Chapitre 3

Probabilités et statistiques

La théorie des probabilités et les méthodes d'analyse statistique constituent des outils indispensables dans toutes les sciences. En chimie, ces outils permettent par exemple d'estimer les incertitudes statistiques lors d'une mesure expérimentale ou de concevoir des méthodes de calcul basées sur un échantillonnage statistique. En outre, ils sont au cœur de la description de la matière par la mécanique quantique et la thermodynamique statistique. Dans ce chapitre, nous adoptons une approche raccourcie de la théorie des probabilités en introduisant directement le concept de variable aléatoire (discrète ou continue), sans passer par le concept d'« univers ». Nous nous intéressons particulièrement aux lois de probabilité continues, à l'estimation de l'espérance et de la variance d'une variable aléatoire sur un échantillon et aux tests d'hypothèse.

3.1 Variable aléatoire

3.1.1 Définition

Nous commençons par définir une variable aléatoire discrète finie, de façon à préparer l'extension à une variable aléatoire continue.

Définition 15. *(Variable aléatoire discrète finie).* On définit une variable aléatoire discrète finie X comme une variable pouvant prendre un nombre N fini de valeurs réelles x dans un ensemble $E = \{x_1, x_2, ..., x_N\}$ suivant une loi de probabilité discrète P déterminée par une fonction de probabilité $p : E \to \mathbb{R}$ de façon que :

1. la fonction p est positive ou nulle : pour tout $x \in E,\ p(x) \geq 0$;
2. la somme des $p(x)$ sur tous les éléments $x \in E$ est égale à 1 :

$$\sum_{x \in E} p(x) = p(x_1) + p(x_2) + \cdots + p(x_N) = 1;$$

3. la probabilité que X prenne une valeur dans un sous-ensemble $S \subset E$ est donnée par la somme des $p(x)$ sur tous les éléments $x \in S$:

$$\mathrm{P}(X \in S) = \sum_{x \in S} p(x).$$

La loi de probabilité discrète P prend des valeurs réelles entre 0 et 1. Formellement, il s'agit d'une fonction qui va de l'ensemble de tous les sous-ensembles de E vers l'intervalle $[0, 1]$.

Par exemple, si le sous-ensemble S ne contient qu'un seul élément, $S = \{x_i\}$, alors $\mathrm{P}(X \in S)$ est la probabilité que X prenne la valeur x_i :

$$\mathrm{P}(X \in S) = \mathrm{P}(X = x_i) = p(x_i).$$

Si le sous-ensemble S contient deux éléments, $S = \{x_i, x_j\}$, alors $\mathrm{P}(X \in S)$ est la probabilité que X prenne la valeur x_i ou x_j :

$$\mathrm{P}(X \in S) = \mathrm{P}(X = x_i \text{ ou } X = x_j) = p(x_i) + p(x_j).$$

Et ainsi de suite.

Remarque : Il est possible d'étendre cette définition à un nombre infini dénombrable de valeurs $x_1, x_2, ...$

Exemple 1 : On lance un dé à 6 faces numérotées de 1 à 6. La variable aléatoire discrète « nombre obtenu » X peut donc prendre les valeurs dans l'ensemble $E = \{1, 2, 3, 4, 5, 6\}$. Si le dé est non truqué, tous les résultats sont équiprobables et on a :

$$p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}.$$

Ceci s'appelle une loi de probabilité discrète uniforme. La probabilité d'obtenir un nombre pair s'obtient par somme des probabilités :

$$\mathrm{P}(X \in \{2, 4, 6\}) = \mathrm{P}(X = 2 \text{ ou } X = 4 \text{ ou } X = 6) = p(2) + p(4) + p(6) = \frac{3}{6}.$$

Exemple 2 : On considère une particule pouvant se trouver dans trois états quantiques, labellés 1, 2 et 3, et d'énergies E_1, E_2, et E_3, respectivement. Appelons X la variable aléatoire discrète « label de l'état où se trouve la particule ». A température fixée T, on montre en thermodynamique statistique que la probabilité que la particule soit dans l'état i dépend de l'énergie E_i de cet état :

$$\mathrm{P}(X = i) = p(i) = \frac{1}{Z}\, e^{-E_i/k_\mathrm{B}T},$$

où k_B est la constante de Boltzmann et Z est un facteur de normalisation pour avoir $\sum_{i=1}^{3} p(i) = 1$. Ici, P s'appelle la loi de probabilité de Boltzmann.

Nous pouvons à présent définir une variable aléatoire continue de façon très similaire.

Définition 16. *(Variable aléatoire continue).* On définit une variable aléatoire continue X comme une variable pouvant prendre une continuité de valeurs réelles $x \in \mathbb{R}$ suivant une loi de probabilité continue P déterminée par une fonction de densité de probabilité $f : \mathbb{R} \to \mathbb{R}$ de façon que :

1. la fonction f est positive ou nulle : pour tout $x \in \mathbb{R}$, $f(x) \geq 0$;
2. la fonction f est intégrable sur $\mathbb{R}$ et son intégrale sur $\mathbb{R}$ est égale à 1 :
$$\int_{-\infty}^{\infty} f(x)dx = 1;$$
3. la probabilité que X prenne une valeur dans un intervalle ou une réunion d'intervalles $I \subset \mathbb{R}$ est donnée par l'intégrale de f sur I :

$$\mathrm{P}(X \in I) = \int_I f(x)dx.$$

La loi de probabilité continue P prend des valeurs réelles entre 0 et 1. Formellement, il s'agit d'une fonction qui va de l'ensemble de tous les intervalles ou réunions d'intervalles de $\mathbb{R}$ vers l'intervalle $[0, 1]$.

Pour un intervalle $I =]a, b] \subset \mathbb{R}$, par exemple, on note habituellement la probabilité :

$$\mathrm{P}(X \in I) = \mathrm{P}(a < X \leq b) = \int_a^b f(x)dx.$$

Cela correspond à l'aire sous la courbe de f entre $x = a$ et $x = b$.

Remarque : Pour une variable aléatoire continue, la probabilité que X prenne exactement une valeur x est nulle : $\mathrm{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$. Pour avoir une probabilité non nulle, il faut obligatoirement considérer la probabilité que la valeur prise par X se trouve dans un intervalle non limité à un point. De fait, pour une variable aléatoire continue, le fait d'inclure ou non les bornes de l'intervalle n'a aucun impact sur la probabilité :

$$\mathrm{P}(a < X < b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a \leq X < b) = \mathrm{P}(a \leq X \leq b).$$

Exemple 1 : Une loi de probabilité continue très importante est la loi normale centrée réduite, notée $\mathcal{N}(0, 1)$, qui a une densité de probabilité $f : \mathbb{R} \to \mathbb{R}$ donnée par :

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

Il s'agit d'une fonction de Gauss ou fonction gaussienne. On peut montrer que f s'intègre bien à 1 :

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}dx = 1.$$

Cette dernière intégrale a été calculée en utilisant la célèbre intégrale de Gauss $\int_{-\infty}^{\infty} e^{-y^2}dy = \sqrt{\pi}$ et en effectuant le changement de variable $y = x/\sqrt{2}$. La courbe représentative de f est tracée sur la figure 3.1. Par exemple, la probabilité $\mathrm{P}(1 < X \leq 2)$ correspond à l'aire sous la courbe entre $x = 1$ et $x = 2$.

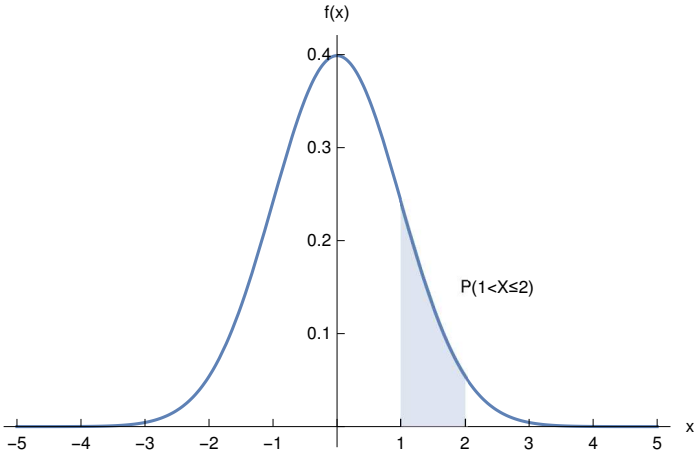

FIGURE 3.1 – Densité de probabilité f de la loi normale centrée réduite $\mathcal{N}(0,1)$. La probabilité $\mathrm{P}(1 < X \leq 2)$ correspond à l'aire sous la courbe entre $x = 1$ et $x = 2$.

Exemple 2 : En mécanique quantique, la position X d'une particule (comme un électron), dans un espace à une dimension pour simplifier, est une variable aléatoire continue. La densité de probabilité associée est $x \mapsto f(x) = |\psi(x)|^2$ où $\psi : \mathbb{R} \to \mathbb{C}$ est la fonction d'onde de l'état dans lequel se trouve la particule. La probabilité de trouver la particule dans l'intervalle $[a, b]$ est donc

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

Dans toute la suite de ce chapitre, nous nous intéressons uniquement au cas de variables aléatoires continues, mais tout ce qui suit peut être étendu sans difficulté au cas des variables aléatoires discrètes.

3.1.2 Fonction de répartition

Définition 17. *(Fonction de répartition).* La fonction de répartition d'une variable aléatoire X est la fonction $F : \mathbb{R} \to \mathbb{R}$ définie par :

$$F(t) = \mathrm{P}(X \leq t).$$

Pour une variable aléatoire continue, cela correspond à intégrer la densité de probabilité de $-\infty$ jusqu'à t :

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

Si f est continue, F est donc une primitive de f, ou inversement f est la dérivée de F, c'est-à-dire $F' = f$.

La fonction de répartition représente les probabilités cumulées de $-\infty$ à t. Elle caractérise entièrement la loi de probabilité.

Théorème 7. *(Propriétés de la fonction de répartition).* La fonction de répartition F d'une variable aléatoire satisfait les propriétés :

— $0 \leq F(t) \leq 1$ pour tout $t \in \mathbb{R}$.

— F est croissante.
— $F(-\infty) = \lim_{t\to-\infty} F(t) = 0 \quad$ et $\quad F(+\infty) = \lim_{t\to+\infty} F(t) = 1.$

L'utilité de la fonction de répartition est qu'elle permet d'exprimer la probabilité que X prenne une valeur dans l'intervalle $]a, b]$ sous la forme :

$$\mathrm{P}(a < X \leq b) = F(b) - F(a).$$

En particulier, on a :

$$\mathrm{P}(X > a) = \mathrm{P}(a < X < +\infty) = F(+\infty) - F(a) = 1 - F(a).$$

Exemple : La fonction de répartition de loi normale centrée réduite $\mathcal{N}(0, 1)$ peut être exprimée sous la forme :

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{t}{\sqrt{2}}\right)\right),$$

où erf : $\mathbb{R} \to \mathbb{R}$ est une fonction spéciale appelée « fonction d'erreur » et définie par :

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-x^2} dx.$$

La fonction de répartition F de la loi normale centrée réduite est représentée sur la figure 3.2.

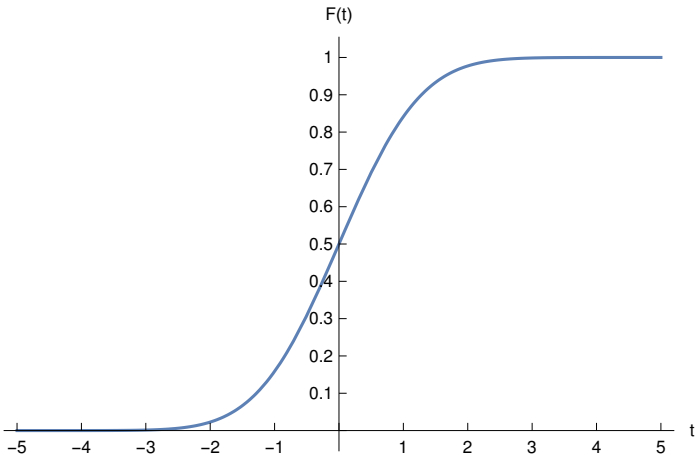

FIGURE 3.2 – Fonction de répartition F de loi normale centrée réduite $\mathcal{N}(0, 1)$.

Remarque : Pour une loi de probabilité continue, la médiane m est définie comme la valeur telle que $F(m) = 1/2$, c'est-à-dire $\mathrm{P}(X \leq m) = 1/2$. En utilisant $F(+\infty) = 1 = \mathrm{P}(X \leq m) + \mathrm{P}(X \geq m)$, on a donc aussi $\mathrm{P}(X \geq m) = 1/2$. La médiane est donc le point séparant la densité de probabilité en deux parties de poids égaux.

3.1.3 Espérance, variance et écart-type

Définition 18. *(Espérance d'une variable aléatoire continue).* Soit une variable aléatoire continue X de densité de probabilité f. On définit l'espérance (ou la moyenne) de X par (si l'intégrale existe) :

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

En pratique, on parle souvent de l'espérance de la loi de probabilité sans faire explicitement référence à une variable aléatoire.

Nous verrons plus tard que l'espérance $\mathrm{E}(X)$ correspond bien à l'interprétation intuitive suivante : si on répète une même expérience aléatoire n fois alors la moyenne des valeurs obtenues pour la variable aléatoire X tend vers $\mathrm{E}(X)$ quand $n \to \infty$.

On peut appliquer une fonction $\varphi : \mathbb{R} \to \mathbb{R}$ à une variable aléatoire X. On obtient une nouvelle variable aléatoire, notée $\varphi(X)$, qui suit une loi de probabilité pouvant être déduite de la loi de probabilité de X :

$$\mathrm{P}(\varphi(X) \in I) = \mathrm{P}(X \in \varphi^{-1}(I)),$$

où $\varphi^{-1}(I)$ désigne l'image réciproque de I par φ, c'est-à-dire l'ensemble des $x \in \mathbb{R}$ tel que $\varphi(x) \in I$. L'espérance de $\varphi(X)$ peut être calculée facilement. Pour une variable aléatoire continue, on a (si l'intégrale existe) :

$$\mathrm{E}(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x) f(x) dx.$$

En choisissant $\varphi(X) = c\,X$ et $\varphi(X) = X + c$ où c est une constante réelle, on peut ainsi obtenir les espérances des variables aléatoires $c\,X$ et de $X + c$.

Théorème 8. *(Espérances de $c\,X$ et de $X + c$).* Soit une variable aléatoire X. Les espérances de $c\,X$ et de $X + c$ où c est une constante réelle sont :

$$\mathrm{E}(c\,X) = c\,\mathrm{E}(X) \quad \text{et} \quad \mathrm{E}(X + c) = \mathrm{E}(X) + c.$$

En choisissant $\varphi(X) = X^2$, on peut aussi définir l'espérance du carré de X. Pour une variable aléatoire continue, on a (si l'intégrale existe) :

$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx.$$

Nous en avons besoin pour définir la variance et l'écart-type de X.

Définition 19. *(Variance et écart-type d'une variable aléatoire).* Soit une variable aléatoire X. La variance de X, si elle existe, est

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2 \right) = \mathrm{E}(X^2) - \mathrm{E}(X)^2,$$

et l'écart-type de X est

$$\sigma(X) = \sqrt{\mathrm{V}(X)}.$$

En pratique, on parle souvent de la variance et de l'écart-type de la loi de probabilité sans faire explicitement référence à une variable aléatoire.

La variance ou l'écart-type nous renseigne sur l'écart moyen des valeurs prises par X par rapport à la moyenne $\mathrm{E}(X)$.

Avec les propriétés de l'espérance, on peut facilement déterminer les variances de $c\,X$ et de $X + c$ où c est une constante réelle.

Théorème 9. *(Variances et écarts-types de $c\,X$ et de $X+c$).* Soit une variable aléatoire X. Les variances de $c\,X$ et de $X + c$ où c est une constante réelle sont :

$$\mathrm{V}(c\,X) = c^2\,\mathrm{V}(X) \quad \text{et} \quad \mathrm{V}(X + c) = \mathrm{V}(X).$$

Les écarts-types correspondants sont donc :

$$\sigma(c\,X) = |c|\,\sigma(X) \quad \text{et} \quad \sigma(X + c) = \sigma(X).$$

Exemple 1 : Pour une variable aléatoire continue X suivant la loi normale centrée réduite $\mathcal{N}(0, 1)$, on trouve que l'espérance est nulle :

$$\mathrm{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-x^2/2}\, dx = 0,$$

puisqu'on intègre une fonction impaire sur un intervalle symétrique autour de $x = 0$. Par ailleurs, on peut calculer $E(X^2)$ par intégration par parties en posant $u(x) = x$ et $v'(x) = x\, e^{-x^2/2}$, et donc $u'(x) = 1$ et $v(x) = -e^{-x^2/2}$:

$$\begin{aligned}
\mathrm{E}(X^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2\, e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, \left(x\, e^{-x^2/2}\right)\, dx \\
&= \frac{1}{\sqrt{2\pi}} \left[-x\, e^{-x^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1.
\end{aligned}$$

Ceci montre que la variance et l'écart-type sont 1 :

$$\mathrm{V}(X) = 1 \quad \text{et} \quad \sigma(X) = 1.$$

Exemple 2 : En mécanique quantique, la position X et la quantité de mouvement (ou impulsion) P d'une particule (dans un espace à une dimension) sont deux variables aléatoires continues. Le célèbre principe d'incertitude de Heisenberg dit que l'on ne peut pas déterminer simultanément X et P avec une précision arbitraire. Il s'agit d'une inégalité qui s'exprime avec les écarts-types de X et P :

$$\sigma(X)\ \sigma(P) \geq \frac{\hbar}{2},$$

où $\hbar$ est la constante de Planck réduite.

3.1.4 Variable aléatoire centrée et réduite

Définition 20. *(Variable aléatoire centrée et réduite).* Soit une variable aléatoire X. On dit que :

— X est centrée si son espérance est nulle : $\mathrm{E}(X) = 0$.

— X est réduite si sa variance est égale à 1 : $\mathrm{V}(X) = 1$.

Nous avons introduit plus haut la loi normale centrée réduite $\mathcal{N}(0,1)$ de densité de probabilité $f : \mathbb{R} \to \mathbb{R}$ donnée par

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

qui a une espérance de 0 et une variance 1. C'est la signification du « (0,1) » dans $\mathcal{N}(0,1)$ et cela explique la qualification « centrée réduite » dans le nom de cette loi.
Il s'agit en fait d'un cas particulier d'une loi de probabilité continue plus générale : la loi normale $\mathcal{N}(\mu, \sigma^2)$ (où $\mu \in \mathbb{R}$ et $\sigma \in]0, +\infty[$ sont deux paramètres) de densité de probabilité $f_{\mu,\sigma} : \mathbb{R} \to \mathbb{R}$ donnée par

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

qui a une espérance de μ et une variance σ^2.

Ces deux lois sont reliées par un simple changement de variables. Si X est une variable aléatoire suivant la loi normale $\mathcal{N}(\mu, \sigma^2)$, alors la variable aléatoire Z définie par

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

suit la loi normale centrée réduite $\mathcal{N}(0,1)$. En effet, la probabilité d'avoir X dans l'intervalle $]x_1, x_2] \subset \mathbb{R}$,

$$\mathrm{P}(x_1 < X \leq x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-(x-\mu)^2/2\sigma^2} dx,$$

peut se réexprimer comme

$$\mathrm{P}\left(\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} \leq \frac{x_2 - \mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-(x-\mu)^2/2\sigma^2} dx,$$

ou, en effectuant le changement de variables $z = (x - \mu)/\sigma$ dans l'intégrale,

$$\mathrm{P}\left(\frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} \leq \frac{x_2 - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1-\mu}{\sigma}}^{\frac{x_2-\mu}{\sigma}} e^{-z^2/2} dz.$$

Finalement, en posant $z_1 = (x_1 - \mu)/\sigma$ et $z_2 = (x_2 - \mu)/\sigma$, on voit que l'on a :

$$\mathrm{P}(z_1 < Z \leq z_2) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-z^2/2} dz,$$

ce qui montre bien que $Z = (X - \mu)/\sigma$ suit la loi normale centrée réduite $\mathcal{N}(0,1)$.

Ceci est général : pour une variable aléatoire suivant n'importe quelle loi de probabilité, on peut toujours se ramener à une variable aléatoire suivant la version centrée et réduite de cette loi.

Définition 21. *(Variable aléatoire centrée et réduite associée).* Soit une variable aléatoire X d'espérance $\mu = \mathrm{E}(X)$ et d'écart-type $\sigma = \sigma(X)$. La variable aléatoire centrée et réduite associée est :

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

On a donc : $\mathrm{E}(Z) = 0$ et $\mathrm{V}(Z) = \sigma(Z) = 1$.

3.1.5 Couple de variables aléatoires

Définition 22. *(Couple de variables aléatoires continues).* On définit un couple de variables aléatoires continues (X, Y) comme une variable pouvant prendre une continuité de valeurs réelles $(x, y) \in \mathbb{R}^2$ suivant une loi de probabilité continue P déterminée par une fonction de densité de probabilité $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ satisfaisant les propriétés habituelles :

1. la fonction $f_{X,Y}$ est positive ou nulle : pour tout $(x, y) \in \mathbb{R}^2$, $f_{X,Y}(x, y) \geq 0$;
2. la fonction $f_{X,Y}$ est intégrable sur $\mathbb{R}^2$ et son intégrale sur $\mathbb{R}^2$ est égale à 1 :
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1;$$
3. la probabilité que (X, Y) prenne une valeur dans un domaine $D \subset \mathbb{R}^2$ est donnée par l'intégrale de $f_{X,Y}$ sur D :
$$\mathrm{P}((X, Y) \in D) = \iint_D f_{X,Y}(x, y) dx dy.$$

Par exemple, pour un domaine donné par le produit de deux intervalles $D =]x_1, x_2] \times]y_1, y_2]$, on a la probabilité :

$$\mathrm{P}((X, Y) \in D) = \mathrm{P}(x_1 < X \leq x_2 \text{ et } y_1 < Y \leq y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X,Y}(x, y) dx dy.$$

La fonction de répartition associée $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ a alors pour expression :

$$F_{X,Y}(t, s) = \mathrm{P}(X \leq t \text{ et } Y \leq s) = \int_{-\infty}^{t} \int_{-\infty}^{s} f_{X,Y}(x, y) dx dy.$$

Définition 23. *(Lois marginales).* Soit (X, Y) un couple de variables aléatoires continues de densité de probabilité $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$. La première variable X est une variable aléatoire continue de densité de probabilité $f_X : \mathbb{R} \to \mathbb{R}$ donnée par

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

De même, la seconde variable Y est une variable aléatoire continue de densité de

probabilité $f_Y : \mathbb{R} \to \mathbb{R}$ donnée par

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

Les lois de probabilité de X et Y sont appelées lois marginales.

En général, la densité de probabilité $f_{X,Y}$ du couple de variables ne peut être exprimée uniquement avec les densités de probabilité f_X et f_Y de chaque variable, sauf si X et Y sont des variables dites indépendantes.

Définition 24. *(Variables aléatoires continues indépendantes).* Soit un couple de variables aléatoires continues (X, Y). Les variables aléatoires X et Y sont dites indépendantes si la densité de probabilité du couple est donnée par le produit des densités de probabilité de chaque variable :

$$\text{pour tout } (x, y) \in \mathbb{R}^2,\ f_{X,Y}(x, y) \ = \ f_X(x)\ f_Y(y).$$

Si X et Y sont indépendantes, la probabilité d'avoir $x_1 < X \leq x_2$ et $y_1 < Y \leq y_2$ est simplement le produit des probabilités :

$$\mathrm{P}(x_1 < X \leq x_2 \text{ et } y_1 < Y \leq y_2) = \mathrm{P}(x_1 < X \leq x_2)\ \mathrm{P}(y_1 < Y \leq y_2).$$

Comme pour le cas d'une seule variable aléatoire, on peut appliquer une fonction $\varphi : \mathbb{R}^2 \to \mathbb{R}$ à un couple de variables aléatoires (X, Y). On obtient une nouvelle variable aléatoire, notée $\varphi(X, Y)$, qui suit une autre loi de probabilité pouvant être déduite de celle de (X, Y). Pour le cas de variables aléatoires continues, l'espérance de $\varphi(X, Y)$ est (si l'intégrale existe) :

$$\mathrm{E}(\varphi(X, Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x, y) f_{X,Y}(x, y)dxdy.$$

Ceci permet en particulier de définir l'espérance de la somme $X + Y$ et du produit $X\,Y$ de variables aléatoires.

Théorème 10. *(Espérance et variance de la somme de deux variables aléatoires).* Pour deux variables aléatoires X et Y, on a :

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) \ \text{ et } \ \mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\,\mathrm{Cov}(X, Y),$$

où $\mathrm{Cov}(X, Y) = \mathrm{E}(X\,Y) - \mathrm{E}(X)\mathrm{E}(Y)$ est la covariance de X et Y.

La covariance de X et Y renseigne sur les corrélations existant entre les deux variables aléatoires. Pour des variables aléatoires indépendantes, elle est nulle et la variance de $X + Y$ est simplement la somme des variances de X et de Y.

Théorème 11. *(Variance de la somme de deux variables aléatoires indépendantes).* Pour deux variables aléatoires X et Y indépendantes, on a $\mathrm{Cov}(X, Y) = 0$ et donc :

$$\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y).$$

Tous ces résultats sur un couple de variables aléatoires se généralisent directement à une suite de n variables aléatoires $(X_1, X_2, ..., X_n)$.

Exemple : Dans l'espace à une dimension $\mathbb{R}$, considérons deux particules quantiques différentes de positions X_1 et X_2 (par exemple, un électron et un proton). Les positions des deux particules constituent un couple de variables aléatoires continues (X_1, X_2). La densité de probabilité associée est $(x, y) \mapsto f_{X_1,X_2}(x, y) = |\psi(x, y)|^2$ où $\psi : \mathbb{R}^2 \to \mathbb{C}$ est la fonction d'onde de l'état dans lequel se trouvent les deux particules. La probabilité de trouver la première particule dans l'intervalle $[a, b]$ et la deuxième particule dans l'intervalle $[c, d]$ est

$$\mathrm{P}(a \leq X_1 \leq b \text{ et } c \leq X_2 \leq d) = \int_a^b \int_c^d f_{X_1,X_2}(x, y)dxdy.$$

La probabilité de trouver la première particule dans l'intervalle $[a, b]$ quel que soit la position de la deuxième particule est

$$\mathrm{P}(a \leq X_1 \leq b) = \int_a^b f_{X_1}(x)dx,$$

où $f_{X_1} : x \mapsto f_{X_1}(x) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x, y)dy$ où est la densité de probabilité marginale de la variable X_1. Dans le cas simple où les particules n'interagissent pas entre elles alors les variables aléatoires X_1 et X_2 sont indépendantes et on a :

$$\mathrm{P}(a \leq X_1 \leq b \text{ et } c \leq X_2 \leq d) = \mathrm{P}(a \leq X_1 \leq b)\ \mathrm{P}(c \leq X_2 \leq d).$$

Dans le cas plus réaliste où les particules interagissent entre elles, les variables aléatoires X_1 et X_2 ne sont pas indépendantes et on a en général :

$$\mathrm{P}(a \leq X_1 \leq b \text{ et } c \leq X_2 \leq d) \neq \mathrm{P}(a \leq X_1 \leq b)\ \mathrm{P}(c \leq X_2 \leq d).$$

Dans ce cas, on dit qu'il y a corrélation entre les particules.

3.1.6 Quelques lois de probabilité continues importantes

Loi normale centrée réduite (ou loi normale standard)
Nous avons déjà discuté de la loi normale centrée réduite (ou loi normale standard) $\mathcal{N}(0, 1)$ de densité de probabilité $f : \mathbb{R} \to \mathbb{R}$ donnée par

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

qui a une espérance de 0 et une variance 1. La loi normale a une importante capitale à cause du théorème central limite que nous verrons plus loin et qui permet en particulier de définir des intervalles de confiance. La fonction de répartition est :

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2}\, dx = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{t}{\sqrt{2}}\right)\right),$$

où $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx$ est la fonction spéciale d'erreur. Rappelons que la fonction de répartition permet de calculer les probabilités (pour a et b réels) :

$$\mathrm{P}(X \leq a) = F(a),$$

$$\mathrm{P}(X > a) = 1 - \mathrm{P}(X \leq a) = 1 - F(a),$$

$$\mathrm{P}(a < X \leq b) = F(b) - F(a).$$

Avec un ordinateur, on peut calculer facilement la valeur de $F(t)$ pour n'importe quel $t \in \mathbb{R}$. Sans ordinateur, on utilise une table de la loi normale où l'on trouve pour plusieurs valeurs de t (disons t_1, t_2, t_3, etc...) les valeurs de $F(t)$ correspondantes ($F(t_1)$, $F(t_2)$, $F(t_3)$, etc...). Si on doit calculer $F(a)$, on cherche donc dans la table la valeur de t la plus proche de a, disons $t_i \approx a$, et on approxime alors $F(a) \approx F(t_i)$.
La table contient uniquement des valeurs de t positives. En effet, grâce au fait que la densité de probabilité de la loi normale est symétrique par rapport à la droite $x = 0$, c'est-à-dire $f(-x) = f(x)$ pour tout $x \in \mathbb{R}$, on peut vérifier que la fonction de répartition de la loi normale a la symétrie suivante :

$$F(-t) = 1 - F(t), \quad \text{pour tout } t \in \mathbb{R}.$$

Si on veut calculer $F(-a)$ pour a positif, on écrira donc $F(-a) = 1 - F(a)$ et on cherchera dans la table la valeur de $F(a)$.
Dans certains problèmes, on se fixe une probabilité $p = F(a)$ et on veut déterminer la valeur de a correspondante. Formellement, a est donné par la fonction inverse de F, c'est-à-dire $a = F^{-1}(p)$. On utilise alors la table en sens inverse, c'est-à-dire que l'on cherche la valeur de $F(t)$ la plus proche de p, disons $F(t_i) \approx p$, et on approxime alors a par la valeur t_i correspondante : $a \approx t_i$.
Enfin, pour traiter le cas d'une variable aléatoire X suivant une loi normale non-centrée et non-réduite $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et d'écart-type σ, on fait le changement de variable $Z = (X - \mu)/\sigma$ où Z suit la loi normale centrée réduite $\mathcal{N}(0, 1)$ (voir section 3.1.4). Par exemple, pour calculer la probabilité $\mathrm{P}(a < X \leq b)$ on peut alors écrire

$$\mathrm{P}(a < X \leq b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right),$$

où F est toujours la fonction de répartition de la loi normale centrée réduite.

Loi du χ^2 (khi-2)
Considérons k (entier strictement positif) variables aléatoires X_1, X_2, ..., X_k indépendantes suivant toutes la loi normale centrée réduit $\mathcal{N}(0, 1)$. Par définition, la variable aléatoire Q définit par la somme de leurs carrés,

$$Q = X_1^2 + X_2^2 + \cdots X_k^2,$$

suit la loi du χ^2 à k degrés de liberté, notée χ_k^2. La loi χ_k^2 est utilisée dans le test d'hypothèse statistique dit « test du χ^2 » que nous verrons plus loin.
Il est possible de déterminer que la densité de probabilité $f_k : \mathbb{R} \to \mathbb{R}$ de la loi χ_k^2 est donnée par

$$f_k(x) = \begin{cases} 0 & \text{si } x < 0 \\ \dfrac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & \text{si } x \geq 0, \end{cases}$$

où $\Gamma :]0, +\infty[\to \mathbb{R}$ est une fonction spéciale appelée « fonction gamma » et définie par :

$$\Gamma(z) = \int_0^{+\infty} y^{z-1} e^{-y} dy.$$

La fonction gamma est la généralisation de la fonction factorielle à n'importe quel nombre réel positif. Dans le cas particulier d'un entier n strictement positif, on a $\Gamma(n) = (n-1)!$. La loi χ^2_k a une espérance k et une variance $2k$. La densité de probabilité f_k est tracée dans la figure 3.3 pour $k = 1$, $k = 4$ et $k = 10$. On voit en effet que, plus k augmente, plus f_k se déplace vers les grandes valeurs de x et s'étale. On peut montrer (avec le théorème central limite) que, dans la limite $k \to \infty$, la loi χ^2_k tend asymptotiquement vers la loi normale d'espérance k et de variance $2k$.

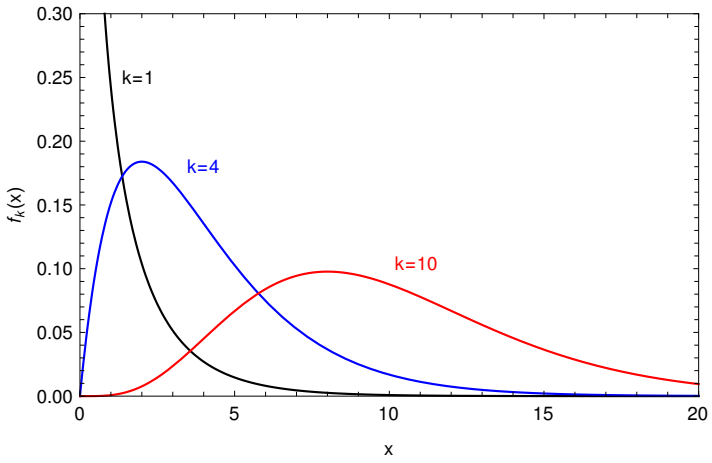

FIGURE 3.3 – Densité de probabilité f_k de la loi χ^2_k pour $k = 1$, $k = 4$ et $k = 10$.

Comme pour la loi normale, il existe des tables donnant la fonction de répartition de loi χ^2_k.

Loi de Student

Considérons une variable aléatoire X suivant la loi normale centrée réduite $\mathcal{N}(0, 1)$ et une variable aléatoire Q, indépendante de X, suivant la loi du χ^2 à k degrés de liberté χ^2_k. Par définition, la variable aléatoire T définit par

$$T = \frac{X}{\sqrt{Q/k}}$$

suit la loi de Student (ou loi t) à k degrés de liberté, notée t_k. La loi t_k est utilisée dans le test d'hypothèse statistique dit « test de Student (ou test t) » que nous verrons plus loin. Il est possible de déterminer que la densité de probabilité $f_k : \mathbb{R} \to \mathbb{R}$ de la loi t_k est donnée par

$$f_k(x) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)} (1 + x^2/k)^{-(k+1)/2}.$$

Dans cette expression, k peut en fait être n'importe quel nombre réel strictement positif. Pour $k > 1$, la loi t_k a une espérance 0, et pour $k > 2$ la loi t_k a une variance $k/(k-2)$. La densité de probabilité f_k est tracée dans la figure 3.4 pour $k = 1$, $k = 3$ et $k = 10$. On vérifie que f_k est symétrique par rapport à $x = 0$ et se stabilise rapidement quand k augmente. On peut montrer que, dans la limite $k \to \infty$, la loi de Student t_k tend asymptotiquement vers la loi normale centrée réduite $\mathcal{N}(0, 1)$.

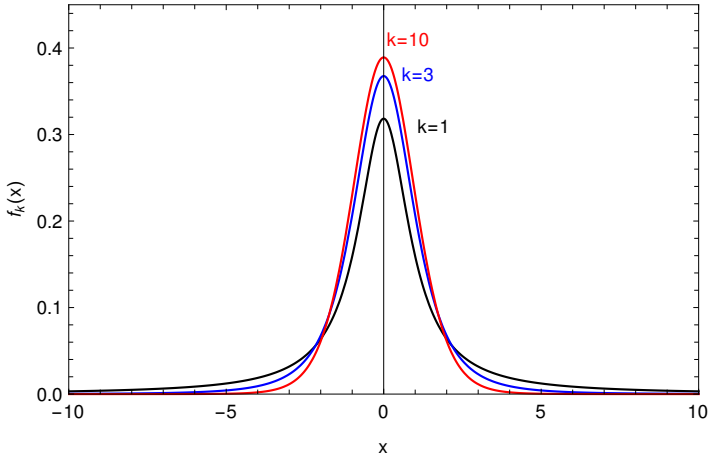

FIGURE 3.4 – Densité de probabilité f_k de la loi de Student t_k pour $k = 1$, $k = 3$ et $k = 10$.

Comme pour la loi normale et la loi du χ^2, il existe des tables donnant la fonction de répartition de loi de Student t_k.

3.2 Théorème central limite et intervalles de confiance

3.2.1 Théorème central limite

Considérons la situation expérimentale commune suivante. On veut mesurer une quantité physique X (par exemple, le volume à l'équivalence lors d'un dosage acido-basique). A cause des erreurs de mesure, si on répète l'expérience n fois, chaque nouvelle mesure de X conduit à des valeurs (légèrement) différentes x_1, x_2, ..., x_n. On peut considérer X comme une variable aléatoire continue pouvant prendre différentes valeurs suivant une loi de probabilité inconnue. La valeur physique que l'on cherche à mesurer peut alors être définie comme l'espérance $\mathrm{E}(X)$ de X. La valeur de l'espérance $\mathrm{E}(X)$ peut être estimée par la valeur moyenne des n valeurs mesurées x_1, x_2, ..., x_n

$$m_n = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

pour n suffisamment grand. Le théorème central limite permet de justifier ce résultat et donne une estimation de l'erreur commise par rapport à $\mathrm{E}(X)$.

Commençons par introduire le concept d'échantillon aléatoire.

Définition 25. *(Échantillon aléatoire).* Soit une variable aléatoire X. Un échantillon aléatoire de X de taille n est une suite de n variables aléatoires $(X_1, X_2, ..., X_n)$ indépendantes deux à deux et suivant chacune la même loi de probabilité que X.

Suivant l'exemple plus haut, ceci correspond à répéter l'expérience n fois, la variable aléatoire X_i correspondant à la $i$ème mesure. Attention à ne pas confondre les variables aléatoires $(X_1, X_2, ..., X_n)$ décrivant un échantillon générique (où les valeurs prises par les variables aléatoires ne sont pas encore déterminées) et la réalisation d'un échantillon particulier où les valeurs des variables aléatoires sont fixées à des valeurs particulières x_1, x_2, ..., x_n.

On peut maintenant énoncer le théorème central limite.

Théorème 12. *(Théorème central limite).* Soit une variable aléatoire X suivant une loi de probabilité quelconque d'espérance $\mu = \mathrm{E}(X)$ et de variance $\sigma^2 = \mathrm{V}(X)$ finies. On considère un échantillon aléatoire $(X_1, X_2, ..., X_n)$ de la variable aléatoire X de taille n. La moyenne des variables aléatoires de l'échantillon

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

est une variable aléatoire qui suit, asymptotiquement dans la limite $n \to \infty$, la loi normale $\mathcal{N}(\mu, \sigma^2/n)$ d'espérance μ et de variance σ^2/n.

Il est important de comprendre que M_n est une variable aléatoire car elle prend différentes valeurs sur différents échantillons. Dans l'exemple plus haut, m_n correspond à la valeur prise par la variable aléatoire M_n sur un échantillon particulier. Le théorème central limite nous dit donc que, pour n suffisamment grand, la variable aléatoire M_n suit une loi normale de même espérance que X, c'est-à-dire $\mathrm{E}(M_n) = \mathrm{E}(X) = \mu$, et dont la variance décroît en $1/n$, c'est-à-dire $\mathrm{V}(M_n) = \sigma^2/n$, ou de manière équivalente dont l'écart-type décroît en $1/\sqrt{n}$, c'est-à-dire $\sigma(M_n) = \sigma/\sqrt{n}$. Pour n suffisamment grand, la densité de probabilité de M_n est donc celle de la loi normale d'espérance μ et d'écart-type $\sigma/\sqrt{n}$

$$f_{\mu,\sigma/\sqrt{n}}(x) = \frac{1}{(\sigma/\sqrt{n})\sqrt{2\pi}} e^{-(x-\mu)^2/2(\sigma/\sqrt{n})^2}.$$

Il s'agit une densité de probabilité centrée sur μ et dont l'écart-type diminue quand n augmente. Cette diminution de l'écart-type est illustrée sur la figure 3.5.

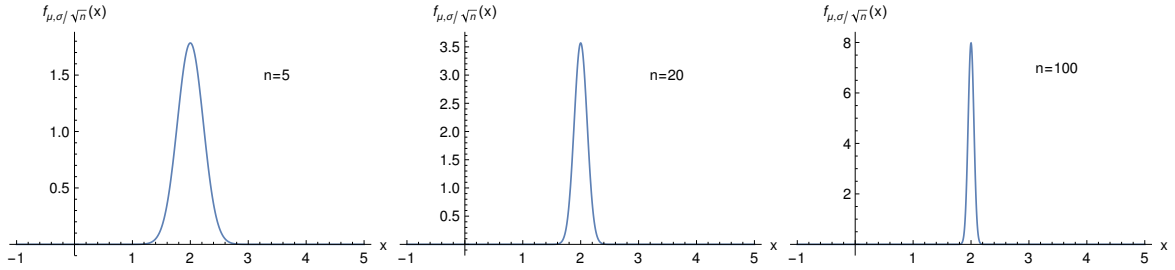

FIGURE 3.5 – Densité de probabilité de la loi normale $f_{\mu,\sigma/\sqrt{n}}$ pour $\mu = 2$, $\sigma = 0.5$, et $n = 5$, $n = 20$, $n = 100$.

On voit donc que, quand n augmente, la variable aléatoire M_n ne peut prendre que des valeurs de plus en plus resserrées autour de la valeur $\mu = \mathrm{E}(X)$ que l'on cherche à déterminer. Dans la limite $n \to \infty$, M_n ne peut prendre que la valeur μ. On dit que M_n est un estimateur statistique de la valeur μ avec des fluctuations décroissant en $1/\sqrt{n}$. Concrètement, cela veut donc dire que la moyenne m_n calculée sur un échantillon particulier (c'est-à-dire la valeur prise par la variable aléatoire M_n sur cet échantillon particulier) est nécessairement une bonne approximation à μ lorsque la taille de l'échantillon n est suffisamment grande. Par ailleurs, nous allons à présent voir que le fait que M_n suive asymptotiquement une loi normale permet d'estimer, pour une taille d'échantillon n fixée suffisamment grande, l'erreur commise par rapport à la valeur exacte μ, sous la forme d'un intervalle de confiance.

Remarque : Nous avons donné le théorème central limite pour la variable aléatoire $M_n = (X_1 + X_2 + \cdots + X_n)/n$ correspondant à la moyenne sur l'échantillon. On peut aussi exprimer le théorème central limite pour la variable aléatoire $\Sigma_n = X_1 + X_2 + \cdots + X_n$ correspondant à la somme des variables aléatoires de l'échantillon. Dans ce cas, Σ_n suit, asymptotiquement dans la limite $n \to \infty$, la loi normale $\mathcal{N}(n\mu, n\sigma^2)$ d'espérance $n\mu$ et de variance $n\sigma^2$.

3.2.2 Intervalles de confiance

Définition 26. *(Intervalle de confiance).* Pour la variable aléatoire M_n représentant la moyenne sur un échantillon de taille n (définie dans le théorème 12) d'espérance μ et d'écart-type $\sigma/\sqrt{n}$, on dit que

$$[M_n - \varepsilon\frac{\sigma}{\sqrt{n}}, M_n + \varepsilon\frac{\sigma}{\sqrt{n}}]$$

est un intervalle de confiance au seuil p (ou, de manière équivalente, au risque $1-p$) si le nombre réel positif ε est tel que

$$\mathrm{P}\left(|M_n - \mu| < \varepsilon\frac{\sigma}{\sqrt{n}}\right) = p.$$

Cela signifie qu'il y a une probabilité p que l'espérance μ (que l'on cherche à déterminer) se trouve dans l'intervalle $[M_n - \varepsilon\sigma/\sqrt{n}, M_n + \varepsilon\sigma/\sqrt{n}]$. Inversement, il y a un risque de probabilité $1-p$ que μ se trouve en dehors de cet intervalle. C'est une façon de donner l'erreur statistique sur l'estimation de l'espérance μ. Notez que l'intervalle de confiance est un intervalle aléatoire puisque M_n est une variable aléatoire prenant différentes valeurs sur différentes réalisations de l'échantillon.

En pratique, on peut soit se donner une valeur ε et calculer la valeur de p correspondante, soit se donner une valeur de p et calculer la valeur ε correspondante. Pour faire ce calcul, il est pratique d'introduire la variable Z_n centrée et réduite

$$Z_n = \frac{M_n - \mu}{\sigma/\sqrt{n}},$$

qui, d'après le théorème central limite, pour n suffisamment grand, suit la loi normale centrée réduite. On peut alors écrire la probabilité $\mathrm{P}(|M_n - \mu| < \varepsilon\sigma/\sqrt{n})$ comme

$$\begin{aligned}\mathrm{P}\left(|M_n - \mu| < \varepsilon\frac{\sigma}{\sqrt{n}}\right) &= P(|Z_n| < \varepsilon)\\ &= P(-\varepsilon < Z_n < \varepsilon)\\ &= F(\varepsilon) - F(-\varepsilon),\end{aligned}$$

où F est la fonction de répartition de la loi normale centrée réduite. Finalement, en utilisant $F(-\varepsilon) = 1 - F(\varepsilon)$ (voir la section 3.1.6), on arrive à une forme pratique de l'équation reliant ε et p

$$2F(\varepsilon) - 1 = p,$$

ou encore

$$F(\varepsilon) = \frac{p+1}{2}.$$

On rappelle que les valeurs de F peuvent être obtenues soit avec un ordinateur soit en consultant une table de la loi normale.

Voici trois exemples importants où l'on fixe la valeur de ε :

— Si on choisit $\varepsilon = 1$, on trouve $p = 2F(1) - 1 \approx 0.68$. L'intervalle à « un sigma » $[M_n - \sigma/\sqrt{n}, M_n + \sigma/\sqrt{n}]$ est donc un intervalle de confiance à environ 68 %.

— Si on choisit $\varepsilon = 2$, on trouve $p = 2F(2) - 1 \approx 0.95$. L'intervalle à « deux sigmas » $[M_n - 2\sigma/\sqrt{n}, M_n + 2\sigma/\sqrt{n}]$ est donc un intervalle de confiance à environ 95 %.

— Si on choisit $\varepsilon = 3$, on trouve $p = 2F(3) - 1 \approx 0.997$. L'intervalle à « trois sigmas » $[M_n - 3\sigma/\sqrt{n}, M_n + 3\sigma/\sqrt{n}]$ est donc un intervalle de confiance à environ 99.7 %.

Inversement, voici des exemples où l'on fixe la valeur de p :

— Si on choisit $p = 0.90$, on trouve $\varepsilon \approx 1.64$. L'intervalle $[M_n - 1.64\,\sigma/\sqrt{n}, M_n + 1.64\,\sigma/\sqrt{n}]$ est donc un intervalle de confiance à 90 %.

— Si on choisit $p = 0.95$, on trouve $\varepsilon \approx 1.96$. L'intervalle $[M_n - 1.96\,\sigma/\sqrt{n}, M_n + 1.96\,\sigma/\sqrt{n}]$ est donc un intervalle de confiance à 95 %.

— Si on choisit $p = 0.99$, on trouve $\varepsilon \approx 2.58$. L'intervalle $[M_n - 2.58\,\sigma/\sqrt{n}, M_n + 2.58\,\sigma/\sqrt{n}]$ est donc un intervalle de confiance à 99 %.

Tout ceci peut être appliqué si on connaît σ. Le problème est que habituellement on ne connaît pas σ puisque c'est l'écart-type de la variable aléatoire initiale X dont la loi de probabilité est inconnue. La solution est d'estimer σ par un estimateur statistique de l'écart-type sur l'échantillon, par exemple [1] :

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

L'estimateur statistique S_n est une variable aléatoire. La variable aléatoire précédemment introduite, $Z_n = (M_n - \mu)/(\sigma/\sqrt{n})$, est alors remplacée par

$$\tilde{Z}_n = \frac{M_n - \mu}{S_n/\sqrt{n}},$$

qui suit encore la loi normale centrée réduite pour n suffisamment grand. L'intervalle de confiance au seuil p devient alors

$$[M_n - \varepsilon\frac{S_n}{\sqrt{n}}, M_n + \varepsilon\frac{S_n}{\sqrt{n}}],$$

avec toujours $2F(\varepsilon) - 1 = p$ où F est la fonction de répartition de la loi normale centrée réduite.

Sur un échantillon particulier, M_n prend une valeur

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

1. La présence du facteur $n - 1$ au lieu de n au dénominateur peut a priori surprendre : il s'agit de l'estimateur de l'écart-type dit non-biaisé. Le facteur $n - 1$ au lieu de n permet de corriger un biais provenant du fait que M_n est lui-même un estimateur de l'espérance μ. Pour n suffisamment grand, on peut bien sûr approcher $n - 1$ par n.

et S_n prend une valeur

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

Cela donnera donc en pratique un intervalle de confiance :

$$[m_n - \varepsilon\frac{s_n}{\sqrt{n}}, m_n + \varepsilon\frac{s_n}{\sqrt{n}}].$$

Sur un graphique, on représente cet intervalle de confiance au seuil p par une barre d'erreur autour de la valeur m_n. Cette barre d'erreur doit être interprétée comme une estimation de l'incertitude statistique sur l'estimation de la valeur μ que l'on cherche à déterminer. On peut considérer que la valeur recherchée μ à une probabilité p de se trouver dans cette barre d'erreur. Souvent les barres d'erreur représentées sur un graphique correspondent aux intervalles de confiance à « un sigma » c'est-à-dire à $p \approx 0.68$ ou à « deux sigmas » c'est-à-dire à $p \approx 0.95$.

Remarque : Pour un échantillon de petite taille (on dit souvent $n < 30$), alors il est plus précis de considérer que la variable aléatoire $\tilde{Z}_n = (M_n - \mu)/(S_n/\sqrt{n})$ suit une loi de Student à $n-1$ degrés de liberté, notée t_{n-1}. On peut donc dans ce cas calculer les intervalles de confiance en utilisant la relation $2F_{n-1}(\varepsilon) - 1 = p$ où F_{n-1} est la fonction de répartition de la loi t_{n-1}. Les différences avec la loi normale soit néanmoins relativement faibles. Voici deux exemples pour un échantillon de petite taille $n = 10$:

- L'intervalle de confiance à « un sigma » ($\varepsilon = 1$) calculé avec la loi de Student t_9 donne un seuil de $p = 2F_9(1) - 1 \approx 0.66$, au lieu de 0.68 avec la loi normale $\mathcal{N}(0, 1)$.
- L'intervalle de confiance au seuil $p = 0.95$ calculé avec la loi de Student t_9 donne $\varepsilon \approx 2.26$, au lieu de 1.96 avec la loi normale $\mathcal{N}(0, 1)$.

La plupart du temps, on utilise donc la loi normale $\mathcal{N}(0, 1)$ pour calculer les intervalles de confiance même dans le cas d'un échantillon de petite taille.

Exemple : On a répété un dosage acido-basique $n = 10$ fois et on a mesuré à chaque fois un volume v à l'équivalence légèrement différent :

v_i (mL)	2.10	1.85	2.00	2.05	1.95	1.90	2.10	2.05	2.00	1.95

La valeur moyenne du volume à l'équivalence sur cet échantillon est

$$\bar{v} = \frac{1}{10}\sum_{i=1}^{10} v_i = 1.995 \text{ mL}.$$

On peut estimer l'écart-type sur le volume à l'équivalence par

$$\sigma \approx s = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(v_i - \bar{v})^2} \approx 0.083 \text{ mL}.$$

On a donc un écart-type sur la valeur moyenne du volume à l'équivalence de $s/\sqrt{10} \approx 0.026$ mL. En utilisant la loi normale $\mathcal{N}(0, 1)$ pour calculer les intervalles de confiance, on trouve par exemple :

- l'intervalle de confiance à « un sigma » :
$$[1.995 - 0.026, 1.995 + 0.026] \approx [1.97, 2.02] \text{ mL};$$
- l'intervalle de confiance au seuil de 90 % :
$$[1.995 - 1.64 \times 0.026, 1.995 + 1.64 \times 0.026] \approx [1.95, 2.04] \text{ mL}.$$

3.3 Tests d'hypothèse statistique

3.3.1 Généralités

Les tests d'hypothèse statistique sont des méthodes des statistiques inférentielles permettant de décider si des données sur un échantillon dont on dispose soutiennent ou pas une hypothèse particulière concernant la loi de probabilité ayant généré cet échantillon.

Un test d'hypothèse statistique teste ce que l'on appelle « l'hypothèse nulle » (notée H_0). Cette hypothèse nulle suppose que les données observées sur un échantillon (ou une quantité calculée avec ces données) sont le résultat d'une loi de probabilité particulière et que les écarts entre les données et cette loi de probabilité sont uniquement dus au hasard, c'est-à-dire aux fluctuations d'échantillonnage. On introduit aussi parfois explicitement « l'hypothèse alternative » (notée H_1) qui est une hypothèse rivale à hypothèse nulle. Le plus souvent, l'hypothèse alternative est simplement la négation de l'hypothèse nulle, auquel cas il n'est pas nécessaire de l'expliciter.

Un test d'hypothèse statistique introduit une « variable aléatoire test » mesurant d'une certaine façon l'écart entre les données et la loi de probabilité ayant supposément générée ces données. Si l'hypothèse nulle est vraie alors la variable aléatoire test doit suivre une certaine loi de probabilité dite « loi de probabilité du test » qui est telle qu'il est peu probable que la variable aléatoire test prenne des valeurs trop éloignées de la valeur 0. On calcule alors la valeur prise par la variable aléatoire test sur l'échantillon particulier considéré et si celle-ci est trop éloignée de la valeur 0 alors on rejette l'hypothèse nulle (et on accepte l'hypothèse alternative). Dans le cas contraire, on ne rejette pas l'hypothèse nulle et on conclut que les données sont compatibles avec la loi de probabilité supposée.

Pour décider si la valeur prise par la variable aléatoire test est trop éloignée ou pas de la valeur 0, on utilise un critère probabiliste. On se donne un risque α (souvent 0.05 ou 0.01) représentant la probabilité de rejeter l'hypothèse nulle alors que celle-ci est vraie. Ceci détermine un écart maximal de la valeur prise par la variable aléatoire test par rapport à la valeur 0 au-delà duquel on rejette l'hypothèse nulle avec le risque α de se tromper.

Ce risque α de rejeter à tort l'hypothèse nulle s'appelle aussi « risque de première espèce ». La règle de décision du test comporte également un deuxième risque implicite, noté β, qui est celui de ne pas rejeter l'hypothèse nulle alors que c'est l'hypothèse alternative qui est vraie. Ce risque β s'appelle le « risque de deuxième espèce ». La probabilité complémentaire du risque de deuxième espèce, $1 - \beta$, définit la « puissance du test » : c'est la probabilité de rejeter correctement l'hypothèse nulle lorsque c'est bien l'hypothèse alternative qui est vraie.

Nous allons voir deux tests d'hypothèse courants : le test de Student et le test du χ^2.

3.3.2 Test de Student

Le test de Student (ou test t) permet de décider si la moyenne d'une distribution statistique obtenue sur un échantillon est compatible avec l'espérance de la loi de probabilité ayant supposément généré cet échantillon.

La situation est proche mais pas identique de celle de la section 3.2. On considère une variable aléatoire X suivant une loi normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 a priori inconnues. On considère un échantillon aléatoire $(X_1, X_2, ..., X_n)$ de cette variable

X de taille n. On rappelle que la moyenne des variables aléatoires de l'échantillon est :

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

On considère alors un échantillon particulier $(x_1, x_2, ..., x_n)$ sur lequel la variable aléatoire M_n prend la valeur

$$m_n = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Dans le cas le plus courant du test de Student dit « bilatéral », on veut tester l'hypothèse nulle suivante :

Hypothèse nulle (H_0) du test de Student bilatéral : la moyenne m_n observée sur cet échantillon est compatible avec le fait que l'espérance de X ait une certaine valeur μ_0, c'est-à-dire $\mu = \mu_0$.

On introduit la variable aléatoire test

$$T_n = \frac{M_n - \mu_0}{S_n/\sqrt{n}},$$

où S_n est toujours l'estimateur statistique de l'écart-type de X suivant

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

Si H_0 est vraie alors on peut montrer que la variable aléatoire T_n suit une loi de Student à $n-1$ degrés de liberté, notée t_{n-1}. Puisque la densité de probabilité $f_{n-1} : x \mapsto f_{n-1}(x)$ de la loi de Student t_{n-1} est centrée en $x = 0$ et décroît rapidement quand x s'écarte de 0 (voir section 3.1.6), la variable aléatoire T_n a une probabilité faible de s'écarter beaucoup de la valeur 0. Plus quantitativement, si on se donne une petite probabilité α (par exemple, $\alpha = 0.05$ ou $\alpha = 0.01$), on peut déterminer avec la loi de Student t_{n-1} la valeur maximale $t_{\max}$ de sorte que la probabilité que la variable aléatoire T_n s'écarte de 0 en valeur absolue d'une valeur supérieure à $t_{\max}$ soit égale à α

$$\mathrm{P}(|T_n| > t_{\max}) = \alpha.$$

En d'autres termes, si H_0 est vraie, alors $|T_n|$ prendra une valeur supérieure à $t_{\max}$ uniquement avec la faible probabilité α. Inversement, toujours si H_0 est vraie, alors $|T_n|$ prendra une valeur inférieure à $t_{\max}$ avec la grande probabilité $1 - \alpha$.

Le calcul de la probabilité $\mathrm{P}(|T_n| > t_{\max})$ donne

$$\begin{aligned}\mathrm{P}(|T_n| > t_{\max}) &= 1 - \mathrm{P}(|T_n| \leq t_{\max}) \\ &= 1 - \left[F_{n-1}(t_{\max}) - F_{n-1}(-t_{\max})\right],\end{aligned}$$

où F_{n-1} est la fonction de répartition de la loi de Student t_{n-1}. En utilisant la symétrie de la densité de probabilité de la loi de Student par rapport à $x = 0$, on a $F_{n-1}(-t_{\max}) = 1 - F_{n-1}(t_{\max})$, ce qui conduit à

$$\mathrm{P}(|T_n| > t_{\max}) = 2 - 2F_{n-1}(t_{\max}).$$

La relation entre $t_{\max}$ et α est donc

$$2 - 2F_{n-1}(t_{\max}) = \alpha,$$

ou encore

$$F_{n-1}(t_{\max}) = 1 - \frac{\alpha}{2}.$$

Le test de Student consiste à calculer la valeur prise par la variable aléatoire T_n sur l'échantillon considéré,

$$t_n = \frac{m_n - \mu_0}{s_n/\sqrt{n}},$$

où s_n est la valeur prise par la variable aléatoire S_n sur l'échantillon, et à comparer la valeur absolue de t_n avec la valeur $t_{\max}$ associée à une petite probabilité α donnée. On a deux possibilités :

1. si $|t_n| > t_{\max}$ alors on rejette l'hypothèse nulle H_0 ;
2. si $|t_n| \leq t_{\max}$ alors on ne rejette pas l'hypothèse nulle H_0.

Dans le premier cas, on rejette H_0 (on décide donc que l'espérance de X n'est pas égale à μ_0, c'est-à-dire $\mu \neq \mu_0$) car on considère que la valeur $|t_n|$ est trop élevée. Cependant, il y a une petite probabilité α de se tromper, c'est-à-dire de rejeter H_0 alors que celle-ci est en fait vraie.

Dans le deuxième cas, on ne rejette pas H_0 (on décide donc que $\mu = \mu_0$) car on considère que la valeur $|t_n|$ est suffisamment basse pour être compatible avec le fait que sa valeur provienne seulement de fluctuations aléatoires dues à l'échantillonnage. Attention : dans ce cas, on ne peut pas être sûr que H_0 soit vraie, on sait juste que les données de l'échantillon dont on dispose ne permettent pas de mettre en doute H_0.

Le test de Student permet donc de prendre une décision mais avec un risque d'erreur. Ce risque est la probabilité α de rejeter à tort H_0. Ce risque α est choisi librement. Plus on choisit un risque α faible, plus la valeur $t_{\max}$ sera élevée et il sera plus difficile de rejeter H_0. Dans le cas extrême où on choisirait un risque nul de se tromper, $\alpha = 0$, on aurait $t_{\max} = +\infty$, on ne rejetterait alors jamais H_0 !

Remarque 1 : Il existe aussi le cas moins courant du test de Student « unilatéral », qui s'utilise si on connaît a priori la direction d'un possible écart entre μ et μ_0. L'hypothèse alternative H_1, qui était de façon implicite « $\mu \neq \mu_0$ » dans le cas du test bilatéral, devient alors :

— « $\mu > \mu_0$ » pour un test unilatéral à droite (le critère de décision devient $\mathrm{P}(T_n > t_{\max}) = \alpha$) ;
— « $\mu < \mu_0$ » pour un test unilatéral à gauche (le critère de décision devient $\mathrm{P}(T_n < t_{\max}) = \alpha$).

Remarque 2 : Pour être exact, le test de Student requière que la variable aléatoire X ayant généré l'échantillon suive une loi normale. Cependant, on applique souvent le test de Student de manière approximative même si X ne suit pas une loi normale. Pour un échantillon de taille n suffisamment grande, ceci est de toute manière justifié par le théorème central limite.

Exemple : Reprenons l'exemple numérique du dosage acido-basique avec un échantillon de taille $n = 10$ de la section 3.2.2. On veut tester par un test de Student bilatéral si la moyenne observée des volumes à l'équivalence sur l'échantillon, $\bar{v} = 1.995$ mL, est compatible avec une espérance du volume à l'équivalence $\mu_0 = 2.05$ mL. On rappelle que l'écart-type sur le volume à l'équivalence a été estimé à $\sigma \approx s \approx 0.083$ mL. On calcule la valeur t du test de Student

$$t = \frac{\bar{v} - \mu_0}{s/\sqrt{n}} = \frac{1.995 - 2.05}{0.083/\sqrt{10}} \approx -2.10.$$

On détermine alors la valeur $t_{\max}$ correspondant à un risque α en utilisant la relation $F_9(t_{\max}) = 1 - \alpha/2$ où F_9 est la fonction de répartition de la loi de Student à 9 degrés de liberté. Par exemple, avec un ordinateur ou une table, on trouve que :

— un risque de $\alpha = 0.10$ (soit 10 %) correspond à $t_{\max} \approx 1.83$;
— un risque de $\alpha = 0.05$ (soit 5 %) correspond à $t_{\max} \approx 2.26$;
— un risque de $\alpha = 0.01$ (soit 1 %) correspond à $t_{\max} \approx 3.25$.

Si on s'autorise un risque de $\alpha = 0.10$, on a $|t| > t_{\max}$: on rejette l'hypothèse nulle H_0, c'est-à-dire que l'on décide que l'espérance μ ne peut pas être égale à $\mu_0 = 2.05$ mL.
Si on s'autorise un risque de $\alpha = 0.05$ ou de $\alpha = 0.01$, on a $|t| < t_{\max}$: on ne rejette pas l'hypothèse nulle H_0, c'est-à-dire que l'on ne peut pas exclure que l'espérance μ soit égale à $\mu_0 = 2.05$ mL.

3.3.3 Test du χ^2

Le test du χ^2 permet de décider si une distribution statistique obtenue sur un échantillon est compatible avec une loi de probabilité donnée. Il est plus clair d'expliquer ce test avec un exemple concret.

Prenons donc l'exemple d'un dé à 6 faces numérotées de 1 à 6. Un lancer de dé a donc $N = 6$ résultats possibles. On lance le dé un grand nombre de fois n. On appelle N_i la variable aléatoire donnant le nombre de fois que l'on obtient le numéro i ($1 \leq i \leq N$). Bien sûr, la somme des N_i est contrainte d'être égale au nombre total de lancers : $\sum_{i=1}^{N} N_i = n$. Pour une série particulière de n lancers (correspondant ici à ce que l'on appelle un échantillon), les variables aléatoires N_1, N_2, ..., N_6 prennent des valeurs particulières appelées n_1, n_2, ..., n_6. Comme d'habitude, on prendra garde à ne pas confondre les variables aléatoires N_1, N_2, ..., N_6 (dont les valeurs ne sont pas encore déterminées) et les valeurs n_1, n_2, ..., n_6 prises par ces variables aléatoires sur un échantillon particulier.

On veut vérifier si le dé est non truqué en utilisant la distribution statistique $(n_1, n_2, ..., n_6)$ obtenue pour un échantillon particulier. Si le dé est non truqué, alors un lancer de dé suit la loi de probabilité discrète uniforme, c'est-à-dire les probabilités d'obtenir les numéros 1, 2, ..., 6 sont toutes égales :

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}.$$

On peut donc traduire le fait que le dé soit non truqué par l'hypothèse nulle suivante :

Hypothèse nulle (H_0) : le dé est non truqué $\Leftrightarrow$ la distribution statistique observée est issue de la loi de probabilité discrète uniforme.

En supposant que H_0 soit vraie, alors on peut calculer la distribution statistique théorique attendue en utilisant la loi de probabilité uniforme : pour n lancers de dé, on s'attend à ce que le nombre de fois qu'on obtient le numéro i soit

$$n_{\text{theo},i} = n\, p_i, \quad \text{pour } 1 \leq i \leq N.$$

Bien sûr, même si H_0 est vraie, on aura en général $N_i \neq n_{\text{theo},i}$ puisque les variables N_i peuvent fluctuer d'un échantillon à l'autre. On veut donc décider si H_0 est vraie, en prenant en compte ces fluctuations d'échantillonnage.

Introduisons alors la variable aléatoire test suivante qui donne une mesure des écarts entre les variables N_i et les valeurs théoriques $n_{\text{theo},i}$:

$$\chi^2 = \sum_{i=1}^{N} \frac{(N_i - n_{\text{theo},i})^2}{n_{\text{theo},i}}.$$

Comme la notation le suggère, si H_0 est vraie, alors la variable aléatoire χ^2 suit une loi du χ^2. En effet, dû au fait que N_i est la somme de n variables aléatoires (prenant les valeurs 0 ou 1 suivant le résultat de chaque lancer de dé), on peut montrer par le théorème central limite que, si H_0 est vraie, la variable aléatoire $(N_i - n_{\text{theo},i})/\sqrt{n_{\text{theo},i}}$ suit asymptotiquement pour grand n une loi normale centrée réduite $\mathcal{N}(0,1)$. Si les variables aléatoires $(N_i - n_{\text{theo},i})/\sqrt{n_{\text{theo},i}}$ étaient indépendantes, la variable aléatoire χ^2 suivrait alors une loi du χ^2 à $k = N$ degrés de liberté (voir section 3.1.6). Seulement ces variables aléatoires ne sont pas indépendantes puisqu'on a la contrainte $\sum_{i=1}^{N} N_i = n$. Dans ces conditions, on peut montrer que la variable aléatoire χ^2 suit alors une loi du χ^2 à $k = N-1$ degrés de liberté, notée χ^2_{N-1}.

Si la variable aléatoire χ^2 suit la loi de probabilité χ^2_{N-1}, alors comme la densité de probabilité associée $f_{N-1} : x \mapsto f_{N-1}(x)$ décroît rapidement quand x augmente, la variable aléatoire χ^2 a une probabilité faible de prendre une valeur très élevée. Plus quantitativement, si on se donne une petite probabilité α (par exemple, $\alpha = 0.05$ ou $\alpha = 0.01$), on peut déterminer avec la loi χ^2_{N-1} la valeur maximale $\chi^2_{\max}$ de sorte que la probabilité que la variable aléatoire χ^2 prenne une valeur supérieure à $\chi^2_{\max}$ soit égale à α

$$\mathrm{P}(\chi^2 > \chi^2_{\max}) = 1 - F_{N-1}(\chi^2_{\max}) = \alpha,$$

où F_{N-1} est la fonction de répartition de la loi χ^2_{N-1}. En d'autres termes, si H_0 est vraie, alors χ^2 prendra une valeur supérieure à $\chi^2_{\max}$ uniquement avec la faible probabilité α. Inversement, toujours si H_0 est vraie, alors χ^2 prendra une valeur inférieure à $\chi^2_{\max}$ avec la grande probabilité $1 - \alpha$.

Le test du χ^2 consiste à calculer la valeur prise par la variable aléatoire χ^2 sur l'échantillon considéré,

$$\chi^2_{\text{calc}} = \sum_{i=1}^{N} \frac{(n_i - n_{\text{theo},i})^2}{n_{\text{theo},i}},$$

et à comparer cette valeur avec la valeur $\chi^2_{\max}$ associée à une petite probabilité α donnée. On a deux possibilités :

1. si $\chi^2_{\text{calc}} > \chi^2_{\max}$ alors on rejette l'hypothèse nulle H_0 ;
2. si $\chi^2_{\text{calc}} \leq \chi^2_{\max}$ alors on ne rejette pas l'hypothèse nulle H_0.

Dans le premier cas, on rejette H_0 (on décide donc que le dé est truqué) car on considère que la valeur χ^2_{calc} est trop élevée. Cependant, il y a une petite probabilité α de se tromper, c'est-à-dire de rejeter H_0 alors que celle-ci est en fait vraie.

Dans le deuxième cas, on ne rejette pas H_0 (on décide donc que le dé est non truqué) car on considère que la valeur χ^2_{calc} est suffisamment basse pour être compatible avec le fait que sa valeur provienne seulement de fluctuations aléatoires dues à l'échantillonnage. Attention : dans ce cas, on ne peut pas être sûr que H_0 soit vraie, on sait juste que les données de l'échantillon dont on dispose ne permettent pas de mettre en doute H_0.

Le test du χ^2 permet donc de prendre une décision mais avec un risque d'erreur. Ce risque est la probabilité α de rejeter à tort H_0. Ce risque α est choisi librement. Plus on choisit un risque α faible, plus la valeur χ^2_{max} sera élevée et il sera plus difficile de rejeter H_0. Dans le cas extrême où on choisirait un risque nul de se tromper, $\alpha = 0$, on aurait $\chi^2_{\text{max}} = +\infty$, on ne rejetterait alors jamais H_0 !

Regardons maintenant un exemple numérique.

Exemple : On considère toujours un dé à 6 faces numérotées de 1 à 6. On effectue $n = 100$ lancers avec ce dé. On appelle toujours n_i le nombre de fois que l'on a obtenu le numéro i. Les nombres n_i sont donnés dans le tableau suivant :

i	1	2	3	4	5	6
n_i	9	14	15	18	16	28

On veut tester l'hypothèse nulle (H_0) suivante : le dé est non truqué, c'est-à-dire la distribution statistique observée est issue de la loi de probabilité discrète uniforme. On commence par calculer les nombres théoriques (qui sont ici tous égaux) $n_{\text{theo},i} = n \times 1/6 \approx 16.666$. On calcule alors la valeur de χ^2 pour l'échantillon considéré :

$$\chi^2_{\text{calc}} = \sum_{i=1}^{6} \frac{(n_i - n_{\text{theo},i})^2}{n_{\text{theo},i}} \approx 11.96.$$

On détermine alors la valeur χ^2_{max} correspondant à un risque α en utilisant la relation $1 - F_5(\chi^2_{\text{max}}) = \alpha$ où F_5 est la fonction de répartition de la loi du χ^2 à 5 degrés de liberté. Par exemple, avec un ordinateur ou une table, on trouve que :

— un risque de $\alpha = 0.05$ (soit 5 %) correspond à $\chi^2_{\text{max}} \approx 11.07$;
— un risque de $\alpha = 0.01$ (soit 1 %) correspond à $\chi^2_{\text{max}} \approx 15.09$.

Si on s'autorise un risque de $\alpha = 0.05$, on a $\chi^2_{\text{calc}} > \chi^2_{\text{max}}$: on rejette l'hypothèse nulle H_0, c'est-à-dire que l'on décide que le dé est truqué (mais avec un risque de 5% de se tromper).
Si on s'autorise un risque de $\alpha = 0.01$, on a $\chi^2_{\text{calc}} < \chi^2_{\text{max}}$: on ne rejette pas l'hypothèse nulle H_0, c'est-à-dire que l'on ne peut pas dire que le dé soit truqué.

Exercice 6. *(Densité électronique).* On considère un système quantique à un électron dans un espace à une dimension. La position X de l'électron est une variable aléatoire. Théoriquement, on pense que la densité de probabilité associée $f : \mathbb{R} \to \mathbb{R}$ pour ce système est donnée par

$$f(x) = e^{-2|x|}.$$

Pour vérifier cela, on mesure la position de l'électron en répétant l'expérience $n = 1000$ fois. On compte le nombre de fois que l'on trouve l'électron dans chacun des 12 intervalles suivants : $I_1 =]-\infty, -1]$, $I_2 =]-1, -0.8]$, $I_3 =]-0.8, -0.6]$, $I_4 =]-0.6, -0.4]$, $I_5 =]-0.4, -0.2]$, $I_6 =]-0.2, 0]$, $I_7 =]0, 0.2]$, $I_8 =]0.2, 0.4]$, $I_9 =]0.4, 0.6]$, $I_{10} =]0.6, 0.8]$, $I_{11} =]0.8, 1]$, $I_{12} =]1, +\infty]$. On appelle n_i le nombre de fois

que l'on trouve électron dans l'intervalle I_i. Les nombres n_i observés sont donnés dans le tableau suivant :

i	1	2	3	4	5	6	7	8	9	10	11	12
n_i	61	35	42	71	118	174	158	115	70	53	23	80

1. Vérifier que la fonction f définie bien une densité de probabilité.
2. En supposant que X suit la loi de densité de probabilité f, calculer les nombres théoriques attendus pour chaque intervalle :
$$n_{\text{theo},i} = n\ \mathrm{P}(X \in I_i) = n \int_{I_i} f(x)dx.$$
3. Utiliser un test du χ^2 avec un risque de 1 % pour tester si la distribution statistique observée est compatible avec la loi de densité de probabilité f.

Corrigé des exercices

Exercice 6

1. Vérifions que la fonction $f : x \mapsto f(x) = e^{-2|x|}$ définit bien une densité de probabilité :
   - La fonctionnelle exponentielle est toujours positive donc f est positive : pour tout $x \in \mathbb{R}$, $f(x) > 0$.
   - f est intégrable sur $\mathbb{R}$ et on a :

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} e^{-2|x|}dx = \int_{-\infty}^{0} e^{+2x}dx + \int_{0}^{\infty} e^{-2x}dx = \left[\frac{e^{+2x}}{2}\right]_{-\infty}^{0} + \left[\frac{e^{-2x}}{-2}\right]_{0}^{\infty} = 1.$$

Donc f définit bien une densité de probabilité.

2. Si la variable aléatoire X suit la loi de densité de probabilité f, alors la probabilité que X prenne des valeurs dans l'intervalle $]a, b]$, où $b \geq a \geq 0$, est donnée par

$$\mathrm{P}(X \in]a, b]) = \int_a^b f(x)dx = \int_a^b e^{-2x}dx = \left[\frac{e^{-2x}}{-2}\right]_a^b = \frac{e^{-2a} - e^{-2b}}{2}.$$

Par ailleurs, du fait que f est symétrique par rapport à $x = 0$, les probabilités sur les intervalles négatifs sont identiques : $\mathrm{P}(X \in]-b, -a]) = \mathrm{P}(X \in]a, b])$. On obtient donc ainsi les probabilités sur les $N = 12$ intervalles :

$$\mathrm{P}(X \in I_1) = \mathrm{P}(X \in I_{12}) \approx 0.0677,$$

$$\mathrm{P}(X \in I_2) = \mathrm{P}(X \in I_{11}) \approx 0.0333,$$

$$\mathrm{P}(X \in I_3) = \mathrm{P}(X \in I_{10}) \approx 0.0496,$$

$$\mathrm{P}(X \in I_4) = \mathrm{P}(X \in I_9) \approx 0.0741,$$

$$\mathrm{P}(X \in I_5) = \mathrm{P}(X \in I_8) \approx 0.1105,$$

$$\mathrm{P}(X \in I_6) = \mathrm{P}(X \in I_7) \approx 0.1648.$$

En multipliant ces probabilités par $n = 1000$, on obtient la distribution statistique théorique $n_{\mathrm{theo},i}$ sur les $N = 12$ intervalles :

i	1	2	3	4	5	6	7	8	9	10	11	12
$n_{\mathrm{theo},i}$	67.67	33.28	49.65	74.07	110.50	164.84	164.84	110.50	74.07	49.65	33.28	67.67

3. Effectuons maintenant le test du χ^2. L'hypothèse nulle (H_0) testée est « la distribution statistique observée n_i est issue de la loi de densité de probabilité f ». On calcule la valeur de χ^2 :

$$\chi^2_{\mathrm{calc}} = \sum_{i=1}^{12} \frac{(n_i - n_{\mathrm{theo},i})^2}{n_{\mathrm{theo},i}} \approx 9.41.$$

On détermine alors la valeur χ^2_{max} correspondant à un risque $\alpha = 0.01$ (1 %) en utilisant la relation $1 - F_{11}(\chi^2_{\mathrm{max}}) = \alpha$ où F_{11} est la fonction de répartition de la loi du χ^2 à $N - 1 = 11$ degrés de liberté. Avec un ordinateur ou une table, on trouve $\chi^2_{\mathrm{max}} \approx 24.73$. On a donc $\chi^2_{\mathrm{calc}} < \chi^2_{\mathrm{max}}$: on ne rejette pas l'hypothèse nulle H_0. On peut donc dire que, au niveau de risque choisi, la distribution statistique observée est bien compatible avec la loi de densité de probabilité f.