Collection
Statistique
et probabilités
appliquées

Valentin Rousson

$$variance(Y) = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{1}{n}\left(\sum_i y_i^2 - n\bar{y}^2\right) = \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 = mean(Y^2) - mean^2($$

Statistique appliquée aux sciences de la vie

$$\sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - 2\bar{y} = \sum_i y_i^2 - 2y = \sum_i y_i^2 - n$$

$$\mathrm{Var}\left(\sum_i Y_i\right) = \sum_i \mathrm{Var}(Y_i).$$

$$\mathrm{Var}\,(\widehat{\mu}) = Var\left(\frac{\sum_i Y_i}{n}\right) = \frac{\sum_i \mathrm{Var}(Y_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma}{n}$$

Springer

Statistique appliquée aux sciences de la vie

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Valentin Rousson

Statistique appliquée aux sciences de la vie

Springer

Valentin Rousson
Institut Universitaire de Médecine Sociale et Préventive (IUMSP)
Centre Hospitalier Universitaire Vaudois et Université de Lausanne
Route de la Corniche 10
1010 Lausanne
Suisse

ISBN 978-2-8178-0393-7 Springer Paris Berlin Heidelberg New York

Springer-Verlag France est membre du groupe Springer Science + Business Media

Maquette de couverture : Jean-François Montmarché

Collection
Statistique et probabilités appliquées

dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Dans la même collection :

– *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004

– *Optimisation appliquée*
Yadolah Dodge, octobre 2004

– *Le choix bayésien. Principes et pratique*
Christian P. Robert, novembre 2005

– *Régression. Théorie et applications*
Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007

– *Le raisonnement bayésien. Modélisation et inférence*
Éric Parent, Jacques Bernier, juillet 2007

– *Premiers pas en simulation*
Yadolah Dodge, Giuseppe Melfi, juin 2008

– *Génétique statistique*
Stephan Morgenthaler, juillet 2008

– *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique, 2e édition*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, septembre 2009

– *Pratique du calcul bayésien*
Jean-Jacques Boreux, Éric Parent, décembre 2009

– *Statistique. La théorie et ses applications, 2e édition*
Michel Lejeune, septembre 2010

– *Le logiciel R*
Pierre Lafaye de Micheaux, Rémy Drouilhet, Benoît Liquet, novembre 2010

– *Probabilités et processus stochastiques*
Yves Caumel, avril 2011

– *Analyse statistique des risques agro-environnementaux*
David Makowski, Hervé Monod, septembre 2011

Préface

Ce texte d'introduction à la statistique a été initialement écrit en tant que support d'un cours intitulé « Statistique pour biologistes » dispensé aux étudiants de deuxième année du Bachelor ès Sciences en biologie à l'Université de Lausanne (UNIL). Comme les méthodes et la philosophie de la statistique ne diffèrent pas vraiment d'un domaine d'application à un autre, nous avons l'intention de réutiliser tout ou partie de ce support pour d'autres cours d'introduction à la statistique destinés à des étudiants en médecine, pharmacie, neuroscience ou méthodologie clinique. C'est pourquoi ce texte est intitulé « Statistique appliquée aux sciences de la vie ».

Nous avons essayé d'écrire un texte qui soit le plus possible autonome et qui ne repose pas sur une littérature statistique abondante. Nous ne faisons également que peu de référence à l'utilisation d'un logiciel statistique. Les quelques fois où nous l'avons fait, il s'agit du logiciel gratuit R (`www.r-project.org`). Il ne s'agit donc pas d'un manuel d'utilisation de la statistique, mais d'un ouvrage qui doit nous aider à comprendre les principes importants de la statistique.

Les exemples présentés proviennent de différents domaines des sciences de la vie. Certaines des données utilisées sont des données réelles (dont la source est alors indiquée dans le texte), d'autres sont des données fictives. Précisons toutefois que la plupart des analyses présentées sont en grande partie sorties de leur contexte, de sorte que les résultats qui en découlent n'ont ici aucune valeur scientifique sérieuse. Ces exemples servent avant tout à illustrer l'application des méthodes statistiques.

Les étudiants des diverses sciences de la vie ont au moins deux points en commun. Le premier est que les sciences qu'ils étudient sont loin d'être exactes et les méthodes élémentaires de la statistique leur sont particulièrement utiles. Le second est qu'ils sont peu habitués au formalisme mathématique, très présent dans la science statistique et qui a malheureusement mauvaise réputation. Ceci ne peut être cependant qu'un malentendu car le formalisme mathématique ne devrait pas venir compliquer un exposé, mais le préciser et le clarifier. Aussi, bien que nous sommes restés un peu informels sur certains points, notamment en ce qui concerne la définition des variables aléatoires, et plus généralement sur les concepts de probabilités, nous n'avons pas renoncé aux formules mathématiques, notre devise étant qu'« une belle formule vaut mieux que mille mots ». De nombreuses notes de bas de page contiennent des développements

et commentaires qui intéresseront peut-être les statisticiens plus expérimentés et qui permettront de faire le lien avec d'autres ouvrages plus avancés ou mathématiquement plus rigoureux.

Que ce soit avec des mots ou avec des formules, notre but principal est resté toutefois de motiver au mieux l'introduction de chaque nouveau concept statistique, de discuter en détail de son interprétation, de son utilité, de sa valeur ajoutée et de sa relation avec les autres concepts. Nous pensons que la clé de la compréhension de la statistique se trouve dans ce que nous avons appelé le « paradigme de la statistique », à savoir le fait qu'un estimateur peut être vu comme une variable, ce qui permet de faire le lien entre la statistique descriptive et la statistique inférentielle et de donner ainsi à un cours de statistique une certaine unité de doctrine.

Je tiens à remercier Yadolah Dodge qui m'a encouragé à écrire cet ouvrage, Alfio Marazzi qui m'a aidé à simplifier certains passages (dont le titre du livre), Patrick Taffé qui m'a suggéré de nombreuses lectures statistiques intéressantes dont quelques-unes sont citées ici, ainsi que Philippe Vuistiner qui a relu et commenté différentes versions de ce texte. Je remercie également Dieter Häring, Oskar Jenni, Remo Largo et Peter Vollenweider qui m'ont mis à disposition des ensembles de données utilisés dans cet ouvrage.

Valentin Rousson
Décembre 2012

Sommaire

Chapitre 1

Premiers concepts

Nous introduisons la statistique comme la science de la *variabilité*. Voici les données d'une expérience en biologie au cours de laquelle on a mesuré la hauteur (en cm) de $n = 60$ spécimens d'*Onobrychis viciifolia* (une plante herbacée vivace) après une culture de six mois[1] :

21	21	23	22	23	29	24	21	18	23	19	18	20	24	20
20	19	19	22	21	18	20	23	17	20	25	23	21	14	18
29	28	28	14	28	26	22	22	22	29	19	26	16	17	23
18	25	22	20	22	18	32	26	21	20	27	20	19	19	18

La biologie étant une science non exacte, on observe de la variabilité : toutes les plantes n'ont pas atteint la même hauteur. Cette variabilité pose un certain nombre de problèmes/questions :

1. **La variabilité implique de la confusion.** On ne peut pas donner le détail de ces données dans l'`abstract` d'une publication scientifique ; il faudra se contenter d'un résumé. Quelle information concise et utile peut-on extraire de ces données ? *Comment décrire la variabilité ?*

2. **La variabilité implique de l'incertitude.** Sachant que les plantes d'une même expérience atteignent des hauteurs différentes, qu'est-ce qui nous permet d'affirmer que les plantes d'une autre expérience similaire atteindront des hauteurs comparables ? Jusqu'à quel point peut-on généraliser les résultats d'une expérience ? *Comment inférer en présence de variabilité ?*

3. **La variabilité implique des questions scientifiques.** L'un des buts de toute science est d'identifier les sources de variabilité afin d'être en me-

[1] Nous remercions le Dr Dieter Häring de nous avoir mis à disposition ces données, qui représentent un sous-ensemble des données décrites et analysées dans Häring *et al.* (2008).

sure de prévoir (voire même de modifier) les phénomènes étudiés. Pourquoi certaines plantes croissent-elles plus que d'autres ? Quelle hauteur atteindra un *Onobrychis* particulier ? *Comment prédire la variabilité ?*

La statistique nous aide à répondre à ces questions. On distingue en particulier les méthodes de *statistique descriptive*, qui traitent de la première problématique, des méthodes de *statistique inférentielle*, qui traitent de la deuxième. Les techniques de *modélisation statistique* permettront dans une certaine mesure d'aborder la troisième de ces problématiques.

1.1 Population, variable et échantillon

Trois concepts essentiels de la statistique sont ceux de population, de variable et d'échantillon :

- la **population** est l'ensemble des individus d'intérêt d'une étude, que ce soient des patients, des plantes, des insectes ou différents lancers d'une pièce de monnaie ; avant d'entreprendre une étude ou une expérience, il s'agit de définir autant précisément que possible *qui nous intéresse*

- une **variable** est une caractéristique d'intérêt mesurable sur les individus de la population, par exemple l'âge d'un patient, la hauteur après une culture de six mois d'un *Onobrychis* ou le résultat, pile ou face, du lancer d'une pièce de monnaie ; il s'agit ici de définir *quoi nous intéresse*.

La répartition des différentes valeurs possibles d'une variable entre les individus de la population est appelée la *distribution de la variable dans la population*. Afin d'obtenir de l'information sur cette distribution, il ne suffit pas de mesurer un seul individu de la population. Pour des raisons pratiques évidentes, on ne peut pas non plus mesurer l'entier de la population, qui contient typiquement un grand nombre, parfois même une infinité d'individus (on ne pourra pas planter tous les *Onobrychis* susceptibles d'être plantés). Entre ces deux extrêmes, il s'agit de trouver un compromis, et le compromis en statistique s'appelle un échantillon. Ainsi :

- un **échantillon** est un ensemble de quelques individus représentatifs de la population pour lesquels une variable est effectivement mesurée[2].

On mentionnera à nouveau ici les rôles différents de la statistique descriptive et de la statistique inférentielle. La statistique descriptive nous permet de résumer les données d'un échantillon. La statistique inférentielle nous permet de faire le lien entre un échantillon et une population, nous informant sur ce

[2]On utilisera le même terme d'*échantillon* pour désigner à la fois l'ensemble des individus mesurés et l'ensemble des mesures (des observations) faites sur ces individus, c'est-à-dire l'ensemble des données récoltées. On verra cependant au chapitre 9 qu'une distinction entre ces deux concepts peut s'avérer utile dans certains cas.

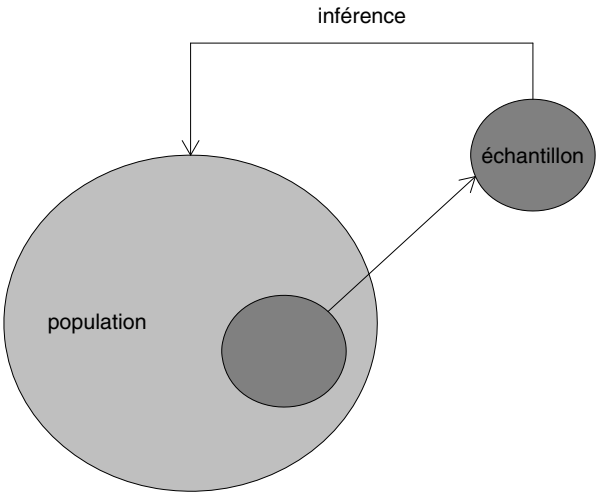

Figure 1.1 – Illustration du contexte général d'une analyse statistique.

que l'on peut savoir de la population à partir des données de l'échantillon. La figure 1.1 illustre le contexte propre à toute analyse statistique.

On s'intéresse à une population, mais on ne dispose que d'un échantillon.

1.2 Échantillonnage et indépendance

Obtenir un échantillon représentatif d'une population n'est en général pas chose aisée. Nous ne discuterons pas ici des différentes techniques d'échantillonnage qui constituent un domaine spécifique de la science statistique. Nous nous contenterons de dire qu'il s'agit de s'assurer que chaque individu de la population ait la même "chance" d'être sélectionné dans l'échantillon afin que toute la diversité de la population y soit représentée.

Afin de constituer notre échantillon, le plus simple serait de disposer d'une liste complète des individus de la population et d'en tirer un certain nombre au hasard (il s'agirait alors d'un *échantillonnage aléatoire*). Ceci sera rarement faisable en pratique et on fera parfois le chemin inverse : on définira la population à partir de l'échantillon, en se demandant de quelle population notre échantillon est représentatif. Si cette population ne correspond pas à celle qui nous

intéresse, on appliquera des critères d'inclusion et d'exclusion. Si par exemple la population d'intérêt ne comprend pas d'enfants, on exclura les enfants de notre échantillon. De même, si une partie spécifique de la population n'est pas représentée dans notre échantillon, on redéfinira la population en conséquence.

Dans ce texte, on supposera aussi (la plupart du temps) que les observations qui constituent un échantillon sont *indépendantes*. On verra plus loin comment définir mathématiquement le concept d'indépendance. On dira pour l'instant que deux observations sont indépendantes si la connaissance de la première ne nous informe en rien sur la valeur de la seconde. Par exemple, deux observations faites sur deux individus tirés au hasard de la population sont indépendantes : savoir que le premier individu mesure 180 cm ne nous informe en rien sur la taille du second. Par contre, si l'échantillonnage est fait de telle sorte que le second individu est l'un des frères du premier, on aura une certaine idée de la taille du second individu avant même de l'avoir mesuré (sachant que le premier individu mesure 180 cm et que deux frères auront tendance à se ressembler). Dans ce cas, les deux observations ne seront pas indépendantes. De même, deux observations faites sur un même individu ne sont en général pas indépendantes.

1.3 Principaux types de variables

Une première étape lors d'une analyse statistique est l'identification des types de variables auxquelles on a affaire. On distingue les variables qualitatives (appelées aussi variables catégorielles) des variables quantitatives. Les valeurs possibles (ou modalités) d'une variable qualitative sont non numériques alors que les valeurs possibles d'une variable quantitative sont numériques. Il existe ensuite essentiellement deux types de variables qualitatives, nominales et ordinales, et deux types de variables quantitatives, discrètes et continues. On a ainsi la classification suivante :

- **variable qualitative nominale**

  → il s'agit d'une variable « purement qualitiative »

  → étant données deux observations a et b d'une variable qualitative nominale, on peut seulement dire si elles sont identiques ou différentes (on aura $a = b$ ou $a \neq b$)

  → exemples : couleur des yeux, espèce animale

- **variable qualitative ordinale**

  → les valeurs possibles d'une variable ordinale peuvent être ordonnées de la plus petite à la plus grande

  → étant données deux observations a et b d'une variable qualitative ordinale, on peut dire si elles sont identiques ou différentes, et si elles sont différentes, on peut dire laquelle des deux est la plus grande (on aura $a = b$, $a < b$ ou $a > b$)

→ exemples : degré de sévérité d'une maladie, satisfaction du patient (sur une échelle qualitative avec les modalités « très insatisfait », « plutôt insatisfait », « plutôt satisfait » et « très satisfait »)

- **variable quantitative discrète**

  → le terme « discret » se réfère au nombre limité (fini) de valeurs possibles

  → étant données deux observations a et b d'une variable quantitative discrète, on peut mesurer la distance $a - b$ qui les sépare (qui pourra être nulle, négative ou positive)

  → exemples : nombre de personnes dans une famille (qui pourra être de $1, 2, 3, \cdots$, mais pas de 1.5, d'où le fait qu'il s'agit d'une variable discrète et non continue), nombre d'espèces animales dans un biotope

- **variable quantitative continue**

  → une variable continue admet une infinité de valeurs possibles[3]

  → étant données deux observations a et b d'une variable quantitative continue, on peut mesurer la distance $a - b$ qui les sépare, et celle-ci ne sera jamais exactement nulle (on aura $a - b \neq 0$; toutes les observations seront différentes, du moins en théorie)

  → exemples : poids, taille (deux individus n'auraient jamais exactement la même taille si les mesures étaient d'une précision infinie).

À l'opposé des variables continues, on a les *variables binaires* :

- **variable binaire**

  → une variable binaire admet seulement deux valeurs possibles, souvent codées par 1 et 0

  → exemples : 1 = homme et 0 = femme, 1 = mort et 0 = vivant, 1 = gauche et 0 = droite, 1 = pile et 0 = face.

Une variable binaire est à première vue un cas particulier de variable qualitative nominale. Une variable binaire peut cependant aussi être considérée comme une variable ordinale (en décrétant que la valeur codée par 1 est plus grande que la valeur codée par 0) et même comme une variable quantitative discrète (en décrétant que la distance entre les deux valeurs possibles est égale à 1).

Dans ce texte, nous allons principalement considérer des variables continues et des variables binaires, qui sont à la fois très courantes dans la pratique et les plus commodes à traiter statistiquement.

[3] Le concept de variable continue est un concept théorique. En pratique, toute variable quantitative est en fait discrète (on aura seulement un nombre fini de valeurs possibles étant donné que les mesures ne sont pas d'une précision infinie), et le choix de la traiter en tant que variable continue ou en tant que variable discrète se posera parfois. Dans notre exemple ci-dessus, on traitera la hauteur des *Onobrychis* comme une variable continue bien que certaines de ces hauteurs soient identiques après les avoir arrondies au centimètre le plus proche.

Chapitre 2

Distribution d'une variable

Lorsque l'on parle de « la distribution d'une variable dans la population », on se réfère à la répartition des différentes valeurs possibles de cette variable entre les individus de la population. Notons cependant que le concept de « distribution d'une variable » peut s'appliquer à n'importe quel ensemble de données, que ce soit une population ou un échantillon. On va voir dans ce chapitre à quoi ressemble une distribution pour les différents types de variables. On va voir également comment on peut résumer une distribution en utilisant des méthodes de statistique descriptive. Ceci se fera à l'aide de *graphiques* et de *caractéristiques numériques*. Une caractéristique numérique fameuse qui s'applique aux variables quantitatives est la *moyenne arithmétique*. La moyenne des hauteurs des 60 *Onobrychis* introduits au chapitre 1 vaut par exemple 21.7 cm. Il s'agit d'un résumé des 60 mesures individuelles que l'on peut aisément reporter dans un abstract de publication scientifique. Pour les variables qualitatives, on remplacera la moyenne par des pourcentages, autre concept fameux de statistique descriptive.

2.1 Distribution d'une variable qualitative

Voici un ensemble de données récoltées sur $n = 158$ individus d'une population (fictive) du nord de l'Europe dont on a mesuré le groupe sanguin :

O	O	O	A	A	A	A	A	O	O	O	O	O	A	A	O
O	O	O	O	O	A	O	O	O	B	O	O	O	O	O	A
A	A	A	O	O	B	O	O	O	O	O	O	A	O	A	O
A	B	O	A	B	O	A	A	A	A	AB	O	A	O	O	O
A	AB	O	AB	O	O	B	A	A	A	O	A	O	O	B	O
O	O	O	A	O	A	A	A	O	O	A	O	A	A	A	O
O	AB	A	A	O	A	O	A	A	A	O	O	A	O	A	O
O	A	O	A	A	O	O	A	A	O	A	O	A	B	A	O
O	A	O	B	O	O	O	A	O	O	A	O	A	A	O	O
O	A	O	A	A	B	A	O	O	O	B	A	A	A		

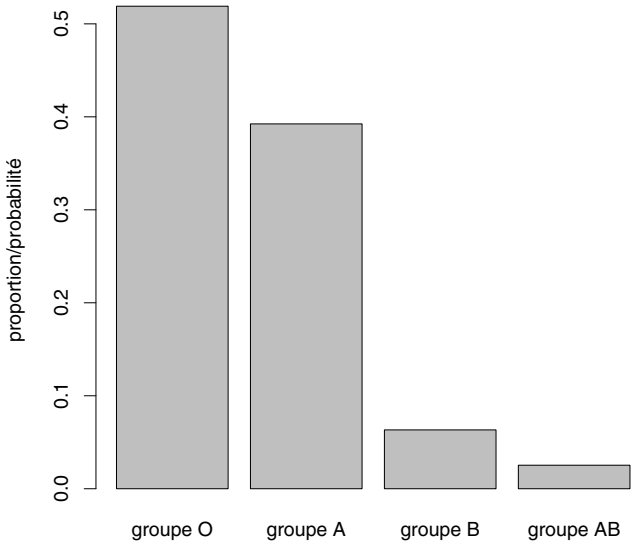

Figure 2.1 – Exemple de diagramme en barres.

Il s'agit d'une variable qualitative nominale avec quatre modalités (groupe O, groupe A, groupe B et groupe AB). Comme ces modalités ne sont pas numériques, on ne peut pas ici calculer de moyenne arithmétique. Au lieu de moyenne, on résume un ensemble de données qualitatives par une table de contingence et par des pourcentages, aussi appelés *proportions* :

groupe O	groupe A	groupe B	groupe AB	total
82	62	10	4	158
52 %	39 %	6 %	3 %	100 %

La proportion des individus avec groupe sanguin O est par exemple de $82/158 = 52$ %. Ces quatre proportions constituent plus qu'un résumé de ces données qualitatives : elles définissent complètement la distribution de la variable (elles donnent l'information complète sur la répartition des valeurs possibles entre les individus). La connaissance des trois premières proportions serait d'ailleurs suffisante, la quatrième pouvant être déduite du total de 100 %. Ces proportions peuvent être représentées graphiquement par un diagramme en barres (en anglais : `barplot`), comme illustré par la figure 2.1. Notons que seule la hauteur de ces barres représente de l'information (la largeur des barres est arbitraire).

On notera qu'une proportion peut aussi être interprétée comme une *probabilité*. Dire que 52 % des individus ont le groupe sanguin O implique que si on tirait au hasard un individu de cet ensemble, il y aurait une probabilité de

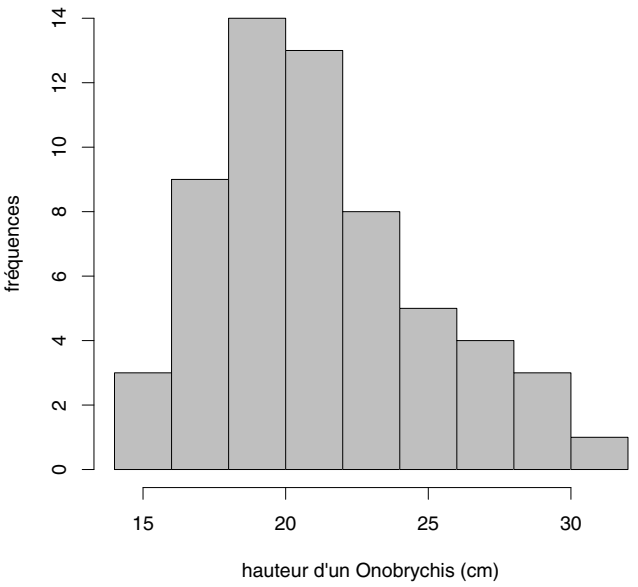

Figure 2.2 – Exemple d'histogramme (fréquences).

52 % (aussi notée 0.52) que l'on observe le groupe sanguin O. En ce sens, les termes de « proportion » et de « probabilité » sont parfaitement synonymes et nous utiliserons indifféremment l'un ou l'autre terme dans ce texte.

Les concepts de table de contingence, de proportion et de diagramme en barres peuvent être utilisés pour résumer/définir la distribution d'une variable qualitative ou d'une variable quantitative discrète. Par contre, ces concepts ne s'appliquent pas aux variables continues, pour lesquelles toutes les observations seront (en théorie) différentes les unes des autres. On aura alors une proportion de $1/n$ pour chacune des n valeurs observées, de sorte qu'un diagramme en barres ne serait pas un résumé de l'information, mais ne ferait que reproduire les données elles-mêmes.

2.2 Distribution d'une variable continue

Afin de résumer graphiquement la distribution d'une variable continue, on peut calculer un *histogramme*. Il s'agit de compartimenter l'ensemble des valeurs possibles en un certain nombre d'intervalles de même longueur et de compter le nombre d'observations dans chaque intervalle. Pour la hauteur des *Onobrychis*, on considère par exemple des intervalles de longueur 2 cm : 14–16 cm, 16–18 cm, etc., jusqu'à 30–32 cm. On compte ainsi 3 *Onobrychis* dans le premier intervalle, 9 dans le deuxième, etc. Un histogramme est une représentation graphique de ces fréquences, comme le montre la figure 2.2.

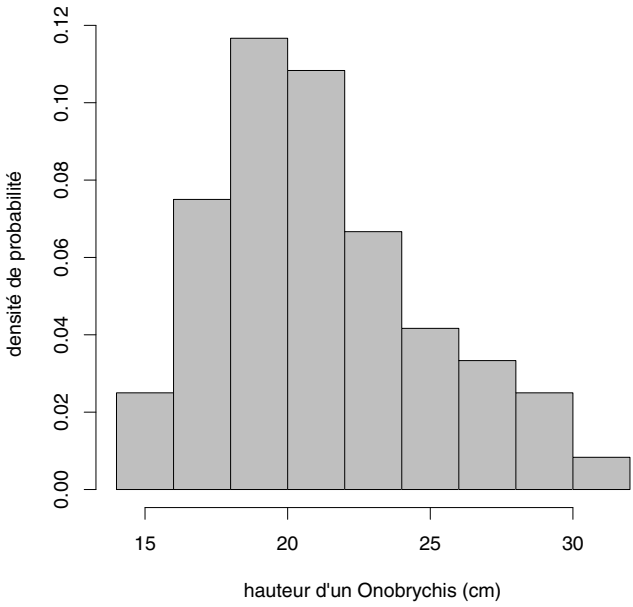

Figure 2.3 – Exemple d'histogramme (densité de probabilité).

Afin de donner une interprétation probabiliste à l'aire des rectangles formant un histogramme, on peut diviser ces fréquences par le nombre total d'observations n et par la longueur des intervalles. Dans notre exemple, on divise les fréquences d'*Onobrychis* par 120 (60×2) et on obtient $3/120 = 0.025$ pour le premier intervalle, $9/120 = 0.075$ pour le deuxième, etc. Un histogramme montrant ces quantités est représenté dans la figure 2.3. La forme de l'histogramme est exactement la même, mais l'échelle sur l'axe vertical a changé, de sorte que les aires des différents rectangles peuvent être ici interprétées comme des proportions ou probabilités. Si on tirait au hasard un *Onobrychis* de cet ensemble, il y aurait une probabilité de $2 \times 0.025 = 0.05$ qu'il mesure entre 14 et 16 cm, une probabilité de $2 \times 0.075 = 0.15$ qu'il mesure entre 16 et 18 cm, etc. L'aire totale de l'histogramme vaut par conséquent 1 (la probabilité qu'un *Onobrychis* tiré au hasard de cet ensemble mesure entre 14 et 32 cm est en effet égale à 1).

Un tel histogramme ne nous donne cependant pas d'information sur la probabilité d'obtenir un *Onobrychis* entre 14 et 15 cm, ou entre 15 et 16 cm. Il faudrait pour cela considérer des intervalles plus petits, ce qui est problématique dans un petit échantillon à cause du manque d'observations. Il faudrait pour cela un échantillon plus grand. Dans une population de taille infinie, on pourrait même considérer des intervalles infiniment petits. L'histogramme aurait alors les apparences d'une courbe que l'on appelle *densité de probabilité*.

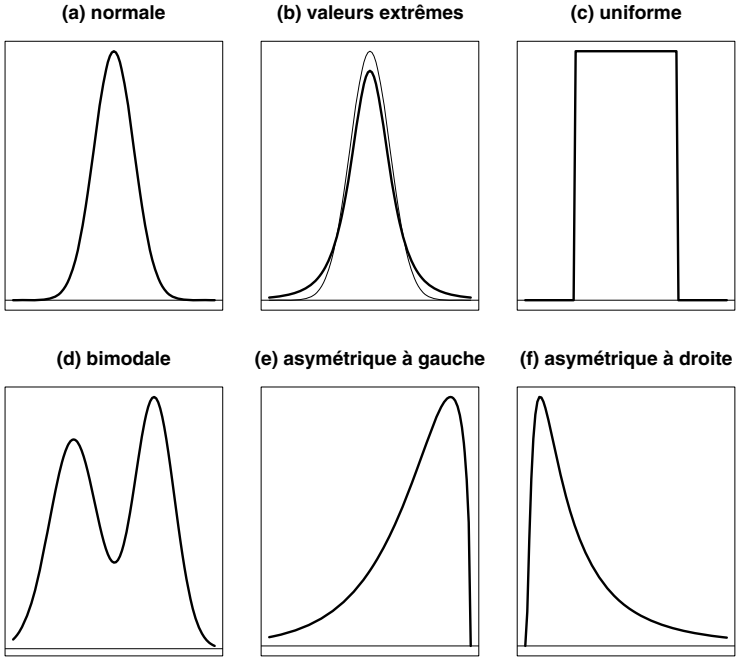

Figure 2.4 – Exemples de densités de probabilité pour une variable continue.

2.3 Densité de probabilité

Une densité de probabilité est une fonction positive, dont l'aire sous la courbe vaut 1, définissant la distribution d'une variable continue dans une population infinie. La figure 2.4 nous montre six exemples de densités de probabilité[1]. Le graphique (a) nous montre une densité en forme de cloche. Il s'agit de la fameuse *densité normale* que l'on rencontre souvent dans un cours de statistique. Lorsque la distribution d'une variable continue peut être décrite par une telle densité, on parlera de *variable normale* et on dira que les observations de cette variable proviennent d'une *distribution normale*.

Une densité de probabilité peut s'interpréter comme suit. L'axe horizontal représente le continuum de l'ensemble des valeurs possibles. La valeur de la densité en un point de ce continuum est proportionnelle à la probabilité d'obtenir des données aux alentours de ce point. Plus précisément, la probabilité

[1]Pour les connaisseurs, les distributions représentées sur ces graphiques sont les suivantes : (a) normale ; (b) Student avec 3 *dl* ; (c) uniforme ; (d) mélange de deux normales avec moyennes et variances différentes (50 % des observations proviennent d'une distribution normale avec moyenne -2 et variance 1.2^2, et 50 % proviennent d'une distribution normale avec moyenne 2 et variance 1) ; (e) khi-deux avec 3 *dl* (multipliée par -1) ; (f) log-normale (distribution de l'exponentielle d'une variable normale standardisée).

d'obtenir une observation dans un intervalle de ce continuum est égale à l'aire sous la courbe dans cet intervalle. Si la distribution d'une variable continue Y est définie par une densité de probabilité $f(y)$, et si a et b sont deux valeurs possibles de Y, la probabilité d'obtenir une observation entre a et b sera donc :

$$\Pr\{a \leq Y \leq b\} = \int_a^b f(y)\,\mathrm{d}y.$$

En particulier, on aura $\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y = 1$. Par ailleurs, la probabilité d'obtenir une observation plus petite ou égale à une valeur a est appelée la *fonction de répartition* de la variable Y. On notera cette probabilité comme suit :

$$F(a) = \Pr\{Y \leq a\} = \int_{-\infty}^{a} f(y)\,\mathrm{d}y.$$

Un fonction de répartition $F(y)$ est donc une fonction croissante (de 0 pour de petites valeurs de y, vers 1 pour de grandes valeurs), dont la dérivée est égale à la densité de probabilité (on a ainsi $F'(y) = f(y)$). On retrouvera le concept de fonction de répartition notamment au chapitre 10 dans le contexte du calcul de la puissance d'un test statistique[2].

Si on connaissait la densité de probabilité de la hauteur des *Onobrychis* après six mois, il serait alors possible de calculer la probabilité d'obtenir un *Onobrychis* entre 14 et 15 cm ou entre 15 et 16 cm. La figure 2.5 nous illustre quelques exemples de calcul de probabilités à partir d'une densité normale. Notons que la probabilité théorique d'obtenir exactement une valeur donnée est nulle. De la même manière que l'on ne verra jamais deux *Onobrychis* mesurant exactement la même hauteur, on ne verra jamais un *Onobrychis* mesurant exactement 15 cm. Par contre, la probabilité d'obtenir un *Onobrychis* mesurant entre 14.9 et 15.1 cm sera certes petite, mais non nulle.

Deux caractéristiques importantes d'une distribution normale sont la *symétrie* de la densité et le fait que cette fonction tende rapidement vers zéro lorsqu'on s'éloigne du centre de symétrie, comme le montre le graphique (a) de la figure 2.4. En ce sens, on observera peu de *valeurs extrêmes* pour une variable normale. Une troisième caractéristique de la normalité est l'*unimodalité*, c'est-à-dire le fait que la densité n'admette qu'un seul *mode*, ou maximum.

La figure 2.4 nous montre cinq exemples de distributions continues non normales. Le graphique (b) nous montre une densité avec plus de valeurs extrêmes qu'une densité normale (représentée en arrière-plan par un trait fin sur ce graphique). Le graphique (c) nous montre une densité *uniforme*, pour laquelle l'ensemble des valeurs possibles est borné, et où la probabilité d'obtenir

[2]Le concept de fonction de répartition est une alternative au concept de densité de probabilité pour définir la distribution d'une variable continue. Comme une égalité n'est en principe pas possible pour une variable continue, on pourrait remplacer l'inégalité large « $\leq$ » par une inégalité stricte « $<$ » sans modifier le sens de cette définition. Cependant, une fonction de répartition (au contraire d'une densité de probabilité) peut également être définie pour une variable quantitative discrète, auquel cas l'inégalité large doit être conservée.

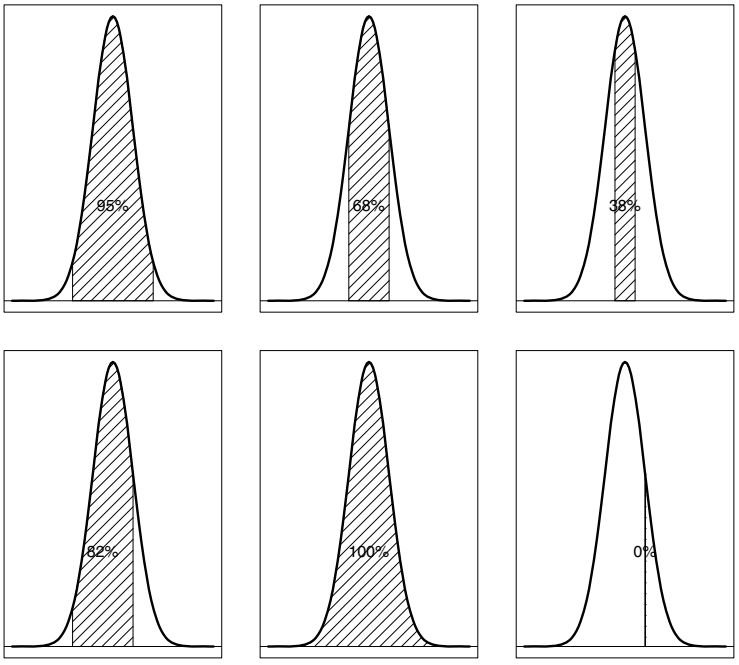

Figure 2.5 – Probabilités associées à une distribution normale.

une observation dans un intervalle de cet ensemble est proportionnelle à la longueur de cet intervalle[3]. Le graphique (d) nous montre une densité *bimodale*, que l'on aura si notre population est constituée de deux sous-populations très différentes, par exemple d'hommes et de femmes. Les deux derniers graphiques nous montrent des densités asymétriques, vers la gauche pour le graphique (e), vers la droite pour le graphique (f). La variabilité s'étale ici plus dans un sens que dans l'autre, ce qui est typique d'une variable admettant un maximum, respectivement un minimum parmi les valeurs possibles. Les distributions asymétriques vers la droite sont courantes dans les sciences de la vie.

Un histogramme calculé dans un échantillon nous donne une idée (une estimation) de la densité de probabilité de la variable dans la population. L'histogramme de la hauteur des *Onobrychis* de la figure 2.3 nous suggère par exemple une distribution relativement proche d'une distribution normale. Un problème pratique de l'histogramme est cependant le choix de la longueur des intervalles (et par conséquent du nombre de ces intervalles). Or, ce choix est important. La figure 2.6 nous montre des histogrammes de ces mêmes *Onobrychis* calculés

[3]La densité de probabilité d'une distribution uniforme entre les valeurs a et b est donnée par $f(y) = 1/(b-a)$ si $a \leq y \leq b$ (et $f(y) = 0$ sinon), alors que la fonction de répartition est donnée par $F(y) = 0$ si $y < a$, $F(y) = (y-a)/(b-a)$ si $a \leq y \leq b$ et $F(y) = 1$ si $y > b$.

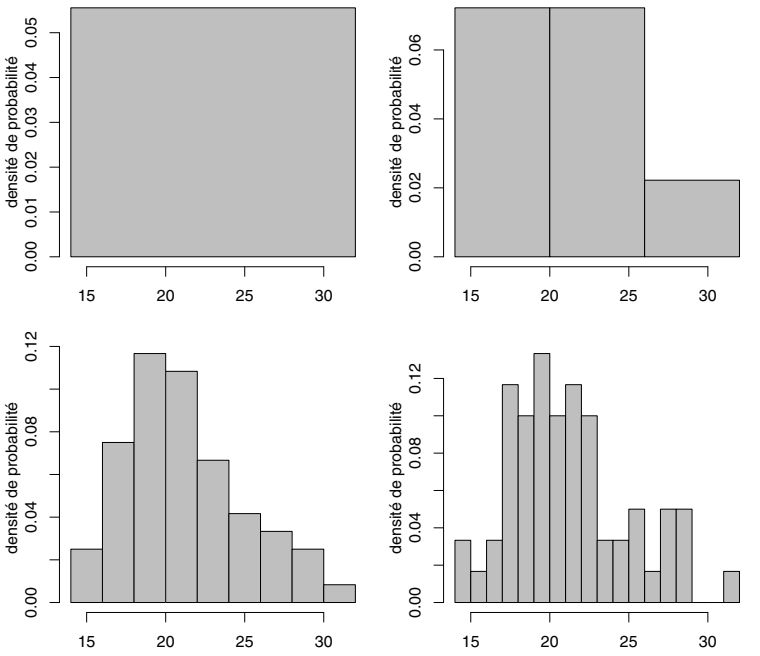

Figure 2.6 – Histogrammes des mêmes données avec différentes longueurs d'intervalle.

avec différentes longueurs d'intervalle. Choisir des intervalles trop grands nous empêche d'identifier correctement la forme de la distribution. Dans le cas extrême où l'on ne considère qu'un seul intervalle, l'histogramme nous suggère (trivialement) une distribution uniforme. Par ailleurs, choisir des intervalles trop petits mettra en évidence certains détails de la distribution dus au hasard de l'échantillonnage, que l'on ne retrouverait pas si l'histogramme était calculé sur l'ensemble de la population.

2.4 Boxplot et quantiles

Un diagramme en boîte, ou pour utiliser le terme anglais, un `boxplot`, est une alternative à l'histogramme pour résumer graphiquement la distribution d'une variable continue, qui a l'avantage de ne pas dépendre d'un choix de l'utilisateur (tel que le choix de la longueur des intervalles pour un histogramme). Le boxplot est parfois appelé *résumé à cinq valeurs*. En effet, les cinq caractéristiques numériques suivantes sont représentées dans un boxplot :

- le **minimum** (sans les valeurs extrêmes)
- le **quantile 25 %** (appelé aussi le 1^{er} *quartile*)

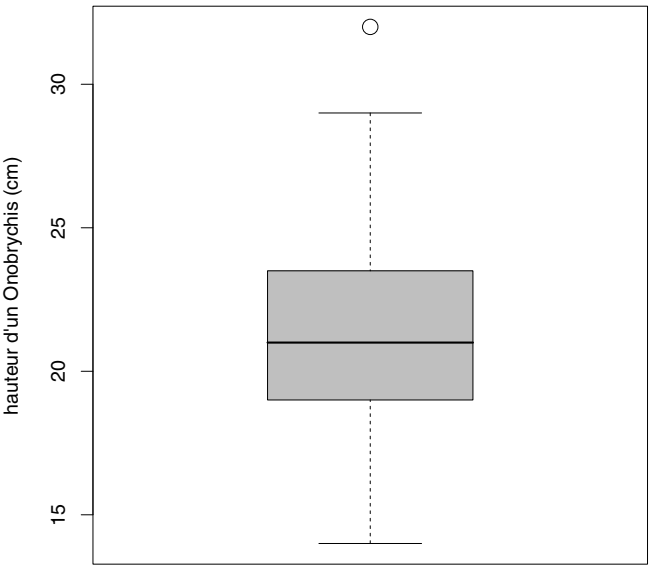

Figure 2.7 – Exemple de boxplot.

- le **quantile 50 %** (appelé aussi le 2^e *quartile* ou la *médiane*)
- le **quantile 75 %** (appelé aussi le 3^e *quartile*)
- le **maximum** (sans les valeurs extrêmes).

Par définition, 25 % des observations sont plus petites et 75 % des observations sont plus grandes que le quantile 25 % ; 50 % des observations sont plus petites et 50 % sont plus grandes que le quantile 50 % ; 75 % des observations sont plus petites et 25 % sont plus grandes que le quantile 75 %[4].

> Par définition, q% des observations sont plus petites et $(100 - q)$% sont plus grandes que le *quantile* q%.

La figure 2.7 nous montre le boxplot de la hauteur des *Onobrychis*. L'ensemble des valeurs possibles est représenté sur l'axe vertical. Le boxplot s'étend du minimum au maximum (dans cet exemple, de 14 à 29 cm). La boîte du boxplot s'étend du quantile 25 % au quantile 75 % (de 19 à 23.25 cm) et contient donc 50 % des observations. Le quantile 50 % (21 cm) est représenté par un

[4]Dans la population, le quantile α d'une variable continue Y avec fonction de répartition $F(y)$ est ainsi la valeur q_α pour laquelle on a $F(q_\alpha) = \alpha$.

trait horizontal à l'intérieur de la boîte. Les *valeurs extrêmes* sont représentées séparément (il y a dans notre exemple un *Onobrychis* extrême, mesurant 32 cm). Une observation est définie comme une valeur extrême si :

- elle est au-delà du quantile 75 % de plus de 1.5 fois la longueur de la boîte
- elle est en deçà du quantile 25 % de plus de 1.5 fois la longueur de la boîte.

Exemple 2.1 *Nous expliquons ici plus en détail comment se calculent les cordonnées du boxplot de la figure 2.7. On ordonne tout d'abord les observations de la plus petite à la plus grande (toujours exprimées en cm) :*

14	14	16	17	17	18	18	18	18	18	18	18	19	19	19
19	19	19	20	20	20	20	20	20	20	20	21	21	21	21
21	21	22	22	22	22	22	22	22	23	23	23	23	23	23
24	24	25	25	26	26	26	27	28	28	28	29	29	29	32

On calcule ensuite les quantiles 25 %, 50 % *et* 75 %. *Comme il y a* $n = 60$ *observations, ces quantiles correspondent ici aux observations avec rangs* 15.75, 30.5 *et* 45.25. *On procède ainsi*[5] *:*

- *les observations des rangs* 15 *et* 16 *étant toutes deux égales à* 19 *cm, l'observation de rang* 15.75 *(et donc le quantile* 25 %*) vaut de même* 19 *cm*
- *les observations des rangs* 30 *et* 31 *étant toutes deux égales à* 21 *cm, l'observation de rang* 30.5 *(et donc le quantile* 50 %*) vaut de même* 21 *cm*
- *les observations des rangs* 45 *et* 46 *étant égales à* 23 *et* 24 *cm, l'observation de rang* 45.25 *(et donc le quantile* 75 %*) est défini par la moyenne pondérée suivante :* $0.75 \cdot 23 + 0.25 \cdot 24 = 23.25$ *cm.*

La longueur de la boîte est donc égale à $23.25 - 19 = 4.25$ *cm. Une observation est alors considérée comme une valeur extrême si elle est supérieure à* $23.25 + 1.5 \cdot 4.25 = 29.625$ *cm ou inférieure à* $19 - 1.5 \cdot 4.25 = 12.625$ *cm. Ainsi, seule l'observation de* 32 *cm satisfait l'un de ces critères. Le minimum et le maximum de l'ensemble de données sans cette unique valeur extrême sont alors de* 14 *et* 29 *cm, complétant les coordonnées du boxplot.*

[5] Le rang correspondant au quantile 75 % est par exemple défini comme suit : il s'agit du rang r pour lequel la distance entre les rangs 1 et r correspond à 75 % de la distance entre les rangs 1 et 60, c'est-à-dire pour lequel on a $r - 1 = 0.75(60 - 1)$. On trouve ainsi $r = 1 + 0.75 \cdot 59 = 45.25$. L'observation y correspondant au rang 45.25 se calcule ensuite comme suit : comme son rang est trois fois plus proche du rang 45 que du rang 46, y sera également trois fois plus proche de l'observation correspondant au rang 45 (qui vaut 23 cm) que de l'observation correspondant au rang 46 (qui vaut 24 cm) ; on résout donc $3(y - 23) = 24 - y$ et on trouve $y = 0.75 \cdot 23 + 0.25 \cdot 24 = 23.25$ cm. Notons toutefois que la manière de calculer les quantiles pourra différer légèrement d'un logiciel statistique à un autre.

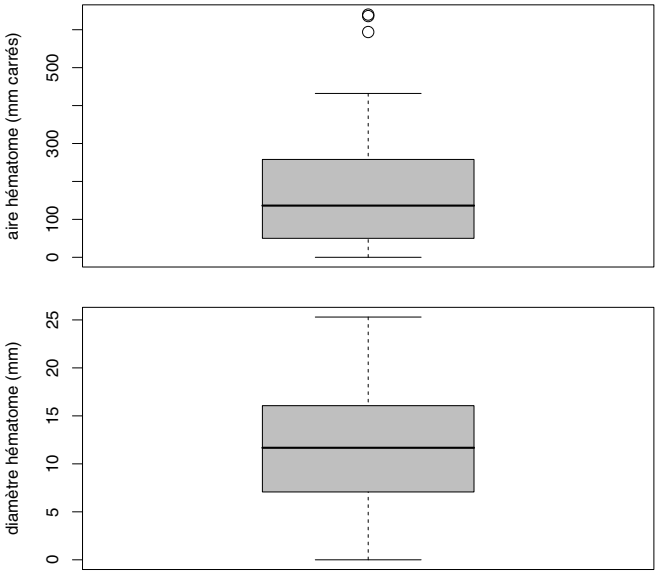

Figure 2.8 – Exemple de transformation (racine carrée) pour approcher la normalité.

La position de la boîte d'un boxplot nous montre où se trouve la partie centrale des données, alors que la longueur de la boîte est une mesure de leur variabilité. Un boxplot nous informe également sur la forme de la distribution. Dans notre exemple, la distribution apparaît à peu près symétrique (le quantile 50 % est proche du centre de la boîte et à mi-chemin entre minimum et maximum) et il y a peu de valeurs extrêmes. On est ainsi proche d'une distribution normale. Dans d'autres exemples, un boxplot nous permettra de reconnaître des signes de non-normalité, notamment l'asymétrie d'une distribution.

Dans le cas où l'on a une distribution non normale, il sera souvent utile de changer d'échelle afin de se rapprocher de la normalité, ce qui aura l'avantage de simplifier l'analyse statistique (on verra en effet que l'utilisation de certaines méthodes statistiques nécessite la normalité des observations). Changer d'échelle signifie transformer les données de façon monotone (alors que changer d'unités signifiera dans ce texte transformer les données de façon linéaire/affine). Une transformation monotone est une transformation qui respecte l'ordre mais modifie les distances entre les observations. On pourra essayer par exemple une transformation logarithmique (que l'on discutera plus loin) ou, comme dans l'exemple ci-dessous, la transformation « racine carrée ».

Exemple 2.2 *Le boxplot du haut de la figure 2.8 résume les aires de* $n = 34$ *hématomes (mesurées en* mm^2*)*[6]*. La distribution est ici clairement asymétrique*

[6]Nous remercions le Dr Valéria Caso de nous avoir mis à disposition ces données.

vers la droite (et donc non normale) : le quantile 50 % *est plus proche du quantile* 25 % *que du quantile* 75 %*, plus proche du minimum que du maximium, et les trois valeurs extrêmes se trouvent à droite de la distribution. Le boxplot du bas de la figure 2.8 résume les racines carrées des aires de ces hématomes (interprétables comme des diamètres exprimés en mm). La distribution est ici parfaitement symétrique et les valeurs extrêmes ont disparu, ou plutôt, elles ont été réconciliées avec les données. On est donc proche de la normalité, ce qui nous suggère que l'échelle adéquate (l'échelle naturelle) pour une analyse statistique de cette variable est le diamètre de l'hématome plutôt que son aire.*

2.5 Mesures de tendance centrale

On considère un ensemble de n observations $y_1, y_2, \cdots, y_n$ d'une variable quantitative Y. En statistique descriptive, on est intéressé à définir des caractéristiques numériques qui résument la distribution d'un ensemble de données. Une mesure de tendance centrale est une caractéristique numérique nous informant sur la position du centre (du milieu) des données. Parmi celles-ci, on a notamment la *moyenne arithmétique* (en anglais : `mean`), aussi appelée simplement « moyenne », et la *médiane* (en anglais : `median`), aussi appelée « quantile 50 % ». Ces quantités sont définies comme suit :

- **la moyenne**[7] :

$$mean(Y) = \frac{1}{n}\sum_i y_i$$

  → la moyenne est la somme des observations divisée par le nombre des observations

[7] Dans ce qui suit, la notation $\sum_i y_i$ (que l'on devrait en fait écrire $\sum_{i=1}^{n} y_i$) désignera la somme des observations y_i :

$$\sum_i y_i = y_1 + y_2 + \cdots + y_n.$$

Notons que ces sommes sont remplacées par des intégrales si on a une infinité d'observations. Pour une variable continue Y avec densité $f(y)$, la moyenne d'une population infinie est ainsi définie par :

$$mean(Y) = \int_{-\infty}^{\infty} y f(y)\,\mathrm{d}y$$

alors que la variance (concept que l'on verra plus loin) de cette population est définie par :

$$variance(Y) = \int_{-\infty}^{\infty} (y - mean(Y))^2 f(y)\,\mathrm{d}y.$$

Pour une variable Y quantitative discrète, on aura par contre :

$$mean(Y) = \sum_y y \cdot \Pr\{Y = y\} \quad \text{et} \quad variance(Y) = \sum_y (y - mean(Y))^2 \cdot \Pr\{Y = y\}$$

où ces sommes s'effectuent sur les différentes valeurs possibles y de la variable Y et où les $\Pr\{Y = y\}$ dénotent les probabilités d'obtenir ces différentes valeurs.

→ on utilisera parfois la notation $\bar{y} = \frac{1}{n}\sum_i y_i$

→ la moyenne est plus fine (admet davantage de valeurs possibles) que la médiane

→ la moyenne est le centre de gravité des observations :

$$\begin{aligned}\sum_i (y_i - \bar{y}) &= \sum y_i - n\bar{y} \\ &= n\bar{y} - n\bar{y} \\ &= 0\end{aligned}$$

→ la moyenne n'est pas toujours représentative de l'ensemble des observations (par exemple dans le cas d'une distribution asymétrique)

→ la moyenne est sensible aux valeurs extrêmes

- **la médiane** :

$$median(Y) = \begin{cases} y_{((n+1)/2)} & \text{si } n \text{ impair} \\ \frac{1}{2}(y_{(n/2)} + y_{(n/2+1)}) & \text{si } n \text{ pair} \end{cases}$$

→ pour calculer la médiane, il s'agit au préalable d'ordonner les observations ($y_{(i)}$ dénotant la i-ième observation ordonnée)

→ la médiane se trouve au milieu des données, dans le sens où 50 % des observations sont plus grandes et 50 % plus petites que la médiane

→ la médiane est donc toujours représentative de l'ensemble des observations (même si la distribution est asymétrique)

→ la médiane est « robuste » aux valeurs extrêmes

→ la médiane peut se calculer aussi pour les variables ordinales.

Exemple 2.3 *Pour l'exemple des Onobrychis, moyenne et médiane sont semblables : on a en effet* $mean(Y) = 21.7$ *cm et* $median(Y) = 21$ *cm. Si on avait toutefois remplacé par erreur la valeur* 32 *cm par la valeur* 3200 *cm, on aurait trouvé* $mean(Y) = 74.5$ *cm et* $median(Y) = 21$ *cm. La valeur de la moyenne est ainsi très affectée par la présence d'une seule valeur extrême ou aberrante, au contraire de la médiane. En ce sens, la médiane est dite « robuste » aux valeurs extrêmes ou aberrantes.*

La similarité entre la moyenne et la médiane est un indicateur de la symétrie d'une distribution, la moyenne étant plus élevée que la médiane pour une distribution asymétrique vers la droite, la médiane étant plus élevée que la moyenne pour une distribution asymétrique vers la gauche.

La question de l'utilisation de la moyenne ou de la médiane comme mesure de tendance centrale est un vieux débat. Pour les variables quantitatives discrètes avec peu de valeurs possibles, la moyenne aura l'avantage d'être plus fine

que la médiane. Par exemple, la médiane de la variable « nombre d'enfants par famille » sera égale à 1 ou 2 dans la plupart des pays occidentaux, alors que la moyenne de cette variable sera différente dans chaque pays, permettant une comparaison beaucoup plus fine entre les pays. Pour les variables continues, la médiane est souvent préférée à la moyenne, en tous cas dans le cadre des sciences de la vie où l'on est en général intéressé à avoir une valeur représentative de notre population, ainsi que robuste aux valeurs extrêmes[8]. Cependant, dans les cas où moyenne et médiane seront semblables, par exemple pour une variable proche de la normalité, on utilisera la moyenne qui est plus commode à analyser et à modéliser statistiquement.

2.6 Mesures de variabilité

La moyenne ou la médiane ne résument qu'un aspect de la distribution d'une variable quantitative, nous informant sur la position du centre des données. Un autre aspect important, on y revient, est la variabilité. Considérons deux espèces d'*Onobrychis*, une première où tous les spécimens atteindraient des hauteurs entre 18 et 22 cm, et une seconde où les hauteurs varieraient beaucoup plus, entre 10 et 30 cm. Les deux distributions seraient très différentes mais auraient une moyenne semblable (aux alentours de 20 cm). En calculant uniquement la moyenne, on ne rendrait pas compte des différences entre ces deux espèces. Pour ce faire, il nous faudra calculer une *mesure de variabilité*.

Il existe de nombreuses mesures de variabilité, les plus connues étant la variance et l'écart type (en anglais : `standard deviation`). Notons qu'une mesure de variabilité admet des valeurs positives ou nulles, une absence totale de variabilité impliquant une valeur nulle. On citera ainsi :

- **la variance** :

$$variance(Y) = mean((Y - mean(Y))^2) = \frac{1}{n}\sum_i (y_i - \bar{y})^2$$

→ la variance est la moyenne des carrés des écarts à la moyenne

→ on peut obtenir une formulation alternative de la variance comme suit ;

[8] Dans les sciences économiques, on s'intéressera par contre à la moyenne plutôt qu'à la médiane lorsqu'il s'agira par exemple d'équilibrer un budget, car la moyenne est liée au total (on peut retrouver un total à partir d'une moyenne, non à partir d'une médiane). Si un impôt total de 5 000 000 euros est budgétisé pour une commune de 5000 habitants imposables, il faudra s'assurer que la moyenne (et non la médiane) des contributions de ces habitants soit de 1000 euros (car $5000 \times 1000 = 5\ 000\ 000$). Le choix entre moyenne et médiane dépendra donc surtout du contexte.

on montre tout d'abord que :

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i y_i^2 - 2\bar{y}\sum_i y_i + n\bar{y}^2 \\ &= \sum_i y_i^2 - 2\bar{y}(n\bar{y}) + n\bar{y}^2 \\ &= \sum_i y_i^2 - n\bar{y}^2\end{aligned}$$

et on obtient :

$$\begin{aligned}variance(Y) &= \frac{1}{n}\sum_i (y_i - \bar{y})^2 \\ &= \frac{1}{n}\left(\sum_i y_i^2 - n\bar{y}^2\right) \\ &= \frac{1}{n}\sum_i y_i^2 - \bar{y}^2 \\ &= mean(Y^2) - mean^2(Y)\end{aligned}$$

→ la variance est donc aussi la moyenne des carrés moins le carré de la moyenne

→ pour calculer la variance, on n'a donc pas besoin de connaître le détail des observations, mais seulement de connaître les deux moyennes suivantes[9] : $\frac{1}{n}\sum_i y_i$ et $\frac{1}{n}\sum_i y_i^2$

→ la variance est une mesure de variabilité exprimée dans le carré des unités des observations

- **l'écart type** : $stdev(Y) = \sqrt{variance(Y)} = \sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2}$

  → l'écart type est la racine carrée de la variance

  → l'écart type est une mesure de variabilité exprimée dans la même unité que les observations

  → l'écart type (tout comme la variance) est sensible aux valeurs extrêmes

- **l'étendue** : $range(Y) = y_{(n)} - y_{(1)}$

  → l'étendue (en anglais : `range`) est égale à la valeur maximale moins la valeur minimale des observations

  → l'étendue est difficile à définir dans une population infinie

  → l'étendue est hypersensible aux valeurs extrêmes

[9] Alternativement, si on connaît le nombre n d'observations, on peut calculer la variance à partir des deux sommes suivantes : $\sum_i y_i$ et $\sum_i y_i^2$.

- **l'intervalle interquartile** : $iqrange(Y) = y_{(3n/4)} - y_{(n/4)}$

  → l'intervalle interquartile (en anglais : `interquartile range`) est égal à la longueur de la boîte du boxplot des observations

  → l'intervalle interquartile est robuste aux valeurs extrêmes

- **le coefficient de variation** : $varcoef(Y) = stdev(Y)/mean(Y)$

  → le coefficient de variation est égal à l'écart type divisé par la moyenne

  → il s'agit d'une mesure de variabilité sans unité

  → le coefficient de variation est défini seulement pour des variables avec valeurs possibles positives (il risquerait sinon d'être négatif, ce qui n'aurait pas de sens pour une mesure de variabilité).

Exemple 2.4 *Dans l'exemple des* $n = 60$ *Onobrychis, on calcule ces différentes mesures de variabilité comme suit :*

- $variance(Y) = 29\ 126/60 - (1302/60)^2 = 14.5\ cm^2$

  → *on aura auparavant calculé* $\sum_i y_i = 1302$ *et* $\sum_i y_i^2 = 29\ 126$

- $stdev(Y) = \sqrt{14.5} = 3.8\ cm$

- $range(Y) = 32 - 14 = 18\ cm$

- $iqrange(Y) = 23.25 - 19 = 4.25\ cm$

- $varcoef(Y) = 3.8/21.7 = 0.18$.

2.7 Changement d'unités

Lorsque l'on change d'unités (lorsque l'on transforme une variable Y de façon linéaire/affine, en la multipliant par une constante b et en lui ajoutant une autre constante a, obtenant $a + bY$), par exemple lorsque l'on passe de centimètres en mètres pour la hauteur d'un *Onobrychis*, ou de degrés Celsius en degrés Fahrenheit pour une température, on a les règles élémentaires suivantes :

- $mean(a + bY) = a + b \cdot mean(Y)$
- $variance(a + bY) = b^2 \cdot variance(Y)$
- $stdev(a + bY) = |b| \cdot stdev(Y)$.

On a en particulier :

$$variance(-Y) = variance(Y) \quad \text{et} \quad stdev(-Y) = stdev(Y)$$

(logique : une mesure de variabilité ne peut pas être négative). On mentionnera également d'autres règles analogues[10] :

- $median(a + bY) = a + b \cdot median(Y)$
- $quantile(a + bY) = a + b \cdot quantile(Y)$ (pour n'importe quel quantile)
- $minimum(a + bY) = a + b \cdot minimum(Y)$
- $maximum(a + bY) = a + b \cdot maximum(Y)$
- $range(a + bY) = |b| \cdot range(Y)$
- $iqrange(a + bY) = |b| \cdot iqrange(Y)$
- $varcoef(bY) = varcoef(Y)$ (avec b positif).

Exemple 2.5 *Un changement d'unités de centimètres en mètres (le remplacement d'une variable Y par $Y/100$) a les conséquences suivantes :*

- $mean(Y) = 25\ cm$ *implique* $mean(Y/100) = 0.25\ m$
- $variance(Y) = 36\ cm^2$ *implique* $variance(Y/100) = 0.0036\ m^2$
- $stdev(Y) = 6\ cm$ *implique* $stdev(Y/100) = 0.06\ m$.

Un changement d'unités de degrés Celsius en degrés Fahrenheit (le remplacement d'une variable Y par $32 + 1.8 \cdot Y$) a les conséquences suivantes :

- $mean(Y) = 30^oC$ *implique* $mean(32 + 1.8 \cdot Y) = 86\ F$
- $variance(Y) = 16^oC^2$ *implique* $variance(32 + 1.8 \cdot Y) = 51.84\ F^2$
- $stdev(Y) = 4^oC$ *implique* $stdev(32 + 1.8 \cdot Y) = 7.2\ F$.

2.8 Distribution normale

On a déjà mentionné le concept de distribution normale. Il est temps de définir mathématiquement la densité de probabilité d'une variable normale. En fait, il n'existe pas une seule distribution normale. Il s'agit d'une famille de distributions indexée par deux paramètres : la moyenne, que l'on notera ici μ, et l'écart type, que l'on notera σ. La densité d'une variable normale avec

[10] On a par ailleurs :

$$median(g(Y)) = g(median(Y))$$

pour toute transformation monotone g (pas seulement pour g linéaire/affine), par exemple pour $g(Y) = \log(Y)$ ou $g(Y) = \sqrt{Y}$, cette propriété étant aussi valable pour n'importe quel quantile, y compris pour le minimum et le maximum, mais pas pour la moyenne.

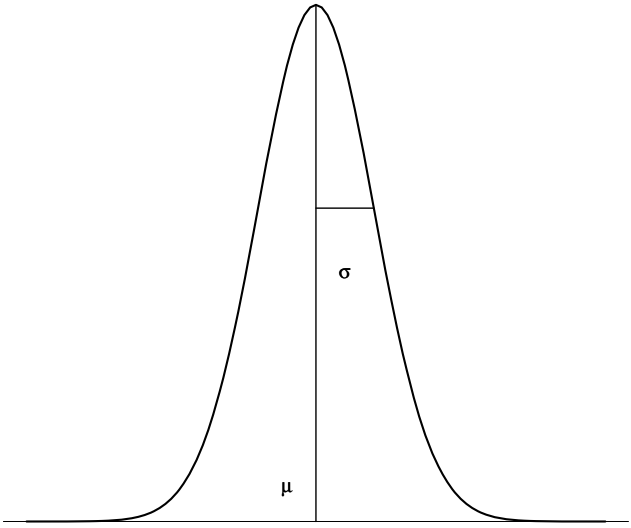

Figure 2.9 – Paramètres μ et σ pour une distribution normale.

moyenne μ et écart type σ est définie par :

$$\phi_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

La figure 2.9 nous en montre un exemplaire. Si une variable Y a une distribution normale avec moyenne μ et écart type σ, la probabilité d'obtenir une observation entre deux valeurs a et b est donnée par :

$$\Pr\{a \leq Y \leq b\} = \int_a^b \phi_{\mu,\sigma}(y)\,\mathrm{d}y.$$

On dénotera par ailleurs la fonction de répartition d'une variable normale avec moyenne μ et écart type σ par $\Phi_{\mu,\sigma}(y)$[11]. Comme le montre la figure 2.9, la moyenne μ est le centre de symétrie, alors que l'écart type σ est le point d'inflexion de la fonction de densité. La figure 2.10 nous montre différentes distributions normales. Le paramètre μ est une mesure de la tendance centrale (on dit aussi de la position) de la distribution. Le paramètre σ est une mesure de la variabilité (de la dispersion) de la distribution.

Pour une variable normale, les paramètres μ et σ ne constituent pas seulement un résumé de la distribution : ils caractérisent la distribution. En connaissant μ et σ, on connaît toute la distribution. En particulier, on connaît tous les

[11] Notons que cette fonction de répartition $\Phi_{\mu,\sigma}(y)$ ne peut pas s'écrire sous forme explicite (elle est définie implicitement comme étant l'intégrale de la densité $\phi_{\mu,\sigma}(y)$).

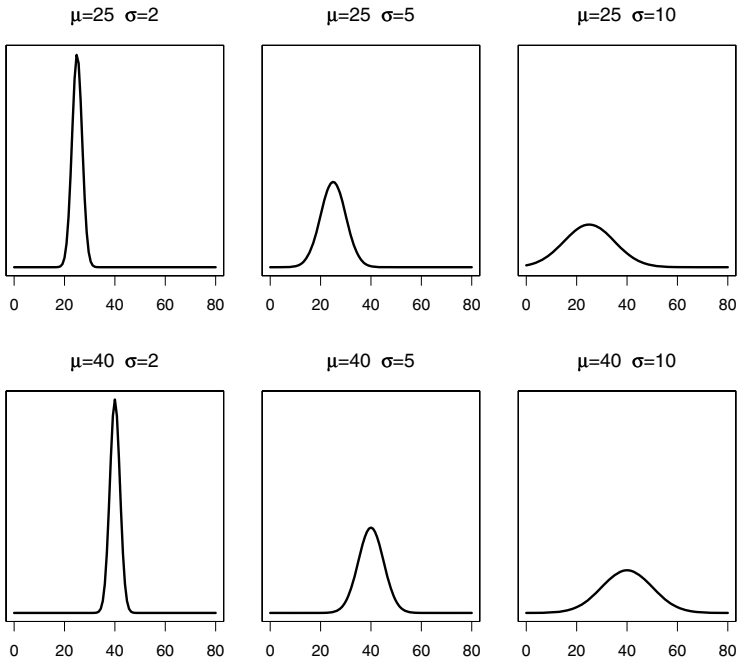

Figure 2.10 – Exemples de distributions normales avec différentes valeurs de μ et σ.

quantiles. La valeur $\mu + 2\sigma$ correspond par exemple au quantile 97.7 % d'une distribution normale. Par symétrie, la valeur $\mu - 2\sigma$ correspond au quantile 2.3 %. On aura par conséquent $97.7\ \% - 2.3\ \% = 95.4\ \%$ des observations dans l'intervalle $\mu \pm 2\sigma$.

> Pour une distribution normale, presque toutes les observations ($\approx 95\ \%$) sont comprises dans l'intervalle : moyenne $\pm$ 2 écarts types.

Ce résultat donne un sens précis au fait qu'il n'y a pas beaucoup de valeurs extrêmes dans une distribution normale. Attention, ce résultat n'est pas forcément valable pour d'autres distributions.

Exemple 2.6 *Pour la hauteur des Onobrychis, on avait une moyenne de* 21.7 *cm et un écart type de* 3.8 *cm. Si la distribution était normale, presque tous les Onobrychis devraient se trouver dans l'intervalle :*

$$21.7 \pm 2 \cdot 3.8 = [14.1; 29.3]\ \textit{cm.}$$

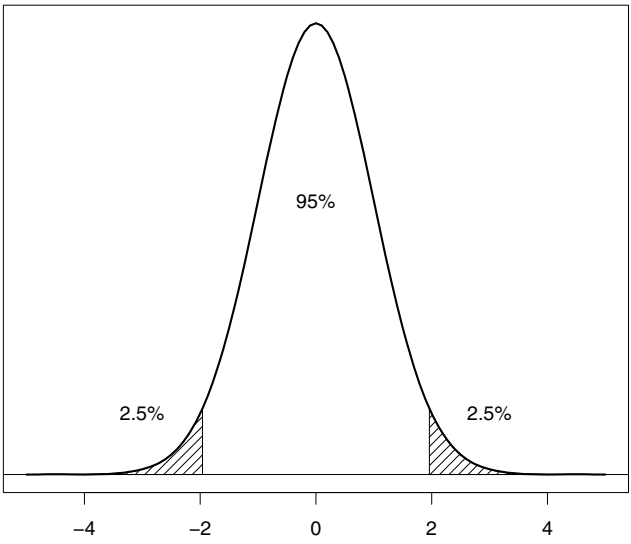

Figure 2.11 – Distribution normale standardisée ($\mu = 0$, $\sigma = 1$).

Or, ceci correspond à peu près aux données observées, ce qui est une indication que l'on est ici proche de la normalité (on confirme ainsi quantitativement les informations visuelles données par le boxplot et l'histogramme).

2.9 Distribution normale standardisée

Un changement d'unités (une transformation linéaire/affine) n'affecte pas la forme de la distribution. Si Y est une variable normale, il en sera de même pour la variable $a + bY$. Par ailleurs, si Y est une variable de moyenne μ et d'écart type σ, la *variable standardisée* $Z = (Y - \mu)/\sigma$ aura une moyenne 0 et un écart type 1. On a en effet :

- $mean(Z) = mean\left(\frac{Y-\mu}{\sigma}\right) = \frac{mean(Y)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$
- $stdev(Z) = stdev\left(\frac{Y-\mu}{\sigma}\right) = \frac{stdev(Y)}{\sigma} = \frac{\sigma}{\sigma} = 1.$

En particulier, si Y est normale de moyenne μ et d'écart type σ, la variable standardisée $Z = (Y - \mu)/\sigma$ sera normale de moyenne 0 et d'écart type 1. Il s'agit d'une *distribution normale standardisée* dont la densité est montrée dans la figure 2.11.

Dans ce texte, on notera par z_α le quantile α d'une distribution normale standardisée[12]. Le tableau A.1 (donné en annexe) nous permet de calculer ces quantiles. Elle contient les valeurs de la fonction de répartition d'une distribution normale standardisée, que l'on note $\Phi_{0,1}(x)$, pour de nombreuses valeurs de $x \geq 0$[13]. Par définition, on a donc $\Phi_{0,1}(z_\alpha) = \alpha$. Par symétrie, on a $\Phi_{0,1}(-x) = 1 - \Phi_{0,1}(x)$ et donc $z_{1-\alpha} = -z_\alpha$. On a ainsi par exemple :

$$\begin{array}{ccc} \Phi_{0,1}(1.645) = 0.95 & \text{et donc} & z_{0.95} = 1.645 \\ \Phi_{0,1}(1.96) = 0.975 & \text{et donc} & z_{0.975} = 1.96 \\ \Phi_{0,1}(-1.96) = 1 - \Phi_{0,1}(1.96) = 0.025 & \text{et donc} & z_{0.025} = -1.96 \\ \Phi_{0,1}(0) = 0.5 & \text{et donc} & z_{0.5} = 0 \end{array}$$

Plus généralement, le tableau A.1 nous permet de calculer les quantiles de n'importe quelle variable Y avec distribution normale de moyenne μ et d'écart type σ. Les valeurs tabulées nous indiquent à quel quantile correspond $\mu + x \cdot \sigma$ pour de nombreuses valeurs de x. Par exemple :

- **la valeur $\mu + 1.645 \cdot \sigma$ correspond au quantile 95 %**
- **la valeur $\mu + 1.96 \cdot \sigma$ correspond au quantile 97.5 %**.

Par symétrie, la valeur $\mu - 1.645 \cdot \sigma$ correspond au quantile 5 % et la valeur $\mu - 1.96 \cdot \sigma$ correspond au quantile 2.5 %. Ainsi, pour une distribution normale, 90 % des observations seront comprises dans l'intervalle $\mu \pm 1.645 \cdot \sigma$, alors que 95 % des observations seront comprises dans l'intervalle $\mu \pm 1.96 \cdot \sigma$. En pratique, un intervalle contenant 95 % des observations est souvent calculé en utilisant l'approximation $\mu \pm 2 \cdot \sigma$ (comme on le fera souvent au cours de ce texte).

Exemple 2.7 *Si on admet que la hauteur des Onobrychis est normalement distribuée avec une moyenne de* 21.7 *cm et un écart type de* 3.8 *cm, le quantile* 95 % *est donné par* $21.7 + 1.645 \cdot 3.8 = 28.0$ *cm et le quantile* 97.5 % *par* $21.7 + 1.96 \cdot 3.8 = 29.1$ *cm. Par symétrie, le quantile* 5 % *est donné par* $21.7 - 1.645 \cdot 3.8 = 15.4$ *cm et le quantile* 2.5 % *par* $21.7 - 1.96 \cdot 3.8 = 14.3$ *cm. On aura ainsi* 90 % *des Onobrychis dans l'intervalle* $[15.4; 28.0]$ *cm et* 95 % *des Onobrychis dans l'intervalle* $[14.3; 29.1]$ *cm.*

Le tableau A.1 nous permet ainsi de calculer la proportion des observations qui se trouvent à l'intérieur d'un intervalle donné (c'est-à-dire la probabilité

[12] Attention, dans certains ouvrages z_α désigne le quantile $1 - \alpha$ (et non le quantile α) d'une distribution normale standardisée.

[13] Ce tableau a été produit en utilisant la commande `pnorm` du logiciel statistique R. Ainsi `pnorm(1.96)` nous donne 0.975. Notons aussi que la fonction réciproque dans R est la commande `qnorm`. Ainsi `qnorm(0.975)` nous donne 1.96.

qu'une observation tirée au hasard de la population se trouve dans cet intervalle) si on peut supposer que ces observations proviennent d'une distribution normale. De cette manière, à partir de la moyenne et de l'écart type, et supposant la normalité de la variable, le lecteur d'un abstract de publication scientifique pourra reconstruire toute la distribution, sans pour autant connaître le détail des données. On aura par exemple :

38.3 %	des observations dans l'intervalle	$\mu \pm 0.5 \cdot \sigma$
50.0 %	des observations dans l'intervalle	$\mu \pm 0.675 \cdot \sigma$
68.3 %	des observations dans l'intervalle	$\mu \pm 1 \cdot \sigma$
86.6 %	des observations dans l'intervalle	$\mu \pm 1.5 \cdot \sigma$
90.0 %	des observations dans l'intervalle	$\mu \pm 1.645 \cdot \sigma$
95.0 %	des observations dans l'intervalle	$\mu \pm 1.96 \cdot \sigma$
95.4 %	des observations dans l'intervalle	$\mu \pm 2 \cdot \sigma$
99.0 %	des observations dans l'intervalle	$\mu \pm 2.58 \cdot \sigma$
99.7 %	des observations dans l'intervalle	$\mu \pm 3 \cdot \sigma$
99.9 %	des observations dans l'intervalle	$\mu \pm 3.29 \cdot \sigma$
99.99 %	des observations dans l'intervalle	$\mu \pm 4 \cdot \sigma$

Plus généralement, on aura (α dénotant la probabilité qu'une observation ne se trouve pas dans l'intervalle) :

$1 - \alpha$	des observations dans l'intervalle	$\mu \pm z_{1-\alpha/2} \cdot \sigma.$

Ces résultats sont connus plus ou moins par cœur par un statisticien. Ils permettent d'une part d'interpréter la valeur d'un écart type dans le cadre d'une distribution normale. Par exemple, $2 \cdot 0.675 = 1.35$ écart type correspond à l'intervalle interquartile (la longueur de la boîte du boxplot). D'autre part, 4 écarts types représentent en pratique (c'est-à-dire « à 95 % ») une mesure de l'étendue d'une variable normale (bien qu'en théorie l'étendue serait infinie). Dans cette optique, un écart type correspond en gros à un quart de l'étendue.

Dans le cadre d'une distribution normale, on mentionnera également qu'une différence de plus de 3 écarts types par rapport à la moyenne sera rarement observée, alors qu'une différence de plus de 4 écarts types par rapport à la moyenne ne sera pratiquement jamais observée. On laissera par ailleurs au lecteur le soin de vérifier (comme exercice) que selon un boxplot, une observation d'une distribution normale sera considérée comme une valeur extrême si elle se trouve à plus de 2.7 écarts types de la moyenne. Il y en aura à peu près 7 sur 1000.

On terminera cette section par insister sur le fait que les concepts de moyenne et d'écart type (ou de variance) sont surtout utilisés dans le cadre d'une distribution normale. Pour d'autres distributions continues, on préférera souvent la médiane à la moyenne, et parfois l'intervalle interquartile à l'écart type (dont l'interprétation hors de la normalité est plus obscure).

2.10 Variable standardisée et qq-plot

Lorsque l'on standardise une variable, en lui soustrayant d'abord sa moyenne (c'est-à-dire en la centrant) puis en la divisant par son écart type, les unités que l'on obtient sont des *nombres d'écarts types par rapport à la moyenne* (en anglais : z-`scores` ou `standard deviation scores`). Il s'agit de l'unité du statisticien. Son interprétation est facilitée si on peut supposer la normalité car la notion d'écart type est bien connue dans un cadre normal.

Exemple 2.8 *Pour les Onobrychis, les observations standardisées* $z_i = (y_i - 21.7)/3.8$ *sont données ci-dessous (par ordre croissant) :*

−2.03	−2.03	−1.50	−1.24	−1.24	−0.97	−0.97	−0.97	−0.97	−0.97
−0.97	−0.97	−0.71	−0.71	−0.71	−0.71	−0.71	−0.71	−0.45	−0.45
−0.45	−0.45	−0.45	−0.45	−0.45	−0.45	−0.18	−0.18	−0.18	−0.18
−0.18	−0.18	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.34
0.34	0.34	0.34	0.34	0.34	0.61	0.61	0.87	0.87	1.13
1.13	1.13	1.39	1.66	1.66	1.66	1.92	1.92	1.92	2.71

Ces valeurs ne correspondent plus à des centimètres, mais à des écarts types au-dessous ou au-dessus de la moyenne. Le plus petit Onobrychis se trouve par exemple à 2.03 *écarts types au-dessous de la moyenne, alors que le plus grand d'entre eux se trouve à* 2.71 *écarts types au-dessus de la moyenne. À première vue, cette nouvelle unité peut paraître moins « parlante » que l'unité originale. Pourtant, pour un non-spécialiste des Onobrychis, une valeur de* 32 *cm ne suggérera pas forcément grand chose, alors qu'une valeur de* 2.71 *écarts types au-dessus de la moyenne, pour autant que l'on puisse supposer la normalité, sera parlante pour un statisticien. Des résultats standardisés de la sorte faciliteront également certaines comparaisons. Il ne serait par exemple pas évident de juger si un Onobrychis mesurant* 30 *cm après six mois est plus ou moins extrême qu'un Onobrychis mesurant* 60 *cm après douze mois ; le jugement deviendra plus facile si les mesures sont exprimées en nombres d'écarts types par rapport à la moyenne.*

Les observations standardisées peuvent être utilisées dans un `qq-plot` (ou `quantile-quantile plot`, connue en français sous le nom de *droite de Henry*), une technique graphique pour vérifier la normalité. Par exemple, $n = 10$ observations standardisées d'une variable normale devraient à peu près coïncider avec les quantiles :

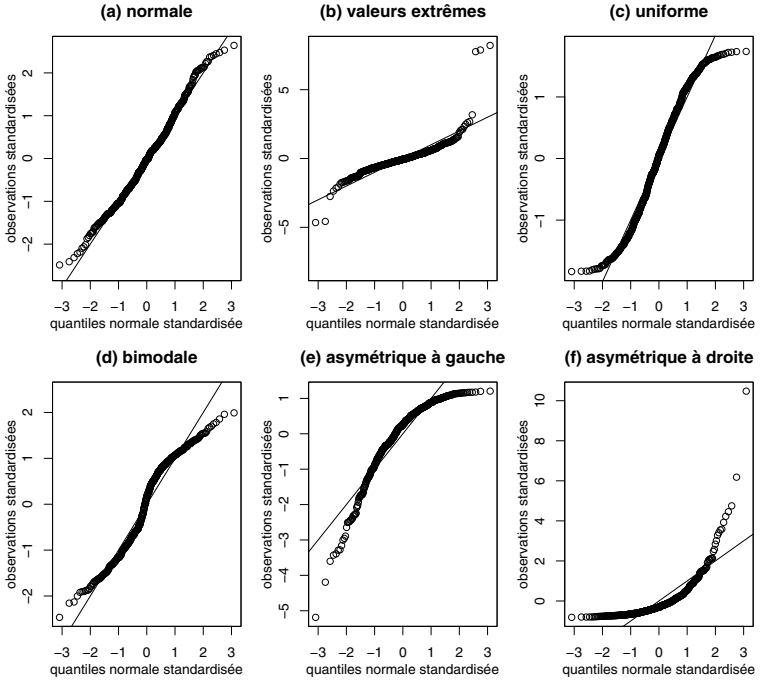

Figure 2.12 – Exemples de qq-plots pour différentes distributions continues.

5 %	15 %	25 %	35 %	45 %	55 %	65 %	75 %	85 %	95 %

d'une variable normale standardisée, c'est-à-dire avec (selon le tableau A.1) :

−1.645	−1.035	−0.675	−0.385	−0.135	0.135	0.385	0.675	1.035	1.645

Plus généralement, n observations standardisées (et rangées par ordre croissant) d'une variable normale devraient à peu près coïncider avec les quantiles $(i - 0.5)/n$ d'une distribution normale standardisée (pour $i = 1, \cdots, n$). Un *qq-plot* est un graphique nous montrant ces observations standardisées en fonction de ces quantiles d'une distribution normale standardisée. Si une variable est proche de la normalité, les points sur ce graphique seront à peu près alignés sur la droite identité.

La figure 2.12 nous montre des qq-plots obtenus à partir de $n = 500$ observations (standardisées) provenant des densités de probabilité montrées dans la figure 2.4. Seul le graphique (a), où la distribution est effectivement normale, nous montre un bel alignement. Les autres graphiques nous montrent à quoi ressemble un qq-plot pour différentes distributions non normales.

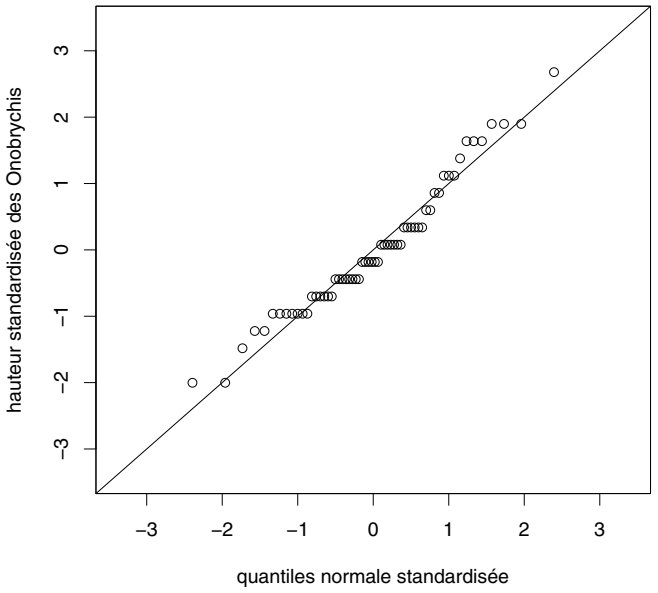

Figure 2.13 – Exemple de qq-plot pour un ensemble d'observations.

Exemple 2.9 *La figure 2.13 nous montre le qq-plot pour la hauteur des Onobrychis où l'on voit un relativement bel alignement, signe de normalité.*

2.11 Mesures de non-normalité

Certaines méthodes statistiques (dont la moyenne et l'écart type) sont avant tout utilisées dans le cadre d'une distribution normale. Il est donc important de pouvoir reconnaître si une variable continue est à peu près normale ou non. On a vu que ceci peut se faire à l'aide d'un boxplot ou d'un qq-plot. On va voir ici comment aborder cette question de façon quantitative, en introduisant des mesures de non-normalité.

On rappelle que deux caractéristiques importantes de la normalité sont la symétrie de la distribution et le fait qu'il y a peu de valeurs extrêmes. Pour quantifier ceci, on va introduire une mesure d'asymétrie (en anglais : `skewness`) et une mesure d'aplatissement (en anglais : `kurtosis`) d'une variable Y, définies de la manière suivante :

- **asymétrie** :

$$skewness(Y) = \frac{mean((Y - mean(Y))^3)}{stdev^3(Y)}$$

- **aplatissement** :

$$kurtosis(Y) = \frac{mean((Y - mean(Y))^4)}{stdev^4(Y)} - 3.$$

Le numérateur de $skewness(Y)$ est la moyenne des déviations au cube par rapport à la moyenne. Dans le cas d'une distribution asymétrique vers la droite, les larges déviations positives seront amplifiées par le cube et la mesure d'asymétrie sera positive. Dans le cas d'une distribution asymétrique vers la gauche, les larges déviations négatives seront amplifiées par le cube et la mesure d'asymétrie sera négative. Dans le cas d'une distribution symétrique, par exemple pour une distribution normale, les déviations positives compenseront les déviations négatives et la mesure d'asymétrie sera nulle. La division par le cube de l'écart type au dénominateur implique par ailleurs :

$$skewness(a + bY) = signe(b) \cdot skewness(Y)$$

autrement dit que cette mesure d'asymétrie est invariante (au signe près) lorsque l'on change d'unités.

Le numérateur du premier terme de $kurtosis(Y)$ est la moyenne des déviations à la puissance quatre par rapport à la moyenne. Cette mesure d'aplatissement sera donc grandement positive en présence de valeurs extrêmes, qui ont par définition de larges déviations positives ou négatives par rapport à la moyenne, et qui seront amplifiées par la puissance quatre. La division par l'écart type à la puissance quatre au dénominateur implique par ailleurs :

$$kurtosis(a + bY) = kurtosis(Y)$$

autrement dit que cette mesure d'aplatissement est invariante lorsque l'on change d'unités. En outre, cette mesure d'aplatissement sera nulle pour une distribution normale grâce à la soustraction au premier terme de la valeur 3. Ainsi, une distribution avec un aplatissement positif aura plus de valeurs extrêmes qu'une distribution normale, alors qu'une distribution avec un aplatissement négatif aura moins de valeurs extrêmes qu'une distribution normale.

Si Y est normale, on aura $skewness(Y) = 0$ et $kurtosis(Y) = 0$.

Précisons toutefois qu'une distribution normale est un concept théorique. Dans la réalité, on n'observera jamais de variable exactement normale. En pratique, il s'agira donc de vérifier qu'une variable est *approximativement normale*, non pas qu'elle est exactement normale. De façon pragmatique, on pourra décider qu'une variable est approximativement normale si les mesures d'asymétrie et d'aplatisement sont comprises entre -0.5 et $+0.5$. Dans le cas contraire,

on essaiera de transformer la variable afin de la rendre approximativement normale. Si on hésite entre deux transformations, on choisira la transformation pour laquelle ces deux mesures de non-normalité sont les plus proches de zéro[14].

Exemple 2.10 *Pour la hauteur Y des Onobrychis, on calcule $skewness(Y) = 0.49$ et $kurtosis(Y) = -0.15$, nous suggérant une fois encore une normalité approximative.*

2.12 Transformation logarithmique

Nous venons de rappeler qu'il est parfois souhaitable de transformer les données dans le but de nous approcher de la normalité afin de faciliter l'analyse statistique. Une transformation qui fonctionne souvent en pratique lorsque la distribution est asymétrique vers la droite, et qui offre certains avantages d'interprétation, est la transformation logarithmique (on parlera de variable « log-transformée »). On a par exemple :

- **log base 2** : +1 unité sur échelle logarithmique = multiplication par 2 sur échelle originale
- **log base 10** : +1 unité sur échelle logarithmique = multiplication par 10 sur échelle originale.

En général, on utilisera pourtant le logarithme naturel (base $e = 2.718...$), comme on le fera systématiquement dans ce texte (la notation « log » dénotant ainsi le logarithme naturel dans ce qui suit). Une variable Y pour laquelle $\log(Y)$ est normale est dite une *variable log-normale*. En utilisant l'approximation $\log(x+1) \approx x$ (excellente pour $x \leq 0.25$), on aura alors :

$$\log(a) - \log(b) = \log(a/b) = \log\left(1 + \frac{a-b}{b}\right) \approx \frac{a-b}{b}.$$

[14]On citera les résultats suivants pour une variable Y avec distribution (a) normale (avec paramètres μ et σ), (b) continue uniforme (entre a et b), (c) binomiale (avec paramètres n et π), (d) de Student (avec dl degrés de liberté), et (e) du khi-deux (avec dl degrés de liberté) :

	$mean(Y)$	$variance(Y)$	$skewness(Y)$	$kurtosis(Y)$
(a)	μ	σ^2	0	0
(b)	$(a+b)/2$	$(b-a)^2/12$	0	-1.2
(c)	$n\pi$	$n\pi(1-\pi)$	$(1-2\pi)/\sqrt{n\pi(1-\pi)}$	$(1-6\pi(1-\pi))/(n\pi(1-\pi))$
(d)	0	$dl/(dl-2)$	0	$6/(dl-4)$
(e)	dl	$2 \cdot dl$	$\sqrt{8/dl}$	$12/dl$

Pour une distribution de Student, ces quatre résultats sont valables si respectivement $dl > 1$, $dl > 2$, $dl > 3$ et $dl > 4$. On aura sinon $mean(Y) = \infty$ si $dl \leq 1$, $variance(Y) = \infty$ si $dl \leq 2$, $skewness(Y) = \infty$ si $dl \leq 3$ et $kurtosis(Y) = \infty$ si $dl \leq 4$. Les distributions avec une mesure d'aplatissement positive sont dites *leptokurtiques*, celles avec une mesure d'aplatissement négative sont dites *platikurtiques*. Par exemple, les distributions de Student sont leptokurtiques alors qu'une distribution uniforme est platikurtique.

Ceci nous indique qu'appliquer une transformation logarithmique revient (en un sens) à exprimer les différences entre observations de façon relative plutôt qu'absolue. Si on compare par exemple deux *Onobrychis* de 24 cm et de 25 cm, leur différence absolue est de 1 cm, alors que leur différence relative est de $\log(25) - \log(24) = 3.22 - 3.18 = 0.04 \approx (25 - 24)/24$.

Il existe de nombreux domaines où l'on préfère exprimer des différences de façon relative plutôt que de façon absolue (exemple : on parlera d'une augmentation d'une prime d'assurances de 10 % plutôt que de 200 euros). Ce sera le cas pour des variables avec des valeurs possibles strictement positives, où la valeur zéro représente en général non seulement une valeur de référence, mais également un minimum absolu (en principe inatteignable)[15]. Pour de telles variables, une analyse statistique se fera le plus naturellement du monde sur l'échelle logarithmique.

Considérons à présent le cas d'une variable Y asymétrique, que l'on transforme de telle sorte que $\log(Y)$ soit à peu près symétrique. On aura ainsi :

$$mean(\log(Y)) \approx median(\log(Y)).$$

Comme on a toujours par ailleurs :

$$median(\log(Y)) = \log(median(Y))$$

on obtiendra finalement :

$$mean(\log(Y)) \approx \log(median(Y))$$

ce qui implique :

$$\exp(mean(\log(Y))) \approx median(Y).$$

On a vu que pour une variable asymétrique, la médiane est souvent plus intéressante que la moyenne car elle est représentative de l'ensemble des données. On voit ici que la quantité $\exp(mean(\log(Y)))$ est une approximation de la médiane de Y, qui a l'avantage d'être calculée à partir d'une moyenne (la moyenne de $\log(Y)$). On combine ainsi la représentativité de la médiane avec la finesse de la moyenne, ce qui n'est pas rien ! Cette quantité est aussi connue sous le nom de *moyenne géométrique*[16] :

$$\exp(mean(\log(Y))) = \exp\left(\frac{1}{n}\sum_i \log(y_i)\right) = \sqrt[n]{\prod_i y_i}.$$

Ainsi, lorsqu'on calcule une moyenne géométrique plutôt qu'une moyenne arithmétique pour mesurer la tendance centrale d'un ensemble de données, on reconnaît implicitement que l'échelle naturelle pour analyser ces données est l'échelle logarithmique, non l'échelle originale.

[15] À l'inverse, il y aura peu de sens à calculer des différences relatives (ou des quotients) impliquant une variable dont le minimum absolu est autre que zéro. Il y a par exemple peu de sens à dire qu'une température de 30 degrés Celsius est trois fois supérieure à une température de 10 degrés Celsius ; car que dire alors d'une température de -10 degrés Celsius ?

[16] La notation $\prod_i y_i$ désigne le produit des observations : $\prod_i y_i = y_1 \cdot y_2 \cdot \cdots \cdot y_n$.

Un autre résultat intéressant est le suivant. En utilisant la formule de Taylor (bien connue des mathématiciens), on obtient[17] :

$$variance(\log(Y)) \approx \frac{variance(Y)}{mean^2(Y)}$$

c'est-à-dire :

$$stdev(\log(Y)) \approx varcoef(Y).$$

Ainsi, lorsqu'on calcule un coefficient de variation plutôt qu'un écart type pour mesurer la variabilité d'un ensemble de données, on reconnaît implicitement que l'échelle naturelle pour analyser ces données est l'échelle logarithmique, non l'échelle originale.

Exemple 2.11 *On introduira plus en détail au chapitre 5 un ensemble de données où l'on a mesuré le taux de créatine dans le sang pour* $n = 31$ *femmes atteintes d'une maladie génétique. Sur l'échelle logarithmique, où l'on sera proche de la normalité, on a une moyenne de* 4.71 *et un écart type de* 0.91*. Cela suggère sur l'échelle originale un taux de créatine médian de* $\exp(4.71) = 111.1$ *et un coefficient de variation de* 0.91*.*

2.13 Distribution d'une variable binaire

Une variable binaire Y, avec ses deux valeurs possibles que l'on notera 1 et 0, est un cas particulier de variable qualitative. Comme il y a deux valeurs possibles, la distribution d'une variable binaire est caractérisée par une seule proportion. Cependant, une variable binaire peut aussi être vue comme un cas particulier de variable quantitative (mesurée sur une échelle où l'une des valeurs possibles vaut 0 et la distance entre les deux valeurs possibles vaut 1). On peut dès lors calculer sa moyenne et sa variance comme suit :

- $mean(Y)$ est la proportion de 1
- $variance(Y) = mean(Y^2) - mean^2(Y) = mean(Y)(1 - mean(Y))$.

Exemple 2.12 *Si* Y *dénote le sexe d'un nouveau-né (avec* $Y = 1$ *pour les garçons et* $Y = 0$ *pour les filles) et si sur* 1000 *naissances on observe* 527 *garçons et* 473 *filles, on aura* $mean(Y) = 0.527$ *(et donc une proportion de* 52.7 % *de garçons) et* $variance(Y) = 0.527 \cdot 0.473 = 0.249$*.*

[17] En appliquant la formule de Taylor :

$$g(Y) \approx g(a) + (Y - a)g'(a)$$

avec $a = mean(Y)$, et en utilisant les propriétés élémentaires de la variance, on obtient le résultat général suivant :

$$variance(g(Y)) \approx variance(Y) \cdot [g'(mean(Y))]^2.$$

La formule ci-dessus s'obtient donc avec $g(Y) = \log(Y)$ (et $g'(mean(Y)) = 1/mean(Y)$).

Notons que l'on aura $variance(Y) = 0$ si $mean(Y) = 0$ ou si $mean(Y) = 1$, c'est-à-dire si toutes les observations sont identiques (toutes égales à 0 ou toutes égales à 1, ce qui correspond effectivement à aucune variabilité). Par ailleurs, la variance d'une variable binaire ne pourra dépasser la valeur maximale de 0.25, qui sera atteinte si $mean(Y) = 0.5$. Lors d'une élection entre deux candidats, la variabilité des avis dans la population des votants sera ainsi maximale (et égale à 0.25) si les deux candidats réunissent chacun 50 % des suffrages.

Une proportion peut ainsi être vue comme un cas particulier de moyenne (la moyenne d'une variable binaire), ce qui aura son importance dans ce qui suit. Par ailleurs, la variance d'une variable binaire est une fonction de sa moyenne et n'apporte aucune information supplémentaire sur sa distribution (en connaissant la moyenne, on connaît automatiquement la variance).

On ne résume pas les données d'une variable binaire par une moyenne et un écart type mais seulement par une proportion.

Chapitre 3

Estimation

Les caractéristiques numériques (moyenne, variance, proportion, etc.) introduites au chapitre précédent nous permettent de résumer une distribution. En pratique, on les calcule sur les données de notre échantillon. En théorie, on pourrait aussi les calculer sur les données de la population, si seulement elles étaient disponibles. Ainsi, bien que l'on calcule en pratique la *moyenne de l'échantillon*, on peut concevoir l'existence de la *moyenne de la population*, et c'est d'ailleurs à cette dernière que l'on s'intéresse en premier lieu. On adopte alors le point de vue suivant :

- les caractéristiques de la population sont dites les « véritables » caractéristiques, souvent appelées *paramètres*
- les caractéristiques de l'échantillon sont vues comme des *estimateurs* (ou des *estimations*) de ces paramètres.

Afin de distinguer entre paramètres et estimateurs, on utilisera des « chapeaux » pour les estimateurs. Une caractéristique $\widehat{\theta}$ calculée dans l'échantillon est donc un estimateur d'un paramètre θ défini sur la population. Pour une variable continue, on utilise les notations suivantes :

- μ dénote la moyenne de la population et $\widehat{\mu}$ la moyenne de l'échantillon (on dira aussi moyenne empirique)
- σ^2 dénote la variance de la population et $\widehat{\sigma}^2$ la variance de l'échantillon (on dira aussi variance empirique).

Pour une variable binaire (avec codage 1/0), on utilise les notations suivantes :

- π dénote la proportion de 1 dans la population et $\widehat{\pi}$ la proportion de 1 dans l'échantillon (on dira aussi proportion empirique).

Ainsi, $\widehat{\mu}$ est un estimateur de μ, $\widehat{\sigma}^2$ est un estimateur de σ^2, et $\widehat{\pi}$ est un estimateur de π.

3.1 Distribution d'un estimateur

Voici à présent le moment clé de ce texte d'introduction à la statistique. L'estimateur $\widehat{\theta}$ d'un paramètre θ se calcule en pratique dans un seul échantillon. On aura donc une seule valeur de $\widehat{\theta}$. Cependant, si on répétait l'expérience, c'est-à-dire si on tirait d'autres échantillons de la même population, on pourrait calculer cet estimateur $\widehat{\theta}$ dans chacun de ces échantillons et on obtiendrait d'autres valeurs de $\widehat{\theta}$ qui seraient différentes les unes des autres. Autrement dit, l'estimateur $\widehat{\theta}$ varierait d'un échantillon à l'autre. Avec un peu d'imagination, on peut ainsi considérer un estimateur $\widehat{\theta}$ comme étant une variable. Ceci constitue le « paradigme de la statistique ».

> Paradigme statistique : un estimateur $\widehat{\theta}$ est vu comme une variable.

Cette variable $\widehat{\theta}$ est définie sur une « population d'échantillons » (de même taille que l'échantillon original et obtenus en utilisant la même technique d'échantillonnage). On pourra ainsi parler de la distribution d'un estimateur. Afin de résumer cette distribution, on appliquera les techniques de statistique descriptive vues au chapitre précédent et on pourra ainsi définir la moyenne, la variance ou l'écart type de $\widehat{\theta}$. On utilise les notations suivantes :

- $\mathrm{E}(\widehat{\theta})$ dénote la moyenne de $\widehat{\theta}$, que l'on appellera aussi son *espérance* (en anglais : `expectation`)
- $\mathrm{SE}(\widehat{\theta})$ dénote l'écart type de $\widehat{\theta}$, que l'on appellera aussi son *erreur type* (en anglais : `standard error`)
- $\mathrm{Var}(\widehat{\theta})$ dénote la variance de $\widehat{\theta}$.

Alors que les variables discutées dans le chapitre précédent (telle la hauteur d'un *Onobrychis*) peuvent s'observer sur plusieurs individus d'une population, une « variable estimateur » $\widehat{\theta}$ a ceci de particulier qu'elle ne s'observera qu'une seule fois (vu que l'on ne disposera en général que d'un seul échantillon de cette population d'échantillons). Dans ces conditions, on aura recours à des arguments théoriques (mathématiques, calcul des probabilités) ou à des simulations pour connaître sa distribution.

3.2 Variable aléatoire

Les probabilistes appellent ces variables qui ne s'observent (ou *se réalisent*) qu'une seule fois des *variables aléatoires*. Un estimateur est donc une variable aléatoire. Plus généralement, toute quantité calculable sur un échantillon peut

être vue comme une variable aléatoire. Cela est aussi valable pour chaque observation y_i d'une variable Y, par exemple pour la première observation y_1 d'un échantillon. En pratique, on observe un seul y_1 (dans notre exemple, le premier *Onobrychis* mesure 21 cm), mais on en observerait d'autres si on répétait l'expérience (on aurait un premier *Onobrychis* dans chaque échantillon).

Il sera utile dans ce qui suit de faire la distinction entre le concept et la réalisation d'une variable aléatoire en utilisant une lettre majuscule pour le premier et une lettre minuscule pour la seconde[1]. On notera ainsi Y_1 la variable aléatoire « première observation de l'échantillon » et y_1 sa réalisation dans notre échantillon. On dira que l'observation $y_1 = 21$ cm est la réalisation de la variable aléatoire Y_1. Si les individus de notre échantillon sont tirés au hasard d'une population dans laquelle une variable Y a une moyenne μ et une variance σ^2, on aura pour chaque observation Y_i :

- $\mathrm{E}(Y_i) = \mu$
- $\mathrm{Var}(Y_i) = \sigma^2$.

On peut ainsi attribuer à chaque observation de l'échantillon les paramètres de la population. Plus généralement, on peut attribuer à chaque observation Y_i de l'échantillon la distribution de la variable Y dans la population. Par exemple, la distribution de Y_i sera normale si la distribution de Y est normale.

On peut par ailleurs définir formellement le concept d'indépendance entre observations comme suit : deux observations Y_1 et Y_2 sont dites indépendantes si la distribution de Y_2, sachant que l'on observe $Y_1 = y_1$, ne dépend pas de la valeur observée y_1. Dans notre exemple des *Onobrychis*, ceci veut dire que si on répète l'échantillonnage, la distribution de la hauteur du second *Onobrychis* Y_2 calculée sur les échantillons où l'on observe un premier *Onobrychis* de $y_1 = 21$ cm sera la même que la distribution de Y_2 calculée sur les échantillons où l'on observe $y_1 = 25$ cm, et sera encore la même que la distribution de Y_2 calculée sur les échantillons où l'on observe $y_1 = 27.4536$ cm, et ainsi de suite[2].

Notons que la somme, la différence ou le produit de deux variables aléatoires Y_1 et Y_2 sont de nouvelles variables aléatoires. On a alors les résultats suivants :

- Y_1 et Y_2 normales implique $Y_1 + Y_2$ et $Y_1 - Y_2$ normales
- $\mathrm{E}(Y_1 + Y_2) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2)$ et $\mathrm{E}(Y_1 - Y_2) = \mathrm{E}(Y_1) - \mathrm{E}(Y_2)$
- $\mathrm{E}(Y_1 \cdot Y_2) = \mathrm{E}(Y_1) \cdot \mathrm{E}(Y_2)$ (si Y_1 et Y_2 sont indépendantes).

[1] On renoncera toutefois à une telle utilisation de minuscules et de majuscules avec les lettres grecques. Ainsi, $\widehat{\theta}$ dénotera dans ce qui suit à la fois le concept d'estimateur vu comme variable aléatoire et sa réalisation dans notre échantillon. On fera par contre (en principe) une distinction linguistique, en appelant *estimateur* le premier et *estimation* la seconde.

[2] Un lecteur attentif nous rappellera qu'il est impossible d'observer un *Onobrychis* mesurant exactement 21 cm, ni 25 cm, ni même 27.4536 cm. En fait, un probabiliste nous dira que cela arrivera quand même de temps en temps, mais que cela arrivera avec une probabilité nulle (la proportion des *Onobrychis* mesurant exactement 21 cm étant nulle par rapport à l'infinité des *Onobrychis* de la population considérée).

Les choses ne seront cependant pas toujours aussi simples, par exemple $\mathrm{E}(Y_1/Y_2)$ sera différent de $\mathrm{E}(Y_1)/\mathrm{E}(Y_2)$ (l'espérance d'un quotient ne sera pas égale au quotient des espérances). On a par ailleurs le résultat fondamental suivant[3] :

- $\mathrm{Var}(Y_1 + Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$ (si Y_1 et Y_2 sont indépendantes).

> La variance de la somme de variables aléatoires indépendantes est égale à la somme de leurs variances.

En rappelant que $\mathrm{Var}(-Y) = \mathrm{Var}(Y)$, on en déduit également :

- $\mathrm{Var}(Y_1 - Y_2) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2)$ (si Y_1 et Y_2 sont indépendantes).

Notons que l'écart type ne possède pas cette propriété. L'écart type de la somme de deux variables aléatoires indépendantes sera en effet plus petit que la somme de leurs écarts types.

3.3 Distribution de la moyenne d'un échantillon

On peut à présent calculer l'espérance de la variable estimateur $\widehat{\mu}$. Si on dénote par $Y_1, Y_2, \cdots, Y_n$ les observations de l'échantillon, on obtient en généralisant un résultat vu dans la section précédente :

$$\mathrm{E}\left(\sum_i Y_i\right) = \sum_i \mathrm{E}(Y_i).$$

On en déduit :

$$\mathrm{E}(\widehat{\mu}) = E\left(\frac{\sum_i Y_i}{n}\right) = \frac{\sum_i \mathrm{E}(Y_i)}{n} = \frac{n\mu}{n} = \mu.$$

Ce résultat n'est pas surprenant. Il nous dit que $\widehat{\mu}$ serait en moyenne égal à μ si on répétait l'expérience. On dit alors que $\widehat{\mu}$ est un *estimateur sans biais* (ou non biaisé) de μ.

[3] Illustrons ce résultat avec un exemple : on considère deux variables Y_1 et Y_2 quantitatives discrètes, la première avec une distribution uniforme sur l'ensemble $\{1, 2, 6\}$ (c'est-à-dire un tiers de 1, un tiers de 2 et un tiers de 6), la seconde avec une distribution uniforme sur l'ensemble $\{3, 6, 12\}$. Le lecteur pourra vérifier que l'on a ici $\mathrm{Var}(Y_1) = 14/3$ et $\mathrm{Var}(Y_2) = 14$. Si ces deux variables sont indépendantes, la variable $Y_1 + Y_2$ aura une distribution uniforme sur l'ensemble $\{4, 5, 7, 8, 9, 12, 13, 14, 18\}$ (constitué de toutes les sommes possibles entre un élément du premier et un élément du second ensemble). Il se trouve alors que la variance de cet ensemble, c'est-à-dire $\mathrm{Var}(Y_1 + Y_2)$, est égale à $56/3$, c'est-à-dire à la somme $14/3 + 14$. Notons que l'on a également $\mathrm{E}(Y_1) = 3$, $\mathrm{E}(Y_2) = 7$ et $\mathrm{E}(Y_1 + Y_2) = 10$ $(3 + 7)$. Par contre, la médiane de $Y_1 + Y_2$ (qui vaut 9) n'est pas égale à la somme des médianes de Y_1 et de Y_2 (qui valent 2 et 6) ; certaines propriétés remarquables de la moyenne et de la variance ne sont pas partagées par d'autres paramètres.

En moyenne, la moyenne de l'échantillon est égale à la moyenne de la population.

Autrement dit, bien que la moyenne de notre échantillon soit en général différente de la moyenne de la population, il est bon de savoir qu'en moyenne on viserait juste si on répétait l'expérience (on ne serait pas systématiquement en dessous ni systématiquement en dessus).

Calculons à présent la variance de $\widehat{\mu}$. Pour autant que les observations de l'échantillon soient indépendantes, on obtient en généralisant un résultat vu dans la section précédente :

$$\mathrm{Var}\left(\sum_i Y_i\right) = \sum_i \mathrm{Var}(Y_i).$$

On en déduit[4] :

$$\mathrm{Var}\,(\widehat{\mu}) = Var\left(\frac{\sum_i Y_i}{n}\right) = \frac{\sum_i \mathrm{Var}(Y_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Il s'agit d'une nouvelle formule fondamentale de la statistique que nous encadrerons sous la forme équivalente suivante :

$$\mathrm{SE}(\widehat{\mu}) = \frac{\sigma}{\sqrt{n}}.$$

Un estimateur avec une petite erreur type (une petite variance) est un estimateur qui ne varierait pas beaucoup d'un échantillon à l'autre si on répétait l'expérience. L'estimation dépend donc peu de l'échantillon particulier dans lequel elle est calculée. Si l'estimateur est par ailleurs non biaisé, l'estimation sera alors proche du paramètre que l'on désire estimer. Autrement dit, une petite erreur type (une petite variance) pour un estimateur non biaisé est un gage de précision.

Un estimateur est précis s'il est non biaisé et si son erreur type est petite.

[4]On notera que si la taille N de la population est finie et si les n observations (avec $n \leq N$) sont tirées au hasard de cette population mais « sans remise », alors les observations ne seront pas tout à fait indépendantes et on aura $\mathrm{Var}(\widehat{\mu}) = (\sigma^2/n) \cdot (N-n)/(N-1)$. Le facteur correctif $(N-n)/(N-1)$ sera cependant négligeable (c'est-à-dire proche de 1) si la taille N de la population est beaucoup plus grande que la taille n de l'échantillon.

Un estimateur non biaisé est donc d'autant plus précis que son erreur type (sa variance) est petite. La formule ci-dessus nous indique que la précision de l'estimateur $\widehat{\mu}$ dépend de deux facteurs : la racine carrée de la taille n de l'échantillon dans lequel il est calculé et la variabilité, mesurée par l'écart type σ, dans la population de laquelle l'échantillon provient (et dont on veut connaître la moyenne μ). Ainsi :

- **l'estimateur est d'autant plus précis que la taille de l'échantillon est grande**

  → logique : le hasard de l'échantillonnage aura moins d'impact dans un grand échantillon que dans un petit échantillon

- **l'estimateur est d'autant plus précis que la variabilité dans la population est petite**

  → logique : le hasard de l'échantillonnage aura moins d'impact si les individus de la population de laquelle provient l'échantillon se ressemblent, que s'ils sont très différents les uns des autres.

Notons que ce résultat est un résultat théorique. En pratique, on ne peut pas calculer l'erreur type de $\widehat{\mu}$ puisqu'elle dépend de σ (que l'on ne connaît pas plus que l'on ne connaît μ). On aura ainsi deux bonnes raisons d'estimer σ. Cela nous donnera d'une part une estimation de la variabilité dans la population, ce qui est intéressant en soi, et cela nous donnera d'autre part une estimation de l'erreur type (et donc de la précision) de l'estimateur $\widehat{\mu}$.

On connaît à présent l'espérance et la variance de $\widehat{\mu}$. Les résultats suivants complètent notre connaissance de la distribution de $\widehat{\mu}$:

- $\widehat{\mu} = \frac{1}{n}\sum_i Y_i$ sera normale si Y_i est normale
- $\widehat{\mu} = \frac{1}{n}\sum_i Y_i$ sera approximativement normale même si Y_i est non normale (pour autant que n soit suffisamment grand).

Le premier de ces résultats se déduit directement d'autres résultats déjà mentionnés. Le second de ces résultats est fameux et connu sous le nom de *théorème central limite*[5]. On a ainsi le résultat général suivant :

La variable standardisée $\dfrac{\widehat{\mu} - \mathrm{E}(\widehat{\mu})}{\mathrm{SE}(\widehat{\mu})} = \sqrt{n} \cdot \dfrac{\widehat{\mu} - \mu}{\sigma}$ aura une distribution normale standardisée si n est suffisamment grand.

Notons qu'il est difficile de répondre à la question « à partir de quelle taille n d'échantillon, la distribution de $\widehat{\mu}$ peut-elle être considérée comme étant ap-

[5]Ce théorème a été publié en 1812 par le mathématicien Pierre-Simon de Laplace et a fait apparaître au grand jour le rôle privilégié de la distribution normale en statistique.

proximativement normale ? ». La réponse dépendra en effet du degré de non-normalité et notamment du degré d'asymétrie de la distribution des observations Y_i. Ainsi, $n = 100$ ne sera pas forcément suffisant si Y_i est asymétrique (par exemple si Y_i est log-normale), alors que $n = 10$ sera en général suffisant si Y_i est symétrique (évidemment, $n = 1$ suffira si Y_i est elle-même normale).

Ces résultats ont été établis sur des bases théoriques. On peut les vérifier en effectuant des simulations dont voici un exemple. On considère une population de nombres entiers répartis uniformément entre 1 et 10 (avec 10 % de 1, 10 % de 2, etc.). On considère de la sorte une variable Y_i quantitative discrète, dont la distribution n'est pas du tout normale (mais uniforme). On laissera vérifier au lecteur que l'on a dans cette population une moyenne de $\mu = 5.5$ et une variance de $\sigma^2 = 8.25$. De cette population, on tire au hasard un échantillon de taille $n = 10$. On obtient par exemple[6] :

3	4	6	10	3	9	10	7	7	1

La moyenne de cet échantillon est égale à $\widehat{\mu} = 6.0$, ce qui est légèrement au-dessus du véritable $\mu = 5.5$. Les résultats ci-dessus nous disent cependant que si on répétait l'expérience, la distribution des $\widehat{\mu}$ obtenus serait à peu près normale avec une moyenne de $\mu = 5.5$ et un écart type de $\sigma/\sqrt{n} = \sqrt{8.25/10} = 0.91$. Vu que l'on connaît ici la population, il est possible de vérifier ces résultats en répétant effectivement l'expérience, ce que nous avons fait 10 000 fois avec les résultats suivants :

1^{er} échantillon	3 4 6 10 3 9 10 7 7 1	$\widehat{\mu} = 6.0$
2^{e} échantillon	3 2 7 4 8 5 8 10 4 8	$\widehat{\mu} = 5.9$
3^{e} échantillon	10 3 7 2 3 4 1 4 9 4	$\widehat{\mu} = 4.7$
4^{e} échantillon	5 6 5 2 9 7 8 2 8 5	$\widehat{\mu} = 5.7$
5^{e} échantillon	9 7 8 6 6 8 1 5 8 7	$\widehat{\mu} = 6.5$
⋮	⋮	⋮
$10\,000^{\text{e}}$ échantillon	5 8 2 5 4 1 3 10 1 2	$\widehat{\mu} = 4.1$

En retenant deux décimales, la moyenne des 10 000 valeurs de $\widehat{\mu}$ ainsi obtenues était de 5.50 et leur écart type était de 0.91, comme prédit par la théorie. La figure 3.1 nous montre l'histogramme de ces 10 000 valeurs. Il suggère effectivement une distribution normale. De plus, les mesures d'asymétrie et d'aplatissement de ces 10 000 valeurs de $\widehat{\mu}$ étaient respectivement de 0.01 et de -0.07, confirmant ainsi la normalité.

[6] Dans R, un échantillon de taille 10 provenant de cette distribution peut être généré en utilisant la commande `sample(1:10,10,replace=T)`. Dans notre exemple, on a par ailleurs initialisé le générateur de nombres aléatoires de R en utilisant `set.seed(1)`.

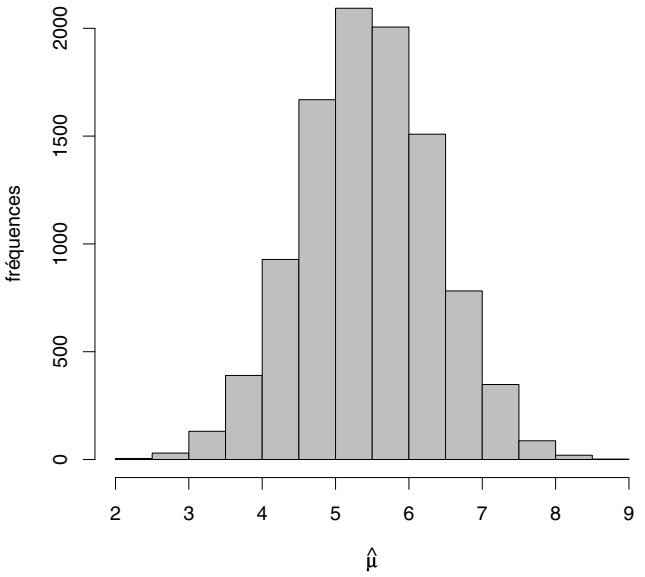

Figure 3.1 – Histogramme de 10 000 valeurs de $\widehat{\mu}$.

Notons qu'avec 10 000 simulations, on obtient une excellente approximation de la distribution de $\widehat{\mu}$. Pour connaître exactement la distribution de $\widehat{\mu}$ par simulation, il faudrait répéter l'expérience une infinité de fois.

3.4 Distribution de la variance d'un échantillon

On a écrit qu'il existe au moins deux raisons de vouloir estimer la variance σ^2 de la population, la seconde étant que cela nous permet d'estimer la précision de l'estimateur $\widehat{\mu}$ de la moyenne μ de cette même population *via* son erreur type. Afin d'estimer σ^2, l'idée naturelle serait de calculer la variance empirique définie par :

$$\widehat{\sigma}^2 = \frac{\sum_i (Y_i - \widehat{\mu})^2}{n}.$$

On va voir cependant que $\widehat{\sigma}^2$ est un *estimateur biaisé* de σ^2 : la valeur de $\widehat{\sigma}^2$ calculée dans notre échantillon serait en moyenne en dessous du véritable paramètre σ^2 si on répétait l'expérience. En utilisant cet estimateur, on aura ainsi tendance à sous-estimer la variabilité, en particulier lorsque n est petit (on va voir que le biais est négligeable pour un grand n). On a en effet (pour autant

que les observations $Y_1, Y_2, \cdots, Y_n$ de l'échantillon soient indépendantes) :

$$\begin{aligned}
\mathrm{E}(\widehat{\sigma}^2) &= E\left(\frac{\sum_i (Y_i - \widehat{\mu})^2}{n}\right) \\
&= E\left(\frac{\sum_i Y_i^2 - n\widehat{\mu}^2}{n}\right) \\
&= \frac{\sum_i \mathrm{E}(Y_i^2) - n\mathrm{E}(\widehat{\mu}^2)}{n} \\
&= \frac{\sum_i \mathrm{Var}(Y_i) + \sum_i E^2(Y_i) - n \cdot \mathrm{Var}(\widehat{\mu}) - n \cdot E^2(\widehat{\mu})}{n} \\
&= \frac{n\sigma^2 + n\mu^2 - n\sigma^2/n - n\mu^2}{n} \\
&= \frac{(n-1)\sigma^2}{n} \\
&< \sigma^2.
\end{aligned}$$

Afin d'obtenir un estimateur sans biais de la variance de la population σ^2, on redéfinit la variance de l'échantillon comme suit :

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \widehat{\mu})^2}{n-1} = \frac{n}{n-1} \cdot \widehat{\sigma}^2.$$

Lorsque l'on calcule une variance dans un échantillon, on divise ainsi par $n-1$ (et non par n) la somme des déviations au carré par rapport à la moyenne. On a alors[7] :

$$\mathrm{E}(\tilde{\sigma}^2) = E\left(\frac{n}{n-1} \cdot \widehat{\sigma}^2\right) = \frac{n}{n-1} \cdot \mathrm{E}(\widehat{\sigma}^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \sigma^2 = \sigma^2.$$

Ainsi, $\tilde{\sigma}^2$ est un estimateur sans biais de σ^2 (en moyenne $\tilde{\sigma}^2$ serait égal à σ^2 si on répétait l'expérience). Ce nouvel estimateur n'est pas défini pour un échantillon de taille $n = 1$, ce qui correspond à une certaine logique (on ne peut pas prétendre estimer une variabilité à partir d'une seule observation). On estimera par ailleurs l'erreur type de $\widehat{\mu}$ par $\tilde{\sigma}/\sqrt{n}$.

Exemple 3.1 *Dans notre exemple des $n = 60$ Onobrychis introduits au chapitre 1, on avait $\widehat{\sigma}^2 = 14.54$ (et ainsi $\widehat{\sigma} = 3.81$). On aura donc $\tilde{\sigma}^2 = (60/59) \cdot 14.54 = 14.79$ (et ainsi $\tilde{\sigma} = 3.85$) et on estimera l'erreur type de $\widehat{\mu}$ par $3.85/\sqrt{60} = 0.50$.*

Ici également, on peut vérifier ce résultat théorique par simulation. Pour les 10 000 échantillons simulés dans la section précédente, on a obtenu par exemple :

[7] Notons toutefois que si la taille N de la population est finie, l'espérance de $\tilde{\sigma}^2$ sera égale à $\sigma^2 \cdot N/(N-1)$ (non pas à σ^2), ce qui motive certains auteurs à utiliser également un dénominateur de $N-1$ pour définir la variance de la population.

1^{er} échantillon	3 4 6 10 3 9 10 7 7 1	$\widehat{\sigma}^2 = 9.0$	$\tilde{\sigma}^2 = 10.0$
2^{e} échantillon	3 2 7 4 8 5 8 10 4 8	$\widehat{\sigma}^2 = 6.3$	$\tilde{\sigma}^2 = 7.0$
3^{e} échantillon	10 3 7 2 3 4 1 4 9 4	$\widehat{\sigma}^2 = 8.0$	$\tilde{\sigma}^2 = 8.9$
4^{e} échantillon	5 6 5 2 9 7 8 2 8 5	$\widehat{\sigma}^2 = 5.2$	$\tilde{\sigma}^2 = 5.8$
5^{e} échantillon	9 7 8 6 6 8 1 5 8 7	$\widehat{\sigma}^2 = 4.7$	$\tilde{\sigma}^2 = 5.2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$10\,000^{\text{e}}$ échantillon	5 8 2 5 4 1 3 10 1 2	$\widehat{\sigma}^2 = 8.1$	$\tilde{\sigma}^2 = 9.0$

La moyenne des 10 000 valeurs de $\widehat{\sigma}^2$ était de 7.45, ce qui est au-dessous du véritable σ^2. Par contre, la moyenne des 10 000 valeurs de $\tilde{\sigma}^2$ était de 8.28, ce qui est très proche de $\sigma^2 = 8.25$. La théorie nous dit que la moyenne des $\tilde{\sigma}^2$ aurait été parfaitement égale à σ^2 si on avait effectué une infinité de simulations.

3.5 Distribution d'une proportion calculée dans un échantillon

Rappelons que la distribution d'une variable binaire est caractérisée par une proportion et qu'une proportion peut être considérée comme un cas particulier de moyenne. Il suffit pour cela de coder les deux catégories de la variable binaire par 1 et 0 et de considérer la variable binaire ainsi codée comme une variable quantitative discrète (ce que nous ferons en général dans ce texte), la proportion de 1 étant alors égale à la moyenne de cette variable. Du coup, les résultats vus dans ce chapitre pour l'estimateur $\widehat{\mu}$ de la moyenne μ de la population sont aussi valables pour l'estimateur $\widehat{\pi}$ de la proportion π de 1 dans la population.

Considérons un échantillon de n observations indépendantes $Y_1, Y_2, \cdots, Y_n$ d'une variable binaire avec $\mathrm{E}(Y_i) = \pi$ et donc $\mathrm{Var}(Y_i) = \pi(1-\pi)$. En appliquant les résultats vus ci-dessus, on obtient :

- $\mathrm{E}(\widehat{\pi}) = \mathrm{E}\left(\frac{\sum_i Y_i}{n}\right) = \pi$

  $\rightarrow \widehat{\pi}$ est un estimateur sans biais de π

- $\mathrm{Var}(\widehat{\pi}) = \frac{\pi(1-\pi)}{n}$ et donc $\mathrm{SE}(\widehat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}$

  $\rightarrow$ la précision de l'estimateur $\widehat{\pi}$ dépend de la véritable proportion π

  $\rightarrow$ il sera plus difficile d'estimer précisément une proportion proche de 0.5 qu'une proportion proche de 0 ou de 1

- la distribution de $\widehat{\pi}$ sera approximativement normale pour autant que n soit suffisamment grand.

Ce dernier point appelle quelques commentaires. Notons tout d'abord que $\widehat{\pi}$ est une variable quantitative discrète avec $n+1$ valeurs possibles : $0, 1/n, 2/n,$

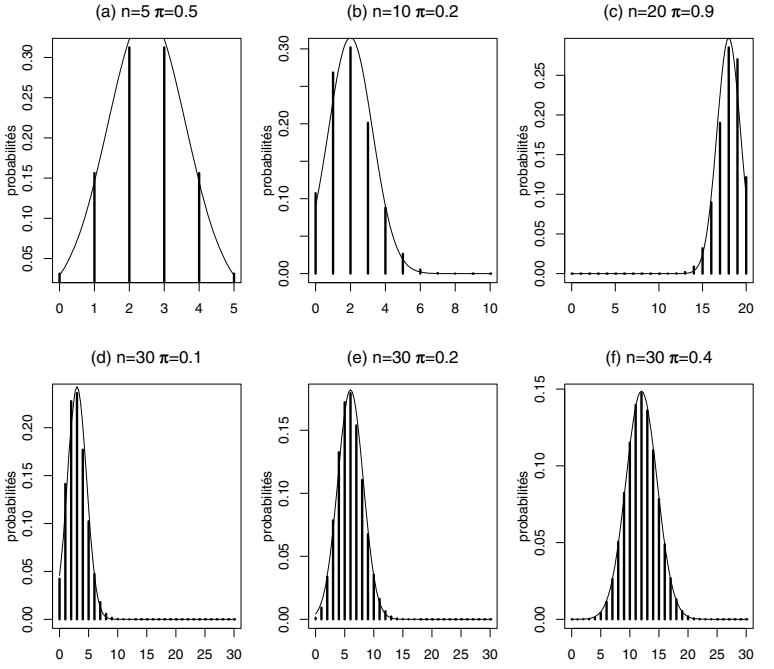

Figure 3.2 – Exemples de distributions binomiales avec différentes valeurs de n et π.

$\cdots, (n-1)/n, 1$ (avec un échantillon de taille $n = 10$, il sera possible d'observer une proportion $\widehat{\pi}$ de 0, 0.1, 0.2, etc., mais il sera impossible d'observer une proportion $\widehat{\pi}$ de 0.15 ou de 0.18). Cette distribution discrète sera pourtant proche d'une distribution normale si n est grand (comme on le verra sur la figure 3.2). En général, l'approximation normale sera bonne si $n\pi \geq 5$ et si $n(1-\pi) \geq 5$. Pour s'assurer d'une normalité approximative, on aura donc besoin d'un plus grand n dans les cas où π est proche de 0 ou de 1 que dans les cas où π est proche de 0.5. Ainsi $n = 10$ suffira si $\pi = 0.5$, alors qu'il faudra $n = 50$ si $\pi = 0.1$ ou $\pi = 0.9$.

3.6 Distribution exacte d'une proportion calculée dans un échantillon

La distribution exacte de $\widehat{\pi}$ est connue. Elle est habituellement tabulée pour la variable aléatoire $K = n\widehat{\pi}$ qui représente le nombre de 1 dans l'échantillon (avec $n = 10$, $K = 0$ correspond à $\widehat{\pi} = 0$, $K = 1$ à $\widehat{\pi} = 0.1$, $K = 2$ à $\widehat{\pi} = 0.2$, etc.). La distribution de K est dite *binomiale avec paramètres n et π*. Il s'agit d'une famille de distributions indexée par deux paramètres : la taille n

de l'échantillon et la proportion π de 1 dans la population de laquelle provient l'échantillon. Rappelons que la distribution d'une variable quantitative discrète peut être caractérisée par les proportions (les probabilités) d'occurence des différentes valeurs possibles. Pour une distribution binomiale avec paramètres n et π, ces probabilités sont données par[8] :

$$\Pr\{K = k\} = \frac{n!}{k!(n-k)!} \cdot \pi^k (1-\pi)^{n-k}$$

et ceci pour les $n+1$ valeurs possibles $k = 0, 1, 2, \cdots, n$. Par exemple, si $n = 10$ et $\pi = 0.2$, on aura les probabilités suivantes[9] :

$$\begin{array}{rcccccc}
\Pr\{K=0\} & = & \frac{10!}{0!\cdot 10!} \cdot 0.2^0 \cdot 0.8^{10} & = & 0.8^{10} & = & 11\ \% \\
\Pr\{K=1\} & = & \frac{10!}{1!\cdot 9!} \cdot 0.2^1 \cdot 0.8^{9} & = & 10 \cdot 0.2 \cdot 0.8^{9} & = & 27\ \% \\
\Pr\{K=2\} & = & \frac{10!}{2!\cdot 8!} \cdot 0.2^2 \cdot 0.8^{8} & = & 45 \cdot 0.2^2 \cdot 0.8^{8} & = & 30\ \% \\
\Pr\{K=3\} & = & \frac{10!}{3!\cdot 7!} \cdot 0.2^3 \cdot 0.8^{7} & = & 120 \cdot 0.2^3 \cdot 0.8^{7} & = & 20\ \% \\
\Pr\{K=4\} & = & \frac{10!}{4!\cdot 6!} \cdot 0.2^4 \cdot 0.8^{6} & = & 210 \cdot 0.2^4 \cdot 0.8^{6} & = & 8.8\ \% \\
\Pr\{K=5\} & = & \frac{10!}{5!\cdot 5!} \cdot 0.2^5 \cdot 0.8^{5} & = & 252 \cdot 0.2^5 \cdot 0.8^{5} & = & 2.6\ \% \\
\Pr\{K=6\} & = & \frac{10!}{6!\cdot 4!} \cdot 0.2^6 \cdot 0.8^{4} & = & 210 \cdot 0.2^6 \cdot 0.8^{4} & = & 0.55\ \% \\
\Pr\{K=7\} & = & \frac{10!}{7!\cdot 3!} \cdot 0.2^7 \cdot 0.8^{3} & = & 120 \cdot 0.2^7 \cdot 0.8^{3} & = & 0.079\ \% \\
\Pr\{K=8\} & = & \frac{10!}{8!\cdot 2!} \cdot 0.2^8 \cdot 0.8^{2} & = & 45 \cdot 0.2^8 \cdot 0.8^{2} & = & 0.0074\ \% \\
\Pr\{K=9\} & = & \frac{10!}{9!\cdot 1!} \cdot 0.2^9 \cdot 0.8^{1} & = & 10 \cdot 0.2^9 \cdot 0.8 & = & 0.00041\ \% \\
\Pr\{K=10\} & = & \frac{10!}{10!\cdot 0!} \cdot 0.2^{10} \cdot 0.8^{0} & = & 0.2^{10} & = & 0.000010\ \%
\end{array}$$

Si la véritable proportion est égale à $\pi = 0.2$ et si la taille de l'échantillon est $n = 10$, on aura ainsi une grande probabilité d'observer $\widehat{\pi} = 0.1$, $\widehat{\pi} = 0.2$ ou $\widehat{\pi} = 0.3$. Par contre, il sera extrêmement rare d'observer $\widehat{\pi} = 1$.

La figure 3.2 nous montre des exemples de distributions binomiales (sous forme de diagrammes en barres, qui est la méthode graphique habituelle pour représenter la distribution d'une variable quantitative discrète). Le graphique (b) nous montre la distribution binomiale avec $n = 10$ et $\pi = 0.2$ que nous venons de calculer. Une distribution normale est également représentée en arrière-plan par un trait fin sur chacun de ces graphiques. Selon notre critère ci-dessus ($n\pi \geq 5$ et $n(1-\pi) \geq 5$), l'approximation normale est bonne pour les distributions binomiales représentées dans les graphiques (e) et (f).

[8] $k!$ représente le produit $k(k-1)\cdots 2 \cdot 1$, par exemple $4! = 24$ et par convention $0! = 1$.

[9] Dans R, ces probabilités peuvent s'obtenir avec la commande `dbinom(k,n,pi)`, où `k` est une valeur possible et `n` et `pi` sont les paramètres de la distribution binomiale en question. Dans cet exemple, on a ainsi utilisé `dbinom(0:10,10,0.2)`.

Chapitre 4

Intervalle de confiance

On va voir à présent comment utiliser les concepts et résultats vus au chapitre précédent pour faire de l'inférence sur les paramètres de la population à partir des données de l'échantillon.

4.1 Méthode de Wald

On se place tout d'abord dans un contexte général. On considère l'estimateur $\widehat{\theta}$ (calculé sur l'échantillon) d'un paramètre θ (défini sur la population). Si $\widehat{\theta}$ est un estimateur sans biais et normalement distribué, on aura donc :

$$\Pr\left\{\theta - 1.96 \cdot \mathrm{SE}(\widehat{\theta}) \leq \widehat{\theta} \leq \theta + 1.96 \cdot \mathrm{SE}(\widehat{\theta})\right\} = 0.95$$

que l'on exprime parfois en standardisant $\widehat{\theta}$ de la sorte :

$$\Pr\left\{-1.96 \leq \frac{\widehat{\theta} - \theta}{\mathrm{SE}(\widehat{\theta})} \leq 1.96\right\} = 0.95.$$

En manipulant ces inéquations de façon à ce que θ se retrouve au milieu, on obtient :

$$\Pr\left\{\widehat{\theta} - 1.96 \cdot \mathrm{SE}(\widehat{\theta}) \leq \theta \leq \widehat{\theta} + 1.96 \cdot \mathrm{SE}(\widehat{\theta})\right\} = 0.95.$$

Cette dernière formulation est particulièrement intéressante. Elle se lit de la façon suivante : « la probabilité que la véritable valeur du paramètre θ se trouve à moins de 2 (ou 1.96) erreurs types de son estimateur $\widehat{\theta}$ est de 95 % ». Voilà exactement le genre d'information que l'on désire connaître lorsque l'on fait de l'inférence. L'intervalle donné par :

$$[\widehat{\theta} - 1.96 \cdot \mathrm{SE}(\widehat{\theta}); \widehat{\theta} + 1.96 \cdot \mathrm{SE}(\widehat{\theta})]$$

est dit un *intervalle de confiance au niveau 95 % pour* θ (en anglais : `95 % confidence interval`, que l'on abrégera parfois par 95 % CI). En pratique,

on remplacera souvent la valeur 1.96 par 2. La méthode qui consiste à calculer un intervalle de confiance en additionnant et soustrayant à l'estimateur un multiple de son erreur type, fondée sur l'absence de biais et sur la normalité de l'estimateur, est appelée *méthode de Wald* (du nom d'un statisticien). L'intervalle ci-dessus est appelé *intervalle de confiance de Wald au niveau 95 % pour* θ.

Un intervalle de confiance au niveau 95 % pour θ contient donc la véritable valeur du paramètre θ avec une probabilité de 95 %. Le sens de cette phrase ne va pourtant pas de soi. En effet, l'intervalle de confiance que l'on calcule sur notre échantillon va soit contenir soit ne pas contenir la véritable valeur du paramètre θ. On ne sait certes pas s'il la contient ou non (vu que l'on ne connaît pas cette véritable valeur), mais il n'y a pas d'autres possibilités. On peut donc se demander ce que l'on veut dire par « contenir une valeur avec une probabilité de 95 % ». De même, dans la vie courante, on peut se demander ce que l'on veut dire lorsque l'on affirme par exemple « il y a une probabilité de 50 % que je passe cet examen », ou « il y a une probabilité de 90 % que je gagne ce match de tennis ».

En fait, lorsque l'on parlera de probabilité dans ce texte, il s'agira toujours d'une proportion, c'est-à-dire d'un numérateur divisé par un dénominateur. Afin de définir un dénominateur dans les exemples ci-dessus, on s'imaginera une fois encore que l'on répète l'expérience. Ainsi, dire « il y a une probabilité de 50 % que je passe cet examen » revient à dire « si je me présente 100 fois à cet d'examen, je le passerai 50 fois », alors que dire « il y a une probabilité de 90 % que je gagne ce match de tennis » revient à dire « si je joue 100 fois ce match, je le gagnerai 90 fois ». De même, dire « un intervalle de confiance contient le véritable paramètre avec une probabilité de 95 % » veut dire « si on tire 100 échantillons de la population et si on calcule dans chacun d'entre eux un intervalle de confiance pour ce paramètre, 95 d'entre eux contiendront sa véritable valeur »[1].

La figure 4.1 illustre le concept d'intervalle de confiance. La véritable valeur du paramètre θ est représentée par la ligne verticale. Les 100 segments horizontaux représentent chacun un intervalle de confiance au niveau 95 % pour θ, ces intervalles étant calculés dans 100 échantillons différents tirés de cette même population. On voit que 95 d'entre eux contiennent la véritable valeur de θ, alors que 5 d'entre eux ne la contiennent pas.

[1] Nous avons commis ici une petite simplification de langage (que l'on commettra encore quelquefois dans ce qui suit). Il faudrait dire en fait « 95 d'entre eux contiendront la véritable valeur de ce paramètre, *en moyenne* ». En effet, le nombre d'intervalles de confiance sur 100 qui contiendront la véritable valeur du paramètre est lui-même une variable aléatoire. S'il s'agit d'un intervalle de confiance exact (concept que l'on verra plus loin), la distribution de cette variable aléatoire sera binomiale avec paramètres $n = 100$ et $\pi = 0.95$. Si on répétait l'expérience (qui consiste ici à tirer 100 échantillons de la population, à calculer un intervalle de confiance au niveau 95 % dans chacun d'entre eux, et à compter combien d'entre eux contiennent effectivement la véritable valeur du paramètre), la réalisation de cette variable aléatoire ne sera donc pas systématiquement 95 (mais pourra être par exemple 94, 96 ou 97). Cependant, la moyenne de cette variable aléatoire sera égale à 95.

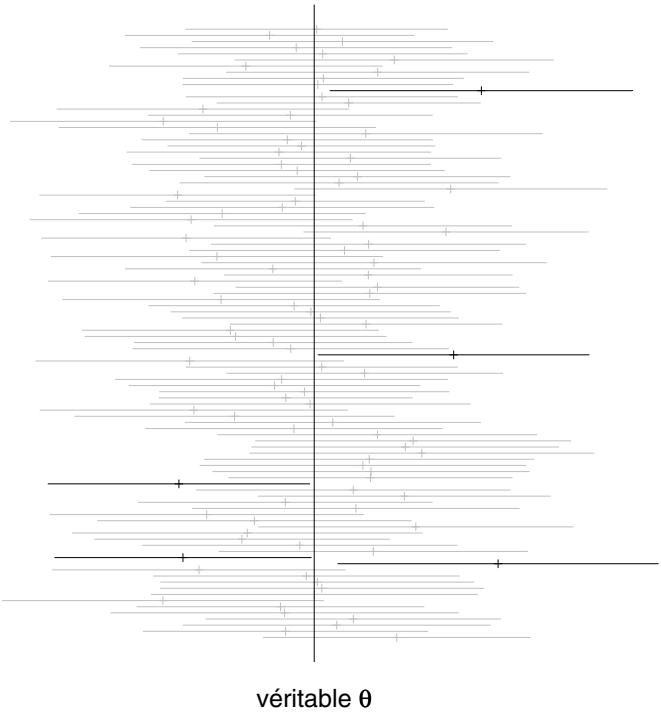

Figure 4.1 – Intervalle de confiance contenant le véritable θ avec probabilité 0.95.

Lorsqu'on calcule à partir de notre échantillon un intervalle de confiance au niveau 95 % pour un paramètre θ de la population, on espère que notre intervalle ne soit pas parmi les 5 % des intervalles qui ne contiennent pas la véritable valeur de θ. Si ce risque de 5 % nous paraît trop important, on pourra élever le *niveau de confiance* à 99 %. En utilisant la méthode de Wald, on calculera alors :

$$[\widehat{\theta} - 2.58 \cdot \mathrm{SE}(\widehat{\theta}); \widehat{\theta} + 2.58 \cdot \mathrm{SE}(\widehat{\theta})].$$

Sur 100 échantillons où l'on calculerait un tel intervalle de confiance, un seul (en moyenne) ne contiendrait pas la véritable valeur de θ. Si ce risque de 1 % nous paraît encore trop grand, on élèvera encore le niveau de confiance, par exemple à 99.9 % comme suit :

$$[\widehat{\theta} - 3.29 \cdot \mathrm{SE}(\widehat{\theta}); \widehat{\theta} + 3.29 \cdot \mathrm{SE}(\widehat{\theta})].$$

Plus généralement, on calcule un intervalle de confiance au niveau $1 - \alpha$ pour θ comme suit (α dénotant la probabilité que θ ne soit pas dans l'intervalle) :

$$[\widehat{\theta} - z_{1-\alpha/2} \cdot \mathrm{SE}(\widehat{\theta}); \widehat{\theta} + z_{1-\alpha/2} \cdot \mathrm{SE}(\widehat{\theta})].$$

On note que la longueur de l'intervalle de confiance augmente avec le niveau de confiance. Or, un intervalle de confiance trop large ne sera pas très utile (car peu informatif). Il s'agit donc de trouver un compromis raisonnable entre niveau de confiance et longueur de l'intervalle de confiance. En pratique, ce compromis est souvent choisi à 95 %.

> On calcule en général un intervalle de confiance au niveau 95 %.

4.2 Intervalle de confiance de Wald pour une moyenne

On a vu que la moyenne $\widehat{\mu}$ de l'échantillon est un estimateur sans biais et (approximativement) normal de la moyenne μ de la population. On peut donc ici utiliser la méthode de Wald. Comme $\mathrm{SE}(\widehat{\mu}) = \sigma/\sqrt{n}$, on aura :

$$\Pr\left\{\mu - 1.96 \cdot \sigma/\sqrt{n} \leq \widehat{\mu} \leq \mu + 1.96 \cdot \sigma/\sqrt{n}\right\} = 0.95$$

que l'on exprime parfois en standardisant $\widehat{\mu}$ de la sorte :

$$\Pr\left\{-1.96 \leq \sqrt{n} \cdot \frac{\widehat{\mu} - \mu}{\sigma} \leq 1.96\right\} = 0.95.$$

En manipulant ces inéquations de façon à ce que μ se retrouve au milieu, on obtient :

$$\Pr\left\{\widehat{\mu} - 1.96 \cdot \sigma/\sqrt{n} \leq \mu \leq \widehat{\mu} + 1.96 \cdot \sigma/\sqrt{n}\right\} = 0.95.$$

L'intervalle donné par :

$$[\widehat{\mu} - 1.96 \cdot \sigma/\sqrt{n}; \widehat{\mu} + 1.96 \cdot \sigma/\sqrt{n}]$$

est donc un intervalle de confiance de Wald au niveau 95 % pour μ. En pratique, il faudra remplacer σ par son estimateur $\tilde{\sigma}$ dans ces formules (ce qui revient à dire : il faudra estimer l'erreur type de $\widehat{\mu}$). On calculera donc un intervalle de confiance de Wald au niveau 95 % pour μ comme suit :

$$[\widehat{\mu} - 1.96 \cdot \tilde{\sigma}/\sqrt{n}; \widehat{\mu} + 1.96 \cdot \tilde{\sigma}/\sqrt{n}].$$

Plus généralement, l'intervalle donné par :

$$[\widehat{\mu} - z_{1-\alpha/2} \cdot \tilde{\sigma}/\sqrt{n}; \widehat{\mu} + z_{1-\alpha/2} \cdot \tilde{\sigma}/\sqrt{n}]$$

est un intervalle de confiance de Wald au niveau $1 - \alpha$ pour μ.

Exemple 4.1 *Dans l'exemple des Onobrychis, on avait* $n = 60$, $\widehat{\mu} = 21.7$ *cm et* $\tilde{\sigma} = 3.85$ *cm. Un intervalle de confiance au niveau* 95 % *pour* μ *s'obtient ainsi par :*

$$21.7 \pm 1.96 \cdot 3.85/\sqrt{60} = [20.7; 22.7] \text{ cm} .$$

Bien que l'on ne connaisse toujours pas la véritable moyenne μ *de la hauteur des Onobrychis après une culture de six mois, on est capable ici de la cerner avec une grande probabilité. On peut dire que cette moyenne se trouve avec une probabilité de* 95 % *entre* 20.7 *et* 22.7 *cm.*

4.3 Intervalle de confiance de Student pour une moyenne

Le remplacement de σ par son estimateur $\tilde{\sigma}$ en fin de section précédente n'est pas totalement anodin. La méthode de Wald se fonde en effet sur la normalité de la variable aléatoire :

$$\frac{\widehat{\mu} - \mathrm{E}(\widehat{\mu})}{\mathrm{SE}(\widehat{\mu})} = \sqrt{n} \cdot \frac{\widehat{\mu} - \mu}{\sigma}$$

(que l'on aura lorsque les observations Y_i sont normales ou lorsque n est assez grand). Or, il se trouve que la variable aléatoire légèrement modifiée :

$$\sqrt{n} \cdot \frac{\widehat{\mu} - \mu}{\tilde{\sigma}}$$

ne sera pas normale lorsque n est petit, et ceci même si les observations Y_i sont normales. Dans ce dernier cas, la distribution de cette variable aléatoire est cependant connue mathématiquement et appelée *distribution de Student* (ou *distribution t*). Il s'agit d'une nouvelle famille de distributions indexée par un paramètre appelé *degrés de liberté*, que nous noterons *dl*[2].

Lorsque les observations Y_i sont normales, la variable aléatoire $\sqrt{n} \cdot \dfrac{\widehat{\mu} - \mu}{\tilde{\sigma}}$ aura une distribution de Student avec $n - 1$ *dl*.

[2]La densité de probabilité d'une distribution de Student avec $n - 1$ degrés de liberté est définie par :

$$f(y) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi} \cdot \Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{y^2}{n-1}\right)^{-\frac{n}{2}}$$

où $\Gamma(x)$ dénote la fonction Gamma. Ce résultat a été publié en 1908 par William Gosset, dont le pseudonyme était Student. On a récemment célébré les 100 ans de ce résultat. Voir par exemple l'article de Zabell (2008).

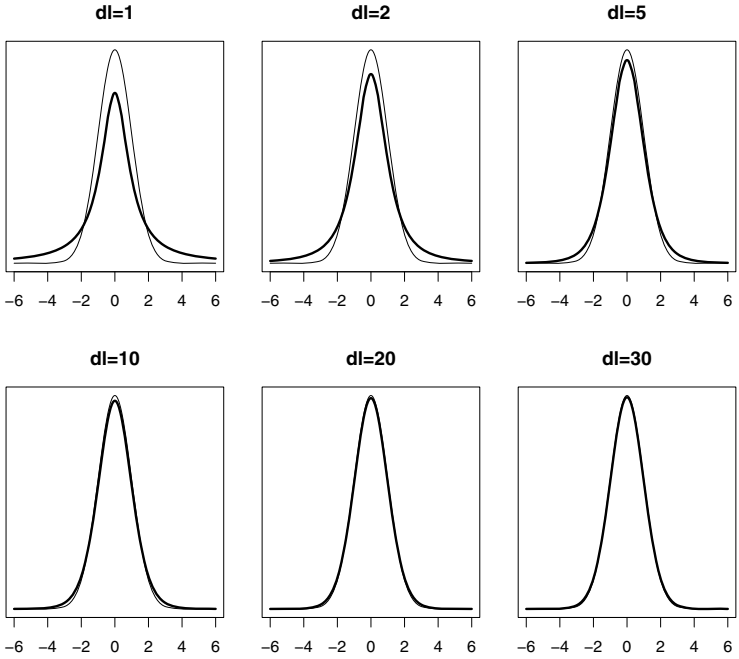

Figure 4.2 – Exemples de distributions de Student avec différentes valeurs de dl.

La figure 4.2 nous montre différentes distributions de Student. Il s'agit de distributions symétriques autour de la valeur 0, avec plus de valeurs extrêmes qu'une distribution normale standardisée, notamment lorsque dl est petit. Pour une grande valeur de dl (on dira à partir de $dl = 30$), une distribution de Student sera cependant très proche d'une distribution normale standardisée (représentée en arrière-plan par un trait fin sur ces graphiques).

Dans ce texte, on notera par $t_{\alpha,dl}$ le quantile α d'une distribution de Student avec paramètre dl. Par symétrie, on a $t_{1-\alpha,dl} = -t_{\alpha,dl}$. Comme les quantiles d'une distribution de Student sont plus éloignés de 0 que ceux d'une distribution normale standardisée, on a $|t_{\alpha,dl}| > |z_\alpha|$. Pour une grande valeur de dl, on a cependant $t_{\alpha,dl} \approx z_\alpha$. Le tableau A.2 (donné en annexe) nous donne quelques quantiles importants d'une distribution de Student pour différentes valeurs de dl. Au besoin, un logiciel statistique nous donnera n'importe quel quantile de ces distributions. Le quantile 97.5 % d'une distribution de Student vaut par exemple 4.30, 2.57, 2.23, 2.09 et 2.04 avec respectivement $dl = 2, 5, 10, 20$ et 30. Pour une valeur de dl plus élevée, ce quantile se rapprochera encore davantage de la valeur 1.96 (le quantile 97.5 % d'une distribution normale standardisée)[3].

[3]Dans R, on calcule par exemple $t_{0.975,2}$ par la commande `qt(0.975,2)` (et on trouve ainsi 4.30, voir le tableau A.2 donné en annexe).

Lorsque les observations Y_i sont normales, on aura ainsi :

$$\Pr\left\{-t_{0.975,n-1} \leq \sqrt{n} \cdot \frac{\widehat{\mu}-\mu}{\tilde{\sigma}} \leq t_{0.975,n-1}\right\} = 0.95$$

c'est-à-dire (en manipulant ces inéquations de façon à ce que μ se retrouve au milieu) :

$$\Pr\left\{\widehat{\mu} - t_{0.975,n-1} \cdot \tilde{\sigma}/\sqrt{n} \leq \mu \leq \widehat{\mu} + t_{0.975,n-1} \cdot \tilde{\sigma}/\sqrt{n}\right\} = 0.95.$$

L'intervalle donné par :

$$[\widehat{\mu} - t_{0.975,n-1} \cdot \tilde{\sigma}/\sqrt{n}; \widehat{\mu} + t_{0.975,n-1} \cdot \tilde{\sigma}/\sqrt{n}]$$

est dit un *intervalle de confiance de Student au niveau 95 % pour* μ. Plus généralement, l'intervalle donné par :

$$[\widehat{\mu} - t_{1-\alpha/2,n-1} \cdot \tilde{\sigma}/\sqrt{n}; \widehat{\mu} + t_{1-\alpha/2,n-1} \cdot \tilde{\sigma}/\sqrt{n}]$$

est un intervalle de confiance de Student au niveau $1-\alpha$ pour μ. Cet intervalle peut être utilisé quelle que soit la taille $n \geq 2$ de l'échantillon, pour autant que les observations Y_i soient normales.

Un intervalle de confiance de Student pour une moyenne se calcule donc comme un intervalle de confiance de Wald pour une moyenne, sauf que l'on remplace $z_{1-\alpha/2}$ par $t_{1-\alpha/2,n-1}$. Un intervalle de confiance de Student sera donc légèrement plus large qu'un intervalle de confiance de Wald (car $t_{1-\alpha/2,dl} > z_{1-\alpha/2}$). La méthode de Student tient compte du fait que $\tilde{\sigma}$ est un estimateur imprécis de σ lorsque n est petit. On ajoute ainsi de l'incertitude à notre calcul, ce qui se traduit logiquement par une augmentation de la longueur de l'intervalle de confiance. Pour l'exemple des *Onobrychis*, cela ne change pas grand chose car l'échantillon est suffisamment grand (il faudra remplacer $z_{0.975} = 1.96$ par $t_{0.975,59} = 2.00$). Pour un échantillon plus petit, cela fera par contre une différence.

Exemple 4.2 *Voici un échantillon de* $n = 10$ *personnes ayant effectué un régime pendant deux mois et chez lesquelles on a observé les pertes de poids suivantes (en kg) :*

−1.6	−0.1	0.2	0.4	0.6	0.8	1.0	1.6	1.7	2.6

*Ces observations sont rangées par ordre croissant. Une valeur négative indique une prise de poids au lieu d'une perte de poids (*2 *personnes sur* 10 *ont pris du poids). On s'intéresse à la moyenne* μ *de ces pertes de poids dans une population de personnes qui suivraient le même régime. On dira dans cet exemple que* μ *représente l'effet (ou l'effet moyen) du régime*[4]*. On a ici :*

[4]On pourra nous rétorquer qu'une perte de poids dans ce contexte ne sera pas forcément l'effet du régime proprement dit ; cela pourrait tout aussi bien être un effet psychologique, ou l'effet du temps qui passe, mais ceci est un autre débat que l'on retrouvera au chapitre 15.

- $\widehat{\mu} = 0.72$ *kg (perte de poids en moyenne)*
- $\tilde{\sigma} = 1.145$ *kg (indicateur de la variabilité des pertes de poids)*
- *on calcule un intervalle de confiance de Student au niveau* 95 % *pour* μ *comme suit (en utilisant* $t_{0.975,9} = 2.26$*) :*

$$0.72 \pm 2.26 \cdot 1.145/\sqrt{10} = [-0.10; 1.54] \text{ kg}$$

→ *le véritable* μ *se trouve avec probabilité* 95 % *entre* -0.10 *et* $+1.54$ *kg*

→ *bien que les personnes de notre échantillon aient en moyenne perdu du poids, on ne peut pas exclure la possibilité que dans la population la moyenne des pertes de poids soit en fait nulle (voire même négative), la valeur* 0 *étant contenue dans l'intervalle*

→ *on dira ici que l'effet du régime n'est pas prouvé statistiquement.*

On notera que l'on ne serait pas arrivé au même résultat si on avait calculé un intervalle de confiance de Wald au niveau 95 % *pour* μ*, car on aurait obtenu l'intervalle* $[0.01; 1.43]$ *(qui ne contient pas la valeur* 0*).*

4.4 Niveau nominal et niveau réel d'un intervalle de confiance

L'analyse statistique de l'exemple précédent est correcte pour autant que la variable « perte de poids » soit distribuée normalement dans la population. Si tel n'est pas le cas, un intervalle de confiance de Student ne sera pas toujours valide : on prétendra calculer un intervalle de confiance au niveau 95 %, alors qu'en réalité, la probabilité que le véritable μ se trouve à l'intérieur de l'intervalle sera différente de 95 % (elle sera par exemple inférieure à 95 %). Dans ce cas, le *niveau nominal* (c'est-à-dire le niveau désiré) de l'intervalle de confiance sera de 95 %, mais son *niveau réel* sera différent de 95 %. On adopte la terminologie suivante :

- **un intervalle de confiance est dit exact** si son niveau réel correspond exactement à son niveau nominal
- **un intervalle de confiance est dit valide** (ou approximativement valide) si son niveau réel correspond approximativement à son niveau nominal
- **un intervalle de confiance est dit non valide** si son niveau réel est (très) différent de son niveau nominal.

Parmi les intervalles de confiances non valides, on fait la distinction suivante[5] :

[5] Dans certains ouvrages, on réserve le terme de « non valide » à un intervalle de confiance libéral, considérant qu'il est plus grave d'avoir une méthode libérale plutôt que conservatrice.

- **un intervalle de confiance est dit conservateur** si son niveau réel est supérieur à son niveau nominal
- **un intervalle de confiance est dit libéral** si son niveau réel est inférieur à son niveau nominal.

Ainsi, un intervalle de confiance de Student pour une moyenne sera exact (et *a fortiori* valide) si les observations sont normales, et ceci quelle que soit la taille $n \geq 2$ de l'échantillon. Par contre, un intervalle de confiance de Wald pour une moyenne sera libéral (et donc non valide) avec un petit n. Par exemple, si les observations sont normales, on peut calculer que le niveau réel d'un intervalle de confiance de Wald au niveau nominal 95 % pour une moyenne sera respectivement de 91.8 %, 87.8 % et 70.0 % avec des échantillons de taille $n = 10$, $n = 5$ et $n = 2$. Toujours en supposant la normalité des observations, le niveau réel de cet intervalle de confiance sera par contre de 93.5 % avec $n = 20$ et de 94.0 % avec $n = 30$ (la validité de l'intervalle de confiance s'améliorant avec un grand n). En résumé, on a la situation suivante :

	CI Wald		CI Student	
	Y_i normal	Y_i non normal	Y_i normal	Y_i non normal
n petit	non valide	non valide	exact	non valide
n grand	valide	valide	exact	valide

Comme son domaine de validité est plus large, on utilise en pratique la formule de Student plutôt que celle de Wald lorsqu'il s'agit de calculer un intervalle de confiance pour une moyenne. Un intervalle de confiance de Student pour une moyenne ne sera cependant pas valide avec n petit et Y_i non normale.

On peut alors revenir à la question posée dans le chapitre précédent, à savoir « que veut dire n petit » ? La réponse, on s'en souvient, était de dire que cela dépend du degré de non-normalité de Y_i et notamment de son asymétrie. Pour illustrer cela, nous avons effectué une série de 36 simulations, en générant à chaque fois 10 000 échantillons de taille $n = 10, 30, 50, 100, 200$ ou 500, et avec des observations Y_i distribuées selon chacune des six distributions montrées dans la figure 2.4[6]. Dans chacun de ces 10 000 échantillons, nous avons calculé un intervalle de confiance de Student au niveau 95 % pour la moyenne μ. Nous avons ensuite calculé la proportion d'entre eux qui contenait effectivement la véritable valeur de μ (que nous connaissions, puisque nous avons nous-mêmes simulé ces données). Ces proportions constituent des estimations du niveau réel d'un intervalle de confiance de Student au niveau nominal 95 %[7].

[6]Dans R, un échantillon de taille n provenant de ces six distributions peut être généré en utilisant respectivement les commandes suivantes : `rnorm(n)`, `rt(n,3)`, `runif(n,-0.5,0.5)`, `c(rnorm(n/2,-2,1.2),rnorm(n/2,2,1))`, `-rchisq(n,3)+3` et `rlnorm(n)-exp(0.5)`. Notons que ces six distributions ont une moyenne nulle.

[7]Avec 10 000 simulations, l'erreur type de l'estimateur d'une proportion est donnée par $\sqrt{0.95 \cdot 0.05}/10\,000 = 0.002$ (du moins si la véritable proportion vaut 0.95). Ainsi, lorsque le niveau réel d'un intervalle de confiance est (proche de) 95 %, on l'estimera avec une précision

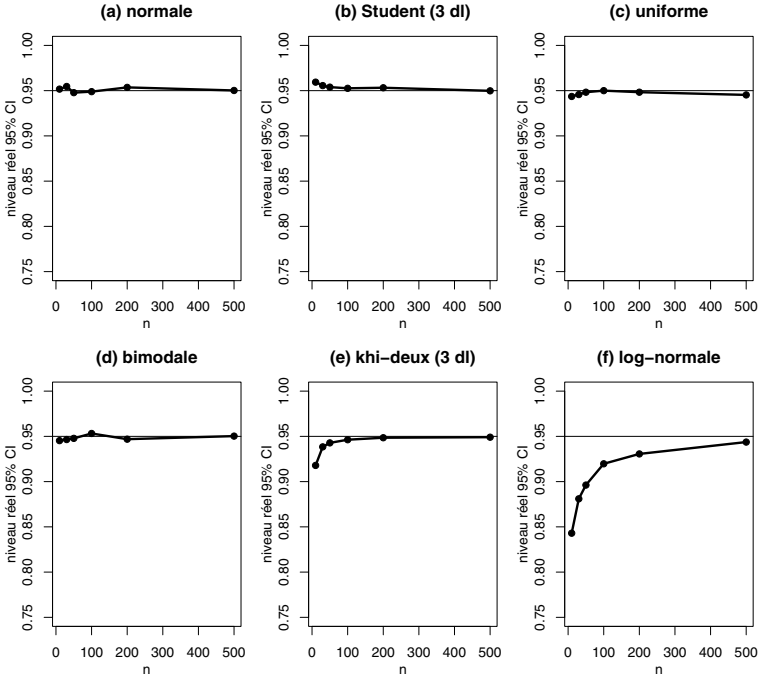

Figure 4.3 – Niveau réel d'un 95 % CI de Student pour μ selon n.

Les résultats sont montrés dans la figure 4.3. Le graphique (a) représente le cas où les observations sont normales, cas pour lequel l'intervalle de confiance de Student est exact, comme le confirment nos simulations. Les graphiques (b) et (c) représentent des cas de distributions symétriques mais non normales. On voit que le niveau réel d'un intervalle de confiance de Student reste proche du nominal 95 %, même avec $n = 10$. Ce constat vaut également pour le graphique (d) représentant le cas d'une distribution bimodale. Par contre, la situation est moins bonne dans les deux derniers graphiques représentant des distributions asymétriques, où le niveau réel s'approche du nominal 95 % seulement à partir de $n = 30$ dans le graphique (e), et seulement à partir de $n = 500$ dans le graphique (f). Des méthodes *bootstrap* nous permettront dans certains cas d'améliorer la validité de nos intervalles de confiance[8].

de $\pm 1.96 \cdot 0.002 = \pm 0.004$, ce qui n'est pas si mal. Pour connaître exactement ce niveau réel, il faudrait une erreur type nulle et donc une infinité de simulations.

[8] Brièvement, les méthodes bootstrap consistent à « ré-échantillonner » (avec remise) par exemple 500 échantillons de taille n à partir de notre échantillon initial. On obtient ainsi 500 « pseudo-échantillons » de taille n. Dans chacun de ces pseudo-échantillons, on calcule une estimation $\widehat{\theta}$ du paramètre d'intérêt θ. On obtient ainsi 500 « pseudo-estimations ». Dans la méthode bootstrap originale, les quantiles 2.5 % et 97.5 % de ces 500 pseudo-estimations constituent les bornes d'un *intervalle de confiance bootstrap au niveau 95 % pour* θ. Cette méthode s'applique à (presque) n'importe quel paramètre et s'avère fort utile dans les cas

Notons par ailleurs qu'il sera difficile de vérifier la normalité des observations avec un n petit. On justifiera parfois la normalité en se référant à la littérature. Une autre étude aura peut-être analysé une variable semblable à la nôtre dans laquelle il y avait suffisamment de données pour vérifier la normalité. Si cette variable a été par exemple log-transformée dans cette étude, on pourra faire de même dans la nôtre.

4.5 Intervalle de confiance et intervalle de prédiction

Il est primordial de ne pas confondre les concepts d'intervalle de confiance et d'intervalle de prédiction. On a vu qu'un intervalle de confiance au niveau 95 % pour une moyenne μ se calcule (approximativement) par moyenne $\pm$ 2 erreurs types de la moyenne, c'est-à-dire par :

$$\widehat{\mu} \pm 2 \cdot \tilde{\sigma}/\sqrt{n}.$$

Rappelons par ailleurs que pour une variable normale, à peu près 95 % des observations se trouveront dans l'intervalle moyenne $\pm$ 2 écarts types, c'est-à-dire dans l'intervalle :

$$\widehat{\mu} \pm 2 \cdot \tilde{\sigma}.$$

Ce second intervalle est parfois appelé un *intervalle de prédiction au niveau 95 % pour une observation* car il y a une probabilité de 95 % qu'une observation tirée au hasard de la population se trouve dans cet intervalle[9].

De même, dans un abstract de publication scientifique, certains auteurs reportent, en plus de la moyenne de l'échantillon $\widehat{\mu}$, l'erreur type de la moyenne $\tilde{\sigma}/\sqrt{n}$, alors que d'autres reportent plus volontiers l'écart type $\tilde{\sigma}$ des observations. Bien que l'on puisse facilement passer de l'un à l'autre en multipliant ou divisant par $\sqrt{n}$, la distinction entre ces deux concepts est cruciale.

> Un intervalle de confiance nous informe sur la précision de l'estimation de la moyenne, un intervalle de prédiction sur la variabilité du phénomène étudié.

où l'on ne connaît pas de formule explicite pour calculer un intervalle de confiance. Des méthodes bootstrap plus sophistiquées nous permettent par ailleurs d'améliorer la validité des intervalles de confiance classiques, tel un intervalle de confiance pour une moyenne dans le cas où la distribution est asymétrique.

[9]Notons que l'on pourra calculer un intervalle de prédiction exact au niveau $1 - \alpha$ par $\widehat{\mu} \pm t_{1-\alpha/2,n-1} \cdot \tilde{\sigma}\sqrt{1 + 1/n}$. On reviendra sur ce point au chapitre 14.

Exemple 4.3 *Pour l'exemple des Onobrychis, où l'on avait* $n = 60$, $\widehat{\mu} = 21.7$ *cm,* $\tilde{\sigma} = 3.85$ *cm et* $\tilde{\sigma}/\sqrt{n} = 0.50$, *on peut calculer :*

- *intervalle de confiance au niveau* 95 % *pour la moyenne :*

$$21.7 \pm 2 \cdot 0.50 = [20.7; 22.7] \text{ cm}$$

- *intervalle de prédiction au niveau* 95 % *pour une observation :*

$$21.7 \pm 2 \cdot 3.85 = [14.0; 29.4] \text{ cm} .$$

Ces deux intervalles sont évidemment très différents l'un de l'autre. Le premier nous dit que la moyenne de la hauteur des Onobrychis est certainement comprise entre 20.7 *et* 22.7 *cm. Le second nous dit que presque tous les Onobrychis mesurent entre* 14.0 *et* 29.4 *cm.*

Rappelons que l'on n'a pas besoin de la normalité des observations pour qu'un intervalle de confiance pour une moyenne soit valide si n est suffisamment grand (grâce au théorème central limite), alors que la normalité des observations sera absolument nécessaire pour assurer la validité d'un intervalle de prédiction (même avec un grand n).

Notons également que la longueur d'un intervalle de confiance est d'autant plus petite que la taille de l'échantillon est grande. On aura en effet une meilleure précision de notre estimateur si on observe 10 000 *Onobrychis* plutôt que 10. Par contre, que l'on observe dix ou dix mille *Onobrychis* ne modifiera pas la variabilité du phénomène étudié, et donc la longueur d'un intervalle de prédiction ne va pas diminuer avec un grand n.

Lorsqu'on planifie une étude, une question importante est celle du choix de la taille n de l'échantillon. Une stratégie possible consiste à choisir n de telle manière que l'on obtienne une estimation suffisamment précise d'un paramètre d'intérêt[10]. On peut définir cette précision par la longueur L d'un intervalle de confiance au niveau 95 % pour ce paramètre, c'est-à-dire en gros par quatre fois l'erreur type de l'estimateur (pour autant qu'il soit sans biais et approximativement normalement distribué). Lorsque le paramètre d'intérêt est la moyenne, il s'agit donc de résoudre l'équation $L = 4\sigma/\sqrt{n}$, ce qui nous donne $n = (4\sigma/L)^2$. Un autre point à relever est que la longueur d'un intervalle de confiance pour une moyenne est proportionnelle non pas à n mais à $\sqrt{n}$. Cela implique que pour doubler la précision de l'estimation, il faut multiplier par quatre la taille de l'échantillon.

Exemple 4.4 *Si on désire avoir une estimation avec une précision de* ± 2 *cm pour la moyenne de la hauteur des Onobrychis (la longueur de l'intervalle de confiance pour cette moyenne étant ainsi de* $L = 4$ *cm), et si par ailleurs on peut penser que la variabilité des Onobrychis pourrait atteindre* $\sigma = 5$ *cm, il s'agira d'en planter* $n = (4 \cdot 5/4)^2 = 25$ *pour atteindre notre objectif. Si on désire par contre une précision de* ± 1 *cm (et donc* $L = 2$ *cm), il s'agira d'en planter* $n = (4 \cdot 5/2)^2 = 100$.

[10] On verra au chapitre 10 une autre stratégie pour le calcul de la taille d'un échantillon.

4.6 Transformation logarithmique

On a mentionné brièvement un exemple au chapitre 2 où l'on a mesuré le taux de créatine dans le sang pour $n = 31$ femmes atteintes d'une maladie génétique (on y reviendra plus en détail dans le chapitre 5). Sur l'échelle originale, on a une moyenne de 167.1 et un écart type de 162.9. En utilisant $t_{0.975,30} = 2.04$, on calcule un intervalle de confiance au niveau 95 % pour le taux moyen de créatine comme suit :

$$167.1 \pm 2.04 \cdot 162.9/\sqrt{31} = [107.4; 226.8].$$

Bien qu'il soit évident que ces données ne sont pas normales (si on soustrait deux écarts types à la moyenne afin de calculer la borne inférieure d'un intervalle de prédiction au niveau 95 %, on obtient une valeur très négative, alors qu'un taux de créatine est par définition positif), cet intervalle de confiance pourrait quand même être valide si l'asymétrie de la distribution n'était pas trop grande car la taille de l'échantillon n'est pas si petite (par contre un intervalle de prédiction au niveau 95 % de la forme $167.1 \pm 2 \cdot 162.9$ n'est donc pas valide).

Nous avions cependant recommandé d'effectuer l'analyse statistique sur l'échelle logarithmique, où l'on avait une moyenne de 4.71 et un écart type de 0.91. En admettant que la normalité soit atteinte sur cette échelle, on calcule un intervalle de confiance au niveau 95 % pour la moyenne de log-créatine comme suit :

$$4.71 \pm 2.04 \cdot 0.91/\sqrt{31} = [4.38; 5.04].$$

Rappelons que la quantité $\exp(4.71) = 111.1$ est ici un estimateur du taux médian de créatine sur l'échelle originale. L'analyse sur l'échelle logarithmique nous permet alors d'obtenir un intervalle de confiance pour le taux médian de créatine sur l'échelle originale en calculant l'exponentielle des bornes de l'intervalle de confiance ci-dessus :

$$[\exp(4.38); \exp(5.04)] = [79.8; 154.5].$$

On remarque en passant qu'en utilisant cette méthode, l'estimateur du taux médian (111.1) ne se trouve pas au centre de l'intervalle de confiance pour ce taux médian (un intervalle de confiance ne se calcule donc pas toujours en utilisant une formule du type $a \pm b$). L'analyse sur l'échelle logarithmique nous permet par ailleurs de calculer un intervalle de prédiction sur l'échelle originale comme suit : si on croit à la normalité sur l'échelle logarithmique, 95 % des valeurs de log-créatine seront dans l'intervalle :

$$4.71 \pm 2 \cdot 0.91 = [2.89; 6.53]$$

et donc 95 % des taux de créatine (mesurés sur l'échelle originale) seront dans l'intervalle :

$$[\exp(2.89); \exp(6.53)] = [18.0; 685.4].$$

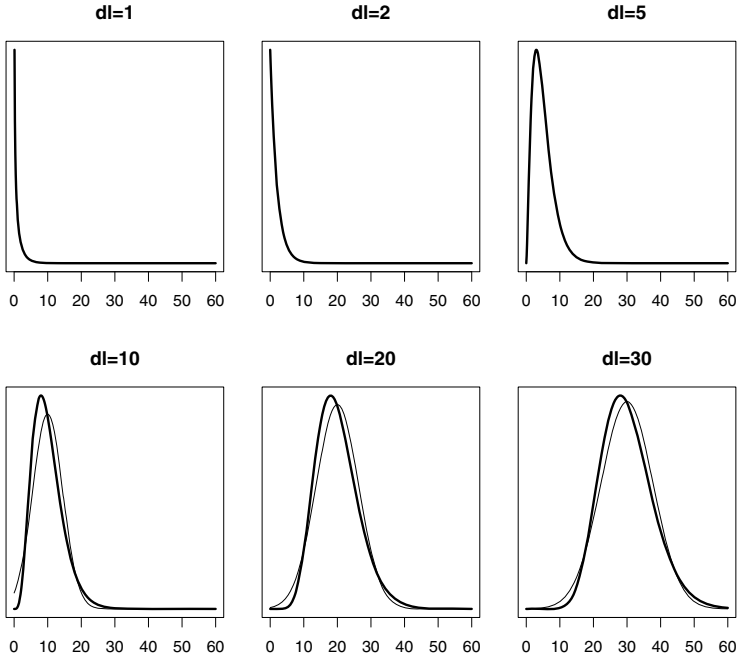

Figure 4.4 – Exemples de distributions du khi-deux avec différentes valeurs de dl.

4.7 Intervalle de confiance pour une variance

Si la distribution des observations Y_i est normale, on pourra non seulement calculer un intervalle de confiance exact pour la moyenne, mais également un intervalle de confiance exact pour la variance, comme expliqué ci-dessous (cette méthode n'étant cependant pas valide si les observations ne sont pas normales, même avec de grands échantillons). Si les observations sont normales, on connaît en effet mathématiquement la distribution de la variable aléatoire $(n-1)\tilde{\sigma}^2/\sigma^2$, que l'on appelle *distribution du khi-deux*. Il s'agit d'une nouvelle famille de distributions, asymétriques avec valeurs possibles positives, et indexée par un paramètre appelé ici aussi *degrés de liberté* et noté dl. Le nombre de degrés de liberté pour cette variable aléatoire $(n-1)\tilde{\sigma}^2/\sigma^2$ est le même que celui que l'on avait pour la variable aléatoire $\sqrt{n}\cdot(\widehat{\mu}-\mu)/\tilde{\sigma}$, c'est-à-dire $dl = n-1$.

La figure 4.4 nous montre des exemples de distributions du khi-deux. Comme on le voit, une distribution du khi-deux est relativement différente d'une distribution normale (représentée en arrière-plan par un trait fin sur ces graphiques), bien qu'elle commence à lui ressembler avec une grande valeur de dl. Dans ce texte, on notera par $\chi^2_{\alpha,dl}$ le quantile α d'une distribution du khi-deux avec dl degrés de liberté (où χ représente la lettre grecque « khi »). Le tableau A.3 (donné en annexe) nous donne quelques quantiles importants d'une distribu-

tion du khi-deux pour différentes valeurs de *dl*. Au besoin, un logiciel statistique nous donnera n'importe quel quantile de ces distributions[11].

On aura ainsi :

$$\Pr\left\{\chi^2_{\alpha/2,n-1} \leq \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}\right\} = 1-\alpha.$$

En manipulant ces inéquations de manière à ce que le paramètre d'intérêt σ^2 se retrouve au milieu, on obtient :

$$\Pr\left\{\frac{(n-1)\tilde{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)\tilde{\sigma}^2}{\chi^2_{\alpha/2,n-1}}\right\} = 1-\alpha.$$

Ainsi, l'intervalle donné par :

$$\left[\frac{(n-1)\tilde{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}; \frac{(n-1)\tilde{\sigma}^2}{\chi^2_{\alpha/2,n-1}}\right]$$

est un intervalle de confiance au niveau $1-\alpha$ pour la variance σ^2 qui sera exact si les observations sont normales. Un intervalle de confiance exact au niveau $1-\alpha$ pour l'écart type σ s'obtient en calculant la racine carrée de ces bornes.

Exemple 4.5 *Pour les Onobrychis, on avait* $\tilde{\sigma}^2 = 14.79$ *et* $n = 60$. *En supposant la normalité des observations (que l'on avait vérifiée au chapitre 2), et en utilisant* $\chi^2_{0.025,59} = 39.66$ *et* $\chi^2_{0.975,59} = 82.12$, *on calcule un intervalle de confiance au niveau* 95 % *pour la variance* σ^2 *comme suit :*

$$\left[\frac{59 \cdot 14.79}{82.12}; \frac{59 \cdot 14.79}{39.66}\right] = [10.63; 22.00].$$

Un intervalle de confiance au niveau 95 % *pour l'écart type* σ *se calcule alors par :*

$$\left[\sqrt{10.63}; \sqrt{22.0}\right] = [3.26; 4.69].$$

4.8 Intervalle de confiance de Wald pour une proportion

On a vu qu'une proportion π est un cas particulier d'une moyenne (la moyenne d'une variable binaire). À partir de l'estimateur $\widehat{\pi}$ de cette proportion, on peut donc appliquer la méthode de Wald et calculer un intervalle de confiance au niveau 95 % pour π comme suit : $\widehat{\pi} \pm 1.96 \cdot \mathrm{SE}(\widehat{\pi})$. Cependant, l'erreur type de $\widehat{\pi}$ dépend du véritable π :

$$\mathrm{SE}(\widehat{\pi}) = \sqrt{\pi(1-\pi)/n}.$$

[11] Dans R, on calcule par exemple $\chi^2_{0.95,1}$ par la commande `qchisq(0.95,1)` (et on trouve ainsi 3.84, voir le tableau A.3 donné en annexe).

En pratique, on remplace π par son estimateur $\widehat{\pi}$ dans la formule de $\mathrm{SE}(\widehat{\pi})$. Un intervalle de confiance de Wald au niveau 95 % pour π se calcule donc comme suit :

$$[\widehat{\pi} - 1.96 \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi})/n}; \widehat{\pi} + 1.96 \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi})/n}].$$

Plus généralement, un intervalle de confiance de Wald au niveau $1-\alpha$ pour π se calcule somme suit :

$$[\widehat{\pi} - z_{1-\alpha/2} \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi})/n}; \widehat{\pi} + z_{1-\alpha/2} \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi})/n}].$$

Exemple 4.6 *Reprenons l'exemple des* 527 *garçons parmi un total de* $n = 1000$ *nouveau-nés. On a ici une proportion empirique de garçons à la naissance de* $\widehat{\pi} = 0.527$*. Peut-on pour autant conclure sur la base de ces données qu'il naît statistiquement plus de garçons que de filles ? Pour répondre à cette question, on calcule un intervalle de confiance au niveau* 95 % *pour la véritable proportion* π *de garçons à la naissance. La solution de Wald nous donne ici :*

$$0.527 \pm 1.96 \cdot \sqrt{0.527 \cdot 0.473/1000} = [0.496; 0.558].$$

On ne peut donc pas exclure statistiquement la possibilité que la véritable proportion de garçons à la naissance soit de $\pi = 0.5$ *(cette valeur se trouvant à l'intérieur de l'intervalle de confiance).*

Lors de la planification d'une étude, on choisit parfois la taille n de l'échantillon dans le but d'obtenir une estimation suffisamment précise d'une proportion π. Si on mesure la précision de cette estimation par la longueur L d'un intervalle de confiance au niveau 95 % pour π, on aura (approximativement) $L = 4\sqrt{\pi(1-\pi)/n}$. Cette longueur sera maximale pour $\pi = 0.5$, égale dans ce cas à $L = 2/\sqrt{n}$, ce qui nous donne $n = (2/L)^2$.

Exemple 4.7 *Si on désire avoir une estimation d'une proportion avec une précision de* ± 0.01 *(la longueur de l'intervalle de confiance pour cette proportion étant alors de* $L = 0.02$*), il faudra inclure* $n = (2/0.02)^2 = 10\ 000$ *individus dans notre échantillon. Il s'agira par exemple d'interroger* $n = 10\ 000$ *personnes représentatives de la population des votants si on désire avoir une bonne chance de savoir qui va gagner les prochaines élections entre deux candidats.*

4.9 Intervalle de confiance de Wilson pour une proportion

En reprenant le raisonnement depuis le début, on va voir qu'il est possible d'obtenir une solution alternative pour calculer un intervalle de confiance pour π. Comme $\widehat{\pi}$ est un estimateur sans biais de π, approximativement normalement distribué et avec une erreur type de $\sqrt{\pi(1-\pi)/n}$, on a en effet :

$$\Pr\left\{\pi - 1.96 \cdot \sqrt{\pi(1-\pi)/n} \leq \widehat{\pi} \leq \pi + 1.96 \cdot \sqrt{\pi(1-\pi)/n}\right\} \approx 0.95$$

que l'on exprime parfois en standardisant $\widehat{\pi}$ de la sorte :

$$\Pr\left\{-1.96 \leq \sqrt{n} \cdot \frac{\widehat{\pi} - \pi}{\sqrt{\pi(1-\pi)}} \leq 1.96\right\} \approx 0.95.$$

Il se trouve que l'on peut manipuler ces inéquations de façon à ce que π se retrouve au milieu sans qu'il soit nécessaire d'estimer l'erreur type de $\widehat{\pi}$ (alors qu'on estimait cette erreur type avec la méthode de Wald). On laissera le lecteur vérifier qu'il s'agit pour cela de résoudre une équation du deuxième degré. On obtient de cette manière l'intervalle de confiance suivant :

$$\frac{\widehat{\pi} + 1.96^2/(2n)}{1 + 1.96^2/n} \pm \frac{1.96 \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi}) + 1.96^2/(4n)}}{\sqrt{n}(1 + 1.96^2/n)}.$$

Il s'agit d'un *intervalle de confiance de Wilson au niveau 95 % pour* π. Plus généralement, un intervalle de confiance de Wilson au niveau $1-\alpha$ pour π se calcule comme suit[12] :

$$\frac{\widehat{\pi} + z^2_{1-\alpha/2}/(2n)}{1 + z^2_{1-\alpha/2}/n} \pm \frac{z_{1-\alpha/2} \cdot \sqrt{\widehat{\pi}(1-\widehat{\pi}) + z^2_{1-\alpha/2}/(4n)}}{\sqrt{n}(1 + z^2_{1-\alpha/2}/n)}.$$

Pour un grand n, un intervalle de confiance de Wilson pour π sera proche d'un intervalle de confiance de Wald pour π. Il est à noter que ni l'un ni l'autre ne sera exact. Ces intervalles de confiance seront cependant valides si n est grand et si π n'est pas trop proche de 0 ou de 1. En pratique, ils seront valides si $n\pi \geq 5$ et si $n(1-\pi) \geq 5$. Si ces conditions ne sont pas respectées, la solution de Wilson sera en général préférable à la solution de Wald. Avec $n = 10$, on peut calculer que le niveau réel d'un intervalle de confiance au niveau nominal 95 % pour π lorsque $\pi = 0.05, 0.1, 0.15$ et 0.2 sera par exemple de 40 %, 65 %, 79 % et 89 % s'il est calculé selon la méthode de Wald, et de 91 %, 93 %, 95 % et 97 % (et donc bien plus proche du nominal 95 %) s'il est calculé selon la méthode de Wilson[13].

Exemple 4.8 *Considérons un exemple où l'on mesure le succès d'une opération avec une variable binaire (1 = succès ; 0 = échec) pour* $n = 10$ *patients. Il peut y avoir ainsi* $0, 1, 2, \cdots, 9$ *ou* 10 *opérations réussies, correspondant aux* 11 *valeurs possibles* $\widehat{\pi} = 0, 0.1, 0.2, \cdots, 0.9, 1$. *Les intervalles de confiance au niveau* 95 % *pour la véritable proportion de succès* π *de cette opération calculés selon les méthodes de Wald ou de Wilson sont donnés dans le tableau ci-dessous pour chaque valeur possible de* $\widehat{\pi}$ *:*

[12] Cet intervalle porte le nom de son inventeur. Voir l'article de Wilson (1927).

[13] Pourtant, au contraire d'un intervalle de confiance de Student pour une moyenne, un intervalle de confiance de Wilson pour une proportion n'est pas très connu du grand public, peut-être parce que les proportions sont souvent estimées à partir de grands échantillons, pour lesquels la solution de Wald est valide. Pour de petits échantillons, la solution de Wilson mérite cependant d'être connue. De plus, il s'agit de l'intervalle de confiance en dualité avec le fameux test du khi-deux pour une proportion, comme on le verra au chapitre 8.

$\widehat{\pi}$	95 % *CI Wald*	95 % *CI Wilson*	95 % *CI Wald modifié*
$0/10 = 0$	$[0; 0]$	$[0.00; 0.28]$	$[-0.04; 0.33]$
$1/10 = 0.1$	$[-0.09; 0.29]$	$[0.02; 0.40]$	$[0.00; 0.43]$
$2/10 = 0.2$	$[-0.05; 0.45]$	$[0.07; 0.51]$	$[0.05; 0.52]$
$3/10 = 0.3$	$[0.02; 0.58]$	$[0.11; 0.60]$	$[0.11; 0.61]$
$4/10 = 0.4$	$[0.10; 0.70]$	$[0.17; 0.69]$	$[0.17; 0.69]$
$5/10 = 0.5$	$[0.19; 0.81]$	$[0.24; 0.76]$	$[0.24; 0.76]$
$6/10 = 0.6$	$[0.30; 0.90]$	$[0.31; 0.83]$	$[0.31; 0.83]$
$7/10 = 0.7$	$[0.42; 0.98]$	$[0.40; 0.89]$	$[0.39; 0.89]$
$8/10 = 0.8$	$[0.55; 1.05]$	$[0.49; 0.94]$	$[0.48; 0.95]$
$9/10 = 0.9$	$[0.71; 1.09]$	$[0.60; 0.98]$	$[0.57; 1.00]$
$10/10 = 1$	$[1; 1]$	$[0.72; 1.00]$	$[0.67; 1.04]$

On notera que la borne inférieure de l'intervalle de confiance calculé selon la méthode de Wald est plus petite que 0 *lorsque* $\widehat{\pi} = 0.1$ *ou* 0.2*, et que sa borne supérieure est plus grande que* 1 *lorsque* $\widehat{\pi} = 0.8$ *ou* 0.9*. Un intervalle de confiance calculé selon la méthode de Wilson ne partage pas ces problèmes, ses bornes étant toujours comprises entre* 0 *et* 1*.*

On voit également que la solution de Wald se résume en un seul point lorsque $\widehat{\pi} = 0$ *ou* $\widehat{\pi} = 1$*, suggérant à tort que l'on a prouvé statistiquement que la véritable proportion* π *est de* 0 %*, respectivement de* 100 %*. Il s'agit de cas où la solution de Wald n'est pas valide : il ne suffit pas que la proportion empirique soit de* 100 % *pour pouvoir conclure statistiquement que la véritable proportion est de* 100 %*. La solution de Wilson donnée par* $[0.72; 1.00]$ *est ici plus crédible : avec* 10 *opérations réussies sur* 10 *patients, on peut conclure statistiquement que la véritable proportion de succès est au moins de* 72 %*.*

Notons qu'un intervalle de confiance de Wilson au niveau 95 % *peut se calculer approximativement en utilisant la formule de Wald, après avoir augmenté l'échantillon de deux fois la valeur* 1 *et de deux fois la valeur* 0*. Par exemple, dans le cas où l'échantillon original est constitué de* 10 *fois la valeur* 1 *(et de* 0 *fois la valeur* 0*), on applique la formule de Wald à un échantillon constitué de* 12 *fois la valeur* 1 *et de* 2 *fois la valeur* 0*, c'est-à-dire à un échantillon de* $n = 14$ *observations pour lequel on a* $\widehat{\pi} = 12/14 = 0.857$*, et on obtient :*

$$0.857 \pm 1.96 \cdot \sqrt{0.857 \cdot 0.143/14} = [0.67; 1.04].$$

Il s'agit de la solution donnée dans la colonne « 95 % CI Wald modifié » du tableau ci-dessus, qui est effectivement relativement proche de la solution de Wilson. Notons que cette approximation fonctionne seulement avec un niveau proche de 95 %[14]*.*

[14] On pourra lire à ce sujet l'article de Agresti et Caffo (2000).

Chapitre 5

Comparaison de deux distributions

Dans certaines études, on ne dispose pas d'un seul échantillon de données mais de deux échantillons, provenant de deux populations différentes. L'intérêt principal de l'étude est alors la comparaison des distributions de deux variables, l'une définie dans la première population, l'autre dans la seconde. Pour illustrer cela, on reprend un exemple d'*Onobrychis*. Les 60 *Onobrychis* introduits au chapitre 1 ont été cultivés avec un niveau nutritif faible et ont atteint les hauteurs suivantes :

21	21	23	22	23	29	24	21	18	23	19	18	20	24	20
20	19	19	22	21	18	20	23	17	20	25	23	21	14	18
29	28	28	14	28	26	22	22	22	29	19	26	16	17	23
18	25	22	20	22	18	32	26	21	20	27	20	19	19	18

Dans cette même étude, 60 autres *Onobrychis* ont été cultivés avec un niveau nutritif élevé et ont atteint les hauteurs suivantes :

34	38	32	29	24	26	30	30	31	32	30	33	39	35	24
30	32	31	36	25	47	26	36	30	35	32	29	37	33	32
29	25	30	28	31	28	30	28	26	30	25	38	25	35	31
26	32	31	29	23	27	40	28	35	32	27	33	29	36	23

Ces deux échantillons proviennent de deux populations d'*Onobrychis*, l'une cultivée avec un niveau nutritif élevé, l'autre avec un niveau nutritif faible. On comparera ainsi les distributions des deux variables suivantes[1] :

[1]Notons que, dans cet exemple, les deux variables ont en fait la même définition (il s'agit de la hauteur d'un *Onobrychis* après une culture de six mois). Dans ce cas, la comparaison des distributions de deux variables peut aussi se voir comme la comparaison de la distribution d'une même variable entre deux populations (deux groupes) d'*Onobrychis*.

- Y_1 : hauteur après six mois dans une population d'*Onobrychis* cultivés avec un niveau nutritif élevé
- Y_0 : hauteur après six mois dans une population d'*Onobrychis* cultivés avec un niveau nutritif faible.

On va voir dans ce chapitre comment on peut comparer statistiquement les distributions de ces deux variables. On abordera comme d'habitude le problème tout d'abord du point de vue de la statistique descriptive (afin de résumer la différence entre ces deux distributions), puis de la statistique inférentielle (afin de généraliser aux deux populations les différences observées entre les deux échantillons). On supposera dans tout ce chapitre que les observations des deux échantillons sont indépendantes. En particulier, les individus du premier échantillon ne seront pas les mêmes que les individus du second.

5.1 Différence de moyenne

On a vu au chapitre 2 que la moyenne est un paramètre couramment utilisé pour résumer la distribution d'une variable continue. Si deux variables continues Y_1 et Y_0 sont résumées par leurs moyennes μ_1 et μ_0, il apparaît dès lors naturel de comparer les distributions de ces deux variables *via* la *différence de moyenne* :

$$\Delta = \mu_1 - \mu_0.$$

On a vu également que l'utilisation de la moyenne pour résumer une distribution continue fait particulièrement sens pour une variable normale. On va voir dans ce chapitre que l'utilisation de la différence de moyenne Δ pour résumer la différence entre deux distributions continues fait particulièrement sens dans le cas où *les deux variables sont normales avec une même variance*.

En effet, si les variances sont les mêmes dans un cadre normal, le paramètre Δ ne sera pas seulement un résumé mais *caractérisera* la différence entre les deux distributions. En particulier, une différence de moyenne nulle ($\Delta = 0$) impliquera alors forcément deux distributions identiques. On n'aura par contre pas de telle implication si les variances sont différentes, cas pour lequel le paramètre Δ ne sera qu'un résumé imparfait et incomplet de la situation.

La figure 5.1 nous montre des situations très différentes avec pourtant une même différence de moyenne (ici $\Delta = 1$) entre deux distributions normales. On voit ainsi que la différence entre les quantiles 95 % des deux distributions vaut par exemple -0.2, 2.6 et 4.3 sur les graphiques (b), (c) et (d). Si on ne suppose pas une variance identique pour les deux distributions, le fait d'avoir $\Delta = 1$ ne nous informe donc en rien sur la différence entre les quantiles 95 % de ces deux distributions, pas même sur le signe de cette différence (le raisonnement étant le même pour un autre quantile). Si on suppose par contre une même variance pour les deux distributions, alors une différence entre moyennes de $\Delta = 1$ impliquera également une différence entre quantiles de $\Delta = 1$, comme illustré sur le graphique (a).

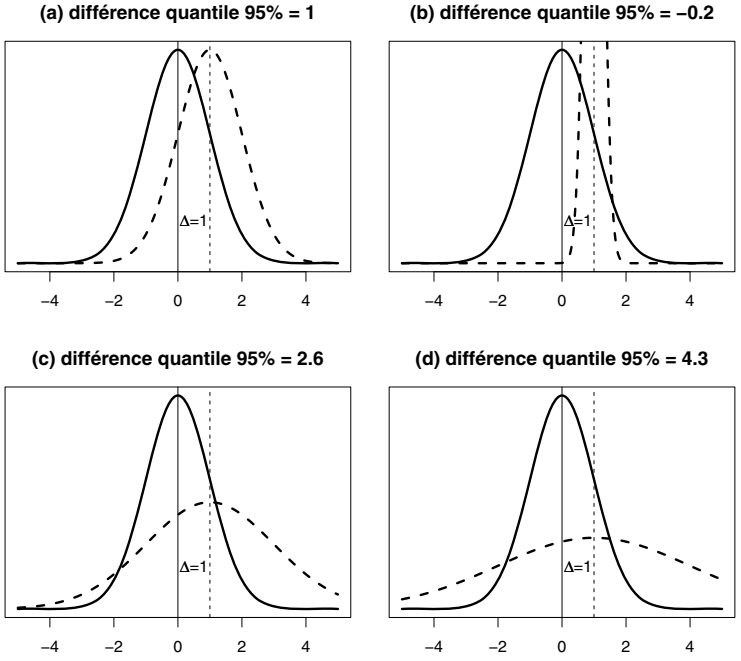

Figure 5.1 – Différentes situations avec deux distributions normales et $\Delta = 1$.

On verra à partir du chapitre 13 qu'un modèle statistique est un ensemble d'hypothèses à propos des distributions desquelles proviennent les observations. Lorsque l'on a deux échantillons de données, le « modèle idéal » consiste donc à supposer que les deux distributions sont normales avec une même variance, comme dans le graphique (a) de la figure 5.1.

> On appellera modèle idéal la situation où l'on peut supposer une distribution normale et une même variance dans les deux populations.

À partir de deux échantillons d'observations des variables Y_1 et Y_0 avec moyennes μ_1 et μ_0 dans les populations, on estimera $\Delta = \mu_1 - \mu_0$ par :

$$\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0$$

où $\widehat{\mu}_1$ et $\widehat{\mu}_0$ dénotent les moyennes empiriques calculées sur ces échantillons. ll se trouve que $\widehat{\Delta}$ est un estimateur sans biais et approximativement normal de Δ. Cela découle du fait que $\widehat{\mu}_1$ et $\widehat{\mu}_0$ sont eux-mêmes des estimateurs sans biais et approximativement normaux de μ_1 et μ_0.

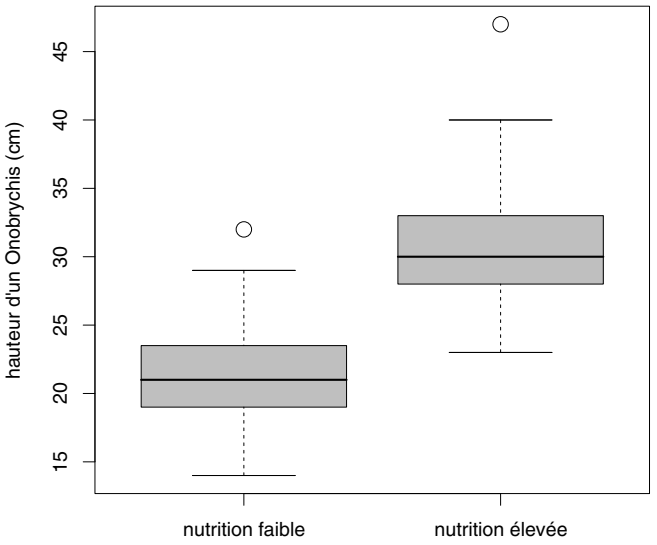

Figure 5.2 – Exemple de données satisfaisant approximativement le modèle idéal.

Exemple 5.1 *Les boxplots de la figure 5.2 nous permettent de comparer la hauteur des deux groupes d'Onobrychis, cultivés avec un niveau nutritif faible et élevé, introduits en début de chapitre. Ces boxplots suggèrent que l'on est proche du modèle idéal : on a une variabilité similaire et une normalité approximative dans les deux groupes (de sorte que* Δ *est un excellent résumé de la situation). On a les résultats suivants :* $\widehat{\mu}_1 = 30.8$ *cm,* $\widehat{\mu}_0 = 21.7$ *cm et donc* $\widehat{\Delta} = 30.8 - 21.7 = 9.1$ *cm. On estime ainsi une différence de moyenne de* 9.1 *cm en faveur du groupe avec niveau nutritif élevé. Notons que dans le contexte de cette étude, ce paramètre* Δ *peut être interprété comme l'effet (moyen) du niveau nutritif sur la croissance d'un Onobrychis.*

5.2 Intervalle de confiance de Wald pour une différence de moyenne

La différence de moyenne empirique $\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0$ est donc un estimateur sans biais et approximativement normal de la véritable différence de moyenne $\Delta = \mu_1 - \mu_0$. On peut dès lors utiliser la méthode de Wald pour calculer un intervalle de confiance pour Δ. L'erreur type de $\widehat{\Delta}$ est égale à :

$$\mathrm{SE}(\widehat{\Delta}) = \sqrt{\sigma_1^2/n_1 + \sigma_0^2/n_0}$$

où σ_1^2 et σ_0^2 dénotent les variances des variables Y_1 et Y_0 dans les populations, et où n_1 et n_0 dénotent les tailles des échantillons. En remplaçant les variances σ_1^2 et σ_0^2 par leurs estimateurs $\tilde{\sigma}_1^2$ et $\tilde{\sigma}_0^2$ calculés dans ces échantillons, on obtient un intervalle de confiance de Wald au niveau $1-\alpha$ pour Δ comme suit :

$$\widehat{\Delta} \pm z_{1-\alpha/2} \cdot \sqrt{\tilde{\sigma}_1^2/n_1 + \tilde{\sigma}_0^2/n_0}.$$

Cet intervalle de confiance sera valide si n_1 et n_0 sont assez grands.

Exemple 5.2 *Dans l'exemple des Onobrychis, on a les résultats suivants :*

- $\widehat{\mu}_1 = 30.8$ *cm,* $\tilde{\sigma}_1^2 = 21.45$ *cm*^{2}*,* $n_1 = 60$
- $\widehat{\mu}_0 = 21.7$ *cm,* $\tilde{\sigma}_0^2 = 14.8$ *cm*^{2}*,* $n_0 = 60$
- $\widehat{\Delta} = 30.8 - 21.7 = 9.1$ *cm*
- 95 % *CI pour* Δ *:*

$$9.1 \pm 1.96 \cdot \sqrt{\frac{21.45}{60} + \frac{14.8}{60}} = [7.6; 10.6] \text{ cm}$$

→ *la véritable différence de moyenne entre les deux groupes d'Onobrychis se trouve avec une probabilité de* 95 % *entre* 7.6 *et* 10.6 *cm*

→ *on a prouvé statistiquement que l'effet du niveau nutritif sur la croissance d'un Onobrychis après six mois est au moins de* 7.6 *cm.*

5.3 Intervalle de confiance de Student pour une différence de moyenne

De façon analogue à ce que l'on a vu au chapitre précédent dans le cas d'un intervalle de confiance de Wald pour une moyenne, le remplacement des variances σ_1^2 et σ_0^2 par leurs estimateurs $\tilde{\sigma}_1^2$ et $\tilde{\sigma}_0^2$ lors du calcul d'un intervalle de confiance de Wald pour une différence de moyenne Δ n'est pas anodin. L'intervalle de confiance obtenu ne sera en effet pas valide avec de petits n_1 ou n_0. Par contre, l'intervalle de confiance de Student pour Δ que l'on introduit ci-dessous sera exact quelles que soient les tailles d'échantillon, pour autant que les hypothèses du modèle idéal soient satisfaites (à savoir, la normalité et une même variance dans les deux populations). On aura alors $\sigma_1^2 = \sigma_0^2 = \sigma^2$ et on utilisera l'estimateur sans biais suivant de cette variance commune :

$$\tilde{\sigma}^2 = \frac{(n_1 - 1)\tilde{\sigma}_1^2 + (n_0 - 1)\tilde{\sigma}_0^2}{n_1 + n_0 - 2}$$

(cet estimateur $\tilde{\sigma}^2$ étant une moyenne des estimateurs $\tilde{\sigma}_1^2$ et $\tilde{\sigma}_0^2$, pondérée par l'importance des échantillons). L'hypothèse d'une variance commune implique par ailleurs :

$$\mathrm{SE}(\widehat{\Delta}) = \sqrt{\sigma^2/n_1 + \sigma^2/n_0} = \sigma \cdot \sqrt{\frac{n_1 + n_0}{n_1 n_0}}.$$

Alors que la variable aléatoire :

$$\frac{\widehat{\Delta} - \mathrm{E}(\widehat{\Delta})}{\mathrm{SE}(\widehat{\Delta})} = \sqrt{\frac{n_1 n_0}{n_1 + n_0}} \cdot \frac{\widehat{\Delta} - \Delta}{\sigma}$$

a une distribution normale standardisée, il se trouve que la variable aléatoire légèrement modifiée :

$$\sqrt{\frac{n_1 n_0}{n_1 + n_0}} \cdot \frac{\widehat{\Delta} - \Delta}{\tilde{\sigma}}$$

a une distribution de Student avec $n_1 + n_0 - 2$ *dl*. On calcule ainsi un intervalle de confiance de Student au niveau $1 - \alpha$ pour Δ comme suit :

$$\widehat{\Delta} \pm t_{1-\alpha/2, n_1+n_0-2} \cdot \tilde{\sigma} \cdot \sqrt{\frac{n_1 + n_0}{n_1 n_0}}.$$

Cet intervalle de confiance sera exact sous les hypothèses du modèle idéal. Si on s'écarte du modèle idéal (non-normalité et/ou variances différentes), cet intervalle de confiance sera valide pour autant que n_1 et n_0 soient à la fois grands et proches l'un de l'autre (auquel cas, un intervalle de Student sera proche d'un intervalle de confiance de Wald)[2].

Exemple 5.3 *Dans l'exemple des Onobrychis, où les boxplots de la figure 5.2 nous suggéraient une situation proche du modèle idéal, on a* $n_1 = n_0 = 60$, $\widehat{\Delta} = 9.1$ *cm*, $\tilde{\sigma}_1^2 = 21.45$ *cm*^{2}, $\tilde{\sigma}_2^2 = 14.8$ *cm*^{2} *et* $\tilde{\sigma}^2 = (21.45 + 14.8)/2 = 18.1$ *cm*^{2}. *On calcule un intervalle de confiance de Student au niveau* 95 % *pour* Δ *comme suit (en utilisant* $t_{0.975,118} = 1.98$*) :*

$$9.1 \pm 1.98 \cdot \sqrt{18.1} \cdot \sqrt{\frac{60 + 60}{60 \cdot 60}} = [7.6; 10.6] \text{ cm} .$$

Cet intervalle de confiance de Student est très proche (et même identique si on ne retient qu'une décimale dans les calculs) à l'intervalle de confiance de Wald calculé ci-dessus (car n_1 *et* n_0 *sont à la fois grands et proches l'un de l'autre).*

[2]Dans le cas $n_1 = n_0$, on aura $\tilde{\sigma}^2 = (\tilde{\sigma}_1^2 + \tilde{\sigma}_0^2)/2$ et l'estimateur de $\mathrm{SE}(\widehat{\Delta})$ sera donc identique que l'on suppose une même variance ou des variances différentes dans les populations. Cela explique pourquoi un intervalle de confiance de Student sera proche d'un intervalle de confiance de Wald lorsque n_1 et n_0 sont à la fois grands et proches l'un de l'autre.

5.4 Intervalle de confiance de Welch pour une différence de moyenne

On va voir à présent une troisième méthode pour calculer un intervalle de confiance pour Δ, à savoir la méthode de Welch qui combine les idées de Wald et de Student. L'idée est d'utiliser la formule de Wald mais avec des quantiles d'une distribution de Student (plutôt que des quantiles d'une distribution normale standardisée). Un *intervalle de confiance de Welch au niveau* $1-\alpha$ *pour* Δ se calcule en effet comme suit[3] :

$$\widehat{\Delta} \pm t_{1-\alpha/2,dl} \cdot \sqrt{\tilde{\sigma}_1^2/n_1 + \tilde{\sigma}_0^2/n_0}$$

où le nombre dl de degrés de liberté est ici défini par :

$$dl = \frac{\left(\frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0}\right)^2}{\frac{\tilde{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\tilde{\sigma}_0^4}{n_0^2(n_0-1)}}.$$

Avec de grands n_1 et n_0, on aura également un grand dl (on a $dl \approx 114$ dans l'exemple des *Onobrychis*), de sorte que la méthode de Welch sera proche de la méthode de Wald. Par ailleurs, dans le cas particulier $\tilde{\sigma}_1^2 = \tilde{\sigma}_0^2$ et $n_1 = n_0$, on aura $dl = n_1 + n_0 - 2$, de sorte que la méthode de Welch coïncidera avec la méthode de Student.

5.5 Validité des intervalles de confiance pour différence de moyenne

La validité d'un intervalle de confiance de Welch pour une différence de moyenne est en général supérieure à la validité d'un intervalle de confiance de Wald, et donc la méthode de Welch est en général préférable à la méthode de Wald. Le choix entre la méthode de Welch et celle de Student est par contre plus délicat. Le tableau ci-dessous nous donne une idée générale de la situation (comme d'habitude, la limite entre petits et grands échantillons dépend de la distribution des observations ; au besoin, des simulations nous permettront de préciser ces limites) :

		normalité		non-normalité	
		$\sigma_1^2 \approx \sigma_0^2$	$\sigma_1^2 \neq \sigma_0^2$	$\sigma_1^2 \approx \sigma_0^2$	$\sigma_1^2 \neq \sigma_0^2$
CI Student	n_1 ou n_0 petits	exact	non valide	non valide	non valide
	n_1 et n_0 grands, $n_1 \neq n_0$	exact	non valide	valide	non valide
	n_1 et n_0 grands, $n_1 \approx n_0$	exact	valide	valide	valide
CI Welch	n_1 ou n_0 petits	valide	valide	non valide	non valide
	n_1 et n_0 grands, $n_1 \neq n_0$	valide	valide	valide	valide
	n_1 et n_0 grands, $n_1 \approx n_0$	valide	valide	valide	valide

[3]Cet intervalle porte le nom de son inventeur. Voir l'article de Welch (1938).

On voit que le domaine de validité de la méthode de Welch est plus large que celui de la méthode de Student. La méthode de Welch est en effet valide dans beaucoup de situations, à la fois avec de petits échantillons (si toutefois la normalité des observations peut être supposée) et/ou avec des variances différentes dans les deux populations. Cette méthode ne sera cependant pas valide avec des distributions non normales et au moins un petit échantillon (cas dans lequel la méthode de Student ne sera pas valide non plus).

Ainsi, la méthode de Welch devrait être la méthode de choix. Pourtant, en pratique et dans la littérature, on a souvent tendance à croire au modèle idéal (que l'on atteindra parfois par transformation des données) et à calculer un intervalle de confiance de Student, qui sera alors exact et qui aura par ailleurs l'avantage d'être plus facilement généralisable à des problèmes de modélisation statistique plus complexes, tels ceux que nous verrons lorsque nous traiterons de régression (à partir du chapitre 13). Lorsque les variances des deux populations sont différentes et que les échantillons ont des tailles différentes, il est toutefois important de savoir qu'un intervalle de confiance de Student ne sera pas valide (même avec de grands échantillons et des observations normales)[4]. Nous reviendrons sur la question du choix entre les méthodes de Welch et de Student au chapitre 11.

5.6 Transformation logarithmique

Lorsque les hypothèses du modèle idéal ne sont pas satisfaites, on peut essayer de s'en rapprocher en appliquant une tranformation logarithmique. Il n'est en effet pas rare qu'une telle transformation nous permette de *stabiliser la variance* (la rendre comparable dans les deux groupes) en même temps que de nous rapprocher de la normalité.

On a vu au chapitre 2 que l'exponentielle d'une moyenne calculée sur l'échelle logarithmique (pour autant que les données soient à peu près symé-

[4]Si les variances et les tailles d'échantillon sont différentes, un intervalle de confiance de Student sera libéral si c'est le petit échantillon qui provient de la population avec la plus grande variabilité, et il sera conservateur dans le cas contraire. Pour être plus précis, si on veut que le niveau réel d'un intervalle de confiance de Student au niveau nominal $1-\alpha$ pour Δ soit au plus de $1-\alpha^-$ et au moins de $1-\alpha^+$ (où $\alpha^- < \alpha < \alpha^+$), il faudra que la condition suivante soit satisfaite (on suppose ici que n_1 et n_0 sont assez grands) :

$$\frac{z_{1-\alpha/2}}{z_{1-\alpha^-/2}} \leq \sqrt{\frac{1+(n_1/n_0)(\sigma_0^2/\sigma_1^2)}{n_1/n_0+\sigma_0^2/\sigma_1^2}} \leq \frac{z_{1-\alpha/2}}{z_{1-\alpha^+/2}}.$$

Dans le cas où l'on choisit $\alpha = 0.05$, $\alpha^- = 0.035$ et $\alpha^+ = 0.065$, ce qui implique $z_{1-\alpha/2} = 1.96$, $z_{1-\alpha^-/2} = 2.11$ et $z_{1-\alpha^+/2} = 1.845$, cette condition sera satisfaite si :

$$0.93 \leq \sqrt{\frac{1+(n_1/n_0)(\sigma_0^2/\sigma_1^2)}{n_1/n_0+\sigma_0^2/\sigma_1^2}} \leq 1.06$$

par exemple si $(1/1.2) \leq n_1/n_0 \leq 1.2$ et si $0.5 \leq \sigma_0/\sigma_1 \leq 2$, c'est-à-dire si la taille du grand échantillon n'excède pas de plus de 20 % la taille du petit, et si le plus grand des écarts types ne vaut pas plus du double du plus petit (dans les populations).

triques sur cette échelle) nous donne une estimation de la médiane sur l'échelle originale. On va voir à présent que l'exponentielle d'une différence de moyenne calculée sur l'échelle logarithmique (pour autant que les données soient à peu près symétriques sur cette échelle) nous donne une estimation du *quotient des médianes* sur l'échelle originale. On a en effet :

$$\begin{aligned}
\exp(\Delta) &= \exp(mean(\log(Y_1)) - mean(\log(Y_0))) \\
&\approx \exp(median(\log(Y_1)) - median(\log(Y_0))) \\
&= \exp(\log(median(Y_1)) - \log(median(Y_0))) \\
&= \exp\left(\log\left(median(Y_1)/median(Y_0)\right)\right) \\
&= median(Y_1)/median(Y_0).
\end{aligned}$$

Exemple 5.4 *On considère les données d'une étude où l'on compare le taux de créatine entre deux groupes de femmes, un groupe de $n_1 = 31$ femmes atteintes d'une maladie génétique et un groupe contrôle de $n_0 = 39$ femmes non atteintes de cette maladie*[5]*. Un des buts de cette étude est de montrer qu'un taux anormalement élevé de créatine dans le sang pourrait être un symptôme de cette maladie. Les boxplots du graphique du haut de la figure 5.3 comparent les deux groupes sur l'échelle originale. On y voit non seulement des distributions manifestement asymétriques vers la droite, mais également une grande différence de variabilité dans les deux groupes, à tel point que l'échelle utilisée sur ce graphique n'apparaît pas du tout adaptée au groupe contrôle (que l'on distingue à peine). On est ici loin du modèle idéal.*

Les boxplots du graphique du bas de cette figure comparent les deux groupes après une transformation logarithmique de ces données. Cette échelle logarithmique convient à présent aux deux groupes, la variance a été en grande partie stabilisée (même si elle demeure plus grande dans le groupe maladie), la normalité a été approchée et les valeurs extrêmes ont été en grande partie réconciliées avec les données. Il est donc ici recommandé d'effectuer l'analyse statistique sur l'échelle logarithmique, où l'on a les résultats suivants :

- $\widehat{\mu}_1 = 4.71$, $\tilde{\sigma}_1^2 = 0.83$, $n_1 = 31$
- $\widehat{\mu}_0 = 3.66$, $\tilde{\sigma}_0^2 = 0.20$, $n_0 = 39$
- $\widehat{\Delta} = 4.71 - 3.66 = 1.05$
- $\tilde{\sigma}^2 = \frac{30 \cdot 0.83 + 38 \cdot 0.20}{68} = 0.48$
- *on calcule un intervalle de confiance de Student au niveau* 95 % *pour* Δ

[5] Il s'agit de l'ensemble de données No 38 du livre de Andrews et Herzberg (1985). Pour le groupe contrôle, nous avons considéré les 39 femmes de la page 224 pour lesquelles quatre paramètres sanguins ont été mesurés. Pour le groupe maladie, nous avons considéré les 31 femmes de la page 227 pour lesquelles quatre paramètres sanguins ont été mesurés, et âgées de moins de 50 ans.

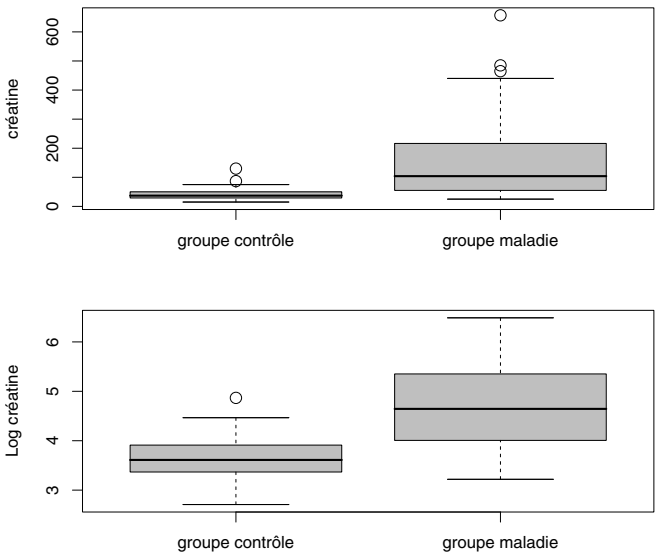

Figure 5.3 – Comparaison de deux échantillons sur échelle originale et logarithmique.

comme suit (en utilisant $t_{0.975,68} = 2.00$*) :*

$$1.05 \pm 2.00 \cdot \sqrt{0.48} \cdot \sqrt{\frac{31+39}{31 \cdot 39}} = [0.72; 1.38]$$

- *pour calculer un intervalle de confiance de Welch au niveau* 95 % *pour* Δ*, il s'agit tout d'abord de calculer :*

$$dl = \frac{\left(\frac{0.83}{31} + \frac{0.20}{39}\right)^2}{\frac{0.83^2}{31^2 \cdot 30} + \frac{0.20^2}{39^2 \cdot 38}} = 41.4$$

(que l'on approximera par 41*) et d'utiliser ensuite* $t_{0.975,41} = 2.02$ *pour obtenir :*

$$1.05 \pm 2.02 \cdot \sqrt{\frac{0.83}{31} + \frac{0.20}{39}} = [0.69; 1.41]$$

$\rightarrow$ *bien que les variances estimées* $\tilde{\sigma}_1^2$ *et* $\tilde{\sigma}_0^2$ *demeurent assez différentes sur l'échelle logarithmique, les calculs de ces deux intervalles de confiance donnent des résultats semblables (car on a* $n_1 \approx n_0$*)*

$\rightarrow$ *que l'on utilise l'un ou l'autre de ces intervalles, on a prouvé statistiquement que le taux de créatine est plus élevé dans le groupe maladie.*

On a par ailleurs :

- $\exp(\widehat{\Delta}) = \exp(1.05) = 2.88$

  → *la différence de moyenne de* 1.05 *sur l'échelle logarithmique correspond à un taux médian de créatine* 2.88 *fois plus élevé dans le groupe maladie que dans le groupe contrôle*

- *un intervalle de confiance au niveau* 95 % *pour* $\exp(\Delta)$ *s'obtient en calculant l'exponentielle des bornes d'un intervalle de confiance au niveau* 95 % *pour* Δ *:*

$$[\exp(0.72); \exp(1.38)] = [2.05; 3.97]$$

  → *notons que l'estimateur du quotient des médianes (*2.88*) ne se trouve pas au centre de l'intervalle de confiance*

  → *on a donc prouvé statistiquement que le taux médian de créatine est au moins* 2.05 *fois plus élevé dans le groupe maladie que dans le groupe contrôle.*

5.7 Différence de moyenne standardisée

On a vu que le modèle idéal (normalité et même variance dans les deux populations) permet d'une part de donner son plein sens au paramètre de statistique descriptive Δ et d'autre part de calculer un intervalle de confiance exact pour Δ, y compris avec de petits échantillons. On va voir dans cette section un troisième avantage du modèle idéal : il nous permet de standardiser une différence de moyenne Δ par l'écart type commun σ, en calculant une *différence de moyenne standardisée* comme suit :

$$\delta = \Delta/\sigma.$$

On estimera δ par $\widehat{\delta} = \widehat{\Delta}/\tilde{\sigma}$. On exprime ainsi une différence de moyenne non plus dans l'unité originale des variables, mais dans l'unité du statisticien : en nombre d'écarts types.

La figure 5.4 nous montre des boxplots (fictifs) calculés dans deux groupes satisfaisant le modèle idéal avec différentes valeurs de δ. Dans le graphique (a), la différence de moyenne entre les deux groupes correspond à $\delta = 0.25$ (un quart d'écart type). Sachant qu'un écart type vaut en gros le quart de l'étendue, la différence entre les deux groupes correspond donc à un seizième de l'étendue dans un groupe, ce qui visuellement ne semble pas grand chose. Une telle différence est souvent considérée comme une *petite différence* dans la pratique et dans la littérature[6]. Dans le graphique (b), on a une différence

[6]Bien que cela dépende du contexte : une différence de gain correspondant à un $\delta = 0.25$ obtenue en choisissant un placement en bourse plutôt qu'un autre pourrait suffire à faire de nous des millionnaires. On pourra également parcourir à ce sujet le livre de Cohen (1988).

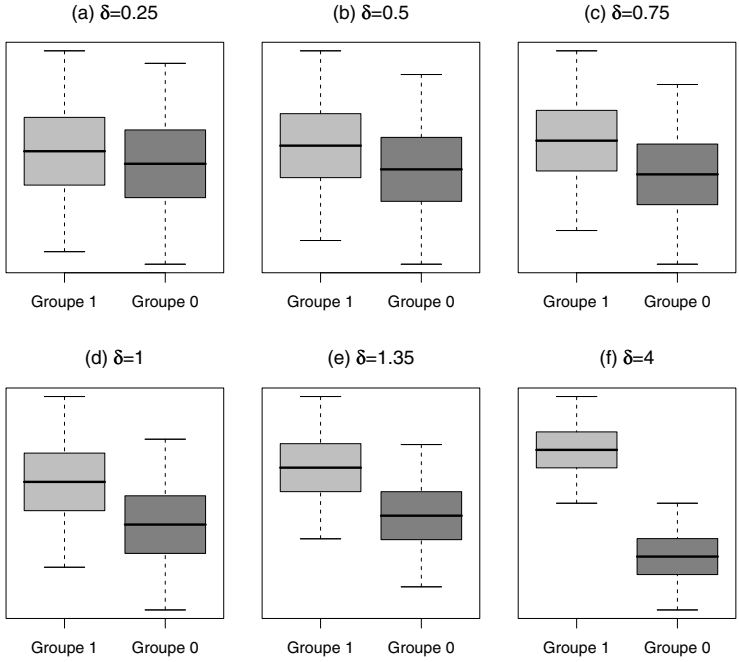

Figure 5.4 – Boxplots représentant différentes valeurs de δ.

de $\delta = 0.5$, un demi écart type ou un huitième de l'étendue, qui est parfois considérée comme une *différence moyenne*, alors que le graphique (c), où l'on a $\delta = 0.75$, exhibe ce que certains commencent à considérer comme une *grande différence*. Les boîtes des deux boxplots ne se chevauchent plus à partir de $\delta = 1.35$ (graphique (e)), alors que les boxplots sont complètement séparés à partir de $\delta = 4$ (graphique (f)).

Exemple 5.5 *Dans l'exemple des Onobrychis, on a une différence moyenne de hauteur correspondant à :*

$$\widehat{\delta} = 9.1/\sqrt{18.1} = 2.14 \quad \textit{écarts types}$$

entre le groupe avec niveau nutritif élevé et le groupe avec niveau nutritif faible. Dans l'exemple de la maladie génétique, on a une différence moyenne de log-créatine correspondant à :

$$\widehat{\delta} = 1.05/\sqrt{0.48} = 1.52 \quad \textit{écart type}$$

entre le groupe maladie et le groupe contrôle. Sans être des spécialistes de ces domaines mais connaissant la valeur d'un écart type dans le cadre normal, on

*est capable de dire que ces différences sont certainement importantes « cliniquement ». On notera aussi que le concept de différence de moyenne standardisée nous permet ici de comparer deux résultats qui seraient sinon difficiles à comparer : on peut dire que l'« effet » de la maladie sur le taux de créatine (*1.52 *écart type) est en un sens moins important que l'effet du niveau nutritif sur la croissance d'un Onobrychis (*2.14 *écarts types)*[7].

5.8 Quotient de variance

Lorsque la distribution des observations est normale dans les deux populations, on pourra également faire de l'inférence sur le quotient des variances σ_1^2/σ_0^2. On connaît en effet mathématiquement la distribution de la variable aléatoire :

$$\frac{\tilde{\sigma}_1^2/\sigma_1^2}{\tilde{\sigma}_0^2/\sigma_0^2}$$

que l'on appelle *distribution de Fisher* (ou *distribution F*). Il s'agit d'une famille de distributions asymétriques avec valeurs possibles positives, indexée par deux paramètres, appelés parfois *degrés de liberté au numérateur* et *degrés de liberté au dénominateur*, que nous noterons par *dln* et *dld*. Pour la variable aléatoire ci-dessus, on aura $dln = n_1 - 1$ et $dld = n_0 - 1$. Il est important de noter que les résultats vus dans cette section ne sont pas valables sans l'hypothèse de normalité des distributions, même avec de grands échantillons.

La figure 5.5 nous montre des exemples de distributions F. Dans ce texte, on notera par $F_{\alpha,dln,dld}$ le quantile α d'une distribution F avec paramètres *dln* et *dld*. Le tableau A.4 donné an annexe nous en donne quelques-uns. Au besoin, un logiciel statistique nous donnera n'importe quel quantile de ces distributions[8]. On aura ainsi :

$$\Pr\left\{F_{\alpha/2,n_1-1,n_0-1} \leq \frac{\tilde{\sigma}_1^2/\sigma_1^2}{\tilde{\sigma}_0^2/\sigma_0^2} \leq F_{1-\alpha/2,n_1-1,n_0-1}\right\} = 1-\alpha.$$

En manipulant ces inéquations de manière à ce que le paramètre d'intérêt σ_1^2/σ_0^2 se retrouve au milieu, on obtient :

$$\Pr\left\{\frac{\tilde{\sigma}_1^2/\tilde{\sigma}_0^2}{F_{1-\alpha/2,n_1-1,n_0-1}} \leq \sigma_1^2/\sigma_0^2 \leq \frac{\tilde{\sigma}_1^2/\tilde{\sigma}_0^2}{F_{\alpha/2,n_1-1,n_0-1}}\right\} = 1-\alpha.$$

Un intervalle de confiance au niveau $1-\alpha$ pour le quotient des variances σ_1^2/σ_0^2 est ainsi calculé par :

$$\left[\frac{\tilde{\sigma}_1^2/\tilde{\sigma}_0^2}{F_{1-\alpha/2,n_1-1,n_0-1}}; \frac{\tilde{\sigma}_1^2/\tilde{\sigma}_0^2}{F_{\alpha/2,n_1-1,n_0-1}}\right].$$

[7] On admettra volontiers que cette comparaison particulière n'intéressera pas grand monde, mais il y aura des cas où il sera scientifiquement intéressant de pouvoir comparer des résultats *a priori* incomparables. Par exemple : est-ce que l'effet du dopage est plus important en cyclisme ou en athlétisme ? Sur un sprint de 100 m ou sur un marathon ?

[8] Dans R, on calcule par exemple $F_{0.95,3,10}$ par la commande `qf(0.95,3,10)` (et on trouve ainsi 3.71, voir le tableau A.4 donné en annexe).

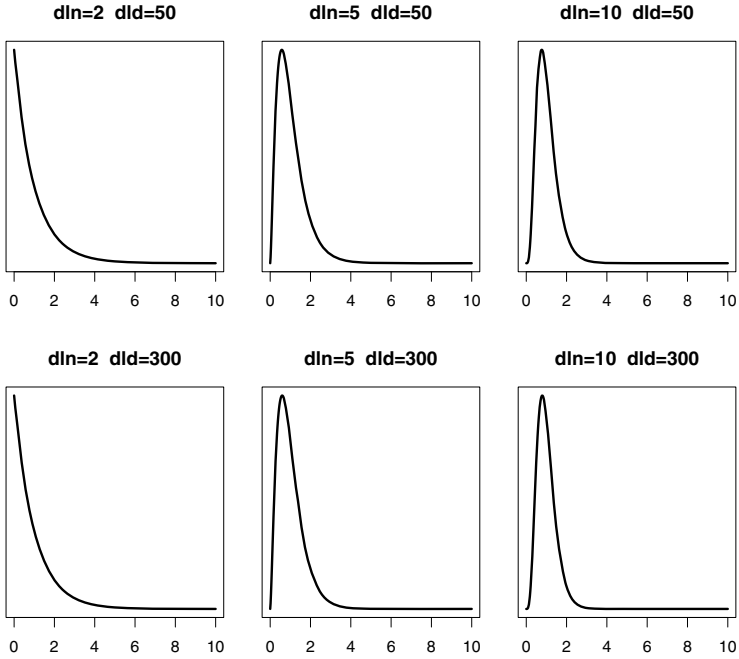

Figure 5.5 – Exemples de distributions F avec différentes valeurs de *dln* et *dld*.

Cet intervalle de confiance sera exact sous l'hypothèse de la normalité des deux distributions. Un intervalle de confiance exact au niveau $1-\alpha$ pour le quotient des écarts types σ_1/σ_0 s'obtient en calculant la racine carrée de ces bornes.

Exemple 5.6 *Pour les Onobrychis cultivés avec un niveau élevé et faible, on avait respectivement* $\tilde{\sigma}_1^2 = 21.45$ *et* $\tilde{\sigma}_0^2 = 14.8$*, avec par ailleurs* $n_1 = n_0 = 60$*. Le quotient des variances observé était donc* $21.45/14.8 = 1.45$*. Comme on a* $F_{0.025,59,59} = 0.60$ *et* $F_{0.975,59,59} = 1.67$*, un intervalle de confiance au niveau* 95 % *pour le véritable quotient des variances* σ_1^2/σ_0^2 *s'obtient par :*

$$[1.45/1.67; 1.45/0.60] = [0.87; 2.42].$$

Un intervalle de confiance au niveau 95 % *pour le quotient des écarts types* σ_1/σ_0 *s'obtient alors par :*

$$[\sqrt{0.87}; \sqrt{2.42}] = [0.93; 1.56].$$

En particulier, la valeur 1 *se trouve à l'intérieur de l'intervalle de confiance, de sorte que l'on n'a pas prouvé statistiquement que la variabilité est différente dans les deux groupes.*

5.9 Différence de proportion

Bien qu'une variable binaire soit *a priori* un concept plus simple qu'une variable continue, il est en un sens plus difficile de comparer deux variables binaires que de comparer deux variables continues. En effet, deux variables binaires Y_1 et Y_0 avec des proportions π_1 et π_0 différentes auront également des variances $\pi_1(1-\pi_1)$ et $\pi_0(1-\pi_0)$ différentes. Il s'ensuit que l'on ne peut pas caractériser la différence entre deux distributions binaires avec un seul paramètre (il n'existe pas de modèle idéal pour les variables binaires). Afin de donner toute l'information, il s'agit de donner les deux paramètres π_1 et π_0.

Le but de la statistique descriptive demeure pourtant de résumer l'information (à défaut de la caractériser) avec un seul paramètre. Lorsqu'il s'agit de comparer deux distributions binaires, on peut considérer par exemple la *différence des proportions* :

$$\Lambda = \pi_1 - \pi_0.$$

Ce n'est cependant pas l'unique possibilité. Certains auteurs préfèrent le *quotient des proportions* π_1/π_0 ou encore l'`odds-ratio` $(\pi_1/(1-\pi_1))/(\pi_0/(1-\pi_0))$ (concept que l'on retrouvera au chapitre 16).

À partir de deux échantillons d'observations de taille n_1 et n_0 des variables Y_1 et Y_0 avec proportions π_1 et π_0, on estimera $\Lambda = \pi_1 - \pi_0$ par :

$$\widehat{\Lambda} = \widehat{\pi}_1 - \widehat{\pi}_0$$

où $\widehat{\pi}_1$ et $\widehat{\pi}_0$ dénotent les proportions empiriques calculées sur ces échantillons. On a par ailleurs les résultats suivants :

- $\widehat{\Lambda}$ est un estimateur sans biais et approximativement normal de Λ
- l'erreur type de cet estimateur est donnée par :

$$\mathrm{SE}(\widehat{\Lambda}) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_0(1-\pi_0)}{n_0}}$$

- en utilisant la méthode de Wald, on peut calculer un intervalle de confiance au niveau $1-\alpha$ pour Λ comme suit :

$$\widehat{\Lambda} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_0(1-\widehat{\pi}_0)}{n_0}}.$$

Un intervalle de confiance de Wald pour Λ sera valide pour autant que les tailles des échantillons n_1 et n_0 soient assez grandes et que les véritables proportions π_1 et π_0 ne soient pas trop proches de 0 ou de 1[9].

[9] De la même manière que la validité d'un intervalle de confiance de Wald pour une proportion peut être améliorée en le calculant à partir d'un échantillon où l'on aura ajouté quatre observations fictives, deux fois la valeur 0 et deux fois la valeur 1, Agresti et Caffo (2000) ont suggéré d'améliorer la validité d'un intervalle de confiance de Wald pour une différence de proportion en le calculant à partir de deux échantillons où l'on aura ajouté un 0 et un 1 dans chaque échantillon (on aura donc ici aussi ajouté en tout quatre observations fictives). Voir également le récent article de Fagerland *et al.* (2011).

Exemple 5.7 *On veut comparer la proportion* π_1 *de gauchers chez les garçons et la proportion* π_0 *de gauchères chez les filles. On observe* 41 *gauchers parmi* $n_1 = 327$ *garçons et* 25 *gauchères parmi* $n_0 = 332$ *filles*[10]*. Cela correspond à des proportions empiriques* $\widehat{\pi}_1 = 41/327 = 0.125$ *et* $\widehat{\pi}_0 = 25/332 = 0.075$*. On estime ainsi une différence de proportion de* $\widehat{\Lambda} = 0.125 - 0.075 = 0.05$ *et on calcule un intervalle de confiance au niveau* 95 % *pour la véritable différence de proportion* $\Lambda = \pi_1 - \pi_0$ *comme suit :*

$$0.05 \pm 1.96 \cdot \sqrt{\frac{0.125 \cdot 0.875}{327} + \frac{0.075 \cdot 0.925}{332}} = [0.004; 0.096].$$

On a donc prouvé statistiquement qu'il y a plus de gauchers chez les garçons que de gauchères chez les filles (en tout cas dans les populations d'où proviennent nos échantillons), la valeur $\Lambda = 0$ *n'étant (de justesse) pas comprise dans l'intervalle de confiance pour* Λ*.*

[10] Ces données proviennent de l'étude publiée dans Rousson *et al.* (2009).

Chapitre 6

Principe d'un test statistique

On a introduit au chapitre 4 le concept d'intervalle de confiance que l'on a utilisé pour faire de l'inférence sur certains paramètres de la population. Dans ce chapitre, nous présentons une approche alternative pour faire de l'inférence *via* le concept de *test statistique*. Nous verrons au chapitre 8 que les concepts d'intervalle de confiance et de test statistique sont intimement liés.

6.1 L'hypothèse nulle et l'hypothèse alternative

Il existe des centaines de tests statistiques. Chacun de ces tests est associé à une *hypothèse nulle* que l'on notera H_0. Le but d'un test statistique est de démontrer qu'une hypothèse nulle est fausse en la confrontant aux données de notre échantillon. Si les données sont incompatibles avec l'hypothèse nulle, on *rejette* l'hypothèse nulle. On ne pourra pas par contre démontrer qu'une hypothèse nulle est vraie. L'hypothèse nulle n'est donc pas l'hypothèse d'intérêt ni l'hypothèse scientifique d'une étude. L'hypothèse scientifique d'une étude est *l'hypothèse alternative*, que l'on notera parfois H_1, et qui sera en quelque sorte le contraire de l'hypothèse nulle. Il s'agira donc de formuler l'hypothèse nulle de telle sorte que son rejet implique l'hypothèse alternative.

On démontre une hypothèse alternative en rejetant une hypothèse nulle.

Par exemple, afin de démontrer l'hypothèse alternative (scientifique) suivante :

H_1 : le niveau nutritif a un effet sur la croissance des *Onobrychis*

on essaiera de rejeter l'hypothèse nulle :

H_0 : le niveau nutritif n'a aucun effet sur la croissance des *Onobrychis*.

En d'autres termes, afin de démontrer que le niveau nutritif a un effet sur la croissance des *Onobrychis*, on essaiera de démontrer qu'il n'est pas possible qu'il n'en ait pas. On recherche en quelque sorte une preuve par l'absurde. On part du contraire de ce que l'on veut démontrer (l'hypothèse nulle) et on essaie d'aboutir à une contradiction entre l'hypothèse nulle et les données, afin de pouvoir conclure ce que l'on veut démontrer (l'hypothèse alternative). Telle est la stratégie d'un test statistique.

6.2 Erreurs de première et de seconde espèce

Le résultat d'un test statistique est donc le rejet ou le non-rejet d'une hypothèse nulle H_0. Ceci peut nous mener à deux types d'erreur : rejeter H_0 alors qu'elle est vraie ou ne pas rejeter H_0 alors qu'elle est fausse. On appelle ces erreurs respectivement l'erreur de première espèce et l'erreur de seconde espèce. Notons que lorsque l'on rejette H_0, la seule erreur que l'on peut commettre est une erreur de première espèce, alors que lorsque l'on ne rejette pas H_0, la seule erreur que l'on peut commettre est une erreur de seconde espèce. On peut résumer la situation dans le tableau suivant :

	rejeter H_0	ne pas rejeter H_0
H_0 vraie	erreur de première espèce	bonne décision
H_0 fausse	bonne décision	erreur de seconde espèce

Évidemment, on ne pourra pas commettre une erreur de première espèce si H_0 est fausse, de même que l'on ne pourra pas commettre une erreur de seconde espèce si H_0 est vraie. Si H_0 est vraie, on note par α la probabilité de commettre une erreur de première espèce. Si H_0 est fausse, on note par β la probabilité de commettre une erreur de seconde espèce. On peut résumer la situation dans le tableau suivant :

	probabilité de rejeter H_0	probabilité de ne pas rejeter H_0
H_0 vraie	α	$1-\alpha$
H_0 fausse	$1-\beta$	β

Lorsque l'on effectue un test statistique, on pourra choisir la valeur de α, que l'on appellera aussi le *seuil du test*. On dira que l'on rejette ou que l'on ne rejette pas une hypothèse nulle au seuil α. Évidemment, on aimerait que α soit le plus petit possible. Il y a cependant un conflit entre α et β. En choisissant α trop petit, on risque d'augmenter considérablement β. Le cas extrême consisterait à choisir $\alpha = 0$, ce qui reviendrait à ne jamais rejeter H_0 (quelles que soient les données observées) et impliquerait alors $\beta = 1$. Il s'agit dès lors d'adopter un compromis, et ce compromis est en général fixé à $\alpha = 5\,\%$ (selon une convention arbitraire mais souvent raisonnable et largement établie)[1].

[1] On verra par ailleurs que le choix de fixer le seuil d'un test à 5 % est lié au choix de fixer le niveau d'un intervalle de confiance à 95 %.

La règle de rejet ou de non-rejet d'une hypothèse nulle doit être ainsi définie de façon à ce que si l'hypothèse nulle était vraie (on dira parfois : sous l'hypothèse nulle), il y aurait une probabilité de 5 % de la rejeter à tort. Cela veut dire que si H_0 était vraie et si on répétait l'expérience (l'échantillonnage) 100 fois, on ne rejetterait H_0 que 5 fois (en moyenne). On aura ainsi :

α = seuil du test = probabilité de rejeter H_0 alors que H_0 est vraie = 5 %.

Lorsque l'on effectue un test statistique, on ne choisit pas par contre la valeur de β. On verra au chapitre 10 comment on peut calculer β, qui dépendra à la fois de la taille de l'échantillon et du « degré de fausseté » de H_0. La quantité $1 - \beta$ est par ailleurs appelée la *puissance du test*, qui est donc une mesure de la capacité du test à rejeter à raison une hypothèse nulle qui est fausse.

En résumé, dans un test statistique, on contrôle α mais on ne contrôle pas β, qui pourra être dans certains cas considérablement plus grand que α. En choisissant α petit, on s'assure contre une erreur de première espèce, mais on n'a aucune garantie contre une erreur de seconde espèce. En fait, on considère implicitement qu'il est plus grave de commettre une erreur de première espèce que de commettre une erreur de seconde espèce.

En général, on se montrera donc plus prudent dans nos conclusions dans le cas d'un non-rejet d'une hypothèse nulle (car on pourrait commette une erreur de seconde espèce qui n'est pas contrôlée), que dans le cas d'un rejet d'une hypothèse nulle (car l'erreur de première espèce que l'on pourrait commettre est contrôlée). Pour cette même raison, nous éviterons dans ce texte l'expression « accepter une hypothèse nulle », que l'on retrouve dans certains ouvrages en lieu et place de l'expression « ne pas rejeter une hypothèse nulle » que nous adoptons[2].

6.3 Concept de valeur p

On a mentionné que choisir $\alpha = 0$ impliquerait de ne jamais rejeter une hypothèse nulle (de façon à n'avoir aucune chance de commettre une erreur de première espèce si H_0 était vraie). À l'autre extrême, choisir $\alpha = 1$ reviendrait à

[2]On fait parfois un parallèle avec la justice, où l'on considère qu'il est plus grave de condamner un innocent que de ne pas condamner un coupable. Dans le contexte d'un test statistique, l'hypothèse nulle est l'accusée et les données sont des preuves potentielles contre l'accusée. Si ces preuves sont suffisamment convaincantes, on condamne H_0 (on rejette H_0). Si ces preuves ne sont pas suffisamment convaincantes, on ne condamne pas H_0. Cependant, des preuves contre l'accusée ne peuvent pas être utilisées pour prouver son innocence, et c'est pourquoi on préfère dire dans ce cas « on ne rejette pas H_0 » plutôt que de dire « on accepte H_0 ». Une hypothèse nulle non rejetée est en quelque sorte une accusée en sursis. Peut-être va-t-on bientôt récolter de nouvelles données qui condamneront l'hypothèse nulle. En attendant, on lui accorde le bénéfice du doute.

systématiquement rejeter une hypothèse nulle. D'une manière générale, plus on augmente α, plus il sera aisé de rejeter une hypothèse nulle. Si deux chercheurs adoptent un seuil différent, il se pourrait donc qu'à partir des mêmes données, l'un rejette H_0 et l'autre ne la rejette pas. Afin que le lecteur d'un article scientifique puisse lui-même choisir son seuil, il est coutume de reporter dans ces articles ce que l'on appelle *une valeur p*. La valeur p peut être définie comme le seuil minimal au-delà duquel on rejette l'hypothèse nulle.

valeur p = seuil minimal au-delà duquel on rejette H_0.

Le résultat d'un test statistique sera donc une valeur p qui nous permet de décider si on rejette ou non l'hypothèse nulle selon le seuil α que l'on s'est préalablement choisi, avec la règle suivante :

On rejette H_0 au seuil α si $p \leq \alpha$.

Lorsque l'on rejette une hypothèse nulle au seuil α ($p \leq \alpha$), on parlera de résultat *significatif* au seuil α. Dans le cas contraire ($p > \alpha$), on parlera de résultat *non significatif* au seuil α. Connaissant la valeur p, un lecteur d'article scientifique peut ainsi décider de lui-même, en fonction du seuil qu'il s'est choisi, s'il rejette ou non l'hypothèse nulle considérée. En pratique cependant, tout le monde ou presque choisit le seuil $\alpha = 5$ %, l'hypothèse nulle étant rejetée si $p \leq 0.05$. Lorsque l'on parle d'un résultat significatif sans préciser le seuil, il s'agira (par convention) d'un résultat significatif au seuil de 5 %.

On peut alternativement définir le concept de valeur p sans passer par les concepts d'erreur de première et de seconde espèce, en terme de *hasard de l'échantillonnage*, comme nous l'expliquons ci-dessous. À cause du hasard de l'échantillonnage, il y aura inévitablement une certaine distance entre les données et l'hypothèse nulle, et cela même si l'hypothèse nulle était vraie. Par exemple, même si le niveau nutritif n'avait aucun effet sur la croissance d'un *Onobrychis* (ce qui constitue ici l'hypothèse nulle), et donc même si la véritable différence de moyenne entre deux groupes d'*Onobrychis* cultivés avec un niveau nutritif faible et élevé était de 0 cm, on observerait quand même une différence de moyenne non nulle dans nos échantillons, due dans ce cas uniquement au hasard de l'échantillonnage. Ainsi, lorsque l'on observe une différence non nulle entre deux moyennes, par exemple une différence de 9.1 cm entre deux groupes d'*Onobrychis*, la question statistique que l'on se pose est la suivante : est-ce que cette différence observée de 9.1 cm pourrait être due uniquement au hasard de l'échantillonnage ou est-ce que le hasard de l'échantillonnage ne peut pas être tenu seul responsable d'une telle différence ? Dans le premier cas, on ne rejette pas l'hypothèse nulle, dans le second cas, on la rejette. Dans ce

contexte, la valeur p peut être définie comme « la probabilité que le hasard de l'échantillonnage puisse produire des données aussi éloignées (ou encore plus éloignées) de l'hypothèse nulle que le sont les données de notre échantillon, si l'hypothèse nulle était vraie ». Dans l'exemple des *Onobrychis*, on aura par exemple $p = 0.0000000000000002$, de sorte qu'il serait extrêmement rare d'observer des données aussi éloignées de l'hypothèse nulle si celle-ci était vraie, les données contredisant ainsi fortement l'hypothèse nulle[3].

Plus la valeur p est petite,
plus l'hypothèse nulle est contredite par les données.

La convention veut donc que l'on parle de résultat significatif si on a $p \leq 0.05$. On peut ainsi résumer la situation comme suit :

- une **différence significative** (entre les données et l'hypothèse nulle) est une différence que le hasard de l'échantillonnage ne pourrait que rarement produire

  → les données sont incompatibles avec H_0

  → on a une preuve statistique contre H_0 (on rejette H_0)

  → notons qu'il s'agit « seulement » d'une preuve statistique contre H_0, non d'une preuve mathématique (bien que contrôlée à 5 %, on pourrait quand même commettre une erreur de première espèce ; sauf dans des cas très spéciaux, on ne pourra pas totalement exclure la possibilité que le hasard soit responsable de quelque chose ; on n'aura pas de valeur p exactement nulle)

- une **différence non significative** (entre les données et l'hypothèse nulle) est une différence que le hasard de l'échantillonnage pourrait produire

  → les données sont compatibles avec H_0

  → on n'a pas de preuve statistique contre H_0 (on ne rejette pas H_0)

  → attention, on n'a pas pour autant de preuve statistique pour H_0 (on pourrait commettre une erreur de seconde espèce, qui n'est pas contrôlée).

[3]Le concept de valeur p, introduit vers 1925 par le fameux statisticien R.A. Fisher, est l'un des plus difficiles et des plus incompris de la statistique. Certaines interprétations courantes de la valeur p sont erronées. En particulier, la valeur p n'est pas « la probabilité que l'hypothèse nulle soit vraie ». Une telle probabilité serait du reste difficile à définir dans notre contexte où nous définissons une probabilité comme une proportion, à savoir un numérateur divisé par un dénominateur. Or, quel serait le dénominateur d'une telle probabilité ? On n'a en principe qu'une seule hypothèse nulle (et il est difficile de s'imaginer « répéter » une hypothèse nulle). De fait, une hypothèse nulle sera soit vraie soit fausse et il n'y aura pas de sens ici de parler de la probabilité qu'elle soit vraie ou qu'elle soit fausse (il faudrait pour cela adopter une approche *bayésienne* de la statistique). La valeur p n'est donc pas la probabilité que l'hypothèse nulle soit vraie, mais la probabilité d'observer des données aussi éloignées (ou plus éloignées) de l'hypothèse nulle que les nôtres si l'hypothèse nulle était vraie.

6.4 Tests multiples

Malheureusement, notre compromis de 5 % qui apparaît raisonnable lorsque l'on effectue un seul test statistique n'est souvent plus raisonnable lorsque l'on en effectue plusieurs. Supposons que l'on considère m hypothèses nulles et que toutes ces hypothèses nulles soient vraies (on aimerait par exemple démontrer l'efficacité de m médicaments qui n'ont tous en réalité aucun effet). Si les données que l'on utilise pour chacun de ces tests sont indépendantes les unes des autres, la probabilité d'obtenir au moins un (faux) résultat significatif sera de $1 - 0.95^m$. Lorsque m augmente, on a ainsi une inflation du risque de première espèce. Si on teste par exemple un médicament sans effet ni chez les hommes ni chez les femmes, on aura une probabilité non pas de 5 % mais de 10 % de rejeter au moins l'une des $m = 2$ hypothèses nulles considérées (et de conclure à tort que le médicament a un effet, soit pour les hommes si on rejette la première hypothèse nulle, soit pour les femmes si on rejette la seconde). Le tableau suivant nous donne d'autres exemples de cette inflation :

m	$1 - 0.95^m$	m	$1 - 0.95^m$
1	5 %	6	26 %
2	10 %	10	40 %
3	14 %	20	64 %
4	19 %	50	92 %
5	23 %	100	99 %

Si on teste un médicament sans effet sur $m = 100$ sous-groupes de patients, on trouvera ainsi presque à coup sûr au moins un résultat significatif même si ce médicament n'a en réalité aucun effet pour personne.

Les tests multiples constituent un sérieux problème de la statistique car il est fréquent que plusieurs hypothèses soient testées dans une même étude. Afin de nous prémunir contre cette inflation du risque de première espèce, on utilise parfois une *correction de Bonferroni*. Cette correction consiste à ne considérer comme significatifs que les résultats pour lesquels on a $p \leq 0.05/m$. Par exemple, un résultat avec $p = 0.02$ sera considéré comme significatif si on a effectué $m = 1$ ou $m = 2$ tests, mais ne sera pas considéré comme significatif si on a effectué $m = 3$ tests ou plus.

D'une manière générale, il est recommandé de limiter le nombre de tests dans une étude. Un test statistique s'effectue en principe lors d'une *analyse confirmatoire* (lorsqu'une hypothèse scientifique a été formulée *a priori*), non lors d'une *analyse exploratoire* (lorsque l'on n'a aucune idée précise quant aux résultats que l'on peut attendre d'une étude). Il s'agit également d'être prudent dans nos conclusions. Si on réussit par exemple à prouver statistiquement qu'un médicament est efficace chez les hommes mais pas chez les femmes, il serait bon d'avoir une explication médicale crédible (et il serait encore mieux de l'avoir formulée avant d'avoir récolté les données). Dans le cas contraire, il pourrait s'agir d'un faux significatif.

6.5 Statistique de test

Afin d'essayer de rejeter une hypothèse nulle, on utilise les données de notre échantillon. Les données d'un échantillon constituent donc une preuve (qui sera convaincante ou non) contre une hypothèse nulle. Plus concrètement, il s'agit de calculer une distance entre les données de l'échantillon et l'hypothèse nulle, que l'on appellera une *statistique de test* et que l'on notera par T_{stat}.

Comme tout ce qui est calculé sur un échantillon, une statistique de test peut se voir comme une variable aléatoire : on l'observe en pratique une seule fois, sur notre seul échantillon, mais en théorie (et avec un peu d'imagination) on pourrait l'observer plusieurs fois si on répétait l'expérience. On peut donc parler de la distribution d'une statistique de test. Plus précisément, on s'intéressera à *la distribution de la statistique de test sous l'hypothèse nulle*. On s'imagine ainsi que l'on répète l'expérience, non pas sous les conditions réelles d'une étude, mais sous les conditions spécifiées par l'hypothèse nulle. Il s'agit d'établir mathématiquement quelle est cette distribution, afin de savoir ce que l'on est en droit d'attendre de cette statistique de test sous l'hypothèse nulle.

On notera par t_{stat} la réalisation de T_{stat} (la valeur de T_{stat} calculée/observée sur notre échantillon). Il s'agit de la distance observée entre les données de notre échantillon et l'hypothèse nulle. On comparera ensuite t_{stat} avec la distribution théorique de T_{stat} établie sous l'hypothèse nulle. Si l'observation (t_{stat}) n'est pas compatible avec la théorie (distribution de T_{stat} sous H_0), on dira que les données ne sont pas compatibles avec l'hypothèse nulle, qui sera ainsi rejetée. Un test statistique s'effectue donc en quatre étapes de la manière suivante :

- définir une statistique de test T_{stat} calculable sur un échantillon
- établir mathématiquement la distribution théorique de T_{stat} sous H_0
- calculer la réalisation t_{stat} de T_{stat} sur notre échantillon
- comparer t_{stat} avec la distribution théorique de T_{stat} sous H_0.

Les deux premières étapes sont des étapes théoriques, fondées sur les mathématiques. Les données entrent en jeu à partir de la troisième étape. La quatrième étape est le calcul de la valeur p, qui mesure à quel point les données sont incompatibles avec l'hypothèse nulle, et qui nous permet de décider si on rejette ou non l'hypothèse nulle. Nous verrons de nombreux exemples dans les chapitres suivants. Notons encore qu'un test statistique sera dit *exact* si on connaît mathématiquement (et si on utilise effectivement) la distribution de la statistique de test sous H_0, alors qu'il sera dit *valide* si à défaut de la connaître exactement, on dispose d'une bonne approximation de cette distribution[4].

[4]Comme tout ce qui se calcule à partir d'un échantillon, la valeur p, qui sera une fonction de la statistique de test t_{stat}, peut être vue comme la réalisation d'une variable aléatoire P (fonction de T_{stat}). Si T_{stat} est continue, un test statistique sera exact si la distribution de P sous H_0 est uniforme entre 0 et 1, c'est-à-dire si $\Pr\{P \leq \alpha\} = \alpha$ (la probabilité de commettre une erreur de première espèce étant ainsi bel et bien exactement égale à α).

Chapitre 7

Tests du khi-deux pour tables de contingence

Nous illustrons dans ce chapitre les principes d'un test statistique en introduisant les *tests du khi-deux*.

7.1 Comparaison de distributions de variables qualitatives

Nous considérons dans cette section un exemple tiré de la littérature[1]. On s'intéresse à la variable « couleur des cheveux » mesurée sur une échelle qualitative avec les $J = 5$ valeurs possibles suivantes : « blond », « roux », « châtain », « brun » et « noir ». On veut comparer cette variable qualitative entre $I = 4$ populations d'écoliers qui ont respectivement les yeux bleus, les yeux verts, les yeux bruns et les yeux noirs, le but étant de montrer qu'il existe un lien entre la couleur des cheveux et la couleur des yeux, autrement dit que la distribution de la couleurs des cheveux diffère d'une population à l'autre. On essaiera donc de rejeter l'hypothèse nulle suivante :

H_0 : la distribution de la variable « couleur des cheveux » est la même dans les 4 populations aux « couleurs des yeux » différentes.

Pour ce faire, on utilise les données récoltées auprès de 4 échantillons représentatifs de ces populations d'écoliers, qui sont résumées dans la table de contingence de dimension 4×5 suivante (contenant les *fréquences observées*) :

[1] Il s'agit d'un exemple publié par R.A. Fisher (1940).

couleur yeux	couleur cheveux blond	roux	châtain	brun	noir	total
bleu	326	38	241	110	3	718
	(45 %)	(5 %)	(34 %)	(15 %)	(0 %)	(100 %)
vert	688	116	584	188	4	1580
	(44 %)	(7 %)	(37 %)	(12 %)	(0 %)	(100 %)
brun	343	84	909	412	26	1774
	(19 %)	(5 %)	(51 %)	(23 %)	(1 %)	(100 %)
noir	98	48	403	681	85	1315
	(7 %)	(4 %)	(31 %)	(52 %)	(6 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

On a vu que la distribution d'une variable qualitative avec J valeurs possibles peut être caractérisée par J proportions (en fait $J-1$ proportions sont suffisantes). Dans cette table, on a donc calculé ces $J=5$ proportions pour chacun des $I=4$ échantillons. La dernière ligne de la table nous donne en outre la « distribution moyenne » de la variable « couleur des cheveux » calculée sur le total des observations (on a ici 27 % de cheveux blonds, 5 % de cheveux roux, 40 % de cheveux châtains, 26 % de cheveux bruns, 2 % de cheveux noirs).

À première vue, ces données semblent effectivement contredire l'hypothèse nulle. La distribution de la couleur des cheveux est par exemple très différente entre les écoliers aux yeux bleus (majorité de cheveux blonds), les écoliers aux yeux bruns (majorité de cheveux châtains) et les écoliers aux yeux noirs (majorité de cheveux bruns). Rappelons toutefois qu'à cause du hasard de l'échantillonnage, on observera inévitablement certaines différences entre ces quatre distributions, et ceci même si l'hypothèse nulle était vraie. Par contre, de trop grosses différences seront suspectes. Toute la question est donc de déterminer si les différences observées entre ces quatre distributions sont suspectes ou non, autrement dit si elles peuvent être mises sur le compte du hasard de l'échantillonnage (auquel cas on ne rejettera pas H_0) ou non (auquel cas on rejettera H_0).

Pour répondre à cette question, on utilise une statistique de test. Rappelons qu'une statistique de test est une mesure de la distance entre les données et l'hypothèse nulle dont on connaît (approximativement) la distribution sous H_0. Afin de définir une telle distance, on notera tout d'abord que si nos données respectaient à la lettre H_0, on aurait la même distribution de la couleur des cheveux dans chacun de ces 4 échantillons, ces 4 distributions étant alors égales à la distribution moyenne (calculée dans la dernière ligne de la table ci-dessus). On aurait alors la table de contingence suivante (contenant les *fréquences attendues* sous l'hypothèse nulle) :

couleur	couleur cheveux					
yeux	blond	roux	châtain	brun	noir	total
bleu	193.9	38.1	284.8	185.4	15.7	718
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
vert	426.7	83.9)	626.8	408.0	34.6	1580
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
brun	479.1	94.2	703.7	458.1	38.9	1774
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
noir	355.2	69.8	521.7	339.6	28.8	1315
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

Les fréquences attendues pour la première ligne de cette table s'obtiennent par exemple de la manière suivante : $718 \cdot 1455/5387 = 193.9$ cheveux blonds, $718 \cdot 286/5387 = 38.1$ cheveux roux, $718 \cdot 2137/5387 = 284.8$ cheveux châtains, $718 \cdot 1391/5387 = 185.4$ cheveux bruns et $718 \cdot 118/5387 = 15.7$ cheveux noirs. Les calculs sont analogues pour les autres lignes de cette table. La statistique de test sera une mesure de la distance globale entre les fréquences observées et les fréquences attendues. Elle se calcule en sommant les 4×5 termes suivants :

$$\begin{aligned} t_{stat} &= \frac{(326-193.9)^2}{193.9} + \frac{(38-38.1)^2}{38.1} + \frac{(241-284.8)^2}{284.8} + \frac{(110-185.4)^2}{185.4} + \frac{(3-15.7)^2}{15.7} \\ &\quad + \frac{(688-426.7)^2}{426.7} + \frac{(116-83.9)^2}{83.9} + \frac{(584-626.8)^2}{626.8} + \frac{(188-408-0)^2}{408.0} + \frac{(4-34.6)^2}{34.6} \\ &\quad + \frac{(343-479.1)^2}{479.1} + \frac{(84-94.2)^2}{94.2} + \frac{(909-703.7)^2}{703.7} + \frac{(412-458.1)^2}{458.1} + \frac{(26-38.9)^2}{38.9} \\ &\quad + \frac{(98-355.2)^2}{355.2} + \frac{(48-69.8)^2}{69.8} + \frac{(403-521.7)^2}{521.7} + \frac{(681-339.6)^2}{339.6} + \frac{(85-28.8)^2}{28.8} \\ &= 1240.0 \end{aligned}$$

Au-delà de notre exemple particulier, si on dénote par O_{ij} les fréquences observées (O_{ij} désignant le nombre de fois que l'on observe la valeur possible j dans le groupe i, pour $i = 1, \cdots, I$ et $j = 1, \cdots, J$), les fréquences attendues E_{ij} sont définies comme suit :

$$E_{ij} = \frac{\sum_{j=1}^{J} O_{ij} \cdot \sum_{i=1}^{I} O_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij}}$$

et la statistique de test est définie comme suit :

$$T_{stat} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

On rappellera ici la distinction de notation entre le concept de la variable aléatoire T_{stat} et sa réalisation t_{stat} dans notre échantillon. Pour un test du khi-

deux appliqué à une table de contingence, les quatre étapes d'un test statistique se présentent alors de la manière suivante :

- la première étape consiste à définir une statistique de test T_{stat} comme on vient de le faire ci-dessus
- la deuxième étape consiste à établir mathématiquement la distribution de cette statistique de test sous l'hypothèse nulle ; il se trouve que la distribution de T_{stat} sous H_0 est ici (approximativement) une *distribution du khi-deux* avec $(I-1)(J-1)$ *dl*, l'approximation étant bonne si la majorité (par exemple 80 %) des fréquences attendues sont supérieures à 5, auquel cas le test du khi-deux est dit valide (dans notre exemple, il s'agit donc de $(4-1)(5-1) = 12$ *dl*)
- la troisième étape consiste à calculer la réalisation t_{stat} de la variable aléatoire T_{stat} dans notre échantillon (dans notre exemple $t_{stat} = 1240.0$)
- la quatrième étape consiste à comparer la valeur observée (dans notre exemple $t_{stat} = 1240.0$) avec la distribution théorique de T_{stat} sous H_0 (dans notre exemple, une distribution du khi-deux avec 12 *dl*), l'idée étant de rejeter hypothèse nulle si la statistique de test observée t_{stat} est incompatible (trop grande) par rapport à la distribution théorique.

En ce qui concerne cette quatrième étape, on adopte la règle de rejet suivante :

> On rejette H_0 au seuil α si $t_{stat} \geq \chi^2_{1-\alpha,(I-1)(J-1)}$.

Rappelons que l'on avait noté par $\chi^2_{1-\alpha,dl}$ le quantile $1-\alpha$ d'une distribution du khi-deux avec *dl* degrés de liberté (on trouvera quelques-uns de ces quantiles dans le tableau A.3 donné en annexe). En adoptant cette règle de rejet, la probabilité de commettre une erreur de première espèce sera bel et bien de α (elle sera du moins approximativement de α si les conditions de validité du test sont respectées). Le graphique du haut de la figure 7.1 nous montre la région de rejet au seuil de 5 % lorsque $dl = 12$. Dans notre exemple, on rejette H_0 au seuil de 5 % car $t_{stat} = 1240.0$ est beaucoup plus grand que $\chi^2_{0.95,12} = 21.03$. On a donc réussi à prouver statistiquement notre hypothèse scientifique, à savoir que la distribution de la couleur des cheveux n'est pas la même dans ces quatre populations (et donc qu'il y a un lien entre couleur des cheveux et couleur des yeux).

Le graphique du bas de la figure 7.1 illustre par ailleurs le calcul de la valeur p qui s'obtient à partir de t_{stat} comme suit :

$$p = \Pr\{T_{stat} \geq t_{stat}\}.$$

La valeur p est donc littéralement « la probabilité d'observer une distance T_{stat} entre les données et l'hypothèse nulle aussi grande (ou encore plus grande) que

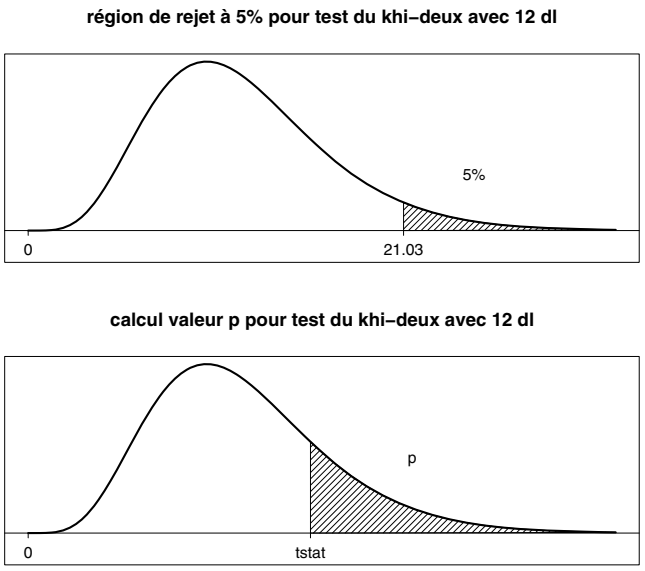

Figure 7.1 – Région de rejet à 5 % et calcul de la valeur p pour test du χ^2 avec 12 dl.

la distance t_{stat} observée dans notre échantillon, si H_0 était vraie ». On a de façon équivalente :

$$t_{stat} = \chi^2_{1-p,(I-1)(J-1)}.$$

Autrement dit, on calcule p de telle sorte que t_{stat} corresponde au quantile $1-p$ d'une distribution du khi-deux avec $(I-1)(J-1)$ dl. Dans notre exemple, un logiciel statistique nous donnera $p < 0.0001$[2]. Cela veut dire que l'hypothèse nulle est rejetée à pratiquement n'importe quel seuil[3].

[2]Dans R, on pourra ici calculer la valeur p par la commande `1-pchisq(1240.0,12)`.

[3]Dans cet exemple, on a comparé $I = 4$ populations d'écoliers aux « couleurs des yeux » différentes par rapport à la distribution de la variable « couleur des cheveux » qui admet $J = 5$ valeurs possibles, et on a essayé de rejeter l'hypothèse nulle :

H_0 : la distribution de la variable « couleur des cheveux » est la même dans les 4 populations aux « couleurs des yeux » différentes.

On pourrait retourner le problème et comparer $J = 5$ populations d'écoliers aux « couleurs des cheveux » différentes par rapport à la distribution de la variable « couleur des yeux » qui admet $I = 4$ valeurs possibles. Dans ce cas, l'hypothèse nulle que l'on essaierait de rejeter serait la suivante :

H_0 : la distribution de la variable « couleur des yeux » est la même dans les 5 populations aux « couleurs des cheveux » différentes.

En inversant les lignes et les colonnes de la table de contingence et en appliquant un test du khi-deux, on obtiendra exactement le même résultat que ci-dessus (même $t_{stat} = 1240.0$, même $dl = 12$, et donc même valeur p). En fait, les deux hypothèses nulles sont équivalentes. Dire que la distribution de la couleur des cheveux est la même dans des populations aux

7.2 Comparaison d'une distribution qualitative avec distribution de référence

Le test du khi-deux vu dans la section précédente s'utilise notamment lorsqu'il s'agit de comparer la distribution d'une variable qualitative entre deux groupes. On va voir dans cette section un test du khi-deux qui nous permet de comparer la distribution d'une variable qualitative avec une distribution de référence, donnée par exemple dans la littérature.

Nous reprenons pour cela l'exemple introduit au chapitre 2 à propos des groupes sanguins récoltés auprès de $n = 158$ individus d'une population du nord de l'Europe. Nous avions les fréquences observées suivantes :

groupe O	groupe A	groupe B	groupe AB	total
82	62	10	4	158
52 %	39 %	6 %	3 %	100 %

L'hypothèse scientifique d'une telle étude sera par exemple que la distribution des groupes sanguins dans cette population du nord de l'Europe diffère de celle de la France, où l'on a (selon la littérature) 43 % de groupe O, 45 % de groupe A, 9 % de groupe B et 3 % de groupe AB. On essaiera alors de rejeter l'hypothèse nulle suivante :

H_0 : il y a 43 % de groupe O, 45 % de groupe A, 9 % de groupe B et 3 % de groupe AB.

Pour ce faire, on utilise une idée similaire à celle vue dans la section précédente. Si nos données respectaient à la lettre l'hypothèse nulle, les fréquences attendues sous l'hypothèse nulle seraient les suivantes :

couleurs des yeux différentes revient à dire que la distribution de la couleur des yeux est la même dans des populations aux couleurs des cheveux différentes (et *vice versa*). Ces deux hypothèses nulles sont également équivalentes à l'hypothèse nulle suivante :

H_0 : les variables « couleur des cheveux » et « couleur des yeux » sont indépendantes.

Cette hypothèse nulle nous dit que le fait de connaître la couleur des yeux d'un écolier ne nous informe en rien sur sa couleur des cheveux, de même que le fait de connaître la couleur des cheveux d'un écolier ne nous informe en rien sur sa couleur des yeux. Ainsi, le test du khi-deux que l'on voit dans cette section est parfois appelé « test d'indépendance pour variables qualitatives ». Notons par ailleurs que ces trois hypothèses nulles nous suggèrent trois façons différentes d'échantillonner les données. La première nous suggère d'échantillonner parmi 4 populations d'écoliers aux couleurs des yeux différentes et de mesurer la couleur des cheveux ; la deuxième d'échantillonner parmi 5 populations d'écoliers aux couleurs des cheveux différentes et de mesurer la couleur des yeux ; la troisième d'échantillonner parmi une population d'écoliers et de mesurer à la fois la couleur des cheveux et la couleur des yeux (ce qui a été fait dans cet exemple). L'équivalence de ces trois hypothèses nulles implique que l'on peut appliquer ce même test du khi-deux, quelle que soit la façon dont on a échantillonné les données.

groupe O	groupe A	groupe B	groupe AB	total
67.9	71.1	14.2	4.7	158
43 %	45 %	9 %	3 %	100 %

Pour juger si les différences entre fréquences observées et fréquences attendues peuvent être mises sur le compte du hasard de l'échantillonnage, on calcule la statistique de test suivante :

$$t_{stat} = \frac{(82-67.9)^2}{67.9} + \frac{(62-71.1)^2}{71.1} + \frac{(10-14.2)^2}{14.2} + \frac{(4-4.7)^2}{4.7} = 5.44.$$

Plus généralement, si on a une variable qualitative avec I valeurs possibles et que l'on dénote par O_i les fréquences observées (le nombre de fois que l'on observe la valeur possible i pour $i = 1, \cdots, I$), et par E_i les fréquences attendues, la statistique de test est définie comme suit :

$$T_{stat} = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}.$$

On peut ici établir mathématiquement que si l'hypothèse nulle était vraie, cette statistique de test aurait (approximativement) une distribution du khi-deux avec $I-1$ *dl*. Cette approximation sera bonne si les fréquences attendues sont supérieures à 5, auquel cas le test du khi-deux est dit valide (notons que cette condition n'est pas tout à fait satisfaite dans notre exemple, où l'on a une fréquence attendue de 4.7). On adopte alors la règle de rejet suivante :

> On rejette H_0 au seuil α si $t_{stat} \geq \chi^2_{1-\alpha,I-1}$.

La valeur p se calcule de manière analogue à ce que l'on a vu dans la section précédente, c'est-à-dire que l'on aura $t_{stat} = \chi^2_{1-p,I-1}$. Dans notre exemple, $t_{stat} = 5.44$ est plus petit que $\chi^2_{0.95,3} = 7.81$, de sorte que la distance entre fréquences observées et fréquences attendues n'est pas assez grande pour rejeter l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera $p = 0.14$; autrement dit, la valeur $t_{stat} = 5.44$ correspond au quantile 86 % d'une distribution du khi-deux avec 3 *dl*)[4]. Il se pourrait donc que la distribution des groupes sanguins dans cette population du nord de l'Europe soit identique à celle que l'on a en France (du moins, on n'a pas réussi à prouver le contraire).

7.3 Comparaison de deux proportions

Dans le cas particulier fréquent d'une table de contingence de dimension $I \times J$ avec $I = J = 2$, l'application d'un test du khi-deux revient à comparer la

[4]Dans R, on pourra ici calculer la valeur p par la commande `1-pchisq(5.44,3)`.

distribution d'une variable binaire entre deux groupes, c'est-à-dire à comparer deux proportions que l'on notera comme d'habitude π_1 et π_0. L'hypothèse nulle sera alors simplement :

$$H_0 : \Lambda = 0$$

où $\Lambda = \pi_1 - \pi_0$. Dans ce qui suit, nous appellerons ce cas particulier de test du khi-deux un *test du khi-deux pour deux proportions.*

Exemple 7.1 *On reprend l'exemple où l'on voulait comparer les proportions de gauchers chez les garçons et de gauchères chez les filles. On considère l'hypothèse nulle donnée par :*

H_0 : *la distribution gauchers/droitiers est la même chez garçons et filles*

qui est équivalente à l'hypothèse nulle suivante :

$$H_0 : \Lambda = 0$$

où $\Lambda = \pi_1 - \pi_0$ est la différence entre la proportion de gauchers π_1 chez les garçons et la proportion de gauchères π_0 chez les filles. À partir de $n_1 = 327$ garçons et $n_0 = 332$ filles, on avait les fréquences observées suivantes :

	gauchers	*droitiers*	
garçons	41 (12.5 %)	286 (87.5 %)	327 (100 %)
filles	25 (7.5 %)	307 (92.5 %)	332 (100 %)
total	66 (10 %)	593 (90 %)	659 (100 %)

Si nos données reproduisaient à la lettre l'hypothèse nulle, on aurait par ailleurs les fréquences attendues suivantes :

	gauchers	*droitiers*	
garçons	32.75 (10 %)	294.25 (90 %)	327 (100 %)
filles	33.25 (10 %)	298.75 (90 %)	332 (100 %)
total	66 (10 %)	593 (90 %)	659 (100 %)

On calcule alors la statistique de test suivante :

$$t_{stat} = \frac{(41 - 32.75)^2}{32.75} + \frac{(286 - 294.25)^2}{294.25} + \frac{(25 - 33.25)^2}{33.25} + \frac{(307 - 298.75)^2}{298.75} = 4.58.$$

La distribution de T_{stat} sous H_0 est une distribution du khi-deux avec 1 dl. On compare donc $t_{stat} = 4.58$ avec $\chi^2_{0.95,1} = 3.84$ et on rejette l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera $p = 0.032$; autrement dit, la valeur $t_{stat} = 4.58$ correspond au quantile 96.8 % d'une distribution du khi-deux avec 1 dl). On conclut que la distribution gauchers/droitiers n'est pas la même chez les garçons et chez les filles (c'est-à-dire que la différence de proportion Λ n'est pas nulle).

Notons que la statistique de test originale d'un test du khi-deux pour deux proportions, que l'on avait définie par :

$$T_{stat} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

peut être alternativement calculée comme suit :

$$T_{stat} = \frac{n_1 n_0}{n_1 + n_0} \cdot \frac{\widehat{\Lambda}^2}{\widehat{\pi}(1-\widehat{\pi})}$$

où n_1 et n_0 représentent les tailles des échantillons dans lesquels on calcule les estimateurs $\widehat{\pi}_1$ et $\widehat{\pi}_0$ de π_1 et π_0, et où l'on définit :

$$\widehat{\pi} = \frac{n_1\widehat{\pi}_1 + n_0\widehat{\pi}_0}{n_1 + n_0}$$

($\widehat{\pi}$ étant ainsi une moyenne de $\widehat{\pi}_1$ et $\widehat{\pi}_0$, pondérée par l'importance des échantillons). Rappelons que l'on rejette H_0 au seuil α si $t_{stat} \geq \chi^2_{1-\alpha,1}$. Or, on peut montrer qu'une distribution normale standardisée pour une variable Y implique une distribution du khi-deux avec $dl = 1$ degré de liberté pour la variable Y^2. On a ainsi :

$$\chi^2_{1-\alpha,1} = z^2_{1-\alpha/2}$$

avec par exemple :

$$\chi^2_{0.95,1} = 3.84 = 1.96^2 = z^2_{0.975}.$$

Dire que :

$$t_{stat} \geq \chi^2_{1-\alpha,1}$$

revient donc à dire que :

$$\sqrt{t_{stat}} \geq z_{1-\alpha/2}.$$

Un test du khi-deux pour deux proportions peut dès lors également s'effectuer en utilisant comme statistique de test la racine carrée de la statistique de test originale (multipliée par le signe de $\widehat{\Lambda}$), c'est-à-dire :

$$T_{stat} = \sqrt{\frac{n_1 n_0}{n_1 + n_0}} \cdot \frac{\widehat{\Lambda}}{\sqrt{\widehat{\pi}(1-\widehat{\pi})}}$$

où l'on rejette H_0 au seuil α si $|t_{stat}| \geq z_{1-\alpha/2}$. Contrairement à la statistique de test originale qui admettait seulement des valeurs positives, cette nouvelle statistique de test admet des valeurs positives ou négatives. Il s'agit d'une *statistique de test signée* (comme on le verra dans le chapitre suivant) qui nous permet d'identifier dans quelle direction on rejette une hypothèse nulle, la règle étant la suivante :

- dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} < 0$ (et donc $\widehat{\Lambda} < 0$), on conclut $\Lambda < 0$

- dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} > 0$ (et donc $\widehat{\Lambda} > 0$), on conclut $\Lambda > 0$.

Exemple 7.2 *On reprend l'exemple ci-dessus où l'on compare la proportion π_1 de gauchers chez les garçons et la proportion π_0 de gauchères chez les filles, avec comme hypothèse nulle $H_0 : \Lambda = 0$, où $\Lambda = \pi_1 - \pi_0$. On avait donc $n_1 = 327$, $n_0 = 332$, $\widehat{\pi}_1 = 41/327 = 0.125$, $\widehat{\pi}_0 = 25/332 = 0.075$ et $\widehat{\Lambda} = 0.125 - 0.075 = 0.05$. On a ainsi :*

$$\widehat{\pi} = \frac{327 \cdot 0.125 + 332 \cdot 0.075}{327 + 332} = \frac{66}{659} = 0.1.$$

On retrouve la statistique de test originale d'un test du khi-deux pour deux proportions en la calculant comme suit :

$$t_{stat} = \frac{327 \cdot 332}{327 + 332} \cdot \frac{0.05^2}{0.1 \cdot 0.9} = 4.58.$$

On comparait $t_{stat} = 4.58$ avec $\chi^2_{0.95,1} = 3.84$. De façon équivalente, on peut calculer une statistique de test signée comme suit :

$$t_{stat} = +\sqrt{4.58} = \sqrt{\frac{332 \cdot 327}{332 + 327}} \cdot \frac{0.05}{\sqrt{0.1 \cdot 0.9}} = 2.14$$

et comparer $|t_{stat}| = 2.14$ avec $z_{0.975} = 1.96$. Comme $|t_{stat}| \geq 1.96$, on rejette H_0 au seuil de 5 %. Comme on a ici $t_{stat} > 0$ (et donc $\widehat{\Lambda} > 0$), on peut en outre conclure $\Lambda > 0$. On a donc prouvé statistiquement non seulement que la distribution gauchers/droitiers n'est pas la même chez les garçons et chez les filles, mais qu'il y a plus de gauchers chez les garçons que de gauchères chez les filles.

7.4 Comparaison d'une proportion avec valeur de référence

L'application d'un test du khi-deux pour la comparaison de la distribution d'une variable qualitative et d'une distribution de référence avec $I = 2$ valeurs possibles revient à comparer une proportion π avec une proportion de référence π^*, donnée par exemple dans la littérature (ou d'un intérêt particulier). L'hypothèse nulle sera alors simplement :

$$H_0 : \pi = \pi^*.$$

Dans ce qui suit, nous appellerons ce cas particulier de test du khi-deux un *test du khi-deux pour une proportion.*

Exemple 7.3 *Prenons l'exemple où l'on aimerait montrer que les deux candidats A et B en lice lors d'une élection n'obtiendront pas le même nombre de voix. On aimerait rejeter l'hypothèse nulle donnée par :*

$$H_0 : \textit{il y a } 50\ \% \textit{ de votes pour } A \textit{ et } 50\ \% \textit{ de votes pour } B$$

qui est équivalente à l'hypothèse nulle suivante :

$$H_0 : \pi = 0.5$$

où π dénote la proportion des voix obtenues par le candidat A, et où l'on a donc $\pi^* = 0.5$. Supposons que parmi $n = 350$ personnes interrogées et représentatives de la population des électeurs, 154 disent qu'ils voteront pour A et 196 disent qu'ils voteront pour B, de sorte que l'on aura les fréquences observées suivantes :

candidat A	*candidat B*	*total*
154	196	350
44 %	56 %	100 %

Si nos données reproduisaient à la lettre l'hypothèse nulle, on aurait par ailleurs les fréquences attendues suivantes :

candidat A	*candidat B*	*total*
175	175	350
50 %	50 %	100 %

On calcule alors la statistique de test suivante :

$$t_{stat} = \frac{(154 - 175)^2}{175} + \frac{(196 - 175)^2}{175} = 5.04.$$

La distribution de T_{stat} sous H_0 est une distribution du khi-deux avec 1 dl. On compare donc t_{stat} avec $\chi^2_{0.05,1} = 3.84$ et on rejette l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera $p = 0.025$; autrement dit, la valeur $t_{stat} = 5.04$ correspond au quantile 97.5 % d'une distribution du khi-deux avec 1 dl). On conclut que les candidats A et B n'obtiendront pas tous deux 50 % des voix (c'est-à-dire que la proportion π des voies obtenues par le candidat A ne sera pas 50 %).

Ici aussi, la statistique de test originale que l'on avait définie par :

$$T_{stat} = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

peut être alternativement calculée comme suit :

$$T_{stat} = \frac{n(\widehat{\pi} - \pi^*)^2}{\pi^*(1 - \pi^*)}$$

où n représente la taille de notre échantillon dans lequel on a calculé l'estimateur $\widehat{\pi}$ de π. Rappelons que l'on rejette H_0 au seuil α si $t_{stat} \geq \chi^2_{1-\alpha,1}$. En utilisant le même raisonnement que dans la section précédente, un test du khi-deux pour une proportion peut également s'effectuer en utilisant comme statistique de test la racine carrée de la statistique originale (multipliée par le signe de $\widehat{\pi} - \pi^*$), qui est une statistique de test signée définie par :

$$T_{stat} = \sqrt{n} \cdot \frac{\widehat{\pi} - \pi^*}{\sqrt{\pi^*(1-\pi^*)}}$$

où l'on rejette H_0 au seuil α si $|t_{stat}| \geq z_{1-\alpha/2}$. L'utilisation de cette statistique de test signée nous permet d'identifier dans quelle direction on rejette H_0 :

- dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} < 0$ (et donc $\widehat{\pi} < \pi^*$), on conclut $\pi < \pi^*$
- dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} > 0$ (et donc $\widehat{\pi} > \pi^*$), on conclut $\pi > \pi^*$.

Exemple 7.4 *On reprend l'exemple ci-dessus où l'on compare la proportion π des voies obtenues par le candidat A avec la proportion de référence $\pi^* = 0.5$, l'hypothèse nulle étant $H_0 : \pi = 0.5$. On avait $n = 350$ et $\widehat{\pi} = 0.44$. On retrouve la statistique de test originale d'un test du khi-deux pour une proportion en la calculant comme suit :*

$$t_{stat} = \frac{350(0.44 - 0.5)^2}{0.5 \cdot 0.5} = 5.04.$$

On comparait $t_{stat} = 5.04$ avec $\chi^2_{0.95,1} = 3.84$. De façon équivalente, on peut calculer une statistique de test signée comme suit :

$$t_{stat} = -\sqrt{5.04} = \sqrt{350} \cdot \frac{0.44 - 0.5}{\sqrt{0.5 \cdot 0.5}} = -2.24$$

et comparer $|t_{stat}| = 2.24$ avec $z_{0.975} = 1.96$. Comme $|t_{stat}| \geq 1.96$, on rejette H_0 au seuil de 5 %. Comme on a ici $t_{stat} < 0$ (et donc $\widehat{\pi} < 0.5$), on peut en outre conclure $\pi < 0.5$. On a donc prouvé statistiquement non seulement que les candidats A et B n'obtiendront pas tous deux 50 % des voix, mais que le candidat A obtiendra moins de 50 % des voix (autrement dit, que c'est le candidat B qui va gagner les élections).

Chapitre 8

Test statistique sur la valeur d'un paramètre

Dans sa formulation originale, la statistique de test utilisée dans un test du khi-deux est une mesure de la distance globale entre les données et l'hypothèse nulle, qui est par définition positive, l'hypothèse nulle étant rejetée si cette distance est trop grande. On va voir dans ce chapitre que lorsque l'hypothèse nulle implique la valeur d'un (et d'un seul) paramètre, on pourra utiliser une statistique de test signée qui aura selon les cas une valeur positive ou une valeur négative. On pourra alors distinguer si la distance entre les données et l'hypothèse nulle est « trop grande positivement » ou « trop grande négativement », ce qui impliquera deux conclusions différentes. On pourra en outre faire dans ce cas la distinction entre un *test unilatéral* et un *test bilatéral*. On verra également dans ce chapitre le principe de dualité entre un test statistique bilatéral sur la valeur d'un paramètre et un intervalle de confiance pour ce paramètre.

8.1 Test unilatéral et test bilatéral

On considère ici le cas général d'un test statistique d'une hypothèse nulle sur la valeur θ^* d'un paramètre θ de la forme :

$$H_0 : \theta = \theta^*.$$

On suppose que l'on dispose d'une statistique de test t_{stat} signée (c'est-à-dire admettant des valeurs positives ou négatives) avec les propriétés suivantes :

- $t_{stat} < 0$ implique $\widehat{\theta} < \theta^*$
- $t_{stat} > 0$ implique $\widehat{\theta} > \theta^*$
- $t_{stat} = 0$ implique $\widehat{\theta} = \theta^*$ (ce qui correspond au cas où les données reproduisent à la lettre l'hypothèse nulle)

- pour simplifier l'exposé, on supposera en outre que la distribution de T_{stat} sous H_0 est symétrique et d'espérance nulle (comme c'est le cas par exemple pour une distribution normale standardisée).

Dans ces conditions, on pourra choisir entre trois règles différentes de rejet de l'hypothèse nulle, qui correspondent à un test unilatéral à gauche (en anglais : `one-sided test to the left`), à un test unilatéral à droite (en anglais : `one-sided test to the right`) et à un test bilatéral (en anglais : `two-sided test`). Ces trois règles de rejet sont définies comme suit :

- **test unilatéral à gauche** : on rejette H_0 uniquement dans des cas où l'on a une trop grande distance négative entre les données et l'hypothèse nulle

  → on rejette H_0 au seuil α si t_{stat} est plus petit que le quantile α de la distribution de T_{stat} sous H_0

  → si la distribution de T_{stat} sous H_0 est normale standardisée, on rejette H_0 au seuil α si $t_{stat} \leq z_\alpha$ (si $t_{stat} \leq -1.645$ au seuil de 5 %)

- **test unilatéral à droite** : on rejette H_0 uniquement dans des cas où l'on a une trop grande distance positive entre les données et l'hypothèse nulle

  → on rejette H_0 au seuil α si t_{stat} est plus grand que le quantile $1-\alpha$ de la distribution de T_{stat} sous H_0

  → si la distribution de T_{stat} sous H_0 est normale standardisée, on rejette H_0 au seuil α si $t_{stat} \geq z_{1-\alpha}$ (si $t_{stat} \geq 1.645$ au seuil de 5 %)

- **test bilatéral** : on rejette H_0 à la fois dans des cas où l'on a une trop grande distance négative, et dans des cas où l'on a une trop grande distance positive entre les données et l'hypothèse nulle

  → on rejette H_0 au seuil α si $|t_{stat}|$ est plus grand que le quantile $1-\alpha/2$ de la distribution de T_{stat} sous H_0

  → si la distribution de T_{stat} sous H_0 est normale standardisée, on rejette H_0 au seuil α si $|t_{stat}| \geq z_{1-\alpha/2}$ (si $|t_{stat}| \geq 1.96$ au seuil de 5 %).

Le calcul de la valeur p (illustré par la figure 8.1) se fait de la manière suivante :

- **test unilatéral à gauche**

  → la valeur p est définie par $p = \Pr\{T_{stat} \leq t_{stat}\}$

  → t_{stat} correspond au quantile p de la distribution de T_{stat} sous H_0

  → si $p \leq \alpha$, on rejette $H_0 : \theta = \theta^*$ au seuil α et on conclut $\theta < \theta^*$

  → il sera par contre impossible de conclure $\theta > \theta^*$

  → lorsque l'on effectue un test unilatéral à gauche, on notera parfois l'hypothèse alternative par $H_1 : \theta < \theta^*$

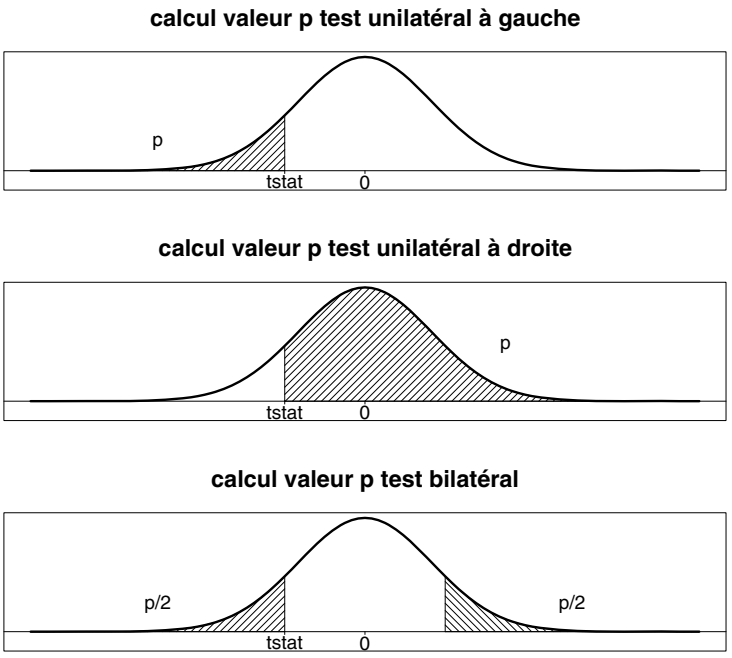

Figure 8.1 – Calcul d'une valeur p pour test unilatéral et bilatéral.

- **test unilatéral à droite**

  → la valeur p est définie par $p = \Pr\{T_{stat} \geq t_{stat}\}$

  → t_{stat} correspond au quantile $1 - p$ de la distribution de T_{stat} sous H_0

  → si $p \leq \alpha$, on rejette $H_0 : \theta = \theta^*$ au seuil α et on conclut $\theta > \theta^*$

  → il sera par contre impossible de conclure $\theta < \theta^*$

  → lorsque l'on effectue un test unilatéral à droite, on notera parfois l'hypothèse alternative par $H_1 : \theta > \theta^*$

- **test bilatéral**

  → la valeur p est définie par $p = \Pr\{|T_{stat}| \geq |t_{stat}|\}$

  → $|t_{stat}|$ correspond au quantile $1 - p/2$ de la distribution de T_{stat} sous H_0

  → la valeur p d'un test bilatéral vaut le double de la plus petite des deux valeurs p obtenues en effectuant des tests unilatéraux à gauche et à droite

  → si $p \leq \alpha$, on rejette $H_0 : \theta = \theta^*$ au seuil α

→ dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} < 0$ (et donc $\widehat{\theta} < \theta^*$), on conclut $\theta < \theta^*$

→ dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} > 0$ (et donc $\widehat{\theta} > \theta^*$), on conclut $\theta > \theta^*$

→ lorsque l'on effectue un test bilatéral, on notera parfois l'hypothèse alternative par $H_1 : \theta \neq \theta^*$.

Exemple 8.1 *On reprend l'exemple où l'on compare la proportion π_1 de gauchers chez les garçons et la proportion π_0 de gauchères chez les filles, où hypothèse nulle était donnée par $H_0 : \Lambda = 0$, avec $\Lambda = \pi_1 - \pi_0$. La statistique de test signée d'un test du khi-deux pour deux proportions était de $t_{stat} = 2.14$. La distribution de la statistique de test signée T_{stat} sous H_0 est ici normale standardisée. Le tableau A.1 donné en annexe nous indique que $t_{stat} = 2.14$ correspond au quantile 98.4 % d'une distribution normale standardisée. Ceci implique les valeurs p et les conclusions suivantes :*

- *dans un test unilatéral à gauche, on a $p = 0.984$, de sorte que l'on ne rejette pas H_0 au seuil de 5 %*
- *dans un test unilatéral à droite, on a $p = 1 - 0.984 = 0.016$, de sorte que l'on rejette H_0 au seuil de 5 % et on conclut $\Lambda > 0$*
- *dans un test bilatéral, on double la plus petite des deux valeurs p ci-dessus et on obtient $p = 2(1 - 0.984) = 0.032$, de sorte que l'on rejette H_0 au seuil de 5 % et on conclut $\Lambda > 0$ (car on a $t_{stat} > 0$ et donc $\widehat{\Lambda} > 0$)*[1].

La valeur p est donc par définition deux fois plus grande dans un test bilatéral que dans un test unilatéral, de sorte qu'il sera plus difficile d'obtenir un résultat significatif dans un test bilatéral que dans un test unilatéral. On peut dès lors se demander pourquoi ne pas opter systématiquement pour un test unilatéral. En fait, on court le risque d'être tenté de choisir la direction du test unilatéral (à gauche ou à droite) après avoir observé les données, ce qui invaliderait la procédure. Dans notre exemple ci-dessus, on pourrait être tenté d'effectuer un test unilatéral à gauche dans les cas où l'on observerait plus de gauchères chez les filles que de gauchers chez les garçons, et d'effectuer un test unilatéral à droite dans les cas où l'on observerait plus de gauchers chez les garçons que de gauchères chez les filles. Dans ces conditions, bien que l'on prétende effectuer un test au seuil de 5 %, on effectuerait en réalité un test au seuil de 10 %. En effet, si l'hypothèse nulle était vraie, c'est-à-dire s'il n'y avait pas de différence entre les proportions de gauchers chez les garçons et de gauchères chez les filles, on aurait une probabilité de 10 % de démontrer

[1] On retrouve ici la valeur $p = 0.032$ que l'on avait calculée pour cet exemple au chapitre 7. Dans un test du khi-deux, la valeur p obtenue en utilisant la statistique de test originale correspond donc à la valeur p obtenue en utilisant la statistique de test signée, calculée dans un test bilatéral (non unilatéral).

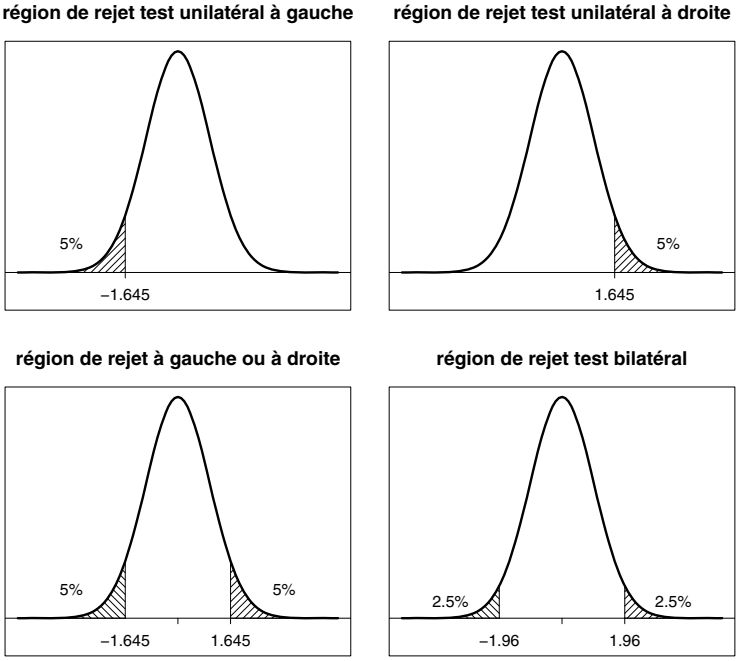

Figure 8.2 – Région de rejet à 5 % avec T_{stat} normale standardisée sous H_0.

à tort qu'il y en a une (5 % de démontrer à tort qu'il y a plus de gauchères que de gauchers, et 5 % de démontrer à tort qu'il y a plus de gauchers que de gauchères, ce qui représente une inflation du risque de première espèce).

En fait, le choix d'effectuer un test unilatéral ou un test bilatéral devrait dépendre de notre hypothèse alternative (scientifique), c'est-à-dire de notre motivation à vouloir rejeter une hypothèse nulle. On a la situation suivante :

- on optera pour un test unilatéral à gauche si notre hypothèse scientifique est $\theta < \theta^*$, et s'il n'y a aucun sens/intérêt à vouloir démontrer $\theta > \theta^*$
- on optera pour un test unilatéral à droite si notre hypothèse scientifique est $\theta > \theta^*$, et s'il n'y a aucun sens/intérêt à vouloir démontrer $\theta < \theta^*$
- si aucune des situations $\theta < \theta^*$ ou $\theta > \theta^*$ n'est *a priori* ni complètement absurde ni complètement inintéressante, on optera pour un test bilatéral.

La figure 8.2 nous montre les régions de rejet au seuil de 5 % (qui correspondent aux surfaces hachurées) dans le cas typique où la statistique de test a une distribution normale standardisée sous H_0. Dans un test unilatéral à gauche, on rejette H_0 au seuil de 5 % si $t_{stat} \leq -1.645$ (graphique en haut à gauche). Dans un test unilatéral à droite, on rejette H_0 au seuil de 5 % si

$t_{stat} \geq 1.645$ (graphique en haut à droite). Si on choisissait la direction du test unilatéral (à gauche ou à droite) selon que t_{stat} soit négatif ou positif, on rejetterait H_0 si $|t_{stat}| \geq 1.645$, ce qui correspondrait à un seuil de 10 % au lieu de 5 % (graphique en bas à gauche). Dans un test bilatéral, afin de garantir un seuil de 5 %, on rejette donc H_0 si $|t_{stat}| \geq 1.96$ (graphique en bas à droite).

8.2 Test statistique *versus* intervalle de confiance

Un avantage d'utiliser un test bilatéral plutôt qu'un test unilatéral est que la conclusion du test peut directement se déduire du calcul d'un intervalle de confiance. En effet, si on considère un test de l'hypothèse nulle $H_0 : \theta = \theta^*$ applicable à n'importe quelle valeur possible θ^* pour θ, on a le principe fondamental suivant :

$$\begin{array}{c} H_0 : \theta = \theta^* \text{ rejetée au seuil } \alpha \text{ (test bilatéral)} \\ \Longleftrightarrow \\ \theta^* \notin \text{CI pour } \theta \text{ au niveau } 1-\alpha. \end{array}$$

En particulier :

$$\begin{array}{c} H_0 : \theta = \theta^* \text{ rejetée au seuil 5 \% (test bilatéral)} \\ \Longleftrightarrow \\ \theta^* \notin \text{95 \% CI pour } \theta. \end{array}$$

Tests statistiques et intervalles de confiance sont donc intimement liés :

- si on dispose d'un test statistique pour une hypothèse nulle de la forme $H_0 : \theta = \theta^*$, on peut définir un intervalle de confiance au niveau $1-\alpha$ pour le paramètre θ comme étant l'intervalle qui contient toutes les valeurs θ^* qui ne sont pas rejetées au seuil α dans un test bilatéral
- si on dispose d'une formule pour un intervalle de confiance pour θ, on peut définir la valeur $1-p$ d'un test statistique bilatéral de l'hypothèse nulle $H_0 : \theta = \theta^*$ comme étant le niveau maximal en deça duquel l'intervalle de confiance ne contient pas la valeur testée θ^*.

On a ainsi une dualité entre un test statistique bilatéral sur la valeur d'un paramètre et un intervalle de confiance pour ce paramètre. Par exemple, le test

du khi-deux pour une proportion que l'on a vu en fin de chapitre précédent est en dualité avec l'intervalle de confiance de Wilson pour une proportion[2].

On pourrait dès lors penser que tests statistiques et intervalles de confiance sont totalement redondants. En fait, ils sont plus complémentaires que redondants, comme on va l'illustrer dans les deux prochains exemples.

Exemple 8.2 *Dans notre exemple où l'on voulait déterminer lequel des deux candidats A et B allait gagner les élections, on avait estimé la proportion π de votes pour le candidat A par $\widehat{\pi} = 154/350 = 0.44$. Un test du khi-deux (bilatéral) pour une proportion nous donne $p = 0.025$, alors qu'un intervalle de confiance de Wilson au niveau 95 % pour π nous donne $[0.39; 0.49]$. En utilisant la dualité entre ce test statistique et cet intervalle de confiance, on obtient les informations suivantes :*

- *la valeur $p = 0.025$ nous indique que le niveau maximal d'un intervalle de confiance de Wilson ne contenant pas la valeur testée $\pi^* = 0.5$ est 97.5 %*

  $\rightarrow$ *on sait qu'un 95 % CI pour π ne contient pas la valeur 0.5*

  $\rightarrow$ *on sait qu'un 99 % CI pour π contient la valeur 0.5*

- *l'intervalle de confiance $[0.39; 0.49]$ contient toutes les valeurs π^* qui ne sont pas rejetées dans un test du khi-deux bilatéral au seuil de 5 %*

  $\rightarrow$ *$H_0 : \pi = 0.5$ est rejetée dans un test bilatéral au seuil de 5 %*

  $\rightarrow$ *$H_0 : \pi = 0.4$ n'est pas rejetée dans un test bilatéral au seuil de 5 %.*

La question est à présent de savoir quel résultat de statistique inférentielle (valeur p et/ou intervalle de confiance) reporter aux côtés de l'estimation du paramètre d'intérêt dans un abstract de publication scientifique. L'exemple suivant nous illustre trois cas de figure (report de la valeur p, report d'un intervalle de confiance, report des deux).

Exemple 8.3 *On reprend l'exemple introduit au chapitre 4 des $n = 10$ personnes qui ont suivi un régime de deux mois. L'effet du régime était défini par la moyenne μ des pertes de poids. On avait observé une perte de poids moyenne de $\widehat{\mu} = 0.72$ kg. Afin de prouver statistiquement que le régime a un effet, il faudrait rejeter l'hypothèse nulle $H_0 : \mu = 0$ (et conclure $\mu > 0$). On verra au chapitre suivant que l'on pourra ici utiliser un test de Student. Dans un test bilatéral, on obtient $p = 0.08$ de sorte que cela ne suffit pas pour rejeter l'hypothèse nulle au seuil de 5 %. On verra aussi que ce test est en dualité avec un intervalle de confiance de Student pour μ. Au niveau 95 %, on avait calculé $[-0.10; 1.54]$*

[2]Par contre, le test du khi-deux pour deux proportions n'est pas en dualité avec un intervalle de confiance pour une différence de proportion, car ce test ne s'applique qu'à l'hypothèse nulle particulière $H_0 : \Lambda = 0$ (la valeur testée est ici forcément 0 et non une autre valeur possible pour le paramètre Λ ; on ne pourra donc pas calculer un intervalle contenant toutes les valeurs non rejetées par un tel test).

kg, ce qui confirme le non-rejet de H_0 *(la valeur testée* 0 *étant à l'intérieur de cet intervalle de confiance). Discutons à présent des trois possibilités de report de cette information dans un abstract de publication (avec dans chaque cas les informations que le lecteur de cet abstract pourra en déduire) :*

- **report de la valeur** p *(ici* $p = 0.08$*) :*

  → *par convention, cette valeur p se réfère à l'hypothèse nulle particulière* $H_0 : \mu = 0$

  → *on sait donc que l'hypothèse nulle* $H_0 : \mu = 0$ *n'est pas rejetée au seuil de* 5 %

  → *on sait par ailleurs que l'hypothèse nulle* $H_0 : \mu = 0$ *serait rejetée au seuil de* 10 %

  → *autrement dit, on sait que la valeur* 0 *appartient à un* 95 % *CI pour* μ*, mais n'appartient pas à un* 90 % *CI pour* μ

  → *la publication d'une valeur p nous permet donc de choisir le seuil du test, mais nous impose la valeur testée (ici* $H_0 : \mu = 0$*)*

- **report d'un** 95 % **CI** *(ici* $[-0.10; 1.54]$ *kg) :*

  → *on sait que l'hypothèse nulle* $H_0 : \mu = 0$ *n'est pas rejetée au seuil de* 5 %

  → *on sait aussi par exemple que l'hypothèse nulle* $H_0 : \mu = 1$ *ne serait pas rejetée au seuil de* 5 %*, alors que l'hypothèse nulle* $H_0 : \mu = -1$ *serait rejetée au seuil de* 5 %

  → *la publication d'un* 95 % *CI nous permet donc de choisir l'hypothèse nulle testée, mais nous impose le seuil du test (*$\alpha = 5$ %*)*

- **report de la valeur** p **et d'un** 95 % **CI**

  → *on sait ce qui se passe lorsque l'on change de seuil pour l'hypothèse nulle particulière* $H_0 : \mu = 0$

  → *on sait ce qui se passe lorsque l'on change d'hypothèse nulle au seuil particulier de* $\alpha = 5$ %

  → *on dispose ici d'une information plus complète que dans les deux premiers abstracts.*

En résumé :

- si les conclusions de notre étude dépendent crucialement du rejet ou du non-rejet d'une hypothèse nulle particulière $H_0 : \theta = \theta^*$, et donc de la conclusion $\theta < \theta^*$ ou $\theta > \theta^*$, on reportera avant tout la valeur p

- si on s'intéresse à la valeur d'un paramètre θ (et non simplement de savoir s'il est supérieur ou inférieur à une valeur particulière), on reportera avant tout l'intervalle de confiance.

Dans notre dernier exemple, s'il est crucial de connaître le signe de l'effet du régime (négatif ou positif), on s'intéressera plutôt à la valeur p. Si on veut connaître l'importance de cet effet, on s'intéressera plutôt à l'intervalle de confiance. En pratique, c'est souvent le second cas de figure qui prévaut, ce qui devrait nous inciter à reporter un intervalle de confiance plutôt qu'une valeur p[3].

8.3 Test d'équivalence

On a mentionné au chapitre 6 qu'il n'est pas possible de démontrer statistiquement qu'une hypothèse nulle est vraie. Si on compare par exemple deux médicaments (un nouveau médicament et un médicament standard) par rapport à leur efficacité mesurée sur une échelle binaire (1 = succès ; 0 = échec), on pourra essayer de démontrer statistiquement $H_1 : \Lambda > 0$, que le nouveau médicament est supérieur au médicament standard, mais on ne pourra pas démontrer statistiquement $H_0 : \Lambda = 0$, que les deux médicaments ont même efficacité (où Λ dénote la différence des proportions de succès entre les deux médicaments). Dans certains cas, il serait pourtant intéressant de pouvoir le faire, par exemple lorsque l'efficacité du médicament standard est déjà établie et que le nouveau médicament a par ailleurs certains avantages secondaires par rapport au médicament standard, tel un coût de production moins élevé.

Afin de pouvoir aborder statistiquement ce genre de question, on la posera de façon un peu différente. On renoncera tout d'abord à vouloir démontrer une égalité parfaite. On se contentera de démontrer que les deux médicaments sont *cliniquement équivalents*, ce qui sera le cas si la différence des proportions de succès Λ se trouve à l'intérieur d'un *domaine d'équivalence* $[\Lambda^-; \Lambda^+]$, par exemple $[-0.05; +0.05]$, que l'on aura préspécifié. Afin de prouver statistiquement l'équivalence des deux médicaments, il s'agira donc de prouver statistiquement que $\Lambda > \Lambda^-$ et que $\Lambda < \Lambda^+$. Ceci peut se faire en utilisant deux tests statistiques classiques :

- on essaiera de rejeter $H_0 : \Lambda = \Lambda^-$ (dans notre exemple $H_0 : \Lambda = -0.05$) dans un test unilatéral à droite, de sorte que l'on puisse conclure $\Lambda > \Lambda^-$
- on essaiera de rejeter $H_0 : \Lambda = \Lambda^+$ (dans notre exemple $H_0 : \Lambda = 0.05$) dans un test unilatéral à gauche, de sorte que l'on puisse conclure $\Lambda < \Lambda^+$.

On laissera le lecteur se convaincre que la dualité entre test statistique et intervalle de confiance implique que ces deux hypothèses nulles seront rejetées au seuil de 5 % si (et seulement si) un intervalle de confiance au niveau 90 % pour Λ est entièrement inclus dans $[\Lambda^-; \Lambda^+]$[4]. Une convention (discutable) nous incite pourtant à utiliser ici aussi un intervalle de confiance au niveau 95 % (au lieu de 90 %) pour Λ. On adopte alors le principe suivant :

[3] Les publications biomédicales favorisent cependant largement les valeurs p au détriment des intervalles de confiance.

[4] Voir par exemple l'article de Schuirmann (1987).

On aura prouvé statistiquement que deux médicaments sont cliniquement équivalents si un 95 % CI pour leur différence Λ est entièrement inclus dans le domaine d'équivalence préspécifié.

Exemple 8.4 *Supposons que deux médicaments sont considérés comme cliniquement équivalents si la différence entre leurs proportions de succès est plus petite que* 0.05 *(autrement dit, le domaine d'équivalence pour* Λ *est fixé à* $[-0.05; +0.05]$*). On traite* $n_1 = 100$ *patients avec un nouveau médicament et* $n_0 = 100$ *patients avec un médicament standard et on observe* 80 *succès dans chaque groupe. On a ainsi* $\widehat{\pi}_1 = \widehat{\pi}_0 = 0.8$ *et* $\widehat{\Lambda} = 0$ *et un intervalle de confiance de Wald au niveau* 95 % *pour* Λ *nous donne :*

$$0 \pm 1.96 \cdot \sqrt{\frac{0.8 \cdot 0.2}{100} + \frac{0.8 \cdot 0.2}{100}} = [-0.11; 0.11].$$

Comme cet intervalle de confiance n'est pas entièrement inclus dans le domaine d'équivalence, on n'a pas réussi à démontrer statistiquement l'équivalence des deux médicaments (bien que l'on ait observé exactement les mêmes proportions de succès pour les deux médicaments dans nos échantillons).

La figure 8.3 illustre différentes situations possibles. Dans notre exemple de médicaments, l'axe horizontal représente l'ensemble des valeurs possibles pour Λ. Une valeur nulle pour Λ indique que les deux médicaments ont exactement la même efficacité, alors qu'une valeur positive favorise le nouveau médicament et qu'une valeur négative favorise le médicament standard. À partir de là, il s'agit de définir un domaine d'équivalence pour Λ, dont les bornes Λ^- et Λ^+ sont représentées par les deux traits verticaux pleins. Notons que le choix de Λ^- et Λ^+ dépend de considérations cliniques et ne constitue pas une question statistique. Il s'agit ensuite à partir de nos données de calculer un intervalle de confiance au niveau 95 % pour Λ et de constater s'il est ou non entièrement inclus dans le domaine d'équivalence. Nous avons ici cinq exemples. Le premier intervalle (depuis le haut) de cette figure recouvre complètement le domaine d'équivalence. On n'aura ici statistiquement rien prouvé. Le deuxième et le dernier de ces intervalles ne sont pas non plus entièrement inclus dans le domaine d'équivalence, de sorte que l'on ne pourra pas non plus conclure à l'équivalence des deux médicaments. Notons cependant que la borne inférieure de ces intervalles de confiance est bel et bien plus grande que la borne inférieure du domaine d'équivalence. On dira dans un tel cas que l'on a prouvé statistiquement la *non-infériorité* du nouveau médicament. Les troisième et quatrième intervalles sont par contre entièrement inclus dans le domaine d'équivalence. On aura ici prouvé statistiquement que les deux médicaments sont cliniquement équivalents.

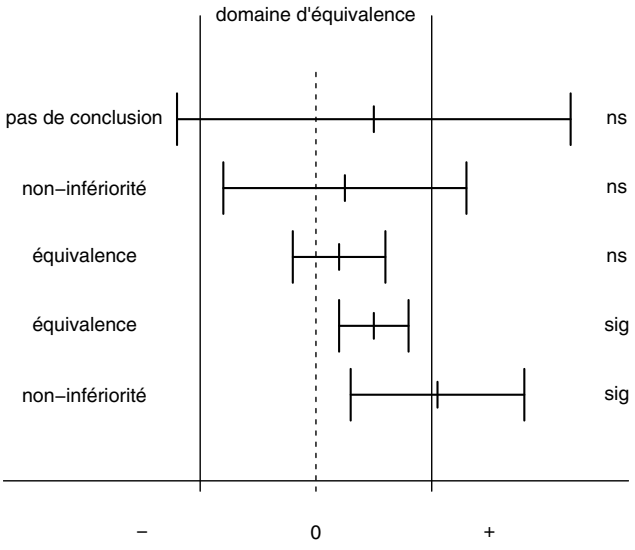

Figure 8.3 – Test d'équivalence *via* le calcul d'un intervalle de confiance.

Nous avons également indiqué à droite de ces intervalles si le résultat d'un test bilatéral de l'hypothèse nulle $H_0 : \Lambda = 0$ est significatif (sig) ou non significatif (ns) au seuil de 5 %. Rappelons que l'on aura un résultat significatif au seuil de 5 % si et seulement si la valeur 0 ne se trouve pas à l'intérieur de l'intervalle de confiance au niveau 95 % pour Λ. Les trois premiers intervalles représentent des cas de résultats non significatifs, les deux derniers des cas de résultats significatifs. Le quatrième intervalle, où l'on a un résultat significatif, nous montre un cas à première vue contradictoire. On aura ici réussi à prouver statistiquement à la fois que le nouveau médicament est plus efficace que le médicament standard ($\Lambda > 0$) et que les deux médicaments sont cliniquement équivalents ($\Lambda \in [\Lambda^-; \Lambda^+]$). Cela illustre le fait qu'un résultat significatif n'implique pas forcément un résultat cliniquement important (on en rediscutera au chapitre 10). De même, un résultat non significatif n'implique pas forcément une équivalence, comme nous le montre le premier intervalle de cette figure.

Chapitre 9

Tests de Wald et de Student

Nous présentons dans ce chapitre des exemples de tests de Wald, ainsi que les tests de Student et le test de Welch. Ce sont des tests sur la valeur d'un paramètre, où l'on pourra appliquer les principes vus dans le chapitre précédent.

9.1 Méthode de Wald

On considère une hypothèse nulle sur la valeur d'un paramètre de la forme :

$$H_0 : \theta = \theta^*$$

où θ représente le paramètre d'intérêt (par exemple une proportion, une différence de proportion, une moyenne ou une différence de moyenne) et θ^* dénote la valeur testée que l'on aimerait rejeter. Un *test de Wald* consiste à utiliser un estimateur $\widehat{\theta}$ de θ pour construire la statistique de test qui est définie comme suit :

$$T_{stat} = \frac{\widehat{\theta} - \theta^*}{\mathrm{SE}(\widehat{\theta})}.$$

Dans le cas particulier (et fréquent) où la valeur testée est $\theta^* = 0$, la statistique de test d'un test de Wald est donc simplement $\widehat{\theta}/\mathrm{SE}(\widehat{\theta})$, l'estimateur divisé par son erreur type. Si l'estimateur est sans biais et (approximativement) normalement distribué, cette statistique de test aura sous H_0 (approximativement) une distribution normale standardisée. On rejette ainsi H_0 au seuil α :

- si $t_{stat} \leq z_\alpha$ (si $t_{stat} \leq -1.645$ au seuil de 5 %) dans un test unilatéral à gauche (auquel cas on conclut $\theta < \theta^*$)
- si $t_{stat} \geq z_{1-\alpha}$ (si $t_{stat} \geq 1.645$ au seuil de 5 %) dans un test unilatéral à droite (auquel cas on conclut $\theta > \theta^*$)
- si $|t_{stat}| \geq z_{1-\alpha/2}$ (si $|t_{stat}| \geq 1.96$ au seuil de 5 %) dans un test bilatéral

→ dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} < 0$ (et donc $\widehat{\theta} < \theta^*$), on conclut $\theta < \theta^*$

→ dans le cas d'un rejet de H_0 avec par ailleurs $t_{stat} > 0$ (et donc $\widehat{\theta} > \theta^*$), on conclut $\theta > \theta^*$.

La valeur p se calcule comme suit[1] :

- $p = \Pr\{T_{stat} \leq t_{stat}\} = \Phi_{0,1}(t_{stat})$ dans un test unilatéral à gauche (t_{stat} correspond au quantile p d'une distribution normale standardisée)
- $p = \Pr\{T_{stat} \geq t_{stat}\} = \Phi_{0,1}(-t_{stat})$ dans un test unilatéral à droite (t_{stat} correspond au quantile $1-p$ d'une distribution normale standardisée)
- $p = \Pr\{|T_{stat}| \geq |t_{stat}|\} = 2\Phi_{0,1}(-|t_{stat}|)$ dans un test bilatéral ($|t_{stat}|$ correspond au quantile $1-p/2$ d'une distribution normale standardisée).

Un test de Wald de l'hypothèse nulle $H_0 : \theta = \theta^*$ est en dualité avec un intervalle de confiance de Wald pour le paramètre θ. Ainsi, un intervalle de confiance de Wald au niveau $1-\alpha$ pour θ contient toutes les valeurs θ^* qui ne sont pas rejetées au seuil α dans un test de Wald bilatéral.

9.2 Test de Wald pour une proportion

Lorsque l'on désire comparer une proportion π avec une valeur de référence π^*, on considère l'hypothèse nulle suivante :

$$H_0 : \pi = \pi^*.$$

La statistique de test d'un test de Wald pour une proportion est donnée par :

$$T_{stat} = \frac{\widehat{\pi} - \pi^*}{\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}}$$

où n représente la taille de l'échantillon et $\widehat{\pi}$ la proportion empirique. Ce test de Wald est en dualité avec un intervalle de confiance de Wald pour π. Comme la distribution de T_{stat} sous H_0 n'est pas exactement normale, il ne s'agit pas d'un test exact. Le test sera néanmoins valide si n est assez grand et si la valeur testée π^* n'est pas trop proche de 0 ou de 1 (ce sont les mêmes conditions de validité que pour l'intervalle de confiance dual).

Exemple 9.1 *On reprend l'exemple où l'on voulait déterminer quel candidat A ou B allait gagner les élections, avec* $H_0 : \pi = 0.5$ *(où* π *dénote la proportion*

[1] Dans R, on pourra ici calculer la valeur p à partir de la statistique de test `tstat` par la commande `pnorm(tstat)`, `pnorm(-tstat)` ou `2*pnorm(-abs(tstat))`, respectivement dans un test unilatéral à gauche, dans un test unilatéral à droite et dans un test bilatéral.

des électeurs votant pour A et où on considère donc $\pi^* = 0.5$*),* $n = 350$ *et* $\widehat{\pi} = 0.44$*. Pour effectuer un test de Wald, on calcule la statistique de test :*

$$t_{stat} = \frac{0.44 - 0.5}{\sqrt{\frac{0.44 \cdot 0.56}{350}}} = -2.26.$$

On rejette H_0 *au seuil de* 5 % *car* $|t_{stat}| \geq 1.96$ *(test bilatéral). Comme on a par ailleurs* $\widehat{\pi} < 0.5$*, on conclut* $\pi < 0.5$ *et donc que le candidat B va gagner les élections. Afin de calculer la valeur p, notons que* $|t_{stat}| = 2.26$ *correspond au quantile* 98.8 % *d'une distribution normale standardisée, de sorte que l'on a* $p = 2(1 - 0.988) = 0.024$*.*

On a vu toutefois que l'hypothèse nulle $H_0 : \pi = \pi^*$ peut également se tester avec un test du khi-deux pour une proportion, qui est en dualité avec un intervalle de confiance de Wilson pour π. De la même manière que la validité d'un intervalle de confiance de Wilson est en général supérieure à la validité d'un intervalle de confiance de Wald, la validité d'un test du khi-deux est en général supérieure à la validité d'un test de Wald[2]. On préférera ainsi le test du khi-deux au test de Wald, les résultats des deux tests étant toutefois similaires avec un grand n (dans l'exemple précédent, on a $t_{stat} = -2.24$ et $p = 0.025$ avec un test du khi-deux et $t_{stat} = -2.26$ et $p = 0.024$ avec un test de Wald).

9.3 Test de Wald pour une différence de proportion

Lorsqu'il s'agit de comparer deux proportions π_1 et π_0, on a vu qu'un test du khi-deux nous permettait de tester l'hypothèse nulle suivante :

$$H_0 : \Lambda = 0$$

où $\Lambda = \pi_1 - \pi_0$ représente la différence des proportions. Un test de Wald pour une différence de proportion nous permet de considérer une hypothèse nulle plus générale de la forme :

$$H_0 : \Lambda = \Lambda^*$$

[2]Il est par ailleurs intéressant de comparer les statistiques de test de ces deux tests. Pour un test de Wald, on a donc :

$$T_{stat} = \frac{\widehat{\pi} - \pi^*}{\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}}$$

alors que pour un test du khi-deux, la statistique de test signée est donnée par :

$$T_{stat} = \frac{\widehat{\pi} - \pi^*}{\sqrt{\frac{\pi^*(1-\pi^*)}{n}}}.$$

La différence entre les deux statistiques de test apparaît au dénominateur, c'est-à-dire au niveau de l'estimation de l'erreur type de $\widehat{\pi}$, qui vaut en réalité $\mathrm{SE}(\widehat{\pi}) = \sqrt{\pi(1-\pi)/n}$. Cette erreur type est estimée empiriquement dans un test du Wald, alors qu'elle est estimée sous l'hypothèse nulle dans un test du khi-deux, ce qui permet d'améliorer la validité du test.

où la valeur testée Λ^* pourra être n'importe quelle valeur possible pour Λ (pas forcément $\Lambda^* = 0$). La statistique de test de ce test de Wald est donnée par :

$$T_{stat} = \frac{\widehat{\Lambda} - \Lambda^*}{\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_0(1-\widehat{\pi}_0)}{n_0}}}$$

où n_1 et n_0 représentent les tailles des deux échantillons, $\widehat{\pi}_1$ et $\widehat{\pi}_0$ les proportions empiriques, et $\widehat{\Lambda} = \widehat{\pi}_1 - \widehat{\pi}_0$ la différence de proportion empirique. Ce test de Wald est en dualité avec un intervalle de confiance de Wald pour Λ. Comme la distribution de T_{stat} sous H_0 n'est pas exactement normale, il ne s'agit pas d'un test exact. Le test sera néanmoins valide si n_1 et n_0 sont assez grands, et si les véritables proportions π_1 et π_0 qui sont comparées ne sont pas trop proches de 0 ou de 1 (ce sont les mêmes conditions de validité que pour l'intervalle de confiance dual).

Exemple 9.2 *On reprend l'exemple où l'on comparait les proportions de gauchers π_1 chez les garçons et de gauchères π_0 chez les filles, avec $H_0 : \Lambda = 0$, $n_1 = 327$, $n_0 = 332$, $\widehat{\pi}_1 = 0.125$, $\widehat{\pi}_0 = 0.075$, et donc $\widehat{\Lambda} = 0.05$. Pour effectuer un test de Wald, on calcule la statistique de test :*

$$t_{stat} = \frac{0.05}{\sqrt{\frac{0.125 \cdot 0.875}{327} + \frac{0.075 \cdot 0.925}{332}}} = 2.14.$$

On rejette H_0 au seuil de 5 % car $|t_{stat}| \geq 1.96$ (test bilatéral). Comme on a par ailleurs $\widehat{\Lambda} > 0$, on conclut $\Lambda > 0$ et donc qu'il y a plus de gauchers chez les garçons que de gauchères chez les filles. Afin de calculer la valeur p, notons que $|t_{stat}| = 2.14$ correspond au quantile 98.4 % d'une distribution normale standardisée, de sorte que l'on aura $p = 2(1 - 0.984) = 0.032$ (en retenant deux décimales dans nos calculs, on obtient exactement le même résultat qu'avec un test du khi-deux).

Dans le cas particulier où la valeur testée est $\Lambda^* = 0$, on aura donc le choix entre un test du khi-deux et un test de Wald. Comme la validité du premier est en général supérieure à la validité du second, on préférera ici aussi le test du khi-deux au test de Wald (les résultats des deux tests étant toutefois similaires pour de grands n_1 et n_0, comme illustré dans l'exemple précédent)[3].

[3]Ici aussi, il est intéressant de comparer les statistiques de test de ces deux tests. Pour un test de Wald, on a donc dans ce cas particulier :

$$T_{stat} = \frac{\widehat{\Lambda}}{\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_0(1-\widehat{\pi}_0)}{n_0}}}$$

alors que pour un test d'un test du khi-deux, la statistique de test signée est donnée par :

$$T_{stat} = \frac{\widehat{\Lambda}}{\sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n_1} + \frac{\widehat{\pi}(1-\widehat{\pi})}{n_0}}}$$

9.4 Test de Student pour une moyenne

Lorsque l'on désire comparer la moyenne μ d'une variable continue avec une valeur de référence μ^*, on considère l'hypothèse nulle suivante :

$$H_0 : \mu = \mu^*.$$

La statistique de test d'un test de Wald est donnée ici par :

$$T_{stat} = \frac{\widehat{\mu} - \mu^*}{\tilde{\sigma}/\sqrt{n}}$$

où n représente la taille de l'échantillon, $\widehat{\mu}$ et $\tilde{\sigma}^2$ dénotant nos estimateurs de la moyenne μ et de la variance σ^2 de la variable en question.

Un test de Wald pour une moyenne est en dualité avec un intervalle de confiance de Wald pour cette moyenne. On a vu cependant au chapitre 4 que lorsqu'il s'agit de calculer un intervalle de confiance pour une moyenne, on préfère calculer un intervalle de confiance de Student, dont la validité est en général supérieure, et qui a l'avantage d'être exact lorsque les observations sont normales. Le test statistique qui est en dualité avec un intervalle de confiance de Student est appelé *test de Student* (en anglais : `one-sample t-test`). Pour effectuer ce test, on compare la statistique de test ci-dessus non pas à une distribution normale standardisée (comme on le ferait pour un test de Wald), mais à une distribution de Student avec $n-1$ *dl*. On rejette ainsi H_0 au seuil α :

- si $t_{stat} \leq t_{\alpha,n-1}$ dans un test unilatéral à gauche
- si $t_{stat} \geq t_{1-\alpha,n-1}$ dans un test unilatéral à droite
- si $|t_{stat}| \geq t_{1-\alpha/2,n-1}$ dans un test bilatéral.

Dans le cas d'un rejet de H_0, on conclut comme d'habitude :

- $\mu < \mu^*$ dans le cas où $t_{stat} < 0$ (et donc $\widehat{\mu} < \mu^*$)
- $\mu > \mu^*$ dans le cas où $t_{stat} > 0$ (et donc $\widehat{\mu} > \mu^*$).

Le calcul de la valeur p se fera le plus souvent à l'aide d'un logiciel statistique car il implique les quantiles d'une distribution de Student de la façon suivante[4] :

où $\widehat{\pi} = (n_1\widehat{\pi}_1 + n_0\widehat{\pi}_0)/(n_1+n_0)$ est une moyenne pondérée de $\widehat{\pi}_1$ et $\widehat{\pi}_0$. La différence apparaît à nouveau au dénominateur, c'est-à-dire au niveau de l'estimation de l'erreur type de $\widehat{\Lambda}$, qui vaut en réalité $\mathrm{SE}(\widehat{\Lambda}) = \sqrt{(\pi_1(1-\pi_1))/n_1 + (\pi_0(1-\pi_0))/n_0}$. Alors que cette erreur type est estimée empiriquement dans un test de Wald, elle est (en quelque sorte) estimée sous l'hypothèse nulle dans un test du khi-deux, où l'on suppose $\pi_1 = \pi_0$, d'où l'estimation d'une « proportion commune » $\widehat{\pi}$, ce qui permet d'améliorer la validité du test.

[4] Dans R, on pourra ici calculer la valeur p à partir de la statistique de test `tstat` par la commande `pt(tstat,n-1)`, `pt(-tstat,n-1)` ou `2*pt(-abs(tstat),n-1)`, respectivement dans un test unilatéral à gauche, dans un test unilatéral à droite et dans un test bilatéral, où `n` dénote la taille de l'échantillon.

- $p = \Pr\{T_{stat} \leq t_{stat}\}$ dans un test unilatéral à gauche (t_{stat} correspond au quantile p d'une distribution de Student avec $n-1$ *dl*)
- $p = \Pr\{T_{stat} \geq t_{stat}\}$ dans un test unilatéral à droite (t_{stat} correspond au quantile $1-p$ d'une distribution de Student avec $n-1$ *dl*)
- $p = \Pr\{|T_{stat}| \geq |t_{stat}|\}$ dans un test bilatéral ($|t_{stat}|$ correspond au quantile $1-p/2$ d'une distribution de Student avec $n-1$ *dl*).

Il s'agit d'un *test exact* si les observations sont normales. En absence de normalité, le test sera valide pour autant que n soit suffisamment grand (ce sont les mêmes conditions de validité que pour l'intervalle de confiance dual).

Exemple 9.3 *On veut démontrer que les enfants nés prématurément commencent à marcher plus tard que les enfants nés à terme, qui eux marchent en moyenne après* $\mu^* = 12$ *mois (selon la littérature). L'hypothèse nulle que l'on essaie de rejeter est donc :*

$$H_0 : \mu = 12$$

où μ *représente la moyenne de la variable continue « âge des premiers pas ». Admettons qu'à partir d'un échantillon de* $n = 10$ *enfants nés prématurément, on observe* $\widehat{\mu} = 12.8$ *mois et* $\tilde{\sigma} = 1.8$ *mois. On calcule alors :*

$$t_{stat} = \frac{12.8 - 12}{1.8/\sqrt{10}} = 1.41$$

Dans un test bilatéral, on compare $|t_{stat}| = 1.41$ *avec* $t_{0.975,9} = 2.26$*, de sorte que* H_0 *n'est pas rejetée au seuil de* 5 % *(un logiciel statistique nous donnera* $p = 0.19$*). Notre hypothèse scientifique n'est donc pas statistiquement démontrée.*

9.5 Test de Welch pour une différence de moyenne

Lorsque l'on désire comparer les moyennes μ_1 et μ_0 de deux variables continues, on considère l'hypothèse nulle suivante :

$$H_0 : \Delta = \Delta^*$$

où $\Delta = \mu_1 - \mu_0$ représente la différence de moyenne et Δ^* la valeur testée (on considérera souvent $\Delta^* = 0$). La statistique de test d'un test de Wald est donnée ici par :

$$T_{stat} = \frac{\widehat{\Delta} - \Delta^*}{\sqrt{\frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0}}}$$

où n_1 et n_0 représentent les tailles des échantillons, $\widehat{\mu}_1$ et $\widehat{\mu}_0$ nos estimateurs des moyennes et $\tilde{\sigma}_1^2$ et $\tilde{\sigma}_0^2$ nos estimateurs des variances des variables en question, et $\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0$ la différence de moyenne empirique.

Un test de Wald pour une différence de moyenne est en dualité avec un intervalle de confiance de Wald pour cette différence de moyenne. On a vu cependant au chapitre 5 que lorsqu'il s'agit de calculer un intervalle de confiance pour une différence de moyenne, on préfère calculer un intervalle de confiance de Welch, dont la validité est supérieure. Le test en dualité avec un intervalle de confiance de Welch est appelé *test de Welch*. Pour effectuer ce test, on compare la statistique de test ci-dessus non pas à une distribution normale standardisée (comme on le ferait pour un test de Wald), mais à une distribution de Student avec *dl* degrés de liberté, où *dl* est donné par :

$$dl = \frac{\left(\frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0}\right)^2}{\frac{\tilde{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\tilde{\sigma}_0^4}{n_0^2(n_0-1)}}.$$

Les règles de rejet de H_0 et les calculs de valeur p sont analogues à ce que l'on a vu dans la section précédente. Un test de Welch n'est pas un test exact. Il s'agit cependant d'un test valide si les observations sont normales ou si n_1 et n_0 sont assez grands (mêmes conditions de validité que pour l'intervalle de confiance dual ; en particulier, on n'a pas besoin d'une hypothèse sur l'égalité des variances σ_1^2 et σ_0^2 des deux variables).

Exemple 9.4 *On reprend l'exemple de l'étude où il s'agissait de montrer que le taux de créatine est plus élevé pour les femmes atteintes d'une maladie génétique que pour les femmes non atteintes de cette maladie. On aimerait ici rejeter l'hypothèse nulle :*

$$H_0 : \Delta = 0$$

où Δ dénote la différence de moyenne du taux de log-créatine mesuré sur ces deux populations de femmes. Rappelons que l'on effectuait l'analyse sur l'échelle logarithmique afin de nous rapprocher de la normalité. À partir des échantillons de taille $n_1 = 31$ et $n_0 = 39$, on avait $\widehat{\mu}_1 = 4.71$, $\widehat{\mu}_0 = 3.66$, $\tilde{\sigma}_1^2 = 0.83$ et $\tilde{\sigma}_0^2 = 0.20$, ainsi que $\widehat{\Delta} = 4.71 - 3.66 = 1.05$. On peut donc calculer :

$$t_{stat} = \frac{1.05}{\sqrt{\frac{0.83}{31} + \frac{0.20}{39}}} = 5.88.$$

On compare t_{stat} avec les quantiles d'une distribution de Student avec :

$$dl = \frac{\left(\frac{0.83}{31} + \frac{0.20}{39}\right)^2}{\frac{0.83^2}{31^2(31-1)} + \frac{0.20^2}{39^2(39-1)}} \approx 41$$

degrés de liberté. Dans un test bilatéral, on compare donc $|t_{stat}| = 5.88$ avec $t_{0.975,41} = 2.02$, de sorte que l'on rejette l'hypothèse nulle au seuil 5 % (un logiciel statistique nous donnera $p < 0.0001$). Notre hypothèse scientifique est ainsi statistiquement démontrée.

9.6 Test de Student pour une différence de moyenne

Lorsque l'on considère l'hypothèse nulle :

$$H_0 : \Delta = \Delta^*$$

et que l'on croit à ce que l'on avait appelé le « modèle idéal », c'est-à-dire à la normalité et à une même variance dans les deux groupes ($\sigma_1^2 = \sigma_0^2$, que l'on notera simplement par σ^2), on a à disposition un test exact, appelé test de Student (en anglais : `two-sample` t-`test`), qui est en dualité avec un intervalle de confiance de Student pour Δ. Pour ce faire, on estime la variance commune par :

$$\tilde{\sigma}^2 = \frac{(n_1 - 1)\tilde{\sigma}_1^2 + (n_0 - 1)\tilde{\sigma}_0^2}{n_1 + n_0 - 2}$$

et on utilise la statistique de test donnée par :

$$T_{stat} = \frac{\widehat{\Delta} - \Delta^*}{\sqrt{\frac{\tilde{\sigma}^2}{n_1} + \frac{\tilde{\sigma}^2}{n_0}}}$$

que l'on peut ré-écrire comme suit :

$$T_{stat} = \sqrt{\frac{n_1 n_0}{n_1 + n_0}} \cdot \frac{\widehat{\Delta} - \Delta^*}{\tilde{\sigma}}.$$

On compare ensuite cette statistique de test à une distribution de Student avec $n_1 + n_0 - 2$ *dl*, les règles de rejet de H_0 et les calculs de valeur p étant analogues à ce que l'on a vu dans les deux sections précédentes. Il s'agit donc d'un test exact sous les hypothèses du modèle idéal. Si le modèle idéal ne peut pas être postulé, ce test sera valide pour autant que n_1 et n_0 soient assez grands et proches l'un de l'autre (ce sont les mêmes conditions de validité que pour l'intervalle de confiance dual).

Exemple 9.5 *En reprenant l'exemple de la section précédente, on calcule ici :*

$$\tilde{\sigma}^2 = \frac{(31 - 1)0.83 + (39 - 1)0.20}{31 + 39 - 2} = 0.48$$

et ainsi :

$$t_{stat} = \sqrt{\frac{31 \cdot 39}{31 + 39}} \cdot \frac{1.05}{\sqrt{0.48}} = 6.30.$$

Dans un test bilatéral, on compare donc $|t_{stat}| = 6.30$ *avec* $t_{0.975,68} = 2.00$*, de sorte que l'on rejette l'hypothèse nulle au seuil* 5 % *(un logiciel statistique nous donnera* $p < 0.0001$*). On a ainsi un résultat semblable à celui obtenu avec un test de Welch (on avait* $t_{stat} = 5.88$ *et* 41 *dl, alors que l'on a ici* $t_{stat} = 6.30$ *et* 68 *dl, avec dans les deux cas* $p < 0.0001$*).*

9.7 Test de Student pour données pairées

Dans les deux sections précédentes, on a implicitement supposé que les données de nos deux échantillons, à partir desquels on estimait les moyennes μ_1 et μ_0 des variables Y_1 et Y_0 que l'on voulait comparer, étaient indépendantes. En particulier, les individus mesurés dans le premier échantillon n'étaient pas les mêmes que les individus mesurés dans le second échantillon. Cela sera forcément le cas si les deux variables sont définies sur des populations différentes, telles une population de femmes avec une maladie génétique et une population de femmes sans cette maladie. Lorsque les deux variables sont définies sur la même population, il sera parfois possible de mesurer ces deux variables sur les mêmes individus. On aura alors deux échantillons de données mais un seul échantillon d'individus, de sorte que les données ne seront pas indépendantes et les tests vus dans les deux sections précédentes ne seront pas valides.

Afin d'illustrer notre propos, nous considérons un exemple où l'on aimerait démontrer l'hypothèse scientifique qui affirme qu'un certain médicament permet d'abaisser le rythme cardiaque chez les personnes diabétiques. Les deux variables d'intérêt sont ici les suivantes :

- Y_1 : rythme cardiaque sans le médicament
- Y_0 : rythme cardiaque avec le médicament.

Ces deux variables sont définies sur la même population (les personnes diabétiques). On a alors deux stratégies différentes pour récolter des données :

- on peut considérer deux échantillons différents de personnes diabétiques, ne pas donner le médicament dans l'un et le donner dans l'autre, mesurer le rythme cardiaque de chaque individu, puis comparer les rythmes cardiaques entre les deux échantillons
- on peut considérer un seul échantillon de personnes diabétiques, mesurer le rythme cardiaque deux fois par individu, une fois sans avoir donné le médicament et une fois après avoir l'avoir donné, puis comparer les deux mesures.

La première stratégie correspond à ce que l'on a vu jusqu'ici. C'était le cas dans l'exemple des *Onobrychis* où l'on ne pouvait pas faire autrement (on ne pouvait pas cultiver un même *Onobrychis* une fois avec un niveau nutritif faible et une fois avec un niveau nutritif élevé), de sorte que les observations étaient considérées indépendantes. Par contre, lorsque l'on adopte la seconde stratégie, on ne pourra pas considérer que les observations sont indépendantes car on aura deux mesures par individu. On aura dans ce cas des *données pairées* et il s'agira d'utiliser un test statistique approprié, tel le *test de Student pour données pairées* (en anglais : `paired` t`-test`) que nous introduisons ci-dessous.

Nous dénotons comme d'habitude par $\Delta = \mu_1 - \mu_0$ la différence des moyennes de Y_1 et Y_0. On aimerait donc rejeter l'hypothèse nulle suivante :

$$H_0 : \Delta = 0$$

à partir de données pairées. Notons qu'avec des données pairées, on observe non seulement les variables Y_1 et Y_0, mais également la variable $Y = Y_1 - Y_0$ et ceci pour chaque individu. Or, la moyenne μ de Y est égale à la différence Δ entre les moyennes de Y_1 et de Y_0. Ainsi, rejeter l'hypothèse nulle $H_0 : \Delta = 0$ revient à rejeter l'hypothèse nulle $H_0 : \mu = 0$, ce que l'on peut essayer de faire avec un test de Student pour une moyenne.

Un test de Student pour données pairées des variables Y_1 et Y_0 est un test de Student sur la moyenne de la variable $Y = Y_1 - Y_0$.

Il s'agit d'un test exact si la variable $Y = Y_1 - Y_0$ est normale. Si cette variable n'est pas normale, ce test sera valide pour autant que le nombre d'individus n dans l'échantillon soit assez grand. Nous illustrons et comparons ces deux stratégies ci-dessous avec notre exemple.

Exemple 9.6 *Si on adopte la première stratégie d'échantillonnage, on aura n_1 individus dans le premier échantillon (sans médicament) et n_0 individus dans le second échantillon (avec médicament). On mesurera sur chaque individu la variable d'intérêt, ici le rythme cardiaque par minute. Nous présentons ci-dessous de telles données avec $n_1 = n_0 = 8$. Une manière classique d'organiser les données dans un tableur est de consacrer une ligne par individu et une colonne par variable. Nous aurons ici* 16 *lignes et* 2 *colonnes, la première pour désigner le groupe (*1 = *sans médicament ;* 0 = *avec médicament), la seconde pour désigner la variable d'intérêt (le rythme cardiaque). On aura par exemple :*

groupe	*rythme cardiaque*
1	74
1	68
1	84
1	53
1	75
1	87
1	69
1	71
0	66
0	67
0	62
0	47
0	56
0	60
0	63
0	72

On a ici $\widehat{\mu}_1 = 72.6$, $\widehat{\mu}_0 = 61.6$, $\widehat{\Delta} = 11.0$ *et* $\tilde{\sigma} = 9.1$. *Si on effectue un test de Student bilatéral pour une différence de moyenne (avec comme hypothèse nulle* $H_0 : \Delta = 0$*), on calcule* $t_{stat} = 2.40$ *et* $p = 0.031$. *On rejette donc l'hypothèse nulle au seuil de* 5 %. *Comme on a par ailleurs* $\widehat{\Delta} > 0$, *on conclut* $\Delta > 0$ *(le médicament permet d'abaisser le rythme cardiaque).*

Si la seconde stratégie d'échantillonnage a été adoptée, nos deux groupes seront constitués des mêmes $n = 8$ *individus. Il s'agira alors d'organiser les données en* 8 *lignes et* 2 *colonnes, la première pour désigner* Y_1 *(rythme cardiaque sans médicament), la seconde pour désigner* Y_0 *(rythme cardiaque avec médicament). La variable d'intérêt sera ici* $Y = Y_1 - Y_0$ *que l'on représentera dans une troisième colonne. En considérant les mêmes données que ci-dessus, on a ainsi :*

Y_1	Y_0	$Y = Y_1 - Y_0$
74	66	8
68	67	1
84	62	22
53	47	6
75	56	19
87	60	27
69	63	6
71	72	−1

Pour la troisième colonne on a $\widehat{\mu} = 11.0$ *et* $\tilde{\sigma} = 10.3$. *Si on effectue un test de Student bilatéral pour une moyenne en utilisant les données de la troisième colonne (avec comme hypothèse nulle* $H_0 : \mu = 0$*), on calcule* $t_{stat} = 3.02$ *et* $p = 0.019$. *On rejette donc l'hypothèse nulle au seuil de* 5 %. *Comme on a par ailleurs* $\widehat{\mu} > 0$, *on conclut* $\mu > 0$ *(et donc* $\Delta > 0$*; le médicament permet d'abaisser le rythme cardiaque)*[5].

[5]Notons que dans ces exemples, on obtient une statistique de test plus élevée (et une valeur p plus petite) si on traite ces données comme étant pairées plutôt qu'indépendantes. Cela sera souvent le cas en pratique et on aura en général avantage, dans le but de rejeter une hypothèse nulle d'égalité de deux moyennes, à utiliser des données pairées plutôt que des données indépendantes (et donc à adopter la seconde stratégie pour récolter les données plutôt que la première lorsque cela sera techniquement possible). On aura en effet dans le cas de données pairées :

$$T_{stat} = \sqrt{n} \cdot \widehat{\Delta}/\tilde{\sigma}$$

alors que l'on aura pour des données indépendantes (avec $n_1 = n_0 = n$, ce qui implique $n_1 n_0/(n_1 + n_0) = n/2$) :

$$T_{stat} = \sqrt{n/2} \cdot \widehat{\Delta}/\tilde{\sigma}.$$

Attention, les $\tilde{\sigma}$ impliqués dans ces deux statistiques de tests ne sont pas les mêmes. Le premier est un estimateur de l'écart type commun de Y_1 et Y_0. Le second est un estimateur de l'écart type de la variable $Y = Y_1 - Y_0$. Sous les hypothèses du modèle idéal, le premier sera égal à $\sqrt{2(1-\rho)}$ fois le second, où ρ dénote le coefficient de corrélation entre Y_1 et Y_0 (un nombre entre -1 et $+1$, concept que nous développerons au chapitre 12). La première statistique de test sera donc supérieure à la seconde si cette corrélation est positive.

Chapitre 10

Calcul de taille d'échantillon

On a vu que la significativité d'un test statistique se calcule à partir d'une statistique de test, cette dernière étant une mesure de la distance entre les données et l'hypothèse nulle. Il est toutefois important de préciser qu'il ne s'agit pas d'une distance exprimée dans les unités originales des données, mais d'une distance qui tient compte également de l'incertitude des estimations, notamment de la taille de l'échantillon. Plus l'échantillon est grand, plus les estimateurs sont précis et en ce sens, plus la distance entre les données et l'hypothèse nulle est grande (car plus certaine). Ceci implique le message suivant :

> Un résultat statistiquement significatif n'implique pas forcément un résultat cliniquement important.

Par exemple, une différence significative entre les effets de deux médicaments n'implique pas forcément qu'un médicament est nettement supérieur à l'autre ; cela veut seulement dire que l'on est statistiquement sûr qu'un médicament est supérieur à l'autre (sans dire de combien). Inversement, une différence non significative entre les effets de deux médicaments n'implique pas forcément qu'il n'y a pas de différence entre les deux médicaments ; cela veut seulement dire que l'on n'est pas sûr qu'il existe réellement une différence (on ne peut pas exclure la possibilité que la différence observée entre les deux médicaments soit due uniquement au hasard). Encore une fois :

- le terme *significatif* en statistique ne veut pas dire « grand » ou « important » ; cela veut dire « non dû au hasard », ou plus précisément, « que le hasard ne pourrait que rarement produire »
- le terme *non significatif* en statistique ne veut pas dire « petit » ou « non important » ; cela veut dire « qui pourrait être dû au hasard ».

10.1 Valeur p *versus* taille de l'échantillon

Nous allons illustrer l'influence de la taille de l'échantillon sur le résultat d'un test statistique avec le test de Student pour une différence de moyenne ($H_0 : \Delta = 0$). Si les tailles des deux échantillons sont identiques, $n_1 = n_0 = n$, on calcule :

$$t_{stat} = \sqrt{n/2} \cdot \widehat{\Delta}/\tilde{\sigma}.$$

Cette statistique de test dépend explicitement de trois facteurs :

- **la différence de moyenne observée** $\widehat{\Delta}$: la statistique de test est d'autant plus élevée (en valeur absolue) que cette différence est grande
- **l'estimation de l'écart type commun** $\tilde{\sigma}$: la statistique de test est d'autant plus élevée (en valeur absolue) que cet écart type est petit
- **la racine carrée de la taille de l'échantillon** $\sqrt{n}$: la statistique de test est d'autant plus élevée (en valeur absolue) que l'échantillon est grand.

Rappelons que la valeur p (et donc la significativité du résultat) dépend de la statistique de test calculée : la valeur p est d'autant plus petite que la statistique de test t_{stat} calculée est élevée.

Exemple 10.1 *Dans l'exemple des Onobrychis, on avait* $n_1 = n_0 = n = 60$, $\widehat{\Delta} = 9.1$ *cm et* $\tilde{\sigma} = \sqrt{18.1} = 4.3$ *cm. On calcule donc :*

$$t_{stat} = \sqrt{60/2} \cdot 9.1/\sqrt{18.1} = 11.7$$

et on obtient un résultat significatif : $p < 0.0001$. *Regardons à présent ce que vaudrait cette valeur p (dans un test bilatéral) si on avait observé ces mêmes quantités* $\widehat{\Delta} = 9.1$ *cm et* $\tilde{\sigma} = \sqrt{18.1} = 4.3$ *cm sur des échantillons n plus petits :*

- *avec* $n = 30$, *on aurait* $t_{stat} = 8.3$ *et* $p < 0.0001$
- *avec* $n = 20$, *on aurait* $t_{stat} = 6.8$ *et* $p < 0.0001$
- *avec* $n = 10$, *on aurait* $t_{stat} = 4.8$ *et* $p = 0.00015$
- *avec* $n = 5$, *on aurait* $t_{stat} = 3.4$ *et* $p = 0.01$
- *avec* $n = 3$, *on aurait* $t_{stat} = 2.6$ *et* $p = 0.06$.

La valeur p augmente lorsque n diminue. Avec seulement $n = 3$ *Onobrychis par groupe, le résultat n'est plus significatif (on ne pourrait dans ce cas pas exclure qu'une différence observée de* 9.1 *cm soit due au hasard de l'échantillonnage).*

Exemple 10.2 *Considérons à présent le cas où l'on observerait une toute petite différence de* $\widehat{\Delta} = 0.2$ *cm entre deux groupes d'Onobrychis (avec par ailleurs* $\tilde{\sigma} = \sqrt{18.1} = 4.3$ *cm comme ci-dessus). On aurait alors les résultats suivants selon la taille n des échantillons (dans un test bilatéral) :*

- *avec* $n = 10$*, on aurait* $t_{stat} = 0.11$ *et* $p = 0.92$
- *avec* $n = 100$*, on aurait* $t_{stat} = 0.33$ *et* $p = 0.74$
- *avec* $n = 1000$*, on aurait* $t_{stat} = 1.05$ *et* $p = 0.29$
- *avec* $n = 5000$*, on aurait* $t_{stat} = 2.35$ *et* $p = 0.02$
- *avec* $n = 10\ 000$*, on aurait* $t_{stat} = 3.32$ *et* $p = 0.001$.

La valeur p diminue lorsque n augmente. Une différence de 0.2 *cm est certes petite, mais si elle était observée sur de grands échantillons, elle serait bel et bien significative, c'est-à-dire qu'elle ne pourrait pas être mise sur le compte du hasard de l'échantillonnage. Plus exactement, elle serait significative au seuil de* 5 % *si :*

$$\sqrt{n/2} \cdot 0.2/\sqrt{18.1} \geq 1.96$$

(on notera que pour simplifier, on a remplacé $t_{0.975,n-2}$ *par* $z_{0.975}$*), c'est-à-dire si :*

$$n \geq 2\left(\frac{1.96}{0.2/\sqrt{18.1}}\right)^2 \approx 3477.$$

Plus généralement, toute différence observée (non exactement nulle) sera significative si n est assez grand. Avec un grand échantillon, il y a en effet peu de place pour le hasard et tout ou presque devient significatif. Ainsi, un résultat significatif établi à partir d'un grand échantillon pourrait selon les cas ne représenter aucun intérêt clinique ou pratique. Une valeur p seule n'est pas suffisante pour juger de l'importance clinique d'un résultat.

10.2 Puissance d'un test statistique

Reprenons l'exemple d'un test de Student pour une différence de moyenne, avec comme hypothèse nulle $H_0 : \Delta = 0$ et deux tailles d'échantillons égales $n_1 = n_0 = n$. L'hypothèse nulle est rejetée dans un test bilatéral au seuil α si :

$$\sqrt{n/2} \cdot |\widehat{\Delta}|/\tilde{\sigma} \geq t_{1-\alpha/2,2n-2}.$$

Pour simplifier un peu l'exposé, nous remplacerons d'une part les quantiles d'une distribution de Student $t_{1-\alpha/2,2n-2}$ par ceux d'une normale standardisée $z_{1-\alpha/2}$. D'autre part, afin de réduire le nombre de paramètres impliqués, on utilisera $\widehat{\delta} = \widehat{\Delta}/\tilde{\sigma}$, qui est une estimation de la différence de moyenne standardisée (voir chapitre 5). L'hypothèse nulle est ainsi rejetée dans un test bilatéral au seuil α si :

$$\sqrt{n/2} \cdot |\widehat{\delta}| \geq z_{1-\alpha/2}.$$

Le résultat du test (rejet ou non-rejet de H_0) dépend donc de n et de $\widehat{\delta}$. Au moment d'entreprendre une étude, il n'est évidemment pas possible de savoir si

l'hypothèse nulle sera rejetée ou non (car on ne connaît pas à l'avance quelle sera la valeur observée de $\widehat{\delta}$). Il est toutefois possible de calculer la probabilité que l'hypothèse nulle soit rejetée (la probabilité d'obtenir un résultat significatif) en fonction de n et d'une hypothèse sur la véritable valeur de δ. Cette probabilité est appelée la *puissance du test* (en anglais : `power`), notée $1 - \beta$[1].

> La puissance d'un test est la probabilité d'obtenir un résultat significatif.

Dans notre exemple, on calcule la puissance du test comme suit :

$$1 - \beta = \Pr\left\{\sqrt{n/2} \cdot \widehat{\delta} \leq z_{\alpha/2}\right\} + \Pr\left\{\sqrt{n/2} \cdot \widehat{\delta} \geq z_{1-\alpha/2}\right\}.$$

Le premier terme est la probabilité de rejeter H_0 et de conclure $\Delta < 0$. Le second terme est la probabilité de rejeter H_0 et de conclure $\Delta > 0$. Rappelons que la variable aléatoire $\widehat{\Delta}$ a une distribution normale avec moyenne Δ et écart type $\sqrt{2/n} \cdot \sigma$. Il s'ensuit que $\widehat{\delta}$ a une distribution approximativement normale avec moyenne δ et écart type $\sqrt{2/n}$, et donc que $\sqrt{n/2} \cdot \widehat{\delta}$ a une distribution approximativement normale avec moyenne $\sqrt{n/2} \cdot \delta$ et écart type 1. Cela implique[2] :

$$1 - \beta = \int_{-\infty}^{z_{\alpha/2}} \phi_{\sqrt{n/2}\cdot\delta,1}(t)\,\mathrm{d}t + \int_{z_{1-\alpha/2}}^{+\infty} \phi_{\sqrt{n/2}\cdot\delta,1}(t)\,\mathrm{d}t.$$

La figure 10.1 illustre ce calcul de puissance en nous montrant la densité de probabilité de la variable aléatoire $\sqrt{n/2} \cdot \widehat{\delta}$ pour différentes valeurs hypothétiques de $\sqrt{n/2} \cdot \delta$, ainsi que le résultat des deux intégrales ci-dessus avec $\alpha = 5$ % (la densité de probabilité de $\sqrt{n/2} \cdot \widehat{\delta}$ sous l'hypothèse nulle est par ailleurs représentée par un trait fin dans chacun de ces graphiques). Dans le cas où $\sqrt{n/2} \cdot \delta = 0$ (c'est-à-dire dans le cas où l'hypothèse nulle est vraie), on a 2.5 % de conclure $\Delta < 0$ et 2.5 % de conclure $\Delta > 0$. Dans le cas où $\sqrt{n/2} \cdot \delta = 1$, on n'a plus que 0.2 % de conclure $\Delta < 0$ mais on a 16.9 % de conclure $\Delta > 0$. En tout, on a donc une probabilité de $0.2 + 16.9 = 17.1$ % d'obtenir un résultat significatif (et donc une puissance de 17.1 %). La puissance du test augmente à 51.6 % dans le cas où $\sqrt{n/2} \cdot \delta = 2$ et à 85.1 % dans le cas où $\sqrt{n/2} \cdot \delta = 3$.

Afin de pouvoir utiliser le tableau A.1 pour le calcul de la puissance de ce test, il est utile de l'exprimer comme suit :

$$1 - \beta = \Phi_{\sqrt{n/2}\cdot\delta,1}(z_{\alpha/2}) + \left(1 - \Phi_{\sqrt{n/2}\cdot\delta,1}(z_{1-\alpha/2})\right)$$

[1] On rappelle que β est l'erreur de seconde espèce, c'est-à-dire la probabilité de ne pas rejeter H_0 alors que H_0 est fausse (l'erreur de première espèce α étant la probabilité de rejeter H_0 alors que H_0 est vraie).

[2] On rappelle que $\phi_{\mu,\sigma}$ et $\Phi_{\mu,\sigma}$ dénotent respectivement la densité de probabilité et la fonction de répartition d'une distribution normale avec moyenne μ et écart type σ.

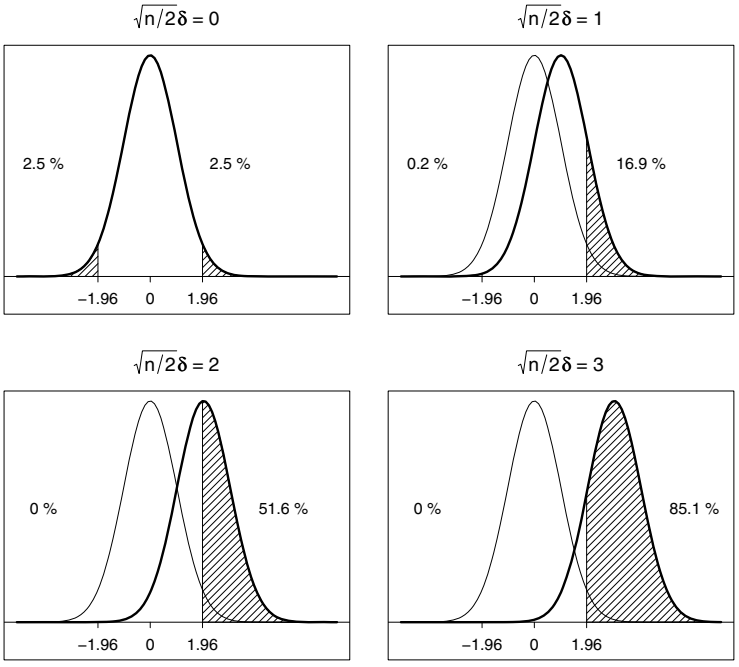

Figure 10.1 – Puissance d'un test de Student pour différentes valeurs de $\sqrt{n/2} \cdot \delta$.

ce qui revient au même que :

$$1 - \beta = \Phi_{0,1}(z_{\alpha/2} - \sqrt{n/2} \cdot \delta) + \Big(1 - \Phi_{0,1}(z_{1-\alpha/2} - \sqrt{n/2} \cdot \delta)\Big)$$

ou encore :

$$1 - \beta = \Phi_{0,1}(z_{\alpha/2} - \sqrt{n/2} \cdot \delta) + \Phi_{0,1}(z_{\alpha/2} + \sqrt{n/2} \cdot \delta).$$

Par exemple, avec $\alpha = 5$ % et $\sqrt{n/2} \cdot \delta = 3$, on obtient une puissance de :

$$1 - \beta = \Phi_{0,1}(-4.96) + \Phi_{0,1}(1.04) = 0 + 0.851 = 85.1\ \%$$

(on retrouve ce que l'on avait dans la figure 10.1). Pour $\delta = 0$, les deux termes ci-dessus sont égaux à $\alpha/2$ et on obtient $1 - \beta = \alpha$ (la puissance d'un test étant égale à son seuil si l'hypothèse nulle est vraie).

Le graphique du bas de la figure 10.2 nous montre la puissance du test en fonction de la valeur hypothétique de $\sqrt{n/2} \cdot \delta$. La droite horizontale à 5 % représente le seuil du test. Les graphiques du haut et du milieu de cette figure nous montrent par ailleurs la probabilité de conclure $\Delta < 0$, respectivement la probabilité de conclure $\Delta > 0$ (la puissance du test, représentée sur ces deux

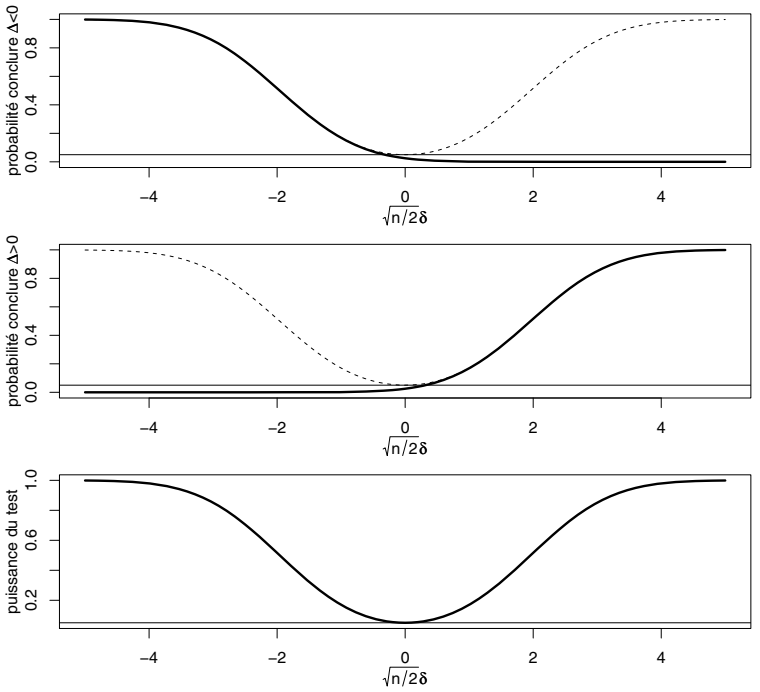

Figure 10.2 – Puissance d'un test de Student selon $\sqrt{n/2} \cdot \delta$.

graphiques par un trait pointillé, étant donc la somme de ces deux probabilités). On voit que la puissance approche 100 % pour de grandes valeurs de $\sqrt{n/2} \cdot |\delta|$.

Étant donné une valeur hypothétique de δ, il existe donc forcément une valeur de n nous permettant d'atteindre n'importe quelle puissance désirée, par exemple 80 % ou 90 %. Ce n représente le nombre d'individus qu'il s'agit d'inclure dans notre étude afin d'avoir une grande probabilité d'obtenir un résultat significatif (c'est-à-dire afin d'avoir une grande probabilité de démontrer statistiquement notre hypothèse scientifique). Au moment de la planification d'une étude, il s'agit donc de postuler une valeur pour la véritable différence de moyenne standardisée δ entre les deux groupes considérés. Si on postule une valeur négative ($\delta < 0$) et que l'on aimerait pouvoir conclure $\Delta < 0$ avec une probabilité $1 - \beta$, on choisira n en résolvant l'équation suivante :

$$\Phi_{0,1}(z_{\alpha/2} - \sqrt{n/2} \cdot \delta) = 1 - \beta.$$

De même, si on postule une valeur positive ($\delta > 0$) et que l'on aimerait pouvoir conclure $\Delta > 0$ avec une probabilité $1 - \beta$, on résoudra :

$$\Phi_{0,1}(z_{\alpha/2} + \sqrt{n/2} \cdot \delta) = 1 - \beta.$$

Dans les deux cas, $z_{\alpha/2} + \sqrt{n/2} \cdot |\delta|$ doit correspondre au quantile $1 - \beta$ d'une

distribution normale standardisée :

$$z_{\alpha/2} + \sqrt{n/2} \cdot |\delta| = z_{1-\beta}.$$

On obtient dans les deux cas :

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}.$$

Exemple 10.3 *Pour atteindre une puissance de* 80 %*, on utilise la formule ci-dessus avec* $z_{1-\beta} = z_{0.8} = 0.84$*. Pour atteindre une puissance de* 90 %*, on utilise* $z_{1-\beta} = z_{0.9} = 1.28$*. Au seuil de* 5 %*, on utilise par ailleurs* $z_{1-\alpha/2} = z_{0.975} = 1.96$ *et la formule est ainsi :*

- $n = 2(1.96 + 0.84)^2/\delta^2 \approx 16/\delta^2$ *pour atteindre une puissance de* 80 %
- $n = 2(1.96 + 1.28)^2/\delta^2 \approx 21/\delta^2$ *pour atteindre une puissance de* 90 %*.*

On calcule alors les n suivants selon la valeur postulée de δ *:*

$1-\beta$	$\lvert\delta\rvert = 1.25$	$\lvert\delta\rvert = 1$	$\lvert\delta\rvert = 0.75$	$\lvert\delta\rvert = 0.5$	$\lvert\delta\rvert = 0.25$
80 %	11	16	28	63	251
90 %	14	21	38	84	336

Rappelons toutefois que l'on a un peu simplifié le problème en remplaçant les quantiles d'une distribution de Student par ceux d'une distribution normale standardisée (les n obtenus seraient sinon légèrement plus grands)[3].

10.3 Exemples de calculs de taille d'échantillon

Le principe de calcul d'une taille d'échantillon présenté ci-dessus s'applique à tous les tests statistiques. En voici quelques exemples (dans tous les cas, on considère un test bilatéral)[4] :

- **test de Student pour une différence de moyenne**

  → $H_0 : \Delta = \Delta^*$, où $\Delta = \mu_1 - \mu_0$ représente la différence des moyennes de deux populations et Δ^* la valeur testée

[3] On notera au passage que pour atteindre une puissance de 50 %, on utilise $z_{0.5} = 0$ et la formule devient $n = 2(1.96/\delta)^2 \approx 8/\delta^2$. Il s'agit de la formule que l'on avait utilisée dans la section précédente pour déterminer à partir de quel n une différence standardisée observée de $\widehat{\delta} = 0.2/\sqrt{18.1} = 0.047$ était significative. On avait en effet calculé $n = 2 \cdot (1.96/0.047)^2 \approx 3477$. Calculer le n nécessaire pour qu'une valeur observée $\widehat{\delta}$ soit significative revient donc à calculer le n nécessaire pour atteindre une puissance de 50 %, en postulant $\delta = \widehat{\delta}$.

[4] La taille d'échantillon n nécessaire pour atteindre une puissance $1 - \beta$ dans un test unilatéral au seuil α s'obtient en remplaçant $z_{1-\alpha/2}$ par $z_{1-\alpha}$ dans les formules données dans cette section.

→ on considère le cas de deux échantillons de même taille $n_1 = n_0 = n$ (observations indépendantes)

→ la statistique de test $T_{stat} = \sqrt{n/2}(\widehat{\Delta} - \Delta^*)/\tilde{\sigma}$ a approximativement une distribution normale avec moyenne $\sqrt{n/2}(\Delta - \Delta^*)/\sigma$ et écart type 1, où σ^2 représente la variance commune aux deux populations

→ puissance du test approximative :

$$1 - \beta = \Phi_{0,1}(z_{\alpha/2} - \sqrt{n/2}(\Delta - \Delta^*)/\sigma) + \Phi_{0,1}(z_{\alpha/2} + \sqrt{n/2}(\Delta - \Delta^*)/\sigma)$$

→ taille d'échantillon n nécessaire (par groupe) pour atteindre une puissance $1 - \beta$:

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{(\Delta - \Delta^*)^2}$$

→ il s'agit de faire une hypothèse sur la véritable valeur de $(\Delta - \Delta^*)/\sigma$

→ dans le cas particulier où $\Delta^* = 0$, on retrouve le calcul de taille d'échantillon présenté dans la section précédente

→ Note : on rappelle que dans le cas $n_1 = n_0$, on obtient un résultat similaire en utilisant un test de Student ou un test de Welch ; ce calcul de taille d'échantillon s'applique donc également au test de Welch (σ^2 représentant dans ce cas la moyenne des variances σ_1^2 et σ_0^2 des deux populations, c'est-à-dire $\sigma^2 = (\sigma_1^2 + \sigma_0^2)/2$)

- **test de Student pour une moyenne**

→ $H_0 : \mu = \mu^*$, où μ représente la moyenne d'une population et μ^* la valeur testée

→ la statistique de test $T_{stat} = \sqrt{n}(\widehat{\mu} - \mu^*)/\tilde{\sigma}$ a approximativement une distribution normale avec moyenne $\sqrt{n}(\mu - \mu^*)/\sigma$ et écart type 1, où σ^2 représente la variance dans la population en question

→ puissance du test approximative :

$$1 - \beta = \Phi_{0,1}(z_{\alpha/2} - \sqrt{n}(\mu - \mu^*)/\sigma) + \Phi_{0,1}(z_{\alpha/2} + \sqrt{n}(\mu - \mu^*)/\sigma)$$

→ taille d'échantillon n nécessaire pour atteindre une puissance $1 - \beta$:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{(\mu - \mu^*)^2}$$

→ il s'agit de faire une hypothèse sur la véritable valeur de $(\mu - \mu^*)/\sigma$

- **test de Wald pour une différence de proportion**

  $\rightarrow H_0 : \Lambda = \Lambda^*$, où $\Lambda = \pi_1 - \pi_0$ représente une différence de proportion et Λ^* la valeur testée

  $\rightarrow$ on considère le cas de deux échantillons de même taille $n_1 = n_0 = n$ (observations indépendantes)

  $\rightarrow$ la statistique de test :

$$T_{stat} = \sqrt{\frac{n}{\widehat{\pi}_1(1-\widehat{\pi}_1) + \widehat{\pi}_0(1-\widehat{\pi}_0)}} \cdot (\widehat{\Lambda} - \Lambda^*)$$

  a approximativement une distribution normale avec moyenne :

$$\sqrt{\frac{n}{\pi_1(1-\pi_1) + \pi_0(1-\pi_0)}} \cdot (\Lambda - \Lambda^*)$$

  et écart type 1

  $\rightarrow$ puissance du test approximative :

$$\begin{aligned} 1-\beta &= \Phi_{0,1}\left(z_{\alpha/2} - \frac{\sqrt{n}\cdot(\Lambda-\Lambda^*)}{\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}}\right) \\ &\quad + \Phi_{0,1}\left(z_{\alpha/2} + \frac{\sqrt{n}\cdot(\Lambda-\Lambda^*)}{\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}}\right) \end{aligned}$$

  $\rightarrow$ taille d'échantillon n nécessaire (par groupe) pour atteindre une puissance $1-\beta$:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\pi_1(1-\pi_1) + \pi_0(1-\pi_0))}{(\Lambda - \Lambda^*)^2}$$

  $\rightarrow$ il s'agit de faire des hypothèses sur les véritables valeurs de π_1 et π_0

- **test du khi-deux pour deux proportions**

  $\rightarrow H_0 : \Lambda = 0$, où $\Lambda = \pi_1 - \pi_0$ représente une différence de proportion (cas particulier de l'hypothèse nulle précédente)

  $\rightarrow$ on considère le cas de deux échantillons de même taille $n_1 = n_0 = n$ (observations indépendantes)

  $\rightarrow$ la statistique de test :

$$T_{stat} = \sqrt{\frac{n}{2\widehat{\pi}(1-\widehat{\pi})}} \cdot \widehat{\Lambda}$$

a approximativement une distribution normale avec moyenne :

$$\sqrt{\frac{n}{2\pi(1-\pi)}} \cdot \Lambda$$

et écart type :

$$\sqrt{\frac{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}{2\pi(1-\pi)}}$$

où $\pi = (\pi_1 + \pi_0)/2$

→ puissance du test approximative :

$$\begin{aligned} 1-\beta &= \Phi_{0,1}\left(\frac{z_{\alpha/2}\sqrt{2\pi(1-\pi)} - \sqrt{n}\cdot\Lambda}{\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}}\right) \\ &\quad + \Phi_{0,1}\left(\frac{z_{\alpha/2}\sqrt{2\pi(1-\pi)} + \sqrt{n}\cdot\Lambda}{\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}}\right) \end{aligned}$$

→ taille d'échantillon n nécessaire (par groupe) pour atteindre une puissance $1-\beta$:

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\pi(1-\pi)} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}\right)^2}{(\pi_1-\pi_0)^2}$$

→ il s'agit de faire des hypothèses sur les véritables valeurs de π_1 et π_0

- **test du khi-deux pour une proportion**

  → $H_0 : \pi = \pi^*$, où π représente une proportion et π^* la valeur testée

  → la statistique de test :

$$T_{stat} = \sqrt{\frac{n}{\pi^*(1-\pi^*)}} \cdot (\widehat{\pi} - \pi^*)$$

a approximativement une distribution normale avec moyenne :

$$\sqrt{\frac{n}{\pi^*(1-\pi^*)}} \cdot (\pi - \pi^*)$$

et écart type :

$$\sqrt{\frac{\pi(1-\pi)}{\pi^*(1-\pi^*)}}$$

→ puissance du test approximative :

$$\begin{aligned} 1-\beta &= \Phi_{0,1}\left(\frac{z_{\alpha/2}\sqrt{\pi^*(1-\pi^*)}-\sqrt{n}(\pi-\pi^*)}{\sqrt{\pi(1-\pi)}}\right) \\ &\quad +\Phi_{0,1}\left(\frac{z_{\alpha/2}\sqrt{\pi^*(1-\pi^*)}+\sqrt{n}(\pi-\pi^*)}{\sqrt{\pi(1-\pi)}}\right) \end{aligned}$$

→ taille d'échantillon n nécessaire pour atteindre une puissance $1-\beta$:

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{\pi^*(1-\pi^*)}+z_{1-\beta}\sqrt{\pi(1-\pi)}\right)^2}{(\pi-\pi^*)^2}$$

→ il s'agit de faire une hypothèse sur la véritable valeur de π.

Exemple 10.4 *Le tableau ci-dessous contient quelques exemples de calcul de taille d'échantillon pour un test du khi-deux pour deux proportions en fonction de diverses valeurs hypothétiques de* π_1 *et* π_0*, afin d'atteindre une puissance de* 80 % *(en dessous de la diagonale) ou* 90 % *(au-dessus de la diagonale) au seuil de* 5 %*. Rappelons qu'un test du khi-deux n'est pas tout à fait valide si* n *est trop petit ou si* π_1 *et* π_0 *sont trop proches de* 0 *ou de* 1*, en principe lorsque* $n\pi$ *ou* $n(1-\pi)$ *est inférieur à 5 (avec donc* $\pi = (\pi_1+\pi_0)/2$*). Dans ces cas-là, la taille d'échantillon calculée est donnée entre parenthèses.*

	π_0								
π_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1		266	82	42	26	17	(12)	(9)	(6)
0.2	199		392	109	52	30	19	13	(9)
0.3	62	294		477	124	56	31	19	12
0.4	32	82	356		519	130	56	30	17
0.5	20	39	93	388		519	124	52	26
0.6	(14)	23	42	97	388		477	109	42
0.7	(10)	15	24	42	93	356		392	82
0.8	(7)	(10)	15	23	39	82	294		266
0.9	(5)	(7)	10	14	20	32	62	199	

Si on veut par exemple démontrer statistiquement à l'aide d'un test du khi-deux pour deux proportions que l'efficacité d'un médicament A (mesurée sur une échelle binaire : $1 =$ *succès ;* $0 =$ *échec) est supérieure à celle d'un médicament B à partir de deux échantillons de* $n_1 = n_0 = n$ *patients, et si on postule une efficacité de* 70 % *pour le médicament A et de* 50 % *pour le médicament B, il s'agit d'inclure deux groupes de* $n = 93$ *patients pour atteindre une puissance de* 80 % *et deux groupes de* $n = 124$ *patients pour atteindre une puissance de* 90 %.

Exemple 10.5 *Le tableau ci-dessous nous donne le n nécessaire pour atteindre une puissance de* 80 % *ou* 90 % *dans un test du khi-deux pour une proportion au seuil de* 5 % *avec la valeur testée particulière* $\pi^* = 0.5$*, en fonction de diverses valeurs hypothétiques de* π*.*

$1-\beta$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.45$	$\pi = 0.47$	$\pi = 0.49$
80 %	10	20	47	194	783	2178	19 620
90 %	12	25	62	259	1047	2915	26 265

Si on veut par exemple démontrer statistiquement à l'aide d'un test du khi-deux pour une proportion qu'il naît plus de garçons que de filles à partir d'un échantillon de n nouveau-nés, et que l'on postule 51 % *de garçons et donc* 49 % *de filles, il s'agit d'inclure dans notre étude* $n = 19\,620$ *nouveau-nés pour atteindre une puissance de* 80 %*, et* 26 265 *nouveau-nés pour atteindre une puissance de* 90 %*.*

Chapitre 11

Tests exacts avec statistique de test discrète

Un test statistique est dit exact si on connaît mathématiquement (et si on utilise effectivement) la distribution de la statistique de test sous H_0. Par exemple, le test de Student pour une moyenne est un test exact sous l'hypothèse de normalité, alors que le test de Student pour une différence de moyenne est un test exact sous le modèle idéal. Par contre, les tests du khi-deux et de Wald ne sont pas des tests exacts car la distribution de la statistique de test utilisée sous H_0 (une distribution du khi-deux, respectivement une distribution normale standardisée) n'est qu'une approximation de la distribution théorique. Notons que la statistique de test d'un test de Student est une variable aléatoire continue. Dans ce chapitre, nous allons introduire des tests exacts où la statistique de test sera une variable aléatoire discrète, à savoir le test binomial, le test exact de Fisher, le test de McNemar, le test du signe, le test de Mann-Whitney et le test de Wilcoxon. Alors que les tests exacts avec statistique de test continue sont également exacts au sens du contrôle de l'erreur de première espèce (seuil réel = seuil nominal), on va voir que les tests exacts avec statistique de test discrète sont en fait conservateurs (seuil réel < seuil nominal).

11.1 Test binomial pour une proportion

Le test binomial est une alternative au test du khi-deux lorsque l'on désire comparer une proportion π avec une valeur de référence π^*, c'est-à-dire lorsqu'il s'agit d'essayer de rejeter l'hypothèse nulle :

$$H_0 : \pi = \pi^*.$$

La statistique de test d'un test binomial est donnée par $T_{stat} = K$, où K est le nombre de 1 dans l'échantillon (on a donc $K = n\widehat{\pi}$, où n est la taille de l'échantillon et $\widehat{\pi}$ la proportion empirique). Il s'agit d'une variable aléatoire

discrète dont la distribution est définie par l'ensemble des probabilités $\Pr\{K = k\}$ pour les $n+1$ valeurs possibles $k = 0, 1, \cdots, n$, où k dénote la réalisation de la variable aléatoire K dans notre échantillon. Il se trouve que l'on connaît mathématiquement la distribution de cette statistique de test sous H_0 : il s'agit d'une distribution binomiale avec paramètres n et π^*. Un test binomial est donc un test exact.

Comme l'hypothèse nulle d'un test binomial fait intervenir la valeur d'un paramètre, on pourra faire la distinction entre un test unilatéral et un test bilatéral. Dans le cas d'un test unilatéral, on procédera de manière similaire à ce que l'on a vu avec une statistique de test continue :

- **test unilatéral à gauche**

  → la valeur p est définie par $p = \Pr\{T_{stat} \leq t_{stat}\}$, c'est-à-dire par $p = \Pr\{K \leq k\}$

  → on rejette H_0 au seuil α si $p \leq \alpha$, c'est-à-dire si $\Pr\{K \leq k\} \leq \alpha$, auquel cas on conclut $\pi < \pi^*$

- **test unilatéral à droite**

  → la valeur p est définie par $p = \Pr\{T_{stat} \geq t_{stat}\}$, c'est-à-dire par $p = \Pr\{K \geq k\}$

  → on rejette H_0 au seuil α si $p \leq \alpha$, c'est-à-dire si $\Pr\{K \geq k\} \leq \alpha$, auquel cas on conclut $\pi > \pi^*$.

La définition de la valeur p, et la décision de rejeter ou de ne pas rejeter l'hypothèse nulle dans le cas d'un test bilatéral, ne va par contre pas de soi car la distribution de la statistique de test sous H_0 n'est en général pas symétrique (sauf dans le cas particulier $\pi^* = 0.5$). En fait, on utilisera la même règle que celle vue avec une statistique de test continue, à savoir on doublera la plus petite des deux valeurs p obtenues avec des tests unilatéraux. Ainsi :

- **test bilatéral**

  → la valeur p est définie comme le double de la plus petite des deux valeurs p obtenues avec des tests unilatéraux à gauche et à droite

  → on rejette H_0 au seuil α si $p \leq \alpha$

  → on rejette ainsi H_0 au seuil α si $\Pr\{K \leq k\} \leq \alpha/2$ ou si $\Pr\{K \geq k\} \leq \alpha/2$

  → les cas où l'on rejette H_0 à cause de $\Pr\{K \leq k\} \leq \alpha/2$ sont des cas où $\widehat{\pi} < \pi^*$, et on conclut donc $\pi < \pi^*$

  → les cas où l'on rejette H_0 à cause de $\Pr\{K \geq k\} \leq \alpha/2$ sont des cas où $\widehat{\pi} > \pi^*$, et on conclut donc $\pi > \pi^*$.

Exemple 11.1 *Nous illustrons ici le test binomial dans le cas d'un échantillon de taille* $n = 10$*, où l'on veut rejeter l'hypothèse nulle* $H_0 : \pi = 0.3$ *(on considère donc* $\pi^* = 0.3$*). La distribution de la statistique de test sous* H_0 *est binomiale avec paramètres* 10 *et* 0.3*. Les valeurs possibles et leurs probabilités d'occurrence sont données dans les deux premières colonnes du tableau ci-dessous*[1]*. À partir de ces probabilités, on peut facilement calculer les valeurs p que l'on obtient dans un test unilatéral à gauche, un test unilatéral à droite ou un test bilatéral, qui sont données dans les trois dernières colonnes du tableau.*

$t_{stat} = k$	$\Pr\{K = k\}$	*p unilatéral gauche*	*p unilatéral droite*	*p bilatéral*
0	0.028	0.028	1.000	0.056
1	0.121	0.149	0.972	0.299
2	0.233	0.383	0.851	0.766
3	0.267	0.650	0.617	1.000
4	0.200	0.850	0.350	0.701
5	0.103	0.953	0.150	0.301
6	0.037	0.989	0.047	0.095
7	0.009	0.998	0.011	0.021
8	0.001	1.000	0.002	0.003
9	0.000	1.000	0.000	0.000
10	0.000	1.000	0.000	0.000

Dans cet exemple, on adopte ainsi les règles de rejet suivantes :

- *on rejette* H_0 *dans un test unilatéral à gauche au seuil de* 5 % *si on observe* $k = 0$ *(auquel cas on conclut* $\pi < 0.3$*)*
- *on rejette* H_0 *dans un test unilatéral à droite au seuil de* 5 % *si on observe* $k \geq 6$ *(auquel cas on conclut* $\pi > 0.3$*)*
- *on rejette* H_0 *dans un test bilatéral au seuil de* 5 % *si on observe* $k \geq 7$ *(auquel cas on conclut* $\pi > 0.3$*).*

Notons qu'avec $n = 10$*, il sera impossible de rejeter* H_0 *dans un test bilatéral au seuil de* 5 % *et de conclure* $\pi < 0.3$ *(la puissance d'un test binomial bilatéral sera donc nulle pour montrer* $\pi < 0.3$ *à partir d'un échantillon de taille* $n = 10$*).*

Exemple 11.2 *On aimerait montrer qu'il y a plus de corbeaux noirs que de corbeaux blancs, autrement dit on aimerait rejeter* $H_0 : \pi = 0.5$*, où* π *dénote la proportion de corbeaux noirs. Si on utilise un test binomial bilatéral et si on observe* 100 % *de corbeaux noirs dans notre échantillon (c'est-à-dire* $k = n$*), on aura* $p = 1/2^{n-1}$ *et donc les valeurs p suivantes (selon n) :*

n	1	2	3	4	5	6	7	8	9	10
p	1	0.5	0.25	0.125	0.06	0.03	0.016	0.008	0.004	0.002

[1] Dans R, on calcule ces probabilités en utilisant la commande `dbinom(0:10,10,0.3)`.

On ne pourra donc pas conclure statistiquement (au seuil de 5 %*) qu'il y a plus de corbeaux noirs que de corbeaux blancs si on observe* $k = 5$ *corbeaux noirs sur* $n = 5$ *(*$p = 0.06$*). On arrivera par contre à une telle conclusion si on observe* $k = 6$ *corbeaux noirs sur* $n = 6$ *(*$p = 0.03$*).*

11.2 Comparaison des tests binomial et du khi-deux

On a fait au chapitre 4 la distinction entre le niveau nominal et le niveau réel d'un intervalle de confiance. On va faire ici une distinction similaire entre le seuil nominal et le seuil réel d'un test statistique. Le seuil nominal d'un test statistique est la probabilité α désirée de commettre une erreur de première espèce (que l'on choisit en principe à $\alpha = 5$ %). Le seuil réel d'un test statistique est la probabilité réelle de commettre une erreur de première espèce.

Lorsque la statistique de test est continue et que sa distribution sous H_0 est connue, il sera possible de définir une région de rejet qui correspond exactement à une probabilité de α, de sorte que le seuil réel coïncide avec le seuil nominal. Par exemple, dans un test de Student bilatéral pour une moyenne, on rejette H_0 au seuil α si t_{stat} est plus petit que le quantile $\alpha/2$, ou si t_{stat} est plus grand que le quantile $1 - \alpha/2$ d'une distribution de Student avec $n - 1$ *dl*, de sorte que la probabilité de rejeter H_0 si H_0 est vraie vaut exactement α (si les observations sont normales). Cela ne sera par contre pas possible en général avec une statistique de test discrète, comme illustré dans le graphique du haut de la figure 11.1, où l'on voit une distribution binomiale avec $n = 10$ et $\pi = 0.3$ (la distribution de la statistique de test sous H_0 pour le test binomial de notre exemple de la section précédente). On ne pourra pas définir ici de région de rejet qui correspond exactement à une probabilité de $\alpha = 5$ % (on définira une région de rejet qui correspond à une probabilité plus petite que 5 %). Par conséquent, le seuil réel du test sera en général plus petit que le seuil nominal.

Exemple 11.3 *On reprend l'exemple où l'on veut rejeter* $H_0 : \pi = 0.3$ *à l'aide d'un test binomial et d'un échantillon de taille* $n = 10$*. Pour un seuil nominal de* 5 %*, on a la situation suivante :*

- *test unilatéral à gauche : on rejette* H_0 *si on observe* $k = 0$*, ce qui arrivera sous* H_0 *avec une probabilité de* 2.8 %
- *test unilatéral à droite : on rejette* H_0 *si on observe* $k \geq 6$*, ce qui arrivera sous* H_0 *avec une probabilité de* $0.037 + 0.009 + 0.001 + 0.000 + 0.000 = 4.6$ %
- *test bilatéral : on rejette* H_0 *si on observe* $k \geq 7$*, ce qui arrivera sous* H_0 *avec une probabilité de* $0.009 + 0.001 + 0.000 + 0.000 = 1.1$ %.

Le seuil réel du test sera donc de 2.8 %*,* 4.6 % *ou* 1.1 % *selon que l'on utilise un test unilatéral à gauche, un test unilatéral à droite ou un test bilatéral. Dans*

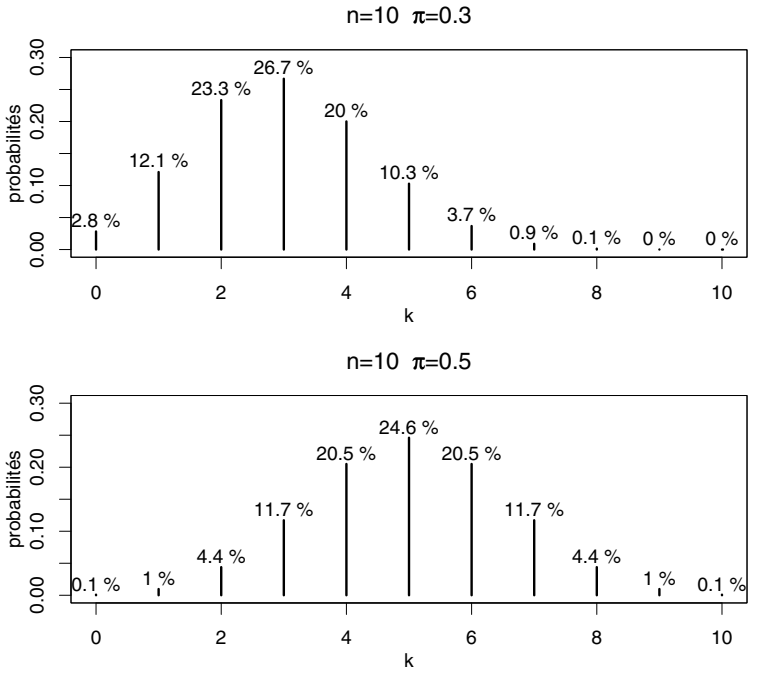

Figure 11.1 – Distribution binomiale avec $n = 10$ et $\pi = 0.3$ ou $\pi = 0.5$.

les trois cas, le seuil réel est en dessous du seuil nominal de 5 %.

On adopte la terminologie suivante :

- un test est dit *exact* (au sens du seuil) si son seuil réel correspond à son seuil nominal
- un test est dit *conservateur* si son seuil réel est plus petit que son seuil nominal
- un test est dit *libéral* si son seuil réel est plus grand que son seuil nominal.

Un test exact, au sens de la connaissance de la distribution de la statistique de test sous H_0, ne sera en général pas un test exact au sens du seuil, mais un test conservateur si la statistique de test est discrète.

Il est intéressant de comparer le seuil réel d'un test binomial avec celui d'un test du khi-deux. La figure 11.2 nous montre le seuil réel atteint par ces deux

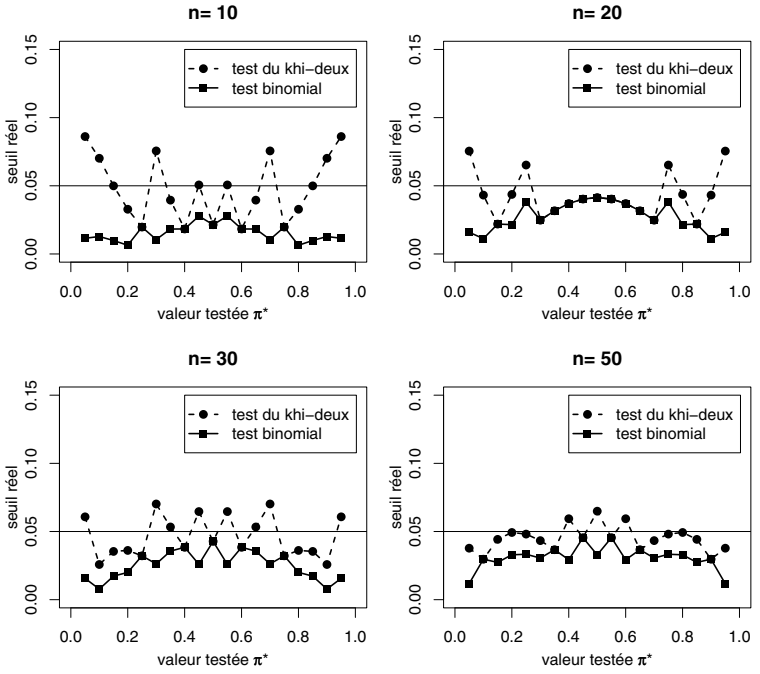

Figure 11.2 – Seuil réel des tests binomial et du khi-deux au seuil nominal de 5 %.

tests au seuil nominal de 5 % selon différentes valeurs testées π^* entre 0 et 1, et pour différentes tailles d'échantillon n. On voit que le seuil réel du test du khi-deux oscille aux alentours (et relativement proche) de 5 %, ce test étant parfois un peu libéral, parfois un peu conservateur selon les valeurs de n et π^*. Par contre, le test binomial est toujours conservateur quelles que soient les valeurs de n et π^*.

Exemple 11.4 *On reprend l'exemple où l'on veut rejeter* $H_0 : \pi = 0.3$ *avec un échantillon de taille* $n = 10$*, cette fois-ci à l'aide d'un test du khi-deux. La statistique de test est donnée par* $t_{stat} = \sqrt{10}(k/10 - 0.3)/\sqrt{0.3 \cdot 0.7}$*, et on aura ainsi (selon* k*) :*

k	0	1	2	3	4	5	6	7	8	9	10
t_{stat}	-2.07	-1.38	-0.69	0.00	0.69	1.38	2.07	2.76	3.45	4.14	4.83

On rejette H_0 *dans un test bilatéral au seuil nominal de* 5 % *si* $|t_{stat}| \geq 1.96$*. Comme on le voit dans ce tableau, ce sera le cas si on observe* $k = 0$ *ou si on observe* $k \geq 6$*, ce qui arrivera sous* H_0 *avec une probabilité de* $0.028 + 0.037 + 0.009 + 0.001 + 0.000 + 0.000 = 7.6$ % *(ces probabilités étant données dans le graphique du haut de la figure 11.1). Ainsi, le seuil réel d'un test du khi-deux*

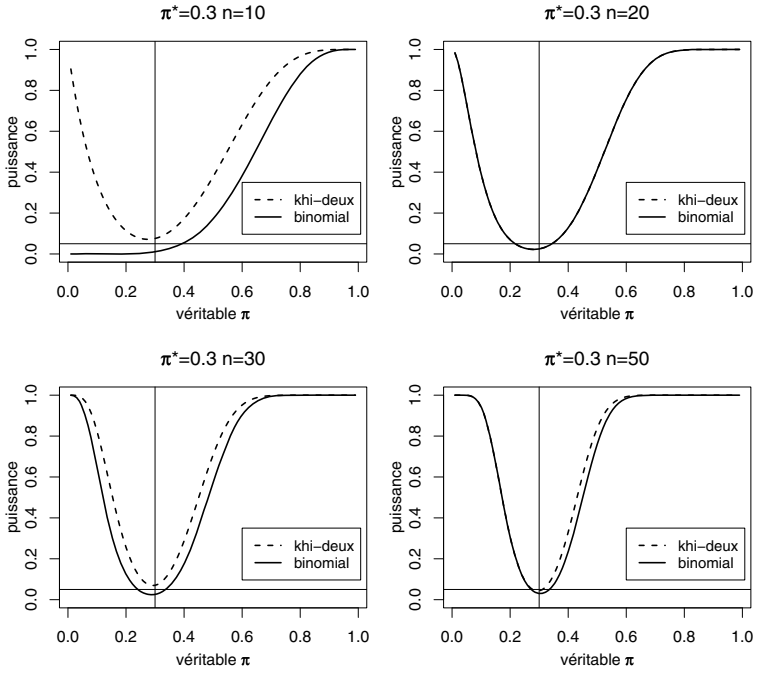

Figure 11.3 – Puissance des tests binomial et du khi-deux au seuil nominal de 5 %.

bilatéral est ici de 7.6 %*, alors que le seuil réel d'un test binomial bilatéral était de* 1.1 %*. Le test du khi-deux est donc un peu libéral, alors que le test binomial est dans le même temps conservateur.*

À cause du conflit entre erreurs de première et de seconde espèce, il est important de noter que plus un test sera conservateur moins il sera puissant. La figure 11.3 nous montre la puissance d'un test binomial en comparaison avec la puissance d'un test du khi-deux lorsqu'il s'agit d'essayer de rejeter l'hypothèse nulle $H_0 : \pi = 0.3$ dans un test bilatéral au seuil nominal de 5 %, selon la véritable valeur de π et selon la taille n de l'échantillon. Rappelons que la puissance d'un test statistique est égale à son seuil réel (représenté par une droite horizontale sur ces graphiques) si la véritable valeur de π est égale à la valeur testée (ici $\pi^* = 0.3$, représentée par une droite verticale). On voit en particulier que le test binomial n'est pas du tout puissant avec $n = 10$ si la véritable valeur de π est plus petite que la valeur testée $\pi^* = 0.3$. D'une manière générale, un test du khi-deux sera souvent plus puissant qu'un test binomial. Le prix à payer pour ce gain de puissance est que son seuil réel dépassera parfois un peu le seuil nominal.

Exemple 11.5 *Calculons la puissance d'un test binomial et la puissance d'un*

test du khi-deux lorsqu'il s'agit d'essayer de rejeter l'hypothèse nulle $H_0 : \pi = 0.3$ *dans un test bilatéral au seuil nominal de* 5 % *à partir d'un échantillon de taille* $n = 10$*, et lorsque la véritable valeur du paramètre testé est* $\pi = 0.5$*. La distribution de* K *est alors binomiale avec paramètres* $n = 10$ *et* $\pi = 0.5$*, représentée dans le graphique du bas de la figure 11.1. En utilisant les probabilités données sur ce graphique, on peut calculer les puissances suivantes :*

- *test binomial : on rejette* H_0 *(et on conclut* $\pi > 0.3$*) si on observe* $k \geq 7$*, ce qui arrivera (si* $\pi = 0.5$*) avec une probabilité de* $0.117 + 0.044 + 0.010 + 0.001 = 17.2$ %
- *test du khi-deux : on rejette* H_0 *(et on conclut* $\pi > 0.3$*) si on observe* $k \geq 6$*, ce qui arrivera (si* $\pi = 0.5$*) avec une probabilité de* $0.205 + 0.117 + 0.044 + 0.010 + 0.001 = 37.7$ %*.*

Ainsi, la puissance du test binomial est seulement de 17.2 %*, alors que celle du test du khi-deux est de* 37.7 %*, ce qui représente une différence non négligeable. La figure 11.3 a été produite en refaisant de tels calculs pour différentes valeurs de* π *et différentes tailles d'échantillon* n*.*

Lorsqu'il s'agit de comparer une proportion avec une valeur de référence et donc d'essayer de rejeter une hypothèse nulle de la forme $H_0 : \pi = \pi^*$, on aura ainsi le choix entre le test binomial, qui est conservateur (on est sûr d'être en dessous du seuil nominal) mais peu puissant, et le test du khi-deux, qui est parfois un peu libéral (on n'est pas certain d'être en dessous du seuil nominal) mais plus puissant. Il s'agit là d'un véritable choix et on trouve parmi les statisticiens des adeptes de chacun de ces deux tests[2].

On peut également définir un intervalle de confiance pour une proportion π en dualité avec un test binomial, que l'on retrouve dans la littérature sous le nom de *intervalle de confiance de Clopper-Pearson*[3]. Au niveau $1 - \alpha$, il s'agit de l'intervalle qui contient toutes les valeurs π^* qui ne sont pas rejetées dans un test binomial bilatéral au seuil α. Cet intervalle de confiance est parfois dit exact parce qu'il est en dualité avec un test statistique exact, mais il est en fait conservateur, au sens où la probabilité que la véritable valeur de π se trouve à l'intérieur de l'intervalle sera plus grande que $1 - \alpha$.

[2]Nous recommandons à ce sujet la lecture de l'article d'Agresti et Coull (1998).

[3]Un intervalle de confiance de Clopper-Pearson au niveau $1 - \alpha$ pour π se calcule explicitement comme suit :

$$\left[\frac{k}{k + (n - k + 1)F_{1-\alpha/2,2(n-k+1),2k}} ; \frac{(k+1)F_{1-\alpha/2,2(k+1),2(n-k)}}{n - k + (k+1)F_{1-\alpha/2,2(k+1),2(n-k)}} \right]$$

où k est le nombre de 1 observé dans l'échantillon, et où $F_{1-\alpha/2,dln,dld}$ est le quantile $1-\alpha/2$ d'une distribution F avec paramètres *dln* et *dld*. Comme il est conservateur, cet intervalle sera parfois bien plus large et donc moins informatif qu'un intervalle de confiance de Wilson. Pour de petits échantillons, on aura donc le choix entre un intervalle de confiance conservateur et moins informatif (Clopper-Pearson) et un intervalle de confiance parfois un peu libéral mais plus informatif (Wilson). Pour de grands échantillons, ces deux intervalles de confiance seront par contre similaires.

11.3 Test exact de Fisher

Le test exact de Fisher est une alternative au test du khi-deux lorsque l'on désire comparer deux proportions π_1 et π_0 caractérisant deux variables binaires Y_1 et Y_0, c'est-à-dire lorsqu'il s'agit d'essayer de rejeter l'hypothèse nulle (en notant $\Lambda = \pi_1 - \pi_0$) :

$$H_0 : \Lambda = 0.$$

Considérons deux échantillons de n_1 et n_0 observations indépendantes de Y_1 et Y_0. On pourra regrouper les données dans une table de contingence :

a	b
c	d

où a et b représentent les fréquences d'observations avec respectivement $Y_1 = 1$ et $Y_1 = 0$, et c et d représentent les fréquences d'observations avec respectivement $Y_0 = 1$ et $Y_0 = 0$. On aura donc $a + b = n_1$, $c + d = n_0$, ainsi que $\widehat{\pi}_1 = a/(a+b)$ et $\widehat{\pi}_0 = c/(c+d)$, et on notera par ailleurs $a+c = m_1$, $b+d = m_0$ et $N = n_1 + n_0 = m_1 + m_0 = a + b + c + d$.

La statistique de test d'un test exact de Fisher est donnée par $t_{stat} = a$, que l'on considère comme étant la réalisation d'une variable aléatoire $T_{stat} = A$ définie sur la population des tables de contingence dont la somme des lignes et la somme des colonnes sont égales à la somme des lignes et à la somme des colonnes de la table de contingence dans notre échantillon. Autrement dit, on s'imagine ici que l'on répète l'échantillonnage, en ne gardant cependant que les échantillons pour lesquels on aura non seulement $a + b = n_1$ et $c + d = n_0$, mais également $a + c = m_1$ et $b + d = m_0$. Ainsi, A est une variable aléatoire discrète avec valeurs possibles $a = \max(0, n_1 + m_1 - N), \cdots, \min(n_1, m_1)$.

Il se trouve que l'on connaît mathématiquement la distribution de cette variable aléatoire sous l'hypothèse nulle. Il s'agit d'une *distribution hypergéométrique* avec paramètres N, n_1 et m_1 définie par les probabilités suivantes[4] :

$$\Pr\{A = a\} = \frac{\frac{m_1!}{a!(m_1-a)!} \cdot \frac{(N-m_1)!}{(n_1-a)!(N-m_1-n_1+a)!}}{\frac{N!}{n_1!(N-n_1)!}}.$$

Dans le cas d'un test unilatéral, le calcul de la valeur p et la décision de rejeter ou de ne pas rejeter l'hypothèse nulle se font comme d'habitude de la manière suivante :

- **test unilatéral à gauche**

  → la valeur p est définie par $p = \Pr\{T_{stat} \leq t_{stat}\}$, c'est-à-dire par $p = \Pr\{A \leq a\}$

[4] Il s'agit de la distribution de la fréquence de boules blanches obtenues si on tirait sans remise n_1 boules d'une urne contenant m_1 boules blanches et $N - m_1$ boules noires. Dans R, on calcule ces probabilités en utilisant la commande `dhyper(a,m1,N-m1,n1)`.

→ on rejette H_0 au seuil α si $p \leq \alpha$, c'est-à-dire si $\Pr\{A \leq a\} \leq \alpha$, auquel cas on conclut $\Lambda < 0$

- **test unilatéral à droite**

  → la valeur p est définie par $p = \Pr\{T_{stat} \geq t_{stat}\}$, c'est-à-dire par $p = \Pr\{A \geq a\}$

  → on rejette H_0 au seuil α si $p \leq \alpha$, c'est-à-dire si $\Pr\{A \geq a\} \leq \alpha$, auquel cas on conclut $\Lambda > 0$.

Le cas d'un test bilatéral est un peu plus compliqué. Au contraire d'un test binomial, la règle ne consiste pas ici à doubler la plus petite des deux valeurs p obtenues avec des tests unilatéraux. On procède de la manière suivante :

- **test bilatéral**

  → la valeur p est définie par la somme des probabilités d'obtenir des valeurs de A moins probables que (ou également probables à) la valeur observée a, à savoir :

$$p = \sum_{a' : \Pr\{A=a'\} \leq \Pr\{A=a\}} \Pr\{A = a'\}$$

  → on rejette H_0 au seuil α si $p \leq \alpha$

  → dans les cas où l'on rejette H_0 à cause d'un a petit, avec donc $\widehat{\pi}_1 < \widehat{\pi}_0$ et $\widehat{\Lambda} < 0$, on conclut $\Lambda < 0$

  → dans les cas où l'on rejette H_0 à cause d'un a grand, avec donc $\widehat{\pi}_1 > \widehat{\pi}_0$ et $\widehat{\Lambda} > 0$, on conclut $\Lambda > 0$.

Exemple 11.6 *Supposons que l'on observe une table de contingence avec $n_1 = 8$, $m_1 = 4$ et $N = 14$. Les valeurs possibles pour a sont donc $0, 1, 2, 3, 4$. Le tableau ci-dessous contient les cinq tables de contingence correspondantes, ainsi que le calcul des valeurs p pour chacune d'entre elles* [5] *:*

cas possibles	*probabilité sous H_0*	*test unilatéral gauche*	*test unilatéral droite*	*test bilatéral*
$\begin{array}{\|cc\|}\hline 0 & 8 \\ 4 & 2 \\ \hline\end{array}$	$\Pr\{a = 0\} = 0.015$	$p = 0.015$	$p = 1.000$	$p = 0.015$
$\begin{array}{\|cc\|}\hline 1 & 7 \\ 3 & 3 \\ \hline\end{array}$	$\Pr\{a = 1\} = 0.160$	$p = 0.175$	$p = 0.986$	$p = 0.245$
$\begin{array}{\|cc\|}\hline 2 & 6 \\ 2 & 4 \\ \hline\end{array}$	$\Pr\{a = 2\} = 0.420$	$p = 0.594$	$p = 0.826$	$p = 1.000$
$\begin{array}{\|cc\|}\hline 3 & 5 \\ 1 & 5 \\ \hline\end{array}$	$\Pr\{a = 3\} = 0.336$	$p = 0.930$	$p = 0.406$	$p = 0.580$
$\begin{array}{\|cc\|}\hline 4 & 4 \\ 0 & 6 \\ \hline\end{array}$	$\Pr\{a = 4\} = 0.070$	$p = 1.000$	$p = 0.070$	$p = 0.085$

[5] Dans R, on calcule ces probabilités en utilisant la commande `dhyper(0:4,4,10,8)`.

On rejette donc l'hypothèse nulle (et on conclut $\Lambda < 0$) dans le cas où l'on observe la première table de contingence ci-dessus (avec $a = 0$), pour laquelle on aura la même valeur p dans un test unilatéral à gauche ou dans un test bilatéral ($p = 0.015$). Par contre, on ne rejette pas l'hypothèse nulle, ni dans un test unilatéral ni dans un test bilatéral, dans les cas où l'on observe l'une des autres tables de contingence.

Exemple 11.7 *On veut montrer que la prise d'aspirine chez les rats protège d'une infection due à une certaine bactérie. On compare deux groupes de rats, l'un à qui l'on n'a pas donné d'aspirine, l'autre à qui l'on a donné de l'aspirine. On aimerait rejeter $H_0 : \Lambda = 0$ avec $\Lambda = \pi_1 - \pi_0$, où π_1 et π_0 dénotent les proportions d'infection, respectivement dans le groupe sans et avec aspirine. Dans le groupe sans aspirine, 11 rats sur 12 ont été infectés par la bactérie. Dans le groupe avec aspirine, seuls 6 rats sur 13 ont été infectés. On a donc $\widehat{\pi}_1 = 11/12 = 0.92$, $\widehat{\pi}_0 = 6/13 = 0.46$, $\widehat{\Lambda} = 0.92 - 0.46 = 0.46$, ainsi que la table de contingence suivante :*

	infection	*pas d'infection*
sans aspirine	11	1
avec aspirine	6	7

Notre statistique de test est $a = 11$. Afin de calculer la valeur p, il s'agit de considérer toutes les tables de contingence avec $n_1 = 12$, $m_1 = 17$ et $N = 25$ (c'est-à-dire toutes les tables avec 12 rats dans un groupe et 13 dans l'autre, et avec en tout 17 rats infectés et 8 rats non infectés). Ces tables correspondent aux valeurs possibles $a = 4, 5, \cdots, 12$. Les probabilités d'observer ces différentes valeurs possibles sous l'hypothèse nulle sont données ci-dessous[6] *:*

a	4	5	6	7	8	9	10	11	12
$\Pr\{A = a\}$	0.000	0.010	0.067	0.209	0.327	0.262	0.105	0.019	0.001

Dans un test bilatéral, on calcule la valeur p comme la somme de ces probabilités qui sont plus petites ou égales à $\Pr\{A = 11\} = 0.019$, et on obtient :

$$p = 0.019 + 0.010 + 0.001 + 0.000 = 0.030$$

de sorte que l'on rejette l'hypothèse nulle $H_0 : \Lambda = 0$ au seuil de 5 %. Comme on a par ailleurs $\widehat{\Lambda} > 0$, on conclut $\Lambda > 0$, c'est-à-dire qu'il y a plus d'infections chez les rats sans aspirine (et donc que l'aspirine protège de l'infection).

Tout comme le test binomial, le test exact de Fisher est en fait conservateur au sens du seuil. La figure 11.4 nous montre le seuil réel atteint par un test exact de Fisher selon les tailles $n_1 = n_0$ d'échantillons, et pour différentes valeurs de $\pi_1 = \pi_0$. Le seuil réel atteint par un test du khi-deux pour deux

[6] Dans R, on calcule ces probabilités en utilisant la commande `dhyper(4:12,17,8,12)`.

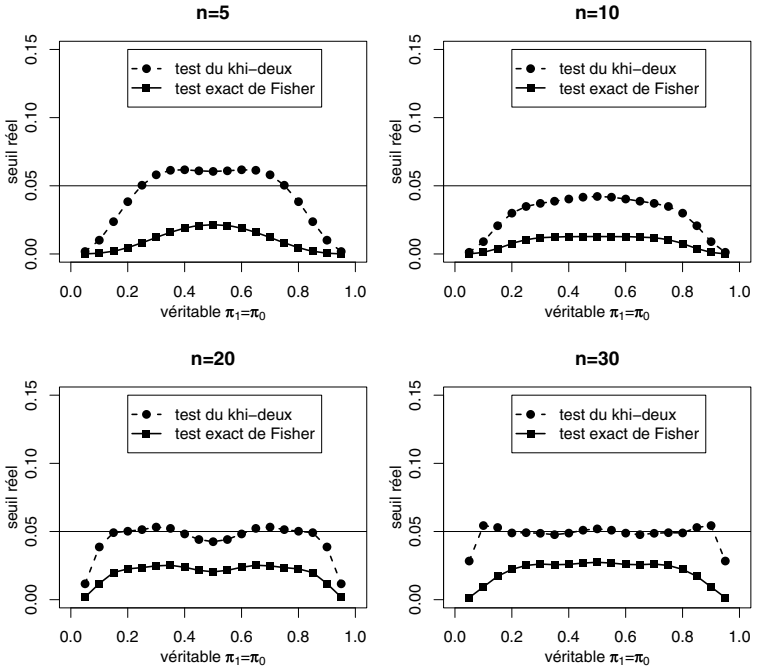

Figure 11.4 – Seuil réel des tests de Fisher et du khi-deux au seuil nominal de 5 %.

proportions est également montré sur ce graphique. De façon analogue à ce que l'on a vu dans la section précédente, le seuil réel d'un test du khi-deux oscille aux alentours (et relativement proche) de 5 %, ce test étant parfois un peu libéral (surtout pour de petits échantillons), alors que le test de Fisher est dans le même temps toujours conservateur, ce qui se traduira par une perte de puissance. La question du choix entre un test de Fisher et un test du khi-deux pour deux proportions ressemble ainsi à celle du choix entre un test binomial et un test du khi-deux pour une proportion. Le premier sera conservateur et donc moins puissant. Le second sera plus puissant mais parfois un peu libéral.

11.4 Test de McNemar

On a vu au chapitre 9 comment comparer les distributions de deux variables continues à partir de données pairées. Le test de McNemar nous permet de faire de même avec des variables binaires. Considérons deux variables binaires Y_1 et Y_0 avec proportions π_1 et π_0 et l'hypothèse nulle (où $\Lambda = \pi_1 - \pi_0$) :

$$H_0 : \Lambda = 0.$$

On a mentionné qu'avec des données pairées, on observe non seulement les variables Y_1 et Y_0 mais aussi la variable $Y = Y_1 - Y_0$, et ceci pour chaque individu. On utilise les observations de cette variable Y pour comparer les distributions de deux variables continues à partir de données pairées. Mais alors que la différence entre deux variables continues est encore une variable continue, la différence entre deux variables binaires n'est plus une variable binaire. Si les variables binaires Y_1 et Y_0 admettent les deux valeurs possibles 1 et 0, la variable Y admet les 3 valeurs possibles : $\{+1, 0, -1\}$. Afin d'obtenir quand même une variable binaire, on éliminera les individus avec $Y = 0$ et on notera par π la proportion de $+1$ parmi les $\{+1, -1\}$ (définie sur la population). Cette proportion π n'est pas égale au paramètre d'intérêt Λ, mais on aura tout de même $\Lambda = 0$ si $\pi = 0.5$, $\Lambda < 0$ si $\pi < 0.5$ et $\Lambda > 0$ si $\pi > 0.5$. Ainsi :

- rejeter l'hypothèse nulle $H_0 : \Lambda = 0$ revient à rejeter l'hypothèse nulle $H_0 : \pi = 0.5$
- cela peut se faire à l'aide d'un test binomial (ou d'un test du khi-deux pour une proportion)
- si on rejette H_0 avec par ailleurs $\widehat{\pi} < 0.5$, on conclut $\pi < 0.5$ et donc $\Lambda < 0$
- si on rejette H_0 avec par ailleurs $\widehat{\pi} > 0.5$, on conclut $\pi > 0.5$ et donc $\Lambda > 0$.

Un test de McNemar pour données pairées de variables Y_1 et Y_0 est un test binomial sur la proportion de $+1$ parmi les $\{+1, -1\}$ pour la variable $Y = Y_1 - Y_0$.

Exemple 11.8 *Une étude veut déterminer si un médicament A est plus efficace qu'un médicament B sur une échelle binaire (1 = succès ; 0 = échec). On considère les variables binaires suivantes :*

- Y_1 *: efficacité du médicament A*
- Y_0 *: efficacité du médicament B.*

On note respectivement π_1 et π_0 les proportions de succès des médicaments A et B ; on définit $\Lambda = \pi_1 - \pi_0$ et on essaie de rejeter l'hypothèse nulle $H_0 : \Lambda = 0$ à partir de n_1 observations de Y_1 et n_0 observations de Y_0. Admettons que les deux échantillons de données soient mesurés sur le même échantillon de $n_0 = n_1 = 40$ patients, de sorte que les données ne sont pas indépendantes mais pairées. On aura par exemple :

	méd. B succès	*méd. B échec*	
méd. A succès	20	5	25
méd. A échec	13	2	15
	33	7	40

Parmi les 40 *patients, la différence* $Y = Y_1 - Y_0$ *vaut* $+1$ *pour* 5 *patients (médicament A plus efficace que médicament B), vaut* -1 *pour* 13 *patients (médicament B plus efficace que médicament A) et vaut* 0 *pour* 22 *patients (pas de différence entre les deux médicaments). Parmi les* 18 *patients avec une différence non nulle, la proportion empirique de* $+1$ *parmi les* $\{+1, -1\}$ *est* $\widehat{\pi} = 5/18 = 0.28$. *Le test de McNemar consiste à tester l'hypothèse nulle* $H_0 : \pi = 0.5$ *à l'aide d'un test binomial (ou d'un test du khi-deux pour une proportion). La statistique de test d'un test binomial est donnée par* $t_{stat} = 5$. *Sous l'hypothèse nulle, la distribution de cette statistique de test est binomiale avec paramètres* $n = 18$ *et* $\pi = 0.5$, *d'où l'on peut calculer* $p = 0.096$ *(test bilatéral). Notons que l'on obtiendrait* $p = 0.061$ *avec un test du khi-deux bilatéral pour une proportion. Dans les deux cas, cela ne suffit pas pour rejeter l'hypothèse nulle* $H_0 : \pi = 0.5$ *au seuil de* 5 %, *et on ne peut donc pas conclure statistiquement à l'efficacité supérieure de l'un des deux médicaments.*

11.5 Test du signe

Lorsque l'on s'intéresse à la médiane ν plutôt qu'à la moyenne μ d'une variable continue non normale, on peut vouloir rejeter l'hypothèse nulle suivante :

$$H_0 : \nu = \nu^*$$

où ν^* est une valeur de référence (donnée par exemple dans la littérature). On peut alors utiliser un test du signe, dont le test de McNemar est un cas particulier[7]. Pour ce faire, on attribue tout d'abord à chaque observation de notre échantillon l'une des 3 valeurs possibles suivantes :

- **la valeur** $+1$ aux observations strictement supérieures à ν^*
- **la valeur** 0 aux observations égales à ν^*
- **la valeur** -1 aux observations strictement inférieures à ν^*.

On élimine ensuite les 0 (il n'y en aura pas si la variable Y est vraiment continue) et on note par π la proportion des $+1$ parmi les $\{+1, -1\}$ (définie sur la population). Ainsi :

- rejeter l'hypothèse nulle $H_0 : \nu = \nu^*$ revient à rejeter l'hypothèse nulle $H_0 : \pi = 0.5$

[7]Un intervalle de confiance au niveau $1 - \alpha$ pour la médiane ν s'obtient par ailleurs en considérant l'ensemble des valeurs ν^* non rejetées dans un test du signe bilatéral au seuil α.

- cela peut se faire à l'aide d'un test binomial (ou d'un test du khi-deux pour une proportion)
- si on rejette H_0 avec par ailleurs $\widehat{\pi} < 0.5$, on conclut $\pi < 0.5$ et donc $\nu < \nu^*$
- si on rejette H_0 avec par ailleurs $\widehat{\pi} > 0.5$, on conclut $\pi > 0.5$ et donc $\nu > \nu^*$.

Exemple 11.9 *Considérons les* 11 *données fictives suivantes d'une variable* Y *continue asymétrique (ordonnées par ordre croissant) :*

100	110	200	400	400	800	1500	3000	5000	10 000	42 195

Si notre but est de rejeter l'hypothèse nulle $H_0 : \nu = 400$*, on attribue à nos observations les valeurs suivantes :*

-1	-1	-1	0	0	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$

Notre proportion empirique de $+1$ *parmi les* 9 *valeurs non nulles est de* $\widehat{\pi} = 6/9 = 0.67$*. Le test du signe consiste à tester l'hypothèse nulle* $H_0 : \pi = 0.5$ *à l'aide d'un test binomial (ou d'un test du khi-deux pour une proportion). La statistique de test d'un test binomial est donnée par* $t_{stat} = 6$*. Sous l'hypothèse nulle, la distribution de cette statistique de test est binomiale avec paramètres* $n = 9$ *et* $\pi = 0.5$*, d'où l'on peut calculer* $p = 0.51$ *(test bilatéral). Notons que l'on obtiendrait* $p = 0.32$ *avec un test du khi-deux bilatéral pour une proportion. Dans les deux cas, cela ne suffit pas pour rejeter l'hypothèse nulle* $H_0 : \pi = 0.5$ *au seuil de* 5 %*, et on ne rejette donc pas non plus l'hypothèse nulle équivalente (et qui nous intéresse)* $H_0 : \nu = 400$*.*

11.6 Test de Mann-Whitney

Le test de Mann-Whitney est aussi appelé test de Wilcoxon (ou plus précisément en anglais : `Wilcoxon rank sum test`)[8]. On considère ici deux variables continues Y_1 et Y_0 avec fonctions de répartition $F_1(y)$ et $F_0(y)$[9] et l'hypothèse nulle suivante :

$$H_0 : \text{les distributions de } Y_1 \text{ et } Y_0 \text{ sont identiques}$$

[8] En fait, F. Wilcoxon a été le premier à publier une version de ce test en 1945 pour deux échantillons de même taille, avant l'article conjoint de H.B. Mann et D.R. Whitney publié en 1947, qui généralisèrent le test aux cas de deux échantillons de tailles différentes.

[9] Rappelons que la fonction de répartition d'une variable continue Y est définie par $F(y) = \Pr\{Y \leq y\}$.

ou autrement dit :

$$H_0 : F_1(y) = F_0(y) \quad \text{pour toute valeur possible } y.$$

À partir de deux échantillons de respectivement n_1 et n_0 observations indépendantes de Y_1 et Y_0, on calcule la statistique de test de la façon suivante :

- on établit les rangs des observations en fusionnant les deux échantillons
- on calcule la somme w des rangs des observations du premier échantillon
- on standardise cette somme de rangs comme suit :

$$\widehat{\psi} = \frac{w - \frac{n_1(n_1+1)}{2}}{n_1 n_0}$$

- on calcule la statistique de test comme suit :

$$t_{stat} = \sqrt{\frac{12 n_1 n_0}{n_1 + n_0 + 1}} \cdot (\widehat{\psi} - 0.5).$$

La distribution de T_{stat} sous H_0 est connue mathématiquement, appelée *distribution de Mann-Whitney-Wilcoxon*, et disponible dans certains logiciels statistiques. Lorsque n_1 et n_0 sont assez grands, cette distribution est cependant proche d'une distribution normale standardisée, que l'on utilisera ici comme approximation[10]. On rejettera donc l'hypothèse nulle d'égalité des distributions au seuil α si la statistique de test calculée t_{stat} est trop extrême par rapport à une distribution normale standardisée, c'est-à-dire si $|t_{stat}| \geq z_{1-\alpha/2}$.

Exemple 11.10 *On aimerait démontrer que les distributions des tailles des hommes et des femmes préhistoriques sont différentes et donc rejeter l'hypothèse nulle que ces deux distributions sont identiques. On a mesuré* $n_1 = 6$ *hommes préhistoriques et* $n_0 = 5$ *femmes préhistoriques, et on a les résultats suivants (en cm) :*

hommes préhistoriques	120	119	130	129	122	124
femmes préhistoriques	118	125	114	102	101	

On remplace ces valeurs par leurs rangs et on obtient :

[10] En utilisant la distribution exacte, le seuil réel du test sera en général inférieur au seuil nominal, de sorte que le test de Mann-Whitney sera conservateur au sens du seuil. Pour $n_1 = n_0 = 3$, la puissance d'un test de Mann-Whitney bilatéral au seuil de 5 % sera même nulle (il ne servira donc à rien d'appliquer un test de Mann-Whitney dans pareil cas). Par ailleurs, des simulations montrent que l'approximation normale de la distribution de la statistique de test sous l'hypothèse nulle est en général valide à partir d'échantillons aussi petits que $n_1 = n_0 = 4$.

hommes préhistoriques	6	5	11	10	7	8
femmes préhistoriques	4	9	3	2	1	

À partir de la somme des rangs des hommes préhistoriques $w = 47$*, on calcule* $\widehat{\psi} = 0.87$ *et* $t_{stat} = 2.01$*. Comme on a* $|t_{stat}| \geq 1.96$*, on rejette l'hypothèse nulle au seuil de* 5 % *et on conclut que les deux distributions sont différentes.*

Le rejet d'une telle hypothèse nulle n'est pourtant pas d'une grande utilité. Formellement, cela nous permet seulement de conclure statistiquement que les deux distributions sont différentes, sans nous informer si ces deux distributions diffèrent par rapport à leur position, leur dispersion ou leur forme[11]. Afin de rendre le rejet de l'hypothèse nulle d'un test de Mann-Whitney un peu plus informatif, il s'agira de faire une hypothèse supplémentaire, qui sera celle d'un *modèle d'ordonnance stochastique*. On supposera que :

$$F_1(y) \leq F_0(y) \text{ ou } F_1(y) \geq F_0(y) \quad \text{pour toute valeur possible } y$$

autrement dit que les fonctions de répartition $F_1(y)$ et $F_0(y)$ ne se croisent pas.

Dans un modèle d'ordonnance stochastique, on suppose que les fonctions de répartition ne se croisent pas.

Le modèle idéal introduit au chapitre 5 (normalité et même variance) est un cas particulier du modèle d'ordonnance stochastique, ce dernier étant donc plus général (moins restrictif). La figure 11.5 nous montre des exemples de distributions qui satisfont le modèle d'ordonnance stochastique (les fonctions de répartition des deux distributions, représentées par un trait plein et par un trait pointillé, ne se croisent pas) sans satisfaire le modèle idéal. En particulier, ce modèle inclut des cas où les variances des deux distributions sont différentes. Cependant, il n'inclut pas le cas de deux distributions normales avec une même moyenne et des variances différentes, car les fonctions de répartition se croiseraient. En fait, un modèle d'ordonnance stochastique implique que si

[11] Le test de Mann-Whitney sera surtout puissant lorsque les deux distributions diffèrent par rapport à leur position (leur moyenne), mais il sera également modérément puissant lorsque les deux distributions ont une même position (une même moyenne) et une dispersion différente, en particulier lorsque les tailles d'échantillon seront différentes et que les données du petit échantillon proviennent de la distribution avec la plus grande variabilité. Par exemple, si la distribution de Y_1 est normale avec moyenne 0 et écart type 1, et si la distribution de Y_0 est normale avec moyenne 0 et écart type 10 (les deux distributions ont donc une même position mais des dispersions différentes), la probabilité d'obtenir un résultat significatif au seuil de 5 % sera de près de 10 % si $n_1 = n_0 = 100$, et de plus de 20 % si $n_1 = 1000$ et $n_0 = 100$ (ceci peut être établi par simulation). Ainsi, si on ne fait pas d'hypothèse supplémentaire, un rejet de l'hypothèse nulle pourrait tout aussi bien être dû à une différence de position qu'à une différence de dispersion.

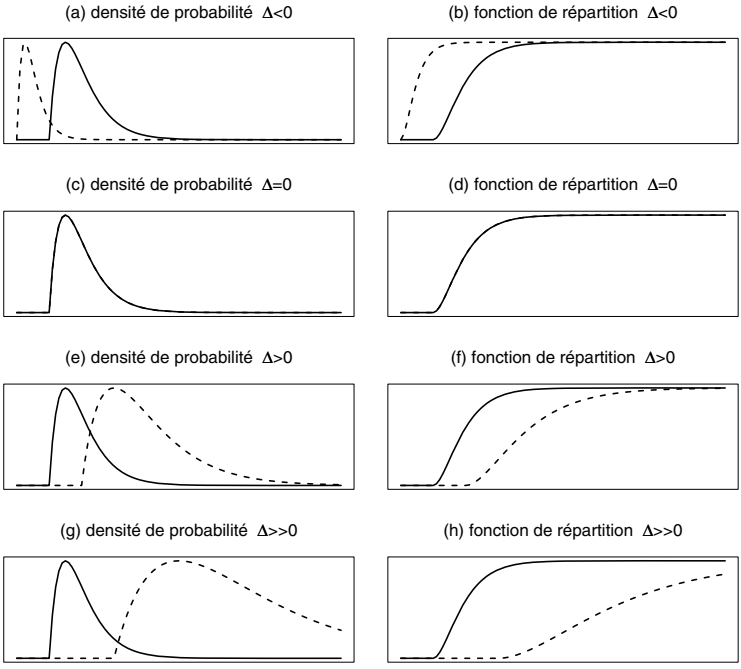

Figure 11.5 – Exemples de distributions sous le modèle d'ordonnance stochastique.

les deux distributions ont une même moyenne, alors elles sont identiques (même forme, même variabilité), ce qui correspond à l'hypothèse nulle. Par contre, si les moyennes diffèrent, alors les formes et les variances des deux distributions peuvent également différer, pour autant que les fonctions de répartition ne se croisent pas.

Sous un modèle d'ordonnance stochastique, l'hypothèse nulle donnée par :

$$H_0 : F_1(y) = F_0(y) \quad \text{pour toute valeur possible } y$$

est équivalente à l'hypothèse nulle suivante :

$$H_0 : \Delta = 0$$

où $\Delta = \mu_1 - \mu_0$ dénote la différence des moyennes μ_1 et μ_0 des deux distributions. Dans ce contexte, le test de Mann-Whitney peut être considéré comme un test sur la nullité de la différence de moyenne. Cependant, comme le test de Mann-Whitney est surtout utilisé dans des cas où les distributions ne sont pas normales (voir la section suivante), la statistique descriptive que l'on associera à ce test ne sera pas toujours la différence de moyenne Δ, mais plutôt le paramètre $\psi = \Pr\{Y_1 > Y_0\}$, la probabilité que si on tire au hasard un individu de la première population et un individu de la seconde population, le premier

ait une valeur plus grande que le second[12]. En effet, la statistique de test d'un test de Mann-Whitney se calcule à partir de $\widehat{\psi}$ qui est un estimateur de ψ. Or sous un modèle d'ordonnance stochastique, l'hypothèse nulle ci-dessus est également équivalente à l'hypothèse nulle suivante :

$$H_0 : \psi = 0.5.$$

On se retrouve dans la situation d'un test sur la valeur d'un paramètre où l'on peut procéder de la manière suivante :

- **test unilatéral à gauche**

  → on rejette H_0 au seuil α si $t_{stat} \leq z_\alpha$

  → dans le cas d'un rejet de H_0, on conclut $\psi < 0.5$ (ou de façon équivalente sous un modèle d'ordonnance stochastique, $\Delta < 0$)

- **test unilatéral à droite**

  → on rejette H_0 au seuil α si $t_{stat} \geq z_{1-\alpha}$

  → dans le cas d'un rejet de H_0, on conclut $\psi > 0.5$ (ou de façon équivalente sous un modèle d'ordonnance stochastique, $\Delta > 0$)

- **test bilatéral**

  → on rejette H_0 au seuil α si $|t_{stat}| \geq z_{1-\alpha/2}$

  → si on rejette H_0 avec par ailleurs $t_{stat} < 0$ (et donc $\widehat{\psi} < 0.5$), on conclut $\psi < 0.5$ (et donc aussi $\Delta < 0$)

  → si on rejette H_0 avec par ailleurs $t_{stat} > 0$ (et donc $\widehat{\psi} > 0.5$), on conclut $\psi > 0.5$ (et donc aussi $\Delta > 0$).

Exemple 11.11 *Dans notre exemple de la comparaison des tailles des hommes et femmes préhistoriques, on avait calculé* $\widehat{\psi} = 0.87$ *et* $t_{stat} = 2.01$*. On estime ainsi à* 87 % *la probabilité qu'un homme préhistorique (pris au hasard dans sa population) soit plus grand qu'une femme préhistorique (prise au hasard dans sa population). On avait par ailleurs un résultat significatif au seuil de* 5 % *car* $|t_{stat}| \geq 1.96$*, ce qui correspond à un test bilatéral sous un modèle d'ordonnance stochastique (on peut calculer par ailleurs* $p = 0.04$*). Sous un tel modèle, comme on a en outre* $\widehat{\psi} > 0.5$*, on conclut* $\psi > 0.5$ *(et donc* $\Delta > 0$*). On a ainsi réussi à montrer statistiquement non seulement que les distributions de la taille des hommes et des femmes préhistoriques sont différentes, mais que les hommes préhistoriques sont plus grands que les femmes préhistoriques.*

[12] Sous le modèle idéal, c'est-à-dire deux distributions normales de même variance, le lecteur intéressé pourra vérifier que ce paramètre ψ et la différence de moyenne standardisée δ sont liés par $\delta = \sqrt{2} \cdot z_\psi$ et $\psi = \Phi_{0,1}(\delta/\sqrt{2})$.

11.7 Comparaison des tests de Welch, Student et Mann-Whitney

Nous avons introduit dans ce texte trois tests statistiques, les tests de Welch, de Student et de Mann-Whitney, nous permettant de tester l'hypothèse nulle :

$$H_0 : \Delta = 0$$

où Δ dénote la différence de moyenne entre deux distributions continues. Une question qui se pose à présent est de savoir lequel des trois utiliser en pratique[13].

Dans un choix entre différents tests statistiques, le premier critère considéré est la validité du test, à savoir le contrôle de l'erreur de première espèce. Parmi les tests qui seront (approximativement) valides, un second critère sera alors la puissance du test, afin de minimiser l'erreur de seconde espèce.

> Si on a plusieurs tests statistiques à disposition dans un contexte donné, on choisira le test le plus puissant parmi les tests valides.

En ce qui concerne la validité de ces trois tests, nous avons la situation suivante (du moins lorsque les tailles d'échantillons n_1 et n_0 sont assez grandes)[14] :

test	modèle idéal	modèle d'ordonnance stochastique	cadre général
Welch	valide	valide	valide
Student	exact	valide	non valide (valide si $n_1 = n_0$)
Mann-Whitney	valide	valide	non valide

Parmi ces trois tests, le test de Welch a donc le domaine de validité le plus large et peut notamment s'appliquer dans des cas où il serait possible d'avoir deux distributions avec une même moyenne mais des variances différentes. Pour appliquer le test de Mann-Whitney, il s'agit par contre de faire une hypothèse supplémentaire, à savoir l'hypothèse d'un modèle d'ordonnance stochastique. Le test de Student est pour sa part exact sous le modèle idéal et valide sous le modèle d'ordonnance stochastique.

Sous l'hypothèse d'un modèle d'ordonnance stochastique, les trois tests sont donc valides et c'est leur puissance qui doit les départager. La figure 11.6 nous

[13] On citera à ce sujet un article assez technique de Fay et Proschan (2010) qui illustre la complexité de cette question apparemment simple.

[14] Notons que le test de Mann-Whitney serait exact (bien que conservateur au sens du seuil) sous le modèle idéal ou le modèle d'ordonnance stochastique si on utilisait la distribution exacte de la statistique de test sous l'hypothèse nulle.

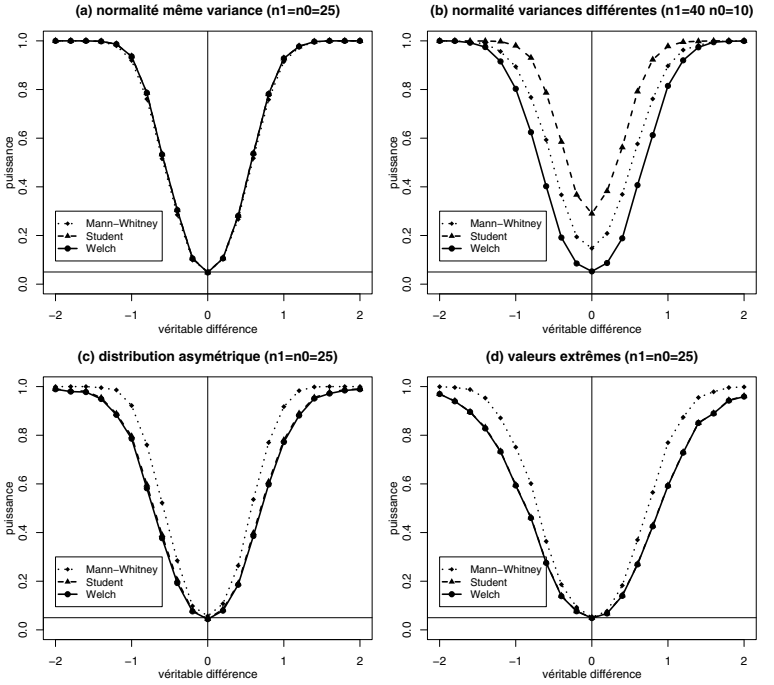

Figure 11.6 – Puissance des tests de Welch, Student et Mann-Whitney.

montre sur l'axe vertical la puissance de ces tests (appliqués de façon bilatérale à un seuil nominal de 5 %) estimée à partir de 2000 échantillons générés selon les modèles suivants : (a) le modèle idéal (distributions normales, même variance) ; (b) un modèle avec distributions normales de variances différentes ; (c) et (d) deux modèles d'ordonnance stochastique où les distributions sont respectivement asymétriques et avec des valeurs extrêmes. Les tailles d'échantillons sont ici $n_1 = n_0 = 25$ pour les situations (a), (c) et (d), et de $n_1 = 40$ et $n_0 = 10$ pour la situation (b) (où le petit échantillon provient de la distribution avec la variance la plus grande). Sur l'axe horizontal, la valeur 0 correspond à des simulations effectuées sous l'hypothèse nulle ($\Delta = 0$), les valeurs négatives correspondent à des cas avec $\Delta < 0$, les valeurs positives à des cas avec $\Delta > 0$[15]. Rappelons que la puissance d'un test sous l'hypothèse nulle correspond à son

[15]Pour être plus précis, les distributions considérées sont les suivantes : (a) normale avec moyenne 0 et écart type 1 *versus* normale avec moyenne d et écart type 1 ; (b) normale avec moyenne 0 et écart type 1 *versus* normale avec moyenne $4d$ et écart type 4 : (c) exponentielle d'une normale avec moyenne 0 et écart type 1 *versus* exponentielle d'une normale avec moyenne d et écart type 1 (il s'agit donc de distributions log-normales) ; (d) Student avec 3 *dl versus* Student avec 3 *dl* translatée de d. Ce paramètre d utilisé dans les simulations (entre -2 et $+2$) est représenté sur l'axe horizontal de la figure 11.6. Il s'agit d'une fonction monotone de la différence de moyenne Δ. Ainsi, plus la valeur de d est différente de 0, plus on s'éloigne de l'hypothèse nulle.

seuil réel. Afin d'être valide, la puissance du test doit être approximativement de 5 % lorsque $\Delta = 0$. On obtient les résultats suivants :

(a) **modèle idéal (normalité et même variance)** : théoriquement, le test de Student est le plus puissant des trois, mais pratiquement, les différences sont minimes entre les trois tests

(b) **modèle avec distributions normales de variances différentes** : le graphique confirme que les tests de Student et de Mann-Whitney sont ici non valides (seuil réel nettement plus grand que 5 %), le test de Welch étant donc le seul choix possible ici

(c) **modèle d'ordonnance stochastique et distributions asymétriques** : les trois tests sont valides, le test de Mann-Whitney est le plus puissant

(d) **modèle d'ordonnance stochastique avec valeurs extrêmes** : les trois tests sont valides, le test de Mann-Whitney est le plus puissant.

Des simulations incluant d'autres modèles et d'autres tailles d'échantillon confirmeront ces tendances générales. Nos conclusions sont ainsi les suivantes :

- si on ne veut faire aucune hypothèse, on choisira le **test de Welch** qui a l'**avantage de la validité** (seule restriction : ce test n'est pas valide dans des cas où l'on a à la fois des petits échantillons et des distributions non normales ; à vérifier par simulations en cas de doutes)

- si on peut faire l'hypothèse d'un modèle d'ordonnance stochastique, on pourra choisir le **test de Mann-Whitney** qui a l'**avantage de la puissance** (sans que l'on ait besoin de nous soucier de la forme des distributions, ni des valeurs aberrantes)

- si notre analyse risque d'être complexe (par exemple, lorsqu'elle nécessite des ajustements pour des variables confondantes), on aura intérêt à rechercher le modèle idéal (en transformant les données, en éliminant les valeurs aberrantes) et à utiliser le **test de Student** qui, outre le fait d'être exact et puissant sous le modèle idéal (y compris avec peu d'observations), a l'**avantage de la généralisation** (il s'inscrit dans le contexte plus général des modèles de régression qui vont au-delà de la simple comparaison entre deux groupes, comme on le verra au chapitre 15).

Exemple 11.12 *Nous reprenons l'exemple de la comparaison des tailles des hommes et des femmes préhistoriques. Avec un test de Mann-Whitney, on avait réussi à démontrer statistiquement que les hommes préhistoriques étaient plus grands que les femmes préhistoriques ($p = 0.04$). On obtient un résultat semblable avec un test de Student ($p = 0.03$) ou avec un test de Welch ($p = 0.06$),*

bien que ce dernier ne soit pas significatif au seuil de 5%. *Si on remplaçait par erreur le plus grand des hommes préhistoriques de notre échantillon qui mesurait* 130 *cm par la valeur aberrante de* 230 *cm, le résultat du test de Mann-Whitney n'en serait pas affecté (*$p = 0.04$*) car les rangs des observations n'en seraient pas modifiés. Par contre, les valeurs p des tests de Student et de Welch ont sensiblement augmenté (avec respectivement* $p = 0.19$ *et* $p = 0.18$*). En effet, bien que l'estimation de la différence de moyenne entre hommes et femmes préhistoriques s'en trouve augmentée (de* 12.0 *cm à* 28.7 *cm), il en est de même de l'estimation de la variabilité, ce qui atténue la statistique de test de ces tests. Cela illustre le fait que les tests de Welch et de Student risquent de manquer de puissance en présence ne serait-ce que d'une seule valeur extrême ou aberrante.*

11.8 Test de Wilcoxon

Le test de Wilcoxon (en anglais : `Wilcoxon signed rank test`) est une variante du test de Mann-Whitney que l'on peut appliquer aux données pairées de deux variables quantitatives Y_1 et Y_0, et que l'on voit brièvement ici. Formellement, l'hypothèse nulle que l'on considère est la suivante :

$$H_0 : \text{les distributions de } Y_1 \text{ et } Y_0 \text{ sont identiques.}$$

Dans le cas où l'on suppose un modèle d'ordonnance stochastique, on pourra considérer l'hypothèse nulle équivalente :

$$H_0 : \Delta = 0$$

où $\Delta = \mu_1 - \mu_0$ dénote la différence des moyennes des variables Y_1 et Y_0, et conclure dans le cas d'un résultat significatif $\Delta < 0$ ou $\Delta > 0$ selon le signe de la statistique de test. Cette dernière se calcule de la manière suivante (à partir d'observations de la variable $Y = Y_1 - Y_0$) :

- on élimine les différences nulles (les observations pour lesquelles $Y = 0$; il n'y en aura pas si les variables sont vraiment continues) et on note par n le nombre de différences non nulles
- on établit les rangs des différences non nulles selon leurs valeurs absolues
- on calcule la somme w des rangs ci-dessus correspondant aux différences positives
- on calcule la statistique de test suivante :

$$t_{stat} = \sqrt{\frac{24}{n(n+1)(2n+1)}} \cdot \left(w - \frac{n(n+1)}{4} \right).$$

La distribution exacte de T_{stat} sous H_0 est connue et proche d'une distribution normale standardisée.

Exemple 11.13 *On reprend l'exemple où l'on compare les rythmes cardiaques Y_1 et Y_0 sans et avec médicament chez les personnes diabétiques. Nous avions les données pairées suivantes :*

Y_1	Y_0	$Y = Y_1 - Y_0$	$\lvert Y\rvert$	*rang*
74	66	8	8	5
68	67	1	1	1.5
84	62	22	22	7
53	47	6	6	3.5
75	56	19	19	6
87	60	27	27	8
69	63	6	6	3.5
71	72	-1	1	1.5

Nous avons ajouté dans les deux dernières colonnes les différences en valeur absolue $|Y| = |Y_1 - Y_0|$ ainsi que les rangs correspondants (les deux plus petites valeurs absolues étant égales à 1, on leur attribue toutes deux le rang 1.5 ; les deux suivantes étant égales à 6, on leur attribue le rang 3.5). Notons qu'aucune différence n'est nulle. Parmi ces $n = 8$ différences non nulles, il y en a 7 positives et 1 négative. La somme des rangs correspondant aux 7 différences positives est de $w = 34.5$. On calcule ainsi $t_{stat} = 2.31$. En utilisant l'approximation normale de la distribution de cette statistique de test sous l'hypothèse nulle, on obtient un résultat significatif dans un test bilatéral au seuil de 5 % ($p = 0.02$). Comme $t_{stat} > 0$, on conclut $\Delta > 0$, c'est-à-dire que le nouveau médicament fait baisser le rythme cardiaque chez les personnes diabétiques (on avait une conclusion similaire avec un test de Student pour données pairées).

11.9 Récapitulatif des tests statistiques

On termine ce chapitre en donnant un tableau récapitulatif des tests statistiques sur la valeur d'un paramètre discutés jusqu'ici dans ce texte.

paramètre	H_0	test statistique approprié
une moyenne	$\mu = \mu^*$	Student
deux moyennes	$\Delta = \Delta^*$	Welch ou Student
une proportion	$\pi = \pi^*$	khi-deux ou binomial
deux proportions	$\Lambda = \Lambda^*$	Wald
une médiane	$\nu = \nu^*$	test du signe
deux moyennes	$\Delta = 0$	Welch, Student ou Mann-Whitney
deux moyennes pairées	$\Delta = 0$	Student ou Wilcoxon
deux proportions	$\Lambda = 0$	khi-deux ou Fisher
deux proportions pairées	$\Lambda = 0$	McNemar

Chapitre 12

Analyse de corrélation

Nous allons à présent nous intéresser à l'analyse de la relation entre deux variables mesurées sur les individus d'une même population. Certains cas particuliers de relations ont déjà été abordés au chapitre 5. Par exemple, lorsque nous avons comparé la croissance des *Onobrychis* cultivés à deux niveaux nutritifs différents, nous avons implicitement analysé la relation entre une variable binaire (le niveau nutritif faible ou élevé) et une variable continue (la hauteur des *Onobrychis* après une culture de six mois). De même, lorsque nous avons comparé les proportions de gauchers chez les garçons et de gauchères chez les filles, nous avons analysé la relation entre deux variables binaires. Dans ce chapitre, nous allons voir comment on peut analyser statistiquement une relation entre deux variables continues. En particulier, nous allons introduire le concept de *corrélation*. Alors que la corrélation dans le langage courant se réfère à une relation au sens large du terme, la corrélation en statistique se réfère en général à un paramètre décrivant/résumant la relation entre deux variables continues.

12.1 Diagramme de dispersion

Nous commençons ce chapitre par des aspects de statistique descriptive. La statistique descriptive, nous l'avons vu, consiste à produire des graphiques et à résumer les données *via* des caractéristiques numériques correspondant à des paramètres dans la population. La question se complique un peu si on considère deux variables continues X et Y au lieu d'une seule, autrement dit si on considère une variable bivariée continue (X, Y). Ces deux variables seront mesurées sur chaque individu de notre échantillon et nos n observations seront des *données bivariées*, que l'on notera $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Le graphique que l'on utilise pour visualiser ces données est un *diagramme de dispersion*.

La figure 12.1 nous montre un exemple de diagramme de dispersion avec sur l'axe horizontal une variable X, ici la longueur de la main, et sur l'axe vertical une variable Y, ici la taille de la personne (toutes deux exprimées en

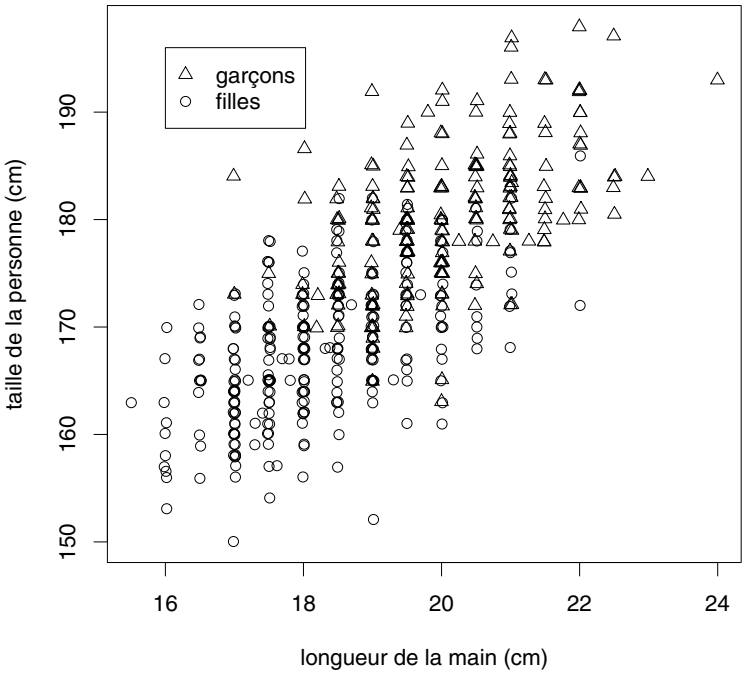

Figure 12.1 – Diagramme de dispersion pour données bivariées continues.

cm) qui ont été mesurées sur $n = 494$ étudiants en médecine (289 filles et 205 garçons)[1]. Ainsi, chaque point sur ce graphique représente un étudiant ou une étudiante (on a utilisé des symboles différents pour les garçons et pour les filles). Ce diagramme de dispersion nous montre tout d'abord que la relation entre les variables X et Y n'est pas une *relation mathématique* (ou *relation exacte*) du type $Y = g(X)$ pour une certaine fonction g. Si on avait une relation mathématique entre X et Y, la connaissance de X (et de g) nous donnerait automatiquement Y. Or, ceci n'est évidemment pas le cas sur ce diagramme. Les individus avec une main de 20 cm n'ont par exemple pas tous la même taille, certains d'entre eux mesurant 160 cm, d'autres 190 cm. On retrouve ici la variabilité propre aux sciences non exactes. Cependant, bien qu'il n'existe pas de relation mathématique entre les variables, ce diagramme de dispersion nous montre quand même une certaine tendance que l'on qualifiera de positive : les individus avec de grandes mains sont en général (en moyenne) plus grands que les individus avec de petites mains (même s'il y a des exceptions). À défaut de relation mathématique, on parlera de *relation statistique* entre les variables.

Un diagramme de dispersion permet ainsi de visualiser la relation statis-

[1]Il s'agit d'étudiants et d'étudiantes de l'Université de Zurich mesurés lors d'un cours de statistique entre 2006 et 2007.

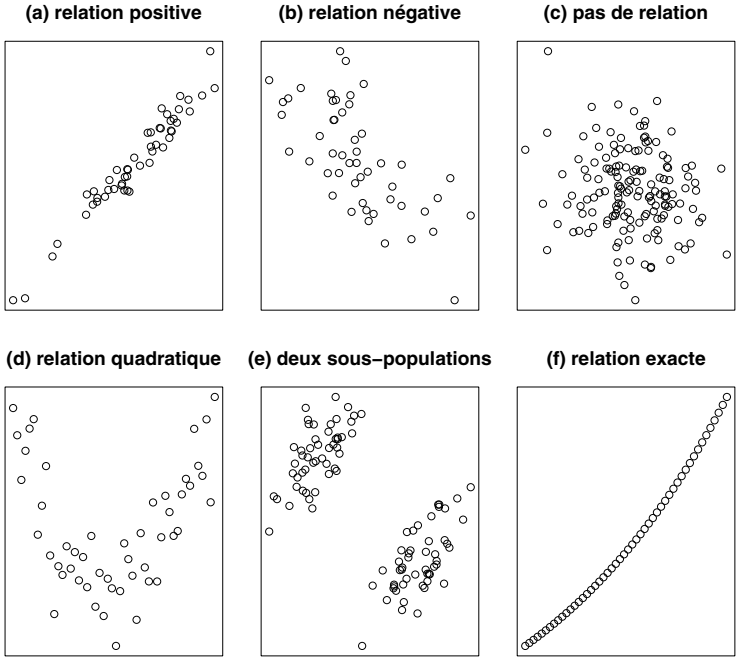

Figure 12.2 – Exemples de relations entre deux variables continues.

tique entre deux variables continues X et Y. La figure 12.2 nous en montre six exemples. Le graphique (a) nous montre un nouvel exemple de relation positive où l'on est plus proche d'une relation exacte que dans l'exemple précédent (il y a moins de variabilité autour de la tendance positive). Le graphique (b) nous montre un exemple de relation négative : si les variables X et Y représentaient comme ci-dessus la taille de la main et la taille d'une personne, les individus avec de grandes mains seraient ici en moyenne plus petits que les individus avec de petites mains. Le graphique (c) nous montre un exemple où l'on n'a pas de relation statistique entre les variables, qui sont donc indépendantes (la connaissance de la longueur de la main ne nous informerait ici en rien sur la taille de la personne). Le graphique (d) nous montre le cas d'une relation plus complexe où la relation n'est ni positive ni négative, sans pour autant que les variables soient indépendantes (les individus avec des mains moyennes étant ici plus petits que les individus avec de petites ou de grandes mains). Le graphique (e) nous montre un exemple encore plus complexe où les données proviennent d'une population visiblement hétérogène formée de deux sous-populations (ce pourrait être par exemple deux espèces animales, l'une grande en taille avec des petites pattes, l'autre petite en taille avec de grandes pattes, ce qui définit une relation globalement négative, au niveau des espèces, bien que la relation entre

longueur de patte et taille de l'animal soit positive à l'intérieur de chaque espèce). Finalement, le graphique (f) nous montre un exemple de relation exacte entre les deux variables, que l'on n'observera jamais en pratique sauf dans des cas triviaux (par exemple lorsque l'on aura deux fois la même variable mesurée dans des unités différentes).

12.2 Covariance

Lorsque l'on s'intéresse à la relation entre deux variables continues, la première étape d'une analyse de statistique descriptive consiste à produire un diagramme de dispersion nous permettant de visualiser le type de relation qui existe entre les variables. La seconde étape consiste à essayer de quantifier ce que l'on voit. Pour ce faire, nous allons introduire les paramètres de statistique descriptive que sont la *covariance* et la *corrélation*.

La covariance entre X et Y est définie de la manière suivante :

$$covariance(X, Y) = mean((X - mean(X))(Y - mean(Y))).$$

La covariance est donc la moyenne des produits des écarts aux moyennes. En notant $\bar{x} = \frac{1}{n}\sum_i x_i$ et $\bar{y} = \frac{1}{n}\sum_i y_i$, on calcule ainsi[2] :

$$\frac{1}{n}\sum_i (x_i - \bar{x})\,(y_i - \bar{y})\,.$$

Il s'agit d'une généralisation de la variance, que l'on retrouve dans le cas particulier $X = Y$:

$$covariance(X, X) = variance(X).$$

En notant que :

$$\begin{aligned}\sum_i (x_i - \bar{x})\,(y_i - \bar{y}) &= \sum_i x_i y_i - \bar{y}\sum_i x_i - \bar{x}\sum_i y_i + n\bar{x}\bar{y} \\ &= \sum_i x_i y_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y} \\ &= \sum_i x_i y_i - n\bar{x}\bar{y}\end{aligned}$$

[2]Comme pour la variance, la covariance calculée dans un échantillon est parfois calculée avec un dénominateur $n-1$ plutôt que n comme suit : $\frac{1}{n-1}\sum_i (x_i - \bar{x})\,(y_i - \bar{y})$, afin d'obtenir un estimateur sans biais de la covariance définie sur la population.

une formulation alternative et équivalente de la covariance est la suivante :

$$\begin{aligned} covariance(X,Y) &= \frac{1}{n}\sum_i (x_i - \bar{x})\,(y_i - \bar{y}) \\ &= \frac{1}{n}\left(\sum_i x_i y_i - n\bar{x}\bar{y}\right) \\ &= \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y} \\ &= mean(XY) - mean(X)\cdot mean(Y) \end{aligned}$$

Autrement dit, la covariance peut également être définie comme « la moyenne du produit moins le produit des moyennes »[3]. Afin de calculer la covariance, il suffit donc de connaître les moyennes suivantes : $\frac{1}{n}\sum_i x_i$, $\frac{1}{n}\sum_i y_i$ et $\frac{1}{n}\sum_i x_i y_i$.

Afin de pouvoir interpréter le signe de la covariance, on peut répartir les individus en quatre quadrants comme suit :

- individus dans le **quadrant 1** : $x_i > \bar{x}$ et $y_i > \bar{y}$, produit $(x_i - \bar{x})\cdot(y_i - \bar{y})$ positif et donc **contribution positive** à la covariance
- individus dans le **quadrant 2** : $x_i < \bar{x}$ et $y_i > \bar{y}$, produit $(x_i - \bar{x})\cdot(y_i - \bar{y})$ négatif et donc **contribution négative** à la covariance
- individus dans le **quadrant 3** : $x_i < \bar{x}$ et $y_i < \bar{y}$, produit $(x_i - \bar{x})\cdot(y_i - \bar{y})$ positif et donc **contribution positive** à la covariance
- individus dans le **quadrant 4** : $x_i > \bar{x}$ et $y_i < \bar{y}$, produit $(x_i - \bar{x})\cdot(y_i - \bar{y})$ négatif et donc **contribution négative** à la covariance.

On a ainsi les résultats suivants :

- $covariance(X,Y) > 0$

  → les contributions positives l'emportent sur les contributions négatives

  → on aura en principe plus d'individus dans les quadrants 1 et 3 que dans les quadrants 2 et 4

  → relation globalement positive entre les variables

- $covariance(X,Y) < 0$

  → les contributions négatives l'emportent sur les contributions positives

  → on aura en principe plus d'individus dans les quadrants 2 et 4 que dans les quadrants 1 et 3

  → relation globalement négative entre les variables

[3] On rappelle que la variance peut être définie comme « la moyenne des carrés moins le carré de la moyenne ».

- $covariance(X, Y) \approx 0$

  → les contributions positives et négatives s'annulent

  → on aura en principe autant d'individus dans les quadrants 1 et 3 que dans les quadrants 2 et 4

  → relation globalement nulle.

Exemple 12.1 *On reprend l'exemple de la relation entre la longueur de la main et la taille de* $n = 494$ *étudiants en médecine. On a ici* $\frac{1}{n}\sum_i x_i = 19.0063$ *(la moyenne des longueurs de mains est de* 19 *cm),* $\frac{1}{n}\sum_i y_i = 173.1144$ *(la moyenne des tailles des étudiants est de* 173 *cm) et* $\frac{1}{n}\sum_i x_i y_i = 3300.17$ *(si on calcule pour chaque individu le produit de la longueur de sa main et de sa taille, la moyenne de ces produits est de* 3300 cm^2*). On obtient ainsi :*

$$covariance(X, Y) = 3300.17 - 19.0063 \cdot 173.1144 = 9.9.$$

Cette covariance positive confirme quantitativement ce que le diagramme de dispersion nous suggère visuellement, à savoir une relation globalement positive entre la longueur de la main et la taille d'une personne. Par ailleurs, la figure 12.3 nous montre la répartition des étudiants dans les quatre quadrants. Il y en a effectivement nettement plus dans les quadrants 1 *et* 3 *que dans les quadrants* 2 *et* 4*, ce qui est le propre des relations globalement positives.*

La covariance satisfait de nombreuses propriétés mathématiques. En voici quelques-unes que nous laisserons le soin au lecteur de vérifier :

$$\begin{aligned} covariance(X, Y) &= covariance(Y, X) \\ covariance(X, Y + Z) &= covariance(X, Y) + covariance(X, Z) \\ covariance(X, a) &= 0 \\ covariance(X, a + bY) &= b \cdot covariance(X, Y). \end{aligned}$$

La première propriété est une propriété de symétrie. La deuxième est une propriété de distributivité. La troisième nous dit que la covariance entre une variable X et une constante a est nulle. La quatrième propriété nous informe sur la conséquence d'un changement d'unités. On retrouve ainsi un résultat déjà mentionné :

$$\begin{aligned} variance(a + bX) &= covariance(a + bX, a + bX) \\ &= b \cdot b \cdot covariance(X, X) \\ &= b^2 \cdot variance(X). \end{aligned}$$

On en déduit également le résultat suivant :

$$\begin{aligned} variance(X + Y) &= covariance(X + Y, X + Y) \\ &= covariance(X, X) + covariance(X, Y) \\ &\quad + covariance(Y, X) + covariance(Y, Y) \\ &= variance(X) + variance(Y) + 2 \cdot covariance(X, Y). \end{aligned}$$

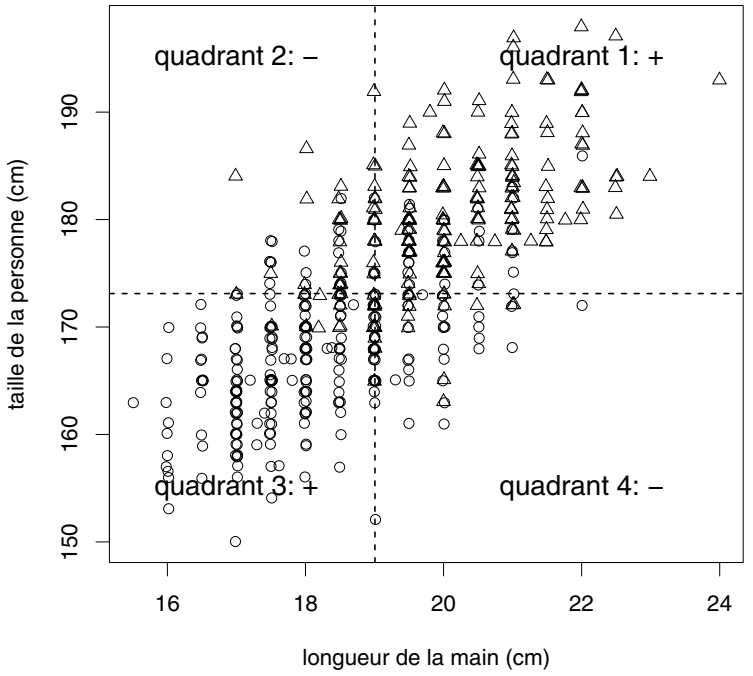

Figure 12.3 – Répartition des individus en quatre quadrants.

Dans le cas où X et Y sont indépendantes, on a $covariance(X, Y) = 0$ et on retrouve le résultat fondamental :

$$variance(X + Y) = variance(X) + variance(Y) \quad \text{si } X \text{ et } Y \text{ indépendantes.}$$

Alors que l'indépendance entre X et Y implique la nullité de leur covariance, la réciproque n'est cependant pas valable : la nullité de la covariance entre X et Y n'implique pas l'indépendance de ces variables. La nullité de la covariance implique seulement que la relation est globalement nulle, c'est-à-dire ni globalement positive, ni globalement négative. Nous reviendrons sur ce point dans la section suivante.

12.3 Corrélation de Pearson

On vient de voir comment interpréter le signe d'une covariance. L'interprétation de la valeur d'une covariance (au-delà de son signe) n'est par contre pas commode, en particulier car la valeur de la covariance est affectée par un changement d'unités (on aura un autre résultat selon que l'on mesure des individus en mètres ou en centimètres, par exemple). C'est la raison pour laquelle on a coutume de « standardiser » la covariance en la divisant par le produit des

écarts types des deux variables. On obtient ainsi *le coefficient de corrélation de Pearson*, souvent simplement appelé *la* corrélation, qui est le paramètre de statistique descriptive le plus utilisé pour décrire quantitativement la relation entre deux variables continues. On a ainsi[4] :

$$\begin{aligned} correlation(X,Y) &= \frac{covariance(X,Y)}{\sqrt{variance(X)\cdot variance(Y)}} \\ &= \frac{\frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_i (x_i-\bar{x})^2 \cdot \frac{1}{n}\sum_i (y_i-\bar{y})^2}} \\ &= \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum_i x_i^2 - \bar{x}^2\right)\cdot\left(\frac{1}{n}\sum_i y_i^2 - \bar{y}^2\right)}}. \end{aligned}$$

Afin de calculer une corrélation, il suffit donc de connaître les moyennes suivantes : $\frac{1}{n}\sum_i x_i$, $\frac{1}{n}\sum_i y_i$, $\frac{1}{n}\sum_i x_i y_i$, $\frac{1}{n}\sum_i x_i^2$ et $\frac{1}{n}\sum_i y_i^2$.

Comme la corrélation conserve le signe de la covariance, l'interprétation du signe de la corrélation se fait de la même manière que l'interprétation du signe de la covariance. Une corrélation est par contre toujours comprise entre -1 et $+1$ (alors qu'une covariance pourrait atteindre n'importe quelle valeur entre $-\infty$ et $+\infty$ selon les unités de nos variables), ce qui facilite son interprétation.

Une corrélation est toujours comprises entre -1 et $+1$.

Exemple 12.2 *On reprend l'exemple de la relation entre la longueur de la main et la taille de* $n = 494$ *étudiants en médecine. On a déjà donné* $\frac{1}{n}\sum_i x_i = 19.0063$, $\frac{1}{n}\sum_i y_i = 173.1144$ *et* $\frac{1}{n}\sum_i x_i y_i = 3300.17$. *On a en outre* $\frac{1}{n}\sum_i x_i^2 = 363.587$ *et* $\frac{1}{n}\sum_i y_i^2 = 30\ 047.21$. *On obtient ainsi :*

$$correlation(X,Y) = \frac{3300.17 - 19.0063\cdot 173.144}{\sqrt{(363.587 - 19.0063^2)(30\ 047.21 - 173.1144^2)}} = 0.73$$

On retrouve le signe positif de la covariance qui indique une relation globalement positive.

[4]Notons que l'on peut ici utiliser indifféremment un dénominateur n ou $n-1$ dans la définition de la covariance et des variances que l'on retrouve au numérateur et au dénominateur de la corrélation, car on a :

$$\frac{\frac{1}{n}\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_i (x_i-\bar{x})^2 \cdot \frac{1}{n}\sum_i (y_i-\bar{y})^2}} = \frac{\frac{1}{n-1}\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_i (x_i-\bar{x})^2 \cdot \frac{1}{n-1}\sum_i (y_i-\bar{y})^2}}.$$

La corrélation possède par ailleurs les propriétés mathématiques suivantes :

- $correlation(X, Y) = correlation(Y, X)$

  → concept symétrique (propriété héritée de la covariance)

- $correlation(X, X) = \frac{covariance(X,X)}{variance(X)} = 1$

  → valeur maximale atteinte lorsque $X = Y$ (logique : une variable ne peut pas être plus en relation avec une autre variable qu'avec elle-même)

- $correlation(X, a) = \frac{covariance(X,a)}{\sqrt{variance(X) \cdot variance(a)}} = \frac{0}{0}$

  → la corrélation entre une variable et une constante n'est pas définie

- $correlation(X, a + bY) = signe(b) \cdot correlation(X, Y)$

  → la corrélation est invariante (au signe près) lorsque l'on change d'unités

  → la corrélation sera identique que l'on prenne nos mesures en mètres ou en centimètres, par exemple

- $correlation(X, a + bX) = signe(b) \cdot correlation(X, X) = \pm 1$

  → les valeurs minimale et maximale d'une corrélation de -1 et $+1$ sont atteintes lorsque l'on a $Y = a + bX$, c'est-à-dire lorsque l'on a une relation exacte linéaire entre les variables.

On verra par ailleurs plus en détail dans le chapitre suivant que l'on peut interpréter le carré d'une corrélation comme le *pourcentage de la variance de l'une des variables que l'on peut prédire linéairement par l'autre*. En attendant, on dira qu'une corrélation mesure « le degré d'exactitude de la relation entre les variables », ou « à quel point la relation est proche d'une relation exacte linéaire », ou encore « l'intensité de la relation ».

> Le signe de la corrélation indique la direction de la relation ;
> la valeur de la corrélation est une mesure de l'intensité de la relation.

La figure 12.4 contient neuf exemples de corrélation. Les graphiques (a) et (b) nous montrent des relations exactes linéaires (respectivement positive et négative) pour laquelle la corrélation atteint ses valeurs maximales et minimales de $+1$ et -1. Pour une relation exacte positive monotone mais non linéaire, la corrélation sera positive mais plus petite que 1, comme le montre le graphique (f) où l'on voit une relation exacte logarithmique entre les variables, correspondant à une corrélation de 0.94. Rappelons toutefois qu'une relation exacte entre variables ne sera jamais observée en pratique (sauf dans des cas triviaux). Par contre, une relation pourra être plus ou moins proche d'une relation exacte

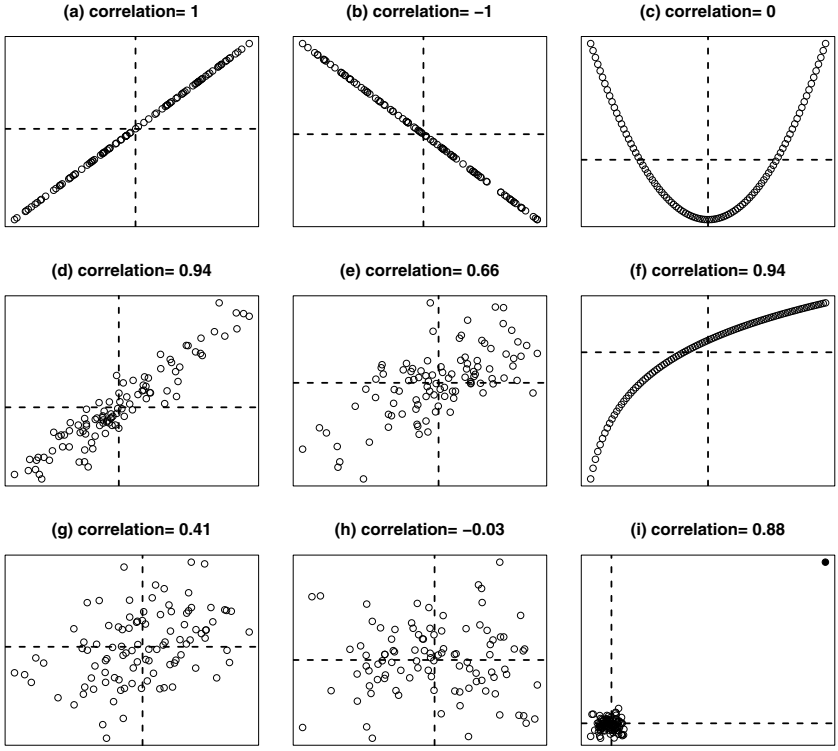

Figure 12.4 – Exemples de corrélations entre variables continues.

linéaire, et c'est précisément cette information que nous donne la corrélation. Les graphiques (d), (e) et (g) nous montrent trois exemples de relations positives avec plus ou moins d'intensité. Alors que l'on est assez proche d'une relation exacte linéaire dans le graphique (d) (avec une corrélation de 0.94), on en est plus loin dans le graphique (g) (avec une corrélation de 0.41). Le graphique (h) nous montre un cas où l'on est proche de l'indépendance entre les variables (corrélation très proche de 0). On rappelle ainsi que l'indépendance implique la nullité de la covariance et donc de la corrélation. Par contre, on a mentionné que la réciproque n'est pas valable. Une covariance/corrélation nulle n'implique pas l'indépendance entre les variables, comme nous le montre le graphique (c). On a ici une corrélation nulle (il y a en particulier autant d'observations dans les quadrants 1 et 3 que dans les quadrants 2 et 4), bien que la relation entre les variables soit exacte quadratique (les variables ne sont donc pas du tout indépendantes ; étant donné la valeur de X, on connaît ici parfaitement la valeur de Y). Finalement, le graphique (i) nous montre que la corrélation est très sensible aux valeurs extrêmes ou aberrantes. On voit sur ce graphique les mêmes données que celles représentées dans le graphique (h),

sauf une qui a été déplacée loin en haut à droite. Alors que la corrélation était presque nulle dans le graphique (h), elle est égale à 0.88 dans le graphique (i) suite au déplacement d'une seule observation. Il est donc utile de rappeler que la première étape d'une analyse de statistique descriptive bivariée consiste à produire un diagramme de dispersion, qui nous permettra de vérifier si une corrélation calculée de 0.88 représente bel et bien une relation positive, proche d'être exacte linéaire, ou si cette corrélation est due à la présence de quelques valeurs extrêmes ou aberrantes.

12.4 Corrélation *versus* causalité

Cette courte section est dédiée à un point essentiel : une corrélation élevée entre deux variables est une indication d'une *association statistique entre les variables*, mais une association statistique entre variables n'implique pas nécessairement une relation de cause à effet.

> La corrélation mesure une association entre variables, non une causalité.

Il se peut en effet qu'une association observée entre deux variables X et Y soit due à l'existence d'une troisième variable Z qui influence à la fois à X et Y. On parle dans ce cas d'une *variable confondante*. On pourrait citer de nombreux exemples dont voici quelques-uns :

- le nombre de boissons fraîches X consommées par jour dans une ville est corrélé avec le nombre de coups de soleil Y répertoriés dans cette ville

  → ceci n'implique pas qu'une forte consommation de ces boissons soit responsable de ces coups de soleil (ni que les coups de soleil donnent soif)

  → la variable confondante est ici la température Z

- le nombre de cigognes X dans un pays est corrélée avec le nombre de naissances Y

  → ceci n'est pas une preuve que les cigognes apportent les bébés (ni que les bébés apportent les cigognes)

  → la variable confondante est ici la grandeur du pays Z (il y aura plus de cigognes et plus de naissances dans un grand pays que dans un petit)

- la rapidité X d'un enfant sur une course de 100 m est corrélée avec sa rapidité de calcul Y

  → ceci n'indique pas forcément que la pratique du sport améliore le quotient intellectuel (ni le contraire)

  → la variable confondante est ici l'âge Z de l'enfant, les performances sportives et intellectuelles s'améliorant naturellement avec l'âge.

Les corrélations obtenues dans ces exemples seront dues uniquement ou partiellement à l'influence évidente de variables confondantes. On pourra alors s'intéresser à la corrélation qui demeure entre X et Y après avoir éliminé l'influence de la variable confondante Z. En particulier dans le troisième exemple, on pourra s'intéresser à la corrélation qui demeure entre performance sportive et performance intellectuelle une fois éliminée l'influence de l'âge. Une telle corrélation sera appelée *corrélation partielle*, comme on le verra au chapitre 14.

En pratique, il ne sera hélas pas toujours aussi facile d'identifier les variables confondantes potentielles, qui demeureront le plus souvent inconnues. Notons également que même si on pouvait être sûr qu'il n'y a aucune variable confondante, le calcul d'une corrélation seule ne nous permettrait pas de trancher si c'est X qui influence Y ou si c'est Y qui influence X, la corrélation étant un concept symétrique. D'une manière générale, il est extrêmement difficile de prouver une relation de cause à effet et on doit le plus souvent se contenter de montrer des associations statistiques entre les variables.

12.5 Corrélation et choix de la population

Lorsque la population contient plusieurs sous-populations dont les distributions diffèrent par rapport aux variables X et/ou Y, la variable qualitative définissant l'appartenance aux différentes sous-populations est un nouvel exemple de variable confondante, qui pourra profondément affecter la corrélation entre X et Y dans un sens ou dans l'autre. La figure 12.5 nous montre quelques exemples fameux de ce qui peut se produire.

Le graphique (a) nous montre une association entre deux variables avec une corrélation de 0.69. Dans le graphique (b), on retrouve cette même corrélation de 0.69 dans deux sous-populations différentes, alors que la corrélation calculée sur l'ensemble des données est négative, égale à -0.91. Si on reprend l'exemple de l'association entre la longueur des pattes et la taille d'un animal, on aura ici une population constituée de deux espèces animales, l'une grande avec de courtes pattes, l'autre petite avec de longues pattes. L'espèce animale représente dans cet exemple une variable confondante. En mélangeant certaines espèces animales, on pourra ainsi selon les cas transformer une corrélation positive en une corrélation négative. En mélangeant d'autres espèces, on pourra de la même manière transformer une corrélation négative en une corrélation positive, ou alors augmenter considérablement une corrélation positive. Le graphique (c) nous montre un mélange de deux nouvelles espèces, l'une grande avec de longues pattes, l'autre petite avec de courtes pattes. La corrélation calculée sur l'ensemble de ces animaux est ici de 0.98 (alors qu'elle est de 0.69 pour chacune des deux espèces). De tels exemples sont assez flagrants et l'utilisation d'un diagramme de dispersion devrait nous garder d'une interprétation erronée d'une corrélation. Parfois, la situation sera cependant plus subtile et il sera difficile de ne pas se laisser tromper. Dans le graphique (d), on a à nouveau affaire à deux espèces animales différentes, l'une en moyenne légèrement plus

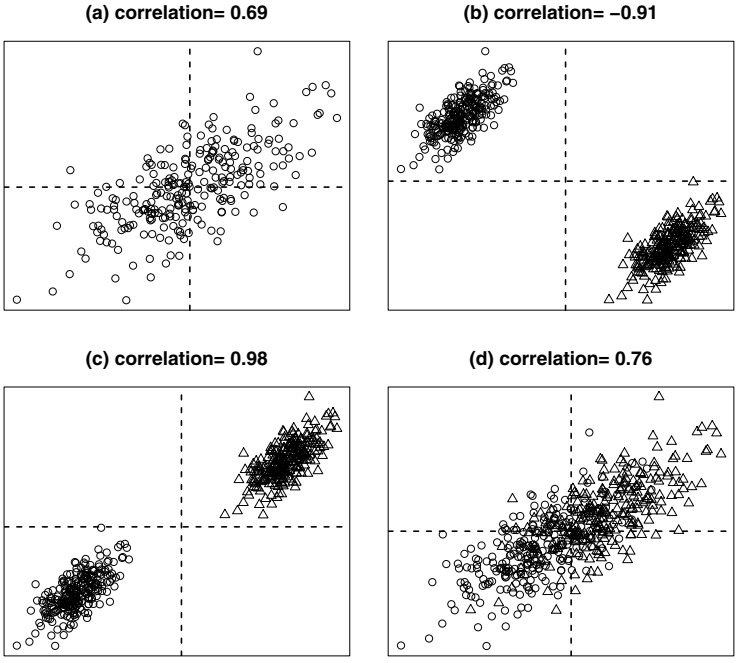

Figure 12.5 – Exemples de corrélation en présence de deux sous-populations.

grande que l'autre et avec des pattes également un peu plus longues. Les deux espèces se recoupent cependant largement, de telle sorte que l'on ne décèlerait pas forcément la présence de ces deux sous-populations sur ce graphique si on n'avait pas utilisé des symboles différents (des ronds et des triangles). La corrélation augmente ici légèrement de 0.69 (pour chacune des espèces) à 0.76 (pour l'ensemble des animaux), ce qui n'est certes pas très spectaculaire, mais ce qui pose tout de même la question de principe suivante : doit-on dans un tel exemple reporter une corrélation de 0.69 ou de 0.76 ? Si la question scientifique considérée est celle de la relation entre la longueur des pattes et la taille au sein d'une espèce animale, on considérera qu'une partie de cette corrélation de 0.76 est due au mélange des espèces et on reportera une corrélation de 0.69.

Exemple 12.3 *On reprend l'exemple de l'association entre la longueur de la main et la taille de* $n = 494$ *étudiants en médecine, où l'on avait calculé une corrélation de* 0.73. *Ces étudiants sont constitués de* 289 *filles et de* 205 *garçons et donc de deux sous-populations distinctes par rapport à la longueur de la main et à la taille de l'étudiant (les garçons étant en moyenne plus grands et avec de plus longues mains que les filles). Or, on vient de voir qu'un tel mélange avait pour effet de gonfler une corrélation. Cela veut dire qu'une partie de cette corrélation de* 0.73 *est en fait due au mélange des sexes. En calculant une*

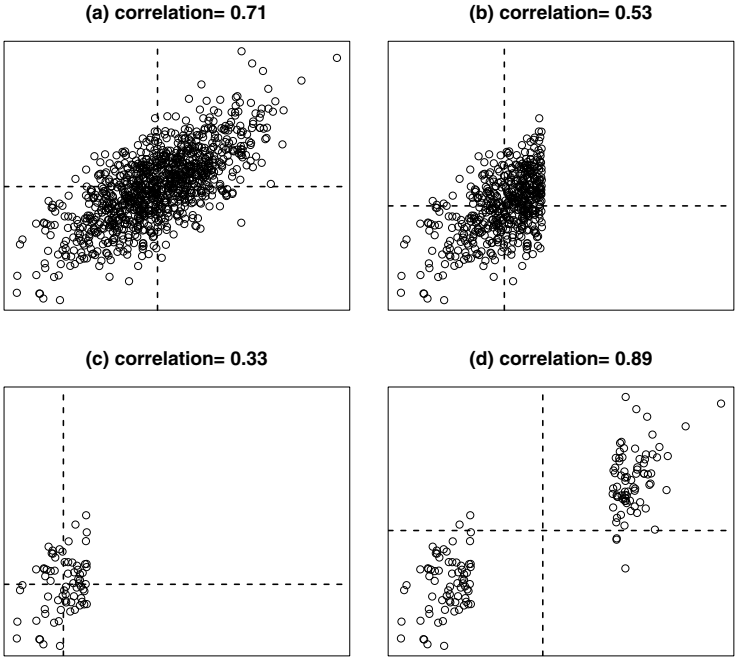

Figure 12.6 – Effet d'une sélection selon X sur une corrélation.

corrélation séparément pour chaque sexe, on trouve 0.57 *chez les filles et* 0.53 *chez les garçons (c'est-à-dire une corrélation d'à peu près* 0.55 *au lieu de* 0.73*). Si notre question (d'anatomie) est la relation qui peut exister entre la longueur de la main et la taille chez l'espèce humaine, on ne voudra sans doute pas y inclure la part confondante due au sexe et on reportera ici une corrélation d'à peu près* 0.55 *plutôt que de* 0.73*.*

On pourra également modifier la valeur d'une corrélation entre deux variables en sélectionnant une sous-population particulière d'une population au départ homogène, selon des critères d'inclusion et d'exclusion parfois artificiels. Les figures 12.6 et 12.7 nous montrent les conséquences possibles d'une telle sélection sur la valeur d'une corrélation. Afin de commenter les graphiques de la figure 12.6, on reprendra l'exemple de la corrélation entre la longueur de la main et la taille d'une personne. Le graphique (a) nous montre un échantillon provenant d'une population homogène avec une corrélation de 0.71. Dans le graphique (b), il s'agit des mêmes données où l'on a éliminé les individus avec de grandes mains. La corrélation s'en trouve diminuée de 0.71 à 0.53. On a en effet éliminé la plupart des individus qui se trouvaient dans le quadrant 1 du graphique (a) et qui contribuaient positivement à la corrélation. La sélection a été encore plus drastique dans le graphique (c), où l'on n'a retenu que les

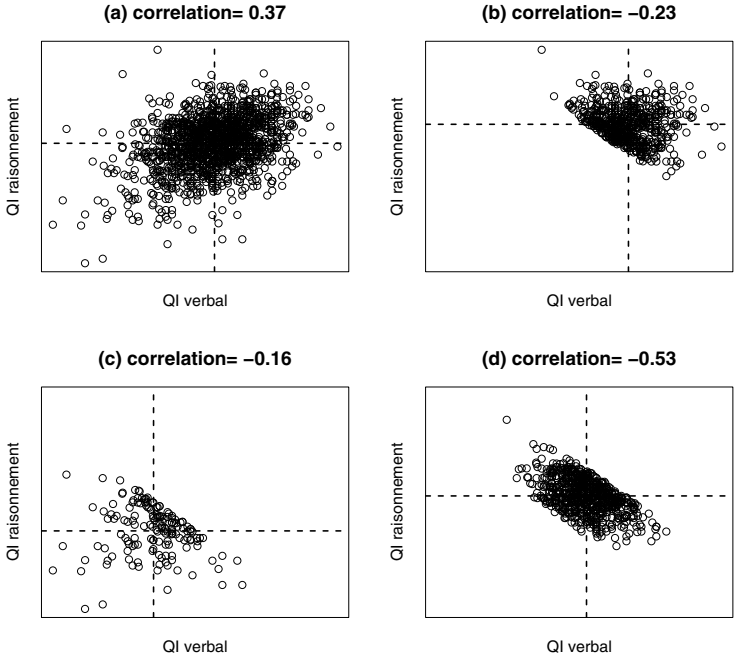

Figure 12.7 – Effet d'une sélection selon X et Y sur une corrélation.

individus avec de petites mains et où la corrélation n'est plus que de 0.33. On trouvera souvent ainsi des corrélations plus petites en restreignant le domaine d'observation. Notons qu'il est aussi possible d'augmenter artificiellement la valeur d'une corrélation, comme illustré dans le graphique (d), où l'on a retenu les individus avec de petites mains et avec de grandes mains, mais où l'on a exclu les mains moyennes. On a ainsi éliminé presque tous les individus qui se trouvaient dans les quadrants 2 et 4 du graphique (a) et qui contribuaient négativement à la corrélation, de sorte que celle-ci augmente de 0.71 à 0.89. Ces exemples nous montrent que l'on peut influencer la valeur d'une corrélation de manière importante en considérant une sous-population plutôt qu'une autre (en sélectionnant nos individus d'une façon plutôt que d'une autre)[5].

Exemple 12.4 *La figure 12.7 nous montre d'autres exemples de sélection artificielle d'individus parmi une population au départ homogène. On s'intéresse ici*

[5]L'assertion « la corrélation entre la longueur de la main et la taille d'une personne est égale à 0.55 » est donc incomplète. Il faut encore préciser à quelle population cette corrélation se réfère. On a vu que cette corrélation sera différente selon que l'on considère une population mélangée de filles et de garçons, ou une population constituée uniquement de filles ou de garçons. Cette corrélation sera également différente selon que l'on considère par exemple une population d'étudiants en médecine ou une population de basketteurs, le choix de ces derniers résultant d'une sélection d'individus en général plutôt grands et avec de grandes mains.

à la relation entre deux types d'intelligence mesurés par deux scores différents de quotient intellectuel (QI), le « QI verbal » et le « QI raisonnement »[6]*. En considérant un échantillon représentatif de jeunes adultes en bonne santé, la corrélation est positive et vaut* 0.37*, comme le montre le graphique (a). Pourtant, si on se restreint aux individus avec un grand QI (ceux pour lesquels la somme du QI verbal et du QI raisonnement surpasse une certaine limite), on se retrouve avec les individus montrés sur le graphique (b) et on obtient une corrélation négative de* -0.23*. On pourrait être ainsi tenté de conclure que chez les surdoués, ces deux composantes de l'intelligence sont corrélées négativement. On obtient également des corrélations négatives de* -0.16 *et de* -0.53 *si on se restreint aux sous-doués ou aux personnes moyennement douées, comme le montrent les graphiques (c) et (d). Ces corrélations négatives ne sont cependant qu'un effet de notre sélection. Dans le graphique (b), les personnes avec un QI verbal modeste auront forcément un QI raisonnement élevé, sans quoi elles n'auraient pas été sélectionnées en tant que « personne surdouée » (la somme des deux QI n'aurait pas surpassé la limite imposée). De même, dans le graphique (d), les personnes avec un QI verbal élevé auront forcément un QI verbal modeste sans quoi elles n'auraient pas été sélectionnées en tant que « personne moyennement douée ».*

Le message principal de cette section est que la valeur d'une corrélation dépend crucialement de la population dans laquelle elle est définie. En considérant une population hétérogène (c'est-à-dire en mélangeant des sous-populations très différentes) ou alors en sélectionnant certains individus plutôt que d'autres à partir d'une population au départ homogène (en observant ainsi une variabilité artificielle et non pas naturelle), on peut modifier à loisir ou presque la valeur d'une corrélation.

12.6 Distribution normale bivariée

En général, on aimera calculer une corrélation dans le cadre d'une population homogène, résultant d'une sélection naturelle et non artificielle des individus. On aura notamment cette situation si la distribution de la variable bivariée (X, Y) est *normale bivariée*, concept que nous introduisons ici. On rappelle que la densité de probabilité d'une variable Y normalement distribuée avec moyenne μ et écart type σ est donnée par :

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

[6]Nous remercions le Dr Oskar Jenni de nous avoir mis à disposition ces données.

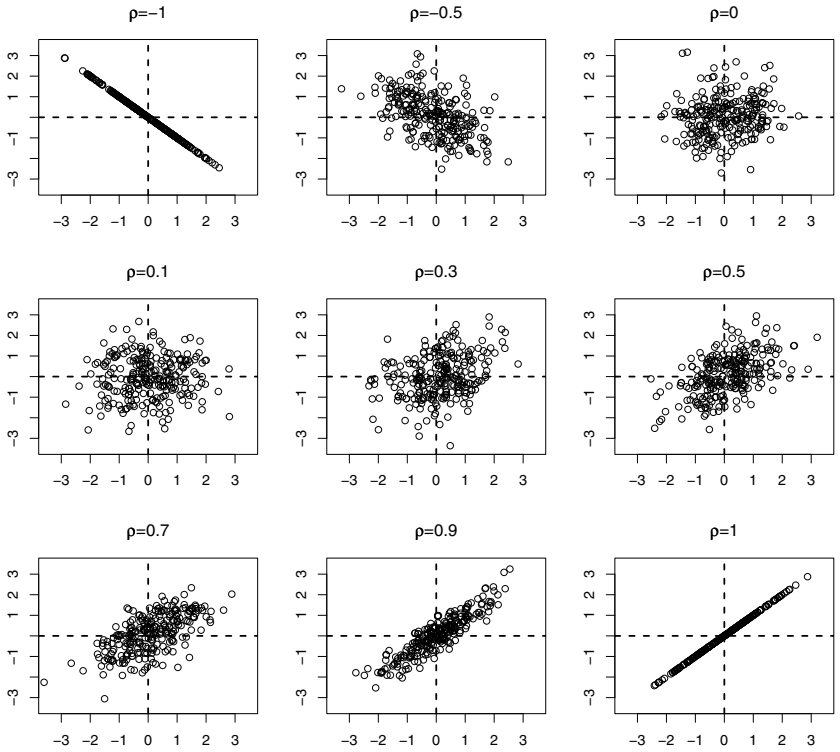

Figure 12.8 – Exemples de données provenant d'une variable normale bivariée.

On dira que la distribution de (X, Y) est *normale bivariée* ou *binormale* si sa densité de probabilité est définie par :

$$f(x,y) = \frac{\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right)}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

où μ_X et μ_Y dénotent les moyennes et σ_X et σ_Y les écarts types des composantes X et Y de la variable bivariée (X, Y), et où ρ dénote leur coefficient de corrélation. La densité de probabilité d'une variable bivariée nous informe sur la probabilité d'obtenir des observations aux alentours des différentes valeurs possible (x, y) de (X, Y) (et qui définit ainsi sa distribution). La famille des distributions normales bivariées est caractérisée par cinq paramètres : μ_X, μ_Y, σ_X, σ_Y et ρ. Alors que les deux premiers sont des mesures de position et les deux suivants des mesures de dispersion de X et de Y, le cinquième est un paramètre d'association entre X et Y. Ce cinquième paramètre ρ caractérise l'association entre X et Y lorsque la distribution de (X, Y) est normale bivariée.

La corrélation ρ n'est pas seulement un résumé,
mais caractérise l'association entre X et Y si (X, Y) est normale bivariée.

On mentionnera en outre les propriétés suivantes valables dans le cas où (X, Y) est normale bivariée :

- une relation exacte entre X et Y est forcément linéaire (et donc une relation exacte entre X et Y implique une corrélation de ± 1)
- une corrélation nulle entre X et Y implique l'indépendance entre ces variables (ce qui on l'a vu n'était pas le cas dans un cadre plus général).

Dans le cas où X et Y sont normales standardisées, c'est-à-dire lorsque $\mu_X = \mu_Y = 0$ et $\sigma_X = \sigma_Y = 1$, la densité de probabilité d'une variable (X, Y) normale bivariée est donnée par :

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

La figure 12.8 nous montre des données provenant de ces distributions binormales pour différentes valeurs de ρ. On reconnaît la binormalité des données aux aspects suivants :

- les composantes X et Y sont normales univariées si (X, Y) est normale bivariée (la réciproque est cependant fausse : la normalité de X et Y est une condition nécessaire mais pas suffisante à la binormalité de (X, Y))
- des données normales bivariées sont homogènes, remplissant une ellipse plus ou moins étirée selon la valeur de la corrélation.

Le calcul d'une corrélation fera donc particulièrement de sens dans un cadre normal bivarié. Afin de nous guider dans notre choix de population, qui on l'a vu aura une grande influence sur la valeur d'une corrélation, une approche possible consiste ainsi à rechercher autant que possible la binormalité des données (par exemple en transformant les données, en éliminant les valeurs extrêmes ou aberrantes, en effectuant des analyses séparées dans différents sous-groupes) avant de calculer une corrélation. En fait, la plupart des exemples potentiellement problématiques vus précédemment (relation exacte mais non linéaire, corrélation nulle sans l'indépendance des variables, présence de valeurs aberrantes, mélange de sous-populations, variabilité artificielle) ne se retrouvent pas dans un cadre normal bivarié. Rechercher la binormalité des données facilitera donc grandement l'interprétation d'une corrélation.

12.7 Corrélation de Spearman

On appelle corrélation de Spearman la corrélation habituelle (c'est-à-dire la corrélation de Pearson) calculée en remplaçant les valeurs originales des

variables par leurs rangs. Si on note par $R = rang(X)$ et $S = rang(Y)$, on a donc par définition :

$$spearman(X, Y) = correlation(R, S).$$

La corrélation de Spearman possède les propriétés avantageuses suivantes :

- au contraire de la corrélation de Pearson, la corrélation de Spearman est robuste par rapport aux valeurs extrêmes ou aberrantes
- comme il s'agit d'un cas particulier de corrélation de Pearson, les valeurs possibles d'une corrélation de Spearman sont comprises entre -1 et $+1$
- les valeurs minimales et maximales de -1 et $+1$ sont atteintes non seulement dans le cas d'une relation exacte linéaire, mais plus généralement dans le cas d'une relation exacte monotone
- la corrélation de Spearman est invariante (au signe près) non seulement suite à un changement d'unités (transformation linéaire/affine), mais plus généralement suite à une transformation monotone de X et/ou de Y

  $\rightarrow$ on aura par exemple $spearman(\log(X), \sqrt{Y}) = spearman(X, Y)$
- Si (X, Y) est normale bivariée avec corrélation ρ, on aura :

  $$spearman(X, Y) = (6/\pi) \arcsin(\rho/2)$$

  $\rightarrow$ comme $(6/\pi) \arcsin(\rho/2) \approx \rho$, on aura ainsi dans un cadre binormal[7] :

  $$spearman(X, Y) \approx correlation(X, Y)$$

  $\rightarrow$ le fait d'avoir des corrélations de Spearman et de Pearson proches l'une de l'autre est compatible avec la binormalité des données.

Admettons qu'il existe des transformations monotones g et h de X et Y telles que la variable bivariée ainsi transformée $(g(X), h(Y))$ soit (approximativement) normale bivariée. Les deux dernières propriétés de la corrélation de Spearman impliquent le résultat remarquable suivant :

$$\begin{aligned} spearman(X, Y) &= spearman(g(X), h(Y)) \\ &\approx correlation(g(X), h(Y)). \end{aligned}$$

En calculant la corrélation de Spearman entre X et Y, on obtient ainsi une approximation de la corrélation de Pearson que l'on aurait si on réussissait à atteindre (ou à approcher) la binormalité à l'aide de transformations g et h, sans pour autant devoir identifier ces transformations.

[7] Dans un cadre binormal, la corrélation de Spearman sera légèrement plus petite que la corrélation habituelle de Pearson (en valeur absolue), mais la différence entre les deux n'excédera pas 0.02.

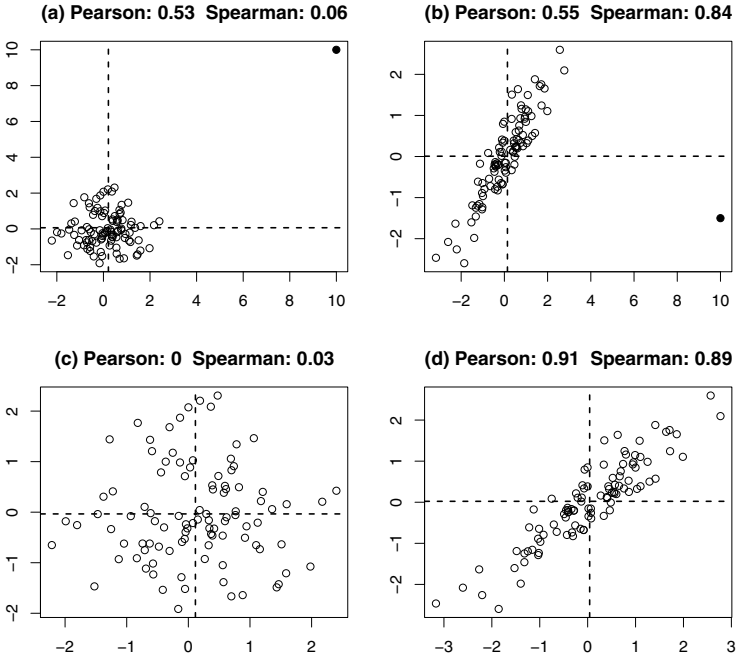

Figure 12.9 – Corrélations avec et sans une valeur aberrante.

Exemple 12.5 *La figure 12.9 nous rappelle que la corrélation de Pearson est très sensible aux valeurs extrêmes ou aberrantes. Dans le graphique (a), la corrélation de Pearson est égale à* 0.53 *alors qu'elle est nulle dans le graphique (c) où une valeur aberrante a été éliminée. À l'inverse, dans le graphique (b), la corrélation de Pearson est égale à* 0.55 *alors qu'elle vaut* 0.91 *dans le graphique (d) où une valeur aberrante a été éliminée. Ainsi, la présence d'une seule valeur aberrante peut gonfler ou dégonfler une corrélation. La corrélation de Spearman est nettement plus robuste à cet égard. Elle est égale à* 0.06 *dans le graphique (a) et à* 0.84 *dans le graphique (b), nous indiquant une faible, respectivement une forte intensité de l'association entre les variables, sans que l'on ait besoin de détecter ni d'éliminer ces observations aberrantes. Notons aussi que les corrélations de Pearson et de Spearman sont proches l'une de l'autre dans les graphiques (c) et (d), où l'on est proche de la binormalité.*

Exemple 12.6 *La figure 12.10 illustre quelques relations entre différents paramètres sanguins mesurés sur* $n = 31$ *femmes atteintes d'une maladie génétique*[8]*. Le graphique (a) nous montre la relation entre la créatine et le pyruvate.*

[8]Il s'agit toujours de l'ensemble de données No 38 du livre de Andrews et Herzberg (1985), où l'on considère ici trois paramètres sanguins au lieu d'un seul.

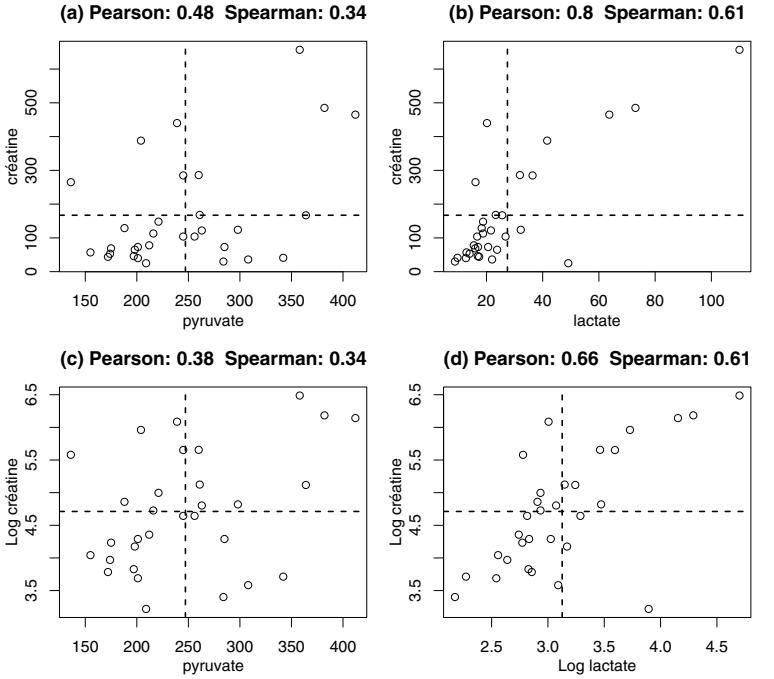

Figure 12.10 – Corrélations avec données originales et transformées.

*Bien qu'il n'y ait pas ici de valeur aberrante évidente, la corrélation de Pearson est bien plus élevée que celle de Spearman (*0.48 *versus* 0.34*). En projetant les données sur l'axe vertical, on remarque cependant une asymétrie prononcée de la distribution de la créatine. Une transformation logarithmique de la créatine nous rapproche de la binormalité, comme illustré sur le graphique (c), où la corrélation de Pearson se trouve relativement proche de celle de Spearman (*0.28 *versus* 0.34*, cette dernière étant invariante par transformation monotone). Ainsi, la corrélation de Spearman calculée sur les données originales du graphique (a) constitue une bonne approximation de la corrélation de Pearson obtenue dans le graphique (c) après transformation des données vers la binormalité. Le graphique (b) nous montre par ailleurs la relation entre la créatine et le lactate. Ici aussi, la corrélation de Pearson est nettement plus élevée que celle de Spearman (*0.80 *versus* 0.61*) alors que les deux corrélations sont plus semblables dans le graphique (d) (*0.66 *versus* 0.61*), où l'on a effectué une transformation logarithmique des deux variables afin de nous rapprocher de la binormalité. Ici aussi, la corrélation de Spearman calculée sur les données originales du graphique (b) constitue une bonne approximation de la corrélation de Pearson obtenue dans le graphique (d) après transformation des données vers la binormalité.*

12.8 Inférence sur la corrélation

On a introduit dans les sections précédentes le concept de corrélation et on a longuement discuté de son interprétation dans différents contextes. Il est temps de rappeler que le calcul d'une corrélation dans notre échantillon n'est qu'un estimateur de la « véritable corrélation » définie sur la population, qui représente ici le paramètre d'intérêt. Nous noterons par ρ cette véritable corrélation et par $\widehat{\rho}$ son estimateur (la corrélation empirique), que l'on supposera calculé à partir de n *observations indépendantes* de la variable bivariée (X, Y). On va voir dans cette section comment faire de l'inférence sur ρ à partir de $\widehat{\rho}$.

La corrélation empirique $\widehat{\rho}$ est un estimateur approximativement sans biais de ρ. Par contre, sa distribution n'est pas normale, de sorte que l'on ne pourra pas directement utiliser la méthode de Wald pour calculer un intervalle de confiance valide pour ρ. On a vu à plusieurs reprises que la transformation d'une variable pouvait nous rapprocher de la normalité. Or, cette technique s'applique également à des estimateurs. Si la variable (X, Y) est normale bivariée, il se trouve que la transformation suivante de l'estimateur :

$$\operatorname{arctanh}(\widehat{\rho}) = \frac{1}{2} \log\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right)$$

aura une distribution approximativement normale. Il s'agit de la transformation « arc-tangente hyperbolique », ou encore *transformation de Fisher*, qui transforme une corrélation comprise entre -1 et $+1$ en un nombre compris entre $-\infty$ et $+\infty$[9].

La corrélation empirique ainsi transformée $\operatorname{arctanh}(\widehat{\rho})$ est un estimateur approximativement sans biais et normalement distribué de la véritable corrélation transformée $\operatorname{arctanh}(\rho)$, son erreur type étant donnée (approximativement) par :

$$\operatorname{SE}(\operatorname{arctanh}(\widehat{\rho})) = \frac{1}{\sqrt{n-3}}.$$

On peut dès lors utiliser la méthode de Wald pour calculer un intervalle de confiance au niveau $1-\alpha$ pour $\operatorname{arctanh}(\rho)$ comme suit :

$$\left[\operatorname{arctanh}(\widehat{\rho}) - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}; \operatorname{arctanh}(\widehat{\rho}) + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right].$$

Afin d'obtenir un intervalle de confiance pour ρ, il suffit d'appliquer aux bornes de cet intervalle la transformation réciproque dite « tangente hyperbolique » :

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$$

[9] Le fait que la variable aléatoire $\operatorname{arctanh}(\widehat{\rho})$ soit plus proche de la normalité que la variable aléatoire $\widehat{\rho}$ n'est donc pas surprenant puisque les valeurs possibles de $\operatorname{arctanh}(\widehat{\rho})$ sont comprises entre $-\infty$ er $+\infty$, ce qui est compatible avec une distribution normale, alors que les valeurs possibles de $\widehat{\rho}$ sont bornées entre -1 et $+1$, ce qui est incompatible avec la normalité (le domaine des valeurs possibles pour une distribution normale n'étant pas borné).

qui transforme un nombre compris entre $-\infty$ et $+\infty$ en une corrélation comprise entre -1 et $+1$. L'intervalle donné par :

$$\left[\tanh\left(\operatorname{arctanh}(\widehat{\rho}) - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right); \tanh\left(\operatorname{arctanh}(\widehat{\rho}) + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}\right)\right]$$

est donc un intervalle de confiance au niveau $1-\alpha$ pour ρ. Cet intervalle est valide pour autant que la distribution de (X, Y) soit proche d'être normale bivariée[10].

En utilisant la dualité entre intervalle de confiance et test statistique, la procédure décrite ci-dessus nous permet également de tester une hypothèse nulle du type :

$$H_0 : \rho = \rho^*.$$

On rejettera H_0 dans un test bilatéral au seuil α si la valeur testée ρ^* n'appartient pas à l'intervalle de confiance ci-dessus[11].

Exemple 12.7 *On reprend l'exemple de l'association entre la longueur de la main et la taille de $n = 289$ filles et $n = 205$ garçons. Dans les deux cas, on est proche de la binormalité. Pour les filles, on avait une corrélation empirique de $\widehat{\rho} = 0.57$. Après transformation, on obtient $\operatorname{arctanh}(\widehat{\rho}) = 0.65$ et donc un intervalle de confiance au niveau 95 % pour $\operatorname{arctanh}(\rho)$ donné par :*

$$0.65 \pm 1.96/\sqrt{286} = [0.53; 0.77].$$

Un intervalle de confiance au niveau 95 % pour ρ est ainsi donné par :

$$[\tanh(0.53); \tanh(0.77)] = [0.49; 0.64].$$

Pour les garçons, on avait une corrélation empirique de $\widehat{\rho} = 0.53$. Après transformation, on obtient $\operatorname{arctanh}(\widehat{\rho}) = 0.59$ et donc un intervalle de confiance au niveau 95 % pour $\operatorname{arctanh}(\rho)$ donné par :

$$0.59 \pm 1.96/\sqrt{202} = [0.45; 0.73].$$

Un intervalle de confiance au niveau 95 % pour ρ est ainsi donné par :

$$[\tanh(0.45); \tanh(0.73)] = [0.42; 0.62].$$

Notons que dans les deux cas, la valeur 0 est loin d'appartenir à l'intervalle de confiance, de sorte que l'on peut conclure que ces corrélations sont significatives au seuil de 5 %.

[10] On pourra utiliser cette même technique pour calculer un intervalle de confiance pour la corrélation de Spearman sans que l'on ait besoin de la binormalité.

[11] La statistique de test calculée sera ici $t_{stat} = \sqrt{n-3} \cdot (\operatorname{arctanh}(\widehat{\rho}) - \operatorname{arctanh}(\rho^*))$ et on aura un résultat significatif dans un test bilatéral au seuil α si $|t_{stat}| \geq z_{1-\alpha/2}$. Comme d'habitude, on conclura dans le cas d'un résultat significatif $\rho < \rho^*$ ou $\rho > \rho^*$ selon que $t_{stat} < 0$ ou $t_{stat} > 0$, c'est-à-dire selon que $\widehat{\rho} < \rho^*$ ou $\widehat{\rho} > \rho^*$.

Toujours dans le cas d'une variable (X, Y) normale bivariée, on mentionnera qu'il existe également un test exact pour tester l'hypothèse nulle particulière :

$$H_0 : \rho = 0.$$

La statistique de test est donnée par :

$$T_{stat} = \frac{\sqrt{n-2} \cdot \widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}}$$

qui sous l'hypothèse nulle aura une distribution de Student avec $n-2$ *dl*. Comme d'habitude, on rejette H_0 dans un test bilatéral au seuil α si $|t_{stat}| \geq t_{1-\alpha/2,n-2}$. Dans le cas d'un résultat significatif, on conclut alors $\rho < 0$ ou $\rho > 0$ selon que $t_{stat} < 0$ ou $t_{stat} > 0$, c'est-à-dire selon que $\widehat{\rho} < 0$ ou $\widehat{\rho} > 0$.

Exemple 12.8 *En reprenant notre exemple de l'association entre la longueur de la main et la taille d'une personne, on calcule pour les filles :*

$$t_{stat} = \frac{\sqrt{287} \cdot 0.57}{\sqrt{1-0.57^2}} = 11.8$$

et pour les garçons :

$$t_{stat} = \frac{\sqrt{203} \cdot 0.53}{\sqrt{1-0.53^2}} = 8.9.$$

Dans les deux cas, on rejette l'hypothèse nulle $H_0 : \rho = 0$ au seuil de 5 % (test bilatéral) car la statistique de test est plus grande que $t_{0.975,287} = 1.97$, respectivement que $t_{0.975,203} = 1.97$ (un logiciel statistique nous donnera dans les deux cas $p < 0.0001$), et on conclut à une corrélation significative et positive entre ces deux variables.

On terminera cette section en rappelant qu'un résultat significatif n'implique pas forcément un résultat cliniquement important. La situation est particulièrement claire dans le cas d'une corrélation, la statistique de test ci-dessus dépendant explicitement des deux seules quantités $\widehat{\rho}$ et n. Ainsi, n'importe quelle corrélation empirique $\widehat{\rho}$ non nulle sera significative si elle est observée sur un échantillon de taille n suffisamment grande[12].

[12]Une corrélation empirique $\widehat{\rho}$ sera en effet significative au seuil de 5 % (dans un test bilatéral) si $n \geq 2 + 1.96^2(1-\widehat{\rho}^2)/\widehat{\rho}^2$ (on a ici remplacé pour simplifier $t_{0.975,n-2}$ par 1.96). Ainsi, même une corrélation empirique aussi petite que $\widehat{\rho} = 0.01$ sera significative si elle a été calculée sur un échantillon de $n = 40\,000$ individus.

Chapitre 13

Régression linéaire simple

Nous voici arrivés à la troisième problématique énoncée en début de texte, celle de la modélisation statistique. On considère dans ce chapitre une variable bivariée continue (X, Y). Alors que le calcul de la corrélation entre X et Y nous informait sur la direction (globalement positive, globalement négative ou globalement nulle) de la relation entre X et Y, ainsi que sur le « degré d'exactitude » de cette relation, la régression linéaire que nous introduisons dans ce chapitre nous permettra de décrire plus précisément *comment* les deux variables sont liées. Ceci nous permettra également d'effectuer des prédictions pour une variable à partir de l'autre. On essaiera ainsi de répondre aux questions suivantes :

1. Comment peut-on prédire Y à partir de X ?
2. Comment se modifie Y en fonction de X ?

Au contraire de la corrélation qui était un concept symétrique des deux variables, les rôles de X et Y différeront en régression, où X sera le *prédicteur* et Y la *variable réponse*.

13.1 Droite de régression

La figure 13.1 montre le diagramme de dispersion de la relation entre la taille X et le poids Y pour $n = 30$ observations bivariées $(x_1, y_1), \cdots, (x_n, y_n)$[1]. La relation est ici globalement positive, les individus les plus grands étant aussi en général (mais pas systématiquement) les plus lourds. Une droite est également présente sur ce graphique. Sa pente positive nous indique la direction de la relation. Il s'agit de la *droite de régression* qui nous aidera à répondre aux deux questions énoncées ci-dessus.

[1] Il s'agit d'adultes de sexe féminin qui ont participé à l'étude CoLaus à Lausanne, décrite dans l'article de Firmann *et al.* (2008).

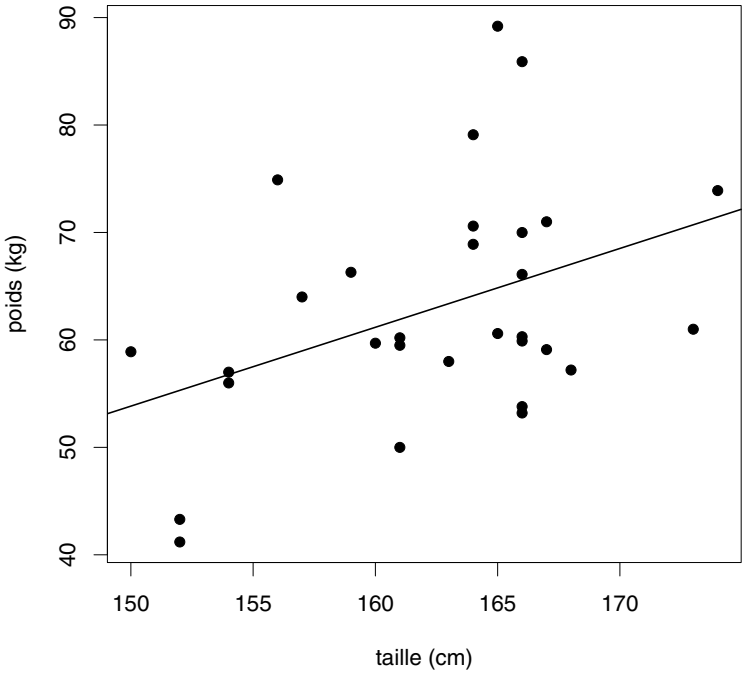

Figure 13.1 – Diagramme de dispersion avec droite de régression.

L'équation de cette droite est notée par :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

où $\widehat{\beta}_0$ représente la constante (en anglais : `intercept`) et $\widehat{\beta}_1$ la pente (en anglais : `slope`) de cette droite. On utilise cette équation pour prédire le poids d'un individu en fonction de sa taille x.

$\widehat{\beta}_0 + \widehat{\beta}_1 x$ est le poids prédit pour un individu avec taille x.

En appliquant ces calculs aux individus de notre échantillon, on peut comparer les poids ainsi prédits avec les poids réellement observés pour ces individus, ce qui nous permet d'avoir une idée de la qualité de ces prédictions. On utilise les notations suivantes :

- $\widehat{y}_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ est la valeur prédite du poids du i-ième individu de notre échantillon
- $\widehat{\varepsilon}_i = y_i - \widehat{y}_i^*$ est son *erreur de prédiction*, aussi appelée *résidu*.

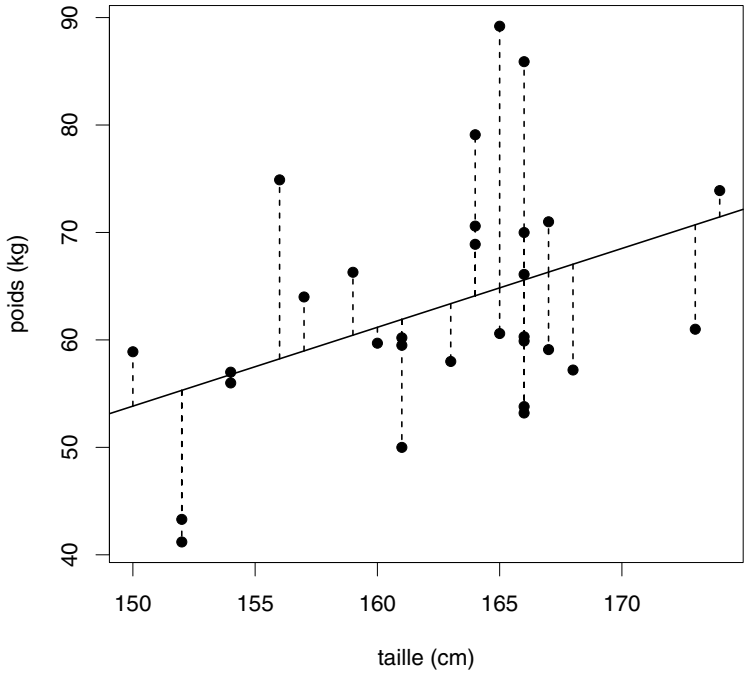

Figure 13.2 – Illustration du critère des moindres carrés.

La droite de régression est calculée de façon à minimiser globalement ces erreurs de prédiction. Le critère de minimisation que l'on utilise est celui des *moindres carrés*, défini par :

$$\sum_i \widehat{\varepsilon}_i^2 = \sum_i (y_i - \widehat{y}_i^*)^2 = \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

La droite de régression, appelée également *droite des moindres carrés*, est donc la droite pour laquelle la somme des carrés des distances verticales entre les données et la droite est minimale. Ces distances verticales (qui représentent les erreurs de prédictions) sont mises en évidence dans la figure 13.2.

Les coefficients $\widehat{\beta}_0$ et $\widehat{\beta}_1$ de la droite de régression sont les solutions des dérivées partielles du critère des moindres carrés par rapport à $\widehat{\beta}_0$ et $\widehat{\beta}_1$. Ils satisfont les deux conditions :

$$\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

et

$$\sum_i x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0.$$

Le résolution de ce système d'équations à deux inconnues nous donne les solutions :

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_i x_i^2 - \bar{x}^2}$$

et

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Afin de calculer la droite de régression, il suffit donc de connaître les moyennes suivantes : $\bar{x} = \frac{1}{n}\sum_i x_i$, $\bar{y} = \frac{1}{n}\sum_i y_i$, $\frac{1}{n}\sum_i x_i^2$ et $\frac{1}{n}\sum_i x_i y_i$. Ainsi, la pente de la droite de régression est égale à la covariance entre X et Y divisée par la variance de X. Notons aussi le lien suivant entre le coefficient de corrélation et la pente de la droite de régression[2] :

$$\widehat{\rho} = \widehat{\beta}_1 \cdot \frac{\tilde{\sigma}_X}{\tilde{\sigma}_Y}.$$

Comme le quotient $\tilde{\sigma}_X/\tilde{\sigma}_Y$ des estimateurs des écarts types de X et de Y est un nombre positif, le signe de la pente de la droite de régression est le même que le signe de la corrélation. Ainsi :

- la relation entre X et Y est **globalement positive** si $\widehat{\beta}_1 > 0$
- la relation entre X et Y est **globalement négative** si $\widehat{\beta}_1 < 0$
- la relation entre X et Y est **globalement nulle** si $\widehat{\beta}_1 = 0$.

La droite de régression possède de nombreuses propriétés mathématiques. En voici quelques-unes :

- la droite de régression passe par le centre de gravité des données $(\bar{x}, \bar{y})$
- la somme (et donc la moyenne) des résidus est nulle

  → on aura en effet :

$$\begin{aligned}
\sum_i \widehat{\varepsilon}_i &= \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) \\
&= \sum_i (y_i - (\bar{y} - \widehat{\beta}_1 \bar{x}) - \widehat{\beta}_1 x_i) \\
&= \sum_i (y_i - \bar{y}) - \widehat{\beta}_1 \sum_i (x_i - \bar{x}) \\
&= 0
\end{aligned}$$

[2] On peut ici utiliser indifféremment les estimateurs des variances avec dénominateur n ou $n-1$, et donc remplacer le quotient $\tilde{\sigma}_X/\tilde{\sigma}_Y$ par le quotient $\widehat{\sigma}_X/\widehat{\sigma}_Y$. De même, on peut utiliser indifféremment un dénominateur n ou $n-1$ pour la covariance et la variance utilisées lors du calcul de $\widehat{\beta}_1$ car on a :

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2} = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}.$$

→ les erreurs de prédictions positives et négatives s'annulent

→ ceci implique également que $\frac{1}{n}\sum_i \widehat{y}_i^* = \frac{1}{n}\sum_i y_i = \bar{y}$

- la moyenne des résidus étant nulle, la variance des résidus est proportionnelle au critère des moindres carrés : $\sum_i \widehat{\varepsilon}_i^2$
- minimiser le critère des moindres carrés revient donc à définir la droite de régression comme étant la droite pour laquelle les erreurs de prédictions ont une *moyenne nulle* et une *variance minimale*.

Exemple 13.1 *Voici le détail des données pour les* $n = 30$ *individus représentés sur les figures 13.1 et 13.2 :*

i	x_i	y_i	$\widehat{y}_i^*$	$\widehat{\varepsilon}_i$	i	x_i	y_i	$\widehat{y}_i^*$	$\widehat{\varepsilon}_i$
1	166	60.3	65.6	−5.3	16	164	79.1	64.1	15.0
2	161	50.0	61.9	−11.9	17	156	74.9	58.2	16.7
3	167	59.1	66.3	−7.2	18	166	85.9	65.6	20.3
4	174	73.9	71.4	2.5	19	166	70.0	65.6	4.4
5	165	89.2	64.8	24.4	20	164	70.6	64.1	6.5
6	164	68.9	64.1	4.8	21	159	66.3	60.4	5.9
7	157	64.0	59.0	5.0	22	167	71.0	66.3	4.7
8	152	43.3	55.3	−12.0	23	166	59.9	65.6	−5.7
9	166	53.2	65.6	−12.4	24	160	59.7	61.2	−1.5
10	161	60.2	61.9	−1.7	25	165	60.6	64.8	−4.2
11	166	66.1	65.6	0.5	26	166	53.8	65.6	−11.8
12	161	59.5	61.9	−2.4	27	152	41.2	55.3	−14.1
13	154	57.0	56.8	0.2	28	154	56.0	56.8	−0.8
14	150	58.9	53.8	5.1	29	163	58.0	63.4	−5.4
15	168	57.2	67.0	−9.8	30	173	61.0	70.7	−9.7

On a ici $\frac{1}{n}\sum_i x_i = 162.43$, $\frac{1}{n}\sum_i y_i = 62.96$, $\frac{1}{n}\sum_i x_i^2 = 26\,419.17$ *et* $\frac{1}{n}\sum_i x_i y_i = 10\,252.18$. *On calcule ainsi :*

$$\widehat{\beta}_1 = \frac{10\,252.18 - 162.43 \cdot 62.96}{26\,419.17 - 162.43^2} = 0.734$$

et

$$\widehat{\beta}_0 = 62.96 - 0.734 \cdot 162.43 = -56.23.$$

La droite de régression est ainsi donnée par l'équation :

$$\widehat{y}^* = -56.23 + 0.734x.$$

Les valeurs prédites $\widehat{y}_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ *des poids des individus de notre échantillon sont données dans le tableau ci-dessus, de même que les erreurs de prédiction (ou résidus)* $\widehat{\varepsilon}_i = y_i - \widehat{y}_i^*$. *On compte* 14 *erreurs positives (sous-estimation du poids) et* 16 *erreurs négatives (surestimation du poids). Le lecteur vérifiera que la somme des* 14 *erreurs positives compense exactement la somme de* 16 *erreurs négatives, la somme (et donc la moyenne) de ces résidus étant nulle.*

13.2 Droite de régression sur la population

Dans la section précédente, on a utilisé des chapeaux pour désigner les coefficients de la droite de régression car ils étaient calculés à partir des données de l'échantillon. Rappelons toutefois que notre intérêt n'est pas l'échantillon, mais toute la population de laquelle provient notre échantillon. Si on connaissait la taille et le poids de tous les individus de notre population, on pourrait en effet calculer une droite de régression sur cette population, en utilisant le même critère des moindres carrés. Cette droite théorique représenterait la « véritable droite de régression ». Comme d'habitude, la droite de régression $\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x$ calculée sur notre échantillon sera une estimation de la véritable droite de régression définie sur la population, que nous noterons par $y^* = \beta_0 + \beta_1 x$.

Il est important de noter que même si on connaissait les véritables paramètres β_0 et β_1 de la véritable droite de régression, et que l'on utilisait cette droite pour faire des prédictions, on ferait quand même des erreurs de prédiction (la relation entre la taille et le poids, qui n'est pas exacte dans notre échantillon, ne sera pas non plus exacte dans la population). Si on utilisait la véritable droite de régression pour faire des prédictions, la prédiction que l'on ferait pour le i-ième individu serait donnée par :

$$y_i^* = \beta_0 + \beta_1 x_i$$

et on aurait ainsi l'erreur de prédiction suivante :

$$\varepsilon_i = y_i - y_i^* = y_i - \beta_0 - \beta_1 x_i.$$

Or, cette « véritable erreur de prédiction » ε_i ne peut pas être observée dans notre échantillon. L'erreur de prédiction $\widehat{\varepsilon}_i$ que nous calculons dans notre échantillon pour le i-ième individu :

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i^* = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

n'est donc qu'une « estimation » de la véritable erreur se prédiction ε_i. Contrairement au poids Y ou à la taille X, la variable que l'on appelle *variable résiduelle* définie par :

$$\varepsilon = Y - \beta_0 - \beta_1 X$$

et qui représente « l'erreur de prédiction que l'on ferait si on connaissait la véritable droite de régression », est une *variable non observable* car elle dépend de paramètres β_0 et β_1 qui sont eux-mêmes inconnus.

La droite de régression définie sur la population a des propriétés similaires à la droite de régression calculée dans notre échantillon. On a ainsi[3] :

$$\beta_1 = \frac{covariance(X, Y)}{variance(X)}$$

[3]On rappellera à cette occasion que la régression n'est pas un concept symétrique des variables X et Y. Les paramètres β_0 et β_1 de la droite de régression ci-dessus sont définis avec X comme prédicteur et Y comme réponse. Si on inversait les rôles (si Y était le prédicteur

et

$$\beta_0 = mean(Y) - \beta_1 \cdot mean(X).$$

La droite de régression est par ailleurs la droite pour laquelle la variable résiduelle $\varepsilon = Y - \beta_0 - \beta_1 X$ est de moyenne nulle et de variance minimale. On a ainsi $mean(\varepsilon) = 0$, alors que $variance(\varepsilon)$ est minimisé. On notera à nouveau le lien entre le coefficient de corrélation et la pente de la droite de régression :

$$correlation(X, Y) = \beta_1 \cdot \sqrt{\frac{variance(X)}{variance(Y)}}.$$

13.3 Variance prédite et variance résiduelle

On peut décomposer une variable Y en une somme de deux termes :

$$Y = (\beta_0 + \beta_1 X) + \varepsilon.$$

Le premier terme $\beta_0 + \beta_1 X$ représente la droite de régression, aussi appelée *meilleur prédicteur linéaire* de Y à partir de X. Le second terme ε représente l'erreur de prédiction, aussi appelée variable résiduelle. Or, l'erreur de prédiction se trouve être non corrélée avec le prédicteur. On a en effet :

$$\begin{aligned} covariance(X, \varepsilon) &= covariance(X, Y - \beta_0 - \beta_1 X) \\ &= covariance(X, Y) - \beta_1 \cdot variance(X) \\ &= covariance(X, Y) - \frac{covariance(X, Y)}{variance(X)} \cdot variance(X) \\ &= 0. \end{aligned}$$

Il s'ensuit :

$$\begin{aligned} variance(Y) &= variance(\beta_0 + \beta_1 X + \varepsilon) \\ &= variance(\beta_0 + \beta_1 X) + variance(\varepsilon) + 2\beta_1 \cdot covariance(X, \varepsilon) \\ &= variance(\beta_0 + \beta_1 X) + variance(\varepsilon). \end{aligned}$$

Il s'agit d'un nouveau résultat fondamental de la statistique que l'on aura plaisir à encadrer :

---

et X la réponse), on obtiendrait une droite de régression avec paramètres donnés par :

$$\beta_1' = \frac{covariance(X, Y)}{variance(Y)}$$

ainsi que :

$$\beta_0' = mean(X) - \beta_1' \cdot mean(Y)$$

qui seront différents des paramètres β_0 et β_1 (sauf si X et Y ont la même moyenne et la même variance, auquel cas ils seront identiques).

$$\boxed{variance(Y) = variance(\beta_0 + \beta_1 X) + variance(\varepsilon).}$$

On peut ainsi décomposer la variance de Y en une somme de deux variances, la variance du meilleur prédicteur linéaire et la variance résiduelle. Autrement dit, on peut décomposer la variance (la variabilité) de Y en une partie prédictible et une partie non prédictible :

- $variance(\beta_0 + \beta_1 X)$, c'est-à-dire la variance de Y prédite linéairement par X, représente la **variabilité prédictible**
- $variance(\varepsilon)$, c'est-à-dire la variance résiduelle, représente la **variabilité non prédictible**.

À ce stade, on mentionnera le résultat remarquable suivant, que nous avions déjà évoqué dans le chapitre précédent :

$$\begin{aligned} \rho^2 &= \beta_1^2 \cdot \frac{variance(X)}{variance(Y)} \\ &= \frac{variance(\beta_0 + \beta_1 X)}{variance(Y)} \\ &= 1 - \frac{variance(\varepsilon)}{variance(Y)}. \end{aligned}$$

Autrement dit :

- le carré ρ^2 de la corrélation entre X et Y représente le *pourcentage de la variance de Y prédite linéairement par X*
- la quantité $1 - \rho^2$ représente le pourcentage de la variance de Y qui n'est pas prédite linéairement par X.

Ainsi, le carré de la corrélation est une mesure de la qualité globale des prédictions que l'on fera en utilisant la droite de régression[4]. On notera les deux cas extrêmes suivants :

- $\rho^2 = 1$: on prédit 100 % de la variance de Y linéairement par X (on a une relation exacte linéaire entre X et Y)
- $\rho^2 = 0$: on prédit 0 % de la variance de Y linéairement par X (la connaissance de X ne nous sert en rien pour prédire Y, du moins si on utilise la droite de régression).

[4]Bien que la droite de régression ne soit en général pas la même pour prédire Y à partir de X ou pour prédire X à partir de Y, la qualité de la prédiction sera la même dans les deux cas, ρ^2 représentant à la fois le pourcentage de la variance de Y prédite linéairement par X et le pourcentage de la variance de X prédite linéairement par Y.

Cette décomposition de la variabilité de Y en une partie prédictible et une partie non prédictible se retrouve également dans notre échantillon. En particulier, le carré $\widehat{\rho}^2$ de la corrélation empirique entre X et Y est un estimateur du pourcentage de la variance de Y prédite linéairement par X.

Exemple 13.2 *Dans notre exemple ci-dessus, on peut calculer* $\widehat{\rho} = 0.405$ *et on estime ainsi à* $\widehat{\rho}^2 = 16.4$ % *le pourcentage de la variance du poids prédite linéairement par la taille (*83.6 % *de la variance étant non prédite).*

Appliquons à présent la méthode des moindres carrés à une situation où l'on prédirait le poids Y d'un individu sans connaître sa taille (ni aucun autre prédicteur X). On fait donc ici la même prédiction quel que soit l'individu, notre équation de prédiction étant simplement $\widehat{y}^* = \widehat{\beta}_0$. L'erreur de prédiction pour le i-ième individu d'un échantillon de n observations $y_1, y_2, \cdots, y_n$ est donnée par $y_i - \widehat{\beta}_0$, de sorte que le critère des moindres carrés que l'on minimise est défini par :

$$\sum_i (y_i - \widehat{\beta}_0)^2.$$

Le lecteur vérifiera que la solution est donnée par $\widehat{\beta}_0 = \bar{y}$. On prédit ainsi le poids d'un individu par la moyenne (estimée) de la variable « poids ». On retrouve le même résultat sur la population où l'on aura $\beta_0 = mean(Y)$. Sans prédicteur, la variable résiduelle est donc simplement $Y - mean(Y)$, de sorte que la variance résiduelle est égale à $variance(Y)$. Comparons alors la situation sans prédicteur, avec la situation où l'on dispose d'un prédicteur X pour prédire le poids Y. Dans ce dernier cas, on dénote comme d'habitude par ε la variable résiduelle. La variance résiduelle est ainsi égale à $variance(\varepsilon)$ si on dispose d'un prédicteur, alors qu'elle est égale à $variance(Y)$ sans prédicteur, ce qui représente une réduction de variance résiduelle de :

$$variance(Y) - variance(\varepsilon).$$

Exprimé en pourcentage (de la variance résiduelle sans prédicteur), ceci représente une réduction de :

$$\frac{variance(Y) - variance(\varepsilon)}{variance(Y)} = 1 - \frac{variance(\varepsilon)}{variance(Y)} = \rho^2.$$

Le carré de la corrélation entre X et Y est donc également interprétable comme « réduction (en %) de la variance résiduelle en utilisant X comme prédicteur de Y, comparé à une situation sans prédicteur ».

Exemple 13.3 *On considère une variable* Y *de variance égale à* 250. *Supposons que la variance prédite (en utilisant une équation de régression du type* $\widehat{y}^* = \beta_0 + \beta_1 x$*) soit égale à* 100 *et la variance résiduelle à* 150. *Ainsi,* $100/250 = 40$ % *de la variance de* Y *est prédite par* X. *Alternativement, on peut dire que la variance résiduelle a été réduite de* 250 *(sans utiliser* X*) à* 100 *(en utilisant* X*), ce qui correspond effectivement à une réduction de* $(250-150)/250 = 40$ %.

13.4 Hypothèse de linéarité

On a jusqu'ici considéré la droite de régression comme un outil de prédiction. On va à présent lui donner une interprétation descriptive plus précise en faisant une *hypothèse de linéarité*. On a vu au chapitre 5 comment comparer deux groupes par rapport à une variable continue *via* leur différence de moyenne. Si on désirait analyser la relation entre la taille X et le poids Y, on pourrait ainsi comparer un « groupe de petits » et un « groupe de grands » par rapport à leur poids. On calculerait la moyenne des poids dans ces deux groupes et la différence de ces moyennes. En régression, on fait mieux. On considère *une infinité de groupes* : le groupe des individus mesurant 160 cm, le groupe des 165 cm, le groupe des 170 cm, le groupe des 171.23 cm, etc, et on s'intéresse au poids moyen dans chacun de ces groupes. On notera par $mean(Y|X = x)$ le poids moyen du groupe mesurant x cm. L'hypothèse de linéarité consiste à supposer que ces moyennes sont alignées sur une droite. Si elles le sont, alors ce sera forcément sur la droite de régression.

L'hypothèse de linéarité peut ainsi être formulée de la façon suivante :

$$mean(Y|X = x) = \beta_0 + \beta_1 x.$$

Sous cette hypothèse, la quantité $\beta_0 + \beta_1 x$ n'est pas seulement la prédiction du poids d'un individu mesurant x cm, c'est également le poids moyen des individus mesurant x cm. Autrement dit, on prédit le poids d'un individu par le poids moyen du groupe auquel cet individu appartient, l'erreur de prédiction étant la différence entre le poids d'un individu et le poids moyen de son groupe. Cette dernière sera également interprétée comme le poids d'un individu *ajusté pour sa taille*. Pour les individus de notre échantillon, $\widehat{y}_i^*$ sera donc une estimation de $mean(Y|X = x_i)$ et $\widehat{\varepsilon}_i$ sera une estimation de $y_i - mean(Y|X = x_i)$.

Exemple 13.4 *Dans notre exemple de la relation entre la taille et le poids, on avait estimé l'équation de régression par* $\widehat{y}^* = -56.23 + 0.734x$. *On avait également calculé les valeurs prédites* $\widehat{y}_i^*$ *et les erreurs de prédictions* $\widehat{\varepsilon}_i$ *pour chacun des* $n = 30$ *individus. Comparons par exemple les* 6^e *et* 7^e *individus :*

- *le* 6^e *individu mesure* $x_6 = 164$ *cm, ce qui lui vaut un poids prédit de* $\widehat{y}_6^* = -56.23 + 0.734 \cdot 164 = 64.1$ *kg ; or, cet individu pèse en réalité* $y_6 = 68.9$ *kg, d'où une erreur de prédiction de* $\widehat{\varepsilon}_6 = 68.9 - 64.1 = 4.8$ *kg*
- *le* 7^e *individu mesure* $x_7 = 157$ *cm, ce qui lui vaut un poids prédit de* $\widehat{y}_7^* = -56.23 + 0.734 \cdot 157 = 59.0$ *kg ; or, cet individu pèse en réalité* $y_7 = 64.0$ *kg, d'où une erreur de prédiction de* $\widehat{\varepsilon}_7 = 64.0 - 59.0 = 5.0$ *kg.*

En faisant l'hypothèse de linéarité, on a en outre les interprétations suivantes :

- *le poids moyen du groupe du* 6^e *individu (ceux mesurant* 164 *cm) est estimé à* 64.1 *kg ; avec ses* 68.9 *kg, cet individu pèse donc* 4.8 *kg de plus que la moyenne de son groupe*

- *le poids moyen du groupe du* 7^e *individu (ceux mesurant* 157 *cm) est estimé à* 59.0 *kg ; avec ses* 64.0 *kg, cet individu pèse donc* 5.0 *kg de plus que la moyenne de son groupe.*

Ainsi, bien que le 6^e *individu soit plus lourd que le* 7^e *(poids de* 68.9 *kg contre* 64.0 *kg), il sera considéré comme moins lourd si on ajuste le poids pour la taille (poids ajusté pour la taille de* +4.8 *kg contre* +5.0 *kg).*

Une façon équivalente de définir l'hypothèse de linéarité est la suivante :

$$\begin{aligned} mean(\varepsilon|X=x) &= mean(Y-\beta_0-\beta_1 X|X=x) \\ &= mean(Y|X=x)-mean(\beta_0+\beta_1 X|X=x) \\ &= \beta_0+\beta_1 x-(\beta_0+\beta_1 x) \\ &= 0. \end{aligned}$$

On a vu que l'on a de toute façon $mean(\varepsilon) = 0$ (que l'hypothèse de linéarité soit satisfaite ou non). Avec l'hypothèse de linéarité, on aura en plus $mean(\varepsilon|X = x) = 0$ quelle que soit la valeur de x. Ainsi, les erreurs de prédiction ne seront pas seulement nulles en moyenne globalement, mais aussi nulles en moyenne localement (c'est-à-dire dans chaque groupe défini par $X = x$). Dans notre exemple, cela veut dire que les erreurs de prédiction seront en moyenne nulles dans le groupe des individus mesurant 160 cm, dans le groupe des 165 cm, dans le groupe des 170 cm, dans le groupe des 171.23 cm, etc.

Comme son nom l'indique, l'hypothèse de linéarité est une *hypothèse*, qui selon les cas sera raisonnable ou non raisonnable, et qu'il s'agira de vérifier, par exemple graphiquement. On vérifiera ainsi que la droite de régression passe bien au milieu des points représentant les données non seulement globalement, mais aussi localement. Comme on l'a dit pour la normalité, l'hypothèse de linéarité est cependant un concept théorique. En pratique, l'hypothèse de linéarité ne sera jamais vraie au sens strict du terme, mais pourra constituer une bonne approximation de la réalité. Il s'agira donc de vérifier graphiquement que l'hypothèse de linéarité soit *approximativement satisfaite* (sachant qu'elle ne sera jamais strictement satisfaite, ni dans un échantillon, ni même dans une population)[5].

La figure 13.3 nous montre quelques exemples. Dans les graphiques (a) et (b), on dispose effectivement de 10 groupes d'individus (avec dans chaque groupe une valeur de X différente)[6]. Dans le graphique (a), la droite de régression passe bien (approximativement) par la moyenne de la variable Y dans

[5]Il y a parfois une ambiguïté de langage lorsque l'on parle de « relation linéaire » entre deux variables. En mathématique, cela veut dire que les *données individuelles* sont alignées sur une droite. En statistique, cela veut dire que des *moyennes* (calculées sur des individus semblables) sont alignées sur une droite.

[6]On aura cette situation dans des études expérimentales, où il sera possible de choisir (au lieu de simplement observer) les valeurs de la variable X pour les individus de notre échantillon. On pourra par exemple choisir les doses d'un médicament administré aux patients. On pourra ainsi former 10 groupes de patients (correspondant à 10 doses différentes), les patients au sein d'un même groupe recevant la même dose.

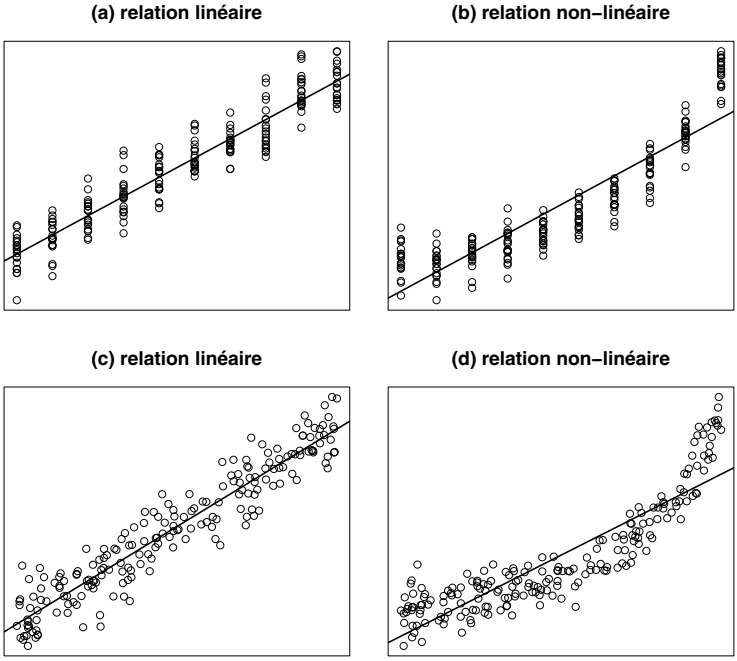

Figure 13.3 – Exemples de relations linéaires et non linéaires entre variables.

chaque groupe. L'hypothèse de linéarité est donc approximativement satisfaite. Cela n'est pas le cas dans le graphique (b) où la droite de régression passe dans certains groupes plus près du minimum ou du maximum que de la moyenne, ce qui contredit l'hypothèse de linéarité. Si les n observations de la variable X dans notre échantillon sont toutes différentes les unes des autres, de sorte que l'on dispose en fait de n groupes avec dans chaque groupe un seul individu comme c'est le cas dans les graphiques (c) et (d), on procédera d'une manière similaire en regroupant localement les individus. On voit ainsi que l'hypothèse de linéarité est approximativement satisfaite dans le graphique (c), alors qu'elle n'est pas du tout satisfaite dans le graphique (d).

Notre droite de régression nous permet donc d'estimer la moyenne d'une variable Y dans une infinité de groupes définis par les différentes valeurs possibles x d'une variable X. Par interpolation, on peut obtenir des estimations y compris pour des valeurs x non représentées dans notre échantillon. En calculant par exemple $\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 170$, on obtient une estimation du poids moyen des individus mesurant 170 cm quand bien même on n'aurait aucun individu mesurant exactement 170 cm dans notre échantillon. Une telle estimation sera raisonnable pour autant que notre échantillon contienne des individus qui soient un peu plus petits et d'autres qui soient un peu plus grands que 170 cm, de manière

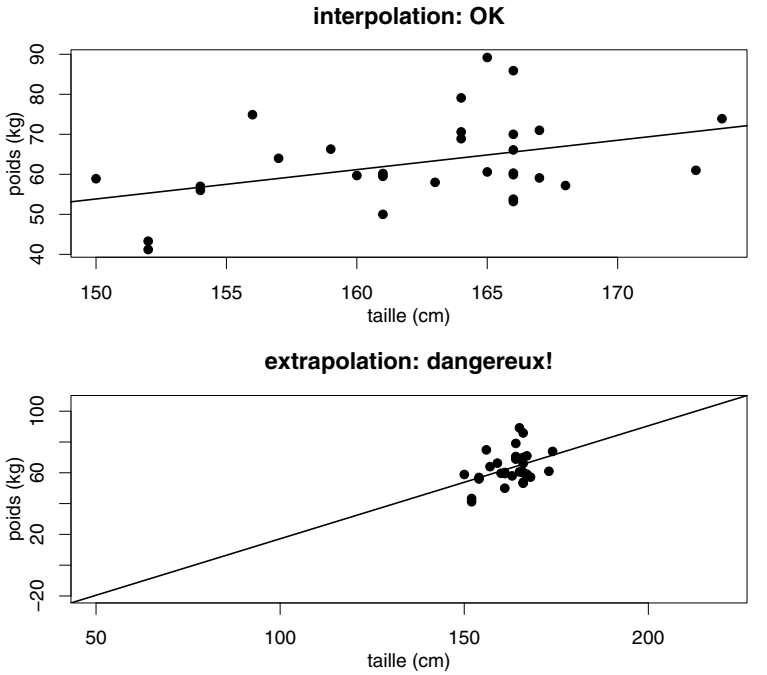

Figure 13.4 – Interpolation et extrapolation d'une régression.

à ce que la valeur 170 cm se trouve à l'intérieur de notre domaine d'observation, où l'on peut effectivement vérifier l'hypothèse de linéarité. Par contre, il sera impossible de vérifier l'hypothèse de linéarité en dehors de notre domaine d'observation, de sorte qu'une extrapolation de notre équation de régression hors de ce domaine se fera à nos risques et périls, sans aucune garantie.

Exemple 13.5 *Dans notre exemple de la relation entre la taille et le poids, nos* $n = 30$ *individus mesurent entre* 150 *et* 174 *cm. L'hypothèse de linéarité semble ici raisonnable (comme le montre le graphique du haut de la figure 13.4, bien qu'il soit parfois difficile de juger avec peu d'observations). On peut ainsi utiliser notre équation de régression donnée par :*

$$\widehat{y}^* = -56.23 + 0.734x$$

pour estimer des poids moyens à l'intérieur de notre domaine d'observation. On estimera par exemple le poids moyen des 170 *cm par :*

$$-56.23 + 0.734 \cdot 170 = 68.5 \text{ kg.}$$

Par contre, il sera dangereux d'utiliser cette équation pour estimer le poids moyen d'individus mesurant 200 *cm ou, pire, pour estimer le poids moyen de*

bébés mesurant 50 *cm. Dans ce dernier cas, notre équation nous donnerait un poids négatif de :*

$$-56.23 + 0.734 \cdot 50 = -19.5 \ kg$$

(comme on le voit sur le graphique du bas de la figure 13.4). Ce résultat absurde démontre qu'une hypothèse de linéarité qui peut apparaître raisonnable à l'intérieur du domaine d'observation n'est parfois plus du tout raisonnable en dehors du domaine d'observation.

13.5 Interprétation des paramètres de la droite de régression

L'hypothèse de linéarité permet non seulement de donner une interprétation à la droite de régression, elle permet également de donner une interprétation aux paramètres β_0 et β_1 de cette droite comme suit :

- **pour la constante**, on a :

  $$\beta_0 = mean(Y|X = 0)$$

  → la constante est la moyenne de Y dans le groupe particulier $X = 0$

  → cette interprétation est intéressante seulement si la valeur particulière $X = 0$ est intéressante

  → en pratique, la constante ne sera pas toujours intéressante, la valeur $X = 0$ étant parfois loin en dehors du domaine d'observation

- **pour la pente**, on a pour n'importe quelle valeur de x :

  $$mean(Y|X = x + 1) - mean(Y|X = x) = \beta_0 + \beta_1(x + 1) - (\beta_0 + \beta_1 x)$$

  et ainsi :

  $$\beta_1 = mean(Y|X = x + 1) - mean(Y|X = x)$$

  → la pente est la différence de moyenne de Y entre deux groupes différant d'une unité par rapport à X

  → la pente d'une droite de régression est ainsi une généralisation du concept de différence de moyenne

  → au contraire de la constante, la pente sera toujours intéressante

  → plus généralement, on aura :

  $$k\beta_1 = mean(Y|X = x + k) - mean(Y|X = x)$$

  → en multipliant par k la pente d'une droite de régression, on obtient la différence de moyenne de Y entre deux groupes différant de k unités par rapport à X

→ la pente de la droite de régression nous aidera à répondre à la seconde question de la régression énoncée en début de chapitre (comment se modifie Y en fonction de X).

Sous l'hypothèse de linéarité, la pente d'une droite de régression représente l'augmentation moyenne de Y lorsque X augmente d'une unité.

Exemple 13.6 *On reprend notre exemple de la relation entre la taille et le poids, où l'on avait estimé l'équation de régression* $\widehat{y}^* = -56.23 + 0.734x$. *Sous l'hypothèse de linéarité, on a les interprétations suivantes :*

- *contante :* $\widehat{\beta}_0 = -56.23$

  → *le poids moyen d'un individu mesurant* 0 *cm est estimé à* -56.23 *kg*

  → *exemple d'interprétation absurde (la valeur* $X = 0$ *n'étant pas une valeur intéressante, ni même sensée pour* X*)*

  → *la constante n'est pas interprétable, mais demeure indispensable pour l'estimation des poids moyens des différents groupes*

- *pente :* $\widehat{\beta}_1 = 0.734$

  → *la différence de moyenne du poids entre deux groupes différant de* 1 *cm en taille est estimée à* 0.734 *kg*

  → *la différence de moyenne du poids entre deux groupes différant de* 10 *cm en taille est estimée à* 7.34 *kg*

  → *on observe une augmentation de* 0.734 *kg par cm.*

On notera que l'interprétation des paramètres se modifie avec un changement d'unités des variables. On peut ainsi rendre la constante intéressante en rendant la valeur $X = 0$ intéressante. Si par exemple une taille de 160 cm est considérée intéressante, on peut soustraire 160 cm à toutes les tailles. De cette manière la valeur $X = 0$ représente le groupe des 160 cm et la constante β_0 sera la moyenne des poids dans ce groupe. On peut alternativement standardiser une variable X en lui soustrayant d'abord sa moyenne et en la divisant ensuite par son écart type, de telle sorte que la valeur $X = 0$ représente la moyenne de cette variable et qu'une différence d'une unité entre deux valeurs de X corresponde à une différence d'un écart type.

Exemple 13.7 *La moyenne des tailles de notre échantillon est de* 162.4 *cm et l'écart type est de* 5.98 *cm. Si on soustrait aux tailles originales* 162.4*, si on les divise ensuite par* 5.98 *(obtenant par exemple pour un individu mesurant* 160

cm la valeur $(160 - 162.4)/5.98 = -0.40$*, ce qui veut dire que cet individu se trouve à* 0.40 *écart type en dessous de la moyenne) et si on recalcule l'équation de régression, on trouve :*

$$\widehat{y}^* = 62.96 + 4.39x.$$

L'interprétation est ici la suivante :

- *constante :* $\widehat{\beta}_0 = 62.96$

  $\rightarrow$ *le poids moyen d'un individu de taille moyenne (c'est-à-dire mesurant* 162.4 *cm) est estimé à* 62.96 *kg*

  $\rightarrow$ *la constante est ici intéressante*

  $\rightarrow$ *on notera que cette valeur de* 62.96 *kg est égale à la moyenne des poids* $\bar{y}$ *de notre échantillon, la droite de régression passant par* $(\bar{x}, \bar{y})$

- *pente :* $\widehat{\beta}_1 = 4.39$

  $\rightarrow$ *la différence de moyenne du poids entre deux groupes différant d'un écart type en taille (c'est-à-dire de* 5.98 *cm) est estimée à* 4.39 *kg*

  $\rightarrow$ *on notera que cette valeur de* 4.39 *kg représente* 5.98 *fois la pente calculée sans standardisation de la taille (qui était de* 0.734*).*

Comme la corrélation ρ, la pente de la droite de régression β_1 est donc un résumé de l'association statistique entre X et Y. Bien que partageant le même signe, ρ et β_1 nous donnent cependant une information différente. La corrélation ρ est un nombre entre -1 et $+1$ qui n'a pas d'unité. La pente de la droite de régression β_1 est exprimée dans les unités de Y (dans notre exemple, en kg/cm). La première nous informe si l'association est plus ou moins proche d'une relation exacte linéaire. La seconde nous informe sur l'importance de la différence de moyenne entre deux groupes différant d'une unité par rapport à X. La figure 13.5 nous montre différentes situations possibles. On a une grosse pente dans les graphiques (a) et (b) et une petite pente dans les graphiques (c) et (d). On a une grosse corrélation dans les graphiques (a) et (c) et une petite corrélation dans les graphiques (b) et (d). Au sens de la corrélation, l'association statistique entre X et Y est donc plus importante dans le graphique (c) que dans le graphique (b). Au sens de la pente de la droite de régression, l'association statistique entre X et Y est par contre plus importante dans le graphique (b) que dans le graphique (c)[7].

Comme on l'avait fait pour la corrélation, on terminera cette section en mentionnant que le calcul d'une droite de régression nous permet de décrire/résumer

[7] Une autre différence importante entre la corrélation et la pente d'une droite de régression est que l'interprétation d'une corrélation aura surtout du sens dans le cadre d'une distribution (X, Y) normale bivariée, alors que l'interprétation d'une droite de régression pourra également se faire dans le cadre d'une étude expérimentale, où l'on pourra choisir nous-mêmes les valeurs de la variable X pour les individus de notre échantillon, et où on s'éloignera parfois de la binormalité (car on aura alors de la variabilité artificielle plutôt que naturelle).

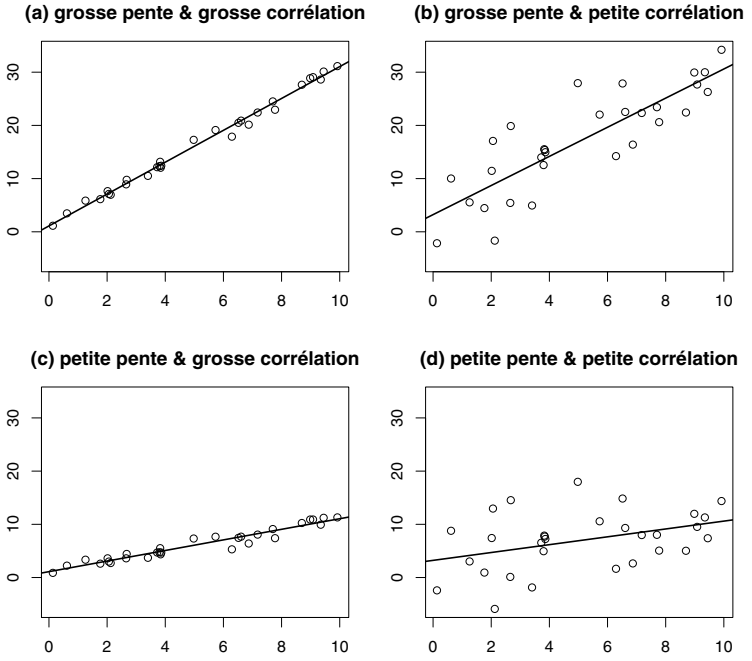

Figure 13.5 – Corrélation *versus* pente de la droite de régression.

l'association statistique entre deux variables X et Y, mais ne nous permet pas de conclure à une relation de cause à effet entre les variables. Si par exemple X représente la dose d'un médicament et Y le rétablissement des patients, une droite de régression avec une grosse pente nous indique que les patients traités avec une forte dose se rétablissent mieux en moyenne que ceux traités avec une faible dose, mais cela n'implique pas forcément que ces différences de rétablissement soient la conséquence des différentes doses utilisées. En effet, les patients avec une forte dose étaient peut-être aussi plus jeunes, plus motivés et/ou traités dans un meilleur hôpital que les patients avec une faible dose. Dans ce cas, (une partie de) l'association observée entre X et Y pourrait être due à ces variables confondantes.

13.6 Modèle de régression linéaire simple

Nous venons de discuter en détail de l'hypothèse de linéarité qui nous permet de donner une interprétation descriptive à la droite de régression, ainsi qu'aux paramètres de cette droite. Dans cette section, nous allons faire des hypothèses supplémentaires qui nous permettront de faciliter l'inférence sur la droite de régression. Ces hypothèses constituent le *modèle de régression linéaire simple*.

L'hypothèse de linéarité concerne la moyenne de la variable Y dans les groupes définis par les différentes valeurs possibles x de X (on suppose que les moyennes des différents groupes sont alignées sur la droite de régression $\beta_0 + \beta_1 x$). La deuxième hypothèse de notre modèle concerne la variance de la variable Y dans ces groupes : on supposera que la variance est la même dans chaque groupe. Si tel est le cas, cela veut dire que la qualité de la prédiction sera identique dans chaque groupe. Cette variance commune sera alors égale à la variance résiduelle, que l'on notera σ^2_ε. Il s'agit de l'*hypothèse d'homoscédasticité* (ou de la « variance constante »). La troisième hypothèse concerne la forme de la distribution de la variable Y dans les différents groupes : on supposera cette distribution normale dans chaque groupe. Ces trois hypothèses spécifient complètement la distribution de la variable Y dans les différents groupes, qui sera donc normale de moyenne $\beta_0 + \beta_1 x$ et de variance σ^2_ε. On a ainsi :

linéarité	:	$mean(Y\|X = x) = \beta_0 + \beta_1 x$
homoscédasticité	:	$variance(Y\|X = x) = \sigma^2_\varepsilon$ (ne dépend pas de x)
normalité	:	la variable $Y\|X = x$ est normale (quel que soit x).

Ces hypothèses peuvent se vérifier graphiquement. La figure 13.6 nous montre quelques exemples. Dans le graphique (a), l'hypothèse de linéarité n'est pas du tout satisfaite (la droite ne passe pas au milieu des points localement). Dans le graphique (b), l'hypothèse de linéarité est (approximativement) satisfaite mais l'hypothèse d'homoscédasticité ne l'est pas : la variabilité des groupes définis par de petites valeurs x est nettement plus petite que la variabilité des groupes définis par de grandes valeurs x (il s'agit d'un cas d'*hétéroscédasticité* : la variance n'est pas constante, mais augmente ici avec x). Dans le graphique (c), la linéarité et l'homoscédasticité sont (approximativement) satisfaites, mais pas la normalité. Dans chaque groupe, la distribution apparaît en effet très asymétrique. Les trois hypothèses sont par contre (approximativement) satisfaites dans le graphique (d).

Lorsque l'une ou l'autre de ces hypothèses ne sera pas satisfaite, on pourra essayer d'améliorer la situation en transformant les variables (X et/ou Y). Notons par ailleurs que ces trois hypothèses seront satisfaites si la variable (X, Y) est normale bivariée. Comme il n'est pas rare d'observer des distributions normales dans la nature, ce modèle linéaire sera souvent utile en pratique. En statistique, normalité rime donc avec linéarité (et avec homoscédasticité). En particulier, les relations entre les variables sont forcément linéaires dans un cadre binormal[8].

[8] Attention, la réciproque n'est pas vraie. Alors que la binormalité implique un modèle de régression linéaire, on peut avoir un modèle de régression linéaire sans avoir nécessairement la binormalité, par exemple dans le cadre d'une étude expérimentale.

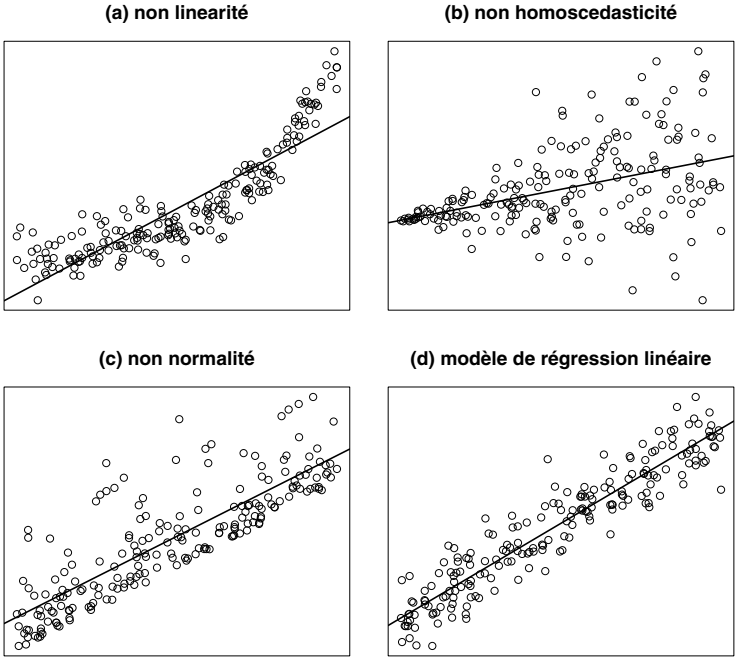

Figure 13.6 – Hypothèses du modèle de régression linéaire simple.

On va voir à présent que l'on peut reformuler les hypothèses d'un modèle de régression simple en fonction des résidus (ce qui facilitera la généralisation de ces hypothèses au cas de la régression multiple dans le chapitre suivant). On a vu que l'hypothèse de linéarité peut être définie de façon équivalente par :

$$mean(Y|X = x) = \beta_0 + \beta_1 x \quad \text{ou par} \quad mean(\varepsilon|X = x) = 0.$$

Par ailleurs, la variable résiduelle dans le groupe $X = x$, que nous notons par $\varepsilon|X = x$, peut s'obtenir à partir de la variable Y dans le groupe $X = x$, que nous notons par $Y|X = x$, par simple soustraction de sa moyenne $mean(Y|X = x)$, de sorte que la variabilité et la forme de ces deux distributions sont identiques. Il s'ensuit que l'hypothèse d'homoscédasticité peut être définie de façon équivalente par :

$$variance(Y|X = x) = \sigma_\varepsilon^2 \quad \text{ou par} \quad variance(\varepsilon|X = x) = \sigma_\varepsilon^2.$$

De même, l'hypothèse de normalité revient de manière équivalente à supposer que la variable $Y|X = x$ est normale ou à supposer que la variable $\varepsilon|X = x$ est normale. Ainsi, les trois hypothèses ci-dessus reviennent à supposer que les résidus $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ de notre échantillon ($i = 1, \cdots, n$) proviennent d'une même distribution, qui est normale avec moyenne nulle et variance σ_ε^2.

On ajoute alors une quatrième hypothèse à ce modèle de régression qui consiste à supposer que ces n résidus sont indépendants les uns des autres[9]. Les quatre hypothèses d'un modèle de régression linéaire simple, que l'on retrouvera en régression multiple, sont donc :

linéarité	:	$mean(\varepsilon_i) = 0$
homoscédasticité	:	$variance(\varepsilon_i) = \sigma_\varepsilon^2$
normalité	:	ε_i normalement distribués
indépendance	:	ε_i indépendants.

13.7 Inférence sur la droite de régression

Nous allons à présent voir comment faire de l'inférence sur la droite de régression[10]. Sous les hypothèses d'un modèle de régression linéaire simple (discutées en détail dans la section précédente), on peut montrer que $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont

[9]Dans certains ouvrages, cette quatrième hypothèse vient en troisième position, la normalité venant en quatrième. Nous l'avons mis en quatrième car contrairement aux trois premières hypothèses, qui concernent la nature du phénomène étudié, cette quatrième hypothèse concerne l'échantillonnage. Notons que ce ne sont pas les observations Y_i de la variable Y qui doivent être indépendantes, mais bel et bien les erreurs de prédiction ε_i. Pour rendre clair cette distinction, notons qu'en régression (contrairement à ce que l'on a vu en corrélation) on pourra choisir les valeurs x_i de la variable X comme il nous plaira (on pourra par exemple sur-représenter les grands et sous-représenter les petits, si cela nous chante). Ainsi, si on décide que nos deux premiers individus (sur les n de notre échantillon) doivent mesurer $x_1 = x_2 = 170$ cm, il s'agira pour ces deux premières observations de tirer au hasard deux individus parmi la population des individus mesurant 170 cm. Leurs poids Y_1 et Y_2 ne seront certes pas indépendants. En effet, la connaissance du poids du premier individu nous informera sur le poids du second (du moins s'il y a une certaine association entre les variables X et Y, de sorte que les poids de deux individus de même taille auront tendance à se ressembler). Par contre, leurs résidus $\varepsilon_1 = y_1 - \beta_0 - \beta_1 x_1$ et $\varepsilon_2 = y_2 - \beta_0 - \beta_1 x_2$ seront indépendants. En effet, la connaissance du résidu du premier individu ne nous informera en rien sur le résidu du second : savoir par exemple que le poids du premier individu se trouve au-dessus de la moyenne de son groupe ne nous donnera aucune indication sur la position du poids du second individu par rapport à la moyenne de son groupe.

[10]Rappelons ici le paradigme de la statistique, qui consiste à considérer un estimateur comme étant une variable en s'imaginant que l'on répète l'échantillonnage. Lorsque les données sont bivariées, il s'agit de préciser comment on s'imagine cette répétition de l'échantillonnage. Il y a au moins deux approches possibles :

- on peut s'imaginer avoir dans chaque échantillon n réalisations indépendantes (X_i, Y_i) de la variable bivariée (X, Y)

  → les X_i et les Y_i seront différents dans chaque échantillon

- on peut s'imaginer avoir n réalisations indépendantes Y_i des variables « conditionnelles »

des estimateurs sans biais et normalement distribués des véritables paramètres β_0 et β_1. Il s'ensuit que $\widehat{\beta}_0 + \widehat{\beta}_1 x$ est un estimateur sans biais et normalement distribué de $\beta_0 + \beta_1 x$ (la véritable moyenne de Y dans le groupe $X = x$). Pour $\widehat{\beta}_0$ et $\widehat{\beta}_1$, on a en outre les variances suivantes :

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{\sigma_\varepsilon^2}{n} \cdot \frac{\sum_i x_i^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{n \cdot \widehat{\sigma}_X^2} \cdot \frac{\sum_i x_i^2}{n}$$

et

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{n \cdot \widehat{\sigma}_X^2}.$$

Il est toujours instructif de regarder de plus près la formule de la variance d'un estimateur (qui on le rappelle nous informe sur la précision de cet estimateur) :

- ces variances diminuent lorsque la taille n de l'échantillon augmente

  → comme d'habitude, les erreurs types (c'est-à-dire les racines carrées de ces variances) sont inversement proportionnelles à $\sqrt{n}$

- ces variances diminuent lorsque σ_ε^2 diminue, c'est-à-dire lorsque la relation est proche d'être exacte linéaire

  → logique : il est plus facile d'estimer précisément une droite lorsque les données sont proches de cette droite

- ces variances diminuent lorsque la variance des x_i augmente

  → logique : il est plus facile d'estimer précisément une droite lorsque le domaine d'observation est large

- la variance de $\widehat{\beta}_0$ diminue lorsque la moyenne des carrés des x_i diminue, c'est-à-dire lorsque les x_i se trouvent proche (aux alentours) de 0

  → logique : il est plus facile d'estimer précisément la moyenne de Y dans le groupe $X = 0$ lorsque ce groupe se trouve au centre du domaine d'observation.

$Y|X = x_i$ $(i = 1, \cdots, n)$

→ seuls les Y_i seront différents d'un échantillon à l'autre, les x_i demeurant les mêmes dans chaque échantillon

→ si par exemple le premier individu mesure 170 cm dans notre échantillon, on s'imagine que le premier individu mesurera de même 170 cm dans tous les échantillons.

Alors que la première approche est en général considérée pour faire de l'inférence sur une corrélation, on adopte souvent la seconde approche pour faire de l'inférence dans le cadre d'un modèle de régression. Cette seconde approche serait naturelle si on choisissait effectivement les valeurs x_i de la variable X dans notre échantillon initial (comme on le fait dans une étude expérimentale). Dans les cas où l'on observe les X_i sans les avoir choisis, on fera avec cette seconde approche de l'inférence *conditionnellement* aux valeurs x_i observées dans notre échantillon. Cela aura le mérite de simplifier les mathématiques sans forcément changer fondamentalement les résultats par rapport à ce que l'on obtiendrait en considérant la plus difficile première approche, comme expliqué par exemple dans l'article de Sampson (1974).

De la même manière que l'on a défini la variance d'un estimateur comme étant la variance calculée sur les observations de cet estimateur que l'on obtiendrait si on répétait l'expérience, on peut définir la covariance entre deux estimateurs comme étant la covariance calculée sur les observations bivariées de ces deux estimateurs que l'on obtiendrait si on répétait cette même expérience. Ainsi par exemple, la covariance entre les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sera donnée par :

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-\bar{x}\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}.$$

Cela nous permet de calculer la variance de $\widehat{\beta}_0 + \widehat{\beta}_1 x$ comme suit :

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x) &= \mathrm{Var}(\widehat{\beta}_0) + x^2 \mathrm{Var}(\widehat{\beta}_1) + 2x\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\
&= \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{\sum_i x_i^2}{n} + x^2 - 2x\bar{x} \right) \\
&= \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{\sum_i x_i^2}{n} - \bar{x}^2 + \bar{x}^2 + x^2 - 2x\bar{x} \right) \\
&= \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \left(\frac{\sum_i (x_i - \bar{x})^2}{n} + (x - \bar{x})^2 \right) \\
&= \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right).
\end{aligned}$$

Ici également, on constate que la variance diminue (et donc la précision de l'estimateur augmente) lorsque n augmente, lorsque σ_ε^2 diminue, lorsque la variance des x_i augmente et lorsque le groupe $X = x$ se trouve proche du centre $\bar{x}$ du domaine d'observation.

Nos estimateurs étant sans biais et normalement distribués, on pourra utiliser la méthode de Wald pour faire de l'inférence. Les variances de ces estimateurs dépendent cependant de la variance résiduelle σ_ε^2 qui est inconnue et qu'il s'agira d'estimer. Si on disposait de n observations indépendantes $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ de la variable résiduelle ε, on pourrait estimer σ_ε^2 sans biais en calculant $\sum_i (\varepsilon_i - \bar{\varepsilon})^2/(n-1)$ avec $\bar{\varepsilon} = \sum_i \varepsilon_i / n$. Or, on a vu que les résidus ε_i ne sont pas observés mais estimés par $\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$. Il se trouve alors que l'estimateur donné par[11] :

$$\tilde{\sigma}_\varepsilon^2 = \frac{\sum_i \widehat{\varepsilon}_i^2}{n-2}$$

est un estimateur sans biais de σ_ε^2. On notera également la formule suivante :

$$\sum_i \widehat{\varepsilon}_i^2 = \sum_i (y_i - \bar{y})^2 - \widehat{\beta}_1^2 \sum_i (x_i - \bar{x})^2$$

[11] On rappelle que la moyenne des $\widehat{\varepsilon}_i$ est nulle, de sorte que $\tilde{\sigma}_\varepsilon^2$ représente la variance des $\widehat{\varepsilon}_i$ calculée non pas avec un dénominateur n, ni $n-1$, mais avec un dénominateur $n-2$. Il faudra donc un minimum de $n = 3$ observations pour estimer la variance résiduelle (logique : avec seulement $n = 2$ observations, la droite de régression passerait par les deux observations et on ne pourrait pas estimer de variance autour de cette droite).

qui nous permet de calculer $\tilde{\sigma}_\varepsilon^2$ à partir des moyennes habituelles (on n'a pas besoin du détail des données, mais seulement de connaître $\frac{1}{n}\sum_i x_i$, $\frac{1}{n}\sum_i y_i$, $\frac{1}{n}\sum_i x_i^2$, $\frac{1}{n}\sum_i x_i y_i$ et $\frac{1}{n}\sum_i y_i^2$).

Afin de faire de l'inférence sur les paramètres β_0, β_1 et $\beta_0 + \beta_1 x$ en utilisant la méthode de Wald, on remplacera la variance résiduelle σ_ε^2 par son estimateur $\tilde{\sigma}_\varepsilon^2$ dans les formules des variances des estimateurs $\widehat{\beta}_0$, $\widehat{\beta}_1$ et $\widehat{\beta}_0 + \widehat{\beta}_1 x$ données ci-dessus. Ici aussi, cependant, ce remplacement ne sera pas totalement anodin et nous amènera à utiliser des quantiles d'une distribution de Student (avec $n-2$ *dl*) plutôt que des quantiles d'une distribution normale standardisée dans le calcul des intervalles de confiance pour ces paramètres, ou pour les tests statistiques correspondants (voir ci-dessous). Toujours sous les hypothèses de notre modèle, ces intervalles de confiance et ces tests statistiques seront exacts, y compris pour de petits échantillons à partir de $n \geq 3$.

Sous les hypothèses du modèle, on pourra donc calculer les intervalles de confiance de Student suivants pour les paramètres d'une régression[12] :

- intervalle de confiance au niveau $1-\alpha$ pour β_0 :

$$\widehat{\beta}_0 \pm t_{1-\alpha/2,n-2} \cdot \sqrt{\frac{\tilde{\sigma}_\varepsilon^2}{n} \cdot \frac{\sum_i x_i^2}{\sum_i (x_i - \bar{x})^2}}$$

- intervalle de confiance au niveau $1-\alpha$ pour β_1 :

$$\widehat{\beta}_1 \pm t_{1-\alpha/2,n-2} \cdot \sqrt{\frac{\tilde{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}}$$

- intervalle de confiance au niveau $1-\alpha$ pour $\beta_0 + \beta_1 x$:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{1-\alpha/2,n-2} \cdot \sqrt{\tilde{\sigma}_\varepsilon^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)}.$$

Le troisième de ces intervalles de confiance peut se calculer pour chaque valeur possible x. La figure 13.7 nous montre ces intervalles calculés avec les données de notre exemple. On y voit la borne inférieure et la borne supérieure d'un intervalle de confiance au niveau 95 % pour $\beta_0 + \beta_1 x$ en fonction de x. On voit

[12] Si les hypothèses du modèle ne sont pas satisfaites, la validité de l'inférence n'est pas garantie. On peut examiner à l'aide de simulations quelles sont les conséquences des violations de ces hypothèses sur la validité de l'inférence. Certains commentaires faits dans les chapitres précédents restent ici valables, par exemple le fait que l'on perde de la puissance statistique lorsque l'on s'éloigne de la normalité, notamment en présence de valeurs extrêmes ou aberrantes pour les ε_i. Une violation de l'hypothèse d'homoscédasticité risque d'être plus problématique, surtout si les ε_i avec une forte variabilité sont moins représentés dans notre échantillon que les ε_i avec une faible variabilité (de la même manière, on a vu qu'un test de Student pour une différence de moyenne sera libéral si c'est le petit groupe qui varie le plus). L'hypothèse d'indépendance est indispensable (sans quoi l'inférence pourrait être libérale, même en cas d'une légère violation) alors qu'une violation grave de l'hypothèse de linéarité rendra la droite de régression moins intéressante (et l'inférence moins indispensable).

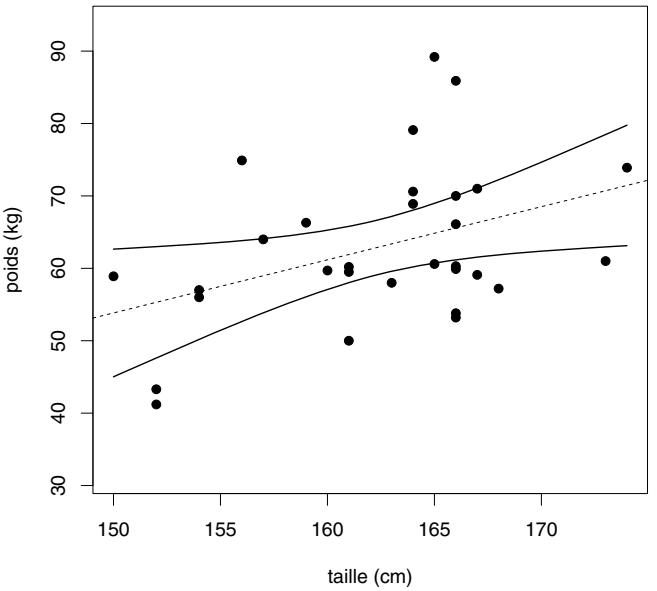

Figure 13.7 – Intervalles de confiance pour les moyennes des différents groupes.

que ces intervalles construits autour de la droite de régression (représentée par un trait pointillé) deviennent plus larges, et donc que l'incertitude à propos de la moyenne du groupe $X = x$ augmente, lorsque x s'éloigne du centre des données.

On appelle tests de Student (ou tests t) les tests en dualité avec ces intervalles de confiance. On testera par exemple l'hypothèse nulle :

$$H_0 : \beta_1 = 0$$

en calculant la statistique de test donnée par :

$$t_{stat} = \frac{\widehat{\beta}_1}{\sqrt{\frac{\tilde{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}}}$$

et en la comparant avec les quantiles d'une distribution de Student avec $n-2$ *dl* (on rejettera ainsi H_0 dans un test bilatéral au seuil α si $|t_{stat}| \geq t_{1-\alpha/2,n-2}$)[13].

Exemple 13.8 *Reprenons depuis le début notre exemple de la relation entre la taille et le poids, que l'on étudie à partir d'un échantillon de* $n = 30$ *individus,*

[13]Le lecteur vérifiera que cette statistique de test est égale à celle utilisée au chapitre 12 pour tester la nullité de la corrélation, de sorte qu'il sera équivalent de tester $H_0 : \beta_1 = 0$ ou de tester $H_0 : \rho = 0$ (ce qui n'est pas surprenant, sachant que la nullité de la pente de la droite de régression est équivalente à la nullité de la corrélation).

où les hypothèses d'un modèle de régression linéaire semblent raisonnables. On va voir comment tout peut être calculé à partir des cinq moyennes suivantes : $\frac{1}{n}\sum_i x_i = \bar{x} = \widehat{\mu}_X = 162.43$, $\frac{1}{n}\sum_i y_i = \bar{y} = \widehat{\mu}_Y = 62.96$, $\frac{1}{n}\sum_i x_i^2 = 26\ 419.17$, $\frac{1}{n}\sum_i x_i y_i = 10\ 252.18$ *et* $\frac{1}{n}\sum_i y_i^2 = 4077.42$. *On peut calculer :*

$$\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2 = 30(26\ 419.17 - 162.43^2) = 1037.37$$

ainsi que :

$$\sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n\bar{y}^2 = 30(4077.42 - 62.96^2) = 3403.81$$

et estimer les variances de X *et* Y *par* $\tilde{\sigma}_X^2 = 1037.37/29 = 35.77$ *et* $\tilde{\sigma}_Y^2 = 3403.81/29 = 117.37$ *(ce qui implique les estimations des écarts types* $\tilde{\sigma}_X = \sqrt{35.77} = 5.98$ *et* $\tilde{\sigma}_Y = \sqrt{117.37} = 10.83$*). On peut aussi calculer :*

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y} = 30(10\ 252.18 - 162.43 \cdot 62.96) = 761.22$$

et estimer la corrélation par :

$$\widehat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}} = \frac{761.22}{\sqrt{1037.37 \cdot 3403.81}} = 0.405.$$

Le pourcentage de la variance de Y *prédite linéairement par* X *est donc estimé par* $\widehat{\rho}^2 = 0.405^2 = 16.4$ %. *On avait également calculé :*

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{761.22}{1037.37} = 0.734$$

ainsi que :

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \cdot \bar{x} = 62.96 - 0.734 \cdot 162.43 = -56.23$$

obtenant l'estimation de la droite de régression :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x = -56.23 + 0.734x.$$

On peut à présent estimer la variance résiduelle :

$$\tilde{\sigma}_\varepsilon^2 = \frac{\sum_i (y_i - \bar{y})^2 - \widehat{\beta}_1^2 \cdot \sum_i (x_i - \bar{x})^2}{n-2} = \frac{3403.81 - 0.734^2 \cdot 1037.37}{28} = 101.62$$

(ce qui implique l'estimation de l'écart type résiduel $\tilde{\sigma}_\varepsilon = \sqrt{101.62} = 10.08$*). Si on veut faire de l'inférence sur* β_1*, on estime son erreur type par :*

$$\frac{\tilde{\sigma}_\varepsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{10.08}{\sqrt{1037.37}} = 0.313$$

et on calcule un intervalle de confiance au niveau 95 % *pour* β_1 *par (en utilisant* $t_{0.975,28} = 2.05$*) :*

$$\widehat{\beta}_1 \pm t_{0.975,28} \frac{\tilde{\sigma}_\varepsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}} = 0.734 \pm 2.05 \cdot 0.313 = [0.093; 1.375].$$

Si on veut tester l'hypothèse nulle $H_0 : \beta_1 = 0$ *avec un test de Student, on calcule la statistique de test suivante :*

$$t_{stat} = \frac{\widehat{\beta}_1}{\frac{\tilde{\sigma}_\varepsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}}} = \frac{0.734}{0.313} = 2.34.$$

Comme on a $|t_{stat}| = 2.34 > 2.05 = t_{0.975,28}$*, on rejette* H_0 *dans un test bilatéral au seuil de* 5 % *et on conclut* $\beta_1 > 0$ *(ce qui est compatible avec le fait que la valeur* 0 *n'appartient pas à l'intervalle de confiance ci-dessus). Un logiciel statistique nous donnera ici* $p = 0.026$*. Notons que l'on obtient la même statistique de test si on teste l'hypothèse nulle équivalente* $H_0 : \rho = 0$ *définie par :*

$$t_{stat} = \frac{\sqrt{n-2} \cdot \widehat{\rho}}{\sqrt{1 - \widehat{\rho}^2}} = \frac{\sqrt{28} \cdot 0.405}{\sqrt{1 - 0.405^2}} = 2.34$$

d'où l'on peut conclure $\rho > 0$ *(avec également* $p = 0.026$*). Le lecteur pourra par ailleurs calculer un intervalle de confiance au niveau* 95 % *pour* β_0 *donné par :*

$$-53.26 \pm 2.05 \cdot 50.87 = [-160.44; 47.97]$$

alors que des intervalles de confiance au niveau 95 % *pour* $\beta_0 + \beta_1 x$ *pour différentes valeurs de* x *sont montrés dans la figure 13.7.*

13.8 Intervalle de prédiction

On a introduit la régression comme une méthode nous permettant d'effectuer des prédictions. On a considéré jusqu'ici uniquement des prédictions ponctuelles. Sous les hypothèses d'un modèle de régression linéaire, il sera possible de calculer en outre des *intervalles de prédiction*.

Le concept d'intervalle de prédiction a été introduit au chapitre 4. À partir de l'estimation $\widehat{\mu}_Y$ de la moyenne μ_Y et de l'estimation $\tilde{\sigma}^2_Y$ de la variance σ^2_Y d'une variable Y, on calcule l'intervalle de prédiction au niveau 95 % suivant :

$$\widehat{\mu}_Y \pm 2 \cdot \tilde{\sigma}_Y.$$

Sous l'hypothèse de la normalité de Y, cet intervalle contient (approximativement) 95 % des observations. Si on considère un individu de la population dont on aimerait prédire la valeur de Y, il y aura ainsi une probabilité (approximativement) de 0.95 pour que cette valeur soit contenue dans cet intervalle. Si on connaît par ailleurs la valeur x d'une caractéristique X de cet individu, et

s'il existe une association statistique entre X et Y, on pourra alors ajuster cet intervalle de prédiction en utilisant notre modèle de régression. Sous les hypothèses du modèle, la variable $Y|X = x$ sera en effet normale avec une moyenne $\beta_0 + \beta_1 x$, estimée par $\widehat{\beta}_0 + \widehat{\beta}_1 x$, et avec une variance σ_ε^2, estimée par $\tilde{\sigma}_\varepsilon^2$. On calculera ainsi l'intervalle de prédiction ajusté suivant :

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm 2 \cdot \tilde{\sigma}_\varepsilon.$$

Si les hypothèses du modèle sont correctes, cet intervalle contiendra la valeur Y de cet individu avec une probabilité (approximativement) de 0.95.

On peut ainsi calculer deux intervalles de prédiction pour la valeur Y de cet individu. En principe, le second intervalle sera cependant plus court que le premier et en ce sens la prédiction sera plus précise. La longueur du premier intervalle sera égale (approximativement) à $4\sigma_Y$, alors que la longueur du second sera égale (approximativement) à $4\sigma_\varepsilon$. On peut quantifier cette réduction de longueur obtenue en utilisant le second intervalle de prédiction plutôt que le premier en l'exprimant en pourcentage de la longueur du premier comme suit (que nous noterons ici PIR pour `prediction interval reduction`) :

$$PIR = \frac{4\sigma_Y - 4\sigma_\varepsilon}{4\sigma_Y} = 1 - \frac{\sigma_\varepsilon}{\sigma_Y}.$$

Cette formule n'est pas sans rappeler celle du carré de la corrélation :

$$\rho^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_Y^2}.$$

On peut calculer ainsi :

$$PIR = 1 - \sqrt{1 - \rho^2}.$$

On a vu que le carré de la corrélation est une mesure de l'intensité de l'association entre X et Y que l'on peut interpréter comme le pourcentage de la variance de Y prédite linéairement par X. Cette quantité PIR représente une mesure alternative de l'intensité de cette association que l'on peut interpréter comme la réduction de longueur de l'intervalle de prédiction lorsque l'on utilise X pour prédire Y, comparé à une situation sans prédicteur.

Par exemple, une corrélation de $\rho = 0.5$ correspond à un pourcentage de variance prédite de 25 % et à une réduction de longueur de l'intervalle de prédiction de seulement $PIR = 1 - \sqrt{1 - 0.5^2} = 13\,\%$. En ce sens, un prédicteur dont la corrélation avec la variable réponse est de 0.5 ne nous sera pas d'une immense utilité pour améliorer nos prédictions. Afin d'obtenir une réduction de $PIR = 50\,\%$ (c'est-à-dire de diminuer par deux la longueur de nos intervalles de prédiction), il s'agira d'avoir à disposition un prédicteur X dont la corrélation avec Y est au moins de $\rho = \pm 0.87$, ce qui n'est pas si courant. Pour atteindre $PIR = 75\,\%$, la corrélation doit être de $\rho = \pm 0.97$, ce qui est encore moins courant. Cela illustre toute la difficulté d'obtenir des prédictions précises en pratique. On essaiera d'améliorer un peu la situation au chapitre suivant avec la régression multiple, mais la tâche restera difficile.

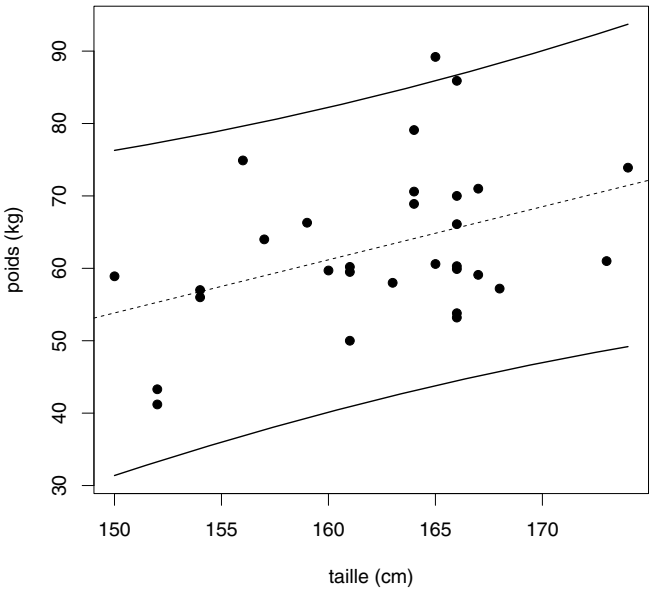

Figure 13.8 – Intervalles de prédiction pour les individus des différents groupes.

Exemple 13.9 *À partir de notre échantillon de* $n = 30$ *individus, on a estimé un poids moyen de* $\widehat{\mu}_Y = 62.96$ *kg et un écart type de* $\tilde{\sigma}_Y = 10.83$ *kg. Si on doit prédire le poids d'un individu choisi au hasard dans la population, on calculera l'intervalle de prédiction suivant (en supposant la normalité des poids) :*

$$62.96 \pm 2 \cdot 10.83 = [41.3; 84.6] \ kg$$

et on dira qu'il y a une probabilité de 0.95 *pour que cet individu pèse entre* 41.3 *et* 84.6 *kg. Si on sait par contre que cet individu mesure* 170 *cm, on pourra utiliser notre modèle de régression pour améliorer la prédiction (réduire la longueur de l'intervalle de prédiction). On a estimé* $\widehat{\beta}_0 = -56.23$, $\widehat{\beta}_1 = 0.734$ *et* $\tilde{\sigma}_\varepsilon = 10.08$. *On pourra ainsi calculer l'intervalle de prédiction ajusté suivant :*

$$-56.23 + 0.734 \cdot 170 \pm 2 \cdot 10.08 = [48.4; 88.7] \ kg$$

et on dira qu'il y a une probabilité de 0.95 *pour que cet individu mesurant* 170 *cm pèse entre* 48.4 *et* 88.7 *kg. Cet intervalle est un peu plus court que le premier. Cependant, le gain n'est pas énorme puisqu'on estime* PIR *à* $1 - \tilde{\sigma}_\varepsilon/\tilde{\sigma}_Y = 1 - 10.08/10.83 = 7$ %. *Ainsi, connaître la taille de l'individu permet de réduire la longueur de l'intervalle de prédiction de seulement* 7 %.

Nous terminerons cette section en rappelant que le concept d'intervalle de prédiction (pour la valeur d'un individu) est très différent du concept d'intervalle de confiance pour une moyenne (définie sur une population d'individus).

La figure 13.8 nous montre les intervalles de prédiction pour le poids Y d'un individu avec une taille $X = x$ calculés dans le cadre de notre exemple ci-dessus pour différentes valeurs possibles de x. On constate que ces intervalles sont évidemment beaucoup plus larges que les intervalles de confiance pour les moyennes $\beta_0 + \beta_1 x$ des poids des individus avec $X = x$ que l'on avait montrés dans la figure 13.7. On constate également que ces intervalles contiennent 29 des 30 observations, ce qui représente effectivement approximativement 95 % des observations[14].

13.9 Régression vers la moyenne

On considère dans cette section un modèle de régression linéaire dans le cas particulier où X et Y ont une même distribution, que l'on supposera normale avec une moyenne μ et une variance σ^2, et où la corrélation entre X et Y est positive. Comme les deux variables ont même variance, la pente de la droite de régression coïncide avec la corrélation :

$$\beta_1 = \rho.$$

Ainsi, la pente de la droite de régression ne peut pas être dans ce cas plus grande que 1. On aura par ailleurs :

$$\beta_0 = (1 - \rho)\mu.$$

Si on suppose en plus que (X, Y) est normale bivariée, l'hypothèse de linéarité sera satisfaite et on aura ainsi :

$$mean(Y|X = x) = \beta_0 + \beta_1 x = (1 - \rho)\mu + \rho x.$$

La moyenne de la variable Y pour le groupe $X = x$ sera donc quelque part entre μ et x (elle sera plus proche de μ si $\rho < 0.5$, et plus proche de x si $\rho > 0.5$). Si $0 < \rho < 1$, on a alors les résultats suivants :

- $x > \mu$ implique $mean(Y|X = x) < x$
- $x < \mu$ implique $mean(Y|X = x) > x$.

[14] En fait, les intervalles de prédiction dans cette figure ont été calculés non pas avec la formule approximative donnée ci-dessus, mais avec la formule exacte donnée par :

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{0.975,n-2} \cdot \tilde{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

qui tient compte de l'incertitude due au fait que les moyennes et les variances ont été estimées (elles ne sont pas connues). De tels intervalles de prédiction deviennent un petit peu plus larges lorsque la valeur x s'éloigne du centre $\bar{x}$ des données, ceci moins nettement toutefois que pour les intervalles de confiance pour les moyennes $\beta_0 + \beta_1 x$. Nous reviendrons sur le concept d'intervalle de prédiction exact dans le chapitre suivant.

Considérons l'exemple où X représente la taille du père et Y la taille du fils. Une même distribution pour X et Y indique ici une stabilité de l'espèce humaine (en ce qui concerne la taille). Une corrélation positive indique que les fils des pères grands en taille seront en moyenne plus grands que les fils des pères petits en taille. Par contre, le résultat ci-dessus nous dit la chose suivante :

- les pères grands en taille auront en moyenne des fils plus petits qu'eux-mêmes
- les pères petits en taille auront en moyenne des fils plus grands qu'eux-mêmes.

> Les pères grands en taille auront des fils plus petits qu'eux-mêmes (bien que plus grands que les fils des pères petits en taille).

Ce phénomène appelé *régression vers la moyenne* est nécessaire à la stabilité d'une espèce : si les pères grands en taille avaient en moyenne des fils plus grands qu'eux-mêmes, et si les pères petits en taille avaient en moyenne des fils plus petits qu'eux-mêmes, on évoluerait vers une population constituée en majorité de géants et de nains[15].

Exemple 13.10 *La figure 13.9 nous montre un exemple de données simulées de la relation entre X la taille du père et Y la taille du fils pour $n = 500$ couples pères-fils, où la distribution de (X, Y) est binormale avec paramètres $\mu_X = \mu_Y = 175$ cm, $\sigma_X = \sigma_Y = 7$ cm et $\rho = 0.5$[16]. On observe dans cet exemple des moyennes de $\widehat{\mu}_X = 175.3$ cm et $\widehat{\mu}_Y = 175.2$ cm, des écarts types de $\tilde{\sigma}_X = 7.3$ cm et $\tilde{\sigma}_X = 7.0$ cm, ainsi qu'une corrélation de $\widehat{\rho} = 0.50$. La droite de régression est donnée par :*

$$\widehat{y}^* = 91.5 + 0.48x.$$

On estime ainsi que les fils des pères mesurant $x = 180$ cm mesureront (en moyenne) :

$$91.5 + 0.48 \cdot 180 = 177.9 \text{ cm}$$

et seront donc (en moyenne) plus petits que leurs pères. Par contre, on estime que les fils des pères mesurant $x = 170$ cm mesureront (en moyenne) :

$$91.5 + 0.48 \cdot 170 = 173.1 \text{ cm}$$

et seront donc plus grands que leurs pères.

[15] Ce phénomène de régression vers la moyenne, qui est donc à l'origine du terme de *régression* utilisé en statistique, a été observé vers la fin du XIX^e siècle par Francis Galton (cousin de Charles Darwin) lors d'une étude sur l'hérédité chez les petits pois.

[16] Pour simuler ces données, on a utilisé la fonction `mvrnorm` de la librairie `MASS` de R de la façon suivante : `mvrnorm(500,c(175,175),cbind(c(7*7,7*7*0.5),c(7*7*0.5,7*7)))`.

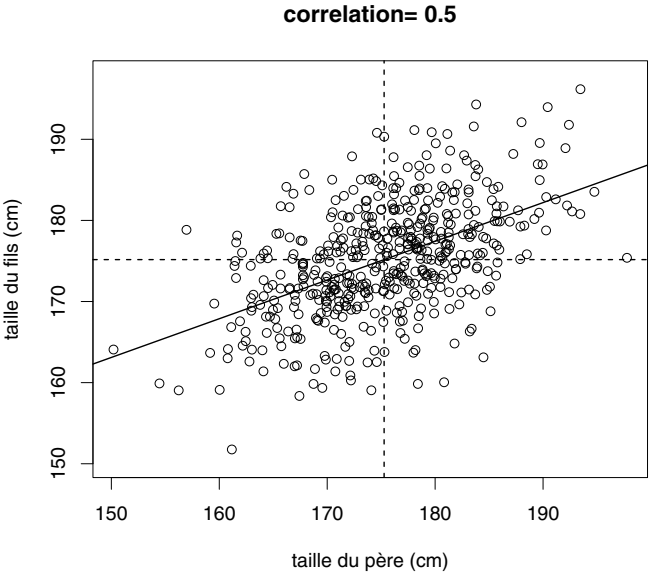

Figure 13.9 – Illustration de la régression vers la moyenne.

Un phénomène de régression vers la moyenne sera typiquement observé dans une situation de test-retest et sera parfois à l'origine de conclusions scientifiques erronées. Supposons par exemple que l'on forme deux groupes d'étudiants en fonction des résultats d'un premier test. Les étudiants qui ont bien réussi le premier test ont droit à un enseignement traditionnel et ceux qui ne l'ont pas bien réussi ont droit à un enseignement spécial. On effectue alors un second test quelques mois plus tard pour juger de la qualité de ces enseignements. Or, indépendamment de la qualité de l'enseignement reçu, les étudiants du second groupe auront tendance à mieux réussir, et les étudiants du premier groupe à moins bien réussir lors du second test que lors du premier test, ce qui pourrait donner à penser que l'enseignement spécial est supérieur à l'enseignement traditionnel, bien qu'un tel résultat pourrait être dû uniquement à la conséquence inévitable d'une régression vers la moyenne.

Chapitre 14

Régression linéaire multiple

La régression multiple est une généralisation de la régression simple. En régression multiple, on essaiera de prédire une variable réponse Y non pas à partir d'un seul prédicteur X, mais à partir de m prédicteurs, que l'on notera $X_1, X_2, \cdots, X_m$[1]. On aura donc des *données multivariées* avec $m+1$ variables mesurées pour chacun des n individus de notre échantillon, que l'on notera $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, où x_{ij} désigne l'observation du j-ième prédicteur et y_i l'observation de la variable réponse pour le i-ième individu (pour $i = 1, \cdots, n$ et $j = 1, \cdots, m$). On essaiera de répondre aux questions suivantes :

1. Comment peut-on prédire Y à partir de $X_1, X_2, \cdots, X_m$?

   → notre but sera d'améliorer la prédiction par rapport à ce que l'on avait en régression simple.

2. Comment se modifie Y en fonction de $X_1, X_2, \cdots, X_m$?

   → ce sera ici plus compliqué qu'en régression simple.

3. Quelles sont les prédicteurs importants de Y ?

   → il s'agira d'identifier parmi les prédicteurs disponibles un sous-ensemble de prédicteurs qui ne sont *ni inutiles, ni redondants* pour prédire Y

   → cette troisième question est nouvelle par rapport à ce que l'on avait en régression simple.

[1] Au lieu de *prédicteurs* et *variable réponse*, certains ouvrages utilisent les termes de *variables explicatives* et *variable expliquée*, ou encore de *variables indépendantes* et *variable dépendante*. Nous préférons pour notre part utiliser des termes se référant à une « prédiction » plutôt qu'à une « explication » car ce dernier peut faire penser à une relation de cause à effet entre les variables, ce qui ne sera pas forcément le cas. Le terme de variables indépendantes pour désigner les prédicteurs peut également prêter à confusion, car ceux-ci seront en général corrélés les uns avec les autres, et donc non indépendants au sens statistique du terme.

14.1 Hyperplan de régression

On répondra aux questions énoncées ci-dessus à l'aide de l'*hyperplan de régression*, qui est une généralisation de la droite de régression. L'hyperplan de régression est défini par une équation de la forme suivante :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_m x_m$$

que l'on utilisera pour prédire la valeur de la variable Y pour un individu avec $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_m = x_m$. Ici aussi, $\widehat{\beta}_0$ est appelée la constante, alors que $\widehat{\beta}_1, \widehat{\beta}_2, \cdots, \widehat{\beta}_m$ sont les pentes associées aux différents prédicteurs. Notons qu'un hyperplan sera un plan dans le cas $m = 2$ et une droite dans le cas $m = 1$. On utilise à nouveau la notation suivante :

- $\widehat{y}_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_m x_{im}$ est la valeur prédite de la variable Y pour le i-ième individu de notre échantillon
- $\widehat{\varepsilon}_i = y_i - \widehat{y}_i^*$ est son *erreur de prédiction*, aussi appelée *résidu*.

Comme au chapitre précédent, les coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ de l'hyperplan de régression sont calculés afin de minimiser le critère des moindres carrés :

$$\sum_i \widehat{\varepsilon}_i^2 = \sum_i (y_i - \widehat{y}_i^*)^2 = \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_m x_{im})^2.$$

La solution pour $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ est explicite dans une notation matricielle[2]. On retrouvera des propriétés semblables à celles vues en régression simple, avec notamment $\sum_i \widehat{\varepsilon}_i = 0$.

L'hyperplan de régression calculé dans notre échantillon :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_m x_m$$

est un estimateur du *véritable hyperplan de régression* :

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

[2]Les données de notre échantillon consisteront en un vecteur y de dimension n qui contient les observations par rapport à la variable réponse, et d'une matrice X de dimension $n \times (m+1)$ (en anglais : `design matrix`) qui contient les observations par rapport aux m prédicteurs (ainsi qu'une colonne de « 1 ») comme suit :

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}.$$

On calcule alors le vecteur $\widehat{\beta}$ de dimension $m+1$ contenant les coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ de l'hyperplan de régression défini par la méthode des moindres carrés comme suit :

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

où X^T désigne la matrice transposée de X.

que l'on calculerait dans la population si on disposait de toutes les données de celle-ci. De même, les erreurs de prédiction calculées sur l'échantillon :

$$\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2} - \cdots - \widehat{\beta}_m x_{im}$$

sont des estimations des « véritables erreurs de prédiction » (que l'on ferait si on utilisait le véritable hyperplan de régression pour faire la prédiction), données par :

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_m x_{im}.$$

La variable résiduelle sur la population est ainsi définie par :

$$\varepsilon = Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_m X_m.$$

C'est une variable non observable car les véritables paramètres $\beta_0, \beta_1, \cdots, \beta_m$ sont inconnus. L'hyperplan de régression est par ailleurs l'hyperplan pour lequel on a $mean(\varepsilon) = 0$ et où $variance(\varepsilon)$ est minimisé.

Comme en régression simple, on pourra décomposer une variable Y en une somme de deux termes :

$$Y = (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m) + \varepsilon$$

le premier terme représentant l'hyperplan de régression ou *le meilleur prédicteur linéaire* de Y à partir des prédicteurs $X_1, X_2, \cdots, X_m$, le second terme représentant l'erreur de prédiction ou variable résiduelle. Il se trouve que l'erreur de prédiction est non correlée avec chacun des prédicteurs, de sorte que l'on peut décomposer la variance de Y en une somme de deux termes :

$$variance(Y) = variance(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m) + variance(\varepsilon).$$

On fera ensuite la même interprétation qu'en régression simple :

- $variance(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)$ représente la variance de Y prédite linéairement par les m prédicteurs
- $variance(\varepsilon)$ représente la variance résiduelle.

Ainsi, le *pourcentage de la variance de Y prédite linéairement par $X_1, X_2, \cdots, X_m$* est défini par :

$$\rho^2 = \frac{variance(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}{variance(Y)} = 1 - \frac{variance(\varepsilon)}{variance(Y)}.$$

Cette quantité est également interprétable comme mesure de réduction de variance résiduelle :

$$\rho^2 = \frac{variance(Y) - variance(\varepsilon)}{variance(Y)}$$

ainsi que comme mesure de la qualité globale de la prédiction. On verra plus loin comment estimer ρ^2.

Exemple 14.1 *Nous considérons un exemple avec* $n = 248$ *adultes de sexe masculin et les* $m = 6$ *prédicteurs suivants*[3] *:*

- X_1 *: poids (en kg)*
- X_2 *: âge (en années)*
- X_3 *: taille (en cm)*
- X_4 *: mesure de l'abdomen (en cm)*
- X_5 *: mesure du biceps (en cm)*
- X_6 *: mesure du poignet (en cm).*

La variable réponse est la suivante :

- Y *: graisse corporelle (exprimée en % du poids du corps).*

La figure 14.1 nous montre les diagrammes de dispersion décrivant les relations bivariées entre toutes ces variables, ainsi que les corrélations correspondantes. On voit que la variable réponse est corrélée positivement avec tous les prédicteurs (surtout l'abdomen et le poids), sauf la taille (avec laquelle la corrélation est quasiment nulle). On voit aussi que les prédicteurs sont corrélés entre eux, avec parfois des corrélations non négligeables, notamment entre le poids, l'abdomen, le biceps et le poignet. Dans cet exemple, on calcule l'hyperplan de régression suivant :

$$\widehat{y}^* = -13.3 - 0.12x_1 + 0.046x_2 - 0.077x_3 + 0.84x_4 + 0.23x_5 - 1.73x_6.$$

Ainsi, pour une personne pesant 80 *kg, âgée de* 40 *ans, mesurant* 175 *cm, avec un abdomen de* 115 *cm, un biceps de* 30 *cm et un poignet de* 19 *cm, on prédira :*

$$-13.3 - 0.12 \cdot 80 + 0.046 \cdot 40 - 0.077 \cdot 175 + 0.84 \cdot 115 + 0.23 \cdot 30 - 1.73 \cdot 19 = 36 \ \%$$

de graisse corporelle. On estimera par ailleurs à 72 % *le pourcentage de la variance de* Y *prédite par cette équation (voir plus loin).*

14.2 Hypothèse de linéarité

Afin de donner une interprétation descriptive à l'hyperplan de régression, nous allons à nouveau faire une hypothèse de linéarité. En régression simple,

[3]Il s'agit d'un exemple fameux que l'on peut trouver sur Internet, par exemple sous `www.amstat.org/publications/jse/datasets/fat.dat.txt`. Ces données ont été récoltées par le Dr A. Garth Fisher et publiées par Roger W. Johnson (1996). L'ensemble de données original comprend $n = 252$ individus. Nous avons ici éliminé 4 d'entre eux avec certaines valeurs extrêmes, à savoir les Nos 39 (poids de 165 kg), 41 (poids de 119 kg), 42 (taille de 75 cm) et 216 (graisse corporelle de 47.5 %), de telle sorte que nous nous retrouvons avec $n = 248$ individus. Nous avons également transformé la taille de pouces en cm (*via* une multiplication par 2.54) et le poids de livres en kilogrammes (*via* une division par 2.205).

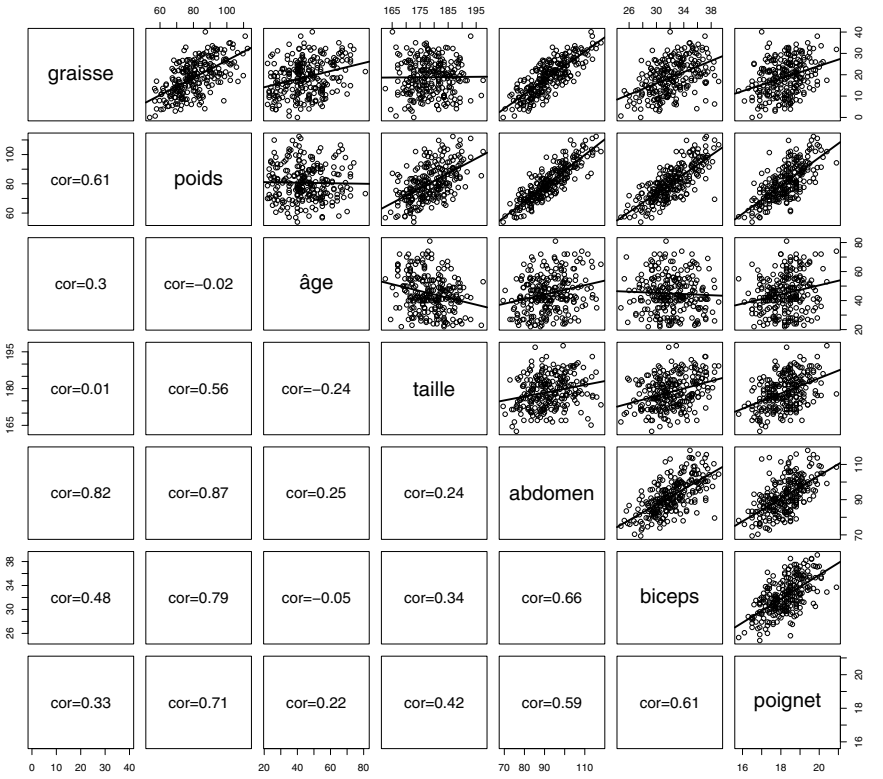

Figure 14.1 – Relations entre six prédicteurs et une variable réponse.

on considérait une infinité de groupes (un groupe pour chacune des valeurs possibles du seul prédicteur) et on s'intéressait aux moyennes de la variable Y dans chacun de ces groupes. Ce sera pareil en régression multiple, sauf que l'on aura une infinité de groupes « encore plus grande ». On aura entre autres (en reprenant l'exemple ci-dessus) :

- les individus pesant 80 kg, âgés de 40 ans, mesurant 175 cm, avec un abdomen de 115 cm, un biceps de 30 cm et un poignet de 19 cm
- les individus pesant 90 kg, âgés de 35 ans, mesurant 192 cm, avec un abdomen de 95 cm, un biceps de 28 cm et un poignet de 17 cm
- les individus pesant 91 kg, âgés de 35 ans, mesurant 192 cm, avec un abdomen de 95 cm, un biceps de 28 cm et un poignet de 17 cm
- etc.

L'hypothèse de linéarité consiste à supposer que les moyennes de la variable Y définies dans chacun de ces innombrables groupes sont « hyper-alignées », autrement dit qu'elles se trouvent sur l'hyperplan de régression.

L'hypothèse de linéarité est ainsi la suivante :

$$mean(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m$$

où $mean(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)$ désigne la moyenne de Y dans le groupe défini par $X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m$. Ainsi, utiliser l'hyperplan de régression pour faire de la prédiction consiste à prédire la valeur de Y d'un individu par la moyenne du groupe auquel cet individu appartient, l'erreur de prédiction étant la différence entre la valeur de Y pour cet individu et la moyenne de son groupe. Cette dernière peut être interprétée comme la valeur de Y *ajustée* pour les caractéristiques de l'individu par rapport aux prédicteurs. De même pour les individus de notre échantillon, la valeur prédite $\widehat{y_i^*}$ est une estimation de $mean(Y|X_1 = x_{i1}, X_2 = x_{i2}, \cdots, X_m = x_{im})$, c'est-à-dire de la moyenne du groupe auquel le i-ième individu appartient, et $\widehat{\varepsilon_i}$ est une estimation de la différence entre la valeur y_i de cet individu et cette moyenne.

L'hypothèse de linéarité est donc encore plus forte en régression multiple qu'en régression simple, impliquant davantage de groupes. Ici aussi, l'hypothèse de linéarité revient à dire que les erreurs de prédictions sont en moyenne nulles non seulement globalement mais également localement (dans chaque groupe). On aura donc non seulement $mean(\varepsilon) = 0$, mais également :

$$mean(\varepsilon|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) = 0$$

quel que soit le groupe défini par $X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m$. Cette condition ne sera cependant pas facile à vérifier dans un espace à $m+1$ dimensions. En plus d'un problème de visualisation, les données seront relativement éloignées les unes des autres dans un tel espace, de sorte qu'il ne sera pas commode de regrouper les individus localement, comme on l'avait fait en régression simple. Les dangers de l'extrapolation nous guetteront ainsi d'autant plus en régression multiple qu'en régression simple car nos prédictions se feront souvent en dehors du domaine d'observation. On voudra par exemple utiliser notre hyperplan de régression pour prédire la graisse corporelle d'un individu pesant 80 kg, âgés de 40 ans, mesurant 175 cm, avec un abdomen de 115 cm, un biceps de 30 cm et un poignet de 19 cm, comme on l'a fait dans notre exemple ci-dessus, mais un tel individu se trouvera peut-être éloigné de chacun des n individus de notre échantillon, de sorte que l'on n'aura aucune garantie de la qualité de la prédiction dans cet endroit de l'espace défini par les m prédicteurs.

14.3 Interprétation des paramètres

Sous l'hypothèse de linéarité, on peut en outre interpréter les paramètres $\beta_0, \beta_1, \cdots, \beta_m$ de l'hyperplan de régression comme suit :

- **pour la constante**, on a :

$$\beta_0 = mean(Y|X_1 = 0, X_2 = 0, \cdots, X_m = 0)$$

→ la constante est la moyenne de Y dans le groupe particulier défini par $X_1 = X_2 = \cdots = X_m = 0$

→ interprétation non intéressante si le groupe défini par $X_1 = X_2 = \cdots = X_m = 0$ est loin du domaine d'observation, ce qui arrivera souvent en pratique

→ on pourra rendre l'interprétation de la constante intéressante en centrant les prédicteurs, c'est-à-dire en leur soustrayant leur moyenne, auquel cas le groupe défini par $X_1 = X_2 = \cdots = X_m = 0$ sera celui des individus qui sont dans la moyenne par rapport à tous les prédicteurs

- **pour la pente associée au prédicteur** X_1, on a pour n'importe quelle valeur possible x_1 :

$$\begin{array}{rl} mean(Y|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_m = x_m) & \\ -mean(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) & = \\ \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots \beta_m x_m & \\ -(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_m x_m) & \end{array}$$

et ainsi :

$$\begin{array}{rcl} \beta_1 & = & mean(Y|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_m = x_m) \\ & & -mean(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) \end{array}$$

→ la pente associée à X_1 est la différence de moyenne de Y entre deux groupes différant d'une unité par rapport à X_1, et identiques par rapport à tous les autres prédicteurs

- **pour la pente associée au prédicteur** X_2, on a de même pour n'importe quelle valeur possible x_2 :

$$\begin{array}{rcl} \beta_2 & = & mean(Y|X_1 = x_1, X_2 = x_2 + 1, \cdots, X_m = x_m) \\ & & -mean(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) \end{array}$$

→ la pente associée à X_2 est la différence de moyenne de Y entre deux groupes différant d'une unité par rapport à X_2 et identiques par rapport à tous les autres prédicteurs

→ interprétation similaire (et intéressante) pour toutes les pentes

→ ces pentes nous aideront à répondre à la deuxième question énoncée en début de chapitre (comment se modifie Y en fonction de $X_1, X_2, \cdots, X_m$).

Exemple 14.2 *Nous reprenons les données de l'exemple précédent avec cette fois-ci (en conservant les mêmes notations pour les différentes variables) le poids X_1 comme variable réponse et avec les trois prédicteurs suivants : l'âge X_2, la taille X_3 et l'abdomen X_4. On considère tout d'abord un exemple de régression simple (qui est un cas particulier de régression multiple), avec comme unique prédicteur l'âge. On obtient la droite de régression :*

$$\widehat{x}_1^* = 81.5 - 0.02x_2.$$

Notons qu'un test de Student sur la pente nous donne ici $p = 0.78$, de sorte que l'on ne peut pas exclure l'hypothèse que la relation entre l'âge et le poids est globalement nulle (le diagramme de dispersion correspondant de la figure 14.1 ne nous montre par ailleurs aucune relation évidente entre les variables).

Ce résultat peut apparaître à première vue surprenant si on pense que les gens ont tendance à prendre du poids avec les années. En fait, la relation entre l'âge et le poids est ici confondue par une variable que l'on appellera « génération ». En effet, les jeunes de notre échantillon ont vécu à une autre époque que les moins jeunes. Or, il est établi que les nouvelles générations sont plus grandes en taille que les anciennes (grâce à une meilleure nutrition et à de meilleures conditions de vie en général). Comme le poids et la taille sont positivement corrélés (voir la figure 14.1), les jeunes seront donc à la base plus lourds que les moins jeunes, ce qui compensera les kilos pris par ces derniers au cours des années, d'où l'association quasiment nulle observée entre l'âge et le poids.

Considérons à présent une régression multiple avec deux prédicteurs : l'âge et la taille. On obtient le plan de régression :

$$\widehat{x}_1^* = -115.7 + 0.12x_2 + 1.07x_3.$$

Ici, le coefficient de l'âge est positif (et il est également significatif, $p = 0.02$). En comparant jeunes et moins jeunes à taille égale, on ne tient plus compte de l'influence confondante de la génération et on trouve ainsi le résultat attendu, à savoir que les jeunes sont moins lourds que les moins jeunes (avec une différence moyenne estimée à 0.12 kg par année, c'est-à-dire 1.2 kg pour 10 ans). Ce résultat est à présent compatible avec l'hypothèse que les gens prennent du poids avec les années (on notera par ailleurs que la relation bien connue entre la taille et le poids est ici confirmée et quantifiée : en comparant à âge égal, on estime une différence moyenne de poids de 1.07 kg par cm de taille).

Considérons finalement une régression multiple avec comme trois prédicteurs l'âge, la taille et l'abdomen. On obtient l'hyperplan de régression :

$$\widehat{x}_1^* = -112.5 - 0.14x_2 + 0.59x_3 + 1.01x_4.$$

Le signe du coefficient de l'âge a encore changé : il est devenu négatif (et la valeur p a encore diminué, avec $p < 0.0001$). Cela veut dire que si on compare non seulement à taille égale, mais également à abdomen égal, les jeunes sont

plus lourds que les moins jeunes (avec une différence moyenne estimée à -0.14 *kg par année, c'est-à-dire* -1.4 *kg pour* 10 *ans). Notre interprétation de ce résultat est la suivante. Il y a essentiellement deux façons de prendre du poids : soit via une augmentation de graisse (que l'on retrouvera surtout au niveau de l'abdomen), soit via une augmentation de muscle. La prise de poids au cours du vieillissement concerne l'augmentation de graisse. Or, en comparant à abdomen égal, on ne tient plus compte de cette augmentation de graisse, de sorte que la différence de poids se décide au niveau du muscle. Le coefficient négatif de l'âge dans cette équation est donc compatible avec un autre résultat attendu : les jeunes ont plus de muscle que les moins jeunes.*

Notons que la constante n'est interprétable dans aucun de ces exemples (il n'existe pas de personnes adultes âgées de 0 *ans, mesurant* 0 *cm et avec un abdomen de* 0 *cm), mais qu'elle est indispensable dans ces équations lorsqu'il s'agit de calculer des prédictions.*

Dans cet exemple, on constate que l'introduction de nouveaux prédicteurs dans une équation de régression modifie la valeur et donc l'interprétation des coefficients des prédicteurs déjà présents dans l'équation. Cela sera toujours le cas, sauf si le nouveau prédicteur se trouve être non corrélé avec les prédicteurs déjà dans l'équation, auquel cas les coefficients de ces derniers ne seront pas modifiés. On retrouvera ce résultat au chapitre 15 lorsque l'on parlera d'essais cliniques[4].

> L'introduction d'un nouveau prédicteur dans l'équation de régression modifie les coefficients des prédicteurs déjà dans l'équation, sauf si ces prédicteurs sont non corrélés avec le nouveau prédicteur.

Exemple 14.3 *Nous reprenons les données de l'exemple ci-dessus avec comme variable réponse* Y *la graisse corporelle et avec deux prédicteurs :* X_1 *le poids*

[4]On citera également le résultat suivant. Considérons une équation de régression simple avec une variable réponse Y et un prédicteur X_1, donnée par :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1.$$

Si on introduit un second prédicteur X_2 dans cette équation, on aura :

$$\widehat{y}^* = \widehat{\beta}'_0 + \widehat{\beta}'_1 x_1 + \widehat{\beta}'_2 x_2$$

où les coefficients $\widehat{\beta}'_0$ et $\widehat{\beta}'_1$ seront en général différents de $\widehat{\beta}_0$ et $\widehat{\beta}_1$. En particulier, on aura :

$$\widehat{\beta}'_1 = \frac{\widehat{\beta}_1 - \widehat{\rho}_{X_1X_2} \cdot \widehat{\rho}_{YX_2} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_{X_1}}}{1 - \widehat{\rho}^2_{X_1X_2}}$$

où $\widehat{\rho}_{X_1X_2}$ et $\widehat{\rho}_{YX_2}$ dénotent les corrélations estimées entre X_1 et X_2, respectivement entre Y et X_2, et $\tilde{\sigma}_{X_1}$ et $\tilde{\sigma}_Y$ dénotent les écarts types estimés de X_1 et Y. Avec $\widehat{\rho}_{X_1X_2} = 0$, on aura en effet $\widehat{\beta}'_1 = \widehat{\beta}_1$, comme mentionné ci-dessus.

et X_2 l'âge. Si on considère une régression simple avec le poids comme unique prédicteur, on obtient :

$$\widehat{y}^* = -13.9 + 0.406x_1.$$

On estime ainsi une augmentation de 0.406 % *de graisse corporelle par kg. Si on introduit l'âge dans l'équation comme second prédicteur, on obtient :*

$$\widehat{y}^* = -23.1 + 0.410x_1 + 0.197x_2.$$

Le coefficient du poids ne s'en trouve quasiment pas modifié (on a 0.410 *au lieu de* 0.406*) car l'âge est quasiment non corrélé avec le poids (corrélation de* -0.02*). Si l'âge et le poids avaient été exactement non corrélés, le coefficient du poids aurait été exactement le même dans les deux équations.*

14.4 Ajustement pour les variables confondantes

Dans l'avant-dernier exemple ci-dessus, la variable taille était jugée confondante pour l'association entre les variables âge et poids, étant négativement corrélée avec la première (*via* un effet de génération, les générations les plus grandes en taille étant les plus récentes et donc les moins âgées), et positivement corrélée avec la seconde. On a vu comment on peut éliminer son influence en l'introduisant (aux côtés de l'âge) dans l'équation de régression pour prédire le poids. L'interprétation de la pente de l'âge dans cette nouvelle équation nous permet en effet de comparer la moyenne du poids entre deux groupes d'âges différents mais de taille égale. On dira dans pareil cas que la relation entre l'âge et le poids a été *ajustée pour la taille*. En fixant la taille, on l'empêche de confondre. Il s'agit d'un principe fondamental de la régression multiple (et sans doute de son utilité principale).

> On introduit une variable confondante dans une équation de régression afin de pouvoir éliminer son influence.

Ce principe se généralise au cas de plusieurs variables confondantes. Évidemment, afin de pouvoir les introduire dans l'équation, il faudra que ces variables confondantes soient identifiées, mesurables et mesurées sur les individus de notre échantillon, ce qui ne sera malheureusement pas toujours le cas.

Exemple 14.4 *Les données ci-dessous ont été récoltées afin d'étudier le développement de la motricité chez les enfants et les adolescents*[5]*. La motricité est*

[5]Nous remercions Remo Largo et Oskar Jenni, du Département Croissance et Développement de l'Hôpital universitaire de Zurich, de nous avoir mis à disposition ces données. Elles ont été récoltées sur des participants à l'Etude Longitudinale de Zurich et concernent un exercice du *Zurich Neuromotor Assessment* (Largo *et al.*, 2007).

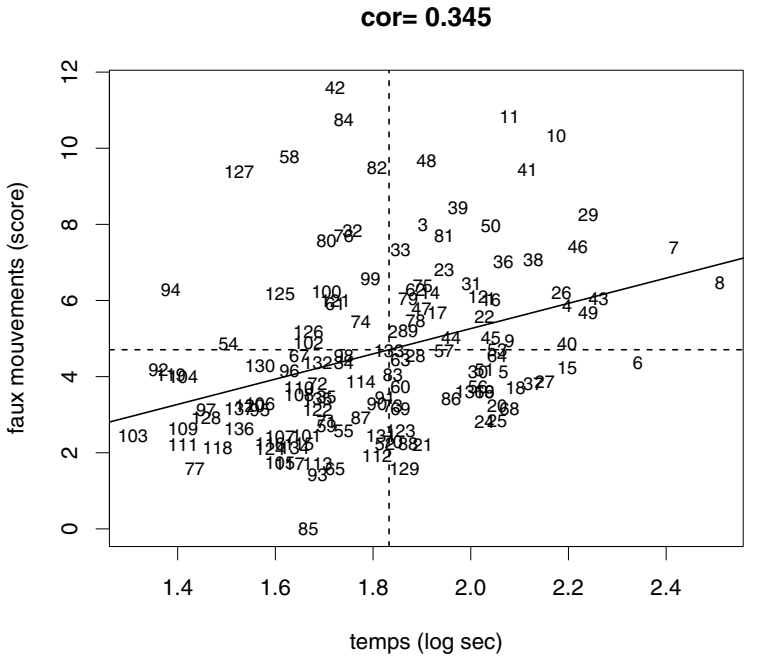

Figure 14.2 – Relation entre qualité et rapidité des mouvements.

un concept complexe qui ne se mesure pas aussi facilement que la taille ou le poids. On pourra par exemple mesurer le temps nécessaire pour accomplir un exercice (impliquant un mouvement particulier). La rapidité des mouvements ne représente toutefois qu'un aspect de la motricité, un autre aspect étant la qualité des mouvements, que l'on pourra mesurer via un score recensant les faux mouvements (l'absence de faux mouvements étant un signe de qualité). La figure 14.2 nous montre la relation entre le temps et le score de faux mouvements pour $n = 137$ *enfants et adolescents âgés entre* 6 *et* 15 *ans (on a choisi ici d'analyser ensemble filles et garçons). Chaque enfant ou adolescent est représenté sur ce graphique par un nombre entre* 1 *et* 137 *afin que le lecteur intéressé puisse les retrouver dans les différentes figures que nous allons présenter. Notons que le temps a été log-transformé (exprimé en log secondes sur ces figures) pour se rapprocher de la normalité. Les enfants se trouvant sur la gauche du graphique sont donc les plus rapides, alors que les enfants se trouvant en bas du graphique sont ceux avec une meilleure qualité de mouvement. On a une corrélation positive (et significative) de* 0.345 *entre les deux variables, suggérant que les mouvements les plus rapides sont aussi en moyenne de meilleure qualité, ce qui serait d'un intérêt scientifique.*

Cependant, cette relation est confondue par l'âge des enfants et adolescents,

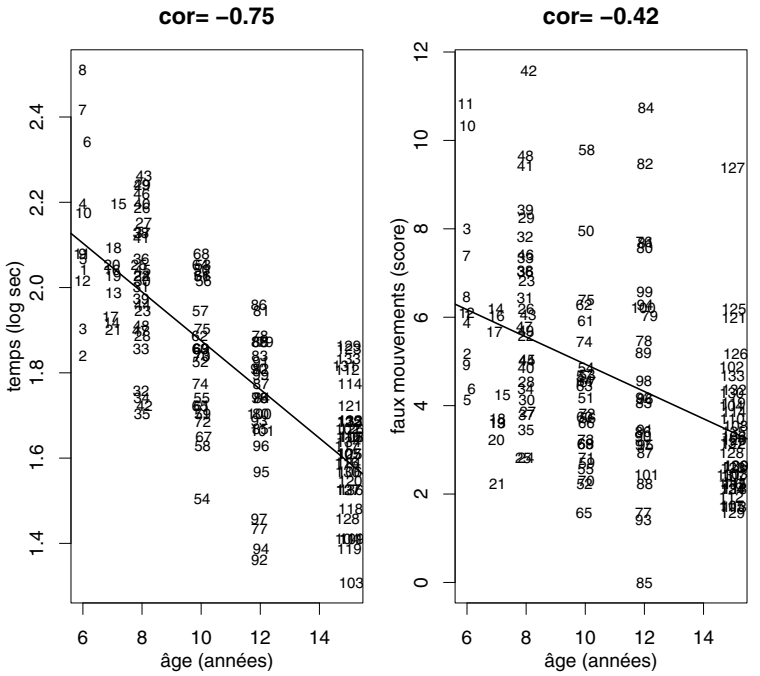

Figure 14.3 – Relation de la qualité et de la rapidité des mouvements avec l'âge.

la rapidité et la qualité des mouvements s'améliorant naturellement avec l'âge. La figure 14.3 nous confirme que la corrélation entre l'âge et le temps est de -0.75 *(les adolescents étant plus rapides que les enfants), et celle entre l'âge et le score de faux mouvements est de* -0.42 *(les mouvements des adolescents étant de meilleure qualité que ceux des enfants).*

Si on considère une équation de régression simple avec le temps X_1 *comme prédicteur et le score de faux mouvements* Y *comme variable réponse, on obtient :*

$$\widehat{y}^* = -1.38 + 3.319x_1.$$

On estime ainsi une augmentation moyenne du score de faux mouvements de 3.32 *points pour une augmentation de* 1 *log seconde qui est significative (*$p < 0.0001$*). En introduisant l'âge* X_2 *dans l'équation, on obtient cependant :*

$$\widehat{y}^* = 6.48 + 0.631x_1 - 0.272x_2.$$

Pour un âge fixé, l'augmentation moyenne estimée du score de faux mouvements n'est plus que de 0.63 *point pour une augmentation de* 1 *log seconde, ce qui représente cinq fois moins que dans la première équation, et ce qui n'est plus significatif (*$p = 0.58$*). Une fois ajustée pour l'âge, on ne peut donc pas exclure*

qu'il n'y ait pas d'association entre la rapidité et la qualité des mouvements[6].

14.5 Corrélation partielle

On a vu au chapitre 13 qu'une pente de régression et une corrélation sont deux résumés possibles (mais apportant une information différente) de l'association statistique entre deux variables continues. On a également vu dans la section précédente comment on peut ajuster une pente de régression, afin d'éliminer l'influence des variables confondantes. On va voir dans cette section comment on peut ajuster de la sorte une corrélation.

Afin de calculer une corrélation entre deux variables X_1 et Y en s'affranchissant de l'influence d'une variable confondante X_2, on procède de la manière suivante :

- on calcule les résidus de la régression avec X_1 comme variable réponse et X_2 comme prédicteur ; on appellera cette variable résiduelle « X_1 ajustée pour X_2 » (notée $X_1|X_2$)
- on calcule les résidus de la régression avec Y comme variable réponse et X_2 comme prédicteur ; on appellera cette variable résiduelle « Y ajustée pour X_2 » (notée $Y|X_2$)
- on calcule la corrélation entre les deux variables résiduelles ci-dessus, à savoir $X_1|X_2$ et $Y|X_2$.

La corrélation obtenue est appelée la corrélation entre X_1 et Y *ajustée pour* X_2 ou *corrélation partielle*, que l'on notera $\rho_{X_1Y|X_2}$. Cette corrélation partielle peut s'obtenir alternativement en calculant directement :

$$\rho_{X_1Y|X_2} = \frac{\rho_{X_1Y} - \rho_{X_1X_2}\rho_{YX_2}}{\sqrt{(1-\rho^2_{X_1X_2})(1-\rho^2_{YX_2})}}$$

où ρ_{X_1Y}, $\rho_{X_1X_2}$ et ρ_{YX_2} dénotent les corrélations habituelles, respectivement entre X_1 et Y, entre X_1 et X_2, et entre Y et X_2. On notera que dans le cas où la variable X_2 n'est corrélée ni avec X_1 ni avec Y, on aura $\rho_{X_1Y|X_2} = \rho_{X_1Y}$ (dans ce cas, X_2 n'est pas une variable confondante)[7].

[6] Notons que l'on peut ici inverser les rôles, et considérer (en gardant les mêmes notations que ci-dessus) Y comme prédicteur et X_1 comme variable réponse. On obtient l'équation de régression simple :

$$\widehat{x}_1^* = 1.66 + 0.0359y.$$

En introduisant X_2 dans l'équation, on obtient :

$$\widehat{x}_1^* = 2.42 + 0.00362y - 0.0560x_2$$

la pente du prédicteur Y étant dix fois plus petite que dans la première équation et n'étant plus significative (avec exactement le même $p = 0.58$ que ci-dessus).

[7] La corrélation partielle $\rho_{X_1Y|X_2}$ est par ailleurs liée à la pente β_1 associée au prédicteur X_1 dans l'équation de régression multiple avec Y comme variable réponse et avec X_1 et X_2

Ce principe se généralise à plusieurs variables confondantes. On peut ajuster la corrélation entre X_1 et Y pour les variables confondantes $X_2, X_3, \cdots, X_m$, en introduisant toutes ces variables confondantes dans les équations de régression calculées lors des deux premières étapes de l'algorithme ci-dessus.

Exemple 14.5 *On reprend l'exemple de la motricité. Rappelons que Y dénote le score de faux mouvements, X_1 le temps nécessaire pour accomplir un exercice et X_2 l'âge des enfants et adolescents. On avait les corrélations suivantes : $\rho_{X_1Y} = 0.345$, $\rho_{X_1X_2} = -0.75$ et $\rho_{YX_2} = -0.42$. On peut ainsi calculer la corrélation partielle*

$$\rho_{X_1Y|X_2} = \frac{0.345 - (-0.75)\cdot(-0.42)}{\sqrt{(1-(-0.75)^2)(1-(-0.42)^2)}} = 0.05.$$

Une fois éliminée l'influence de l'âge, il n'y a donc pratiquement plus de corrélation entre la rapidité et la qualité des mouvements. On retrouvera cette même corrélation partielle de 0.05 *en calculant la corrélation habituelle entre les résidus de la régression où X_1 est la variable réponse et X_2 le prédicteur, et les résidus de la régression où Y est la variable réponse et X_2 le prédicteur, c'est-à-dire la corrélation entre le temps ajusté pour l'âge et le score de faux mouvements ajusté pour l'âge, comme illustré dans la figure 14.4.*

14.6 Modèle de régression linéaire multiple

Après avoir défini l'hypothèse de linéarité dans le cadre de la régression multiple, ce qui nous a permis de donner une interprétation descriptive à l'hyperplan de régression, nous allons à présent introduire les autres hypothèses de ce qui constituera le *modèle de régression linéaire multiple*, qui nous permettront de faciliter l'inférence sur l'hyperplan de régression.

Tout se passe de façon similaire à ce que l'on avait en régression simple. L'hypothèse de linéarité concerne l'hyper-alignement des moyennes de la variable Y pour les différents groupes définis par les innombrables combinaisons de valeurs possibles des prédicteurs. On fait ensuite une hypothèse d'homoscédasticité et une hypothèse de normalité, la première nous informant sur la variance (qui sera constante, égale à la variance résiduelle, que l'on notera σ^2_ε), la seconde sur la forme (qui sera normale) de la distribution de la variable Y dans les différents groupes. On a ainsi :

---

comme prédicteurs, par la formule suivante :

$$\rho_{X_1Y|X_2} = \beta_1 \cdot \frac{\sigma_{X_1|X_2}}{\sigma_{Y|X_2}}$$

où $\sigma_{X_1|X_2}$ et $\sigma_{Y|X_2}$ dénotent les écarts types des variables résiduelles définies ci-dessus. On retrouve ainsi le lien habituel entre une corrélation et une pente de régression.

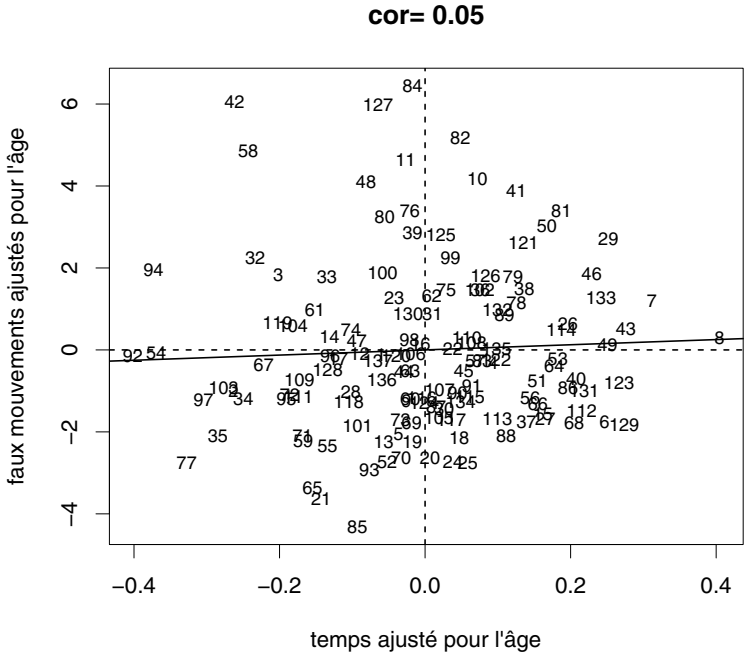

Figure 14.4 – Relation entre qualité et rapidité des mouvements ajustés pour l'âge.

linéarité	:	$mean(Y\|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
homoscédasticité	:	$variance(Y\|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m) = \sigma_\varepsilon^2$
normalité	:	$Y\|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m$ normale.

Ces hypothèses spécifient complètement la distribution de la variable Y dans les différents groupes. Comme déjà mentionné pour l'hypothèse de linéarité, ces hypothèses sont plus difficiles à vérifier en régression multiple qu'en régression simple (nous reviendrons sur ce point dans la section suivante)[8].

Comme en régression simple, on peut reformuler ces trois hypothèses en fonction des résidus et on y ajoutera également une quatrième hypothèse concernant l'indépendance de ces résidus[9]. Ainsi, les quatre hypothèses d'un modèle

[8] Nous mentionnerons également que ces trois hypothèses sont satisfaites lorsque la variable $(X_1, X_2, \cdots, X_m, Y)$ est *normale multivariée* (concept que nous ne développons pas ici et qui généralise les concepts de normalité et de binormalité au cas multivarié).

[9] On rappellera que l'hypothèse d'indépendance concerne l'échantillonnage, alors que les

de régression multiple se formulent exactement de la même manière que les quatre hypothèses d'un modèle de régression simple, à savoir :

linéarité	:	$mean(\varepsilon_i) = 0$
homoscédasticité	:	$variance(\varepsilon_i) = \sigma_\varepsilon^2$
normalité	:	ε_i normalement distribués
indépendance	:	ε_i indépendants.

14.7 Analyse des résidus

Les hypothèses d'un modèle de régression linéaire peuvent donc être formulées en fonction des résidus ε_i. Afin de vérifier ces hypothèses, on utilisera nos résidus estimés $\widehat{\varepsilon}_i$ et on procédera à une *analyse des résidus*. Il sera cependant impossible de vérifier ces hypothèses pour toutes les combinaisons des valeurs possibles de $X_1, X_2, \cdots, X_m$. On effectuera alors des regroupements. On regardera par exemple si les hypothèses du modèle sont (approximativement) satisfaites aux alentours des observations avec $X_1 = x_1$ (on regroupera toutes les observations avec une valeur proche de $X_1 = x_1$, quelles que soient leurs valeurs pour les autres prédicteurs $X_2, \cdots, X_m$)[10]. Ceci peut se faire à l'aide d'un diagramme de dispersion montrant les résidus estimés $\widehat{\varepsilon}_i$ en fonction du prédicteur X_1. On procédera ensuite de la même manière pour les autres prédicteurs $X_2, \cdots, X_m$, obtenant m diagrammes de dispersion. Les hypothèses du modèle ne sont pas contredites si dans aucun de ces graphiques n'apparaît un signe de non-linéarité, de non-homoscédasticité ou de non-normalité[11].

trois autres hypothèses concernent la nature du phénomène étudié. Ici aussi, ce ne sont pas les observations Y_i de la variable Y qui doivent être indépendantes, mais les erreurs de prédiction ε_i. Ceci implique que l'on pourra (si on le désire) choisir les valeurs $x_{i1}, x_{i2}, \cdots, x_{im}$ des différents individus de notre échantillon, c'est-à-dire que l'on pourra choisir les groupes qui sont représentés dans notre échantillon, pour autant que les individus soient alors tirés au hasard (et indépendamment les uns des autres) de ces différents groupes.

[10] La distribution des résidus étant la même, c'est-à-dire normale de moyenne nulle et de variance σ_ε^2, dans chacun des innombrables groupes définis par toutes les combinaisons de valeurs possibles de $X_1, X_2, \cdots, X_m$, elle sera encore normale de moyenne nulle et de variance σ_ε^2 lorsque l'on mettra ensemble les résidus de certains de ces groupes, par exemple tous ceux avec $X_1 = x_1$. De même, la distribution sera toujours normale de moyenne nulle et de variance σ_ε^2 lorsque l'on regroupera l'ensemble des résidus de notre échantillon (mettre ensemble des observations avec une même distribution ne modifiera pas la distribution).

[11] L'hypothèse d'indépendance (qui concerne l'échantillonnage) sera en principe plus difficile à vérifier empiriquement, sauf dans des cas particuliers. Par exemple, lorsque les observations sont récoltées au cours du temps, on pourra utiliser des méthodes du domaine des *séries chronologiques* afin de vérifier que les résidus de deux observations récoltées proches dans le

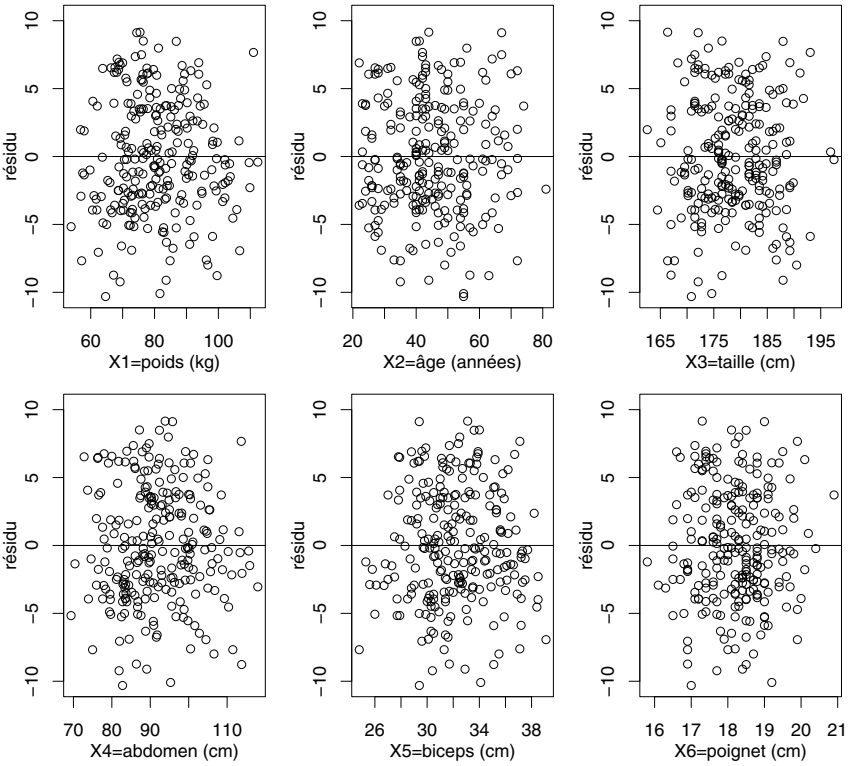

Figure 14.5 – Résidus *versus* prédicteurs.

Exemple 14.6 *On reprend l'exemple de la prédiction de graisse corporelle. La figure 14.5 nous montre les résidus estimés en fonction des* $m = 6$ *différents prédicteurs. Il s'agit de vérifier les points suivants :*

- *la moyenne des résidus doit être nulle partout (linéarité)*
- *la variance des résidus autour de* 0 *doit être la même partout (homoscédasticité)*
- *la distribution des résidus autour de* 0 *doit être partout symétrique et avec peu de valeurs extrêmes (normalité).*

Le terme « partout » veut dire ici « aux alentours de chaque valeur possible de chacun des prédicteurs ». Autrement dit, afin de ne pas contredire les hypo-

temps ne soient pas en moyenne plus semblables que les résidus de deux observations récoltées loin dans le temps. Dans le cas contraire, on pourra essayer de modéliser cette dépendance. De même, on pourra essayer de modéliser la dépendance que l'on aura dans les cas où plusieurs observations sont faites sur un même individu. Ces méthodes de détection et de modélisation de la dépendance entre observations dépassent cependant le cadre de ce texte.

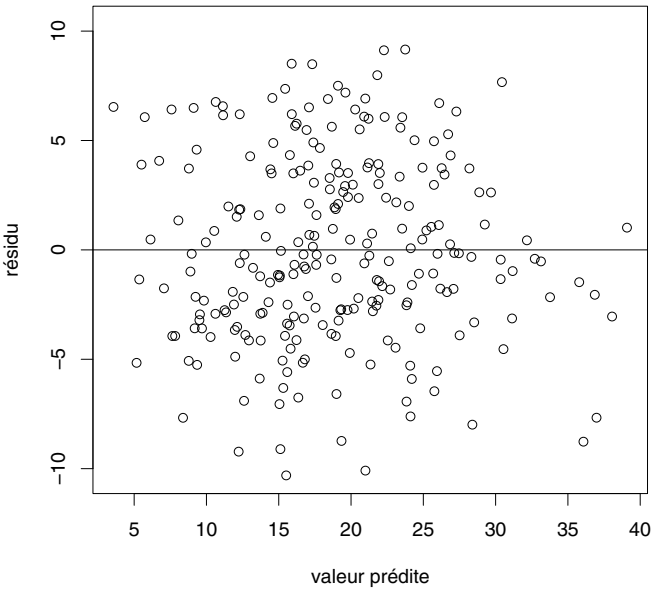

Figure 14.6 – Résidus *versus* valeurs prédites.

thèses du modèle, aucune structure ne doit apparaître sur ces graphiques. Dans notre exemple, on ne détecte aucune anomalie notoire et on dira ainsi que les hypothèses du modèle semblent raisonnables.

Afin de réduire le nombre de diagrammes de dispersion à inspecter, on peut regarder le diagramme de dispersion montrant les résidus estimés en fonction des valeurs prédites (une sorte de résumé des m diagrammes de dispersion précédents). Ici encore, il s'agit de vérifier que les résidus sont en moyenne à peu près nuls, à peu près de même variance et approximativement normalement distribués, et cela aux alentours de chaque valeur prédite possible. En ce qui concerne la normalité, qui n'est pas toujours facile à vérifier à partir d'un diagramme de dispersion, on pourra en outre regarder un boxplot ou un qq-plot calculé sur l'ensemble des résidus.

Exemple 14.7 *La figure 14.6 nous montre les résidus estimés en fonction des valeurs prédites dans notre exemple de la prédiction de graisse corporelle. Aucune structure flagrante n'est visible sur ce graphique, à part peut-être vers la droite du graphique, où les individus avec une très grande valeur prédite de graisse corporelle ont des résidus en moyenne négatifs (et donc non nuls), ce qui pourrait indiquer que le modèle surestime la graisse corporelle pour ces individus. La figure 14.7 nous montre par ailleurs un boxplot et un qq-plot calculés sur l'ensemble des résidus, où l'on ne voit pas de contradiction flagrante avec l'hypothèse de normalité. Notons que l'on a ici standardisé ces résidus en les*

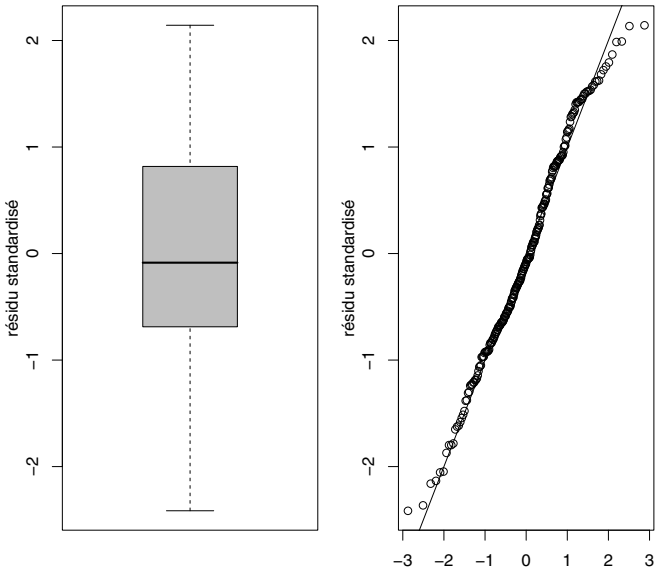

Figure 14.7 – Boxplot et qqplot des résidus standardisés.

divisant par une estimation de l'écart type résiduel (que l'on introduira dans la section suivante). Sous les hypothèses du modèle, ces résidus standardisés auront (approximativement) une distribution normale standardisée (en particulier, 95 % *d'entre eux devraient se trouver entre* -2 *et* $+2$*).*

Nous terminerons cette section en rappelant qu'un modèle de régression linéaire n'est au mieux qu'une approximation de la réalité. À strictement parler, aucune relation n'est exactement linéaire, aucune variance n'est exactement constante et aucune distribution n'est exactement normale. Dans cette optique, nous n'encourageons pas à tester formellement les hypothèses d'un modèle de régression au moyen de tests statistiques dont les hypothèses nulles seraient la linéarité, l'homoscédasticité ou la normalité (tel le *test de Kolmogorov-Smirnov*). De tels tests seraient utiles si le but était de rejeter une hypothèse nulle. Cependant, notre but sera ici d'« accepter » l'hypothèse nulle (de montrer que le modèle est correct). Or, nous avons vu d'une part que le non-rejet d'une hypothèse nulle n'implique pas son « acceptation » (on ne pourra donc pas prouver statistiquement que le modèle est correct). D'autre part, comme notre hypothèse nulle est à strictement parler fausse (notre modèle n'étant qu'une approximation de la réalité), on la rejettera inévitablement si notre échantillon est suffisamment grand (et si on utilise un test raisonnablement puissant). Ainsi, la pratique qui autoriserait l'utilisation d'un modèle de régression linéaire si et seulement si il n'est pas rejeté statistiquement reviendrait à favoriser les petits échantillons (pour lesquels on n'aura pas assez de puissance statistique

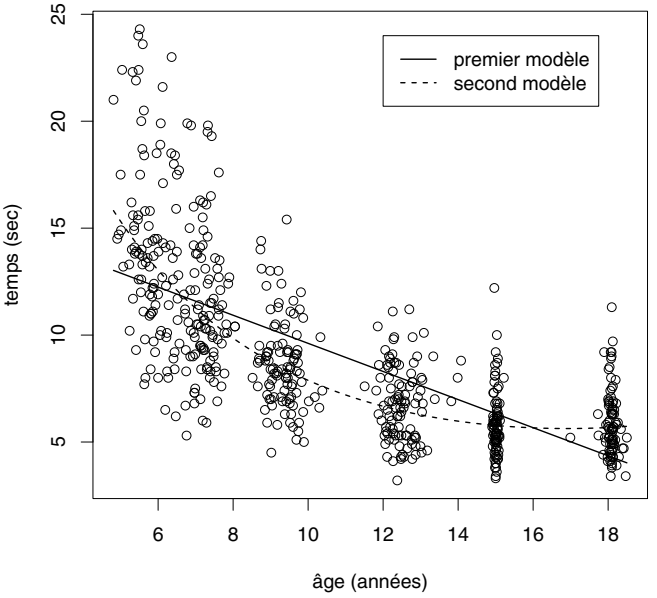

Figure 14.8 – Comparaison d'un modèle linéaire et d'un modèle quadratique.

pour détecter même de grands écarts par rapport au modèle), au détriment des grands échantillons (pour lesquels on aura assez de puissance statistique pour détecter même de petits écarts par rapport au modèle), ce qui est absurde. Il s'agira donc de *vérifier* (par exemple graphiquement) plutôt que de *tester* les hypothèses d'un modèle de régression linéaire, afin de s'assurer qu'elles sont raisonnables. Si elles ne le sont pas, on essaiera de transformer les variables ou de transformer le modèle afin d'améliorer la situation. Si on hésite entre deux modèles, on choisira celui pour lequel l'analyse des résidus est la plus satisfaisante[12].

Exemple 14.8 *Nous présentons un nouvel exemple de motricité. Le but est ici la modélisation de la rapidité des mouvements Y (mesurée via le temps nécessaire à accomplir un exercice) en fonction de l'âge X_1, à partir d'un échantillon de $n = 593$ enfants et adolescents âgés ici entre 5 et 18 ans. Ces données sont présentées dans la figure 14.8, où l'on voit que les enfants deviennent plus rapides avec l'âge, avec un « plateau » atteint aux alentours de 15 ans*[13]*. La droite de régression représentée sur ce graphique nous montre qu'un modèle de régression linéaire simple ne serait pas approprié, les hypothèses de linéarité et d'homoscédasticité n'étant pas du tout satisfaites (d'une part la droite ne passe*

[12] À ce sujet, on citera la phrase célèbre de Georges Box « All models are wrong but some are useful » qui résume bien la philosophie de la modélisation statistique.

[13] Ce sont à nouveau des données du *Zurich Neuromotor Assessment* (Largo *et al.*, 2007).

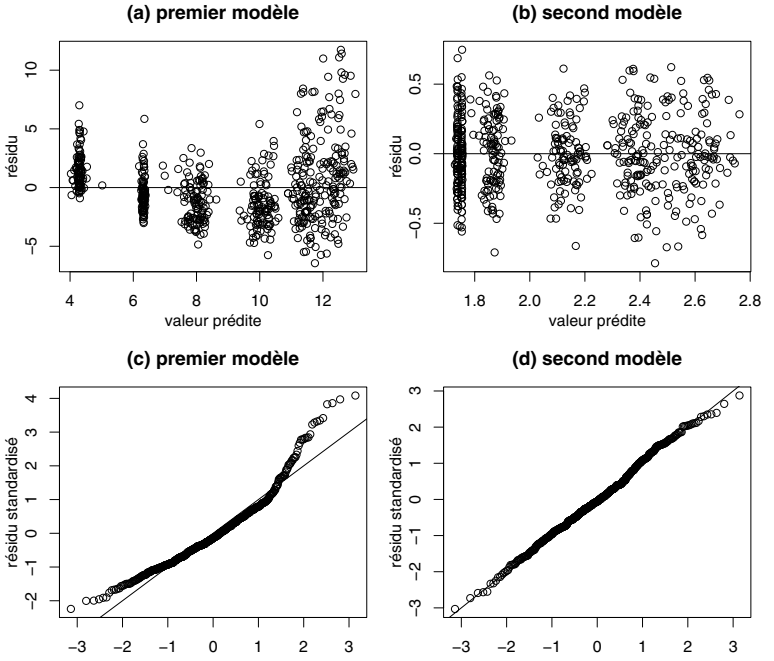

Figure 14.9 – Analyse des résidus pour deux modèles.

pas au milieu des données à chaque âge, d'autre part la variabilité des observations diminue visiblement avec l'âge). Le graphique (a) de la figure 14.9, où l'on voit les résidus en fonction des valeurs prédites pour ce premier modèle, nous montre une information équivalente (les grandes valeurs prédites correspondant aux jeunes âges). Le qq-plot des résidus sur le graphique (c) de cette figure nous montre en outre que l'hypothèse de normalité n'est pas non plus satisfaite.

Afin de « stabiliser la variance » et de nous approcher de la normalité, on a ici log-transformé la variable réponse, c'est-à-dire que l'on a modélisé $Z = \log(Y)$ *au lieu de* Y*. Comme la relation entre* Z *et* X_1 *n'est pas non plus linéaire, on a en outre considéré un modèle quadratique, qui est un cas particulier de modèle de régression linéaire multiple, obtenu en introduisant dans le modèle un second prédicteur égal au carré du premier (on a donc* $X_2 = X_1^2$*). On a ainsi calculé l'hyperplan de régression suivant :*

$$\widehat{z}^* = 3.745 - 0.239x_1 + 0.00706x_1^2.$$

Une analyse graphique des résidus pour ce second modèle nous est présentée dans les graphiques (b) et (d) de la figure 14.9, où l'on voit que les hypothèses de linéarité, d'homoscédasticité et de normalité sont à présent raisonnablement satisfaites. Si on désire représenter les résultats de ce second modèle sur l'échelle

de Y (et non de Z), il s'agit de calculer la courbe suivante :

$$\exp(\widehat{z}^*) = \exp(3.745 - 0.239x_1 + 0.00706x_1^2)$$

qui est également représentée dans la figure 14.8. Nous laisserons le lecteur se convaincre que cette courbe peut s'interpréter comme la médiane (et non comme la moyenne) de la variable Y en fonction de l'âge X_1 (nous reviendrons sur cet exemple un peu plus loin).

14.8 Inférence sur l'hyperplan de régression

Comme en régression simple, les hypothèses d'un modèle de régression linéaire multiple nous permettent de faire de l'inférence sur l'hyperplan de régression. Sous ces hypothèses, les coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ calculés sur notre échantillon sont des estimateurs sans biais et normalement distribués des véritables paramètres $\beta_0, \beta_1, \cdots, \beta_m$ définis sur la population[14]. Des formules explicites sont par ailleurs disponibles pour les variances de ces estimateurs, de sorte que l'on pourra utiliser la méthode de Wald pour faire de l'inférence[15]. Cependant, les variances de ces estimateurs dépendent de la variance résiduelle σ_ε^2. Cette variance résiduelle sera estimée sans biais par :

$$\tilde{\sigma}_\varepsilon^2 = \frac{\sum_i \widehat{\varepsilon}_i^2}{n - m - 1}.$$

On divise ici la somme des carrés des résidus estimés non pas par n, ni par $n-1$, ni même par $n-2$, mais par $n-m-1$, qui constitue le nombre de *degrés de liberté* de notre modèle de régression. On procédera ensuite de la manière habituelle, en remplaçant la variance résiduelle σ_ε^2 par son estimateur $\tilde{\sigma}_\varepsilon^2$ dans les formules des variances des estimateurs. Ici aussi, on utilisera des quantiles d'une distribution de Student (avec $n-m-1$ *dl*) plutôt que des quantiles d'une distribution normale standardisée dans le calcul d'un intervalle de confiance pour les paramètres de la régression ou pour les tests statistiques correspondants (voir ci-dessous). Sous les hypothèses du modèle, ces intervalles de confiance et ces tests statistiques seront exacts, y compris pour de petits échantillons à partir de $n \geq m + 2$[16].

[14] Il s'ensuit que $\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_m x_m$ est également un estimateur sans biais et normalement distribué de $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ (la véritable moyenne de Y dans le groupe défini par $X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m$).

[15] En reprenant notre notation matricielle, la matrice donnée par :

$$\mathrm{Var}(\widehat{\beta}) = \sigma_\varepsilon^2 \cdot (X^T X)^{-1}$$

contient sur sa diagonale les variances des estimateurs $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ (les éléments hors diagnonale de cette matrice étant les covariances entre ces estimateurs).

[16] Bien que l'inférence soit exacte sous les hypothèses du modèle pour une taille d'échantillon aussi petite que $n = m+2$, on ne pourra pas vérifier ces hypothèses à partir d'un échantillon aussi petit. Afin de pouvoir vérifier (au moins partiellement) ces hypothèses, il est souvent recommandé d'avoir $n \geq 10 \cdot (m+1)$, c'est-à-dire d'avoir (au moins) dix fois plus d'observations dans l'échantillon que de paramètres à estimer dans le modèle.

En plus des estimations des paramètres $\widehat{\beta}_j$ de l'hyperplan de régression, un logiciel statistique nous donnera donc les erreurs types (estimées) $\mathrm{SE}(\widehat{\beta}_j)$ de ces estimateurs (pour $j = 0, 1, \cdots, m$). À partir de là, on pourra calculer un intervalle de confiance de Student au niveau $1 - \alpha$ pour β_j comme suit :

$$\widehat{\beta}_j \pm t_{1-\alpha/2,n-m-1} \cdot \mathrm{SE}(\widehat{\beta}_j).$$

On pourra également tester l'hypothèse nulle $H_0 : \beta_j = 0$ en calculant la statistique de test d'un test de Student (ou test t) donnée par :

$$t_{stat} = \frac{\widehat{\beta}_j}{\mathrm{SE}(\widehat{\beta}_j)}$$

et en la comparant avec les quantiles d'une distribution de Student avec $n - m - 1$ *dl* (on rejettera ainsi H_0 dans un test bilatéral au seuil α si $|t_{stat}| \geq t_{1-\alpha/2,n-m-1}$). La plupart des logiciels statistiques nous donneront directement les valeurs calculées t_{stat} de ces statistiques de tests, ainsi que les valeurs p correspondantes[17].

Exemple 14.9 *Dans notre exemple de la prédiction de la graisse corporelle, un logiciel statistique nous donnera un tableau de la forme suivante :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−13.254	14.784	−0.896	0.371
poids	−0.120	0.097	−1.234	0.219
âge	0.046	0.029	1.596	0.112
taille	−0.077	0.066	−1.178	0.240
abdomen	0.841	0.086	9.833	0.000
biceps	0.234	0.160	1.460	0.146
poignet	−1.725	0.494	−3.490	0.001

Dans la première colonne, on retrouve les estimations des paramètres de l'hyperplan de régression. On remarque par exemple le signe négatif de la pente du poids pour prédire la graisse corporelle (−0.12 *: à âge, taille, abdomen, biceps et poignet égal, les individus plus lourds ont en moyenne moins de graisse corporelle, avec une diminution estimée à* 0.12 % *de graisse par kg). Dans la deuxième colonne, on a les erreurs types (estimées) de ces estimateurs. L'erreur type de l'estimateur de la pente du poids est par exemple estimée à* 0.097*. Les statistiques de test données dans la troisième colonne sont obtenues en divisant les estimations de la première colonne par les erreurs types de la deuxième. Pour le poids, on obtient ainsi* $t_{stat} = -0.12/0.097 = -1.234$*. Les valeurs p de la quatrième colonne sont obtenues en comparant ces statistiques de test aux quantiles d'une distribution de Student avec* $248-6-1 = 241$ *dl. Pour le poids,*

[17] Dans R, on obtient les résultats d'un modèle de régression multiple avec la variable réponse `y` et les six prédicteurs `x1, x2, x3, x4, x5` et `x6` en utilisant la commande `summary(lm(y∼x1+x2+x3+x4+x5+x6))`.

on obtient $p = 0.219$, de sorte que l'on ne rejette pas l'hypothèse de la nullité de la pente du poids au seuil de 5 %. Par ailleurs, un intervalle de confiance au niveau 95 % pour la véritable pente du poids s'obtient comme suit (en utilisant $t_{0.975,241} = 1.97$) :

$$0.012 \pm 1.97 \cdot 0.097 = [-0.179; 0.203].$$

Notons qu'afin de calculer ces erreurs types, il aura fallu auparavant calculer une estimation de la variance résiduelle, donnée ici par $\tilde{\sigma}^2_\varepsilon = 18.2$.

14.9 Estimation du pourcentage de la variance prédite

Nous allons voir à présent comment estimer le pourcentage ρ^2 de la variance de Y prédite par l'ensemble des m prédicteurs d'un modèle de régression linéaire. Notons tout d'abord que la décomposition de la variance de Y en une somme de deux termes, représentant la variance prédite et la variance résiduelle, est non seulement valable dans la population, mais également dans l'échantillon (pour autant que l'on utilise la définition originale de la variance avec un dénominateur n). On a en effet :

$$\frac{\sum (y_i - \bar{y})^2}{n} = \frac{\sum_i (\widehat{y}^*_i - \bar{y})^2}{n} + \frac{\sum_i (y_i - \widehat{y}^*_i)^2}{n}.$$

L'*estimateur empirique de ρ^2* est donc le suivant[18] :

$$\widehat{\rho}^2 = \frac{\sum_i (\widehat{y}^*_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}.$$

Pour un grand nombre m de prédicteurs, ce ne sera pourtant pas un bon estimateur de ρ^2. Rappelons que l'on a par définition :

$$\rho^2 = 1 - \frac{\sigma^2_\varepsilon}{\sigma^2_Y}.$$

Par ailleurs, l'estimateur $\widehat{\rho}^2$ peut s'écrire par :

$$\widehat{\rho}^2 = 1 - \frac{\sum_i (y_i - \widehat{y}^*_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\widehat{\sigma}^2_\varepsilon}{\widehat{\sigma}^2_Y}.$$

Dans cette dernière formulation, on utilise $\widehat{\sigma}^2_\varepsilon = \sum \widehat{\varepsilon}^2_i / n$ et $\widehat{\sigma}^2_Y = \sum (y_i - \bar{y})^2 / n$, qui sont des estimateurs biaisés, sous-estimant σ^2_ε et σ^2_Y. Comme le biais du

[18] Rappelons qu'en régression simple, l'estimateur $\widehat{\rho}^2$ ainsi défini est égal au carré de la corrélation empirique entre l'unique prédicteur et la variable réponse. Plus généralement, $\widehat{\rho}^2$ sera le carré de la corrélation empirique entre les valeurs observées y_i et les valeurs prédites $\widehat{y}^*_i$, ce résultat étant valable en régression simple comme en régression multiple.

premier est plus important que le biais du second (surtout avec un grand m), l'estimateur $\widehat{\rho}^2$ aura ainsi tendance à surestimer ρ^2. Afin de corriger ce biais, il suffira d'utiliser dans la formule ci-dessus des estimateurs sans biais de σ^2_ε et σ^2_Y, obtenant l'*estimateur corrigé de* ρ^2 :

$$\tilde{\rho}^2 = 1 - \frac{\tilde{\sigma}^2_\varepsilon}{\tilde{\sigma}^2_Y}.$$

Notons que l'on peut au besoin facilement passer de l'un à l'autre en utilisant la formule[19] :

$$\tilde{\rho}^2 = \frac{(n-1)\widehat{\rho}^2 - m}{n-m-1}.$$

Exemple 14.10 *Dans notre exemple de la prédiction de la graisse corporelle, on estime la variance de la graisse corporelle* Y *par* $\tilde{\sigma}^2_Y = 65.1$ *et la variance résiduelle par* $\tilde{\sigma}^2_\varepsilon = 18.2$. *On estime donc le pourcentage de la variance de* Y *prédite par les* $m = 6$ *prédicteurs par* $\tilde{\rho}^2 = 1 - 18.2/65.1 = 72.0$ %. *Notons que l'estimateur empirique de ce pourcentage de variance est ici de* $\widehat{\rho}^2 = 72.6$ % *(voir plus loin).*

14.10 Tests sur la nullité de plusieurs paramètres

En régression multiple, on pourra non seulement tester la nullité d'une certaine pente dans le modèle, mais également la nullité simultanée de toutes les pentes. L'hypothèse nulle d'un tel test sera ainsi :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0.$$

Notons que si toutes les pentes étaient nulles, le modèle de régression se réduirait à la seule constante et on aurait avec un tel modèle 0 % de variance prédite. Cette hypothèse nulle peut donc également s'écrire :

$$H_0 : \rho^2 = 0.$$

Autrement dit, l'hypothèse nulle implique que l'ensemble des m prédicteurs du modèle ne nous sert à rien pour prédire Y.

[19] Dans la littérature, ces estimateurs $\widehat{\rho}^2$ et $\tilde{\rho}^2$ de ρ^2 sont souvent notés respectivement par R^2 et R^2_{adj} (en anglais : `R-square` et `adjusted R-square`). Notons qu'ils seront très proches l'un de l'autre en régression simple, raison pour laquelle on utilisera souvent $\widehat{\rho}^2$, qui a l'avantage d'être le carré de l'estimateur habituel de la corrélation entre les deux variables et de ne jamais être négatif (alors que $\tilde{\rho}^2$ pourra dans certains cas être légèrement négatif, bien que l'on ait par définition $\rho^2 \geq 0$). Il sera par contre important d'utiliser $\tilde{\rho}^2$ (au lieu de $\widehat{\rho}^2$) dans une régression multiple avec beaucoup de prédicteurs, car la surestimation de ρ^2 par $\widehat{\rho}^2$ pourra s'avérer grossière. Notons aussi que $\widehat{\rho}^2$ (au contraire de $\tilde{\rho}^2$) augmente systématiquement lors de l'introduction de nouveaux prédicteurs dans le modèle et atteindrait même $\widehat{\rho}^2 = 100$ % avec $m = n - 1$ prédicteurs (cas où $\tilde{\sigma}^2_\varepsilon$ et $\tilde{\rho}^2$ ne seraient pas définis).

Afin d'essayer de rejeter cette hypothèse nulle, on utilise la statistique de test suivante :

$$T_{stat} = \frac{n-m-1}{m} \cdot \frac{\widehat{\rho}^2}{1-\widehat{\rho}^2}.$$

Il se trouve que l'on connaît mathématiquement la distribution de T_{stat} sous H_0 (et sous les hypothèses d'un modèle de régression linéaire). Il s'agit en effet d'une distribution de Fisher (ou distribution F) avec paramètres $dln = m$ et $dld = n - m - 1$[20]. On rejette ainsi l'hypothèse nulle au seuil α si on a :

$$t_{stat} \geq F_{1-\alpha,m,n-m-1}$$

où $F_{1-\alpha,m,n-m-1}$ dénote le quantile $1-\alpha$ d'une distribution F avec paramètres $dln = m$ et $dld = n - m - 1$. Le tableau A.4 donné en annexe nous en donne quelques-uns (avec $\alpha = 0.05$). Il s'agit d'un test exact sous les hypothèses du modèle. Dans le cas du rejet de l'hypothèse nulle, on conclut qu'au moins une pente est non nulle (sans que l'on sache ni laquelle, ni le signe de cette pente) et donc que $\rho^2 > 0$. Dans le cas $m = 1$, ce test F est équivalent à un test de Student bilatéral sur la nullité de l'unique pente d'un modèle de régression simple[21].

Bien que cela ne soit pas essentiel à notre propos, notons qu'un logiciel statistique nous donnera parfois le résultat de ce test sous la forme d'un *tableau d'analyse de variance* de la forme suivante :

variabilité	somme des carrés	dl	moyenne des carrés
régression	$SS_{reg} = \sum_i (\widehat{y}_i^* - \bar{y})^2$	m	$MS_{reg} = \frac{SS_{reg}}{k}$
résiduelle	$SS_{res} = \sum_i (y_i - \widehat{y}_i^*)^2$	$n-m-1$	$MS_{res} = \frac{SS_{res}}{n-m-1} = \tilde{\sigma}_\varepsilon^2$
totale	$SS_{tot} = \sum_i (y_i - \bar{y})^2$	$n-1$	$MS_{tot} = \frac{SS_{tot}}{n-1} = \tilde{\sigma}_Y^2$

Dans ce tableau, on donne le détail de la décomposition de la variance empirique de la variable réponse (appelée ici la variance totale) en une somme de deux variances, la première prédite par le modèle de régression, la seconde résiduelle. La première colonne nous donne les numérateurs de ces variances empiriques, qui sont donc des sommes de carrés, notées SS_{reg}, SS_{res} et SS_{tot} (en anglais : `sum of squares`). La deuxième colonne nous donne le nombre de degrés de liberté dl par lequel il s'agit de diviser ces sommes de carrés pour obtenir des estimateurs sans biais des variances σ_ε^2 et σ_Y^2 (le nombre de degrés de liberté associé à la première somme de carrés étant obtenu par soustraction). La troisième colonne, obtenue en divisant la première par la deuxième, nous donne ce que l'on appelle ici des moyennes de carrés, notées MS_{reg}, MS_{res} et

[20] Nous avions introduit cette famille de distributions au chapitre 5 lorsqu'il s'agissait de faire de l'inférence sur le quotient des variances de deux distributions normales.

[21] Dans le cas $m = 1$, ce test F est également équivalent au test exact bilatéral sur la nullité de la corrélation entre l'unique prédicteur et la variable réponse introduit au chapitre 12. La statistique de test du test F est en effet égale au carré de la statistique de test du test sur la nullité de la corrélation, alors que l'on a par ailleurs $F_{1-\alpha,1,dl} = t^2_{1-\alpha/2,dl}$.

MS_{tot} (en anglais : `mean squares`). La statistique de test du test F peut alors se calculer comme suit :

$$T_{stat} = \frac{MS_{reg}}{MS_{res}}.$$

Exemple 14.11 *Dans notre exemple de la prédiction de la graisse corporelle, on obtient le tableau d'analyse de variance suivant :*

variabilité	*somme des carrés*	*dl*	*moyenne des carrés*
régression	11 674.9	6	1945.8
résiduelle	4396.4	241	18.2
totale	16 071.3	247	65.1

On a ainsi $\widehat{\rho}^2 = 11\ 674.9/16\ 071.3 = 72.6\ \%$. *Afin de tester l'hypothèse nulle*

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

on calcule la statistique de test suivante :

$$t_{stat} = \frac{241}{6} \cdot \frac{0.726}{1 - 0.726} = \frac{1945.8}{18.2} = 106.7.$$

que l'on compare avec $F_{0.95,6,241} = 2.14$. *Comme* $t_{stat} \geq F_{0.95,6,241}$, *on rejette l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera* $p < 0.0001$*)*[22]. *On conclut que les véritables pentes associées à nos six prédicteurs ne sont pas toutes nulles (on a prouvé statistiquement que* $\rho^2 > 0$, *c'est-à-dire que dans l'ensemble, nos* $m = 6$ *prédicteurs ne sont pas complètement inutiles pour prédire la variable réponse).*

Il est également possible de tester la nullité simultanée d'un sous-ensemble préspécifié de r pentes d'un modèle de régression linéaire (avec $1 \leq r \leq m$) à l'aide d'un *test du F partiel*. On pourra par exemple tester la nullité simultanée des r premières pentes, l'hypothèse nulle s'écrivant ainsi :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0.$$

Si ces pentes étaient nulles, le pourcentage de la variance prédite par le modèle de régression incluant l'ensemble des m prédicteurs ne serait pas plus élevé que le pourcentage de la variance prédite que l'on aurait après avoir éliminé ces r prédicteurs de notre modèle, que l'on notera $\rho^2(H_0)$. Autrement dit, l'hypothèse nulle implique que les r premiers prédicteurs n'améliorent pas la prédiction de Y (sachant que l'on a déjà les autres $m - r$ prédicteurs dans le modèle). On utilise ici la statistique de test suivante :

$$T_{stat} = \frac{n - m - 1}{r} \cdot \frac{\widehat{\rho}^2 - \widehat{\rho}^2(H_0)}{1 - \widehat{\rho}^2}$$

[22] Dans R, on pourra ici calculer la valeur p par la commande `1-pf(106.7,6,241)`.

où $\widehat{\rho}^2(H_0)$ dénote l'estimateur empirique de $\rho^2(H_0)$. La distribution de T_{stat} sous H_0 (et sous les hypothèses d'un modèle de régression linéaire) est une distribution F avec paramètres $dln = r$ et $dld = n - m - 1$. On rejette ainsi l'hypothèse nulle au seuil α si on a :

$$t_{stat} \geq F_{1-\alpha,r,n-m-1}.$$

Il s'agit d'un test exact sous les hypothèses du modèle. Dans le cas $r = m$, ce test sera identique au test F ci-dessus pour tester la nullité simultanée de toutes les pentes du modèle (on aura alors $\widehat{\rho}^2(H_0) = 0$). Dans le cas $r = 1$, ce test sera équivalent à un test de Student bilatéral pour tester la nullité de la pente considérée.

Exemple 14.12 *On reprend l'exemple de la prédiction de la graisse corporelle. On aimerait tester si l'introduction du poids, de l'âge et de la taille dans un modèle où l'on aurait déjà comme prédicteurs l'abdomen, le biceps et le poignet, nous permet d'améliorer la prédiction de la graisse corporelle. L'hypothèse nulle est ainsi :*

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

En utilisant l'ensemble des $m = 6$ prédicteurs, on avait $\widehat{\rho}^2 = 0.726$. En éliminant les $r = 3$ premiers d'entre eux, on obtient $\widehat{\rho}^2(H_0) = 0.702$. On calcule alors la statistique de test suivante :

$$t_{stat} = \frac{241}{3} \cdot \frac{0.726 - 0.702}{1 - 0.726} = 7.04.$$

que l'on compare avec $F_{0.95,3,241} = 2.64$. Comme $t_{stat} \geq F_{0.95,3,241}$, on rejette l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera $p = 0.0001$)[23]*. On conclut que l'introduction (simultanée) du poids, de l'âge et de la taille dans le modèle de régression permet d'améliorer significativement (bien qu'ici modestement) le pourcentage de variance prédite et donc la qualité globale de la prédiction de la graisse corporelle.*

14.11 Multicolinéarité

Comme écrit dans un chapitre précédent, il est toujours instructif de regarder de plus près la formule de la variance d'un estimateur. En ce qui concerne la variance de l'estimateur $\widehat{\beta}_j$ de la pente associée au j-ième prédicteur, on a le résultat intéressant suivant (pour $j = 1, \cdots, m$) :

$$\mathrm{Var}(\widehat{\beta}_j) = \frac{\sigma_\varepsilon^2}{n\widehat{\sigma}_{X_j}^2} \cdot \frac{1}{1 - \widehat{\rho}_j^2}$$

où $\widehat{\sigma}_{X_j}^2$ représente la variance empirique de X_j et $\widehat{\rho}_j^2$ représente l'estimateur empirique du pourcentage de la variance de X_j prédite linéairement par les

[23] Dans R, on pourra ici calculer la valeur p par la commande `1-pf(7.04,3,241)`.

autres $m-1$ prédicteurs (calculée à partir d'un modèle de régression avec X_j comme variable réponse et avec $m-1$ prédicteurs). Cette quantité $1/(1-\widehat{\rho}_j^2)$ est appelée *le facteur d'inflation de la variance* dû aux corrélations entre les prédicteurs, autrement dit, à ce qu'on appelle la *multicolinéarité*. Dans le cas d'un seul prédicteur ($m=1$), on aura $\widehat{\rho}_j^2 = 0$ et ce facteur sera égal à 1 (on retrouve alors la formule habituelle de l'erreur type de la pente d'une droite de régression). Dans le cas extrême où X_j est une combinaison exacte linéaire des autres prédicteurs, on aura $\widehat{\rho}_j = 1$, de sorte que la variance de l'estimateur de la pente sera infinie (en fait, l'hyperplan de régression ne pourra dans ce cas pas être calculé). D'une manière générale, plus un prédicteur sera corrélé avec les autres prédicteurs, moins l'estimation de sa pente sera précise. La multicolinéarité entre prédicteurs implique donc le problème suivant :

Il est difficile d'estimer précisément une association entre X_1 et Y ajustée pour des variables confondantes $X_2, \cdots, X_m$ très corrélées avec X_1.

On verra dans la section suivante un autre problème impliqué par la multicolinéarité entre prédicteurs :

Introduire des prédicteurs très corrélés dans un modèle de régression nuit à la qualité de la prédiction.

Exemple 14.13 *Nous présentons un exemple impliquant $n = 44$ adultes de sexe masculin qui ont participé à un régime*[24]*. On essaie ici de prédire la perte de poids (en kg) Y connaissant la perte de glucose (en mg/dL) X_1 et la perte de cholestérol (en mg/dL) X_2 mesurées durant ce régime. Les relations entre X_1 et Y et entre X_2 et Y sont globalement positives et approximativement linéaires, comme le montrent les graphiques (a) et (b) de la figure 14.10. Un modèle de régression simple avec X_1 comme prédicteur nous donne :*

$$\widehat{y}^* = 3.56 + 0.0490x_1$$

avec une erreur type pour l'estimateur de la pente de 0.0153 *($p = 0.003$). Un modèle de régression simple avec X_2 comme prédicteur nous donne :*

$$\widehat{y}^* = 3.60 + 0.0812x_2$$

[24]Il s'agit de données récoltées dans le cadre d'une étude publiée par Cocco *et al.* (2005), bien que la problématique que nous regardons ici n'ait pas été abordée dans cette publication.

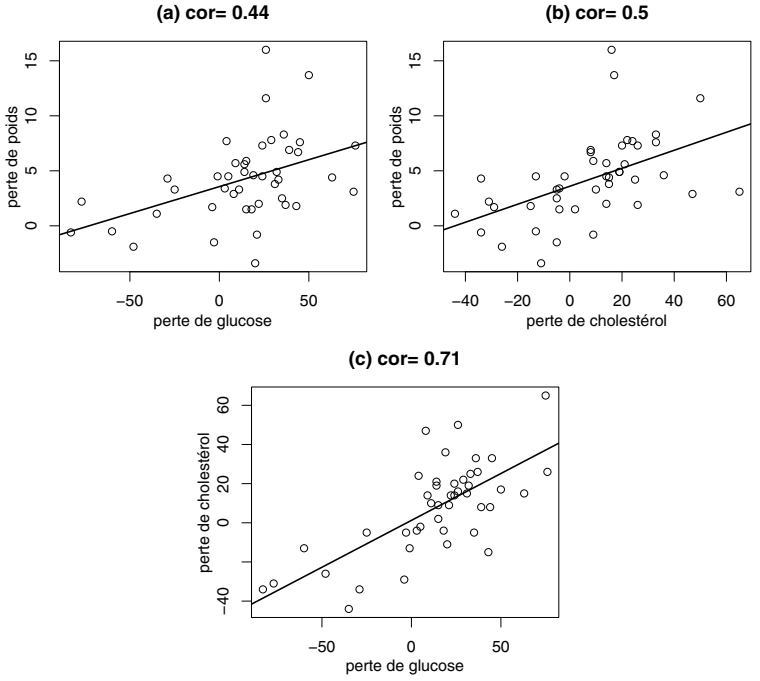

Figure 14.10 – Exemple de données avec deux prédicteurs corrélés.

avec une erreur type pour l'estimateur de la pente de 0.0212 *(*$p = 0.0005$*). Dans les deux cas, on a donc un résultat significatif. Si on considère à présent un modèle de régression multiple avec* X_1 *et* X_2 *comme prédicteurs, on trouve :*

$$\widehat{y}^* = 3.49 + 0.0196x_1 + 0.0614x_2$$

avec des erreurs types pour les estimateurs des pentes de 0.0209 *(*$p = 0.36$*) et de* 0.0308 *(*$p = 0.053$*), respectivement pour* X_1 *et* X_2*. Par rapport à la régression simple, les estimations des pentes ont diminué, les erreurs types ont augmenté et les pentes ne sont plus significatives.*

Cette diminution des estimations des pentes et cette augmentation des erreurs types sont dues à la corrélation non négligeable de 0.71 *entre les deux prédicteurs, que l'on peut voir sur le graphique (c) de la figure 14.10. En effet, les individus qui ont beaucoup perdu de glucose ont en moyenne également beaucoup perdu de cholestérol. Dans ces conditions, les pentes du modèle de la régression multiple sont difficiles à estimer. On notera également que le test* F *sur la nullité simultanée des deux pentes dans notre modèle de régression multiple est significatif (*$p = 0.002$*), bien que les deux tests de Student sur la nullité de chacune des pentes ne le soient pas (*$p = 0.36$ *et* $p = 0.053$*). Le résultat du test* F *nous dit que l'on a prouvé statistiquement que la perte de glucose et*

la perte de cholestérol ne sont dans l'ensemble pas complètement inutiles pour prédire la perte de poids. Les résultats des tests de Student nous disent que l'on n'a pas réussi à prouver statistiquement que l'introduction de l'un de ces prédicteurs dans le modèle nous permet d'améliorer la prédiction sachant que l'autre prédicteur est déjà dans le modèle. Autrement dit, on a prouvé statistiquement une association entre ces prédicteurs et la réponse, mais on ne peut pas exclure que ces deux prédicteurs soient redondants (un seul suffirait). En fait, si notre critère était la qualité de la prédiction, on choisirait le modèle avec la perte de cholestérol comme unique prédicteur (comme on va le voir un peu plus loin).

14.12 Intervalle de prédiction

Sous les hypothèses d'un modèle de régression linéaire, on pourra non seulement effectuer des prédictions ponctuelles mais également calculer des intervalles de prédiction, comme on l'a fait en régression simple. Dans ce cas, on ne se contente donc pas de prédire la valeur Y d'un individu dont on connaît les caractéritiques $x_1, x_2, \cdots, x_m$ par rapport aux prédicteurs $X_1, X_2, \cdots, X_m$, mais on calcule un intervalle de prédiction qui contiendra cette valeur Y avec une probabilité de (approximativement) 0.95 de la manière suivante :

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_m x_m \pm 2\tilde{\sigma}_\varepsilon.$$

Comme en régression simple, on peut quantifier par :

$$PIR = 1 - \frac{\sigma_\varepsilon}{\sigma_Y} = 1 - \sqrt{1-\rho^2}$$

la réduction de la longueur de l'intervalle de prédiction obtenue grâce à notre modèle de régression (en comparaison avec un intervalle de prédiction calculé sans ce modèle de régression).

Exemple 14.14 *On reprend l'exemple de la prédiction de la graisse corporelle. Dans une section précédente, pour une personne pesant* 80 *kg, âgée de* 40 *ans, mesurant* 175 *cm, avec un abdomen de* 115 *cm, un biceps de* 30 *cm et un poignet de* 19 *cm, on avait prédit (en utilisant notre équation de régression) un pourcentage de graisse corporelle de :*

$$-13.3 - 0.12 \cdot 80 + 0.046 \cdot 40 - 0.077 \cdot 175 + 0.84 \cdot 115 + 0.23 \cdot 30 - 1.73 \cdot 19 = 36.1.$$

En rappelant que l'on avait estimé la variance résiduelle par $\tilde{\sigma}_\varepsilon^2 = 18.2$*, on peut alors calculer l'intervalle de prédiction suivant :*

$$36.1 \pm 2\sqrt{18.2} = [27.6; 44.6].$$

Sous les hypothèses du modèle, le pourcentage de graisse corporelle de cet individu sera donc avec une probabilité de 0.95 *entre* 27.6 % *et* 44.6 %.

Si on devait prédire la graisse corporelle d'un individu sans connaître ses caractéristiques ci-dessus, c'est-à-dire sans utiliser ce modèle de régression, on utiliserait des estimations de la moyenne et de la variance de cette variable, avec dans notre cas $\widehat{\mu}_Y = 18.9$ *et* $\tilde{\sigma}^2_Y = 65.1$. *En supposant la normalité de* Y, *on calculerait l'intervalle de prédiction suivant :*

$$18.9 \pm 2\sqrt{65.1} = [2.8; 35.0]$$

qui est plus large que le premier. La réduction de la longueur de l'intervalle de prédiction obtenue grâce à notre modèle de régression correspond à un PIR *estimé de* $1 - \sqrt{18.2/65.1} = 0.47$. *On aura ainsi réussi à diminuer de* 47 % *(c'est-à-dire pratiquement d'un facteur* 2*) la longueur de l'intervalle de prédiction.*

Les bornes inférieure et supérieure d'un intervalle de prédiction au niveau 95 % sont en fait des estimations des quantiles 2.5 % et 97.5 % de la variable réponse Y pour une certaine combinaison de valeurs possibles des prédicteurs. Plus généralement, on pourra grâce à notre modèle de régression linéaire estimer n'importe quel quantile de cette distribution. On s'intéressera à ces quantiles lorsqu'il s'agira d'établir des normes d'une variable Y en fonction d'autres variables (les prédicteurs), comme illustré dans l'exemple suivant.

Exemple 14.15 *On revient sur notre exemple de motricité où il s'agissait de modéliser le temps* Y *nécessaire pour accomplir un exercice en fonction de l'âge* X_1 *à partir d'un échantillon de* $n = 593$ *enfants et adolescents. On avait log-transformé la variable* Y, *modélisant ainsi* $Z = \log(Y)$, *et on avait considéré un modèle quadratique, obtenant :*

$$\widehat{z}^* = 3.745 - 0.239x_1 + 0.00706x_1^2.$$

On estime par ailleurs pour ce modèle une variance résiduelle de $\tilde{\sigma}^2_\varepsilon = 0.0676$. *On peut ainsi estimer le quantile* $q\%$ *de la variable* $Z = \log Y$ *à un âge* x_1 *(en log secondes) par :*

$$3.745 - 0.239x_1 + 0.00706x_1 + z_{q/100} \cdot \sqrt{0.0676}.$$

À partir de là, on peut estimer le quantile $q\%$ *de la variable originale* Y *à un âge* x_1 *(en secondes) par*[25] *:*

$$\exp(3.745 - 0.239x_1 + 0.00706x_1 + z_{q/100} \cdot \sqrt{0.0676}).$$

On estime par exemple les quantiles 2.5 %, 10 %, 25 %, 50 %, 75 %, 90 % *et*

[25] On utilise ici le fait que $quantile(\log(Y)) = \log(quantile(Y))$, ce qui implique $\exp(quantile(\log(Y))) = quantile(Y)$.

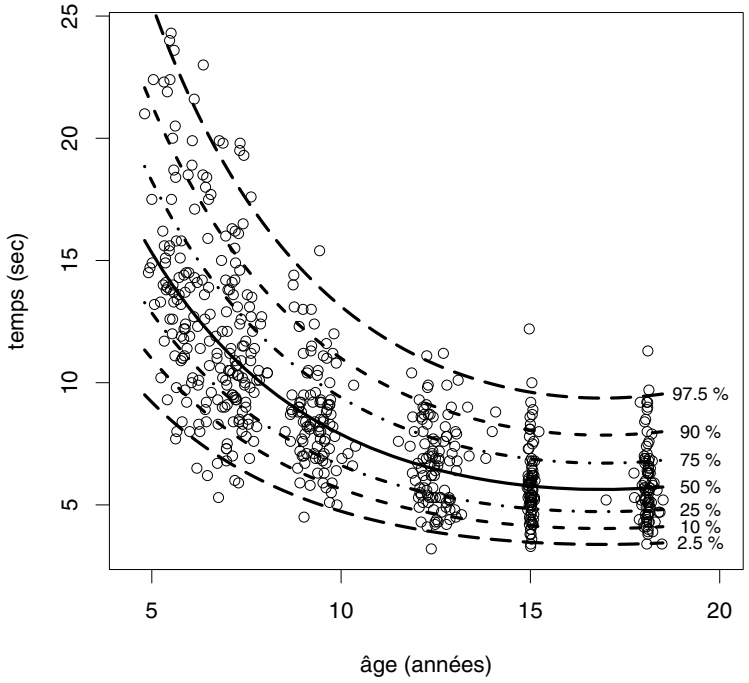

Figure 14.11 – Exemple de normes calculées à l'aide d'un modèle de régression.

97.5 % *de la variable* Y *à l'âge* $x_1 = 10$ *ans (en secondes) par :*

$$\begin{aligned}
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 - 1.96 \cdot \sqrt{0.0676}) &= 4.7 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 - 1.28 \cdot \sqrt{0.0676}) &= 5.6 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 - 0.67 \cdot \sqrt{0.0676}) &= 6.6 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 + 0.00 \cdot \sqrt{0.0676}) &= 7.9 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 + 0.67 \cdot \sqrt{0.0676}) &= 9.3 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 + 1.28 \cdot \sqrt{0.0676}) &= 11.0 \\
\exp(3.745 - 0.239 \cdot 10 + 0.00706 \cdot 10^2 + 1.96 \cdot \sqrt{0.0676}) &= 13.1.
\end{aligned}$$

De tels quantiles calculés entre 5 *et* 18 *ans sont présentés dans la figure 14.11 (on y voit également les données utilisées pour cette modélisation). Il sera ainsi possible de comparer via ces normes les performances de deux enfants d'âge différent ou celles d'un même enfant à deux âges différents (afin de pouvoir juger si par exemple un enfant en retard dans son développement a fait des progrès ou non).*

Nous terminerons cette section en mentionnant que la formule d'un intervalle de prédiction au niveau 95 % donnée ci-dessus n'est en fait qu'une approxi-

mation. Elle ne tient pas compte de l'incertitude liée au fait que les paramètres de l'hyperplan de régression, de même que la variance résiduelle, ne sont pas connus mais estimés. En fait, la probabilité que la valeur Y de l'individu pour lequel on fait la prédiction se trouve à l'intérieur de cet intervalle sera un peu plus petite que 0.95. Sous les hypothèses du modèle, il existe cependant une formule explicite pour calculer un intervalle de prédiction exact[26]. La longueur d'un intervalle de prédiction approximatif est de $4\tilde{\sigma}_\varepsilon$. La longueur d'un intervalle de prédiction exact dépendra légèrement des caractéristiques de l'individu pour lequel on fait la prédiction et sera égale « en moyenne » à :

$$2t_{0.975,n-m-1} \cdot \tilde{\sigma}_\varepsilon \cdot \sqrt{1+(m+1)/n}.$$

L'introduction de nouveaux prédicteurs dans le modèle aura ainsi les conséquences suivantes :

- une diminution de l'écart type résiduel σ_ε ainsi que (le plus souvent) de son estimation $\tilde{\sigma}_\varepsilon$[27]
- une augmentation du facteur $\sqrt{1+(m+1)/n}$
- une augmentation de la valeur du quantile $t_{0.975,n-m-1}$, négligeable cependant si n est suffisamment grand.

Dans le but d'améliorer la prédiction (c'est-à-dire ici de diminuer la longueur des intervalles de prédiction), il s'agira d'introduire un nouveau prédicteur dans le modèle si et seulement si la diminution de $\tilde{\sigma}_\varepsilon$ compense l'augmentation du facteur $\sqrt{1+(m+1)/n}$. En particulier, l'introduction de nouveaux prédicteurs très corrélés avec ceux déjà dans le modèle sera nuisible à la prédiction (car on augmentera m et donc $\sqrt{1+(m+1)/n}$ sans beaucoup diminuer $\tilde{\sigma}_\varepsilon$).

14.13 Choix du modèle

Dans une situation avec m prédicteurs potentiels à disposition, on pourra choisir d'inclure ou de ne pas inclure chacun de ces prédicteurs dans notre

[26] En reprenant notre notation matricielle, un intervalle de prédiction exact au niveau $1-\alpha$ pour la valeur Y d'un individu avec les caractéristiques $x_1, \cdots, x_m$ par rapport aux prédicteurs $X_1, \cdots, X_m$ est donné par :

$$x\widehat{\beta} \pm t_{1-\alpha/2,n-m-1} \cdot \tilde{\sigma}_\varepsilon \sqrt{1+x^T(X^TX)^{-1}x}$$

où $x^T = (1, x_1, \cdots, x_m)$. Cela veut dire que si on répétait l'échantillonnage et si dans chaque échantillon on estimait notre modèle de régression et on calculait cet intervalle de prédiction, exactement 95 % d'entre eux contiendraient la valeur Y de cet individu.

[27] En fait, l'estimateur sans biais $\tilde{\sigma}^2_\varepsilon$ de la variance résiduelle σ^2_ε ne diminue pas forcément lorsque l'on introduit de nouveaux prédicteurs dans le modèle, au contraire de la véritable variance résiduelle σ^2_ε ou de son estimateur empirique $\widehat{\sigma}^2_\varepsilon$ (de même, l'estimateur corrigé $\tilde{\rho}^2$ du pourcentage de variance prédite ρ^2 n'augmente pas forcément lorsque l'on introduit de nouveaux prédicteurs dans le modèle, au contraire du véritable pourcentage de variance prédite ρ^2 ou de son estimateur empirique $\widehat{\rho}^2$).

modèle, ce qui définit 2^m modèles linéaires différents. Admettons que chacun d'entre eux satisfasse (approximativement) les hypothèses d'un modèle de régression linéaire, lequel doit-on choisir en pratique ? La réponse à cette question dépend de notre motivation à effectuer une régression multiple. Voici trois motivations possibles, en lien avec les trois questions énoncées en début de chapitre :

1. On aimerait **prédire au mieux** Y (c'est-à-dire minimiser la longueur des intervalles de prédiction) en fonction des prédicteurs à disposition.

2. On aimerait **estimer une association entre** Y **et un prédicteur** X_1 **ajustée pour les autres prédicteurs** (considérés comme des variables confondantes).

3. On aimerait **identifier un modèle avec uniquement des prédicteurs dont on a prouvé statistiquement qu'ils ne sont ni inutiles, ni redondants pour prédire** Y (afin d'obtenir une description parcimonieuse de nos données).

On choisira alors le modèle de la façon suivante :

- dans le premier cas, une stratégie possible consiste à choisir le modèle qui minimise la quantité $\tilde{\sigma}_\varepsilon \cdot \sqrt{1+(m+1)/n}$ discutée en fin de section précédente[28]

  → les valeurs p associées aux pentes des prédicteurs dans un tel modèle seront toutes plus petites que 0.157, du moins pour un grand n[29]

  → afin d'optimiser la prédiction, on pourra donc avoir dans notre modèle des prédicteurs qui ne sont pas significatifs au seuil de 5 % (mais qui sont significatifs au seuil de 15.7 %)

- dans le deuxième cas, c'est à nous de décider (en principe sans avoir regardé les données au préalable) quels sont les prédicteurs (quelles sont les variables confondantes) à introduire dans le modèle

  → on gardera dans notre modèle les prédicteurs jugés confondants indépendamment des valeurs p

[28] Minimiser cette quantité est équivalent (pour de grands n) à minimiser le critère dit AIC (`Akaike Information Criterion`) introduit par Hirotsugu Akaike en 1973. On pourra lire à ce sujet le livre de Burnham et Anderson (2004) ou l'article de Rousson et Goşoniu (2007).

[29] Pour se convaincre de cela, on montrera tout d'abord à l'aide d'algèbre élémentaire que l'introduction d'un prédicteur supplémentaire dans le modèle diminue notre critère $\tilde{\sigma}_\varepsilon \cdot \sqrt{1+(m+1)/n}$ si et seulement si $t_{stat} \geq 2n/(n+m)$, où t_{stat} dénote la statistique de test d'un test du F partiel sur la nullité de la pente du nouveau prédicteur. Pour un grand n, on diminuera donc notre critère si $t_{stat} \geq 2$. Rappelons que l'on compare t_{stat} à une distribution F avec paramètres $dln = 1$ et $dld = n - m - 1$. Or, pour un grand n, cette distribution sera proche d'une distribution du khi-deux avec 1 dl. Par ailleurs, la valeur 2 correspond au quantile 84.3 % d'une telle distribution. Dire que $t_{stat} \geq 2$ revient donc à dire que $p \leq 0.157$.

- dans le troisième cas, on choisit généralement un modèle où chaque prédicteur est significatif au seuil de 5 %.

Exemple 14.16 *On reprend l'exemple des individus qui ont suivi un régime, avec comme variable réponse Y la perte de poids, et comme prédicteurs potentiels X_1 la perte de glucose et X_2 la perte de cholestérol. Si notre but est de prédire au mieux la perte de poids des individus, on choisira le modèle avec la perte de cholestérol comme seul prédicteur (car on obtient $p = 0.36$ pour la perte de glucose lorsqu'on l'introduit dans ce modèle, ce qui est largement au-dessus de 0.157). Par contre, si notre but est d'étudier l'association entre la perte de cholestérol et la perte de poids, et que l'on juge la perte de glucose comme étant une variable confondante pour cette association, alors on n'aura pas d'autre choix que le modèle incluant les deux prédicteurs.*

Si notre motivation est d'optimiser la prédiction, une stratégie possible consiste à calculer les 2^m modèles envisageables et à choisir celui pour lequel notre critère $\tilde{\sigma}_\varepsilon \cdot \sqrt{1 + (m+1)/n}$ est minimisé. Afin d'économiser du temps de calcul, on ne considère parfois qu'un sous-ensemble de ces 2^m modèles. Une méthode souvent utilisée consiste à calculer tout d'abord le modèle avec l'ensemble des m prédicteurs, puis à éliminer le prédicteur associé à la plus grande valeur p et à continuer ainsi jusqu'à ce que toutes les valeurs p soient plus petites que 0.157. En anglais, il s'agit de la méthode `backward elimination`[30]. En utilisant cette méthode, on n'est pas sûr d'identifier le meilleur modèle, mais on trouvera souvent « un bon modèle ». On pourra utiliser la même méthode en utilisant un seuil de 0.05 au lieu de 0.157 si notre motivation est l'identification d'un ensemble de prédicteurs importants (ni inutiles, ni redondants). Par contre, une telle méthode ne doit pas être utilisée si notre motivation est l'estimation d'une association ajustée pour les variables confondantes (où le choix des prédicteurs à inclure dans le modèle dépend de la question scientifique posée et non du résultat de tests statistiques).

Exemple 14.17 *Appliquons la méthode « backward elimination » décrite ci-dessus à notre exemple de la prédiction de la graisse corporelle. Dans le modèle avec les $m = 6$ prédicteurs, le prédicteur associé à la plus grande valeur p était la taille ($p = 0.240$). Si on élimine la taille, on obtient le modèle suivant :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−28.080	7.764	−3.617	0.000
poids	−0.196	0.073	−2.671	0.008
âge	0.044	0.029	1.527	0.128
abdomen	0.899	0.070	12.755	0.000
biceps	0.292	0.152	1.913	0.057
poignet	−1.721	0.495	−3.479	0.001

[30] Il existe de nombreuses variantes de cette méthode, par exemple, la méthode dite en anglais `forward selection`.

Dans ce modèle, tous les prédicteurs ont une valeur $p \leq 0.157$*, de sorte que l'on choisirait ce modèle si le but était d'optimiser la prédiction (de minimiser la longueur des intervalles de prédiction). Par contre, certaines valeurs p sont plus grandes que* 0.05*. Si notre but était d'identifier un ensemble de prédicteurs qui sont tous significatifs, on éliminerait l'âge (*$p = 0.128$*) et on obtiendrait :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−32.895	7.114	−4.624	0.000
poids	−0.258	0.061	−4.244	0.000
abdomen	0.963	0.057	17.038	0.000
biceps	0.277	0.153	1.818	0.070
poignet	−1.373	0.440	−3.119	0.002

À partir de là, on éliminerait encore le biceps ($p = 0.070$*) pour obtenir :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−29.169	6.844	−4.262	0.000
poids	−0.204	0.053	−3.827	0.000
abdomen	0.956	0.057	16.870	0.000
poignet	−1.286	0.440	−2.926	0.004

On a ici identifié un modèle avec trois prédicteurs significatifs. Si ce modèle avait été préspécifié, on aurait ici prouvé statistiquement que chacun de ces trois prédicteurs améliore la prédiction par rapport aux deux autres (et donc que ces trois prédicteurs ne sont ni inutiles ni redondants)[31].

14.14 Valeurs aberrantes et points leviers

On va s'intéresser dans cette dernière section à la distribution des résidus empiriques $\widehat{\varepsilon}_i$. Sous les hypothèses du modèle, ils sont d'espérance nulle, normalement distribués, alors que leur variance est donnée par :

$$\mathrm{Var}(\widehat{\varepsilon}_i) = \sigma_\varepsilon^2(1 - h_i)$$

[31] D'une manière générale, ces méthodes de sélection de modèle sont souvent critiquées. Un problème est celui des tests multiples. Même si aucun des m prédicteurs potentiels n'est utile pour prédire Y, la probabilité de sélectionner un modèle contenant au moins l'un de ces prédicteurs sera bien plus grande que le seuil de 0.05 (si m est grand). Un autre problème est que les méthodes d'inférence que nous avons présentées ne sont plus tout à fait valables lorsque le modèle est sélectionné en compétition avec d'autres modèles. En effet, bien que nos estimateurs soient sans biais, on aura tendance à choisir un modèle qui est meilleur dans notre échantillon que dans la population (les modèles qui sont moins bons dans notre échantillon que dans la population seront moins souvent choisis). Du coup, on aura tendance à surestimer la performance du modèle choisi (les valeurs absolues des pentes et le pourcentage de variance prédite seront surestimés). Pour remédier à cela, une méthode possible (mais coûteuse en puissance statistique) consiste à diviser l'échantillon en deux parties, la première partie étant utilisée pour la sélection du modèle, la seconde pour l'estimation et l'inférence.

où h_i est une quantité comprise entre 0 et 1[32]. Dans un exemple précédent, nous avions calculé des *résidus standardisés* définis par $\widehat{\varepsilon}_i/\tilde{\sigma}_\varepsilon$, en mentionnant qu'ils avaient (approximativement) une distribution normale standardisée, ce qui nous permet de détecter des valeurs extrêmes ou aberrantes. Le résultat ci-dessus nous suggère une standardisation plus fine des résidus, à savoir $\widehat{\varepsilon}_i/(\tilde{\sigma}_\varepsilon\sqrt{1-h_i})$ (parfois appelés *résidus studentisés*).

Ainsi, la variance des résidus empiriques n'est pas la même pour tous. Elle est d'autant plus petite que h_i est grand. Une petite variance pour $\widehat{\varepsilon}_i$ indique que le i-ième résidu empirique varierait peu d'un échantillon à l'autre (si on répétait l'échantillonnage). Une petite variance, associée à une espérance nulle, signifie par ailleurs que ce résidu empirique sera proche de zéro. Ainsi, ce résidu attirera l'hyperplan de régression à lui, quelle que soit la valeur y_i observée. Autrement dit, les observations avec une grande valeur h_i sont des observations qui ont une grande influence sur l'estimation de l'hyperplan de régression. On appelle de telles observations des *points leviers* (en anglais : `leverage points`). Un point levier est une observation loin des autres observations dans l'espace à m dimensions défini par les prédicteurs. On a par ailleurs $\sum_i h_i = m+1$, la moyenne des h_i étant égale à $(m+1)/n$. En pratique, une observation est parfois considérée comme un point levier si sa valeur h_i vaut plus du double de cette moyenne, c'est-à-dire si $h_i > 2(m+1)/n$.

Un point levier aura une mauvaise influence sur l'estimation de l'hyperplan de régression s'il est par ailleurs également une valeur aberrante (c'est-à-dire une observation qui ne devrait pas faire partie de la population décrite par le modèle linéaire). On notera par contre qu'une observation aberrante qui n'est pas un point levier n'aura pas une grande influence sur l'estimation des paramètres de l'hyperplan de régression. Elle en aura cependant sur l'estimation de la variance résiduelle et des erreurs types, qui seront surestimées, ce qui nous coûtera de la puissance statistique.

Exemple 14.18 *La figure 14.12 illustre l'influence possible d'un point levier en régression simple (les mêmes principes étant valables en régression multiple). Dans les deux graphiques de cette figure, on a une observation aberrante qui ne fait pas partie de la même population que les autres observations (pour lesquelles un modèle linéaire est adéquat). Dans le graphique du haut, cette observation aberrante n'est cependant pas un point levier (se trouvant en plein milieu des autres observations par rapport au prédicteur). Dans le graphique du bas, l'observation aberrante est par contre un point levier (se trouvant loin des autres, tout à droite du graphique). On a dans les deux cas calculé la droite de régression en utilisant l'ensemble des observations (trait plein) et en excluant cette observation aberrante (trait pointillé). On voit de quelle manière une observation aberrante qui est aussi un point levier influence dramatiquement l'estimation de la droite de régression (graphique du bas), alors qu'une observation aberrante qui n'est pas un point levier n'a que peu d'influence sur l'estimation de cette droite (graphique du haut).*

[32] On aura $h_i = x_i^T(X^TX)^{-1}x_i$ où $x_i^T = (1, x_{i1}, \cdots, x_{im})$ $(i = 1, \cdots, n)$.

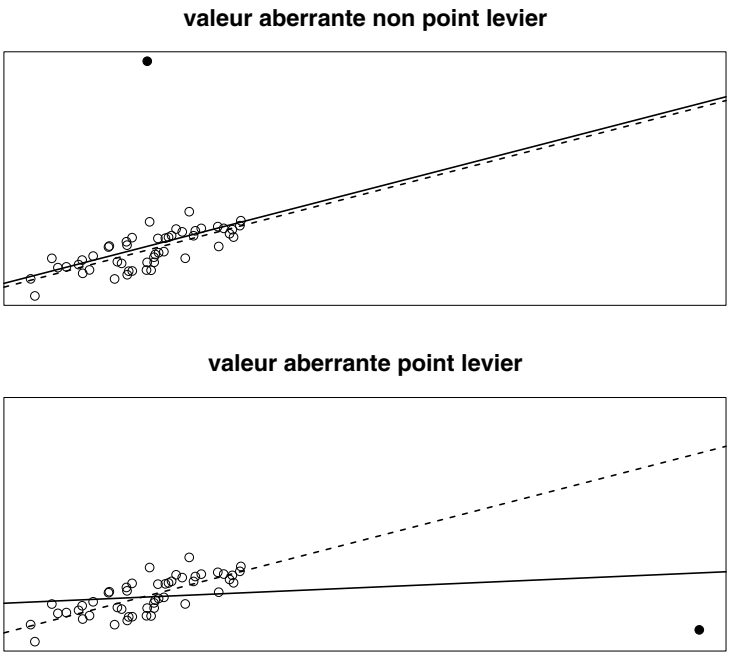

Figure 14.12 – Influence d'une observation aberrante sur la régression.

Exemple 14.19 *Nous allons pour la dernière fois considérer l'exemple de la prédiction de la graisse corporelle. La figure 14.13 nous montre les résidus studentisés (sur l'axe vertical) et les quantités h_i (sur l'axe horizontal) que l'on obtient en calculant un modèle de régression linéaire incluant l'ensemble des $m = 6$ prédicteurs. Sur l'axe vertical, on voit peu de résidus extrêmes (la plupart se trouvent comme il se doit entre les limites -2 et $+2$, représentées par des droites horizontales). Sur l'axe horizontal, on voit par contre un certain nombre d'observations qui dépassent la limite autorisée de $h_i = 2 \cdot 7/248 = 0.056$ (représentée par une droite verticale). Il y a notamment un individu avec une très grosse valeur de h_i qui a donc une grande influence sur l'estimation de l'hyperplan de régression. Il s'agit d'un individu loin des autres dans l'espace défini par les $m = 6$ prédicteurs, avec notamment un gros biceps mais un petit poids et un petit abdomen (on le distingue sur certains des diagrammes de dispersion de la figure 14.1). Il serait en principe plus prudent de refaire les calculs sans cette observation. Se faisant, on obtient le résultat suivant :*

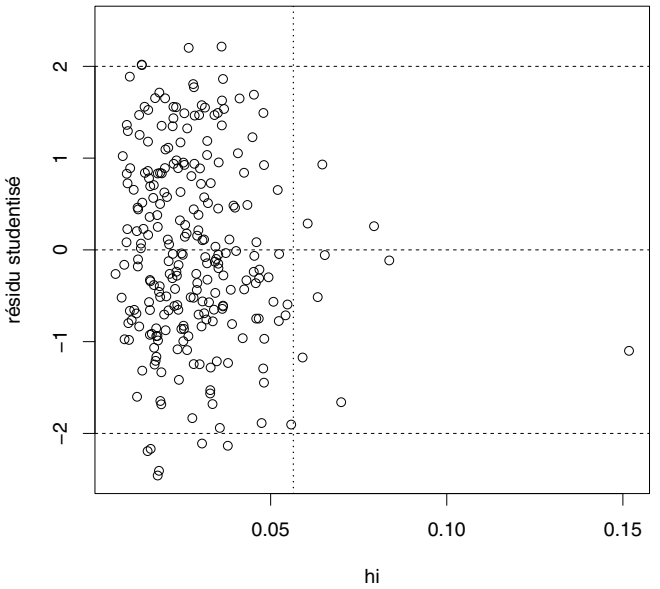

Figure 14.13 – Exemple de détection des points leviers.

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−17.167	15.280	−1.124	0.262
poids	−0.145	0.100	−1.443	0.150
âge	0.047	0.029	1.625	0.105
taille	−0.061	0.067	−0.907	0.365
abdomen	0.850	0.086	9.886	0.000
biceps	0.294	0.171	1.722	0.086
poignet	−1.708	0.495	−3.454	0.001

Ainsi, l'élimination d'un seul individu, si elle n'a pas changé dramatiquement les résultats, a tout de même modifié certains d'entre eux, notamment l'estimation de la pente du biceps (qui était de 0.234 *et qui est désormais de* 0.294*), mais également certaines valeurs p dont celle de la taille, qui est à présent nettement la plus grande d'entre toutes (ce qui peut avoir son importance si on utilise un algorithme du type « backward elimination » pour choisir un modèle).*

Chapitre 15

Régression avec prédicteurs binaires

On a introduit la régression comme une méthode permettant de prédire une variable réponse continue à partir d'un ensemble de prédicteurs qui étaient dans tous nos exemples également des variables continues. D'une manière générale cependant, les prédicteurs dans un modèle de régression linéaire doivent être des variables quantitatives mais pas forcément des variables continues. Dans ce chapitre, nous allons présenter des modèles de régression qui incluent des prédicteurs binaires (cas particuliers de variables quantitatives). Ceci nous permettra d'introduire également les méthodes connues sous le nom d'*analyse de variance* et d'*analyse de covariance*. Comme d'habitude, les deux valeurs possibles d'une variable binaire seront codées par 1 et 0.

15.1 Comparaison de deux groupes

On considère dans cette section un modèle de régression linéaire simple avec un prédicteur X binaire (la variable réponse Y étant comme d'habitude continue). Dans un cadre général, on a vu que la régression nous permet de comparer la moyenne de Y entre des groupes définis par les différentes valeurs possibles des prédicteurs. Avec un seul prédicteur binaire, la comparaison se réduit à deux groupes, le groupe défini par $X = 1$ et le groupe défini par $X = 0$. Comme il existe toujours une droite qui passe par deux points, l'hypothèse de linéarité sera dans ce cas forcément satisfaite. La droite de régression :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

sera donc celle qui passe par les moyennes des deux groupes. L'interprétation des paramètres se fait comme d'habitude de la manière suivante :

- **pour la constante** : $\beta_0 = mean(Y|X = 0)$

  $\rightarrow$ la constante est la moyenne de Y dans le groupe (de référence) $X = 0$

- **pour la pente** : $\beta_1 = mean(Y|X = 1) - mean(Y|X = 0)$

  $\rightarrow$ la pente est la différence de moyenne de Y entre les deux groupes.

Un modèle de régression linéaire est donc ici constitué uniquement de trois hypothèses :

- on suppose que la variance de Y est la même dans les deux groupes (homoscédasticité)
- on suppose que la distribution de Y est normale dans les deux groupes
- on suppose que les observations sont indépendantes.

On retrouve ainsi les mêmes hypothèses que celles faites dans le cadre du *modèle idéal* introduit au chapitre 5.

> Le modèle idéal (comparaison de deux distributions normales et de même variance) est un cas particulier de modèle de régression linéaire.

De même, les tests et intervalles de confiance de Student pour une différence de moyenne sont équivalents à des tests et intervalles de confiance de Student pour la pente d'une droite de régression avec prédicteur binaire. En effet, la droite de régression $\widehat{\beta}_1$ introduite ci-dessus est égale à la différence de moyenne empirique $\widehat{\Delta}$ entre les deux groupes, alors que l'estimateur de l'écart type résiduel $\tilde{\sigma}_\varepsilon$ de cette régression correspond à l'estimateur de l'écart type commun $\tilde{\sigma}$ dans les deux groupes. Si n_1 et n_0 dénotent les tailles des groupes avec respectivement $X = 1$ et $X = 0$, on a $\bar{x} = n_1/(n_1 + n_0)$, ce qui implique :

$$\sum_i (x_i - \bar{x})^2 = n_1 \cdot \left(1 - \frac{n_1}{n_1 + n_0}\right)^2 + n_0 \cdot \left(0 - \frac{n_1}{n_1 + n_0}\right)^2 = \frac{n_1 n_0}{n_1 + n_0}.$$

On teste ainsi la nullité de la pente de la droite de régression ($H_0 : \beta_1 = 0$) en calculant la statistique de test

$$t_{stat} = \frac{\widehat{\beta}_1}{\frac{\tilde{\sigma}_\varepsilon}{\sqrt{\sum_i (x_i - \bar{x})^2}}} = \sqrt{\frac{n_1 n_0}{n_1 + n_0}} \cdot \frac{\widehat{\Delta}}{\tilde{\sigma}}$$

qui est effectivement la même statistique de test que celle utilisée dans un test de Student pour tester la nullité de la différence de moyenne entre les deux groupes ($H_0 : \Delta = 0$). La distribution de la statistique de test sous H_0 est également dans les deux cas une distribution de Student avec $n_1 + n_0 - 2$ *dl*. Une illustration de cette analogie est donnée dans l'exemple suivant.

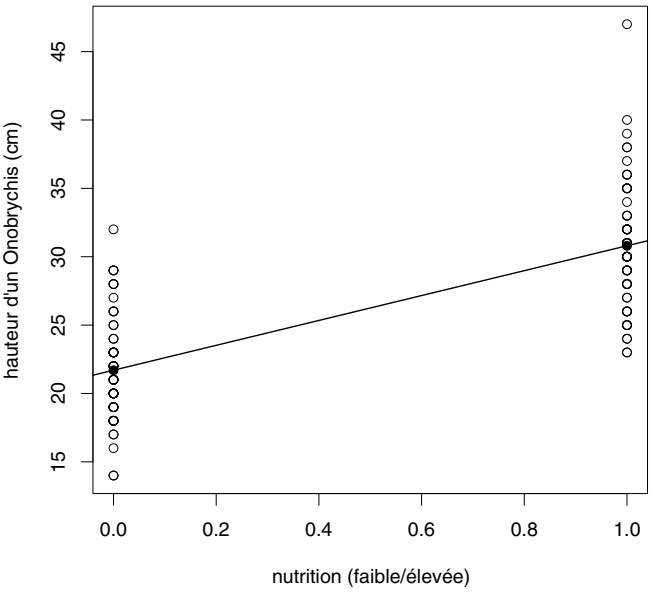

Figure 15.1 – Relation entre variable réponse continue et prédicteur binaire.

Exemple 15.1 *On reprend l'exemple du chapitre 5 où l'on comparait la hauteur des Onobrychis cultivés avec différents niveaux nutritifs. On avait estimé une hauteur moyenne de* $\widehat{\mu}_1 = 30.8$ *cm dans le groupe de* $n_1 = 60$ *Onobrychis cultivés avec niveau nutritif élevé et de* $\widehat{\mu}_0 = 21.7$ *cm dans le groupe de* $n_0 = 60$ *Onobrychis cultivés avec niveau nutritif faible, et ainsi une différence de moyenne de* $\widehat{\Delta} = 9.1$ *cm. On avait par ailleurs estimé une variance commune de* $\tilde{\sigma}^2 = 18.1$ cm^2. *Le graphique habituel pour montrer ces données est un boxplot (voir la figure 5.2). En notant par* Y *la hauteur d'un Onobrychis et par* X *le niveau nutritif (où* $X = 1$ *désigne un niveau nutritif élevé et* $X = 0$ *un niveau nutritif faible), on peut également représenter ces données avec un diagramme de dispersion, comme celui de la figure 15.1. On y voit la droite de régression qui passe par les moyennes des deux groupes représentées par les deux points noirs sur ce graphique. L'hypothèse de linéarité est ici forcément satisfaite. Les autres hypothèses (homoscédasticité, normalité) semblent raisonnables sur ce graphique. L'équation de la droite de régression est la suivante :*

$$\widehat{y}^* = 21.7 + 9.1x.$$

On retrouve la moyenne du groupe $X = 0$ *dans le rôle de la constante et la différence de moyenne entre les deux groupes dans le rôle de la pente de la droite de régression (on a* $\widehat{\beta}_0 = \widehat{\mu}_0$ *et* $\widehat{\beta}_1 = \widehat{\Delta}$*). Par ailleurs, la variance résiduelle est estimée par* $\tilde{\sigma}^2_\varepsilon = 18.1$, *ce qui correspond à la variance commune estimée*

ci-dessus (on a $\tilde{\sigma}^2_\varepsilon = \tilde{\sigma}^2$*). On a par ailleurs :*

$$\sum_i (x_i - \bar{x})^2 = \frac{n_1 n_0}{n_1 + n_0} = \frac{60 \cdot 60}{60 + 60} = 30.$$

Afin de tester la nullité de la pente de la régression ($H_0 : \beta_1 = 0$*) ou la nullité de la différence de moyenne entre les deux groupes (*$H_0 : \Delta = 0$*), on calcule* $t_{stat} = \sqrt{30} \cdot 9.1/\sqrt{18.1} = 11.7$*. Comme* $|t_{stat}|$ *est plus grand que* $t_{0.975,118} = 1.98$*, on rejette* H_0 *dans un test bilatéral au seuil de 5 % (un logiciel statistique nous donnera* $p < 0.0001$*).*

15.2 Comparaison de deux groupes dans une étude observationnelle

Comparer deux groupes par rapport à une variable continue est donc un cas particulier de modèle de régression. Il est dès lors possible d'ajuster cette comparaison pour des variables confondantes comme on va l'illustrer dans cette section. Rappelons à ce sujet qu'une régression nous permet de décrire une association statistique entre deux variables X et Y, et non de prouver une relation de cause à effet. Ceci est aussi valable si X est un prédicteur binaire. Supposons que l'on étudie la relation entre X le tabac (avec $X = 1$ pour les fumeurs et $X = 0$ pour les non-fumeurs) et Y une mesure continue de la performance physique, et que l'on observe une moins bonne performance physique en moyenne chez les fumeurs que chez les non-fumeurs. Ceci ne constitue pas pour autant une preuve que le tabac est la cause de cette moins bonne performance. Si le groupe des fumeurs est par exemple plus âgé que le groupe des non-fumeurs, la cause de la moins bonne performance physique chez les fumeurs pourrait tout aussi bien être l'âge que le tabac. Plus généralement, dans une étude dite *observationnelle*, les fumeurs et les non-fumeurs différeront inévitablement par rapport à de nombreuses caractéristiques, que ce soient des facteurs génétiques, des facteurs psychologiques ou des habitudes comportementales, et certains de ces facteurs pourraient également expliquer la moins bonne performance physique observée dans le groupe des fumeurs.

Si de telles caractéristiques confondantes sont identifiées, mesurables et mesurées sur les individus de notre échantillon, on peut les introduire dans notre modèle de régression (en plus du tabac). On peut ainsi effectuer une comparaison de la performance physique entre fumeurs et non-fumeurs *ajustée pour l'âge* (ou pour d'autres caractéristiques que l'on jugera potentiellement confondantes pour une telle association). Si la moins bonne performance physique chez les non-fumeurs est confirmée dans un tel modèle, on n'aura certes toujours pas prouvé définitivement une relation de cause à effet entre tabac et performance physique, mais on aura au moins réussi à exclure certaines raisons possibles (autres que le tabac) qui auraient pu expliquer l'association observée.

Exemple 15.2 *On reprend l'exemple illustré dans la figure 5.3 où l'on comparait le taux de créatine entre un groupe de* $n_1 = 31$ *femmes atteintes d'une maladie génétique et un groupe contrôle de* $n_0 = 39$ *femmes non atteintes de cette maladie. L'objectif était de montrer qu'un taux de créatine élevé pouvait constituer un symptôme et donc être la conséquence de cette maladie. On avait log-transformé le taux de créatine afin de nous rapprocher du modèle idéal (et donc d'un modèle de régression linéaire). On avait des taux moyens de log-créatine de* $\widehat{\mu}_1 = 4.71$ *et* $\widehat{\mu}_0 = 3.66$*, et ainsi une différence moyenne de* 1.05 *en faveur du groupe maladie qui était significative avec un test de Student (*$p < 0.0001$*). En calculant une régression simple avec comme prédicteur binaire* X_1 *la maladie (où l'on a* $X_1 = 1$ *pour le groupe malade et* $X_1 = 0$ *pour le groupe non malade) et comme variable réponse* Y *le taux de log-créatine, on obtient donc :*

$$\widehat{y}^* = 3.66 + 1.05x_1.$$

Cependant, de tels résultats ne suffisent pas pour prouver que la maladie est la cause d'un taux élevé de créatine. Comme il s'agit d'une étude observationnelle, les deux groupes de femmes diffèrent inévitablement par rapport à de nombreuses caractéristiques. Il se trouve par exemple que le groupe malade est nettement plus âgé que le groupe non malade (en moyenne 32.6 *contre* 27.3 *ans). La cause du taux de créatine plus élevé dans le groupe malade pourrait donc être l'âge plus élevé observé dans ce groupe. En introduisant l'âge* X_2 *dans notre modèle de régression, on trouve cependant :*

$$\widehat{y}^* = 2.74 + 0.87x_1 + 0.034x_2.$$

Cela veut dire qu'en comparant à âge égal, on a encore une différence de moyenne de log-créatine de 0.87 *en faveur du groupe malade, qui est certes plus petite que dans notre premier modèle, mais qui est toujours significative (*$p < 0.0001$*). Ainsi, une partie de la différence de moyenne de* 1.05 *observée entre les deux groupes peut être mise sur le compte de l'âge, mais pas toute cette différence. Cela ne suffit toujours pas pour prouver une relation de cause à effet entre la maladie et un taux de créatine élevé, mais on peut au moins exclure que la différence restante de* 0.87 *soit due à l'âge.*

La figure 15.2 illustre les résultats obtenus avec nos deux modèles. Ces diagrammes de dispersion nous montrent la relation entre l'âge et le taux de créatine (log-transformé), les individus du groupe malade étant représentés par des triangles, ceux du groupe non malade par des cercles. Les résultats du premier modèle sont illustrés dans le graphique du haut. On y voit deux droites horizontales (car l'association entre l'âge et le taux de créatine n'est pas prise en compte dans ce modèle), l'une pour les malades (trait plein), l'autre pour les non-malades (trait pointillé), qui se trouvent au niveau des moyennes calculées dans ces deux groupes. La distance verticale entre les deux droites vaut 1.05 *et représente la différence de moyenne entre les deux groupes non ajustée pour l'âge. Les résultats du second modèle sont illustrés dans le graphique du bas. Les deux droites ont une pente de* 0.034*, représentant l'association entre l'âge*

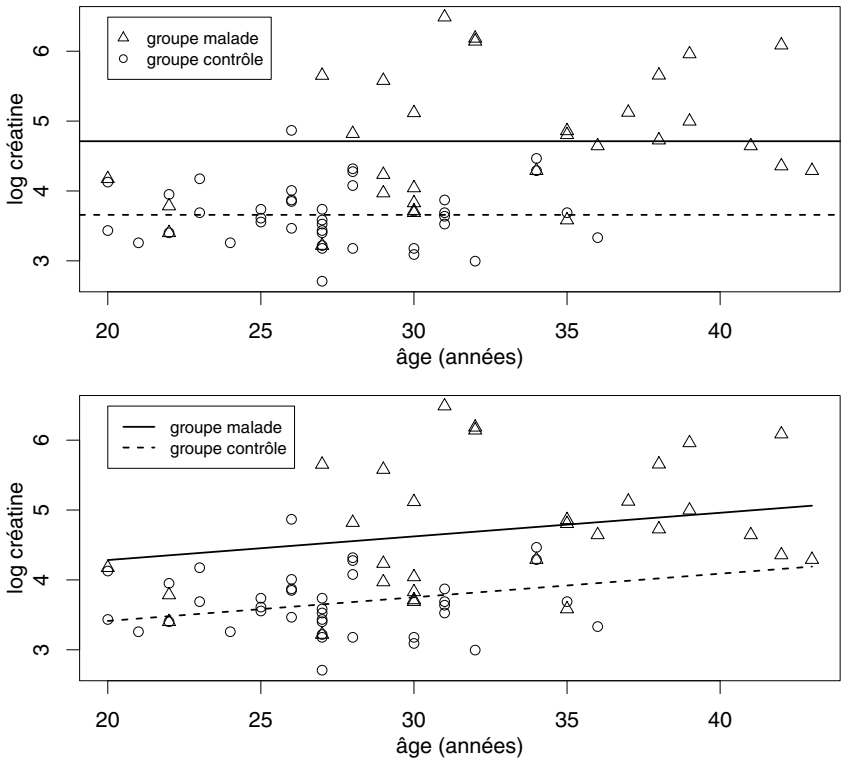

Figure 15.2 – Comparaison de deux groupes dans une étude observationnelle.

et le taux de créatine. La distance verticale entre les deux droites vaut 0.87 *et représente la différence de moyenne entre les deux groupes ajustée pour l'âge, qui est donc plus petite que la différence non ajustée.*

15.3 Comparaison de deux groupes dans un essai clinique

Si notre but était de prouver statistiquement une relation de cause à effet entre deux variables, par exemple l'effet d'un médicament sur l'état de santé des patients, il faudrait effectuer un essai clinique (en anglais : `clinical trial`). À partir d'un échantillon d'individus représentatif de notre population d'intérêt, par exemple les patients souffrant d'une certaine maladie, le principe d'un essai clinique consiste à répartir aléatoirement, on dit aussi *randomiser*, ces individus en deux groupes : on donne le médicament aux individus du premier groupe et on donne un placebo aux individus du second groupe. Grâce à la randomisation, les deux groupes de patients seront comparables en moyenne

par rapport à leur état de santé, mais aussi par rapport à toutes les variables confondantes potentielles (qu'elles soient connues ou inconnues, mesurables ou non mesurables). La seule différence entre les deux groupes au début de l'étude est *contrôlée* : il s'agit du médicament que les individus du premier groupe reçoivent et que les individus du second groupe ne reçoivent pas. Dans ces conditions, une différence entre l'état de santé des patients des deux groupes à la fin de l'étude pourra légitimement être interprétée comme un *effet causal* du médicament[1].

D'un point de vue statistique, l'analyse des données d'un essai clinique consistera simplement en une comparaison de deux groupes par rapport à une variable d'intérêt, par exemple une variable continue. Si X représente le groupe (avec dans notre exemple $X = 1$ pour le médicament et $X = 0$ pour le placebo) et Y la variable d'intérêt (par exemple l'état de santé d'un patient après six mois de traitement), on pourra effectuer un test de Student sur la différence de moyenne de Y entre les deux groupes, ou de façon équivalente un test de Student sur la pente de la droite de régression dans un modèle avec X comme prédicteur et Y comme variable réponse, cette pente étant donc ici interprétée comme l'effet causal de X sur Y. Ici aussi, on pourra introduire (en plus du groupe) d'autres prédicteurs dans ce modèle de régression. Cela nous permettra d'une part de corriger certains défauts éventuels de la randomisation par rapport à ces prédicteurs (au cas où le hasard aurait mal fait les choses, par exemple au cas où un groupe serait nettement plus âgé que l'autre malgré la randomisation). D'autre part, cela nous permettra de diminuer l'écart type résiduel de la régression et donc l'erreur type de l'estimateur de notre effet causal (on reviendra sur ce point dans la section suivante).

Exemple 15.3 *Nous présentons les résultats d'un essai clinique où l'on compare deux groupes de* $n_1 = n_0 = 40$ *patients en surpoids, volontaires pour faire un régime de trois mois. On note par* X_1 *la variable définissant les groupes. Dans le groupe* $X_1 = 1$*, les patients ont bénéficié d'un médicament pour accompagner le régime et les aider à perdre du poids. Dans le groupe* $X_1 = 0$*, les patients ont reçu à la place un placebo. La variable réponse* Y *est la perte de poids après trois mois. En moyenne, les patients du groupe médicament ont perdu* $\widehat{\mu}_1 = 4.72$ *kg, alors que les patients du groupe placebo ont perdu* $\widehat{\mu}_0 = 2.47$ *kg. La différence de moyenne, qui peut être ici interprétée comme l'effet causal du médicament, est ainsi estimée à* $4.72 - 2.47 = 2.25$ *kg. L'erreur type de cet estimateur est de* 0.70 *kg. La statistique de test d'un test de Student pour comparer ces deux moyennes est donnée par* $t_{stat} = 2.25/0.70 = 3.21$*, l'effet causal du médicament étant donc significatif (*$p = 0.002$*). Un intervalle de confiance au niveau* 95 *% pour le véritable effet causal du médicament est par ailleurs*

[1]Il ne sera malheureusement pas toujours possible d'effectuer un essai clinique, que ce soit pour des raisons pratiques, techniques ou éthiques. Pour la question du tabac évoquée dans la section précédente, on ne pourra pas forcer les individus d'un groupe à fumer et les individus de l'autre groupe à ne pas fumer.

donné par (en utilisant $t_{0.975,78} = 1.99$*) :*

$$2.25 \pm 1.99 \cdot 0.70 = [0.86; 3.64] \ kg \ .$$

On a donc prouvé statistiquement que le médicament permet de perdre en moyenne au moins 0.86 *kg de plus qu'un placebo après trois mois de régime. Ce résultat peut également s'obtenir en considérant un modèle de régression avec* X_1 *comme prédicteur et* Y *comme variable réponse. On obtient ainsi :*

$$\widehat{y}^* = 2.47 + 2.25x_1$$

avec la même inférence que ci-dessus en ce qui concerne la pente de cette droite de régression, interprétable comme l'effet causal du médicament.

Un autre prédicteur potentiel de Y *est le poids des patients au début du régime, que nous noterons par* X_2*. Les patients avec un poids initial élevé sont en effet susceptibles de perdre plus de poids que les patients avec un poids initial bas. Dans cet exemple fictif, nos deux groupes ont en moyenne (presque) le même poids initial (*105.36 *versus* 105.38 *kg). En introduisant* X_2 *dans le modèle, on ne change donc quasiment pas la pente de* X_1*. On obtient en effet :*

$$\widehat{y}^* = -47.67 + 2.25x_1 + 0.48x_2.$$

Par contre, l'erreur type de l'estimateur de cette pente est à présent de 0.58 *(alors qu'elle était de* 0.70 *dans le premier modèle). La statistique de test d'un test de Student sur la nullité de cette pente est donnée par* $t_{stat} = 2.25/0.58 = 3.88$*, la valeur p étant plus petite que dans le premier modèle (*$p = 0.0002$*). Un intervalle de confiance au niveau* 95 % *pour le véritable effet causal du médicament est par ailleurs donné par (en utilisant* $t_{0.975,77} = 1.99$*) :*

$$2.25 \pm 1.99 \cdot 0.58 = [1.10; 3.40] \ kg \ .$$

L'intervalle de confiance calculé avec ce second modèle est donc plus étroit que celui calculé avec le premier modèle, l'estimation de l'effet causal étant plus précise. En particulier, on a ici réussi à prouver statistiquement que le médicament permet de perdre en moyenne au moins 1.10 *kg (et pas seulement* 0.86 *kg) de plus qu'un placebo après trois mois de régime.*

La figure 15.3 illustre les résultats de ces deux modèles. Ces diagrammes de dispersion nous montrent la relation entre le poids initial et la perte de poids, les individus du groupe médicament étant représentés par des triangles, ceux du groupe placebo par des cercles. Les résultats du premier modèle sont illustrés dans le graphique du haut. On y voit deux droites horizontales (car l'association entre le poids initial et la perte de poids n'est pas prise en compte dans ce modèle), l'une pour le groupe médicament (trait plein), l'autre pour le groupe placebo (trait pointillé), qui se trouvent au niveau des moyennes calculées dans ces deux groupes. La distance verticale entre les deux droites vaut 2.25 *et représente l'estimation de l'effet causal du médicament. Les résultats du second modèle sont illustrés dans le graphique du bas. Les deux droites ont une pente*

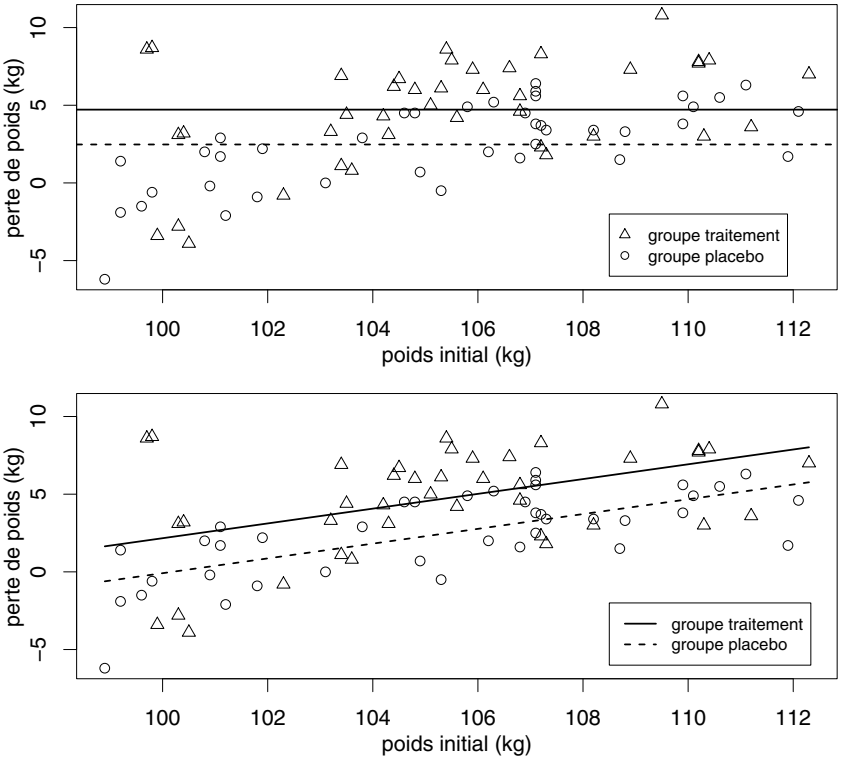

Figure 15.3 – Comparaison de deux groupes dans un essai clinique.

de 0.48*, représentant l'association entre le poids initial et la perte de poids, alors que la distance verticale entre ces deux droites vaut toujours* 2.25*. Ainsi, l'estimation de l'effet causal du médicament est la même dans les deux modèles, l'intérêt de considérer le second modèle étant ici une amélioration de la précision et non un ajustement de l'estimation.*

15.4 Planification d'une expérience

On distingue en général les études expérimentales des études observationnelles. Dans une étude expérimentale, on est capable de manipuler (choisir) les valeurs des prédicteurs pour les individus de notre échantillon, alors que dans une étude observationnelle, on se contente d'observer les individus. Un essai clinique est un cas particulier d'étude expérimentale avec un seul prédicteur. Dans certaines expériences, on manipule les valeurs de plusieurs prédicteurs.

Une question importante est celle de la planification d'une étude ou d'une expérience afin d'obtenir les meilleures estimations possibles des paramètres

d'intérêt[2]. On considère ci-dessous le cas simple et fréquent où l'on s'intéresse au coefficient β_1 d'un prédicteur X_1 dans un modèle de régression linéaire avec une variable réponse Y continue, par exemple lorsqu'il s'agit de comparer deux groupes par rapport à cette variable Y, X_1 étant alors un prédicteur binaire (on a vu de tels exemples dans les deux sections précédentes, respectivement pour une étude observationnelle et pour un essai clinique). Une formule particulièrement importante à ce sujet, vue au chapitre 14 dans le cadre général d'un modèle de régression linéaire multiple, est celle de la variance de l'estimateur $\widehat{\beta}_1$ de β_1 donnée par :

$$\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma_\varepsilon^2}{n\widehat{\sigma}_{X_1}^2} \cdot \frac{1}{1-\widehat{\rho}_1^2}$$

où n représente la taille totale de l'échantillon, σ_ε^2 la variance résiduelle du modèle de régression, $\widehat{\sigma}_{X_1}^2$ la variance empirique du prédicteur d'intérêt X_1 et $\widehat{\rho}_1^2$ le pourcentage empirique de la variance de X_1 prédite linéairement par les autres prédicteurs du modèle (on aura en particulier $\widehat{\rho}_1^2 = 0$ s'il n'y a pas d'autres prédicteurs dans le modèle). Il s'agit ainsi de planifier notre étude ou notre expérience afin de minimiser cette variance, c'est-à-dire de :

- choisir n suffisamment grand
- réduire le plus possible σ_ε^2
- choisir les valeurs de notre prédicteur d'intérêt afin de minimiser $\widehat{\rho}_1^2$
- choisir les valeurs de notre prédicteur d'intérêt afin de maximiser $\widehat{\sigma}_{X_1}^2$.

En ce qui concerne le choix de la taille n de l'échantillon, on peut adopter une stratégie similaire à celle évoquée au chapitre 10, à savoir choisir n de façon à atteindre une certaine puissance statistique. Afin de rejeter l'hypothèse nulle $H_0 : \beta_1 = 0$ dans un test de Student bilatéral au seuil α avec une probabilité (ou puissance) de $1-\beta$, on choisit ainsi :

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\beta_1^2} \cdot \frac{\sigma_\varepsilon^2}{\widehat{\sigma}_{X_1}^2} \cdot \frac{1}{1-\widehat{\rho}_1^2}.$$

Il s'agit donc de postuler une valeur pour le véritable β_1, mais également pour σ_ε^2, ainsi que pour $\widehat{\rho}_1^2$ et $\widehat{\sigma}_{X_1}^2$ (ces deux dernières quantités étant en général contrôlées et donc connues dans le cadre d'une étude expérimentale).

Afin de réduire la variance résiduelle σ_ε^2, il s'agit d'introduire dans notre modèle des prédicteurs importants de Y, autrement dit d'*ajuster* notre modèle pour ces prédicteurs (en s'assurant toutefois que les hypothèses d'un modèle de régression linéaire demeurent raisonnables avec l'introduction de ces prédicteurs). On a vu que cet ajustement est souvent nécessaire dans une étude observationnelle (car ces prédicteurs sont alors souvent considérés comme des

[2]On parle en anglais de `design of experiment` qui constitue un autre domaine spécifique de la science statistique.

variables confondantes), alors qu'un tel ajustement est également souhaitable dans un essai clinique (dans le but de réduire σ_ε^2). Au moment de la planification d'une étude, il faut penser aux prédicteurs à introduire dans le modèle afin de pouvoir les mesurer sur les individus de notre échantillon.

La quantité $\widehat{\rho}_1^2$ sera nulle (et donc minimisée) si le prédicteur d'intérêt X_1 est non corrélé avec les autres prédicteurs introduits dans le modèle. Dans le cas où X_1 est un prédicteur binaire, X_1 sera non corrélé avec les autres prédicteurs si ces derniers ont même moyenne dans chacun des groupes définis par $X_1 = 1$ et $X_1 = 0$. Dans un essai clinique, cela sera approximativement le cas grâce à la randomisation. Si on redoute que la randomisation fasse mal les choses (surtout avec de petits échantillons), on peut lui forcer la main de façon à ce que les deux groupes soient effectivement exactement comparables par rapport aux prédicteurs jugés particulièrement importants. On répartira par exemple nos individus dans les deux groupes de façon à ce que la proportion de femmes (et donc d'hommes) y soit exactement la même, avec par exemple 60 femmes et 40 hommes dans chaque groupe, de sorte que la corrélation entre la variable groupe et la variable sexe soit exactement nulle. Notre randomisation sera ainsi *bloquée* pour le sexe (en anglais : `randomized block design`), ce que l'on oppose parfois à une randomisation *complète* (en anglais : `completely randomized design`). Si la randomisation est bloquée pour tous les prédicteurs introduits dans notre modèle de régression, on aura effectivement $\widehat{\rho}_1^2 = 0$.

Dans une étude observationnelle, on n'a pas de randomisation et donc pas de randomisation bloquée, mais on pourra parfois effectuer un pairage entre les individus des deux groupes (en anglais : `matching`)[3]. Dans notre exemple où l'on comparait le taux de créatine entre un « groupe malade » et un « groupe contrôle », ce dernier aurait pu être constitué de façon à ce que chaque femme de ce groupe soit « pairée » avec une femme du groupe malade, de telle sorte que les femmes des deux groupes se ressemblent par rapport à certaines caractéristiques jugées importantes (on aurait pu ainsi inclure des femmes dans le groupe contrôle qui ont le même âge que les femmes du groupe malade). Dans une étude observationnelle, on ne peut certes pas choisir les valeurs des prédicteurs pour les individus, mais on peut sélectionner certains individus plutôt que d'autres dans notre échantillon afin de minimiser la quantité $\widehat{\rho}_1^2$.

Finalement, il s'agit de choisir les valeurs du prédicteur X_1 dans notre échantillon (en les manipulant dans une étude expérimentale, en sélectionnant certains individus plutôt que d'autres dans une étude observationnelle) afin de maximiser sa variance $\widehat{\sigma}_{X_1}^2$. En particulier, si les valeurs possibles pour X_1 sont comprises entre 0 et 1, cette variance sera maximale si on a $X_1 = 1$ pour la moitié des individus et $X_1 = 0$ pour l'autre moitié. Si par exemple X_1 représente

[3]Notons que les techniques d'ajustement d'une part et de randomisation bloquée/pairage d'individus d'autre part ne sont pas redondantes mais participent toutes deux de façon différente à réduire la variance de l'estimateur du paramètre d'intérêt. L'une n'exclut donc pas l'autre et ces deux techniques peuvent être utilisées dans une même étude. En fait, un ajustement pour les prédicteurs utilisés dans un algorithme de blocage/pairage est même nécessaire en principe, sans quoi l'hypothèse d'indépendance des résidus (la quatrième hypothèse d'un modèle de régression linéaire) ne serait pas vérifiée.

la proportion de jours dans l'année où un individu fume, on aura une plus grande puissance statistique si on inclut uniquement des individus qui fument tous les jours ($X_1 = 1$) et des individus qui ne fument jamais ($X_1 = 0$) que si on inclut également des individus qui fument occasionnellement ($0 < X_1 < 1$).

15.5 Analyse de variance

Un modèle de régression peut donc inclure à la fois des prédicteurs continus et des prédicteurs binaires, qui sont tous des prédicteurs quantitatifs. On va voir à présent comment traiter les prédicteurs qualitatifs. Afin d'introduire dans un modèle de régression un prédicteur qualitatif admettant q valeurs possibles (définissant q groupes), on procède de la manière suivante :

- **on définit $q-1$ variables binaires**, notées ici par $X_1, X_2, \cdots, X_{q-1}$, X_j désignant l'appartenance au groupe j (pour $j = 1, 2, \cdots, q-1$)

  $\rightarrow$ on aura ainsi[4] :

  - groupe 1 : $X_1 = 1 \quad X_2 = 0 \quad \cdots \quad X_{q-1} = 0$
  - groupe 2 : $X_1 = 0 \quad X_2 = 1 \quad \cdots \quad X_{q-1} = 0$

    $\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$
  - groupe $q-1$: $X_1 = 0 \quad X_2 = 0 \quad \cdots \quad X_{q-1} = 1$
  - groupe q : $X_1 = 0 \quad X_2 = 0 \quad \cdots \quad X_{q-1} = 0$

- **on introduit ces $q-1$ variables binaires comme prédicteurs dans notre modèle de régression linéaire** (qui sera donc multiple si $q > 2$)

  $\rightarrow$ une variable qualitative avec q valeurs possibles « coûte » $q-1$ paramètres dans un modèle de régression.

> Dans un modèle de régression, un prédicteur qualitatif avec q modalités est représenté par $q-1$ prédicteurs binaires.

Considérons à présent un modèle de régression avec une variable réponse Y et un prédicteur qualitatif représenté par ces $m = q-1$ prédicteurs binaires. L'équation de notre hyperplan de régression est donc la suivante :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_{q-1} x_{q-1}.$$

[4]Notons que, selon les logiciels statistiques, ce sera parfois le premier groupe (et non le dernier groupe comme nous le faisons ici) qui sera le groupe de référence, c'est-à-dire le groupe pour lequel chacune des $q-1$ variables binaires seront nulles.

Notons que l'hypothèse de linéarité est ici forcément satisfaite, de sorte que l'on prédit la valeur Y d'un individu par la moyenne du groupe auquel cet individu appartient[5]. On interprète par ailleurs les paramètres de la façon suivante :

- $\beta_0 = mean(Y|X_1 = \cdots = X_{q-1} = 0)$

  $\rightarrow$ la constante est la moyenne de Y dans le groupe (de référence) q

- $\beta_1 = mean(Y|X_1 = 1, X_2 = \cdots = X_{q-1} = 0) - mean(Y|X_1 = \cdots = X_{q-1} = 0)$

  $\rightarrow$ la pente associée à X_1 est la différence de moyenne de Y entre les groupes 1 et q

- $\beta_2 = mean(Y|X_2 = 1, X_1 = \cdots = X_{q-1} = 0) - mean(Y|X_1 = \cdots = X_{q-1} = 0)$

  $\rightarrow$ la pente associée à X_2 est la différence de moyenne de Y entre les groupes 2 et q

  $\rightarrow$ interprétation similaire pour toutes les pentes.

L'hypothèse nulle d'aucune association entre cette variable qualitative avec q valeurs possibles et la variable réponse Y revient à dire que la moyenne de Y est la même dans les q groupes, autrement dit que :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{q-1} = 0.$$

Cette hypothèse nulle peut se tester avec un test F sur la nullité simultanée de l'ensemble de ces $m = q - 1$ pentes. Si l'hypothèse nulle ci-dessus est rejetée, on conclut que les q groupes n'ont pas tous la même moyenne de Y. Afin de savoir quels sont les groupes qui diffèrent significativement des autres, on pourra effectuer des tests dits *post hoc*, en comparant deux à deux les différents groupes[6]. Différentes méthodes existent pour traiter de la problématique des tests multiples, l'une d'elles étant la correction de Bonferroni introduite au chapitre 6.

Exemple 15.4 *On a introduit au chapitre 5 deux échantillons d'Onobrychis cultivés avec un niveau nutritif faible et avec un niveau nutritif élevé. En fait, parmi les* 60 *Onobrychis cultivés avec un niveau nutritif faible, les* 30 *premiers ont été cultivés avec le niveau nutritif* 1 *et les* 30 *suivants avec le niveau nutritif*

[5] En effet, seules q parmi les 2^{q-1} combinaisons théoriques des valeurs possibles de ces $q-1$ prédicteurs définissent un groupe qui a du sens (une combinaison avec $X_1 = X_2 = 1$ n'aurait par exemple pas de sens). L'hypothèse de linéarité est donc satisfaite car un hyperplan de régression défini par q paramètres passe forcément par les moyennes de ces q groupes. Un modèle de régression linéaire est donc ici constitué uniquement de trois hypothèses, à savoir que la variance est la même dans chaque groupe, que la distribution est normale dans chaque groupe et que les observations sont indépendantes.

[6] On pourra comparer les groupes 1 et q, les groupes 2 et q, $\cdots$, ou les groupes $q-1$ et q en effectuant des tests de Student sur la nullité des pentes $\beta_1, \beta_2, \cdots$, ou β_{q-1} du modèle ci-dessus. Afin d'effectuer des comparaisons impliquant d'autres groupes que le groupe de référence q, par exemple une comparaison entre les groupes 1 et 2, on pourra recalculer la régression en codant les groupes différemment (en changeant ainsi le groupe de référence).

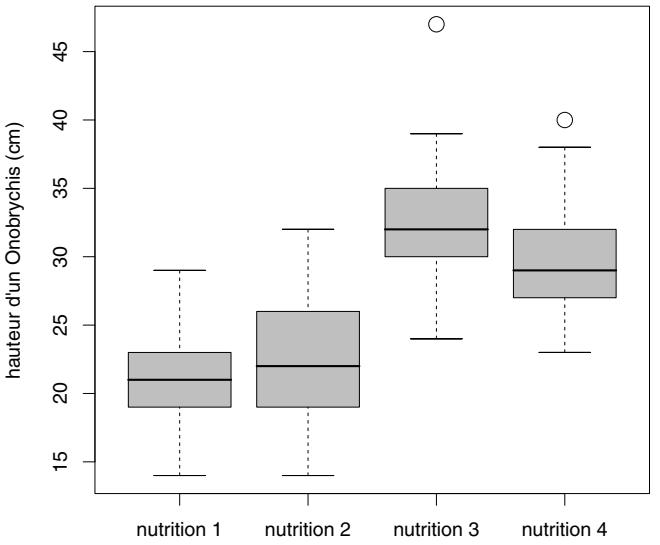

Figure 15.4 – Boxplots comparant quatre groupes.

2. *De même, parmi les* 60 *Onobrychis cultivés avec un niveau nutritif élevé, les* 30 *premiers ont été cultivés avec le niveau nutritif* 3 *et les* 30 *suivants avec le niveau nutritif* 4. *On va ici comparer la hauteur* Y *entre ces* $q = 4$ *groupes d'Onobrychis. Les boxplots de la figure 15.4 nous résument ces données. Les hauteurs moyennes de ces Onobrychis sont de* 20.9, 22.5, 31.9 *et* 29.7 *cm, pour respectivement les niveaux nutritifs* 1, 2, 3 *et* 4. *Un modèle linéaire avec comme variable réponse* Y *et comme prédicteur le niveau nutritif considéré comme une variable quantitative serait ici non approprié (la relation entre le niveau nutritif et la hauteur d'un Onobrychis n'étant visiblement pas linéaire). On considérera donc le niveau nutritif comme une variable qualitative. En notant par* X_1 *l'appartenance au groupe* 1, *par* X_2 *l'appartenance au groupe* 2 *et par* X_3 *l'appartenance au groupe* 3, *on obtient le modèle de régression suivant :*

$$\widehat{y}^* = 29.7 - 8.8x_1 - 7.1x_2 + 2.3x_3.$$

On calcule par ailleurs $\widehat{\rho}^2 = 0.563$. *La statistique de test d'un test* F *pour tester la nullité simultanée de ces trois pentes (et donc l'égalité des moyennes de ces quatre groupes) est donnée par* $t_{stat} = (116/3) \cdot 0.563/(1 - 0.563) = 49.8$, *qui est plus grand que* $F_{0.95,3,116} = 2.68$. *On a ainsi prouvé statistiquement que ces quatre moyennes ne sont pas les mêmes et donc que la hauteur d'un Onobrychis dépend du niveau nutritif.*

Notons que les pentes associées à X_1, X_2 *et* X_3 *nous informent sur les différences entre les groupes* 1 *et* 4, 2 *et* 4, *et* 3 *et* 4. *Afin d'obtenir de l'in-*

formation sur d'autres différences, on pourra calculer les modèles de régression équivalents (c'est-à-dire donnant les mêmes prédictions), en dénotant par X_4 l'appartenance au groupe 4 :

$$\widehat{y}^* = 31.9 - 11.1x_1 - 9.4x_2 - 2.3x_4$$

$$\widehat{y}^* = 22.5 - 1.7x_1 + 9.4x_3 + 7.1x_4$$

ou

$$\widehat{y}^* = 20.9 + 1.7x_2 + 11.1x_3 + 8.8x_4.$$

On retrouve en effet la même valeur de $\widehat{\rho}^2$ et le même résultat du test F pour chacun de ces modèles, bien que les pentes diffèrent d'un modèle à l'autre (le groupe de référence étant différent dans chaque modèle). À l'aide de tests de Student sur la nullité de ces pentes (et en appliquant une correction de Bonferroni), on obtient des différences significatives entre les groupes 1 et 3, 1 et 4, 2 et 3, ainsi que 2 et 4, les différences n'étant pas significatives entre les groupes 1 et 2, ni entre les groupes 3 et 4 (raison pour laquelle nous avions regroupé les groupes 1 et 2, ainsi que les groupes 3 et 4 au chapitre 5).

15.6 Analyse de covariance

Lorsqu'un modèle de régression contient à la fois des prédicteurs binaires (représentant éventuellement des variables qualitatives, appelées parfois des *facteurs*) et des prédicteurs continus (appelées en anglais : `covariates`), on parle d'*analyse de covariance*. Dans ces modèles, on peut également introduire des *interactions* entre les différents prédicteurs.

Considérons un modèle où l'on prédit une variable continue Y (par exemple une mesure de la performance physique) à partir des trois prédicteurs suivants :

- une variable binaire X_1 (par exemple le sexe, avec $X_1 = 1$ pour les hommes et $X_1 = 0$ pour les femmes)
- une variable continue X_2 (par exemple l'âge)
- une interaction $X_3 = X_1 \cdot X_2$ (obtenue par multiplication des deux premiers prédicteurs).

La variable X_3 est ainsi définie de telle sorte que :

$$\begin{aligned} X_3 = X_2 \quad &\text{si} \quad X_1 = 1 \\ X_3 = 0 \quad &\text{si} \quad X_1 = 0 \end{aligned}$$

Dans notre exemple, X_3 représente donc l'âge de l'individu s'il s'agit d'un homme et la valeur 0 s'il s'agit d'une femme. L'équation de notre hyperplan de régression est donnée comme d'habitude par :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3.$$

Pour les femmes, on a $X_1 = X_3 = 0$, ce qui implique :

$$\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_2 x_2.$$

Pour les hommes, on a $X_1 = 1$ et $X_3 = X_2$, ce qui implique :

$$\widehat{y}^* = (\widehat{\beta}_0 + \widehat{\beta}_1) + (\widehat{\beta}_2 + \widehat{\beta}_3) x_2.$$

Ce modèle de régression avec trois prédicteurs nous permet donc de décrire la relation entre l'âge X_2 et la performance physique Y avec deux droites de régression différentes, l'une pour les femmes et l'autre pour les hommes. On interprète les paramètres de la façon suivante :

- β_0 et β_2 définissent la droite de régression dans le groupe (de référence) $X_1 = 0$
- β_1 est la différence entre les constantes de ces deux droites de régression
- β_3 est la différence entre les pentes de ces deux droites de régression.

Les deux droites sont identiques si $\beta_1 = \beta_3 = 0$. Sous les hypothèses d'un modèle de régression linéaire, cela peut se tester à l'aide d'un test du F partiel. Un test de l'hypothèse nulle $H_0 : \beta_3 = 0$ à l'aide d'un test de Student sera également d'intérêt. En effet, la nullité de β_3 dans ce modèle aura les implications suivantes[7] :

- les deux droites de régression sont parallèles (ont une même pente)
- la relation entre X_1 et Y peut être résumée par la même différence de moyenne β_1 quelle que soit la valeur de la variable X_2
- la relation entre X_2 et Y peut être résumée par la même pente β_2 dans le groupes $X_1 = 0$ et dans le groupe $X_1 = 1$
- on dit alors qu'il n'y a pas d'interaction entre les prédicteurs X_1 et X_2. (dans le cas contraire, on aura une interaction entre X_1 et X_2).

La figure 15.5 nous montre des exemples d'interactions. Dans chacun de ces graphiques, on a deux droites de régression décrivant la relation entre X_2 (sur l'axe horizontal) et Y (sur l'axe vertical) dans les groupes $X_1 = 1$ (trait plein)

[7]Par contre, les tests des hypothèses nulles $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$ ou $H_0 : \beta_2 = 0$ seront souvent d'un intérêt moindre (en présence d'une interaction). La raison à cela est que le signe de ces paramètres (en particulier le fait qu'ils soient nuls ou non nuls) dépendra des unités choisies pour X_1 et X_2, et en ce sens seront arbitraires (les résultats changeront si on mesure par exemple une température en degrés Fahrenheit plutôt qu'en degrés Celsius, ou si on échange les codes des deux groupes définis par une variable binaire). D'une manière générale, il est conseillé d'introduire les prédicteurs X_1 et X_2 dans un modèle de régression contenant leur interaction $X_1 \cdot X_2$ (même si X_1 et X_2 ne sont pas significatifs). De même, il est conseillé d'introduire une constante dans un modèle de régression, sans laquelle certaines propriétés évoquées dans les chapitres précédents ne sont plus valables (par exemple le lien entre corrélation et pente d'une droite de régression).

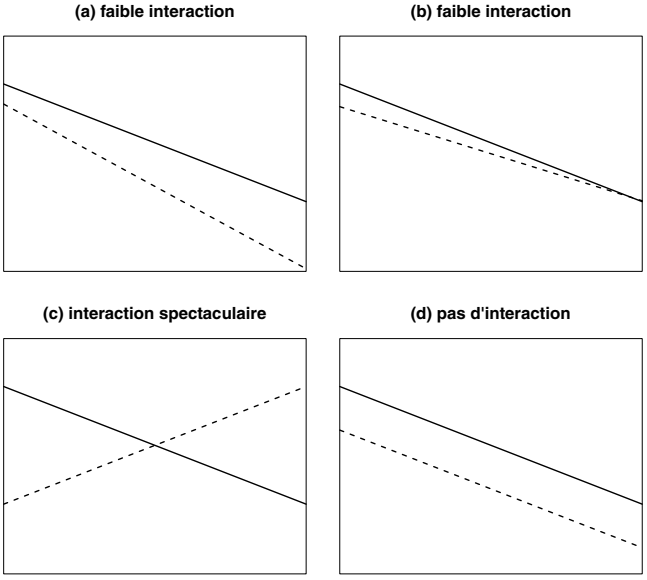

Figure 15.5 – Interactions entre prédicteur binaire et prédicteur continu.

et $X_1 = 0$ (trait pointillé). Dans les graphiques (a) et (b), on a une faible interaction entre les deux prédicteurs. En reprenant notre exemple ci-dessus, ces graphiques nous montrent que la performance physique est en général moins bonne chez les femmes que chez les hommes et qu'elle baisse avec l'âge pour les deux sexes, mais pas à la même vitesse. Alors qu'elle baisse plus vite chez les femmes dans le graphique (a), la différence entre hommes et femmes augmentant ainsi avec l'âge, elle baisse plus vite chez les hommes dans le graphique (b), la différence entre hommes et femmes diminuant avec l'âge (disparaissant même complètement pour de grands âges). Dans le graphique (c), on a une interaction spectaculaire entre les deux prédicteurs, la performance physique baissant avec l'âge chez les hommes et augmentant avec l'âge chez les femmes, ces dernières dépassant les hommes à partir d'un certain âge. Dans le graphique (d), on retrouve le cas décrit par $\beta_3 = 0$ où l'on n'a pas d'interaction entre les prédicteurs. La performance physique baisse ici avec l'âge à la même vitesse pour les deux sexes. De même, la différence entre les sexes est la même à chaque âge.

Exemple 15.5 *On reprend l'exemple de notre essai clinique ci-dessus, où Y représente la perte de poids après un régime de trois mois, X_1 le groupe (avec $X_1 = 1$ pour le groupe médicament et $X_1 = 0$ pour le groupe placebo) et X_2 le poids initial des patients. On avait calculé le modèle de régression suivant :*

$$\widehat{y}^* = -47.67 + 2.25x_1 + 0.48x_2.$$

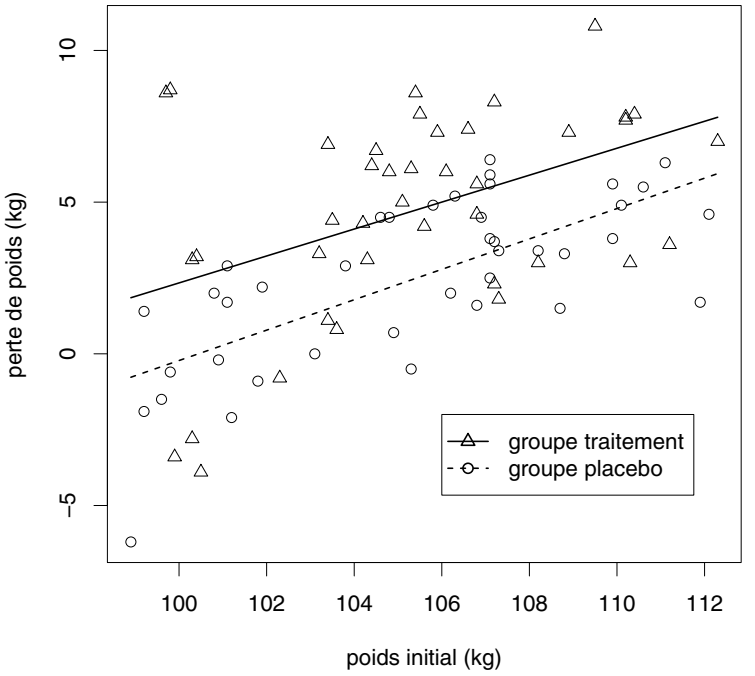

Figure 15.6 – Comparaison de deux groupes avec interaction.

Ce modèle implique les relations suivantes entre poids initial et perte de poids pour les patients du groupe placebo :

$$\widehat{y}^* = -47.67 + 0.48x_2$$

et pour les patients du groupe médicament :

$$\widehat{y}^* = -45.42 + 0.48x_2.$$

Autrement dit, les deux droites sont parallèles, de sorte que la distance verticale entre ces deux droites est la même à chaque poids initial x_2 et vaut 2.25 kg, cette valeur étant une estimation de l'effet du médicament sur la perte de poids, valable dans ce modèle quel que soit le poids initial.

Si on introduit une interaction $X_3 = X_1 \cdot X_2$ dans le modèle, on obtient :

$$\widehat{y}^* = -50.32 + 8.26x_1 + 0.501x_2 - 0.057x_3.$$

Ce modèle implique les relations suivantes entre poids initial et perte de poids pour les patients du groupe placebo :

$$\widehat{y}^* = -50.32 + 0.501x_2$$

et pour les patients du groupe médicament :

$$\widehat{y}^* = -42.06 + 0.444x_2.$$

Ces deux droites de régression sont montrées dans la figure 15.6 (avec un trait plein pour le groupe médicament et un trait pointillé pour le groupe placebo). À cause de l'interaction, ces deux droites ne sont pas tout à fait parallèles, la pente étant légèrement moins importante dans le groupe médicament que dans le groupe placebo. De même, la distance verticale entre les deux droites n'est pas tout à fait la même à chaque poids initial x_2. Elle est par exemple de :

- $8.26 - 0.057 \cdot 100 = 2.56$ *kg pour un poids initial de* 100 *kg*
- $8.26 - 0.057 \cdot 110 = 1.99$ *kg pour un poids initial de* 110 *kg.*

Cela pourrait suggérer que l'effet du médicament est un peu moins important pour un poids initial élevé que pour un poids initial bas. Cependant, cette interaction négative n'est pas significative avec un test de Student ($p = 0.73$). On n'a donc pas prouvé statistiquement que l'effet du médicament diffère selon le poids initial, ce qui nous permet de résumer cet effet par l'unique valeur de 2.25 *kg (ce qui simplifie également le message de cette étude).*

En résumé, un modèle de régression linéaire multiple peut inclure :

- des prédicteurs continus
- des prédicteurs binaires
- des prédicteurs qualitatifs codés sous forme de prédicteurs binaires
- des interactions entre prédicteurs continus, binaires ou qualitatifs.

Une interaction entre deux prédicteurs correspond à un nouveau prédicteur obtenu des deux premiers par multiplication.

Attention, une interaction entre deux prédicteurs qualitatifs avec respectivement q_1 et q_2 valeurs possibles nécessite l'introduction de $(q_1-1)(q_2-1)$ prédicteurs dans le modèle, obtenus par multiplication des $q_1 - 1$ prédicteurs binaires définissant le premier prédicteur qualitatif et des $q_2 - 1$ prédicteurs binaires définissant le second prédicteur qualitatif, et donc « coûte » $(q_1 - 1)(q_2 - 1)$ paramètres. De même, une interaction entre un prédicteur binaire ou continu avec un prédicteur qualitatif avec q valeurs possibles coûte $q - 1$ paramètres. Par contre, une interaction entre deux prédicteurs binaires, entre deux prédicteurs continus, ou entre un prédicteur binaire et un prédicteur continu ne coûte qu'un seul paramètre. On retrouve ici le fait que les variables binaires et les variables continues sont les plus commodes à traiter dans une analyse statistique.

Chapitre 16

Régression logistique

On a vu comment décrire la relation entre des prédicteurs, continus ou binaires, et une variable réponse continue à l'aide d'un modèle de régression linéaire. On va voir dans ce chapitre comment décrire la relation entre des prédicteurs, continus ou binaires, et une variable réponse binaire à l'aide d'un *modèle de régression logistique*. Alors que les paramètres d'un modèle de régression linéaire s'interprètent comme des moyennes et des différences de moyenne, les paramètres d'un modèle de régression logistique s'interprètent en terme d'*odds* et d'*odds-ratio*, concepts clés que nous commençons par introduire ci-dessous.

16.1 Odds et odds-ratio

La distribution d'une variable binaire (avec valeurs possibles 1 et 0) est caractérisée par la proportion π de 1. En fait, n'importe quelle transformation monotone de π, en particulier la quantité $\pi/(1-\pi)$ que l'on appelle en français une cote et en anglais un `odds` (terme que nous utiliserons ci-dessous), caractérise également cette distribution. On peut passer d'une proportion à un odds et *vice versa* en utilisant les formules suivantes :

$$odds = \frac{proportion}{1 - proportion} \quad \text{et} \quad proportion = \frac{odds}{odds + 1}.$$

Le tableau suivant nous donne quelques exemples de passage de l'un à l'autre :

proportion	*odds*	*proportion*	*odds*
0	0	0.5	1
0.001	0.001001	0.67	2
0.01	0.0101	0.9	9
0.05	0.053	0.95	19
0.1	0.11	0.99	99
0.2	0.25	0.999	999
0.33	0.5	1	∞

Alors qu'une proportion est un nombre entre 0 et 1, un odds est un nombre entre 0 et ∞. Les concepts d'odds et de proportion sont pourtant très proches l'un de l'autre lorsque les valeurs sont petites.

Exemple 16.1 *Lorsque l'on dit qu'il naît* 105 *garçons pour* 100 *filles, cela veut dire que l'odds d'avoir un garçon à la naissance est de* 1.05. *La proportion de garçons à la naissance (la probabilité d'avoir un garçon) est alors égale à* $1.05/2.05 = 0.512$. *Alors que cette proportion est calculée en divisant le nombre de garçons nés par le nombre total de naissances, l'odds correspondant est calculé en divisant le nombre de garçons nés par le nombre de filles nées. Les deux quantités représentent toutefois la même information, qui est simplement exprimée différemment.*

Afin de comparer les distributions de deux variables binaires Y_1 et Y_0, on a introduit au chapitre 5 le concept de différence de proportion :

$$\Lambda = \pi_1 - \pi_0$$

où π_1 et π_0 dénotent les proportions de 1 pour les variables Y_1 et Y_0. Alors que l'on compare les proportions π_1 et π_0 *via* leur différence, on compare les odds $\pi_1/(1-\pi_1)$ et $\pi_0/(1-\pi_0)$ *via* leur quotient :

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}$$

que l'on appelle un `odds-ratio`. Pour de petites valeurs de π_1 et π_0, l'odds-ratio est très proche du quotient (ou ratio) des proportions π_1/π_0. Pour de grandes valeurs de π_1 ou π_0, les deux concepts sont cependant différents[1].

Un odds-ratio est donc une alternative à une différence de proportion lorsqu'il s'agit de comparer deux distributions binaires. On aura notamment :

$$\begin{aligned} \omega < 1 \quad &\text{si} \quad \Lambda < 0 \\ \omega = 1 \quad &\text{si} \quad \Lambda = 0 \\ \omega > 1 \quad &\text{si} \quad \Lambda > 0. \end{aligned}$$

Ces deux concepts ne sont cependant pas équivalents, comme nous le montre la figure 16.1. L'axe horizontal représente ici l'ensemble des valeurs possibles pour la proportion π_0 et l'axe vertical l'ensemble des valeurs possibles pour la proportion π_1. Chaque courbe sur ce graphique représente l'ensemble des couples (π_0, π_1) correspondant à un même odds-ratio (la courbe représentée

[1] Un avantage de l'odds-ratio est qu'il est invariant si on échange les codes 1 et 0 des deux valeurs possibles des variables binaires que l'on compare, alors que le ratio des proportions ne l'est pas. Si on compare par exemple deux techniques d'opération avec respectivement 90 % et 80 % de réussites, et donc 10 % et 20 % d'échecs, le ratio des proportions sera de $0.9/0.8 = 1.125$ si on compare les réussites, et de $0.2/0.1 = 2$ si on compare les échecs, alors que le ratio des odds sera de $(0.9/0.1)/(0.8/0.2) = (0.2/0.8)/(0.1/0.9) = 2.25$ que l'on compare les réussites ou que l'on compare les échecs.

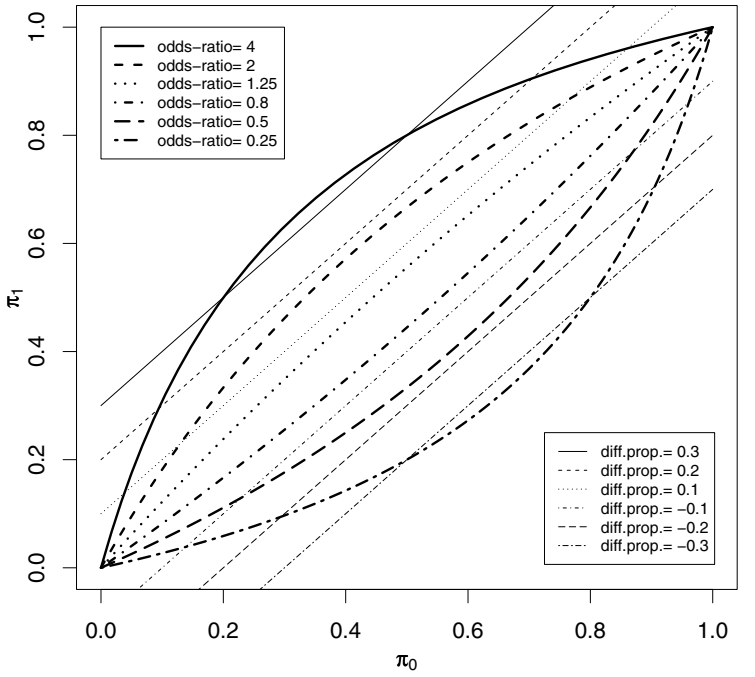

Figure 16.1 – Odds-ratio *versus* différence de proportion.

par un trait plein correspondant par exemple à un odds-ratio de 4). De même, chaque droite sur ce graphique représente l'ensemble des couples (π_0, π_1) correspondant à une même différence de proportion (la droite représentée par un trait plein correspondant par exemple à une différence de proportion de 0.3). On remarque que chaque courbe croise plusieurs droites et que chaque droite croise plusieurs courbes. Ainsi, à un odds-ratio de 4 pourra correspondre par exemple une différence de proportion de 0.1, 0.2 ou 0.3. De même, à une différence de proportion de 0.1 pourra correspondre un odds-ratio de 4 ou de 2 (mais pas de 1.25)[2].

Exemple 16.2 *Pour convaincre le lecteur que les concepts d'odds-ratio et de différence de proportion ne sont pas équivalents, on considère un exemple où l'on compare la proportion π_1 de guérison obtenue avec un médicament et la*

[2]Le lecteur pourra vérifier en exercice qu'à un odds-ratio $\omega > 1$ pourra correspondre n'importe quelle différence de proportion entre 0 (non compris) et $(\sqrt{\omega}-1)^2/(\omega-1)$, alors qu'à une différence de proportion $\Lambda > 0$ pourra correspondre n'importe quel odds-ratio entre $(1+\Lambda)^2/(1-\Lambda)^2$ et ∞. À un odds-ratio de 4 pourra par exemple correspondre une différence de proportion entre 0 (non compris) et 1/3, alors qu'à une différence de proportion de 0.2 pourra correspondre un odds-ratio entre 2.25 et ∞. Si les proportions π_1 et π_0 sont petites, il sera ainsi possible d'avoir dans le même temps une différence de proportion qui apparaîtra petite et un odds-ratio qui apparaîtra grand.

proportion π_0 de guérison obtenue avec un placebo, et ceci séparément chez les hommes et chez les femmes. Admettons que l'on ait les résultats suivants :

- *chez les hommes :* $\pi_1 = 0.6$ *et* $\pi_0 = 0.4$
- *chez les femmes :* $\pi_1 = 0.95$ *et* $\pi_0 = 0.85$.

On a ainsi les différences de proportion et odds-ratios suivants :

- *chez les hommes :* $\Lambda = 0.2$ *et* $\omega = 2.25$
- *chez les femmes :* $\Lambda = 0.1$ *et* $\omega = 3.35$.

À la question « est-ce que le médicament (en comparaison avec le placebo) a plus d'effet chez les hommes ou chez les femmes ? », on répondra donc « chez les hommes » si on mesure cet effet via une différence de proportion, et on répondra « chez les femmes » si on mesure cet effet via un odds-ratio.

Une différence de proportion et un odds-ratio sont deux concepts non équivalents pour comparer deux variables binaires.

En résumé, lorsqu'il s'agit de comparer deux distributions binaires, certains préfèrent calculer une différence de proportion et d'autres un odds-ratio. Lorsqu'il s'agit de modéliser une variable binaire, on va voir cependant que l'odds-ratio est un paramètre plus commode que la différence de proportion.

16.2 Étude cas-témoins

Une raison de la popularité de l'odds-ratio est son utilité dans l'analyse des données d'une *étude cas-témoins* (en anglais : `case control study`). On oppose parfois une étude cas-témoins, qui est une *étude rétrospective*, à une *étude de cohorte* qui est une *étude prospective*. Alors que le principe d'une étude de cohorte consiste à suivre au cours du temps des groupes d'individus avec différentes caractéristiques, par exemple un groupe de fumeurs et un groupe de non-fumeurs, et à établir de façon prospective si (par exemple) le groupe fumeur développe plus souvent une maladie que le groupe non fumeur après un certain nombre d'années, on procède inversement dans une étude cas-témoins. On part d'un groupe de malades et d'un groupe de non-malades et on établit rétrospectivement si les individus du groupe malade ont plus souvent fumé dans le passé que les individus du groupe non malade. Dans une étude cas-témoins, on n'a donc pas besoin d'attendre que les individus tombent malades pour pouvoir effectuer des comparaisons (on inclut dans notre étude des individus qui sont déjà malades).

Dans une étude cas-témoins, la proportion de malades que l'on inclut dans l'étude ne correspond pas forcément à la proportion de malades dans la population (ce que l'on appelle la *prévalence* de la maladie). En fait, les malades seront souvent sur-représentés. On inclura par exemple 100 individus avec une certaine maladie et 100 individus sans cette maladie même si la prévalence de cette maladie dans la population est beaucoup plus petite que 50 %. On ne pourra donc pas estimer à partir des données d'une étude cas-témoins la proportion de malades (la probabilité de développer la maladie) parmi les fumeurs et la proportion de malades parmi les non-fumeurs. Par contre, on pourra estimer la proportion de fumeurs parmi les malades et la proportion de fumeurs parmi les non-malades, et donc les comparer en calculant une différence de proportion ou un odds-ratio.

Il se trouve pourtant que l'odds-ratio comparant la distribution binaire « fumeur/non-fumeur » entre le groupe malade et le groupe non malade est le même que l'odds-ratio comparant la distribution binaire « malade/non-malade » entre le groupe fumeur et le groupe non fumeur. En notant par X le tabac (avec $X = 1$ pour les fumeurs et $X = 0$ pour les non-fumeurs) et par Y la maladie (avec $Y = 1$ pour les malades et $Y = 0$ pour les non-malades), et en considérant la table de contingence suivante :

	$Y = 1$	$Y = 0$
$X = 1$	a	b
$X = 0$	c	d

ces deux odds-ratios sont en effet égaux à :

$$\omega = \frac{\frac{a/(a+c)}{c/(a+c)}}{\frac{b/(b+d)}{d/(b+d)}} = \frac{\frac{a/(a+b)}{b/(a+b)}}{\frac{c/(c+d)}{d/(c+d)}} = \frac{ad}{bc}.$$

Il s'agit là d'une propriété de symétrie remarquable de l'odds-ratio, que ne partage pas la différence de proportion[3].

Dans une étude cas-témoins, on a une estimation de l'odds-ratio comparant les odds de développer une maladie chez les fumeurs et les non-fumeurs, quand bien même on n'a pas d'estimation de ces odds.

Exemple 16.3 *Une étude cas-témoins célèbre est celle conduite par Richard Doll dans les années 1950, qui a considéré deux groupes de* 1357 *malades du*

[3]On notera au passage que grâce à cette propriété de symétrie, l'odds-ratio est parfois considéré comme une mesure de corrélation entre deux variables binaires.

cancer du poumon (il s'agit des cas) et de 1357 *non-malades (il s'agit des témoins) et recherché (rétrospectivement) qui avait été fumeur et qui avait été non fumeur. Les résultats étaient les suivants :*

	malade	*non-malade*
fumeur	1350	1296
non-fumeur	7	61

À partir de ces données, on ne peut pas estimer les probabilités de développer la maladie (ni donc les odds correspondants) chez les fumeurs et chez les non-fumeurs, les malades étant manifestement sur-repésentés dans cette étude. Par contre, on peut estimer les proportions de fumeurs parmi les malades et parmi les non-malades par $1350/1357 = 0.995$ *et* $1296/1357 = 0.955$*, et donc les odds correspondants par* $0.995/0.005$ *et* $0.955/0.045$*, ainsi que l'odds-ratio les comparant par :*

$$\frac{0.995/0.005}{0.955/0.045} = \frac{1350 \cdot 61}{1296 \cdot 7} = 9.1.$$

Or, cet odds-ratio est également un estimateur de l'odds-ratio comparant les odds de développer la maladie chez les fumeurs et chez les non-fumeurs. On estime ainsi que l'odds de développer la maladie chez les fumeurs est 9.1 *plus élevé que l'odds de développer la maladie chez les non-fumeurs, bien que l'on n'ait pas d'estimation de ces deux odds.*

16.3 Inférence sur l'odds-ratio

On va voir à présent comment faire de l'inférence sur l'odds-ratio. On notera tout d'abord que la distribution d'un odds empirique $\widehat{\pi}/(1-\widehat{\pi})$ est loin d'être normale (où $\widehat{\pi}$ dénote la proportion empirique correspondante). Par contre, une transformation logarithmique le rapprochera de la normalité. On rappellera ensuite qu'en utilisant la formule de Taylor, on a :

$$variance(\log(Y)) \approx \frac{variance(Y)}{mean^2(Y)}.$$

De façon plus générale, on a :

$$covariance(\log(X), \log(Y)) \approx \frac{covariance(X,Y)}{mean(X) \cdot mean(Y)}.$$

On obtient ainsi (en rappelant que l'on a $\mathrm{Var}(\widehat{\pi}) = \pi(1-\pi)/n$, où n dénote la taille de l'échantillon et π la véritable proportion) :

$$\begin{aligned}
\mathrm{Var}(\log(\widehat{\pi}/(1-\widehat{\pi}))) &= \mathrm{Var}(\log(\widehat{\pi}) - \log(1-\widehat{\pi})) \\
&= \mathrm{Var}(\log(\widehat{\pi})) + \mathrm{Var}(\log(1-\widehat{\pi})) \\
&\quad -2\mathrm{Cov}(\log(\widehat{\pi}), \log(1-\widehat{\pi})) \\
&\approx \frac{\pi(1-\pi)/n}{\pi^2} + \frac{\pi(1-\pi)/n}{(1-\pi)^2} - 2\frac{-\pi(1-\pi)/n}{\pi(1-\pi)} \\
&= \frac{1-\pi}{n\pi} + \frac{\pi}{n(1-\pi)} + \frac{2}{n} \\
&= \frac{1}{n\pi} + \frac{1}{n(1-\pi)}.
\end{aligned}$$

Considérons alors un odds-ratio empirique :

$$\widehat{\omega} = \frac{\widehat{\pi}_1(1-\widehat{\pi}_0)}{\widehat{\pi}_0(1-\widehat{\pi}_1)}$$

où $\widehat{\pi}_1$ et $\widehat{\pi}_0$ dénotent les proportions empiriques de 1 calculées à partir de deux échantillons de tailles n_1 et n_0 d'observations indépendantes des variables binaires Y_1 et Y_0 (avec véritables proportions π_1 et π_0) que l'on aimerait comparer. On a ainsi :

$$\log(\widehat{\omega}) = \log\left(\frac{\widehat{\pi}_1/(1-\widehat{\pi}_1)}{\widehat{\pi}_0/(1-\widehat{\pi}_0)}\right) = \log(\widehat{\pi}_1/(1-\widehat{\pi}_1)) - \log(\widehat{\pi}_0/(1-\widehat{\pi}_0))$$

et par conséquent :

$$\begin{aligned}
\mathrm{Var}(\log(\widehat{\omega})) &= \mathrm{Var}(\log(\widehat{\pi}_1/(1-\widehat{\pi}_1)) + \mathrm{Var}(\log(\widehat{\pi}_0/(1-\widehat{\pi}_0)) \\
&\approx \frac{1}{n_1\pi_1} + \frac{1}{n_1(1-\pi_1)} + \frac{1}{n_0\pi_0} + \frac{1}{n_0(1-\pi_0)}.
\end{aligned}$$

Si k_1 et k_0 dénotent le nombre de 1 dans ces échantillons, de sorte que $\widehat{\pi}_1 = k_1/n_1$ et $\widehat{\pi}_0 = k_0/n_0$, on estime cette variance par :

$$\frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_0} + \frac{1}{n_0 - k_0}.$$

Afin d'obtenir un estimateur précis d'un odds-ratio, il faudra donc des fréquences k_1, $n_1 - k_1$, k_0 et $n_0 - k_0$ élevées dans chacune des quatre cases de la table de contingence contenant les données. En dénotant par :

$$\omega = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}$$

le véritable odds-ratio, on calcule un intervalle de confiance de Wald au niveau $1-\alpha$ pour $\log(\omega)$ comme suit :

$$\log(\widehat{\omega}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_0} + \frac{1}{n_0 - k_0}}.$$

Un intervalle de confiance au niveau $1-\alpha$ pour ω s'obtient en calculant l'exponentielle de ces bornes.

Exemple 16.4 *Dans notre étude cas-témoins ci-dessus, on a un odds-ratio empirique de* $\widehat{\omega} = 9.1$*, ce qui implique* $\log(\widehat{\omega}) = 2.206$*. L'erreur type de cet estimateur est estimée par :*

$$\sqrt{\frac{1}{1350} + \frac{1}{61} + \frac{1}{1296} + \frac{1}{7}} = 0.401.$$

On calcule un intervalle de confiance au niveau 95 % *pour* $\log(\omega)$ *comme suit :*

$$2.206 \pm 1.96 \cdot 0.401 = [1.420; 2.290].$$

Un intervalle de confiance au niveau 95 % *pour le véritable odds-ratio* ω *est alors donné par :*

$$[\exp(1.420); \exp(2.290)] = [4.1; 19.9].$$

On a donc prouvé statistiquement que l'odds de développer la maladie chez les fumeurs est au moins 4.1 *fois plus grand que l'odds de développer la maladie chez les non-fumeurs. Comme il s'agit ici d'une maladie rare (même chez les fumeurs), les probabilités de développer cette maladie seront petites, les odds seront proches des probabilités et l'odds-ratio sera proche du ratio des probabilités. On peut donc dire que l'on a prouvé statistiquement que les fumeurs ont un risque au moins* 4.1 *plus élevé que les non-fumeurs de développer la maladie (bien que l'on ne connaisse pas la valeur absolue de ce risque).*

16.4 Régression logistique simple

On va voir à présent comment définir un modèle de régression simple pour prédire une variable réponse Y binaire à partir d'un prédicteur X. Rappelons tout d'abord que dans un modèle de régression linéaire, l'hypothèse de linéarité consiste à supposer que les moyennes de la variable Y calculées dans les différents groupes définis par les différentes valeurs possibles x de X sont alignées sur une droite d'équation $\beta_0 + \beta_1 x$ (la droite de régression). Rappelons ensuite que la moyenne d'une variable binaire est une proportion. En utilisant un modèle de régression linéaire pour une variable réponse Y binaire, on aurait ainsi l'interprétation suivante pour les paramètres β_0 et β_1 de la droite de régression :

- **pour la constante** : $\beta_0 = proportion(Y|X = 0)$

  $\rightarrow$ la constante serait la proportion (d'individus avec $Y = 1$) dans le groupe particulier défini par $X = 0$

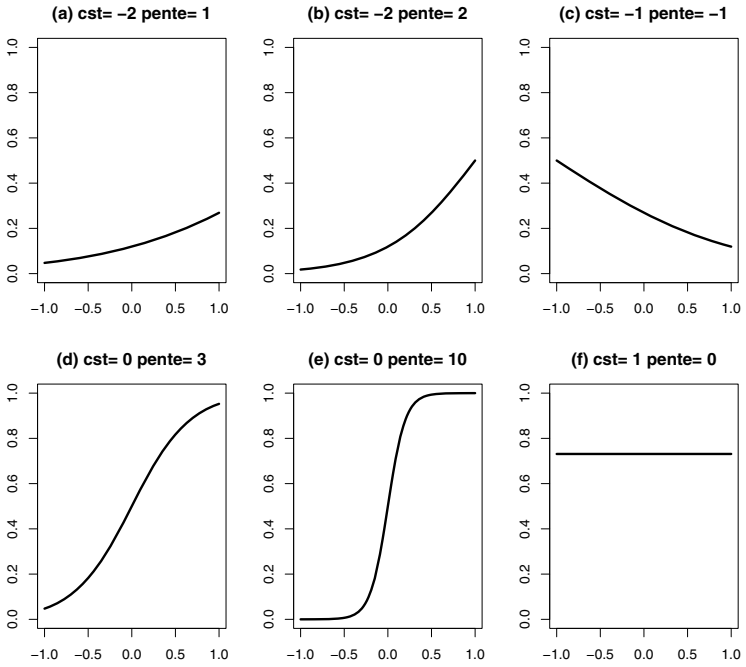

Figure 16.2 – Exemples de fonctions logistiques.

- **pour la pente** : $\beta_1 = proportion(Y|X = x+1) - proportion(Y|X = x)$

  $\rightarrow$ la pente serait une différence des proportions entre deux groupes qui diffèrent d'une unité par rapport à X.

Notons cependant que la proportion $\beta_0 + \beta_1 x$ prédite pour le groupe $X = x$ pourrait être plus petite que 0 ou plus grande que 1 (selon les valeurs de β_0, β_1 et x), ce qui est absurde. Cela est la raison pour laquelle l'hypothèse de linéarité n'est souvent pas réaliste pour une variable binaire.

Afin de définir un modèle qui soit réaliste, il faudra remplacer l'hypothèse de linéarité ci-dessus par une autre hypothèse. On supposera ainsi que les moyennes (ou proportions) de Y dans les différents groupes définis par $X = x$ se trouvent non pas sur une droite, mais sur une courbe bornée entre 0 et 1. En régression logistique, on considère une courbe dite logistique d'équation $\exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_1 x))$. La figure 16.2 nous en montre quelques exemples (avec différentes valeurs pour β_0 et β_1). Une fonction logistique est une fonction monotone croissante de 0 vers 1 si $\beta_1 > 0$, monotone décroissante de 1 vers 0 si $\beta_1 < 0$, et constante si $\beta_1 = 0$.

Dans un modèle de régression logistique, on fait ainsi l'hypothèse suivante :

$$proportion(Y|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Cette hypothèse peut également s'écrire par :

$$odds(Y|X = x) = \exp(\beta_0 + \beta_1 x)$$

ou encore par :

$$\log(odds(Y|X = x)) = \beta_0 + \beta_1 x.$$

Il s'agit à nouveau d'une hypothèse de linéarité. Ce ne sont cependant pas les proportions des différents groupes qui sont supposées être alignées sur une droite, mais les logarithmes des odds correspondants (que l'on appelle *logits*). On a alors l'interprétation suivante pour les paramètres d'un modèle de régression logistique (que l'on continuera à appeler la constante et la pente) :

- **pour la constante**, on a :

  $$\beta_0 = \log(odds(Y|X = 0))$$

  ce qui revient à dire :

  $$\exp(\beta_0) = odds(Y|X = 0)$$

  → la constante est le logarithme de l'odds d'avoir $Y = 1$ dans le groupe particulier $X = 0$

  → cet odds s'obtient en calculant l'exponentielle de cette constante

- **pour la pente**, on a pour n'importe quelle valeur de x :

  $\log(odds(Y|X = x+1)) - \log(odds(Y|X = x)) = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x)$

  et ainsi (en rappelant qu'une différence de logarithmes est égale au logarithme d'un quotient) :

  $$\beta_1 = \log\left(\frac{odds(Y|X = x+1)}{odds(Y|X = x)}\right)$$

  c'est-à-dire :

  $$\exp(\beta_1) = \frac{odds(Y|X = x+1)}{odds(Y|X = x)}$$

  → la pente est le logarithme de l'odds-ratio comparant les odds d'avoir $Y = 1$ entre deux groupes qui diffèrent d'une unité par rapport à X

  → cet odds-ratio s'obtient en calculant l'exponentielle de cette pente

  → $\exp(k\beta_1)$ est l'odds-ratio comparant les odds d'avoir $Y = 1$ entre deux groupes qui diffèrent de k unités par rapport à X.

Alors que les paramètres d'un modèle de régression linéaire pour une variable réponse binaire seraient interprétables comme proportions et différences de proportion, les paramètres d'un modèle de régression logistique pour une variable réponse binaire sont donc interprétables en terme d'odds et d'odds-ratio.

Notons que les hypothèses d'homoscédasticité et de normalité que l'on fait habituellement dans un modèle de régression linéaire ne sont pas non plus satisfaites avec une variable réponse binaire. D'une part, la variance d'une variable binaire dépend par définition de sa moyenne et donc ne sera pas constante si la moyenne ne l'est pas. D'autre part, une variable binaire n'est évidemment pas normale. Un modèle de régression logistique sera donc uniquement constitué d'une hypothèse de linéarité (sur l'échelle du logit)[4], ainsi que d'une hypothèse sur l'indépendance des observations[5].

Les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ des véritables paramètres β_0 et β_1 d'un modèle de régression logistique sont calculés à partir des données observées dans notre échantillon non pas en utilisant la méthode des moindres carrés, mais la méthode dite du *maximum de vraisemblance* que nous introduirons brièvement en fin de chapitre. La probabilité d'avoir $Y = 1$ dans le groupe $X = x$, d'après le modèle égale à $\exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_1 x))$, est donc estimée par $\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)/(1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x))$. Contrairement à la méthode des moindres carrés, on n'a pas de formule explicite pour $\widehat{\beta}_0$ et $\widehat{\beta}_1$ (à part dans certains cas particuliers). Un logiciel statistique sera donc ici indispensable. Il nous donnera également des estimations des erreurs types des estimateurs des paramètres du modèle. Grâce à la normalité approximative de ces estimateurs, on pourra utiliser la méthode de Wald pour calculer des intervalles de confiance pour ces paramètres et effectuer des tests statistiques, qui ne seront pas exacts cependant. Des tests statistiques dits *tests du rapport de vraisemblance*, qui auront en général une meilleure validité que les tests de Wald, sont également disponibles.

Exemple 16.5 *On considère un exemple avec $n = 100$ individus adultes dont on a mesuré l'âge X (en années) et la présence ou l'absence Y d'une maladie*

[4]Alors que l'on considère dans un modèle de régression linéaire une variable $Y|X = x$ normale avec moyenne $\beta_0 + \beta_1 x$ et variance σ^2, on considère dans un modèle de régression logistique une variable $Y|X = x$ binaire avec moyenne (ou proportion) $\exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_1 x))$. Il s'agit de deux cas particuliers de ce que l'on appelle un *modèle linéaire généralisé* (en anglais : `generalized linear model`). Dans un tel modèle, on postule une forme générale pour la distribution de la variable $Y|X = x$ et on suppose qu'une fonction de sa moyenne dépend linéairement de x.

[5]Il y a plusieurs façons de spécifier cette hypothèse d'indépendance en régression logistique, qui correspondent à différentes façons d'échantillonner les données. On peut supposer n observations indépendantes de la variable Y pour des individus avec des valeurs x_i choisies par nous-mêmes, comme on l'a vu en régression linéaire (on aura alors une indépendance *conditionnelle aux valeurs* x_i de notre échantillon). Dans une étude cas-témoins, on suppose par contre n observations indépendantes de la variable X pour des individus avec des valeurs y_i choisies par nous-mêmes, ce que nous n'avions pas en régression linéaire (indépendance *conditionnelle aux valeurs* y_i de notre échantillon). Finalement, on peut supposer n observations indépendantes de la variable bivariée (X, Y).

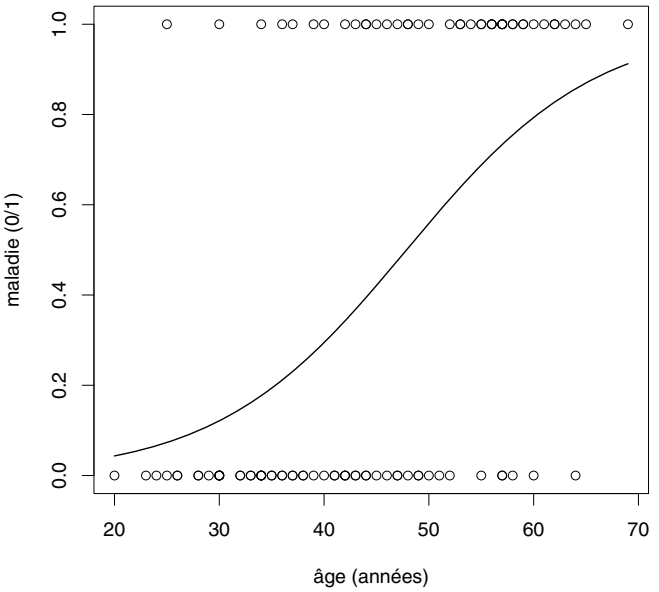

Figure 16.3 – Exemple de régression logistique simple.

coronarienne (avec $Y = 1$ pour les malades et $Y = 0$ pour les non-malades)[6]*. Les données sont représentées dans la figure 16.3. En calculant une régression logistique à l'aide d'un logiciel statistique, on trouve les résultats suivants*[7] *:*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	-5.309	1.134	-4.683	0.000
âge	0.111	0.024	4.610	0.000

On a ici $\widehat{\beta}_0 = -5.309$, $\widehat{\beta}_1 = 0.111$, *et donc* $\exp(\widehat{\beta}_0) = \exp(-5.309) = 0.005$ *et* $\exp(\widehat{\beta}_1) = \exp(0.111) = 1.117$. *On estime ainsi l'odds d'avoir la maladie pour les individus âgés de* $X = 0$ *ans à* 0.005 *(et la probabilité d'avoir la maladie à cet âge à* $0.005/1.005 = 0.005$ *; il s'agit toutefois d'une extrapolation dangereuse, les individus les plus jeunes dans cet échantillon étant âgés de* 20 *ans). L'odds-ratio comparant les odds d'avoir la maladie entre deux groupes d'individus avec une année d'écart est par ailleurs estimé à* 1.117. *Cet odds-ratio est significatif au seuil de* 5 % *avec un test de Wald (*$p < 0.0001$*). Un intervalle de confiance de Wald au niveau* 95 % *pour cet odds-ratio est donné par :*

$$\exp(0.111 \pm 1.96 \cdot 0.024) = [1.066; 1.171].$$

[6] Nous utilisons ici les données publiées à la page 3 du livre de Hosmer et Lemeshow (2000), qui représente une excellente introduction à la régression logistique.

[7] Dans R, on obtient les résultats d'un modèle de régression logistique avec la variable réponse `y` et le prédicteur `x` en utilisant la commande `summary(glm(y~x,family=binomial))`.

Notons que l'odds-ratio comparant les odds d'avoir la maladie entre deux groupes d'individus avec 10 ans d'écart est estimé à $\exp(10 \cdot 0.111) = 3.034$ *avec un intervalle de confiance de :*

$$\exp(10 \cdot 0.111 \pm 1.96 \cdot 10 \cdot 0.024) = [1.890; 4.857].$$

La proportion de malades (la probabilité d'être malade) en fonction de l'âge x *est par ailleurs estimée par la fonction* $\exp(-5.309+0.111x)/(1+\exp(-5.309+0.111x))$ *qui est également représentée dans la figure 16.3. Par exemple, la probabilité d'être malade pour un individu âgé de* 30 *ans est estimée par :*

$$\frac{\exp(-5.309 + 0.111 \cdot 30)}{1 + \exp(-5.309 + 0.111 \cdot 30)} = 0.12$$

et la probabilité d'être malade pour un individu âgé de 60 *ans par :*

$$\frac{\exp(-5.309 + 0.111 \cdot 60)}{1 + \exp(-5.309 + 0.111 \cdot 60)} = 0.79.$$

Exemple 16.6 *Si on calcule un modèle de régression logistique pour les données de notre étude cas-témoins ci-dessus avec comme prédicteur* X *le tabac (*$X = 1$ *pour les fumeurs et* $X = 0$ *pour les non-fumeurs) et comme variable réponse* Y *la maladie (*$Y = 1$ *pour les malades et* $Y = 0$ *pour les non-malades), on obtient :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	-2.165	0.399	-5.425	0.000
tabac	2.206	0.401	5.502	0.000

L'odds-ratio comparant les odds d'être malade entre les fumeurs et les non-fumeurs est donc estimé à $\exp(2.206) = 9.1$*, avec une erreur type de* 0.401 *pour son logarithme, ce qui correspond exactement à ce que l'on a calculé avec la méthode de la section précédente (qui était donc un cas particulier de régression logistique avec un prédicteur binaire). Attention, dans une étude cas-témoins, seule la pente d'un modèle de régression logistique aura du sens. Si les individus avec* $Y = 1$ *sont sur-représentés dans l'échantillon, on ne pourra pas interpréter la constante du modèle, ni estimer de probabilités d'avoir la maladie.*

16.5 Régression logistique multiple

On va à présent généraliser le modèle de régression logistique simple de la section précédente à un modèle de régression logistique multiple, où l'on considère plusieurs prédicteurs (continus ou binaires) notés $X_1, X_2, \cdots, X_m$. L'hypothèse de linéarité consiste à supposer que les logarithmes des odds d'avoir

$Y = 1$ dans les différents groupes, définis par les différentes combinaisons de valeurs possibles des prédicteurs, sont alignés sur un hyperplan. On a ainsi :

$$\log(odds(Y|X_1 = x_1, \cdots, X_m = x_m)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

ce qui implique :

$$proportion(Y|X_1 = x_1, \cdots, X_m = x_m) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}.$$

L'interprétation des paramètres est la suivante :

- **pour la constante**, on a :

  $$\beta_0 = \log(odds(Y|X_1 = 0, \cdots, X_m = 0))$$

  ce qui revient à dire :

  $$\exp(\beta_0) = odds(Y|X_1 = 0, \cdots, X_m = 0)$$

  → la constante est le logarithme de l'odds d'avoir $Y = 1$ dans le groupe particulier $X_1 = \cdots = X_m = 0$

  → cet odds s'obtient en calculant l'exponentielle de cette constante

- **pour la pente associée au prédicteur** X_1, on a pour n'importe quelle valeur de x_1 :

  $$\begin{aligned} &\log(odds(Y|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_m = x_m)) \\ &\quad - \log(odds(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)) \quad = \\ &\qquad \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2 + \cdots + \beta_m x_m \\ &\qquad -(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m) \end{aligned}$$

  et ainsi :

  $$\beta_1 = \log\left(\frac{odds(Y|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_m = x_m)}{odds(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)}\right)$$

  c'est-à-dire :

  $$\exp(\beta_1) = \frac{odds(Y|X_1 = x_1 + 1, X_2 = x_2, \cdots, X_m = x_m)}{odds(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)}$$

  → la pente associée à X_1 est le logarithme de l'odds-ratio comparant les odds d'avoir $Y = 1$ entre deux groupes qui diffèrent d'une unité par rapport à X_1, et qui sont identiques par rapport à tous les autres prédicteurs

  → cet odds-ratio s'obtient en calculant l'exponentielle de cette pente

- **pour la pente associée au prédicteur** X_2, on a pour n'importe quelle valeur de x_2 :

$$\beta_2 = \log\left(\frac{odds(Y|X_1 = x_1, X_2 = x_2 + 1, \cdots, X_m = x_m)}{odds(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)}\right)$$

c'est-à-dire :

$$\exp(\beta_2) = \frac{odds(Y|X_1 = x_1, X_2 = x_2 + 1, \cdots, X_m = x_m)}{odds(Y|X_1 = x_1, X_2 = x_2, \cdots, X_m = x_m)}$$

→ la pente associée à X_2 est le logarithme de l'odds-ratio comparant les odds d'avoir $Y = 1$ entre deux groupes qui diffèrent d'une unité par rapport à X_2, et qui sont identiques par rapport à tous les autres prédicteurs

→ interprétation similaire pour toutes les pentes.

Ici aussi, les estimateurs $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ des véritables paramètres $\beta_0, \beta_1, \cdots, \beta_m$ d'un modèle de régression logistique multiple sont calculés à partir des données observées dans notre échantillon, que l'on supposera indépendantes, en utilisant la méthode non explicite du maximum de vraisemblance. La probabilité d'avoir $Y = 1$ dans le groupe $X_1 = x_1, \cdots, X_m = x_m$, d'après le modèle égale à $\exp(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)/(1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m))$, est donc estimée par $\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m)/(1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_m x_m))$. On aura également des estimations des erreurs types des estimateurs des paramètres du modèle, et on pourra utiliser la méthode de Wald pour calculer des intervalles de confiance pour ces paramètres et effectuer des tests statistiques (qui ne seront pas exacts). Alternativement, on pourra effectuer des tests du rapport de vraisemblance, que l'on pourra également utiliser pour tester la nullité simultanée de plusieurs paramètres dans le modèle.

Exemple 16.7 *On considère un exemple de régression logistique calculé à partir de* $n = 3141$ *individus dont on a mesuré l'âge* X_1 *(en années), le cholestérol* X_2 *(en mg/dL), la pression systolique* X_3 *(en mmHg), le bmi* X_4 *(le poids divisé par la taille au carré, en kg/m*^{2}*) et le tabac* X_5 *(*$X_5 = 1$ *pour les fumeurs et* $X_5 = 0$ *pour les non-fumeurs). La variable réponse* Y *est à nouveau la présence ou l'absence d'une maladie coronarienne (*$Y = 1$ *pour les malades et* $Y = 0$ *pour les non-malades). On a les résultats suivants*[8] *:*

[8] Nous reproduisons ici le tableau 6.6 du livre de Vittinghoff *et al.* (2005). Notons aussi que dans R, on peut obtenir les résultats d'un modèle de régression logistique multiple avec une variable réponse binaire `y` et cinq prédicteurs `x1, x2, x3, x4` et `x5` en utilisant la commande `summary(glm(y~x1+x2+x3+x4+x5,family=binomial))`.

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−12.311	0.977	−12.598	0.000
âge	0.064	0.012	5.412	0.000
cholestérol	0.011	0.002	7.080	0.000
pression systolique	0.019	0.004	4.716	0.000
bmi	0.057	0.026	2.179	0.029
tabac	0.634	0.140	4.526	0.000

Ces prédicteurs sont tous significatifs (on a donc montré statistiquement qu'ils ne sont ni inutiles ni redondants). En calculant l'exponentielle de ces cinq pentes, on obtient des estimations d'odds-ratio. Par exemple, l'odds-ratio comparant les odds de développer la maladie entre un groupe de fumeurs et un groupe de non-fumeurs, qui ont par ailleurs même âge, même cholestérol, même pression systolique et même bmi, est estimé à $\exp(0.634) = 1.885$*, alors qu'un intervalle de confiance au niveau* 95 % *pour cet odds-ratio est donné par :*

$$\exp(0.634 \pm 1.96 \cdot 0.140) = [1.433; 2.480].$$

Par ailleurs, l'exponentielle de la constante $\exp(-12.311) = 0.000$ *nous donne une estimation de l'odds de développer la maladie dans le groupe de référence où tous les prédicteurs sont égaux à* 0*, qui n'a cependant pas de sens ici (vu qu'il n'existe personne avec un cholestérol, une pression systolique et un bmi de* 0*). Par contre, cette constante est importante pour estimer les odds (et donc les probabilités) de développer la maladie dans les différents groupes. Par exemple, l'odds de développer la maladie pour un individu non fumeur âgé de* 50 *ans, avec un cholestérol de* 220 *mg/dL, une pression systolique de* 150 *mmHg et un bmi de* 25 *kg/m*^{2} *est estimé par :*

$$\exp(-12.311 + 0.064 \cdot 50 + 0.011 \cdot 220 + 0.019 \cdot 150 + 0.057 \cdot 25) = 0.089$$

ce qui correspond à une probabilité de développer la maladie de $0.089/1.089 = 0.082$*, alors que pour un individu semblable fumeur, on estime un odds de :*

$$\exp(-12.311 + 0.064 \cdot 50 + 0.011 \cdot 220 + 0.019 \cdot 150 + 0.057 \cdot 25 + 0.634) = 0.168$$

(qui est effectivement 1.885 *plus élevé que pour le non-fumeur ci-dessus), ce qui correspond à une probabilité de développer la maladie de* $0.168/1.168 = 0.144$*.*

Exemple 16.8 *Nous présentons ici les données d'une étude cas-témoins, avec* 483 *cas de personnes atteintes du cancer de la cavité buccale et* 447 *témoins, pour lesquels on connaît la consommation de cigarettes (le nombre de cigarettes par jour, avec quatre modalités) et la consommation d'alcool (le nombre de grammes par jour, avec quatre modalités)*[9]*. Le tabac et l'alcool sont ici les deux prédicteurs de la présence ou non de la maladie dans un modèle de régression logistique. Le tableau ci-dessous nous donne les fréquences des cas et*

[9]Ces données ont été décrites dans l'article de Rothman et Keller (1972).

des témoins (ces derniers entre parenthèses) dans les 16 *groupes définis par les différentes combinaisons de valeurs possibles des deux prédicteurs :*

	tabac							
alcool	0		1–19		20–39		40+	
0	10	(38)	11	(26)	13	(36)	9	(8)
1–9	7	(27)	16	(35)	50	(60)	16	(19)
10–44	4	(12)	18	(16)	60	(49)	27	(14)
45+	5	(8)	21	(20)	125	(52)	91	(27)

En considérant les prédicteurs comme des variables qualitatives, on obtient les résultats suivants (les non-fumeurs et les non-buveurs constituant le groupe de référence) :

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−1.609	0.265	−6.063	0.000
tabac 1–19	0.589	0.284	2.070	0.038
tabac 20–39	1.026	0.255	4.017	0.000
tabac 40+	1.409	0.282	4.988	0.000
alcool 1–9	0.290	0.233	1.245	0.213
alcool 10–44	0.844	0.238	3.540	0.000
alcool 45+	1.378	0.226	6.109	0.000

Un test du rapport de vraisemblance sur la nullité simultanée des trois pentes associées au prédicteur « tabac » nous donnera un résultat significatif ($p < 0.0001$)*, et de même pour un test sur la nullité simultanée des trois pentes associées au prédicteur « alcool »* ($p < 0.0001$)*. On en conclut que le tabac et l'alcool sont deux prédicteurs ni inutiles ni redondants de la présence de la maladie. Par ailleurs, à consommation d'alcool égale, on estime les odds-ratios suivants (comparant les odds de développer la maladie entre deux groupes) :*

- $\exp(0.589) = 1.802$ *entre les non-fumeurs et les fumeurs avec* 1–19 *cigarettes par jour*
- $\exp(1.026) = 2.790$ *entre les non-fumeurs et les fumeurs avec* 20–39 *cigarettes par jour*
- $\exp(1.409) = 4.092$ *entre les non-fumeurs et les fumeurs avec plus de* 40 *cigarettes par jour.*

On en déduit également les odds-ratios suivants :

- $\exp(1.026 - 0.589) = 1.548$ *entre les fumeurs avec* 1–19 *cigarettes par jour et les fumeurs avec* 20–39 *cigarettes par jour*
- $\exp(1.409 - 1.026) = 1.467$ *entre les fumeurs avec* 20–39 *cigarettes par jour et les fumeurs avec plus de* 40 *cigarettes par jour.*

De même, à consommation de tabac égale, on estime les odds-ratios suivants :

- $\exp(0.290) = 1.336$ *entre les non-buveurs et les buveurs avec* 1–9 *grammes par jour*
- $\exp(0.844 - 0.290) = 1.740$ *entre les buveurs avec* 1–9 *grammes et les buveurs avec* 10–44 *grammes par jour*
- $\exp(1.378 - 0.844) = 1.706$ *entre les les buveurs avec* 10–44 *et les buveurs avec plus de* 45 *grammes par jour.*

On remarque que les odds-ratios comparant deux modalités adjacentes sont assez proches les uns des autres, à la fois pour le tabac et pour l'alcool. Ceci nous suggère d'utiliser un modèle plus simple, où l'on considère le tabac et l'alcool comme prédicteurs quantitatifs (avec les valeurs 0, 1, 2 *et* 3 *pour leurs quatre modalités), pour lequel on obtient :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	−1.629	0.186	−9.350	0.000
tabac	0.452	0.083	5.420	0.000
alcool	0.490	0.068	7.253	0.000

Dans ce modèle, on estime les odds-ratios suivants :

- $\exp(0.452) = 1.571$ *si on compare deux groupes avec une même consommation d'alcool et qui diffèrent d'une modalité par rapport au tabac*
- $\exp(0.490) = 1.632$ *si on compare deux groupes avec une même consommation de tabac et qui diffèrent d'une modalité par rapport à l'alcool*
- $\exp(0.452 + 0.490) = 2.565$ *si on compare deux groupes qui diffèrent d'une modalité par rapport au tabac et d'une modalité par rapport à l'alcool.*

Notons que ce second modèle avec prédicteurs quantitatifs n'a que 3 *paramètres alors que le premier modèle avec prédicteurs qualitatifs en avait* 7*. En adoptant une certaine paramétrisation, on peut montrer que le second modèle est un cas particulier du premier avec* 4 *paramètres nuls. À l'aide d'un test du rapport de vraisemblance, on peut tester la nullité simultanée de ces* 4 *paramètres dans le premier modèle et on trouve un résultat non significatif (*$p = 0.89$*). On en conclut que le premier modèle n'améliore pas significativement la prédiction par rapport au second, ce qui constitue une motivation pour utiliser le second.*

16.6 Ajustement pour les variables confondantes

Comme on l'a vu dans le cadre d'un modèle de régression linéaire, la régression multiple nous permet de décrire une association entre deux variables en s'affranchissant de l'influence potentielle de variables confondantes. Nous

ré-introduisons ci-dessous la problématique des variables confondantes à l'aide d'un exemple provenant de la littérature[10].

Dans une étude rétrospective, on veut comparer les taux de succès de deux techniques d'opération (une ancienne et une nouvelle technique) pour éliminer un calcul rénal. En recherchant dans les dossiers médicaux, on retrouve les résultats suivants pour 350 opérations effectuées avec chaque technique :

	succès	échec	total
nouvelle technique	289	61	350
ancienne technique	273	77	350

On estime ainsi des probabilités de succès de $289/350 = 0.83$ pour la nouvelle technique et de $273/350 = 0.78$ pour l'ancienne technique, ce qui correspond à un odds-ratio de $(289 \cdot 77)/(61 \cdot 273) = 1.34$ en faveur de la nouvelle technique.

En regardant les dossiers médicaux plus en détail, on s'aperçoit cependant que dans la plupart des cas la nouvelle technique a été utilisée pour éliminer de petits calculs et l'ancienne technique pour éliminer de gros calculs (plus difficiles à éliminer). En comparant les deux techniques séparément pour petits et gros calculs, on obtient alors les résultats suivants :

	petit calcul			gros calcul		
	succès	échec	total	succès	échec	total
nouvelle technique	234	36	270	55	25	80
ancienne technique	81	6	87	192	71	263

Pour les petits calculs, on estime ainsi des probabilités de succès de $234/270 = 0.87$ avec la nouvelle technique et de $81/87 = 0.93$ avec l'ancienne technique. Pour les gros calculs, on estime des probabilités de succès de $55/80 = 0.69$ avec la nouvelle technique et de $192/263 = 0.73$ avec l'ancienne technique. Cela correspond à des odds-ratios de $(234 \cdot 6)/(36 \cdot 81) = 0.48$ et de $(55 \cdot 71)/(25 \cdot 192) = 0.81$, cette fois-ci en défaveur de la nouvelle technique. Ainsi, lorsque les deux techniques d'opération sont comparées à taille de calcul égale, la nouvelle technique est moins performante que l'ancienne et ceci à la fois dans le sous-groupe des petits calculs et dans le sous-groupe des gros calculs, alors qu'elle était plus performante dans la comparaison globale. Un tel retournement de situation est connu sous le nom de *paradoxe de Simpson*.

La raison de ce paradoxe est que l'association entre la technique d'opération et la réussite de cette opération est confondue par la taille du calcul. Afin d'éliminer l'influence de cette variable confondante, on a les possibilités suivantes :

- on peut effectuer une **analyse séparée dans chaque sous-groupe** (aussi appelée *analyse stratifiée*, comme on l'a fait ci-dessus)

[10] Il s'agit d'une étude publiée par Charig *et al.* (1986).

→ on compare les odds d'avoir un succès entre les deux techniques d'opération séparément pour les petits et pour les gros calculs

→ on estime ainsi deux odds-ratios, l'un dans le sous-groupe des petits calculs, l'autre dans le sous-groupe des gros calculs

→ on perd cependant de la puissance statistique, les estimations étant faites sur des échantillons plus petits

→ on est également confronté à la problématique des tests multiples

- on peut considérer un **modèle de régression logistique multiple** avec la technique d'opération et la taille du calcul comme prédicteurs de la réussite de l'opération

  → on compare les odds d'avoir un succès entre les deux techniques d'opération en *ajustant pour la taille du calcul*

  → on estime un seul odds-ratio, valable à la fois dans le sous-groupe des petits calculs et dans le sous-groupe des gros calculs

  → cet unique odds-ratio sera un bon résumé de la situation s'il n'y a pas d'interaction importante entre les prédicteurs (c'est-à-dire si les deux odds-ratios calculés avec la première approche, dans les sous-groupes des petits et des gros calculs, ne diffèrent pas trop l'un de l'autre)

  → cette seconde approche correspond à un modèle de régression logistique sans interaction entre les prédicteurs (alors que la première approche correspond à un modèle avec interaction, comme on va le voir ci-dessous).

Exemple 16.9 *On reprend l'exemple ci-dessus où l'on s'intéresse à l'association entre la technique X_1 de l'opération ($X_1 = 1$ pour la nouvelle technique, $X_1 = 0$ pour l'ancienne technique) et la réussite Y de l'opération ($Y = 1$ pour un succès, $Y = 0$ pour un échec). Une comparaison globale correspond à un modèle de régression logistique simple où l'on prédit Y à partir de X_1. On obtient ainsi :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	1.266	0.129	9.809	0.000
technique	0.290	0.191	1.517	0.129

L'odds-ratio comparant les odds de réussir l'opération entre les deux techniques est donc estimé à $\exp(0.290) = 1.34$*, ce qui correspond à l'odds-ratio empirique calculé ci-dessus, suggérant une meilleure réussite pour la nouvelle technique d'opération bien que ce résultat soit non significatif ($p = 0.129$).*

On a vu cependant que cette association était confondue par la taille X_2 du calcul (avec $X_2 = 1$ pour un gros calcul, $X_2 = 0$ pour un petit calcul). En introduisant X_2 dans le modèle, on obtient les résultats suivants :

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	2.294	0.247	9.289	0.000
technique	−0.357	0.229	−1.559	0.119
taille	−1.261	0.239	−5.274	0.000

L'odds-ratio ajusté, comparant les odds de réussir l'opération entre les deux techniques à taille de calcul égale, est donc estimé à $\exp(-0.357) = 0.70$*, suggérant cette fois-ci une moins bonne réussite pour la nouvelle technique d'opération, bien que ce résultat ne soit pas non plus significatif (*$p = 0.119$*). Avec ce modèle, on estime par ailleurs les odds suivants de réussir une opération :*

- *ancienne technique, petit calcul :* $\exp(2.294) = 9.914$
- *nouvelle technique, petit calcul :* $\exp(2.294 - 0.357) = 6.94$
- *ancienne technique, gros calcul :* $\exp(2.294 - 1.261) = 2.81$
- *nouvelle technique, gros calcul :* $\exp(2.294 - 0.357 - 1.261) = 1.97$.

Ces odds correspondent à des probabilités de succès estimées respectivement à 0.91*,* 0.87*,* 0.74 *et* 0.66*, qui sont assez proches des probabilités empiriques calculées ci-dessus.*

Le modèle précédent est un modèle sans interaction entre les prédicteurs. On suppose ainsi un même odds-ratio que l'on compare les deux techniques d'opération dans le sous-groupe des petits ou dans celui des gros calculs. Si on introduit une interaction dans le modèle, on obtient :

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	2.603	0.423	6.152	0.000
technique	−0.731	0.459	*−1.591*	0.112
taille	−1.608	0.445	−3.611	0.000
technique:taille	0.525	0.537	0.976	0.329

Dans ce nouveau modèle, on estime les odds-ratios suivants (comparant les odds de réussir l'opération entre les deux techniques) :

- *pour les petits calculs :* $\exp(-0.731) = 0.48$
- *pour les gros calculs :* $\exp(-0.731 + 0.525) = 0.81$.

On retrouve de la sorte l'approche consistant à estimer les odds-ratios séparément dans les sous-groupes des petits et des gros calculs (on note cependant que l'interaction n'est pas significative dans ce modèle, ce qui nous motivera à utiliser le modèle précédent et donc à utiliser comme résumé de la situation un unique odds-ratio ajusté de 0.70 *plutôt que ces deux odds-ratios différents de* 0.48 *et* 0.81*). Avec ce nouveau modèle, on estime par ailleurs les odds suivants de réussir une opération :*

- *ancienne technique, petit calcul :* $\exp(2.603) = 13.50$
- *nouvelle technique, petit calcul :* $\exp(2.603 - 0.731) = 6.50$
- *ancienne technique, gros calcul :* $\exp(2.603 - 1.608) = 2.70$
- *nouvelle technique, gros calcul :* $\exp(2.603 - 0.731 - 1.608 + 0.525) = 2.20$.

Ces odds correspondent à des probabilités de succès estimées respectivement à 0.93*,* 0.87*,* 0.73 *et* 0.69*. Comme le modèle contient quatre paramètres et comme il y a seulement quatre probabilités à estimer, ces estimations correspondent exactement aux probabilités empiriques calculées ci-dessus.*

16.7 Comparaison de deux groupes dans un essai clinique

L'exemple de la section précédente était une étude observationnelle rétrospective. Si notre but était de montrer une relation de cause à effet entre la technique de l'opération et la réussite de cette opération, il faudrait effectuer un essai clinique, c'est-à-dire randomiser les patients en deux groupes, les patients du premier groupe étant opérés avec la nouvelle technique, ceux du second groupe avec l'ancienne technique. La randomisation nous assurerait que l'on ait (approximativement) la même répartition de gros et de petits calculs dans les deux groupes, et plus généralement que les deux groupes soient comparables par rapport à toutes les variables confondantes possibles et imaginables.

Voici les résultats d'une tel essai clinique (fictif), où l'on compare 350 opérations effectuées avec chaque technique. Dans cet exemple, on a dans chaque groupe 250 petits calculs et 100 gros calculs avec les résultats suivants :

	petits calculs			gros calculs		
	succès	échec	total	succès	échec	total
nouvelle technique	225	25	250	50	50	100
ancienne technique	125	125	250	10	90	100

Pour les petits calculs, on estime des probabilités de succès de $225/250 = 0.9$ avec la nouvelle technique et de $125/250 = 0.5$ avec l'ancienne technique. Pour les gros calculs, on estime des probabilités de succès de $50/100 = 0.5$ avec la nouvelle technique et de $10/100 = 0.1$ avec l'ancienne technique. On a ainsi plus de réussite avec la nouvelle technique à la fois dans le sous-groupe des petits et dans celui des gros calculs. Qui plus est, on estime ici un odds-ratio identique dans les deux sous-groupes, donné par $(225 \cdot 125)/(25 \cdot 125) = (50 \cdot 90)/(50 \cdot 10) = 9$. Comme il s'agit d'un essai clinique, il est ici légitime d'interpréter cet odds-ratio comme l'effet causal de la technique d'opération sur la réussite de l'opération, la nouvelle technique améliorant d'un facteur 9 l'odds de réussir l'opération (que ce soit pour un petit ou pour un gros calcul).

Bien que les deux groupes soient ici parfaitement comparables et bien que l'on ait un odds-ratio identique dans chaque sous-groupe, il est intéressant (et même troublant) de constater qu'une comparaison globale ne nous donne pas le même résultat. En regroupant petits et gros calculs, on a en effet :

	succès	échec	total
nouvelle technique	275	75	350
ancienne technique	135	215	350

On estime ici des probabilités de succès de $275/350 = 0.79$ pour la nouvelle technique et de $135/350 = 0.39$ pour l'ancienne technique, ce qui correspond à un odds-ratio de $(275 \cdot 215)/(75 \cdot 135) = 5.84$. Cet odds-ratio est certes toujours en faveur de la nouvelle technique, mais il est bien plus petit que l'odds-ratio de 9 calculé dans chacun des sous-groupes[11].

Ce phénomène est lié au résultat important suivant : les pentes associées aux prédicteurs dans un modèle de régression logistique se modifient lorsque l'on introduit un nouveau prédicteur dans le modèle, *même si ce dernier est non corrélé avec les prédicteurs déjà présents dans le modèle*. Rappelons que ce n'était pas le cas dans un modèle de régression linéaire.

> L'introduction d'un nouveau prédicteur dans un modèle de régression logistique modifie les pentes des prédicteurs déjà dans le modèle même si ces prédicteurs sont non corrélés avec le nouveau prédicteur.

[11] Dans un essai clinique, on n'aura pas de retournement de situation aussi spectaculaire que celui noté dans l'étude observationnelle de la section précédente, mais il est donc possible d'obtenir un odds-ratio plus élevé dans une comparaison au sein des sous-groupes que dans une comparaison globale. En fait, il ne s'agit pas du même odds-ratio qui est estimé dans ces deux analyses. Dans la seconde analyse, on compare les probabilités de réussite de l'opération entre deux patients pris au hasard dans toute la population, l'un opéré avec la nouvelle, l'autre avec l'ancienne technique. Dans la première analyse, on compare les probabilités de réussite de l'opération entre deux patients opérés avec des techniques différentes, mais avec la même taille de calcul. L'odds-ratio de la seconde analyse est appelé *l'odds-ratio au niveau de la population*. L'odds-ratio de la première analyse est appelé *l'odds-ratio au niveau d'un sous-groupe*. En poussant la logique plus loin, on pourrait également s'intéresser à *l'odds-ratio au niveau d'un individu*, où l'on comparerait les probabilités de réussite de l'opération pour un même individu opéré avec la nouvelle ou avec l'ancienne technique (que l'on ne pourra cependant pas estimer faute de données, car il ne sera pas possible en général d'opérer un même individu avec deux techniques différentes). Notons qu'un odds-ratio calculé dans un sous-groupe sera en général plus élevé qu'un odds-ratio calculé dans toute la population. De même, un odds-ratio calculé au niveau d'un individu sera en général plus élevé qu'un odds-ratio calculé dans un sous-groupe. Dans le cadre d'un essai clinique, lorsqu'on estime un odds-ratio au niveau de la population ou au niveau des sous-groupes, on aura ainsi tendance à sous-estimer les odds-ratios que l'on aurait au niveau des individus.

Exemple 16.10 *On reprend l'exemple de l'essai clinique ci-dessus où l'on s'intéresse à l'effet de la technique X_1 de l'opération ($X_1 = 1$ pour la nouvelle technique, $X_1 = 0$ pour l'ancienne technique) sur la réussite Y de l'opération ($Y = 1$ pour un succès, $Y = 0$ pour un échec). En calculant un modèle de régression logistique où l'on prédit Y à partir de X_1, on obtient :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	-0.465	0.110	-4.238	0.000
technique	1.765	0.170	10.357	0.000

On estime ainsi l'effet de la technique d'opération sur la réussite de l'opération via un odds-ratio de $\exp(1.765) = 5.84$. *En introduisant toutefois la taille X_2 du calcul (avec $X_2 = 1$ pour un gros calcul, $X_2 = 0$ pour un petit calcul) dans le modèle, on obtient :*

	estimation	*erreur type*	t_{stat}	*valeur p*
(constante)	0.000	0.122	0.000	1.000
technique	2.197	0.208	10.575	0.000
taille	-2.197	0.225	-9.755	0.000

En comparant à taille de calcul égale, on estime à présent un effet de la technique d'opération sur la réussite de l'opération via un odds-ratio de $\exp(2.197) = 9$, *ce qui correspond à un plus grand effet que dans le premier modèle (bien que le prédicteur X_2 introduit dans ce second modèle soit non corrélé avec X_1).*

16.8 Sensibilité, spécificité et courbe ROC

Afin d'obtenir une mesure de la qualité globale de la prédiction en régression linéaire, on calculait (une version corrigée du carré de) la corrélation entre les y_i et les $\widehat{y}_i^*$, où y_i dénotait la valeur observée et $\widehat{y}_i^*$ la valeur prédite de la variable réponse Y pour le i-ième individu de notre échantillon ($i = 1, \cdots, n$). En régression logistique, les valeurs observées y_i sont binaires. Afin d'obtenir des valeurs prédites $\widehat{y}_i^*$ qui sont également binaires, on calculera tout d'abord les « probabilités estimées » données par :

$$\widehat{\pi}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_m x_{im})}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_m x_{im})}.$$

La quantité $\widehat{\pi}_i$ est une estimation de la probabilité d'avoir $Y = 1$ pour des individus semblables au i-ième individu de notre échantillon par rapport aux valeurs des prédicteurs (où x_{ij} dénote la valeur du j-ième prédicteur pour le i-ième individu de notre échantillon, pour $i = 1, \cdots, n$ et $j = 1, \cdots, m$). On dichotomisera ensuite ces probabilités estimées de la façon suivante :

- $\widehat{y}_i^* = 1$ si $\widehat{\pi}_i \geq C$

- $\widehat{y}_i^* = 0$ si $\widehat{\pi}_i < C$

où C désigne une limite critique (par exemple $C = 0.5$). On va voir ci-dessous comment on peut comparer les y_i et les $\widehat{y}_i^*$ en régression logistique afin de mesurer la qualité globale de la prédiction.

Supposons (pour contextualiser notre discussion) que la variable réponse Y représente une maladie (avec $Y = 1$ pour les malades et $Y = 0$ pour les non-malades). Une valeur prédite peut alors être interprétée comme le résultat d'un diagnostic que l'on établirait en utilisant notre modèle de régression logistique, le diagnostic étant positif pour les individus avec $\widehat{y}_i^* = 1$ et négatif pour les individus avec $\widehat{y}_i^* = 0$. Par ailleurs, les individus réellement malades sont ceux avec $y_i = 1$, les individus avec $y_i = 0$ étant les non-malades. On peut alors répartir nos individus dans la table de contingence suivante :

	$y_i = 1$ (malade)	$y_i = 0$ (non-malade)
$\widehat{y}_i^* = 1$ (diagnostic positif)	a	b
$\widehat{y}_i^* = 0$ (diagnostic négatif)	c	d

Les a individus malades avec diagnostic positif sont les « vrais positifs », les b individus non malades avec diagnostic positif sont les « faux positifs », les c individus malades avec diagnostic négatif sont les « faux négatifs », et les d individus non malades avec diagnostic négatif sont les « vrais négatifs ». À partir de cette table de contingence, on peut estimer les probabilités suivantes[12] :

- la **sensibilité** du diagnostic (en anglais : `sensitivity`) : la probabilité d'établir un diagnostic positif pour les individus réellement malades

  → on estime cette probabilité par $a/(a + c)$

- la **spécificité** du diagnostic (en anglais : `specificity`) : la probabilité d'établir un diagnostic négatif pour les individus réellement non malades

  → on estime cette probabilité par $d/(b + d)$

- la **valeur prédictive positive** du diagnostic : la probabilité d'être malade si le diagnostic est positif

  → on estime cette probabilité par $a/(a + b)$

[12] Notons toutefois que la validité de ces estimations dépend de la façon dont les données sont échantillonnées. Nous avions évoqué trois possibilités. Si on choisit nous-mêmes les valeurs de certains prédicteurs pour les individus de notre échantillon (où certaines valeurs pourront être sur-représentées et d'autres sous-représentées), les estimations de la prévalence, de la sensibilité et de la spécificité ne seront en général pas valides, alors que les estimations des valeurs prédictives positives et négatives le seront. Si on choisit nous-mêmes les valeurs de la variable réponse pour les individus de notre échantillon (étude cas-témoins), les estimations de la prévalence, de la valeur prédictive positive et de la valeur prédictive négative ne seront en général pas valides, alors que les estimations de la sensibilité et de la spécificité le seront. Si on ne choisit ni les valeurs des prédicteurs, ni celles de la variable réponse, mais que l'on observe un échantillon représentatif de la population, les estimations de ces cinq probabilités seront valides.

- la **valeur prédictive négative** du diagnostic : la probabilité d'être non malade si le diagnostic est négatif

  → on estime cette probabilité par $d/(c+d)$

- la **prévalence** de la maladie : la proportion d'individus avec la maladie (la probabilité d'être malade) dans la population d'intérêt

  → on estime cette probabilité par $(a+c)/(a+b+c+d)$.

Notons qu'à partir de la sensibilité, de la spécificité et de la prévalence, on peut calculer la valeur prédictive positive par[13] :

$$\frac{\text{sensibilité} \cdot \text{prévalence}}{\text{sensibilité} \cdot \text{prévalence} + (1 - \text{specificité}) \cdot (1 - \text{prévalence})}$$

et la valeur prédictive négative par :

$$\frac{\text{spécificité} \cdot (1 - \text{prévalence})}{\text{spécificité} \cdot (1 - \text{prévalence}) + (1 - \text{sensibilité}) \cdot \text{prévalence}}.$$

Ces valeurs prédictives sont ainsi d'autant plus grandes que l'on a une grande sensibilité et une grande spécificité. La valeur prédictive positive est par ailleurs d'autant plus grande que la prévalence est grande, alors que la valeur prédictive négative est d'autant plus grande que la prévalence est petite.

Exemple 16.11 *Si on utilise un diagnostic avec une sensibilité de* 0.9 *et une spécificité de* 0.8*, et si la prévalence de la maladie est de* 0.005*, la valeur prédictive positive (la probabilité d'être effectivement malade pour un individu avec un diagnostic positif) sera seulement de :*

$$\frac{0.9 \cdot 0.005}{0.9 \cdot 0.005 + (1 - 0.8)(1 - 0.005)} = 0.02$$

alors qu'elle sera de :

$$\frac{0.9 \cdot 0.2}{0.9 \cdot 0.2 + (1 - 0.8)(1 - 0.2)} = 0.53$$

si la prévalence est de 0.2*. La valeur prédictive négative (la probabilité d'être effectivement non malade pour un individu avec un diagnostic négatif) sera par ailleurs de :*

$$\frac{0.8 \cdot (1 - 0.005)}{0.8 \cdot (1 - 0.005) + (1 - 0.9) \cdot 0.005} = 0.999$$

[13] Ces formules sont des cas particuliers du *Théorème de Bayes*, qui dit que :

$$\Pr\{A|B\} = \frac{\Pr\{B|A\} \cdot \Pr\{A\}}{\Pr\{B|A\} \cdot \Pr\{A\} + \Pr\{B|\bar{A}\} \cdot \Pr\{\bar{A}\}}$$

où A et B sont des « événements », $\bar{A}$ dénote l'événement contraire de A, $\Pr\{A\}$ la probabilité d'avoir A et $\Pr\{A|B\}$ la probabilité d'avoir A sachant que l'on a B. On obtient la première formule ci-dessus avec A le fait d'être malade, $\bar{A}$ le fait d'être non malade et B le fait d'avoir un diagnostic positif. On obtient la seconde formule ci-dessus avec A le fait d'être non malade, $\bar{A}$ le fait d'être malade et B le fait d'avoir un diagnostic négatif.

si la prévalence est de 0.005 *et de :*

$$\frac{0.8 \cdot (1 - 0.2)}{0.8 \cdot (1 - 0.2) + (1 - 0.9) \cdot 0.2} = 0.97$$

si la prévalence est de 0.2.

Un bon diagnostic est un diagnostic avec une grande sensibilité et une grande spécificité. Si un diagnostic est établi à partir d'un modèle de régression logistique, la sensibilité et la spécificité dépendront de la limite critique C utilisée ci-dessus. On pourra augmenter la sensibilité en abaissant cette limite critique, mais cela se fera au détriment de la spécificité qui dans le même temps baissera. Dans le cas extrême $C = 0$, on diagnostiquera la maladie chez tout le monde, de sorte que la sensibilité sera de 100 % et la spécificité de 0 %. De même, on pourra augmenter la spécificité en élevant cette limite critique, mais cela se fera au détriment de la sensibilité qui dans le même temps baissera. Dans le cas extrême $C = 1$, on ne diagnostiquera la maladie chez personne, de sorte que la spécificité sera de 100 % et la sensibilité de 0 %. Le choix d'une limite critique C est donc un compromis entre sensibilité et spécificité. Dans certaines applications, la sensibilité sera plus importante que la spécificité et on choisira ainsi une limite critique basse. Dans d'autres, la spécificité sera plus importante que la sensibilité et on choisira une limite critique élevée.

On peut mesurer la qualité globale de la prédiction d'un modèle de régression logistique en mesurant la qualité globale de ces différents diagnostics que l'on peut établir à l'aide de ce modèle. Une mesure souvent utilisée est l'*aire sous la courbe ROC*, que l'on calcule comme suit[14] :

- à partir des données de l'échantillon, on estime **la sensibilité et la spécificité** que l'on peut atteindre en choisissant les différentes limites critiques données par $C = \widehat{\pi}_i$ (pour $i = 1, \cdots, n$)

  → on obtient des estimations de (au plus) n couples possibles (1 − spécificité, sensibilité)

  → on y ajoute les deux couples suivants : (1,1) (obtenu avec $C = 0$) et (0,0) (obtenu avec $C = 1$)

- on représente ces **différents couples possibles** par des points sur un graphique (où l'axe horizontal représente 1− la spécificité et l'axe vertical représente la sensibilité) et on relie ces points par des segments

  → la courbe ainsi obtenue est une estimation de la courbe ROC (de l'anglais : `Receiver Operating Characteristic`)

  → d'une manière générale, les diagnostics que l'on peut établir à l'aide du modèle seront d'autant meilleurs que cette courbe sera loin (au-dessus) de la diagonale

[14] D'autres mesures de la qualité globale de la prédiction en régression logistique on été proposées, voir par exemple l'article de Mittlböck et Schemper (1996).

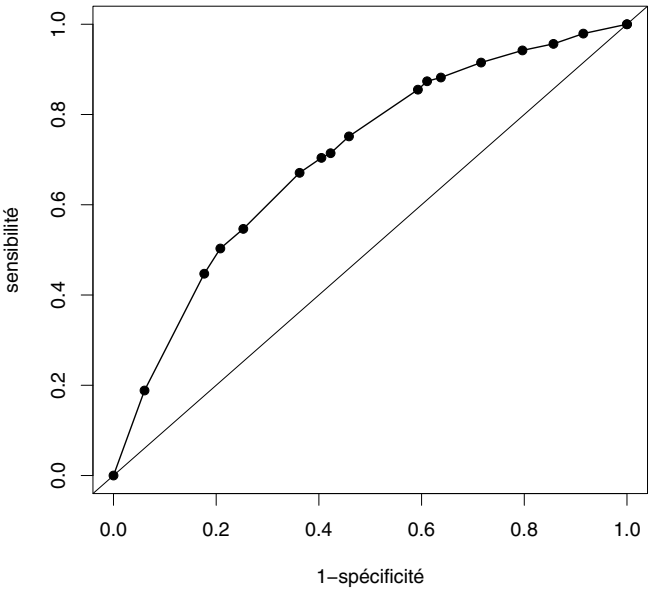

Figure 16.4 – Exemple de courbe ROC.

→ un diagnostic parfait serait atteint si la courbe atteignait le coin gauche supérieur du graphique (correspondant à 100 % de sensibilité et à 100 % de spécificité)

- on calcule l'**aire sous cette courbe** qui est une mesure globale de la qualité des diagnostics que l'on peut établir à l'aide du modèle et donc de la qualité globale de la prédiction pour ce modèle.

Cette approche sera valide si on a des estimations valides de la sensibilité et de la spécificité. Notons que la somme de la sensibilité et de la spécificité obtenues à partir d'un modèle de régression logistique est toujours supérieure ou égale à 1, quelle que soit la limite critique C utilisée, de sorte qu'une courbe ROC ne se retrouve jamais en dessous de la diagonale et que l'aire sous la courbe ROC est toujours égale au moins à 0.5[15].

Exemple 16.12 *On reprend les données de l'étude cas-témoins où l'on veut prédire la présence d'un cancer de la cavité buccale en fonction de la consommation de tabac et d'alcool. On a* 16 *combinaisons de valeurs possibles des*

[15]Notons également que l'aire sous cette courbe correspond à la probabilité que la probabilité estimée $\widehat{\pi}_i$ d'un individu malade soit supérieure à la probabilité estimée $\widehat{\pi}_i$ d'un individu non malade (si on choisissait au hasard un malade parmi les malades et un non-malade parmi les non-malades de notre échantillon). L'aire sous la courbe ROC est donc égale à la quantité $\widehat{\psi}$ utilisée dans un test de Mann-Whitney, où l'on comparerait le groupe des malades et le groupe des non-malades par rapport aux probabilités estimées $\widehat{\pi}_i$.

prédicteurs, pour lesquels on a les valeurs de $\widehat{\pi}_i$ suivantes (on utilise ici le modèle avec prédicteurs quantitatifs) :

	tabac			
alcool	0	1–19	20–39	40+
0	0.164	0.236	0.326	0.432
1–9	0.243	0.335	0.441	0.554
10–44	0.343	0.451	0.563	0.670
45+	0.460	0.573	0.678	0.768

On estime ainsi les couples possibles $(1 - \textit{spécificité}, \textit{sensibilité})$ *suivants (en utilisant ces valeurs de* $\widehat{\pi}_i$ *comme valeur critique* C*) :*

C	$1 -$ *spécificité*	*sensibilité*	C	$1 -$ *spécificité*	*sensibilité*
0.000	1.000	1.000	0.451	0.459	0.752
0.164	1.000	1.000	0.460	0.423	0.714
0.236	0.915	0.979	0.554	0.405	0.704
0.243	0.857	0.957	0.563	0.362	0.671
0.326	0.796	0.942	0.573	0.253	0.547
0.335	0.716	0.915	0.670	0.208	0.503
0.343	0.638	0.882	0.678	0.177	0.447
0.432	0.611	0.874	0.768	0.060	0.188
0.441	0.593	0.855	1.000	0.000	0.000

La courbe ROC de la figure 16.4 nous montre ces quantités. On peut ensuite calculer l'aire sous cette courbe, qui est égale à 0.706*. Cette aire sous la courbe représente une mesure de la qualité globale de la prédiction pour ce modèle de régression logistique*[16].

16.9 Vérification du modèle

Il est plus difficile de vérifier l'unique hypothèse d'un modèle de régression logistique que de vérifier les trois hypothèses d'un modèle de régression linéaire (on met de côté ici les hypothèses d'indépendance). En régression linéaire, on utilise les résidus $y_i - \widehat{y}_i^*$ $(i = 1, \cdots, n)$ qui ont tous sous les hypothèses du modèle la même distribution (normale, d'espérance nulle et de même variance). En régression logistique, on pourrait aussi calculer des résidus définis par $y_i - \widehat{\pi}_i$ $(i = 1, \cdots, n)$, mais ces résidus auraient sous les hypothèses du modèle des distributions différentes, de sorte qu'une analyse de ces résidus serait

[16] Comme il s'agit d'une étude cas-témoins où les malades sont sur-représentés, les $\widehat{\pi}_i$ calculés ici ne sont pas des estimations valides des probabilités d'avoir la maladie dans les différents groupes, l'estimation de la constante du modèle n'étant pas valide. Néanmoins, les couples $(1 - \text{spécificité}, \text{sensibilité})$ obtenus seraient les mêmes avec une constante différente dans le modèle, de sorte que l'estimation de la courbe ROC est valide.

problématique. Bien que différentes façons de standardiser ces résidus aient été proposées dans la littérature, il n'est pas évident de reconnaître des résidus qui sont en accord avec les hypothèses du modèle, de résidus qui ne le sont pas.

Dans le but de vérifier l'hypothèse de linéarité en régression logistique, une possibilité est de regrouper les individus selon leurs probabilités estimées $\widehat{\pi}_i$ (les individus avec des $\widehat{\pi}_i$ semblables se retrouvant dans un même groupe). On pourra alors comparer les proportions empiriques (la moyenne des y_i) et les proportions prédites (la moyenne des $\widehat{\pi}_i$) de $Y = 1$ dans ces différents groupes. Une différence trop grande entre ces proportions dans l'un ou l'autre de ces groupes indiquera un défaut du modèle[17].

Exemple 16.13 *On reprend les données de l'étude cas-témoins où l'on prédit la présence d'un cancer de la cavité buccale en fonction de la consommation de tabac et d'alcool. On a ici* 16 *groupes naturels (correspondant aux* 16 *combinaisons de valeurs possibles des prédicteurs). Dans chacun de ces groupes, on compare la proportion empirique de malades (la moyenne des y_i) avec la proportion de malades prédite par le modèle (la moyenne des $\widehat{\pi}_i$). On a les résultats suivants (les proportions prédites sont données entre parenthèses et sont calculées à partir du modèle avec prédicteurs quantitatifs) :*

	tabac			
alcool	0	1–19	20–39	40+
0	0.21 (0.16)	0.30 (0.24)	0.27 (0.33)	0.53 (0.43)
1–9	0.21 (0.24)	0.31 (0.33)	0.45 (0.44)	0.46 (0.55)
10–44	0.25 (0.34)	0.53 (0.45)	0.55 (0.56)	0.66 (0.67)
45+	0.38 (0.46)	0.51 (0.57)	0.71 (0.68)	0.77 (0.77)

La plus grande différence entre proportion empirique et proportion prédite de malades est ainsi de 10 %*, que l'on trouve dans le groupe des non-buveurs et grands fumeurs où l'on observe* 53 % *de malades, alors que le modèle n'en prédit que* 43 %*. Cette différence de* 10 % *n'est cependant pas énorme et ne se retrouve pas dans les groupes adjacents, de sorte que notre modèle apparaît être une approximation satisfaisante de la réalité*[18]*.*

On terminera cette section en mentionnant un résultat remarquable. Un modèle de régression logistique pour une variable réponse binaire Y avec comme

[17] Des tests statistiques ont été proposés pour nous aider à décider si une telle différence est acceptable ou non (par exemple le *test de Hosmer-Lemeshow*). On rappelle toutefois qu'un modèle de régression ne sera en général qu'une approximation de la réalité, et donc faux au sens strict du terme, de sorte que même un bon modèle pourra être rejeté statistiquement si l'échantillon dans lequel il est calculé est suffisamment grand.

[18] Rappelons que dans le cadre d'une étude cas-témoins, les $\widehat{\pi}_i$ ne sont certes pas des estimations valides des probabilités d'avoir la maladie dans les différents groupes. On a donc ici vérifié le modèle dans le cadre d'un échantillon non représentatif de la population d'intérêt, où les malades sont sur-représentés. Cela est cependant suffisant pour obtenir des estimations valides des odds-ratios dans la population d'intérêt (comme expliqué en début de chapitre).

prédicteur une variable continue X est impliqué par un modèle de régression linéaire avec Y comme prédicteur binaire et X comme variable réponse. Autrement dit, le modèle idéal du chapitre 5 qui était un cas particulier de modèle de régression linéaire, est également un cas particulier de modèle de régression logistique. Dans cette optique, un modèle de régression logistique apparaît comme un choix naturel lorsqu'il s'agit de modéliser une variable réponse binaire[19].

16.10 Méthode du maximum de vraisemblance

On a mentionné que les paramètres d'un modèle de régression logistique sont estimés à partir des données de l'échantillon en utilisant la méthode du maximum de vraisemblance (en anglais : `maximum likelihood`). Dans cette dernière section, nous introduisons brièvement cette méthode d'estimation qui s'applique à de nombreux modèles statistiques.

Afin d'illustrer comment fonctionne cette méthode, nous considérerons le problème classique qui consiste à estimer la probabilité π qu'une pièce de monnaie tombe sur pile (autrement dit, la proportion de piles dans la population constituée des différents lancers de cette pièce). On considère ainsi une variable Y binaire avec $Y = 1$ pour « pile » et $Y = 0$ pour « face ». Supposons que l'on lance la pièce $n = 4$ fois avec les résultats suivants : 1 1 0 1. La probabilité que le premier lancer donne $Y = 1$ est donc π. La probabilité que le deuxième lancer donne $Y = 1$ est également π. La probabilité que le troisième lancer donne $Y = 0$ est $1 - \pi$. La probabilité que le quatrième lancer donne $Y = 1$ est à nouveau π. Comme ces observations sont indépendantes, la probabilité d'obtenir cet échantillon (si on répétait l'expérience) est donnée par la multiplication des quatre probabilités ci-dessus, c'est-à-dire par :

$$L = \pi \cdot \pi \cdot (1 - \pi) \cdot \pi = \pi^3 \cdot (1 - \pi).$$

[19] Si X dénote la variable continue et Y la variable binaire, on a (théorème de Bayes) :

$$\Pr\{Y = 1|X = x\} = \frac{\Pr\{X = x|Y = 1\} \Pr\{Y = 1\}}{\Pr\{X = x|Y = 1\} \Pr\{Y = 1\} + \Pr\{X = x|Y = 0\} \Pr\{Y = 0\}}.$$

On définit par ailleurs $\Pr\{X = x|Y = 1\}/\Pr\{X = x|Y = 0\}$ comme le quotient des densités $f_1(x)$ et $f_0(x)$ de la variable X dans les deux groupes définis par $Y = 1$ et $Y = 0$. Sous l'hypothèse d'un modèle de régression linéaire avec X comme réponse et Y comme prédicteur ces deux densités seront normales, avec une variance commune σ^2 et des moyennes μ_1 et μ_0, c'est-à-dire $f_j(x) = (1/\sqrt{2\pi\sigma^2}) \exp(-(x - \mu_j)^2/(2\sigma^2))$ (pour $j = 0, 1$). On aura donc :

$$\frac{\Pr\{X = x|Y = 1\}}{\Pr\{X = x|Y = 0\}} = \frac{f_1(x)}{f_0(x)} = \exp\left(\frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \frac{\mu_1 - \mu_0}{\sigma^2} \cdot x\right)$$

et ainsi :

$$\Pr\{Y = 1|X = x\} = \frac{\frac{\Pr\{X=x|Y=1\}}{\Pr\{X=x|Y=0\}} \cdot \frac{\Pr\{Y=1\}}{\Pr\{Y=0\}}}{1 + \frac{\Pr\{X=x|Y=1\}}{\Pr\{X=x|Y=0\}} \cdot \frac{\Pr\{Y=1\}}{\Pr\{Y=0\}}} = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

avec $\beta_0 = (\mu_0^2 - \mu_1^2)/(2\sigma^2) + \log(\Pr\{Y = 1\}/\Pr\{Y = 0\})$ et $\beta_1 = (\mu_1 - \mu_0)/\sigma^2$, ce qui correspond à un modèle de régression logistique avec Y comme réponse et X comme prédicteur.

On appelle la *vraisemblance de l'échantillon* (en anglais : `likelihhod`) cette probabilité L. On a ainsi (par exemple) :

- $L = 0$ si $\pi = 0$
- $L = 0.0009$ si $\pi = 0.1$
- $L = 0.0625$ si $\pi = 0.5$
- $L = 0.1024$ si $\pi = 0.8$
- $L = 0$ si $\pi = 1$.

Cette vraisemblance est donc une fonction de la véritable probabilité (inconnue) π. Elle est plus élevée pour certaines valeurs de π que pour d'autres. Dans notre exemple, la probabilité d'observer un échantillon semblable au nôtre serait ainsi plus grande si on avait $\pi = 0.8$ que si on avait $\pi = 0.5$. L'estimateur du maximum de vraisemblance consiste à estimer π par la valeur qui maximise cette vraisemblance. Ce principe se généralise à des modèles plus compliqués.

Avec la méthode du maximum de vraisemblance, on estime les paramètres d'un modèle par les valeurs maximisant la vraisemblance de l'échantillon.

Dans notre exemple, on recherche ainsi la valeur $\widehat{\pi}$ qui maximise la fonction donnée par :

$$L(\pi) = \pi^3 \cdot (1 - \pi).$$

Notons que le maximum de cette fonction est également le maximum du logarithme de cette fonction (que l'on appellera en anglais le `log-likelihood`) :

$$\log L(\pi) = 3 \log(\pi) + \log(1 - \pi).$$

La transformation logarithmique nous permet de transformer un produit en une somme, qui est plus facile à dériver mathématiquement. Dans notre exemple, la dérivée de cette fonction est égale à $3/\pi - 1/(1 - \pi)$ et la solution de notre problème est donc la solution de l'équation suivante :

$$3/\widehat{\pi} - 1/(1 - \widehat{\pi}) = 0$$

Notre estimation de π est donc $\widehat{\pi} = 3/4 = 0.75$, valeur pour laquelle on obtient la vraisemblance maximale de $L = 0.1055$.

Dans cet exemple, l'estimateur du maximum de vraisemblance de la probabilité que la pièce tombe sur pile est simplement la proportion de piles observée dans notre échantillon. Ceci justifie en quelque sorte l'utilisation de la proportion empirique comme estimateur d'une proportion. De même, l'estimateur du maximum de vraisemblance de la moyenne d'une population sera la moyenne

de l'échantillon, pour autant que l'on postule une distribution normale. Plus généralement, les estimateurs du maximum de vraisemblance des paramètres d'un hyperplan de régression seront les estimateurs calculés en utilisant la méthode des moindres carrés, pour autant que l'on postule un modèle de régression linéaire (avec notamment la normalité des résidus)[20].

En régression logistique, la méthode du maximum de vraisemblance consiste à rechercher les valeurs de $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$ qui maximisent la vraisemblance :

$$L = \prod_i \widehat{\pi}_i^{y_i} (1 - \widehat{\pi}_i)^{1-y_i}$$

et donc :

$$\log L = \sum_i \left(y_i \log(\widehat{\pi}_i) + (1 - y_i) \log(1 - \widehat{\pi}_i) \right).$$

Rappelons que l'on a :

$$\widehat{\pi}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_m x_{im})}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_m x_{im})}.$$

Ainsi, la vraisemblance L est (*via* les $\widehat{\pi}_i$) une fonction des $m+1$ coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \cdots, \widehat{\beta}_m$, dont le maximum est donné par les solutions des dérivées partielles par rapport à ces coefficients qui doivent ainsi satisfaire les $m+1$ conditions suivantes :

$$\sum_i (y_i - \widehat{\pi}_i) = 0 \quad \text{et} \quad \sum_i x_{ij}(y_i - \widehat{\pi}_i) = 0 \quad (\text{pour } j = 1, \cdots, m).$$

La première condition implique que la somme des résidus $y_i - \widehat{\pi}_i$ est nulle (comme en régression linéaire). Dans un modèle de régression logistique sans prédicteur, on estime ainsi l'unique paramètre β_0 par la proportion empirique de $Y = 1$. Dans un modèle de régression logistique plus complexe, les solutions de ces équations ne seront en général pas explicites.

Les estimateurs du maximum de vraisemblance des paramètres d'un modèle statistique seront approximativement sans biais (du moins lorsque la taille de l'échantillon est suffisamment grande) et approximativement normalement distribués. On pourra donc utiliser la méthode de Wald pour faire de l'inférence sur les paramètres du modèle. On a cependant mentionné qu'il existe également des tests dits du rapport de vraisemblance (en anglais : `likelihood ratio tests`), dont la validité est supérieure à celle des tests de Wald (bien que l'on aura des résultats similaires avec de grands échantillons)[21]. Afin de

[20] On notera toutefois que l'estimateur du maximum de vraisemblance de la variance résiduelle σ_ε^2 dans un modèle de régression linéaire est donné par la variance empirique des résidus, que l'on avait notée $\widehat{\sigma}_\varepsilon^2$, et non par l'estimateur sans biais de cette variance, noté $\tilde{\sigma}_\varepsilon^2$.

[21] Les tests du rapport de vraisemblance s'appliquent également à un modèle de régression linéaire, mais ils ne sont pas très utiles dans ce cadre où l'on dispose déjà de tests exacts (les tests de Student et les tests F).

tester la nullité simultanée de r paramètres dans un modèle de régression logistique avec un test du rapport de vraisemblance, on calcule la statistique de test suivante :

$$t_{stat} = 2(\log L - \log L_0).$$

où L dénote la valeur de la vraisemblance qui est maximisée par les estimateurs du maximum de vraisemblance dans notre modèle et L_0 la valeur de la vraisemblance qui est maximisée par les estimateurs du maximum de vraisemblance dans un modèle qui ne contient pas les r prédicteurs associés aux paramètres dont on aimerait tester la nullité. Notons que la vraisemblance ne peut qu'augmenter lorsqu'on introduit de nouveaux prédicteurs dans le modèle, de sorte que l'on aura nécessairement $L \geq L_0$. Si cette augmentation est trop importante, on rejette l'hypothèse nulle de la nullité des r paramètres en question. Sous cette hypothèse nulle, la distribution de cette statistique de test est approximativement une distribution du khi-deux avec r *dl*. On rejette ainsi l'hypothèse nulle au seuil α si $t_{stat} \geq \chi^2_{1-\alpha,r}$.

On mentionnera finalement que la quantité $-2\log L$ est appelée la *déviance* du modèle[22]. Maximiser la vraisemblance revient donc à minimiser la déviance. Par ailleurs, le critère AIC que l'on avait brièvement mentionné au chapitre 14 en relation avec la problématique du choix d'un modèle de régression pour optimiser la prédiction, consiste à « pénaliser » la déviance d'un modèle en lui additionant deux fois le nombre de ses paramètres. On aura ainsi :

$$AIC = -2\log L + 2(m+1).$$

Dans le but d'optimiser la prédiction en régression logistique, une stratégie consiste à choisir un modèle minimisant AIC. L'introduction d'un prédicteur dans un modèle diminuera la valeur de AIC si la statistique de test d'un test du rapport de vraisemblance sur la nullité de la pente associée à ce prédicteur est plus grande que 2 (et donc la valeur p plus petite que 0.157). On avait un résultat semblable en utilisant la stratégie qui consistait à minimiser la longueur des intervalles de prédiction pour le choix d'un modèle en régression linéaire.

[22]La statistique de test d'un test du rapport de vraisemblance est donc la différence des déviances des deux modèles que l'on compare. Dans R, on obtient la déviance d'un modèle de régression logistique multiple avec une variable réponse binaire `y` et des prédicteurs `x1, x2, x3, x4` et `x5` en utilisant la commande `glm(y~x1+x2+x3+x4+x5,family=binomial)$deviance`.

Annexe A

Tableaux

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862
1.1	0.864	0.867	0.869	0.871	0.873	0.875	0.877	0.879	0.881	0.883
1.2	0.885	0.887	0.889	0.891	0.893	0.894	0.896	0.898	0.900	0.901
1.3	0.903	0.905	0.907	0.908	0.910	0.911	0.913	0.915	0.916	0.918
1.4	0.919	0.921	0.922	0.924	0.925	0.926	0.928	0.929	0.931	0.932
1.5	0.933	0.934	0.936	0.937	0.938	0.939	0.941	0.942	0.943	0.944
1.6	0.945	0.946	0.947	0.948	0.949	0.951	0.952	0.953	0.954	0.954
1.7	0.955	0.956	0.957	0.958	0.959	0.960	0.961	0.962	0.962	0.963
1.8	0.964	0.965	0.966	0.966	0.967	0.968	0.969	0.969	0.970	0.971
1.9	0.971	0.972	0.973	0.973	0.974	0.974	0.975	0.976	0.976	0.977
2.0	0.977	0.978	0.978	0.979	0.979	0.980	0.980	0.981	0.981	0.982
2.1	0.982	0.983	0.983	0.983	0.984	0.984	0.985	0.985	0.985	0.986
2.2	0.986	0.986	0.987	0.987	0.987	0.988	0.988	0.988	0.989	0.989
2.3	0.989	0.990	0.990	0.990	0.990	0.991	0.991	0.991	0.991	0.992
2.4	0.992	0.992	0.992	0.992	0.993	0.993	0.993	0.993	0.993	0.994
2.5	0.994	0.994	0.994	0.994	0.994	0.995	0.995	0.995	0.995	0.995
2.6	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
2.7	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997
2.8	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.998
2.9	0.998	0.998	0.998	0.998	0.998	0.998	0.998	0.999	0.999	0.999

Tableau A.1 – Fonction de répartition $\Phi_{0,1}(x)$ d'une distribution normale standardisée pour différentes valeurs de $x \geq 0$. Pour des valeurs de $x < 0$, on utilise $\Phi_{0,1}(-x) = 1 - \Phi_{0,1}(x)$. Ce tableau nous permet de déterminer le quantile α d'une distribution normale standardisée (z_α) grâce à la relation $\Phi_{0,1}(z_\alpha) = \alpha$ (par exemple $\Phi_{0,1}(z_{0.975}) = 0.975$; comme on a dans ce tableau $\Phi_{0,1}(1.96) = 0.975$, on trouve ainsi $z_{0.975} = 1.96$). Ce tableau nous permet aussi de déterminer à quel quantile correspond la valeur $\mu + x \cdot \sigma$ lorsque la distribution est normale de moyenne μ et écart type σ (par exemple, $\mu + 1.96 \cdot \sigma$ correspond au quantile 97.5%). Les éléments de ce tableau ont été produits avec la commande `pnorm(x)` de R (par exemple `pnorm(1.96)` nous donne 0.975).

							α						
dl	55%	60%	65%	70%	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.9%
1	0.16	0.33	0.51	0.73	1.00	1.38	1.96	3.08	6.31	12.71	31.82	63.66	318.31
2	0.14	0.29	0.45	0.62	0.82	1.06	1.39	1.89	2.92	4.30	6.96	9.93	22.33
3	0.14	0.28	0.42	0.58	0.77	0.98	1.25	1.64	2.35	3.18	4.54	5.84	10.21
4	0.13	0.27	0.41	0.57	0.74	0.94	1.19	1.53	2.13	2.78	3.75	4.60	7.17
5	0.13	0.27	0.41	0.56	0.73	0.92	1.16	1.48	2.02	2.57	3.37	4.03	5.89
6	0.13	0.27	0.40	0.55	0.72	0.91	1.13	1.44	1.94	2.45	3.14	3.71	5.21
7	0.13	0.26	0.40	0.55	0.71	0.90	1.12	1.42	1.90	2.37	3.00	3.50	4.79
8	0.13	0.26	0.40	0.55	0.71	0.89	1.11	1.40	1.86	2.31	2.90	3.35	4.50
9	0.13	0.26	0.40	0.54	0.70	0.88	1.10	1.38	1.83	2.26	2.82	3.25	4.30
10	0.13	0.26	0.40	0.54	0.70	0.88	1.09	1.37	1.81	2.23	2.76	3.17	4.14
11	0.13	0.26	0.40	0.54	0.70	0.88	1.09	1.36	1.80	2.20	2.72	3.11	4.03
12	0.13	0.26	0.40	0.54	0.69	0.87	1.08	1.36	1.78	2.18	2.68	3.06	3.93
13	0.13	0.26	0.39	0.54	0.69	0.87	1.08	1.35	1.77	2.16	2.65	3.01	3.85
14	0.13	0.26	0.39	0.54	0.69	0.87	1.08	1.34	1.76	2.15	2.62	2.98	3.79
15	0.13	0.26	0.39	0.54	0.69	0.87	1.07	1.34	1.75	2.13	2.60	2.95	3.73
16	0.13	0.26	0.39	0.54	0.69	0.86	1.07	1.34	1.75	2.12	2.58	2.92	3.69
17	0.13	0.26	0.39	0.53	0.69	0.86	1.07	1.33	1.74	2.11	2.57	2.90	3.65
18	0.13	0.26	0.39	0.53	0.69	0.86	1.07	1.33	1.73	2.10	2.55	2.88	3.61
19	0.13	0.26	0.39	0.53	0.69	0.86	1.07	1.33	1.73	2.09	2.54	2.86	3.58
20	0.13	0.26	0.39	0.53	0.69	0.86	1.06	1.32	1.73	2.09	2.53	2.84	3.55
21	0.13	0.26	0.39	0.53	0.69	0.86	1.06	1.32	1.72	2.08	2.52	2.83	3.53
22	0.13	0.26	0.39	0.53	0.69	0.86	1.06	1.32	1.72	2.07	2.51	2.82	3.50
23	0.13	0.26	0.39	0.53	0.69	0.86	1.06	1.32	1.71	2.07	2.50	2.81	3.48
24	0.13	0.26	0.39	0.53	0.69	0.86	1.06	1.32	1.71	2.06	2.49	2.80	3.47
25	0.13	0.26	0.39	0.53	0.68	0.86	1.06	1.32	1.71	2.06	2.48	2.79	3.45
26	0.13	0.26	0.39	0.53	0.68	0.86	1.06	1.31	1.71	2.06	2.48	2.78	3.44
27	0.13	0.26	0.39	0.53	0.68	0.85	1.06	1.31	1.70	2.05	2.47	2.77	3.42
28	0.13	0.26	0.39	0.53	0.68	0.85	1.06	1.31	1.70	2.05	2.47	2.76	3.41
29	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.31	1.70	2.04	2.46	2.76	3.40
30	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.31	1.70	2.04	2.46	2.75	3.38
35	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.31	1.69	2.03	2.44	2.72	3.34
40	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.30	1.68	2.02	2.42	2.70	3.31
45	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.30	1.68	2.01	2.41	2.69	3.28
50	0.13	0.26	0.39	0.53	0.68	0.85	1.05	1.30	1.68	2.01	2.40	2.68	3.26
60	0.13	0.25	0.39	0.53	0.68	0.85	1.04	1.30	1.67	2.00	2.39	2.66	3.23
70	0.13	0.25	0.39	0.53	0.68	0.85	1.04	1.29	1.67	1.99	2.38	2.65	3.21
80	0.13	0.25	0.39	0.53	0.68	0.85	1.04	1.29	1.66	1.99	2.37	2.64	3.19
90	0.13	0.25	0.39	0.53	0.68	0.85	1.04	1.29	1.66	1.99	2.37	2.63	3.18
100	0.13	0.25	0.39	0.53	0.68	0.84	1.04	1.29	1.66	1.98	2.36	2.63	3.17
200	0.13	0.25	0.39	0.53	0.68	0.84	1.04	1.29	1.65	1.97	2.34	2.60	3.13
300	0.13	0.25	0.39	0.53	0.68	0.84	1.04	1.28	1.65	1.97	2.34	2.59	3.12
400	0.13	0.25	0.39	0.53	0.68	0.84	1.04	1.28	1.65	1.97	2.34	2.59	3.11
500	0.13	0.25	0.39	0.53	0.68	0.84	1.04	1.28	1.65	1.96	2.33	2.59	3.11
∞	0.13	0.25	0.39	0.52	0.67	0.84	1.04	1.28	1.65	1.96	2.33	2.58	3.09

Tableau A.2 – Quantile α d'une distribution de Student avec dl degrés de liberté ($t_{\alpha,dl}$) pour différentes valeurs de $\alpha > 50\%$ et dl. Pour des valeurs de $\alpha < 50\%$, on utilise $t_{1-\alpha,dl} = -t_{\alpha,dl}$ (avec en outre $t_{0.5,dl} = 0$). Les éléments de ce tableau ont été produits avec la commande `qt(alpha,dl)` de R (par exemple `qt(0.975,10)` nous donne 2.23).

	α												
dl	0.5%	1%	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%	99.5%
1	0.00	0.02	0.00	0.00	0.02	0.10	0.46	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.21	0.05	0.10	0.21	0.57	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.58	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.35	12.84
4	0.21	1.06	0.48	0.71	1.06	1.92	3.36	5.38	7.78	9.49	11.14	13.28	14.86
5	0.41	1.61	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	2.20	1.24	1.64	2.20	3.46	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	2.83	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	3.49	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.95
9	1.73	4.17	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	4.87	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	5.58	3.82	4.58	5.58	7.58	10.34	13.70	17.27	19.68	21.92	24.73	26.76
12	3.07	6.30	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.56	7.04	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.08	7.79	5.63	6.57	7.79	10.16	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	8.55	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	9.31	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.84	32.00	34.27
17	5.70	10.09	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	10.87	8.23	9.39	10.87	13.68	17.34	21.61	25.99	28.87	31.53	34.80	37.16
19	6.84	11.65	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	12.44	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	13.24	10.28	11.59	13.24	16.34	20.34	24.93	29.61	32.67	35.48	38.93	41.40
22	8.64	14.04	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	14.85	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	15.66	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.41	39.36	42.98	45.56
25	10.52	16.47	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	17.29	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.88	41.92	45.64	48.29
27	11.81	18.11	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.20	46.96	49.65
28	12.46	18.94	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	19.77	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	20.60	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
35	17.19	24.80	20.57	22.46	24.80	29.05	34.34	40.22	46.06	49.80	53.20	57.34	60.27
40	20.71	29.05	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	66.77
45	24.31	33.35	28.37	30.61	33.35	38.29	44.34	50.98	57.51	61.66	65.41	69.96	73.17
50	27.99	37.69	32.36	34.76	37.69	42.94	49.34	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	46.46	40.48	43.19	46.46	52.29	59.34	66.98	74.40	79.08	83.30	88.38	91.95
70	43.27	55.33	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	64.28	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	73.29	65.65	69.13	73.29	80.62	89.33	98.65	107.56	113.14	118.14	124.12	128.30
100	67.33	82.36	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17

Tableau A.3 – Quantile α d'une distribution du khi-deux avec dl degrés de liberté ($\chi^2_{\alpha,dl}$) pour différentes valeurs de α et dl. Les éléments de ce tableau ont été produits avec la commande `qchisq(alpha,dl)` de R (par exemple `qchisq(0.95,1)` nous donne 3.84).

	dln											
dld	1	2	3	4	5	6	7	8	9	10	11	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.91	1.88
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.90	1.86
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.89	1.85
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.82	1.78
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.81	1.78
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.77
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75

Tableau A.4 – Quantile 95% d'une distribution F avec degrés de liberté *dln* et *dld* ($F_{0.95,dln,dld}$) pour différentes valeurs de *dln* et *dld*. Les éléments de ce tableau ont été produits avec la commande `qf(0.95,dln,dld)` de R (par exemple `qf(0.95,1,30)` nous donne 4.17).

Bibliographie

[1] A. Agresti and B. Caffo. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician*, 54 :280–288, 2000.

[2] A. Agresti and B.A. Coull. Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician*, 52 :119–126, 1998.

[3] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.V. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, 1973.

[4] D.F. Andrews and A.M. Herzberg. *Data - A Collection of Problems from Many Fields for the Student and Research Worker*. Springer, 1985.

[5] G.E.P. Box. Robustness in the strategy of scientific model building. In R.L. Launer and G.N. Wilkinson, editors, *Robustness in Statistics : Proceedings of Workshop*. Academic Press, 1979.

[6] K. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*. Springer, second edition, 2004.

[7] C.R. Charig, D.R. Webb, S.R. Payne, and J.E.A. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave. *British Medical Journal*, 292 :879–882, 1986.

[8] G. Cocco, S. Pandolfi, and V. Rousson. Sufficient weight reduction decreases cardiovascular complications in diabetic patients with the metabolic syndrome. *HeartDrug*, 5 :68–74, 2005.

[9] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, second edition, 1988.

[10] P.A. Cornillon and E. Matzner-Lober. *Régression - Théorie et applications*. Springer, 2007.

[11] M.J. Crawley. *The R Book*. Wiley, 2007.

[12] Y. Dodge and V. Rousson. *Analyse de régression appliquée*. Dunod, second edition, 2004.

[13] R. Doll and B. Hill. A study of the aetiology of carcinoma of the lung. *British Medical Journal*, 4797 :1271–1286, 1952.

[14] O.J. Dunn and V.A. Clark. *Basic Statistics - A Primer for the Biomedical Sciences*. Wiley, 2009.

[15] M.W. Fagerland, S. Lydersen, and P. Laake. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, 2011.

[16] M.P. Fay and M.A. Proschnan. Wilcoxon-Mann-Whitney or t-test ? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4 :1–39, 2010.

[17] M. Firmann, V. Mayor, P. Marques-Vidal, M. Bochud, Pécoud A., D. Hayoz, F. Paccaud, M. Preisig, K.S. Song, X. Yuan, T.M. Danoff, H.A. Stirnadel, D. Waterworth, V. Mooser, G. Waeber, and P. Vollenweider. The CoLaus study : a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders*, 8(6), 2008.

[18] R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[19] R.A. Fisher. The precision of discriminant functions. *Annals of Eugenics*, 10 :422–429, 1940.

[20] D.A. Häring, M.J. Huber, D. Suter, P.J. Edwards, and A. Lüscher. Plant enemy-derived elicitors increase the foliar tannin concentration of onobrychis viciifolia without a trade-off to growth. *Annals of Botany*, 102 :979–987, 2008.

[21] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, second edition, 2000.

[22] R.W. Johnson. Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education*, 4(1), 1996.

[23] B.R. Kirkwood and J.A.C. Sterne. *Essential Medical Statistics*. Blackwell Science, second edition, 2003.

[24] R. Largo, V. Rousson, J. Caflisch, and O.G. Jenni. *Zurich Neuromotor Assessment*. AWE Verlag, 2007.

[25] M. Lejeune. *Statistique - La théorie et ses applications*. Springer, 2004.

[26] H.B. Mann and D.R. Whitney. On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18 :50–60, 1947.

[27] M. Mittlböck and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15 :1987–1997, 1996.

[28] K. Rothman and A. Keller. The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx. *Journal of Chronic Diseases*, 25 :711–716, 1972.

[29] V. Rousson, T. Gasser, Caflisch J., and O.G. Jenni. Neuromotor performance of normally developing left-handed children and adolescents. *Human Movement Science*, 28 :809–817, 2009.

[30] V. Rousson and N.F. Goşoniu. An R-square coefficient based on final prediction error. *Statistical Methodology*, 4 :331–340, 2007.

[31] A.R. Sampson. A tale of two regressions. *Journal of the American Statistical Association*, 69 :682–689, 1974.

[32] D.J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15 :657–680, 1987.

[33] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society (Series B)*, 13 :238–241, 1951.

[34] Student. The probable error of a mean. *Biometrika*, 6 :1–25, 1908.

[35] G. van Belle, L.D. Fisher, P.J. Heagerty, and T. Lumley. *Biostatistics - A Methodology for the Health Sciences*. Wiley, second edition, 2004.

[36] E. Vittinghoff, D.V. Glidden, S.C. Shiboski, and C.E. McCulloch. *Regression Methods in Biostatistics*. Springer, 2005.

[37] B.L. Welch. The significance of the difference between two means when the population are unequal. *Biometrika*, 29 :350–362, 1938.

[38] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1 :80–83, 1945.

[39] E.B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22 :209–212, 1927.

[40] S.L. Zabell. On Student's 1908 article 'the probable error of a mean'. *Journal of the American Statistical Association*, 103 :1–7, 2008.