Exemple : Avec $\bar{x}_{150} = 103$ et $s_{x,150} = 37$, obtenus pour un échantillon de taille $n = 150$, on déduit qu'avec grande confiance, le temps moyen μ_0 passé devant les écrans par l'ensemble des enfants de 6 ans est au plus de

$$103 + 1.645\frac{37}{\sqrt{150}} = 107.9696... \approx 108 \text{ minutes}\,.$$

3

Synthèse — Intervalles de confiance, résultats plus avancés

1. Intervalles de confiance simultanés

Toute exploitation simultanée de deux intervalles de confiance au niveau 95% (ce qui est le cas de tous les intervalles donnés ci-dessus) vaut avec confiance seulement 90%.

Exemple : En reprenant les exemples du chapitre précédent, avec grande confiance (mais de niveau seulement 90%), on peut simultanément dire que la part p_0 de l'ensemble des habitants trouvant que la crise est bien gérée vaut $53\% \pm 3.1\%$ et que le temps moyen μ_0 passé devant les écrans par l'ensemble des enfants de 6 ans est de 103 ± 6 minutes.

2. Planification

Le problème de la planification consiste à se demander quelle taille d'échantillon constituer pour parvenir à une précision donnée. On procède en deux temps : on effectue un premier coup de sonde (sur $m \geqslant 30$ individus), et on calcule ensuite le nombre de données $n - m$ supplémentaires à récupérer. On applique les formules d'intervalles de confiance avec une marge d'erreur (les formules symétriques avec $\pm \ldots$), et uniquement celles-ci.

Plus précisément, avec les $m \geqslant 30$ données préliminaires, on constate qu'on a pour l'instant une précision (une marge d'erreur) de

$$\gamma = 1.96 \, \frac{s_{x,m}}{\sqrt{m}} \qquad \text{ou} \qquad \gamma = 1.96 \sqrt{\frac{\overline{x}_m(1-\overline{x}_m)}{m}} \, .$$

On se fixe un vœu de précision à $\pm\varepsilon$, c'est-à-dire qu'il faut réduire la précision actuelle (la marge d'erreur actuelle) d'un facteur $a = \gamma/\varepsilon$. Pour cela, il faudra un échantillon de taille au total de $n = m\,a^2$.

En clair, toute division de la marge d'erreur d'un facteur a requiert une multiplication de la taille d'échantillon par un facteur a^2.

Exemple : Pour l'estimation du temps moyen μ_0 passé chaque jour par l'ensemble des enfants de 6 ans devant des écrans, on a, avec les $m = 150$ enfants observés actuels, une marge d'erreur à ± 6 minutes. Supposons que l'on veuille plutôt une précision de l'estimation à ± 3 minutes, c'est-à-dire, une division de la marge d'erreur par un facteur $a = 2$. Cela requerrait au total $n = 150 \times 2^2 = 600$ observations d'enfants, soit $600 - 150 = 450$ de plus qu'actuellement.

Exemple : Pour l'estimation de la proportion de population p_0 de l'ensemble des habitants trouvant que la crise est bien gérée, on a, avec les $m = 1\,000$ sondés actuels, une marge d'erreur à $\pm 3.1\%$. Supposons que l'on veuille plutôt une précision de l'estimation à $\pm 1.5\%$, c'est-à-dire, une division de la marge d'erreur par un facteur $a = 3.1/1.5$. Cela requerrait au total

$$n = 1\,000\, a^2 = 1\,000\,(3.1/1.5)^2 = 4\,271.11... \approx 4\,272 \quad \text{sondés,}$$

soit $4\,272 - 1\,000 = 3\,272$ sondés de plus.

4

Synthèse — Tests d'hypothèses : généralités et comparaison à une valeur de référence

Dissymétrie des hypothèses en jeu. Les tests d'hypothèses permettent de répondre à une question. Mais la réponse n'est pas strictement binaire : soit on pourra exclure l'assertion de départ (on la tiendra pour fausse et on valider l'hypothèse alternative) ; soit on ne pourra pas l'exclure (mais sans la valider pour autant : on ne saura pas si cette assertion de départ est vraie ou fausse, simplement, avec les données disponibles, on ne peut pas prouver qu'elle soit fausse). On note donc une grande dissymétrie entre l'hypothèse de départ H_0, que l'on ne pourra jamais valider, et l'hypothèse alternative H_1, qu'on pourra valider. Le vocabulaire est éclairant : soit on rejette H_0, soit on conserve H_0 (mais peut-être faute d'avoir suffisamment de données pour la rejeter). Les tests ont fondamentalement tendance à conserver H_0 ; ils ne la rejettent que si elle est gravement contredite par les données.

1. Choix des hypothèses

Plusieurs critères peuvent être mis en jeu, tous vont dans le sens de l'introduction ci-dessus mais certains s'appliquent mieux que d'autres pour chaque cas concret :

- H_0 est l'hypothèse de laquelle il est raisonnable de partir ou à laquelle on est attaché (l'opinion commune, l'avis du manager, « comme on a toujours pensé », etc.), et H_1 est l'hypothèse contraire, qui susciterait l'étonnement ou la réflexion.
- Lorsque l'on veut établir un fait, H_1 est ce que l'on veut prouver, et on prend simplement pour H_0 le contraire de H_1.
- Quand une action risquée est envisagée et qu'évidemment, on ne veut l'entreprendre qu'à bon escient, on prend pour H_1 la situation correspondant à cette action risquée et pour H_0, son contraire, qui correspond à une vision prudente de la situation.

Attention à bien placer l'égalité dans H_0 ! Le cas d'égalité est toujours dans H_0, pour des raisons mathématiques profondes. En pratique, nos H_0 seront toujours d'une des formes suivantes :

- $H_0 : p_0 = p_{ref}$ ou $H_0 : \mu_0 = \mu_{ref}$
- $H_0 : p_0 \leqslant p_{ref}$ ou $H_0 : \mu_0 \leqslant \mu_{ref}$
- $H_0 : p_0 \geqslant p_{ref}$ ou $H_0 : \mu_0 \geqslant \mu_{ref}$

Les hypothèses H_1 seront elles toujours d'une des formes suivantes :

- $H_1 : p_0 \neq p_{ref}$ ou $H_1 : \mu_0 \neq \mu_{ref}$
- $H_1 : p_0 > p_{ref}$ ou $H_1 : \mu_0 > \mu_{ref}$
- $H_1 : p_0 < p_{ref}$ ou $H_1 : \mu_0 < \mu_{ref}$

Cas limite. Tester $H_0 : p_0 \leqslant p_{ref}$ contre $H_1 : p_0 > p_{ref}$ revient à tester $H_0 : p_0 = p_{ref}$ contre $H_1 : p_0 > p_{ref}$. Et de même pour les autres cas. C'est ce que l'on appellera passer au cas limite : se ramener à une égalité pour H_0.

Exemple de mise en œuvre du choix d'hypothèses

Partons d'une situation abstraite dans laquelle on veut effectuer un test sur une proportion (inconnue) p_0 de la population. Des raisons externes nous amènent à considérer une proportion de référence p_{ref}. Supposons qu'on teste $H_0 : p_0 = p_{ref}$ contre $H_1 : p_0 > p_{ref}$. On fixe ces hypothèses et on collecte ensuite (et ensuite seulement !) des données $x_1, \dots, x_n$ sur n sondés. Notons $\overline{x}_n$ la proportion d'échantillon observée. On se demande si la valeur $\overline{x}_n$, qui est proche de p_0, est compatible avec p_{ref} (auquel cas on conservera H_0) ou si elle est le signe que p_0 est plus grande que p_{ref} (auquel cas on rejettera H_0).

Exemple : On s'interroge sur la proportion p_0 de la population trouvant que la crise est bien gérée et on se demande si cette proportion est plus grande que la valeur de référence $p_{ref} = 50\%$. Si l'on est le parti au pouvoir, on choisit les hypothèses $H_0 : p_0 \leqslant 50\%$ contre $H_1 : p_0 > 50\%$ parce que l'on veut montrer que la majorité de la France est derrière nous (ce qui correspond à H_1). Les données de sondage sur $n = 1\,000$ sondés sont que $\overline{x}_{1000} = 53\%$ des sondés trouvent la crise bien gérée. On se demande si cette valeur d'échantillon de 53% nous permet d'affirmer qu'en effet, la majorité de la France est derrière nous, c'est-à-dire de rejeter l'hypothèse $H_0 : p_0 \leqslant 50\%$. On peut reformuler cela comme : on se demande si la valeur d'échantillon $\overline{x}_{1000} = 53\%$ est significativement supérieure à 50%, auquel cas on rejettera H_0, ou si au contraire, l'écart de 3 points peut n'être dû qu'à l'aléa d'échantillonnage (l'aléa ayant donné lieu au recrutement de cet échantillon-là et pas un autre). Dans ce dernier cas, on conservera H_0 et on dira que l'écart de 3 points n'est pas significatif. Il y aurait alors contradiction entre les données et l'hypothèse H_0, mais trop légère pour qu'on rejette H_0.

2. Test de comparaison à une proportion de référence : théorie

Statistique de test = écart renormalisé. On s'intéresse à ce que l'on appelle une statistique de test, qui normalise l'écart entre $\overline{x}_n$ et p_{ref} afin de lui donner un comportement plus universel : selon une loi normale standard sous l'hypothèse $H_0 : p_0 = p_{ref}$. Cette statistique de test a pour expression

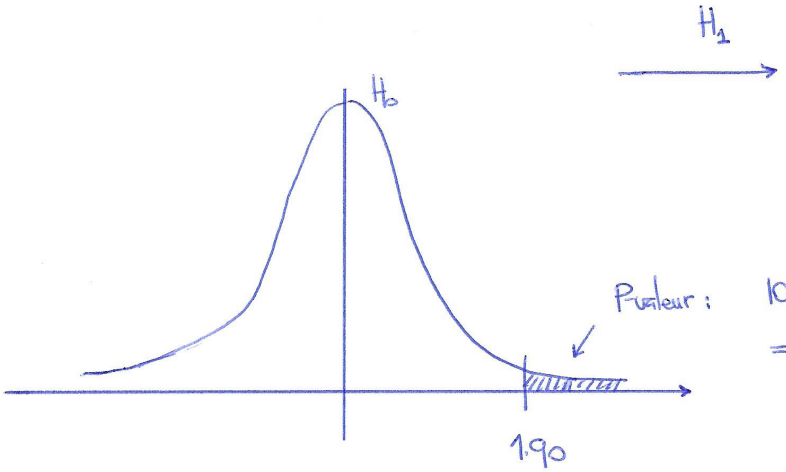
$$t_n = \sqrt{n}\,\frac{\overline{x}_n - p_{ref}}{\sqrt{p_{ref}(1-p_{ref})}}\,.$$

Son comportement sous H_0 est représenté sur les dessins par une courbe en cloche, qui représente les probabilités des valeurs prises par la loi normale standard. Cette statistique de test t_n prend donc sous H_0 des valeurs autour de 0, typiquement entre -2 et $+2$.
Sous H_1, elle suit un comportement différent (qui provient du fait que $\overline{x}_n$ est proche de p_0) :

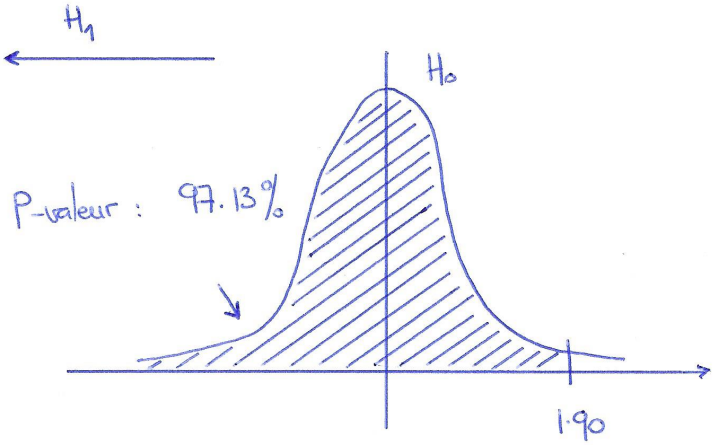
- Sous $H_1 : p_0 > p_{ref}$, la statistique de test prend des valeurs positives grandes. On place une flèche $\rightarrow$ sur le dessin permettant de calculer la P-valeur, de même sens que le signe $>$ dans l'hypothèse $p_0 > p_{ref}$.
- Sous $H_1 : p_0 < p_{ref}$, ce sont des valeurs négatives, représentées par une flèche $\leftarrow$ de même sens que le signe $<$.
- Sous $H_1 : p_0 \neq p_{ref}$, les deux sont possibles, on place deux flèches $\leftarrow$ et $\rightarrow$.

P-valeur. Pour calculer la P-valeur, on calcule d'abord la valeur numérique (sur les données) de la statistique de test. La P-valeur est alors la probabilité sous la loi normale standard de la zone

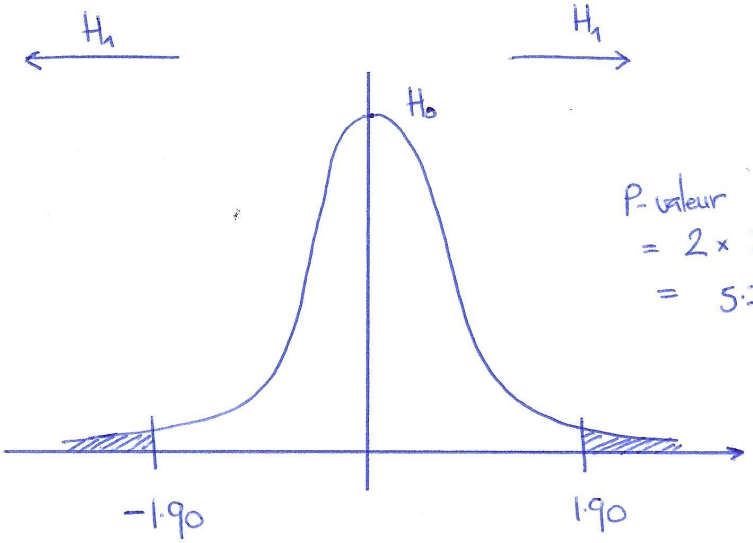
Cas du parti au pouvoir



Cas d'un parti d'opposition



Cas d'une observatrice neutre



correspondant à H_1 et déterminée par la valeur numérique de t_n. Voir les dessins (que nous commentons plus en détails ci-dessous).

Il faut voir la P-valeur comme un index de crédibilité de H_0 au vu des données et de H_1 : si la P-valeur est faible, c'est que la valeur de la statistique de test t_n, qui est rappelons-le, une mesure normalisée de l'écart de $\overline{x}_n$ à p_{ref}, est trop grande pour être une valeur probable sous H_0. C'est donc que l'écart entre $\overline{x}_n$ et p_{ref} est significatif. On rejettera donc H_0. Au contraire, une P-valeur plus grande indiquera un accord ou un désaccord léger mais pas rhédibitoire entre H_0 et les données : l'écart normalisé t_n n'est pas une valeur suffisamment improbable. On conservera donc H_0.

En pratique, et au moins dans le monde du business, on détermine le côté probable / improbable au seuil de 5%. Une P-valeur plus grande que 5% invite à la conservation de H_0, une P-valeur plus petite que 5% invite au rejet de H_0.

Ecriture des conclusions. Lorsque la P-valeur est plus grande que 5%, on conserve H_0 et la conclusion doit traduire qu'on conserve l'hypothèse de départ sans toutefois la valider. On écrira une phrase sur les modèles suivants (à adapter à chaque cas d'espèce en remplaçant H_0 ou H_1 par des mots) :

Ces données ne permettent pas d'exclure H_0.
Ces données ne permettent pas de mettre en évidence H_1.

Il s'agit donc de conclusions prudentes et vagues, souvent formulées avec des doubles négations, et peu exploitables côté business.

Lorsque la P-valeur est plus petite que 5%, on rejette H_0 et la conclusion doit traduire que nous nous sommes départis de l'hypothèse initiale, que nous l'avons invalidée avec une certitude raisonnable. On peut cette fois-ci être affirmatif :

Ces données excluent H_0.
Ces données mettent en évidence H_1.

Il s'agit donc de conclusions nettes et exploitables côté business.

3. Test de comparaison à une proportion de référence : exemple

Nous reprenons notre exemple fil rouge de satisfaction face à la gestion d'une crise et allons varier les commanditaires de l'étude. Ci-dessus, nous avions déjà discuté le cas du parti au pouvoir.

Cas du parti au pouvoir. Il veut montrer que la crise est bien gérée et choisit, avant d'effectuer le sondage, les hypothèses $H_0 : p_0 \leqslant 50\%$ contre $H_1 : p_0 > 50\%$. Son espoir est que les données permettent de rejeter H_0. Pour traiter ces hypothèses, on passe au cas limite et on se ramène à tester $H_0 : p_0 = 50\%$ contre $H_1 : p_0 > 50\%$. On rappelle qu'on interroge $n = 1000$ sondés. Le comportement de la statistique de test

$$t_{1000} = \sqrt{1000}\,\frac{\overline{x}_{1000} - 0.5}{\sqrt{0.5(1-0.5)}}$$

est représenté sur la figure du haut : loi normale standard sous H_0, grandes valeurs positives sous H_1. On calcule sa valeur numérique :

$$t_{1000} = \sqrt{1000}\,\frac{0.53 - 0.5}{\sqrt{0.5(1-0.5)}} \approx 1.90\,.$$

Dit autrement, l'écart « brut » de $\overline{x}_{1000} - p_{ref} = 53\% - 50\% = 3\%$ est normalisé en un écart valant 1.90. La P-valeur est ici la probabilité qu'une loi normale standard soit plus grande que 1.90 et qu'on

peut calculer comme valant 2.87% à l'aide de la table de la loi normale standard en dernière page du polycopié. Cette P-valeur est inférieure à 5% : on rejette H_0 et on a établi H_1, ce que l'on traduit en clair de la manière suivante :

> Ces données montrent que plus de 50% de la population trouve que la crise est bien gérée.

Cas d'un parti d'opposition. Il veut montrer que la crise est mal gérée et choisit, avant d'effectuer le sondage, les hypothèses $H_0 : p_0 \geqslant 50\%$ contre $H_1 : p_0 < 50\%$. Son espoir est que les données permettent de rejeter H_0 et de passer à H_1. Pour traiter ces hypothèses, on passe au cas limite et on se ramène à tester $H_0 : p_0 = 50\%$ contre $H_1 : p_0 < 50\%$. Alors seulement on effectue le sondage et on recueille les données. Zut, ces dernières montrent que sur l'échantillon, une proportion $\overline{x}_{1000} = 53\%$ des sondés trouvent la crise bien gérée. Les données d'échantillon ne sont pas en désaccord avec l'hypothèse $H_0 : p_0 \geqslant 50\%$ et on ne pourra donc pas la rejeter, il faudra la conserver. Le mieux est de dire cela et de le traduire en langage grand public par :

> Ces données ne permettent pas d'exclure que plus de 50% de la population trouve que la crise est bien gérée.

Si l'on veut toutefois faire les calculs et obtenir une P-valeur, c'est possible : la figure du milieu indique que l'on obtient alors une P-valeur de 97.13%. Cela confirme la conservation de H_0 (mais on ne valide pas pour autant H_0, même avec cette grande P-valeur : nous sommes liés par les hypothèses choisies avant le recueil des données).

Cas d'un observateur neutre, ouvert à tous les possibles. Un observateur neutre est ouvert à toutes les alternatives possibles : que le taux de satisfaction p_0 dans la population soit plus grand ou plus petit que 50%. Il n'en privilégie aucune (ce qui, d'ailleurs, lui évitera d'être frustré s'il s'est lié par avance avec des hypothèses unilatères et le regrette). Il considérera donc toujours l'hypothèse $H_0 : p_0 = 50\%$ mais cette fois-ci contre l'hypothèse bilatère $H_1 : p_0 \neq p_{\text{ref}}$. Dit autrement, on se demande si la proportion d'échantillon $\overline{x}_{1000}$ est significativement différente de 50% (alors que dans les cas précédents, on se demandait respectivement si elle était significativement supérieure ou inférieure à 50%). Les déviations sous H_1 ont lieu des deux côtés et il faut en tenir compte ; la figure du bas l'illustre. On obtient une P-valeur doublée par rapport au premier cas, de 5.74%. Elle est très légèrement supérieure à 5%, nous sommes vraiment dans une situation limite où nous pourrions dire que les données prouvent presque que le taux p_0 est différent de 50%, en l'occurrence (vu la valeur de $\overline{x}_{1000}$), qu'il est supérieur à 50%. Bien évidemment, si à l'examen vous dites simplement que la P-valeur est supérieure à 5% et que donc on conserve H_0, et que rien ne montre que le taux de satisfaction soit différent de 50%, vous aurez tous les points. Mais ici, grâce à la P-valeur, on voit que cette conservation se fait sur le fil. La P-valeur est une information plus riche sur l'accord entre H_0 et les données qu'une simple réponse binaire de conservation ou rejet de H_0.

4. Test de comparaison à une moyenne de référence

Les choses se déroulent de manière tout à fait similaire que pour le cas d'une comparaison à une proportion. On teste $H_0 : \mu_0 = \mu_{\text{ref}}$ contre une hypothèse H_1 de la forme indiquée page 7. La seule différence par rapport au cas des proportions est la statistique de test, dont la formule est

$$t_n = \sqrt{n}\,\frac{\overline{x}_n - \mu_{\text{ref}}}{s_{x,n}},$$

où $s_{x,n}$ désigne l'écart-type des données (sa valeur toujours indiquée dans l'énoncé).

6

Synthèse — Tests de comparaison de deux échantillons

On part de deux séries de données $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$, parfois de même longueur, parfois de longueurs différentes, et on veut déterminer si les 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 moyennes de population μ_1 et μ_2, ou les proportions de population p_1 et p_2 correspondantes sont différentes.

Types de données. Comme précédemment, les données peuvent porter sur des proportions ou des moyennes générales. Une seconde distinction est également à opérer, selon que les données sont

- appariées, c'est-à-dire obtenues en effectuant deux mesures x_j et y_j sur chaque individu statistique j ; dans ce cas, on a nécessairement $n = m$;
- indépendantes, c'est-à-dire que les individus choisis pour obtenir les x_j n'ont rien à voir avec les individus choisis pour obtenir les y_j ; dans ce cas, les deux séries de données peuvent en particulier être de longueurs n et m différentes.

Dans un tableur, les données se présenteraient ainsi :

Données appariées

ID	Variable 1	Variable 2
1	16	12
2	18	16
3	30	29
4	29	32
5	18	17
6	22	25
7	46	42
8	32	38
9	33	34
...	...	...

Ex : ID = Produits
Variable 1 = prix dans magasin 1
Variable 2 = prix dans magasin 2

→ On compare des colonnes

vs.

Données indépendantes

ID	Group	Variable
1	0	12
2	1	16
3	0	29
4	0	32
5	1	17
6	1	25
7	0	42
8	0	38
9	1	34
...	...	...

Ex : ID = Clients
Group = Avec ou sans promotion
Variable = Montant d'achat

→ On compare des lignes

Nous allons traiter trois cas seulement (car le quatrième cas, avec les proportions appariées, est trop compliqué à notre niveau) :

— les proportions indépendantes,
— les moyennes indépendantes,
— les moyennes appariées.

Dans tous ces cas, la statistique de test est donnée par la différence $\overline{x}_n - \overline{y}_m$ entre les deux moyennes d'échantillon, correctement renormalisée. Plus précisément, elle sera de la forme

$$t_{n,m} = \frac{\overline{x}_n - \overline{y}_m}{\text{normalisation}},$$

où [normalisation] est une quantité telle que la statistique de test $t_{n,m}$ suive (approximativement) une loi normale ou une loi proche sous H_0. Nous allons bien entendu vous dire dans chaque cas quelle est sa formule.

1. Proportions indépendantes

Rappel. Dans le cas d'une série de données et d'une comparaison à une proportion de référence, nous utilisions la statistique de test

$$t_n = \sqrt{n}\,\frac{\overline{x}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1-p_{\text{ref}})}} = \frac{\overline{x}_n - p_{\text{ref}}}{\sqrt{p_{\text{ref}}(1-p_{\text{ref}})/n}}.$$

Extension au test de deux proportions indépendantes. Ici, nous allons remplacer le numérateur par $\overline{x}_n - \overline{y}_m$ et adapter le dénominateur pour tenir compte du fait que nous n'avons pas de proportion de référence p_{ref} et que nous avons deux tailles d'échantillon n et m : la statistique de test sera donnée par

$$t_{n,m} = \frac{\overline{x}_n - \overline{y}_m}{\sqrt{\widehat{p}(1-\widehat{p})\,(1/n + 1/m)}},$$

où $\widehat{p}$ est une proportion que nous définissons maintenant. Nous allons considérer l'hypothèse H_0 : $p_1 = p_2$. Sous cette hypothèse, la meilleure manière d'estimer la proportion de population commune $p_1 = p_2$, c'est de mettre toutes les données dans le même sac :

$$\widehat{p} = \frac{x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_m}{n+m}.$$

Le principe du test est conclu par l'indication que sous H_0 : $p_1 = p_2$, la statistique de test suit approximativement une loi normale, tandis que sous H_1 : $p_1 > p_2$, elle prend de grandes valeurs positives, sous H_1 : $p_1 < p_2$, de grandes valeurs négatives, et sous H_1 : $p_1 \neq p_2$, de grandes valeurs négatives ou positives.

Exemple : Nous reprenons notre exemple des sondages politiques sur la gestion de crise. La perception de la population a-t-elle changé entre les deux crises ? Un premier sondage, effectué lors de la crise de mars-avril, indique que sur $n = 1000$ habitants sondés, une part de $\overline{x}_{1000} = 530/1000 = 53\%$ pense que cette première crise est bien gérée. (Ce sont les données que nous considérions jusqu'à présent.) Un second sondage, effectué 7 mois plus tard lors de la crise d'octobre-novembre, indique que sur $m = 1002$ habitants sondés, une part de $\overline{y}_{1002} = 492/1002 \approx 49.1\%$ pense que cette deuxième crise est bien gérée. Notez bien que les deux échantillons utilisés pour les sondages n'ont rien à voir entre eux et ont été choisis indépendamment au sein de la population ; c'est cela qui justifie le caractère de données indépendantes. La question initialement posée étant neutre (« La perception de

la population a-t-elle changé ? ») et n'indiquant pas d'a priori, nous allons recourir à une hypothèse H_1 bilatère et tester $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$, où p_1 et p_2 sont respectivement les parts d'habitants satisfaits de la gestion de la première et de la deuxième crise. La proportion commune sous H_0 admet pour estimation numérique

$$\widehat{p} = \frac{530 + 492}{1000 + 1002} = 51.05\%,$$

de sorte que la statistique de test prend la valeur

$$t_{1000,1002} = \frac{530/1000 - 492/1002}{0.5105\,(1 - 0.5105)\,(1/1000 + 1/1002)} \approx 1.74.$$

Un schéma (que nous vous laissons dessiner) montre que la P-valeur vaut alors

$$P = 2 \times (100\% - 95.91\%) \approx 8.2\% > 5\%.$$

On conserve $H_0 : p_1 = p_2$. On traduit cela en une phrase « grand public » ; voici deux exemples de telles phrases :

> Ces données ne permettent pas d'exclure que les taux de satisfaction soient identiques lors de la gestion des deux crises.
>
> Les taux de de satisfaction des actions de gestion de crise ne sont pas significativement différents entre les deux crises.

2. Moyennes indépendantes

On discute ici de deux échantillons de données obtenus indépendamment mais visant à comparer non plus des proportions mais des moyennes. Ce qui suit explique comment tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.

Principe (plus compliqué). Le test repose toujours sur une statistique de test de la forme

$$t_{n,m} = \frac{\overline{x}_n - \overline{y}_m}{\text{normalisation}},$$

mais la normalisation est beaucoup plus compliquée à énoncer ; sa formule dépend du fait que les écarts-types (variances) des deux échantillons sont égaux ou différents. Dans le premier cas, on effectuera une estimation conjointe de la valeur commune ; dans le second cas, on estimera chaque écart-type (chaque variance) séparément. Dans les deux cas, les formules sont longues, compliquées et impossibles à mettre en œuvre à la main : nous avons besoin de l'assistance d'un logiciel statistique ! Qui fournit des sorties similaires à celles reproduites à la page suivante. Les tableaux du haut sont les tableaux originels et les tableaux du bas expliquent quelles cases sont utiles.

Notez en particulier que nous avons changé le texte de nombreuses cases pour mieux expliquer (en français !) ce qui y est reporté. En l'occurrence, on lit, dans cet ordre :

1. La P-valeur du pré-test d'égalité des variances (pour savoir quelle formule utiliser ensuite) ;
2. Si cette P-valeur est supérieure à 5%, les variances ne sont pas significativement différentes ; si elle est inférieure à 5%, les variances sont significativement différentes ; pour la suite, on utilise la bonne ligne (parmi les deux lignes) du grand tableau ;
3. On passe au test d'égalité des moyennes proprement dit : on lit la P-valeur dans la colonne dédiée ;

Version complète en anglais

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Monthly net salary	Men	179	3431,46	3895,437	291,159
	Women	147	2434,90	1282,947	105,816

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Monthly net salary	Equal variances assumed	8,195	,004	2,972	324	,003	996,560	335,371	336,781	1656,339
	Equal variances not assumed			3,217	223,375	,001	996,560	309,791	386,074	1607,047

Version simplifiée en français

	Catégorie	Tailles d'échant.	Moyennes d'échantillon		
	1ère	179	3431,46		
	2nde	147	2434,90		

	Pré-test d'égalité des variances (H0 : égalité / H1 : var. différentes)		Test d'égalité des moyennes (H0 : égalité / H1 : moyennes différentes)						
								Intervalle de confiance à 95% sur la différence des moyennes	
		P-valeur	t		P-valeur (bilatère)			Borne basse	Borne haute
Variances pas significativement différentes (H0 conservée)		,004 (soit 0.4%)	Statistique de test		,003 soit 0.3%				
Variances significativement différentes (H0 rejetée)					,001 soit 0.1%				

4. Si besoin, un intervalle de confiance symétrique sur la différence des moyennes est fourni dans la partie droite du tableau.

Exemple : Les tableaux reproduits traitent des salaires moyens pour les hommes et les femmes dans la CSP 37 (cadres supérieurs) et en région parisienne. On se fonde sur deux échantillons indépendants, composés respectivement de 179 hommes et de 147 femmes. Les salaires moyens lus dans l'échantillon sont de $\bar{x}_{179} = 3431.46$ euros pour les hommes et $\bar{y}_{147} = 2434.90$ euros pour les femmes. Pour voir s'ils sont significativement différents, on effectue donc un pré-test d'égalité des variances (P-valeur de 0.4%, on rejette l'hypothèse d'égalité des variances et on lit la seconde ligne du tableau), puis le test d'égalité des moyennes proprement dit. Sa P-valeur est de 0.1%, on rejette donc l'hypothèse H_0 d'égalité des moyennes, c'est-à-dire que l'on a montré que les salaires moyens des hommes et des femmes sont significativement différent (on a mis en évidence un effet). En l'occurrence, lorsque l'on regarde les salaires moyens d'échantillon, on a prouvé que les hommes gagnaient davantage en moyenne que les femmes dans cette catégorie (c'est ce que l'on appelle une direction d'effet). On pourrait même ajouter une phrase mettant en jeu l'intervalle de confiance [386.074 – 1 607.047], en arrondissant ce dernier en l'intervalle [350 – 1 650] pour plus de lisibilité ; cette quantification est ce qu'on appelle une mesure de la taille d'effet. Résumons notre conclusion :

> Ces données montrent que les salaires moyens des hommes et des femmes de cette catégorie sont différents [*effet*], en l'occurrence, les hommes gagnent davantage en moyenne [*direction d'effet*] : avec grande confiance, l'écart moyen de salaires mensuels se situe quelque part entre 350 et 1 650 euros [*taille d'effet*].

3. Moyennes appariées

Deux échantillons sont appariés lorsqu'ils sont obtenus à partir des mêmes individus statistiques, sur lesquels deux mesures ont été effectuées. Ces deux mesures sont très liées. On pourrait par exemple :

— Comparer les prix de produits (= individus statistiques) dans deux magasins différents ; les deux prix relevés sont très liés, car il y a des ordres de grandeur raisonnables pour le prix d'un pack de yaourts ou d'une bouteille de vodka.

— Comparer les salaires de l'enfant aîné et du suivant : les individus statistiques sont les fratries, la première mesure est le salaire de l'aîné, la seconde mesure, celui du cadet. Les deux mesures sont liées car l'aîné et le cadet sont extraits du même milieu social et ont donc des trajectoires sociales comparables.

— De même pour la comparaison des salaires femmes–hommes au sein des couples hétérosexuels (= individus statistiques).

Puisque les mesures sont très liées, le mieux pour les comparer est de s'intéresser à leurs différences. Ainsi, on peut mettre en évidence l'effet des conditions (magasin 1 vs. magasin 2, aîné vs. cadet) par rapport à l'ordre de grandeur commun. On rajoute donc une colonne dans le tableau de principe des données appariées, qui était en première page de ce chapitre (voir figure en haut de la page suivante). Et on va tester si la moyenne des différences est significativement différente de 0 ou non. Comme on n'a plus qu'une seule série de données, à savoir les $z_j = x_j - y_j$ pour $j = 1, \ldots, n$, il suffit d'appliquer un test de comparaison à une valeur de référence, en l'occurrence, $\mu_{\text{ref}} = 0$. On se ramène au chapitre précédent ! En particulier, la statistique de test considérée est

$$t_n = \sqrt{n}\,\frac{\bar{z}_n - 0}{s_{z,n}} = \frac{\bar{x}_n - \bar{y}_n}{s_{x-y,n}/\sqrt{n}}\,;$$

la normalisation est ici donnée par $s_{x-y,n}/\sqrt{n}$, où $s_{x-y,n}$ est l'écart-type $s_{x-y,n}$ de la suite des différences $z_j = x_j - y_j$. (Sa valeur sera donnée par l'énoncé.)

ID	Variable 1	Variable 2	Différence
1	16	12	4
2	18	16	2
3	30	29	1
4	29	32	-3
5	18	17	1
6	22	25	-3
7	46	42	4
8	32	38	-6
9	33	34	-1
...	...	...	

Exemple : Reprenons notre example sur les temps moyens d'écrans. Les individus statistiques sont donnés par les 150 familles considérées. On effectue des mesures initiales $x_1, \dots, x_{150}$ de temps d'écrans ; on obtient une moyenne de $\overline{x}_{150} = 103$ minutes (et un écart-type de $s_{x,150} = 37$ minutes). Les parents effectuent un stage de sensibilisation aux côtés néfastes des écrans, et on effectue une seconde campagne de mesures de temps d'écrans, donnant lieu aux observations $y_1, \dots, y_{150}$, de moyenne $\overline{y}_{150} = 99$ minutes (et d'écart-type $s_{y,150} = 34$ minutes). On calcule les écarts $z_j = x_j - y_j$: leur moyenne est $\overline{z}_{150} = 4$ minutes et leur écart-type vaut $s_{z,150} = 15$ minutes. Le stage a-t-il été efficace ?

Ici, on considère un test unilatère H_0 : moyennes pas significativement différentes contre H_1 : la moyenne a significativement baissé après le stage (personne n'ira penser que le stage de sensibilisation puisse encourager les temps d'écrans !). C'est-à-dire que l'on va tester $H_0 : \mu_0 = 0$ contre $H_1 : \mu_0 > 0$, où μ_0 est le temps moyen (sur l'ensemble des familles de France) de la différence du temps d'écrans qu'on observerait grâce au stage (dans le sens « avant moins après »). La statistique de test suit une loi normale sous H_0, elle prend des valeurs plus grandes sous H_1, et sa valeur numérique est

$$t_{150} = \sqrt{150}\,\frac{103 - 99}{15} \approx 3.26\,.$$

La P-valeur est très faible, on rejette fermement H_0, et on conclut :

> Ces données montrent que le stage de sensibilisation réduit bien le temp moyen passé devant les écrans.

7

Synthèse — Tests du χ^2

[Prononcer « ki-deux » ou « ki-carré » et écrire avec la lettre grecque χ]

Hypothèses testées. Nous voyons deux tests du χ^2 : celui d'ajustement à une loi de référence (cette loi étant précisée dans l'énoncé) et celui d'indépendance, aussi appelé test d'homogénéité. Les hypothèses qu'ils testent respectivement sont les suivantes.

Test d'ajustement—On dispose d'une seule série de données. On teste H_0 [conformité :] les données recueillies sont réparties selon la loi de référence contre H_1 [non-conformité :] les données recueillies sont réparties selon une autre loi.

Test d'indépendance/homogénéité—On dispose de couples de données, typiquement, l'indication de la catégorie à laquelle appartient l'individu statistique, d'une part, et d'autre part, une variable d'intérêt. On teste H_0 [indépendance/homogénéité :] la valeur de la variable d'intérêt ne dépend pas de la catégorie, i.e., les lois de répartition de la variable d'intérêt sont homogènes entre les catégories contre H_1 [dépendance/absence d'homogénéité :] la catégorie influe sur la valeur de la variable d'intérêt, i.e., les lois de répartition de la variable d'intérêt ne sont pas toutes égales selon les catégories.

Nous vous donnons ci-dessous des traductions concrètes de ces phrases générales sur des exemples.

Conditions d'application. Les tests du χ^2 requièrent que la taille totale d'échantillon n soit plus grande que 30 et que tous les effectifs attendus (calculés par le logiciel statistique) soient plus grands que 5. Une note de bas de tableau vous indique si la seconde condition est vérifiée, mais n'oubliez pas d'indiquer également la première.

[Exemples : voir pages suivantes]

Exemple : χ^2 d'ajustement. Considérons la sortie statistique suivante.

Waiting time

	Observed N	Expected N	Residual
<= 2 min	58	52,0	6,0
> 2 min & <= 5 min	44	41,6	2,4
> 5 min	2	10,4	-8,4
Total	104		

Test Statistics

	Waiting time
Chi-Square	7,615[a]
df	2
Asymp. Sig.	,022

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,4.

On dispose d'une seule série de données : 104 temps d'attente téléphonique. La loi de référence (qui était donnée dans l'énoncé correspondant) est que 50% de ceux-ci sont compris entre 0 et 2 minutes, 40% entre 2 et 5 minutes, et 10% au-delà de 5 minutes. D'où les effectifs attendus : $52 = 104 \times 50\%$, puis $41.6 = 104 \times 40\%$ et enfin, $10.4 = 104 \times 10\%$.

On teste donc H_0 : les temps d'attente observés se répartissent selon cette loi de référence contre H_1 : ils suivent une autre loi. On confronte les effectifs attendus aux effectifs observés, valant respectivement 58, 44 et 2. On le fait validement car la taille totale d'échantillon $n = 104$ est plus grande que 30 et que tous les effectifs attendus sont plus grands que 5. (Le logiciel le signale d'ailleurs en note de bas de tableau.) On fait cela en calculant un écart, selon une formule compliquée appelée la divergence du χ^2 et valant ici 7.615. La loi limite est une loi du χ^2 à 2 degrés de liberté ("df" veut dire "degrees of freedom"). On en déduit la P-valeur $0.022 = 2.2\%$. Elle est plus petite que 5%, on rejette H_0. Les données ne suivent pas la loi de référence. Attention ! on compare alors les effectifs observés et attendus pour comprendre où se situent les plus grandes déviations. On se rend compte qu'il y a plus de temps d'attente courts qu'attendus, et moins de temps longs qu'attendus. On en conclut :

> Les temps d'attente observés ne suivent pas la loi de référence prescrite, en l'occurrence, les consommateurs attendent moins longtemps que loi de référence.

A l'examen—Nous vous demanderons simplement de
- formuler H_0 et H_1,
- constater que les conditions d'application sont vérifiées,
- lire la P-valeur
- et d'écrire une conclusion non technique et compréhensible par le grand public (et indiquant par ailleurs où les écarts significatifs à la loi de référence ont lieu lorsqu'il y en a).

Exemple : χ^2 d'indépendance (d'homogénéité). Considérons la sortie statistique suivante.

Group * Number of glasses Crosstabulation

			Number of glasses			
			3 glasses or less	3 - 6 glasses	6 glasses or more	Total
Group	8am	Count	8	10	5	23
		Expected Count	8,8	7,5	6,6	23,0
	10am	Count	12	7	10	29
		Expected Count	11,2	9,5	8,4	29,0
Total		Count	20	17	15	52
		Expected Count	20,0	17,0	15,0	52,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,335[a]	2	,311
Likelihood Ratio	2,343	2	,310
Linear-by-Linear Association	,072	1	,789
N of Valid Cases	52		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 6,63.

On dispose de couples de données : 52 couples avec le nombre de verres bus la veille et le groupe de cours (8h ou 10h). Les données portent sur 23 étudiants du groupe de 8h et 29 étudiants de celui de 10h. On compare les profils de consommation d'alcool (= les lois de répartition du nombre de verres bus) par groupe. On se demande si H_0 [indépendance/homogénéité] : les profils de consommation sont identiques entre les deux groupes (i.e., la consommation est indépendante du groupe) contre H_1 [dépendance/absence d'homogénéité] : les profils de consommation sont différents dans les deux groupes (i.e., la consommation dépend du groupe).

Des effectifs attendus sont calculés sous l'hypothèse d'homogénéité, et confrontés aux effectifs observés. Le test est validement appliqué car la taille totale d'échantillon vaut $n = 52 \geqslant 30$ et tous les effectifs attendus sont plus grands que 5. On lit la P-valeur (et la divergence du χ^2, qui est une mesure de l'écart) dans la ligne "Pearson Chi-Square" : $0.311 = 31.1\%$. Elle est plus grande que 5%, on conserve H_0 et on conclut :

> Ces données ne mettent en évidence aucune différence significative entre les profils de consommation d'alcool des deux groupes. / Ces données ne permettent pas d'exclure que la consommation d'alcool soit la même dans les deux groupes.

8

Synthèse — Régression linéaire simple

Dans ce chapitre, nous étudions l'explication statistique d'une variable dite variable à expliquer en fonction d'une variable explicative, selon une relation du type $y = a + bx +$ erreur.

Exemple : Prix du forfait de ski en fonction de la taille du domaine. L'exemple suivi tout au long de ce chapitre est reproduit à la page suivante : la variable explicative est la taille du domaine skiable en km, et la variable à expliquer le prix du forfait en euros. On verra que la relation proposée ici est statistiquement significative, de très bonne qualité, et s'écrit :

Prix forfait (euros) = 92.366 + 0.434 × Taille domaine (km) + résidu (d'écart-type 20.594)

Vue d'ensemble. Il vous suffira, à l'examen, de savoir réaliser l'analyse de régression en 5 points :

1. Validité statistique du modèle
2. Qualité du modèle
3. Ecriture de la relation proposée
4. Validation économique / Interprétation de la relation proposée
5. Prévision d'une nouvelle valeur / Vérification de compatibilité d'une observation avec le modèle

Etape 1 : Validité statistique. Il s'agit de voir si le coefficient de pente 0.434 est H_0 : Pas significativement différent de 0 contre H_1 : Significativement différent de 0. La P-valeur se lit dans la colonne "Sig." de la ligne du coefficient et est bien $< 5\%$ (elle est même quasi-nulle). On rejette l'hypothèse de nullité du coefficient et on a bien prouvé que ce coefficient 0.434 était significativement différent de 0.

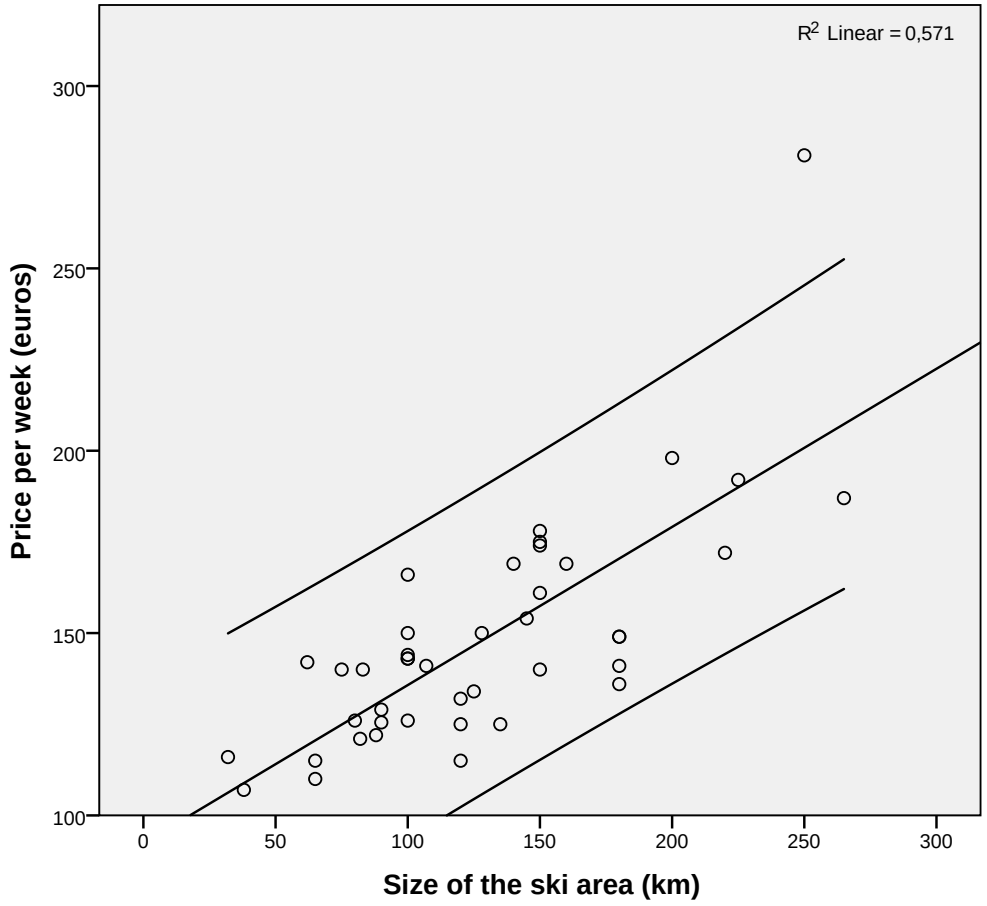
Si cela n'avait pas été le cas, on aurait déclaré le modèle non statistiquement valide et on aurait arrêté là l'étude (le reste des étapes est sans objet).

Etape 2 : Qualité du modèle. On lit l'indicateur $r^2 = 0.571 = 57.1\%$ dans le premier tableau. La phrase exprimant le sens de cette valeur est :

> 57.1% (des variations) du prix de forfait de ski sont expliqués par (les variations de) la taille du domaine skiable.

Un r^2 supérieur à 20% est bon en régression linéaire simple, surtout s'agissant de phénomènes économiques.

Attention ! La validité statistique ne s'apprécie pas au vu de la valeur nominale du r^2 ; des r^2 bas de l'ordre de 2% ou 3% peuvent correspondre à des modèles statistiquement valides, lorsque la



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,756[a]	,571	,560	20,594

a. Predictors: (Constant), Size of the ski area (km)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22596,541	1	22596,541	53,280	,000[b]
	Residual	16964,418	40	424,110		
	Total	39560,958	41			

a. Dependent Variable: Price per week (euros)

b. Predictors: (Constant), Size of the ski area (km)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	92,366	8,247		11,200	,000
	Size of the ski area (km)	,434	,059	,756	7,299	,000

a. Dependent Variable: Price per week (euros)

taille d'échantillon n est grande. Pour savoir si un modèle est ou non statistiquement valide, on lit la P-valeur indiquée plus haut.

Etape 3 : Ecriture de la relation. Ici, c'est :

$$\text{Prix forfait (euros)} = 92.366 + 0.434 \times \text{Taille domaine (km)} + \text{résidu (d'écart-type 20.594)}$$

Il ne faut pas oublier les unités ni surtout, le terme de résidu (car la régression n'explique pas tout le phénomène, juste 57.1% du phénomène ici ; ce terme de résidu correspond aux 42.9% restants).

Etape 4 : Validation économique / interprétation. La validation économique consiste à se demander si la relation semble logique ; plus précisément, on se demander si le signe du coefficient de pente est logique. Ici, il est positif, ce qui veut dire que plus le domaine est de grande taille, plus le forfait hebdomadaire est cher, ce qui est parfaitement logique.

L'interprétation des coefficients consiste à commenter leur valeur numérique. Le coefficient 0.434 indique que lorsque la taille du domaine skiable augmente d'1 km, le prix du forfait augmente de 0.434 euros (approximativement et en moyenne). Pour le coefficient 92.366, on ne peut pas l'interpréter en soi, car la taille du domaine ne peut être nulle ; quand on regarde les données, on lit une taille minimale de domaine autour de 25 km. On refuse alors d'interpréter 92.366 et on interprète plutôt $92.366 + 0.434 \times 25 = 103.216$ comme la valeur minimale du forfait de ski (approximative et en moyenne).

Etape 5 : Prévision / Compatibilité. Cette dernière étape s'attache au terme de résidu : c'est une erreur, un écart au modèle déterministe formé par la droite de régression, mais dont on connaît l'ordre de grandeur de la taille. En effet, un résidu d'écart-type s prend typiquement (disons 95% du temps) des valeurs individuelles entre -2 et $+2s$.

Cette dernière étape procure donc des fourchettes de valeurs plausibles pour des valeurs individuelles de forfaits de ski (on parle d'intervalle de prévision : car il s'agit de valeurs individuelles, par opposition aux intervalles de confiance qui portent sur des moyennes).

Par exemple, quelle est la fourchette de tarifs attendue pour un domaine skiable de taille 100 km ? C'est

$$92.366 + 0.434 \times 100 \pm 2 \times 20.594 = 135.766 \pm 41.188 \approx 136 \pm 42 \text{ euros.}$$

Le test d'une compatibilité consiste à se demander si une observation y se situe bien dans l'intervalle de prévision $92.366 + 0.434x \pm 41.188$ associé à la valeur x de la variable explicative. Si oui, c'est parfait, l'observation est déclarée compatible avec le modèle et sinon, on parle de valeur atypique. Sur le nuage de points où nous avons représenté la droite de régression $y = 92.366 + 0.434x$, nous avons également également représenté les droites $y = 92.366 + 0.434x - 41.188$ et $y = 92.366 + 0.434x + 41.188$. Nous voyons une observation située hors de son propre intervalle de prévision (tandis que toutes les autres observations sont compatibles avec le modèle).

9

Synthèse — Régression linéaire multiple

Nous prenons comme exemple filé tout au long de ce chapitre la modélisation de prix de forfaits de ski (variable à expliquer) en fonction de diverses caractéristiques du domaine (variables explicatives) : altitude maximal, nombre de remontées, nombre de pistes.

1. Etude d'un modèle donné

Nous commençons par l'étude d'un modèle donné—voir les deux sorties de la page suivante—et suivons la même analyse en 5 étapes, avec quelques changements. Ce qui suit va surtout souligner ces changements et passera plus rapidement en revue ce qui ne change pas.

Etape 1 : Validité statistique. Elle se décompose ici en deux sous-points : (1.a) validité globale et (1.b) validités marginales.

La validité globale consiste à s'assurer que (H_1 :) au moins une variable explicative est utile pour la modélisation linéaire en rejetant (H_0 :) aucune variable n'est utile. La P-valeur pour ce test se lit dans le deuxième tableau.

Les tests de validité marginale sont menés dans le troisième tableau : pour chaque variable explicative on teste si (H_0 :) elle ne contribue pas significativement à la modélisation linéaire étant donné les autres variables déjà présentes dans le modèle contre (H_1 :) elle y contribue significativement. On teste ici l'apport incrémental d'une variable. Si au moins une variable ne contribue pas significativement au modèle linéaire, ce dernier est plus compliqué que nécessaire et pourrait être simplifié.

Si le modèle n'est pas globalement valide (rare) ou si des problèmes de validités individuelles se posent pour une ou plusieurs variables (plus fréquent), on déclare le modèle statistiquement non valide et on s'arrête là, on ne continue pas l'étude.

Exemple. La première régression linéaire multiple (avec les trois variables) n'est pas statistiquement valide : si elle est bien globalement valide, un problème de validité marginale se pose pour la variable Nombre de remontées (P-valeur 68.3%). Ce modèle est utile mais plus compliqué que nécessaire, on ne le retient pas.

En revanche, on peut le simplifier et supprimer la variable dont la contribution marginale n'est pas significative. On calcule ainsi le modèle Prix du forfait / Altitude maximal, Nombre de pistes et on obtient la seconde sortie de régression. Ce modèle est bien statistiquement valide, tant globalement qu'au niveau des validités marginales (toutes les P-valeurs sont quasi-nulles). C'est ce second modèle, et lui seul, que nous exploiterons dans la suite.

***Not* statistically valid**

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,847[a]	,717	,708	20,500

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	100300,605	3	33433,535	79,553	,000[b]
	Residual	39505,283	94	420,269		
	Total	139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort, Number of ski lifts

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,626	9,685		2,956	,004
	Maximum altitude of the resort	,031	,005	,437	6,657	,000
	Number of ski lifts	-,029	,071	-,032	-,410	,683
	Number of slopes	,630	,101	,548	6,265	,000

a. Dependent Variable: Ski pass price (for 7 days)

Statistically valid

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,847[a]	,717	,711	20,411

a. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	100229,959	2	50114,979	120,298	,000[b]
	Residual	39575,929	95	416,589		
	Total	139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,252	9,599		2,943	,004
	Maximum altitude of the resort	,031	,005	,441	6,825	,000
	Number of slopes	,602	,074	,524	8,114	,000

a. Dependent Variable: Ski pass price (for 7 days)

Anticipation sur l'étape 4 : Validité économique. Souvent les sujets de régression vous demanderons dans la même question de valider statistiquement puis économiquement le modèle de régression. La validation économique consiste (comme en régression simple) à se demander si les signes des coefficients de pente sont logiques. Ici, on observe des signes positifs pour les coefficients associés à l'altitude maximale et au nombre de pistes, ce qui veut dire que les prix de forfait augmentent avec l'altitude maximal et le nombre de pistes : c'est parfaitement logique ! On valide donc économiquement le modèle.

Etape 2 : Qualité du modèle. Deux indicateurs de qualité sont proposés : le r^2, qui représente la proportion (de variations) de la variable à expliquer retrouvée par le modèle de régression ; et sa version dite ajustée r^2_{adj} est une légère correction à la baisse, d'autant plus importante qu'il y plus de variables explicatives.

Exemple. On lit $r^2 = 71.7\%$ et $r^2_{adj} = 71.1\%$. Grâce à $r^2 = 71.7\%$ on peut écrire : 71.7% (des variations) du prix du forfait sont expliqués par les (variations des) nombres de pistes et (de) l'altitude maximale. On verra plus loin à quoi sert le r^2_{adj}.

Etape 3 : Ecriture de la relation. Ce point ne change pas fondamentalement par rapport à la régression linéaire simple, il y a simplement davantage de lignes à écrire !

Exemple. On lit ici la relation :

$$\begin{aligned}\text{Prix du forfait (euros)} &= 28.252 \\ &+ 0.031 \times \text{Altitude maximale du domaine (mètres)} \\ &+ 0.602 \times \text{Nombre de pistes (sans unités)} \\ &+ \text{Résidu (d'écart-type : 20.411 euros)}\end{aligned}$$

Etape 4 : Validation économique / interprétation. Nous avons déjà rappelé en quoi consiste la validation économique plus haut. Pour l'interprétation des coefficients, il s'agit de dire, pour les coefficients de pente, que lorsque l'altitude maximale augmente de 100 mètres (disons), le prix du forfait augmente de $0.031 \times 100 = 3.1$ euros (approximativement et en moyenne), et lorsqu'il y a (par exemple) 10 pistes de plus, l'augmentation est de $0.602 \times 10 = 6.02$ euros (approximativement et en moyenne).

C'est un peu plus délicat pour le terme constant 28.252, et comme en régression simple, il faut identifier des valeurs minimales (en regardant le jeu de données) pour l'altitude et le nombre de pistes. On peut obtenir le tableau de statistiques descriptives suivant et on y lit les valeurs minimales 1110 mètres (arrondie à 1100 mètres) et 4 pistes :

		Ski pass price (for 7 days)	Maximum altitude of the resort	Number of ski lifts	Number of slopes
N	Valid	98	98	98	98
	Missing	0	0	0	0
Mean		132,66	2395,26	37,90	50,61
Median		126,00	2450,00	26,50	39,00
Std. Deviation		37,964	542,464	41,308	33,051
Minimum		70	1110	4	4
Maximum		241	3840	377	135

La valeur minimale (approximative et moyenne) pour le prix du forfait résultant de ce modèle est donc de

$$28.252 + 0.031 \times 1100 + 0.602 \times 4 = 64.76 \text{ euros.}$$

La valeur minimale de 70 euros lue dans les données est compatible avec cette valeur minimale moyenne (approximative) de 64.76 euros, voir point suivant.

En ajoutant et en soustrayant (opération blanche !) $0.031 \times 1100 + 0.602 \times 4$ à la relation précédemment écrite, on pourrait la réécrire comme :

$$\begin{aligned}\text{Prix du forfait (euros)} \quad &= 64.76 \\ &+ 0.031 \times (\text{Altitude maximale du domaine [mètres]} - 1100) \\ &+ 0.602 \times (\text{Nombre de pistes [sans unités] - 4}) \\ &+ \text{Résidu (d'écart-type : 20.411 euros)}\end{aligned}$$

Les interprétations des coefficients de pente 0.031 et 0.602 restent identiques, et 64.76 euros apparaît directement comme la valeur minimale moyenne (approximative) du prix du forfait.

Etape 5 : Prévision / Compatibilité. On calcule un intervalle de prévision (où les valeurs individuelles doivent se trouver, avec grande confiance) comme en régression simple : en remplaçant le terme de résidu d'écart-type s par $\pm 2s$.

Exemple. Dans les données, la station de Bessans propose un forfait à 70 euros, a une altitude maximale de 2200 euros et dispose de 4 pistes. Le prix de son forfait est-il compatible avec le modèle ? Le modèle propose une fourchette de prix de forfait de

$$28.252 + 0.031 \times 2200 + 0.602 \times 4 \pm 2 \times 20.411 = 98.86 \pm 40.822\,,$$

qui contient bien la valeur 70. Le prix du forfait à Bessans est compatible avec le modèle.

2. Comparaison de modèles / sélection de modèles

Comparer deux modèles entre eux. Le r^2 augmente mécaniquement avec l'ajout de variables explicatives. Pour comparer de manière plus juste deux modèles avec des nombres de variables explicatives différent, il faut pondérer le r^2 par le nombre de variables. C'est ce que font le r^2_{adj} (plus il est grand, meilleur est le modèle) et l'estimation de l'écart-type des résidus (plus elle est petite, meilleur est le modèle). Vous modulerez toutefois cela en pratique à l'aune de considérations « stratégiques » (= de commentaires subjectifs sur la manière dont vous recevez et pensez pouvoir exploiter les modèles).

Méthodes de sélection automatique. Les sujets d'examen reportent souvent le résultat de deux méthodes de sélection.

La méthode « forward » part de la meilleure variable explicative prise isolément (celle pour laquelle la régression linéaire simple a le plus grand r^2) et enrichit itérativement le modèle d'une variable à la fois, la meilleure addition possible, tant que cela est possible : tant que la variable ajoutée reste individuellement significative.

La méthode « backward » part du modèle complet (= avec toutes les variables explicatives) et le simplifie itérativement en supprimant une variable à la fois, la moins individuellement significative (= celle avec la plus grande P-valeur), tant que cela est nécessaire.

Pourquoi d'ailleurs supprime-t-on (ou ajoute-t-on) seulement une variable à la fois ? Car un manque de significativité individuelle peut venir de deux causes :

— une inutilité intrinsèque (si on prenait pour variable explicative du prix du forfait l'âge du maire de la commune, elle ne jouerait pas !) ;

— une redondance (une variable peut mesurer la même information qu'une autre variable ou une combinaison de variables).

A cause de la possibilité de redondance, il arrive que deux variables ne soient pas simultanément utiles dans le même, mais qu'il soit utile d'en garder une des deux. Dans le modèle complet, les deux variables seraient non individuellement significatives, mais dès qu'une d'elle est supprimée, l'autre devient individuellement significative.

Voici ce que donne la sélection « forward » sur notre exemple :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,760[a]	,578	,574	24,787
2	,847[b]	,717	,711	20,411

a. Predictors: (Constant), Number of slopes

b. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	80825,821	1	80825,821	131,558	,000[b]
	Residual	58980,067	96	614,376		
	Total	139805,888	97			
2	Regression	100229,959	2	50114,979	120,298	,000[c]
	Residual	39575,929	95	416,589		
	Total	139805,888	97			

a. Dependent Variable: Ski pass price (for 7 days)

b. Predictors: (Constant), Number of slopes

c. Predictors: (Constant), Number of slopes, Maximum altitude of the resort

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	88,459	4,596		19,248	,000
	Number of slopes	,873	,076	,760	11,470	,000
2	(Constant)	28,252	9,599		2,943	,004
	Number of slopes	,602	,074	,524	8,114	,000
	Maximum altitude of the resort	,031	,005	,441	6,825	,000

a. Dependent Variable: Ski pass price (for 7 days)