Mathématiques de la Modélisation II
Ecologie et Statistiques
Année 2023/2024

1 Rappels de probabilités

1.1 Axiomes de probabilité

À une expérience aléatoire est associé un ensemble Ω qui est l'ensemble de tous les résultats possibles de cette expérience. Un sous-ensemble de Ω est appelé *événement*. On dit que $\mathbb{P}$ est une *probabilité* sur Ω si pour tous événements A et B de Ω, on a :

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$;
2. $\mathbb{P}(\Omega) = 1$;
3. $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$;
4. si $A \cap B = \emptyset$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

À partir de ces quatre axiomes, on peut montrer d'autres propriétés, en particulier :

5. si $A_i \cap A_j = \emptyset$ pour $i \neq j$ alors $\mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$;
6. $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathrm{c}})$;
7. si $A \subset B$ alors $\mathbb{P}(A^{\mathrm{c}} \cap B) = \mathbb{P}(B) - \mathbb{P}(A)$;
8. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

1.2 Variable aléatoire

On appelle *variable aléatoire* toute application de Ω dans E :

$$X \ : \ \Omega \to E \ .$$

L'ensemble E est souvent plus maniable que l'ensemble Ω qui peut être très compliqué.

Deux exemples :

1. On envoie une particule sur une cible circulaire de rayon R ; Ω est l'ensemble des trajectoires possibles d'une particule entre une source, et X est la distance du point d'impact au centre de la cible, $X : \Omega \longrightarrow [0, R]$.
   L'ensemble Ω n'est pas descriptible, par contre il y a plusieurs modélisations possibles pour X, par exemple :

$$\mathbb{P}(X \leqslant t) = \begin{cases} 0 & \text{si } t < 0, \\ \dfrac{\pi t^2}{\pi R^2} & \text{si } t \in [0, R], \\ 1 & \text{si } t > R; \end{cases} \quad \text{ou } \mathbb{P}(X \leqslant t) = \begin{cases} 0 & \text{si } t < 0, \\ \dfrac{1 - \mathrm{e}^{-t^2/2}}{1 - \mathrm{e}^{-R^2/2}} & \text{si } t \in [0, R], \\ 1 & \text{si } t > R. \end{cases}$$

2. Un oiseau (d'une race donnée dans un environnement donné) a au plus trois descendants arrivant à maturité par an ; Ω est l'ensemble des descendances possibles de l'oiseau au bout de deux ans, et X est le nombre de petit-descendants de l'oiseau arrivant à maturité au bout des deux ans, $X : \Omega \longrightarrow \{0, \ldots, R\}$.
   On peut calculer $\mathbb{P}(X = 0), \ldots, \mathbb{P}(X = 9)$.

Les ensembles $\{X \leqslant b\}$ ou $\{X \geqslant a\}$ ou $\{a \leqslant X \leqslant b\}$ sont des événements de Ω, les axiomes de probabilité permettent de les manipuler. Par exemple :

$$\mathbb{P}(|X| > a) = \mathbb{P}(X < -a \text{ ou } X > +a) = \mathbb{P}(\{X < -a\} \cup \{X > +a\}) = \mathbb{P}(X < -a) + \mathbb{P}(X > +a).$$

1.3 Loi de probabilité d'une variable aléatoire

1.3.1 Fonction de répartition

La loi de probabilité d'une variable aléatoire X peut toujours être décrite par sa *fonction de répartition*

$$\begin{array}{rll} F : & \mathbb{R} & \to [0,1] \\ & t & \to F(t) = \mathbb{P}(X \leqslant t) \end{array}$$

La fonction F est croissante, continue à droite, avec $\lim_{t \to -\infty} F(t) = 0$ et $\lim_{t \to +\infty} F(t) = 1$.

Exercice. Montrer que $\mathbb{P}(X > a) = 1 - F(a)$ et que $\mathbb{P}(a < X \leqslant b) = F(b) - F(a)$.

1.3.2 Variables aléatoires discrètes

Si l'ensemble E est fini ou dénombrable, on dit que X est une *variable aléatoire discrète*. Sa loi de probabilité peut alors être décrite par les probabilités aux points $\mathbb{P}(X = x)$, $x \in E$ (souvent plus simple que la fonction de répartition). Dans l'exemple 2, $E = \{0, 1, 2, \ldots, 9\}$ et la loi de X est donnée par $\mathbb{P}(X = 0), \mathbb{P}(X = 1), \ldots, \mathbb{P}(X = 9)$.
Si A est un sous-ensemble de E, on obtient $\mathbb{P}(X \in A)$ par la formule

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).$$

Dans l'exemple 2, la probabilité que l'oiseau ait au plus deux petit-descendants arrivant à maturité est $\mathbb{P}(X \leqslant 2) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2)$.

Exemples de lois discrètes :

— La *loi uniforme discrète* sur $E = \{1, \ldots, N\}$, notée $\mathcal{U}(\{1, \ldots, N\})$: $\mathbb{P}(X = k) = \dfrac{1}{N}$ pour tout $k \in \{1, \ldots, N\}$. C'est le cas où toutes les réalisations de X sont équiprobables.

— La *loi de Bernoulli* de paramètre $p \in [0, 1]$, notée $\mathcal{B}(p)$: $E = \{0, 1\}$, $\mathbb{P}(X = 0) = 1 - p$, $\mathbb{P}(X = 1) = p$. C'est la loi associée à une expérience aléatoire à deux issues *succès* et *échec*, par exemple $X =$ résultat du lancer d'une pièce de monnaie pour laquelle la probabilité d'obtenir face est p (la pièce n'est pas forcément équilibrée).

— La *loi binomiale* de paramètre $n \in \mathbb{N}^*$ et $p \in [0, 1]$, notée $B(n, p)$: $E = \{0, 1, 2, \ldots, n\}$ et pour tout $k \in E$,

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

C'est la loi du nombre total de succès lorsque l'on répète n fois de manière identique et indépendante une expérience aléatoire à deux issues *succès* et *échec*, par exemple $X =$ nombre de faces obtenu lors de n lancers d'une pièce de monnaie pour laquelle la probabilité d'obtenir face est p.

— La *loi géométrique* de paramètre $p \in [0,1]$, notée $\mathcal{G}(p)$: $E = \{1, 2, \ldots\}$ et pour tout $k \geq 1$,

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

C'est la loi du nombre d'épreuves nécessaires à l'obtention d'un succès, lorsque l'on répète de manière identique et indépendante une expérience aléatoire à deux issues *succès* et *échec*, par exemple X = numéro du lancer donnant face pour la première fois lors de lancers successifs d'une pièce pour laquelle la probabilité d'obtenir face est p.

Une variante de cette loi, notée $\mathcal{G}'(p)$, est $E = \{0, 1, 2, \ldots\} = \mathbb{N}$ et pour tout $k \geq 0$,

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

C'est la loi du nombre d'échecs précédant un succès lorsque l'on répète de manière identique et indépendante une expérience aléatoire à deux issues *succès* et *échec*, par exemple X = nombre de piles avant le premier face lors de lancers successifs d'une pièce pour laquelle la probabilité d'obtenir face est p.

— La *loi de Poisson* de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$: $E = \{0, 1, 2, \ldots\}$ et pour tout $k \geq 0$,

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

C'est la loi des événements rares, par exemple le nombre d'enfants dans une famille.

1.3.3 Variables aléatoires à densité

Si l'on peut mesurer X avec une précision infinie (par exemple si X est un instant, une distance, *etc.*), l'ensemble E n'est pas dénombrable. C'est le cas de l'exemple 1.

Définition. *S'il existe une fonction positive f telle que $\mathbb{P}(a \leqslant X \leqslant b) = \int_a^b f(u)du$ pour tous réels a, b, on dit que X admet la densité de probabilité f.*

Exemple. Dans le cas de l'exemple 1, $f(t) = \dfrac{2t}{R^2}$ si $t \in [0, R]$ et $f(t) = 0$ si $t \notin [0, R]$.

Proposition 1. *Si X admet la densité f, $F(t) = \int_{-\infty}^{t} f(u)du$ (la fonction de répartition est l'aire sous la courbe de la densité), et $f(t) = F'(t)$. La loi de X est donc décrite par la densité de probabilité f.*

Remarque. Une densité est une fonction f positive telle que $\int_{-\infty}^{+\infty} f(u)du = 1$.

Remarque. Si X admet la densité f, alors

$$F(b) - F(a) = \mathbb{P}\left(a < X \leqslant b\right) = \mathbb{P}\left(a \leqslant X \leqslant b\right) = \mathbb{P}\left(a \leqslant X < b\right) = \mathbb{P}\left(a < X < b\right) = \int_a^b f(u)du\ .$$

Exemples de lois à densité :

— La *loi uniforme* sur $[a, b]$, notée $\mathcal{U}([a, b])$: c'est l'analogue de l'équiprobabilité. La densité f est constante sur tout un intervalle. Par exemple, la loi uniforme sur l'intervalle $[a, b]$ a pour densité $f(t) = 1/(b-a)$ si $t \in [a, b]$ et une densité nulle en dehors de cet intervalle, si $t \notin [a, b]$. Le coefficient $1/(b-a)$ assure que $\int_a^b du/(b-a) = 1$.

— La *loi exponentielle* de paramètre $\lambda > 0$, notée $\mathcal{E}(\lambda)$: la densité f d'une loi exponentielle de paramètre λ, est $f(t) = \lambda \mathrm{e}^{-\lambda t}$ lorsque $t > 0$ (et zéro lorsque $t \leq 0$).

— Les *lois gaussiennes (ou lois normales)* présentent une importance particulière pour la suite de ce cours. Nous leur consacrons le paragraphe suivant.

1.3.4 Variables aléatoires suivant la loi normale, ou loi gaussienne

La loi normale centrée réduite. On appelle loi normale centrée réduite, ou loi gaussienne standard, la loi de probabilité dont la densité est donnée par :

$$f(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \text{ pour tout } t \in \mathbb{R}.$$

Le terme $1/(\sqrt{2\pi})$ assure que l'intégrale de la densité sur $]-\infty, \infty[$ vaut 1.

Lorsque X suit une loi normale centrée réduite (c'est-à-dire lorsque X admet la densité de probabilité f), on note $X \sim \mathcal{N}(0,1)$. Pour obtenir les valeurs de la fonction de répartition de cette loi, on utilise une table de valeurs numériques. Toutefois, la table de valeurs numériques ne donne pas les valeurs de $F(t)$ lorsque t est négatif. Il faut alors utiliser la parité de f pour établir la formule

$$F(-t) = 1 - F(t) \text{ pour tout } t \text{ positif.}$$

Exemple. On modélise une erreur de mesure par une variable aléatoire X que l'on suppose de loi normale centrée réduite. Quelle est la probabilité que l'erreur soit comprise entre -1 et 1 ?
Réponse : $\mathbb{P}(-1 \leqslant X \leqslant 1)$, qui s'écrit comme l'intégrale entre -1 et 1 de f :

$$\mathbb{P}(-1 \leqslant X \leqslant 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} dt.$$

Cette intégrale ne peut se calculer que numériquement car la fonction f n'a pas de primitive explicite. Commençons par exprimer $\mathbb{P}(-1 \leqslant X \leqslant 1)$ à partir de la fonction de répartition F : $\mathbb{P}(-1 \leqslant X \leqslant 1) = F(1) - F(-1)$. Dans la table de valeurs numériques, on lit $F(1) = 0.8413$ mais on ne lit pas $F(-1)$. Utilisons alors les propriétés de symétrie de la densité

$$F(-1) = \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} dt = \int_{1}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} dt$$

(on peut soit démontrer cette formule en effectuant le changement de variable $u = -t$ dans la première intégrale, soit simplement s'en convaincre en comparant les aires sous la courbe de la densité de $-\infty$ à -1 et de 1 à $+\infty$). On a obtenu pour le moment : $F(-1) = \int_1^{+\infty} f(t)dt = \mathbb{P}(X \geqslant 1)$. Utilisons alors la propriété de complémentarité : $\mathbb{P}(X \geqslant 1) = 1 - \mathbb{P}(X < 1)$ (les inégalités larges ou strictes n'ont pas d'importance pour les variables à densité) : finalement nous trouvons : $F(-1) = 1 - \mathbb{P}(X \leqslant 1) = 1 - F(1)$. La réponse à la question posée est donc $\mathbb{P}(-1 \leqslant X \leqslant 1) = F(1) - F(-1) = F(1) - (1 - F(1)) = 2F(1) - 1 = 2 \times 0.8413 - 1 = 0.6826$.

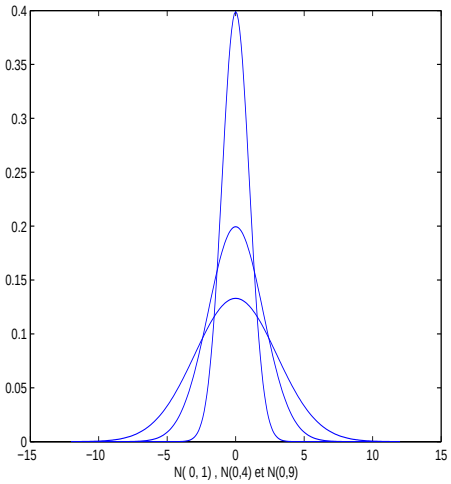
En conclusion, la fonction de répartition de la loi normale centrée réduite est tabulée mais il ne faut pas se contenter de savoir lire la table de valeurs numériques dans des cas particuliers (la lecture de $F(1)$ est directe, il en est de même pour la lecture des valeurs $F(a)$ avec $a \geqslant 0$), il faut également savoir manipuler les probabilités et utiliser les propriétés de symétrie de la densité f pour se ramener dans tous les cas à une somme, ou une différence, de termes du type $F(a)$ avec $a \geqslant 0$.

Exercice. Montrer que si X suit la loi normale centrée réduite, $\mathbb{P}(|X| \leqslant t) = 2F(t) - 1$ et $\mathbb{P}(|X| \geqslant t) = 2(1 - F(t))$ pour tout t positif.

Les autres lois normales, ou lois gaussiennes. À partir de la loi normale centrée réduite, on peut construire les autres lois normales. Soit X une variable de loi normale centrée réduite. La variable Y définie par $Y = \sigma X$ possède la densité de probabilité f :

$$f(t) = \frac{1}{\sqrt{2\pi}|\sigma|} \mathrm{e}^{-\frac{t^2}{2\sigma^2}} \text{ pour tout } t \in \mathbb{R}.$$

On appelle cette loi de probabilité la loi normale (ou loi gaussienne) de paramètres 0 et σ^2, et on note $Y \sim \mathcal{N}(0, \sigma^2)$. Le paramètre σ augmente (s'il augmente) ou diminue (s'il diminue) l'étalement de la courbe de densité, comme l'illustre la figure ci-après (densité des lois $\mathcal{N}(0,1)$, $\mathcal{N}(0,4)$, $\mathcal{N}(0,9)$).

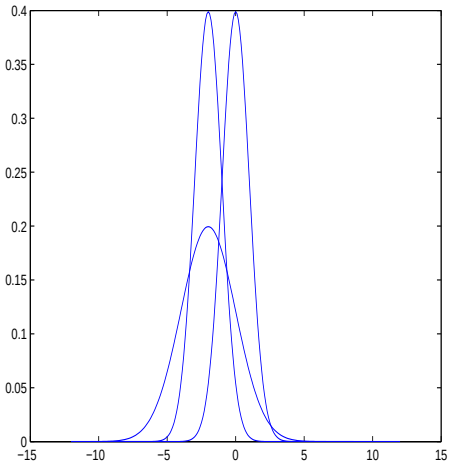


Nous décentrons maintenant la loi : la densité devient symétrique autour d'une valeur μ et non plus autour de 0. Soit X une variable de loi normale centrée réduite. La variable Z définie par $Z = \mu + \sigma X$, possède la densité de probabilité f :

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \text{ pour tout } t \in \mathbb{R}.$$

On appelle cette loi de probabilité la loi normale ou loi gaussienne de paramètres μ et σ^2, et on note $Z \sim \mathcal{N}(\mu, \sigma^2)$. Le paramètre μ est un paramètre qui indique la *position* de la courbe densité alors que le paramètre σ^2 est un paramètre qui indique l'étalement, ou la *dispersion* de la courbe densité.

Exercice. Sur la figure suivante, identifier les densités respectives des lois $\mathcal{N}(0,1)$, $\mathcal{N}(-1,1)$, $\mathcal{N}(-1,4)$:



Passage de la loi $\mathcal{N}(\mu, \sigma^2)$ à la loi $\mathcal{N}(0,1)$. Seule la loi normale centrée réduite est tabulée. Comment alors trouver les valeurs de la fonction de répartition pour les autres lois ? Par exemple, si une mesure est modélisée par une variable Z de loi $\mathcal{N}(1.5\ ;\ 4)$, quelle est la probabilité que cette mesure dépasse 3.5 ? La réponse est bien sûr $\mathbb{P}(Z > 3.5)$, mais l'on ne dispose pas de tables de valeurs numériques de la loi $\mathcal{N}(1.5\ ;\ 4)$ qui permettraient de lire directement cette valeur.

On a vu que si X suit la loi $\mathcal{N}(0,1)$, alors $\mu + \sigma X$ suit la loi $\mathcal{N}(\mu, \sigma^2)$.
De même, lorsque Z suit la loi $\mathcal{N}(\mu, \sigma^2)$ alors

$$X = \frac{Z - \mu}{\sigma} \text{ suit la loi } \mathcal{N}(0,1).$$

Lorque l'on transforme une variable Z de loi $\mathcal{N}(\mu, \sigma^2)$ en une variable X de loi $\mathcal{N}(0,1)$, on dit que l'on centre (passage de Z à $Z - \mu$) et que l'on réduit (passage de $Z - \mu$ à $X = \frac{Z-\mu}{\sigma}$) la variable Z. Revenons à notre exemple :

$$\mathbb{P}(Z > 3.5) = \mathbb{P}(Z - 1.5 > 3.5 - 1.5) = \mathbb{P}\left(\frac{Z - 1.5}{2} > \frac{3.5 - 1.5}{2}\right) = \mathbb{P}(X > 1) = 1 - F(1)$$

où X suit la loi $\mathcal{N}(0,1)$ et où F est la fonction de répartition de la loi normale centrée réduite. Finalement, la réponse est : $\mathbb{P}(Z > 3.5) = 1 - 0.8413 = 0.1587$.

1.4 Indépendance

Définition. *Deux événements A et B sont indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.*
Deux variables aléatoires discrètes X et Y sont indépendantes si $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \times \mathbb{P}(Y = y)$ pour toutes valeurs x et y prises par ces variables.
Les variables aléatoires discrètes $X_1, \ldots, X_n$ sont indépendantes si

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \times \cdots \times \mathbb{P}(X_n = x_n)$$

pour toutes valeurs $x_1, \ldots, x_n$ prises par ces variables.
Deux variables aléatoires à densité X et Y sont indépendantes si $f_{(X,Y)}(x,y) = f_X(x) f_Y(y)$ pour toutes valeurs x et y prises par ces variables (la densité du couple est égale au produit des densités).
Les variables aléatoires à densité $X_1, \ldots, X_n$ sont indépendantes si

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_{X_1}(x_1) \times \cdots f_{X_n}(x_n)$$

pour toutes valeurs $x_1, \ldots, x_n$ prises par ces variables.

La définition mathématique n'est pas forcément parlante, alors qu'elle recouvre une notion assez intuitive, du moins en ce qui concerne l'indépendance d'événements et l'indépendance de variables aléatoires discrètes : deux événements sont indépendants si la connaissance de la réalisation de l'un ne donne pas d'information sur la probabilité de réalisation de l'autre. Autrement dit, A et B sont indépendants si la probabilité que A se réalise sachant que B est réalisé est égale à la probabilité que A se réalise (la réalisation de B n'apporte aucune connaissance supplémentaire sur la réalisation de A), et ceci s'écrit mathématiquement : $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$. En remplaçant dans cette formule $\mathbb{P}(A/B)$ par sa définition : $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ on retrouve : A et B sont indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. Il en est de même pour les variables aléatoires discrètes : deux variables aléatoires discrètes X et Y sont indépendantes si la connaissance de la valeur x prise par X ne donne pas d'information sur la probabilité pour Y de prendre la valeur y, et cela quelles que soient les valeurs x et y : $\mathbb{P}(Y = y \,|\, X = x) = \mathbb{P}(Y = y)$ pour tous x et y. En remplaçant dans cette formule $\mathbb{P}(X = x \,|\, Y = y)$ par sa définition $\mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$ on retrouve également : X et Y sont indépendantes si $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \times \mathbb{P}(Y = y)$ pour toutes valeurs x et y prises par ces variables.

Exemple. Le pelage d'un animal peut être de type A,B ou C. On observe 1000 individus distincts, on obtient le tableau suivant :

	A	B	C
mâle	160	170	158
femelle	120	262	130

Pensez-vous que les caractères sexe et pelage sont indépendants ? On peut évidemment donner une réponse intuitive, qui ne peut être que subjective, la réponse statistique se fait par un test du chi-deux d'indépendance (section 7.3).

1.5 Espérance, variance

1.5.1 Espérance (ou moyenne théorique)

Définition. *L'espérance d'une v.a. discrète X est*

$$\mathbb{E}(X) = \sum_{x \text{ valeur possible pour } X} x\,\mathbb{P}(X = x)$$

si cette quantité est bien définie.
L'espérance d'une v.a. X admettant la densité f est

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

si cette quantité est bien définie.

L'espérance de X peut s'interpréter comme la valeur moyenne de X. Dans le cas discret, la formule $\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x)$ est une moyenne pondérée par le poids de chaque valeur.

Exemple. On note X le résultat d'un lancer d'un dé à 6 faces, X prend les valeurs $1, 2, 3, 4, 5, 6$. Si le dé est non truqué, $\mathbb{P}(X = x) = 1/6$ pour toutes les valeurs de x, et l'on obtient :

$$\mathbb{E}(X) = \sum_{x=1}^{6} x\,\mathbb{P}(X = x) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3,5 :$$

la valeur *moyenne* d'un lancer est 3,5. Par contre si le dé est truqué et ne tombe que sur les faces 4,5,6 avec $\mathbb{P}(X = x) = 1/3$ pour $x = 4, 5, 6$ et $\mathbb{P}(X = x) = 0$ pour $x = 1, 2, 3$, on obtient :

$$\mathbb{E}(X) = \sum_{x=1}^{6} x\,\mathbb{P}(X = x) = \frac{1}{3}(4 + 5 + 6) = 5 :$$

la valeur *moyenne* d'un lancer est 5.

L'expression valeur *moyenne* appliquée à un unique lancer peut paraître surprenante. Il s'agit d'une moyenne théorique. Par contre, si le dé est lancé 1000 fois, la notion de valeur moyenne de ces mille lancers est assez naturelle (c'est la moyenne arithmétique des 1000 lancers) et l'on s'attend, de manière intuitive, à ce que cette moyenne soit proche de 3,5 si le dé n'est pas truqué, et de 5 si le dé est truqué. Cette coïncidence entre l'espérance (ou moyenne théorique) d'un lancer unique et la moyenne arithmétique d'un grand nombre de lancers (ou moyenne empirique) n'en est justement pas une : il s'agit d'une loi très importante en probabilités dénommée la loi des grands nombres (*cf.* section 2.2 pour la loi faible des grands nombres). La loi forte des grands nombres, non énoncée dans ce polycopié, permet de raccorder le point de vue des mathématiciens avec le point de vue de l'utilisateur : cette loi montre que l'espérance mathématique est presque toujours égale à la moyenne arithmétique (ou moyenne empirique) sur une infinité de lancers. L'expression *presque toujours* nécessite une formalisation mathématique qui n'a pas sa place dans ce cours, aussi seule la loi faible (qui énonce le même phénomène, mais en probabilité) est donnée en section 2.2 (théorème 8).

Proposition 2. *Soit Φ une fonction de $\mathbb{R}$ dans $\mathbb{R}$. Si X est une v.a. discrète,*

$$\mathbb{E}(\Phi(X)) = \sum_{x \text{ valeur possible pour } X} \Phi(x)\mathbb{P}(X = x)$$

lorsque cette quantité est bien définie.
Si X est une v.a. admettant la densité f,

$$\mathbb{E}(\Phi(X)) = \int_{-\infty}^{+\infty} \Phi(x) f(x) dx$$

lorsque cette quantité est bien définie.

Cette proposition indique que la connaissance de la loi de X permet de calculer l'espérance (ou valeur moyenne, ou moyenne théorique) de $\Phi(X)$. En particulier on pourra calculer $\mathbb{E}\left(X^2\right)$.

Exemple. Dans le cas du lancer du dé vu ci-dessus,

$$\mathbb{E}\left(X^2\right) = \sum_{x=1}^{6} x^2\,\mathbb{P}\left(X=x\right) = \frac{1}{6}(1+4+9+16+25+36) = \frac{91}{6} \simeq 15.17$$

si le dé n'est pas truqué, et

$$\mathbb{E}\left(X^2\right) = \sum_{x=1}^{6} x^2\,\mathbb{P}\left(X=x\right) = \frac{1}{3}(16+25+36) = \frac{77}{3} \simeq 25.67$$

si le dé est truqué.

Théorème 3 (Propriétés de l'espérance)**.**

- *On a* $\mathbb{E}\left(aX+b\right) = a\mathbb{E}\left(X\right)+b$ *pour tous réels* a *et* b*.*
- *On a également* $\mathbb{E}\left(X+Y\right) = \mathbb{E}\left(X\right)+\mathbb{E}\left(Y\right)$*.*
- *Si* $X \geqslant 0$*, alors* $\mathbb{E}\left(X\right) \geqslant 0$
- *Si* X *et* Y *sont indépendantes alors* $\mathbb{E}\left(XY\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$*; l'espérance du produit est égale au produit des espérances. Attention : la réciproque est fausse.*

La dernière propriété se généralise comme suit : si $X_1, X_2, \ldots, X_n$ sont indépendantes, alors

$$\mathbb{E}\left(X_1 X_2 \cdots X_n\right) = \mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right)\cdots\mathbb{E}\left(X_n\right);$$

autrement dit, l'espérance de leur produit est égale au produit de leurs espérances. La réciproque est fausse.

1.5.2 Variance

Définition. *La variance d'une v.a.* X *est* $\mathbb{V}\left(X\right) = \mathbb{E}\left(\left(X-\mathbb{E}\left(X\right)\right)^2\right)$*.*
L'écart-type d'une v.a. X *est* $\sigma(X) = \left(\mathbb{V}\left(X\right)\right)^{1/2}$*.*

L'écart-type est une mesure de l'écart théorique moyen entre X et $\mathbb{E}\left(X\right)$. Une autre mesure de cet écart aurait pu être $\mathbb{E}\left(\left|X-\mathbb{E}\left(X\right)\right|\right)$, mais la valeur absolue complique les calculs. Par souci de simplicité, $\sigma(X) = \sqrt{\mathbb{E}\left(\left(X-\mathbb{E}\left(X\right)\right)^2\right)}$ a été choisi comme indicateur de l'écart théorique entre X et $\mathbb{E}\left(X\right)$. Cet indicateur mesure *la dispersion* théorique moyenne de X autour de $\mathbb{E}\left(X\right)$.

Exemple. Pour mesurer la concentration m d'une solution qu'il manipule, un chimiste utilise un appareil qui possède une précision de 2 mmol.L^{-1}. Le chimiste effectue 3 mesures sur des prélèvements de la solution. On note X_1, X_2, X_3 les variables aléatoires représentant les résultats des mesures. Les variables X_1, X_2, X_3 sont supposées indépendantes (on suppose que les mesures ne s'influencent pas, ce qui revient en réalité à supposer que les erreurs de mesure d'une mesure à l'autre n'ont aucun lien entre elles), et de même loi (les prélévements sont identiques, l'appareil également, le chimiste réalise ses trois mesures de la même manière).

On suppose que la moyenne théorique d'une mesure est m, où m est la véritable concentration de la solution, ce qui s'écrit mathématiquement $\mathbb{E}\left(X_i\right) = m$ pour $i = 1, 2, 3$ (les variables ont même loi, donc même espérance). Si ce n'était pas le cas, il faudrait soit changer l'appareil de mesure, soit changer de chimiste.

On suppose que la dispersion moyenne de la mesure autour de la véritable concentration m est 2 mmol.L^{-1} : seule l'imprécision de la machine joue. Ceci s'écrit mathématiquement $\sigma(X_i) = 2$ pour

$i = 1, 2, 3$ ou encore $\mathbb{V}(X_i) = 4$ pour $i = 1, 2, 3$ (les variables ont même loi, donc même variance). En réalité, dans l'écart-type d'une mesure il faut intégrer tous les facteurs de variabilité, la précision de la machine en est un, on peut aussi penser que les prélévements peuvent être légérement différents, *etc.* La notion de variance est liée à la notion d'aléatoire : lorsque la variance d'une variable X est nulle, la variable X ne produit qu'un seul résultat (alors fatalement égal à $\mathbb{E}(X)$). Dans l'exemple envisagé, si la variance était nulle, toutes les mesures seraient identiques.

Il reste alors à proposer une loi de probabilité pour ces mesures : une loi à densité (on pourrait descendre à une précision infinie), symétrique autour de la véritable concentration m (si ce n'était pas le cas, l'appareil ou le chimiste ou les prélévements seraient à modifier). Classiquement, on commence par envisager pour ce type d'expérience une loi de probabilité gaussienne.
En conclusion, une modélisation possible pour cette expérience est : les variables X_1, X_2, X_3 sont indépendantes et suivent toutes trois la même loi normale $\mathcal{N}(m, 4)$.

Proposition 4. $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2$.

Exemple. Utilisons la formule de la proposition précédente pour calculer l'écart-type de X, variable aléatoire représentant le résultat d'un lancer d'un dé à 6 faces. On a vu en section 1.5.1 que :

- $\mathbb{E}(X) = 7/2$ et $\mathbb{E}\left(X^2\right) = 91/6$ si le dé n'est pas truqué, ce qui donne $\mathbb{V}(X) = (91/6) - (49/4) \simeq 2.9167$ et $\sigma(X) \simeq 1.7078$.
- $\mathbb{E}(X) = 5$ et $\mathbb{E}\left(X^2\right) = 77/3$ si le dé est truqué, ce qui donne $\mathbb{V}(X) = (77/3) - 25 \simeq 0.6667$ et $\sigma(X) \simeq 0.8165$.

Théorème 5 (Propriétés de la variance)**.**

- *On a $\mathbb{V}(X) \geqslant 0$, et si $\mathbb{V}(X) = 0$ alors X prend une seule valeur avec probabilité 1 (autrement dit, X n'est pas aléatoire).*
- *Pour tous réels a et b, $\mathbb{V}(aX + b) = a^2\mathbb{V}(X)$.*
- *Si X et Y sont indépendantes alors $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$: la variance de la somme est égale à la somme des variances. La réciproque est fausse.*

La dernière propriété se généralise comme suit : si $X_1, X_2, \ldots, X_n$ sont indépendantes alors

$$\mathbb{V}(X_1 + X_2 + \cdots + X_n) = \mathbb{V}(X_1) + \mathbb{V}(X_2) + \cdots + \mathbb{V}(X_n)\,;$$

autrement dit, la variance de la somme est alors égale à la somme des variances. La réciproque est fausse.

Exercice. Dans le cas général, montrer que $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,Cov(X, Y)$ où $Cov(X, Y)$ est définie par $Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$. On appelle cette quantité la covariance de X et de Y.

1.5.3 Espérance et variance des lois de probabilité usuelles

- $X \sim \mathcal{B}(p) : \mathbb{E}(X) = p$, $\mathbb{V}(X) = p(1 - p)$.
- $X \sim \mathcal{B}(n, p) : \mathbb{E}(X) = np$, $\mathbb{V}(X) = np(1 - p)$.
- $X \sim \mathcal{G}(p) : \mathbb{E}(X) = \dfrac{1}{p}$, $\mathbb{V}(X) = \dfrac{1 - p}{p^2}$.
- $X \sim \mathcal{G}'(p) : \mathbb{E}(X) = \dfrac{1 - p}{p}$, $\mathbb{V}(X) = \dfrac{1 - p}{p^2}$.
- $X \sim \mathcal{P}(\lambda) : \mathbb{E}(X) = \lambda$, $\mathbb{V}(X) = \lambda$.
- $X \sim \mathcal{U}(\{1, \ldots, N\}) : \mathbb{E}(X) = \dfrac{N + 1}{2}$, $\mathbb{V}(X) = \dfrac{(N - 1)(N + 1)}{12}$.

— $X \sim \mathcal{U}([a,b]) : \mathbb{E}(X) = \dfrac{a+b}{2}, \mathbb{V}(X) = \dfrac{(b-a)^2}{12}$.

— $X \sim \mathcal{E}(\lambda) : \mathbb{E}(X) = \dfrac{1}{\lambda}, \mathbb{V}(X) = \dfrac{1}{\lambda^2}$.

— $X \sim \mathcal{N}(m, \sigma^2) : \mathbb{E}(X) = m, \mathbb{V}(X) = \sigma^2$. Ainsi, les paramètres de la loi normale sont respectivement son espérance et sa variance.

2 Le n-échantillon et la moyenne empirique

2.1 Le n-échantillon

Un certain nombre d'études sont basées sur la répétition d'expériences identiques et indépendantes. Dans l'exemple de la section 1.5.2, un chimiste utilise un appareil pour mesurer la concentration d'une solution qu'il manipule, supposons qu'il réalise 10 mesures sur des prélévements de cette solution. On a vu qu'une modélisation possible est la suivante : les variables $X_1, X_2, \ldots, X_{10}$ sont indépendantes et suivent toutes la même loi normale $\mathcal{N}(m, 4)$. On dit que $X_1, X_2, \ldots, X_{10}$ forment un 10-*échantillon* de loi $\mathcal{N}(m, 4)$, où m représente la véritable concentration de la solution. Pour le second tour des élections présidentielles, un institut de sondage interroge 1000 personnes choisies au hasard dans l'ensemble des électeurs français, chaque individu est prié de choisir sa réponse : A, B ou NSP. L'institut modélise ce sondage par une suite $Y_1, Y_2, \ldots, Y_{1000}$ de variables indépendantes et de même loi discrète dont les valeurs possibles sont A, B ou NSP : $\mathbb{P}(Y_i = A) = p_A$, $\mathbb{P}(Y_i = B) = p_B$, $\mathbb{P}(Y_i = NSP) = p_{NSP}$ avec $p_A + p_B + p_{NSP} = 1$, et $i = 1, 2, \ldots, 1000$ (toutes les variables ont même loi). On appelle $\mathcal{L}$ la loi commune de ces variables. On dit que $Y_1, Y_2, \ldots, Y_{1000}$ forme un 1000-*échantillon* de loi $\mathcal{L}$. Un n-*échantillon* d'une loi $\mathcal{L}$ est donc tout simplement la donnée de n variables aléatoires indépendantes et de même loi $\mathcal{L}$, ce que nous répétons dans la définition suivante.

Définition. *On dit que $X_1, \ldots, X_n$ est un n-échantillon d'une loi $\mathcal{L}$ lorsque $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes et de même loi $\mathcal{L}$. Par la suite, on appelle m leur espérance commune et σ^2 leur variance commune.*

Proposition 6. *Soit $X_1, \ldots, X_n$ un n-échantillon de moyenne m et de variance σ^2.*

1. $\mathbb{E}(X_1 + \cdots + X_n) = n \times m$ *(vrai même si les v.a. ne sont pas indépendantes),*
2. $\mathbb{V}(X_1 + \cdots + X_n) = n \times \sigma^2$ *(faux en général si les v.a. ne sont pas indépendantes).*

Par la suite, on note Σ_n la variable aléatoire $X_1 + \cdots + X_n$.

Remarque. Attention à ne pas confondre un n-échantillon $X_1, \ldots, X_n$ et une réalisation de $X_1, \ldots, X_n$, notée $x_1, \ldots, x_n$.

2.2 Moyenne empirique d'un n-échantillon

Souvent, le statisticien dispose d'un n-échantillon d'une loi $\mathcal{L}$, dont un ou plusieurs paramètres sont inconnus et suscitent l'intérêt du statisticien. Par exemple, le chimiste du paragraphe 2.1, devant son échantillon de 10 mesures de loi $\mathcal{N}(m, 4)$, souhaiterait l'exploiter pour extraire des informations, voire une connaissance assez précise du paramètre m, puisque ce paramètre représente la véritable concentration de la solution. Dans l'exemple du sondage pour le second tour des élections présidentielles, introduit également en section 2.1, l'institut de sondage dispose d'un 1000-échantillon et est évidemment très intéressé par les paramètres de la loi de probabilité de ce 1000-échantillon, à savoir les valeurs de p_A, p_B, p_{NSP}. Ces paramètres d'intérêt (m pour le chimiste ; p_A, p_B, p_{NSP} pour l'institut de sondage) ont un rapport avec les paramètres de la loi $\mathcal{L}$. Plus précisément, dans l'exemple du chimiste, $\mathbb{E}(X_i) = m$ (moyenne théorique). Lorsque le paramètre d'intérêt est l'espérance de la loi $\mathcal{L}$, la variable aléatoire $\overline{X}_n$, définie par

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

est l'outil adéquat pour étudier m. Cette variable est appelée *moyenne empirique* de l'échantillon. L'exemple du sondage est un peu plus compliqué : il faut commencer par coder les réponses A, B, NSP par $-1, +1, 0$. L'espérance de la loi $\mathcal{L}$ ainsi codée est alors $\mathbb{E}(Y_i) = (-1 \times p_A) + (1 \times p_B) + (0 \times p_{NSP}) = p_B - p_A$ et la moyenne empirique

$$\bar{Y}_{1000} = \frac{Y_1 + \cdots + Y_{1000}}{1000}$$

est l'outil adéquat pour étudier $p_B - p_A$ (cette différence est bien ce qui intéresse l'institut de sondage). Nous allons maintenant préciser pourquoi la moyenne empirique est un outil adéquat pour étudier l'espérance de la loi $\mathcal{L}$.

Définition. *Soit $X_1, \ldots, X_n$ un n-échantillon de moyenne m et de variance σ^2. La variable aléatoire*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} = \frac{\Sigma_n}{n}$$

est appelée moyenne empirique.

Proposition 7 (Propriétés de la moyenne empirique)**.** $\mathbb{E}\left(\overline{X}_n\right) = m$, $\mathbb{V}\left(\overline{X}_n\right) = \dfrac{\sigma^2}{n}$.

La proposition suivante établit le lien entre la moyenne empirique $\overline{X}_n$ et la moyenne théorique m.

Théorème 8 (Loi faible des grands nombres)**.** *Soit $X_1, \ldots, X_n$ un n-échantillon de moyenne m et de variance σ^2. Pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(|\overline{X}_n - m| > \epsilon\right) \leqslant \frac{\sigma^2}{n\epsilon^2}.$$

Lorsque n devient grand, $\overline{X}_n$ a tendance à être de plus en plus près de m : $\mathbb{P}\left(|\overline{X}_n - m| > \epsilon\right) \longrightarrow_{n \to +\infty} 0$. On dit que $\overline{X}_n$ est un *estimateur* de m, et on dit que la valeur observée de $\overline{X}_n$, notée $\bar{x}_n$, est une *estimation* de m. Il ne faut pas confondre la variable aléatoire (l'estimateur) $\overline{X}_n$ et sa valeur observée (l'estimation). Si l'on réalise une seconde fois le n-échantillon, la valeur observée $\bar{x}_n$ change, alors que l'estimateur est toujours le même.

2.3 Le n-échantillon gaussien

Les lois gaussiennes possèdent les propriétés remarquables suivantes :

Théorème 9 (Propriétés des échantillons gaussiens)**.**

1. *Si $X_1, \ldots, X_n$ sont n variables gaussiennes indépendantes alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est encore une loi gaussienne.*
2. *Si $X_1, \ldots, X_n$ sont n variables gaussiennes indépendantes de lois $\mathcal{N}(m_1, \sigma_1^2), \ldots, \mathcal{N}(m_n, \sigma_n^2)$ alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est la loi gaussienne $\mathcal{N}(m_1 + \cdots + m_n, \sigma_1^2 + \cdots + \sigma_n^2)$.*
3. *Si $X_1, \ldots, X_n$ est un n-échantillon gaussien de loi $\mathcal{N}(m, \sigma^2)$ alors :*
   - *(a) la loi de $\Sigma_n = X_1 + \cdots + X_n$ est la loi gaussienne $\mathcal{N}(nm, n\sigma^2)$;*
   - *(b) la loi de $\overline{X}_n$ est la loi gaussienne $\mathcal{N}(m, \dfrac{\sigma^2}{n})$;*
   - *(c) la loi de U_n, où U_n est la v.a. définie par $U_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$, est la loi $\mathcal{N}(0,1)$. Notons que U_n peut aussi s'écrire $U_n = \dfrac{\Sigma_n - nm}{\sqrt{n}\sigma}$.*

2.4 Le n-échantillon poissonien

Les lois de Poisson possèdent des propriétés remarquables analogues :

Théorème 10 (Propriétés des échantillons poissoniens). *1. Si $X_1, \ldots, X_n$ sont n variables de Poisson indépendantes alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est encore une loi de Poisson.*

2. Si $X_1, \ldots, X_n$ sont n variables de Poisson indépendantes de lois respectives $\mathcal{P}(\lambda_1)$, $\ldots$, $\mathcal{P}(\lambda_n)$ alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est la loi de Poisson $\mathcal{P}(\lambda_1 + \cdots + \lambda_n)$.

3. Si $X_1, \ldots, X_n$ est un n-échantillon de Poisson de loi $\mathcal{P}(\lambda)$ alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est la loi de Poisson $\mathcal{P}(n\lambda)$.

2.5 Le n-échantillon de Bernoulli : Approximations de la loi binomiale

Proposition 11. *Si $X_1, \ldots, X_n$ sont n variables de Bernoulli indépendantes et de même loi $\mathcal{B}(p)$ alors la loi de $\Sigma_n = X_1 + \cdots + X_n$ est la loi binomiale $\mathcal{B}(n, p)$.*

Lorsque n est grand, les coefficients $\binom{n}{k}$ sont difficiles à manipuler. On utilise alors des approximations de la loi binomiale, qui sont valables dans des régimes différents selon que p soit petit (de l'ordre de $\frac{1}{n}$: approximation poissonienne), ou non (p fixé, pas trop petit : approximation gaussienne).

Proposition 12 (Approximation de la loi binomiale par une loi de Poisson). *Lorsque le nombre n est grand, que le réel p est petit et que $\lambda = np$ n'est pas trop grand, (en pratique, n* **supérieur ou égal à** *50, p* **inférieur ou égal à** *0.1* **et** *np* **strictement inférieur à** *15), alors la loi de Σ_n est proche d'une loi de Poisson de paramètre $\lambda = np$.*

Notons que l'espérance de la loi de Poisson $\mathcal{P}(\lambda)$ vaut $\lambda = np$ et est égale à l'espérance de la loi de Σ_n. La variance de la loi de Poisson $\mathcal{P}(\lambda)$ vaut aussi $\lambda = np$, et est presque *égale à la variance $np(1-p)$ de la loi de Σ_n puisque p est petit.*

Remarque. Une autre condition, peut-être plus naturelle, pour approximer la loi de Σ_n par la loi de Poisson de paramètre $\lambda = np$ est : n **supérieur ou égal à 50 et** $\mathbb{E}(\Sigma_n) \simeq \mathbb{V}(\Sigma_n)$.

Prenons un exemple : $p = 0.001$ et $n = 1000$. Que vaut $\mathbb{P}(\Sigma_{1000} = 3)$? Comme il n'y a pas d'ambiguïté, on note Σ pour Σ_{1000}. On approxime la loi de Σ par la loi de Poisson de paramètre $1000 \times 0.001 = 1$.
Alors : $\mathbb{P}(\Sigma = 3) \simeq \mathrm{e}^{-1}1^3/3! = 0.06131$. La valeur exacte est $\mathbb{P}(\Sigma = 3) = 0/06129$.
On remarque que $\mathbb{E}(\Sigma) = 1$ et $\mathbb{V}(\Sigma) = 0.999$: ces deux quantités sont presque égales.

Proposition 13 (Approximation de la loi binomiale par une loi gaussienne). *Lorsque n est grand (en pratique, n* **supérieur ou égal à** *30), et lorsque np et $n(1-p)$ ne sont pas trop petits (en pratique :* **tous deux supérieurs ou égaux à** *15), alors la loi de Σ_n est proche de la loi gaussienne $\mathcal{N}(np, np(1-p))$. Ce résultat s'énonce également sur la variable recentrée et renormalisée : lorsque n est grand (en pratique, n* **supérieur ou égal à** *30), et lorsque np et $n(1-p)$ ne sont pas trop petits (en pratique :* **tous deux supérieurs ou égaux à** *15), alors la loi de*

$$U_n = \frac{\Sigma_n - np}{\sqrt{np(1-p)}}$$

est proche d'une loi Gaussienne $\mathcal{N}(0, 1)$.
Avec la notation $\overline{X}_n$, ceci s'écrit également : la loi de

$$U_n = \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}}$$

est proche d'une loi Gaussienne $\mathcal{N}(0, 1)$.

Notons que l'espérance de $(\Sigma_n - np)/\sqrt{np(1-p)}$ est nulle tout comme celle de la loi $\mathcal{N}(0,1)$. De plus, la variance de Σ_n est égale à $np(1-p)$ (section 1.5.3). La variance de $(\Sigma_n - np)/\sqrt{np(1-p)}$ est donc égale à 1, tout comme la variance de la loi $\mathcal{N}(0,1)$.
En pratique, dire que la loi de $(\Sigma_n - np)/\sqrt{np(1-p)}$ est proche d'une loi $\mathcal{N}(0,1)$, revient à dire que pour tout $a < b$ fixés, la probabilité

$$\mathbb{P}\left(\frac{\Sigma_n - nm}{\sqrt{np(1-p)}} \in [a,b]\right)$$

peut être approximée par

$$\mathbb{P}\left(Z \in [a,b]\right),$$

où Z suit la loi $\mathcal{N}(0,1)$. Rappelons que

$$\mathbb{P}\left(Z \in [a,b]\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2} dx,$$

et que la fonction $\mathrm{e}^{-x^2/2}$ n'a pas de primitive connue. On a donc recours à la table de valeurs numériques de la loi $\mathcal{N}(0,1)$ pour lire $\mathbb{P}\left(Z \in [a,b]\right)$.

Par exemple, supposons que $p = 0.5$ et que $n = 10000$. Ceci se produit par exemple lorsque l'on joue 10000 fois à pile ou face. $\Sigma_{10000} = X_1 + \cdots + X_n$ représente alors le nombre de fois où est apparu face. Comme il n'y a pas d'ambiguïté, on note Σ pour Σ_{10000}. Que vaut alors $\mathbb{P}\left(\Sigma \in [4900, 5100]\right)$?
Pour déterminer cela, on note que $\Sigma \in [4900, 5100]$ est équivalent à $(\Sigma - 5000)/50 \in [-2, 2]$:

$$\mathbb{P}\left(\Sigma \in [4900, 5100]\right) = \mathbb{P}\left(\frac{\Sigma - 5000}{50} \in [-2, 2]\right).$$

Mais la proposition 13 nous assure que la loi de $\frac{\Sigma - 5000}{50}$ est proche de la loi gaussienne $\mathcal{N}(0,1)$. Alors :

$$\mathbb{P}\left(\Sigma \in [4900, 5100]\right) \simeq \mathbb{P}\left(Z \in [-2, 2]\right),$$

où Z suit la loi $\mathcal{N}(0,1)$. On lit dans la table cette probabilité :

$$\begin{aligned} \mathbb{P}\left(Z \in [-2,2]\right) &= F(2) - F(-2) \text{ où } F \text{ est la fonction de répartition de la loi } \mathcal{N}(0,1) \\ &= F(2) - (1 - F(2)) \\ &= 2F(2) - 1 = 0.9545. \end{aligned}$$

On obtient donc $\mathbb{P}(\Sigma \in [4900, 5100]) \simeq 0.9545$. Le calcul exact donne $\mathbb{P}(\Sigma \in [4900, 5100]) \simeq 0.95557$. On peut encore améliorer cette approximation avec une *correction de continuité* :

$$\begin{aligned} \mathbb{P}(\Sigma \in [4900; 5100]) = \mathbb{P}(4899.5 \leqslant \Sigma \leqslant 5100.5) &= \mathbb{P}\left(\frac{4899.5 - 5000}{50} \leqslant \Sigma \leqslant \frac{4899.5 - 5000}{50}\right) \\ &\simeq \mathbb{P}(-2.01 \leqslant Z \leqslant 2.01) = 2\,F(2.01) - 1 \simeq 0.95557. \end{aligned}$$

Le résultat par approximation ne diffère plus du résultat exact !

Remarque. Quand on peut les employer, les approximations par des lois de Poisson ou gaussiennes donnent des informations bien plus précises que l'inégalité de la loi des grands nombres (théorème 8).

2.6 Cas général : le théorème central limite

En fait, l'approximation par une variable aléatoire gaussienne de U_n est valable pour n'importe quelle loi :

Théorème 14 (Théorème central limite). *Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi $\mathcal{L}$ de moyenne m et de variance σ^2. On pose $\Sigma_n = X_1 + \cdots + X_n$. Alors, lorsque n tend vers l'infini, la loi de*

$$U_n = \frac{\Sigma_n - nm}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$$

converge vers la loi $\mathcal{N}(0,1)$.

Ce résultat est remarquable et précis. D'une part, il permet de travailler sur tout échantillon *de grande taille* comme sur un échantillon gaussien, dès lors que l'on travaille sur la moyenne empirique $\overline{X}_n$. D'autre part, il donne une indication sur l'écart entre $\overline{X}_n = \dfrac{\Sigma_n}{n}$ et m bien plus précise que celle obtenue par la loi des grands nombres (théorème 8).

Par exemple, si l'on étudie un 100-échantillon dont on sait à priori que $\sigma^2 = 1$, et que l'approximation TLC est licite, alors :

$$\mathbb{P}\left(-1.96 \leqslant \sqrt{100}(\overline{X}_{100} - m) \leqslant 1.96\right) \simeq \mathbb{P}\left(Z \in [-1.96, +1.96]\right)$$

où Z suit la loi $\mathcal{N}(0,1)$. On lit sur la table de valeurs numériques $\mathbb{P}\left(Z \in [-1.96, +1.96]\right) = 95\%$. On en déduit :

$$\mathbb{P}\left(\overline{X}_{100} - \frac{1.96}{10} \leqslant m \leqslant \overline{X}_{100} + \frac{1.96}{10}\right) = \mathbb{P}\left(-1.96 \leqslant \sqrt{100}(\overline{X}_{100} - m) \leqslant 1.96\right) \simeq 95\%.$$

En particulier, si m est inconnue, on voit qu'avec une probabilité de l'ordre de 95%, m est dans l'intervalle (aléatoire car $\overline{X}_{100}$ est une v.a.) : $[\overline{X}_{100} \pm 0.196]$. À partir d'une observation $\overline{x}_{100}$ de $\overline{X}_{100}$, on a un encadrement statistique de la moyenne théorique m.

Ce théorème confirme également que la moyenne empirique $\overline{X}_n$ d'un n-échantillon de moyenne m et de variance σ^2 est un bon outil (on dit un bon estimateur) pour évaluer m lorsque σ^2 est connue. Par la loi des grands nombres, on savait que $\overline{X}_n$ était proche de m pour n grand, le TLC précise la distance entre $\overline{X}_n$ et m, toujours pour n grand, et permet d'obtenir un encadrement de m (on dit un intervalle de confiance). Si σ^2 est inconnue, ce qui est généralement le cas, on utilise une version un peu plus compliquée du TLC que nous verrons ultérieurement.

La principale limitation de ce théorème réside dans la condition *n grand*. Dans le cas de la loi binomiale, cette condition est évaluée à np et $n(1-p)$ supérieurs ou égaux à 15 ; lorque l'on ne connaît rien sur la loi de l'échantillon, la situation est plus délicate. Plus la loi de l'échantillon est dissymétrique, comme par exemple la loi exponentielle, plus la convergence vers la loi $\mathcal{N}(0,1)$ est lente, et plus l'utilisation du TLC nécessite un grand nombre d'observations. Au contraire, si la loi de l'échantillon est la loi $\mathcal{U}[-1,+1]$, le TLC est licite pour de petites valeurs de n. Pour avoir une idée de la loi de l'échantillon sur lequel on travaille, on peut par exemple tracer un histogramme.

3 Tests d'hypothèses et intervalles de confiance

3.1 Principe des tests

Jusqu'à présent, nous avons supposé que nous connaissions la loi d'une ou plusieurs variables aléatoires, et nous avons montré comment, en utilisant cette information, on pouvait calculer des probabilités d'événements. En pratique, on observe souvent des résultats de variables aléatoires dont on ne connait pas la loi. On cherche en fait justement, à partir des observations des résultats, à obtenir des informations sur la loi. Compte tenu du fait que ces observations sont le résultat d'expériences aléatoires, les conclusions que l'on peut en tirer ne sont jamais des certitudes.
Commençons par un exemple simple. Un fabricant annonce que parmi les nombreux produits qu'il vend, moins de 10% sont défectueux. On en choisit par exemple 10 au hasard et on constate qu'ils ont tous un

défaut. On a envie d'en conclure que le fabricant a menti. Evidement, il se peut que le fabricant n'ait pas menti et qu'on ait eu une malchance terrible de tomber dix fois de suite sur un produit défectueux. Cependant, on sent bien que cela est fort peu probable, et on peut déclarer avec relative certitude qu'il a menti.
Maintenant, supposons que parmi les 10 produits testés, 4 s'avèrent défectueux. On est à ce moment moins confiant lorsque l'on affirme que le fabricant a menti. Le principe des tests statistiques que nous allons décrire dans la suite du cours est d'utiliser les calculs de probabilités que nous avons présentés jusqu'à maintenant pour pouvoir déterminer exactement à partir de quel nombre de produits défectueux on peut affirmer (avec une confiance de 95% disons) que le producteur a menti.

Le même type de raisonnenent s'applique par exemple lorsque l'on se pose les questions suivantes :

1. Quelle est la proportion de graines qui germeront dans telle production de graines ?
2. Dans les confitures de fraise de la marque XXX, quel est le taux de sucre ?

On sème des graines et on observe si elles germent, ou on ouvre quelques pots de confiture, on analyse le contenu et on pèse la quantité de sucre. On souhaite répondre aux questions à l'aide des expériences.

Étudions plus en détail la question 2. On peut se demander : le taux de sucre est-il de 50% ? Pour cela, on fait n analyses, qui donnent pour résultats $x_1, \ldots, x_n$. Bien entendu, tous ces résultats sont différents, d'une part à cause des erreurs de mesure, d'autre part du fait que la fabrication de confiture contient une part de variabilité. On va alors supposer que les résultats des expériences sont les valeurs observées de variables gaussiennes indépendantes $X_1, \ldots, X_n$ de loi $\mathcal{N}(m, \sigma^2)$. Choisir la forme gaussienne de la loi, c'est choisir un modèle. Le paramètre m est celui sur lequel la question se pose. Le paramètre σ indique la précision que l'on pense avoir sur la mesure. On peut alors reformuler la question : "est-ce que m vaut 50 ?"

Pour cela, on va chercher une variable $Z(X_1, \ldots, X_n)$ donnant une bonne approximation de m ; ici, la moyenne empirique $Z = \overline{X}$ convient. On répondra selon la valeur observée de $Z = \overline{X}$: si z_{obs} est proche de 50, on dira "oui, le taux m est 50". Si la valeur observée est éloignée de 50, on dira "non, le taux m n'est pas 50". Autrement dit, on choisira un domaine I de valeurs proches de 50 et on répondra "m vaut 50" si $Z \in I$, et "m ne vaut pas 50" si $Z \notin I$. Evidemment, Z peut prendre toutes les valeurs possibles, donc la réponse que l'on donne n'est pas certaine : on appelle *niveau du test* la probabilité de répondre "m ne vaut pas 50 " alors que la réponse vraie est "m vaut 50" : c'est *l'erreur de première espèce.*
Il y a une autre manière de se tromper : si on répond "m vaut 50" alors que la réponse vraie est "m ne vaut pas 50". C'est ce que l'on appelle *l'erreur de deuxième espèce.*

Le niveau du test est souvent noté α. *C'est le statisticien qui fixe le niveau du test.* Le choix de α permet de déterminer, par des techniques de calcul de probabilités, l'intervalle I. On calcule l'intervalle I de telle sorte que si la vraie réponse est "m vaut 50", alors la probabilité pour que Z ne tombe pas dans I est α. La façon dont on choisit la forme du domaine I dépend aussi de la façon dont on craint de se tromper : si l'on craint que le taux m soit plus grand que 50, on décidera de répondre "non" seulement si Z est plus grand qu'une valeur seuil fixée.

Résumons encore le principe général : on observe des résultats de n variables aléatoires $X_1, \ldots, X_n$ dont on ne connait pas la loi. On cherche, à partir des observations $x_1, \ldots, x_n$ des résultats, à obtenir des informations sur la loi. Compte tenu du fait que ces observations sont le résultat d'expériences aléatoires, les conclusions que l'on peut en tirer ne sont jamais des certitudes.

On souhaite, à partir de l'observation $x_1, \ldots, x_n$, décider si une hypothèse, notée $\mathcal{H}_0$, est plausible et compatible avec cette observation. On choisit une variable *d'intérêt* Z fonction de $X_1, \ldots, X_n$ et on traduit l'hypothèse $\mathcal{H}_0$ par une hypothèse sur la loi de Z. L'hypothèse $\mathcal{H}_0$ est vraie ssi la loi de Z est $\mathcal{L}_0$.

On exprime aussi l'hypothèse alternative à $\mathcal{H}_0$, notée $\mathcal{H}_1$, qui peut être une hypothèse précise sur la loi de Z (par exemple *la loi $\mathcal{L}$ de Z est précisément la loi $\mathcal{L}_1$*) ou une hypothèse un peu floue (*la loi $\mathcal{L}$ de Z n'est pas la loi $\mathcal{L}_0$*).

On décide que l'hypothèse $\mathcal{H}_0$ est vraie si l'on pense que la loi de Z est $\mathcal{L}_0$, et l'on décide que l'hypothèse $\mathcal{H}_0$ est fausse si l'on pense que la loi de Z n'est pas $\mathcal{L}_0$. Pour cela, l'idée est de déterminer un intervalle I de valeurs tel que, si la loi de Z est $\mathcal{L}_0$, $Z \in I$ avec une grande probabilité :

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_0}(Z \in I) &= 1-\alpha \text{ (souvent } 95\%) \\ \mathbb{P}_{\mathcal{H}_0}(Z \notin I) &= \alpha \text{ (souvent } 5\%) \end{aligned}$$

On peut alors déterminer I en utilisant les techniques de calcul de probabilités développées jusqu'à maintenant. Si la valeur observée z_{obs} tombe dans cet ensemble de valeurs plausibles pour $\mathcal{L}_0$ (autrement dit, si $z_{obs} \in I$), alors on ne peut pas invalider l'hypothèse selon laquelle la loi de Z est $\mathcal{L}_0$. Si par contre la valeur observée z_{obs} ne tombe pas dans cet ensemble de valeurs plausibles pour $\mathcal{L}_0$ (autrement dit, si $z_{obs} \notin I$), cela signifie que l'observation est tombée dans l'ensemble des valeurs très peu plausibles pour $\mathcal{L}_0$, et on peut invalider l'hypothèse selon laquelle la loi de Z est $\mathcal{L}_0$. Remarquons que dans le premier cas ($z_{obs} \in I$) on garde $\mathcal{H}_0$ en quelque sorte par défaut, alors que dans le second cas ($z_{obs} \notin I$) on décide réellement de rejeter $\mathcal{H}_0$. Pour cette raison, on choisit souvent comme hypothèse $\mathcal{H}_1$ l'hypothèse que l'on veut confirmer, et pas celle que l'on consent à garder par défaut. Lorsqu'un test conduit à garder l'hypothèse $\mathcal{H}_1$, on dit qu'il est significatif.

En résumé, pour faire un test, il faut :

- Une hypothèse prioritaire notée $\mathcal{H}_0$ sur la loi de Z : *La loi $\mathcal{L}$ de Z est exactement la loi $\mathcal{L}_0$* où $\mathcal{L}_0$ est une loi explicitement décrite.
- Une hypothèse de remplacement $\mathcal{H}_1$, qui sera la conclusion si on décide que l'hypothèse $\mathcal{H}_0$ est fausse.
- Le niveau α du test, qui est un petit nombre (souvent $1\%, 2.5\%, 5\%$).

Vocabulaire et notations : (voir aussi le paragraphe 3.5)

- L'erreur de première espèce (ou niveau) d'un test, notée α, est la probabilité de rejeter l'hypothèse $\mathcal{H}_0$ alors que l'hypothèse $\mathcal{H}_0$ est vraie : $\alpha = \mathbb{P}_{\mathcal{H}_0}(Z \notin I)$. Elle est fixée par le statisticien.
- L'erreur de seconde espèce d'un test, notée β, est la probabilité de conserver l'hypothèse $\mathcal{H}_0$ alors que l'hypothèse $\mathcal{H}_0$ est fausse : $\beta = \mathbb{P}_{\mathcal{H}_1}(Z \in I)$. Elle n'est pas contrôlée par le statisticien.
- L'événement qui nous conduit à rejeter $\mathcal{H}_0$ est souvent appelé zone de rejet de $\mathcal{H}_0$ et notée $\mathcal{R}$. Si la statistique de test est nommée Z, et que la règle est : on conserve $\mathcal{H}_0$ si $Z \in I$, alors que l'on rejette $\mathcal{H}_0$ si $Z \notin I$, la région de rejet de $\mathcal{H}_0$ est $\mathcal{R} = \{Z \notin I\}$.
- L'équation de niveau est l'équation qui permet de déterminer la région de rejet de $\mathcal{H}_0$, cette équation s'écrit $\alpha = \mathbb{P}_{\mathcal{H}_0}(\mathcal{R})$. Si la statistique de test est nommée Z, et que la règle est : on conserve $\mathcal{H}_0$ si $Z \in I$, alors $\alpha = \mathbb{P}_{\mathcal{H}_0}(Z \notin I)$.

Maintenant, le choix à priori d'un niveau est relativement arbitraire. Suposons que dans notre exemple, ayant fixé le niveau à 5%, on ait obtenu pour I l'intervalle $[49; 51]$. On répond par "oui, le taux m est 50" ou "non, le taux m n'est pas 50" après avoir comparé la valeur observée de z_{obs} avec les valeurs seuils 49 et 51, on n'indique pas si la valeur observée est proche ou loin de ces valeurs seuils. Or cette information a un sens : si la valeur observée est au centre de I, (par exemple $z_{obs} = 50.1$), c'est que le résultat est hautement probable sous $\mathcal{H}_0$; si la valeur observée est en dehors de I la signification n'est pas la même selon que $z_{obs} = 51.1$ ou que $z_{obs} = 80$, bien que dans les deux cas l'on rejette $\mathcal{H}_0$. On préfère alors souvent répondre à l'aide du niveau de signification (ou p-valeur).

Définition. *On appelle degré de signification ou probabilité critique ou p-valeur d'un test la valeur notée p_{value} ou α_{obs} l'erreur commise si on rejette $\mathcal{H}_0$ avec les observations dont on dispose.*

Cette définition entraîne que si α est le niveau du test,

- Pour $\alpha < \alpha_{obs}$, le test de niveau α accepte $\mathcal{H}_0$,

— Pour $\alpha > \alpha_{obs}$, le test de niveau α rejette $\mathcal{H}_0$.

En conclusion, voici le protocole de construction d'un test statistique utilisé dans ce cours :

Schéma de la construction d'un test de niveau α (en 7 points) :

1. Déterminer dans quel cadre on se situe, c'est-à-dire préciser le modèle.
2. Au vu de la question posée, déterminer quelles seront les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test. Attention : il faut choisir pour $\mathcal{H}_0$ une hypothèse sous laquelle on sait faire des calculs.
3. Déterminer une statistique (une variable aléatoire fonction de l'échantillon observé) dont on connaît la loi sous $\mathcal{H}_0$, et dont la loi sous $\mathcal{H}_0$ est différente de la loi sous $\mathcal{H}_1$.
4. À partir de la loi de la statistique sous $\mathcal{H}_0$, on établit une règle qui permet de décider quand on rejette et quand on accepte $\mathcal{H}_0$. On en déduit la forme de la région de rejet de $\mathcal{H}_0$.
5. On calcule le ou les seuil(s) de la région de rejet du test de niveau α grâce à la loi de la statistique sous $\mathcal{H}_0$ et à l'équation de niveau.
6. On calcule la valeur observée de la statistique et on donne la conclusion du test avec les deux méthodes suivantes :

   Première méthode : On compare la valeur observée de la statistique au(x) seuil(s) de la région de rejet. Si cette valeur observée tombe dans la région de rejet, on rejette $\mathcal{H}_0$ et on décide $\mathcal{H}_1$; si cette valeur observée ne tombe pas dans la région de rejet, on conserve $\mathcal{H}_0$.

   Seconde méthode : On détermine le seuil de signification, ou probabilité critique, ou p-valeur (p-value en anglais), c'est à dire l'erreur commise si on rejette $\mathcal{H}_0$ avec les observations dont on dispose. Si $\alpha < \alpha_{obs}$, on conserve $\mathcal{H}_0$; si $\alpha > \alpha_{obs}$, on rejette $\mathcal{H}_0$.
7. On conclut finalement en restituant en français les conclusions du test (on rejette $\mathcal{H}_0$ on on conserve cette hypothèse), et en donnant leur interprétation concrète (biologique lorsque le problème est issu de la biologie). Éventuellement, on commente les p-values si l'on a rejeté l'hypothèse $\mathcal{H}_0$: par exemple, si la p-value est très petite, alors ceci indique que l'hypothèse $\mathcal{H}_0$ est extrêmement improbable au vu des observations.

Dans la suite de ce polycopié, pour rendre les choses plus faciles à lire, le septième point de conclusion sera parfois intégré au sixième.

3.2 Cas des variables gaussiennes de variance connue

L'outil utilisé dans ce cadre est le théorème de la section 2.3 qui précise la loi de $\overline{X}_n$ et de $U_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$:

— la loi de $\overline{X}_n$ est la loi gaussienne $\mathcal{N}(m, \dfrac{\sigma^2}{n})$;

— la loi de U_n, où U_n est la v.a. définie par $U_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$ est la loi $\mathcal{N}(0,1)$.

Reprenons l'exemple des pots de confiture : on dispose de 100 variables aléatoires $X_1, \ldots, X_{100}$ représentant les taux de sucre.

1. Déterminer dans quel cadre on se situe, c'est-à-dire préciser le modèle.

   Les v.a. X_i sont indépendantes et de même loi $\mathcal{N}(m,1)$ (on suppose que la variance est connue et vaut 1). Le paramètre m est le taux de sucre moyen réel.
   On pose la question : *Le taux de sucre moyen m est-il égal à 50 ?*
2. Au vu de la question posée, déterminer quelles seront les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test. Attention : il faut choisir pour $\mathcal{H}_0$ l'hypothèse sous laquelle on sait faire des calculs.

   $\mathcal{H}_0 : m = 50$ et $\mathcal{H}_1 : m \neq 50$.

3. Déterminer une statistique (une variable aléatoire fonction de l'échantillons observé) dont on connaît la loi sous $\mathcal{H}_0$ et dont la loi sous $\mathcal{H}_0$ est différente de la loi sous $\mathcal{H}_1$.

   Pour faire un test sur m, on choisit la variable $\overline{X}_{100}$ (moyenne empirique) que l'on note $\overline{X}$ (il n'y a pas d'ambiguïté possible, la taille de l'échantillon étant fixée) :

   $$\overline{X} = \frac{X_1 + \cdots + X_{100}}{100}.$$

   Sa loi est connue sous $\mathcal{H}_0$: $\overline{X} \sim \mathcal{N}(50.1/100)$ et sa loi sous $\mathcal{H}_1$ est différente de sa loi sous $\mathcal{H}_0$: sous $\mathcal{H}_1$, $\overline{X} \sim \mathcal{N}(m \neq 50.1/100)$.

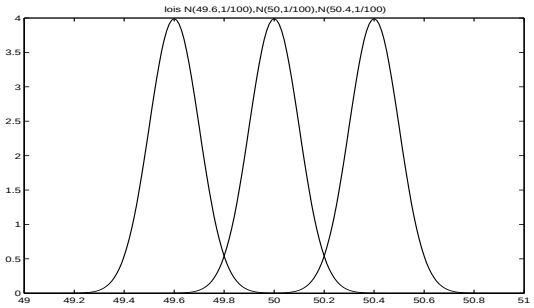
   Remarque importante. Il existe une alternative au choix de $\overline{X}$ comme statistique de test, moins intuitive mais rendant les calculs plus faciles. On peut choisir comme variable de test

   $$U_0 = \frac{\overline{X} - 50}{\sqrt{1/100}}$$

   c'est à dire la version centrée et réduite de $\overline{X}$ sous $\mathcal{H}_0$. La loi de U_0 est connue sous $\mathcal{H}_0$: $U_0 \sim \mathcal{N}(0,1)$ et sa loi sous $\mathcal{H}_1$ est différente de sa loi sous $\mathcal{H}_0$: sous $\mathcal{H}_1$, $U_0 \sim \mathcal{N}(\mu \neq 0,1)$. Choisir $\overline{X}$ ou U_0 comme statistique de test revient strictement au même : la décision et la p-valeur sont les mêmes dans les deux cas.

4. À partir de la loi de la statistique sous $\mathcal{H}_0$ et sous $\mathcal{H}_1$, on établit une règle qui permet de décider quand on rejette et quand on accepte $\mathcal{H}_0$. On en déduit la forme de la région de rejet de $\mathcal{H}_0$.

   On remarque que sous $\mathcal{H}_1$, la loi de $\overline{X}$ est à gauche (si $m < 50$) ou à droite (si $m > 50$) de la loi de $\overline{X}$ sous $\mathcal{H}_0$.



   On rejettera donc $\mathcal{H}_0$ si l'on observe une valeur $\overline{X}^{obs}$ trop à gauche ou trop à droite de 50. Autrement dit, I est de la forme $I = [50 - a\ ;\ 50 + a]$.

   Si on a choisi U_0 comme statistique de test, on rejettera $\mathcal{H}_0$ si l'on observe une valeur U_0^{obs} trop à gauche ou trop à droite de 0. Autrement dit, dans ce cas I est de la forme $I = [-b\ ;\ b]$.

5. On calcule le(s) seuil(s) de la région de rejet grâce à la loi de la statistique sous $\mathcal{H}_0$.

   On calcule le seuil a grâce à l'équation de niveau :

   $$\mathbb{P}_{\mathcal{H}_0}(\overline{X} \notin [50 - a\ ;\ 50 + a]) = \alpha$$

   ou, ce qui revient au même, grâce à l'équation :

   $$\mathbb{P}_{\mathcal{H}_0}(\overline{X} \in [50 - a\ ;\ 50 + a]) = 1 - \alpha.$$

   En prenant $\alpha = 5\%$ et en effectuant quelques petits calculs, on trouve $a = 0.196$.

Détails des calculs :

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_0}(50-a \leqslant \overline{X} \leqslant 50+a) &= 95\% \\ \Leftrightarrow \mathbb{P}_{\mathcal{H}_0}(\sqrt{100}(50-a-50) \leqslant \sqrt{100}(\overline{X}-50) \leqslant \sqrt{100}(50+a-50)) &= 95\% \\ \Leftrightarrow \mathbb{P}_{\mathcal{H}_0}(10\,a \leqslant \sqrt{100}(\overline{X}-50) \leqslant 10\,a) &= 95\% \\ \Leftrightarrow F(10\,a) - F(-10\,a) &= 95\% \end{aligned}$$

où F est la fonction de répartition de la loi normale centrée réduite. Les propriétés de symétrie de cette loi donnent :

$$F(10\,a) - F(-10\,a) = 2F(10\,a) - 1$$

et l'équation $F(10\ a) - F(-10\ a) = 95\%$ équivaut donc à l'équation $2F(10\ a) - 1 = 95\%$, et finalement à l'équation :

$$F(10\,a) = 97.5\% = 0.975$$

que l'on résout en lisant sur la table de la loi normale centrée réduite : $10\,a = 1.96$, soit $a = 0.196$.

À la fin de cette étape, nous connaissons la région de rejet de $\mathcal{H}_0$ du test de niveau 5% :

$$\mathcal{R} = \{\overline{X} \notin [50 - 0.196\ ;\ 50 + 0.196]\}.$$

Si l'on a choisi U_0 comme statistique de test : La loi de U_0 sous $\mathcal{H}_0$ étant la loi normale centrée réduite, l'équation de niveau

$$\mathbb{P}_{\mathcal{H}_0}(U_0 \notin [-b\ ;\ b]) = \alpha$$

donne immédiatement avec $\alpha = 5\%$:

$$2(1 - F(b)) = 5\%$$

soit $b = 1.96$. En ayant choisi de mener le test avec la statistique U_0, la région de rejet de $\mathcal{H}_0$ du test de niveau 5% s'écrit sous la forme :

$$\mathcal{R} = \{U_0 \notin [-1.96\ ;\ 1.96]\}.$$

6-7. On calcule la valeur observée de la statistique et on conclut.

Première méthode. On compare la valeur observée de la statistique au(x) seuil(s) de la région de rejet. Si cette valeur observée tombe dans la région de rejet, on rejette $\mathcal{H}_0$ et on décide $\mathcal{H}_1$; si cette valeur observée ne tombe pas dans la région de rejet, on conserve $\mathcal{H}_0$.
La règle de décision est donc :

— si $\overline{X}^{obs} \in [49.804\ ;\ 50.196]$, on décide de conserver $\mathcal{H}_0$;

— si $\overline{X}^{obs} \notin [49.804\ ;\ 50.196]$, on décide de rejeter $\mathcal{H}_0$.

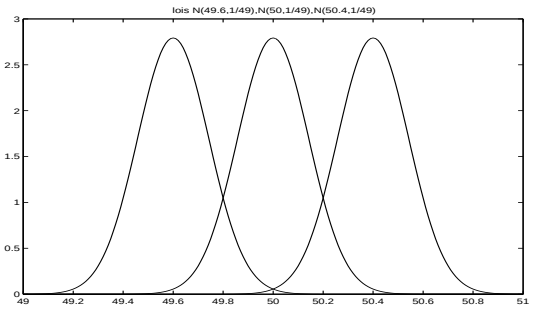
Par exemple, $\overline{X}^{obs} = 50.25$ conduit à rejeter $\mathcal{H}_0$ et à décider que le taux moyen de sucre est supérieur à 50% tandis $\overline{X}^{obs} = 50.1$ conduit à conserver $\mathcal{H}_0$ et à conclure que le taux moyen de sucre est égal à 50%.

Si l'on a choisi U_0 comme statistique de test :

— si $U_0^{obs} \in [-1.96\ ;\ 1.96]$, on décide de conserver $\mathcal{H}_0$,

— si $U_0^{obs} \notin [-1.96\ ;\ 1.96]$, on décide de rejeter $\mathcal{H}_0$.

Par exemple, $\overline{X}^{obs} = 50.25$ conduit à $U_0^{obs} = 2.5$ donc à rejeter $\mathcal{H}_0$ et à décider que le taux moyen de sucre est supérieur à 50% tandis que $\overline{X}^{obs} = 50.1$ conduit à $U_0^{obs} = 1$ donc à à conserver $\mathcal{H}_0$ et à conclure que le taux moyen de sucre est égal à 50%.



Remarque. La règle de décision est d'autant meilleure que la taille de l'échantillon est grande. Si nous ne disposons que d'un échantillon de 49 observations, la loi de $\overline{X}$ est $\mathcal{N}(m, 1/49)$:
Il est plus difficile de reconnaître $\mathcal{H}_1$: en faisant le test avec la statistique $\overline{X}$ on trouve $I = [49.72\ ;\ 50.28]$, intervalle plus grand que celui obtenu pour l'échantillon de taille 100. L'observation $\overline{X}^{obs} = 50.25$ calculée à partir de 49 observations ne permet pas de rejeter $\mathcal{H}_0$, alors que la même observation $\overline{X}^{obs} = 50.25$ calculée à partir de 100 observations permet de rejeter $\mathcal{H}_0$.
Seconde méthode. On détermine le seuil de signification, ou probabilité critique, ou p-valeur (p-value en anglais), c'est à dire l'erreur commise si on rejette $\mathcal{H}_0$ avec les observations dont on dispose. Si $\alpha < \alpha_{obs}$, on conserve $\mathcal{H}_0$; si $\alpha > \alpha_{obs}$, on rejette $\mathcal{H}_0$.

Calculons le niveau de signification du test réalisé sur un échantillon de taille 49, lorsque la moyenne empirique observée vaut $\overline{X}^{obs} = 50.25$. Pour rejeter $\mathcal{H}_0$ avec cette observation, l'intervalle I doit être égal à $[50 - 0.25\ ; 50 + 0.25]$ (c'est le plus petit intervalle permettant de conserver $\mathcal{H}_0$ avec l'observation $\overline{X}^{obs} = 50.25$, le niveau qui lui correspond est bien le niveau de signification α_{obs} : tout niveau $\alpha < \alpha_{obs}$ donnera un plus grand intervalle et conduira à conserver $\mathcal{H}_0$, et tout niveau $\alpha > \alpha_{obs}$ donnera un plus petit intervalle et conduira à rejeter $\mathcal{H}_0$). Effectuons le calcul :

$$\begin{aligned}
\alpha_{obs} &= \mathbb{P}_{\mathcal{H}_0}(Z \notin [50 - 0.25\ ; 50 + 0.25]) = \mathbb{P}_{\mathcal{H}_0}(\overline{X} \notin [50 - 0.25\ ; 50 + 0.25]) \\
&= \mathbb{P}_{\mathcal{H}_0}(7(\overline{X} - 50) \notin [-1.75\ ; +1.75]) = \mathbb{P}(U \notin [-1.75\ ; +1.75]) \text{ où } U \sim \mathcal{N}(0,1) \\
&= \mathbb{P}(U < -1.75) + \mathbb{P}(U > +1.75) = 2\mathbb{P}(U > +1.75) = 2(1 - F(1.75))
\end{aligned}$$

où F est la fonction de répartition de la loi normale centrée réduite. On lit sur la table de cette loi $\alpha_{obs} = 8.02\%$. Cela signifie que pour rejeter $\mathcal{H}$, avec un échantillon de taille 49, et une observation égale à $\overline{X}^{obs} = 50.25$, il faut un niveau supérieur à 8.02%. Avec cette observation $\overline{X}^{obs} = 50.25$, au niveau 5%, on conserve $\mathcal{H}_0$; au niveau 10%, on rejette $\mathcal{H}_0$.

Si l'on a choisi U_0 comme statistique de test : sur un échantillon de taille 49, la statistique U_0 n'est pas la même que sur un échantillon de taille 100. Sur un échantillon de taille 49, U_0 est définie par

$$U_0 = \frac{\overline{X} - 50}{\sqrt{1/49}}$$

c'est à dire la version centrée et réduite de $\overline{X}$ sous $\mathcal{H}_0$. La valeur $\overline{X}^{obs} = 50.25$ donne $U_0^{obs} = 1.75$. L'intervalle I n'a pas changé (c'est un autre avantage lorsque l'on utilise U_0 au lieu de $\overline{X}$ comme statistique de test) et on retrouve

$$\alpha_{obs} = \mathbb{P}_{\mathcal{H}_0}(U_0 \notin [-1.75\ ; +1.75] = 8.02\%.$$

3.3 Cas des sommes de variables de Bernoulli

Reprenons le premier exemple. On se demande si la probabilité, pour une graine, de germer est de 80% ; on sème n graines et on regarde si elle germent. On va choisir le modèle suivant : on dira $x_i = 1$ si

la $i-$ème graine a germé, et $x_i = 0$ sinon. Autrement dit, on suppose que les valeurs $x_1, \ldots, x_n$ sont le résultat de l'observation de variables indépendantes $X_1, \ldots, X_n$ de loi de Bernoulli $\mathcal{B}(p)$, et la question que l'on se pose porte sur la valeur de p (p est la probabilité qu'une graine germe). La statistique pour effectuer le test est $Z = X_1 + \cdots + X_n$, qui représente le nombre de graines germant parmi les n graines semées. La loi de Z est la loi $\mathcal{B}(n,p)$ dont on connaît deux approximations (par une loi de Poisson ou par une loi normale : propositions 12 et 13). Pour réaliser un test sur une somme de Bernoulli, il faudra donc soit utiliser directement la loi binomiale, soit les théorèmes d'approximation de cette loi vus en section 2.5. Dans l'exemple de la germination des graines détaillé ci-dessous nous prendrons $n = 100$ et $p = 0.8$ et utiliserons dons la proposition 13 qui précise non pas la loi asymptotique (c'est-à-dire lorsque n est grand) de $\overline{X}_n$ mais celle de la variable $U_n = \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} = \frac{\Sigma_n - np}{\sqrt{np(1-p)}}$: la loi de cette variable, pour $n = 100$ et $p = 0.8$, pourra être approximée par la loi Gaussienne $\mathcal{N}(0,1)$.

Déroulons le schéma vu précédemment :

1. Déterminer dans quel cadre on se situe, c'est-à-dire préciser le modèle.

   Les v.a. $X_1, \ldots, X_n$ définies par

   $$X_i = \begin{cases} 1 & \text{si la } i\text{-ème graine germe} \\ 0 & \text{si la } i\text{-ème graine ne germe pas} \end{cases}$$

   sont des variables de Bernoulli indépendantes et de même paramètre p. Le paramètre p est la probabilité qu'une graine germe : $p = \mathbb{P}(X_i = 1)$.
   On pose la question : *La probabilité qu'une graine germe p est-elle égale à 80%, ou inférieure à 80 % ?*

2. Au vu de la question posée, déterminer quelles seront les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test. Attention : il faut choisir pour $\mathcal{H}_0$ l'hypothèse sous laquelle on sait faire des calculs.

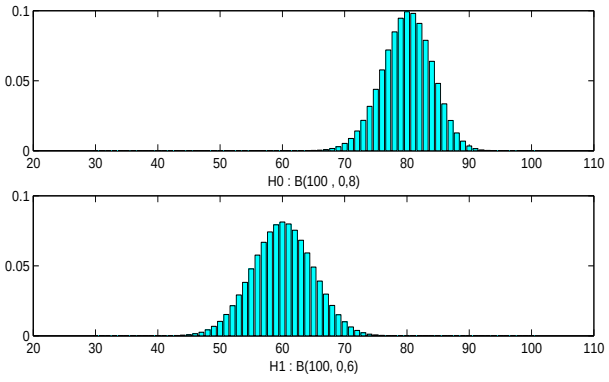
   $\mathcal{H}_0 : p = 0.8$ et $\mathcal{H}_1 : p < 0.8$.

3. Déterminer une statistique (une variable aléatoire fonction de l'échantillon observé) dont on connaît la loi sous $\mathcal{H}_0$ et dont la loi sous $\mathcal{H}_0$ est différente de la loi sous $\mathcal{H}_1$.

   Pour faire un test sur p, on choisit la variable $\Sigma_n = X_1 + \cdots + X_n$ (nombre de graines germant parmi les n graines semées). Sa loi est connue sous $\mathcal{H}_0$: $\Sigma_n \sim \mathcal{B}(n\ ;\ 0.8)$ et sa loi sous $\mathcal{H}_1$ est différente de sa loi sous $\mathcal{H}_0$: sous $\mathcal{H}_1$, $\Sigma_n \sim \mathcal{B}(n\ ;\ p < 0.8)$.

4. À partir de la loi de la statistique sous $\mathcal{H}_0$ et sous $\mathcal{H}_1$, on établit une règle qui permet de décider quand on rejette et quand on accepte $\mathcal{H}_0$. On en déduit la forme de la région de rejet de $\mathcal{H}_0$.

   On remarque que sous $\mathcal{H}_1$, Σ_n a tendance à prendre de plus petites valeurs que sous $\mathcal{H}_0$. Pour s'en convaincre, représentons les histogrammes dans le cas où $n = 100$:



On rejettera donc $\mathcal{H}_0$ si l'on observe une valeur trop petite pour Σ_n. Autrement dit, pour $n = 100$, I est de la forme $I = [A\ ;\ 100]$. On rejette $\mathcal{H}_0$ si $\Sigma_n^{obs} < A$; on conserve $\mathcal{H}_0$ si $\Sigma_n^{obs} \geqslant A$.

5. On calcule le(s) seuil(s) de la région de rejet grâce à la loi de la statistique sous $\mathcal{H}_0$.

   On calcule le seuil A grâce à l'équation :

   $$\mathbb{P}_{\mathcal{H}_0}(\Sigma_n \geqslant A) = 1 - \alpha$$

   ou, ce qui revient au même, grâce à l'équation :

   $$\mathbb{P}_{\mathcal{H}_0}(\Sigma_n < A) = \alpha$$

   Toujours lorsque $n = 100$, en prenant $\alpha = 5\%$, en approximant la loi de $(\Sigma_{100} - 80)/4$ sous $\mathcal{H}_0$ par la loi normale $\mathcal{N}(0, 1)$ (proposition 12) et en effectuant quelques petits calculs, on trouve $A = 73.42$.

6-7. On calcule la valeur observée de la statistique et on conclut.

   Première méthode. La règle de décision est :

   — si $\Sigma_n^{obs} \leqslant 73$, on décide de rejeter $\mathcal{H}_0$;

   — si $\Sigma_n^{obs} \geqslant 74$, on décide de conserver $\mathcal{H}_0$.

   Par exemple, pour $n = 100$, $\Sigma_{100}^{obs} = 70$ conduit à rejeter $\mathcal{H}_0$ et à décider que $p < 0.8$ tandis que $\Sigma_{100}^{obs} = 78$ conduit à conserver $\mathcal{H}_0$ et à décider que $p = 0.8$.

   Seconde méthode. Calculons le niveau de signification du test réalisé sur un échantillon de taille 100, lorsque le nombre de graines ayant germé vaut 78 : $\Sigma_{100}^{obs} = 78$. Le plus petit intervalle I conduisant à conserver $\mathcal{H}_0$ avec cette observation est égal à $[78\ ;\ +\infty[$ (c'est le plus petit intervalle permettant de conserver $\mathcal{H}_0$, le niveau qui lui correspond est bien le niveau de signification α_{obs} : tout niveau $\alpha < \alpha_{obs}$ donnera un plus grand intervalle et conduira à conserver $\mathcal{H}_0$, et tout niveau $\alpha > \alpha_{obs}$ donnera un plus petit intervalle et conduira à rejeter $\mathcal{H}_0$). Effectuons le calcul en approximant la loi de $(\Sigma_{100} - 80)/4$ sous $\mathcal{H}_0$ par la loi normale $\mathcal{N}(0, 1)$ et en notant F la fonction de répartition de la loi normale centrée réduite : $\alpha_{obs} = \mathbb{P}_{\mathcal{H}_0}(\Sigma_{100} \in [78\ ;\ +\infty[) = \mathbb{P}_{\mathcal{H}_0}(\Sigma_{100} \geqslant 78) = \mathbb{P}_{\mathcal{H}_0}((\Sigma_{100} - 80)/4 \geqslant -0.5) \simeq F(-0.5) = 1 - F(0.5) \simeq 30.85\%$. Pour rejeter $\mathcal{H}_0$ lorsque 78 graines parmi 100 ont germé, il faut réaliser un test de niveau $30,85\%$, ce qui n'est pas vraiment raisonnable : une telle observation doit nous conduire à conserver $\mathcal{H}_0$.

   Calculons maintenant le niveau de signification du test réalisé sur un échantillon de taille 100 lorsque le nombre de graines ayant germé vaut 70 : $\Sigma_{100}^{obs} = 70$. Le plus petit intervalle I conduisant à conserver $\mathcal{H}_0$ avec cette observation est égal à $[70\ ;\ +\infty[$. Effectuons le calcul comme précedemment : $\alpha_{obs} = \mathbb{P}_{\mathcal{H}_0}(\Sigma_{100} \in [70\ ;\ +\infty[) = \mathbb{P}_{\mathcal{H}_0}(\Sigma_{100} \geqslant 70) = \mathbb{P}_{\mathcal{H}_0}((\Sigma_{100} - 80)/4 \geqslant -2.5) \simeq F(-2.5) = 1 - F(2.5) \simeq 0.62\%$. Pour conserver $\mathcal{H}_0$ lorsque seulement 70 graines parmi 100 ont germé, il faut réaliser un test de niveau 0.62%, ce qui est trop exigeant : une telle observation doit donc nous conduire à rejeter $\mathcal{H}_0$.

3.4 Hypothèses simples, hypothèses multiples

Définition. *On dit qu'une hypothèse sur un paramètre θ est une hypothèse simple si elle est de la forme $\theta = \theta_0$. Dans les autre cas ($\theta \neq \theta_0$; $\theta <$ ou $\leqslant \theta_0$; $\theta >$ ou $\geqslant \theta_0$), on dit que c'est une hypothèse multiple.*

Dans les exemples 1 (germination des graines) et 2 (pots de confiture), l'hypothèse $\mathcal{H}_0$ était une hypothèse simple : $\mathcal{H}_0 : p = 0.8$ pour l'exemple 1 et $\mathcal{H}_0 : m = 50$ pour l'exemple 2, ce qui permettait de connaître la loi de la statistique sous $\mathcal{H}_0$, et donc de calculer le seuil de la région de rejet de $\mathcal{H}_0$. La situation se complique quelque peu lorsque l'hypothèse $\mathcal{H}_0$ est multiple. Imaginons par exemple que, dans le cadre de l'exemple 1, nous ayons voulu tester l'hypothèse : *La probabilité qu'une graine germe est au moins* 80% contre l'hypothèse *La probabilité qu'une graine germe est inférieure à 80%*. Les hypothèses de ce

test s'écrivent alors $\mathcal{H}'_0 : p \geqslant 0.8$ contre $\mathcal{H}_1 : p < 0.8$. Le même raisonnement qu'en section 3.3 nous conduit à adopter la règle suivante : on rejette $\mathcal{H}'_0$ si $\Sigma_n^{obs} < A$; on conserve $\mathcal{H}'_0$ si $\Sigma_n^{obs} \geqslant A$. L'équation de niveau permettant de trouver le seuil A du test de niveau 5% est $\mathbb{P}_{\mathcal{H}'_0}(\Sigma_n < A) = 0.05$. Cette équation se résout lorque l'on connaît la loi de Σ_n sous $\mathcal{H}'_0$. Mais dans le cas d'une hypothèse multiple, comme ici $\mathcal{H}'_0 : p \geqslant 0.8$, quelle valeur de p prendre sous $\mathcal{H}'_0$? Heureusement, les tests considérés dans ce polycopié sont des tests dits *monotones*, c'est à dire qu'ils conserveront plus facilement notre hypothèse multiple $\mathcal{H}'_0$ si p vaut 0.9 que si p vaut 0.8. Cette remarque en apparence anodine permet de tester $\mathcal{H}'_0 : p \geqslant 0.8$ contre $\mathcal{H}_1 : p < 0.8$ en remplaçant $\mathcal{H}'_0 : p \geqslant 0.8$ par $\mathcal{H}_0 : p = 0.8$.

Conclusion. Le test de $\mathcal{H}'_0 : p \geqslant 0.8$ contre $\mathcal{H}_1 : p < 0.8$ se fait en testant $\mathcal{H}_0 : p = 0.8$ contre $\mathcal{H}_1 : p < 0.8$. De manière plus générale, lors d'un test statistique, lorsque l'hypothèse $\mathcal{H}'_0$ est multiple, on effectue le test en la remplaçant par l'hypothèse simple $\mathcal{H}_0$ associée.

3.5 Erreur de première espèce, erreur de seconde espèce, puissance : α, β et π.

Soit un test statistique de $\mathcal{H}_0$ contre $\mathcal{H}_1$, où $\mathcal{H}_0$ et $\mathcal{H}_1$ sont deux hypothèses simples :

$$\mathcal{H}_0 \ : \ \theta = \theta_0 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta = \theta_1.$$

On note $\mathcal{R}$ la région de rejet de $\mathcal{H}_0$ et on note $\mathcal{R}^c$ la région de conservation de $\mathcal{H}_0$.

1. L'erreur de première espèce α est définie par $\alpha = \mathbb{P}_{\mathcal{H}_0}(\mathcal{R})$.
   L'erreur de première espèce est donc la probabilité de rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie. Cette erreur est fixée par le statisticien. Elle ne peut être ni trop proche de zéro (une erreur de première espèce égale à 0 implique de ne jamais rejeter $\mathcal{H}_0$, il est donc inutile de faire un test statistique) ni trop élevée. Un compromis acceptable est souvent de la choisir égale à 5%.
2. L'erreur de seconde espèce β est définie par $\beta = \mathbb{P}_{\mathcal{H}_1}(\mathcal{R}^c)$.
   L'erreur de seconde espèce est donc la probabilité de ne pas rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie. Cette erreur n'est pas fixée par le statisticien. Elle augmente lorsque α diminue car plus α diminue, plus le test conserve l'hypothèse $\mathcal{H}_0$ et moins le test reconnaît $\mathcal{H}_1$. la seule manière de faire diminuer β lorsque l'on a fixé α est d'augmenter le nombre d'observations.
3. La puissance π est définie par $\pi = \mathbb{P}_{\mathcal{H}_1}(\mathcal{R})$.
   Par passage au complémentaire, remarquons que $\pi = 1 - \mathbb{P}_{\mathcal{H}_1}(\mathcal{R}^c) = 1 - \beta$.
   La puissance est donc la probabilité de rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_1$ est vraie. Cette qualité n'est pas fixée par le statisticien. Elle diminue lorsque α diminue car plus α diminue, plus le test conserve l'hypothèse $\mathcal{H}_0$ et moins le test reconnaît $\mathcal{H}_1$. la seule manière de faire augmenter π lorsque l'on a fixé α est d'augmenter le nombre d'observations.

Exemple. La puissance mesure donc la capacité d'un test à détecter $\mathcal{H}_1$. Dans le cas de l'exemple 2 (pots de confiture), $\mathcal{H}_1$ est l'hypothèse $m \neq 50$, on calcule π en un point précis de $\mathcal{H}_1$. Par exemple, pour $n = 100$, on se demande quelle est la capacité du test de niveau 5%à détecter $m = 50.3$:

$$\pi_{m=50.3} = \mathbb{P}_{m=50.3}(\overline{X} \notin I) = \mathbb{P}_{m=50.3}(\overline{X} \notin [49.804 \ ; \ 50.196]).$$

Sous l'hypothèse $m = 50.3$ la loi de $\overline{X}$ est la loi normale $\mathcal{N}(50.3 \ ; \ \dfrac{1}{100})$. On peut alors finir le calcul :

$$\begin{aligned} \pi_{m=50.3} &= \mathbb{P}_{m=50.3}(10(\overline{X} - 50.3) \notin [-4.96 \ ; \ -1.04]) \\ &= \mathbb{P}(U < -4.96) + \mathbb{P}(U > -1.04) \text{ où } U \sim \mathcal{N}(0,1) \\ &= F(-4.96) + F(+1.04) \end{aligned}$$

où F est la fonction de répartition de la loi $\mathcal{N}(0,1)$. On lit sur la table de cette loi $\pi_{m=50.3} = 85.08\%$. Le test de $m = 50$ contre $m \neq 50$ de niveau 5% et réalisé sur un échantillon de taille 100, a une probabilité 85.08% de détecter $m = 50.3$.

Complément sur la puissance lorsque $\mathcal{H}_1$ est multiple. Supposons que $\mathcal{H}_1$ soit multiple.

— *premier exemple* : supposons que l'on teste

$$\mathcal{H}_0 \ : \ \theta = 5 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta > 5$$

c'est à dire

$$\mathcal{H}_0 \ : \ \theta = 5 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta \in]5 \ ; \ +\infty[.$$

Dans ce cas on peut calculer la puissance du test en tous les $\theta \in]5 \ ; \ +\infty[$. On définit alors la fonction puissance :

$$\theta \to \pi(\theta) = \mathbb{P}_\theta(\mathcal{R}) \qquad \text{pour tout } \theta \in]5 \ ; \ +\infty[.$$

Un test est intéressant si cette fonction est croissante de θ et tend vers 1 quand θ tend vers $+\infty$, cela signifie que plus θ s'éloigne de θ_0 (ici $\theta_0 = 5$), plus la probabilité de rejeter $\mathcal{H}_0$ augmente.

— *second exemple* : supposons que l'on teste

$$\mathcal{H}_0 \ : \ \theta = 5 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta < 5$$

c'est à dire

$$\mathcal{H}_0 \ : \ \theta = 5 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta \in]-\infty \ ; 5[.$$

Dans ce cas on peut calculer la puissance du test en tous les $\theta \in]-\infty \ ; 5[$. On définit alors la fonction puissance :

$$\theta \to \pi(\theta) = \mathbb{P}_\theta(\mathcal{R}) \qquad \text{pour tout } \theta < 5.$$

Un test est intéressant si cette fonction est décroissante de θ et tend vers 1 quand θ tend vers $-\infty$, cela signifie que plus θ s'éloigne de θ_0 (ici $\theta_0 = 5$), plus la probabilité de rejeter $\mathcal{H}_0$ augmente.

— *troisième exemple* : supposons que l'on teste

$$\mathcal{H}_0 \ : \ \theta = 5 \qquad \text{contre} \qquad \mathcal{H}_1 \ : \ \theta \neq 5.$$

Dans ce cas on peut calculer la puissance du test en tous les $\theta \in]-\infty \ ; 5[\cup]5 \ ; \ +\infty[$. On définit alors la fonction puissance :

$$\theta \to \pi(\theta) = \mathbb{P}_\theta(\mathcal{R}) \qquad \text{pour tout } \theta \neq 5.$$

Un test est intéressant si cette fonction est décroissante de θ sur $]-\infty \ ; 5[$, croissante de θ sur $]5 \ ; \ +\infty[$ et tend vers 1 quand θ tend vers $-\infty$ et quand θ tend vers $+\infty$, cela signifie que plus θ s'éloigne de θ_0 (ici $\theta_0 = 5$), plus la probabilité de rejeter $\mathcal{H}_0$ augmente.

3.6 Intervalles de confiance

Souvent, on sait que la loi de la variable Z est d'un certain type (loi normale, loi de Poisson), mais on ne privilégie pas une loi particulière que l'on cherche à tester. On a donc une famille $\mathcal{L}_\theta$ de lois possibles : par exemple $\mathcal{L}_\theta$ serait la loi de Poisson de paramètre θ. On souhaite alors "connaitre" la valeur de θ. Evidemment on ne peut connaitre cette valeur avec certitude, il s'agit donc de donner une fourchette dans laquelle se trouve θ avec forte probabilité. Construire cette fourchette, c'est construire un intervalle de confiance. On se donne un niveau de confiance $1-\alpha$ (la forte probabilité avec laquelle on souhaite que la fourchette contienne θ), et l'on construit un intervalle $J = [A, B]$ tel que la probabilité que θ soit dans J soit au moins $1-\alpha$. Evidemment, A et B dépendent de l'expérience. D'ailleurs, si c'étaient des nombres fixes, comme θ est aussi un nombre fixe, parler de la probabilité que θ soit dans J n'aurait pas de sens. Autrement dit :

Définition. *Un intervalle de confiance J de niveau de confiance $1-\alpha$ pour un paramètre θ est un intervalle $J=[A,B]$ dont les bornes A et B sont des variables aléatoires qui dépendent de l'expérience, dont on peut calculer des valeurs observées à l'aide des observations, et qui vérifie : pour toute valeur possible de θ,*

$$\mathbb{P}(\theta \in J) \geqslant 1-\alpha.$$

Il faut noter que, comme un test, un intervalle de confiance est une procédure avec laquelle on calcule une valeur observée ; si on refait l'expérience, la valeur observée change, mais pas la procédure : ce n'est pas l'intervalle de confiance qui change, c'est sa valeur observée.

Dans l'exemple 2, imaginons que l'on ne cherche plus à faire un test sur le taux de sucre moyen, mais à trouver un encadrement du taux de sucre moyen. La taille de l'échantillon est $n=100$, on cherche un intervalle de confiance au niveau de confiance $1-\alpha=95\%$ pour le taux de sucre moyen réel m. On sait que, quel que soit le taux de sucre moyen m, la variable $10(\overline{X}-m)$ suit la loi normale $\mathcal{N}(0,1)$. De là, pour toute valeur possible de m, on a :

$$\mathbb{P}(-1.96 \leqslant 10(\overline{X}-m) \leqslant +1.96) = 95\%.$$

On en déduit que quelle que soit la valeur de m,

$$\mathbb{P}(\overline{X}-0.196 \leqslant m \leqslant \overline{X}+0.196) = 95\%.$$

L'intervalle $[\overline{X}-0.196\,;\overline{X}+0.196]$ est un intervalle de confiance au niveau de confiance 95 % pour m. Si $\overline{x}_{obs}=49.4$, l'intervalle observé est $J_{obs}=[49.204\ ;\ 49.596]$.

Remarque. Il existe une relation entre l'intervalle de confiance au niveau de confiance $1-\alpha=95\%$ pour m et le test de $\mathcal{H}_0 : m=m_0$ contre $\mathcal{H}_1 : m\neq m_0$ de niveau $\alpha=5\%$. Le test conduit à conserver $\mathcal{H}_0$ si et seulement si $m_0 \in J_{obs}$. Vérifions-le dans le cadre de l'exemple 2, lorsque la taille de l'échantillon est 100 : la valeur m_0 est dans $J_{obs}=[\overline{x}_{obs}-0.196\,;\overline{x}_{obs}+0.196]$ (réalisation de l'intervalle de confiance de niveau de confiance 95%) si et seulement si $\overline{x}_{obs} \in [m_0-0.196\ ;m_0+0.196]$. D'autre part, le test de $\mathcal{H}_0 : m=m_0$ contre $\mathcal{H}_1 : m\neq m_0$ de niveau $\alpha=5\%$ a pour région d'acceptation de $\mathcal{H}_0$ l'intervalle $I=[m_0-0.196\,;m_0+0.196]$ (revoir la section 3.2). Ce test conduit donc à conserver $\mathcal{H}_0$ si et seulement si $\overline{x}_{obs} \in [m_0-0.196\,;m_0+0.196]$.

Remarque. La taille de l'intervalle de confiance varie avec α. Plus on choisit α petit, plus la taille de l'intervalle augmente, donc moins l'encadrement obtenu est précis. Le cas extrême qui consisterait à choisir $\alpha=0\%$ donnerait un intervalle de confiance égal à $]-\infty\ ;\ +\infty[$. Comme pour les test statistiques, le choix $\alpha=5\%$ offre un bon compromis entre un intervalle de taille raisonnable, et une bonne probabilité de tomber dans cet intervalle.

4 Test d'ajustement du chi-deux

4.1 La loi du chi-deux

Définition. *Soient $Y_1,\ldots,Y_d$ des v.a. indépendantes toutes de loi normale centrée réduite $\mathcal{N}(0,1)$. On pose $Q=Y_1^2+Y_2^2+\cdots+Y_d^2$. La loi de Q est une loi à densité, appelée loi du chi-deux à d degrés de liberté, et est notée $\chi^2(d)$. La loi du chi-deux ne possède pas de propriétés de symétrie.*

Proposition 15. $\mathbb{E}(Q)=d$, $\mathbb{V}(Q)=2d$.

Proposition 16. *Supposons que Q suive la loi $\chi^2(d)$. Alors*

$$\mathbb{P}\left(\frac{Q-d}{\sqrt{2d}} \leqslant t\right) \longrightarrow_{d\to+\infty} F(t),$$

où F est la fonction de répartition de la loi normale centrée réduite. En d'autres termes, pour d suffisamment grand, on peut approximer la loi de $\dfrac{Q-d}{\sqrt{2d}}$ par la loi $\mathcal{N}(0,1)$.

4.2 Le modèle étudié

On suppose que l'on a un n-échantillon $X_1, \ldots, X_n$ d'une loi discrète $\mathcal{L}$. On note $a_1, \ldots, a_d$ les valeurs possibles pour X_1. La loi $\mathcal{L}$ est donc caractérisée par la donnée de $\mathbb{P}(X_1 = a_1)$, $\mathbb{P}(X_1 = a_2)$, $\ldots$, $\mathbb{P}(X_1 = a_d)$. On souhaite tester une hypothèse sur cette loi à partir des observations de $X_1, \ldots, X_n$.

Soit l'hypothèse $\mathcal{H}_0$ suivante sur la loi $\mathcal{L}$:

$$\mathbb{P}(X_1 = a_1) = p_1, \ldots, \mathbb{P}(X_1 = a_d) = p_d$$

où $p_1, \ldots, p_d$ sont des nombres donnés explicites tels que $0 \leqslant p_1, \ldots, p_d \leqslant 1$ et $p_1 + \ldots + p_d = 1$.

Pour tester $\mathcal{H}_0$, on veut tirer de l'expérience des informations sur les probabilités $\mathbb{P}(X_1 = a_1)$, $\mathbb{P}(X_1 = a_2)$, $\ldots$, $\mathbb{P}(X_1 = a_d)$. Pour $j = 1, \ldots, d$, on note N_j le nombre de variables X_i prenant la valeur a_j lors des n expériences. Autrement dit N_j est égal au nombre de fois où l'on observe la valeur a_j au cours des n expériences, et N_j/n est la fréquence observée de la valeur a_j au cours des n expériences. Notons que $N_1 + \cdots + N_d = n$, et que la somme des fréquences observées vaut 1. Si l'hypothèse $\mathcal{H}_0$ est vraie, on s'attend à ce que N_j/n soit proche de p_j (c'est la loi des grands nombres : la fréquence empirique tend vers la fréquence théorique (la probabilité) lorsque le nombre d'observations devient grand). Pour faire le test, il est naturel de comparer chacun des p_j avec la fréquence observée N_j/n, ou de comparer chacun des np_j avec N_j.

Par exemple, on relève le groupe sanguin de n individus indépendants et originaires du pays basque. La v.a. X_i représente le groupe sanguin du i-ème individu :

$X_i = 1$ si le groupe sanguin du i-ème individu est O ;

$X_i = 2$ si le groupe sanguin du i-ème individu est A ;

$X_i = 3$ si le groupe sanguin du i-ème individu est B ;

$X_i = 4$ si le groupe sanguin du i-ème individu est AB.

La loi commune des X_i est caractérisée par $\mathbb{P}(X_i = 1)$, $\mathbb{P}(X_i = 2)$, $\mathbb{P}(X_i = 3)$, $\mathbb{P}(X_i = 4)$. On veut tester l'hypothèse $\mathcal{H}_0$:

$$\mathbb{P}(X_i = 1) = 0.43,\ \mathbb{P}(X_i = 2) = 0.45,\ \mathbb{P}(X_i = 3) = 0.09,\ \mathbb{P}(X_i = 4) = 0.03.$$

qui est la répartition classique en France. Parmi les n personnes interrogées, on note N_1 le nombre de personnes de groupe O, N_2 le nombre de personnes de groupe A, N_3 le nombre de personnes de groupe B et N_4 le nombre de personnes de groupe AB. Si l'hypothèse $\mathcal{H}_0$ est vraie, on s'attend à ce que N_1/n soit proche de 0.43 ; N_2/n soit proche de 0.45 ; N_3/n soit proche de 0.09 ; N_4/n soit proche de 0.03.

4.3 Le théorème limite

On pose $n_1 = np_1, n_2 = np_2, \ldots, n_d = np_d$. Si l'hypothèse $\mathcal{H}_0$ est vérifiée, on a : $\mathbb{E}(N_1) = n_1, \ldots, \mathbb{E}(N_d) = n_d$.

Proposition 17. *Lorsque l'hypothèse $\mathcal{H}_0$ est vérifiée et lorsque $n_1 \geq 5, \ldots, n_d \geq 5$, on peut approximer la loi de*

$$Z = \frac{(N_1 - n_1)^2}{n_1} + \cdots + \frac{(N_d - n_d)^2}{n_d}$$

par la loi du chi-deux à $d-1$ degrés de liberté $\chi^2(d-1)$.

Il est important de noter que l'on utilise la loi à $d-1$ degrés de liberté (et non d degrés de liberté). De plus, lorsque la loi $\mathcal{L}$ des X_i n'est pas donnée par $p_1, \ldots p_d$ alors les N_j auront tendance à s'éloigner des n_j et donc la variable Z aura tendance à prendre de plus grandes valeurs que sous $\mathcal{H}_0$.

4.4 Le test

Déroulons le schéma vu précédemment :

1. Déterminer dans quel cadre on se situe, c'est-à-dire préciser le modèle.

   Les v.a. $X_1, \dots, X_n$ sont des v.a. discrètes indépendantes et de même loi admettant d valeurs possibles $a_1, \dots, a_d$.
   Dans notre exemple, $X_1, \dots, X_n$ sont les groupes sanguins de n individus indépendants et originaires du pays basque, les valeurs possibles pour ces v.a. sont 1 (groupe O) ; 2 (groupe A) ; 3 (groupe B) ; 4 (groupe AB).

2. Au vu de la question posée, déterminer quelles seront les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ du test. Attention : il faut choisir pour $\mathcal{H}_0$ l'hypothèse sous laquelle on sait faire des calculs.

   $\mathcal{H}_0 : \mathbb{P}(X_i = a_1) = p_1, \mathbb{P}(X_i = a_2) = p_2, \dots, \mathbb{P}(X_i = a_d) = p_d$
   $\mathcal{H}_1$: l'hypothèse $\mathcal{H}_0$ est fausse
   Dans notre exemple, $\mathcal{H}_0 : \mathbb{P}(X_i = 1) = 0.43, \mathbb{P}(X_i = 2) = 0.45, \mathbb{P}(X_i = 3) = 0.09, \mathbb{P}(X_i = 4) = 0.03$.

3. Déterminer une statistique (une variable aléatoire fonction de l'échantillon observé) dont on connaît la loi sous $\mathcal{H}_0$ et dont la loi sous $\mathcal{H}_0$ est différente de la loi sous $\mathcal{H}_1$.

   On utilise la statistique $Z = \dfrac{(N_1 - n_1)^2}{n_1} + \cdots + \dfrac{(N_d - n_d)^2}{np_d}$ avec $n_j = np_j$ dont la loi sous $\mathcal{H}_0$ (pourvu que $n_1 \geq 5, \dots, n_d \geq 5$) peut être approximée par la loi $\chi^2(d-1)$, et dont la loi sous $\mathcal{H}_1$ est *à droite* de la loi sous $\mathcal{H}_0$. En effet, Z a tendance à prendre des valeurs plus grandes sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$, car sous $\mathcal{H}_1$, les N_j auront tendance à s'éloigner des n_j.
   Dans notre exemple, si $n = 1000$, la statistique utilisée est

   $$Z = \frac{(N_1 - 430)^2}{430} + \frac{(N_2 - 450)^2}{450} + \frac{(N_3 - 90)^2}{90} + \frac{(N_4 - 30)^2}{30}.$$

   Puisque $430, 450, 90, 30$ sont supérieurs à 5, la loi de Z sous $\mathcal{H}_0$ peut être approximée par la loi $\chi^2(3)$.

4. À partir de la loi de la statistique sous $\mathcal{H}_0$ et sous $\mathcal{H}_1$, on établit une règle qui permet de décider quand on rejette et quand on accepte $\mathcal{H}_0$. On en déduit la forme de la région de rejet de $\mathcal{H}_0$.

   Puisque Z a tendance à prendre des valeurs plus grandes sous $\mathcal{H}_0$ que sous $\mathcal{H}_1$, la règle de décision du test est du type :

   — si $z_{obs} > z_0$, on rejette $\mathcal{H}_0$;
   — si $z_{obs} \leqslant z_0$, on conserve $\mathcal{H}_0$.

5. On calcule le(s) seuil(s) de la région de rejet grâce à la loi de la statistique sous $\mathcal{H}_0$.

   Si $n_1 \geq 5, \dots, n_d \geq 5$, on calcule le seuil z_0 grâce à l'équation :

   $$\alpha = \mathbb{P}_{\mathcal{H}_0}(Z > z_0) \simeq \mathbb{P}(Q > z_0) \text{ où } Q \sim \chi^2(d-1).$$

   Dans notre exemple, si $n = 1000$, le seuil du test de niveau 5% de $\mathcal{H}_0$ contre $\mathcal{H}_1$ se calcule en utilisant :

   $$5\% = \mathbb{P}_{\mathcal{H}_0}(Z > z_0) \simeq \mathbb{P}(Q > z_0) \text{ où } Q \sim \chi^2(3).$$

   On lit dans la table de la loi du chi-deux à 3 degrés de liberté : $z_0 = 7.8147$.

6-7. On calcule la valeur observée de la statistique et on conclut.

   Première méthode. Dans notre exemple, si $n = 1000$, la règle de décision est donc :
   si $z_{obs} > 7.8147$, on décide de rejeter $\mathcal{H}_0$;
   si $z_{obs} \leqslant 7.8147$, on décide de conserver $\mathcal{H}_0$.
   Calculons la valeur observée de Z à partir des observations suivantes :

j	1	2	3	4	
N_j	740	250	5	5	1000
n_j	430	450	90	30	1000
$(N_j - n_j)^2$	96100	40000	7225	625	
$\frac{(N_j - n_j)^2}{n_j}$	223.49	88.89	80.28	20.83	$z_{obs} = 413.49$

On rejette $\mathcal{H}_0$: la répartition des groupes sanguins de ces individus n'est pas la même que celle de la population de référence (ici, la population française).

Seconde méthode. La p-valeur est $\alpha_{obs} = \mathbb{P}_{\mathcal{H}_0}(Z > 413.49) \simeq 0$: avec la valeur observée $z_{obs} = 413.49$, on ne peut pas conserver $\mathcal{H}_0$.

4.5 Aménagements

4.5.1 Regroupement par classes

Que faire lorsque pour l'une ou plusieurs des valeurs de j, $n_j < 5$? Ceci ne doit pas être un obstacle pour faire un test qui élimine $\mathcal{H}_0$ si les observations sont trop éloignées des moyennes théoriques. Ce que l'on fait est que l'on regroupe les observations correspondantes aux 'petits n_j' comme une seule et même réponse (qui a une moyenne théorique supérieure à 5), puis on fait le test décrit précédemment. Par exemple, si $d = 5$ et si $n_1 = 50$, $n_2 = 25$, $n_3 = 18$, $n_4 = 3$, $n_5 = 4$, on regroupe les réponses 4 et 5. On a alors

$$n_1 = 50,\ n_2 = 25,\ n_3 = 18,\ n_{4 \text{ ou } 5} = 7.$$

On fait alors le test précédent (avec $d = 4$ puisqu'il n'y a alors que 4 choix possibles).

4.5.2 Hypothèse $\mathcal{H}_0$ non explicite

Un autre problème possible est que l'hypothèse sur $\mathcal{H}_0$ ne donne parfois pas de manière explicite la loi $\mathcal{L}$ et qu'il manque un paramètre pour la déterminer. Par exemple, $\mathcal{H}_0$: "La loi $\mathcal{L}$ est une loi de Poisson". Ici, on ne connait pas le paramètre de la loi de Poisson. On va alors estimer le(s) paramètre(s) manquant(s) – supposons qu'il y en ait r – de manière la plus simple possible, puis on fera le test avec ces valeurs de paramètre(s). Cependant, à la fin, au lieu d'utiliser la table du $\chi^2(d-1)$, on utilise celle du $\chi^2(d-1-r)$: il faut enlever un degré de liberté par paramètre estimé.

Exemple numérique. Sur 100 jours, on observe $N_0 = 21$ jours sans accident sur la $N118$. Il y a $N_1 = 30$ jours avec 1 accident, 29 jours avec 2 accidents, 14 jours avec 3 accidents et 6 jours avec 4 accidents. L'échantillon $X_1, \ldots, X_{1000}$ sous-jacent est défini par : $X_i =$ nombre d'accidents sur la $N118$ le i-ème jour. Ces variables sont supposées indépendantes et de même loi. On souhaite tester au niveau 5% l'hypothèse $\mathcal{H}_0$: "Le nombre d'accidents dans une journée suit une loi de Poisson".

On commence par estimer le paramètre λ de la loi par la moyenne empirique observée. Rappelons que l'espérance d'une variable de Poisson de paramètre λ est λ et que, par la loi des grands nombres, la moyenne empirique

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

est proche de $\mathbb{E}(X_i) = \lambda$ pour n grand. On observe $\overline{x}_{obs} = 1.54$, donc on va prendre 1.54 comme valeur du paramètre λ. Par conséquent : $\mathrm{e}^{-\lambda} = 0.21$. Rappelons aussi que pour une variable de Poisson de paramètre λ, la probabilité de prendre la valeur k est $\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}$. Les probabilités attendues sous $\mathcal{H}_0$ sont donc :

$$p_0 = 0.21,\ p_1 = 0.33,\ p_2 = 0.26,\ p_3 = 0.13,\ p_{4 \text{ ou plus}} = 0.07.$$

Ici, on regroupe toutes les observations supérieures à 4 afin d'avoir une moyenne théorique supérieure à 5 :

$$n_0 = 21, n_1 = 33, n_2 = 26, n_3 = 13, n_{4 \text{ ou plus}} = 7.$$

On observe $z_{obs} = 0/21 + 3^2/33 + 3^2/26 + 1^2/13 + 1^2/7 = 0.83$.
On regarde dans la table de la loi du chi-deux : soit $Q \sim \chi^2(3)$, on lit $\mathbb{P}(Q > 7.8147) = 5\%$.
Comme $0.83 < 7.8147$, on conserve l'hypothèse $\mathcal{H}_0$. La p-valeur est $\alpha_{obs} = \mathbb{P}_{\mathcal{H}_0}(Z > 0.83) = \mathbb{P}(Q > 0.83) = 84.23\%$: avec la valeur observée $z_{obs} = 0.83$, on ne peut pas rejeter $\mathcal{H}_0$.

5 Modèles Gaussiens sur une population

5.1 Exemple introductif

Imaginons que l'on s'intéresse à la quantité de matière grasse présente dans un produit donné. Par une méthode particulière, on peut mesurer cette quantité de matière grasse, et si l'on répète n fois l'expérience, on obtient n résultats $x_1, \ldots, x_n$. Les résultats ne sont pas tous identiques : la variabilité (faible...) étant due à la méthode de mesure. Deux questions se posent alors : quelle est cette variabilité ? Et quelle est, finalement, la quantité de matière grasse présente dans le produit ?

On peut choisir pour ces questions la modélisation gaussienne, c'est-à-dire supposer que les résultats des n expériences sont les valeurs observées de n variables aléatoires $X_1, \ldots, X_n$ indépendantes de loi gaussienne $\mathcal{L} = \mathcal{N}(m, \sigma^2)$. Ce type de modélisation est licite lorsque l'on pense que les variations observées sont dues à l'accumulation de beaucoup de facteurs indépendants (*cf.* le théorème central limite). On veut alors répondre à des questions sur les paramètres m (représentant la quantité "réelle" de matière grasse) et σ (représentant la variabilité de la mesure), c'est-à-dire construire des tests ou des intervalles de confiance. Les situations sont les suivantes :

- m **est inconnu et** σ^2 **est connue** (la variabilité de la méthode a été établie par ailleurs) : cette situation a déjà été vue en section 3.2.
- m **est connu et** σ^2 **est inconnue** (la quantité de matière grasse est connue pour ce produit, on veut connaitre la variabilité d'une nouvelle méthode de mesure) : section 5.2.
- m **est inconnu et** σ^2 **est inconnue** (on ne connait ni la quantité de matière grasse, ni la variabilité de la méthode de mesure) : section 5.3.

5.2 Moyenne connue, variance inconnue

5.2.1 La variance empirique V_n^2

On dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathcal{L} = \mathcal{N}(m_0, \sigma^2)$, où m_0 est connue et σ^2 est inconnue. Puisque

$$\sigma^2 = \mathbb{E}\left((X_1 - m_0)^2\right),$$

on peut penser (méthode empirique) avoir une idée de σ^2 en utilisant la moyenne empirique des écarts quadratiques :

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_0)^2.$$

En effet, de même que $\overline{X}_n$ était la moyenne empirique des $X_1, \ldots, X_n$, la variable V_n^2 est la moyenne empirique du n-échantillon $Y_1, \ldots, Y_n$ où Y_i est définie par : $Y_i = (X_i - m_0)^2$.
De même que l'on avait $\mathbb{E}\left(\overline{X}_n\right) = m$, on a $\mathbb{E}\left(V_n^2\right) = \sigma^2$. Enfin, la relation

$$\frac{nV_n^2}{\sigma^2} = \left(\frac{X_1 - m_0}{\sigma}\right)^2 + \left(\frac{X_2 - m_0}{\sigma}\right)^2 + \ldots + \left(\frac{X_n - m_0}{\sigma}\right)^2,$$

nous donne la loi de nV_n^2/σ^2 : les variables $(X_i - m_0)/\sigma$ sont indépendantes de loi $\mathcal{N}(0,1)$, donc la variable nV_n^2/σ^2 est la somme de n v.a. normales centrées réduites indépendantes, on reconnait la loi du chi-deux à n degrés de liberté (définie en section 4.1) et on note $nV_n^2/\sigma^2 \sim \chi^2(n)$. Pour mémoire, rappelons que dans le cas de l'échantillon gaussien, nous connaissions la loi de la variable $\overline{X}_n$, et de la variable $\sqrt{n}(\overline{X}_n - m)/\sigma$ (théorème 9).

En conclusion, de même que la moyenne empirique $\overline{X}_n$ est l'outil adéquat pour étudier l'espérance m de l'échantillon (par étudier nous entendons d'une part estimer ce paramètre et d'autre part construire des tests et intervalles de confiance pour ce paramètre), la variable V_n^2, appelée *variance empirique*, est l'outil adéquat pour étudier la variance σ^2 de l'échantillon. À partir de la variable V_n^2, nous allons pouvoir construire tests et intervalles de confiance sur σ lorsque l'espérance m_0 est connue. Nous résumons tout ceci dans la proposition suivante, à retenir pour construire tests et intervalles de confiance sur σ lorsque m_0 est connue, et qui est pour la variance empirique l'analogue de la proposition 7 et du théorème 9 pour la moyenne empirique.

Proposition 18 (Définition et propriétés de la variance empirique dans le cas où la moyenne est connue)**.** *La variable aléatoire V_n^2 est appelée variance empirique à moyenne connue. C'est la variable utilisée pour construire tests et intervalles de confiance sur σ dans le cas gaussien à moyenne connue et variance inconnue.*
Elle vérifie $\mathbb{E}\left(V_n^2\right) = \sigma^2$, $\mathbb{V}\left(V_n^2\right) = \dfrac{2\sigma^4}{n}$ *et* $\dfrac{nV_n^2}{\sigma^2} \sim \chi^2(n)$.

Méthodes :

- **Pour tester $\sigma = \sigma_0$ ou $\sigma \leqslant \sigma_0$ ou $\sigma \geqslant \sigma_0$, on utilisera la statistique $Z = \dfrac{nV_n^2}{\sigma_0^2}$.**
- **Pour construire un intervalle de confiance sur σ, on utilisera directement la variable $\dfrac{nV_n^2}{\sigma^2}$.**

Ainsi, pour tester l'hypothèse *la valeur de σ^2 est* 1, on regardera si la valeur observée de $nV_n^2/1$ est plausible pour la loi $\chi^2(n)$. Pour obtenir un encadrement fiable de σ, on cherchera deux valeurs a et b telles que $nV_n^2/\sigma^2 \in [a,b]$ avec grande probabilité. Nous détaillons ces deux approches dans les paragraphes 5.2.2 et 5.2.3.

5.2.2 Tests sur σ

Soit à tester $\mathcal{H}_0$: "$\sigma = \sigma_0$" contre $\mathcal{H}_1$: "$\sigma > \sigma_0$". On choisit comme variable de test $Z = nV_n^2/\sigma_0^2$, qui a tendance à être plus grande sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$. On décide donc de rejeter $\mathcal{H}_0$ quand Z sera plus grand qu'une valeur seuil z_0. Puisque sous $\mathcal{H}_0$, Z suit une loi $\chi^2(n)$, on lit z_0 dans la table du χ^2 à n degrés de liberté, de telle sorte que si on choisit un niveau α, on doit avoir

$$\mathbb{P}_{\mathcal{H}_0}(Z > z_0) = \alpha.$$

On calcule ensuite z_{obs} la valeur observée de Z et on la compare à z_0 : si $z_{obs} > z_0$, on rejette $\mathcal{H}_0$; si par contre, $z_{obs} \leqslant z_0$, on conserve $\mathcal{H}_0$.

Exemple. On veut connaitre la variabilité de la mesure du taux de matière grasse. Pour cela, on effectue des mesures (en grammes pour 1000 grammes) d'un produit dont on sait qu'il contient 60.9 grammes de matière grasse par kilogramme . On obtient les résultats suivants : 60.85 ; 61.05 ; 60.95 ; 60.80 ; 60.90. On veut tester $\mathcal{H}_0$: "$\sigma = 0.1$" contre $\mathcal{H}_1$: "$\sigma > 0.1$". Avec $n = 5$ et au niveau $\alpha = 5\%$, on lit dans la table $\chi^2(5)$: $z_0 = 11.070$. On calcule :

$$\begin{aligned} z_{obs} &= \frac{1}{0.01} \times \left((60.85-60.9)^2 + (61.05-60.9)^2 + (60.95-60.9)^2 + (60.80-60.9)^2 + (60.90-60.9)^2\right) \\ &= 3.75 \end{aligned}$$

Comme $3.75 < z_0$, on ne rejette pas $\mathcal{H}_0$.

5.2.3 Intervalles de confiance pour σ

L'objectif ici est de donner un intervalle pour lequel on puisse affirmer, avec une probabilité connue, que la variabilité de la mesure, c'est-à-dire σ, y appartient.

Rappelons que quelle que soit la véritable valeur de la variabilité σ, on a $Z = \frac{nV_n^2}{\sigma^2} \sim \chi^2(n)$ (proposition 18). On peut donc lire dans la table du χ^2 à n degrés de liberté deux valeurs z_0 et z_1 telles que :

$$\mathbb{P}\left(z_0 \leqslant \frac{nV_n^2}{\sigma^2} \leqslant z_1\right) = 1 - \alpha. \tag{1}$$

On remarque que l'équation (1) se réécrit en : quelle que soit la véritable valeur de la variabilité σ, on a

$$\mathbb{P}\left(\sqrt{\frac{nV_n^2}{z_1}} \leqslant \sigma \leqslant \sqrt{\frac{nV_n^2}{z_0}}\right) = 1 - \alpha.$$

Ceci correspond à la définition d'un intervalle de confiance donnée en section 3.5. Un intervalle de confiance pour σ au niveau de confiance $1 - \alpha$ est donc

$$J = \left[\sqrt{\frac{nV_n^2}{z_1}}; \sqrt{\frac{nV_n^2}{z_0}}\right].$$

Bien remarquer que dans la définition, un intervalle de confiance est un *intervalle aléatoire*, pour lequel on calcule une *valeur observée* J_{obs}.

Exemple. Au niveau de confiance 95%, pour $n = 5$ on lit $z_0 = 0.831$ et $z_1 = 12.833$. On a l'intervalle de confiance

$$J = [\sqrt{\frac{5V_n^2}{12.833}}; \sqrt{\frac{5V_n^2}{0.831}}].$$

On calcule $v_{obs}^2 = 0.0075$. La valeur observée est $J_{obs} = [0.054; 0.213]$.

Remarque. La relation entre l'intervalle de confiance au niveau de confiance $1 - \alpha = 95\%$ pour σ et le test de $\mathcal{H}_0 : \sigma = \sigma_0$ contre $\mathcal{H}_1 : \sigma \neq \sigma_0$ de niveau $\alpha = 5\%$ est la suivante : le test conduit à conserver $\mathcal{H}_0$ si et seulement si $\sigma_0 \in J_{obs}$.

5.3 Moyenne inconnue, variance inconnue

5.3.1 La variance empirique S_n^2 et le test sur la variance

Lorsque la moyenne m est inconnue, la variable V_n^2 ne peut plus être utilisée comme variance empirique car elle contient le paramètre m. Nous allons définir une autre variable, notée S_n^2, qui permettra d'étudier la variance σ^2 lorsque le paramètre m est inconnu. Cette variable est obtenue en remplaçant dans l'expression de V_n^2 le paramètre m (qui est maintenant un paramètre de nuisance pour estimer la variance σ^2) par son estimateur $\overline{X}_n$. On appelle cette nouvelle variable *variance empirique dans le cas où la moyenne est inconnue* (rapidement on dira simplement *variance empirique*, même si ce terme désigne alors à la fois l'estimateur de la variance dans le cas où m est connue, et dans le cas où m est inconnue). La proposition 19 est l'analogue de la proposition 18 dans le cas où m est inconnue. Elle est à retenir pour construire tests et intervalles de confiance sur σ lorsque m est inconnue.

Proposition 19 (Définition et propriétés de la variance empirique dans le cas où la moyenne est inconnue). *La variable S_n^2 définie par :*

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

est appelée variance empirique à moyenne inconnue. C'est la variable utilisée pour construire tests et intervalles de confiance sur σ dans le cas d'un échantillon gaussien de moyenne et variance inconnues.
Elle vérifie $\mathbb{E}\left(S_n^2\right) = \sigma^2$, $\mathbb{V}\left(S_n^2\right) = \dfrac{2\sigma^4}{n-1}$ *et* $\dfrac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$.

Méthodes :

- **Pour tester $\sigma = \sigma_0$ ou $\sigma \leqslant \sigma_0$ ou $\sigma \geqslant \sigma_0$, on utilisera la statistique $Z = \dfrac{(n-1)S_n^2}{\sigma_0^2}$.**
- **Pour construire un intervalle de confiance sur σ, on utilisera directement la variable $\dfrac{(n-1)S_n^2}{\sigma^2}$.**

Remarque. Notons que la somme est divisée par $n-1$ et non par n, et que le nombre de degrés de liberté de la loi du chi-deux est $n-1$ et non plus n. En statistique, il est très fréquent de perdre un degré de liberté dans les les lois à chaque fois que l'on remplace un paramètre par son estimateur (ici, on a remplacé m par $\overline{X}_n$).

5.3.2 La loi de Student

La variable $\overline{X}_n$ peut toujours être utilisée pour estimer m, que σ soit connu ou non, puisque son expression ne fait pas apparaitre le paramètre σ. La loi de $U_n = \frac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$ est toujours la loi $\mathcal{N}(0,1)$, mais cette propriété ne peut plus être utilisée pour construire tests et intervalles de confiance sur le paramètre m puisque U_n fait intervenir le paramètre σ.
En effet, essayons de construire un intervalle de confiance sur m à partir de $U_n = \frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \sim \mathcal{N}(0,1)$:

$$\mathbb{P}(-1.96 \leqslant U_n \leqslant 1.96) = 0.95$$

donc

$$\mathbb{P}\left(-1.96 \leqslant \frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \leqslant 1.96\right) = 95\%$$

ce qui donne

$$\mathbb{P}\left(\overline{X}_n - \frac{1.96\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X}_n + \frac{1.96\sigma}{\sqrt{n}}\right) = 95\%$$

L'intervalle $[\overline{X}_n - \frac{1.96\sigma}{\sqrt{n}}\ ;\ \overline{X}_n + \frac{1.96\sigma}{\sqrt{n}}]$ est un intervalle qui contient m avec une probabilité 95 %, mais ce n'est pas un intervalle de confiance pour m au niveau de confiance 95% puisque l'on ne peut pas calculer sa valeur numérique à partir des données (il manque la valeur de σ). Pour construire tests et intervalles de confiance sur le paramètre m, il est nécessaire de connaitre la loi, soit de $\overline{X}_n$, soit d'une variable construite sur $\overline{X}_n$ *dont on puisse calculer une valeur observée* (une telle variable s'appelle d'ailleurs une statistique). Assez naturellement, on remplace la valeur inconnue σ par son estimateur S_n, mais ce faisant, on passe d'une loi $\mathcal{N}(0,1)$ à une loi de Student.

Proposition 20.

$$\frac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \sim \mathcal{N}(0,1) \quad \textit{devient} \quad \frac{\sqrt{n}(\overline{X}_n - m)}{S_n} \sim \mathcal{T}(n-1)$$

La loi $\mathcal{T}(d)$ est appelée loi de Student à d degrés de liberté.

À retenir : Cette loi possède les mêmes propriétés de symétrie autour de zéro que la loi normale centrée réduite. En particulier, sa fonction de répartition F_T vérifie $F_T(-t) = 1 - F_T(t)$ pour tout t positif, et cela quel que soit le nombre de degrés de liberté de la loi de Student. Lorsque les degrés de liberté augmentent, la loi de Student se rapproche de la loi normale centrée réduite, en général on approxime la loi $\mathcal{T}(d)$ par la loi $\mathcal{N}(0,1)$ dès que $d \geqslant 100$.

5.3.3 Le test de Student d'adéquation à une moyenne théorique, ou test de Student de conformité

Rappelons les résultats obtenus dans le cas σ connu (théorème 9) :

— La loi de $\overline{X}_n$ est la loi gaussienne $\mathcal{N}(m, \frac{\sigma^2}{n})$.

— La loi de U_n, où U_n est la v.a. définie par $U_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma}$, est la loi $\mathcal{N}(0,1)$.

Les lois de $\overline{X}_n$ et de U_n sont toujours celles décrites par ce théorème ; le problème est que ces lois dépendent du paramètre σ qui, contrairement au cas où *m est inconnu et σ^2 connue* (traité en section 3.2), devient ici un paramètre de nuisance. Nous venons de voir en section 5.3.2 comment nous *débarrasser* de σ en le remplaçant par son estimateur S_n, auquel cas la loi $\mathcal{N}(0,1)$ se tranforme alors en une loi de Student :

— La loi de T_n, où T_n est la v.a. définie par $T_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{S_n}$, est la loi $\mathcal{T}(n-1)$.

La proposition 21 et les méthodes qui suivent sont à retenir pour construire tests et intervalles de confiance sur m lorsque σ est inconnu.

Proposition 21. *1. Les variables $\overline{X}_n$ et S_n^2 sont indépendantes.*

2. La loi de $T_n = \dfrac{\sqrt{n}(\overline{X}_n - m)}{S_n}$ est la loi de Student $\mathcal{T}(n-1)$.

Méthodes :

— **Pour tester $m = m_0$ ou $m \leqslant m_0$ ou $m \geqslant m_0$, lorsque σ^2 est inconnue, on utilisera la statistique**

$$T_n = \frac{\sqrt{n}(\overline{X}_n - m_0)}{S_n}.$$

Le test basé sur cette statistique est appelé test de Student d'adéquation à une moyenne théorique, ou test de Student de conformité.

— **Pour construire un intervalle de confiance sur m lorsque σ^2 est inconnue, on utilisera directement la variable $\dfrac{\sqrt{n}(\overline{X}_n - m)}{S_n}$.**

Exemple. Soient $n = 16$ et $X_1, \ldots, X_{16}$ un 16-échantillon de loi $\mathcal{N}(m, \sigma)$ de paramètres m et σ tous deux inconnus. On cherche un intervalle de confiance de niveau de confiance 95% pour m.

On va utiliser $\dfrac{\sqrt{16}(\overline{X}_{16} - m)}{S_{16}} \sim \mathcal{T}(n-1) = \mathcal{T}(15)$.

On lit sur la table de valeurs numériques $\mathbb{P}(-2.131 \leqslant Z_{16} \leqslant 2.131) = 0.95$, donc

$$\mathbb{P}\left(-2.131 \leqslant \frac{\sqrt{16}(\overline{X}_{16} - m)}{S_{16}} \leqslant 2.131\right) = 95\%$$

ce qui donne

$$\mathbb{P}\left(\overline{X}_n - \frac{2.131 S_{16}}{\sqrt{16}} \leqslant m \leqslant \overline{X}_n + \frac{2.131\sigma}{\sqrt{16}}\right) = 95\%$$

L'intervalle $[\overline{X}_{16} - \frac{2.131 S_{16}}{\sqrt{16}}\ ;\ \overline{X}_{16} + \frac{2.131 S_{16}}{\sqrt{16}}]$ est un intervalle qui contient m avec une probabilité 95 %, et c'est bien un intervalle de confiance pour m au niveau de confiance 95% puisque l'on peut calculer sa valeur numérique à partir des données ($\overline{X}_{16}$ et S_{16} sont calculables à partir des données).

6 Modèles Gaussiens sur deux populations

On s'intéresse maintenant à la situation où l'on cherche à comparer deux quantités. Par exemple, on considère deux produits dont on veut comparer les taux de matière grasse. On fait une série de mesures pour le premier produit, que l'on modélise comme étant les valeurs observées d'un n-échantillon d'une loi gaussienne $\mathcal{L}_1 = \mathcal{N}(m_1, \sigma^2)$, et une série de mesures pour le deuxième produit, que l'on modélise comme étant les valeurs observées d'un m-échantillon d'une loi gaussienne $\mathcal{L}_2 = \mathcal{N}(m_2, \sigma^2)$. Si l'on procède avec la même méthode de mesure dans les deux cas, on peut considérer que les deux lois ont la même variance inconnue ou non σ^2. Dans ce cas, on cherchera à comparer m_1 et m_2 (section 6.1 ou section 6.2).

Enfin, dans la situation où l'on dispose d'un n-échantillon de loi $\mathcal{L}_1 = \mathcal{N}(m_1, \sigma_1^2)$ issu d'une première population (par exemple les volumes sanguins de n coureurs de fond) et d'un m-échantillon de loi $\mathcal{L}_2 = \mathcal{N}(m_2, \sigma_2^2)$ (par exemple les volumes sanguins de m haltérophiles), on peut se demander si d'une part $m_1 = m_2$ (dans le célèbre exemple des coureurs et des haltérophiles, la réponse est oui) mais aussi si $\sigma_1 = \sigma_2$ (toujours dans ce même exemple, la réponse est non). Dans ce cas, on cherchera à comparer σ_1 et σ_2 (section 6.3).

6.1 Comparaison des moyennes de deux échantillons gaussiens indépendants et de même variance connue

Soient $X_1, \ldots, X_n$ un n-échantillon de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \ldots, Y_m$ un m-échantillon de loi $\mathcal{N}(m_2, \sigma^2)$. Les deux échantillons sont supposés indépendants, c'est-à-dire que pour tous i et j, X_i est indépendant de Y_j. Ceci entraine l'indépendance des moyennes empiriques $\overline{X}_n$ et $\overline{Y}_m$. De plus, la variable $\overline{X}_n - \overline{Y}_m$ est un estimateur de la différence $m_1 - m_2$. On pose

$$\theta = m_1 - m_2.$$

Comparer les moyennes des deux échantillons revient à faire un test sur la différence θ, et pour cela on utilisera la variable qui estime cette différence, c'est-à-dire la variable $\overline{X}_n - \overline{Y}_m$. La loi de $\overline{X}_n - \overline{Y}_m$ est calculable en utilisant l'indépendance de $\overline{X}_n$ et $\overline{Y}_m$ et en calculant l'espérance et la variance de $\overline{X}_n - \overline{Y}_m$ (utiliser les propriétés des variables aléatoires gaussiennes vues en section 2.3). On obtient :

$$\overline{X}_n - \overline{Y}_m \sim \mathcal{N}\left(\theta\ ;\ \sigma^2(\frac{1}{n} + \frac{1}{m})\right)$$

On peut alors dérouler la mécanique de construction d'un test sur θ. Notons qu'une fois de plus, pour réaliser un test sur un paramètre (ici, le paramètre θ) il faut :

a) choisir un estimateur de θ (ici l'estimateur $\overline{X}_n - \overline{Y}_m$) ;

b) savoir calculer la loi de cet estimateur sous une hypothèse sur θ (ici, sous l'hypothèse $\theta = 0$, l'estimateur $\overline{X}_n - \overline{Y}_m$ suit la loi connue $\mathcal{N}\left(0\ ;\ \sigma^2(\frac{1}{n} + \frac{1}{m})\right)$).

La proposition et les méthodes qui suivent donnent les résultats théoriques nécessaires à la construction de tests ou d'intervalles de confiance sur θ.

Proposition 22. *La loi de* $\dfrac{(\overline{X}_n - \overline{Y}_m) - \theta}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ *est la loi normale* $\mathcal{N}(0, 1)$.

Méthodes :

— **Pour tester $\theta = 0$ (égalité des moyennes des deux échantillons) ou $\theta \leqslant 0$ ou $\theta \geqslant 0$ (comparaison des moyennes des deux échantillons) dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance σ^2 connue, on utilisera la statistique**

$$Z = \frac{(\overline{X}_n - \overline{Y}_m)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

— **Pour construire un intervalle de confiance sur θ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance σ^2 connue, on utilisera directement la variable**

$$\frac{(\overline{X}_n - \overline{Y}_m) - \theta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

6.2 Comparaison des moyennes de deux échantillons gaussiens indépendants et de même variance inconnue

De même qu'en section 5.3, lorsque la variance σ^2 est inconnue, elle doit être remplacée par un estimateur, que nous noterons $\mathcal{S}^2$, qui possède un intérêt propre pour étudier σ^2 via des tests et des intervalles de confiance. À un facteur multiplicatif près, la loi de $\mathcal{S}^2$ est une loi du chi-deux (*cf.* la proposition 23 ci-dessous).

Pour effectuer un test ou construire un intervalle de confiance sur θ, la variable

$$\frac{(\overline{X}_n - \overline{Y}_m) - \theta}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}},$$

de loi $\mathcal{N}(0,1)$, est remplacée par la variable

$$\frac{(\overline{X}_n - \overline{Y}_m) - \theta}{\mathcal{S}\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

dont la loi est une loi de Student.

La section 6.2.1 précise les propriétés statistiques de $\mathcal{S}^2$ et résume les résultats théoriques nécessaires à la construction de tests ou d'intervalles de confiance sur σ ; la section 6.2.2 résume les résultats théoriques nécessaires à la construction de tests ou d'intervalles de confiance sur θ.

6.2.1 La variance empirique et le test sur la variance commune

Proposition 23 (Définition et propriétés de la variance empirique dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance inconnue)**.** *La variable $\mathcal{S}^2$ définie par :*

$$\mathcal{S}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2\right)$$

est l'estimateur de la variance lorsque l'on dispose de deux échantillons gaussiens de moyennes inconnues et de même variance (inconnue). On utilisera cette variable pour construire tests et intervalles de confiance sur σ dans le dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance inconnue. Elle vérifie $\mathbb{E}\left(\mathcal{S}^2\right) = \sigma^2$ et la loi de $\dfrac{(n+m-2)\mathcal{S}^2}{\sigma^2}$ est la loi du chi-deux $\chi^2(n+m-2)$.

Méthodes :

— **Pour tester $\sigma = \sigma_0$ ou $\sigma \leqslant \sigma_0$ ou $\sigma \geqslant \sigma_0$ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance inconnue, on utilisera la statistique $Z = \dfrac{(n+m-2)\mathcal{S}^2}{\sigma_0^2}$.**

— **Pour construire un intervalle de confiance sur σ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance inconnue, on utilisera directement la variable $\dfrac{(n+m-2)\mathcal{S}^2}{\sigma^2}$.**

6.2.2 Le test de Student de comparaison de deux moyennes théoriques, ou test de Student d'homogénéité

Proposition 24. *1. Les variables $\overline{X}_n - \overline{Y}_m$ et $\mathcal{S}^2$ sont indépendantes.*

2. La loi de $\dfrac{(\overline{X}_n - \overline{Y}_m) - \theta}{\mathcal{S}\sqrt{\frac{1}{n} + \frac{1}{m}}}$ est la loi $\mathcal{T}(n+m-2)$.

Méthodes :

— **Pour tester $\theta = 0$ (égalité des moyennes des deux échantillons) ou $\theta \leqslant 0$ ou $\theta \geqslant 0$ (comparaison des moyennes des deux échantillons) dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance σ^2 inconnue, on utilisera la statistique**

$$Z = \frac{(\overline{X}_n - \overline{Y}_m)}{\mathcal{S}\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

— **Pour construire un intervalle de confiance sur θ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de même variance σ^2 inconnue, on utilisera directement la variable**

$$\frac{(\overline{X}_n - \overline{Y}_m) - \theta}{\mathcal{S}\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

6.3 Comparaison des variances de deux échantillons gaussiens indépendants

Soient $X_1, \ldots, X_n$ un n-échantillon de loi $\mathcal{N}(m_1, \sigma_1^2)$ et $Y_1, \ldots, Y_m$ un m-échantillon de loi $\mathcal{N}(m_2, \sigma_2^2)$. On suppose que les deux échantillons sont indépendants, c'est-à-dire que pour tous i et j, X_i est indépendant de Y_j. On pose

$$\lambda = \sigma_2/\sigma_1.$$

Comparer les variances des deux échantillons revient à faire un test sur λ. Si les moyennes m_1, m_2 sont connues, les estimateurs de σ_1^2 et σ_2^2 sont respectivement V_1^2 et V_2^2 définis en section 5.2.1. Si les moyennes m_1, m_2 sont inconnues, les estimateurs de σ_1^2 et σ_2^2 sont respectivement S_1^2 et S_2^2 définis en section 5.3.1 :

$$V_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_1)^2 \quad ; \quad V_2^2 = \frac{1}{m}\sum_{j=1}^{m}(Y_j - m_2)^2 \quad ;$$

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad ; \quad S_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2.$$

L'estimateur du rapport λ^2 sera $\dfrac{V_2^2}{V_1^2}$ lorsque les moyennes m_1 et m_2 sont connues, et sera $\dfrac{S_2^2}{S_1^2}$ lorsque les moyennes m_1 et m_2 sont inconnues. La loi de ces rapports est connue : on l'appelle la loi de Fisher. La mécanique des tests peut de nouveau être appliquée.

La proposition suivante résume les résultats théoriques nécessaires à la construction de tests ou d'intervalles de confiance sur λ.

Proposition 25. *1. La loi de $\dfrac{V_2^2/\sigma_2^2}{V_1^2/\sigma_1^2}$, que l'on peut également écrire $\left(\dfrac{V_2^2}{V_1^2}\right)/\lambda^2$, est la loi de Fisher à m et n degrés de liberté, notée $\mathcal{F}(m,n)$. Cette loi est tabulée.*

— *Pour tester $\lambda = 1$ (égalité des variances des deux échantillons) ou $\lambda \leqslant 1$ ou $\lambda \geqslant 1$ (comparaison des variances des deux échantillons) dans le cas de deux échantillons gaussiens indépendants de moyennes connues et de variances inconnues, on utilisera la statistique $\dfrac{V_2^2}{V_1^2}$.*

— *Pour construire un intervalle de confiance sur λ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de variances inconnues, on utilisera directement la variable* $\left(\frac{V_2^2}{V_1^2}\right)/\lambda^2$.

2. *La loi de* $\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$, *que l'on peut également écrire* $\left(\frac{S_2^2}{S_1^2}\right)/\lambda^2$, *est la loi de Fisher à* $m-1$ *et* $n-1$ *degrés de liberté, notée* $\mathcal{F}(m-1, n-1)$. *Cette loi est tabulée.*

— *Pour tester* $\lambda = 1$ *(égalité des variances des deux échantillons) ou* $\lambda \leqslant 1$ *ou* $\lambda \geqslant 1$ *(comparaison des variances des deux échantillons) dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de variances inconnues, on utilisera la statistique* $\frac{S_2^2}{S_1^2}$.

— *Pour construire un intervalle de confiance sur λ dans le cas de deux échantillons gaussiens indépendants de moyennes inconnues et de variances inconnues, on utilisera directement la variable* $\left(\frac{S_1^2}{S_2^2}\right)/\lambda^2$.

7 Couples de variables aléatoires

7.1 Couple de variables aléatoires discrètes

7.1.1 Lois de probabilité et moments dans le cas discret

Soient X et Y deux variables aléatoires discrètes.

Définition. *La loi de probabilité du couple (X, Y) est donnée par*

$$\{\mathbb{P}((X,Y) = (x,y)) \ ; \ (x,y) \text{ valeur possible pour } (X,Y)\}.$$

En général on note $\mathbb{P}(X = x, Y = y)$ *pour* $\mathbb{P}((X,Y) = (x,y))$.

Si l'on connaît la loi du couple (X, Y), on peut retrouver la loi de X et la loi de Y qui sont appelées *lois marginales* : $\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y)$ et $\mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y)$. Par contre, si l'on connaît seulement les deux lois marginales on ne peut pas retrouver la loi du couple.

Définition. *L'espérance et la matrice de covariance de (X, Y) sont :*

$$\mathbb{E}((X,Y)) = (\mathbb{E}(X), \mathbb{E}(Y)) \text{ et } \Gamma((X,Y)) = \begin{pmatrix} \mathbb{V}(X) & Cov(X,Y) \\ Cov(X,Y) & \mathbb{V}(Y) \end{pmatrix} \tag{2}$$

avec $Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Nous interpréterons cette quantité en section 7.3.

Soit Φ une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$, alors $\Phi(X, Y)$ est une variable aléatoire discrète et on a

$$\mathbb{E}(\Phi(X,Y)) = \sum_x \sum_y \Phi(x,y)\mathbb{P}(X = x, Y = y).$$

En particulier : $\mathbb{E}(XY) = \sum_x \sum_y xy\mathbb{P}(X = x, Y = y)$.

7.1.2 Indépendance dans le cas discret

On dit que deux variables aléatoires discrètes X et Y sont *indépendantes* ssi :

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$$

pour tout x et tout y. Autrement dit, X et Y sont indépendantes ssi la loi jointe de (X, Y) est égale au produit des lois marginales de X et Y. Une définition plus intuitive de l'indépendance de X et Y est :

$$\mathbb{P}(X = x \,|\, Y = y) = \mathbb{P}(X = x)$$

pour tout x et tout y tel que $\mathbb{P}(Y = y) \neq 0$; la connaissance de la valeur prise par Y n'influe pas sur la probabilité que X prenne la valeur x.

Définition. *Soient $X_1, \ldots, X_n$ n variables aléatoires discrètes. On dit que ces variables sont* indépendantes entre elles *si et seulement si pour tous $x_1, \ldots, x_n$:*

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1)\,\mathbb{P}(X_2 = x_2)\ldots\mathbb{P}(X_n = x_n)$$

.

Proposition 26 (Propriétés des variables aléatoires indépendantes (rappel))**.**

- *Si X et Y sont indépendantes alors $Cov(X, Y) = 0$ mais la réciproque est fausse.*
- *Soient $X_1, \ldots, X_n$ des variables aléatoires* indépendantes *alors $\mathbb{E}(\prod_{i=1}^n X_i) = \prod_{i=1}^n \mathbb{E}(X_i)$.*
- *Soient $X_1, \ldots, X_n$ des variables aléatoires* indépendantes *alors $\mathbb{V}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbb{V}(X_i)$.*

7.1.3 Loi conditionnelle dans le cas discret

Soient X et Y deux variables aléatoires et y une valeur possible pour Y (i.e telle que $\mathbb{P}(Y = y) > 0$).

Définition. *La loi conditionnelle de X sachant $\{Y = y\}$ est donnée par*

$$\{\mathbb{P}(X = x \,|\, Y = y) \ ; \ x \text{ valeur possible pour } X\}.$$

On peut noter $\mathbb{P}^{Y=y}(X = x)$ pour $\mathbb{P}(X = x \,|\, Y = y)$. Rappelons la formule des probabilités composées :

$$\mathbb{P}^{Y=y}(X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

À partir de la loi de Y et de la loi conditionnelle de X sachant $Y = y$ pour tout y, on retrouve la loi du couple (X, Y) : pour tout (x, y) valeur possible de (X, Y), $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \,|\, Y = y)\,\mathbb{P}(Y = y)$.

7.2 Couple de variables aléatoires admettant une densité

7.2.1 Lois de probabilité et moments dans le cas à densité

C'est le cas où l'ensemble des valeurs que peut prendre (X, Y) est une partie de $\mathbb{R}^2$. La loi de (X, Y) peut alors être identifiée par sa densité de probabilité $f(x, y)$. Comme dans le cas d'une variable X, on a :

$$f(x, y) \geqslant 0 \text{ et } \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)dxdy = 1.$$

La loi permet de calculer les probabilités d'événements relatifs à (X, Y) : si A est une partie de $\mathbb{R}^2$,

$$\mathbb{P}(((X, Y) \in A) = \int\int_{(x,y)\in A} f(x, y)dxdy.$$

Exemple. Soit (X, Y) un couple de densité de probabilité f définie par

$$f(x,y) = \begin{cases} 6xy^2 & \text{si } 0 \leqslant x \leqslant 1 \text{ et } 0 \leqslant y \leqslant 1 \\ 0 & \text{sinon} \end{cases}$$

On a bien $f(x,y) \geqslant 0$ pour tous x et y, et

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy &= \int_0^1\int_0^1 6xy^2 dxdy \\ &= \int_0^1 6x\left[\frac{y^3}{3}\right]_0^1 dx \\ &= \int_0^1 2x dx \\ &= 1. \end{aligned}$$

Si on veut par exemple calculer $\mathbb{P}(X < Y)$:

$$\begin{aligned} \mathbb{P}(X<Y) &= \int\int_{x<y} f(x,y)dxdy \\ &= \int_0^1 y^2 dy \int_0^y 6x dx \\ &= \int_0^1 y^2 dy \left[3x^2\right]_0^y \\ &= \int_0^1 3y^4 dy \\ &= 3/5. \end{aligned}$$

Comme dans le cas discret, si l'on connaît la loi de (X, Y), on peut retrouver la loi de X et la loi de Y (lois marginales) : la loi de X sera la loi de densité de probabilité $g(x)$ définie par

$$g(x) = \int_{-\infty}^{+\infty} f(x,y)dy \ .$$

On remarque que g est bien une densité de probabilité : $g(x) \geqslant 0$ pour tout x et

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

De même, la loi marginale de Y peut être identifiée par sa densité de probabilité $h(y)$ où l'on pose :

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

Par contre, si l'on connaî t seulement les deux lois marginales on ne peut pas retrouver la loi du couple.

Exemple. La densité marginale de X est

$$g(x) = \begin{cases} \int_0^1 6xy^2 dy = 6x[\frac{y^3}{3}]_0^1 = 2x & \text{si } 0 \leqslant x \leqslant 1 \\ 0 & \text{sinon} \end{cases} .$$

La densité marginale de Y est

$$h(y) = \begin{cases} \int_0^1 6xy^2 dx = 6y^2[\frac{x^2}{2}]_0^1 = 3y^2 & \text{si } 0 \leqslant y \leqslant 1 \\ 0 & \text{sinon} \end{cases} .$$

Les définitions de l'espérance et de la matrice de variance-covariance de (X, Y) sont les mêmes que dans le cas discret (*cf.* l'équation (2)). Lors des calculs, les sommes deviennent des intégrales.
Si Φ une fonction *intégrable* de $\mathbb{R}^2$ dans $\mathbb{R}$, alors $\Phi(X, Y)$ est une variable aléatoire et on a

$$\mathbb{E}\left(\Phi(X,Y)\right) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(x,y)f(x,y)dxdy.$$

En particulier, on a : $\mathbb{E}\left(XY\right) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy.$

7.2.2 Indépendance dans le cas à densité

Comme dans le cas discret, X et Y sont indépendantes ssi la loi jointe de (X, Y) est égale au produit des lois marginales de X et Y. Autrement dit, X et Y sont indépendantes ssi pour tout x et pour tout y :

$$f(x, y) = g(x)h(y)$$

où f est la densité du couple (X, Y), g la densité de X et h la densité de Y.

On peut définir de manière analogue l'indépendance de $X_1, X_2, \ldots, X_n$ lorsque $(X_1, X_2, \ldots, X_n)$ admet la densité de probabilité $f(x_1, x_2, \ldots, x_n)$: $X_1, X_2, \ldots, X_n$ sont indépendantes ssi $f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\ldots f_n(x_n)$ où f_1 est la densité de X_1, f_2 est la densité de X_2, ..., f_n est la densité de X_n. Autrement dit, la densité de la loi jointe est égale au produit des densités des lois marginales.

Les propriétés des variables aléatoires indépendantes sont les mêmes que dans le cas discret. Nous les rappelons encore une fois :

- Si X et Y sont indépendantes alors $\mathrm{Cov}(X, Y) = 0$ mais la réciproque est fausse.
- Soient $X_1, \ldots, X_n$ des variables aléatoires *indépendantes* alors $\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}\left(X_i\right)$.
- Soient $X_1, \ldots, X_n$ des variables aléatoires *indépendantes* alors $\mathbb{V}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{V}\left(X_i\right)$.

7.3 Le problème de l'indépendance

Un individu fournit deux réponses X et Y. Par exemple, l'individu est un champ de blé, X représente le nombre de quintaux par hectare et Y la quantité d'eau de pluie tombée pendant la germination ; ou encore, si tous les pommiers d'un verger ont reçu aléatoirement un traitement A, B, C ou D, à un "individu", c'est-à-dire un rameau choisi au hasard dans ce verger, on peut associer X le traitement reçu par ce rameau et Y le nombre de fruits portés par ce rameau. Dans les deux cas un individu fournit en réponse un couple de v.a. (X, Y) : dans le premier cas le couple admet une densité, dans le second cas, le couple est un couple de v.a. discrètes. Le problème est de déterminer s'il existe une liaison entre X et Y, ou bien si X et Y sont indépendantes. La loi jointe du couple (X, Y) décrit complètement comment les variables X et Y sont liées, et leur indépendance éventuelle. Lorsque le couple (X, Y) admet une densité, la loi jointe est difficilement estimable. Aussi on a recours à deux *indicateurs de liaison* entre X et Y, qui sont la covariance entre X et Y, et la corrélation entre X et Y. Lorsque le couple (X, Y) est un couple de v.a. discrètes, le test du chi-deux d'indépendance permet de tester directement l'indépendance des deux caractères X et Y et le test du chi-deux d'homogénéité permet de tester que le fait d'appartenir à un groupe plutôt qu'à un autre n'influe pas sur les résultats. Sur l'exemple des pommiers, le test du chi-deux d'indépendance testera l'indépendance entre le nombre de fruits par rameau Y et le traitement X reçu par le rameau, tandis que le test du chi-deux d'homogénéité testera que la loi du nombre de fruits par rameau Y ne dépend pas du groupe (A, B, C ou D) auquel appartient le rameau. Ces deux points de vue conduisent en fait à la même règle statistique, même s'il n'est pas évident au premier abord que les statistiques de test coïncident.

7.3.1 Covariance et corrélation

Ces deux indicateurs sont particulièrement utiles lorsque le couple (X, Y) admet une densité.

Définition. *La covariance entre X et Y est le nombre*

$$Cov(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).$$

La covariance entre X et Y "mesure", dans un sens précisé plus loin, la liaison affine entre X et Y, mais elle est dans la même unité que le produit XY. Pour supprimer cette dépendance à l'unité de mesure, on définit le coefficient de corrélation entre X et Y :

Définition. *La corrélation entre X et Y (ou coefficient de corrélation linéaire entre X et Y) est le nombre*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}(X)\,\mathbb{V}(Y)}}.$$

On montre que si Y est une fonction affine de X de coefficient directeur positif, c'est-à-dire s'il existe une liaison déterministe exacte entre X et Y de la forme $Y = aX + b$ où a et b sont deux réels tels que $a > 0$, on a :

$$\rho(X, aX + b) = 1.$$

De même si Y est une fonction affine de X de coefficient directeur négatif, c'est-à-dire s'il existe une liaison déterministe exacte entre X et Y de la forme $Y = aX + b$ où a et b sont deux réels tels que $a < 0$, on a :

$$\rho(X, aX + b) = -1.$$

En fait, ces deux cas sont extrêmes (le cas où il existe une liaison déterministe exacte entre X et Y n'est pas fréquent). On a toutefois le résultat général suivant :

Proposition 27. *On a toujours $-1 \leqslant \rho(X,Y) \leqslant 1$. De plus, $\rho(X,Y) = 1$ si et seulement si Y est une fonction affine de X de coefficient directeur positif, et $\rho(X,Y) = -1$ si et seulement si Y est une fonction affine de X de coefficient directeur négatif.*

On peut interpréter le coefficient de corrélation comme un indice de liaison affine entre deux variables. Si ce coefficient est positif, cela signifie que le fait que X soit plutôt plus grand que d'ordinaire a tendance à indiquer que Y est aussi plutôt grand que d'ordinaire. De plus, plus ce coefficient est proche de 1, plus cette tendance est forte. Inversement, si ce coefficient est négatif, cela signifie que le fait que X soit plutôt plus grand que d'ordinaire a tendance à indiquer que Y est plutôt plus petit que d'ordinaire, et plus ce coefficient est proche de -1, plus cette tendance est forte. Si ce coefficient est proche de zéro, cela signifie qu'il n'y a pratiquement pas de liaison affine entre X et Y. Mais ceci n'est qu'une "indication" : seule la loi jointe décrit complètement le lien entre les deux variables. On peut avoir un coefficient de corrélation nul alors que les variables ne sont pas indépendantes. Si le coefficient de corrélation est proche de zéro, cela signifie qu'il n'y a pratiquement pas de liaison affine entre X et Y, mais il peut y avoir une liaison d'un autre type, par exemple de type sinusoïdale, ou parabolique ...

Prenons un exemple très simple : soient ϵ un signe aléatoire $\mathbb{P}(\epsilon = +1) = \mathbb{P}(\epsilon = -1) = 1/2$, et U une variable aléatoire indépendante de ϵ. Dans ce cas, on a $\rho(U, \epsilon U) = 0$ et pourtant U n'est pas indépendante de ϵU.

Proposition 28. *Si X et Y sont indépendantes, alors $Cov(X,Y) = 0$ et $\rho(X,Y) = 0$. Attention, la réciproque est fausse : on peut avoir $Cov(X,Y) = 0$ alors que les variables X et Y ne sont pas indépendantes.*

Pour évaluer le coefficient de corrélation entre X et Y, on recueille des observations indépendantes dont la loi est la loi de (X, Y). Rappelons que les estimateurs de $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$ sont $\overline{X}$, $\overline{Y}$, S_X^2, S_Y^2. Les estimateurs de $\mathrm{Cov}(X,Y)$ et de $\rho(X,Y)$ sont définis comme suit :

Définition. *Soit $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ un n-échantillon dont la loi commune est la loi de (X, Y).*
La covariance empirique de X et de Y, notée $\hat{\gamma}(X, Y)$ est définie par :

$$\hat{\gamma}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \bar{Y})$$

Le coefficient de corrélation empirique de X et de Y, noté $\hat{\rho}(X, Y)$ est défini par :

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}.$$

Remarque. Puisque $\rho(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\,\mathbb{V}(Y)}}$, il n'est pas très surprenant que $\hat{\rho}(X, Y) = \dfrac{\hat{\gamma}(X, Y)}{\sqrt{S_X^2\, S_Y^2}}$.
Nous n'aborderons pas dans ce cours les tests sur la corrélation de deux v.a. à densité.

7.3.2 Indépendance de deux variables qualitatives

Test du chi-deux d'indépendance.

Soit (X, Y) un couple de variables aléatoires prenant ses valeurs dans $\{a_1, \ldots, a_p\} \times \{b_1, \ldots, b_q\}$. On note π_{ij}, p_j, q_j les probabilités théoriques suivantes :

$$\pi_{ij} = \mathbb{P}(X = a_i, Y = b_j), p_j = \mathbb{P}(X = a_i), q_j = \mathbb{P}(Y = b_j).$$

On tire un n-échantillon de ce couple : $(X_1, Y_1), \ldots, (X_n, Y_n)$. On note N_{ij} le nombre de couples prenant la valeur (a_i, b_j), $N_{i.}$ le nombre de couples pour lesquels X prend la valeur a_i et $N_{.j}$ le nombre de couples pour lesquels Y prend la valeur b_j.
Les estimateurs des probabilités π_{ij}, p_j, q_j sont

$$\hat{\pi}_{ij} = \frac{N_{ij}}{n} \quad , \quad \hat{p}_i = \frac{N_{i.}}{n} \quad , \quad \hat{q}_j = \frac{N_{.j}}{n}.$$

Proposition 29. *Quand n est grand, la variable*

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(N_{ij} - n\pi_{ij})^2}{n\pi_{ij}}$$

se comporte pratiquement comme une variable de loi $\chi^2(pq - 1)$.

Sous l'hypothèse $\mathcal{H}_0$: "X et Y sont indépendantes", lorsque n est grand, la variable

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(N_{ij} - np_jq_j)^2}{np_jq_j}$$

se comporte donc pratiquement comme une variable de loi $\chi^2(pq - 1)$. Sous l'hypothèse $\mathcal{H}_1$, la variable Z a tendance à prendre de plus grandes valeurs que sous $\mathcal{H}_0$.

En pratique, on ne connaît ni p_j, ni q_j, la variable Z n'est donc pas utilisable pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$. On remplace les p_j, q_j par leurs estimateurs $\hat{p}_i$, $\hat{q}_j$ et, ce faisant, on perd $p + q - 2$ degrés de liberté.

Proposition 30 (Test du chi-deux d'indépendance)**.** *Lorsque la condition*

$$\frac{N_{i.}N_{.j}}{n} \geqslant 5 \text{ pour tous } i \text{ et } j$$

est vérifiée, sous l'hypothèse $\mathcal{H}_0$: "X et Y sont indépendantes", la variable

$$Z = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(N_{ij} - \frac{N_{i.}N_{.j}}{n})^2}{\frac{N_{i.}N_{.j}}{n}}$$

se comporte pratiquement comme une variable de loi $\chi^2((p-1)(q-1))$. De plus, sous l'hypothèse $\mathcal{H}_1$, la variable Z a tendance à prendre de plus grandes valeurs que sous $\mathcal{H}_0$.

On utilise ce résultat pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$: X et Y ne sont pas indépendantes. La région de rejet du test de $\mathcal{H}_0$ contre $\mathcal{H}_1$ de niveau α est $\mathcal{R} = \{Z_{obs} \geqslant k\}$ avec $\mathbb{P}(Q \geqslant k) = \alpha$, où Q suit la loi $\chi^2((p-1)(q-1))$.

Exemple (Effet de différents traitements sur la fructification du pommier Golden Delicious)**.** La variable "traitement" prend les valeurs A, B, C, D. La variable "nombre de fruits par rameau" est classée en catégories 0 fruit, 1 fruit, 2 fruits et plus.

Y/X	0	1	2 et plus	
A	203	150	6	359
B	266	112	1	379
C	258	126	2	386
D	196	168	17	381
	923	556	26	1505

Il faut regrouper les catégories (1 fruit) et (2 fruits et plus) en une unique catégorie (1 fruit et plus) (donc $p = 2$). On calcule Z_{obs} :

$$\begin{aligned} Z_{obs} &= \frac{(203-220.17)^2}{220.17} + \frac{(156-138.83)^2}{138.83} + \frac{(266-232.44)^2}{232.44} \\ &+ \frac{(113-146.56)^2}{146.56} + \frac{(258-236.73)^2}{236.73} + \frac{(128-149.27)^2}{149.27} \\ &+ \frac{(196-233.66)^2}{233.66} + \frac{(185-147.34)^2}{147.34} \simeq 36.63 \end{aligned}$$

Le nombre de degrés de liberté est $(2-1)\times(4-1) = 3$ donc

- $\mathcal{R} = \{Z_{obs} \geqslant 9.3484\}$ au niveau 2.5%,
- $\mathcal{R} = \{Z_{obs} \geqslant 11.3449\}$ au niveau 1%,
- $\mathcal{R} = \{Z_{obs} \geqslant 16.2662\}$ au niveau 0.1%.

On rejette $\mathcal{H}_0$: la fructification n'est pas indépendante du traitement ($\alpha_{obs} \simeq 6 \times 10^{-8}$).

Test du chi-deux d'homogénéité.

Soit U une variable aléatoire prenant ses valeurs dans $\{a_1, \ldots, a_d\}$, soit V une variable aléatoire prenant aussi ses valeurs dans $\{a_1, \ldots, a_d\}$, et supposons que les v.a. U et V sont indépendantes. On note p_j et q_j les probabilités théoriques suivantes :

$$p_j = \mathbb{P}(U = a_j) \text{ et } q_j = \mathbb{P}(V = a_j).$$

On veut tester l'hypothèse $\mathcal{H}_0$: *les variables U et V ont même loi* contre l'hypothèse $\mathcal{H}_1$: *les variables U et V n'ont pas même loi.*

Par exemple, la v.a. U représente le résultat (mauvais, moyen ou bon) d'un individu ayant reçu un traitement A, la v.a. V représente le résultat (mauvais, moyen ou bon) d'un individu ayant reçu un traitement B. L'indépendance des individus entre eux assure l'indépendance de U et de V. Tester $\mathcal{H}_0$: *les variables U et V ont même loi* contre $\mathcal{H}_1$: *les variables U et V n'ont pas même loi* teste que la loi du résultat de l'individu ne dépend pas du traitement qu'il a reçu (on dit que les deux groupes formés par les traitements A et B sont homogènes). Sous $\mathcal{H}_0$, on a alors indépendance entre le résultat obtenu et le traitement reçu, même si la modélisation est très différente de celle utilisée pour réaliser un test du chi-deux d'indépendance, où pour chaque individu, on aurait défini un couple de variables aléatoires (X, Y) avec X = résultat au test de l'individu et Y = traitement reçu par l'individu. Si le nombre d'individus ayant reçu le traitement A vaut n et le nombre d'individus ayant reçu le traitement B vaut m, la formalisation du test du chi-deux d'homogénéité permet de disposer de deux échantillons indépendants entre eux $U_1, U_2, \ldots, U_n$ et $V_1, V_2, \ldots, V_m$, tandis que la formalisation du test du chi-deux d'indépendance ne nous permet de disposer que d'un unique échantillon de couples $(X_1, Y_1), \ldots, (X_N, Y_N)$ avec $N = n+m$ (le traitement est alors considéré comme la valeur prise par une variable aléatoire). Lorsque l'on veut effectuer des tests du chi-deux avec un logiciel de statistique, il est important d'avoir les idées claires sur la manière de rentrer les données, on ne rentre pas de la même façon les données pour un test du chi-deux d'indépendance et pour un test du chi-deux d'homogénéité.

Décrivons le test du chi-deux d'homogénéité. On tire un n-échantillon $U_1, U_2, \ldots, U_n$ de même loi que U et un m-échantillon de $V_1, V_2, \ldots, V_m$ de même loi que V. On note N_j le nombre de fois où U_i prend la valeur a_j et M_j le nombre de fois où V_i prend la valeur a_j.
Les estimateurs des probabilités p_j, q_j sont

$$\hat{p}_j = \frac{N_j}{n} \text{ et } \hat{q}_j = \frac{M_j}{m}.$$

Si les variables U et V ont même loi, alors $p_j = q_j$. Nous notons alors cette probabilité commune π_i, que l'on peut alors estimer par :

$$\hat{\pi}_j = \frac{N_j + M_j}{n+m}$$

Proposition 31. *Quand n et m sont grands, la variable*

$$\sum_{j=1}^{d} \frac{(N_j - np_j)^2}{np_j} + \sum_{j=1}^{d} \frac{(M_j - mq_j)^2}{mq_j}$$

se comporte pratiquement comme une variable de loi $\chi^2(2(d-1))$.
Sous l'hypothèse $\mathcal{H}_0$: "X et Y ont même loi", lorsque n et m sont grands, la variable

$$\begin{aligned} Z &= \sum_{j=1}^{d} \frac{(N_j - n\pi_j)^2}{n\pi_j} + \sum_{j=1}^{d} \frac{(M_j - m\pi_j)^2}{m\pi_j} \\ &= \sum_{j=1}^{d} n\, \frac{(\frac{N_j}{n} - \pi_j)^2}{\pi_j} + \sum_{j=1}^{d} m\, \frac{(\frac{M_j}{m} - \pi_j)^2}{\pi_j} \end{aligned}$$

se comporte donc pratiquement comme une variable de loi $\chi^2(2(d-1))$. Sous l'hypothèse $\mathcal{H}_1$, la variable Z a tendance à prendre de plus grandes valeurs que sous $\mathcal{H}_0$.

En pratique, on ne connaît pas π_j : la variable Z n'est donc pas utilisable pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$. On remplace alors les π_j par leurs estimateurs $\hat{\pi}_j$; ce faisant, on perd $d-1$ degrés de liberté.

Proposition 32 (Test du chi-deux d'homogénéité)**.** *Sous l'hypothèse $\mathcal{H}_0$:"X et Y ont même loi", si les conditions*

$$\frac{n}{n+m}(N_j + M_j) \geqslant 5 \text{ et } \frac{m}{n+m}(N_j + M_j) \geqslant 5$$

sont vérifiées pour tout j, la variable

$$\begin{aligned} Z &= \sum_{j=1}^{d} \frac{\left(N_j - n(\frac{N_j+M_j}{n+m})\right)^2}{n(\frac{N_j+M_j}{n+m})} + \sum_{j=1}^{d} \frac{\left(M_j - m(\frac{N_j+M_j}{n+m})\right)^2}{m(\frac{N_j+M_j}{n+m})} \\ &= \sum_{j=1}^{d} n\,\frac{\left(\frac{N_j}{n} - \left(\frac{N_j+M_j}{n+m}\right)\right)^2}{\frac{N_j+M_j}{n+m}} + \sum_{j=1}^{d} m\,\frac{\left(\frac{M_j}{m} - \left(\frac{N_j+M_j}{n+m}\right)\right)^2}{\frac{N_j+M_j}{n+m}} \end{aligned}$$

se comporte pratiquement comme une variable de loi $\chi^2(d-1)$.
De plus, sous l'hypothèse $\mathcal{H}_1$, la variable Z a tendance à prendre de plus grandes valeurs que sous $\mathcal{H}_0$.

On utilise ce résultat pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$: *X et Y n'ont pas même loi.* La région de rejet du test de $\mathcal{H}_0$ contre $\mathcal{H}_1$ de niveau α est $\mathcal{R} = \{Z_{obs} \geqslant k\}$ avec $\mathbb{P}(Q \geqslant k) = \alpha$, où Q suit la loi $\chi^2(d-1)$.

Deux tests qui n'en font qu'un.

Tout test du chi-deux d'homogénéité peut être traité comme un test du chi-deux d'indépendance mais la formalisation du problème est différente selon que l'on décide d'appliquer l'un ou l'autre de ces deux tests. Prenons un exemple : des élèves sont séparés en deux groupes I et II pour suivre un cours d'apprentissage de la lecture selon deux méthodes. À la fin du cours, les élèves passent un test dont les résultats sont notés A (très bon), B (bon), C (moyen), D (mauvais). On peut, pour comparer les deux méthodes d'apprentissage, soit tester l'indépendance du résultat au test final et du groupe, soit tester l'homogénéité des deux groupes en ce qui concerne les résultats du test. Si les deux groupes sont homogènes, il n'y a pas de différence entre les deux méthodes d'apprentissage, et si le résultat au test final est indépendant du groupe, il n'y a pas non plus de différence entre les deux méthodes d'apprentissage ! En fait, les deux tests mènent aux mêmes résultats numériques (même région de rejet, même valeur de la statistique de test). Le test du chi-deux d'indépendance est plus général que le test du chi-deux d'homogénéité vu dans ce cours car il permet de comparer plus de deux groupes d'individus (dans l'exemple, s'il existait trois méthodes d'apprentissage de la lecture, testées sur trois groupes d'élèves, on pourrait encore tester l'indépendance entre le résultat au test final et le groupe) contrairement au test du chi-deux d'homogénéité. Il existe toutefois une généralisation, que nous n'abordons pas dans ce cours, du test du chi-deux d'homogénéité à plus de deux populations.

8 Quelques récapitulatifs du polycopié

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi qu'une variable aléatoire de référence X (on dit que $X_1, \ldots, X_n$ forme un n-échantillon de même loi que X).

8.1 Loi de $\Sigma_n = X_1 + \cdots + X_n$ et approximations

Loi de X	Loi de Σ_n	Approximations de la loi de Σ_n
$\mathcal{P}(\lambda)$	$\mathcal{P}(n\lambda)$	
$\mathcal{B}(p)$	$\mathcal{B}(n\ ;\ p)$	a) Si $n \geqslant 50$, $np \geqslant 15$, $n(1-p) \geqslant 15$: $\mathcal{N}(np\ ;\ np(1-p))$. b) Si $n \geqslant 50$, $p \leqslant 0.1$, $np < 15$: $\mathcal{P}(np)$.
$\mathcal{N}(m\ ;\ \sigma^2)$	$\mathcal{N}(nm\ ;\ n\sigma^2)$	

8.2 Estimateurs pour un n-échantillon gaussien

On suppose que $X \sim \mathcal{N}(m\ ;\ \sigma^2)\,;\, \mathbb{E}(X) = m\,;\, \mathbb{V}(X) = \sigma^2$.

	σ connu	σ inconnu
m connu		a) L'estimateur de σ^2 est $V_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - m)^2$. b) Sa loi est donnée par $\dfrac{nV_n^2}{\sigma^2} \sim \chi^2(n)$.
m inconnu	a) L'estimateur de m est $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. b) Sa loi est donnée par $\overline{X}_n \sim \mathcal{N}(m\ ;\ \dfrac{\sigma^2}{n})$. c) En version centrée réduite : $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \sim \mathcal{N}(0\ ;\ 1)$.	a) L'estimateur de m est $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ b) Les estimateurs de σ^2 sont $S_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ ou $\widehat{\sigma_n^2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. c) Leurs lois sont données par $\overline{X}_n \sim \mathcal{N}(m\ ;\ \dfrac{\sigma^2}{n})$ (inutilisable) ; $\dfrac{(n-1)S_n^2}{\sigma^2} = \dfrac{n\widehat{\sigma_n^2}}{\sigma^2} \sim \chi^2(n-1)$. d) La loi centrée réduite : $\dfrac{\sqrt{n}(\overline{X}_n - m)}{\sigma} \sim \mathcal{N}(0\ ;\ 1)$ est inutilisable, on utilise à la place : $\dfrac{\sqrt{n}(\overline{X}_n - m)}{S_n} \sim \mathcal{T}(n-1)$.

8.3 Données quantitatives, données qualitatives

Données	Variables étudiées	Test statistique
Quantitatives	a) $\overline{X}$ pour un échantillon	a) si σ connu, test gaussien d'adéquation à une moyenne théorique si σ inconnu, test de Student d'adéquation à une moyenne théorique
	b) $\overline{X}$ et $\overline{Y}$ pour deux échantillons	b) si σ connu, test gaussien d'homogénéité de deux moyennes si σ inconnu, test de Student d'homogénéité de deux moyennes
	c) pour un échantillon : V^2 ou S^2 selon que la variance est connue ou non	c) test d'adéquation à une variance théorique par une variable de loi du chi-deux (ne pas confondre avec un test du chi-deux)
	d) pour deux échantillons : (V_X^2 et V_Y^2) ou (S_X^2 et S_Y^2) selon que les variances sont connues ou non	d) test d'homogénéité de deux variances par une variable de loi de Fisher
Qualitatives	a) Effectifs par classe $N_1, N_2, \ldots, N_k$ pour k classes et un échantillon	a) test du chi-deux d'adéquation à une loi théorique
	b) Effectifs par classe $N_1, N_2, \ldots, N_k$ $M_1, M_2, \ldots, M_k$ pour k classes et deux échantillons	b) test du chi-deux d'homogénéité de deux lois
	c) Effectifs par classe $N_{11}, N_{12}, \ldots, N_{1q}, \ldots,$ $N_{p1}, N_{p2}, \ldots, N_{pq}$ pour $p \times q$ classes et un échantillon	c) test du chi-deux d'indépendance