Université Paris-Saclay – Masters BEE, BIP, BS

Biostatistiques

P. Brabant, C. Dillmann, J. Legrand, D. Manicacci, E. Marchadier, S. Ollier, D. Sicard, D. de Vienne

2020-2021

Table des matières

Avertissement

Ce polycopié est une synthèse des notes de cours des enseignants des UE de biostatistiques de M1 (masters Biologie-Santé, BEE, BIP) et M2 (optométrie, ergonomie) de l'UFR sciences. Il est destiné à accompagner les cours, TD et TP de ces UE. Merci d'adresser vos questions ou commentaires permettant d'améliorer ce document à `christine.dillmann@universite-paris-saclay.fr`.

Chapitre 1

Introduction : statistiques, variables aléatoires, échantillons

1.1 Statistiques

Statistique : *ensemble des méthodes permettant d'obtenir, de décrire et d'analyser des observations (ou données). Ces observations consistent généralement en la mesure d'une ou plusieurs caractéristiques sur un ensemble d'individus.*

La *théorie des probabilités* est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude ; la *statistique* est l'activité qui consiste à recueillir, traiter et interpréter un ensemble de données. Les probabilités et la statistique forment les sciences de l'aléatoire.

Individu statistique : *unité de base sur laquelle la mesure est réalisée.*

Les individus peuvent être des personnes, des lunettes (exemple : on veut tester la qualité d'une monture), des bactéries. L'unité statistique peut aussi être un groupe d'individus (exemple : une classe de 25 élèves dont on étudie le comportement, ou toutes les plantes d'*Arabidopsis thaliana* contenues dans une terrine, etc.)

Population : *ensemble des individus sur lesquels on souhaite obtenir des informations*

Cela peut être par exemple, la population française des hommes adultes, ou l'ensemble des lunettes qui sortent d'une chaîne de montage sur une période donnée, les patients d'un hôpital.
En général, quand la population est de très grande taille, il est impossible de faire des mesures sur l'ensemble des individus de la population. Par contre, on peut faire des mesures sur un petit nombre d'individus tirés au hasard dans la population, qui seront considérés comme représentatifs de la population.
Pour les populations de taille restreinte, il est possible de faire des mesures sur l'ensemble des individus, et la population sera décrite sans erreur due à l'échantillonnage.

Echantillon : *sous-ensemble d'individus d'une population tirés au hasard parmi tous les individus de la population. S'ils sont bien tirés au hasard, sans biais, l'échantillon sera dit "représentatif" de la population.*

On distingue deux grand types d'approches statistiques, les *statistiques descriptives* et les *statistiques inférentielles.*

Statistiques descriptives. L'objectif est de dégager les caractéristiques d'un échantillon à l'aide de données chiffrées et de graphiques. On peut vouloir décrire la distribution des variables, variable par variable. On peut aussi souhaiter dégager les caractéristiques principales de l'échantillon, en résumant l'information par certains paramètres (exemple la moyenne, la variance) et en déterminant quelles sont les relations principales entre des variables. Par exemple, on peut donner la moyenne et la variance de la taille des individus d'un échantillon. On peut également quantifier dans quelle mesure la taille et le poids des individus sont liés.

Statistiques inférentielles. On souhaite « inférer » les propriétés d'une population (par exemple la moyenne de la taille des individus de la population) à partir des données d'un échantillon. On veut estimer la valeur de cette moyenne et aussi déterminer « une marge d'erreur ». On veut pouvoir tester des hypothèses, par exemple est-ce que la taille moyenne augmente d'une génération sur l'autre dans une population humaine ?

1.2 Variables aléatoires et lois de probabilité

1.2.1 Variables aléatoires

Variable aléatoire : *on appelle variable aléatoire (v.a.) toute variable dont la valeur dépend du résultat d'une expérience probabiliste.*

On notera X la variable aléatoire, et x une réalisation de cette variable aléatoire, c'est-à-dire une valeur particulière prise par cette variable, pour un tirage aléatoire.

Exemples de variables aléatoires

- On regarde le résultat d'un lancer de dé à six faces non pipé. L'aléa vient de la façon de lancer le dé. La variable est la valeur affichée sur la face du dé (un entier entre un et 6). Lors du lancer de dé, chacune des six valeurs a une probabilité de 1/6 d'être obtenue.
- Dans un échantillon de papillons de l'espèce *Biston bistularia* capturés dans le sud de l'Angleterre, on observe si les papillons sont de couleur claire ou de couleur foncée. La variable aléatoire est ici la couleur, qui a deux modalités, *claire* ou *foncée*. La probabilité d'observer un papillon clair est en lien avec le niveau de pollution et la couleur des écorces des arbres.
- Dans un échantillon d'individus de la population islandaise, on observe si le groupe sanguin des individus est $O+$. La variable aléatoire est le groupe sanguin d'un individu, qui prend deux modalités, $O+$ ou *autre*. La probabilité qu'un individu soit de groupe $O+$ dépend de facteurs génétiques, en particulier de l'histoire de la colonisation de l'Islande par l'homme.
- Sur un échantillon d'épis de maïs, on compte le nombre de grains de chaque épi. L'aléa provient de l'environnement dans lequel est cultivée la plante et de son génotype. Les valeurs possibles seront des nombres entiers, avec une forte variation entre épis.
- On mesure la taille des filles de 6 ans à l'entrée en primaire échantillonnées dans différentes écoles en France. L'aléa vient à la fois de l'environnement et des différences génétiques entre les enfants. Les tailles sont des valeurs réelles, avec une variation continue.
- La grenouille rousse *Rana temporaria* est la grenouille du genre *Rana* la plus commune en Europe. Elle est largement répandue dans le nord de la France et en Belgique. Elle vit dans n'importe quel type d'habitat humide : bois (ornières des chemins forestiers, mares), landes, dépressions prairiales, terres cultivées, parcs et jardins. Il s'agit d'une espèce poïkilotherme, c'est-à-dire que sa température corporelle s'ajuste avec celle du milieu extérieur. On mesure la température corporelle de grenouilles échantillonnées dans une mare. L'aléa vient des variations de température à la surface de la mare, due à la position des zones ombragées par rapport au soleil. Ici aussi, les températures mesurées seront des nombres réels, avec une variation continue.

Une variable aléatoire est caractérisée par :

1. *Les valeurs qu'elle peut prendre, que l'on appelle le* ***support*** *de la variable aléatoire.*
2. *La probabilité d'observer chaque valeur dans la population ou* ***loi de probabilité****.*

On peut distinguer plusieurs catégories de variables aléatoires :

Variables quantitatives : caractéristiques numériques (taille, âge, etc.) qui résultent de mesures sur des individus. Elles s'expriment par des nombres sur lesquels les opérations arithmétiques de base (somme, moyenne, etc.) ont un sens. Une variable aléatoire quantitative peut être ***discrète*** (nombre de descendants d'un individu, nombre de soies thoraciques chez la drosophile, etc.) ou ***continue*** (poids, taille, etc.).

Variables qualitatives : caractéristiques non numériques. Elles peuvent être ***nominales*** (comme la couleur des yeux) ou ***ordinales*** lorsque l'ensemble des catégories est muni d'un ordre (exemple : peu infecté, moyennement infecté, très infecté). Les différentes valeurs d'une variable aléatoire qualitative s'appellent des ***modalités*** ou ***niveaux***. Attention, une variable qualitative peut être codée sous forme

d'une valeur numérique, mais s'il s'agit d'une variable nominale cela n'a pas de sens de faire une opération sur ces valeurs.

Le tableau ci-dessous récapitule les catégories des variables aléatoires pour les exemples cités, et la loi de probabilité associée.

Variable	Catégorie	Support	Loi de probabilité
Lancer de dés	qualitative ou quantitative	1, 2, ..., 6	$\{1/6, 1/6, ...\}$
Couleur papillons	qualitative nominale	clair/foncé	$\{p_c, 1-p_c\}$
Groupe sanguin en Islande	qualitative nominale	$O+$/autre	$\{p_{O+}, 1-p_{O+}\}$
Nombre de grains/plante	quantitative discrète	entiers positifs	approchée par loi normale
Taille des filles	quantitative continue	réels positifs	approchée par loi normale
Température des grenouilles	quantitative continue	entiers positifs	approchée par loi normale

1.2.2 Loi de probabilité

1.2.3 Variable aléatoire qualitative

Pour une variable aléatoire qualitative, on peut établir la liste de toutes les modalités possibles. Soit X une variable aléatoire ayant les modalités $\{a_1, a_2, ..., a_J\}$. On peut calculer la probabilité $P(X = a_j)$ que X prenne la valeur a_j $(j = 1, ..., J)$. La loi de probabilité de X est définie par l'ensemble des $P(X = a_j)$.
Une réalisation x de X est un tirage dans la population. x aura une valeur parmi toutes celles possibles. Autrement dit, X peut prendre les valeurs a_1, ou a_2, ou a_3, etc., mais x ne prendra qu'une seule valeur. Puisque les modalités sont mutuellement exclusives, on a :

$$P(X = a_i \quad \text{et} \quad X = a_j) = 0; \qquad i \neq j$$

De ce qui précède, on déduit que la probabilité que X appartienne à l'une ou l'autre des J modalités vaut 1.

$$\sum_{j=1}^{J} P\left(X = a_j\right) = 1$$

1.2.4 Variable aléatoire quantitative discrète

Les variables aléatoires quantitatives discrètes sont essentiellement des variables de dénombrement. Pour les variables de dénombrement, le support de ces variables aléatoires sont des valeurs entières, soit l'ensemble des entiers soit une partie finie (exemple 2,4,7,12) ou infinie (exemple tous les entiers paires). On peut énumérer ces valeurs et les appeler comme précédemment a_1, a_2, ...,.... On peut associer une probabilité $P(X = a_j)$ à chaque valeur a_j. Une réalisation x de X ne pourra prendre qu'une des valeurs du support a_1,a_2,.... Notons, que l'on peut avoir une variable quantitative discrète qui ne prend pas des valeurs entières. Les variables aléatoires quantitatives discrètes peuvent avoir un support fini ou infini.

Exemple. Soit X le nombre de mâles parmi les 396 descendants d'un taureau de la race Blanc-Bleu-Belge dont le sperme est disponible sur le catalogue de la race. Si l'on considère un descendant, il y a une chances sur deux qu'il soit mâle $(p = 0,5)$. Le nombre total de descendants mâles du taureau peut prendre toutes les valeurs entières entre 0 et 396. On peut calculer exactement la probabilité $P(X = a_j)$ pour chaque valeur entière a_j de l'ensemble $\{0, ..., 396\}$ (en utilisant la loi Binomiale $\mathcal{B}(396; 0,5)$; *cf.* chapitre 2).
Pour une variable aléatoire quantitative discrète avec un support fini, on a également :

$$\sum_{j=1}^{J} P(X = a_j) = 1.$$

Notons que si le support est infini, on aura une somme infinie (on remplace J par ∞).
Dans le cas où les valeurs a_1, a_2, ..., a_J sont ordonnées, on a :

$$P(X \leq a_k) = P(X = a_1 \text{ ou } X = a_2 \text{ ou..ou } X = a_k) = \sum_{j=1}^{k} P(X = a_j).$$

On utilise pour montrer cela la propriété suivante : si A et B sont deux événements disjoints, alors $P(A \quad \text{ou} \quad B) = P(A) + P(B)$. Par exemple, si $A : X = 0$ et $B : X = 1$, on a :

$$P(X \leq 1) = P(X = 0 \quad \text{ou} \quad X = 1) = P(X = 0) + P(X = 1).$$

1.2.4.1 Variable aléatoire quantitative continue

Les variables aléatoires quantitatives continues sont des variables à valeurs dans $\mathbb{R}$, ou bien, le plus souvent en biologie, dans un intervalle inclus dans $\mathbb{R}^+$ (mesures biométriques, concentrations, etc.). La probabilité pour qu'une variable aléatoire X (exemple, le taux de glucose du sang) prenne très exactement la valeur $x = 0,3846$ mg/L) est quasiment nulle. Par contre, on peut calculer la probabilité $F(x)$ pour que X soit plus petit qu'une certaine valeur x :

$$F(x) = P(X \leq x),$$

que l'on appelle ***fonction de répartition*** de X. La fonction de répartition permet de calculer la probabilité que X se trouve dans un intervalle compris entre x et $x + dx$: $P(X \in]x, x + dx]) = F(x + dx) - F(x)$.
La ***fonction de densité***, notée $f(x)$, est la dérivée de F :

$$f(x) = lim_{dx \to 0} \frac{(F(x + dx) - F(x))}{dx}.$$

Rappel : la fonction intégrale (notée $\int$) associe à chaque intervalle $]a; b]$ l'aire sous la courbe de $f(x)$, pour les valeurs de x dans l'intervalle $]a; b]$. Elle s'écrit $F_{a,b}(x) = \int_a^b f(x)dx$. Si l'on dérive $F(X)$, on retrouve la fonction $f(x)$.
Ainsi, la fonction de répartition de la variable aléatoire X peut être définie comme l'aire sous la courbe f entre $-\infty$ et x (figure 1.1). C'est la probabilité pour que X soit plus petit que x :

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

A noter que l'aire totale sous la densité vaut 1 (c'est la somme des probabilités) :

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Pour les lois de probabilité usuelles, les fonctions de répartition $F(x)$ sont tabulées.
Une façon plus intuitive de comprendre le lien entre fonction de répartition $F(x)$ et fonction de densité $f(x)$ est de considérer un n-échantillon (X_1, ..., X_n) de la variable aléatoire X, avec n très grand. On peut faire un histogramme des x_i en les regroupant en J classes de taille Δx. La classe j correspond aux x_i qui appartiennent à l'intervalle $]x_{min} + (j - 1)\Delta x; x_{min} + j\Delta x]$.
On peut définir la probabilité $P(X \in]x; x + \Delta x])$ et l'estimer, comme pour une variable discrète, par la fréquence observée de la classe $P_{\text{obs}}(]x; x + \Delta x])$ dans l'échantillon. Si le nombre d'observations est très grand, il devient possible de définir un très grand nombre de classes de toute petite taille dx, et l'on peut dessiner une fonction continue qui « enveloppe » la distribution empirique de X : c'est la fonction de densité de X.

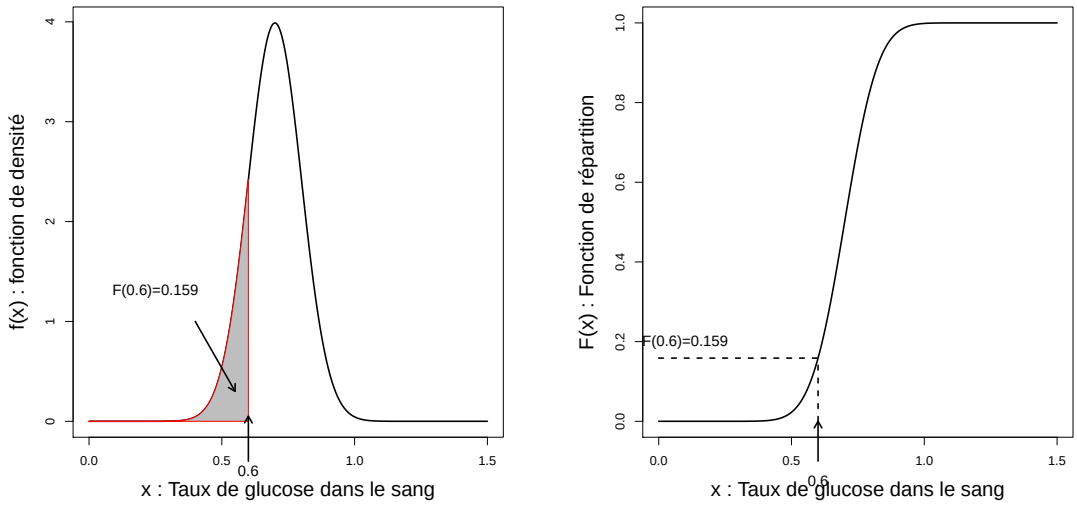


FIGURE 1.1 – **Loi de probabilité d'une variable aléatoire quantitative.** Sur cet exemple, X est la variable aléatoire associée à une mesure de la glycémie à jeun chez l'homme. $F(0,6)$ est la probabilité pour que cette mesure soit plus petite que 0,6. A gauche la fonction de densité, à droite la fonction de répartition.

Quantile. Pour chaque valeur x, on peut calculer la fonction de répartition $F(x)$. Inversement, on peut être intéressé par la valeur de x telle que $F(x)$ ait une certaine valeur. On appelle quantile d'ordre α, noté q_α, la valeur telle que

$$F(q_\alpha) = \alpha.$$

On utilise couramment, pour les tests statistiques, les quantiles $q_{0,025}$, $q_{0,05}$, $q_{0,95}$ et $q_{0,975}$.
La **médiane** d'une distribution est par définition le quantile à 50% : la moitié de la population a une valeur inférieure à la médiane.

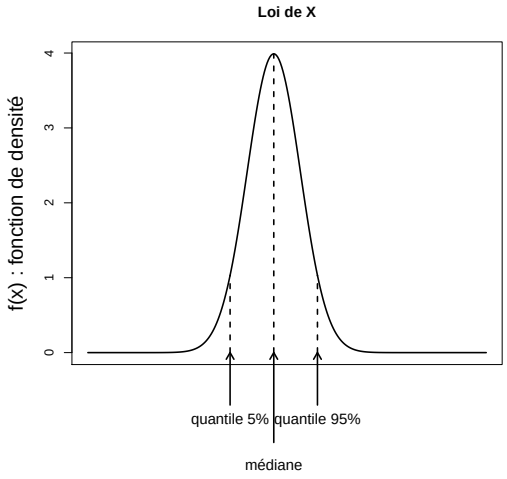


FIGURE 1.2 – **Quantiles usuels d'une loi de distribution.**

1.2.5 Espérance et variance

Les deux paramètres les plus couramment utilisés pour caractériser la loi de probabilité d'une variable aléatoire sont l'espérance et la variance.

Espérance (paramètre de centrage) : c'est la valeur que l'on s'attendrait à trouver, en moyenne, si l'on répétait une infinité de fois la même expérience aléatoire. Elle se note $E(X)$ et se lit « espérance de X ». Elle correspond à une moyenne pondérée des valeurs que peut prendre cette variable :

— *Pour les variables aléatoires discrètes,* $E(X) = \mu_X = \sum_j (P(X = a_j) \cdot a_j)$

— *Pour les variables aléatoires continues,* $E(X) = \mu_X = \int_{-\infty}^{+\infty} (xf(x)dx)$.

Notons que si l'on étudie deux variables aléatoires, X et Y, l'espérance de la somme de deux variables aléatoires est égale à la somme des espérances.

$$E(X + Y) = E(X) + E(Y)$$

Variable centrée : une variable est centrée lorsqu'on retire l'espérance à chacune des valeurs de X. $X_{\text{centrée}} = X - E(X)$. En conséquence, on peut montrer que $E(X_{\text{centrée}}) = 0$.

Variance (paramètre de dispersion) : en statistiques et en théorie des probabilités, la variance est une mesure servant à caractériser la dispersion d'une variable. Elle indique de quelle manière la série statistique ou la variable aléatoire se disperse autour de sa moyenne ou son espérance. Une variance de zéro signale que toutes les valeurs sont identiques. Une petite variance est signe que les valeurs sont proches les unes des autres alors qu'une variance élevée est signe que celles-ci sont très écartées. On définit la variance comme une somme pondérée des carrés des écarts à la moyenne.

— *Pour les v.a. discrètes,* $V(X) = \sigma_X^2 = \sum_j \left[P(X = a_j) \cdot (a_j - E(X))^2\right]$

— *Pour les v.a. continues,* $V(X) = \sigma_X^2 = \int_{-\infty}^{+\infty} \left[(x - E(X))^2 \cdot f(x)dx\right]$

On remarque que la variance s'exprime comme l'espérance de $(X - E(X))^2$:

$$\begin{aligned} V(X) &= \sigma_X^2 \\ &= E\left[(X - E(X))^2\right] \\ &= E(X^2) - [E(X)]^2 . \end{aligned} \tag{1.1}$$

Sous cette dernière forme, qui est commode pour les calculs, on voit que la variance est la différence entre l'espérance du carré de la variable et le carré de l'espérance de la variable.

Il est également important de remarquer, et de retenir, que la variance d'une variable X multipliée par une constante k est :

$$V(kX) = k^2 V(X)$$

et que la variance d'une variable à laquelle on ajoute une constante k est :

$$V(X + k) = V(X).$$

Exemple : mesure du nombre de garçons (X) dans une famille de deux enfants. Si la probabilité d'avoir un garçon est de $1/2$ alors la loi de probabilité de X s'écrit :

a_j	$P(X = a_j)$
0	1/4
1	1/2
2	1/4

L'espérance de X vaut $E(X) = (0/4 + 1/2 + 2/4) = 1$, et la variance $\sigma^2 = \frac{1}{4}(0-1)^2 + \frac{1}{2}(1-1)^2 + \frac{1}{4}(2-1)^2 = 1/2$

Ecart-type : quand on regarde une distribution continue, ce que l'on visualise n'est pas directement la variance (qui s'exprime dans le carré de l'unité de mesure). Le paramètre qui s'exprime dans l'unité de mesure est la racine carrée de la variance, que l'on nomme l'**écart-type** :

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

Variable réduite : Une variable divisée par son écart-type est appelée variable réduite. Par construction, la variance d'une variable réduite vaut un.

1.2.6 Couple de variables aléatoires

Couple de variables aléatoires : (X, Y) *est un couple de variables aléatoires lorsque deux mesures différentes* X *et* Y *sont observées chez le même individu statistique lors d'une expérience probabiliste.*

Lorsque l'on réalise des mesures différentes (eg taille et poids à la naissance) sur le même individu statistique, celles-ci peuvent dépendre l'une de l'autre. Par exemple, il est clair que pour des raisons biologiques (allométrie), on s'attend à une relation positive entre la taille et le poids à la naissance chez l'Homme. Ainsi, deux variables aléatoires mesurées chez les mêmes individus ne sont pas forcément indépendantes entre elles.

Notion de covariance : *On définit la covariance entre deux variables aléatoires quantitatives* X *et* Y *comme* $cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right]$.

Notion d'indépendance : *On dira qu'il y a indépendance entre les deux variables aléatoires* X *et* Y

lorsque, quelle que soit la valeur prise par X*, la loi de* Y *ne change pas, et réciproquement. De cette définition découle, pour les variables aléatoires discrètes :* $P(X = a \quad \text{et} \quad Y = b) = P(X = a) \cdot P(Y = b)$ *et pour les variables aléatoires continues* $f(x, y) = f(x) \cdot f(y)$.

Variance d'une somme. Pour tout couple de variables aléatoires X et Y, la variance de $X + Y$ s'écrit :

$$V(X + Y) = V(X) + V(Y) + 2cov(X, Y),$$

Si X et Y, sont indépendantes, alors $V(X + Y) = V(X) + V(Y)$.
Attention, l'indépendance implique bien une covariance nulle. Cependant, l'inverse n'est pas toujours vérifié.
La covariance des variables X et Y *réduites* est :

$$\rho_{X,Y} = cov\left(\frac{X}{\sqrt{V(X)}}, \frac{Y}{\sqrt{V(Y)}}\right) = \frac{cov(X, Y)}{\sqrt{[V(X)V(Y)]}}. \tag{1.2}$$

C'est le **coefficient de corrélation** de Pearson, une mesure de la covariance qui s'affranchit des unités dans lesquelles s'expriment X et Y.
Il est important de noter que la covariance d'une variable avec elle-même est sa variance : $cov(X, X) = V(X)$. On peut en déduire les bornes de ρ_{XY}. Si X et Y sont strictement liées ($X = Y$), on a

$$\rho_{XY} = \frac{cov(X, X)}{\sqrt{V(X)V(X)}} = 1.$$

Si X et Y sont strictement anti-corrélées ($X = -Y$), on a

$$\rho_{XY} = \frac{cov(X, -X)}{\sqrt{V(X)V(X)}} = -1.$$

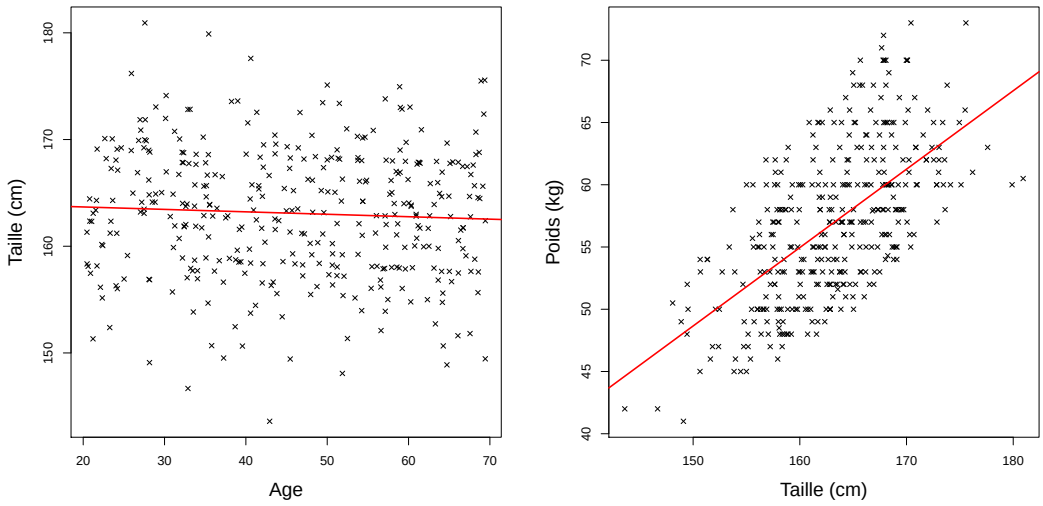


FIGURE 1.3 – **Couples de variables aléatoires quantitatives.** On mesure l'âge, la taille et le poids dans un échantillon de 350 femmes entre 20 et 70 ans, ayant terminé leur croissance. Chaque point représente une femme. A gauche : relation entre l'âge et la taille. On trouve une légère dépendance négative, la corrélation est de $-0,04$. A droite : la relation entre la taille et le poids chez les mêmes individus montre une forte liaison positive, avec un coefficient de corrélation de $0,63$.

1.3 Notion de n-échantillon

Lorsqu'une population est de très grande taille, on ne peut pas faire les mesures sur l'ensemble des individus. On va donc utiliser un sous-ensemble de la population pour inférer des informations sur la population.

***n*-échantillon** : sous-ensemble de n individus tirés au hasard et indépendamment dans la population de référence. On appelle X_i la variable aléatoire associée au tirage de l'individu i ($i = 1, ..., n$), et x_i la valeur observée chez l'individu i. La façon de constituer l'échantillon (tirages aléatoires indépendants) permet de faire l'hypothèse que les variables aléatoires X_i sont indépendantes et de même loi. Si le tirage est bien aléatoire, sans biais, alors l'échantillon sera dit "représentatif".

Les observations x_i peuvent servir à faire des hypothèses concernant la loi commune des X_i, c'est-à-dire la loi de X.

Exemple : A partir de l'échantillon de 350 femmes de la figure (1.3) on peut se demander si les deux variables observées sont indépendantes. A partir de cet échantillon, on peut également estimer la moyenne et la variance des variables observées. C'est ce que nous verrons au chapitre 1.5.

1.4 Statistiques descriptives

On considère les réalisations $(x_1, ..., x_i, ..., x_n)$ des variables aléatoires X_i d'un n-échantillon de la population de référence.
On peut réaliser une représentation graphique des x_i en les regroupant par classes de valeurs. On peut aussi utiliser des mesures pour décrire la distribution des x_i (par exemple moyenne ou variance). On appelle ces mesures des *statistiques résumées*.

1.4.1 Variable qualitative

Si on étudie une variable qualitative, alors les x_i ne peuvent prendre qu'un nombre fini de modalités $(a_1, ..., a_j, ..., a_J)$.

Effectif : *effectif observé pour chaque modalité*

Fréquence relative : *effectif de chaque modalité rapporté au nombre total d'individus* n.

On peut représenter les effectifs ou les fréquences relatives par un diagramme en bâtons.

1.4.2 Variable quantitative

Variable aléatoire discrète. Si on étudie une v.a. discrète, on peut calculer l'effectif n_j de chaque valeur a_j dans l'échantillon, ou sa fréquence relative $\frac{n_j}{n}$ et réaliser un *diagramme en bâtons*.
Par exemple, le tableau ci-dessous est issu d'une enquête de l'INSEE et donne le nombre d'enfants par femme, pour les femmes nées entre 1961 et 1965, ainsi que la fréquence de chacun des cas possibles (zéro, un, deux, trois ou plus de trois enfants) :

Nombre d'enfants/femme	0	1	2	3	> 3
fréquence (%)	13,5	18,2	38,9	20,2	9,2

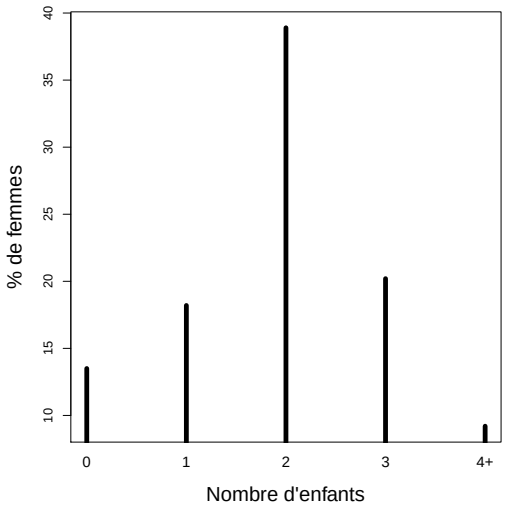


FIGURE 1.4 – **Diagramme en bâtons.** Nombre d'enfants par femme.

Variable aléatoire continue. Si on étudie est une v.a. quantitative continue, les x_i sont tous différents, mais on peut les regrouper par classe, en définissant des intervalles adjacents de valeurs. On compte alors les effectifs observés dans chaque intervalle. L'**histogramme** ou distribution empirique est une représentation des effectifs de chaque classe. De façon équivalente, chaque classe peut aussi être représentée par sa fréquence. Dans un histogramme, en ordonnée, on trouve soit l'effectif, soit la fréquence, soit la valeur telle que l'aire du bâton soit égale à la fréquence. L'allure de l'histogramme peut varier selon les bornes des classes choisies.

Mode d'une distribution empirique : *c'est la classe qui regroupe le plus grand nombre d'individus de l'échantillon. Ce mode peut varier en fonction des bornes des classes que l'on aura choisies.*

A noter que certaines distributions peuvent présenter 2 ou plusieurs pics, on parle dans ce cas de distribution bimodale ou multimodale. C'est le cas en particulier lorsque les individus de l'échantillon appartiennent à des populations différentes (figure 1.5).
Une représentation alternative à l'histogramme est la boîte à moustache ou boxplot (1.6), qui utilise les quantiles de la distribution empirique. Cette représentation permet de représenter sur un même graphique les distributions de plusieurs échantillons différents.

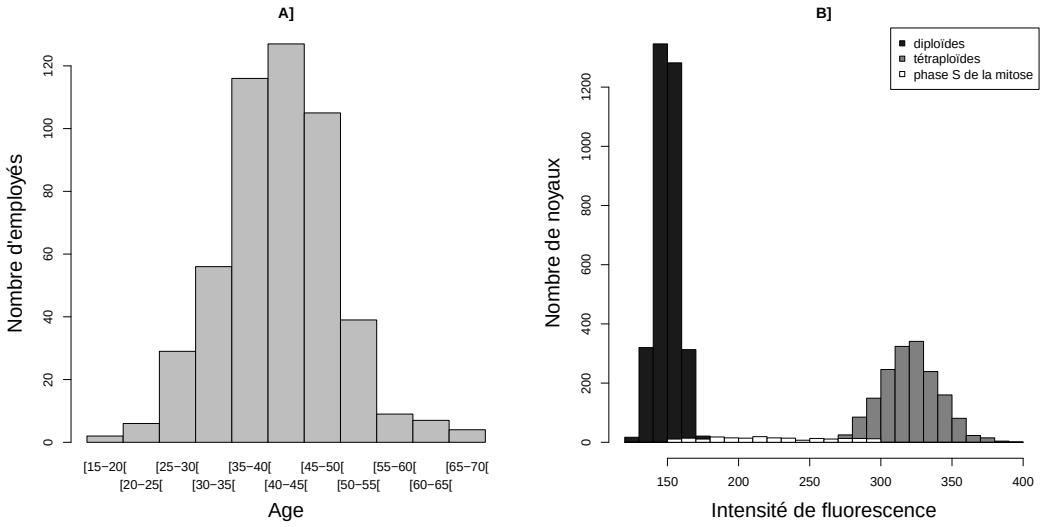


FIGURE 1.5 – **Histogrammes.** Les données continues sont regroupées en classes. A] Pyramide des âges pour les employés d'une banque. Le mode correspond à la classe 40-45 ans. B] Intensité de fluorescence de noyaux de cellules de feuilles de maïs mesurée au cytomètre de flux. On distingue trois populations de noyaux, qui diffèrent par leur degré de fluorescence, indiquant des niveaux de ploïdie différents. En noir, les noyaux de cellules diploïdes. En gris, les noyaux de cellules tétraploïdes. En blanc, des noyaux de cellules en phase S de la mitose.

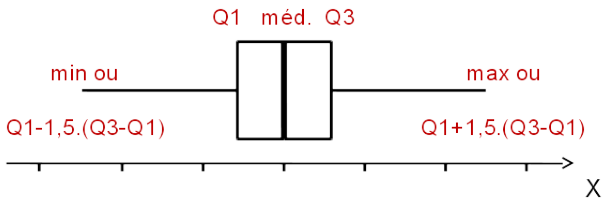


FIGURE 1.6 – **Boîte à moustache ou boxplot.** Une représentation usuelle de la distribution empirique d'un échantillon. La boîte correspond à l'intervalle défini par les quantiles 25% et 75%. L'extrémité des barres horizontales ont une signification variable selon les logiciels et peuvent représenter soit les valeurs minimales/maximales obervées dans l'échantillon soit des valeurs dépendant des quartiles. Dans ce dernier cas, les observations de l'échantillon au dela de ces extrémités sont représentées par des points ou des croix. La barre verticale donne la médiane de la distribution empirique.

Moyenne et variance. On peut utiliser l'analogie avec l'espérance et la variance d'une variable aléatoire pour calculer la moyenne et la variance d'un échantillon.
La moyenne de l'échantillon est un indicateur de position :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Elle est souvent proche du mode, c'est-à-dire de la valeur la plus fréquente.
La variance de l'échantillon donne une indication sur l'ampleur des variations autour de la moyenne. C'est un indicateur de dispersion :

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Pour un couple de variables aléatoires, la covariance donne une idée de la relation linéaire entre les deux variables :

$$s_{xy_n} = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \overline{x})(y_i - \overline{y})\right].$$

Quantiles. Si l'on classe les x_i par ordre croissant, on peut facilement calculer la proportion des observations inférieures à une certaine valeur a, $F_{\text{obs}}(a)$.
La ***médiane*** est la valeur a telle que $F_{\text{obs}}(a) = 0,5$, c'est-à-dire que la moitié des x_i ont une valeur inférieure à la médiane, et la moitié des x_i ont une valeur supérieure.
Le quantile α est la valeur a telle que $F_{\text{obs}}(a) = \alpha$. On utilise souvent les quantiles à 5% et 95%, qui donnent un intervalle contenant 90% de l'échantillon. Les quantiles à 25% et 75% donnent un intervalle contenant la moitié des individus de l'échantillon.

1.5 Statistiques inférentielles

Maintenant, on ne s'intéresse plus à l'échantillon mais à la population. On a une variable aléatoire quantitative et on va utiliser un échantillon pour approcher, on dit *estimer*, les descripteurs de la distribution de la variable aléatoire comme par exemple la moyenne ou la variance.
Chaque x_i est une réalisation d'une variable aléatoire X_i. Les X_i sont indépendantes et de même loi, caractérisées par leur espérance $E(X_i) = \mu_X$ et leur variance $V(X_i) = \sigma_X^2$.

1.5.1 Estimation de la moyenne d'une population

On peut définir la variable aléatoire suivante :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

qui est un calcul de la moyenne des tirages. On utilise pour cela la propriété de sommation des espérances de plusieurs variables aléatoires :

$$E(X + Y) = E(X) + E(Y)$$

On trouve donc facilement que $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu_X$, puisque chaque X_i a la même espérance μ_X.

Estimateur sans biais : *un estimateur sans biais est une variable aléatoire dont l'espérance est égale à la valeur exacte de la quantité que l'on souhaite estimer.*

On peut donc dire que $\overline{X}$ est un estimateur sans biais de μ_X, que l'on note $\widehat{\mu}_X$. En utilisant les réalisations x_i des X_i dans un échantillon, on peut donc proposer la moyenne empirique de l'échantillon comme estimation de la moyenne de la population :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

L'estimateur de μ_X est $\overline{X}$, c'est une variable aléatoire. L'estimation de μ_X est $\overline{x}$, c'est une réalisation de la variable aléatoire $\overline{X}$.

Précision d'un estimateur. Le fait qu'un estimateur soit sans biais ne donne pas d'idée sur sa précision. Pour cela, on peut se demander comment l'estimateur varie autour de sa moyenne, c'est-à-dire calculer la variance de l'estimateur.

Pour la variance de l'estimateur $\overline{X}$ de la moyenne, si on se rappelle que les X_i sont indépendants et de même loi, on a :

$$\begin{aligned} V(\overline{X}) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} n\sigma_X^2 \\ &= \frac{\sigma_X^2}{n}. \end{aligned}$$

Plus la taille de l'échantillon est grande, plus la variance de l'estimateur sera faible. A noter que la précision se mesure au travers de l'écart-type et diminue donc en $1/\sqrt{n}$ avec n. Pour augmenter la précision d'un facteur 10, il faut augmenter la taille de l'échantillon d'un facteur 100.

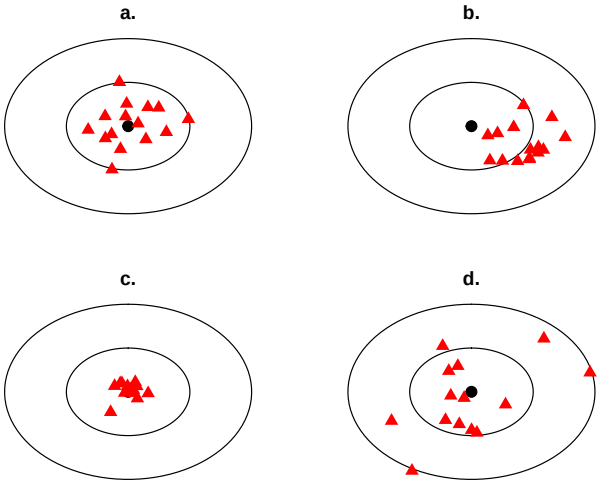


FIGURE 1.7 – **Propriétés d'un estimateur.** Le point noir est le paramètre à estimer dans la population. Chaque triangle rouge est une estimation à partir d'un n-échantillon. **a.** Estimateur sans biais. La moyenne des points rouges est proche du paramètre. **b.** Estimateur biaisé. La moyenne des estimations est différente de la valeur du paramètre. **c.** Estimateur sans biais et précis. Les points rouges sont centrés sur le paramètre et de faible variance. **d.** Estimateur sans biais mais imprécis, la variance de l'estimateur est importante.

1.5.2 Estimation de la variance

En utilisant la propriété de sommation des espérances, on a

$$E\left(\sum_{i=1}^{n}(X_i - \mu_X)^2\right) = n\sigma_X^2$$

Si μ_X est inconnu, on ne peut pas l'utiliser pour estimer la variance, on va donc remplacer μ_X par son estimateur $\overline{X}$, ce qui va modifier l'espérance. En effet on peut écrire :

$$\begin{aligned} E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) &= E\left(\sum_{i=1}^{n}\left[(X_i - \mu_X) - (\overline{X} - \mu_X)\right]^2\right) \\ &= E\left(\sum_{i=1}^{n}(X_i - \mu_X)^2 - n(\overline{X} - \mu_X)^2\right) \\ &= n\sigma_X^2 - \sigma_X^2 \\ &= (n-1)\sigma_X^2. \end{aligned} \tag{1.3}$$

Comme estimateur sans biais de la variance, on peut donc proposer :

$$S^2_{X_{n-1}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

que l'on note aussi $\widehat{\sigma}^2_X$.
En utilisant les réalisations x_i des X_i dans un échantillon, on calcule l'estimation de la variance de la population, que l'on note s^2_X :

$$s^2_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Variance empirique corrigée : *s^2_X est appelée la variance empirique corrigée de l'échantillon. C'est une estimation de la variance de la population.*

Covariance empirique corrigée : *de même, s_{XY} est appelée la covariance empirique de l'échantillon. C'est une estimation sans biais de la covariance dans la population.*

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}((x_i - \overline{x})(y_i - \overline{y}))$$

1.6 Récapitulatif

On se pose des questions sur les paramètres de la loi de distribution d'une variable aléatoire, ou d'un couple de variables aléatoires dans une population. On observe un n-échantillon d'individus de la population. Il est possible de calculer un certain nombre de **statistiques résumées** sur cet échantillon. On peut aussi utiliser l'échantillon pour faire des inférences sur les paramètres de la population. On utilise alors les valeurs mesurées dans l'échantillon pour calculer des **estimations** des paramètres de la population :

	Population	**Estimateur**	**Estimation**
	v.a. X et Y	v.a. $X_1, ..., X_n$, indépendantes et de même loi	observations $x_1, ..., x_n$
	.		
		v.a. $Y_1, ..., Y_n$, indépendantes et de même loi	observations $y_1, ..., y_n$
Moyenne	$E(X) = \mu_X$	$\overline{X}$	$\overline{x}$
Variance	$V(X) = \sigma^2_X$	$S^2_{X_{n-1}}$	s^2_X
Covariance	$Cov(X,Y) = \sigma_{XY}$	S_{XY}	s_{XY}
Corrélation	ρ_{XY}	$\hat{\rho}_{XY}$	r_{XY}

avec :

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
s^2_X &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \overline{x})(y_i - \overline{y})] \\
r_{XY} &= \frac{s_{XY}}{\sqrt{s^2_X s^2_Y}}.
\end{aligned}$$

Chapitre 2

Lois et tests

2.1 Lois de probabilité usuelles

2.1.1 Cas des variables quantitatives discrètes

Pour une variable discrète X ayant comme support $a_1, ..., a_J$, la distribution est donnée par la probabilité que la variable prenne la valeur a_j, $P(X = a_j)$, pour toute valeur de j.
La somme de tous les $P(X = a_j)$ est égale à 1. On peut toujours représenter la loi de probabilité par un diagramme en bâtons (figure 2.1).

2.1.1.1 Loi de Bernoulli

On considère une variable avec deux modalités possibles (exemple : pile/face, port de lunette/pas de lunettes, homme/femme, ségrégation d'un caractère monogénique dominant dans une descendance F2). On peut toujours donner les codes 0 et 1 aux deux modalités. Par exemple, on peut écrire $X = 1$ si pile et $X = 0$ si face.
La loi de la variable est donnée par $P(X = 1) = p$ et $P(X = 0) = 1 - p$. On dit alors que X suit une loi de Bernoulli, notée $\mathcal{B}(p)$. Son espérance est $E(X) = p$ et sa variance est $V(X) = p(1 - p)$ (voir équation (1.1)).

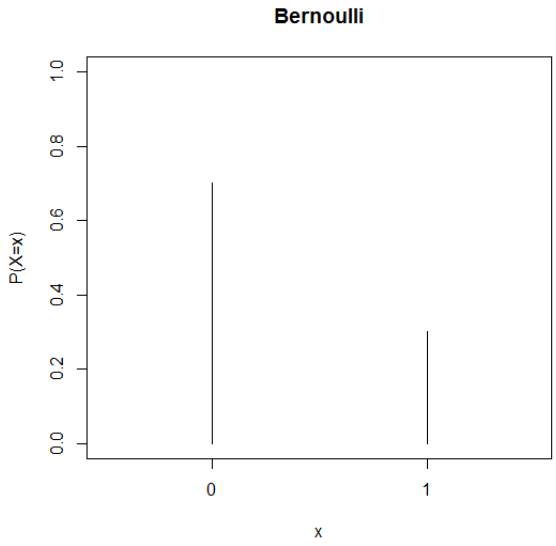


FIGURE 2.1 – **Diagramme en bâtons** représentant la loi de probabilité d'une variable de Bernoulli de probabilité $p = 1/4$.

2.1.1.2 Loi Binomiale

Imaginons que l'on répète n fois le tirage de Bernoulli. On a donc n variables de Bernoulli indépendantes X_i pour $i = 1$ à n, c'est un n-échantillon. On compte le nombre de fois où l'on tombe sur pile, c'est-à-dire que l'on fabrique une nouvelle variable Y égale à la somme des X_i.

$$Y = \sum_{i=1}^{n} X_i$$

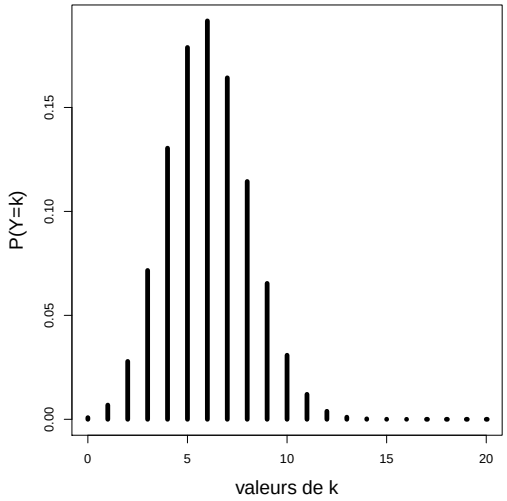


FIGURE 2.2 – **Loi de probabilité d'une variable Binomiale.** $Y \sim \mathcal{B}(20; 0,3)$.

La v.a. Y suit une loi binomiale de paramètres n, le nombre de tirages, et p, la probabilité de tomber sur pile. La loi binomiale est notée : $\mathcal{B}(n; p)$.
La loi de Y est donnée par $P(Y = k)$ pour $k = 0, ..., n$.

$$P(Y = k) = C_n^k p^k (1-p)^{n-k},$$

où C_n^k est la combinaison de k parmi n.
L'espérance de Y est

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np$$

et sa variance est

$$V(Y) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = np(1-p).$$

Toutes les variables de comptage d'événements indépendants suivent des lois Binomiales. Une caractéristique nominale d'un individu peut toujours être décrite comme une variable de Bernoulli (présence ou non de la caractéristique). Lorsque l'on s'intéresse au nombre d'individus possédant la caractéristique dans un n-échantillon, on a affaire à une loi Binomiale, dont le support est $\{0, 1, ..., n\}$.

Estimation de *p*. Souvent, on ne connaît pas la probabilité de succès p, c'est-à-dire la probabilité qu'un individu de la population choisi au hasard possède la caractéristique. On va chercher à l'estimer à partir des observations dans l'échantillon.
On peut déduire de la loi de Y la loi de $Z = \frac{Y}{n}$, qui décrit la proportion de succès pour n tirages indépendants. Z prend les valeurs $\{\frac{0}{n}, \frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\}$.

$$P\left(Z = \frac{k}{n}\right) = P(Y = k) = C_n^k p^k (1-p)^{n-k}.$$

On peut calculer l'espérance de Z :

$$E(Z) = E\left(\frac{Y}{n}\right) = \frac{E(Y)}{n} = p$$

et sa variance :

$$V(Z) = V\left(\frac{Y}{n}\right) = \frac{V(Y)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

On peut remarquer que Z est un estimateur sans biais de p, et que la variance de l'estimateur diminue avec la taille de l'échantillon n. On peut donc proposer l'estimation :

$$\widehat{p} = \frac{y}{n}$$

où y est le nombre observé d'individus possédant la caractéristique dans l'échantillon.

Exemple : on souhaite connaître la probabilité p de souffrir d'un problème de dos chez les hommes de plus de quarante ans. On réalise une enquête auprès de 20 hommes de plus de quarante ans. Sept déclarent souffrir d'un problème de dos. On peut donc estimer la probabilité par

$$\widehat{p} = 7/20 = 0,35$$

.

2.1.1.3 Loi de Poisson

Soit une variable aléatoire Y pouvant prendre des valeurs discrètes 0, 1, 2, 3, ..., $+\infty$ (en théorie). Si Y suit une loi de Poisson on a

$$P(Y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

et

$$E(Y) = V(Y) = \lambda$$

Pour une loi de Poisson, la moyenne et la variance sont égales.

On appelle parfois la loi de Poisson la loi des évenèments rares car si on a une variable aléatoire X suivant une loi binomiale de paramètres n très grand et p petit ($n \geq 50$, $p \leq 0,1$ et $np < 15$), on peut approximer cette loi Binomiale par une loi de Poisson de paramètre np. Par exemple, le nombre d'accidents à un carrefour un jour donné. Si un grand nombre de voitures passent chaque jour et que la probabilité qu'il y ait un accident à chaque passage de voiture est faible, le nombre d'accidents suivra une loi de Poisson.
Dans ce polycopié lorsqu'on étudiera une variable aléatoire X, si on connaît sa loi, on l'indiquera avec le signe $\sim$, si c'est une loi approchée on la notera avec le signe $\approx$. Par exemple si $X \sim \mathcal{B}(n; p)$ avec n et p remplissant les conditions énumérés ci-dessus, $X \approx \mathcal{P}(np)$.
Si l'on compare deux variables aléatoires ayant la même moyenne, l'une suivant une loi Binomiale (moyenne np), et l'autre une loi de Poisson (moyenne $\lambda = np$), on peut remarquer que la variable aléatoire Binomiale aura une variance plus faible. En effet, $np(1-p) < np$.
Lorsque p devient très petit et que n est grand, on a $np(1-p) \approx np$, ce qui justifie l'approximation de la loi Binomiale par une loi de Poisson.

2.1.2 Cas des variables quantitatives continues

Il existe de nombreuses lois de probabilité permettant de *modéliser* une variable quantitative continue. La plus connue est la loi normale. Cependant, toutes les variables quantitatives continues ne suivent pas une loi normale. Nous verrons d'autres lois continues dans ce cours.
On rappelle qu'une variable continue X est décrite par sa fonction de densité $f(x)$ ou par sa fonction de répartition $F(x) = P(X \leq x)$.

2.1.2.1 Loi uniforme

La loi uniforme est caractérisée par la propriété suivante : tous les intervalles de même longueur inclus dans le support de la loi ont la même probabilité. Par conséquent, la fonction de densité associée est une constante.

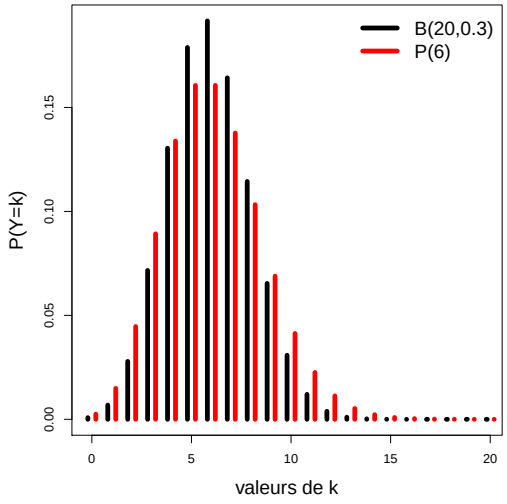


FIGURE 2.3 – **Lois de probabilité d'une variable Binomiale et d'une variable de Poisson de même espérance $E(Y) = np = \lambda = 6$.** Pour la loi de Poisson (en rouge), la probabilité des événements extrêmes (à droite ou à gauche) est plus grande, conduisant à une variance plus importante.

La fonction de densité d'une loi uniforme $\mathcal{U}(a, b)$ dont le support est l'intervalle $[a, b]$ est

$$f(x) = \frac{1}{b-a}$$

et la fonction de répartition $F(x)$ est une fonction linéaire :

$$F(x) = \frac{x-a}{b-a}.$$

L'espérance vaut $E(X) = (a+b)/2$, et la variance $V(X) = \frac{(b-a)^2}{12}$.

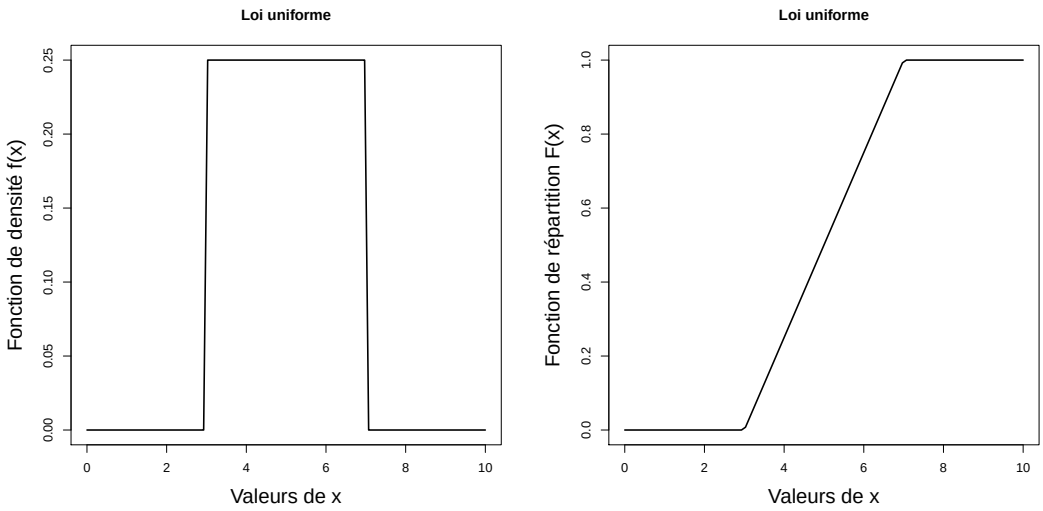


FIGURE 2.4 – **Exemple de loi uniforme $\mathcal{U}(3, 7)$.** La fonction de densité est représentée à gauche, la fonction de répartition à droite.

2.1.2.2 Distribution gaussienne

Une variable X qui suit une loi normale (ou loi gaussienne) de moyenne μ et de variance σ^2 a la fonction de densité suivante :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Loi normale centrée réduite. Cette loi correspond à celle d'une variable aléatoire qui suit une loi normale d'espérance $\mu = 0$ et de variance $\sigma^2 = 1$. On peut toujours passer d'une variable normale $X \sim \mathcal{N}(\mu; \sigma^2)$ à une variable normale centrée et réduite en utilisant une transformation. Ainsi, la variable centrée $X - \mu$ suit la loi $\mathcal{N}(0; \sigma^2)$, et la variable centrée réduite $\frac{X-\mu}{\sigma}$ suit la loi $\mathcal{N}(0; 1)$.
En conséquence, seule la loi normale centrée réduite est en général implantée dans les logiciels de statistiques.

Quelles variables sont souvent modélisées commme des variables gaussiennes ?

- De nombreuses variables biologiques, comme par exemple le poids, la taille, etc.
- Une somme de variables aléatoires indépendantes suivant des lois normales suit une loi normale. Donc notamment si X_1, ..., X_n sont des gaussiennes indépendantes de même loi $\mathcal{N}(\mu; \sigma^2)$ alors leur moyenne est gaussienne de loi $\mathcal{N}(\mu; \frac{\sigma^2}{n})$.
- Soit X_1, X_2, ..., X_n des variables indépendantes de même loi, de moyenne μ et de variance σ^2. On s'intéresse à la moyenne de X, $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$. Quelle que soit la loi des X_i, la loi de $\overline{X}$ *tend* vers une loi normale quand n tend vers l'infini. C'est le **théorème central limite**. En pratique, si n est assez grand (souvent on prend $n \geq 30$), on peut faire l'hypothèse que $\overline{X}$ suit asymptotiquement une loi normale de moyenne μ et de variance $\frac{\sigma^2}{n}$. On notera donc cela $\bar{X} \approx \mathcal{N}(\mu; \sigma^2/n)$ et non $\bar{X} \sim \mathcal{N}(\mu; \sigma^2/n)$
- Une application directe du théorème central limite est l'approximation de la loi binomiale par une loi normale. La loi binomiale est une somme de variables de Bernoulli de même loi et indépendantes. Donc, si Y suit une loi binomiale $\mathcal{B}(n; p)$ et que $n > 30$, $np > 5$ et $n(1-p) > 5$, alors Y suit approximativement $\mathcal{N}(np; np(1-p))$.

Il existe d'autres lois que la loi normale pour décrire la distribution de variable quantitative, comme la loi exponentielle ou la loi gamma. Par exemple, les variables aléatoires mesurant des temps d'attente avant qu'un événement se produise ont typiquement des lois de probabilité exponentielle, dont la fonction de densité est donnée par :

$$f(x) = \lambda e^{-\lambda x}$$

2.2 Principe des tests

On considère une variable aléatoire X qui suit une loi dont la forme est connue (par exemple une loi de Poisson ou une loi normale) mais dont les paramètres sont inconnus. On peut estimer ces paramètres ou chercher à les situer dans une gamme de valeurs, ou encore les comparer à une référence ou entre eux. On parle alors de **tests d'hypothèse**.

Exemples

- La population de la métropole est-elle représentative de la population française ?
- Une opération contre la cataracte modifie-t-elle l'hypermétropie ?
- La taille des ailes des hirondelles est-elle la même chez les mâles et les femelles ?
- L'angle articulaire du bras lors de la pause de membrane d'étanchéité sur des chantiers est supérieure aux recommandations de l'INRS ?

Une difficulté est de choisir le bon test. Il faut bien identifier la question posée et la nature des variables que l'on étudie, dans certains cas valider *a posteriori* des hypothèses complémentaires, etc. Nous vous proposons ici, à l'aide d'un exemple simple, une démarche à suivre qui peut être appliquée pour tous les tests d'hypothèse.

2.2.1 Démarche à suivre pour réaliser un test statistique

Supposons que l'on souhaite vérifier si un appareil qui mesure la longueur focale de lentilles est bien réglé. Le constructeur indique que la précision de la mesure est telle que $\sigma = 0,1$. On a un étalon dont on connaît la distance focale, qui vaut 16 cm. On réalise $n = 10$ mesures de longueur focale de l'étalon à l'aide de l'appareil.

2.2.1.1 Modèle

On appelle X la mesure de la longueur focale de l'étalon. X est une variable aléatoire car, comme l'indique le constructeur, si l'on répète la mesure, on ne trouvera pas exactement le même résultat, cela dépend de la précision de l'appareil. On dispose d'un 10-échantillon de mesures indépendantes X_i. Si l'appareil est bien réglé, on s'attend à ce que chaque X_i suive une loi normale $\mathcal{N}(16; \sigma^2 = 0,01)$. Si l'appareil est mal réglé, l'espérance de X_i devrait être différente de 16. C'est ce que l'on voudrait savoir. On appelle μ la valeur inconnue de l'espérance de X_i et l'on pose le modèle suivant : X_i suit une loi normale $\mathcal{N}(\mu; 0,01)$. Ce modèle suppose que, pour l'instant, on fasse confiance au constructeur concernant la précision de l'appareil ($\sigma = 0,1$ est supposé connu).

2.2.1.2 Hypothèses H0 et H1

Dans ce test, on peut faire deux hypothèses mutuellement exclusives :
- H0 : en espérance, la distance focale mesurée est de 16 ($\mu = 16$)
- H1 : en espérance, la distance focale mesurée est différente de 16 ($\mu \neq 16$)

Attention, les hypothèses portent toujours sur une caractéristique de la loi de probabilité de la variable aléatoire. On ne fait pas d'hypothèse sur la réalisation (x_1, ..., x_n) qui est donnée. Même si H0 est vraie, il y a peu de chances pour que la moyenne de l'échantillon $\overline{x}$ soit exactement égale à 16.
Le principe du test est de trouver une variable aléatoire :
- dont on peut calculer une réalisation à partir des observations
- dont on connaît la loi sous H0.
- pour laquelle la loi sous H1 (même si on ne la connaît pas) est différente.

On appelle cette variable aléatoire **la statistique de test**. H0 est appelée **hypothèse nulle** : c'est l'hypothèse que l'on veut mettre à l'épreuve, ce qui va conduire à la réfuter ou pas. En général, c'est celle qui est « chiffrée », c'est à dire qu'on connaît la valeur des paramètres et qu'on peut donc connaître précisément la loi de la statistique de test. H1 est appelée "l'hypothèse alternative".

2.2.1.3 Choix d'une statistique de test

Pour choisir la statistique de test, on peut revenir au modèle. On sait que si les X_i suivent une loi normale $\mathcal{N}(\mu; 0,01)$, alors $\overline{X}$ suit une loi normale $\mathcal{N}(\mu; 0,01/10)$, et $\overline{X}$ est un estimateur sans biais de μ. On peut calculer une réalisation de $\overline{X}$ à partir d'un échantillon : $\overline{x}$. Par contre, la loi de $\overline{X}$ est inconnue, car la moyenne μ est inconnue. On peut donc proposer comme statistique de test la variable aléatoire $Z = \frac{\overline{X}-16}{0,1/\sqrt{10}}$, dont la loi sous H0 est connue. En effet, on sait que

$$\frac{\overline{X} - \mu}{0,1/\sqrt{10}} \sim \mathcal{N}(0; 1).$$

Donc comme sous H0, $\mu = 16$, on a :

$$Z = \frac{\overline{X} - 16}{0,1/\sqrt{10}} \sim \mathcal{N}(0; 1).$$

Il est également possible de calculer une réalisation de Z dans le n-échantillon :

$$z = \frac{\overline{x} - 16}{0,1/\sqrt{10}}.$$

La variable aléatoire Z est donc un bon candidat comme statistique de test pour tester H0 vs H1. On appelle $P_{\text{H0}}(z) = P_{\text{H0}}(Z \leq z)$ la fonction de répartition de la variable aléatoire Z sous l'hypothèse H0.

Si H0 est vraie, Z est bien une variable aléatoire qui ne dépend que de $\overline{X}$ car les autres paramètres sont des constantes. On notera z_{obs} la réalisation de Z dans un échantillon.

2.2.1.4 Zone de rejet de H0 et choix du risque α

Définition d'une zone de rejet On peut comparer z_{obs} à la distribution théorique de Z sous H0. Si H0 est vraie, on s'attend à ce que la plupart des réalisations de Z soient proches de 0 ($\overline{X}$ proche de 16). Il est peu probable que Z soit très grand en valeur absolue. Si l'on trouve une valeur élevée pour z_{obs}, en valeur absolue, on va décider qu'il y avait peu de chances de l'observer sous H0, et on rejettera H0. **On définit la zone de rejet comme un intervalle :**

$$]-\infty; -z_{\text{seuil}}] \cup [z_{\text{seuil}}; +\infty[.$$

Cet intervalle, illustré figure 2.5, contient les valeurs de la statistique de test qui ont peu de chances de se produire si l'hypothèse H0 est vraie. Pour trouver où se situe la zone de rejet, il faut réfléchir à la localisation de la loi de la statistique de test sous H1 : à droite de la loi sous H0 ? à sa gauche ? à droite ou à gauche ?

Choix du risque α Même si H0 est vraie, la probabilité que Z soit dans la zone de rejet n'est pas nulle. Il existe donc une certaine probabilité de se tromper en rejetant H0 alors que H0 est vraie, c'est ce qu'on appelle le **risque de première espèce**, noté α. Le principe des tests est que l'on fixe α *a priori*, c'est-à-dire *avant* de réaliser l'expérience. On va dire qu'on est prêt à accepter un risque α de se tromper en rejetant H0 alors qu'H0 est vraie. En notant, $P_{\text{H0}}()$ la probabilité d'un évènement sous l'hypothèse qu'H0 est vraie, on peut écrire :

$$\alpha = P(\text{rejeter H0} \mid \text{H0 vraie}) = P_{\text{H0}}(\text{rejeter H0})$$

Pour le test présenté ci-dessus, on a : $\alpha = P_{\text{H0}}(Z > z_{\text{seuil}} \quad \text{ou} \quad Z < -z_{\text{seuil}})$
Dans la pratique, on choisit souvent $\alpha = 5\%$ ou $\alpha = 1\%$. Graphiquement, α est la probabilité représentée par les aires grisées sous la fonction de densité de la statistique de test sous H0 (voir Figure 2.5), dans la zone de rejet de H0. On recherche donc la valeur de z_{seuil} correspondante. A noter que $-z_{\text{seuil}}$ est le quantile d'ordre $\alpha/2$ de la loi de Z, et z_{seuil} est le quantile d'ordre $(1-\alpha/2)$.
La zone de rejet se calcule donc en recherchant des quantiles dans la loi de Z. Ici, on cherche z_{seuil} tel que :

$$F_{\text{H0}}(z_{\text{seuil}}) = 1 - \alpha/2.$$

Les fonctions inverses des fonctions de répartition des lois usuelles sont implémentées dans la plupart des logiciels de statistique et les tableurs. Dans notre exemple, en fixant un risque $\alpha = 5\%$, on trouve $z_{\text{seuil}} = 1,96$ (*fonction R* : `qnorm(p=0.975)`). La zone de rejet de l'hypothèse H0 est donc ici :

$$]-\infty; -1,96] \cup [1,96; +\infty[.$$

Si z_{obs} se trouve dans la zone de rejet, on rejettera H0 et on retiendra H1 au risque d'erreur de première espèce α. En revanche, si z_{obs} n'est pas dans la zone de rejet, on ne pourra pas rejeter H0. Attention, on ne pourra pas pour autant valider H0 car on n'aura pas contrôlé le risque de seconde espèce d'accepter H0 à tort.

2.2.1.5 Réalisation du test : calcul de la valeur observée de la statistique

Jusqu'à présent, les valeurs numériques de l'échantillon n'ont pas été utilisées. La réalisation du test consiste à calculer z_{obs} à partir de l'échantillon.
Les valeurs des 10 mesures sont les suivantes : 16,03 15,89 16,33 16,44 16,07 16,41 16,12 15,94 16,45 16,08.

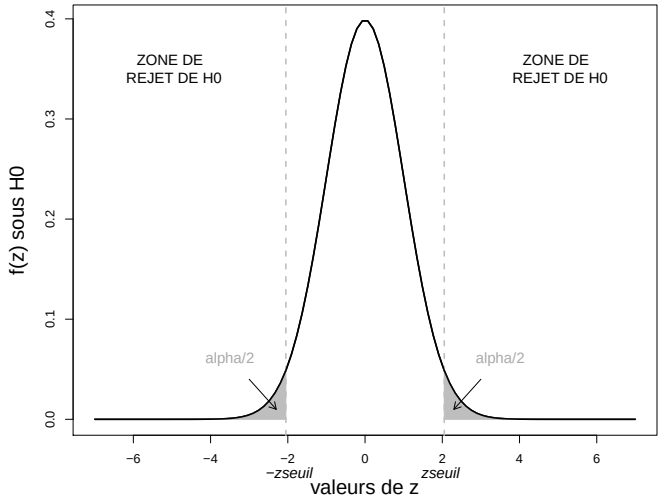


FIGURE 2.5 – **Illustration de la zone de rejet.** La courbe noire est la fonction de densité de Z sous H0. La zone de rejet de l'hypothèse H0 est définie comme un intervalle calculé de façon à ce que l'aire sous la courbe dans l'intervalle de la zone de rejet soit égale au seuil choisi α, appelé *alpha* sur la figure.

On peut calculer la moyenne de l'échantillon : $\overline{x} = 16,176$, et on connaît $n = 10$ et $\sigma = 0,1$. On trouve donc

$$z_{\text{obs}} = \frac{16,176 - 16}{0,1/\sqrt{10}} = 5,56.$$

Compte tenu de l'erreur de mesure annoncée par le constructeur, si la machine est bien réglée, il y a peu de chance de tomber sur des "grandes valeurs" (en valeur absolue). Le principe du test est de rejeter H0 au delà d'une valeur seuil déterminée en fixant le risque de première espèce α. Si l'on choisit $\alpha = 0,05$, le seuil est $1,96$. Comme $5,56$ est plus grand que $1,96$, on est dans la zone de rejet, c'est-à-dire qu'on rejette l'hypothèse H0. Ici, on en conclut que l'appareil est probablement mal réglé.

2.2.1.6 *p*-value

Pour finir le test, on calcule la p-value, ou *pval*, ou niveau observé.

***p*-value** : *C'est la probabilité que la statistique de test soit au delà (dans la zone de rejet) de z_{obs} sous H0. Autrement dit,c'est la valeur de α qu'il aurait fallu prendre pour qu'une limite de la zone de rejet soit z_{obs}. Si cette probabilité est inférieure au risque α choisi, on rejettera H0. La p-value varie entre 0 et 1.*

$$pval = P_{\text{H0}}(|Z| > |z_{\text{obs}}|)$$

Ici z_{obs} est positif, donc $pval = 2P_{\text{H0}}(Z > z_{\text{obs}}) = 2,70 \cdot 10^{-8}$ (*fonction R :* `2*(1-pnorm(q=5.56))`). Cette p-value indique que la probabilité de trouver une moyenne empirique au delà de 16,176 sous l'hypothèse H0 était extrêmement faible.
On peut voir sur la figure 2.6 les valeurs de z_{seuil} et de z_{obs} trouvées dans l'exemple traité. On constate aussi que **lorsque** $pval < \alpha$**, alors** z_{obs} **se situe dans la zone de rejet du test**.

Réalisation du test. Si $z_{\text{obs}} = z_{\text{seuil}}$, alors $pval = \alpha$. Lorsque z_{obs} s'éloigne de zéro, la p-value diminue et devient de plus en plus petite. En dehors de la zone de rejet, on a toujours $pval > \alpha$. Il existe donc une autre façon de réaliser un test statistique :

- Choisir le risque α
- Calculer z_{obs}
- Utiliser z_{obs} pour calculer la p-value

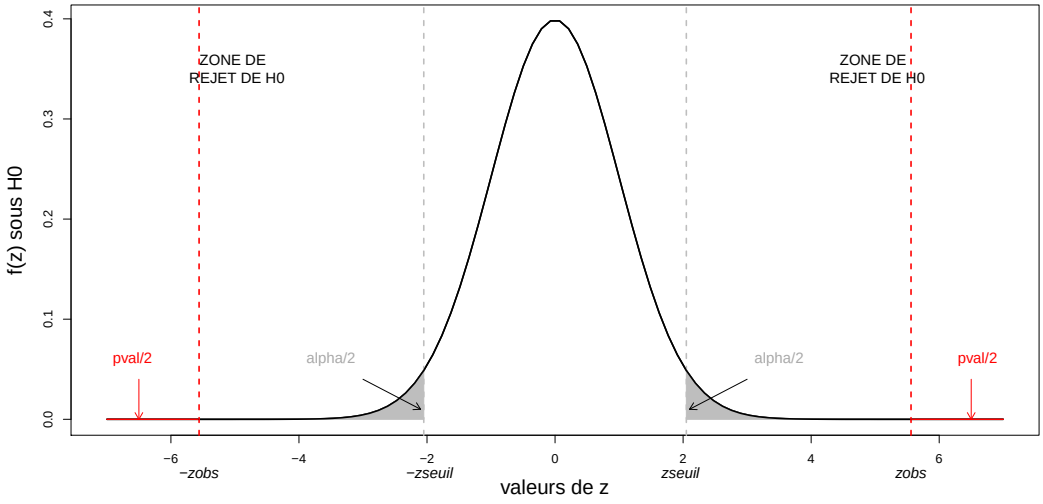


FIGURE 2.6 – **Illustration de la notion de *p*-value.** La courbe noire est la fonction de densité de Z sous H0. La p-value est définie comme l'aire sous la courbe dans l'intervalle des valeurs de z telles que $|Z| > |z_{\text{obs}}|$.

— Si $pval < \alpha$, on rejette H0, si $pval > \alpha$, on ne peut pas rejeter H0.

C'est ce que font aujourd'hui la plupart des logiciels.

2.2.1.7 Conclusion

La conclusion du test comprend deux parties. La conclusion statistique consiste à décider si l'on accepte ou si l'on rejette H0. La conclusion biologique demande à revenir à la question de départ. Lorsque l'on rejette H0, on dira que l'on a une différence statistiquement significative au niveau α.

2.2.2 Risques de première et deuxième espèce, puissance

Lorsque l'on réalise un test statistique, on ne connaît pas la réalité. On ne sait pas si l'hypothèse H0 est vraie ou fausse. Ainsi, lorsque l'on prend une décision à l'issue du test, on peut se tromper de deux façons différentes :

— On rejette H0 alors qu'elle est vraie. On parle alors de *rejet à tort* de H0, ou fausse découverte, ou encore *faux positif*. La probabilité de rejeter H0 alors qu'elle est vraie est appelée **risque de première espèce**, noté α. Ce risque α, que l'on appelle aussi niveau α et parfois seuil α, est toujours choisi *a priori*.

$$\alpha = P\,(\text{rejet de H0} \mid \text{H0 vraie})$$

— On conserve H0 alors qu'elle est fausse. Il s'agit d'un faux négatif. La probabilité d'un faux négatif est appelée **risque de deuxième espèce**, noté β. Pour H1, on peut avoir une hypothèse simple (par exemple $\mu = 18$) mais, en général, on a une hypothèse composite (par exemple $\mu \neq 18$).

2.2.2.1 Notion de puissance

Dans le cas d'une hypothèse simple, on définit la puissance par

$$1 - \beta = 1 - P\,(\text{non rejet de H0} \mid \text{H1 vraie})\,.$$

C'est la probabilité de rejeter H0 si H0 est fausse. Dans le cas d'une hypothèse composite, on définit la fonction puissance dont la valeur dépend de la vraie valeur du paramètre.

En général, β ne peux pas être calculée, car on ne connait pas la loi de la statistique de test sous H1, elle dépend de μ qui est inconnue. Dans l'exemple précédent, la statistique de test se calculait comme $Z = (\overline{X} - 16)/\frac{0,1}{\sqrt{10}}$. Si H0 est fausse, alors $E\left(\overline{X}\right) = \mu \neq 16$. L'espérance de Z vaut donc :

$$E(Z) = \frac{\mu - 16}{0,1/\sqrt{10}}$$

En moyenne, Z sera d'autant plus loin de zéro que μ est différent de 16. Ici, la puissance du test augmente lorsque μ s'éloigne de 16.

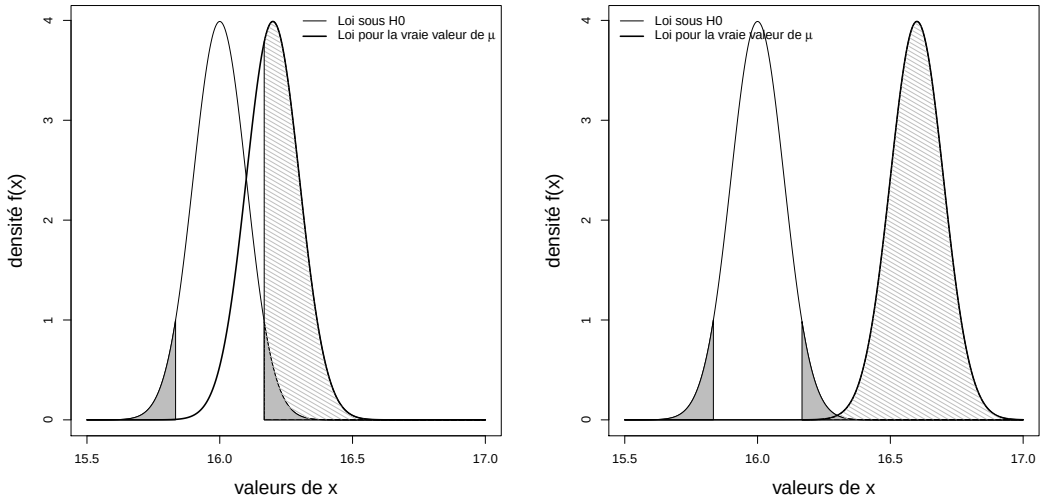


FIGURE 2.7 – **Puissance d'un test statistique.** En trait fin, distribution de X sous H0 ($\mu = 16$). En trait gras, deux exemples de de distribution de X pour la vraie valeur de μ ($\mu = 16,2$ à gauche, et $\mu = 16,7$ à droite). Pour ces tests, l'hypothèse alternative est H1 : $\mu \neq 16$. Sur chaque graphique, l'aire des zones en grisé est le niveau du test. Ces zones correspondent à l'aire sous la densité sous H0 dans la zone de rejet de H0. L'aire hachurée sous la densité pour la vraie valeur de μ est la puissance du test. On voit que la puissance augmente quand la vraie valeur de μ s'écarte de la moyenne supposée sous H0. On parle de "fonction puissance" car la puissance dépend de μ dont la valeur est inconnue.

De façon plus générale, on peut remarquer que la statistique de ce test sur la moyenne peut s'écrire :

$$Z = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{\sigma}$$

Comme on l'a vu précédemment, on s'attend sous H0 à ce que Z fluctue autour zéro. Si on est sous H1, alors l'espérance de Z est

$$E(Z) = \sqrt{n}\frac{(\mu - \mu_{\text{H0}})}{\sigma} \neq 0.$$

Ainsi, la puissance du test sera d'autant plus forte que μ est différent de μ_{H0} (figure 2.7).
Cependant, pour une valeur donnée de μ, cette valeur moyenne sera d'autant plus grande que la taille de l'échantillon n est grande, et que la variance σ^2 est petite. L'expérimentateur ne peut pas contrôler μ. *Par contre, il peut choisir d'augmenter la puissance du test en augmentant la taille de l'échantillon ou en réalisant des expériences dans les conditions les plus homogènes possibles, pour diminuer la variance.* Un graphe de puissance montre la relation entre la taille de l'échantillon et la puissance du test, pour différentes valeurs fixées des autres paramètres (figure 2.8).
La planification expérimentale est une sous-discipline des statistiques qui permet de proposer, en fonction de la question posée et des contraintes liées à l'expérimentation, les plans d'expériences qui maximisent la puissance des tests.

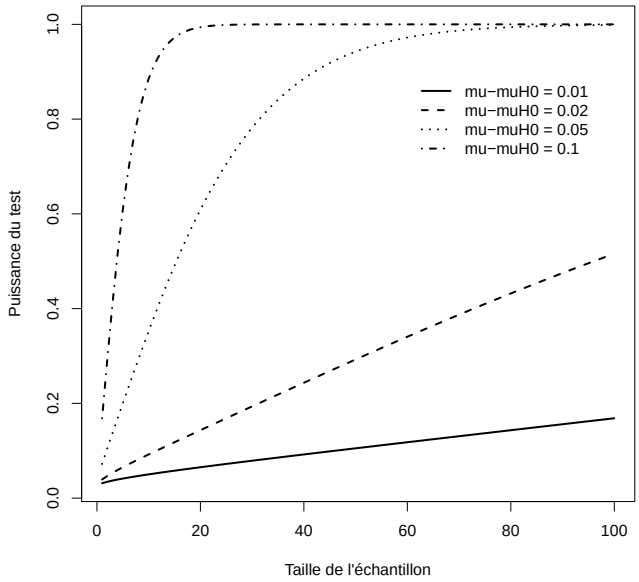


FIGURE 2.8 – **Courbes de puissance.** On représente la puissance du test en fonction de la taille de l'échantillon pour différentes valeurs de la différence de moyenne $(\mu - \mu_{\text{H0}})$, pour $\sigma = 0,1$. Avec un échantillon de taille $n = 20$, on est quasiment sûr de trouver des différences de 0,1 unités. Il faut un échantillon de taille $n = 80$ pour avoir la même chance de détecter des différences de moyenne de 0,05, alors qu'une différence de 0,01 ne sera détectée que dans 10% des cas environ.

2.2.2.2 Test unilatéral contre test bilatéral

Dans notre exemple, on pourrait avoir une hypothèse alternative qui ne soit pas une différence mais une inégalité :
H0 : en moyenne, la distance focale mesurée est de 16 cm
H1 : en moyenne, la distance focale mesurée est supérieure à 16 cm

Il existe donc deux façons de poser l'hypothèse H1 :

— **Test bilatéral :** lorsque l'on n'a pas d'*a priori* sur la valeur du paramètre, on choisit

$$\text{H1} : \mu \neq \mu_{\text{H0}}.$$

Dans ce cas, on rejettera H0 à la fois pour des valeurs positives et négatives de la statistique de test.

— **Test unilatéral :** lorsque l'on sait *a priori* que la valeur du paramètre ne peut être que plus grande, ou alors que plus petite, que la valeur choisie sous H0. On définit alors

$$\text{H1} : \mu > \mu_{\text{H0}}$$

dans le premier cas, et

$$\text{H1} : \mu < \mu_{\text{H0}}$$

dans le deuxième cas. On rejettera H0 pour des valeurs positives et négatives, respectivement, de la statistique de test.

Il faut noter que quelle que soit la forme de H1, l'hypothèse H0 est toujours la même et consiste à proposer une ou des valeurs numériques pour les paramètres du modèle. Ainsi, dans un test unilatéral à droite (H1 : $\mu > \mu_{\text{H0}}$), il n'est pas possible de tester l'hypothèse H0 $\mu < \mu_{\text{H0}}$ car on ne saurait pas chiffrer la statistique de test. On choisit donc l'hypothèse H0 la plus « défavorable », c'est-à-dire celle qui a le moins de chances d'être rejetée, à savoir $\mu = \mu_{\text{H0}}$.

Le choix de l'hypothèse H1 définit la forme de la zone de rejet du test. Certains tests n'existent qu'avec une seule formulation possible de l'hypothèse H1. (figure 2.9).

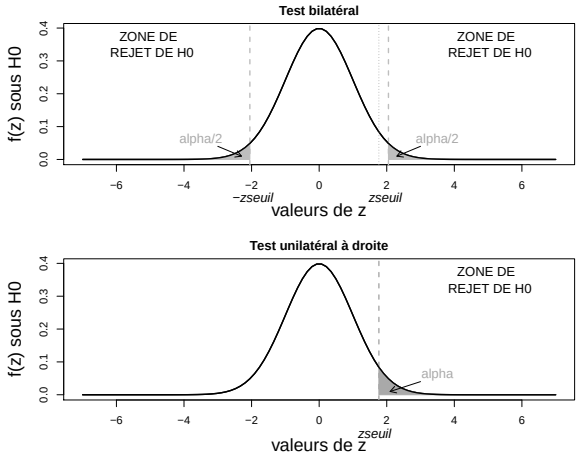


FIGURE 2.9 – **Zone de rejet dans le cas d'un test bilatéral et d'un test unilatéral (à droite).** La courbe noire est la fonction de densité de Z sous H0. Les aires grisées correspondent au risque α. Pour chaque type de test, les traits pointillés montrent la limite de la zone de rejet. Pour le test unilatéral, la zone de rejet se trouve uniquement à droite et la limite zseuil est située légèrement à gauche de la limite supérieure pour le test bilatéral.

2.2.2.3 Démarche à suivre pour réaliser un test statistique

Pour résumer, nous venons de voir, à l'aide d'un exemple, la démarche générale à suivre pour la réalisation de n'importe quel test statistique. Elle peut se décomposer en sept étapes :

1. Poser un modèle : quelle(s) variable(s) est(sont) étudiée(s) ? Quelle est sa loi (quelles sont leurs lois) ? Comment traduire la question en termes de paramètres du modèle ?
2. Formulation des hypothèses H0 et H1.
3. Choix d'une statistique de test et détermination de sa loi sous H0.
4. Choix du risque de première espèce α (appelé aussi niveau) et définition de la zone de rejet
5. Calcul de la valeur observée de la statistique
6. Calcul de la p-value
7. Conclusion statistique (rejet ou non rejet de H0) et biologique (réponse à la question posée).

Les seules étapes qui nécessitent des connaissances au-delà d'un niveau de base en statistiques sont les étapes 1 et 3.
L'étape de modélisation consiste à déterminer les paramètres qui influencent la loi de distribution des observations. Nous allons voir dans ce cours un certain nombre de cas standard qui pourront vous servir dans la plupart des cas que vous rencontrerez. De façon plus générale, l'étape de modélisation consiste à savoir décrire comment les données expérimentales sont produites. Si vous êtes capable de simuler une expérience, vous serez capable de la modéliser.
L'étape de choix d'une statistique de test est un domaine de recherche en soi. Il faut trouver **une variable aléatoire qui résume les données expérimentales et qui peut se calculer uniquement à partir de ces données, dont la loi sous H0 est connue**. Le choix d'une statistique de test dépend intimement du modèle. Ici encore, nous verrons dans le cours les statistiques de tests appropriées pour les modèles les plus courants.

2.3 Tests multiples

2.3.1 Le problème lié à la réalisation de tests multiples

Il arrive souvent qu'à partir d'un même échantillon on réalise le même type de test statistique, avec la même hypothèse H0, sur un grand nombre de variables aléatoires différentes mesurées sur les mêmes individus ou que l'on fasse plusieurs tests sur une même variable observées sur un grand nombre d'échantillons.

Exemples

- On réalise un scan génomique dans un échantillon d'une population de lignées recombinantes pour déterminer s'il y a de la distorsion de ségrégation dans la population. On dispose, pour un grand nombre ($K = 1500$) de positions dans le génome (par exemple des SNP) de la fréquence observée des deux allèles parentaux dans l'échantillon. Pour chacune des K positions, on va tester l'hypothèse H0 qu'il n'y a pas de distorsion de ségrégation, c'est-à-dire que la fréquence de l'allèle du parent 1 est 0,5. On réalise 1500 tests statistiques à partir du même échantillon de lignées.
- L'enquête santé de l'INSEE collecte des variables biométriques (taille, poids, sexe) mais aussi des indicateurs sociaux (salaire, catégorie socio-professionelle, métier). L'inventaire des métiers comprend $n = 125$ catégories de métiers différents. On cherche à savoir s'il est possible de classer les métiers selon la taille moyenne des travailleurs. Pour cela, on va comparer entre elles les tailles moyennes pour chacun des métiers, c'est-à-dire que l'on va faire $K = n(n-1)/2$ comparaisons 2 à 2.

Lorsque l'on réalise K tests statistiques au niveau α, on s'attend à rejeter H0 à tort (faux positif) avec une probabilité α pour chaque test.

Nombre de faux positifs : Dans le cas de K tests au niveau α, V est le nombre de cas de rejets à tort de l'hypothèse H0. V est une variable aléatoire. Si H0 est vraie pour tous les tests, alors $V \sim \mathcal{B}(K; \alpha)$. Si K est grand, on peut avoir un nombre assez important de faux positifs. Par exemple, si l'on réalise 100 tests au seuil 5%, on s'attend en moyenne à $100 \times 0,05 = 5$ faux positifs (si H0 est vraie pour tous les tests).

2.3.2 Des méthodes pour contrôler le risque α global ou le taux de faux positifs

2.3.2.1 Correction de Bonferroni

On peut remarquer que si l'on diminue le seuil du test, le nombre de faux positifs va forcément diminuer. Ainsi, si l'on réalise 100 tests au seuil 5‰ et non plus 5%, et que l'hypothèse H0 est vraie pour tous les tests, on s'attend en moyenne à $100 \times 0,005 = 0,5$, soit zéro ou un faux positif.

Correction de Bonferroni : : *la correction de Bonferroni dans le cas de K tests multiples consiste à réaliser chaque test au seuil α/K.*

A noter que lorsque le seuil d'un test diminue, il devient plus difficile de rejeter l'hypothèse H0, même si elle est fausse. En ce sens, la correction de Bonferroni est très conservatrice.

2.3.2.2 *False Discovery Rate*

Le raisonnement utilisé dans la correction de Bonferroni est que l'hypothèse H0 est vraie pour l'ensemble des tests. Dans la réalité, l'hypothèse H0 peut être fausse pour certains des tests. Il faut alors considérer quatre cas possibles :

- **Faux positif :** V est le nombre de cas où H0 est rejetée lorsqu'elle est vraie,
- **Faux négatif :** T est le nombre de cas où H0 est conservée lorsqu'elle est fausse,
- **Vrai positif :** S est le nombre de cas où H0 est rejetée lorsqu'elle est fausse,
- **Vrai négatif :** U est le nombre de cas où H0 est conservée lorsqu'elle est vraie,

avec $K = V + T + S + U$. Le taux de rejet est $\frac{V+S}{K}$ et les hypothèses rejetées comprennent des faux positifs et des vrais positifs.

False Discovery Rate : *le taux de fausses découvertes, $FDR = V/(V + S)$, est la proportion de cas où H0 est rejetée à tort parmi les cas où H0 est rejetée.*

Loi de probabilité de la *p*-value d'un test sous H0 On rappelle qu'une **statistique de test** est une variable aléatoire associée à un n-échantillon, qui se calcule à partir des variables X_i du n-échantillon. Par exemple, on a vu que, si X est une variable aléatoire gaussienne d'espérance μ et de variance σ^2, alors la statistique $Z = \frac{\overline{X}-\mu_{\text{H0}}}{\sigma/\sqrt{n}}$ suit une loi $\mathcal{N}(0; 1)$ sous H0.
Un test statistique consiste à définir des hypothèses H0 et H1 de façon à ce que la statistique de test :

— puisse se calculer numériquement à partir d'une réalisation du n-échantillon,
— suive une loi de paramètres connus sous l'hypothèse H0.

Dans l'exemple considéré jusqu'à présent (test de conformité sur la moyenne avec variance connue), on connaît μ_{H0} et σ, et l'on peut calculer

$$z_{\text{obs}} = \frac{\overline{x} - \mu_{\text{H0}}}{\sigma/\sqrt{n}}$$

à partir de l'échantillon. On peut aussi calculer la p-value associée à l'échantillon :

$$p_{\text{obs}} = P_{\text{H0}}(|Z| > |z_{\text{obs}}|).$$

Si l'on réalise une autre expérience, on trouvera une valeur différente de z_{obs} et de p_{obs}. Ainsi, la *p-value associée à un test statistique est une variable aléatoire*, que l'on peut appeler P.
La loi de probabilité de P se calcule facilement. Le support de P est l'intervalle $[0, 1]$ car P est une probabilité. Par ailleurs, si l'on connaît la loi de la statistique de test, on peut calculer, pour toute valeur de z, la probabilité

$$p = P_{\text{H0}}(|Z| > |z|).$$

On peut en déduire la fonction de répartition de P en calculant

$$F(p) = P(P \leq p) = P_{\text{H0}}(|Z| > |z|) = p,$$

et en déduire la fonction de densité :

$$f(p) = 1$$

en utilisant $F(p) = \int_0^p f(p)dp$.
Ainsi, la loi de probabilité de la p-value d'un test statistique est une loi uniforme $\mathcal{U}(0, 1)$.
On retrouve bien, en utilisant la loi de P, la notion de risque statistique. En effet, si l'on fait un test au niveau α, on conserve H0 si $p > \alpha$, et on rejette H0 si $p \leq \alpha$. Donc, la probabilité de prendre la mauvaise décision si H0 est vraie (faux positif) est $P_{\text{H0}}(P \leq \alpha) = F(\alpha) = \alpha$.

Le FDR peut être calculé *a posteriori*, à partir de la distribution des p-values de K tests réalisés, en prenant en compte que la distribution attendue des p-values sous H0 est une distribution uniforme. Par contre, si H1 est vraie, on s'attend à un excès de valeurs très faibles de la p-value. La distribution observée des p-values résulte donc d'un mélange de plusieurs distributions : la distribution des tests sous H0 (uniforme) et les distributions des tests sous H1 (dépendant de la valeur de H1 mais avec un excès de faibles valeurs). α_{FDR} est le niveau auquel réaliser chaque test pour s'assurer, globalement, d'une valeur de FDR donnée. La plupart des logiciels de statistiques proposent aujourd'hui le calcul de α_{FDR} à partir de la distribution observée des p-values des K tests réalisés.

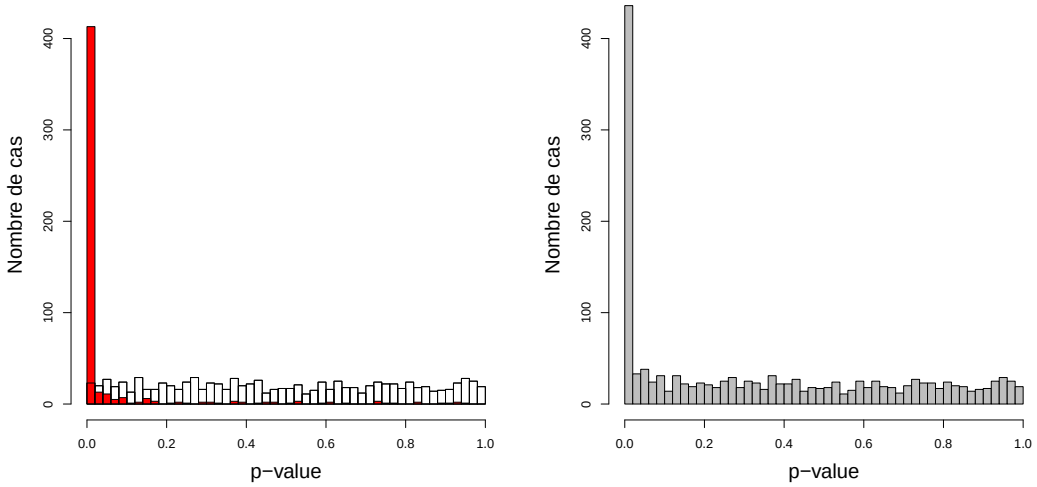


FIGURE 2.10 – **Principe du calcul du *FDR*.** Pour illustrer le principe du calcul, on reprend l'exemple du test présenté dans le chapitre 2.2.1. On effectue ce test pour 1500 machines. Pour chaque machine on fait $n = 30$ mesures. On réalise un test bilatéral donc H0 : $\mu = 16$ contre H1 : $\mu \neq 16$. On va simuler un jeu de données, en considérant que 1000 machines sont correctement réglées donc sous H0 ($\mu = 16$) et 500 machines ne sont pas correctement réglées donc sous H1($\mu \neq 16$, on prend une moyenne entre 16,01 et 16,2). **A gauche :** distributions des p-values pour les machines sous H0 (en blanc) et sous H1 (en rouge). **A droite :** distribution des p-values pour l'ensemble des machines testées. Au niveau 5%, le FDR calculé est de 11,5%, c'est à dire que 11,5% des machines pour lesquels on a rejeté H0 sont des faux positifs. Au niveau 1%, on trouve un FDR de 3,4%. Si l'on veut appliquer la correction de Bonferroni, il faut appliquer à chaque test un niveau de 0,00003 pour un niveau global de 5%. Dans ce cas on ne trouve aucun faux positif, ce niveau est un trop "stringent" car le calcul fait pour la correction de Bonferroni suppose que toutes les machines sont bien réglées ce qui n'est pas le cas.

Chapitre 3

Les tests de conformité

Lorsque l'on s'intéresse à la distribution d'une variable aléatoire dans une population, il arrive que l'on se demande si **un des paramètres de la loi de distribution de la variable aléatoire est égal à une certaine valeur**. Le test statistique est alors appelé de façon générale un **test de conformité**, et change selon le type de variable étudiée et la question posée.
Voici quelques exemples de cas :

- On souhaite vérifier si un appareil de mesure de la longueur focale de lentilles est bien réglé. Le constructeur indique que la précision de la mesure est telle que $\sigma = 0,1$. On a un étalon dont on connaît la distance focale, qui vaut 16 cm. On réalise $n = 10$ mesures de longueur focale de l'étalon à l'aide de l'appareil. Dans ce cas, on étudie une v.a quantitative dont on connaît la variance mais pas la moyenne. Sous certaines conditions, on pourra appliquer un test gaussien décrit au chapitre précédent.
- Les drosophiles de la souche sauvage ont les ailes longues. On dispose d'une souche pure mutante, appelée *miniature*, dont les individus possèdent des ailes de petite taille. Le croisement entre des mouches de la souche *miniature* et des mouches de la souche sauvage donne des descendants F1 de phénotype sauvage, quel que soit le sens du croisement. Si le caractère est contrôlé par un locus dont l'allèle dominant est *aile longue*, on s'attend, lorsque l'on croise entre elles les mouches F1, à 3/4 d'individus de phénotype sauvage et à 1/4 d'individus *miniature* dans la descendance. On étudie une v.a qualitative. Sous certaines conditions, on pourra appliquer le test de conformité du $\mathcal{X}^2$ (chi-deux).
- On veut connaître l'effet d'une injection d'insuline sur la glycémie de patients diabétiques. On mesure, sur un échantillon de patients, la différence de glycémie avant et deux heures après l'injection. Si le traitement n'a pas d'effet, alors en moyenne, la différence sera nulle. Sous certaines conditions, on pourra utiliser le test de conformité de Student.

Le choix de la statistique de test dépend de la nature de la variable aléatoire étudiée.

3.1 Variables aléatoires discrètes : conformité à une loi connue

3.1.1 Le test du $\mathcal{X}^2$ (chi-deux) de conformité

3.1.1.1 Exemple

On génotype un échantillon de 300 lignées recombinantes de blé, issues au départ du croisement entre deux lignées pures, pour un ensemble de marqueurs microsatellites. Les lignées recombinantes sont issues d'une dizaine de générations d'autofécondations à partir du produit du croisement initial, ce qui les rend quasiment homozygotes. On attend donc une ségrégation 1 : 1 dans la descendance pour chaque marqueur. Un écart à la ségrégation 1 : 1 est appelé une distorsion de ségrégation, qui peut provenir de biais lors de la méiose ou alors d'un effet de sélection (l'allèle d'un des parents ayant un avantage sélectif sur l'allèle de l'autre parent). On appelle a_1 l'allèle du parent femelle du croisement initial, et a_2 l'allèle du parent mâle du croisement initial. Les données récoltées sont sous la forme d'un tableau de comptage :

Génotype	a_1a_1	a_2a_2	Total
Nombre de lignées	n_1	n_2	n

1. **Modèle.** On appelle X la variable aléatoire associée à un locus microsatellite d'une lignée de la population recombinante. Comme il n'y a que deux lignées parentales de départ, le support de X est $\{a_1, a_2\}$, où a_1 et a_2 sont les allèles portés par chacun des deux parents. La loi de probabilité de X est une loi discrète, décrite par :

$$P(X = a_1) = p \quad ; \quad P(X = a_2) = q = 1 - p.$$

En l'absence de distorsion de ségrégation, on s'attend à $p = 0,5$. En présence de distorsion de ségrégation, on s'attend à $p \neq 0,5$, mais la valeur de p dépend du type de mécanisme en jeu et n'est pas connue. On considère un n-échantillon $\{X_1, ..., X_n\}$ correspondant au tirage au hasard de n lignées recombinantes.

Si l'on appelle Y_{a_1} la variable aléatoire qui mesure le nombre de fois où $X_i = a_1$ dans l'échantillon, on sait que cette variable aléatoire suit une loi Binomiale $\mathcal{B}(n; p)$, et que $E(Y_{a_1}) = np$. On peut dire que l'effectif *théorique* de la classe a_1 est $m_1 = np$. Avec le même raisonnement, on montre que l'effectif théorique de la classe a_2 est $m_2 = nq$.

On dispose d'une réalisation d'un n-échantillon $\{x_1, ..., x_n\}$ de X.

2. **Hypothèses H0 et H1.** On choisit une hypothèse H0 chiffrable.
   H0 : $p = 0,5$
   H1 : $p \neq 0,5$

3. **Choix d'une statistique de test.** La statistique de test est la statistique du $\mathcal{X}^2$ (prononcer khi-deux), qui dépend de la différence entre les effectifs attendus sous H0 et les effectifs observés. Dans cet exemple, il n'y a que deux modalités. La statistique de test est une somme de deux termes :

$$Z = \sum_{j=1}^{2} \frac{(n_j - m_j)^2}{m_j}. \tag{3.1}$$

   Sous l'hypothèse H0, on s'attend à de petites valeurs de Z, c'est-à-dire à ce que les effectifs observés soient proches des effectifs théoriques. Sous H1, on s'attend à trouver plus de différences entre effectifs observés et théoriques, et donc les valeurs de Z seront plutôt plus grandes que sous H0. A noter que, d'après l'équation 3.1, Z est toujours positive.

   Sous H0, la loi de Z est une loi du $\mathcal{X}^2$ dont le paramètre est le nombre de degrés de liberté. **Dans le cas d'un test de conformité, le nombre de degrés de liberté est le nombre de classes de la variable qualitative étudiée moins 1**. Dans cet exemple, il y a deux classes (a_1 et a_2), donc un seul degré de liberté :

$$Z \sim_{\text{H0}} \mathcal{X}^2_1.$$

   Attention, la loi du $\mathcal{X}^2$ pour la statistique Z est une loi asymptotique. En d'autres termes, la loi de Z sous H0 se rapproche d'une loi du $\mathcal{X}^2$ quand les effectifs *théoriques* des classes sont suffisamment grands. On considère cette approximation comme valable pour des effectifs théoriques tous supérieurs ou égaux à 5. On peut donc appliquer ce test uniquement si tous les effectifs théoriques sont supérieurs ou égaux à 5.

4. **Zone de rejet.** D'après l'expression de Z (équation 3.1), tout écart entre les effectifs observés et les effectifs théoriques va augmenter la valeur de Z. On rejette donc H0 pour de trop grandes valeurs de Z. La zone de rejet est de la forme :

$$[z_{seuil}; +\infty[$$

   On veut trouver la valeur de z_{seuil} telle que $P_{H0}(Z > z_{seuil}) = \alpha$ donc z_{seuil} est le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{X}^2_1$.

5. **Réalisation du test.** Les valeurs de z_{seuil} sont tabulées pour n'importe quel niveau α. La valeur de z_{obs} se calcule en utilisant l'équation 3.1 dans laquelle les effectifs théoriques des classes sont calculés sous H0, et la p-value est calculée en utilisant la fonction de répartition :

$$pval = 1 - F_{\mathcal{X}^2_1}(z_{\text{obs}}).$$

   Fonction R : `chisq.test(x, p=prob)`, où `x` est le vecteur des n_i et `prob` le vecteur des probabilités sous H0.

Validation des conditions d'application. On vérifie *a posteriori* que tous les effectifs théoriques m_j sont supérieurs ou égaux à cinq. A noter que si une ou plusieurs classes ont un effectif théorique inférieur à cinq, il est possible de refaire le test en fusionnant des classes et réduisant ainsi le nombre de modalités et de degrés de liberté.

3.1.1.2 Cas général

On considère un n-échantillon d'une variable aléatoire X de support $\{a_1, ..., a_J\}$, dont la loi de probabilité est connue sous H0 :

$$P_{\text{H0}}(X = a_j) = p_j$$

Sous H0, les effectifs théoriques de la classe j peuvent se calculer :

$$m_j = np_j$$

et la statistique de test

$$Z = \sum_{j=1}^{J} \frac{(n_j - m_j)^2}{m_j}$$

suit approximativement une loi $\mathcal{X}^2_{J-1}$ sous l'hypothèse H0, pourvu que les effectifs théoriques soient suffisamment grands ($m_j \geq 5$).
Dans le cas d'un test de conformité, les degrés de liberté sont le nombre de classes moins 1. En effet, les effectifs observés n_j de chaque classe sont reliés entre eux par la relation $\sum_{j=1}^{J} n_j = n$. De même pour les effectifs théoriques. Ainsi, la statistique de test est une somme de J termes, mais le $J^{\text{ième}}$ terme de la somme se déduit des autres termes.

3.1.2 Test exact de Fisher

Lorsque les conditions d'application du test du $\mathcal{X}^2$ ne sont pas réunies (une classe au moins ayant un effectif théorique < 5) et que l'on ne souhaite pas faire de regroupement de classes, il est possible de faire un test exact de Fisher.
Le principe du test consiste à simuler des données sous l'hypothèse H0 et à comparer la statistique observée z_{obs} à la distribution de Z simulée sous H0. Dans le cas d'un test de conformité, on simule T tirages de n-échantillons dans la loi de probabilité de X sous H0. Pour chaque tirage t, on calcule la valeur z_t de Z. On estime ensuite la p-value du test par la proportion de cas où $z_t > z_{\text{obs}}$:

$$pval = \frac{\text{Nombre de cas où } z_t > z_{\text{obs}}}{T}$$

Plus le nombre de tirages est élevé, plus la distribution simulée sous H0 sera proche de la *vraie* distribution de Z sous H0. Dans la pratique, on peut prendre $T = 2000$.

3.2 Variables aléatoires continues : tests sur la moyenne

On considère un n-échantillon $X_1, ..., X_n$ d'une variable aléatoire continue X. On va considérer deux cas de figure :

1. **Cas gaussien** Les variables aléatoires X_i sont indépendantes, gaussiennes de même loi $\mathcal{N}(\mu; \sigma^2)$. Dans ce cas, la moyenne empirique $\bar{X}$ suit une loi normale $\bar{X} \sim \mathcal{N}(\mu; \frac{\sigma^2}{n})$
2. **Cas des grands échantillons** Les variables aléatoires X_i sont indépendantes, de même loi, de moyenne μ et variance σ^2. On ne fait pas d'hypothèse sur la loi des X_i mais la taille de l'échantillon est suffisamment grande ($n \geq 30$) pour que le théorème central limite s'applique. Dans ce cas $\bar{X} \approx \mathcal{N}(\mu; \frac{\sigma^2}{n})$

On veut savoir si la moyenne μ est égale à une valeur de référence connue μ_{H0}. Selon l'a priori que l'on a, on peut choisir l'un de ces trois tests :
H0 : $\mu = \mu_{\text{H0}}$ contre H1 : $\mu \neq \mu_{\text{H0}}$
ou
H0 : $\mu = \mu_{\text{H0}}$ contre H1 : $\mu < \mu_{\text{H0}}$
ou
H0 : $\mu = \mu_{\text{H0}}$ contre H1 : $\mu > \mu_{\text{H0}}$

Attention, le choix de l'hypothèse H1 ne doit pas être fondé sur les valeurs de l'échantillon mais sur ce que vous souhaitez tester, et qui résulte d'études antérieures. De même, H0 doit être chiffrable donc μ_{H0} doit avoir une valeur numérique.

3.2.1 Cas gaussien, variance connue

Si les X_i suivent une loi normale $\mathcal{N}(\mu; \sigma^2)$ et que l'on connaît σ, la statistique de test Z ci-dessous suit sous H0 une loi normale centrée réduite :

$$Z = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{\sigma} \sim_{\text{H0}} \mathcal{N}(0; 1) \tag{3.2}$$

Le test statistique s'effectue comme indiqué dans l'exemple du chapitre 2. Il faut ensuite vérifier les conditions d'application en observant graphiquement que les valeurs observées ont une distribution qui ressemble à une distribution gaussienne (voir ci-dessous).

3.2.2 Cas gaussien, variance inconnue : le test de Student

Si les X_i suivent une loi normale $\mathcal{N}(\mu; \sigma^2)$ et que l'on ne connaît pas la valeur de σdans la population, ce qui correspond à la plupart des cas, on peut construire une statistique de test en remplaçant σ dans l'équation 3.2 par son estimateur S_{n-1}.
Dans ce cas, la statistique

$$T = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{S_{n-1}} \sim_{\text{H0}} \mathcal{T}_{n-1} \tag{3.3}$$

suit une loi de Student à $n-1$ degrés de libertés sous l'hypothèse H0. Ici, le degré de liberté en moins vient du fait que l'on a remplacé σ par son estimateur S_{n-1}. La loi de Student est une loi symétrique en zéro. Elle tend asympotiquement vers une loi normale $\mathcal{N}(0; 1)$ quand le nombre de degrés de liberté augmente. Dans la pratique, on peut utiliser la loi $\mathcal{N}(0; 1)$ à la place de la loi $\mathcal{T}_{\text{ddl}}$ lorsque $ddl \geq 30$.
Le reste du test statistique se déroule de la même façon que précédemment. On appelle t_{obs} la valeur de la statistique de test dans l'échantillon.

Zone de rejet et *p*-value. Si H0 est vraie, alors la moyenne de l'échantillon $\overline{X}$ doit être proche de la moyenne théorique μ_{H0}, et la statistique de test T sera proche de zéro. Si H0 est fausse, on s'attend à des valeurs absolues élevées pour T. La forme de la zone de rejet dépend de l'hypothèse H1.

— H1 : $\mu \neq \mu_{\text{H0}}$: on rejette H0 pour des grandes valeurs de T, positives ou négatives. La zone de rejet est constituée de deux intervalles :

$$]-\infty, -t_{\text{seuil}}] \cup [t_{\text{seuil}}, +\infty[.$$

En utilisant les propriétés de symétrie de la loi de Student, la p-value se calcule comme :

$$pval = 2(1 - F_{\mathcal{T}_{\text{ddl}}}(|t_{\text{obs}}|)).$$

— H1 : $\mu > \mu_{\text{H0}}$: on rejette H0 pour des grandes valeurs positives de T. La zone de rejet est donc l'intervalle :

$$[t_{\text{seuil}}, +\infty[.$$

La p-value se calcule comme :

$$pval = 1 - F_{\mathcal{T}_{\text{ddl}}}(t_{\text{obs}}).$$

Attention, on utilise dans ce cas t_{obs}, et non sa valeur absolue. Des valeurs négatives de t_{obs} correspondent à une p-value supérieure à 0,5 et au non rejet de H0.

— H1 : $\mu < \mu_{\text{H0}}$: on rejette H0 pour des grandes valeurs négatives de T. La zone de rejet est donc l'intervalle :

$$]-\infty, -t_{\text{seuil}}].$$

La p-value se calcule comme :

$$pval = F_{\mathcal{T}_{\text{ddl}}}(t_{\text{obs}}).$$

Attention, on utilise également dans ce cas t_{obs}, et non pas sa valeur absolue. Des valeurs positives de t_{obs} correspondent à une p-value supérieure à 0,5 et au non rejet de H0.

Validation des conditions d'application. Si la taille de l'échantillon est petite ($n < 30$), il faut vérifier que la loi de distribution de X est gaussienne. On peut utiliser pour cela une représentation graphique particulière, le graphe quantile-quantile, appelé aussi quantile-quantile plot ou Q-Q plot. On sait que si $X \sim \mathcal{N}(\mu; \sigma^2)$, alors $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0; 1)$, et les quantiles de la distribution de X sont reliés à ceux de la distribution de $\frac{X-\mu}{\sigma}$ par la relation :

$$Q_\alpha(X) = \mu + \sigma Q_\alpha\left(\frac{X-\mu}{\sigma}\right)$$

On peut donc comparer graphiquement les quantiles observés dans l'échantillon $(x_1, ..., x_n)$ aux quantiles théoriques d'une loi normale $\mathcal{N}(0; 1)$ (figure 3.1). On vérifie que les points du Q-Q plot s'alignent bien sur une droite, ce qui est à peu près le cas ici.

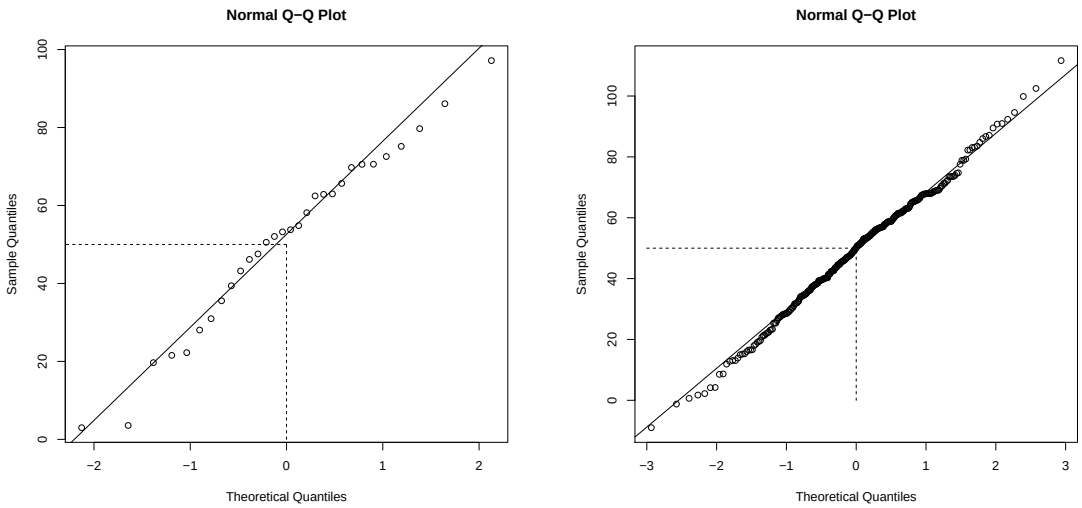


FIGURE 3.1 – **Quantile-quantile plots.** Les deux graphiques ont été réalisés à partir de données simulées avec $X \sim \mathcal{N}(\mu = 50; \sigma^2 = 20^2)$. Deux tailles d'échantillon sont utilisées, $n = 30$ (à gauche) et $n = 300$ (à droite). Les lignes pointillées correspondent aux moyennes théoriques (0 en abscisse et 50 en ordonnée). La pente des droites est égale à l'écart-type théorique ($\sigma = 20$).

3.2.3 Cas non gaussien, n grand

Même si les X_i ne suivent pas une loi normale, le théorème centrale limite garantit que si la taille d'échantillon est suffisamment grande ($n > 30$), alors la moyenne $\overline{X}$ suit, de façon approchée, une loi normale $\mathcal{N}(\mu; \frac{\sigma^2}{n})$. De plus, la loi des grands nombres garantit que l'estimateur S^2 est très proche de σ^2. Ainsi, que l'on connaisse σ^2 ou non, la statistique

$$T = \sqrt{n}\frac{(\overline{X} - \mu_{\text{H0}})}{S_{n-1}} \approx_{\text{H0}} \mathcal{N}(0, 1) \tag{3.4}$$

suit approximativement une loi normale $\mathcal{N}(0,1)$. Le test statistique se déroule comme précédemment, en utilisant la loi normale $\mathcal{N}(0,1)$ pour le calcul du seuil et de la p-value. La seule condition d'application est la taille de l'échantillon, qui doit être suffisamment grande.

3.2.4 Cas non gaussien, n petit

Si l'on a peu d'observations, il faudra réaliser un test non paramétrique (voir section 3.5).

3.3 Données appariées

Il peut arriver que l'on veuille comparer deux variables aléatoires correspondant à des caractéristiques différentes d'un même individu d'une population. Par exemple :

- On souhaite comparer l'efficacité de deux pommades hydratantes A et B. On applique chez des patients une pommade sur la main droite, et l'autre sur la main gauche et on mesure le résultat au bout de quelques heures.
- On cherche à comparer deux méthodes de prélèvements (prise de sang et tests salivaires) pour estimer la prévalence du vecteur du paludisme. On applique les deux méthodes sur chaque patient d'un hôpital et on compare les prévalences estimées par chaque méthode.
- On se demande si les pattes avant et arrière des chevreuils ont la même longueur. On mesure la différence de longueur entre pattes avant et arrière sur un échantillon de 354 chevreuils de la forêt de Fontainebleau pour déterminer si en moyenne la longueur des pattes avant et arrière est la même.
- Pour savoir si un médicament permettant de traiter le cholesterol est efficace, on mesure de taux de cholestérol de 60 patients avant et après le traitement.

Modèle. On considère un n-échantillon d'un couple de variables aléatoires $\{(X_{1i}, X_{2i}), \quad i = 1, ..., n\}$. On se demande si X_1 et X_2 ont la même moyenne. On calcule la différence

$$Y_i = X_{1i} - X_{2i},$$

et on s'intéresse à la variable aléatoire $\overline{Y} = \frac{\sum_i Y_i}{n}$. Si Y suit une loi normale alors $\overline{Y} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$. Si Y ne suit pas une loi normale mais que la taille de l'échantillon est suffisamment grande ($n \geq 30$), alors $\overline{Y} \approx \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$. Ici μ est l'espérance des Y_i et σ^2 la variance dans la population. Si le traitement n'a pas d'effet, on s'attend à ce que les Y_i soient distribués autour de zéro, et en particulier à ce que la moyenne soit égale à zéro : $\mu = 0$. Par contre, on ne connaît pas forcément la variance σ^2, que l'on peut estimer par S^2_{n-1} dans l'échantillon.

Hypothèses H0 et H1. On pose les hypothèses suivantes :
H0 : $\mu = 0$
H1 : $\mu \neq 0$
ou bien, selon la question posée,
H1 : $\mu > 0$ ou $\mu < 0$

Statistique de test : on est donc ramené à un test de conformité sur la moyenne. Si la variance de Y, σ^2, est inconnue, la statistique de test est de la forme

$$T = \sqrt{n}\frac{\overline{Y}}{S_{n-1}}.$$

Il y a deux cas possibles pour la loi de la statistique de test sous H0 :

- Y suit une loi normale. Sous H0, $T \sim_{\text{H0}} \mathcal{T}_{n-1}$
- n est suffisamment grand. Sous H0, $T \sim_{\text{H0}} \mathcal{N}(0,1)$

3.4 Intervalle de confiance sur la moyenne

3.4.1 Définition

Problématique. On considère un n-échantillon d'une variable aléatoire $X_1, ..., X_n$ qui décrit une population. On ne connaît ni la moyenne μ, ni la variance σ^2 de X dans la population. On peut utiliser la loi de probabilité des X_i pour proposer un intervalle probable pour la moyenne μ inconnue, sachant les observations.

Intervalle de confiance : *c'est un intervalle contenant, avec une certaine probabilité fixée au préalable, la vraie valeur d'un paramètre. Ainsi, un intervalle de confiance au risque α contient la valeur inconnue du paramètre avec une probabilité de $1 - \alpha$.*

3.4.2 Calcul de l'intervalle de confiance d'une moyenne

Estimateur de la moyenne. On peut estimer la moyenne inconnue par la moyenne empirique de l'échantillon :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Si X suit une loi normale, alors la variable aléatoire

$$T = \frac{\overline{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} \tag{3.5}$$

suit une loi de Student à $n - 1$ degrés de liberté, et on peut calculer la valeur t_α telle que

$$P_{\mathrm{H0}}\left(-t_\alpha \leq T \leq t_\alpha\right) = 1 - \alpha.$$

En utilisant la relation 3.5, on peut écrire :

$$P_{\mathrm{H0}}\left(-t_\alpha \leq \frac{\overline{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}} \leq t_\alpha\right) = 1 - \alpha,$$

et l'on voit qu'il est possible d'encadrer la valeur de μ inconnue par deux valeurs qui dépendent de $\overline{X}$, S^2_{n-1} et t_α :

$$P_{\mathrm{H0}}\left(-t_\alpha \frac{S_{n-1}}{\sqrt{n}} + \overline{X} \leq \mu \leq t_\alpha \frac{S_{n-1}}{\sqrt{n}} + \overline{X}\right) = 1 - \alpha.$$

Dans cet exemple, l'intervalle de confiance pour μ au rique α est donc :

$$IC_{1-\alpha} = \left]\overline{X} - t_\alpha \frac{S_{n-1}}{\sqrt{n}}, \overline{X} + t_\alpha \frac{S_{n-1}}{\sqrt{n}}\right[. \tag{3.6}$$

On rappelle que S^2_{n-1} est l'estimateur de la variance σ^2 de la population, et qu'il se calcule de la façon suivante :

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Il peut être pratique de se rappeler aussi que lorsque n devient grand, la valeur $t_{0,05}$ tend vers 1,96. On a donc :

$$IC_{0,95} = \left]\overline{X} - 1,96\frac{S_{n-1}}{\sqrt{n}}, \overline{X} + 1,96\frac{S_{n-1}}{\sqrt{n}}\right[.$$

Dans la pratique, l'intervalle de confiance se calcule en remplaçant $\overline{X}$ et S^2_{n-1} par leurs réalisations $\overline{x}$ et s^2_{n-1} dans l'échantillon considéré.

Si X ne suit pas une loi normale, mais que l'échantillon est assez grand, on peut appliquer la même démarche en se rappelant que dans ce cas, la loi de T (défini en 3.5) se rapproche d'une une loi normale $\mathcal{N}(0,1)$.

3.5 Variables aléatoires continues : tests non paramétriques

Les tests non paramétriques ne font pas d'hypothèse sur la loi de distribution de la variable aléatoire. Leur nom vient du fait qu'on ne formule pas d'hypothèse sur les paramètres de la loi. Dans la pratique, on les utilise lorsque la loi de la variable aléatoire est inconnue, et/ou que l'effectif n'est pas suffisamment grand pour faire une approximation par une loi connue. Dans une population, on s'intéresse à un trait quantitatif auquel on associe une variable aléatoire X. On appelle $\mathcal{L}$ la loi de probabilité de X, inconnue, ayant comme densité f. On considère un n-échantillon tiré de la population $\{X_1, X_2, ..., X_n\}$.

3.5.1 Test sur la médiane

Hypothèses testées. On cherche à déterminer si la médiane de la distribution est égale à une valeur donnée m (par exemple $m = 115$). On pose les hypothèses suivantes :
H0 : la médiane de X est m, autrement dit $P(X \leq m) = 0,5$
H1 : la médiane de X est différente de m.
On pourrait aussi utiliser une formulation unilatérale pour l'hypothèse H1, soit H1 : la médiane de X est supérieure à m, soit H1 : la médiane de X est inférieure à m.

Statistique de test. Sous l'hypothèse H0, on s'attend à ce que la moitié des valeurs $\{X_1, X_2, ..., X_n\}$ soient inférieures à m. On peut considérer les variables aléatoires Y_i telles que $Y_i = 1$ si $X_i > m$ et zéro sinon. On appelle Z le nombre de X_i supérieures à m :

$$Z = \sum_{i=1}^{n} Y_i.$$

Sous l'hypothèse H0, la loi de Z est une loi binomiale $\mathcal{B}(n; 0,5)$.
NB : Pour calculer la statistique de test, on ne considérera pas les réalisations de X égales à m. Si on supprime des valeurs, il faudra en ternie compte pour calculer n.

Exemple. La lignée de maïs F252 est, entre autres, caractérisée par la médiane de la hauteur des plantes, qui est $m = 115$ cm. On dispose d'un lot de grains. On aimerait vérifier qu'il s'agit bien de grains de la lignée F252. On fait pousser un échantillon de $n = 10$ plantes, et on mesure leur hauteur après la floraison. On note X la variable aléatoire "hauteur après floraison". On va tester :
H0 : la hauteur médiane des plantes issues du lot de grains est de 115.
H1 : la hauteur médiane des plantes issues du lot de grains est différente de 115.
On trouve les valeurs suivantes pour X et Y :

Plante	x_{obs}=Hauteur (cm)	y_{obs}
1	119,68	1
2	113,22	0
3	108,49	0
4	105,11	0
5	112,52	0
6	120,18	1
7	125,15	1
8	114,13	0
9	136,34	1
10	114,11	0

Sous H0, la loi de Z est $\mathcal{B}(10; 0, 5)$. Ici, on a choisi un test bilatéral. La zone de rejet du test est symétrique par rapport à 5. Elle est de la forme $[0, 5-a] \cup [5+a, 10]$. La valeur de a est fixée selon le risque α choisi de façon à ce que $2 * P_{\text{H0}}(Z \leq 5-a) \leq \alpha$.
On trouve $z_{\text{obs}} = 4$. La p-value se calcule comme

$$pval = P_{\text{H0}}(Z \leq 4) + P_{\text{H0}}(Z \geq 6) = 2 * P_{\text{H0}}(Z \leq 4) = 2 \cdot F_{\mathcal{B}(10;0,5)}(4) = 0,754$$

On ne peut donc pas rejeter H0 au seuil $\alpha = 0,05$.
Fonction R : `2*pbinom(4,size=10,prob=0.5)`

3.5.2 Le test signé des rangs : test de Wilcoxon

Modèle. On cherche à déterminer si la densité est symétrique par rapport à une valeur donnée a. Une des façons de regarder cela est de comparer les lois de $X - a$ et de $a - X$. Pour cela, on définit deux nouvelles variables. La première,

$$R_i = \text{rang}|X_i - a|,$$

mesure de combien X_i s'écarte de a lorsque l'on considère les rangs. La seconde, S_i vaut $+1$ ou -1 selon que X_i s'écarte de a en valeurs positives ($S_i = +1$) ou négatives ($S_i = -1$).
Si $X - a$ et $a - X$ ont la même loi, alors on s'attend à trouver autant d'écarts négatifs que d'écarts positifs. En particulier, la somme des rangs négatifs doit être proche de la somme des rangs positifs.

Hypothèses H0 et H1. H0 : $X - a$ et $a - X$ ont la même loi contre H1 : $X - a$ et $a - X$ n'ont pas la même loi.

Statistique du test. On appelle T^+ la somme des rangs positifs (somme des R_i pour lesquels $S_i = +1$) et T^- la somme des rangs négatifs (somme des R_i pour lesquels $S_i = -1$). Sous l'hypothèse H0, la statistique

$$W = \min(T^+, T^-)$$

suit la loi de Wilcoxon, qui est tabulée.
On peut remarquer que si les rangs sont bien répartis de part et d'autre de a, on s'attend à ce que $T^+ = T^-$. De plus, la somme de tous les rangs $(T^+ + T^-)$ vaut :

$$1 + 2 + \quad ... \quad + n = \frac{n(n+1)}{2}.$$

Sous l'hypothèse H0, on a donc

$$E(W) = \frac{n(n+1)}{4}.$$

Exemple. On peut reprendre l'exemple précédent, avec $a = 115$:

Plante	Hauteur (cm)	$X - a$	R_i	S_i
1	119,68	+4,68	5	+1
2	113,22	–1,78	3	–1
3	108,49	–6,51	7	–1
4	105,11	–9,89	8	–1
5	112,52	–2,48	4	–1
6	120,18	+5,18	6	+1
7	125,15	+10,15	9	+1
8	114,13	–0,87	1	–1
9	136,34	+21,34	10	+1
10	114,11	–0,89	2	–1

On trouve donc $T^+_{\text{obs}} = 5 + 6 + 9 + 10 = 30$ et $T^-_{\text{obs}} = 3 + 7 + 8 + 4 + 1 + 2 = 25$, donc $w_{\text{obs}} = 25$. La p-value se calcule sous R en utilisant `2*psignrank(25,10)`. On trouve $pval = 0,846$. Donc on ne peut pas rejeter l'hypothèse H0.

3.6 Bilan

Les tests de conformité permettent de comparer un ou plusieurs paramètres d'une loi de distribution à des valeurs de référence connues. Selon la nature de la variable aléatoire et la question posée, on utilisera des tests statistiques différents.

Variable aléatoire discrète

On connaît pour chaque valeur possible $\{a_1, ..., a_J\}$ de X la probabilité $P_{\text{H0}}(X_i = a_j) = p_j$. La statistique de test se calcule à partir des écarts entre les effectifs observés dans chaque classe n_j et les effectifs attendus sous H0 (effectifs dits théoriques) $m_j = np_j$. Sous H0, cette statistique suit une loi de $\mathcal{X}^2$ (chi-deux) :

$$Z = \sum_{j=1}^{J} \frac{(n_j - m_j)^2}{m_j} \approx \mathcal{X}^2_{J-1}$$

L'approximation de la loi de Z par la loi du $\mathcal{X}^2$ n'est valide que si tous les m_j sont supérieurs ou égaux à 5. Sinon, on peut faire un test exact de Fisher.

Variable aléatoire continue, test sur la moyenne

On veut savoir si l'espérance de X dans la population est égale à une valeur connue a.

1. X **suit une loi normale**,
   — **Variance connue.** Test gaussien.
   $$Z = \sqrt{n}\frac{\overline{X} - a}{\sigma} \sim_{H0} \mathcal{N}(0; 1)$$
   — **Variance inconnue.** Test de Student.
   $$T = \sqrt{n}\frac{\overline{X} - a}{s_{n-1}} \sim_{H0} \mathcal{T}_{n-1}$$
2. X **ne suis pas une loi normale mais** $n \geq 30$,
   — Test gaussien
   $$T = \sqrt{n}\frac{\overline{X} - a}{s_{n-1}} \approx_{H0} \mathcal{N}(0, 1)$$
3. **Petits échantillons, tests non paramétriques**
   — **Test sur la médiane.** Sous H0, le nombre de valeurs de X plus grandes que la médiane a suit une loi binomiale $\mathcal{B}(n; 0.5)$.
   — **Test signé de rang.** Test de Wilcoxon.
   A noter que l'on peut toujours utiliser un test non paramétrique, mais ces tests sont moins puissants (il est plus difficile de rejeter H0). Donc si on a une variable gaussienne ou un grand échantillon, on préfèrera un test paramétrique.

Chapitre 4

Les tests d'homogénéité

Les tests d'homogénéité correspondent aux cas où l'on dispose de deux ou plusieurs échantillons, et l'on se demande si les populations dont ils sont issus ont des caractéristiques communes, sans connaître forcément les paramètres des lois de distribution de la variable aléatoire dans ces populations.
Voici quelques exemples de cas :

- Dans la population humaine, les suédois sont-ils en moyenne plus grands que les pygmées ?
- La conductance du nerf optique est-elle en moyenne plus faible chez un groupe d'individus atteints d'une inflammation du système nerveux que chez un groupe d'individus sans inflammation ?
- Existe-t-il un dimorphisme sexuel sur la longueur de la queue dans la population d'hirondelles françaises ?
- Dans une population de lignées recombinantes de maïs, les génotypes en un locus marqueur sont-ils liés à une différence de largeur des feuilles ?
- La répartition des naissances selon les mois de l'année est-elle la même dans tous les départements français ?

Ici encore, le choix de la statistique de test dépend de la nature de la variable aléatoire étudiée.

4.1 Variables aléatoires discrètes : le test du $\mathcal{X}^2$

4.1.1 Cas général

On dispose cette fois-ci de plusieurs n_i-échantillons, chacun issu d'une population différente, d'une variable aléatoire X de support $\{a_1, ..., a_j, ..., a_J\}$, et on voudrait savoir si cette variable suit la même loi dans chacune des populations. On appelle X_i la variable aléatoire associée à un tirage dans la $i^{\text{ème}}$ population. A l'issue de l'échantillonnage, on peut résumer les données par un tableau de contingence dont chaque ligne correspond à un échantillon et chaque colonne à l'une des valeurs possible de X_i :

Support	a_1	...	a_j	...	a_J	Total
Echantillon 1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
...	...	...	...	...	...	...
Echantillon i	n_{i1}	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...
Echantillon I	n_{I1}	...	n_{Ij}	...	n_{IJ}	$n_{I.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

où n_{ij} est l'effectif observé dans l'échantillon i pour la classe a_j. Les **effectifs marginaux** sont les sommes sur les lignes ou les colonnes du tableau, notés respectivement $n_{i.}$ et $n_{.j}$. La taille de l'échantillon i est donc $n_{i.}$, le nombre total d'individus appartenant à la classe a_j est $n_{.j}$ et l'effectif total est n.

1. **Modèle.** Chaque échantillon est issu d'une population. On peut donc associer une loi de probabilité à chacun des échantillons :

$$P(X_i = a_j) = p_{ij}$$

   où p_{ij} est la fréquence de la classe a_j dans la population i.

2. **Hypothèses H0 et H1.** On pose comme hypothèse H0 que les échantillons suivent tous la même loi de probabilité :

$$\text{H0} : P(X_i = a_j) = p_j$$

c'est-à-dire que la fréquence de chaque classe a_j est la même dans chaque population. On a aussi $\sum_{j=1}^{J} p_j = 1$. L'hypothèse alternative H1 est qu'il existe au moins un échantillon qui ne suit pas cette loi de probabilité :

$$\text{H1} : \quad \exists i, \quad P\left(X_i = a_j\right) = p_{ij} \neq p_j$$

3. **Statistique de test.** Sous l'hypothèse H0, on peut estimer la fréquence d'une classe a_j en utilisant les effectifs marginaux :

$$\widehat{p}_j = \frac{n_{.j}}{n}$$

Les effectifs théoriques pour la classe a_j dans l'échantillon i se calculent comme

$$m_{ij} = n_{i.}\widehat{p}_j = \frac{n_{i.}n_{.j}}{n}.$$

La statistique de test est la statistique du $\mathcal{X}^2$:

$$Z = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{4.1}$$

Sous l'hypothèse H0, la statistique de test suit approximativement une loi du $\mathcal{X}^2$, sous réserve que les effectifs théoriques soient suffisants :

$$Z \sim_{H0} \mathcal{X}^2_{(I-1)(J-1)}$$

Le nombre de **degrés de liberté** (abrégé en *ddl*) se calcule de la façon suivante : c'est le nombre de termes indépendants de la somme, c'est-à-dire le nombre de termes de la somme moins le nombre de contraintes.
Pour calculer les effectifs théoriques, on utilise toutes les sommes marginales, soit :
$\sum_j n_{ij} = n_{i.}$ ($I - 1$ sommes indépendantes)
$\sum_i n_{ij} = n_{.j}$ ($J - 1$ sommes indépendantes)
$\sum_i n_{i.} = \sum_j n_{.j} = n$ (1 somme).
Le nombre total de termes de la somme est IJ. On a donc :

$$ddl = IJ - (I-1) - (J-1) - 1 = (I-1)(J-1).$$

Validation des conditions d'application. Tout comme pour le test du $\mathcal{X}^2$ de conformité, il faut vérifier que les effectifs théoriques sont supérieurs ou égaux à 5 dans chaque classe. Si les conditions d'application ne sont pas réunies, on peut faire des regroupements de classe (si cela a un sens biologique) ou un test exact de Fisher.

4.1.2 Test exact

Comme pour les tests de conformité, on peut réaliser un test de Fisher en simulant la distribution de Z sous H0. Ce test ne sera pas détaillé ici.

4.1.3 Exemples

Voici quelques exemples de données qui se traitent par un test de $\mathcal{X}^2$ d'homogenéité.

- L'INSEE recence pour chaque mois de l'année le nombre de naissances dans chaque département. On se demande si la distribution des naissances sur l'année varie d'un département à l'autre.
- On mesure par reséquençage les fréquences d'haplotypes dans une région du génome dans différentes populations de chardons dans des champs cultivés d'Île-de-France. On se demande si les fréquences haplotypiques sont les mêmes dans toutes les populations (dispersion par les graines), ou si elles sont différentes d'un champ à l'autre (dispersion par marcottage).

— Une forme d'examen qui facilite le travail du correcteur est le questionnaire à choix multiple (QCM). Pour une question, chaque réponse a la même chance d'être choisie si les étudiants répondent au hasard, alors que la bonne réponse a plus de chances d'être cochée si les étudiants ont appris leur cours. En analysant les réponses à un QCM, on peut savoir si les étudiants ont travaillé ou non.

4.2 Variables aléatoires continues : tests sur la moyenne

On dispose de deux échantillons indépendants d'une variable aléatoire X. On a donc des n_i-échantillons et on notera X_{ij} la variable correspondant au j-ème individu de l'échantillon issu de la population i ($i = 1, 2$). Plusieurs cas de figure peuvent être considérés :

— Chaque échantillon est gaussien et de même variance :

$$X_{ij} \sim \mathcal{N}(\mu_i; \sigma^2)$$

— Chaque échantillon est gaussien avec des variances différentes :

$$X_{ij} \sim \mathcal{N}(\mu_i; \sigma_i^2) \quad \text{et} \quad \sigma_1^2 \neq \sigma_2^2$$

— La taille de chaque échantillon est suffisamment grande ($n_i \geq 30$) pour que l'on puisse appliquer le théorème central limite : la moyenne empirique $\overline{X}_i$ suit approximativement une loi normale. On ne fait pas d'hypothèse sur l'égalité des variances. On a

$$\overline{X}_i \approx \mathcal{N}\left(\mu_i; \frac{\sigma_i^2}{n_i}\right).$$

4.2.1 Comparaison de deux moyennes, cas gaussien, variances égales

4.2.1.1 Test d'hypothèses

Modèle. On dispose de deux échantillons indépendants d'une variable aléatoire X. Chaque échantillon est gaussien et de même variance σ^2, $X_{ij} \sim \mathcal{N}(\mu_i; \sigma^2)$. Les X_{ij} sont indépendants.

Hypothèses H0 et H1. On pose les hypothèses suivantes :

$$\text{H0} : \mu_1 = \mu_2$$

contre

$$\text{H1} : \mu_1 \neq \mu_2 \quad \text{(test bilatéral)}$$

ou

$$\text{H1} : \mu_1 < \mu_2 \quad \text{ou} \quad \text{H1} : \mu_1 > \mu_2 \quad \text{(test unilatéral)}$$

ou

$$\text{H1} : \mu_1 > \mu_2 \quad \text{(test unilatéral).}$$

Statistique de test. On peut utiliser la statistique de test :

$$T = \frac{(\overline{X_1} - \overline{X_2})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.2}$$

Sous l'hypothèse H0, la statistique de test suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté :

$$T \sim_{\text{H0}} \mathcal{T}_{n_1+n_2-2} \tag{4.3}$$

En effet, on a

$$\overline{X}_1 \sim \mathcal{N}\left(\mu_1; \frac{\sigma^2}{n_1}\right)$$

et

$$\overline{X}_2 \sim \mathcal{N}\left(\mu_2; \frac{\sigma^2}{n_2}\right).$$

donc, comme on a deux échantillons indépendants,

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2; \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

d'où

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0;1).$$

Sous H0 $\mu_1 = \mu_2$, donc on a simplement :

$$\frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim_{H0} \mathcal{N}(0;1).$$

Comme σ n'est pas connu, on le remplace par son estimateur S, d'où l'équation 4.2. Cela modifie la loi de la statistique car au dénominateur, on remplace un paramètre avec une valeur donnée par une variable aléatoire, d'où le résultat 4.3.
Pour trouver l'expression de S^2, on va utiliser les deux estimations de la variance de la population (à partir de chacun des échantillons). On peut écrire $S_1^2 = \frac{SC_1}{n_1-1}$ et $S_2^2 = \frac{SC_2}{n_2-1}$ (SC pour Somme des Carrés). Donc la SC totale est $(n_1-1)S_1^2 + (n_2-1)S_2^2$ et le nombre de ddl est $n_1 + n_2 - 2$. D'où :

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}. \tag{4.4}$$

Zone de rejet et p-value. Si H0 est vraie, alors la différence des deux moyennes empiriques $\overline{X}_1$ et $\overline{X}_2$ fluctuera autour de 0. Si H0 est fausse, la distribution de T ne sera pas centrée autour de 0 et sera donc à gauche ou à droite de la distribution sous H0. La forme de la zone de rejet dépend de l'hypothèse H1. Pour un test bilatéral, comme la loi de Student est symétrique, elle sera de la forme :

$$]-\infty, -t_{\text{seuil}}] \cup [t_{\text{seuil}}, +\infty[$$

Pour un test bilatéral, la p-value se calcule en utilisant la propriété de symétrie de la loi de Student :

$$pval = 2\left(1 - F_{\mathcal{T}_{n_1+n_2-2}}(|t_{\text{obs}}|)\right)$$

Validation des conditions d'application. On vérifie graphiquement (Q-Q plot) que les deux variables aléatoires X_1 et X_2 suivent une distribution normale. Si ce n'est pas le cas mais que les échantillons sont grands ($n_i > 30$) on pourra utiliser le test avec approximation gaussienne de la moyenne. Si ce n'est pas le cas et qu'on a un petit échantillon ou que les échantillons sont trop petits pour que la vérification graphique soit possible, il faudra utiliser un test non paramétrique.

4.2.1.2 Intervalle de confiance

On peut construire un intervalle de confiance pour la différence inconnue entre les deux moyennes, $\mu_1 - \mu_2$. En effet, la variable aléatoire

$$T = \frac{\left((\overline{X_1} - \mu_1) - (\overline{X_2} - \mu_2)\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

suit aussi une loi $\mathcal{T}_{n_1+n_2-2}$. On ne peut pas calculer sa valeur car μ_1 et μ_2 sont inconnus, mais on peut calculer la valeur t_α telle que

$$P(-t_\alpha \leq T \leq t_\alpha) = \alpha.$$

En utilisant un peu d'algèbre, comme dans la section sur les tests de conformité, on trouve :

$$IC_{1-\alpha}(\mu_1 - \mu_2) =](\overline{X_1} - \overline{X_2}) - t_\alpha s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\overline{X_1} - \overline{X_2}) + t_\alpha s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}[. \tag{4.5}$$

Exemple. On mesure le nombre moyen de graines par fruit sur 50 plantes femelles et 50 plantes hermaphrodites de gypsophile. Les observations sont synthétisées dans le tableau ci-dessous :

n	$\overline{x}_i$	$s^2_{i_{n_i-1}}$
50	15,64	135,37
50	17,30	149,26

La variance commune peut être estimée comme (*cf.* equation 4.4) :

$$s = (49 \times 135,37 + 49 \times 149,26)/98 = 142,31.$$

On trouve l'intervalle de confiance pour la différence $(\mu_2 - \mu_1)$:

$$IC_{95\%} =]-6,23, +2,90[.$$

Remarque : Zéro est compris dans l'intervalle de confiance, on retrouve la conclusion du test : il n'y a pas de différence statistiquement significative entre les deux moyennes au niveau α choisi.

4.2.2 Comparaison de deux moyennes, cas gaussien, variances inégales

Dans la plupart des cas, on ne connaît ni la moyenne, ni la variance des deux échantillons. On peut faire un test d'égalité des variances (voir plus loin). Il peut arriver que l'on ne puisse pas faire l'hypothèse d'égalité des variances. Dans ce cas, on peut utiliser le test de Student avec correction de Welch mais c'est un test approché.

Modèle. On dispose de deux échantillons indépendants d'une variable aléatoire X. Chaque échantillon est gaussien, $X_{ij} \sim \mathcal{N}(\mu_i; \sigma_i^2)$. On a $\sigma_1^2 \neq \sigma_2^2$. Les X_{ij} sont indépendants.

Hypothèses H0 et H1. On pose H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$.
On peut également utiliser un test unilatéral en prenant comme hypothèse alternative : H1 : $\mu_1 > \mu_2$ ou H1 : $\mu_1 < \mu_2$.

Statistique de test. Sous l'hypothèse H0, la statistique de test

$$W = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}}$$

sous H0, cette statistique suit asymptotiquement une loi de Student, mais avec un nombre de degrés de liberté différent ν, que l'on appelle *degrés de liberté efficaces*, qui dépend des estimations des variances de X_1 et X_2 :

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-2)}}$$

Donc,

$$W \approx_{H0} \mathcal{T}(\nu)$$

Par défaut, la fonction `t.test` du logiciel R suppose l'inégalité des variances et réalise un test de Welch. Attention, car la loi sous H0 est une loi *asymptotique*. En d'autres termes, la loi sous H0 est connue uniquement lorsque les effectifs des deux populations sont grands.

4.2.3 Comparaison de deux moyennes, cas non gaussien, grands échantillons

Modèle. On suppose que l'on a deux échantillons indépendants. La moyenne des X_{ij} est notée μ_i et leur variance σ_i^2. On ne fait pas d'hypothèse sur la loi des X_{ij} mais on suppose que les échantillons sont suffisamment grands ($n_i \geq 30$) pour que l'on puisse faire les hypothèses que

$$\overline{X}_i \approx \mathcal{N}\left(\mu_i; \frac{\sigma_i^2}{n_i}\right).$$

.

Hypothèses H0 et H1. On pose H0 : $\mu_1 = \mu_2$ contre H1 : $\mu_1 \neq \mu_2$.
Comme pour les autres tests, on peut aussi poser une hypothèse alternative unilatérale.

Statistique de test. On utilisera la statistique de test

$$Z = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}}$$

Comme on a de grands échantillons, on considère que la différence des moyennes suit asymptotiquement une loi gaussienne et que les estimateurs des variances convergent vers les vraies valeurs des variances. Aussi, sous l'hypothèse H0, Z suit asymptotiquement une loi normale centrée réduite :

$$Z \approx_{H0} \mathcal{N}(0;1)$$

4.2.4 Test non paramétrique de Mann-Whitney

Si l'on ne connaît rien sur les variables aléatoires X_1 et X_2, et que la taille des échantillons est trop petite pour appliquer le théorème central limite (on n'est pas sûr que $\overline{X}_1$ et $\overline{X}_2$ soient gaussiens), on peut toujours réaliser un test non paramétrique, basé sur les rangs.

Modèle. L'idée de ce test est de comparer deux valeurs prises au hasard, une dans le premier échantillon, et l'autre dans le second échantillon. Si les variables aléatoires suivent la même loi de probabilité, alors il y a une chance sur deux que la première valeur soit inférieure à la seconde :

$$P(X_1 \leq X_2) = P(X_2 \leq X_1) = 0,5.$$

.
On peut ordonner les $n = n_1 + n_2$ éléments des deux échantillons et définir, pour chaque individu, son rang dans la séquence ainsi formée. Si les lois de probabilité sont les mêmes, les rangs des individus des deux échantillons devraient être comparables. A noter que la somme de tous les rangs vaut

$$1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}.$$

Hypothèses H0 et H1. On pose :

$$\text{H0} : P(X_1 \leq X_2) = P(X_2 \leq X_1) = 0,5.$$

H0 sera toujours vraie si les deux populations suivent la même loi de probabilité.

On peut réaliser un test bilatéral en posant :

$$\text{H1} : P(X_1 \leq X_2) \neq P(X_2 \leq X_1),$$

ou bien un test unilatéral, par exemple :

$$\text{H1} : P(X_1 \leq X_2) < P(X_2 \leq X_1).$$

Dans les deux cas, H1 implique que les lois de probabilité des deux populations sont différentes.

Statistique de test. On peut calculer la somme $R1$ des rangs des individus issus du premier échantillon, et la somme $R2$ des rangs des individus issus du second échantillon puis calculer, pour chaque échantillon, la différence avec la valeur minimale attendue (si les individus de l'échantillon occupent les plus petits rangs de 1 à n_1 ou n_2 selon l'échantillon considéré) :

$$U1 = R1 - n_1(n_1 + 1)/2$$

$$U2 = R2 - n_2(n_2 + 1)/2$$

Puisque $R1 + R2 = n(n + 1)/2$, on a $U_1 + U_2 = n_1 n_2$. Donc, si les rangs des deux échantillons sont comparables (hypothèse H0), on attend une valeur moyenne de $\frac{n_1 n_2}{2}$ pour chacune de ces deux variables aléatoires. Chacune de ces variables suit une loi qui peut être tabulée sous H0. En pratique, les tables donnent la loi de la plus petite de ces deux variables, donc la statistique de test est : $U = \min(U1, U2)$.

Exemple. Un laboratoire d'analyse médicales utilise deux appareils différents pour mesurer la glycémie des patients. Le technicien voudrait savoir si les deux appareils donnent les mêmes mesures. Il utilise pour cela les données de glycémie post-prandiale (1h30 après un repas) obtenues sur une journée par le laboratoire. La norme est une valeur inférieure à 1,40 g/L. Les données peuvent se présenter de la façon suivante :

Glycémie	1,64	1,05	1,58	1,20	1,43	1,35	1,10	1,65	1,25	1,63	1,51	1,43	1,32	1,27
Appareil	1	1	1	1	1	1	2	2	2	2	2	2	2	2
Rang R	13	1	11	3	8,5*	7	2	14	4	12	10	8,5*	6	5

* Valeur non entière car il y a deux $8^{\text{èmes}}$ ex aequo.

Les tailles des échantillons sont $n_1 = 6$ et $n_2 = 8$. On peut utiliser le tableau pour calculer $R1 = 43,5$ et $R2 = 61,5$. On trouve donc

$$U_1 = 43,5 - 6 \times 7/2 = 22,5$$

$$U_2 = 61,5 - 8 \times 9/2 = 25,5$$

La statistique de test vaut $U = 22,5$. La p-value peut se calculer en utilisant la fonction `wilcox.test(x1,x2)` de R, où x1 est le vecteur des valeurs issues de l'appareil 1, et x2 le vecteur des valeurs issues de l'appareil 2. On trouve ici une p-value de 0,8972 et on ne peut pas rejeter H0.

4.3 Test d'homogénéité sur la variance

Modèle. On dispose de deux échantillons d'une variable aléatoire X. On appelle X_i la variable aléatoire correspondant à l'échantillon i. On suppose que chaque échantillon est Gaussien :

$$X_i \sim \mathcal{N}(\mu_i; \sigma_i^2)$$

On cherche à savoir si les variances des deux populations sont égales.

Hypothèses H0 et H1. On pose les hypothèses suivantes :

$$H0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$$

contre

$$H0 : \sigma_1^2 \neq \sigma_2^2$$

Statistique de test. Pour trouver une statistique de test, on utilise la propriété suivante : si X suit une loi gaussienne, alors

$$\frac{S_{n-1}^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2.$$

Cette propriété est vraie pour chacun des deux échantillons. Donc, sous l'hypothèse H0,

$$F = \frac{S^2_{1_{n_1-1}}}{S^2_{2_{n_2-1}}}$$

est un rapport de deux variables aléatoires qui suivent chacune une loi de $\mathcal{X}^2$. La résultante est une loi de Fisher $\mathcal{F}^{n_1-1}_{n_2-1}$.

Forme de la zone de rejet et *p*-value. Contrairement à la loi normale et à la loi de Student, la loi de Fisher n'est pas symétrique. Le quantile de gauche n'est pas l'opposé du quantile de droite. Par contre, si $F \sim \mathcal{F}^{n_1-1}_{n_2-1}$, alors $1/F \sim \mathcal{F}^{n_2-1}_{n_1-1}$. Par ailleurs, les valeurs de F sont forcément positives.
Sous H0, on s'attend à ce que S_1^2 et S_2^2 estiment la même quantité σ^2, et leur rapport doit donc être proche de 1. Les valeurs très grandes ou très petites du rapport sont moins probables sous l'hypothèse H0. La forme de la zone de rejet est donc :

$$[0, f_{\text{inf}}] \cup [f_{\text{sup}}, +\infty[.$$

Pour un test au niveau α, les valeurs de f_{inf} et f_{sup} sont les quantiles à $\alpha/2$ et $1-\alpha/2$ de la loi $\mathcal{F}^{n_1-1}_{n_2-1}$.
Pour calculer la p-value, on va calculer la probabilité sous H0 d'une zone de la forme $[0, a] \cup [b, +\infty[$ où a et b sont situés respectivement à gauche et à droite de la médiane de la loi $\mathcal{F}^{n_2-1}_{n_1-1}$ que l'on notera ici m.

— Si $f_{\text{obs}} > m$, alors $b = f_{\text{obs}}$ et la p-value se calcule comme

$$pval = 2 * P(F \geq f_{\text{obs}}) = 2\left(1 - F_{\mathcal{F}^{n_1-1}_{n_2-1}}(f_{\text{obs}})\right)$$

— Si $f_{\text{obs}} < m$, alors $a = f_{\text{obs}}$ et la p-value se calcule comme

$$pval = 2 * P(F \leq f_{\text{obs}}) = 2F_{\mathcal{F}^{n_1-1}_{n_2-1}}(f_{\text{obs}})$$

Sous R, la fonction de répartition de la loi de Fisher s'obtient par la fonction `pf(fobs,n1-1,n2-2)`.

4.4 Bilan

Les tests d'homogénéité permettent de comparer entre eux les paramètres des lois de distribution de deux variables aléatoires pour savoir s'ils sont égaux. On ne réalise pas le même test selon la nature de la variable aléatoire et la question posée.

1. **v.a. discrète :** tests du $\mathcal{X}^2$. Le support contient J classes. On peut comparer I échantillons. Les effectifs théoriques m_{ij} sont calculés par $\frac{n_{i.}n_{.j}}{n}$ où $n_{i.}$ est la taille de l'échantillon i et $n_{.j}$ est le nombre d'individus de la classe j dans les I échantillons.

$$Z = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \sim_{H0} \mathcal{X}^2_{(I-1)(J-1)}$$

   Les effectifs théoriques doivent être supérieurs ou égaux à 5 dans chaque classe.

2. **v.a. continue, test sur la moyenne**
   — **Echantillons gaussiens :** $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_i^2)$ test de Student.
      — **Variances égales**

$$T = \frac{(\overline{X_1} - \overline{X_2})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim_{H0} \mathcal{T}_{n_1+n_2-2}$$

      — **Variances inégales**

$$W = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}} \approx_{H0} \mathcal{T}_\nu$$

— **Echantillons non gaussiens mais grands** ($n_i \geq 30$)

$$Z = \frac{(\overline{X_1} - \overline{X_2})}{\sqrt{\frac{S^2_{1_{n_1-1}}}{n_1} + \frac{S^2_{2_{n_2-1}}}{n_2}}} \approx_{H0} \mathcal{N}(0;1)$$

— **Petits échantillons sans information sur la loi de X_1 et de X_2 :** test de rang de Mann-Whitney.

3. **v.a. continues gaussiennes, test sur la variance :** test de Fisher

$$F = \frac{S^2_{1_{n_1-1}}}{S^2_{2_{n_2-1}}} \sim_{H0} \mathcal{F}^{n_1-1}_{n_2-1}.$$

Chapitre 5

Couples de variables aléatoires et relations de dépendance

5.1 Le $\mathcal{X}^2$ d'indépendance

Modèle. Dans une population, on considère un n-échantillon d'un couple de variables aléatoires qualitatives :

$$\{(X_1, Y_1), (X_2, Y_2), \quad ..., \quad (X_n, Y_n)\}$$

Chacune de ces variables est définie par son support et sa loi de probabilité. Cela veut dire que l'on caractérise chaque individu de la population par deux variables. Par exemple, dans une population de chats, on peut caractériser chaque individu par la couleur dominante de son pelage ($a_1 = gris, a_2 = roux, ..., a_k = blanc$) et son sexe ($b_1 = F, b_2 = M$).

Variable aléatoire	Support	Loi de probabilité
X_i	$(a_1, \quad ... \quad , a_K)$	$P(X_i = a_k) = p_k$ (inconnus)
Y_i	$(b_1, \quad ... \quad , b_L)$	$P(Y_i = b_l) = q_l$ (inconnus)

On compte le nombre de réalisations n_{kl} pour lesquels $X_i = a_k$ et $Y_i = b_l$ dans l'échantillon. Les résultats peuvent se présenter sous la forme d'un tableau de contingence :

Support	b_1	...	b_l	...	b_L	Total
a_1	n_{11}	...	n_{1l}	...	n_{1K}	$n_{1.}$
...	...	...	...	...	...	...
a_k	n_{k1}	...	n_{kl}	...	n_{kL}	$n_{k.}$
...	...	...	...	...	...	...
a_K	n_{K1}	...	n_{Kl}	...	n_{KL}	$n_{K.}$
Total	$n_{.1}$	...	$n_{.l}$	...	$n_{.L}$	n

On peut estimer $P(X_i = a_k) = p_k$ ou $P(Y_i = b_l) = q_k$ en utilisant les effectifs marginaux $n_{k.}$ et $n_{.l}$:

$$\widehat{p}_k = \frac{n_{k.}}{n} \tag{5.1}$$

$$\widehat{q}_l = \frac{n_{.l}}{n}. \tag{5.2}$$

On se demande si X et Y sont indépendants. Si c'est le cas, alors

$$P(X_i = a_k \quad \text{et} \quad Y_i = b_l) = P(X_i = a_k) \times P(Y_i = b_l) = p_k q_l$$

Sous l'hypothèse d'indépendance des deux variables, on peut donc calculer des effectifs théoriques (moyens) pour chaque case du tableau :

$$m_{kl} = n\widehat{p}_k\widehat{q}_l = \frac{n_{k.}n_{.l}}{n}.$$

Hypothèses H0 et H1. On pose
H0 : X et Y sont indépendants : $P(X_i = a_k \quad \text{et} \quad Y_i = b_l) = p_k q_l$, pour tout k, l
H1 : X et Y ne sont pas indépendants : il existe k, l tels que $P(X_i = a_k \quad \text{et} \quad Y_i = b_l) \neq p_k q_l$
Ici, il n'y a qu'une seule façon de poser l'hypothèse alternative H1.

Statistique de test et loi sous H0. On a calculé les effectifs attendus dans chaque classe sous H0. Si H0 est vraie, les effectifs fluctuent autour des ces valeurs. Sous H0, la statistique de test :

$$Z = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(n_{kl} - m_{kl})^2}{m_{kl}} \approx_{\text{H0}} \mathcal{X}^2_{(K-1)(L-1)} \tag{5.3}$$

suit asymptotiquement une loi du $\mathcal{X}^2_{(K-1)(L-1)}$. Aussi, si les effectifs théoriques sont suffisamment grands, sous H0, Z suit approximativement la loi $\mathcal{X}^2_{(K-1)(L-1)}$.

Zone de rejet : Sous H1, il y aura en moyenne plus d'écart que sous H0 entre les effectifs théoriques et observés. La distribution de la statistique de test sous H1 se situe donc à droite de sa distribution sous H0. Aussi, on rejettera H0 pour les grandes valeurs de Z_{obs}. Le seuil sera choisi en fonction du α choisi.

Conditions d'application : les effectifs théoriques doivent être supérieurs ou égaux à 5 dans chaque classe : $\forall k, \forall l, m_{kl} \geq 5$.

5.2 Corrélation

5.2.1 Test paramétrique de Pearson

Modèle. On considère un n-échantillon d'un couple gaussien de variables aléatoires

$$\{(X_1, Y_1), (X_2, Y_2), \quad ..., \quad (X_n, Y_n)\}$$

couple gaussien : *Lorsque l'on observe deux variables aléatoires sur chaque individu, on parle de couple de v.a. On dit qu'un couple de v.a quantitatives est un couple gaussien si chaque combinaison linéaire $a * X_i + b * Y_i$ est gaussien.*

En pratique, on vérifiera la normalité des deux variables (cas $a = 0$ et $b = 0$) et on vérifiera graphiquement que la relation entre les deux variables semble linéaire.

Remarque : Si X et Y sont indépendantes, alors $Cov(X, Y) = 0$, la réciproque n'est vraie que dans certains cas. Pour un couple gaussien, une covariance nulle entre X et Y ($Cov(X_i, Y_i) = 0$) indique l'indépendance des variables. Ainsi, pour un couple gaussien, pour vérifier l'indépendance, on teste la nullité du coefficient de corrélation ρ_{XY} qui est, on le rappelle, une mesure sans unité de la covariance.

Hypothèses H0 et H1. On va supposer que l'on a un couple gaussien. On souhaite tester l'indépendance des deux variables. L'hypothèse H0 est que X et Y sont indépendants que l'on peut traduire par :

$$\text{H0} : \rho_{XY} = 0 \quad \text{(la corrélation est nulle)}.$$

L'hypothèse H1 est que X et Y ne sont pas indépendantes, et l'on peut choisir soit un test bilatéral :

$$\text{H1} : \rho_{XY} \neq 0 \quad \text{(la corrélation est non nulle)}$$

soit un test unilatéral :

$$\text{H1} : \rho_{XY} < 0 \quad \text{(la corrélation est négative)}$$

ou

$$\text{H1} : \rho_{XY} > 0 \quad \text{(la corrélation est positive)}.$$

Statistique de test et loi sous H0. On peut estimer la corrélation par l'estimateur ci-dessous qui mesure la relation linéaire entre deux variables aléatoires :

$$r_{XY} = \frac{S_{XY}}{S_{X_{n-1}} S_{Y_{n-1}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Sous l'hypothèse H0, la statistique de test est :

$$Z = \sqrt{n-2}\frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sim_{\text{H0}} \mathcal{T}_{n-2}$$

suit une loi de Student à $n-2$ ddl.

Conditions d'applications. Le test est valable pour des couples de v.a. gaussiennes. Cependant, pour un grand échantillon ($n \geq 30$), on a un résultat approché similaire à condition de ne pas s'éloigner d'une relation linéaire.

5.2.2 Remarque sur le coefficient de corrélation

Le coefficient de corrélation de deux variables aléatoires mesure le degré de dépendance linéaire entre ces deux variables (figure 5.1a), du type

$$Y = a + bX + \epsilon \tag{5.4}$$

où a et b définissent les relations de dépendance, et ϵ est une variable aléatoire représentant les erreurs de mesure. On verra dans le chapitre modèle linéaire comment on modélise ces erreurs.
On peut montrer que b, la pente de la droite, est :

$$b = \frac{cov(X,Y)}{V(X)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

Il faut noter qu'il existe une relation entre b et le coefficient de corrélation de Pearson (équation 1.2). En effet

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{[V(X)V(Y)]}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}},$$

donc

$$b = \rho_{XY}\sqrt{\frac{V(Y)}{V(X)}} = \rho_{XY}\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}.$$

5.2.3 Précautions à prendre

L'analyse des corrélations entre variables aléatoires peut prêter à des interprétations erronées. La mesure de la corrélation linéaire (ou de la covariance) n'est pas toujours un bon indicateur de la dépendance entre deux variables aléatoires. Une analyse de corrélation doit toujours être accompagnée d'une analyse graphique. Les points suivants **ne doivent jamais être oubliés** :

— **Indépendance et corrélation.** Si deux variables aléatoires sont indépendantes, leur covariance est nulle. Par contre l'inverse n'est pas toujours vrai (figure 5.1b). Deux variables aléatoires peuvent être reliées l'une à l'autre mais présenter une covariance nulle.

— **Dépendance non linéaire**. Les cas de dépendance non linéaire peuvent conduire à une mauvaise estimation de la relation de dépendance (figure 5.1c).

— **Dépendance et causalité**. Une covariance non nulle entre X et Y (figure 5.1a) n'implique pas forcément de relation de causalité entre les deux variables. Il peut exister une variable aléatoire Z qui détermine à la fois X et Y et qui va conduire à une relation linéaire entre X et Y. La relation de causalité est entre Z et X et entre Z et Y, mais pas entre X et Y.

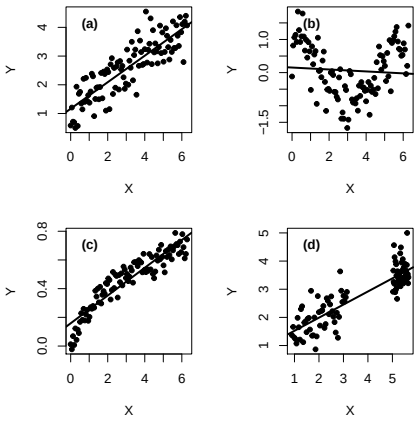


FIGURE 5.1 – **Exemple de corrélations entre caractères.** On modélise une situation où la valeur de Y dépend de X et d'une erreur aléatoire qui est la même dans les quatre situations. **(a)** Modèle linéaire : $Y = 1 + 0.5X + \epsilon$. **(b)** Modèle non linéaire : $Y = cos(X) + \epsilon$. **(c)** Modèle non linéaire : $Y = \frac{X}{3+X} + \epsilon$. **(d)** Population structurée pour X : même relation qu'en **(a)** ($Y = 1 + 0.5X + \epsilon$), mais les valeurs de X sont groupées autour de 2 et de 6.

Modèle	Type	$\widehat{r}_{XY}$	p-value
(a)	linéaire	$0,90$	$< 2.2e^{-16}$
(b)	cos	$0,04$	$0,071$
(c)	hyperbolique	$0,25$	$0,01$
(d)	X structuré	$0,88$	$< 2.2e^{-16}$

TABLE 5.1 – **Résultats du test de Pearson pour les données simulées de la figure 5.1.**

— **Population structurée.** Supposons que la variable X permette de séparer clairement la population en deux grandes classes. A l'intérieur de chaque classe de valeurs de X, la corrélation entre X et Y est faible à nulle. La corrélation entre Y et X reflète la structuration de X (figure 5.1d) et devient dans ce cas plus complexe à interpréter.

Le Tableau 5.1 montre le résultat du test de Pearson pour les jeux de données simulés dans la figure 5.1 et correspondant à quatre cas différents de dépendance entre Y et X.

Attention, il faut toujours accompagner le test statistique d'un examen visuel des données pour vérifier que les conditions d'application du test sont vérifiées.

5.2.4 Test non paramétrique de Spearman

Lorsque la taille de l'échantillon est trop faible (< 30), ou que les variables aléatoires ne sont pas gaussiennes, ou encore que la relation n'est pas linéaire, on peut remplacer le test paramétrique par un test non paramétrique en travaillant sur les corrélations de rang. Il s'agit de remplacer les valeurs des observations par leur rang, et de s'intéresser à la différence de rang D_i. Si X et Y sont parfaitement corrélées, alors le rang de X_i suivra celui de Y_i et la différence de rang sera très faible. Au contraire, s'il n'existe aucune relation entre X et Y, la différence de rang sera aléatoire.

Coefficient de corrélation de Spearman. Le coefficient de corrélation de rang de Spearman est donné par

$$r_S = 1 - \frac{6 \sum_i D_i^2}{n^2(n-1)}.$$

Exemple. On considère la relation entre la consommation annuelle de chocolat (kg/an/personne) et le nombre cumulé de prix Nobel par 10 millions d'habitants pour l'ensemble des pays d'Europe.

Pays	Consommation chocolat	Prix Nobel	X_{rang}	Y_{rang}	d_i
Grèce	2,5	2	1	1	0
Pologne	3,9	3	2	2	0
Pays-Bas	4,2	11	3	4	–1
France	6,2	9,8	4	3	+1
Royaume-Uni	9,8	19	5	5	0
Suisse	12	32	6	6	0

TABLE 5.2 – **Relation entre la consommation de chocolat et le nombre de prix Nobel.**

L'idée est que la consommation annuelle de chocolat reflète le niveau social. On s'attend à une relation positive, mais on ne dispose que de valeurs mesurées dans quelques pays ($n < 30$). On fabrique de nouvelles variables aléatoires discrètes, correspondant au classement de chaque variable, et on mesure d_i, la différence de rang entre X_i et Y_i.

Hypothèses H0 et H1. On pose H0 : il n'y a pas de corrélation entre X et Y, contre H1 : X et Y sont corrélées.

Statistique de test. Pour les petits échantillons, la loi de r_S sous l'hypothèse H0 est tabulée (elle ne correspond pas à une loi connue). Pour les grands échantillons, on a un résultat asymptotique et on peut considérer que la transformation de r_S suivante suit approximativement une loi de Student :

$$r_S\sqrt{\frac{n-2}{1-r_S^2}} \approx \mathcal{T}_{n-2}.$$

5.2.5 Bilan

Les tests d'indépendance permettent de tester l'hypothèse de nullité de la covariance entre un couple de variables aléatoires. Il faut pour cela disposer d'une mesure pour chacune des variables aléatoires pour chaque individu du n-échantillon.

1. **Couple de v.a. discrètes : test du $\mathcal{X}^2$ d'indépendance**
   On dispose d'un tableau de contingence donnant les effectifs observés dans chacune des classes définies par la combinaison des deux v.a. Les effectifs théoriques m_{ij} sont calculés en utilisant les fréquences marginales.

$$Z = \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{(n_{kl}-m_{kl})^2}{m_{kl}} \sim_{\text{H0}} \mathcal{X}^2_{(K-1)(L-1)}$$

   Les effectifs théoriques doivent être supérieurs ou égaux à 5 dans chaque classe.

2. **Couple de v.a. continues**

   — **Test de Pearson** Pour un couple gaussien :

$$Z = \sqrt{n-2}\frac{r_{XY}}{\sqrt{1-r_{XY}^2}} \sim_{H0} \mathcal{T}_{n-2}$$

   Si on a un grand échantillon ($n \geq 30$), on a un résultat asymptotique :

$$Z = \sqrt{n-2}\frac{r_{XY}}{\sqrt{1-r_{XY}^2}} \approx_{H0} \mathcal{T}_{n-2}$$

— **Test de rang de Spearman.** Le coefficient de corrélation de rang se calcule comme :

$$r_S = 1 - \frac{6 \sum_i D_i^2}{n^2(n-1)},$$

où D_i est la différence de rang entre les deux v.a. pour l'individu i. La statistique de test est

$$r_S \sqrt{\frac{n-2}{1-r_S^2}} \approx_{H0} \mathcal{T}(n-2).$$

Cete loi n'est valable que pour de grands échantillons, sinon la loi est tabulée.

Chapitre 6

Le modèle linéaire

Le modèle linéaire regroupe une famille de modèles dont le modèle d'analyse de variance, le modèle de régression et l'analyse de covariance. La problématique générale est la suivante : on s'intéresse aux relations entre une variable aléatoire continue Y et un certain nombre de variables descriptives $X^{(1)}$, $X^{(2)}$, ..., $X^{(P)}$. On veut savoir si la moyenne de Y change selon la valeur des descripteurs. Les variables descriptives peuvent être qualitatives ou quantitatives. Dans tous les cas, on va modéliser les relations de dépendance entre Y et X par un modèle linéaire. En d'autres termes, on suppose que l'on peut expliquer Y comme la somme de plusieurs facteurs, comprenant l'effet moyen des variables $X^{(p)}$ et une variable aléatoire ϵ, qui résume l'effet des facteurs inconnus ou non contrôlables pouvant faire varier Y.
Voici quelques exemples de problèmes que l'on peut traiter dans le cadre général du modèle linéaire :

— On se demande si la production moyenne de blé (en quintaux par hectare) change selon les régions de France. On effectue une enquête auprès des départements pour recueillir les rendements obtenus dans l'année.
La variable à expliquer, Y, est le rendement en blé d'une exploitation agricole en fonction de la région, notée $(a_1, ..., a_i, ..., a_{13})$, où a_i est le nom de la $i^{\text{ème}}$ région de France (13 régions au total). On réalise un échantillonnage de n exploitations agricoles en France. On peut indicer les exploitations issues de de la région i par un j. Et on peut modéliser le rendement de l'exploitation j de la région i, noté Y_{ij} par :

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

où μ_i est la moyenne de la région i et où ϵ_{ij} est l'écart aléatoire entre cette moyenne et le rendement de l'exploitation j. On se demande si les moyennes par région sont différentes. Ce type de problème est traité dans la section *Analyse de variance à 1 facteur.*

— On s'intéresse à la diversité spécifique des coléoptères selon le type de sol (pâturé ou non) et la région géographique (montagne ou plaine). La variable Y à expliquer est la diversité spécifique des coléoptères, qui peut être décrite comme l'addition d'un effet cumulatif du type de sol, $X^{(1)}$, de la région, $X^{(2)}$, et d'une variable aléatoire résumant les spécificités du site d'échantillonnage. Le modèle est le suivant :

$$Y_k = \mu + t_{X_k^{(1)}} + r_{X_k^{(2)}} + \epsilon_k$$

où μ est une moyenne générale. Si l'on indice chaque échantillon par type de sol (i) et par région (j), on obtient :

$$Y_{ijl} = \mu + t_i + r_j + \epsilon_{ijl}$$

Ici, l'indice l indique le numéro de réplicat pour le sol (i) et la région (j). On se demande s'il existe des différences moyennes entre types de sols $(t_1 \neq t_2)$ ou entre régions $(r_1 \neq r_2)$. Ce type de problème est traité dans la section *Analyse de variance à 2 facteurs.*

— On se demande s'il existe chez les mammifères une relation entre le poids à la naissance et la taille adulte. La variable Y à expliquer est la taille adulte, et la variable descriptive X est le poids à la naissance. On étudie un échantillon de souris nées en captivité, mais dont les parents ont été prélevés dans le milieu naturel, dans différentes régions du globe. Cette fois-ci X est une variable continue. On peut modéliser une relation linéaire entre Y et X par

$$Y_k = a + bX_k + \epsilon_k$$

où b est le coefficient de proportionnalité entre Y et X et où ϵ mesure l'ensemble des événements non contrôlés (que l'on considérera aléatoires) pouvant se produire au cours de la vie et susceptibles de modifier la relation entre Y et X. On se demande si le coefficient de proportionalité b est différent de zéro. Ce type de problème est traité dans la section *Régression linéaire.*

— La relation de proportionnalité entre poids à la naissance et taille adulte peut varier selon l'espèce de mammifères considérée. On réalise la même étude que précédemment en considérant trois espèces : la souris, l'homme et le cochon. On dispose de deux descripteurs, l'espèce (variable qualitative) et le poids à la naissance (X, variable continue). En indiçant les individus échantillonnés selon leur espèce (i), on peut écrire

$$Y_{ik} = a_i + b_i X_{ik} + \epsilon_{ik}$$

On se demande s'il existe des différences de taille entre espèces, et s'il existe des différences entre espèces pour le coefficient de proportionnalité entre le poids à la naissance et la taille adulte. Ce type de problème est traité dans la section *Analyse de covariance*.

6.1 ANOVA à 1 facteur, tests de comparaison de moyennes

Problématique. On considère plusieurs n-échantillons de la même variable aléatoire gaussienne Y, mais issus de I populations différentes. On appelle Y_{ik} la variable aléatoire pour le $k^{\text{ième}}$ individu de l'échantillon issu de la population i. On aimerait savoir si la moyenne de cette variable dépend de la population dans laquelle on l'observe.Attention, l'ANOVA ne permet pas d'établir un lien de causalité mais un lien statistique.
On note n_i la taille de l'échantillon issu de la population i (on posera $N = \sum_{i=1}^{I} n_i$). On suppose que la variance de la variable est la même dans chacune des I populations, mais que les moyennes sont potentiellement différentes, c'est-à-dire que chaque Y_{ik} suit une loi normale $\mathcal{N}(\mu_i; \sigma^2)$.

Modèle linéaire. On peut réécrire l'énoncé ci-dessus à l'aide d'une formule mathématique qui permet de condenser les informations :

$$Y_{ik} = \mu_i + \epsilon_{ik}, \qquad \epsilon_{ik} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2) \tag{6.1}$$

Les ϵ_{ik} sont appelés les résidus du modèle et correspondent aux déviations des individus statistiques par rapport à la moyenne de leur population. σ^2 est appelée la variance résiduelle. iid signifie identiquement et indépendamment distribué et rappelle que les Y_{ik} sont issus de n-échantillons, c'est-à-dire qu'ils sont indépendants et de même loi pour un même échantillon.
On peut également reformuler le modèle en définissant μ, la moyenne générale et $\alpha_i = \mu_i - \mu$, l'écart entre la moyenne pour la population i et la moyenne générale.

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \qquad \epsilon_{ik} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2) \tag{6.2}$$

Exemples

— On veut savoir si le taux de cholestérol moyen des français dépend de la région dans laquelle ils vivent. μ_i est le taux de cholestérol moyen dans la région i, et ϵ_{ik} représente l'écart entre μ_i et la valeur de l'individu k, écart dû à des différences génétiques ou à des différences d'habitudes alimentaires qui ne peuvent pas être expliquées par l'influence de la région dans laquelle ils vivent. Par construction, l'espérance des ϵ_{ik} est nulle.

— On cherche à comparer la prise de poids d'une race bovine durant une saison à l'estive dans différents pâturages pour conseiller les éleveurs. μ_i est la prise de poids moyenne après une saison dans le pâturage i, et ϵ_{ik} représente les variations entre deux bovins dues à des différences génétiques ou à des différences individuelles sur la façon de parcourir le pâturage, qui ne peuvent pas être expliquées par la qualité moyenne du pâturage. Par construction, l'espérance des ϵ_{ik} est nulle.

Estimateurs pour les moyennes. Pour estimer la moyenne dans chaque population μ_i, on va utiliser un estimateur. Pour alléger les écritures, on utilise généralement le symbole *point* ($\cdot$) à la place d'un indice pour noter que l'on fait une moyenne (attention : dans le chapitre 4, ce symbole signifiait une somme). Ainsi :

$$Y_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik}$$

permet d'estimer la moyenne dans la population i. De même

$$Y_{..} = \frac{1}{\sum_i n_i} \sum_i \sum_k Y_{ik}$$

permet d'estimer la moyenne générale μ. A noter que, si I est le nombre de populations, la relation

$$Y_{..} = \frac{1}{I} \sum_{i=1}^{I} Y_{i.}$$

n'est vraie que si tous les n_i sont égaux.
On remarque qu'un estimateur de α_i est $\hat{\alpha_i} = Y_{i.} - Y_{..}$.

Estimateur de la variance résiduelle. On peut utiliser le modèle linéaire (6.2) pour estimer la variance résiduelle. On a :

$$\epsilon_{ik} = Y_{ik} - \mu_i$$

Comme les moyennes μ_i sont inconnues, on les remplace par les moyennes empiriques $Y_{i.}$. La variance résiduelle peut alors être estimée comme la variance empirique de ϵ_{ik} :

$$\widehat{\sigma^2} = \frac{1}{\sum_i (n_i - 1)} \sum_i \sum_k (Y_{ik} - Y_{i.})^2 = \frac{1}{\sum_i (n_i - 1)} SSR.$$

SSR est la *somme des carrés d'écart résiduelle* (de l'anglais *Sum of Squared Residuals*). Le terme correctif $\sum_i (n_i - 1) = N - I$ qui permet d'estimer la variance résiduelle à partir de la SSR sont les degrés de liberté associés, notés aussi ddl_R. On a donc, $\widehat{\sigma^2} = \frac{SSR}{ddl_R}$.

Erreur standard : *on appelle erreur standard SE l'écart-type résiduel estimé à partir de plusieurs échantillons :*

$$SE = \sqrt{\frac{SSR}{ddl_R}}. \tag{6.3}$$

On peut utiliser cette estimation de l'erreur standard pour construire un intervalle de confiance pour la moyenne de chaque population.

Test de comparaison de moyenne, ANOVA 1
On cherche à déterminer si la moyenne est la même dans toutes les populations ou si au moins une population diffère des autres.

- **Modèle.** Le modèle est le modèle linéaire (6.2)
- **Hypothèses H0 et H1.** $H0 : \forall (i,j) \quad \mu_i = \mu_j, \quad H1 : \exists (i,j) \quad \mu_i \neq \mu_j$.
  On peut les reformuler : $H0 : \forall i \quad \alpha_i = 0, \quad H1 : \exists i \quad \alpha_i \neq 0$
- **Statistique de test.** Pour écrire la statistique de test, on part de la décomposition de la variance décrite ci-dessous. La variance empirique des Y se calcule à partir de la somme totale des carrés des écarts :

  $$SST = \sum_i \sum_k (Y_{ik} - Y_{..})^2$$

  que l'on peut décomposer en utilisant

$$\begin{aligned} SST &= \sum_i \sum_k ((Y_{ik} - Y_{i.}) + (Y_{i.} - Y_{..}))^2 \\ &= \sum_i \sum_k \big((Y_{ik} - Y_{i.})^2 + (Y_{i.} - Y_{..})^2 + 2(Y_{ik} - Y_{i.})(Y_{i.} - Y_{..})\big) \\ &= \sum_i \sum_k (Y_{i.} - Y_{..})^2 + \sum_i \sum_k (Y_{ik} - Y_{i.})^2 + \sum_i \sum_k 2(Y_{ik} - Y_{i.})(Y_{i.} - Y_{..}) \end{aligned}$$

En développant le dernier terme on trouve :

$$\begin{aligned}\sum_i \sum_k (Y_{ik} - Y_{i.})(Y_{i.} - Y_{..}) &= \sum_i \sum_k \left(Y_{ik}Y_{i.} - Y_{ik}Y_{..} - Y_{i.}^2 + Y_{i.}Y_{..}\right) \\ &= \sum_i Y_{i.} \sum_k Y_{ik} - Y_{..} \sum_i \sum_k Y_{ik} - \sum_i n_i Y_{i.}^2 + \sum_i Y_{i.} n_i Y_{..} \\ &= \sum_i Y_{i.} n_i Y_{i.} - Y_{..} \sum_i n_i Y_{i.} - \sum_i n_i Y_{i.}^2 + \sum_i Y_{i.} n_i Y_{..} \\ &= \sum_i n_i Y_{i.}^2 - N Y_{..}^2 - \sum_i n_i Y_{i.}^2 + N Y_{..}^2 \\ &= 0\end{aligned}$$

On en déduit

$$\begin{aligned}SST &= \sum_i \sum_k (Y_{i.} - Y_{..})^2 + \sum_i \sum_k (Y_{ik} - Y_{i.})^2 \\ &= SSA + SSR\end{aligned}$$

Le principe de la statistique du test est de comparer l'ordre de grandeur de SSA et SSR, elle s'écrit :

$$F = \frac{\frac{SSA}{ddl_A}}{\frac{SSR}{ddl_R}}$$

où ddl_A est le nombre de degrés de liberté de SSA, c'est à dire le nombre de terme indépendants dans cette somme. On remarque que $SSA = \sum_i \sum_k (Y_{i.} - Y_{..})^2 = \sum_i \sum_k \hat{\alpha_i}^2$. Il y a I terme α_i reliés par la relation $\sum_i \alpha_i = 0$ donc $ddl_A = I - 1$.
Sous l'hypothèse H0, la statistique de test suit une loi de Fisher $\mathcal{F}_{ddlR}^{ddlA}$.

— **Choix du risque.** On choisit le risque α de se tromper en rejetant H0.

— **Règle de décision.** Sous H1, on s'attend à ce que la statistique de test soit plutôt plus grande sous H1 que sous H0 donc on rejettera H0 pour des "grandes valeurs" de la statistique de test $\mathcal{R} = \{F \in [f_{lim}; \infty[\}$ avec f_{lim} le quantile d'ordre $1 - \alpha$.

— **Vérification des conditions d'application.** En écrivant le modèle, on a fait l'hypothèse que les résidus sont indépendants, gaussiens et identiquement distribués selon la loi $\mathcal{N}(0, \sigma^2)$. On va vérifier graphiquement que cette hypothèse est acceptable. La démarche pour vérifier graphiquement cela est expliqué dans le chapitre 6.2 Analyse de la variance à deux facteurs.

On récapitule souvent les résultats de ce test dans une table d'ANOVA :

Facteur	ddl	SC*	SCM**	F	p-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
R	$N-I$	SSR	$SSR/ddlR$		

* Somme des carrés.
* Somme des Carrés moyens.

Le test HSD de Tukey-Kramer.
Si on rejette l'hypothèse H0, cela signifie qu'au moins une population diffère des autres (en moyenne). Dans ce cas, on cherchera à comparer les moyennes 2 à 2. S'il y a I population, cela revient à faire $I(I-1)/2$ tests. On peut le faire grâce au test HSD(pour *Honestly Significant Difference*) de Tukey-Kramer.

— **Modèle.** Le modèle est le modèle linéaire (6.2)

— **Hypothèses multiples (une par couple** (i, j)**)** H0 : $\mu_i = \mu_j$, H1 : $\mu_i \neq \mu_j$.

— **Statistique de test pour chaque test**

$$T = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{(\frac{1}{n_i} + \frac{1}{n_j}) SE^2}}$$

On voit que cete statistique ressemble à celle du test de Student mais que la variance est calculée avec l'ensemble des observations et pas uniquement celles des échantillons issus des populations i et j.

— **Choix du risque et règle de décision.** La loi sous H0 est tabulée et prend en compte le fait que l'on fait plusieurs tests donc que l'on augmente artificiellement la valeur du risque de première espèce.

6.2 Analyse de variance à 2 facteurs

Nous avons vu précédemment que l'ANOVA à un facteur est une méthode statistique utilisée pour tester si la valeur moyenne d'une variable quantitative dépend des variations d'une variable qualitative (nommée facteur). Si l'on met en évidence que la valeur moyenne de la variable dépend du facteur, on peut ensuite comparer les moyennes selon les groupes définis par ce facteur. Les groupes peuvent être définis par plus d'un facteur. Nous verrons ici le cas où les groupes sont définis par deux facteurs. On rappelle que l'ANOVA ne permet pas d'établir un lien de causalité mais un lien statistique.

6.2.1 Analyse de variance à deux facteurs sans interaction, cas équilibré

Problématique. Il peut arriver que les populations que l'on étudie puissent être mises dans des catégories multiples. Par exemple, on cherche à connaître la capacité d'engraissage de plusieurs races bovines dans plusieurs pâturages différents. On peut toujours décrire les données par un modèle linéaire :

Modèle

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2). \tag{6.4}$$

Cette fois-ci, le nombre de moyennes inconnues est égal au produit du nombre de niveaux de chaque facteur (race et pâturage). On peut vouloir séparer l'effet des deux facteurs. Par exemple, on peut se demander s'il existe des différences entre races d'une part, et entre pâturage d'autre part. On peut vouloir comparer l'ordre de grandeur des variations entre pâturages ou entre races. Dans ce cas il est plus simple d'identifier chacun des facteurs en réécrivant le modèle (6.4) :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0; \sigma^2), \qquad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0. \tag{6.5}$$

Le passage du modèle (6.4) au modèle (6.5) fait apparaître des termes supplémentaires. μ est la moyenne générale. α_i est l'écart à la moyenne pour les individus de la race i. β_j est l'écart à la moyenne pour les individus du pâturage j. Les paramètres α_i et β_j sont définis comme des écarts à la moyenne et donc la somme des α_i, comme la somme des β_j, est nulle. Les α_i, comme les β_j, ne sont donc pas indépendants entre eux. Les résidus sont ici toujours supposés indépendants et de même loi. On notera n_{ij} le nombre d'observations dans le groupe défini par la combinaison de facteurs (i, j) : pour cette combinaison, $k = 1, ..., n_{ij}$.
On appelle ce modèle un modèle additif puisque l'on suppose que les effets de la race et du pâturage s'additionnent. Cela signifie que l'on suppose que les écarts entre races sont constants dans tous les pâturages, et que les écarts entre pâturages sont constants entre les races.

Estimateurs des paramètres On peut estimer plusieurs paramètres :

— $Y_{...}$ est l'estimateur de la moyenne générale μ.
— $Y_{i..} - Y_{...}$ est l'estimateur de α_i, l'écart à la moyenne générale des individus de la race i.
— $Y_{.j.} - Y_{...}$ est l'estimateur de β_j, l'écart la moyenne générale des individus du pâturage j.
— On peut aussi calculer les résidus estimés : $Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}$

Variances. La variance empirique des Y se calcule à partir de la somme totale des carrés des écarts :

$$SST = \sum_i \sum_j \sum_k (Y_{ijk} - Y_{...})^2$$

Décomposition de la variance : cas du dispositif équilibré

Dans le cas d'un dispositif équilibré, c'est à dire lorsque le nombre d'observations est le même pour chaque combinaison de facteurs ($\forall(i,j) \quad n_{ij} = n$), on peut décomposer SST de la façon suivante :

$$\begin{aligned} SST &= \sum_i \sum_j \sum_k ((Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...}) + (Y_{i..} - Y_{...}) + (Y_{.j.} - Y_{...}))^2 \\ &= \sum_i \sum_j \sum_k (Y_{i..} - Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{.j.} - Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2 \\ &= SSA + SSB + SSR \end{aligned}$$

SSA dépend des paramètres α et ϵ, *SSB* dépend des paramètres β et ϵ, *SSR* ne dépend que de ϵ.

Degrés de liberté. On appelle I le nombre de niveaux du facteur A, J le nombre de niveaux du facteur B et N le nombre total d'observations. Si l'on traduit les termes des sommes de carrés d'écart SSA, SSB et SSR en fonction des paramètres et des contraintes du modèle (6.5), on peut s'apercevoir qu'il y a $I-1$ termes indépendants pour SSA ($ddl_A = I-1$), $J-1$ termes indépendants pour SSB ($ddl_B = J-1$), et $N-1-ddlA-ddlB$ termes indépendants pour SSR ($ddl_R = N-I-J+1$).

Tests ANOVA
On rappelle le modèle :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2), \qquad \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

On utilise la décomposition de la somme des carrés des écarts pour réaliser plusieurs tests statistiques indépendants.

— **Effet du facteur α**

1. **Modèle.** Le modèle linéaire (6.5).
2. **Hypothèses H0 et H1.** H0 : tous les α_i sont nuls, il n'y a pas d'effet du facteur. H1 : au moins un des α_i est différent de zéro.
3. **Choix d'une statistique de test.** Sous l'hypothèse H0, la variable aléatoire $F = \frac{\frac{SSA}{ddlA}}{\frac{SSR}{ddlR}}$ suit une loi de Fisher $\mathcal{F}_{ddlR}^{ddlA}$.
4. **Règle de décision.** On rejettera H0 pour des grandes valeurs de F, c'est-à-dire si l'ordre de grandeur des variations inter-groupes définis par A ($\frac{SSA}{ddlA}$) est plus important que l'ordre de grandeur des variations résiduelles $\frac{SSR}{ddlR}$. La limite sera le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}_{ddlR}^{ddlA}$.

— **Effet du facteur β**

1. **Modèle.** Le modèle linéaire (6.5).
2. **Hypothèses H0 et H1.** H0 : tous les β_j sont nuls, il n'y a pas d'effet du facteur. H1 : au moins un des β_j est différent de zéro.
3. **Choix d'une statistique de test.** Sous l'hypothèse H0, la variable aléatoire $F = \frac{\frac{SSB}{ddlB}}{\frac{SSR}{ddlR}}$ suit une loi de Fisher $\mathcal{F}_{ddlR}^{ddlB}$.
4. **Règle de décision.** On rejettera H0 pour des grandes valeurs de F, c'est-à-dire si l'ordre de grandeur des variations inter-groupes définis par B ($\frac{SSB}{ddlB}$) est plus important que l'ordre de grandeur des variations résiduelles $\frac{SSR}{ddlR}$. La limite sera le quantile d'ordre $1-\alpha$ de la loi $\mathcal{F}_{ddlR}^{ddlB}$.

Table d'ANOVA. La table d'ANOVA résume l'ensemble des tests statistiques réalisés.

Facteur	*ddl*	*SS*	CM^*	F	p-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
B	$J-1$	SSB	$SSB/ddlB$	$F_{B_{\text{obs}}}$	$P_{\text{H0}}(F_B > F_{B_{\text{obs}}})$
R	$N-I-J+1$	SSR	$SSR/ddlR$		

* Carré Moyen.

Elle permet de conclure directement pour chacune des hypothèses.

Conditions d'application. Les tests décrits ci-dessus s'appliquent dans le cas d'un dispositif équilibré. Pour tester les effets des facteurs dans le cas d'un dispositif déséquilibré, on se reportera à 6.2.3. Par ailleurs, il faut vérifier les hypothèses de base du modèle, à savoir que les résidus sont indépendants et de même loi. On utilise pour cela des outils graphiques.

— **Graphe des résidus.** Pour vérifier que les résidus sont indépendants, on représente graphiquement la relation entre les valeurs prédites :

$$\widehat{Y}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j$$

et les résidus observés :

$$\widehat{\epsilon}_{ijk} = Y_{ijk} - \widehat{Y}_{ij}$$

donc pour chaque valeur prédite par la moyenne estimée du groupe ij, $\widehat{Y}_{ij}$, on a n_{ij} résidus. Sous l'hypothèse d'indépendance, on s'attend à ne trouver aucune relation entre ces deux mesures.
La figure 6.1 montre le graphe des résidus obtenus avec un modèle à deux facteurs sans interaction dans l'exemple sur les bovins.Pour chaque combinaison des facteurs milieu et race, on a une valeur prédite et cinq valeurs différentes pour les résidus. On constate que les écarts entre les valeurs résiduelles sont du même ordre de grandeur quelle que soit la combinaison.

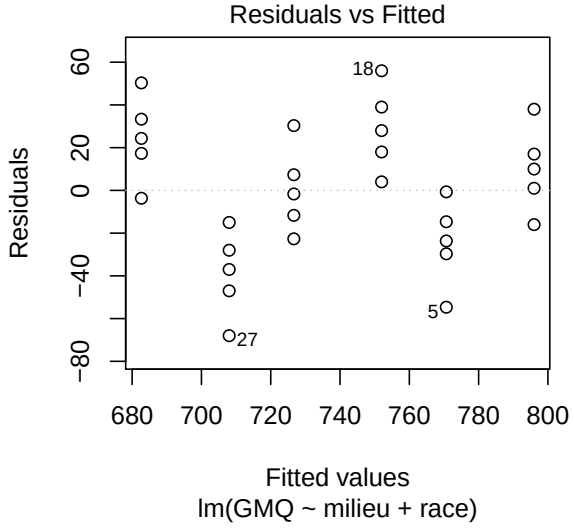


FIGURE 6.1 – **Graphe des résidus.** On représente les résidus en fonction des valeurs prédite pour le modèle à deux facteurs sans interaction. Chaque point correspond à une observation pour laquelle on a calculé la valeur prédite par le modèle, en abscisse, et l'écart à la prédiction, c'est à dire le résidu observé, en ordonnée.

— **Quantile-quantile plot.** Si l'hypothèse de normalité des résidus est vérifiée, alors $\frac{\epsilon_{ijk}}{\sigma} \sim \mathcal{N}(0;1)$. Pour chaque individu statistique du jeu de données, on peut calculer un résidu réduit :

$$e_{ijk} = \frac{\widehat{\epsilon}_{ijk}}{\widehat{\sigma}}$$

Pour voir si ces résidus réduits suivent bien une loi normale centrée réduite, on peut calculer des quantiles empiriques (ou quantiles observés) de leur distribution et les comparer aux quantiles théoriques de la loi $\mathcal{N}(0;1)$. On rappelle que les quantiles des lois usuelles sont tabulés dans les logiciels d'analyse statistique et correspondent à la fonction inverse de la fonction de répartition. Si les résidus suivent bien une loi normale centrée réduite, on s'attend à ce que les points correspondant aux différents couples de quantile (observés, théorique) s'alignent sur la droite $y = x$. Si on n'utilise les résidus non standardisés, alors on s'attend à observer une relation linéaire entre les quantiles observés et le quantiles théoriques (figure 6.2).

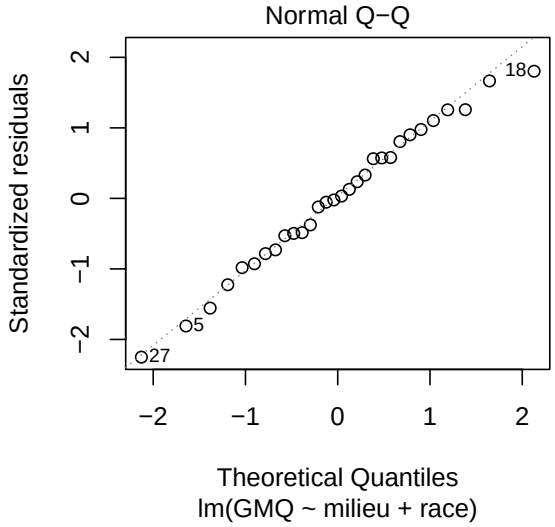


FIGURE 6.2 – **Quantile-quantile plot.** Si les résidus sont Gaussiens, alors les points du Q-Q plot doivent s'aligner sur une droite, ce qui est le cas dans cet exemple.

6.2.2 Analyse de variance à deux facteurs avec interaction, cas équilibré

Modèle. Le modèle (6.5) suppose que les différences entre races sont constantes, de même que les différences entre pâturages. On peut choisir de complexifier le modèle en introduisant un terme d'interaction entre les deux facteurs, pour prendre en compte le fait que les differences entre races ne sont pas forcément identiques dans tous les pâturages :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk}, \qquad \epsilon_{ijk} \quad \text{iid} \quad \mathcal{N}(0;\sigma^2)$$
$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \theta_{ij} = \sum_j \theta_{ij} = 0.$$

Dans l'exemple des bovins, ce modèle inclut le fait que l'effet du paturâge peut dépendre de la race (tableau ci-dessous). L'effet d'interaction est la différence entre l'espérance prédite par le modèle additif et l'espérance de chaque population.

Pâturage	Race 1	Race 2	Total
P1	$\mu + \alpha_1 + \beta_1 + \theta_{11}$	$\mu + \alpha_2 + \beta_1 + \theta_{21}$	$\mu + \beta_1$
P2	$\mu + \alpha_1 + \beta_2 + \theta_{12}$	$\mu + \alpha_2 + \beta_2 + \theta_{22}$	$\mu + \beta_2$
P3	$\mu + \alpha_1 + \beta_3 + \theta_{13}$	$\mu + \alpha_2 + \beta_3 + \theta_{23}$	$\mu + \beta_3$
Total	$\mu + \alpha_1$	$\mu + \alpha_2$	μ

Moyennes. On peut calculer plusieurs moyennes

- $Y_{...}$ est un estimateur de la moyenne générale
- $Y_{i..} - Y_{...}$ est un estimateur de α_i l'écart entre la moyenne générale et la moyenne des individus de la race i.
- $Y_{.j.} - Y_{...}$ est un estimateur de β_j l'écart entre la moyenne générale et la moyenne des individus du pâturage j.
- $Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$ est un estimateur de θ_{ij} l'effet d'interaction entre i et j c'est-à-dire l'écart à l'additivité des effets des deux facteurs.
- $Y_{ijk} - Y_{ij.}$ sont les résidus pour un n-échantillon

Décomposition de la variance : cas du dispositif équilibré.
Comme précédemment, dans le cas d'un dispositif équilibré, on peut décomposer la somme des carrés des écarts totaux :

$$\begin{aligned} SST &= \sum_i \sum_j \sum_k (Y_{ijk} - Y_{...})^2 \\ &= \sum_i \sum_j \sum_k [(Y_{i..} - Y_{...}) + (Y_{.j.} - Y_{...}) + (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}) + (Y_{ijk} - Y_{ij.})]^2 \\ &= \sum_i \sum_j \sum_k (Y_{i..} - Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{.j.} - Y_{...})^2 \\ &\quad + \sum_i \sum_j \sum_k (Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - Y_{ij.})^2 \\ &= SSA + SSB + SSI + SSR \end{aligned}$$

où SSI est la somme des carrés liée à l'interaction.

Degrés de liberté. Comme précédemment, on appelle I le nombre de niveaux du facteur A, J le nombre de niveaux du facteur B et N le nombre total d'observations. Si l'on traduit les termes des sommes de carrés d'écart SSA, SSB, SSI et SSR en fonction des paramètres et des contraintes du modèle (6.5), on peut s'apercevoir qu'il y a $I-1$ termes indépendants pour SSA ($ddl_A = I-1$), $J-1$ termes indépendants pour SSB ($ddl_B = J-1$), $(I-1)(J-1)$ termes indépendants pour SSI ($ddl_I = (I-1)(J-1)$) et $N-1-ddlA-ddlB-ddlI$ termes indépendants pour SSR ($ddl_R = N-IJ$).

Tests ANOVA
Comme précédemment, on utilise la décomposition de la somme des carrés des écarts pour réaliser plusieurs tests statistiques indépendants. En plus des effets des facteurs A et B, on réalise un test supplémentaire pour l'**effet d'interaction**. Ci-dessous, on décrit uniquement ce dernier :

1. **Modèle.** Le modèle linéaire (6.5)
2. **Hypothèses H0 et H1.** H0 : tous les θ_{ij} sont nuls, il n'y a pas d'effet d'interaction. H1 : au moins un des θ_{ij} est différent de zéro.
3. **Choix d'une statistique de test.** Sous l'hypothèse H0, la variable aléatoire $F = \frac{\frac{SSI}{ddlI}}{\frac{SSR}{ddlR}}$ suit une loi de Fisher $\mathcal{F}_{ddlR}^{ddlI}$, avec $ddlI = (I-1)(J-1)$.
4. **Règle de décision.** On rejette H0 pour les grandes valeurs de F. En effet, s'il n'y a pas d'effet d'interaction (effet θ), $\frac{SSI}{ddlI}$ et $\frac{SSR}{ddlR}$ sont deux estimations possibles pour la variance résiduelle. A l'inverse s'il existe, l'ordre de grandeur de $\frac{SSI}{ddlI}$ sera plus important que celui de $\frac{SSR}{ddlR}$.

Table d'ANOVA. La table d'ANOVA résume l'ensemble des tests statistiques réalisés :

Facteur	ddl	SS	CM	F	p-value
A	$I-1$	SSA	$SSA/ddlA$	$F_{A_{\text{obs}}}$	$P_{\text{H0}}(F_A > F_{A_{\text{obs}}})$
B	$J-1$	SSB	$SSB/ddlB$	$F_{B_{\text{obs}}}$	$P_{\text{H0}}(F_B > F_{B_{\text{obs}}})$
I	$(I-1)(J-1)$	SSI	$SSI/ddlI$	$F_{I_{\text{obs}}}$	$P_{\text{H0}}(F_I > F_{I_{\text{obs}}})$
R	$N-IJ$	SSR	$SSR/ddlR$		

Elle permet de conclure directement pour chacune des hypothèses. Attention, l'introduction de paramètres supplémentaires dans le modèle a réduit le nombre de degrés de liberté de la résiduelle.

Conditions d'application. Il faut vérifier les hypothèses de base du modèle, à savoir que les résidus sont indépendants et de même loi. On utilise pour cela des outils graphiques.

Interprétation. Dans l'exemple de la figure 6.3, les variations de la variable X sont expliquées par deux facteurs, A (deux niveaux, $A1$ et $A2$) et B (deux niveaux, $B1$ et $B2$). On peut calculer quatre moyennes $X_{ij.}$. L'espérance de ces moyennes est égale à :

$$E(X_{ij.}) = \mu + \alpha_i + \beta_j + \theta_{ij}$$

On peut représenter graphiquement ces moyennes en fonction des niveaux du facteur B, en choisissant une couleur différente selon les niveaux du facteur A.
On peut constater plusieurs choses :

- Lorsque l'interaction n'est pas significative, les différences entre $A1$ et $A2$ sont constantes, quel que soit le niveau du facteur B. De même, les différences entre $B1$ et $B2$ sont constantes, quel que soit le niveau du facteur A. Les droites sont parallèles.
- Lorsque aucun facteur n'est significatif, toutes les moyennes sont égales.
- Un effet d'interaction significatif correspond à des écarts entre les moyennes d'un facteur qui varient selon le niveau de l'autre facteur. Les droites ne sont plus parallèles. Autrement dit : l'effet du facteur A dépend du niveau de B et inversement.

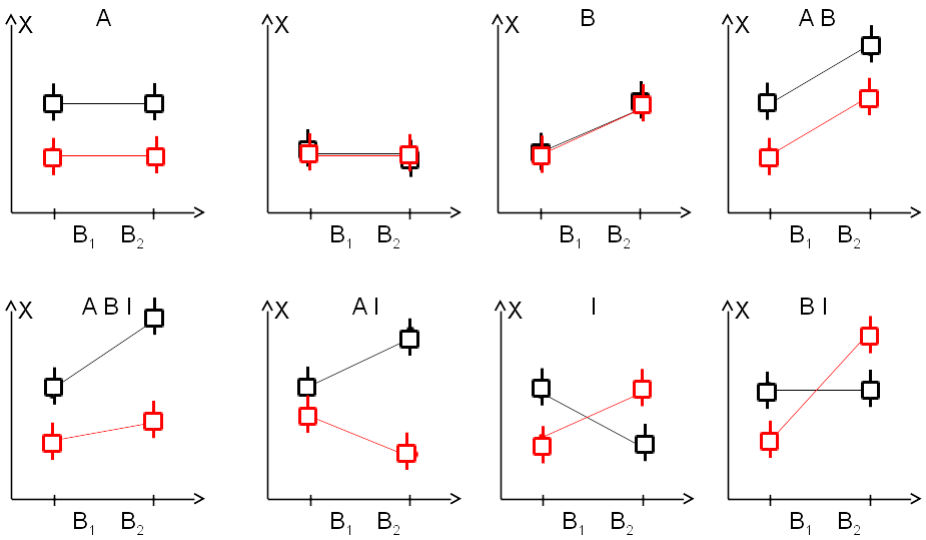


FIGURE 6.3 – **Représentation graphique de l'interaction dans un cas à deux facteurs.** Chaque graphique correspond à un résultat possible de l'ANOVA. Les facteurs qui sont significatifs (A,B,I) sont indiqués. Un résultat de l'ANOVA correspond à une représentation des moyennes de chaque combinaison de facteurs. L'axe des abcisses est le niveau du facteur B. L'axe des ordonnées est la moyenne calculée dans l'échantillon. Pour chaque situation, les moyennes sont indiquées par des carrés. La ligne verticale correspond à l'écart-type résiduel $\widehat{\sigma}_R$. Les quatre graphiques du haut correspondent à l'absence d'interaction. Les quatre graphiques du bas correspondent à des situations où l'interaction est significative.

6.3 Régression linéaire

Lorsque l'on étudie un système biologique (réseau métabolique, système nerveux, écosystème forestier, etc.), on le caractérise par plusieurs variables dont on recueille les valeurs par une étude spécifique (expérience, étude de terrain, observations cliniques, etc.). Exemple : on souhaite étudier les stratégies d'histoire de vie de la levure *Saccharomyces cerevisiae*. On va réaliser une expérience pour mesurer le taux de croissance, la taille des cellules, la capacité biotique, la vitesse de consommation des ressources et le taux de mortalité de différentes souches de levure. Chaque souche sera caractérisée par ces différentes grandeurs. Souvent, la question posée est : « Quelles sont les relations entre ces variables ? ». Une des méthodes utilisées pour répondre à cette question est la régression linéaire. Cette méthode permet d'étudier les variations linéaires d'une variable quantitative en fonction des variations d'une ou de plusieurs autres variables quantitatives. Nous commencerons par décrire la relation entre deux variables, c'est-à-dire la *régression linéaire simple*, puis nous généraliserons en présentant la *régression linéaire multiple*.

6.3.1 La régression linéaire simple

Si l'on reprend l'exemple présenté en introduction, on peut mesurer chez différentes souches de levure la taille moyenne d'une cellule et la capacité biotique, c'est-à-dire la taille maximum d'une population à niveau donnée de ressources.
Sur la figure 6.4, si l'on observe chaque milieu (1% ou 15% de glucose) séparément, on constate que la capacité biotique décroit linéairement avec la taille des cellules. Il semble « naturel » de tracer une droite au milieu des deux nuages de points. Dire qu'il y a une relation linéaire signifie que si l'on connaît la taille moyenne d'une cellule pour une souche donnée, alors on va pouvoir prédire sa capacité biotique à partir d'une équation de droite. Bien sûr, cette prédiction ne sera pas égale à la valeur observée, mais la prédiction devrait être meilleure que si l'on utilise comme prédiction la capacité biotique moyenne

de toutes les souches. La régression linéaire permet donc de proposer une équation pour cette droite et de quantifier la marge d'erreur sur les prédictions.

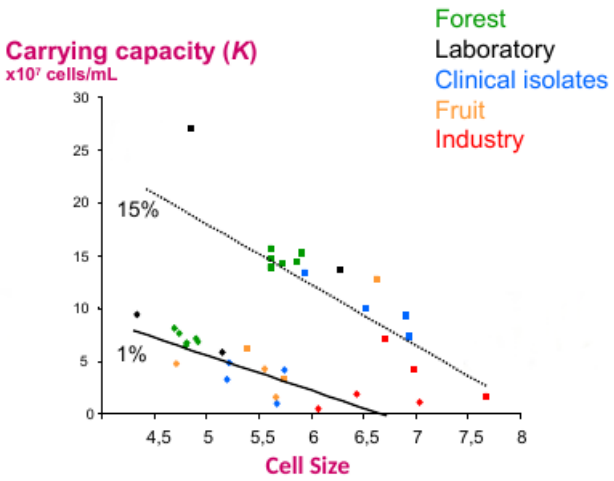


FIGURE 6.4 – **Capacité biotique et taille moyenne des cellules** de différentes souches de la levure *S. cerevisiae* mesurées dans des cultures en batch contenant initialement 1% (losanges) ou 15% (carrés) de glucose . La couleur des points dépend de la niche écologique dans laquelle les souches ont été échantillonnées. (d'après Spor *et al.*, 2009, BMC Evol. Biol. 9 :296)

6.3.1.1 Modèle

Pour présenter le formalisme, on utilisera ici les données d'une étude visant à estimer l'effet des apports en eau sur le rendement de parcelles de blé. Sur la figure 6.5 représentant ces données, on observe 30 points correspondant à 30 parcelles ayant reçu des apports en eau différents et pour lesquelles on a observé le rendement.
On va noter :

- n le nombre d'observations ($n = 30$)
- Y_i le rendement de la $i^{\text{ème}}$ parcelle échantillonnée ($i = 1$ à n)
- x_i l'apport en eau de la $i^{\text{ème}}$ parcelle échantillonnée ($i = 1$ à n)

On considère que l'on peut prédire le rendement Y_i par l'apport en eau x_i moyennant une erreur ϵ_i. On peut donc écrire :

$$Y_i = a + bx_i + \epsilon_i$$

où les résidus ϵ_i sont des variables aléatoires gaussiennes indépendantes et identiquement distribuées selon $\mathcal{N}(0; \sigma^2)$ (σ^2 inconnu). La partie $a + bx_i$ est appelée la *prédiction*. On dit que le rendement est la *variable expliquée* et l'apport en eau est la *variable explicative*. Une autre façon d'écrire ce modèle est de dire que les variables Y_i sont des variables aléatoires gaussiennes indépendantes et distribuées selon $\mathcal{N}(a + bx_i; \sigma^2)$.

6.3.1.2 Estimateurs des paramètres

On cherche à estimer les valeurs inconnues a et b à partir des données de l'échantillon. Pour cela, on peut rechercher la droite « la plus proche » de tous les points en définissant un critère permettant de quantifier la distance entre les points et la droite (figure 6.6). On appelle ce critère l'*erreur quadratique*, définie comme la somme des distances au carré entre Y_i et sa prédiction $a + bx_i$.

$$\text{Erreur quadratique} = SSR = \sum_{i=1}^{n} (Y_i - (a + bx_i))^2$$

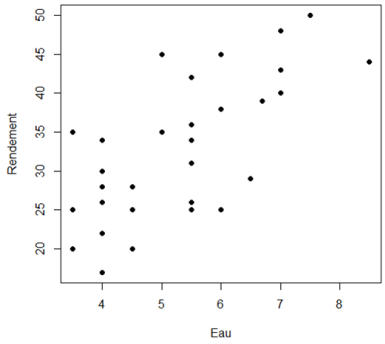


FIGURE 6.5 – **Rendement de parcelles de blé en fonction des apports en eau.** Chaque point relie l'apport en eau d'une parcelle de blé au rendement mesuré sur cette même parcelle.

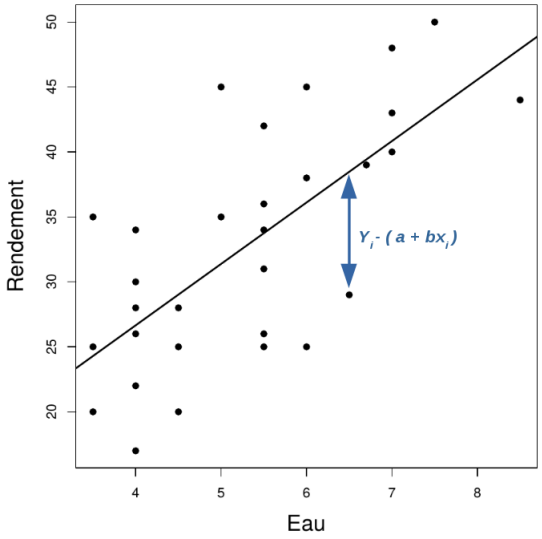


FIGURE 6.6 – **Représentation de la distance entre une mesure et la droite de régression.** Les estimateurs des paramètres a et b de la droite de régression sont choisis de façon à minimiser la somme du carré de ces distances.

L'objectif est de trouver les valeurs de a et b qui minimisent l'erreur quadratique. Par ailleurs, on peut définir la moyenne de l'échantillon pour la variable à expliquer $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, et la moyenne de l'échantillon pour la variable explicative $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
Les valeurs de a et b qui minimisent l'erreur quadratique sont les valeurs qui annulent les dérivées $\frac{\partial SSR}{\partial a}$ et $\frac{\partial SSR}{\partial b}$ (à condition que les dérivées secondes soient positives). En résolvant le système de deux équations à deux inconnues ainsi défini, on trouve que les estimateurs de a et de b qui minimisent l'erreur quadratique sont :

$$\widehat{a} = \overline{Y} - \widehat{b}\overline{x} \tag{6.6}$$

$$\widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{6.7}$$

$\widehat{a}$ et $\widehat{b}$ se calculent à partir des observations de x et y dans l'échantillon.

6.3.1.3 Test d'hypothèse sur le modèle

On cherche à tester l'hypothèse H0 : $b = 0$ contre H1 : $b \neq 0$. On peut noter que sous H0, la valeur prédite de Y est une constante a qui s'estime comme la moyenne empirique des Y, $\overline{Y}$.

Décomposition de la variance. Le modèle de régression linéaire propose d'expliquer les variations de Y à partir de celle de x. Pour cela, on peut décomposer les variations de Y en deux sources de variation, les variations expliquées par les variations de la variable explicative et les variations résiduelles. Ci-dessous, on note $\widehat{Y}_i = \widehat{a} + \widehat{b}x_i$ la prédiction du modèle, et $\overline{Y}$ la prédiction du modèle sous l'hypothèse H0.

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

Si l'on note $SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSres = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ et $SSreg = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$, on peut écrire :

$$SST = SSreg + SSres.$$

Statistique de test et loi sous H0 . Sous l'hypothèse H0, la statistique de test

$$F = \frac{\frac{SSreg}{1}}{\frac{SSres}{n-2}}$$

suit une loi de Fisher $\mathcal{F}_{1,n-2}$.

6.3.1.4 Adéquation du modèle aux données

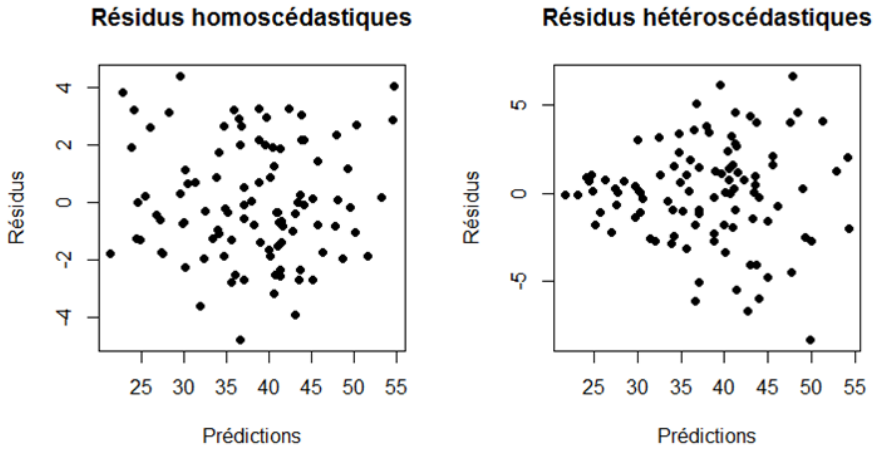


FIGURE 6.7 – **Vérification l'homoscédasticité des résidus**. On doit vérifier que la variance des résidus ne dépend pas des valeurs prédites. Sur ces graphiques, on ne lit pas directement la variance mais on peut se faire une idée de la variance en observant l'écartement des résidus à la moyenne (0). Sur le graphique de gauche, les résidus sont homoscédastiques. Sur le graphique de droite, la variance des résidus augmente avec les valeurs prédites.

Afin d'évaluer la qualité du modèle, il est nécessaire : 1) de vérifier *a posteriori* si les hypothèses faites par le modèle sont vérifiées ; 2) d'estimer la part de la variance de Y expliquée par le modèle.

Analyse des résidus. Une des hypothèses sous-jacentes au modèle est que les résidus sont indépendants, gaussiens et identiquement distribués. On fait notamment l'hypothèse que la moyenne et la variance des résidus ne dépendent pas des valeurs prédites. Lorsque la variance ne dépend pas des valeurs prédites, on parle d'*homoscédasticité*. Dans le cas inverse, on parle d'*hétéroscédasticité*. Pour cela, on pourra examiner un Q-Q plot des résidus standardisés ainsi que le graphique représentant les résidus observés $e_i = y_i - (\widehat{a} + \widehat{b}x_i)$ en fonction des valeurs prédites $\widehat{y_i} = \widehat{a} + \widehat{b}x_i$ où $\widehat{a}$ et $\widehat{b}$ sont les estimations calculées selon les équations (6.6) et (6.7).

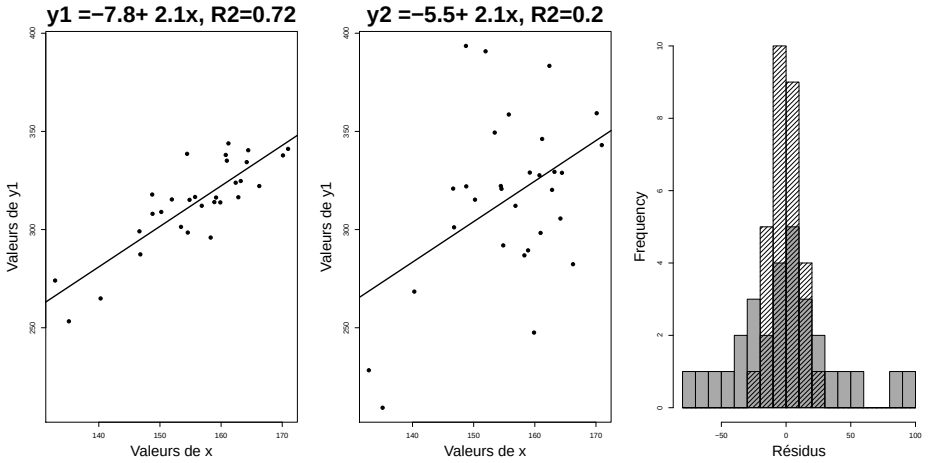


FIGURE 6.8 – **Modèles de régression pour deux jeux de données simulées.** Les équations des deux droites sont très proches, mais les points observés sont moins dispersés pour $y1$ que pour $y2$. La part de la variance expliquée est plus élevée dans le premier cas. Le graphe de droite représente la distribution empirique des résidus pour la variable $y1$ (hachuré) et pour $y2$ (gris).

Coefficient de détermination. La part de la variance expliquée par le modèle est :

$$R^2 = \frac{SSreg}{SST}.$$

On l'appelle le *coefficient de détermination*. Il varie entre 0 et 1.
La figure 6.8 présente deux modèles de régression linéaire, estimés à partir de deux jeux de données différents.

Estimateur de la variance résiduelle. Un estimateur de la variance résiduelle du modèle est :

$$\widehat{\sigma^2} = \frac{SSres}{n-2}$$

6.3.1.5 Tests et intervalle de confiance des paramètres

Loi des estimateurs $\widehat{a}$ et $\widehat{b}$, intervalles de confiance et tests sur les paramètres. Les estimateurs des paramètres de a et de b sont des variables aléatoires. La loi de ces estimateurs est donnée respectivement par :

$$\frac{\widehat{a} - a}{\widehat{Var}(\widehat{a})} \sim \mathcal{T}_{n-2} \quad \text{et} \quad \frac{\widehat{b} - b}{\widehat{Var}(\widehat{b})} \sim \mathcal{T}_{n-2},$$

où $\widehat{Var}(\widehat{a})$ et $\widehat{Var}(\widehat{b})$ sont les variances de $\widehat{a}$ et $\widehat{b}$.
On peut en déduire un intervalle de confiance d'ordre $1-\alpha$ pour chacun des paramètres :

$$IC(a) = [\widehat{a} - t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{a})}; \widehat{a} + t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{a})}]$$

$$IC(b) = [\widehat{b} - t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{b})}; \widehat{b} + t_{1-\alpha/2}\sqrt{\widehat{Var}(\widehat{b})}].$$

On peut également tester la nullité de chacun des paramètres. Par exemple, pour la nullité du paramètre a :

$$\text{H0} : a = 0 \text{ contre H1} : a \neq 0$$

On peut prendre comme statistique de test : $T = \dfrac{\widehat{a}}{\sqrt{\widehat{Var}(\widehat{a})}}$. Sous l'hypothèse H0, la statistique de test suit une loi de Student $\mathcal{T}_{n-2}$.

Sortie R

La figure 6.9 représente la sortie obtenue avec le logiciel R en utilisant la fonction `lm(rendement~eau,data=tab)`.

Dans la table `tab`, `rendement` est une colonne contenant les observations du rendement et `eau` une colonne contenant les apports en eau. La ligne `intercept` correspond à ce qui concerne le paramètre a. La ligne `eau` correspond à ce qui concerne la paramètre b (le paramètre qui multiplie l'apport en eau dans le modèle). La colonne `Estimate` donne les estimations des deux paramètres a et b, la colonne `Std.Error` donne l'estimation des erreurs standards des estimateurs $\sqrt{\widehat{Var}(\widehat{a})}$ et $\sqrt{\widehat{Var}(\widehat{b})}$. Les colonnes `t-value` et `p-value` donnent respectivement la valeur observée de la statistique de test et de sa p-value pour les tests de nullité de a et de b décrits ci-dessus. Dans le second encadré, `Residual standard error` est une estimation de σ et `Multiple R-squared` fourni le coefficient de détermination R^2.

L'estimation de a (7,75) correspond à l'ordonnée à l'origine. L'estimation de b (4,73) correspond à la pente de la droite. En observant les p-value, on constate que le paramètre a n'est pas significativement différent de 0 alors que l'on met en évidence que la paramètre b est significativement différent de 0. On peut donc interpréter sa valeur : si l'on augmente d'une unité les apports en eau, le rendement augmentera de 4,73. Indépendamment du fait que a n'est pas significativement différent de 0, on peut remarquer que l'ordonnée à l'origine n'a pas de sens biologique : le modèle a été validé pour des valeurs de l'apport en eau entre 3 et 9 et la relation linéaire n'est probablement plus vraie si les apports en eau sont trop faibles. D'après ce modèle, les variations des apports en eau expliquent $R^2 = 48\%$ de la variance du rendement. On peut donc se demander si d'autres variables ne pourraient pas apporter une information supplémentaire qui permettrait d'améliorer la qualité de notre prédiction.

```
> reg <- lm(rendement~eau,data=tab)
> summary(reg)

Call:
lm(formula = rendement ~ eau, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-11.1109  -4.2246  -0.0836   3.5482  13.6164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7473     5.1099   1.516    0.141
eau           4.7273     0.9352   5.055 2.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.667 on 28 degrees of freedom
Multiple R-squared:  0.4771, Adjusted R-squared:  0.4585
F-statistic: 25.55 on 1 and 28 DF,  p-value: 2.388e-05
```

FIGURE 6.9 – **Sortie R pour la régression linéaire**

6.3.2 Régression linéaire multiple

Souvent les variations de la variable expliquée peuvent dépendre linéairement des variations de plusieurs variables et non d'une seule comme dans le chapitre précédent. Par exemple, le rendement peu dépendre non seulement des apports en eau mais également des apports azotés, de la température ou de l'acidité du sol. On utilisera cet exemple pour présenter le formalisme du modèle de régression multiple.

6.3.2.1 Le modèle

Comme dans le cas de la régression simple, on note Y_i le rendement de la $i^{\text{ème}}$ parcelle échantillonnée ($i = 1$ à n). Cette fois-ci, on considère que le rendement dépend linéairement de l'apport en eau dans la parcelle x_{1i}, du régime azoté x_{2i}, de l'acidité du sol x_{3i} et de la température x_{4i}. Le modèle précédent peut donc être modifié comme suit :

$$Y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + \epsilon_i$$

où, comme dans le cas du modèle de régression simple, les résidus ϵ_i sont des variables aléatoires gaussiennes indépendantes et identiquement distribuées selon $\mathcal{N}(0; \sigma^2)$.

6.3.2.2 Test de Fisher

Les hypothèses testées sont :

$$\text{H0} : b_1 = b_2 = ... = b_p = 0 \text{ contre H1} : \exists j, \quad b_j \neq 0$$

Ce test repose sur la décomposition de la variance présentée à la section précédente, qui reste valide dans le cas du modèle de régression multiple. La composante $SSreg$ a p degrés de liberté, la composante $SSres$ a $n - p - 1$ degrés de liberté. La statistique de test est :

$$F = \frac{SSreg/p}{SSres/(n-p-1)}.$$

Sous l'hypothèse H0, cette statistique suit une loi de Fisher $F_{p,n-p-1}$. Pour choisir la zone de rejet et effectuer le test, on procède comme pour les tests ANOVA présentés plus haut.

Validation des conditions d'application. Comme précédemment, il faut vérifier que les résidus sont bien indépendants, gaussiens et identiquement distribués en examinant les résidus observés. De plus, on peut faire un test de Fisher pour vérifier si les prédictions du modèle sont meilleures que si l'on prédisait le rendement uniquement par le rendement moyen.

6.3.2.3 Estimateurs des paramètres et tests de nullité de chacun des paramètres

Nous ne détaillerons pas ici les estimateurs de chacun des paramètres. Si l'on note $b_0 = a$ et p le nombre de variables explicatives (ici $p = 4$), la loi de ces estimateurs est donnée par :

$$T = \frac{\widehat{b_j} - b_j}{\widehat{Var}(\widehat{b_j})} \sim \mathcal{T}_{n-p} \text{ avec } j = 1, ..., p$$

On peut donc construire les p tests :

$$\text{H0} : b_j = 0 \text{ contre H1} : b_j \neq 0 \text{ avec } j = 1, ..., p$$

6.3.2.4 Sélection de modèle

Problématique. On peut chercher à trouver la combinaison de variables explicatives qui prédisent au mieux les observations. Il s'agit d'un problème compliqué puisque le nombre de combinaisons possibles augmente très rapidement avec le nombre de descripteurs.
On peut par exemple vouloir comparer un modèle qui prédit le rendement à partir de la teneur en eau et en azote du sol (Ma), à un modèle qui prédit le rendement à partir de la seule température (Mb) :

$$\text{Ma} \quad : \quad Y_i = a + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$$

$$\text{Mb} \quad : \quad Y_i = a + b_4 x_{4i} + \epsilon'_i$$

Comme les variables explicatives ne sont pas les mêmes, les résidus seront différents entre les deux modèles. Un critère de choix pourrait être le modèle ayant la plus faible variance résiduelle, ou le meilleur coefficient de détermination. Cependant, lorsque l'on rajoute des descripteurs ayant individuellement un effet significatif sur le caractère d'intérêt, la variance résiduelle diminue forcément puisque l'on décrit de mieux en mieux les observations. Ainsi, si l'on connaissait absolument tous les facteurs qui déterminent le rendement du blé, on serait capable de prédire Y sans erreur. *Un bon compromis consiste à choisir le modèle qui décrit au mieux les données avec un minimum de paramètres.*

Vraisemblance. La vraisemblance d'un modèle est la probabilité d'observer les valeurs d'un échantillon pour des valeurs données des paramètres. On peut la calculer en utilisant une estimation des paramètres du modèle et la loi de probabilité des variables aléatoires. Dans le cadre d'un modèle linéaire, si les résidus sont gaussiens, il existe une relation entre la vraisemblance du modèle et la somme des carrés d'écart résiduels.

Critère d'Akaike. Le critère AIC se calcule comme une différence entre la log-vraisemblance du modèle et une pénalité proportionnelle au nombre de paramètres à estimer :

$$AIC = 2p - 2\ln(L)$$

où p est le nombre de paramètres du modèle et L la vraisemblance. Le meilleur modèle est celui qui a le plus faible AIC.
Une stratégie pour la régression linéaire multiple consiste donc à tester tous les modèles et à choisir le meilleur sur la base du critère AIC.
Il existe d'autre stratégies, qui consistent à réaliser la sélection de modèles de façon itérative. On peut partir du modèle le plus simple, et essayer à chaque étape d'enlever ou de rajouter un descripteur, en se basant sur le critère AIC pour le choix du descripteur à retirer ou à rajouter.
On aboutit en principe au sous-ensemble de descripteurs qui prédisent au mieux les observations sur la variables d'intérêt.

Exemple. Dans l'exemple du blé, voici les résultat de la dernière étape de sélection. La procédure a rajouté succesivement toutes les variables explicatives. La dernière étape consiste à essayer d'en retirer une :

```
Step: AIC=109.11
rendement = eau + azote + temp + aci
```

Model	Df	Sum of Sq	RSS	Cp
<none>		816.31	388.15	109.11
- aci	1	39.59	855.90	108.53
- temp	1	80.59	896.90	109.93
- eau	1	232.35	1048.66	114.62
- azote	1	384.21	1200.52	118.68

Dans la colonne `Model`, le cas `<none>` correspond au modèle choisi à l'étape précédente. Dans ce cas, c'est le modèle complet. Les autres lignes correspondent au modèle complet moins un des descripteurs. Pour chaque sous-modèle, `Df` est le nombre de degrés de liberté gagnés ou perdus, `Sum of Sq` est la somme des carrés des écarts du modèle ($SSreg$), `RSS` est la somme des carrés des écarts résiduels ($SSres$) et `Cp` donne le critère d'Akaike pour chaque sous-modèle.
On constate que le meilleur modèle est celui qui comporte l'ensemble des descripteurs sauf l'acidité du sol ($AIC = 108,53$).
L'étape suivante consiste à examiner les coefficients estimés.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-19.98903	9.58214	-2.086	0.04693 *
eau	3.54548	1.05064	3.375	0.00233 **
azote	0.07036	0.02078	3.386	0.00226 **
temp	1.26031	0.85770	1.469	0.15372

On constate que tous les coefficients sont positifs. Les descripteurs sont donc corrélés positivement au rendement. On peut remarquer que le coefficient de la variable température n'est pas significativement différent de zéro. Il est probable que le fait d'ajouter cette variable permet d'augmenter la vraisemblance du modèle (donc diminuer l'AIC) même si on ne peut pas mettre en évidence que le coefficient est significativement différent de 0 avec un test.

6.4 Analyse de covariance

6.4.1 Exemple : relations d'allométrie chez les poissons

L'article de Yu et al. (BMC Evolutionary Biology, 2014, 14 :178) présente une étude sur les relations allométriques entre la taille du corps et celle du cerveau chez les vertébrés. A l'intérieur d'une espèce, on observe typiquement une relation allométrique de la forme :

$$W = CP^b$$

où W est la masse du cerveau, P la masse corporelle et C une constante. b est appelé le *coefficient d'allométrie*. On peut effectuer un changement de variables pour passer sur une échelle logarithmique :

$$\ln(W) = \ln(C) + b\ln(P)$$

Sur l'échelle log, on voit que la masse du cerveau s'accroît de façon proportionnelle à la masse corporelle. Les auteurs s'intéressent aux facteurs pouvant expliquer des changements entre espèces du coefficient d'allométrie et permettre de comprendre la différence importante de taille du cerveau entre par exemple les mammifères et les oiseaux. Leur hypothèse est qu'il y a un coût énergétique à l'accroissement de la taille du cerveau. Ils montrent que l'on peut différencier les espèces endothermes (mammifères, oiseaux, insectivores) des espèces ectothermes (poissons, reptiles, amphibiens), avec des coefficients plus élevés chez les espèces endothermes ($C = 0,078$, $b = 0,689$) que chez les ectothermes ($C = 0,014$, $b = 0,578$). Pour vérifier le rôle de la température dans les relations allométriques corps-cerveau, ils échantillonnent des données de la base (`http://fishbase.org`) pour représenter les différentes espèces de poissons et des environnements contrastés pour la température de l'eau (Table 6.1).

Environnement	Température moyenne	Nb espèces
Polaire	1°C	34
Tempéré	15°C	70
Tropical	25°C	88
Sub-tropical*	20–30°C	17

* Les espèces prises en compte dans cet environnement sont 17 espèces de requins qui maintiennent leur température corporelle entre 20 et 30 °C grâce à un système de contractions musculaires.

TABLE 6.1 – **Echantillonnage de 209 espèces de poissons vivant à différentes températures.**

La figure 6.10 est tirée de l'article et montre les relations d'allométrie pour chaque groupe d'espèces, sur une échelle logarithmique.
On observe pour chaque environnement une relation linéaire entre poids corporel et masse du cerveau. Cependant, les espèces vivant dans des environnement différents semblent se comporter de façon différente, avec des différences pour la constante C et pour le coefficient d'allométrie b.

6.4.2 Modèles emboîtés et tests d'hypothèses

Pour simplifier, on appelle Y la variable aléatoire correspondant au logarithme de la masse du cerveau, X le logarithme du poids corporel, et $a = \ln(C)$. On dispose d'un échantillon aléatoire d'espèces dans chaque environnement. On indice les environnements par i ($i = 1, ..., p$, $p = 4$), et les espèces au sein d'un environnement par j. Le couple (Y_{ij}, X_{ij}) constitue un couple de variables aléatoires continues. On peut poser le modèle suivant, qui prend en compte une différence selon les environnements pour la constante a et le coefficient d'allométrie b :

$$Y_{ij} = a_i + b_i X_{ij} + \epsilon ij \tag{6.8}$$

avec ϵ_{ij} la variable aléatoire décrivant les résidus du modèle, que l'on suppose indépendants et de même loi $\mathcal{N}(0; \sigma^2)$.

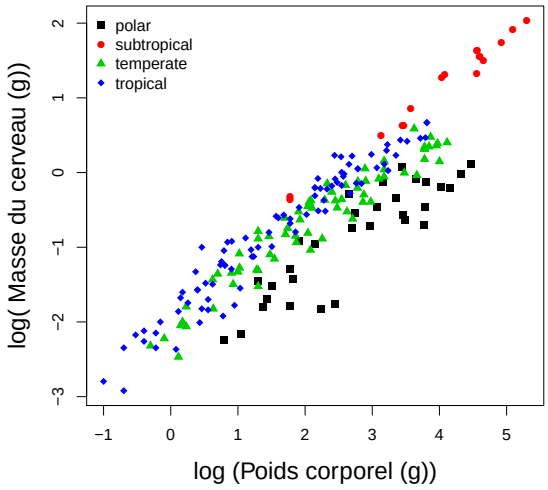


FIGURE 6.10 – **Relations d'allométrie corps-cerveau chez les poissons.** Chaque point correspond à une espèce, et chaque couleur à un type d'habitat, comme décrit dans le Tableau 6.1. On rappelle que l'environnement "sub-tropical" correspond à un échantillonnage de 17 espèces de requins.

6.4.2.1 Test d'hypothèses

Les questions que l'on se pose sont les suivantes :

— Existe-t-il des différences entre environnements pour l'ordonnée à l'origine a ? Les deux hypothèses sont

$$\text{H0} \quad : \quad a_1 = a_2 = ... = a_p = a$$

$$\text{H1} \quad : \quad \exists (i, i'), \quad a_i \neq a_{i'}$$

— Existe-t-il des différences entre environnements pour le coefficient d'allométrie b ? Les deux hypothèses sont

$$\text{H0} \quad : \quad b_1 = b_2 = ... = b_p = b$$

$$\text{H1} \quad : \quad \exists (i, i'), \quad b_i \neq b_{i'}$$

Considérons la seconde question. On peut décrire chaque hypothèse par un modèle :

$$\text{M0} \quad : \quad Y_{ij} = a_i + bX_{ij} + \epsilon_{ij}$$

et

$$\text{M1} \quad : \quad Y_{ij} = a_i + b_i X_{ij} + \epsilon_{ij}^{'}$$

Si H0 est vraie, alors la variance résiduelle σ^2 est la même pour les deux modèles. Par contre, si H0 est fausse, on s'attend à ce que la variance résiduelle du modèle M0 soit plus grande que celle du modèle M1, car il décrit moins bien les données.
M0 et M1 sont deux modèles emboîtés. En effet, le modèle M0 est un sous-modèle de M1, avec $b_i = \text{Cte}$. On peut donc construire une statistique de test qui permettra de comparer les variances résiduelles des deux modèles.
Dans un modèle linéaire, la variance résiduelle est estimée par la somme des carrés des écarts résiduels : Pour le modèle M0, $SSR_0 = \sum_i \sum_j (Y_{ij} - \widehat{Y}_{ij_{M0}})^2$, où $\widehat{Y}_{ij_{M0}} = \widehat{a}_i + \widehat{b} X_{ij}$ est la valeur prédite par le modèle M0.
Pour le modèle M1, $SSR_1 = \sum_i \sum_j (Y_{ij} - \widehat{Y}_{ij_{M1}})^2$, où $\widehat{Y}_{ij_{M1}} = \widehat{a}_i + \widehat{b}_i X_{ij}$ est la valeur prédite par le modèle M1.
Si H0 est vraie, alors la différence $(SSR_0 - SSR_1)$ doit être proche de zéro. On peut remarquer que $SSR_0 - SSR_1$ correspond bien à la somme des carrés des écarts correspondant aux effets b_i.

Dans ce cas, la statistique de test est

$$F = \frac{\frac{SSR_0 - SSR_1}{p-1}}{\frac{SSR_1}{n-p-1}} \sim_{H0} \mathcal{F}_{p-1,n-p-1}.$$

6.4.2.2 Comparaison de deux modèles : tableau de résultat.

Le tableau de résultats se présente de la façon suivante. Les chiffres donnés ici correspondent aux données d'allométrie décrites dans le Tableau 6.1 pour $n = 209$ espèces.

Model	res.df	SSR	model.df	SSM	F	pvalue
M0	204	11.30				
M1	201	10.78	3	0.52	0.23	0.023

Pour chaque modèle, on donne la somme des carrés d'écarts résiduels (`SSR`) et les degrés de liberté correspondant (`res.df`). Pour le modèle le plus complet, on indique le nombre de paramètres supplémentaires (`model.df`) et l'écart $SSM = SSR_{\text{M0}} - SSR_{\text{M1}}$, puis la valeur de la statistique de test `F` et la `pvalue` associée. On retrouve bien ici que le modèle M1 a trois paramètres de plus que le modèle M0 (`model.df`). Le fait de considérer des pentes différentes permet de réduire légèrement la somme des carrés d'écarts résiduels (comparer la colonne `SSR` pour M0 et M1). Ici, sous H0, la statistique F suit une loi $\mathcal{F}_{3,201}$. La p-value est de $0,023$. On rejette donc H0 au seuil 5%.

6.4.2.3 Comparaison de plusieurs modèles emboîtés et critère AIC

Dans l'exemple considéré, on peut écrire plusieurs sous-modèles emboîtés, pour tester différentes hypothèses, sur les coefficients d'allométrie ou sur les constantes. Comme précédemment, on peut alors comparer les modèles deux à deux à l'aide d'un test de Fisher. On peut aussi choisir le modèle qui décrit le mieux les données à l'aide d'un critère comme le critère d'Akaike (voir la section sur la régression multiple).
Le tableau ci-dessous montre différents modèles dérivés du modèle (6.8), le nombre de paramètres estimés (incluant σ^2) et la valeur obtenue pour l'AIC.

Modèle	Valeur prédite	Nb paramètres	AIC
M111	$\widehat{Y}_{ij} = \widehat{a}_i + \widehat{b}_i X_{ij}$	9	–8,50
M110	$\widehat{Y}_{ij} = \widehat{a}_i + \widehat{b} X_{ij}$	6	–4,66
M011	$\widehat{Y}_{ij} = \widehat{a} + \widehat{b}_i X_{ij}$	6	8,32
M100	$\widehat{Y}_{ij} = \widehat{a}_i$	5	508,3
M001	$\widehat{Y}_{ij} = \widehat{a} + \widehat{b} X_{ij}$	3	202,6

TABLE 6.2 – **Analyse de covariance : comparaison de modèles emboîtés.** Pour chaque modèle, la variance résiduelle peut-être différente. Le modèle complet (M111) est ici celui qui a le plus faible AIC.

La figure 6.11 montre à titre d'exemple les valeurs prédites pour trois des modèles testés : le modèle complet (M111), le modèle M110 et le modèle M100. On voit bien que le meilleur modèle est le modèle le plus complet, M111, qui prédit à la fois des différences entre environnements pour le coefficient d'allométrie et pour la constante. La représentation graphique du modèle M110 montre les valeurs prédites en supposant un unique coefficient d'allométrie pour toutes les espèces. Visuellement, on a l'impression d'un ajustement correct, ce qui explique la p-value relativement élevée de 0,02 lorsque l'on compare les modèles M111 et M110. Enfin, la représentation graphique M100 met en évidence les différences d'ordonnées à l'origine, avec surtout une différence importante entre les requins subtropicaux et les autres espèces, quel que soit leur environnement.

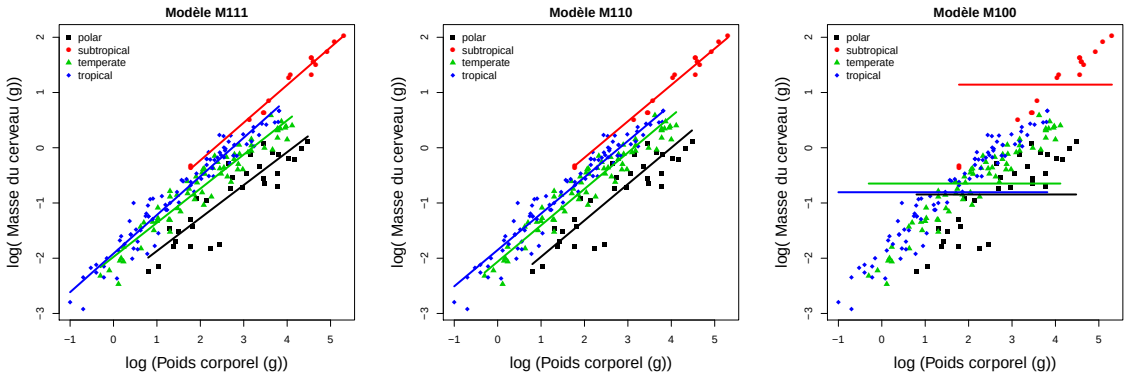


FIGURE 6.11 – **Analyse de covariance : comparaison de modèles.** Chaque graphique représente les données (relation entre le poids corporel et la masse du cerveau) et les valeurs prédites pour un modèle. Les points sont les valeurs observées, les segments de droite les valeurs prédites dans chaque environnement. Le code de couleurs correspond aux différents environnements.

6.4.2.4 Validation des conditions d'application

Une fois que l'on a choisi le meilleur modèle, il reste à valider les conditions d'application. Ici, comme toujours dans le cadre du modèle linéaire, il faut vérifier l'indépendance et la normalité des résidus. Pour cela, on représente graphiquement la relation entre valeurs prédites et résidus, et on réalise un Q-Q plot des résidus.

6.4.3 Estimation des paramètres du modèle

Dans notre exemple, le meilleur modèle est le modèle complet. On peut ensuite se demander, pour chaque type de paramètre (constante et coefficients d'allométrie), quels coefficients sont significativement différents des autres. On peut utiliser les estimations des coefficients et leur intervalle de confiance pour réaliser des tests de conformité (par exemple, $a_1 = 0$ contre $a_1 \neq 0$) ou des tests de comparaison deux à deux (par exemple $a_1 - a_2 = 0$ contre $a_1 - a_2 \neq 0$).
Le logiciel R propose un certain nombre de tests statistiques sur les coefficients estimés. Il est important de savoir déchiffrer les sorties de la fonction `summary`. Par défaut, R choisit d'indicer les coefficients en fonction de l'ordre alphabétique du niveau des facteurs. Ici, il y a quatre environnements qui s'appellent « polar », « subtropical »,« tropical » et « temperate » dans le fichier de données. L'environnement polaire est le premier dans l'ordre alphabétique et sera considéré comme niveau de référence. Les paramètres du modèle M111 sont :

Environnement	Coefficients
Polaire	a_1, b_1
Subtropical	a_2, b_2
Tempéré	a_3, b_3
Tropical	a_4, b_4

R donne la valeur estimée des coefficients $\widehat{a}_1$ et $\widehat{b}_1$, puis les différences entre les autres coefficients et le niveau de référence. Une différence entre deux coefficients est appelée un contraste. Pour chaque contraste, R donne la valeur estimée, l'erreur standard, et effectue un test de conformité par rapport à une valeur de zéro. Le logiciel affiche la valeur de la statistique de test (test de Student) et la p-value.

Conclusion biologique de l'étude. En interprétant les résultats donnés dans le tableau 6.3, on voit que toutes les constantes sont plus faibles dans l'environnement polaire. Pour une même masse corporelle, la taille moyenne du cerveau sera plus grande chez des espèces vivant en milieu tropical ou tempéré que chez celles vivants dans un milieu polaire. Pour les requins vivant en milieu subtropical, elle sera beaucoup plus grande.

Estimation	Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
a_1	(Intercept)	-2.46539	0.11	-21.53	$< 2e-16$***
b_1	X	0.59653	0.04	15.07	$< 2e-16$***
$a_2 - a_1$	envsubtropical	0.87230	0.25	3.39	0.00085**
$a_3 - a_1$	envtemperate	0.48939	0.13	3.81	0.00018***
$a_4 - a_1$	envtropical	0.54941	0.12	4.52	0.00001***
$b_2 - b_1$	envsubtropical:X	0.08587	0.07	1.26	0.21085
$b_3 - b_1$	envtemperate:X	0.02137	0.05	0.463	0.64363
$b_4 - b_1$	envtropical:X	0.10262	0.04	2.31	0.02167*

TABLE 6.3 – **modèle 111 : tests sur les coefficients estimés obtenu avec la fonction *lm* du logiciel R.**

On voit que l'on ne met pas de différence en évidence entre les coefficients d'allométrie des espèces polaires et des espèces subtropicales ou tempérées (les différences $b2 - b1$ et $b3 - b1$ ne sont pas significatives). Par contre, les espèces vivant en milieu tropical ont un coefficient d'allométrie plus élevé que les espèces des milieux polaires. Pour ces espèces, on peut voir effectivement sur la figure 6.11 que la pente de la droite de régression est plus forte.

Chapitre 7

Analyse en composantes principales

7.1 Introduction

En biologie, on dispose souvent de données quantitatives associées à des individus. Pour étudier ce type de données, on représente classiquement les individus par des points placés dans un repère orthogonal selon les valeurs prises par chacune des variables (figure 7.1).

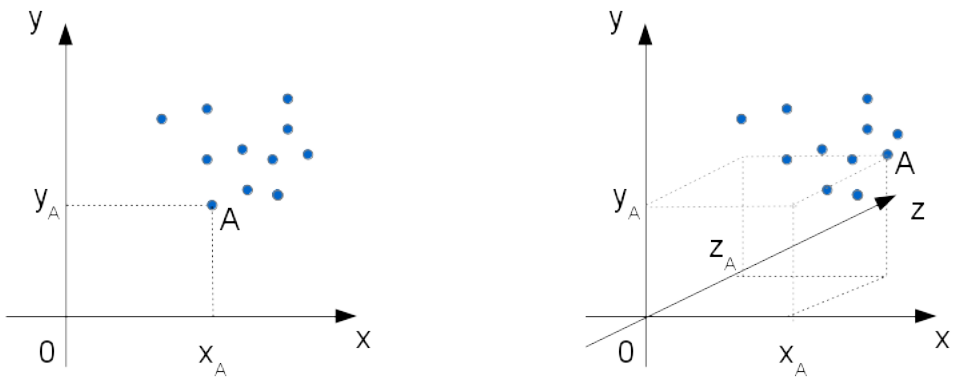


FIGURE 7.1 – **Repères à deux et trois dimensions**

Si cette représentation est parfaitement adaptée aux jeux de données composés de deux à trois variables, elle s'avère plus complexe lorsque l'on s'intéresse à un plus grand nombre de variables. Dans ce cas, on va se tourner vers des méthodes d'analyse multivariée qui vont prendre en compte la distribution de plusieurs variables. Elles vont permettre de simplifier le jeu de données en identifiant les combinaisons de variables les plus informatives : les composantes principales.

En réalisant une analyse en composantes principales, on va pouvoir
- Visualiser les corrélations entre les variables
- Visualiser les relations entre les individus
- Repérer les variables qui peuvent discriminer les individus.

7.2 Principe

L'analyse en composantes principales est une méthode d'analyse factorielle de statistique multivariée. A partir de données composées d'individus (en ligne) décrits par des variables quantitatives (en colonne), la méthode va calculer des variables latentes, c'est-à-dire des combinaisons linéaires des variables observées, et ainsi identifier celles qui résument au mieux la variance contenue dans le jeu de données initial. L'ACP permet aussi de réduire le nombre de variables descriptives tout en limitant la perte d'information.

On peut aborder l'ACP par une approche géométrique. A partir de données positionnées dans un référentiel de départ, on va les représenter dans un nouveau système de coordonnées maximisant les distances entre les individus.
L'exemple de la figure 7.2 illustre le changement de repère permettant de maximiser la dispersion de l'objet selon le premier axe Y_1, puis Y_2, etc., du nouveau repère Y. L'ACP ne déforme pas l'objet initial car les distances entre les points sont conservées, mais modifie l'angle de vue sur cet objet.

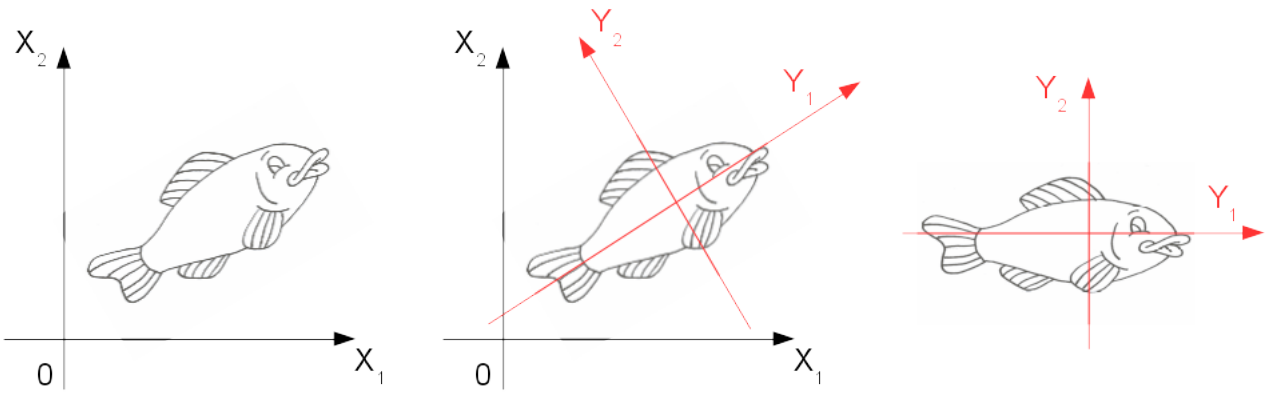


FIGURE 7.2 – **Principe du changement de repère réalisé lors d'une ACP.**

7.3 Méthode

7.3.1 Rappels sur les distances entre deux points

L'ACP est une méthode basée sur les distances qui utilise la distance euclidienne entre les individus du jeu de données.
La distance euclidienne $d(A, B)$ entre deux points A et B situés dans un espace à deux dimensions (figure 7.3) est donnée par la relation :

$$d(A, B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

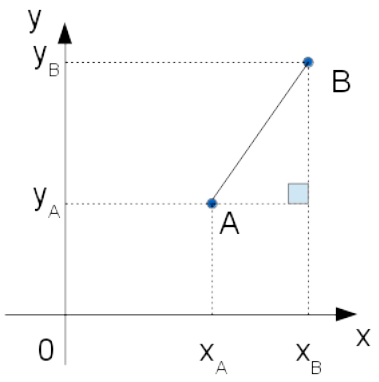


FIGURE 7.3 – **Distance entre deux points dans un repère en deux dimensions**

Dans un espace à J dimensions, on peut généraliser cette expression :

$$d(A_i, A_{i'}) = \sqrt{\sum_{j=1}^{J} (x_{i'j} - x_{ij})^2}$$

- A_i et $A_{i'}$ sont deux points dans le repère X
- x_{ij}, la coordonnée du point A_i sur l'axe X_j
- $x_{i'j}$, la coordonnée du point $A_{i'}$ sur l'axe X_j

7.3.2 Le repère initial

On considère un jeu de données composé de n individus et de J variables. On peut représenter ces données par un nuage de points en J dimensions sur lequel chaque point représente un individu.

Les individus

On peut décrire le nuage de points des individus par son centre de gravité G (*i.e.* le point dont les coordonnées sont les valeurs moyennes des variables) et son inertie I_G (*i.e.* sa dispersion).

Les coordonnées du **centre de gravité** sont :

$$G = [x_{.1}, x_{.2}, x_{.3}, ...x_{.J}]$$

où $x_{.j}$ est la moyenne des coordonnées des n points sur l'axe X_j.

L'**inertie** d'un nuage de points se définit par la somme des carrés des distances de chaque point au centre de gravité, pondérés par le poids m_i de chaque point :

$$I_G = \sum_{i=1}^{n} m_i d^2(G, x_i).$$

Ici, on considère que chaque individu a le même poids et on pose $\sum_{i=1}^{n} m_i = 1$, donc $m_i = \frac{1}{n}$. On obtient alors :

$$I_G = \sum_{i=1}^{n} \frac{1}{n} d^2(G, x_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} (x_{ij} - x_{.j})^2 = \sum_{j=1}^{J} \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_{.j})^2 = \sum_{j=1}^{J} s_j^2$$

On retrouve la variance des individus selon l'axe X_j, notée s_j^2. Dans le cas de variables centrées réduites, la variance de chaque variable est égale à 1, donc $I_G = J$ (nombre de variables).

Les variables

On peut décrire les variables et leurs relations en calculant les variances de chacune d'elles et les covariances de chaque couple de variables.
Ces données sont stockées dans une matrice de variances-covariances (matrice carrée de dimension $J \times J$). La diagonale de cette matrice contient les variances (toujours positives) et les autres cases contiennent les covariances entre paires de variables. La matrice est donc symétrique. Pour s'affranchir des unités de mesure, on utilise souvent les variables réduites, la matrice de variances-covariances devient dans ce cas la matrice des corrélations. Cela permet de donner le même poids à toutes les variables et que le poids ne dépendent pas de l'ordre de grandeur de la variable.

7.3.3 Calcul des nouveaux axes

A partir de la représentation des données selon les J dimensions, on va réaliser un changement de repère. On définit G comme origine du nouveau repère. Les axes $Y_1, Y_2, ..., Y_k, ..., Y_K$ vont être calculés de façon séquentielle, ils passeront tous par G et seront orthogonaux car indépendants. A noter que $K = J$, puisque le changement de repère ne modifie pas le nombre de dimensions de l'espace. La construction des nouveaux axes revient à construire de nouvelles variables, appelées variables latentes ou virtuelles, qui sont des combinaisons linéaires des variables observées du jeu de données initial, telles que :

$$Y_{ik} = a_{k1}x_{i1} + a_{k2}x_{i2} + ... + a_{kJ}x_{iJ}$$

avec :

- Y_{ik}, la coordonnée du $i^{\text{ème}}$ point sur le $k^{\text{ème}}$ axe Y
- a_{kj}, le coefficient qui associe la variable initiale j à la variable latente k du nouveau repère.

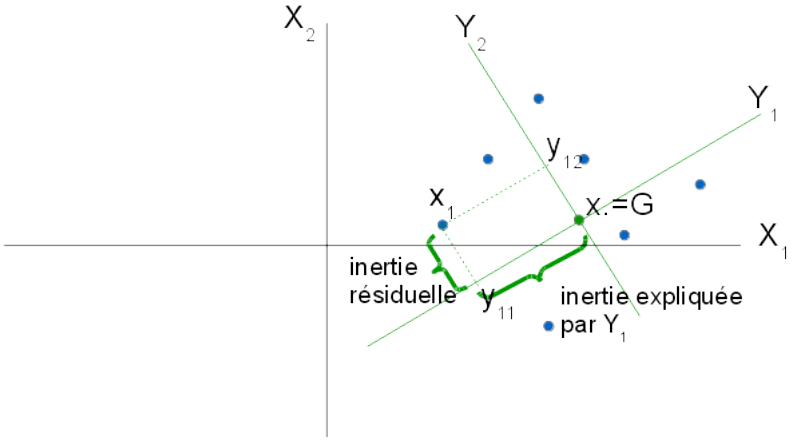


FIGURE 7.4 – **Repères X et Y et projection d'un point x_i sur le nouveau repère Y.**

7.3.3.1 L'axe principal

L'axe principal est noté $Y1$. Lors du changement de repère, l'inertie totale I_G du nuage de points reste la même mais on va chercher à maximiser l'inertie du nuage de points le long du nouvel axe Y_1, tout en minimisant l'inertie résiduelle autour de celui-ci (figure 7.4).

$$I_G = \frac{1}{n}\sum_{i=1}^{n} d^2(G, x_i) = \frac{1}{n}\sum_{i=1}^{n} d^2(G, y_{i1}) + \frac{1}{n}\sum_{i=1}^{n} d^2(y_{i1}, x_i)$$

$$I_G = \text{inertie totale} = \text{inertie expliquée par } Y_1 + \text{inertie résiduelle}$$

Il s'agit de maximiser la distance $d^2(G, y_{ik})$ sachant que

$$\sum_{k=1}^{K} I_{Y_k} = I_G = \text{cte.}$$

et que les Y_k sont indépendants. La solution passe par du calcul vectoriel et de l'optimisation sous contrainte (*cf.* Annexe). On obtient comme solution :

— les valeurs propres $\lambda_k = I_{Y_k}$

— les vecteurs propres $\overrightarrow{a_1} = (a_{11}, a_{12}, ...a_{1J}), ..., \overrightarrow{a_k} = (a_{k1}, a_{k2}, ...a_{kJ}), ..., \overrightarrow{a_K} = (a_{K1}, a_{K2}, ...a_{KJ})$.

Chaque variable virtuelle (ou axe de l'ACP) a une **valeur propre** λ_k qui représente sa variance empirique ($\lambda_k = I_{Y_k} = \sigma^2_{Y_k}$). Les $\overrightarrow{a_k}$ sont les **vecteurs propres**. Pour le premier axe, on a donc la valeur propre λ_1 et le vecteur propre $\overrightarrow{a_1}$. Les valeurs propres et vecteurs propres sont ceux de la matrice de variance-covariance des variables de départ.

7.3.3.2 Les axes suivants

Le maximum d'inertie non portée par le premier axe Y_1 sera portée par l'axe suivant Y_2, etc. Les nouvelles variables passeront par G et, comme elles sont indépendantes les unes des autres, Y_2 sera perpendiculaire à Y_1. Les axes seront donc classés par ordre d'inertie décroissante, et la somme de leurs inerties est égale à l'inertie totale du nuage de points.

$$I_G = I_{Y_1} + ... + I_{Y_k} + ... + I_{Y_K}$$

$I_G = \sum_{k=1}^{K} \sigma_k^2$ dans le cas d'une ACP sur variables non réduites
$I_G = K$ dans le cas d'une ACP sur variables réduites car chaque variable standardisée a une variance égale à 1
I_{Y_k} : l'inertie expliquée par l'axe Y_k
La part d'inertie expliquée par les p premières composantes peut donc s'écrire $\frac{\sum_{k=1}^{p} I_{Y_k}}{I_G}$.

7.3.4 Interprétation des résultats d'une ACP

7.3.4.1 Sélection des composantes principales

Les variables latentes calculées précédemment sont au nombre de K et sont ordonnées selon la décroissance de leurs valeurs propres. Pour interpréter les résultats d'une ACP, on va rechercher les variables latentes qui expliquent le mieux la dispersion du nuage de points.
Pour ce faire, on trace l'histogramme des valeurs propres, appelé aussi graphe des éboulis.

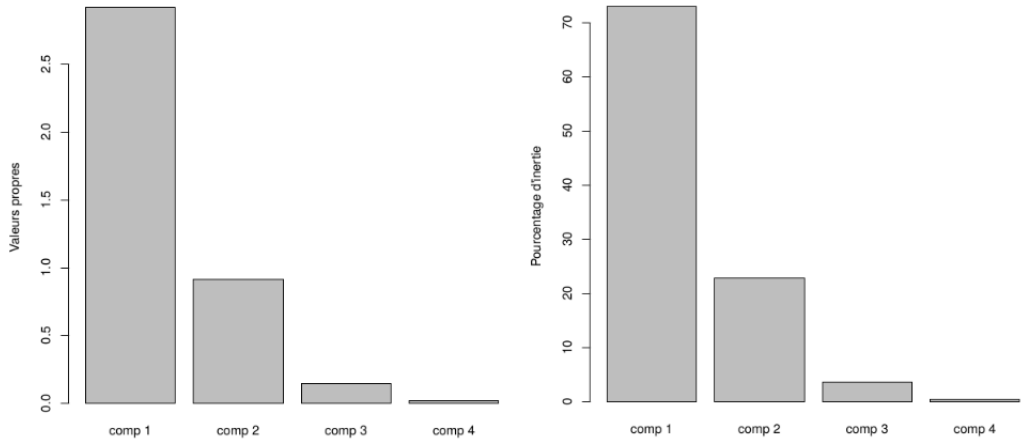


FIGURE 7.5 – **Eboulis des valeurs propres (à gauche), présenté en pourcentage d'inertie (à droite)**

En abscisse les variables latentes et en ordonnée les valeurs propres (ou le pourcentage d'inertie de chacune des composantes). L'histogramme est toujours décroissant et l'allure de l'histogramme va permettre de déterminer les v variables à prendre en considération. On peut par exemple choisir v telle que $> 50\%$ de l'inertie soit expliquée. Parfois, on observe « un coude » dans l'histogramme, c'est-à-dire que les valeurs propres diminuent brusquement. Ceci permet de repérer la composante à partir de laquelle on observe un décrochement. Sur la figure 7.5, ce critère reviendrait à sélectionner la première composantes de l'ACP. En pratique, on considérera au moins les deux premières composantes pour pouvoir représenter les variables et les individus sur un graphe 2D.

7.3.4.2 Représentation des variables

Nous avons vu précédemment que les composantes principales étaient des combinaisons linéaires des variables initiales. On représente les relations entre les différentes variables initiales et les composantes principales sur un cercle des corrélations (figure 7.6).

Interpréter les relations entre les variables initiales et composantes principales

Les deux axes perpendiculaires de ce plan représentent deux composantes principales de l'ACP : Y_1 et Y_2. Chaque variable initiale est représentée par un vecteur ayant pour origine l'origine du repère. L'extrémité de ces vecteurs est le point dont les coordonnées sur les axes principaux sont associées à la mesure de la corrélation avec chacune des composantes principales.
Cette représentation en 2D permet d'interpréter les relations entre les variables initiales et les deux composantes. En réalité, dans l'espace de dimension K, les vecteurs correspondant aux variables initiales sont tous de longueur 1 et décrivent une sphère de rayon égal à 1 (dans le cas de variables réduites).
Plus une variable sera bien représentée dans le plan de l'espace 2D, plus l'extrémité du vecteur sera proche du cercle des corrélations (figure 7.7a)

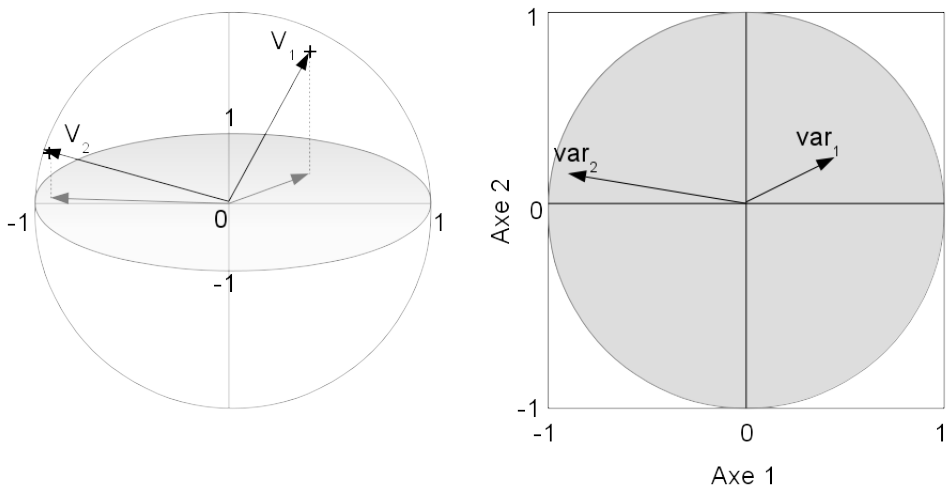


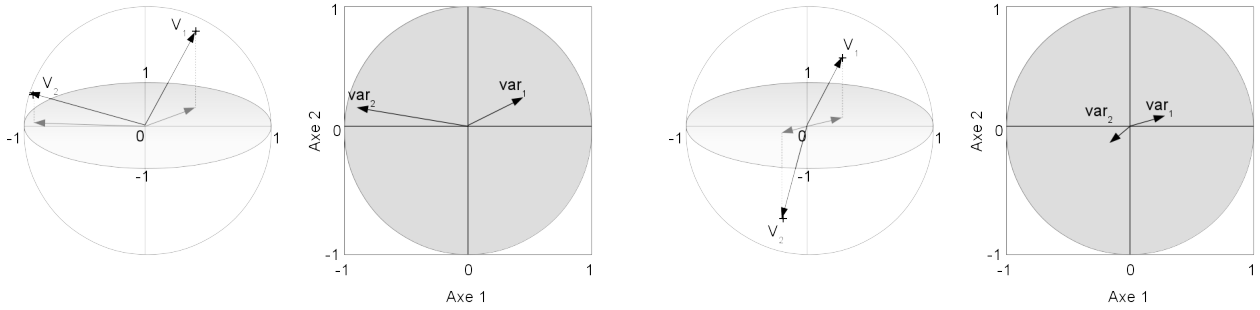
FIGURE 7.6 – **Cercle des corrélations**

Interpréter les relations entre les variables initiales

Sur le cercle des corrélations, on peut également interpréter la relation entre deux variables. Les relations entre variables projetées ne sont interprétables que lorsque les deux variables sont proches du cercle des corrélations (figure 7.7). C'est le cas pour l'exemple de la figure 7.7a mais pas pour celui de la figure 7.7b.
Dans l'espace, l'angle formé par deux vecteurs rend compte de la corrélation entre les variables qu'ils représentent suivant la relation :

$$\cos(\overrightarrow{X1}, \overrightarrow{X2}) = r(X1, X2)$$

Si les deux vecteurs se superposent la corrélation est maximale (= 1), lorsqu'ils forment un angle à 180° les variables qu'ils représentent sont parfaitement anticorrélées et lorsqu'ils sont orthogonaux les variables ne sont pas corrélées.



(a) (b)

FIGURE 7.7 – **Deux exemples de cercles des corrélations**

7.3.4.3 Représentation des individus

Les composantes principales correspondent à des nouvelles variables quantitatives que l'on va utiliser pour placer les individus du jeu de données dans un repère formé par deux composantes principales de l'ACP.
L'origine du repère est le point $G(0, 0)$ et chaque individu i est représenté par un point du plan ayant pour coordonnées sur l'axe Y_k les valeurs y_{ik}. Par projection sur chacun des axes de l'ACP, on pourra identifier les modalités des individus ayant été isolées selon chacun des axes.

Attention

- *La représentation sous forme de cercle des corrélations ne permet d'interpréter que le comportement des variables dont le vecteur est proche du bord du cercle car ce sont les points les mieux projetés dans l'espace factoriel.*
- *La projection en 2D peut faire apparaitre une proximité entre deux points qui sont en réalité distants dans l'espace.*
- *Dans le cas où on identifie plus de trois composantes principales, il sera nécessaire d'interpréter les représentations graphiques pour chacun des couples de composantes.*

7.4 Exemple d'interprétation d'une ACP

On dispose de données relatives à la morphologie des fleurs chez trois espèces d'iris (*Iris setosa*, *I. versicolor* et *I. virginica)* (en cm).

Espèce	Longueur de sépale	Largeur de sépale	Longueur de pétale	Largeur de pétale
I. versicolor	5	3,1	1,7	0,3
I. virginica	4	3,7	1,5	0,2
I. versicolor	4,2	2,8	1,8	0,2
I. setosa	4,7	3,1	1,6	0,2
...	...	...	...	...

TABLE 7.1 – **Jeu de données relatives à la morphologie des fleurs d'iris**. Sur 150 individus appartenant à trois espèces d'iris, on dispose d'une variable qualitative (l'espèce) et des mesures pour quatre variables quantitatives.

On cherche à savoir si les variables permettent de distinguer les trois espèces.
Après avoir calculé les composantes principales, on regarde le graphe des valeurs propres afin d'identifier les composantes principales à prendre en compte (figure 7.8).

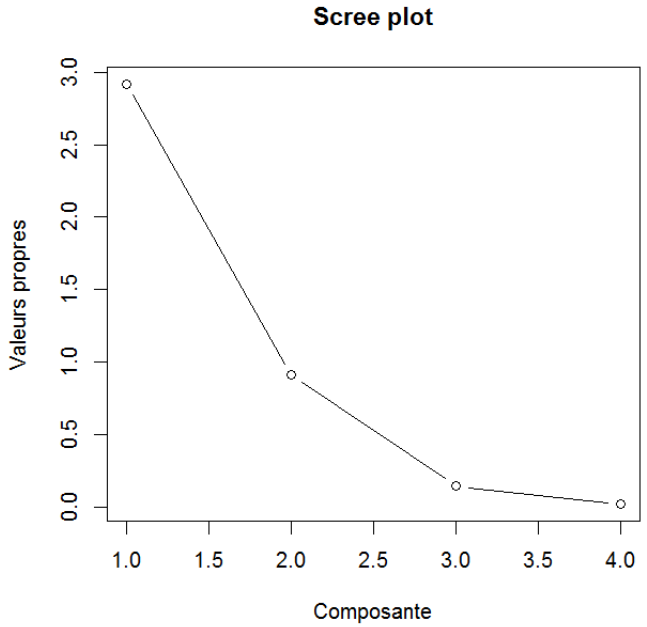


FIGURE 7.8 – **Graphe des valeurs propres sur le jeu de données Iris.**

La première composante principale semble expliquer la majeure partie de la variance du jeu de données initial. On va toutefois projeter les individus sur le repère décrit par les deux première composantes.

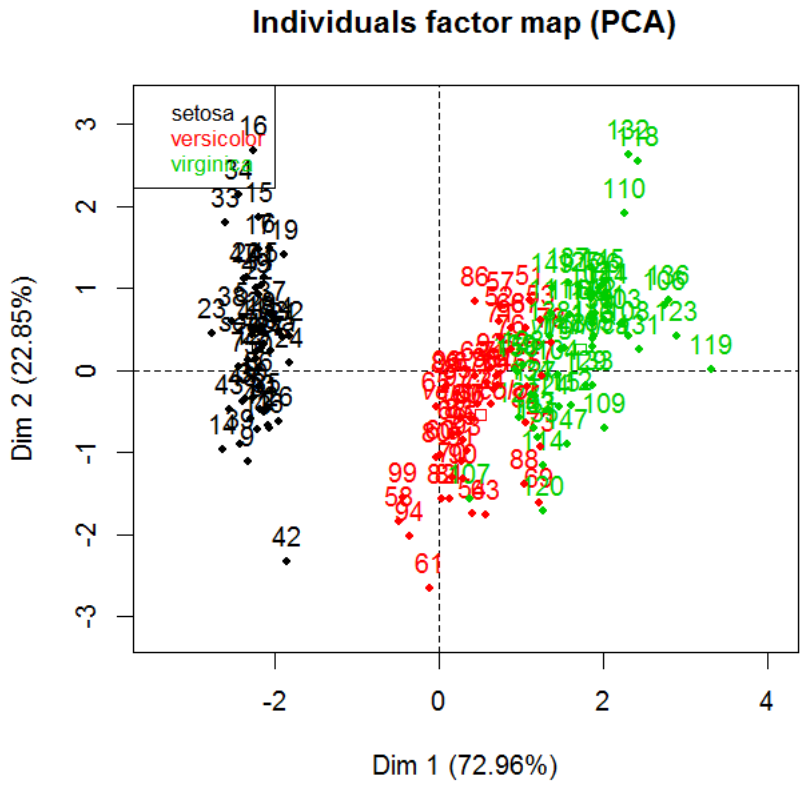


FIGURE 7.9 – **Graphe des individus dans le repère** $(PC1, PC2)$

En coloriant les points selon l'espèce de plante et en les projetant sur les axes PC1 (Y1) et PC2 (Y2), on voit que la PC1 parvient à distinguer presque parfaitement les trois espèces alors que ce n'est pas le cas de l'axe PC2. Les axes du repère indiquent également le pourcentage de variance du jeu de données porté par chacune des composantes. On remarque que l'axe 1 porte 73% et l'axe 2 23% de la variance du jeu de données, soit un total de 96% pour le plan 1-2.
On voudrait mieux comprendre quelles variables initiales sont représentées sur chacune des deux composantes. Le cercle des corrélations nous montre que les quatre variables se projettent à proximité du cercle des corrélations, on peut donc se fier aux angles entre les variables pour en déduire les corrélations. L'axe 1 représente 3 des 4 variables initiales tandis que l'axe 2 représente principalement la largeur des sépales.

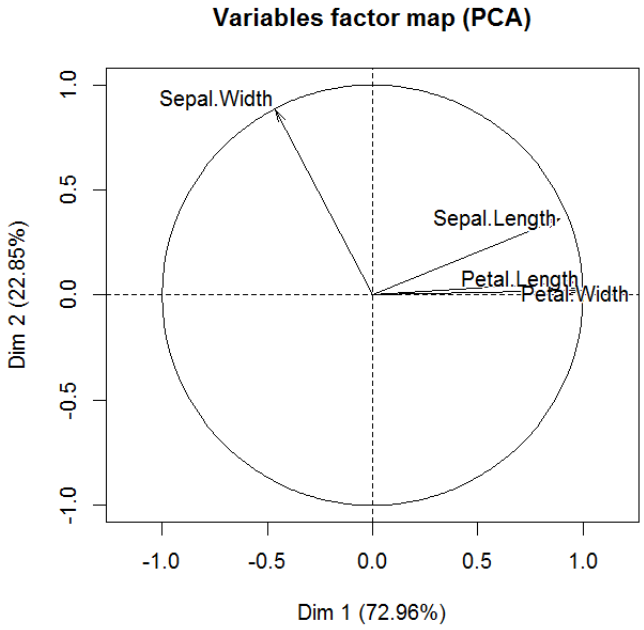


FIGURE 7.10 – **Cercle des corrélations pour les données Iris**

7.5 Annexe : calcul de l'inertie portée par le premier axe

Pour calculer l'inertie I_{Y_1} portée par le premier axe, on va tout d'abord calculer la distance G-y_{i1}

$$\begin{aligned} d(G, y_{i1}) &= \sqrt{(\overrightarrow{Ga_1}.\overrightarrow{Gx_i})^2} \\ &= \sqrt{(\overrightarrow{Ga_1}.\overrightarrow{Gx_i})(\overrightarrow{Gx_i}.\overrightarrow{Ga_1})} \text{ (puisque le produit scalaire est symétrique)} \\ &= \sqrt{a_1^T X_i X_i^T a_1} \text{ (écriture du produit scalaire sous forme matricielle)} \end{aligned}$$

Soit :

- y_{i1}, le projeté othogonal du $i^{\text{ème}}$ point sur l'axe Y_1
  On en déduit alors I_{Y_1},

$$\begin{aligned} I_{Y_1} &= \frac{1}{n}\sum_{i=1}^{n} d^2(G, y_{i1}) \\ &= \frac{1}{n}\sum_{i=1}^{n} a_1^T X_i X_i^T a_1 \\ &= a_1^T \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i^T\right] a_1 \\ &= a_1^T \Sigma a_1 \end{aligned}$$

- Σ, la matrice de variance / covariance

On va donc chercher à maximiser $a_1^T \Sigma a_1$. On sait également que $\|\overrightarrow{Ga_1}\|^2 = a_1^T a_1 = 1$ puisque $\overrightarrow{Ga_1}$ est le vecteur unitaire de l'axe Y_1.
Il s'agit d'un problème d'optimisation sous contrainte pouvant être traité par la méthode d'optimisation de Lagrange par laquelle on obtient l'équation :

$$a_1^T \Sigma a_1 - \lambda_1 a_1^T a_1 = 0$$

En utilisant $a_1^T a_1 = 1$, on obtient

$$a_1^T \Sigma a_1 = \lambda_{Y_1}$$

Chaque variable latente (ou axe de l'ACP) a une **valeur propre** λ_k qui représente sa variance empirique ($\lambda_k = I_{Y_k} = \sigma^2_{Y_k}$). Les $\overrightarrow{a_k}$ sont les **vecteurs propres**.