Licence de Biologie - ENS PSL

Introduction aux biostatistiques

Octobre-novembre 2023

Benoît Perez-Lamarque - benoit.perez@ens.psl.eu

(précédent contributeur : Jérémy Boussier)

Sommaire

Prérequis : Probabilités

0.1 Probabilités discrètes

On s'intéresse à des expériences aléatoires dont l'issue n'est pas déterministe (c'est-à-dire qu'on ne peut pas la connaître à l'avance), mais dont on connaît la probabilité de chaque issue. Dans le cadre des probabilités discrètes, on s'intéresse aux probabilités qui n'ont qu'un nombre dénombrable (*i.e.* fini ou en bijection avec $\mathbb{N}$) d'issues.

0.1.1 Espaces de probabilité

Commençons par des exemples.

Exemples.
1. On considère un lancer de dé équilibré. Les issues sont $1, 2, ..., 6$ et la probabilité de chaque issue est $\frac{1}{6}$. Elles sont dites équiprobables.
2. On considère deux lancers consécutifs d'une pièce équilibrée. Les quatre issues sont « Pile puis Pile », « Face puis Face », « Pile puis Face », « Face puis Pile », et la probabilité de chaque issue est $\frac{1}{4}$. Encore une fois, les issues sont équiprobables.

Définitions. Plus formellement, on définit ici une **expérience aléatoire** comme un couple (Ω, P) où :
— Ω est l'ensemble des issues possibles d'une expérience, qu'on appelle l'univers de l'expérience.
— P la mesure ou loi de probabilité associée à l'expérience. Elle doit vérifier les propriétés suivantes :

$$\begin{cases} \forall x \in \Omega, P(x) \in [0,1] \\ \\ \sum\limits_{x \in \Omega} P(x) = 1 \end{cases}$$

Un événement E est alors un sous-ensemble d'Ω dont on peut déterminer la probabilité de se produire. On définit $P(E) = \sum\limits_{x \in E} P(x)$

Une expérience aléatoire est donc caractérisée par (1) l'ensemble de ses issues et (2) la probabilité associée à chaque issue.

Exemples. Reprenons les exemples précédents :
1. Ici, $\Omega = \{1, ..., 6\}$ et $P(x) = \frac{1}{6}$ pour tout $x \in \Omega$.
2. Ici, $\Omega = \{Pile, Face\}^2$ et $P(x) = \frac{1}{4}$ pour tout $x \in \Omega$. Écrire mathématiquement l'événement « Obtenir au moins une fois Face ». Quelle est sa probabilité ?

Propriétés. On a les propriétés (relativement intuitives) suivantes : soient (Ω, P) un espace de probabilité, et A et B deux événements de cet espace.

(i) $P(A) \in [0, 1]$;

(ii) $P(\varnothing) = 0$ et $P(\Omega) = 1$;

(iii) L'**événement complémentaire** de A, noté $\overline{A}$ or A^c, a pour probabilité :

$$P(\overline{A}) = 1 - P(A)$$

(iv) La probabilité d'avoir A et B est notée $P(A \cap B)$;

(v) La probabilité d'avoir A ou B est notée $P(A \cup B)$ et

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

(vi) Si A and B sont **incompatibles**, *i.e.* $A \cap B = \varnothing$, alors

$$P(A \cup B) = P(A) + P(B)$$

0.1.2 Lien avec le dénombrement

Si les éléments de Ω (dits événements élémentaires) sont équiprobables, alors pour tout événement $A \in \Omega$,

$$P(A) = \frac{card(A)}{card(\Omega)}$$

où le cardinal de A, noté $card(A)$, est le nombre d'éléments de l'ensemble A.
Autrement dit, la probabilité de l'événement A est le nombre d'issues favorables divisé par le nombre total d'issues possibles.

De type de problèmes se transcrit souvent en des problèmes de dénombrements, cependant trouver le nombre de cas favorables peut se révéler fastidieux.

Quelques rappels de dénombrement :

(i) $\forall n \in \mathbb{N}$, le **nombre de permutations** des n éléments, dénoté $\boldsymbol{n!}$ (n *factoriel*), est égal à :

$$n! = \begin{cases} 1 \times 2 \times \ldots \times (n-1) \times n & \text{si } n > 0 \\ 1 & \text{si } n = 0. \end{cases}$$

(ii) Un **arrangement** est un sous ensemble ordonné de k éléments parmi n. Le **nombre d'arrangements** A_n^k de k éléments parmi n vaut :

$$A_n^k = \frac{n!}{(n-k)!}$$

(iii) Une **combination** est un sous ensemble (non ordonné) de k éléments parmi n. Le **nombre de combinaison** C_n^k est :

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

0.1.3 Probabilité conditionnelle et indépendance

A. Probabilité conditionnelle

Soit (Ω, P) un espace de probabilité, A et B deux événements, avec $P(B) \neq 0$.
La **probabilité conditionnelle** de A sachant B, notée $P(A|B)$ ou $P_B(A)$, est définie par

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Elle vérifie donc :

$$P(A \cap B) = P(A|B)P(B)$$

On en déduit le **théorème de Bayes** :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

B. Indépendance

Deux événements A et B sont dits **indépendants** si

$$P(A \cap B) = P(A)P(B)$$

c'est-à-dire, de façon équivalente si $P(B) \neq 0$,

$$P(A|B) = P(A)$$

0.1.4 Variables aléatoires

Une **variable aléatoire** est une fonction définie sur l'ensemble des issues d'une expérience aléatoire.

Par exemple, on lance successivement deux dés et on considère la somme des deux dés. Dans ce cas, $\Omega = \{1, ..., 6\}^2$, $P(x) = \frac{1}{36}$ pour tout $x \in \Omega$,

$$\forall (i, j) \in \Omega, X(i, j) = i + j$$

est la variable aléatoire considérée. Plus généralement donc, une variable aléatoire dans E est une fonction $X : \Omega \to E$. Nous nous restreindrons dans ce cours à $E = \mathbb{R}$. Cela peut être toute fonction qui donne une valeur qui dépend de l'issue de l'expérience.

A. Loi de la variable aléatoire

Soit X une variable aléatoire sur (Ω, P). La **loi de la variable aléatoire X,** notée P_X, est la fonction qui donne la probabilité de chaque résultat possible de X :

$$\forall x \in X(\Omega),\ P_X(x) = P(X = x) = \sum_{X(\omega)=x,\ \omega \in \Omega} P(\omega).$$

Tout comme une expérience aléatoire est caractérisée par (1) l'univers des issues possibles et (2) les probabilités associées à chacune de ces issues (la loi de probabilité), une variable aléatoire est caractérisée par (1) l'ensemble de ses résultats possibles, et (2) les probabilités associées à chacun de ces résultats (la loi de la variable aléatoire).

B. Fonction de répartition

Si X est une variable aléatoire dans $\mathbb{R}$, on appelle **fonction de répartition** de X, notée F_X, la fonction qui à tout réel x associe

$$F_X(x) = P(X \leq x).$$

Cette fonction caractérise entièrement la variable aléatoire dans les cas ordinaires. Elle est croissante et a pour limites 0 et 1 en $-\infty$ et $+\infty$ respectivement.

C. Espérance et variance

L'**espérance** d'une variable aléatoire X, notée $E(X)$ est sa valeur moyenne sur toutes les issues possibles :

$$E(X) = \sum_{\omega \in \Omega} P(\omega)X(\omega) = \sum_{x \in X(\Omega)} xP(X = x).$$

L'espérance est aussi dit « gain moyen ». Un jeu dont l'espérance vaut 0 est équitable.

La **variance** d'une variable aléatoire X, notée $V(X)$, est la moyenne des carrés des écarts à la moyenne :

$$V(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$$

L'**écart type** de X est $\sigma(X) = \sqrt{V(X)}$.

La variance et l'écart type mesurent l'écart moyen à la moyenne, c'est à dire la dispersion des valeurs prises par la variable aléatoire. Si la variance est grande, cela veut dire que les valeurs prises par la variable aléatoire sont dispersées sur un intervalle plus large par rapport à la moyenne, ce qui entraîne une plus grande incertitude quant à la valeur attendue. À l'inverse, si la variance est petite, on a de grandes chances que les valeurs tombent proches de la moyenne.

Soit a et b deux réels, on peut facilement démontrer que $E(aX + b) = aE(X) + b$ et $V(aX + b) = a^2V(X)$.

D. Indépendance

Deux variables aléatoires X et Y à valeurs dans E_1 et E_2 sont **indépendantes** si pour tous $i \in E_1$ et $j \in E_2$,

$$P((X = i) \cap (Y = j)) = P(X = i)P(Y = j)$$

Propriétés. Soient X et Y deux variables aléatoires, $E(X + Y) = E(X) + E(Y)$.
Si X et Y sont indépendantes, on a aussi : $V(X + Y) = V(X) + V(Y)$.
De plus, si X et Y sont indépendantes et à espérance finie : $E(XY) = E(X)E(Y)$.

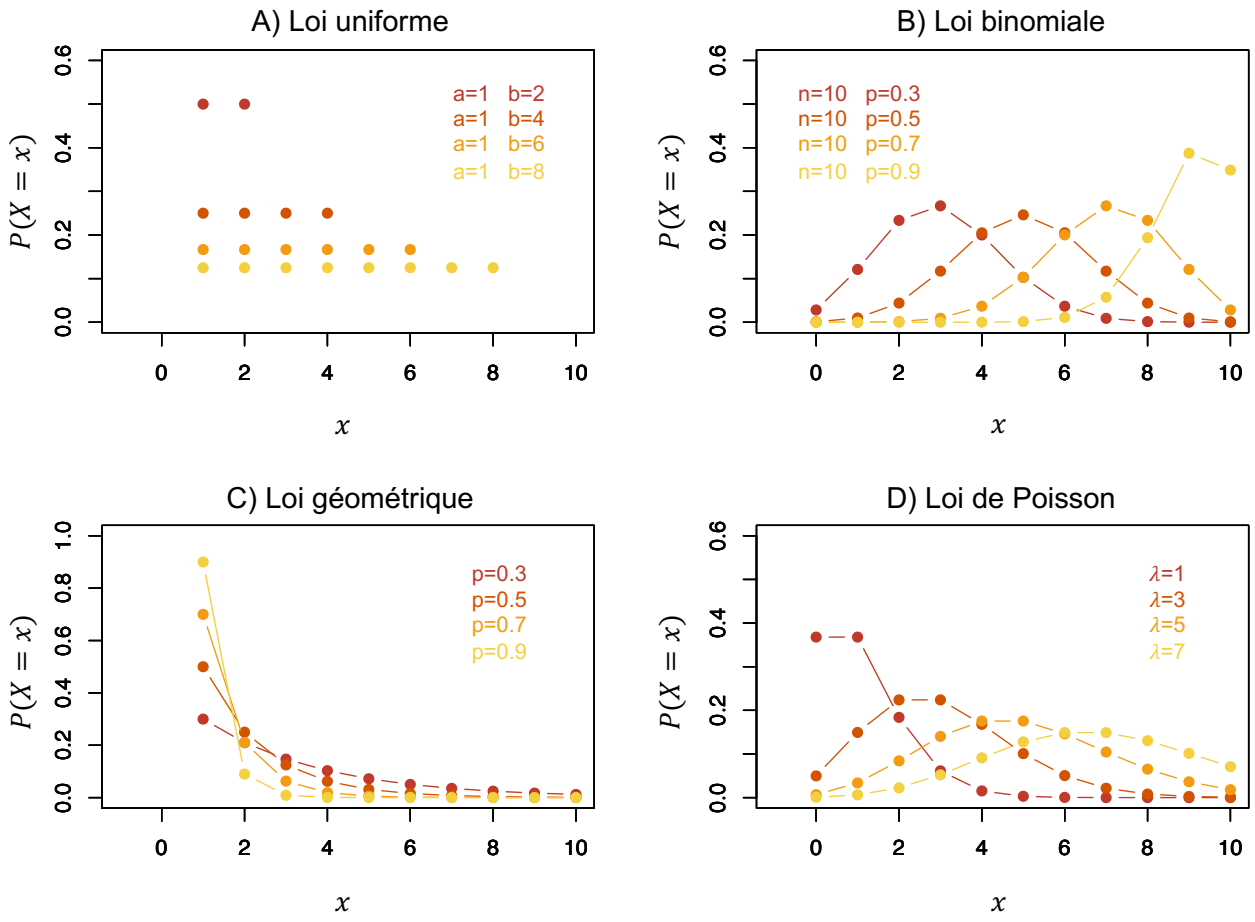

Figure 1: Lois de probabilité discrètes usuelles

0.1.5 Lois de probabilité discrètes usuelles

Les lois de probabilité discrètes classiques présentées dans cette section sont illustrées Figure 1.

A. Loi uniforme $\mathcal{U}(n)$

Si X suit la loi uniforme, $X \sim \mathcal{U}(n)$, elle prend les valeurs $1, 2, ..., n$ de manière équiprobable, *i.e.* pour tout $i \in \{1, ..., n\}$,

$$P_X(i) = P(X = i) = \frac{1}{n}$$

Son espérance et sa variance sont respectivement

$$E(X) = \frac{n+1}{2} \quad \text{et} \quad V(X) = \frac{n^2 - 1}{12}$$

B. Loi de Bernoulli $\mathcal{B}(p)$

Si X suit la loi de Bernoulli de paramètre p, $X \sim \mathcal{B}(p)$, elle prend la valeur 1 avec la probabilité p et 0 avec la probabilité $q = 1 - p$, *i.e.*

$$P(X = x) = \begin{cases} p & \text{si} \quad x = 1 \\ 1 - p & \text{si} \quad x = 0 \end{cases}$$

Son espérance et sa variance valent respectivement

$$E(X) = p \quad \text{et} \quad V(X) = pq = p(1 - p)$$

C. Loi binomiale $\mathcal{B}(n, p)$

On renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p. Soit X la variable aléatoire qui donne le nombre de succès obtenus à l'issue des n épreuves. On a alors $X(\Omega) = \{1, ..., n\}$ et pour tout $k \in X(\Omega)$,

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

On dit que X suit la loi binomiale de paramètres n et p, notée $X \sim \mathcal{B}(n, p)$. Ses espérance et variance sont

$$E(X) = np \quad \text{et} \quad V(X) = np(1 - p)$$

D. Loi géométrique $\mathcal{G}(p)$

Considérons une épreuve de Bernoulli dont la probabilité de succès est p et celle d'échec $q = 1 - p$. On joue de manière indépendante jusqu'au premier succès. On appelle X la variable aléatoire donnant le rang du premier succès. X peut prendre comme valeurs tous les entiers naturels non nuls. Pour tout $k \in \mathbb{N}^*$,

$$P(X = k) = q^{k-1} p$$

On dit dans ce cas que X suit une loi géométrique de paramètre p, notée $X \sim \mathcal{G}(p)$.
Ses espérance et variance sont

$$E(X) = \frac{1}{p} \quad \text{and} \quad V(X) = \frac{1 - p}{p^2}$$

E. Loi de Poisson $\mathcal{P}(\lambda)$

La loi de Poisson de paramètre λ, ou loi des événements rares, notée $\mathcal{P}(\lambda)$, décrit la probabilité d'occurence d'événements rares au cours d'un certain intervalle de temps (événement ayant lieu typiquement moins de 30 fois en moyenne dans l'espace considéré). Elle donne la probabilité d'un nombre certain d'occurrences de l'événement rare sachant le nombre moyen d'occurrences $\lambda \in \mathbb{R}$. Elle s'applique si la probabilité qu'un événement survienne est la même pour chaque unité de temps et si le nombre d'événements qui survient dans une unité de temps est indépendant du nombre d'événements qui survient dans une autre unité.

Soit $X \sim P(\lambda)$, X ne prend que des valeurs entières, *i.e.* $X(\Omega) = N$. Pour tout $k \in \mathbb{N}$,

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

λ est à la fois l'espérance et la variance de la distribution :

$$E(X) = \lambda \quad \text{and} \quad V(X) = \lambda$$

La loi de Poisson est la « version temps continu » de la loi binomiale : si X suit une loi binomiale de paramètres $(N, \lambda/N)$, X tend en loi vers une loi de Poisson de paramètre λ lorsque $N \to \infty$. En biologie, cette loi est notamment utilisée pour décrire les fréquences de mutation de l'ADN.

0.1.6 Remarque sur la notion de variable aléatoire

Lorsque l'on connaît l'ensemble d'arrivée d'une variable aléatoire, c'est-à-dire l'ensemble des valeurs qu'elle peut prendre, ainsi que sa loi, c'est-à-dire la probabilité qu'elle prenne chacune de ces valeurs, connaître l'univers Ω (c'est-à-dire l'espace de départ de la variable aléatoire) n'est plus d'une grande utilité, et on on omettra souvent de le préciser. Autrement dit, une fois qu'on a caractérisé la loi d'une variable aléatoire, l'espace de probabilité initial n'est plus d'un grand intérêt. Par exemple, mettons que je sache que lorsque je mets 1 euro dans une machine à sous, j'ai une chance sur 100 d'en gagner 50, une chance sur 10 d'en gagner 8, et le reste de perdre mon euro. Si je cherche à savoir mon gain sur plusieurs parties, je me fiche de savoir quelles sont les combinaisons (deux images égales, un as et deux images différentes) qui font gagner 8 euros, maintenant que je sais que la probabilité de gagner 8 euros est de $\frac{1}{10}$. De façon générale, la notion de loi de variable aléatoire va occulter celle, antérieure, de loi d'espace de probabilité. En fait, même dans le cas où l'on veut étudier des événements directement sur l'espace Ω, on peut les considérer comme des événements de la variable aléatoire X. Par exemple, pour décrire le problème initial « on lance un dé » : au lieu de définir Ω et P, on dira simplement que la valeur du dé est une variable aléatoire qui suit une loi uniforme sur $\{1, ..., 6\}$.

0.2 Probabilités continues sur $\mathbb{R}$

0.2.1 Introduction à la densité de probabilité

On appelle X la variable aléatoire qui correspond au tirage aléatoire d'un nombre réel entre 0 et 1. L'espace d'arrivée est $[0, 1]$ qui n'est pas dénombrable : on sort donc du cadre

des probabilités discrètes. On veut exprimer la loi de X. Intuitivement, si $x \in \mathbb{R}$, par « équiprobabilité »,

$$P(x) = \frac{1}{Card(\Omega)} = \frac{1}{+\infty} = 0.$$

c'est-à-dire que la probabilité d'obtenir exactement x vaut 0. Ce qui nous intéresse ne va donc pas être la probabilité d'obtenir exactement x, mais la probabilité d'obtenir un résultat proche de x. Plus clairement, si $x \in \mathbb{R}$, lorsque $\mathrm{d}x \to 0$, au premier ordre en $\mathrm{d}x$,

$$P(X \in [x, x + \mathrm{d}x]) = f_X(x)\,\mathrm{d}x$$

La fonction f_X est appelée **densité de la variable aléatoire** X.

Dans notre exemple de tirage aléatoire entre 0 et 1, f_X vaut 1 pour tout x, cela veut dire que la probabilité d'obtenir un résultat entre x et $x + \mathrm{d}x$ tend vers $1 \times \mathrm{d}x$ (la longueur du segment $[x, x + \mathrm{d}x]$).

Pour obtenir la probabilité d'un événement du type $\{X \in [a, b]\}$ avec $a < b$, il faut alors sommer ces petites probabilités, ce qui revient (comme nous sommes dans le cas continu) à les intégrer. On a alors

$$P(X \in [a, b]) = P(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Une variable aléatoire continue est donc caractérisée par sa densité, ce qui constitue la seule différence entre les lois continues et les lois discrètes (pour qui on n'a pas besoin d'avoir recours à cette notion, vue que la probabilité d'un événement est la somme de la probabilité de singletons).

0.2.2 Fonction de répartition

Si X est une variable aléatoire sur $\mathbb{R}$, sa **fonction de répartition** (ou distribution cumulative), notée F_X, est définie, comme dans le cas discret, par : pour tout $x \in X(\Omega)$,

$$F_X(x) = P(X \leq x)$$

ce qui vaut dans le cas continu, comme nous l'avons vu,

$$\forall x \in X(\Omega),\ F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

0.2.3 Espérance et variance

L'espérance d'une variable aléatoire continue réelle X, notée $E(X)$ est définie par

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x$$

La variance d'une variable aléatoire continue réelle X, notée $V(X)$ est définie par

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f_X(x)\,\mathrm{d}x$$

L'écart type de X est $\sigma(X) = \sqrt{V(X)}$.

Soit a et b deux réels, on peut facilement démontrer que $E(aX+b) = aE(X)+b$ et $V(aX+b) = a^2V(X)$.

Soit X et Y deux variables aléatoires, $E(X+Y) = E(X) + E(Y)$. Si X et Y sont indépendantes, on a aussi : $V(X+Y) = V(X) + E(Y)$.

0.2.4 Lois de probabilités continues usuelles

Les lois de probabilité continues présentées dans cette section sont illustrées Figure 2.

A. Loi uniforme $\mathcal{U}(a,b)$

Une variable aléatoire continue X suit une loi uniforme entre a et b (avec $a < b$), $X \sim \mathcal{U}(a,b)$, si X prend la valeur $x \in [a,b]$ avec equiprobabilité et f_X est définie comme:

$$\forall x \in \mathbb{R},\ f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } \ x \in [a,b] \\ 0 & \text{sinon.} \end{cases}$$

$$\forall x \in \mathbb{R},\ F_X(x) = \begin{cases} 0 & \text{si } \ x < a \\ \dfrac{x-a}{b-a} & \text{si } \ x \in [a,b] \\ 1 & \text{si } \ b < x \end{cases}$$

$$E(X) = \frac{a+b}{2} \quad \text{et} \quad V(X) = \frac{(a-b)^2}{12}$$

B. Loi exponentielle $\mathcal{E}(\lambda)$

La loi exponentielle est la « version temps continu » de la loi géométrique : elle mesure le temps qu'il faut pour qu'un événement (sans mémoire) se produise. On dit qu'une variable aléatoire réelle X suit la loi exponentielle de paramètre λ (appelée taux), $X \sim \mathcal{E}(\lambda)$, si X admet pour densité de probabilité la fonction f_X définie, pour tout nombre réel positif $x \geq 0$, par

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

$$\forall x \in \mathbb{R}^+,\ F_X(x) = 1 - e^{-\lambda x}$$

Son espérance et sa variance sont

$$E(X) = \frac{1}{\lambda} \quad \text{et} \quad V(X) = \frac{1}{\lambda^2}$$

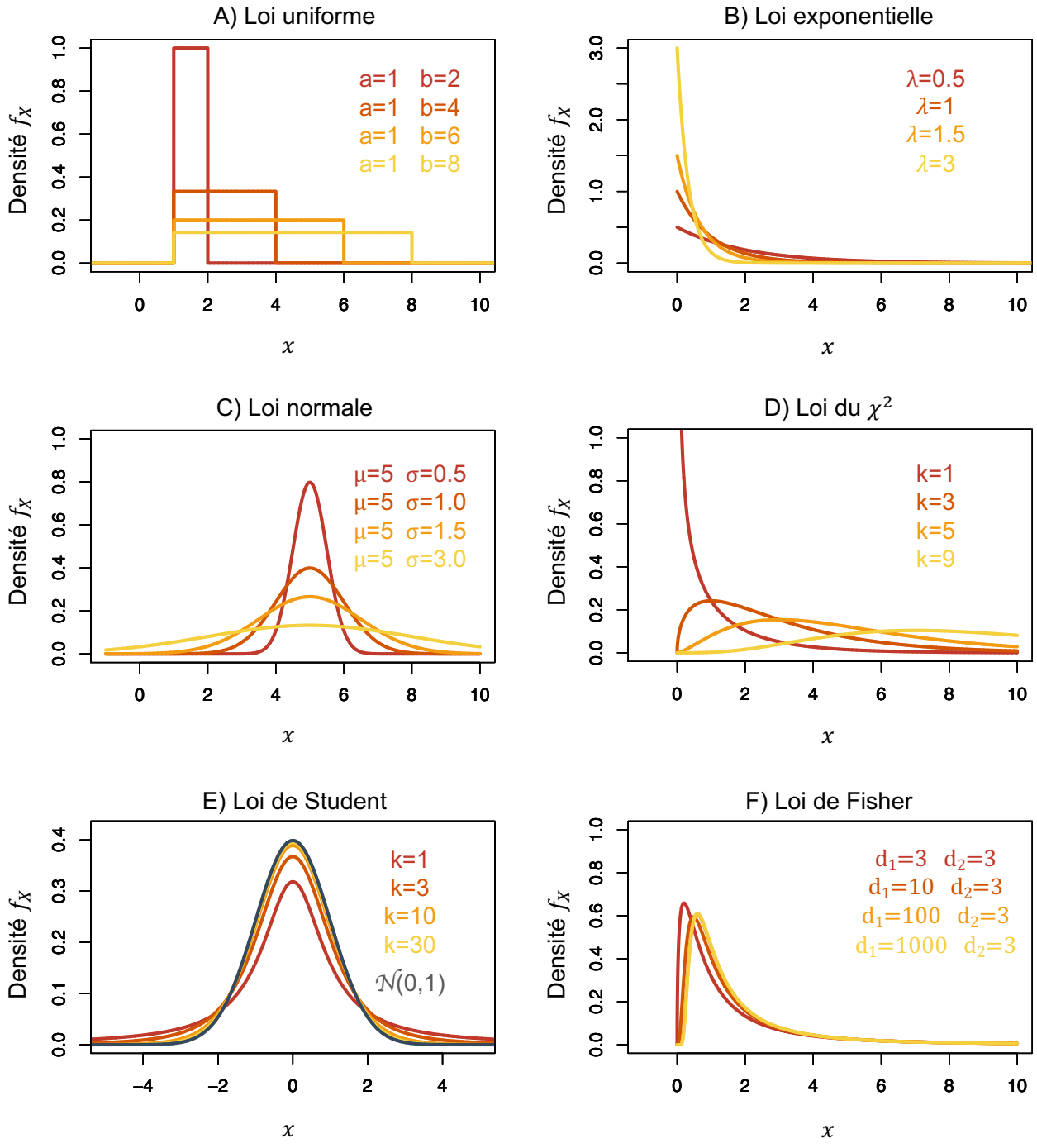

Figure 2: Loi de probabilités continues usuelles

C. Loi normale ou gaussienne $\mathcal{N}(\mu, \sigma^2)$

On dit qu'une variable aléatoire réelle X suit la loi normale d'espérance μ et de variance σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, si X admet pour densité de probabilité la fonction f_X définie, pour tout nombre réel x, par

$$\forall x \in \mathbb{R},\ f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

La fonction de répartition de $\mathcal{N}(\mu, \sigma^2)$ n'a pas de forme simple.

Par définition, $E(X) = \mu$ et $V(X) = \sigma^2$.

Si X et Y sont deux variables aléatoires *indépendantes* de loi respectives $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$, alors $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

On appelle **loi normale centrée réduite** la loi $\mathcal{N}(0, 1)$, noté Z. Z peut être obtenue à partir de n'importe quelle loi normale $X \sim \mathcal{N}(\mu, \sigma^2)$ grâce à la **transformation de centrage et de réduction** suivante :

$$Z = \frac{X - \mu}{\sigma}$$

Z est une fonction paire, donc $f_Z(-z) = f_Z(z)$. Sa fonction de répartition est noté Φ et on a ainsi $\Phi(-z) = 1 - \Phi(z)$. Les valeurs de Φ peuvent être consultées dans la table de la loi normale centrée réduite.

La loi normale est très importante, car c'est la loi vers laquelle tend la valeur moyenne d'une variable aléatoire quelconque (voir paragraphe suivant sur le théorème central limite). Les trois lois suivantes dérivent de la loi normale centrée réduite.

D. Loi du $\chi^2(k)$

Soient $X_1, ..., X_k$ des variables aléatoires indépendantes et de même loi normale centrée réduite (on dit que les variables X_i sont indépendantes et identiquement distribuées - noté *i.i.d*). La loi Y définie comme la somme des carrés des variables aléatoires X_i suit une loi du χ^2 (prononcé « khi-deux ») avec k degrés de liberté :

$$Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$$

Son espérance et variance valent

$$E(Y) = k \quad \text{et} \quad V(Y) = 2k$$

Si X et Y sont deux variables aléatoires indépendantes de loi respectives $\chi^2(k_1)$ et $\chi^2(k_2)$, alors $X + Y \sim \chi^2(k_1 + k_2)$.

E. Loi t de Student $t(k)$

Soient Z une variable aléatoire de loi normale centrée et réduite, $Z \sim \mathcal{N}(0,1)$, et Y une variable indépendante de Z et distribuée suivant la loi du χ^2 à k degrés de liberté, $Y \sim \chi^2(k)$. Soit la variable aléatoire T définie comme :

$$T = \frac{Z}{\sqrt{Y/k}}$$

La loi de T est appelée loi t de Student à k degrés de liberté, notée $t(k)$.

$$\text{Pour } k > 2,\ E(T) = 0 \text{ et } V(T) = \frac{k}{k-2}$$

Remarque. La distribution de la loi de Student tend vers la loi normale centrée réduite lorsque k augmente. Pour $k > 30$, on considère que les deux distributions sont quasi identiques et on peut approximer la loi de Student par $\mathcal{N}(0,1)$.

F. Loi de Fisher $\mathcal{F}(d_1, d_2)$

Soient Y_1 et Y_2 deux variables aléatoires indépendantes de lois de χ^2 à d_1 et d_2 degrés de liberté respectivement. Alors la loi de

$$F = \frac{Y_1/d_1}{Y_2/d_2}$$

est appelée loi de Fisher de paramètres d_1 et d_2, notée $\mathcal{F}(d_1, d_2)$.

0.2.5 Loi des grands nombres et théorème central limite

On considère $X_1, X_2, ...X_n$ variables aléatoires i.i.d. (c'est-à-dire indépendantes et suivant la même loi). Soient μ l'espérance et σ l'écart type de la loi commune aux X_i. On suppose que μ et σ sont finis. Soit M_n la **moyenne empirique** des X_i :

$$M_n = \frac{X_1 + ... + X_n}{n}$$

La **loi des grands nombres** stipule que cette moyenne M_n converge vers l'espérance de la loi commune : pour $n \to \infty$,

$$M_n \to \mu$$

Le **théorème central limite** précise à quelle vitesse a lieu cette convergence :

$$\sqrt{n}(M_n - \mu) \longrightarrow \mathcal{N}(0, \sigma^2)$$

$$\text{ou} \quad \frac{M_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \longrightarrow \mathcal{N}(0,1)$$

Donc pour un n grand, on peut faire l'approximation suivante :

$$M_n \longrightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Le théorème central limite énonce que, dans le contexte de la loi des grands nombres, les fluctuations autour de la moyenne tendent à suivre une distribution normale et deviennent de plus en plus resserrées autour de la moyenne à mesure que la taille de l'échantillon n augmente. Ce phénomène signifie que la loi des fluctuations ne dépend pas de la distribution initiale des variables aléatoires X_i. Cette universalité des fluctuations explique pourquoi la distribution normale est omniprésente dans la modélisation des phénomènes aléatoires.

Exemple : Si une variable aléatoire X suit une loi binomiale, $X \sim \mathcal{B}(n, p)$, qui compte le nombre de succès d'un grand nombre d'essais (en général lorsque $np \geq 5$ et $n(1-p) \geq 5$), le théorème central limite s'applique et la distribution discrète de X peut être approximée par une distribution normale, avec une moyenne np et une variance $np(1-p)$: $X \sim \mathcal{N}(\mu = np, \sigma^2 = np(1-p))$.

Chapitre 1 : Estimation et intervalle de confiance

1.1 Introduction à la théorie statistique

1.1.1 Probabilités *versus* statistiques : deux cadres de pensée pas vraiment opposés

Le **domaine des probabilités** s'intéresse aux propriétés des variables aléatoires. Étant données des variables aléatoires et leurs lois, on veut savoir quelles sont les probabilités qu'un événement se réalise. Par exemple, si l'on lance une pièce équilibrée 10 fois, on peut calculer la probabilité d'obtenir (dans un ordre quelconque) 5 fois Pile et 5 fois Face.

Le **domaine des statistiques** s'intéresse a priori à l'inverse : **on veut donner des informations sur les lois étant données des informations sur une réalisation**, c'est-à-dire un événement qui est arrivé. Par exemple, on a lancé 10 fois une pièce (et on a les résultats de ces lancers), et on voudrait des informations sur la loi qui caractérise le lancer. Evidemment, même avec des informations sur beaucoup de réalisations, on ne pourra jamais décréter, de façon certaine, la loi qui régit cette expérience : ces réalisations ne fournissent que des exemples, et de même qu'un grand nombre d'exemples ne suffit pas à prouver un théorème, un grand nombre de réalisations ne suffira jamais à déterminer la loi que suit l'expérience.
Cependant, on peut tenter, à partir de réalisations d'une expérience, de répondre à des questions plus particulières sur sa loi et s'autoriser une certaine marge d'erreur dans nos conclusions.

1.1.2 Vocabulaire

En statistiques, l'objet d'étude est appelé **population**. Ce peut être un ensemble (fini) d'objets existants (appelées individus, par exemple les êtres humains) ou l'ensemble (infini) des réalisations d'une expérience aléatoire (par exemple les lancers d'un dé). Typiquement, on suppose qu'une caractéristique spécifique des individus (par exemple leur taille) sont des réalisations x_i de variables aléatoires X_i qui suivent une loi commune P_X, c'est-à-dire $x_i = X_i(\omega)$ où $P_{X_i} = P_X$.

L'objet de la **statistique inférentielle** est d'estimer des caractéristiques de la population à partir de la connaissance d'un **échantillon**, c'est-à-dire d'un sous-ensemble (fini, et souvent petit) de la population. Plus rigoureusement, le but est d'estimer la loi P_X, ou plus modestement d'estimer certaines de ses caractéristiques (moyenne, variance), à partir de la connaissance d'un nombre fini de réalisations $x_1, ..., x_n$.

Pour obtenir des estimations valides, l'**échantillonnage** - c'est à dire la sélection des individus qui constituent l'échantillon - doit être aléatoire (on dit aussi **non biaisé**). Si on ne sélectionne que des individus qui partagent une caractéristique supplémentaire A (on parle de sous-population), on biaise notre échantillon, et on n'estime plus la loi P_X, mais $P_{X|A}$.

Dans la suite de ce chapitre, n désigne un entier naturel non nul, X_1, ..., X_n des variables aléatoires indépendantes et identiquement distribuées, et x_1, ..., x_n des réalisations de X_1, ..., X_n.

1.2 Estimateurs (estimation ponctuelle)

Lorsque l'on possède un échantillon suivant une loi inconnue, on peut tenter d'estimer certains des paramètres de cette dernière, comme sa moyenne, sa variance, etc. On fait alors appel à la notion d'**estimateur**.

1.2.1 Définitions

Un **estimateur du paramètre** θ est une variable aléatoire $\hat{\theta}_n$ fonction f de X_1, ..., X_n. Cet estimateur ne doit donc pas dépendre directement du paramètre inconnu θ que l'on souhaite estimer. La réalisation de cet estimateur, notée $u_n = f(x_1, ..., x_n)$, est une fonction des données (*i.e.* de l'échantillon $x_1, ..., x_n$) et est appelée **estimation de** θ.

Deux propriétés font d'un estimateur un « **bon estimateur** » :

(i) un estimateur est dit **non biaisé** si son espérance est le paramètre qu'il estime, soit

$$E(\hat{\theta}_n) = \theta$$

(ii) un estimateur est dit **convergent** s'il converge (en probabilité) vers le paramètre qu'il estime, c'est-à-dire lorsque $n \to \infty$,

$$\hat{\theta}_n \to \theta$$

1.2.2 Estimateur de l'espérance

Soit X une variable aléatoire, telle que $E(|X|) < \infty$. Mettons que l'on cherche à estimer son espérance de la population, $E(X) = \mu$, sachant que l'on connaît n réalisations indépendantes $(x_1, ..., x_n)$ de cette variable aléatoire. Intuitivement, on prend la moyenne empirique de l'échantillon $m_n = \dfrac{x_1 + ... + x_n}{n}$.

L'estimateur correspondant (appelé aussi moyenne de l'échantillon) est :

$$M_n = \frac{X_1 + ... + X_n}{n}$$

Cet estimateur est non biaisé et convergent.

1.2.3 Estimateur de la variance

Soit X une variable aléatoire, telle que $E(|X|) < \infty$ et $V(X) < \infty$. On cherche désormais à estimer la variance de la population, notée $V(X) = \sigma^2$, connaissant encore une fois n réalisations indépendantes $(x_1, ..., x_n)$ de cette variable aléatoire. On a envie de prendre la variance de l'échantillon définie comme :

$$v_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_n)^2$$

l'estimateur correspondant (appelé aussi variance de l'échantillon) étant

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$$

Cependant, V_n est un estimateur biaisé. En effet, on peut démontrer que :

$$E(V_n) = \frac{n-1}{n}\sigma^2$$

Un estimateur non biaisé (et toujours convergent) de σ^2 est donc :

$$S_n^2 = \frac{n}{n-1}V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

On dénomme l'estimateur S_n^2 la **variance empirique**; on dit qu'on passe de V_n à S_n^2 en appliquant la correction de Bessel.

Intuitivement, on utilise une première fois les n observations pour calculer m_n puis on les réutilise pour calculer v_n : vu que m_n intervient dans le calcul de v_n, v_n n'est donc pas déterminé par n observations mais par seulement $n-1$ observations. On dit qu'on perd un **degré de liberté**. Cela se traduit par une **sous estimation de la variance** σ^2 de la population.

Résumé : Estimateurs de l'espérance et de la variance.

Un estimateur non biaisé et convergent de l'espérance $\mu = E(X)$ est :

$$M_n = \frac{X_1 + ... + X_n}{n}$$

Un estimateur non biaisé et convergent de la variance $\sigma^2 = V(X)$ est :

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

1.2.4 Remarque : terminologie

Comme M_n et S_n^2 sont des estimateurs convergents, la moyenne empirique et la variance empirique (fictives) d'un échantillon très grand tend vers l'espérance et la variance de loi que la population suit. Quand la taille de population est très grande (par exemple la population humaine), par abus de langage, on parle souvent de la moyenne (resp. variance) de la population au lieu de l'espérance (resp. variance) de la loi dont les individus sont des réalisations. Il ne faut donc pas confondre la moyenne de la population (qui est en fait un paramètre fixe — l'espérance de la loi) avec la moyenne de l'échantillon (qui est la réalisation d'une variable

aléatoire, et qui dépend donc de l'échantillon).

1.3 Intervalle de confiance

En plus de s'assurer qu'un estimateur est non biaisé (c'est-à-dire qui va, en moyenne, donner la bonne réponse) et convergent (c'est-à-dire qui va converger vers la bonne réponse lorsque $n \to \infty$), on peut vouloir évaluer plus quantitativement sa qualité : on peut par exemple chercher à savoir **à quelle vitesse l'estimateur converge** vers le paramètre θ à estimer (en fonction de la taille de l'échantillon n), ou encore **quelle est la probabilité de se tromper en disant que l'estimateur est proche de sa cible**. La notion d'**intervalle de confiance** intervient dans ce contexte.

1.3.1 Définition

Soit $\alpha \in]0,1[$. Un intervalle $I_n = I(X_1, ..., X_n)$ est dit **intervalle de confiance pour θ de niveau** $1-\alpha$ si :

$$P(\theta \in I(X_1, ..., X_n)) = 1 - \alpha$$

Souvent, on choisit le risque $\alpha = 5\%$, et la propriété signifie : si on faisait un grand nombre de n-échantillonnages E_k (et donc chaque échantillonnage donnerait naissance à un intervalle I_k), alors θ se trouverait dans 95% des I_k.

Pour obtenir le maximum d'information, il faut s'efforcer de construire l'intervalle de confiance le moins large possible qui satisfait la condition de minoration donnée dans la définition.

1.3.2 Exemple 1 : loi normale de variance connue

Supposons que $X \sim \mathcal{N}(\mu, \sigma^2)$, où σ est connu, mais μ à déterminer. On dispose d'un échantillon $(x_1, ..., x_n)$ constitué de réalisations de variables aléatoires i.i.d. $X_1, ..., X_n$ de même loi que X. On sait qu'un bon estimateur de μ est la moyenne empirique de l'échantillon, soit

$$M_n = \frac{X_1 + ... + X_n}{n}$$

On a alors

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$

d'où en sommant les variables

$$\frac{X_1 + ... + X_n - \mu n}{\sigma} \sim \mathcal{N}(0,n)$$

ou encore en divisant par $\sqrt{n}$

$$Z_n = \sqrt{n}\frac{M_n - \mu}{\sigma} \sim \mathcal{N}(0,1) \tag{1}$$

On est donc ramené à une loi normale centrée réduite; le plus petit intervalle I tel que $P(Z_n \in I) = 1 - \alpha$ est centré en 0.
Soit β tel que $P(|Z_n| \leq \beta) = 1 - \alpha$ (si $\alpha = 0.05$, $\beta \approx 1,96$). On a donc :

$$P\left(Z_n \in [-\beta, \beta]\right) = 1 - \alpha$$

On en déduit donc :

$$P\left(\mu \in \left[M_n - \frac{\beta\sigma}{\sqrt{n}}, M_n + \frac{\beta\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$$

Autrement dit, $I_n = \left[M_n - \frac{\beta\sigma}{\sqrt{n}}, M_n + \frac{\beta\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre μ.

Remarque : I_n ne correspond donc pas à l'intervalle qui contient la valeur μ avec forte probabilité. Mais cela indique que la probabilité d'avoir observé une moyenne empirique M_n aussi éloignée de μ était très faible ($< 5\%$) si μ n'est pas dans I_n.

1.3.3 Exemple 2 : loi normale de variance inconnue

Reprenons l'exemple précédent, mais en supposant cette fois-ci que la variance σ^2 n'est pas connue ; l'intervalle de confiance tel qu'il a été défini plus haut n'est donc pas exploitable.

En revanche, l'équation (1) est toujours vraie. Un bon estimateur (non biaisé) de σ^2 est $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$. Dans l'idée, on aimerait substituer σ par S_n dans l'équation (1).
On peut alors montrer (mais ce n'est pas trivial) que :

$$Y_n = (n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

On a alors :

$$\frac{Z_n}{\sqrt{Y_n/(n-1)}} = \sqrt{n}\frac{M_n - \mu}{S_n} \sim t(n-1)$$

Soit β tel que $P\left(\left|\sqrt{n}\frac{M_n - \mu}{S_n}\right| \leq \beta\right) = 1 - \alpha$. On en déduit que :

$$P\left(\mu \in \left[M_n - \frac{\beta S_n}{\sqrt{n}}, M_n + \frac{\beta S_n}{\sqrt{n}}\right]\right) = 1 - \alpha$$

Autrement dit, $\left[M_n - \frac{\beta S_n}{\sqrt{n}}, M_n + \frac{\beta S_n}{\sqrt{n}}\right]$ est un intervalle de confiance de niveau $1 - \alpha$ pour le paramètre μ.

1.3.4 Exemple 3 : loi quelconque

Dans le cas où la loi de X n'est pas normale, il n'est pas possible de faire ce genre de calculs exacts. Cependant, si les hypothèses du théorème central limite sont vérifiées, à savoir les variables aléatoires X_1, ..., X_n indépendantes et suivent la même loi d'espérance μ et de variance σ^2. On sait alors que la moyenne empirique proprement normalisée est asymtotiquement gaussienne, ce qui signifie que l'équation (1) est asymptotiquement valide (*i.e.* pour n suffisamment grand) pour X presque quelconque.

Sachant de plus que S est un estimateur convergent de σ, on est ramené au cas de l'exemple 2 (loi normale de variance inconnue) :

Si β est tel que $P\left(|\sqrt{n}\frac{M_n-\mu}{S_n}| \leq \beta\right) = 1-\alpha$, alors $\left[M_n - \frac{\beta S_n}{\sqrt{n}}, M_n + \frac{\beta S_n}{\sqrt{n}}\right]$ est un **intervalle de confiance asymptotique** de niveau $1-\alpha$ pour le paramètre μ.

Remarque. À partir de $n = 30$, on considère que la distribution de la loi de Student à $n-1$ degrés de liberté est très proche de la distribution de la loi normale centrée réduite (*cf* Figure 2 du Chapitre de prérequis) : on peut donc approximer la loi de Student par la loi normale centrée réduite lorsque l'on utilise les tables des lois usuelles.

Chapitre 2 : Test d'hypothèse

Dans ce chapitre, on s'intéresse à une sous-branche de la **statistique inférentielle : le test d'hypothèse**. Le but n'est plus d'estimer la valeur de paramètres, mais de décider de la valeur de vérité d'une assertion sur la loi : tel médicament a-t-il un effet ? Le risque de cancer est-il plus élevé dans le nord de la France que dans le sud ? Le taux de globules rouges et l'âge sont-ils corrélés ?

2.1 Exemples introductifs

2.1.1 Exemple 1 : loi normale

On observe une réalisation x d'une variable aléatoire X, et l'on veut **tester l'hypothèse** : $X \sim \mathcal{N}(0,1)$. Intuitivement, si $x = 100$, on aura tendance à rejeter l'hypothèse, alors que si $x = 1$, on ne la rejettera pas.

L'idée de la méthode statistique est d'utiliser le même type de raisonnement que pour les problèmes d'intervalles de confiance. Supposons que X suit effectivement une loi normale centrée réduite. Alors $P(-1.96 < X < 1.96) = 0.95$, c'est-à-dire qu'une réalisation aura 95% de chance de se trouver dans l'intervalle $[-1.96, 1.96]$. Comme $100 > 1.96$, l'événement $x = 100$ est « très improbable » sous l'hypothèse $X \sim \mathcal{N}(0,1)$; on aura donc tendance à rejeter l'hypothèse $X \sim \mathcal{N}(0,1)$.

C'est donc une sorte de **raisonnement par l'absurde** : on suppose que l'hypothèse potentiellement à réfuter est vraie, puis on en déduit l'intervalle dans lequel doit se trouver 95% des réalisations (on le choisit en général le plus petit possible pour augmenter la puissance du test), et on regarde si la réalisation (unique) que l'on observe se situe dans cet intervalle ou dans l'ensemble complémentaire (appelé **zone de rejet**). Si l'observation est dans la zone de rejet, on va rejeter l'hypothèse.

Ce raisonnement n'est pas un raisonnement par l'absurde *stricto sensu* : un résultat $x = 100$ peut très bien être issu d'une loi normale centrée réduite (qui peut prendre n'importe quelle valeur de $\mathbb{R}$); simplement, un résultat si extrême est très peu probable.

2.1.2 Exemple 2 : lancers de pièce

On se place dans la situation suivante : on lance 10 fois une pièce et on obtient 10 fois Pile. On peut se demander si cette pièce est bien équilibrée.

Pour cela, une idée qui semble naturelle est de supposer que la pièce est équilibrée et de montrer que n'obtenir que des Piles est très improbable. Supposons que l'on fasse n lancers $X_1, ..., X_n$ qui suivent une loi de Bernoulli de paramètre $p = 1/2$, et on fait correspondre 1 à Pile et 0 à Face. Alors le nombre de fois qu'on a obtenu Pile correspond à $Y_n = X_1 + ... + X_n$, dont la loi est représentée pour certains n dans la Figure 3.

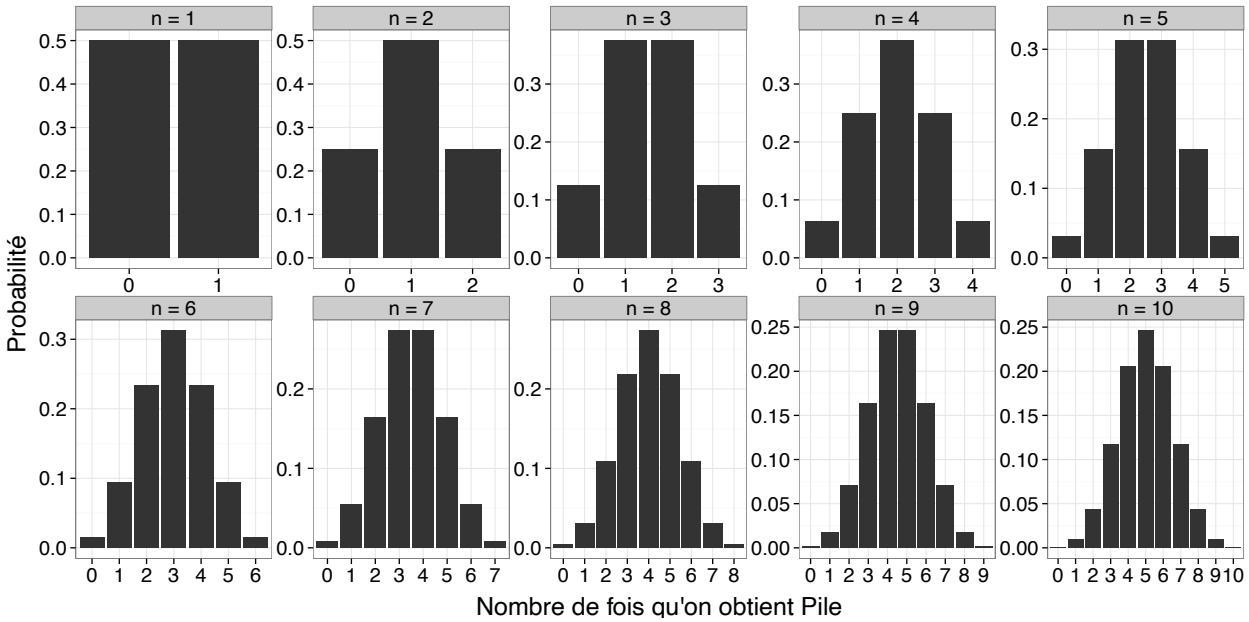

Figure 3: Loi de Y_n en fonction de n

La pièce ayant été lancée dix fois, regardons la loi de Y_{10}. Cherchons le plus petit intervalle de probabilité au moins 95% d'être atteint par Y_n. C'est $[2, 8]$ puisque $P(3 \leq Y_n \leq 7) \approx 0,945$ et $P(2 \leq Y_n \leq 8) \approx 0,989$.

La zone de rejet est donc $\{0, 1, 9, 10\}$, et $y = 10$ dans notre cas se situe dans la zone de rejet.

Mettons que l'on ait lancé la pièce que deux fois. On se place dans le cas $n = 2$, et l'on voit que la zone de rejet pour Y_2 est vide : les résultats extrêmes sont déjà de probabilité > 0.05. Même si l'on obtient $y = 2$ (c'est-à-dire que des Piles), on ne pourra pas rejeter l'hypothèse. Ceci est conforme à l'intuition que nous avons : si on lance 2 ou 3 fois une pièce et qu'on n'obtient que des Piles, cela peut être dû au hasard bien qu'elle soit équilibrée (avec une probabilité non négligeable) ; au contraire, si on la lance 10 fois et qu'on n'obtient que des Piles, la probabilité est si faible que ce soit dû au hasard si elle est équilibrée qu'on va rejeter cette hypothèse.

Remarques.

(i) Encore une fois, on a procédé à une sorte de raisonnement par l'absurde : pour montrer que la pièce n'était pas équilibrée, on a supposé qu'elle l'était. Cette hypothèse est appelée **hypothèse nulle** et est notée H_0. Ayant supposé H_0, on en a déduit la loi de probabilité que suivait le nombre de fois que l'on trouve Pile au cours de 10 lancers indépendants (c'est la partie « probas »). La partie « stats » consiste à désigner un intervalle comme probable et à définir son complémentaire comme zone de rejet.

(ii) Comme dans l'exemple 1, ce raisonnement n'est pas un raisonnement par l'absurde *stricto sensu*, puisque le résultat rejeté n'est pas en soi impossible. Il est possible qu'une pièce équilibrée lancée dix fois d'affilée donne dix fois Pile. C'est juste improbable, et cette probabilité est même de 1/1024. Le rejet de l'hypothèse est donc effectué avec un risque, que l'on appelle **risque de première espèce ou risque** α, et il est bon de

le choisir a priori. En biologie, on choisit souvent $\alpha = 0.05 = 5\%$. Ce risque, ou niveau du test, quantifie la probabilité de **rejeter H_0 alors qu'elle est vraie, c'est-à-dire d'obtenir un faux positif**. Attention : il ne nous donne aucune probabilité sur la véracité de H_0. Si au lieu de choisir $\alpha = 0.05$ on avait choisi $\alpha = 0.005$, alors on aurait pu rejeter $y = 10$ mais pas $y = 9$.

(iii) Ne pas rejeter H_0 ne signifie pas qu'elle est vraie. Dans le premier exemple, on peut très bien trouver 7 fois Pile sur 10 (et dans ce cas on n'aurait pas rejeté H_0) et pourtant être dans le cas où H_0 fausse, c'est-à-dire avec une pièce déséquilibrée avec $p > 1/2$. Ce deuxième type d'erreur, qui est donc le fait de **ne pas rejeter H_0 alors qu'elle est fausse**, c'est-à-dire d'avoir un **faux négatif**, est appelé **erreur de seconde espèce**, et est inversement corrélée à l'erreur de première espèce : moins on a de faux positifs, plus on a de faux négatifs, et inversement.

(iv) Illustrons le danger de ce type de raisonnement, s'il est mal compris : si l'on obtient 2 comme réalisation d'une variable aléatoire X, on rejette que $X \sim \mathcal{N}(0, 1)$ (c'est l'exemple 1). Cependant, si on réalise plusieurs fois cette expérience, on peut très bien trouver 2 une fois alors que la loi est bien $\mathcal{N}(0, 1)$. Avec 20 réalisations d'une telle expérience, on obtient même en moyenne une fois quelque chose de supérieur à 2. Le problème est donc d'avoir fait n tests (sur chacune des valeurs trouvées) : on tombe alors dans l'écueil du test multiple. Si l'on a effectué plusieurs fois une expérience (et c'est très bien !), il faut inclure les n variables aléatoires dans une seule variable aléatoire, et comparer la réalisation (cette fois unique) de cette variable aléatoire avec sa distribution théorique. Dans notre deuxième exemple, on a effectué n expériences $X_1, ..., X_n$ et on a introduit la variable aléatoire $Y_n = X_1 + ... + X_n$. Pour ces n lancers de pièce, on a comparé qu'une valeur (celle de Y_n, autrement dit le nombre de fois qu'on a obtenu Pile) avec sa distribution théorique. Et l'on voit bien que ce n'est pas le nombre de tests considérés séparément (X_1, ..., X_n) qui donne de la puissance au test statistique lorsque n est plus grand, mais c'est la forme de la distribution des Y_n, qui, lorsque n croît, permet de considérer de plus en plus d'événements comme improbables.

(v) Cet exemple illustre également l'importance du théorème central limite. On a vu que la loi suivie par la variable aléatoire « nombre de fois que l'on a obtenu Pile » se rapproche, lorsque n augmente, d'une loi normale.

2.2 Statistiques : théorie et pratique

2.2.1 Formalisme : définitions et premières remarques

Dans la suite, on supposera que les observations recueillies $x_1, ..., x_n$ sont les réalisations de variables aléatoires $X_1, ..., X_n$ de loi(s) inconnue(s). On pose $Y_n = f(X_1, ..., X_n)$ une variable aléatoire fonction des X_i et $y_n = f(x_1, ..., x_n)$ sa réalisation.

A. Hypothèse nulle et hypothèse alternative

Soit H_0 une hypothèse à tester sur la loi des X_i (appelée **hypothèse nulle**) et $H_1 = \bar{H}_0$ l'**hypothèse alternative**.

On appelle **test d'hypothèse** une règle de décision qui au vu de l'observation y permet de décider si on peut ou non rejeter H_0.

Un test est déterminé par sa **région critique (ou zone de rejet)** W qui constitue un sous-ensemble des valeurs possibles de Y_n. Lorsque l'on observe y,

1. si $y_n \in W$, alors on rejette H_0 et on accepte H_1;
2. si $y_n \notin W$, alors on ne rejette pas H_0 (ni H_1).

Remarque. La zone de rejet est, en règle générale, choisie pour être la plus grande possible (tout comme l'intervalle de confiance était choisi le plus petit possible). Ainsi, par exemple, sous H_0,

1. si la loi de Y_n est une loi de $\chi^2(k)$ avec k petit, la zone de rejet sera $[\gamma, +\infty[$, où γ est tel que
$$P_{H_0}(0 \leq Y_n \leq \gamma) = 1 - \alpha \quad i.e. \quad P_{H_0}(Y_n \geq \gamma) = \alpha$$
on parle alors de **test unilatéral** ;

2. si la loi de Y_n est une loi normale centrée, la zone de rejet sera $]-\infty, -\gamma] \cup [\gamma, +\infty[$, où γ est tel que
$$P_{H_0}(-\gamma \leq Y_n \leq \gamma) = 1 - \alpha \quad i.e. \quad P_{H_0}(Y_n \geq \gamma) = \frac{\alpha}{2}$$
on parle alors de **test bilatéral**.

Cependant, il arrive que l'hypothèse H_1 nous pousse à faire un autre choix pour la zone de rejet. Par exemple, si l'on sait que Y_n suit une loi normale d'espérance $\mu \geq 0$ et que $H_0 : \mu = 0$, alors $H_1 : \mu > 0$. Ainsi, on sait que si y_n est très petit, ce sera le fruit du hasard, donc on ne veut pas mettre de valeurs de Y_n négatives dans la zone de rejet. On va alors choisir un test unilatéral (dans ce cas, la zone de rejet ne sera plus maximale).

B. Erreurs

On appelle **erreur de première espèce** (ou **erreur de type I**, ou improprement erreur α) le rejet de H_0 à tort ; elle est mesurée par le risque de première espèce :

$$\alpha = P(Y_n \in W | H_0)$$

Le risque de première espèce est aussi appelé **niveau de signification** d'un test.

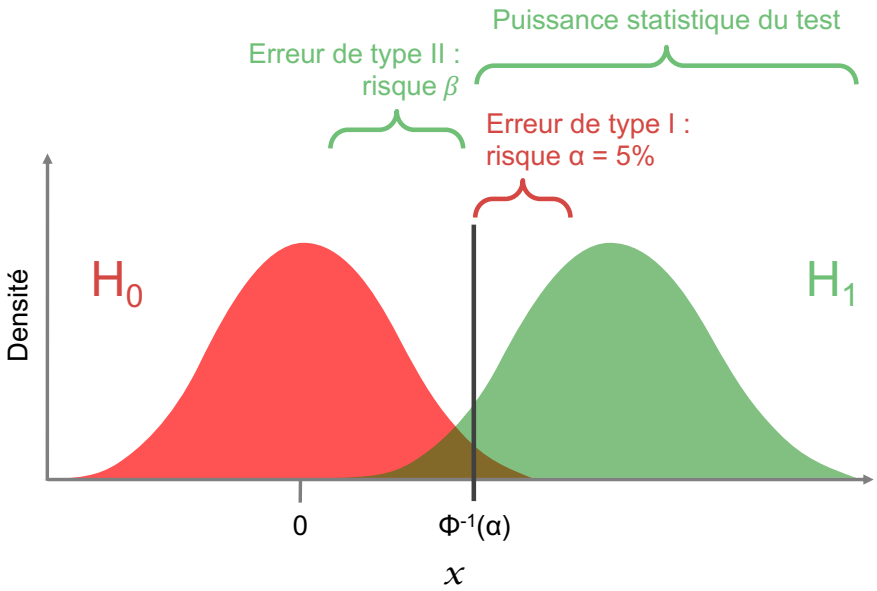

Figure 4: Un test statistique avec les distributions de son hypothèse nulle associée H_0 et de l'hypothèse alternative H_1

On appelle **erreur de seconde espèce** (ou **erreur de type II**, ou improprement erreur β) l'acceptation de H_0 à tort ; elle est mesurée par le risque de seconde espèce :

$$\beta = P(Y_n \notin W | H_1)$$

Remarque. Par convention, lors d'un test, on minimise en priorité le risque de première espèce, aussi les rôles de l'hypothèse nulle H_0 et de l'hypothèse alternative H_1 ne sont pas symétriques. Par conséquent :

1. Le choix de H_0 dépend donc du problème considéré : on choisit comme hypothèse nulle celle que l'on ne souhaite surtout pas voir rejetée à tort ;
2. Souvent, le risque de seconde espèce n'est pas quantifié, donc si y_n n'est pas dans la zone de rejet, on dira non pas que H_0 est vraie, mais simplement qu'on ne la rejette pas au niveau α.

La **puissance statistique d'une expérience** (ou pouvoir statistique) est la capacité de détecter une différence lorsque cette dernière existe, c'est-à-dire la probabilité de rejeter H_0 lorsque H_0 est fausse (Figure 4). On remarque donc que la puissance statistique égale $1 - \beta$.

$$\text{Puissance} = 1 - \beta = P(Y \in W | H_1)$$

En résumé, on a donc :

Réalité / Résultat du test	H_0 est vraie	H_0 est fausse
Non rejet de H_0	$1 - \alpha$	β
Rejet de H_0	α	$1 - \beta$

C. P-value (valeur p)

La **p-value (valeur p)** est la probabilité, sous H_0, que Y_n prenne une valeur plus extrême que la réalisation y_n, dans un sens défini par le choix de la zone de rejet.

Pour un test unilatéral, la p-value p est définie par :

$$p = P_{H_0}(Y_n \geq y_n) \text{ ou } \; p = P_{H_0}(Y_n \leq y_n)$$

Pour un test bilatéral, la p-value p est définie par

$$p = P_{H_0}(Y_n \leq -|y_n| \text{ ou } \; Y_n \geq |y_n|).$$

Il suit que

1. H_0 est rejetée si $p \leq \alpha$;
2. H_0 n'est pas rejetée si $p > \alpha$.

D. Tests paramétriques *versus* tests non paramétriques

Lorsque l'on suppose que les échantillons sont issus d'une **distribution paramétrée** (c'est-à-dire qui ne dépend que d'un nombre fini de paramètres), on parle de **test paramétrique**.

Sinon, on parle de **test non paramétrique**. Les tests non paramétriques nécessitent donc des hypothèses beaucoup moins fortes, mais sont généralement moins puissants que les tests paramétriques, c'est-à-dire qu'à risque α donné, un test paramétrique aura souvent un risque β plus petit que son équivalent non paramétrique.

E. Variables quantitatives *versus* qualitatives

En probabilités, on regroupe les variables aléatoires entre variables discrètes (à valeurs dans un espace fini ou dénombrable) et variables continues (à valeurs dans un ensemble non dénombrable, et admettant une densité de probabilité). Dans le prérequis, nous n'avons cependant vu que des variables aléatoires à valeurs dans des sous-ensemble de $\mathbb{R}$; on appelle une telle variable aléatoire une **variable quantitative**, car elle représente une quantité.

Mais il se peut qu'une variable aléatoire prenne des valeurs qui ne représentent pas des quantité : c'est le cas de la variable « groupe sanguin », qui est à valeur dans {A, B, AB, O}. Cet ensemble n'est pas ordonné, et n'admet même pas d'addition. Une telle variable est dite **variable qualitative**, et les éléments de l'espace d'arrivée sont appelées des **niveaux** de cette variable aléatoire.

2.2.2 La statistique en pratique

Un raisonnement statistique se décompose en plusieurs étapes :

1. On se fixe un risque α (typiquement 0.05).

2. On réalise n expériences dont les résultats sont notés $x_1, ..., x_n$.
3. On fait des hypothèses (auxquelles on croit) sur les lois $X_1, ..., X_n$ que suivent les expériences (par exemple que la loi des X_i est normale, que les X_i sont i.i.d.).
4. On pose une hypothèse que l'on veut tester, l'hypothèse nulle H_0, et on précise l'hypothèse alternative H_1.
5. On introduit une variable aléatoire Y_n qui dépend des $X_1, ..., X_n$ (par exemple, leur moyenne $\frac{X_1 + ... + X_n}{n}$).
6. On calcule la loi de Y_n (partie « probas » du raisonnement) sous H_0.
7. On déduit la zone de rejet en fonction de la loi de Y_n, le niveau α et l'hypothèse alternative H_1.
8. On calcule la réalisation de Y_n, que l'on note y_n.
9. On regarde où elle se situe : si elle est dans la zone de rejet, on rejette H_0, sinon on ne la rejette pas (au niveau α).

Dans la pratique, on n'a pas besoin de faire tout ça, car on utilise des logiciels qui calculent directement les p-values. Il reste les étapes suivantes indiquées dans l'encadré suivant.

Méthode : Analyse statistique.

1. On se fixe un risque α (typiquement 0.05).
2. On réalise n expériences dont les résultats sont notés $x_1, ..., x_n$.
3. On fait des hypothèses (auxquelles on croit) sur les lois $X_1, ..., X_n$ que suivent les expériences.
4. On pose une hypothèse que l'on veut tester, l'hypothèse nulle H_0, et son hypothèse alternative H_1.
5. On choisit un test adéquat (parmi les tests statistiques usuels).
6. On compare la p-value p à α :
   — si $p \leq \alpha$, H_0 est rejetée au niveau α;
   — si $p > \alpha$, H_0 n'est pas rejetée au niveau α.

Chapitre 3 : Tests usuels

3.1 Tests paramétriques — variables quantitatives

3.1.1 T-test à un échantillon

Exemple. Il est de notoriété publique que la température corporelle indiquant une bonne santé est de 37°C. On décide de vérifier la validité de cette assertion; pour cela, on prend la température de 50 personnes adultes en bonne santé. On peut modéliser les températures obtenues comme des réalisations de variables aléatoires i.i.d. $X_1, ..., X_{50}$. La question est donc de savoir si $E(X_i) \neq 37$, autrement dit si la température « moyenne » sur toute la population (à laquelle nous n'avons pas accès) est de 37°C.

Théorème (t-test à un échantillon). Soient $X_1, ..., X_n$ des variables aléatoires i.i.d., dont l'espérance est $E(X_i) = \mu$, sous H_0. On suppose de plus que l'une des deux conditions suivantes est vérifiée :

1. les X_i suivent une loi normale,
2. n est grand (typiquement $n \geq 30$).

Soient $M_n = \dfrac{X_1 + ... + X_n}{n}$ et $S_n^2 = \dfrac{1}{n-1}\sum_i (X_i - M_n)^2$ les estimateurs de la moyenne et de la variance respectivement. Alors, sous H_0, le test statistique du t-test est :

$$Y_n = \sqrt{n}\frac{M_n - \mu}{S_n} \sim t(n-1)$$

Interprétation. Sous H_0, vu que l'espérance de X_i vaut μ, Y_n a peu de chance d'être numériquement élevée en valeur absolue (*i.e.* éloigné de 0).

Remarque. Il n'est jamais superflu de rappeler qu'il ne faut pas confondre l'espérance de la variable aléatoire (dite aussi moyenne de la population) $\mu = E(X)$ avec son estimateur (la moyenne empirique de l'échantillon) $M_n = \dfrac{X_1 + ... + X_n}{n}$. La moyenne de la population est fixée et est indépendante de la réalisation, alors que son estimateur est une variable aléatoire, qui a elle aussi son espérance, sa variance, etc. De même, il ne faut pas confondre la variance de la population $\sigma^2 = V(X)$ avec son estimateur $S_n^2 = \dfrac{1}{n-1}\sum_i (X_i - M_n)^2$.

T-test à un échantillon
— Hypothèses : Soit X une variable aléatoire à valeurs dans $\mathbb{R}$. On réalise n expériences i.i.d. de loi P_X. L'une des deux conditions suivantes est vérifiée :

1. X suit une loi normale ;
2. n est « grand » (typiquement $n > 30$).

— Hypothèse nulle à tester : $H_0 : E(X) = \mu$, c'est-à-dire la moyenne de la population est μ.
— Test statistique : H_0 peut être testée par un t-test à un échantillon.
— Dépendance : Le résultat du test (p-value) dépend de la moyenne de l'échantillon M_n, de son écart type S_n et de la taille de l'échantillon n.
— Fonction dans R : On utilise la fonction `t.test(x=x, mu=mu, alternative="two.sided")` où `x` est le vecteur des résultats x_i, `mu` la valeur μ à tester, et `alternative` indique si on souhaite réaliser un test bilatéral (`two.sided`) ou unilatéral (`less` ou `greater`).

Remarque. Pour tester si les observations suivent une loi normale, il existe différents tests statistiques dont le test de Shapiro (dans R `shapiro.test(x=x)`), dont l'hypothèse nulle correspond à la normalité. En pratique, les tests usuels sont assez robustes à de petites déviations de normalité : une distribution des observations approximativement normale (distribution symétrique et unimodale sans valeurs extrêmes) est généralement un condition suffisante pour appliquer un T-test. Sinon, on peut envisager une transformation des données (*e.g.* passage au logarithme ou à la racine carrée) ou bien l'utilisation un test non paramétrique.

3.1.2 T-test à deux échantillons appariés

Exemple. On veut évaluer l'efficacité d'un médicament pour réduire la température corporelle. Une cohorte de 100 patient·e·s, sélectionné·e·s aléatoirement, prend ce médicament. Leur température est mesurée avant et après la prise. On peut donc considérer les résultats avant la prise comme des réalisations de variables aléatoires i.i.d. X_1, ..., X_{100} et les résultats après la prise comme des réalisations de variables aléatoires i.i.d. Y_1, ..., Y_{100}. La question est de savoir s'il y a une différence significative entre les X_i et les Y_i, mais ces variables aléatoires sont liées puisque à chaque X_i correspond un Y_i (même patient·e) : on dit que les séries sont appariées et on réalise donc un t-test à deux échantillons appariés. En fait, ce que l'on peut faire est de regarder de nouvelles variables aléatoires $Z_i = Y_i - X_i$. On veut savoir si $E(Z) = 0$ (auquel cas l'effet du médicament sur la température corporelle est nul), ou $E(Z) \neq 0$ (auquel cas le médicament a un effet positif ou négatif sur la température). On est donc ramené à un t-test à un échantillon, puisque l'on veut tester $E(Z) = 0$.

T-test à deux échantillons appariés.
— Hypothèses : Soient X et Y deux variables aléatoires à valeurs dans $\mathbb{R}$. On réalise n expériences i.i.d. de loi $P(X, Y)$. L'une des deux conditions suivantes est vérifiée :

1. la loi de $Z = Y - X$ est normale ;
2. n est « grand » (typiquement $n > 30$).

— Hypothèse nulle à tester : $H_0 : E(Z) = 0$.
— Test statistique : H_0 peut être testée par un t-test à deux échantillons appariés, ou, ce qui est équivalent, par un t-test à un échantillon.
— Dépendance : Le résultat du test dépend des moyenne M_n et écart type S_n de Z, et de la taille de l'échantillon n.
— Fonction dans R : On utilise la fonction `t.test(x=x, y=y, paired=TRUE, alternative="two.sided")` ou `t.test(x=y-x, mu=0, alternative="two.sided")` où `x` (resp. `y`) est le vecteur des x_i (resp. y_i) et `alternative` indique si on souhaite réaliser un test bilatéral (`two.sided`) ou unilatéral (`less` ou `greater`). Attention, R calcule `x-y` (et non pas `y-x`).

3.1.3 T-test à deux échantillons non appariés

Exemple. On veut tester si la température corporelle est significativement différente entre des individus sains et des individus présentant une hépatite C chronique. On mesure la température corporelle de 50 patient·e·s sain·e·s et de 50 patient·e·s malades. On peut encore une fois considérer les températures des individus sains comme des réalisations de variables aléatoires i.i.d. X_1, ..., X_{50} et celles des individus chroniques comme réalisations de variables aléatoires i.i.d. Y_1, ..., Y_{50}. Cette fois-ci, on n'a pas affaire à des séries appariées (X_1 et Y_1 n'ont aucun rapport).

Théorème (t-test à deux échantillons non appariés). Soient X_1, ..., X_n des variables aléatoires i.i.d., d'espérance μ et de variance σ^2. Soient Y_1, ..., Y_n des variables aléatoires i.i.d., d'espérance μ et de variance σ^2, tels que les X_i sont indépendants des Y_i (non appariés). Sous H_0, on considère l'égalité des espérances des X_i et Y_i. On suppose de plus que l'une des deux conditions suivantes est vérifiée :

1. les X_i et les Y_i suivent des lois normales,
2. n est grand (typiquement $n \geq 30$).

Soient $M_X = \dfrac{X_1 + ... + X_n}{n}$ et $S_X^2 = \dfrac{1}{n-1}\sum_i (X_i - M_X)^2$ (et de même pour M_Y et S_Y^2). Alors, sous H_0, on a :

$$Z_n = \sqrt{n}\frac{M_X - M_Y}{\sqrt{S_X^2 + S_Y^2}} \sim t(2(n-1))$$

Remarque. Il existe une version de ce test pour des tailles d'échantillons n_X et n_Y différentes pour les deux séries d'échantillons :

$$Z_n = \frac{M_X - M_Y}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim t(n_X + n_Y - 2)$$

Interprétation. Sous H_0, vu que X et Y ont la même espérance, Z_n a peu de chance d'être numériquement élevée en valeur absolue (*i.e.* éloigné de 0).

T-test à deux échantillons non appariés.
— Hypothèses : Soient X et Y deux variables aléatoires indépendantes à valeurs dans $\mathbb{R}$. On réalise n_X expériences i.i.d. de loi P_X, et n_Y expériences i.i.d. de loi P_Y. L'une des deux conditions suivantes est vérifiée :

1. X et Y suivent des lois normales ;
2. n_X et n_Y sont « grands » (typiquement > 30).

— Hypothèse nulle à tester : $H_0 : E(X) = E(Y)$.
— Test statistique : H_0 peut être testée par un t-test à deux échantillons non appariés.
— Dépendance : Le résultat du test dépend de M_X, M_Y, S_X, S_Y, n_X et n_Y.
— Fonction dans R : On utilise la fonction `t.test(x=x, y=y, alternative="two.sided")`, où `x` (resp. `y`) est le vecteur des x_i (resp. y_j) et `alternative` indique si on souhaite réaliser un test bilatéral (`two.sided`) ou unilatéral (`less` ou `greater`).

Remarque. Le t-test à deux échantillons permet de tester l'égalité des moyennes entre deux groupes. Un test équivalent pour tester l'égalité des variances entre deux groupes est le test de Fisher d'égalité de deux variances (ou F-test), dont la variable aléatoire F_n définie comme le ratio des variances empiriques estimées pour les deux groupes suit une loi de Fisher : sous H_0, les variances sont identiques et $F_n = S_X^2/S_Y^2 \sim \mathcal{F}(n_X - 1, n_Y - 1)$.

Test de Fisher d'égalité de deux variances (ou F-test).
— Hypothèses : Soient X et Y deux variables aléatoires indépendantes à valeurs dans $\mathbb{R}$. On réalise n_X expériences i.i.d. de loi P_X, et n_Y expériences i.i.d. de loi P_Y. L'une des deux conditions suivantes est vérifiée :

1. X et Y suivent des lois normales ($X \sim \mathcal{N}(\mu_X, \sigma_X)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$) ;
2. n_X et n_Y sont « grands » (typiquement > 30).

— Hypothèse nulle à tester : $H_0 : V(X) = V(Y)$.
— Test statistique : H_0 peut être testée par un test de Fisher d'égalité de deux variances.
— Dépendance : Le résultat du test dépend de M_X, M_Y, S_X, S_Y, n_X et n_Y.
— Fonction dans R : On utilise la fonction `var.test(x=x, y=y, alternative="two.sided")`, où `x` (resp. `y`) est le vecteur des x_i (resp. y_j) et `alternative` indique si on souhaite réaliser un test bilatéral (`two.sided`) ou unilatéral (`less` ou `greater`).

3.1.4 ANOVA à un facteur

Exemple. On cherche à savoir si la température corporelle varie en fonction de la classe d'âge chez l'être humain. On mesure la température d'individus sélectionnés aléatoirement parmi les classes d'âge suivantes : les enfants, les adolescent·e·s, les adultes, et les personnes âgées. Plus précisément, si on considère que les températures mesurées dans la classe d'âge i sont des réalisations de variables aléatoires i.i.d. $X_1^i,...,X_{n_i}^i$, et en notant $\mu_i = E(X_j^i)$, on veut tester $H_0 : \mu_1 = ... = \mu_4$.

S'il n'y avait que deux classes d'âge, on pourrait utiliser un t-test. Mais pour tester que les moyennes sont toutes les mêmes, il faudrait faire plusieurs tests, et l'on risque d'avoir des erreurs liées au test multiple. Sous des bonnes hypothèses, il faudra utiliser le test d'analyse de la variance (ANOVA, pour ANalysis Of VAriance) à un facteur.

Théorème (ANOVA à un facteur (mesures indépendantes)). Pour tout $1 \leq i \leq p$, on suppose que $X_1^i,...,X_{n_i}^i$ sont des variables aléatoires i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$, que les X_j^i sont tous deux à deux indépendants. Sous H_0, on considère l'égalité des espérances des X_j^i. Soit $n = \sum_{i=1}^{p} n_i$. Soient les variables aléatoires

$$M_i = \frac{X_1^i + ... + X_{n_i}^i}{n_i}, \quad M_n = \frac{\sum_{i,j} X_j^i}{n}$$

$$SCE_{\text{modèle}} = \sum_i n_i (M_i - M_n)^2 \quad \text{et} \quad SCE_{\text{résiduelle}} = \sum_j \sum_i (X_j^i - M_i)^2.$$

Alors, sous H_0, on a :

$$F_n = \frac{SCE_{\text{modèle}}/(p-1)}{SCE_{\text{résiduelle}}/(n-p)} \sim F(p-1, n-p)$$

Interprétation. SCE est l'acronyme de « somme des carrés des écarts » (voir Figure 5). On voit bien ici qu'on compare les valeurs de deux types de variances : la variance expliquée par le modèle (entre les groupes) *versus* la variance non expliquée par le modèle (résiduelle), d'où le nom d'ANOVA (ANalysis Of VAriance). Sous H_0, F_n a peu de chance d'être numériquement élevée (la variance expliquée par le modèle est sensée être du même ordre que la variance résiduelle car il n'y a pas d'effet des groupes testés).

Remarque. Il existe une version de ce test pour des mesures répétées (généralisation de la notion d'appariement pour trois mesures ou plus), appelée ANOVA à mesures répétées. Il existe également une version de ce test lorsque l'on teste deux facteurs en interaction : par exemple l'effet de la classe d'âge et de la prise d'un médicament sur la température corporelle; on parle alors d'**ANOVA à deux facteurs**.

ANOVA à un facteur (mesures indépendantes).
— Hypothèses : Pour chaque $1 \leq i \leq p$, on réalise n_i expériences indépendantes de loi $\mathcal{N}(\mu_i, \sigma^2)$; on suppose que les variables aléatoires sont toutes indépendantes. En d'autres termes, il faut :

1. l'indépendance des lois ;
2. la normalité des lois ;
3. l'égalité des variances entre séries d'expériences : on appelle cette condition l'**homoscédasticité**.

— Hypothèse nulle à tester : $H_0 : \mu_1 = ... = \mu_p$.
— Test statistique : H_0 peut être testée par une ANOVA à un facteur (à mesures indépendantes).
— Dépendance : Le résultat du test dépend de $SCE_{\text{modèle}}$, $SCE_{\text{résiduelle}}$, n et p.
— Fonction dans R : On utilise la fonction `aov(y~f, data=df)`, où `df` est un data frame avec pour colonnes : `y` le vecteur des x_j^i (observations) et `f` le vecteur contenant les facteurs i correspondants aux différents niveaux.

Remarque. Pour tester l'égalité des variances, il existe différents tests statistiques dont le test de Bartletts et de Levene (dans R `bartlett.test(y~f, data=df)` et `car::leveneTest(y~f, data=df)`), dont les hypothèses nulles correspondent à l'homoscédasticité. En cas de rejet, on peut envisager une transformation des données (*e.g.* passage au logarithme ou à la racine carrée) ou bien l'utilisation un test non paramétrique.

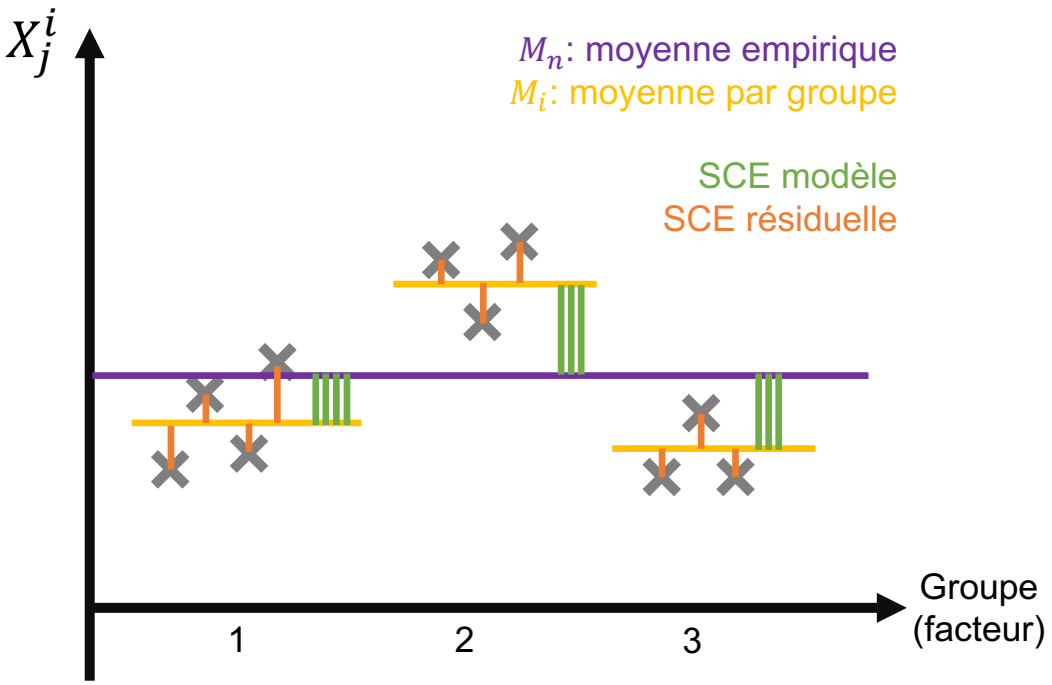

Figure 5: Illustration de l'ANOVA.

3.1.5 Régression linéaire simple

Exemple. On cherche à étudier si la température corporelle tend à augmenter linéairement au cours de la journée. On considère le temps X_i comme une variable continue qui indique le nombre de minutes écoulées depuis 9h du matin. On mesure la température de plusieurs individus à différents moments de la journée. Plus précisément, si on considère que les températures mesurées sont des réalisations de variables aléatoires i.i.d. Y_1,...,Y_n au temps X_1, ..., X_n, on cherche à tester si $Y_i = aX_i + b$ avec a et b non nuls.

Théorème (régression linéaire simple). On considère les variables aléatoires i.i.d. Y_1, ..., Y_n appariées aux variables aléatoires i.i.d. X_1, ..., X_n et y_1, ..., y_n et x_1, ..., x_n les observations correspondantes. On pose $y_i = ax_i + b + \epsilon_i$ où a correspond à la pente (« slope »), b correspond à l'ordonnée à l'origine (« intercept ») et ϵ_i correspond à l'erreur résiduelle aléatoire que l'on considère comme i.i.d. : $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

En d'autres termes, on a $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ avec $\mu_i = aX_i + b$.

On peut déterminer (en utilisant par exemple la méthode des moindres carrés) les valeurs de a et b qui minimisent le carré des termes d'erreurs de la régression :

$$(\hat{a}, \hat{b}) = argmin_{(a,b)} \left(\sum_{i=1}^{n} \epsilon_i^2 \right) = argmin_{(a,b)} \left(\sum_{i=1}^{n} (y_i - ax_i - b)^2 \right)$$

Pour chaque paramètre de la régression, l'hypothèse nulle H_0 correspond à la nullité du paramètre et sous ces H_0, les valeurs respectives des paramètres a et b sont alors données par les lois suivantes :

$$T_a = \frac{\hat{a}}{SE(\hat{a})} \sim t(n-2) \quad \text{et} \quad T_b = \frac{\hat{b}}{SE(\hat{b})} \sim t(n-2)$$

calculées à partir des erreurs types (« standard errors », notées SE) :

$$SE(\hat{a}) = \sqrt{S_n^2 \left(\frac{1}{n} + \frac{M_X}{S_{xx}} \right)} \quad \text{et} \quad SE(\hat{b}) = \sqrt{\frac{S_n^2}{S_{xx}}}$$

avec $M_X = \dfrac{X_1 + ... + X_n}{n}$, $S_n^2 = \dfrac{\sum_{i=1}^{n} \epsilon_i^2}{n-2}$ et $S_{xx} = \displaystyle\sum_{i=1}^{n} (X_i - M_X)^2$.

Remarque. On peut s'intéresser au partitionnement de la variance totale par le modèle en calculant différentes sommes de carrés des écarts (SCE; voir Figure 6):

$$SCE_{\text{totale}} = \sum_{i=1}^{n} (Y_i - M_Y)^2, \; SCE_{\text{modèle}} = \sum_{i=1}^{n} (\hat{Y}_i - M_Y)^2 \text{ et } SCE_{\text{résiduelle}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

avec $M_Y = \dfrac{Y_1 + ... + Y_n}{n}$ et $\hat{Y}_i = \hat{a}X_i + \hat{b}$.

On a donc $SCE_{\text{totale}} = SCE_{\text{modèle}} + SCE_{\text{résiduelle}}$.

On peut alors définir **le coefficient de determination (R^2)**, à partir de la partition de variance expliquée par le modèle :

$$R^2 = \frac{SCE_{\text{modèle}}}{SCE_{\text{totale}}} = 1 - \frac{SCE_{\text{résiduelle}}}{SCE_{\text{totale}}}$$

R^2 varie entre 0 et 1 et indique la proportion de la variabilité de Y qui peut être expliquée par la régression linéaire. De plus, ce partitionnement de la variance permet de retomber sur le cadre des analyses de variance (ANOVA; *cf* paragraphe précédent) et on peut définir la variable aléatoire suivante afin d'évaluer la significativité de la regression linéaire dans son ensemble :

$$F_n = \frac{SCE_{\text{modèle}}/(p-1)}{SCE_{\text{résiduelle}}/(n-p)} \sim \mathcal{F}(p-1, n-p)$$

Interprétation. Sous H_0, F_n a peu de chance d'être numériquement élevée (la variance expliquée par la régression linéaire est sensée être du même ordre que la variance résiduelle si il n'y a pas de relation linéaire entre les deux variables).

On remarque ici la forte similarité entre la régression linéaire et l'ANOVA : elles appartiennent toutes deux à la famille des **modèles linéaires** : la variable à expliquer Y_i est continue et la variable explicative est soit continue (régression linéaire), soit catégorielle (ANOVA). Dans R, on peut utiliser la même fonction dans les deux cas : la fonction `lm()` (pour *linear model*).

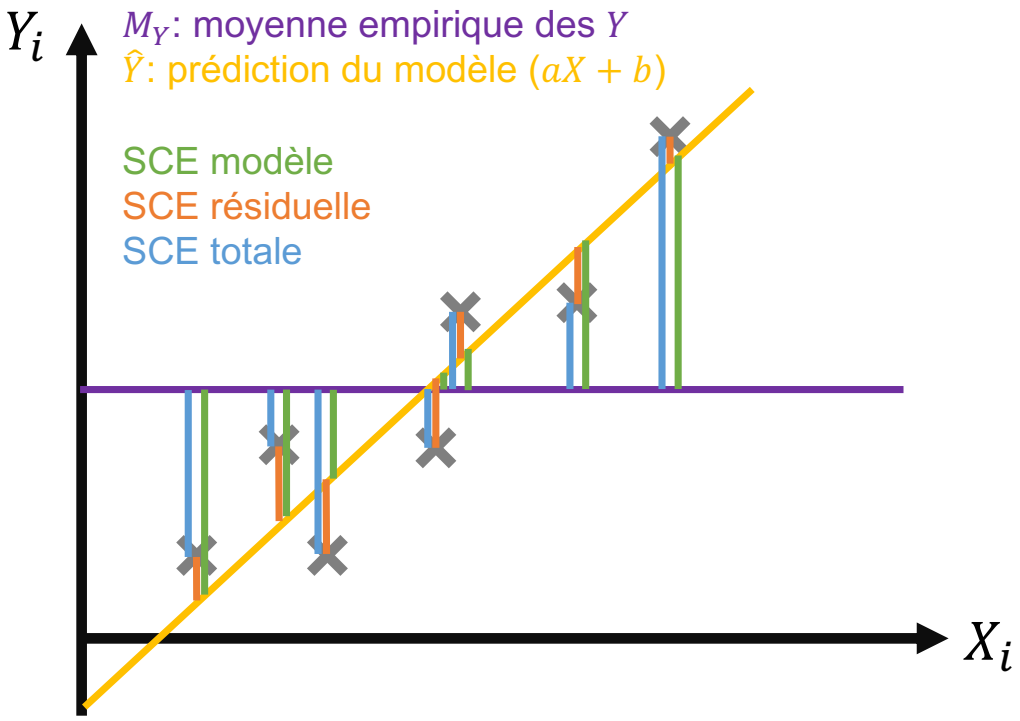

Figure 6: Illustration de la régression linéaire.

Régression linéaire simple.
— Hypothèses : Pour chaque échantillon $1 \leq i \leq n$, on réalise des mesures des x_i et y_i des variables aléatoires X_i et Y_i. On pose $Y_i = aX_i + b$. On vérifie les hypothèses suivantes :

1. X_i et Y_i sont des variables continues ;
2. il existe une relation affine entre les X_i et Y_i ;
3. les termes d'erreurs $\epsilon_i = y_i - ax_i - b$ sont indépendants et identiquement distribuées: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
4. la variance σ^2 des termes d'erreurs est constante pour toutes valeurs de x_i (homoscédasticité).

— Hypothèse nulle à tester pour évaluer la significativité des coefficients :

1. pour a : $H_0^a : a = 0$
2. pour b : $H_0^b : b = 0$

— Test statistique : Les hypothèses nulles H_0 peuvent être testées par des t-tests (à $n - 2$ degrés de liberté).
— Dépendance : Le résultat du test dépend de n, $\hat{a}$, $\hat{b}$ et de leur erreur type associée $SE(\hat{a})$ et $SE(\hat{b})$.
— Fonction dans R : On utilise la fonction `summary(lm(y~x, data=df))`, où `df` est un data frame avec pour colonnes : `y` le vecteur des y_i et `x` le vecteur des x_i correspondants.

3.2 Tests non paramétriques — variables quantitatives

Les tests que nous avons vus jusqu'à présent supposent soit des lois normales (ce qu'il est dur de vérifier), soit de grands échantillons ($n > 30$ typiquement). Si ces **hypothèses ne sont a priori pas vérifiées**, il vaut mieux utiliser, lorsqu'il existe, un **test équivalent non paramétrique**. Ces tests reposent sur l'analyse des rangs des variables aléatoires quantitatives. Ils sont moins puissants (mais souvent presque aussi puissants sur des lois normales) mais devraient bien souvent être privilégiés. Nous revenons sur les tests paramétriques présentés précédemment et indiquons leur équivalent non paramétrique.

Comparaison	Test paramétrique	Test non paramétrique	Fonction sur R
moyenne d'un échantillon	t-test à un échantillon	Wilcoxon signed rank (Wilcoxon à un échantillon)	`wilcox.test (x=x, mu=mu)`
2 échantillons appariés	t-test à 2 échantillons appariés	Wilcoxon signed rank (Wilcoxon à 2 échantillons appariés)	`wilcox.test(x=x, y=y, paired=TRUE)`
2 échantillons non appariés	t-test à 2 échantillons non appariés	Mann–Whitney U (Wilcoxon à 2 échantillons non appariés)	`wilcox.test(x=x, y=y, paired=FALSE)`
$p \geq 3$ échantillons indépendants	ANOVA à un facteur	Kruskal–Wallis (nécessite $n_i \geq 5$ dans chaque groupe)	`kruskal.test (y ~ f, data=df)`

3.3 Tests — variables qualitatives

Les tests que nous avons étudiés jusqu'à présent ne sont pas aisément applicables aux **variables qualitatives**. Ces dernières ne pouvant suivre une loi normale, les t-tests, ANOVA et régressions linéaires ne sont pas adaptés. De plus, l'ensemble d'arrivée ne pouvant a priori être ordonné, les tests de Wilcoxon ne sont pas applicables non plus. L'astuce, pour la mise en place de tests pour des variables qualitatives, est de se ramener à l'étude de variables quantitatives.

3.3.1 Test du χ^2 d'adéquation

Exemple. On cherche à étudier si le risque d'avoir de la fièvre dépend de la classe d'âge. On reporte la classe d'âge de 1000 personnes sélectionnées au hasard dans la population française ayant présenté un épisode fiévreux au cours des 3 derniers mois : enfant, adolescent, adulte ou personne âgée. L'idée est la suivante : on suppose que l'appartenance à une certaine classe d'âge pour les 1000 individus fiévreux provient de la réalisation des **variables aléatoires qualitatives** i.i.d. X_1, ..., X_{1000} à valeurs dans {enfant, ado, adulte, sénior} (variable à $m = 4$ niveaux). L'hypothèse nulle H_0 à tester est la suivante : pour toute classe d'âge,

$$P(X_i = \text{classe d'âge}) = (\text{proportion d'individus de cette classe d'âge dans la population})$$

On va alors comparer les effectifs théoriques (proportions théoriques d'enfants, d'adolescents, d'adultes et de personnes âgées dans la population totale) et les effectifs observés (nombres observées d'enfants, d'adolescents, d'adultes et de personnes âgées présentant de la fièvre).

Théorème (test du χ^2 d'adéquation). Soient X_1, ..., X_n des variables aléatoires i.i.d. à valeurs dans un ensemble fini $F = \{a_1, ..., a_m\}$. On suppose que pour tout $1 \leq j \leq m$, $P(X_i = a_j) = p_j$, appelée proportion théorique. Soient, pour tout $1 \leq j \leq m$, $O_j = card(\{i, X_i = a_j\})$ l'effectif observé, $E_j = p_j n$ l'effectif théorique (*expected* en anglais) pour a_j.
Alors, sous H_0, lorsque $n \to \infty$,

$$U_n = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j} \to \chi^2(m-1)$$

Interprétation. Sous H_0, U_n a peu de chance d'être numériquement élevée (les effectifs observés O_j sont proches des effectifs théoriques E_j).

Test du χ^2 d'adéquation
— Hypothèses : Soit X une variable aléatoire à valeurs dans un ensemble fini $F = \{a_1, ..., a_m\}$. On réalise n expériences i.i.d. de loi P_X. On suppose que n est grand, c'est-à-dire en pratique que les effectifs théoriques E_j doivent tous être ≥ 1 et 80% d'entre eux doivent être ≥ 5.
— Hypothèse nulle à tester : $H_0 : \forall 1 \leq j \leq m, P(X = a_j) = p_j$.
— Test statistique : H_0 peut être testée par un test du χ^2 d'adéquation.
— Dépendance : Le résultat du test dépend des effectifs théoriques E_j et observés O_j.
— Fonction dans R : On utilise la fonction `chisq.test(x=x, p=p)`, où `x` le vecteur des observations (effectifs observés O_j) et `p` le vecteur des probabilités théoriques (a_j).

Remarque. Ce test nécessite des comptes absolus et non des pourcentages. C'est conforme à l'intuition, car 200 *versus* 100 doit être plus significatif que 2 *versus* 1.

3.3.2 Test du χ^2 d'indépendance

Exemple. On cherche maintenant à savoir si le risque de présenter un épisode fiévreux en hiver est diminué par la prise préventive de vitamines. On interroge ainsi 10000 individus à la fin de l'hiver et on note si chaque individu a présenté ou non un épisode fiévreux et si il a suivi ou non une cure de vitamines en début d'hiver. Sous l'hypothèse nulle, le risque de fièvre est indépendant de la prise de vitamines. Afin de la tester, on note les résultats dans un **tableau de contingence** :

	Prise de vitamines	Pas de prise
Fièvre	$O_{V,F}$	$O_{P,F}$
Absence de fièvre	$O_{V,A}$	$O_{P,A}$

L'idée est alors la même que précédemment : on va comparer les effectifs observés avec les effectifs théoriques calculés dans l'hypothèse d'une indépendance des variables « fièvre » et « vitamines ».

Théorème (test du χ^2 d'indépendance). Soient $X_1, ..., X_n$ des variables aléatoires i.i.d. à valeurs dans un ensemble fini $F = \{a_1, ..., a_{m_X}\}$, $Y_1, ..., Y_n$ des variables aléatoires i.i.d. à valeurs dans un ensemble fini $G = \{b_1, ..., b_{m_Y}\}$. On suppose que les X_i sont indépendants des Y_i. Soient, pour tous $1 \leq j \leq m_X, 1 \leq k \leq m_Y, O_{j,k} = card(\{i, X_i = a_j \text{ et } Y_i = b_k\})$ l'effectif observé, et

$$E_{j,k} = \frac{O_{j,+} \times O_{+,k}}{n}$$

l'effectif théorique pour (a_j, b_k), où $O_{j,+} = \sum_k O_{j,k}$ (somme des effectifs pour $X = j$) et $O_{+,k} = \sum_j O_{j,k}$ (somme des effectifs pour $Y = k$).
Alors, sous H_0, lorsque $n \to \infty$,

$$U_n = \sum_{j,k} \frac{(O_{j,k} - E_{j,k})^2}{E_{j,k}} \to \chi^2((m_X - 1)(m_Y - 1))$$

Interprétation. Sous H_0, U_n a peu de chance d'être numériquement élevée (les effectifs observés $O_{j,k}$ sont proches des effectifs théoriques $E_{j,k}$ lorsque les X_i et Y_i sont indépendants).

Test du χ^2 d'indépendance.
— Hypothèses : Soient deux variables aléatoires X et Y à valeurs dans des ensembles finis. On réalise n expériences i.i.d. de loi $P(X, Y)$. On suppose que n est grand, c'est- à-dire en pratique que les effectifs théoriques $E_{j,k}$ doivent tous être ≥ 1 et 80% d'entre eux doivent être ≥ 5.
— Hypothèse nulle à tester (H_0) : les lois de X et Y sont indépendantes.
— Test statistique : H_0 peut être testée par un test du χ^2 d'indépendance.
— Dépendance : Le résultat du test dépend des effectifs théoriques $E_{j,k}$ et observés $O_{j,k}$.
— Fonction dans R : On utilise la fonction `chisq.test(x=x)`, où `x` est la tableau de contingence (tableau des effectifs observés $O_{j,k}$).

Remarque. Comme pour le test précédent, la p-value est une valeur approchée, car n'est exacte qu'asymptotiquement (lorsque $n \to \infty$). Si le tableau de contingence est un tableau 2×2 (c'est-à-dire $card(F) = card(G) = 2$, *i.e.* X et Y n'ont que deux niveaux), il existe un test exact (c'est-à-dire qui est valable pour tout n) : c'est le **test exact de Fischer d'indépendance**. Ce test est particulièrement utile lorsque les effectifs sont trop faible pour réaliser un test du χ^2. Dans R : on utilise la fonction `fisher.test(x=x)`, où `x` est la tableau de contingence.

3.4 Étude de corrélation

Le test du χ^2 d'indépendance nous permet de toucher du doigt une autre question : lorsque nous avons deux mesures sur un même échantillon, ces mesures sont elles corrélées ? Dans cette section, nous revenons à l'étude des variables quantitatives.

Exemple. On cherche à étudier s'il existe une corrélation entre la température corporelle et la charge virale sanguine du SARS-CoV-2. On mesure ces deux variables quantitatives pour 100 individus sélectionnés au hasard parmi les individus présentant un test antigénique positif au SARS-CoV-2. Sous l'hypothèse nulle, les deux variables sont indépendantes.

3.4.1 Coefficient de corrélation de Pearson

Si X et Y sont deux variables aléatoires sur $\mathbb{R}$, alors leur corrélation est définie par

$$\rho = cor(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E((X - E(X))(Y - E(Y)))}{\sigma(X)\sigma(Y)}$$

Il vient que $\rho \in [-1, 1]$, et que $\rho = 1$ ou -1 si et seulement si X et Y sont linéairement dépendants (positivement si $\rho = 1$, négativement si $\rho = -1$).

Pour un échantillon, soient $(x_1, y_1), ..., (x_n, y_n)$ des réalisations indépendantes de lois (X, Y) et leurs moyennes empiriques m_X et m_Y. Le **coefficient de corrélation de Pearson** est :

$$r_n = r_{x,y} = \frac{\sum_{i=1}^n (x_i - m_X)(y_i - m_Y)}{\sqrt{\sum_{i=1}^n (x_i - m_X)^2}\sqrt{\sum_{i=1}^n (y_i - m_Y)^2}}$$

dont l'estimateur correspondant est :

$$R_n = R_{X,Y} = \frac{\sum_{i=1}^n (X_i - M_X)(Y_i - M_Y)}{\sqrt{\sum_{i=1}^n (X_i - M_X)^2}\sqrt{\sum_{i=1}^n (Y_i - M_Y)^2}}$$

R_n est un estimateur de ρ qui est convergent, mais il n'est pas possible de calculer son espérance pour des variables aléatoires quelconques. On ne peut donc dire s'il est biaisé.

Interprétation. En tant qu'estimation de la corrélation entre X et Y, le coefficient de corrélation de Pearson quantifie la **corrélation linéaire** entre les x_i et les y_i. Le coefficient R_n de Pearson correspond aussi à la racine carré du coefficient de détermination R^2 d'une

regression linéaire simple (*cf* paragraphe précédent).

Hypothèse nulle. Un test de significativité existe pour ce coefficient : sous H_0, $\rho = 0$ et on peut montrer que, dans le cas où X et Y suivent des lois normales (hypothèse à vérifier),

$$R_n \sqrt{\frac{n-2}{1-R_n^2}} \sim t(n-2)$$

ce qui permet de tester si X et Y sont statistiquement (linéairement) corrélées. La valeur du coefficient de Pearson, ainsi que la p-value associée, peuvent être obtenues sur R via la commande `cor.test(x, y, method="pearson")`.

Remarque. On constate la forte similarité de la régression linéaire simple avec le test de coefficient de Pearson. Les deux tests reposent sur des hypothèses fortes (normalités des résidus de la régression *versus* distributions normales des X et Y pour la corrélation) mais ne réalisent pas la même chose : le test de corrélation de Pearson mesure simplement la force de la relation linéaire entre deux variables (test symétrique), tandis que la relation linéaire modélise la variable Y comme une fonction linéaire de la variable X et permet ainsi de faire des prédictions (pour toute nouvelle valeur de X on peut prédire quelle serait la valeur de Y sous ce modèle). La régression linéaire est donc plus informative, mais il est important de noter qu'elle correspond à un test asymétrique ($Y \sim X$).

3.4.2 Coefficient de corrélation de Spearman

Pour quantifier des **corrélations monotones** (quand X croît, Y croît/décroît), mais non nécessairement linéaires, le coefficient de corrélation de Pearson n'est pas un bon outil. À la place, on travaille sur les **rangs** et on utilise le **coefficient de corrélation de Spearman**.

Si $(x_1, y_1), ..., (x_n, y_n)$ sont des réalisations de variables i.i.d. de loi (X, Y), alors le coefficient de corrélation de Spearman est le coefficient de Pearson appliqué aux rangs de x et y, c'est-à-dire

$$r_s = r_{a,b}$$

où a_i (resp. b_i) est le rang de x_i (resp. y_i). On peut montrer que, dans le cas où tous les rangs sont distincts,

$$r_s = 1 - \frac{6 \sum_i (a_i - b_i)^2}{n(n^2-1)}$$

Un test de significativité existe aussi pour ce coefficient : si X et Y sont indépendantes, alors on connaît la loi que suit R_s, ce qui permet de tester que X et Y sont statistiquement monotonement corrélées. La valeur du coefficient de Spearman, ainsi que la p-value associée, peuvent être obtenues sur R via la commande `cor.test(x,y,method="spearman")`.

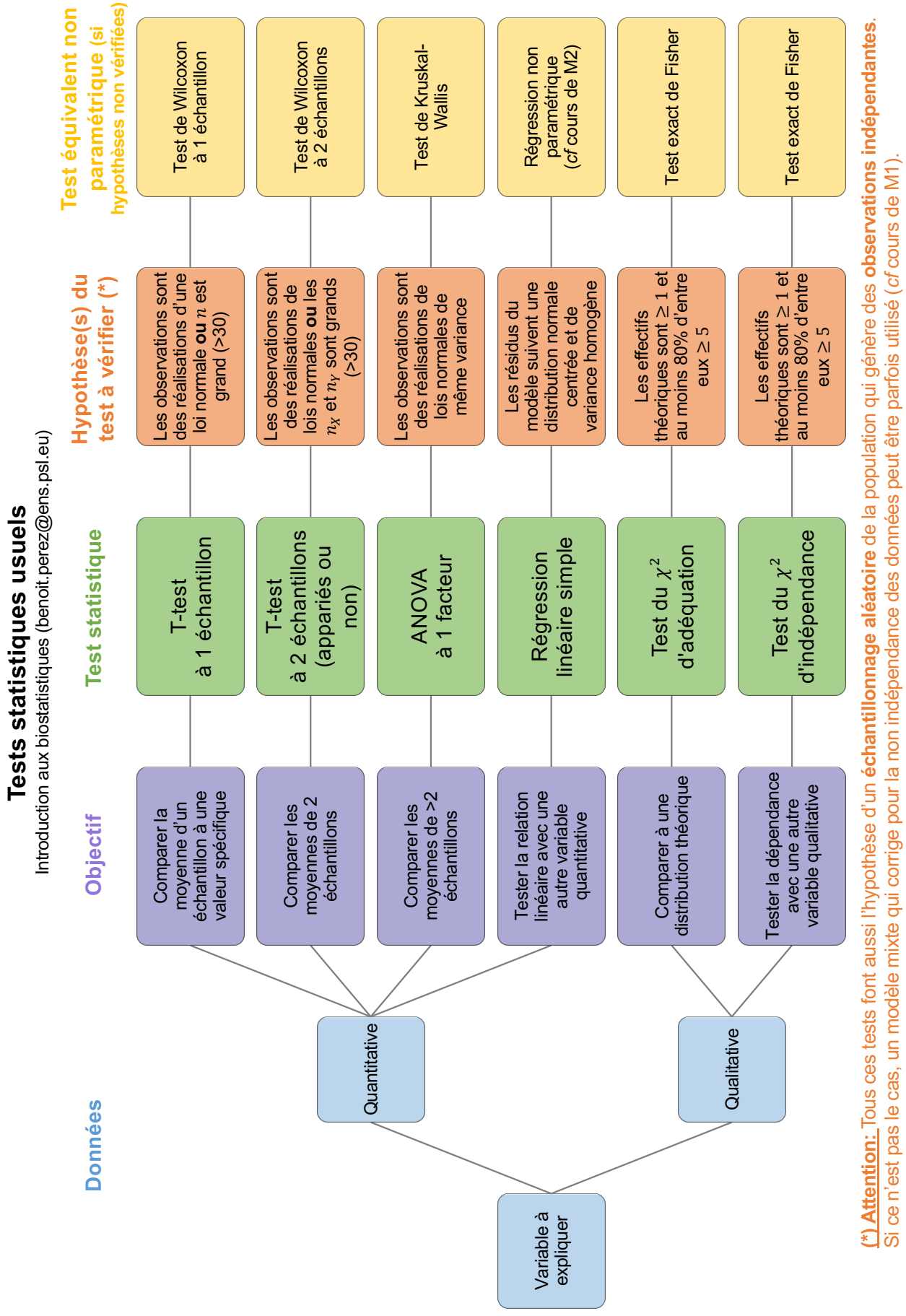

Figure 7: Synthèse des tests statistiques usuels du Chapitre

Chapitre 4 : Quelques réflexions autour des biostatistiques

4.1 Construire un plan d'expérience et puissance statistique

En statistiques, un point essentiel est l'importance de l'**échantillonnage non biaisé**, c'est-à-dire que l'on doit échantillonner les populations considérées au hasard, et non selon un quelconque critère, encore moins un critère a posteriori. Par conséquent, avant même de choisir un test statistique adéquat, c'est la collecte de données qui doit être faite de manière rigoureuse. Par exemple, aussi tentant que cela puisse paraître, il ne faut évidemment pas retirer de l'échantillon ce que l'on considère a posteriori comme un outlier. De même, il faut faire attention à ne pas biaiser son échantillonnage par un préjugé (ou un espoir) sur les résultats attendus. En microscopie par exemple, il est facile de biaiser de façon inconsciente ses résultats, si le choix de la zone à imager est faite par l'humain ; il est donc important de réaliser ce type d'expériences « en aveugle », c'est-à-dire sans savoir à quel groupe appartient un échantillon au moment de l'imager. De façon générale, gardez en tête qu'une expérience mal pensée ou réalisée de façon biaisée ne donnera jamais d'information intéressante, même avec les meilleures statistiques du monde...

Avant de commencer les manipulations ou de collecter les données, il est indispensable de construire un **plan d'expérience**. Par exemple, si l'on souhaite mesurer l'effet d'un certain facteur par rapport à un certain contrôle, m'objectif est de mettre en place une expérience telle que l'éventuelle différence mesurée entre le condition test et la condition contrôle est vraiment causée par le facteur testée et non par une autre source de variation non contrôlée.

En particulier, **avant l'expérience il est indispensable de** :

1. Définir quelle est la **question biologique** posée (*e.g.* quelle population ? quel paramètre à estimer ?) et s'assurer que l'expérience proposée y répond bien (et qu'elle contient tous les contrôles nécessaires).

2. Réfléchir aux possibles **effets confondants** (*e.g.* apparentement, age, sexe) et prévoir une échantillonnage représentatif de la population étudiée qui minimise ces possibles biais.

3. Limiter les **sources de variations techniques** (*e.g.* quel matériel ? qui fait l'expérience ? à quel moment ? dans quel ordre les échantillons sont-ils traités ?). Par exemple, au laboratoire ou sur le terrain, des **réplicats techniques** doivent être réalisés pour s'assurer de la validité des mesures et il est toujours préférable de « **randomiser** » **le traitement des échantillons** (*i.e.* traiter de manière aléatoire les échantillons: ne pas commencer par traiter tous les échantillons contrôles puis tous les échantillons tests). Dans tous les cas, il est indispensable de reporter les **conditions de l'expérience**, par exemple qui a fait l'expérience, à quel moment et dans quel ordre.

4. Considérer **la variation biologique** : prévoir suffisamment d'échantillons (**réplicats biologiques**). Pour cela il est possible de calculer la puissance statistique attendue de

l'expérience en faisant quelques hypothèses sur les échantillons qui seront collectés.

La **puissance statistique** d'une expérience est la capacité de détecter une différence lorsque cette dernière existe (*i.e.* la probabilité de rejeter H_0 lorsque H_0 est fausse). En particulier, en faisant quelques hypothèses sur l'effet attendu et sur la variabilité des observations, il est possible de répondre aux questions suivantes en faisant des analyses de la puissance statistique :

1. De combien d'échantillons n, ai-je besoin si je veux détecter une difference significative de x entre le contrôle et le facteur testé ?
2. Sachant que je dispose de n échantillons, quel est la différence minimale x entre le contrôle et le facteur testé que je pourrai détecter et considérer comme significative ?

Exemple. On veut tester l'effet d'un traitement médicamenteux sur l'augmentation de la température corporelle. On prévoit donc de mesurer la température de n individus après traitement. On considère les variables aléatoires i.i.d. $X_1, ..., X_n$. Soit M_X la moyenne des X_i et S_n leur écart type empirique. Sous H_0, le traitement n'a pas d'effet, c'est-à-dire $E(X_i) = \mu_0 = 37°\text{C}$. Sous l'hypothèse alternative H_1, $E(X_i) = \mu_1 > \mu_0$.

La statistique du test sous H_0 est donnée par : $T_n = \sqrt{n}\dfrac{M_X - \mu_0}{S_n} \sim t(n-1)$ avec un risque de première espèce α.

Sous H_1, on considère la taille de l'effet (*effect size* en anglais), défini comme $d = \dfrac{\mu_1 - \mu_0}{\sigma}$. La distribution de T_n est donc centrée sur $\sqrt{n}\dfrac{\mu_1 - \mu_0}{\sigma} = d\sqrt{n}$. On peut donc définir Y_n telle que : $Y_n = T_n - d\sqrt{n} \sim t(n-1)$.

On peut alors exprimer la puissance du test comme :

$$\begin{aligned}
Puissance &= P(\text{Rejet de } H_0 | H_1 \text{ est vraie}) \\
&= P\left(\sqrt{n}\frac{M_X - \mu_0}{S_n} > t_{k=n-1,\alpha=\alpha\%} \mid H_1 \text{ est vraie}\right) \\
&= P\left(\sqrt{n}\frac{M_X - \mu_0}{S_n} - d\sqrt{n} > t_{k=n-1,\alpha=\alpha\%} - d\sqrt{n} \mid H_1 \text{ est vraie}\right) \\
&= P\left(Y_n > t_{k=n-1,\alpha=\alpha\%} - d\sqrt{n} \mid H_1 \text{ est vraie}\right) \text{ avec } Y_n \sim t(n-1)
\end{aligned}$$

La puissance statistique dépend donc de α, n et d. En faisant des hypothèses sur d, on peut notamment calculer quel est le nombre d'échantillons n nécessaire afin d'avoir une puissance d'au moins 0.95.

Par exemple, si $\mu_1 = 37.5$ et que $\sigma = 0.5$, alors $d = 1$, et on cherche n tel que :

$$P\left(Y_n > t_{k=n-1,\alpha=5\%} - d\sqrt{n}\right) = 0.95$$

En prenant $n = 10$ comme première approximation, on a : $1.83 - \sqrt{n} = -1.83$, donc $n = 3.66^2 \approx 13$.

4.2 Tests multiples

Nous avons déjà évoqué ce point dans le Chapitre 2 : si l'on a 20 réalisations d'une loi normale centrée réduite, et que l'on calcule une p-value pour chacune de ces réalisations, on aura en moyenne une p-value < 0.05, bien que l'hypothèse H_0 soit vraie. Il ne faut pas pour autant rejeter l'hypothèse parce qu'une des p-values est en dessous du niveau de significativité.

On peut même calculer la probabilité p_n d'avoir une p-value significative lorsque H_0 est vraie, en fonction du nombre n de tests indépendants effectués : si $n = 1$, p_1 vaut évidemment 0.05. Dans le cas $n \geq 1$,

$$\begin{aligned} p_n &= P_{H_0}(\text{au moins une p-value significative}) \\ &= 1 - P_{H_0}(\text{aucune p-value significative}) \\ &= 1 - (1 - 0.05)^n \end{aligned}$$

Si l'on fait deux tests à 0.05, on a en réalité fait un test de niveau $p_2 \approx 0.10$; avec 10 tests, le niveau atteint $p_{10} \approx 0.40$, et avec 100 tests, on ne teste en fait plus le niveau 0.05 mais $p_{100} \approx 0.99$, c'est-à-dire qu'on est quasi-sûr de rejeter H_0 lorsqu'elle est vraie, ce qui est problématique.

Cela paraît évident lorsqu'on répète le même test successivement sur chaque échantillon, mais le problème peut être plus complexe : imaginons qu'on mesure les quantité d'ARNm de 20 gènes sur 50 cultures non stimulées, et 50 cultures stimulées avec du LPS. On cherche les gènes qui sont significativement différentiellement transcrits entre les cellules stimulées et non stimulées. On serait tenté de faire un t-test (ou Mann–Whitney) pour chacun des gènes, mais on trouverait alors beaucoup de faux positifs, c'est-à-dire des p-values < 0.05 qui apparaissent avec un risque plus grand que 5% ; on augmenterait fortement le risque de première espèce.

Pour remédier à ce problème, il existe deux solutions :

1. Soit **changer le niveau de significativité individuel** (limite en dessous de laquelle une p-value est considérée significative) pour garder un niveau de significativité global identique égal à 0.05. On parle de **correction de Bonferroni** (ou de FWER; family-wise error rate) : on divise alors le risque $\alpha = 0.05$ par n pour obtenir le nouveau risque α à utiliser : par exemple, on rejettera l'hypothèse nulle uniquement si $p < 0.05/n$;

2. Soit **corriger la p-value** (l'augmenter artificiellement), pour garder un niveau de significativité individuel de 0.05, au dessous duquel on rejette H_0. C'est le rôle des procédures FDR (false-discovery rate). Ces procédures corrigent les p-values en fonction de n, et donnent ce qu'on appelle des q-values (p-values corrigées). La signification exacte des q-values diffèrent selon les méthodes, mais cela dépasse le cadre de ce cours.

En conclusion, il faut limiter au maximum le nombre de tests que l'on fait; idéalement, pour chaque jeu de données (réalisations de variables aléatoires i.i.d.), on ne devrait faire qu'un test. Si l'on effectue trop de tests, il faut soit changer de test, soit adapter le risque α.

4.3 Significativité *versus* « grande » différence

La p-value peut être un indicateur trompeur : si elle quantifie quelque chose, c'est une sorte de confiance dans le rejet de l'hypothèse H_0, qui est une hypothèse qualitative, par exemple que deux populations ont une moyenne différente. Le test ne donne en revanche **aucune information quantitative sur cette différence**.

Exemple. Considérons une drogue anti-virale : sans cette drogue, la charge virale est de moyenne 10^6 PFU/mL et d'écart type 100 PFU/mL ; avec la drogue, la charge virale est de 9.10^5 et d'écart type 100 également. Comme il y a une différence de moyenne au niveau de la population, un grand nombre d'échantillons dont les mesures sont assez précises donnera une p-value très significative entre des cellules traitées et non traitées. Pour autant, la différence entre les moyennes est faible par rapport à l'écart type et à la moyenne elle-même : l'intérêt de cette drogue est donc limitée.

Une p-value significative obtenue doit être interprétée comme « il y a une différence significative entre cellules traitées et non traitées », mais pas « il y a une grande différence entre cellules traitées et non traitées ».

4.4 Non-indépendance des données et modèles mixtes

Exemple. Supposons que l'on veuille tester l'effet de l'âge sur la taille des lézards. On mesure la taille X_i de différents lézards à différents moments de leur vie chez deux populations de lézards qui s'ont suivies annuellement : une en France et l'autre en Allemagne. On a un total de 50 mesures pour chacune des populations, pour autant nous ne pouvons pas directement tester si il y a un effet de l'âge sur la taille des lézards (en utilisant par exemple une ANOVA si l'âge est une variable catégorielle ou bien une régression linéaire simple si l'âge est une variable continue). En effet, l'hypothèse nulle à tester est que l'âge n'a pas d'effet sur la taille et on considérerait que tous les individus étudiés proviennent d'une même population homogène, c'est à dire que chaque observation est indépendante. Or, sous H_0, dans notre expérience, les X_i ne sont pas des variables aléatoires identiquement distribuées, car les X_i proviennent soit de la population française, soit de la population allemande, et donc n'ont potentiellement pas les même espérance, variance et lien avec l'âge (à cause de variations environnementales ou génétiques). On peut le voir en disant que, sous H_0, deux individus français ne partagent pas le même « **degré d'indépendance** » qu'un individu français et un individu allemand. On mesure donc à la fois la différence d'espérance due à la variable « âge » et celle liée à la variable « population ».

Afin de corriger pour ces **problèmes de non indépendance** dans les données, il est possible de construire des **modèles mixtes** qui sont composés à la fois de **variables à effet fixe** (variable(s) dont on cherche à mesurer l'effet biologique) et de **variables à effet aléatoire** (variable(s) dont on cherche à corriger l'effet).

Dans notre expérience, on cherche à expliquer la variable « taille » en fonction de la variable « âge » à effet fixe tout en corrigeant pour la variable « population » à effet aléatoire. En pratique, on peut introduire de tels effets aléatoires en ajoutant des termes de covariances non nulles entre échantillons non-indépendants (ici issus de la même population). La description détaillée de ces modèles mixtes sera l'object de cours de statistiques plus avancés.

Index