Emmanuel Busato
Laboratoire de Physique Corpusculaire
Bureau : 8207
Tél : 04-73-40-72-97
Email : ebusato@cern.ch

version 2

Statistiques

M2 physique des particules

Table des matières

Références bibliographiques

Afin de compléter ce cours, le lecteur pourra consulter les ouvrages suivants :

- **Kendall's Advanced Theory of Statistics** en 3 volumes [1] : le volume 1 (par A. Stuart et K. Ord) traite de la théorie des distributions, le volume 2A (par A. Stuart, K. Ord et S. Arnold) traite de l'inférence fréquentiste et le volume 2B (par A. O'Hagan et J. Forster) de l'inférence bayésienne. C'est l'ouvrage le plus complet et le plus détaillé. C'est aussi celui dont le niveau est le plus élevé et un des plus difficiles à lire.

- **Statistical Methods in Experimental Physics** par W.T. Eadie, D. Drijard, F.E. James, M. Roos, B. Sadoulet [2]. C'est, comme le Kendall, un ouvrage de référence. Un peu plus abordable que celui-ci, il reste tout de même d'un niveau assez élevé. Le lecteur désirant se lancer dans une lecture riche et profonde mais ne disposant que d'un temps limité y trouvera satisfaction.

- **Statistical Data Analysis** par G. Cowan [3]. Il s'agit d'un ouvrage écrit par un physicien pour les physiciens. Il est pédagogique mais beaucoup moins complet que les précédents (il ne décrit quasiment pas l'inférence bayésienne par exemple). Le lecteur désirant juste se familiariser avec quelques-unes des méthodes statistiques les plus courantes pourra certainement trouver son bonheur dans cet ouvrage.

- **Statistics for nuclear and particle physicists** par L. Lyons [4]. Comme le précédent, il s'agit d'un ouvrage écrit par un physicien pour les physiciens. Les notions de base sont très bien expliquées mais il ne va pas très loin dans la description des méthodes complexes qui sont aujourd'hui utilisées en physique.

Introduction

Ce cours a pour objectif de décrire quelques-unes des méthodes les plus utilisées aujourd'hui en physique (des particules notamment) pour l'interprétation des données. Les chapitres 1 à 4 introduisent les notions de base des probabilités. Les chapitres suivant traitent des inférences fréquentiste (ou classique) et bayésienne. L'annexe A décrit succintement les distributions apparaissant dans tous ces chapitres.

Chapitre 1

Notions de base sur la théorie des probabilités

1.1 Variable et processus aléatoire

Une variable aléatoire est une variable dont il est impossible de prédire la valeur. Une variable aléatoire peut être discrète (exemple : valeurs obtenues lors d'un lancé de dé) ou continue (exemple : taille des individus dans une population donnée).

Un processus aléatoire (on parle aussi d'expérience aléatoire ou plus simplement d'expérience) est un processus au cours duquel une ou plusieurs variables aléatoires prennent une valeur (exemple : lancé de dé). L'ensemble des valeurs que peut prendre une variable aléatoire s'appelle l'univers (Ω) ou la population. L'univers peut être fini ou infini. Les valeurs effectivement obtenues lors d'une ou plusieurs expériences forment un échantillon. Un échantillon est donc une partie de l'univers et est toujours de taille finie. Lorsqu'on réalise une expérience, on dit qu'on effectue un tirage aléatoire de l'échantillon dans l'univers (ou la population).

Dans la suite nous noterons $\{X_i\}$ un échantillon (i désigne les éléments de l'échantillon). Si l'expérience consiste à tirer une variable aléatoire une seule fois, alors $\{X_i\}$ est de taille unité. Si l'expérience consiste à tirer plusieurs variables aléatoires une seule fois, alors l'échantillon $\{X_i\}$ a une taille égale au nombre de variables tirées. Si l'expérience consiste à tirer une variable aléatoire plusieurs fois, alors l'échantillon $\{X_i\}$ a une taille égale au nombre de tirages. D'une manière générale, si l'expérience consiste à tirer plusieurs variables aléatoires plusieurs fois, alors l'échantillon $\{X_i\}$ a une taille égale au produit du nombre de variables par le nombre de tirages.

1.2 Événement et partition

On appelle événement une condition logique sur le résultat d'une expérience (c'est-à-dire sur l'échantillon). Un événement peut être vrai ou faux. Un événement sépare donc l'univers en deux sous-ensembles :

- un sous-ensemble qui contient tous les résultats pour lesquels l'événement est vrai (nous l'appelerons A).

- un sous-ensemble, complémentaire du précédent, qui contient tous les résultats pour lesquels l'événement est faux (nous l'appelerons $\overline{A}$).

Par exemple, si on note d la valeur prise par un dé, un événement peut être $d \leq 3$ ou $d = 6$. De même, si on note t la taille d'un individu, un événement peut être $1,60 < t < 1,70$ m.

Par la suite, nous noterons de manière identique l'événement et le sous-ensemble associé. Ainsi, A désignera à la fois un événement et l'ensemble des résultats pour lesquelles cet événement est vrai.

Les notations suivantes seront utilisées :

- $\emptyset$: ensemble vide (événement impossible).
- $A \cup B$: ensemble formé par la réunion de A et B (il contient tous les éléments qui appartiennent à A ou B).
- $A \cap B$: ensemble formé par l'intersection de A et B (il contient tous les éléments qui appartiennent à A et B). Deux événements A et B sont dit incompatibles si $A \cap B = \emptyset$.

Un ensemble d'événements B_i incompatibles ($B_i \cap B_j = \emptyset,\ \forall\ i \neq j$) couvrant tout l'univers ($\cup B_i = \Omega$) s'appelle une partition.

1.3 Axiomes de la théorie des probabilités

À chaque événement A est associé une probabilité $P(A)$. Les axiomes de A.N. Kolmogorov (voir [5] pour le papier original et [6] pour une traduction) définissent la probabilité comme étant une application associant à chaque événement une valeur entre 0 et 1

$$\begin{array}{rccc} P: & \{\text{Événements}\} & \rightarrow & [0,1] \\ & A & \rightarrow & P(A) \end{array}$$

vérifiant :

1. $P(\Omega) = 1$
2. $P(A \cup B) = P(A) + P(B)$ si $A \cap B = \emptyset$ (c'est-à-dire si A et B sont incompatibles).

Il est possible de déduire de cette définition les propriétés suivantes :

- $P(\overline{A}) = 1 - P(A)$
- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) = P(A \cap B) + P(A \cap \overline{B})$ ou, plus généralement, pour un ensemble d'événements B_i formant une partition :

$$P(A) = \sum_i P(A \cap B_i)$$

1.4 Probabilité conditionnelle

Une notion essentielle en théorie des probabilités est celle de probabilité conditionnelle. La probabilité conditionnelle d'un événement A sachant qu'un autre événement B est vrai est :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1.1}$$

La grandeur $P(A|B)$ se lit "probabilité de A sachant B". Calculer des probabilités conditionnelles par rapport à B consiste à redéfinir l'univers $\Omega = B$. La relation 1.1 montre que, comme il se doit, $P(B|B) = 1$.

Il est important de comprendre que $P(A|B)$ et $P(B|A)$ sont des grandeurs fondamentalement distinctes. Cela semble être une banalité mais on observe encore beaucoup de confusion entre les deux. C'est une confusion de ce type qui est à l'origine de la mauvaise interprétation des méthodes d'inférences fréquentiste et bayésienne (nous reviendrons sur ce point dans quelques chapitres lorsque ces méthodes seront décrites).

Exemple 1.1: La vie quotidienne offre de multiples exemples montrant la différence entre $P(A|B)$ et $P(B|A)$. Considérons en guise d'illustration un ensemble d'individus caractérisés par deux attribus : le fait d'être français ou non et le fait d'habiter à Clermont-Ferrand ou non. La probabilité qu'un individu a d'habiter à Clermont-Ferrand sachant qu'il est français n'est pas la même que celle qu'il a d'être français sachant qu'il habite à Clermont-Ferrand.

Deux événements sont indépendants si la probabilité de l'un ne dépend pas de la réalisation de l'autre : $P(A|B) = P(A)$ et $P(B|A) = P(B)$. La relation 1.1 conduit donc à :

$$P(A \cap B) = P(A)P(B) \quad \text{si } A \text{ et } B \text{ indépendants}$$

1.5 Théorème de Bayes

Le théorème de Bayes permet de relier les probabilités conditionnelles $P(A|B)$ et $P(B|A)$. Puisque $P(A \cap B) = P(B \cap A)$, nous déduisons de l'équation 1.1 que

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

et donc

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

Cette relation fondamentale est le théorème de Bayes. Il est souvent pratique de l'écrire de manière différente en utilisant une partition A_i :

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

Exemple 1.2: Trois détenus sont passés au détecteur de mensonges afin de savoir lequel a commis le meurtre dont on les accuse (on est sûr que le meurtier se trouve parmi les trois). Supposons que le détecteur a déclenché une fois alors que les trois ont prétendu être innocents. Quelle est la probabilité

que la personne qui a fait déclencher le détecteur soit le vrai meurtrier ? Quelle est la probabilité que ce soit un innocent qui ait fait déclencher le détecteur ? Pour répondre à ces questions, il faut connaître la réponse du détecteur à un mensonge et à la vérité. Supposons que le détecteur de mensonge déclenche 70% des fois lorsqu'un mensonge est proféré et 3% des fois lorsque la vérité est proférée. Nous avons donc (on notant 1 une réponse positive du détecteur et 0 une réponse négative) :

$$\begin{aligned} P(\text{coupable}|1) &= \frac{P(1|\text{coupable})P(\text{coupable})}{P(1|\text{coupable})P(\text{coupable}) + P(1|\text{innocent})P(\text{innocent})} \\ &= \frac{0,7 \times 1/3}{0,7 \times 1/3 + 0,03 \times 2/3} = 92,1\% \end{aligned}$$

et donc

$$P(\text{innocent}|1) = 1 - P(\text{coupable}|1) = 7,9\%$$

Le même genre de raisonnement peut être fait avec plus de deux événements. Considérons par exemple le cas de trois événements A, B et C. Nous avons :

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

Ceci se démontre de la manière suivante :

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

Un autre relation utile est :

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

En effet, nous avons :

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

1.6 Différentes interprétations des probabilités

La définition axiomatique des probabilités qui a été donnée précédemment est très imprécise quant à la signification même de ce qu'est une probabilité. Il y a deux grandes manières de se représenter une probabilité. La première, dite fréquentiste, définit la probabilité d'un événement comme étant le rapport entre le nombre de fois ou l'événement se produit et le nombre total d'expérience, lorsque ce dernier tend vers l'infini :

$$P(A) = \lim_{N \to \infty} \frac{N(A)}{N}$$

où $N(A)$ est le nombre de fois ou l'événement A se produit. Cette définition suppose que nous répétions la même expérience dans des conditions strictement identique un nombre infini de fois. Dans le cas du

lancé de dé par exemple, la probabilité d'obtenir le chiffre 2 est donnée par la proportion des lancés dans lesquels un 2 est obtenu, les lancés étant tous réalisés dans les mêmes conditions et en nombre infini.

La deuxième, dite bayésienne, définit la probabilité comme un degré de crédibilité. Nous utilisons souvent cette interprétation des probabilités pour parler des événements de la vie courante. Par exemple, on peut se demander quelle est la probabilité qu'il neige la nuit prochaine, de gagner lors du prochain tirage du loto, de se prendre la foudre au moins une fois dans sa vie, etc. Dans aucun de ces cas la définition fréquentiste n'est possible car il est impossible de reproduire la même expérience dans des conditions strictement identiques (la nuit prochaine et le prochain tirage du loto ne se produisent qu'une seule fois, nous n'avons qu'une seule vie, etc.). Cette interprétation des probabilités reflète donc un degré de confiance que l'on a dans l'occurence d'un certain événement. La probabilité calculée dans l'exemple de la section précédente ($P(\text{coupable}|1)$) s'interprète clairement de manière bayésienne et non pas fréquentiste.

L'interprétation bayésienne comporte plus de subjectivité que l'interprétation fréquentiste. Pour cette raison, certaines personnes en viennent à la rejeter totalement et n'acceptent pour seule définition possible que la définition fréquentiste. Cette attitude peut sembler la plus raisonnable au premier abord. Il faut toutefois noter que la définition fréquentiste n'est pas exempte de difficultés car elle suppose que nous puissions répéter la même expérience un nombre infini de fois, ce qui est bien sûr impossible.

1.7 Loi et densité de probabilité

Pour les variables aléatoires discrètes, chaque résultat possible d'une expérience est identifiable à un événement. La probabilité que la variable aléatoire X prennent la valeur x_i correspond à la probabilité que l'événement $X = x_i$ soit vrai :

$$p_i = p(x_i) = P(X = x_i)$$

L'ensemble des p_i forment la loi de probabilité pour la variable X. Dans le cas où le nombre de variable aléatoire est supérieur à un, nous avons :

$$p(x_{1_i}, x_{2_i}, ..., x_{n_i}) = P(X_1 = x_{1_i} \cap X_2 = x_{2_i} \cap ... \cap X_n = x_{n_i})$$

Pour les variables aléatoires continues, la probabilité de prendre une valeur donnée est nulle. On définit donc un événement par une inégalité. Soit par exemple l'événement $x < X < x + \mathrm{d}x$, où $\mathrm{d}x$ est une variation infinitésimale. Nous noterons la probabilité associée à cet événement de la manière suivante :

$$f_X(x)\mathrm{d}x = P(x < X < x + \mathrm{d}x)$$

$f_X(x)$ s'appelle la densité de probabilité. Dans le cas où le nombre de variable aléatoire est supérieur à un, nous avons :

$$f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)\mathrm{d}x_1\mathrm{d}x_2...\mathrm{d}x_n = P(x_1 < X_1 < x_1 + \mathrm{d}x_1 \cap ... \cap x_n < X_n < x_n + \mathrm{d}x_n)$$

$f_{X_1X_2...X_n}(x_1, x_2, ..., x_n)$ s'appelle la densité de probabilité conjointe des variables x_1, x_2, ..., x_n.

Dans le suite, nous utiliserons le terme distribution pour désigner soit une loi de probabilité pour une variable discrète soit une densité de probabilité pour une variable continue.

Plutôt que d'utiliser les distributions il est souvent plus judicieux d'utiliser une autre fonction appelée fonction de répartition. La fonction de répartition d'une variable aléatoire X est, par définition :

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

La probabilité que X se trouve dans l'intervalle $]a, b]$ $(a < b)$ est donc :

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

Si X est une variable aléatoire discrète, alors

$$F_X(x) = \sum_{x_i \leq x} P(X = x_i)$$

Si X est une variable aléatoire continue, alors

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \tag{1.2}$$

La généralisation au cas où le nombre de variable aléatoires est supérieur à un est immédiat. Dans le cas de variables aléatoires continues :

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}...\int_{-\infty}^{x_n} f_{X_1,X_2,...,X_n}(t_1, t_2, ..., t_n)\mathrm{d}t_1\mathrm{d}t_2...\mathrm{d}t_n$$

La relation 1.2 montre que la densité de probabilité est la dérivée de la fonction de répartition :

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

Chapitre 2

Caractérisation des distributions

Dans ce chapitre nous ne donnerons les formules que pour le cas des variables aléatoires continues. Les formules dans le cas des variables discrètes s'obtiennent en remplaçant les intégrales par des sommes :

$$\int f_X(x)\mathrm{d}x \to \sum P_X(x)$$

2.1 Mesures de localisation

L'espérance d'une variable aléatoire est :

$$\mathbb{E}[X] = \int x f_X(x)\mathrm{d}x \tag{2.1}$$

Afin d'alléger les notations, nous noterons parfois $\mathbb{E}[X] = \mu_X$ ou plus simplement $\mathbb{E}[X] = \mu$ lorsqu'il n'y a pas d'ambiguïté sur la variable aléatoire. L'espérance de la somme Y de plusieurs variables aléatoires X_i est égale à la somme des espérances des X_i :

$$\mathbb{E}[Y] = \sum_i \mathbb{E}[X_i]$$

La valeur médiane d'une variable aléatoire, med$[X]$, est donnée par :

$$F_X(\mathrm{med}[X]) = \int_{-\infty}^{\mathrm{med}[X]} f_X(t)\mathrm{d}t = \frac{1}{2}$$

Le mode d'une variable aléatoire est la valeur de la variable pour laquelle la distribution est maximale. On le note mod$[X]$:

$$f_X(\mathrm{mod}[X]) = \max_{x\in\mathbb{R}} f_X(x)$$

Remarques :

- L'espérance et la valeur médiane sont égales si la distribution est symétrique.
- Le mode leur est aussi égal si la distribution est en plus unimodale.
- Dans le cas général d'une distribution asymétrique, mode, espérance et médiane n'ont aucune raison d'être égaux.

2.2 Mesure de dispersion

La variance d'une variable aléatoire est :

$$\text{var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{2.2}$$

L'écart-type est la racine carrée de la variance :

$$\sigma[X] = \sqrt{\text{var}[X]}$$

2.3 Inégalité de Bienaymé-Tchebichef

Soit X une variable aléatoire de variance $\text{var}[X] = \sigma[X]^2$ finie. L'inégalité de Bienaymé-Tchebychev dit que, pour tout $\alpha > 0$,

$$P(|X - \mathbb{E}[X]| \geq \alpha) \leq \frac{\text{var}[X]}{\alpha^2}$$

ou de manière équivalente

$$P(|X - \mathbb{E}[X]| \geq \alpha\sigma[X]) \leq \frac{1}{\alpha^2} \tag{2.3}$$

Cette inégalité se démontre de la manière suivante. Soit D la région telle que $|X - \mathbb{E}[X]| \geq \alpha\sigma[X]$. Nous avons, dans la région D, l'inégalité suivante :

$$\int_{X\in\text{D}} (x - \mathbb{E}[X])^2 f_X(x)\text{d}x \geq (\alpha\sigma[X])^2 \underbrace{\int_{X\in\text{D}} f_X(x)\text{d}x}_{P(|X-\mathbb{E}[X]|\geq\alpha\sigma[X])}$$

Or le membre de gauche de cette inégalité vérifie

$$\int_{X\in\text{D}} (x - \mathbb{E}[X])^2 f_X(x)\text{d}x \leq \int_{X\in\mathbb{R}} (x - \mathbb{E}[X])^2 f_X(x)\text{d}x = \text{var}[X]$$

Ce qui donne au final

$$\text{var}[X] \geq (\alpha\sigma[X])^2 P(|X - \mathbb{E}[X]| \geq \alpha\sigma[X])$$

2.4 Quantiles

Le p-ième q-quantile de la distribution de la variable aléatoire X est la plus petite valeur x telle que

$$F_X(x) \geq \frac{p}{q} \tag{2.4}$$

Dans le cas d'une variable continue, cette inégalité se transforme en égalité. Certains q-quantiles ont des noms spéciaux :

- 100-quantiles : centiles ou percentiles
- 10-quantiles : déciles

- 4-quantiles : quartiles
- 2-quantile : médiane

L'obtention du p-ième q-quantile se fait donc en intégrant la densité de probabilité en partant de $-\infty$ jusqu'à ce que son intégrale soit supérieure ou égale à p/q. L'obtention du $78^{\text{ème}}$ percentile dans le cas d'une variable continue est illustrée sur la figure 2.1.

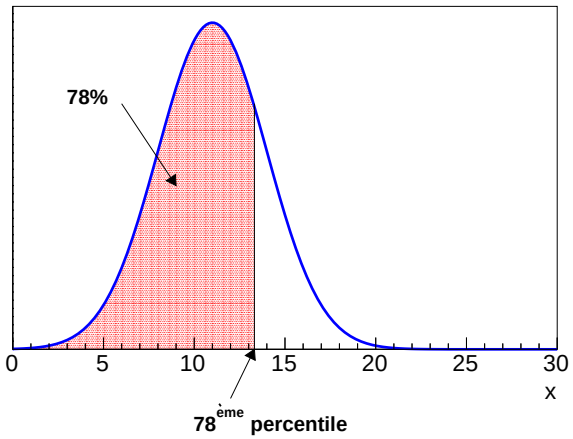


FIGURE 2.1 – Illustration de la relation entre la fonction de répartition et les quantiles.

2.5 Moments d'une variable aléatoire

Le moment d'ordre n d'une variable aléatoire X est :

$$M_n[X] = \mathbb{E}\left[X^n\right]$$

Le moment d'ordre 1 est l'espérance.
Le moment centré d'ordre n d'une variable aléatoire X est :

$$M'_n[X] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^n\right]$$

Le moment centré d'ordre 2 est la variance.
Le moment centré réduit d'ordre n d'une variable aléatoire X est :

$$M''_n[X] = \mathbb{E}\left[\left(\frac{X - \mathbb{E}\left[X\right]}{\sigma\left[X\right]}\right)^n\right] = \frac{M'_n[X]}{\sigma\left[X\right]^n}$$

Le moment centré réduit d'ordre 3 s'appelle le coefficient de dissymétrie (*skewness*) :

$$\gamma_1\left[X\right] = M''_3[X]$$

La valeur de ce coefficient renseigne sur la forme de la distribution :

- $\gamma_1[X] > 0$: la distribution est dissymétrique avec une queue à droite.
- $\gamma_1[X] < 0$: la distribution est dissymétrique avec une queue à gauche.
- $\gamma_1[X] = 0$: la distribution est symétrique.

Le moment centré réduit d'ordre 4 s'appelle le coefficient d'aplatissement (*kurtosis*) :

$$\beta_2[X] = M_4''[X]$$

Plus le coefficient d'aplatissement est élevé plus la distribution est pointue.
Pour des raisons pratiques, le kurtosis normalisé

$$\gamma_2[X] = \beta_2[X] - 3$$

est plus souvent utilisé car sa valeur est nulle pour la distribution gaussienne. Ainsi, une valeur de $\gamma_2[X]$ positive signifie que la distribution est plus pointue qu'une gaussienne alors qu'une valeur négative signifie que la distribution est moins pointue.

2.6 Fonction caractéristique

Soit X une variable aléatoire continue de densité de probabilité $f_X(x)$. Sa fonction caractéristique est, par définition, la transformée de Fourier inverse de la densité de probabilité :

$$\Phi_X(k) = \mathbb{E}\left[e^{ikX}\right] = \int e^{ikx} f_X(x)\mathrm{d}x$$

Il est possible de montrer qu'il y a une correspondance univoque entre densité de probabilité et fonction caractéristique. Deux variables ayant la même fonction caractéristique sont donc distribuées de la même façon.

La fonction caractéristique permet, par dérivation, de calculer les moments :

$$\left.\frac{\mathrm{d}^n \Phi_X(k)}{\mathrm{d}k^n}\right|_{k=0} = i^n \int x^n f_X(x)\mathrm{d}x = i^n M_n[X]$$

Elle permet aussi de démontrer de nombreux théorème plus facilement que si nous avions à les démontrer à partir de la densité de probabilité. Nous rencontrerons quelques exemples par la suite.

2.7 Covariance et fonction de corrélation

2.7.1 Covariance, inégalité de Cauchy-Schwarz

La covariance de deux variables aléatoires est :

$$\mathrm{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

De cette définition, il découle que :

- $\mathrm{cov}(X,X) = \mathrm{var}[X]$

- $\text{cov}(X, Y) = \text{cov}(Y, X)$
- si α et β sont des constantes, alors $\text{cov}(\alpha X, \beta Y) = \alpha\beta\text{cov}(X, Y)$
- si α, β, γ et δ sont des constantes, alors $\text{cov}(\alpha X + \beta Y, \gamma Z + \delta W) = \alpha\gamma\text{cov}(X, Z) + \alpha\delta\text{cov}(X, W) + \beta\gamma\text{cov}(Y, Z) + \beta\delta\text{cov}(Y, W)$

Deux variables X et Y sont décorrélées si $\text{cov}(X, Y) = 0$.
Un résultat important sur la covariance est l'inégalité de Cauchy-Schwarz :

$$\text{cov}(X, Y)^2 \leq \text{var}[X]\,\text{var}[Y]$$

Cette inégalité se démontre de la manière suivante. Soit Z la variable aléatoire définie par

$$Z = X - \frac{\text{cov}(X, Y)}{\text{var}[Y]} Y$$

Il est immédiat de voir que Z et Y sont décorrélés :

$$\text{cov}(Z, Y) = \text{cov}(X, Y) - \frac{\text{cov}(X, Y)}{\text{var}[Y]}\text{cov}(Y, Y) = 0$$

Donc :

$$\begin{aligned}
\text{var}[X] = \text{cov}(X, X) &= \text{cov}\left(Z + \frac{\text{cov}(X, Y)}{\text{var}[Y]} Y, Z + \frac{\text{cov}(X, Y)}{\text{var}[Y]} Y\right) \\
&= \text{var}[Z] + 2\frac{\text{cov}(X, Y)}{\text{var}[Y]} \underbrace{\text{cov}(Y, Z)}_{0} + \frac{\text{cov}(X, Y)^2}{\text{var}[Y]^2}\text{cov}(Y, Y) \\
&= \text{var}[Z] + \frac{\text{cov}(X, Y)^2}{\text{var}[Y]} \geq \frac{\text{cov}(X, Y)^2}{\text{var}[Y]}
\end{aligned}$$

2.7.2 Fonction de corrélation

La fonction de corrélation [1] $\rho(X, Y)$ de deux variables aléatoires X et Y est, par défintion,

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma[X]\,\sigma[Y]}$$

L'inégalité de Cauchy-Schwarz impose que $|\rho(X, Y)| \leq 1$. Trois cas sont particulièrement intéressants :

- $\rho(X, Y) = 1$ lorsque deux variables sont linéairement corrélées (exemple : $y = \alpha x + \beta$, $\alpha > 0$).
- $\rho(X, Y) = -1$ lorsque deux variables sont linéairement anti-corrélées (exemple : $y = \alpha x + \beta$, $\alpha < 0$).
- $\rho(X, Y) = 0$ lorsque deux variables sont non linéairement corrélées.

La figure 2.2 montre quelques distributions conjointes avec leurs corrélations.

1. On l'appelle aussi coefficient de corrélation, facteur de corrélation ou plus simplement corrélation.

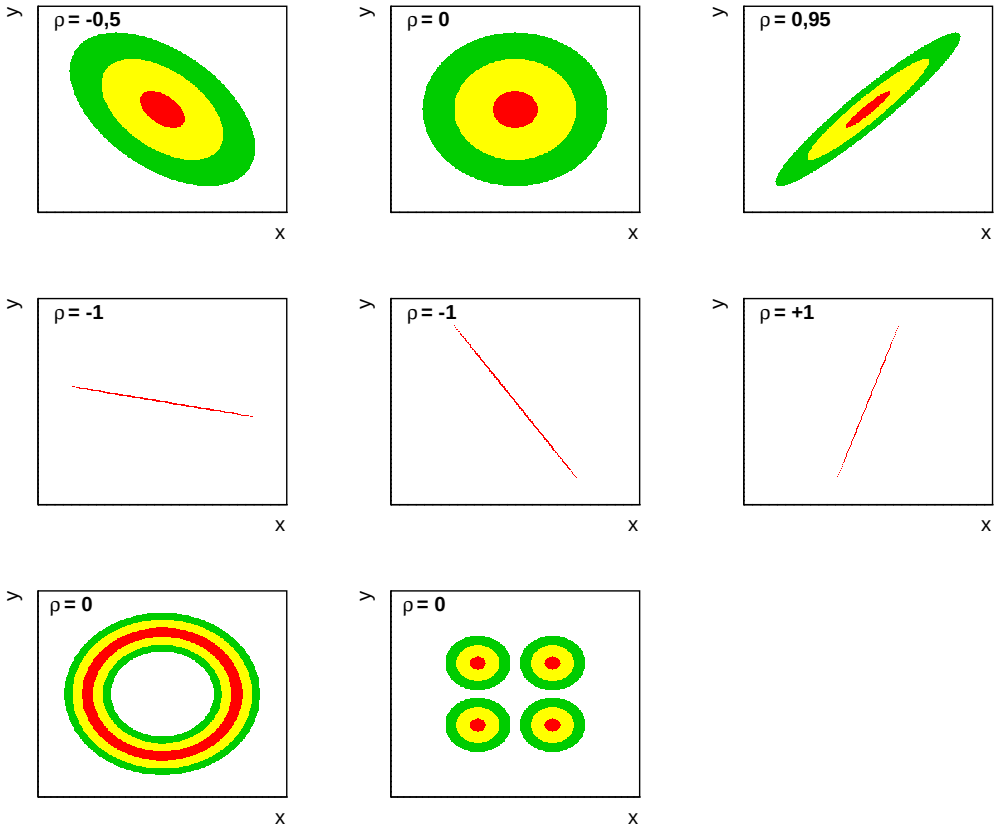


FIGURE 2.2 – Fonction de corrélation pour quelques distributions.

Dans le cas de N variables X_i, on définit la matrice de covariance de la manière suivante :

$$\Sigma = \begin{pmatrix} \sigma[X_1]^2 & \rho(X_1, X_2)\,\sigma[X_1]\,\sigma[X_2] & \dots & \rho(X_1, X_N)\,\sigma[X_1]\,\sigma[X_N] \\ \rho(X_2, X_1)\,\sigma[X_2]\,\sigma[X_1] & \sigma[X_2]^2 & \dots & \rho(X_2, X_N)\,\sigma[X_2]\,\sigma[X_N] \\ \vdots & \vdots & \ddots & \vdots \\ \rho(X_N, X_1)\,\sigma[X_N]\,\sigma[X_1] & \dots & \dots & \sigma[X_N]^2 \end{pmatrix}$$

La matrice de covariance est symétrique ($\rho(X_i, X_j) = \rho(X_j, X_i)$). Pour des variables décorrélées, elle est diagonale.

2.7.3 Décorrélation de variables aléatoires

Lorsque plusieurs variables aléatoires sont corrélées, il peut être pratique de faire un changement de variable de telle sorte à travailler avec des variables décorrélées. Soit X_i un ensemble de n variables aléatoire corrélées de matrice de covariance Σ_X (non diagonale). Décorréler les X_i consister à chercher des fonctions $Y_i(X_1, X_2, \dots, X_n)$ ($i \in [1, n]$) ayant une matrice de covariance Σ_Y diagonale. Décorréler

revient donc à diagonaliser une matrice. Puisque Σ_X est une matrice réelle et symétrique, il est toujours possible de la diagonaliser par une transformation linéaire

$$Y_i = \sum_{j=1}^{n} A_{ij} X_j$$

ou, sous forme matricielle,

$$Y = AX$$

La matrice de transformation A est orthogonale. Ses lignes sont les vecteurs propres (orthonormés) de Σ_X.

Considérons en guise d'illustration le cas bidimensionnel. Σ_X peut s'écrire comme ceci :

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Les valeurs propres de Σ_X s'obtiennent en résolvant l'équation caractéristique $\det(\Sigma_X - \lambda I) = 0$:

$$\det \begin{pmatrix} \sigma_1^2 - \lambda & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \lambda \end{pmatrix} = \lambda^2 - \left(\sigma_1^2 + \sigma_2^2\right)\lambda + \sigma_1^2\sigma_2^2\left(1 - \rho^2\right) = 0$$

Les solutions de cette équation sont (le discriminant est positif)

$$\lambda_\pm = \frac{\sigma_1^2 + \sigma_2^2 \pm \sqrt{\sigma_1^4 + \sigma_2^4 + 2\sigma_1^2\sigma_2^2\left(2\rho^2 - 1\right)}}{2}$$

Notons v_+ et v_- les vecteurs propre de Σ_X. Nous pouvons écrire

$$v_+ = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \text{et} \quad v_- = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$$

et donc

$$A = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

θ est déterminé en résolvant les équations aux valeurs propres $(\Sigma_X - \lambda_\pm I)\, v_\pm = 0$. La première équation est

$$\begin{pmatrix} \sigma_1^2 - \lambda_+ & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \lambda_+ \end{pmatrix} \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

et la deuxième

$$\begin{pmatrix} \sigma_1^2 - \lambda_- & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 - \lambda_- \end{pmatrix} \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Ceci conduit à quatre équations identiques pouvant s'écrire :

$$\tan\theta = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2 + \sqrt{\sigma_1^4 + \sigma_2^4 + 2\sigma_1^2\sigma_2^2\left(2\rho^2 - 1\right)}}$$

Plutôt que cette équation, il est courant de voir dans la littérature l'expression suivante qui est plus simple [2]

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right)$$

2. Nous laissons au lecteur le soin de démontrer que ces deux expressions sont équivalentes (indice : utiliser la relation $\tan 2\theta = \frac{2\tan\theta}{1-\tan^2\theta}$).

2.8 Indépendance et corrélation

Deux variables sont indépendantes si $f_{XY}(x,y) = f_X(x)f_Y(y)$. Deux variables indépendantes sont décorrélees :

$$\mathrm{cov}\,(X,Y) = \iint (x - \mathbb{E}\,[X])\,(y - \mathbb{E}\,[Y])\,f_{XY}(x,y)\mathrm{d}x\mathrm{d}y = \int (x - \mathbb{E}\,[X])\,f_X(x)\mathrm{d}y \int (y - \mathbb{E}\,[Y])\,f_Y(y)\mathrm{d}y = 0$$

Attention, deux variables décorrélées ne sont pas forcément indépendantes ($\mathrm{cov}\,(X,Y) = 0$ n'implique pas forcément que la densité de probabilité puisse se factoriser).

2.9 Distribution marginale et conditionnelle

La densité de probabilité marginale de X est :

$$f_X(x) = \int f_{XY}(x,y)\mathrm{d}y$$

La densité de probabilité conditionnelle de X est :

$$f_X(x|y) = \frac{f_{XY}(x,y)}{\int f_{XY}(x,y)\mathrm{d}x} = \frac{f_{XY}(x,y)}{f_Y(y)}$$

De même, la densité de probabilité conditionnelle de Y est :

$$f_Y(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Des deux expressions précédentes, nous déduisons le théorème de Bayes :

$$f_X(x|y)f_Y(y) = f_Y(y|x)f_X(x)$$

Dans le cas où les deux variables X et Y sont indépendantes, nous avons :

$$f_X(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

Chapitre 3

Fonctions de variables aléatoires

3.1 Changement de variable aléatoire

3.1.1 Cas unidimensionnel

Soit $Y = g(X)$ une fonction de la variable aléatoire X. Si g est une fonction monotone, alors la densité de probabilité de y est donnée par :

$$f_Y(y)\,|\mathrm{d}y| = f_X(x)\,|\mathrm{d}x|$$

Soit

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|$$

Si la fonction g n'est pas monotone, il faut la décomposer en parties monotones et sommer les contributions des différentes parties.

Exemple 3.1: Soit $f(\lambda)$ le spectre en longueur d'onde d'une source de rayon X (un tube à rayons X par exemple). Le spectre en énergie est donné par :

$$g(E) = f(\lambda)\left|\frac{\mathrm{d}\lambda}{\mathrm{d}E}\right| = f(\lambda)\left|\frac{\mathrm{d}}{\mathrm{d}E}\left(\frac{hc}{E}\right)\right| = \frac{hc}{E^2} f\left(\lambda(E)\right)$$

3.1.2 Cas multidimensionnel

Soit $X_1, X_2, ..., X_n$ un ensemble de n variables aléatoires de densité de probabilité conjointe $f_{\{X_i\}}(x_1, x_2, ..., x_n)$ et soit $Y_1(X_1, X_2, ..., X_n), Y_2(X_1, X_2, ..., X_n), ..., Y_n(X_1, X_2, ..., X_n)$ un ensemble de n variables aléatoires fonctions des X_i. La densité de probabilité conjointe des Y_i est donnée par :

$$f_{\{Y_i\}}(y_1, y_2, ..., y_n)\mathrm{d}y_1\mathrm{d}y_2...\mathrm{d}y_n = f_{\{X_i\}}(x_1, x_2, ..., x_n)\mathrm{d}x_1\mathrm{d}x_2...\mathrm{d}x_n$$

Soit,

$$f_{\{Y_i\}}(y_1, y_2, ..., y_n) = f_{\{X_i\}}(x_1, x_2, ..., x_n)\det(J)$$

où $\det(J)$ est le jacobien :

$$\det(J) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

3.2 Somme, produit et rapport de deux variables aléatoires indépendantes

En suivant le même raisonnement que précédemment, il est possible d'établir la densité de probabilité de la somme, du produit et du rapport de deux variables aléatoires indépendantes.

3.2.1 Somme de deux variables aléatoires indépendantes

Soit Y la somme de deux variables aléatoires indépendantes X_1 et X_2 : $Y = X_1 + X_2$. Nous avons :

$$f_Y(y)\mathrm{d}y = \iint f_{X_1}(x_1) f_{X_2}(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

où l'intégrale porte sur l'ensemble des valeurs de x_1 et x_2 tel que $x_1 + x_2 \in [y, y + \mathrm{d}y]$. Pour chaque valeur de x_1, les valeurs de x_2 remplissant cette condition sont dans l'intervalle $[y - x_1, y + \mathrm{d}y - x_1]$. Donc :

$$\begin{aligned} f_Y(y)\mathrm{d}y = \int_{-\infty}^{\infty} \int_{y-x_1}^{y+\mathrm{d}y-x_1} f_{X_1}(x_1) f_{X_2}(x_2)\mathrm{d}x_1\mathrm{d}x_2 &= \int_{-\infty}^{\infty} f_{X_1}(x_1) \left(F_{X_2}(y + \mathrm{d}y - x_1) - F_{X_2}(y - x_1)\right) \mathrm{d}x_1 \\ &= \int f_{X_1}(x_1) f_{X_2}(y - x_1)\mathrm{d}x_1\mathrm{d}y \end{aligned}$$

Finalement :

$$f_Y(y) = \int f_{X_1}(x_1) f_{X_2}(y - x_1)\mathrm{d}x_1$$

La distribution de la somme de deux variables indépendantes s'obtient donc par convolution de Fourier.

3.2.2 Produit de deux variables aléatoires indépendantes

Soit Y le produit de deux variables aléatoires indépendantes X_1 et X_2 : $Y = X_1 X_2$. Nous avons :

$$f_Y(y)\mathrm{d}y = \iint f_{X_1}(x_1) f_{X_2}(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

où l'intégrale porte sur l'ensemble des valeurs de x_1 et x_2 tel que $x_1 x_2 \in [y, y+\mathrm{d}y]$. Pour chaque valeur de x_1, les valeurs de x_2 remplissant cette condition sont dans l'intervalle $[y/x_1, (y+\mathrm{d}y)/x_1]$. Donc :

$$\begin{aligned} f_Y(y)\mathrm{d}y &= \int_{-\infty}^{\infty} \int_{y/x_1}^{(y+\mathrm{d}y)/x_1} f_{X_1}(x_1) f_{X_2}(x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} f_{X_1}(x_1) \left(F_{X_2}\left(\frac{y+\mathrm{d}y}{x_1}\right) - F_{X_2}\left(\frac{y}{x_1}\right)\right) \mathrm{d}x_1 \\ &= \int f_{X_1}(x_1) f_{X_2}(y/x_1) \mathrm{d}x_1 \frac{\mathrm{d}y}{x_1} \end{aligned}$$

Finalement :

$$f_Y(y) = \int f_{X_1}(x_1) f_{X_2}(y/x_1) \frac{\mathrm{d}x_1}{x_1}$$

La distribution du rapport de deux variables indépendantes s'obtient donc par convolution de Mellin.

3.2.3 Rapport de deux variables aléatoires indépendantes

Soit Y le rapport de deux variables aléatoires indépendantes X_1 et X_2 : $Y = X_2/X_1$. Nous avons :

$$f_Y(y)\mathrm{d}y = \iint f_{X_1}(x_1) f_{X_2}(x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

où l'intégrale porte sur l'ensemble des valeurs de x_1 et x_2 tel que $x_2/x_1 \in [y, y+\mathrm{d}y]$. Pour chaque valeur de x_1, les valeurs de x_2 remplissant cette condition sont dans l'intervalle $[x_1 y, x_1(y+\mathrm{d}y)]$. Donc :

$$\begin{aligned} f_Y(y)\mathrm{d}y &= \int_{-\infty}^{\infty} \int_{x_1 y}^{x_1(y+\mathrm{d}y)} f_{X_1}(x_1) f_{X_2}(x_2) \mathrm{d}x_1 \mathrm{d}x_2 \\ &= \int_{-\infty}^{\infty} f_{X_1}(x_1) \left(F_{X_2}(x_1(y+\mathrm{d}y)) - F_{X_2}(x_1 y)\right) \mathrm{d}x_1 \\ &= \int f_{X_1}(x_1) f_{X_2}(x_1 y) \mathrm{d}x_1 (x_1 \mathrm{d}y) \end{aligned}$$

Finalement :

$$f_Y(y) = \int f_{X_1}(x_1) f_{X_2}(x_1 y) x_1 \mathrm{d}x_1$$

3.3 Caractérisation partielle d'une fonction de variable aléatoire

Les calculs précédents ont montré que pour connaître la distribution d'une fonction il est indispensable de connaître les distributions des variables dont elle dépend. Si ces distributions sont inconnues, il est quand même possible de caractériser la distribution de la fonction, mais de manière incomplète. Les calculs suivant montrent comment calculer l'espérance et la variance de la fonction à partir des espérances et variances des variables.

Dans la suite nous appelerons G une variable aléatoire fonction de deux variables aléatoires X et Y. Nous noterons μ_X et μ_Y les espérances de X et Y et $g(x, y)$ la fonction qui donne la valeur de G en fonction de x et y.

3.3.1 Espérance

Pour calculer l'espérance de G, faisons le développement limité de $g(x, y)$ autour des valeurs moyennes de X et Y :

$$g(x,y) \simeq g(\mu_X,\mu_Y) + \left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}(x-\mu_X) + \left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}(y-\mu_Y) + \frac{1}{2}\left[\left.\frac{\partial^2 g}{\partial x^2}\right|_{\mu_X,\mu_Y}(x-\mu_X)^2 + \left.\frac{\partial^2 g}{\partial y^2}\right|_{\mu_X,\mu_Y}(y-\mu_y)^2 + 2\left.\frac{\partial^2 g}{\partial x \partial y}\right|_{\mu_X,\mu_Y}(x-\mu_X)(y-\mu_Y)\right]$$

Donc

$$\mathbb{E}[G] \simeq g(\mu_X,\mu_Y) + \frac{1}{2}\left[\left.\frac{\partial^2 g}{\partial x^2}\right|_{\mu_X,\mu_Y}\mathrm{var}[X] + \left.\frac{\partial^2 g}{\partial y^2}\right|_{\mu_X,\mu_Y}\mathrm{var}[Y] + 2\left.\frac{\partial^2 g}{\partial x \partial y}\right|_{\mu_X,\mu_Y}\mathrm{cov}(X,Y)\right]$$

Au premier ordre, nous avons $\mathbb{E}[G] \simeq g(\mu_X, \mu_Y)$.

3.3.2 Variance, formule de propagation des incertitudes

La variance de G est donnée par :

$$\mathrm{var}[G] = \mathbb{E}\left[G^2\right] - \mathbb{E}[G]^2 \tag{3.1}$$

Nous avons :

$$g(x,y) \simeq g(\mu_X,\mu_Y) + \left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}(x-\mu_X) + \left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}(y-\mu_Y)$$

Les deux termes apparaissant dans le membre de droite de 3.1 sont donc

$$\mathbb{E}\left[G^2\right] \simeq g(\mu_X,\mu_Y)^2 + \left(\left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}\right)^2 \mathrm{var}[X] + \left(\left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}\right)^2 \mathrm{var}[Y] + 2\left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}\left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}\mathrm{cov}(X,Y)$$

et

$$\mathbb{E}[G]^2 \simeq g(\mu_X,\mu_Y)^2$$

La variance est donc

$$\mathrm{var}[G] \simeq \left(\left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}\right)^2 \mathrm{var}[X] + \left(\left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}\right)^2 \mathrm{var}[Y] + 2\left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y}\left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y}\mathrm{cov}(X,Y)$$

Cette relation se généralise sans difficulté au cas ou G dépend de n variables aléatoires X_i ($i \in [1, n]$) :

$$\mathrm{var}[G] \simeq \sum_{i=1}^{n}\left(\left.\frac{\partial g}{\partial x_i}\right|_{\mu_{X_i}}\right)^2 \mathrm{var}[X_i] + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left.\frac{\partial g}{\partial x_i}\right|_{\mu_{X_i}}\left.\frac{\partial g}{\partial x_j}\right|_{\mu_{X_i}}\mathrm{cov}(X_i,X_j) \tag{3.2}$$

ou, de manière équivalente :

$$\mathrm{var}\left[G\right] \simeq \sum_{i,j=1}^{n} \left.\frac{\partial g}{\partial x_i}\right|_{\mu_{X_i}} \left.\frac{\partial g}{\partial x_j}\right|_{\mu_{X_i}} \mathrm{cov}\left(X_i, X_j\right) \tag{3.3}$$

Ces deux dernières équations portent le nom de "formule de propagation des incertitudes" car elles permettent de propager les incertitudes (variances) des variables X_i à la variable G qui en dépend.

Cas particuliers :

- $\rho\left(X,Y\right) = 1 : \mathrm{cov}\left(X,Y\right) = \sigma\left[X\right]\sigma\left[Y\right] \Rightarrow \sigma\left[G\right] = \left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y} \sigma\left[X\right] + \left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y} \sigma\left[Y\right]$
- $\rho\left(X,Y\right) = -1 : \mathrm{cov}\left(X,Y\right) = -\sigma\left[X\right]\sigma\left[Y\right] \Rightarrow \sigma\left[G\right] = \left|\left.\frac{\partial g}{\partial x}\right|_{\mu_X,\mu_Y} \sigma\left[X\right] - \left.\frac{\partial g}{\partial y}\right|_{\mu_X,\mu_Y} \sigma\left[Y\right]\right|$
- Produit de variables décorrélées : $G = \frac{X\times Y}{Z} \Rightarrow \mathrm{var}\left[G\right] = \left(\frac{y}{z}\right)^2 \mathrm{var}\left[X\right] + \left(\frac{x}{z}\right)^2 \mathrm{var}\left[Y\right] + \left(\frac{x\times y}{z^2}\right)^2 \mathrm{var}\left[Z\right]$. Donc :
$$\frac{\mathrm{var}\left[G\right]}{g^2} = \frac{\mathrm{var}\left[X\right]}{x^2} + \frac{\mathrm{var}\left[Y\right]}{y^2} + \frac{\mathrm{var}\left[Z\right]}{z^2}$$
- Somme de variables décorrélées : $G = X + Y \Rightarrow \mathrm{var}\left[G\right] = \mathrm{var}\left[X\right] + \mathrm{var}\left[Y\right]$
- Somme de variables corrélées à 100% : $G = X + Y \Rightarrow \sigma\left[G\right] = \sigma\left[X\right] + \sigma\left[Y\right]$
- Somme de variables anti-corrélées à 100% : $G = X + Y \Rightarrow \sigma\left[G\right] = \sigma\left[X\right] - \sigma\left[Y\right]$

Exemple 3.2: Supposons que nous mesurions la masse d'une particule par la mesure de son énergie (avec un calorimètre) et de son impulsion (avec un trajectographe). Nous avons $m = \sqrt{E^2 - p^2}$, soit :

$$\begin{aligned}\mathrm{var}\left[M\right] &= \left(\frac{E}{\sqrt{E^2-p^2}}\right)^2 \mathrm{var}\left[E\right] + \left(\frac{-p}{\sqrt{E^2-p^2}}\right)^2 \mathrm{var}\left[p\right] \\ &= \frac{E^2}{m^2}\mathrm{var}\left[E\right] + \frac{p^2}{m^2}\mathrm{var}\left[p\right] = \gamma^2\mathrm{var}\left[E\right] + \left(\beta\gamma\right)^2 \mathrm{var}\left[p\right]\end{aligned}$$

3.4 Distribution composée

Soit X une variable aléatoire décrite par la distribution $f_X(x;\alpha)$, où α désigne un paramètre de la distribution (par exemple la moyenne d'une gaussienne). Supposons que α soit lui-même fonction d'une variable aléatoire Y : $\alpha(Y)$. Dans ce cas, X n'est pas distribué suivant $f_X(x;\alpha)$ mais suivant une distribution dite composée, puisqu'elle tient compte des fluctuations de α.

Puisque α est une variable aléatoire, nous pouvons utiliser les notations des probabilités conditionnelles : $f_X(x;\alpha) = f_X(x|y)$. Notons $f_Y(y;\beta)$ la distribution de Y (β désigne les paramètres de cette distribution). La distribution composée de X s'obtient par marginalisation :

$$f_X(x;\beta) = \int f_X(x|y) f_Y(y;\beta) \mathrm{d}y$$

Exemple 3.3: Soit une variable K distribuée suivant une loi de Poisson de paramètre λ

$$p(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Supposons que λ soit distribué suivant une distribution Gamma de paramètres a et b

$$f(\lambda;a,b) = \frac{a(a\lambda)^{b-1}e^{-a\lambda}}{\Gamma(b)}$$

La distribution composée de K est

$$\begin{aligned} p(k;a,b) &= \int_0^\infty p(k;\lambda)f(\lambda;a,b)\mathrm{d}\lambda \\ &= \int_0^\infty \frac{\lambda^k}{k!}e^{-\lambda} \times \frac{a(a\lambda)^{b-1}e^{-a\lambda}}{\Gamma(b)}\mathrm{d}\lambda \\ &= \frac{a^b}{k!\Gamma(b)}\int_0^\infty \lambda^{k+b-1}e^{-(a+1)\lambda}\mathrm{d}\lambda \\ &= \frac{a^b}{k!\Gamma(b)(a+1)^{k+b}}\underbrace{\int_0^\infty \lambda^{k+b-1}e^{-\lambda}\mathrm{d}\lambda}_{\Gamma(k+b)} \end{aligned}$$

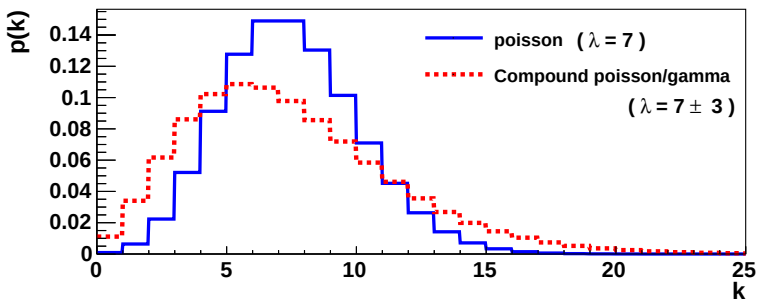
Dans le cas où b est un entier positif, les fonctions Gamma peuvent s'écrire comme des factorielles

$$p(k;a,b) = \frac{a^b}{(a+1)^{k+b}}\frac{(k+b-1)!}{k!(b-1)!} = \binom{k+b-1}{k}\left(\frac{a}{a+1}\right)^b\left(\frac{1}{a+1}\right)^k$$

Posons $k' = k+b$ et $p = a/(a+1)$

$$p(k;a,b) = \binom{k'-1}{k'-b}p^b\,(1-p)^{k'-b} = \binom{k'-1}{b-1}p^b\,(1-p)^{k'-b}$$

Pour la dernière égalité nous avons utilisé $\binom{a}{b} = \binom{a}{a-b}$. Nous reconnaissons la distribution binomiale négative (voir A.7). La figure ci-dessous montre la distribution de Poisson pour $\lambda = 7$ et la distribution composée que nous venons d'obtenir pour $\mathbb{E}[\lambda] = 7$ et $\sigma[\lambda] = 3$ ($a = 0,77$ et $b = 5,44$).



Chapitre 4

Caractérisation d'un échantillon

À l'issu d'une expérience il peut être utile de caractériser l'échantillon obtenu de manière quantitative, sans référence à la population dont il provient. Pour cela, nous pouvons définir plusieurs variables ou grandeurs que nous qualifierons d'empiriques puisqu'elles sont obtenues sur le résultat d'une expérience.

4.1 Moyenne et variance empirique

Considérons un échantillon composé des valeurs obtenues pour une variable aléatoire X lors de n expériences identiques : $\{X_i\} = (X_1, \ldots, X_n)$. Trois grandeurs sont particulièrement importantes pour le caractériser :

- Moyenne empirique :

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.1}$$

  La moyenne empirique n'est rien d'autre que la moyenne arithmétique des x_i.
- Médiane empirique : si les valeurs x_i sont ordonnées dans l'ordre croissant ($x_1 < x_2 < \ldots < x_n$), la médiane empirique est donnée par
  - n pair : $\frac{x_{n/2}+x_{1+n/2}}{2}$
  - n impair : $x_{(n+1)/2}$
- Variance empirique : les deux définitions les plus fréquentes sont

$$S_b^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - M^2 \tag{4.2}$$

  et

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2 \tag{4.3}$$

Nous pourrions de la même manière définir les moment empiriques d'ordre supérieur.

Puisque les X_i sont des variables aléatoires, les moments empiriques sont aussi aléatoires. Nous pouvons donc calculer leurs espérances et leurs variances. Par exemple, l'espérance de la moyenne empirique est :

$$\mathbb{E}[M] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X] = \mathbb{E}[X]$$

et sa variance est :

$$\mathrm{var}[M] = \sum_{i=1}^{n}\left(\frac{\partial M}{\partial X_i}\right)^2 \sigma[X_i]^2 = \sum_{i=1}^{n}\frac{1}{n^2}\sigma[X]^2 = \frac{\sigma[X]^2}{n}$$

soit :

$$\sigma[M] = \frac{\sigma[X]}{\sqrt{n}} \tag{4.4}$$

Nous obtenons ainsi un résultat extrêmement important suivant lequel l'incertitude (ou l'écart-type) sur la moyenne empirique décroît avec la racine carrée de la taille de l'échantillon.

4.2 Densité de probabilité et fonction de répartition empirique

La densité de probabilité empirique d'un échantillon $\{X_i\} = (X_1, \ldots, X_n)$ est :

$$f_{\text{sample}}(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x_i)$$

Si nous calculons l'espérance et la variance de X en utilisant cette densité de probabilité dans les relations 2.1 et 2.2, nous trouvons :

$$\mathbb{E}[X] = M \quad \text{et} \quad \mathrm{var}[X] = S_b^2$$

La fonction de répartition empirique est :

$$F_{\text{sample}}(x) = \frac{\text{Nombre d'éléments dans l'échantillon} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

où $I(A)$ est la fonction indicatrice de l'événement A [1]. Pour un x fixé, $I(X_i \leq x)$ est une variable aléatoire de Bernouilli de paramètre $p = F(x)$. $nF_{\text{sample}}(x)$ est donc distribué suivant la loi binomiale de paramètres n et $p = F(x)$.

4.3 Histogramme

On représente très souvent un échantillon par un histogramme. Pour construire un histogramme, il faut choisir des intervalles (ou classes ou *bins*) pour la variable X et associer chaque éléments de

1. $I(A) = 1$ si A est vrai et 0 si A est faux.

l'échantillon à un intervalle donné. Un histogramme est entièrement défini par la donnée des intervalles (C_k) et des nombres d'éléments (on parle aussi de nombre d'entrées) dans chaque intervalle (N_k) :

$$\text{histogramme} = \{C_k; N_k\}$$

où k est un entier qui va de 1 au nombre total d'intervalles m et $\sum_{k=1}^{m} N_k = n$.

La figure 4.1 montre l'histogramme pour l'échantillon composé par l'ensemble des valeurs données au-dessus de celui-ci.

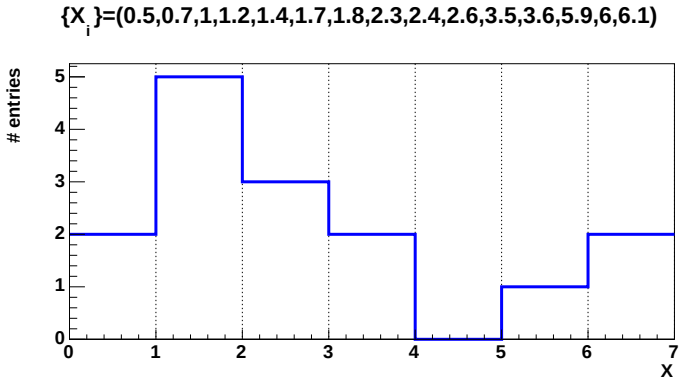


FIGURE 4.1 – Exemple d'histogramme.

Représenter un échantillon sous la forme d'un histogramme induit inévitablement une perte d'information puisque nous ne connaissons plus les valeurs exactes que la variable aléatoire a prise au cours des différents tirages.

La valeur moyenne et la variance de l'échantillon calculées à partir d'un histogramme sont :

$$M = \frac{\sum_{k=1}^{m} x_k N_k}{n} \quad \text{et} \quad S_b = \frac{\sum_{k=1}^{m} x_k^2 N_k}{n} - M^2$$

où x_k est la valeur de X au centre du *bin* k. Dans le cas où les *bins* sont infiniment petits (chaque *bin* contient soit zéro soit une entrée) les expressions ci-dessus se ramènent aux expressions 4.1 et 4.2. Dans le cas (extrême) où il n'y a qu'un seul *bin* contenant toutes les entrées, nous avons

$$M = x_1 \quad \text{et} \quad S = x_1^2 - x_1^2 = 0$$

ce qui ne reflète bien sûr pas les propriétés de l'échantillon.

4.3.1 Loi de probabilité

En terme de probabilité, un histogramme peut être décrit soit par la loi multinomiale soit par la loi de Poisson en fonction du contexte. Dans le cas où la taille totale de l'échantillon n est fixée,

l'histogramme est décrit par la loi multinomiale (voir section A.6) :

$$p(n_1, \ldots, n_m) = \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m} \tag{4.5}$$

où p_i est la probabilité pour que la variable X tombe dans l'intervalle C_i. Si la densité de probabilité de la population $f_X(x)$ dont l'échantillon est issu est connue, alors les p_i sont donnés par :

$$p_i = \int_{C_i} f_X(x) \mathrm{d}x$$

où l'intégrale porte sur l'ensemble des valeurs contenues dans l'intervalle C_i. L'expression 4.5 peut être réécrite de la manière suivante :

$$p(n_1, \ldots, n_m) = \frac{n!}{n_1! \ldots n_m!} \left(\frac{\nu_1}{n}\right)^{n_1} \ldots \left(\frac{\nu_m}{n}\right)^{n_m} = \frac{n!}{n_1! \ldots n_m!} \frac{\nu_1^{n_1} \ldots \nu_m^{n_m}}{n^n}$$

où $\nu_i = np_i$ est le nombre d'entrées attendues dans l'intervalle i ($\nu_i = \mathbb{E}[N_i]$).

Dans le cas où le nombre total d'éléments n dans l'échantillon n'est pas fixé mais distribué suivant une loi de Poisson de paramètres $\nu = \sum_{i=1}^{m} \nu_i$ ($\nu_i = \nu p_i$), l'histogramme est décrit par :

$$\begin{aligned} p(n, n_1, \ldots, n_m) &= \frac{\nu^n}{n!} e^{-\nu} \times \frac{n!}{n_1! \ldots n_m!} p_1^{n_1} \ldots p_m^{n_m} \\ &= \frac{\nu^n}{n!} e^{-\nu} \times \frac{n!}{n_1! \ldots n_m!} \left(\frac{\nu_1}{\nu}\right)^{n_1} \ldots \left(\frac{\nu_m}{\nu}\right)^{n_m} \\ &= e^{-\nu} \frac{\nu_1^{n_1} \ldots \nu_m^{n_m}}{n_1! \ldots n_m!} \\ &= e^{-\nu_1} \ldots e^{-\nu_m} \frac{\nu_1^{n_1} \ldots \nu_m^{n_m}}{n_1! \ldots n_m!} \\ &= \prod_{i=1}^{m} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} \end{aligned} \tag{4.6}$$

La probabilité de l'histogramme dans ce cas est donnée par le produit des probabilités de Poisson dans chaque *bin*.

4.3.2 Histogramme et densité de probabilité

Comme nous l'avons vu dans la section précédente, les nombres d'entrées N_k sont décrits par la loi multinomiale. Les paramètres p_k de cette loi sont donnés par

$$p_k = \int_{C_k} f_X(x) \mathrm{d}x$$

En notant δ la taille des *bins* et x_k la valeur de X au centre du *bin* k, alors :

$$f_X(x_k) = \lim_{\delta \to 0} \frac{\mathrm{d}p_k}{\delta}$$

De plus,

$$\mathrm{d}p_k = \lim_{n\to\infty} \frac{n_k}{n}$$

Donc :

$$f_X(x_k) = \lim_{\substack{\delta\to 0\\ n\to\infty}} \frac{n_k}{n}$$

La densité de probabilité est donc la limite de l'histogramme lorsque la taille des *bins* tend vers zéro et le nombre total d'entrées vers l'infini. Il est également intéressant de noter que, dans ces mêmes limites, les variances et covariances des N_k sont données par :

$$\lim_{\substack{\delta\to 0\\ n\to\infty}} \sigma\left[N_k\right]^2 = \lim_{p_k\to 0} np_k(1-p_k) = np_k = \nu_k$$

et

$$\lim_{\substack{\delta\to 0\\ n\to\infty}} \mathrm{cov}\left(N_k, N_l\right) = \lim_{p_k,p_l\to 0} -np_kp_l = 0$$

Nous retrouvons ainsi le fait que les éléments d'un échantillon sont décorrélés (ce que nous avions utilisé pour dériver l'expression 4.4).

4.4 Limite asymptotique

Il est souvent utile de chercher à déterminer comment se comporte une fonction donnée de l'échantillon lorsque le nombre d'observations tend vers l'infini.

4.4.1 Différents types de convergence

Dans cette section nous noterons Y_n une fonction donnée de l'échantillon : $Y_n = Y_n(X_1, \ldots, X_n) \in \mathbb{R}$.

Convergence en loi (ou en distribution)

Soit $F_n(x)$ la fonction de répartition de Y_n. Soit X une variable aléatoire de fonction de répartition $F(x)$. Y_n tend en loi (ou en distribution) vers X si

$$F_n(x) \xrightarrow[n\to\infty]{} F(x), \text{ pour tout } x \text{ où } F \text{ est continue}$$

Nous noterons

$$Y_n \xrightarrow{L} X$$

ou

$$F_n \xrightarrow{L} F$$

Convergence en probabilité

Soit X une variable aléatoire. Y_n converge en probabilité vers X si, pour tout $\alpha > 0$,

$$P\left(|Y_n - X| \geq \alpha\right) \xrightarrow[n\to\infty]{} 0$$

Nous noterons

$$Y_n \xrightarrow{P} X$$

On montre que si $Y_n \xrightarrow{P} X$, alors $Y_n \xrightarrow{L} X$ (la réciproque est fausse, sauf si X est une constante).

Théorème de Slutsky

Soit $Z_n \in \mathbb{R}$ une autre fonction de l'échantillon $\{X_i\}$ $(i \in [1, n])$ et soit $f(x, y)$ une fonction définie sur $\mathbb{R}^2$ continue pour tout (x, c), où c est une constante. Si Y_n converge en loi vers une variable aléatoire X et si Z_n converge en loi vers c, alors

$$f(Y_n, Z_n) \xrightarrow{L} f(X, c)$$

Remarques :

- Au lieu de la condition "Z_n converge en loi vers c" nous aurions pu dire "Z_n converge en probabilité vers c" (ces deux affirmations sont équivalentes lorsque c est une constante).
- Le théorème n'est plus valable si Z_n ne converge pas vers une constante mais vers une variable aléatoire non dégénérée.

4.4.2 Loi des grands nombres

La loi faible des grands nombres [2] stipule que la moyenne empirique converge en probabilité vers l'espérance :

$$\forall \varepsilon > 0,\ P\left(|M - \mathbb{E}\left[X\right]| \geq \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

En utilisant les notations introduites à la section précédente :

$$M \xrightarrow{P} \mathbb{E}\left[X\right]$$

Cette loi se démontre grâce à l'inégalité de Bienaymé-Tchebichev (voir section 2.3) :

$$P(|M - \mathbb{E}\left[M\right]| \geq \varepsilon) \leq \frac{\mathrm{var}\left[M\right]}{\varepsilon^2}$$

Nous avons montré dans la section 4.1 que $\mathbb{E}\left[M\right] = \mathbb{E}\left[X\right]$ et $\mathrm{var}\left[M\right] = \mathrm{var}\left[X\right]/n$, donc :

$$P(|M - \mathbb{E}\left[X\right]| \geq \varepsilon) \leq \frac{\mathrm{var}\left[X\right]}{n\varepsilon^2}$$

La probabilité que l'écart entre la moyenne empirique et l'espérance de X excède ε est donc bornée supérieurement par une grandeur qui tend vers 0 lorsque n tend vers l'infini.

2. Il existe également une loi forte des grands nombres mais nous aurons jamais à l'utiliser. Nous la passons donc sous silence.

Chapitre 5

Fonction de vraisemblance (ou *likelihood*)

Une grandeur centrale en statistique est celle de vraisemblance (*likelihood*). Dans ce chapitre nous nous contenterons de la définir et de donner son expression dans quelques cas standards. Il faudra patienter jusqu'aux chapitres suivants pour voir comment elle est utilisée en pratique dans les problèmes d'inférences fréquentiste et bayésienne.

5.1 Définitions

Considérons pour commencer le cas où les variables aléatoires sont continues. Soit un échantillon $\{X_i\} = (X_1, \ldots, X_n)$ obtenu à l'issu d'une expérience décrite par la densité de probabilité conjointe $f_{\{X_i\}}(x_1, \ldots, x_n; \theta_1, \ldots \theta_q)$. $\theta_1, \ldots, \theta_q$ sont les paramètres dont dépend la densité de probabilité. Afin d'alléger les notations, nous noterons cette densité de probabilité de manière condensée $f(x; \theta)$. Le *likelihood* est, par définition, égal à la densité de probabilité conjointe, non pas vu comme une fonction de l'échantillon x mais comme une fonction des paramètres θ :

$$\mathcal{L}(\theta; x) = f(x; \theta)$$

Ainsi, le *likelihood* est égal à la densité de probabilité où les rôles de variable et de paramètre sont inversés. En général, le *likelihood* n'est pas normalisé :

$$\int \mathcal{L}(\theta; x) \mathrm{d}\theta \neq 1$$

Dans le cas où les X_i sont indépendants, le *likelihood* peut se factoriser :

$$\mathcal{L}(\theta; x) = f(x_1; \theta) \times \ldots \times f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Le cas où les variables aléatoires sont discrètes est totalement analogue au cas des variables continues. Il suffit de remplacer la densité de probabilité par la probabilité :

$$\mathcal{L}(\theta; x) = P(X_1 = x_1 \cap \ldots \cap X_n = x_n)$$

Il existe de nombreuses situations où la taille de l'échantillon n'est pas fixe mais correspond elle aussi une variable aléatoire, distribuée suivant la loi de Poisson de paramètre μ. Dans ces situations, il faut inclure le terme de Poisson pour n dans la probabilité conjointe et donc dans le *likelihood*. On forme ainsi ce que l'on appelle la vraisemblance étendue (*extended likelihood*) :

$$\mathcal{L}_{\text{ext}}(\mu, \theta; n, x) = \frac{\mu^n}{n!} e^{-\mu} \mathcal{L}(\theta; x)$$

Dans la suite, pour ne pas alourdir les expressions nous noterons souvent le *likelihood* $\mathcal{L}$ et le *likelihood* étendu $\mathcal{L}_{(\text{ext})}(\theta)$ ou même $\mathcal{L}_{(\text{ext})}$. Cela simplifie aussi souvent les calculs de considérer $\ln \mathcal{L}_{(\text{ext})}$ (ou $-\ln \mathcal{L}_{(\text{ext})}$) au lieu de $\mathcal{L}_{(\text{ext})}$ et puisque seule la dépendance en θ de $\ln \mathcal{L}_{(\text{ext})}$ est importante, nous ignorerons les termes qui ne dépendent pas de θ.

5.2 Exemples élémentaires

5.2.1 Loi de Poisson

Considérons une expérience qui consiste à tirer une variable aléatoire N distribuée suivant une loi de Poisson de paramètre μ. Le *likelihood* est donné par

$$\mathcal{L}(\mu; n) = \frac{\mu^n}{n!} e^{-\mu}$$

et

$$-\ln \mathcal{L} = -n \ln \mu + \mu$$

La figure 5.1 montre la fonction bi-dimensionnelle $\mathcal{L}(\mu; n)$. Le likelihood est donné par les courbes le long de μ pour n fixé.

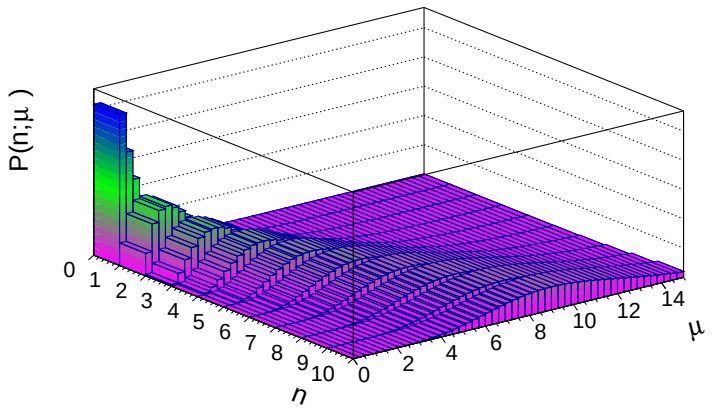


FIGURE 5.1 – Représentation bi-dimensionnelle de la fonction $\mathcal{L}(\mu; n)$.

Si l'expérience consiste à tirer m fois la variable n, alors

$$\mathcal{L}(\mu; n_1, \ldots, n_m) = \prod_{i=1}^{m} \frac{\mu^{n_i}}{n_i!} e^{-\mu}$$

et

$$-\ln \mathcal{L} = -\ln \mu \sum_{i=1}^{m} n_i + m\mu$$

5.2.2 Loi normale

Considérons une expérience qui consiste à tirer une variable aléatoire X distribuée suivant une loi normale de moyenne μ et écart-type σ. Le *likelihood* est donné par

$$\mathcal{L}(\mu, \sigma; x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Si l'expérience consiste à tirer n fois la variable X, alors

$$\mathcal{L}(\mu, \sigma; x_1, \ldots, x_n) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \ldots \times e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

et

$$-\ln \mathcal{L} = n \ln(\sigma) + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$

La figure 5.2 montre la fonction $\mathcal{L}(\mu, \sigma)$ pour $n = 2$ avec $x_1 = 2$ et $x_2 = 10$.

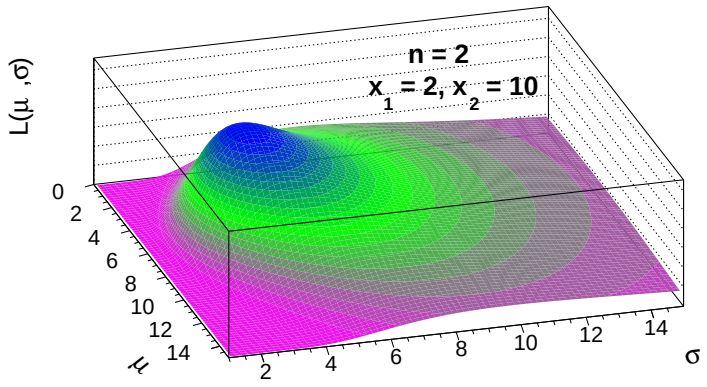


FIGURE 5.2 – Représentation du likelihood dans le cas gaussien pour $n = 2$, $x_1 = 2$ et $x_2 = 10$.

5.3 *Likelihood* pour un mélange

Il arrive souvent que les échantillons soient composés d'un mélange d'éléments issus de plusieurs processus aléatoires (typiquement un processus de signal et un ou plusieurs processus de bruit de fond). Dans ce cas les expressions des *likelihood* sont un peu plus compliquées que précédemment car il faut tenir compte des densités de probabilités pour chacun des processus. Dans la suite, nous noterons P le nombre total de processus et nous les indicerons par la lettre p ($p \in [1, P]$).

5.3.1 *Likelihood non binné*

Considérons une expérience au cours de laquelle une variable aléatoire X est tirée N fois (les tirages sont indépendants). L'échantillon est donc composé des N valeurs $\{X_i\} = (X_1, \ldots, X_N)$. Notons N_p le nombre d'éléments dans l'échantillon issus du processus p

$$\sum_{p=1}^{P} N_p = N$$

et $f_p(x;\theta)$ la densité de probabilité de la variable X pour le processus p. La densité de probabilité globale de X, incluant tous les processus, est donnée par

$$f(x;\{N_p\},\theta) = \frac{\sum\limits_{p=1}^{P} N_p f_p(x;\theta)}{\sum\limits_{p=1}^{P} N_p}$$

Nous avons donc

$$\mathcal{L}(\{N_p\},\theta) = \prod_{i=1}^{N} f(x_i;\{N_p\},\theta) = \prod_{i=1}^{N} \frac{\sum\limits_{p} N_p f_p(x_i;\theta)}{\sum\limits_{p} N_p} \tag{5.1}$$

soit

$$\begin{aligned}
\ln \mathcal{L} &= \sum_{i=1}^{N} \left(\ln \left[\sum_{p=1}^{P} N_p f_p(x_i;\theta) \right] - \ln \left[\sum_{p=1}^{P} N_p \right] \right) \\
&= \sum_{i=1}^{N} \ln \left[\sum_{p=1}^{P} N_p f_p(x_i;\theta) \right] - N \ln \left[\sum_{p=1}^{P} N_p \right]
\end{aligned}$$

D'un point de vue pratique, la contrainte $\sum_p N_p = N$ est prise en compte dans le *likelihood* non étendu que nous venons d'établir en remplaçant un des N_p, par exemple N_P, par $1 - \sum\limits_{p=1}^{P-1} N_p$.

Le *likelihood* étendu s'obtient en supprimant la contrainte $\sum_p N_p = N$ et en ajoutant le terme de Poisson décrivant la probabilité d'avoir N observations. Il est donné par

$$\mathcal{L}_{\text{ext}}(\{N_p\},\theta) = \frac{\left(\sum_p N_p\right)^N}{N!} e^{-\sum_p N_p} \prod_i f(x_i;\{N_p\},\theta)$$

donc

$$\begin{aligned}\ln \mathcal{L}_{\text{ext}} &= N \ln\left[\sum_p N_p\right] - \sum_p N_p + \ln \mathcal{L} \\ &= \sum_i \ln\left[\sum_p N_p f_p(x_i;\theta)\right] - \sum_p N_p\end{aligned}$$

5.3.2 *Likelihood binné*

Dans le cas où la taille de l'échantillon N est très grande, il peut être préférable d'utiliser un histogramme pour le représenter. Nous utiliserons la lettre b pour désigner les *bins* de l'histogramme ($b \in [1, B]$, où B est le nombre total de *bins*) et nous noterons δ_b la largeur du *bin* b.

Pour exprimer le *likelihood binné*, partons de l'expression donnée en 5.1 pour le *likelihood non binné* :

$$\mathcal{L}(\{N_p\},\theta) = \prod_{i=1}^{N} \frac{\sum_p N_p f_p(x_i;\theta)}{\sum_p N_p}$$

Soit $N_{p,b}$ le nombre d'entrées attendues pour le processus p dans le *bin* b ($\sum_{b=1}^{B} N_{p,b} = N_p$) :

$$N_{p,b} = N_p \int_{x\in b} f_p(x;\theta)dx$$

La densité de probabilité pour le processus p peut s'approximer, si δ_b est suffisamment petit, par

$$f_p(x_i;\theta) = \frac{N_{p,b_i}}{N_p \delta_{b_i}}$$

où b_i est le *bin* qui contient la valeur x_i de X. En reportant cette expressions dans le *likelihood*, nous trouvons :

$$\mathcal{L}(\{N_p\},\theta) = \prod_{i=1}^{N} \frac{\sum_p \frac{N_{p,b_i}}{\delta_{b_i}}}{\sum_p N_p} = \prod_{i=1}^{N} \frac{1}{\delta_{b_i}} \frac{\sum_p N_{p,b_i}}{\sum_p N_p}$$

Nous pouvons nous débarasser des facteurs $1/\delta_{b_i}$ puisqu'ils ne dépendent pas des paramètres, ce qui donne :

$$\mathcal{L}(\{N_p\},\theta) = \prod_{i=1}^{N} \frac{\sum_p N_{p,b_i}}{\sum_p N_p}$$

En changeant le produit sur les éléments de l'échantillon en un produit sur les *bins*, le *likelihood* devient

$$\mathcal{L}(\{N_p\},\theta) = \prod_{b=1}^{B} \left(\frac{\sum\limits_p N_{p,b}}{\sum\limits_p N_p} \right)^{N_b^{\text{obs}}}$$

où N_b^{obs} est le nombre d'entrées dans le *bin* b de l'histogramme ($\sum\limits_b N_b^{\text{obs}} = N$). Le terme entre parenthèse dans cette expression n'est rien d'autre que la probabilité pour que X soit dans le *bin* b, en incluant tous les processus. Nous voyons donc que le *likelihood binné* est donné par la loi multinomiale de paramètres $p_b = \frac{\sum\limits_p N_{p,b}}{\sum\limits_p N_p}$. Ceci n'est pas surprenant puisque nous savons qu'un histogramme est décrit par la loi multinomiale (voir section 4.3.1). Notons que l'expression précédente peut aussi s'écrire de la manière suivante :

$$\mathcal{L}(\{N_p\},\theta) = \frac{\prod\limits_{b=1}^{B} \left(\sum\limits_p N_{p,b} \right)^{N_b^{\text{obs}}}}{\left(\sum\limits_p N_p \right)^{N}}$$

Le *likelihood* étendu est donné par

$$\mathcal{L}_{\text{ext}}(\{N_p\},\theta) = \frac{\left(\sum\limits_p N_p \right)^{N}}{N!} e^{-\sum\limits_p N_p} \times \frac{\prod\limits_{b=1}^{B} \left(\sum\limits_p N_{p,b} \right)^{N_b^{\text{obs}}}}{\left(\sum\limits_p N_p \right)^{N}}$$

En supprimant $N!$ au dénominateur,

$$\begin{aligned}
\mathcal{L}_{\text{ext}}(\{N_p\},\theta) = e^{-\sum\limits_p N_p} \times \prod_{b=1}^{B} \left(\sum_p N_{p,b} \right)^{N_b^{\text{obs}}} &= e^{-\sum\limits_p \sum\limits_b N_{p,b}} \times \prod_{b=1}^{B} \left(\sum_p N_{p,b} \right)^{N_b^{\text{obs}}} \\
&= \prod_{b=1}^{B} e^{-\sum\limits_p N_{p,b}} \prod_{b=1}^{B} \left(\sum_p N_{p,b} \right)^{N_b^{\text{obs}}} \\
&= \prod_{b=1}^{B} \left[\left(\sum_p N_{p,b} \right)^{N_b^{\text{obs}}} e^{-\sum\limits_p N_{p,b}} \right]
\end{aligned}$$

Le *likelihood* étendu est donc donné par le produit de la loi de Poisson dans chaque *bin*. Comme précédemment, ceci n'est pas surprenant puisque nous savons que c'est précisement la loi de probabilité suivie par un histogramme avec un nombre total d'entrées distribué suivant une loi de Poisson (voir section 4.3.1).

5.4 *Likelihood* conjoint - expériences auxilliaires

Il arrive parfois que la situation soit plus compliquée que celles considérées jusqu'ici car l'expérience consiste non pas en une seule mais en plusieurs mesures. Dans ce cas, le *likelihood* correspond à la distribution conjointe des différentes expériences. Afin d'illustrer la construction d'un tel *likelihood*, considérons le problème célèbre appelé problème *on/off*. Dans ce problème, deux types de processus sont considérés (un signal et un bruit de fond) et deux expériences de comptage sont réalisées. La première, appelée expérience principale, consiste à mesurer le nombre d'événements dans une région où l'on n'attend à la fois du signal et du bruit de fond (nous appellerons cette région la région de signal). La deuxième, appellée expérience auxilliaire, consiste à mesurer le nombre d'événements dans une région où l'on attend que du bruit de fond (nous appellerons cette région la région de contrôle). Le nombre d'événements de bruit de fond attendus dans la région de contrôle est à priori différent de celui attendu dans la région de signal. Nous noterons τ le rapport entre ces deux nombres :

$$\tau = \frac{\text{nombre d'événements de bruit de fond attendu dans la région de contrôle}}{\text{nombre d'événements de bruit de fond attendu dans la région de signal}}$$

Soit s et b les nombres d'événements de signal et de bruit de fond attendus dans la région de signal, n_{on} le nombre d'événements observés dans la région de signal et n_{off} le nombre d'événements observés dans la région de bruit de fond. Le *likelihood* s'écrit

$$\mathcal{L} = P\left(n_{\text{on}}, n_{\text{off}}; s, b, \tau\right) = \underbrace{\frac{(s+b)^{n_{\text{on}}}}{n_{\text{on}}!} e^{-(s+b)}}_{\text{exp. principale}} \times \underbrace{\frac{(\tau b)^{n_{\text{off}}}}{n_{\text{off}}!} e^{-\tau b}}_{\text{exp. auxilliaire}}$$

Le problème *on/off* est particulièrement intéressant car il permet d'illustrer une procédure très courante en statistique : la reformulation. Reformuler un problème consiste à imaginer une expérience différente de l'expérience initiale mais ayant le même *likelihood* et conduisant donc à la même inférence. Dans le cas du problème *on/off*, nous voyons que le *likelihood* peut s'écrire, en notant $\mu_{\text{tot}} = s + b + \tau b$ et $n_{\text{tot}} = n_{\text{on}} + n_{\text{off}}$ les nombres totaux d'événements attendu et observés dans les deux régions (signal et contrôle),

$$\mathcal{L} = \frac{\mu_{\text{tot}}^{n_{\text{tot}}}}{n_{\text{tot}}!} e^{-\mu_{\text{tot}}} \times \binom{n_{\text{tot}}}{n_{\text{on}}} \rho^{n_{\text{on}}} \left(1-\rho\right)^{n_{\text{tot}} - n_{\text{on}}}$$

où $\rho = (s+b)/\mu_{\text{tot}}$ est la fraction d'événements attendus dans la région de signal. La réalisation des deux expériences de Poisson (principale et auxilliaire) est donc équivalente à la réalisation d'une seule expérience de Poisson globale dans laquelle le nombre d'événements attendu est μ_{tot} et d'une expérience binomiale de paramètre ρ et n_{tot} dans laquelle le nombre d'événements dans la région de signal est mesuré. Autrement dit,

$$\mathcal{L} = \text{Poisson}\left(n_{\text{on}}; s, b\right) \times \text{Poisson}\left(n_{\text{off}}; b, \tau\right) = \text{Poisson}\left(n_{\text{tot}}; \mu_{\text{tot}}\right) \times \text{Binomiale}\left(n_{\text{on}}; n_{\text{tot}}, \rho\right)$$

Chapitre 6

Estimation des paramètres

Ce chapitre traite d'un problème extrêmement fréquent et important en statistique qui est celui de l'estimation des paramètres. Le problème peut se formuler de la manière suivante. Supposons que l'on dispose d'une théorie, régie par un certain nombre de paramètres θ, prédisant la distribution d'une variable aléatoire X (nous noterons $f_X(x;\theta)$ sa densité de probabilité). Supposons de plus que nous réalisions une expérience au cours de laquelle nous recueillons un échantillon de données $\{X_i\}$ ($i \in [1, n]$). Comment peut-on estimer les vraies valeurs des paramètres à partir de l'échantillon ?

Le problème consiste donc à construire, à partir de l'échantillon $\{X_i\}$, un estimateur de θ. Nous noterons l'estimateur avec le même symbole que le paramètre à estimer, auquel nous ajouterons un accent circonflexe. L'estimateur de θ est donc $\hat{\theta}$.

Un estimateur est une fonction de l'échantillon : $\hat{\theta} = \hat{\theta}(\{X_i\})$. C'est donc une variable aléatoire caractérisée par une distribution.

Nous commencerons ce chapitre par un exemple simple permettant de se familiariser avec la problématique liée à l'estimation de paramètres. Nous verrons par la suite de manière un peu plus formelle comment caractériser un estimateur et quelques méthodes populaires permettant de les calculer.

6.1 Exemple préliminaire

Considérons une variable X distribuée suivant une loi gaussienne d'espérance inconnue μ et variance égale à 1. Supposons que nous disposions d'un échantillon de n valeurs de X : $(X_1, \ldots, X_n)$. Comment peut-on, à partir de ces n valeurs, estimer μ ? Un estimateur possible de μ est la moyenne empirique rencontrée au chapitre 4 :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La question qui se pose est de savoir si $\hat{\mu}$ est un bon estimateur de μ, c'est-à-dire si $\hat{\mu} \simeq \mu$. Répondre à cette question n'est en général pas facile. Cela dépend notamment de ce qui est entendu par "bon estimateur" ou $\hat{\mu} \simeq \mu$. Comme $\hat{\mu}$ est une variable aléatoire, sa valeur fluctue si nous répétons l'expérience plusieurs fois. Il se peut par exemple que dans une expérience $\hat{\mu}$ soit très proche de μ alors que dans une autre expérience $\hat{\mu}$ soit beaucoup plus grand que μ. Dans les deux cas nous ne pouvons pas savoir si nous sommes proche ou non de la réalité puisque nous ne connaissons pas μ. La figure 6.1

illustre ceci en montrant les résultats de trois expériences au cours desquelles quatre valeurs de X sont tirées. Nous ne pouvons evidemment pas savoir si c'est le résultat obtenu lors de la première, deuxième ou troisième expérience qui est le plus proche de la vraie valeur.

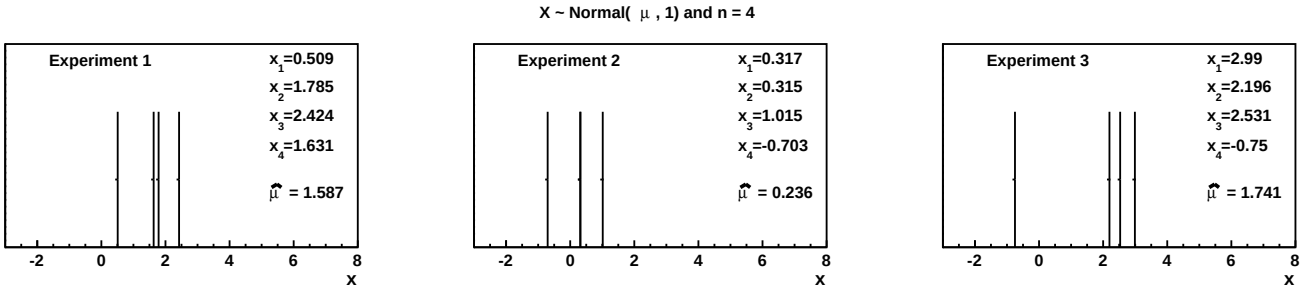


FIGURE 6.1 – Résultats de trois expériences visant à estimer l'espérance d'une loi normale de déviation standard unité. Pour chaque expérience l'échantillon est constitué de quatre valeurs. L'estimateur $\hat{\mu}$ est la moyenne empirique.

Dans l'approche fréquentiste étudiée dans ce chapitre, la réponse à la question posée ci-dessus s'obtient en examinant la distribution de l'estimateur. Puisque X est gaussien, $\hat{\mu}$ l'est aussi. Les résultats obtenus sur la moyenne empirique dans la section 4.1 permettent de déterminer l'espérance et l'écart-type de $\hat{\mu}$:

$$\mathbb{E}[\hat{\mu}] = \mu \quad \text{et} \quad \sigma[\hat{\mu}] = \frac{1}{\sqrt{n}}$$

La figure 6.2 montre la distribution de $\hat{\mu}$ obtenue en répétant l'expérience 1000 fois.

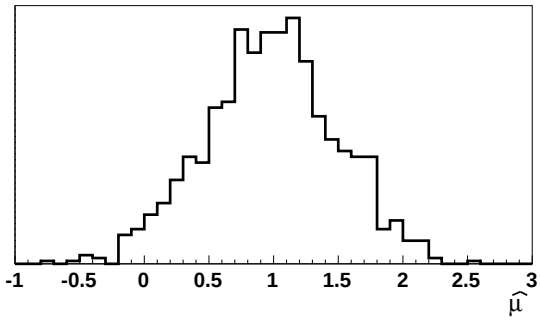


FIGURE 6.2 – Distribution de $\hat{\mu}$ obtenue en répétant l'expérience 1000 fois.

$\hat{\mu}$ est donc une variable aléatoire gaussienne ayant pour espérance la vraie valeur. $\hat{\mu}$ peut donc être vu comme un bon estimateur puisque, en moyenne, il donne la bonne valeur. De plus, lorsque la taille de l'échantillon tend vers l'infini, sa variance tend vers zéro, ce qui constitue aussi une propriété que l'on peut doit attendre d'un bon estimateur. Il n'est en revanche par certain que $\hat{\mu}$ soit le meilleur

estimateur en terme de variance pour n fini. Il se pourrait en effet qu'il existe un autre estimateur ayant, pour une même taille d'échantillon, une variance plus petite. Nous verrons par la suite que ce n'est pas le cas et que $\hat{\mu}$ est l'estimateur de plus petite variance. En somme, la moyenne empirique est, à tout point de vue, un bon estimateur pour l'espérance d'une loi normale de variance connue.

Nous verrons par la suite qu'il n'existe pas toujours d'estimateur qui soit bon sur tous les critères que nous venons de lister et qu'il est en nécessaire soit de favoriser tel ou tel critère soit de trouver un compromis entre tous les critères.

6.2 Propriétés des estimateurs

6.2.1 Biais

Le biais b d'un estimateur est par définition :

$$b = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

où $\mathbb{E}\left[\hat{\theta}\right]$ est l'espérance de l'estimateur donnée par (en notant $g(\hat{\theta})$ la densité de probabilité de $\hat{\theta}$)

$$\mathbb{E}\left[\hat{\theta}\right] = \int \hat{\theta} g(\hat{\theta}) \mathrm{d}\hat{\theta} = \int \hat{\theta}(\{x_i\}) \prod_{i=1}^{n} f_X(x_i;\theta) \mathrm{d}x_i$$

Un estimateur est non biaisé si $b = 0$. La figure 6.3 montre ce que peut être la distribution d'estimateur non-biaisé et d'un estimateur biaisé.

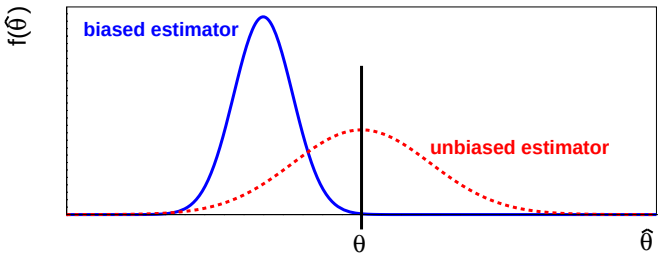


FIGURE 6.3 – Exemple de distribution d'un estimateur sans et avec biais.

Un estimateur non biaisé très couramment utilisé est la moyenne empirique $M = \left(\sum_{i=1}^{n} X_i\right)/n$. M est un estimateur non biaisé de l'espérance de X :

$$\mathbb{E}[M] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X] = \mathbb{E}[X]$$

Pour estimer la variance nous pouvons utiliser les formules de variance empirique données en 4.2 et 4.3. Le biais de S_b^2 se calcule aisément :

$$\begin{aligned}\mathbb{E}\left[S_b^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[M^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X^2\right] - \mathbb{E}\left[M^2\right] \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[M^2\right]\end{aligned}$$

Or

$$\mathbb{E}\left[M^2\right] = \frac{1}{n^2}\sum_i\sum_j \mathbb{E}\left[X_iX_j\right] = \frac{1}{n^2}\sum_i\left(\mathbb{E}\left[X_i^2\right] + \sum_{j\neq i}\mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]\right)$$

car $\operatorname{cov}(X_i, X_j) = 0$ si $i \neq j$ (les X_i étant indépendants les uns des autres, ils sont aussi décorrélés). Donc :

$$\mathbb{E}\left[M^2\right] = \frac{1}{n^2}\sum_i\left(\mathbb{E}\left[X^2\right] + (n-1)\mathbb{E}\left[X\right]^2\right) = \frac{1}{n}\left(\mathbb{E}\left[X^2\right] + (n-1)\mathbb{E}\left[X\right]^2\right)$$

Finalement,

$$\mathbb{E}\left[S_b^2\right] = \mathbb{E}\left[X^2\right] - \frac{1}{n}\mathbb{E}\left[X^2\right] - \frac{n-1}{n}\mathbb{E}\left[X\right]^2 = \frac{n-1}{n}\operatorname{var}\left[X\right]$$

S_b^2 est donc un estimateur biaisé de la variance, d'où le b en indice. Nous voyons aussi que la variance empirique S donnée par l'équation 4.3 est un estimateur non biaisé de la variance (c'est ce qui fait son intérêt).

6.2.2 Convergence

Un estimateur est convergent s'il converge en probabilité vers la vraie valeur lorsque la taille de l'échantillon tend vers l'infini :

$$P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall\varepsilon > 0$$

Du fait de la convergence en probabilité vers la vraie valeur, un estimateur convergent est asymptotiquement non biaisé. La figure 6.4 montre les distributions d'estimateurs convergents pour différentes tailles d'échantillon n_1, n_2 et n_3 (avec $n_3 > n_2 > n_1$). L'estimateur sur la figure de gauche est biaisé pour les petits échantillons alors que celui de droite est non biaisé, quelque soit la taille de l'échantillon.

Un exemple d'estimateur convergent est la moyenne empirique (d'après la loi faible des grands nombres décrite en 4.4.2, la moyenne empirique converge en probabilité vers l'espérance)

6.2.3 Efficacité

Soit $\mathcal{L}$ le *likelihood* :

$$\mathcal{L} = \mathcal{L}(\theta; x_i) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

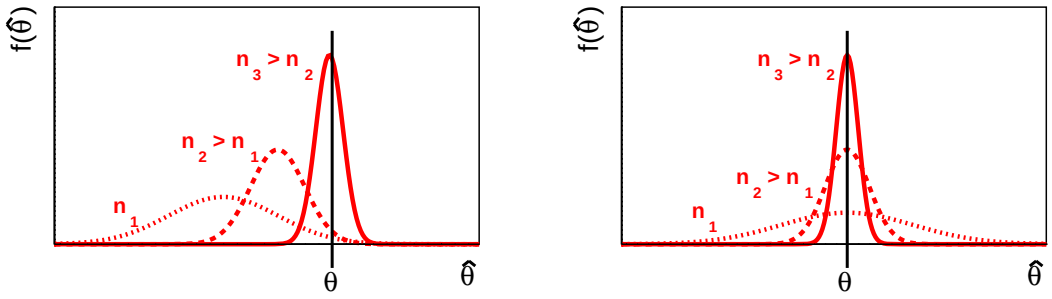


FIGURE 6.4 – Exemples de distributions d'un estimateur convergent pour différentes tailles d'échantillon.

Nous avons bien sûr la condition de normalisation

$$\int \dots \int \mathcal{L} \mathrm{d}x_1 \dots \mathrm{d}x_n = 1$$

En dérivant par rapport à θ,

$$\int \dots \int \frac{\partial \mathcal{L}}{\partial \theta} \mathrm{d}x_1 \dots \mathrm{d}x_n = 0$$

Cette expression peut s'écrire de manière équivalente comme ceci [1] :

$$\mathbb{E}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = 0 \tag{6.1}$$

car

$$\mathbb{E}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = \int \dots \int \frac{\partial \ln \mathcal{L}}{\partial \theta} \mathcal{L} \mathrm{d}x_1 \dots \mathrm{d}x_n = \int \dots \int \frac{\partial \mathcal{L}}{\partial \theta} \mathrm{d}x_1 \dots \mathrm{d}x_n$$

En dérivant cette expression, nous obtenons

$$\int \dots \int \left\{ \frac{\partial}{\partial \theta} \left(\frac{\partial \ln \mathcal{L}}{\partial \theta} \right) \mathcal{L} + \frac{\partial \ln \mathcal{L}}{\partial \theta} \frac{\partial \mathcal{L}}{\partial \theta} \right\} \mathrm{d}x_1 \dots \mathrm{d}x_n = 0$$

$$\Rightarrow \int \dots \int \left\{ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} + \left(\frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \theta} \right)^2 \right\} \mathcal{L} \mathrm{d}x_1 \dots \mathrm{d}x_n = 0$$

Soit,

$$\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = -\mathbb{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right] \tag{6.2}$$

1. Cette écriture n'est pas très rigoureuse. Il serait plus correct d'écrire

$$\mathbb{E}_\theta \left[\left. \frac{\partial \ln \mathcal{L}}{\partial \theta} \right|_\theta \right] = 0,$$

le point important ici étant que $\frac{\partial \ln \mathcal{L}}{\partial \theta}$ et l'espérance doivent être évalués au même point pour que cette grandeur soit nulle.

L'espérance de $\hat{\theta}$ est

$$\mathbb{E}\left[\hat{\theta}\right] = \int \dots \int \hat{\theta}\mathcal{L}\mathrm{d}x_1 \dots \mathrm{d}x_n = \theta + b$$

où b est le biais. En dérivant par rapport à θ, nous obtenons

$$\int \dots \int \hat{\theta}\frac{\partial \ln \mathcal{L}}{\partial \theta}\mathcal{L}\mathrm{d}x_1 \dots \mathrm{d}x_n = 1 + \frac{\partial b}{\partial \theta}$$

En utilisant 6.1, ceci peut s'écrire :

$$\int \dots \int \left(\hat{\theta} - (\theta + b)\right)\frac{\partial \ln \mathcal{L}}{\partial \theta}\mathcal{L}\mathrm{d}x_1 \dots \mathrm{d}x_n = 1 + \frac{\partial b}{\partial \theta}$$

Cette expression est la covariance de $\hat{\theta}$ et $\frac{\partial \ln \mathcal{L}}{\partial \theta}$:

$$\mathrm{cov}\left(\hat{\theta}, \frac{\partial \ln \mathcal{L}}{\partial \theta}\right) = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\left(\frac{\partial \ln \mathcal{L}}{\partial \theta} - \mathbb{E}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right]\right)\right] = \mathbb{E}\left[\left(\hat{\theta} - (\theta + b)\right)\frac{\partial \ln \mathcal{L}}{\partial \theta}\right]$$

D'après l'inégalité de Cauchy-Schwarz,

$$\mathrm{cov}\left(\hat{\theta}, \frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2 \leq \mathrm{var}\left[\hat{\theta}\right]\mathrm{var}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = \mathrm{var}\left[\hat{\theta}\right]\mathbb{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right]$$

Nous trouvons finalement le résultat important qui est que la variance d'un estimateur est bornée inférieurement :

$$\mathrm{var}\left[\hat{\theta}\right] \geq \frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right]}$$

En utilisant 6.2, nous pouvons aussi écrire :

$$\mathrm{var}\left[\hat{\theta}\right] \geq -\frac{\left(1 + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]}$$

La valeur minimale que peut prendre la variance s'appelle la limite de Rao-Cramer-Frechet (RCF). Pour un estimateur non biaisé ou dont le biais ne dépend pas de θ, nous avons :

$$\mathrm{var}\left[\hat{\theta}\right] \geq \frac{-1}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]}$$

Lorsque la borne inférieure sur la variance est atteinte, on dit que l'estimateur est efficace. L'inégalité dans les formules précédentes provient de l'inégalité de Cauchy-Schwarz. Or on sait que l'inégalite de Cauchy-Schwarz devient égalité lorsque les variables sont linéairement corrélées. La limite RCF est donc atteinte si

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = A(\theta)\hat{\theta} + B(\theta) \tag{6.3}$$

où $A(\theta)$ et $B(\theta)$ sont des fonctions de θ uniquement (elles ne dépendent pas de l'échantillon). Cette dernière expression est souvent écrite de façon légèrement différente. En utilisant la propriété 6.1, nous voyons que $B(\theta)/A(\theta) = -\mathbb{E}\left[\hat{\theta}\right]$. Nous pouvons donc écrire la condition 6.3 comme ceci :

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = A(\theta)\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right) = A(\theta)\left(\hat{\theta} - (\theta + b)\right) \tag{6.4}$$

Dans le cas où la grandeur à estimer n'est pas θ mais une fonction $\tau(\theta)$ de θ, la limite RCF s'écrit (en notant $\hat{\tau}$ l'estimateur de $\tau(\theta)$)

$$\mathrm{var}\left[\hat{\tau}\right] \geq -\frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]} \tag{6.5}$$

où b est le biais de $\hat{\tau}$ ($b = \mathbb{E}\left[\hat{\tau}\right] - \tau$). La condition pour que la limite RCF soit atteinte s'écrit quant à elle :

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = A(\theta)\left(\hat{\tau} - \mathbb{E}\left[\hat{\tau}\right]\right) \tag{6.6}$$

La limite RCF n'est souvent pas facile à calculer car elle fait intervenir l'espérance de la dérivée seconde du *likelihood*. Lorsque cette dernière n'est pas calculable aisément, une bonne approximation peut être obtenue en évaluant la dérivée en $\tau = \hat{\tau}$:

$$\mathrm{var}\left[\hat{\tau}\right] \geq -\frac{\left(\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\tau=\hat{\tau}}}$$

En multipliant les deux membres de 6.6 par $\hat{\tau} - \mathbb{E}\left[\hat{\tau}\right]$ et en prenant l'espérance, nous trouvons que la variance d'un estimateur efficace est donnée par :

$$\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta} = A(\theta)\mathrm{var}\left[\hat{\tau}\right]$$

soit

$$\mathrm{var}\left[\hat{\tau}\right] = \frac{\frac{\partial \tau}{\partial \theta} + \frac{\partial b}{\partial \theta}}{A(\theta)} \tag{6.7}$$

Cette expression est souvent plus pratique à manipuler que le second membre de 6.5.

Il est important de noter que, si un estimateur efficace existe, alors il existe pour une seule fonction de θ.

Exemple 6.1: Considérons une expérience de comptage dans laquelle le nombre d'événements est distribué suivant une loi de Poisson. Le *likelihood* est, en notant μ le paramètre de la loi de Poisson :

$$\mathcal{L} = \frac{\mu^N}{N!}e^{-\mu}$$

Donc

$$\ln \mathcal{L} = N \ln \mu - \mu$$

et

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{N}{\mu} - 1 = \frac{1}{\mu}(N - \mu)$$

Cette expression est de la forme 6.4 avec $A(\mu) = 1/\mu$ et $\hat{\mu} = N$ (nous savons par ailleurs que N est un estimateur non biaisé de μ : $\mathbb{E}[N] = \mu$). N est donc un estimateur efficace de μ. On vérifie bien que la limite RCF est égale à la variance de la loi de Poisson μ :

$$-\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2}\right]} = -\frac{1}{\mathbb{E}[-N/\mu^2]} = \frac{\mu^2}{\mathbb{E}[N]} = \mu$$

Exemple 6.2: Considérons une expérience dans laquelle on tire n fois une variable aléatoire X distribuée suivant une loi gaussienne. Nous avons (cf section 5.2.2) :

$$-\ln \mathcal{L} = n \ln(\sigma) + \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}$$

Donc

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = \frac{n}{\sigma^2}(M - \mu)$$

où M est la moyenne empirique. Cette expression est de la forme 6.4 avec $A(\mu) = n/\sigma^2$ et $\hat{\mu} = M$ (nous savons par ailleurs que M est un estimateur non biaisé de μ). M est donc un estimateur efficace de μ. On vérifie bien que la limite RCF est égale à la variance connue de M (σ^2/n) :

$$-\frac{1}{\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \mu^2}\right]} = -\frac{1}{\mathbb{E}[-n/\sigma^2]} = \frac{\sigma^2}{n}$$

Il découle de la condition 6.6 que, pour qu'un estimateur efficace existe, la distribution de X doit appartenir à la famille exponentielle :

$$f_X(x;\theta) = A(\theta)B(x)e^{C(\theta)D(x)} \tag{6.8}$$

où A et C (B et D) sont des fonctions de θ (x) uniquement. Nous laissons au lecteur le soin de vérifier que les différentes distributions considérées dans les exemples précédents appartiennent bien à cette famille.

6.2.4 Erreur quadratique moyenne

Un estimateur de petite variance n'est pas forcement meilleur qu'un estimateur de grande variance s'il est biaisé. Il peut être utile de tenir compte à la fois de la variance et du biais lors de l'évaluation de la qualité d'un estimateur. Nous définissons ainsi l'erreur quadratique moyenne (MSE pour *Mean Square Error*) de la manière suivante :

$$\mathrm{MSE} = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$$

L'erreur quadratique moyenne s'exprime simplement en fonction de la variance et du biais :

$$\begin{aligned}\mathrm{MSE} = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\ &\quad + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)\right] \\ &= \mathrm{var}\left[\hat{\theta}\right] + b^2\end{aligned}$$

où $b = \mathbb{E}\left[\hat{\theta}\right] - \theta$ est le biais de l'estimateur. Nous voyons ainsi que l'intérêt d'un estimateur de faible variance peut être moindre si son biais est important.

6.2.5 Exhaustivité

Soit $T(X_1, \ldots, X_n)$ une fonction de l'échantillon (ou statistique) et soit $p(x_1, \ldots, x_n; \theta)$ la probabilité conjointe :

$$p(\{x_i\}; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

La statistique T est dite exhaustive pour θ (ou par rapport à θ) si elle contient toute l'information nécessaire à l'estimation de ce paramètre. Ceci signifie qu'il suffit de connaître T pour calculer une valeur de l'estimateur $\hat{\theta}$ (la connaissance de l'échantillon complet n'apporte aucune information supplémentaire).

L'exhaustivité est définie formellement de la manière suivante. T est exhaustive si la probabilité conditionnelle d'observer l'échantillon $\{X_i\}$ sachant T ne dépend pas de θ :

$$p(\{x_i\}|T(\{x_i\}); \theta) = p(\{x_i\}|T(\{x_i\})) \tag{6.9}$$

Il existe une autre définition, équivalente et peut-être plus facile à comprendre, qui dit qu'une statistique est exhaustive si la probabilité conjointe de $\{X_i\}$ et $T(\{X_i\})$ peut se factoriser comme ceci (factorisation de Fisher-Neyman) :

$$p(\{x_i\}, T(\{x_i\}); \theta) = p(\{x_i\}; \theta) = g(T(\{x_i\}); \theta) \times h(\{x_i\}) \tag{6.10}$$

L'équivalence entre ces deux définitions se démontre, dans le cas discret, de la manière suivante. La probabilité conditionnelle d'observer $\{X_i\}$ sachant T est :

$$p(\{x_i\}|T(\{x_i\}); \theta) = \frac{p(\{x_i\}, T(\{x_i\}); \theta)}{p(T(\{x_i\}); \theta)} = \frac{p(\{x_i\}; \theta)}{p(T(\{x_i\}); \theta)}$$

Or,

$$p(T(\{x_i\});\theta) = \sum_{\{y_i\}:T(\{y_i\})=T(\{x_i\})} p(\{y_i\};\theta)$$

Si la probabilité peut se factoriser comme en 6.10, alors :

$$\begin{aligned} p(\{x_i\}|T(\{x_i\});\theta) &= \frac{g(T(\{x_i\});\theta)h(\{x_i\})}{\sum_{\{y_i\}:T(\{y_i\})=T(\{x_i\})} g(T(\{y_i\});\theta)h(\{y_i\})} \\ &= \frac{g(T(\{x_i\});\theta)h(\{x_i\})}{g(T(\{x_i\});\theta)\sum_{\{y_i\}:T(\{y_i\})=T(\{x_i\})} h(\{y_i\})} \\ &= \frac{h(\{x_i\})}{\sum_{\{y_i\}:T(\{y_i\})=T(\{x_i\})} h(\{y_i\})} \end{aligned}$$

Dans cette dernière expression toute dépendance en θ a disparu.

Exemple 6.3: Soit X une variable aléatoire de Bernouilli de paramètre θ. Soit un échantillon $\{X_i\}$ ($i \in [1,n]$) de n tirages indépendants de cette variable aléatoire. La probabilité conjointe s'écrit :

$$p(\{x_i\};\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Définissons la statistique $T = \sum_{i=1}^{n} X_i$. Nous avons donc :

$$p(\{x_i\};\theta) = \theta^t(1-\theta)^{n-t}$$

T est donc une statistique exhaustive pour le paramètre θ de la loi de Bernouilli ($g = \theta^t(1-\theta)^{n-t}$ et $h = 1$).

Exemple 6.4: Soit X une variable aléatoire gaussienne de moyenne μ et écart-type σ (nous supposons ce dernier connu). Soit un échantillon $\{X_i\}$ ($i \in [1,n]$) de n tirages indépendants de cette variable aléatoire. La densité de probabilité conjointe s'écrit :

$$f(\{x_i\};\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(x_i-\mu)^2}{2\sigma^2}}$$

Soit $M = \frac{\sum_{i=1}^{n} X_i}{n}$ la moyenne empirique. Introduisons cette moyenne dans f :

$$f(\{x_i\};\mu,\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(x_i-m+m-\mu)^2}{2\sigma^2}} = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}} e^{-\frac{\sum(m-\mu)^2}{2\sigma^2}}$$
$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}} e^{-\frac{n(m-\mu)^2}{2\sigma^2}}$$

Nous avons utilisé le fait que $\sum_{i=1}^{n}(x_i - m)(m - \mu) = 0$. M est donc une statistique exhaustive pour μ ($g = e^{-\frac{n(m-\mu)^2}{2\sigma^2}}$ et $h = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}}$).

6.2.6 Exhaustivité et efficacité

Dans la section précédente nous avons vu qu'une statistique $T(\{X_i\})$ est exhaustive si (nous utilisons ici des notations un peu allégées) :

$$\mathcal{L} = g(t;\theta)h(\{x_i\})$$

Dans ce cas, nous avons :

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \frac{\partial \ln g(t;\theta)}{\partial \theta}$$

$\partial \ln \mathcal{L}/\partial\theta$ ne dépend donc de $\{X_i\}$ qu'à travers T. Or nous avons vu dans la section 6.2.3 que cette condition est nécessaire pour que T soit un estimateur efficace de θ. Rappelons en effet que t est efficace si :

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = A(\theta)\left(t - (\theta + b)\right)$$

Nous voyons ainsi que, d'une manière générale :

- s'il n'existe pas de statistique exhaustive pour θ, alors il n'existe pas d'estimateur efficace de ce paramètre
- même s'il n'existe pas d'estimateur efficace de θ, il peut exister une statistique exhaustive pour ce paramètre
- s'il existe un estimateur efficace de θ, alors c'est une statistique exhaustive pour ce paramètre

En fait, il est possible de montrer que, la plupart du temps (voir [1] pour les détails), la condition pour qu'une statistique exhaustive existe est que les observations soient distribuées suivant une loi appartenant à la famille exponentielle (6.8). Or, c'est cette même condition qui détermine l'existence d'un estimateur efficace. Il y a donc, la plupart du temps, une correspondance biunivoque entre statistique exhaustive et estimateur efficace.

6.3 Méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance consiste à prendre comme estimateur la valeur du paramètre qui maximise le *likelihood* (ou, de manière équivalente, qui minimise $-\ln\mathcal{L}$). $\hat{\theta}$ est donc solution des équations :

$$\left.\frac{\partial(-\ln\mathcal{L})}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0 \quad \text{et} \quad \left.\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta^2}\right|_{\theta=\hat{\theta}} > 0$$

Les estimateurs obtenus par cette méthode sont souvent appelés "estimateurs ML" ou "MLE" (*Maximum Likelihood Estimator*). Nous les indicerons d'un ML (par exemple $\hat{\theta}_{\mathrm{ML}}$).

Afin de justifier intuitivement cette méthode, considérons l'exemple représenté sur la figure 6.5. Le graphique du haut montre la distribution d'une variable aléatoire X pour trois valeurs d'espérance $\theta = \mathbb{E}[X]$ ainsi qu'un échantillon de dix valeurs de X. Il parait clair d'après ce graphique que la vraie valeur de θ est probablement plus proche de 4 que de 1 ou de 6. En effet, les valeurs $\theta = 1$ et $\theta = 6$ paraissent peu à même d'expliquer l'échantillon observé car la probabilité d'observer l'échantillon avec ces valeurs de θ est faible (les valeurs observées se trouvent dans les queues de distribution, là ou la densité de probabilité est faible). La valeur $\theta = 4$ semble beaucoup plus plausible car, dans ce cas, les valeurs observées correspondent à des densités de probabilités relativement grandes (la plupart d'entres-elles se trouvent proche du pic de la distribution). Il parait ainsi naturel d'accorder plus de crédit à une valeur du paramètre qui conduit à une grande probabilité d'observation plutôt qu'à une valeur conduisant à une petite probabilité d'observation. Le likelihood étant une mesure de la probabilité d'observation, la méthode du maximum de vraisemblance s'en trouve ainsi justifiée. Le graphique du bas montre $-\ln\mathcal{L}$ en fonction de θ. La valeur $\theta = 4$ correspond au minimum de cette fonction. Donc $\hat{\theta}_{\mathrm{ML}} = 4$.

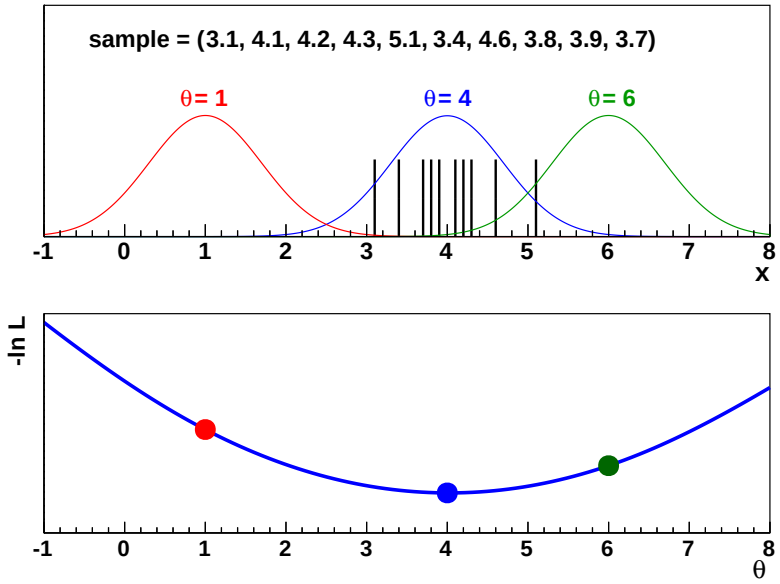


FIGURE 6.5 – Illustration de la méthode du maximum de vraisemblance.

Au-delà de son caractère intuitif, la méthode du maximum de vraisemblance présente beaucoup d'intérêt car les estimateurs qu'elle fournit possèdent, comme nous le verrons dans la section 6.3.1, des propriétés souvent désirables. C'est, pour ces raisons, une des méthodes les plus utilisées.

Exemple 6.5: Considérons une variable aléatoire X distribuée suivant une loi gaussienne d'espérance θ inconnue et variance σ^2 connue. Nous savons, d'après la section 6.2, que

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \theta) = \frac{n}{\sigma^2}(M - \theta)$$

Nous voyons immédiatement que $\hat{\theta}_{\mathrm{ML}} = M$. La méthode du maximum de vraisemblance fournit donc dans ce cas le meilleur estimateur possible puisque M est convergent, non biaisé et efficace. L'exemple discuté précédemment et représenté sur la figure 6.5 correspond à la situation gaussienne considérée ici. Nous vérifions sans difficulté que le calcul de la moyenne empirique à partir des valeurs données sur la figure du haut conduit à $\hat{\theta}_{\mathrm{ML}} = 4$, ce qui correspond bien à la valeur trouvée en examinant la courbe $-\ln \mathcal{L}$ en fonction de θ.

Exemple 6.6: Pour la loi binomiale, nous avons

$$\frac{\partial \ln \mathcal{L}}{\partial p} = \frac{n}{p(1-p)} \left(\frac{k}{n} - p \right)$$

Donc $\hat{p}_{\mathrm{ML}} = k/n$. Cette équation montre également que k/n est un estimateur efficace de p. De plus, k/n est un estimateur non biaisé et convergent de p. Nous voyons donc que la méthode du maximum de vraisemblance trouve, comme dans le cas gaussien considéré dans l'exemple 6.5, le meilleur estimateur possible.
Cet exemple ainsi que le précédent laisse entrevoir une propriété générale des estimateurs ML qui porte sur la relation qu'ils entretiennent avec les estimateurs non biaisés et efficaces. Nous pouvons en effet voir à travers ces exemples que, si un estimateur efficace et non biaisé existe, alors il est trouvé par la méthode du maximum de vraisemblance. Nous reviendrons sur cette propriété générale dans la section 6.3.1.

Il peut arriver que nous voulions estimer non pas θ mais une fonction de ce paramètre $\tau(\theta)$. Dans ce cas, l'estimateur $\hat{\tau}_{\mathrm{ML}}$ de $\tau(\theta)$ est obtenu en résolvant

$$\left. \frac{\partial(-\ln \mathcal{L})}{\partial \tau} \right|_{\tau = \hat{\tau}_{\mathrm{ML}}} = 0 \quad \text{et} \quad \left. \frac{\partial^2(-\ln \mathcal{L})}{\partial \tau^2} \right|_{\tau = \hat{\tau}_{\mathrm{ML}}} > 0$$

Or

$$\frac{\partial(-\ln\mathcal{L})}{\partial\tau} = \frac{\partial(-\ln\mathcal{L})}{\partial\theta}\frac{\mathrm{d}\theta}{\mathrm{d}\tau} \quad \text{et} \quad \frac{\partial^2(-\ln\mathcal{L})}{\partial\tau^2} = \frac{\partial^2 - \ln\mathcal{L}}{\partial\theta^2}\left(\frac{\mathrm{d}\tau}{\mathrm{d}\theta}\right)^{-2} + \frac{\partial - \ln\mathcal{L}}{\partial\theta}\left(\frac{\mathrm{d}^2\theta}{\mathrm{d}\tau^2}\right)$$

Ces équations montrent que la valeur de θ qui maximise le *likelihood* en fonction de θ est aussi celle qui le maximise en fonction de τ. Autrement dit, si $\hat{\theta}_{\mathrm{ML}}$ est l'estimateur ML de θ, alors $\tau(\hat{\theta}_{\mathrm{ML}})$ est l'estimateur ML de τ : $\hat{\tau}_{\mathrm{ML}} = \tau(\hat{\theta}_{\mathrm{ML}})$.

6.3.1 Propriétés des MLE

Propriété 1 : s'il existe une statistique exhaustive pour θ, alors $\hat{\theta}_{\mathrm{ML}}$ en est fonction (ceci se voit directement à partir de l'équation 6.10).

Propriété 2 : si un estimateur efficace non biaisé existe, alors il est trouvé par la méthode du maximum de vraisemblance, il est unique et sa variance est :

$$\mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right] = -\frac{\left(\partial\tau/\partial\theta\right)^2}{\left.\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}}} \tag{6.11}$$

En effet, soit $\hat{\tau}$ un estimateur non biaisé et efficace de $\tau(\theta)$. D'après 6.6 nous avons, pour ce type d'estimateurs, la relation suivante :

$$\frac{\partial\ln\mathcal{L}}{\partial\theta} = A(\theta)\left(\hat{\tau} - \tau(\theta)\right)$$

Nous voyons donc que $\hat{\tau}_{\mathrm{ML}} = \tau(\hat{\theta}_{\mathrm{ML}}) = \hat{\tau}$. Pour montrer que cet estimateur est unique, dérivons l'équation précédente :

$$\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2} = \frac{\partial A(\theta)}{\partial\theta}\left(\hat{\tau} - \tau(\theta)\right) - A(\theta)\frac{\partial\tau}{\partial\theta}$$

En utilisant 6.7,

$$\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2} = \frac{\partial A(\theta)}{\partial\theta}\left(\hat{\tau}_{\mathrm{ML}} - \tau(\theta)\right) - A(\theta)^2\mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right]$$

Evaluons cette expression en $\theta = \hat{\theta}_{\mathrm{ML}}$ (et donc $\tau = \hat{\tau}_{\mathrm{ML}}$) :

$$\left.\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2}\right|_{\tau=\hat{\tau}_{\mathrm{ML}}} = -A(\hat{\theta}_{\mathrm{ML}})^2\mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right] < 0 \tag{6.12}$$

Toutes les valeurs de $\hat{\tau}_{\mathrm{ML}}$ correspondent à des maximums du *likelihood*. Ce dernier est donc unique (s'il y avait plusieurs maximums, il y aurait forcément un minimum entre deux maximums). Enfin, pour démontrer l'équation 6.11, il faut se souvenir que pour un estimateur efficace non biaisé nous avons les relations suivantes :

$$\mathbb{E}\left[\frac{\partial^2\ln\mathcal{L}}{\partial\theta^2}\right] = -\frac{\left(\partial\tau/\partial\theta\right)^2}{\mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right]}$$

et

$$\frac{\partial\tau}{\partial\theta} = A(\hat{\theta}_{\mathrm{ML}})\mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right]$$

Donc

$$\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = -A(\hat{\theta}_{\mathrm{ML}})^2 \mathrm{var}\left[\hat{\tau}_{\mathrm{ML}}\right]$$

Ainsi, d'après 6.12

$$\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right] = \left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\tau=\hat{\tau}_{\mathrm{ML}}}$$

Ce qui démontre bien 6.11.

Propriété 3 : les estimateurs ML sont convergents.

Propriété 4 : les estimateurs ML sont asymptotiquement efficaces.

Propriété 5 : les estimateurs ML sont asymptotiquement gaussiens.

Les propriétés 3, 4 et 5 se démontrent de la façon suivante. Écrivons le développement de Taylor pour la dérivée de $\ln \mathcal{L}$ (θ_0 désigne la vraie valeur de θ) :

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} = \left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} + \left(\hat{\theta}_{\mathrm{ML}} - \theta_0\right) \left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta=\theta^\star}$$

où $\theta^\star$ est compris entre θ_0 et $\hat{\theta}_{\mathrm{ML}}$. Par définition de $\hat{\theta}_{\mathrm{ML}}$, le membre de gauche de cette équation est nul :

$$\hat{\theta}_{\mathrm{ML}} - \theta_0 = -\frac{\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0}}{\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta=\theta^\star}} \tag{6.13}$$

Les numérateur et dénominateur dans le membre de droite peuvent s'écrire comme des moyennes empiriques de variables aléatoires (au facteur $1/n$ près) :

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} = \sum_{i=1}^{n} \left.\frac{\partial \ln f(x_i;\theta)}{\partial \theta}\right|_{\theta=\theta_0}$$

et

$$\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta=\theta^\star} = \sum_{i=1}^{n} \left.\frac{\partial^2 \ln f(x_i;\theta)}{\partial \theta^2}\right|_{\theta=\theta^\star}$$

Ils convergent donc, d'après la loi des grands nombres, vers leurs espérances :

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0} \xrightarrow{P} \mathbb{E}_{\theta_0}\left[\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\theta_0}\right] = 0$$

et

$$\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta=\theta^\star} \xrightarrow{P} \mathbb{E}_{\theta_0}\left[\left.\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right|_{\theta=\theta^\star}\right]$$

Nous voyons donc que si $\mathbb{E}_{\theta_0}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\Big|_{\theta=\theta^\star}\right] \neq 0$, nous devons avoir

$$\hat{\theta}_{\mathrm{ML}} \xrightarrow{P} \theta_0$$

pour satisfaire à l'équation 6.13. Autrement dit, $\hat{\theta}_{\mathrm{ML}}$ est une estimateur convergent de θ_0. Notons que le choix du point autour duquel nous avons fait le développement limité peut sembler à première vue arbitraire. Nous aurions à priori pu choisir un point $\theta_c \neq \theta_0$. Si tel avait été le cas, nous aurions eu

$$\frac{\partial \ln \mathcal{L}}{\partial \theta}\Big|_{\theta=\theta_c} \xrightarrow[n\to\infty]{} \mathbb{E}_{\theta_0}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\Big|_{\theta=\theta_c}\right] \neq 0$$

ce qui aurait permis de montrer que $\hat{\theta}_{\mathrm{ML}}$ ne converge pas vers θ_c mais pas qu'il converge vers θ_0 [2].

Pour prouver que $\hat{\theta}_{\mathrm{ML}}$ est asymptotiquement efficace et gaussien il suffit de remarquer que :

$$\frac{\partial \ln \mathcal{L}}{\partial \theta}\Big|_{\theta=\theta_0} \xrightarrow{L} \mathcal{N}\left(0, I_n(\theta_0)\right)$$

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\Big|_{\theta=\theta^\star} \xrightarrow{P} -I_n(\theta_0)$$

où $I_n(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\big|_{\theta=\theta_0}\right)^2\right] = -\mathbb{E}_{\theta_0}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\Big|_{\theta=\theta_0}\right]$ est l'information de Fisher. Pour la première équation nous avons appliqué le théorème central limite et pour la deuxième nous avons utilisé le fait que $\theta^\star \xrightarrow{P} \theta_0$ (puisque $\theta^\star$ est entre θ_0 et $\hat{\theta}_{\mathrm{ML}}$). Nous sommes donc dans les conditions où nous pouvons appliquer le théorème de Slutsky (voir section 4.4.1) à l'équation 6.13 :

$$\hat{\theta}_{\mathrm{ML}} - \theta_0 \xrightarrow{L} -\frac{\mathcal{N}(0, I_n(\theta_0))}{-I_n(\theta_0)}$$

Finalement,

$$\hat{\theta}_{\mathrm{ML}} - \theta_0 \xrightarrow{L} \mathcal{N}(0, I_n(\theta_0)^{-1})$$

Ceci prouve que $\hat{\theta}_{\mathrm{ML}}$ est asymptotiquement gaussien et asymptotiquement efficace, puisque sa variance est égale à la limite RCF.

6.3.2 Méthode graphique pour l'estimation de l'écart-type des estimateurs ML

Cas unidimensionnel

Faisons le développement limité du *likelihood* autour de $\hat{\theta}_{\mathrm{ML}}$:

$$-\ln \mathcal{L}(\theta) = -\ln \mathcal{L}(\hat{\theta}_{\mathrm{ML}}) + (\theta - \hat{\theta}_{\mathrm{ML}})\left.\frac{\partial(-\ln \mathcal{L})}{\partial \theta}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} + \frac{1}{2}(\theta - \hat{\theta}_{\mathrm{ML}})^2\left.\frac{\partial^2(-\ln \mathcal{L})}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} + \dots$$

2. Il est important de voir ici que $\mathbb{E}\left[\frac{\partial \ln \mathcal{L}}{\partial \theta}\right] = 0$ seulement si $\frac{\partial \ln \mathcal{L}}{\partial \theta}$ et l'espérance sont évalués à la même valeur de θ.

Le second terme dans le membre de droite est nul (par définition de $\hat{\theta}_{\mathrm{ML}}$), donc :

$$-\ln\mathcal{L}(\theta) \simeq -\ln\mathcal{L}(\hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}(\theta-\hat{\theta}_{\mathrm{ML}})^2 \left.\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}}$$

En supposant que $\hat{\theta}_{\mathrm{ML}}$ est efficace, nous pouvons écrire de manière approximative $\left.\frac{\partial^2 \ln\mathcal{L}}{\partial\theta^2}\right|_{\theta=\hat{\theta}_{\mathrm{ML}}} = -\widehat{\mathrm{var}}\left[\hat{\theta}_{\mathrm{ML}}\right]^{-1}$ (nous utilisons la notation $\widehat{\mathrm{var}}\,[\ldots]$ pour faire apparaître le fait que c'est une estimation de la limite RCF) :

$$-\ln\mathcal{L}(\theta) \simeq -\ln\mathcal{L}(\hat{\theta}_{\mathrm{ML}}) + \frac{(\theta-\hat{\theta}_{\mathrm{ML}})^2}{2\widehat{\mathrm{var}}\left[\hat{\theta}_{\mathrm{ML}}\right]} \tag{6.14}$$

Ainsi nous voyons que si $\theta = \hat{\theta}_{\mathrm{ML}} \pm \sqrt{\widehat{\mathrm{var}}\left[\hat{\theta}_{\mathrm{ML}}\right]}$, alors $-\ln\mathcal{L}(\theta) \simeq -\ln\mathcal{L}(\hat{\theta}_{\mathrm{ML}}) + 1/2$. Graphiquement, il suffit de tracer une ligne horizontale à 0,5 au dessus du minimum de $-\ln\mathcal{L}$ pour avoir une estimation de l'écart-type (voir figure 6.6).

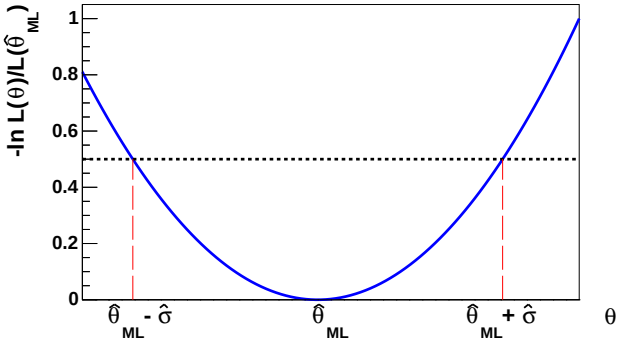


FIGURE 6.6 – Méthode graphique pour la détermination des incertitudes sur les estimateurs ML.

Il existe un cas particulièrement intéressant qui est celui dans lequel le *likelihood* est gaussien d'écart-type connu :

$$-\ln\mathcal{L}(\mu) = \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}$$

Introduisons la moyenne empirique $M = (1/n)\sum_{i=1}^{n} X_i$:

$$-\ln\mathcal{L}(\mu) = \sum_{i=1}^{n}\frac{(X_i-M+M-\mu)^2}{2\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-M)^2}{2\sigma^2} + \sum_{i=1}^{n}\frac{(M-\mu)^2}{2\sigma^2}$$

où nous avons utilisé $\sum_i (X_i-M)(M-\mu) = 0$. Donc :

$$-\ln\mathcal{L}(\mu) = -\ln\mathcal{L}(M) + \frac{(\mu-M)^2}{2(\sigma/\sqrt{n})^2}$$

L'approximation 6.14 est donc, dans le cas gaussien, une égalité exacte. C'est quelque chose que nous aurions pu voir directement car nous savons que le développement limité de $-\ln\mathcal{L}$ dans le cas gaussien s'arrête à l'ordre 2, que M est un estimateur efficace de μ (voir 6.2) et que sa variance est indépendante de μ et égale à σ^2/n (voir 4.4). Il est important toutefois de se souvenir que, d'une manière générale, cette méthode est approximative et ne doit être utilisée que s'il n'est pas possible de faire autrement.

Cas bidimensionnel

Répétons le raisonnement précédent au cas bidimensionnel : $\theta = \begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}$. Le développement limité de $-\ln\mathcal{L}(\theta)$ devient (en notant $\hat{\theta}$ l'estimateur ML afin d'alléger les notations)

$$-\ln\mathcal{L}(\theta) \simeq -\ln\mathcal{L}(\hat{\theta}) + \frac{1}{2}\sum_{i,j=1}^{2}\left(\theta_i-\hat{\theta}_i\right)\left(\theta_j-\hat{\theta}_j\right)\left.\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}}$$

ou, sous matricielle

$$-\ln\mathcal{L}(\theta) \simeq -\ln\mathcal{L}(\hat{\theta}) + \frac{1}{2}\left(\theta-\hat{\theta}\right)^T U^{-1}\left(\theta-\hat{\theta}\right) \tag{6.15}$$

avec

$$U^{-1} = \begin{pmatrix}\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_1^2} & \frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_1\partial\theta_2}\\ \frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_1\partial\theta_2} & \frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_2^2}\end{pmatrix}_{\theta=\hat{\theta}}$$

Dans la suite nous ferons l'hypothèse que le *likelihood* est gaussien en θ[3]. La matrice U est donc indépendante de θ (la condition $\theta=\hat{\theta}$ peut par conséquent être supprimée) et nous pouvons montrer que c'est la matrice de covariance des estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$. En effet, $\hat{\theta}_1$ et $\hat{\theta}_2$ sont des estimateurs non biaisés et efficaces. Donc (nous généralisons ici directement à deux dimensions le résultat unidimensionnel trouvé dans la section 6.3.1, propriété 2)

$$\mathbb{E}\left[\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_i\partial\theta_j}\right] = \left.\frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_i\partial\theta_j}\right|_{\theta=\hat{\theta}} = \frac{\partial^2(-\ln\mathcal{L})}{\partial\theta_i\partial\theta_j}$$

pour tout $i,j=1,2$. Nous pouvons donc écrire

$$U = \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix} \qquad \text{et} \qquad U^{-1} = \frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}$$

avec $\sigma_1^2 = \mathrm{var}\left[\hat{\theta}_1\right]$, $\sigma_2^2 = \mathrm{var}\left[\hat{\theta}_2\right]$ et $\rho = \mathrm{cov}\left(\hat{\theta}_1,\hat{\theta}_2\right)/(\sigma_1\sigma_2)$.

Cherchons l'ensemble des valeurs θ pour lesquelles $-\ln\mathcal{L}(\theta)$ est constant (nous noterons dorénavant cette quantité $-\ln\mathcal{L}$). D'après 6.15, nous voyons que ces valeurs sont définies par

$$\theta^T U^{-1}\theta - 2\hat{\theta}^T U^{-1}\theta + \hat{\theta}^T U^{-1}\hat{\theta} - 2\left(\ln\mathcal{L}(\hat{\theta}) - \ln\mathcal{L}\right) = 0$$

3. On peut montrer que c'est le cas dans la limite asymptotique.

Cette équation est de la forme

$$\theta^T U^{-1}\theta + d^T\theta + c = 0 \tag{6.16}$$

avec

- $d^T = -2\hat{\theta}^T U^{-1}$
- $c = \hat{\theta}^T U^{-1}\hat{\theta} - 2\left(\ln\mathcal{L}(\hat{\theta}) - \ln\mathcal{L}\right)$ est une constante.

6.16 est l'équation cartésienne générale d'une conique. Il s'agit ici d'une ellipse centrée sur $\left(\hat{\theta}_1, \hat{\theta}_2\right)$ et faisant un angle ϕ donné par [4]

$$\tan 2\phi = \left|\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right|$$

Cette ellipse s'appelle l'ellipse des uncertitudes car elle permet, en traçant les tangentes verticales et horizontales, de déterminer les incertitudes σ_1 et σ_2 respectivement. En voici la démonstration dans le cas des tangentes verticales (la démonstration est totalement similaire dans le cas des tangentes horizontales). Considérons la droite verticale passant par $\begin{pmatrix}\hat{\theta}_1 + m\sigma_1 \\ \hat{\theta}_2\end{pmatrix}$, où m est un nombre réel arbitraire. Cherchons le point appartenant à l'ellipse qui intersecte cette droite et notons $\underline{\theta}_2$ sa coordonnée suivant l'axe des ordonnées. Ce point appartient à l'ellipse si

$$\frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}\begin{pmatrix}m\sigma_1 & \underline{\theta}_2 - \hat{\theta}_2\end{pmatrix}\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}\begin{pmatrix}m\sigma_1 \\ \underline{\theta}_2 - \hat{\theta}_2\end{pmatrix} - 2\left(\ln\mathcal{L}(\hat{\theta}) - \ln\mathcal{L}\right) = 0$$

Ceci conduit à l'équation du second degré

$$\underline{\theta}_2^2 - 2\left(m\rho\sigma_2 + \hat{\theta}_2\right)\underline{\theta}_2 + m^2\sigma_2^2 + 2m\rho\sigma_2\hat{\theta}_2 + \hat{\theta}_2^2 - 2\sigma_2^2\left(1-\rho^2\right)\left(\ln\mathcal{L}(\hat{\theta}) - \ln\mathcal{L}\right) = 0$$

dont le discriminant est

$$\Delta = 4\sigma_2^2\left(1-\rho^2\right)\left[2\left(\ln\mathcal{L}(\hat{\theta}) - \ln\mathcal{L}\right) - m^2\right]$$

4. Pour montrer ces résultats, il est utile de se souvenir que l'équation cartésienne d'une conique sous forme matricielle est

$$\theta^T G\theta + d^T\theta + c = 0$$

avec

- $\theta = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}$
- $G = \begin{pmatrix}\alpha & \beta/2 \\ \beta/2 & \gamma\end{pmatrix}$, où α, β et γ sont des réels
- $d = \begin{pmatrix}d_1 \\ d_2\end{pmatrix}$ où d_1 et d_2 sont des réels
- c constant

et que c'est une ellipse si $\Delta = \beta^2 - 4\alpha\gamma < 0$, le centre de l'ellipse étant donné par

$$\begin{pmatrix}\theta_1^c \\ \theta_2^c\end{pmatrix} = \frac{1}{\Delta}\begin{pmatrix}2\gamma d_1 - \beta d_2 \\ 2\alpha d_2 - \beta d_1\end{pmatrix}$$

et son angle ϕ par

$$\tan 2\phi = \left|\frac{\beta}{\gamma - \alpha}\right|$$

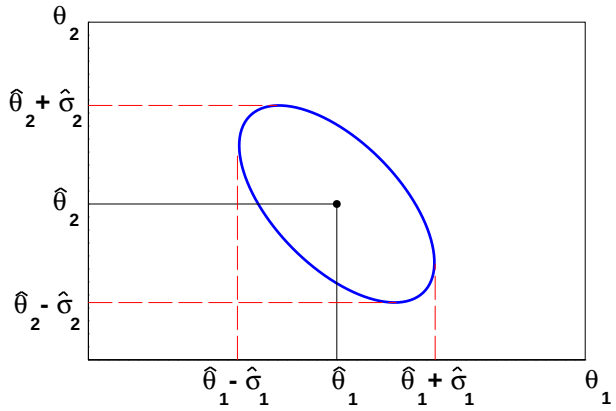


FIGURE 6.7 – Méthode graphique pour la détermination des incertitudes sur les estimateurs ML.

Puisque $1-\rho^2 \leq 1$, le signe de Δ est complètement déterminé par le signe du terme entre crochet. Les trois cas possibles sont

- $-\ln\mathcal{L} > -\ln\mathcal{L}(\hat{\theta}) + \frac{1}{2}m^2$: le discriminant est positif et il y a deux points sur la droite qui intersectent l'ellipse
- $-\ln\mathcal{L} = -\ln\mathcal{L}(\hat{\theta}) + \frac{1}{2}m^2$: le discriminant est nul et il y a un point sur la droite qui intersecte l'ellipse
- $-\ln\mathcal{L} < -\ln\mathcal{L}(\hat{\theta}) + \frac{1}{2}m^2$: le discriminant est négatif et il n'y a aucun point sur la droite qui intersecte l'ellipse

La droite dans le deuxième cas est donc la tangente verticale à l'ellipse. Dans le cas particulier où $m = 1$, la droite est tangente à l'ellipse $-\ln\mathcal{L} = -\ln\mathcal{L}(\hat{\theta}) + 1/2$. Dans la pratique il suffit donc de tracer cette ellipse puis ses tangentes pour remonter aux incertitudes σ_1 et σ_2.

6.3.3 Estimateur ML dans le cas d'un mélange

Considérons le cas d'un mélange (voir section 5.3). Dans le cas non *binné*, nous avions trouvé pour le *likelihood* étendu l'expression suivante :

$$\ln\mathcal{L}_{\text{ext}}\left(\{N_p\},\theta\right) = \sum_{i=1}^{n} \ln\left[\sum_{p=1}^{P} N_p f_p(x_i;\theta)\right] - \sum_{p} N_p$$

où n est la taille totale de l'échantillon, P le nombre total de processus, N_p le nombre d'éléments dans l'échantillon issus du processus p et $f_p(x_i;\theta)$ la densité de probabilité de la variable X pour le processus p.

Soit $N = \sum_p N_p$ le nombre total d'événements attendus et $\nu_p = N_p/N$ ($\sum \nu_p = 1$) la fraction issue

du processus p. Introduisons ces variables dans le *likelihood* :

$$\ln \mathcal{L}_{\text{ext}}\left(N, \{\nu_p\}, \theta\right) = \sum_{i=1}^{n} \ln \left[N \sum_{p=1}^{P} \nu_p f_p(x_i;\theta) \right] - N$$

Les estimateurs $\hat{N}_{\text{ML}}$, $\hat{\nu_p}_{\text{ML}}$ et $\hat{\theta}_{\text{ML}}$ sont solution de

$$\frac{\partial \ln \mathcal{L}_{\text{ext}}}{\partial N} = 0, \quad \frac{\partial \ln \mathcal{L}_{\text{ext}}}{\partial \nu_p} = 0 \quad \text{et} \quad \frac{\partial \ln \mathcal{L}_{\text{ext}}}{\partial \theta} = 0$$

La première équation conduit à $\hat{N}_{\text{ML}} = n$. Les deux autres équations conduisent aux mêmes $\hat{\nu_p}_{\text{ML}}$ et $\hat{\theta}_{\text{ML}}$ que ce que l'on obtiendrait avec le *likelihood* non étendu (nous laissons au lecteur le soin de le vérifier). Ces conclusions, obtenues dans le cas non *binné*, sont aussi valables dans le cas *binné*.

6.4 Méthode des moindres carrés

La méthode des moindres carrés est très souvent utilisée dans les problèmes où les observations $\{Y_i\}$ ($i \in [1, n]$) ont :

- des espérances $m_i(\theta) = \mathbb{E}\left[Y_i\right]$ différentes et inconnues (puisqu'elles dépendent des paramètres à estimer θ),
- une matrice de covariance V connue (à défaut de connaître la matrice de covariance de la population, nous pouvons souvent nous contenter d'une estimation faite à partir de l'échantillon).

$$V = \begin{pmatrix} \sigma_1^2 & & \dots & & \text{cov}\left(Y_1, Y_n\right) \\ & \ddots & & & \\ \vdots & & \text{cov}\left(Y_i, Y_j\right) & & \vdots \\ & & & \ddots & \\ \text{cov}\left(Y_n, Y_1\right) & & \dots & & \sigma_n^2 \end{pmatrix}$$

  en notant $\sigma_i^2 = \text{var}\left[Y_i\right]$

Dans la suite nous noterons parfois $m_i = m(x_i, \theta)$ pour faire apparaître explicitement la variable x_i dont dépend l'espérance.

Le lecteur veillera à ne pas confondre les notations utilisées dans cette section avec celles utilisées dans les autres sections. Ici, x_i n'est pas une variable aléatoire. Afin de limiter la confusion, nous ne notons donc pas l'échantillon $\{X_i\}$ mais $\{Y_i\}$.

La méthode des moindres carrés consiste à prendre pour estimateur les valeurs telles que la grandeur suivante est minimale

$$\chi^2(\theta) = (y - m)^T V^{-1} (y - m) \tag{6.17}$$

Les estimateurs obtenus par cette méthode seront appelés "estimateurs MC" ou "MCE". Nous les indicerons d'un "MC" (e.g. $\hat{\theta}_{\text{MC}}$). Nous avons donc, par définition,

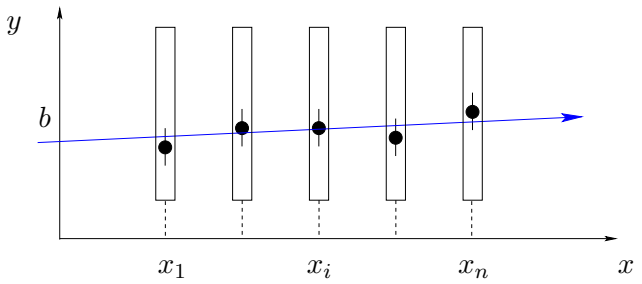
$$\left.\frac{\partial \chi^2}{\partial \theta}\right|_{\theta=\hat{\theta}_{\text{MC}}} = 0 \quad \text{et} \quad \left.\frac{\partial^2 \chi^2}{\partial \theta^2}\right|_{\theta=\hat{\theta}_{\text{MC}}} > 0$$

Dans la suite, nous considérerons souvent le cas particulier où les observations sont indépendantes. La matrice de covariance est alors diagonale et l'expression du χ^2 devient

$$\chi^2(\theta) = \sum_{i=1}^{n} \frac{(y_i - m(x_i, \theta))^2}{\sigma_i^2} \tag{6.18}$$

L'utilisation du symbole χ n'est pas anodine. Il provient du fait que lorsque les Y_i sont gaussiens le χ^2 est distribué suivant une loi de khi carré. Il est important de noter toutefois que cette méthode est employée également dans le cas non gaussien (χ^2 n'est alors plus distribué suivant une loi de khi carré).

Exemple 6.7: La méthode des moindres carrés peut être utilisée par exemple pour reconstruire la trajectoire d'une particule. Considérons une trajectoire rectiligne mesurée par un détecteur composé de n plans comme représenté sur la figure suivante.



Soit x_i la position du plan de détection i ($i = 1, 2, ..., n$) et Y_i la position mesurée dans ce plan. Nous faisons l'hypothèse que les Y_i sont indépendants et distribués de manière gaussienne. De plus, nous supposons que l'incertitude de mesure (notée σ) est la même dans tous les plans de détection. La pente et l'ordonnée à l'origine de la droite recherchée sont notées a et b respectivement. L'espérance de Y_i est :

$$m(x_i, a, b) = ax_i + b$$

Les estimateurs de a et b s'obtiennent en minimisant

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - m(x_i, a, b))^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(y_i - ax_i - b)^2}{\sigma^2}$$

Nous obtenons

$$\left.\frac{\partial \chi^2}{\partial a}\right|_{a=\hat{a}_{\mathrm{MC}}, b=\hat{b}_{\mathrm{MC}}} = 0 \Rightarrow \hat{a}_{\mathrm{MC}} = \frac{\sum_{i=1}^{n} \left(x_i y_i - \hat{b}_{\mathrm{MC}} x_i\right)}{\sum_{i=1}^{n} x_i^2}$$

$$\left.\frac{\partial \chi^2}{\partial b}\right|_{a=\hat{a}_{\mathrm{MC}}, b=\hat{b}_{\mathrm{MC}}} = 0 \Rightarrow \hat{b}_{\mathrm{MC}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{a}_{\mathrm{MC}} x_i)$$

En résolvant ce système, nous trouvons

$$\hat{a}_{\mathrm{MC}} = \frac{n\sum_i x_i y_i - \sum_i \sum_j x_i y_j}{n\sum_i x_i^2 - \sum_i \sum_j x_i x_j}$$

$$\hat{b}_{\mathrm{MC}} = \frac{\sum_i \sum_k x_i^2 y_k - \sum_i \sum_k x_i y_i x_k}{n\sum_i x_i^2 - \sum_i \sum_j x_i x_j}$$

6.4.1 Lien avec la méthode du maximum de vraisemblance

Considérons le cas où les données $\{Y_i\}$ sont distribuées de manière gaussienne. Le *likelihood* est

$$\mathcal{L} = e^{-\frac{1}{2}(y-m)^T V^{-1}(y-m)}$$

dans le cas général et

$$\mathcal{L} = \prod_{i=1}^{n} e^{-\frac{\left(y_i - m(x_i,\theta)\right)^2}{2\sigma_i^2}}$$

dans le cas où les observations sont indépendantes.

Nous voyons immédiatement que $-2\ln\mathcal{L} = \chi^2$. Dans le cas gaussien, les méthodes du maximum de vraisemblance et des moindres carrés sont donc équivalentes.

6.4.2 Méthode des moindres carrés avec un échantillon *binné*

La méthode des moindres carrés peut être utilisée avec un échantillon *binné*, représenté par un histogramme $\{C_k; Y_k\}$ ($k \in [1, n]$) (voir section 4.3) issu du tirage d'une variable aléatoire X distribuée suivant $f_X(x;\theta)$. Soit $m_k(\theta)$ l'espérance de Y_k : $m_k(\theta) = \mathbb{E}[Y_k]$. $m_k(\theta)$ tire sa dépendance en θ de $p_k(\theta) = \int_{C_k} f_X(x;\theta)\mathrm{d}x$. Dans le cas où la taille totale de l'échantillon N ($N = \sum_{k=1}^{n} Y_k$) est considérée comme fixe nous avons

$$m_k(\theta) = N p_k(\theta)$$

alors que dans le cas où elle est considérée comme une variable aléatoire d'espérance ν nous avons

$$m_k(\theta) = \nu p_k(\theta)$$

Dans la suite nous faisons l'hypothèse que les Y_k sont indépendants et suivent chacun une loi de Poisson de paramètre m_k. Cette hypothèse est, comme nous l'avons vu en 4.3, valable lorsque la taille de l'échantillon N est considérée comme une variable aléatoire (plutôt que comme un paramètre fixe). Elle est aussi valable lorsque N est fixe et que le nombre d'entrées attendues dans chaque *bin*

est beaucoup plus petit que le nombre total d'entrées (en effet, dans ce cas la loi multinomiale qui gouverne les Y_k peut être approximée par une loi de Poisson pour chaque Y_k). Nous avons

$$\chi^2 = \sum_{k=1}^{n} \frac{(y_k - m_k(\theta))^2}{m_k(\theta)} \tag{6.19}$$

puisque dans ce cas $\sigma_k = \sqrt{m_k}$. Ce χ^2 est appelé traditionnellement le χ^2 de Pearson.

L'équation 6.19 peut être difficile à manipuler et il est souvent plus pratique d'utiliser l'approximation

$$\chi^2 = \sum_{k=1}^{n} \frac{(y_k - m_k(\theta))^2}{y_k} \tag{6.20}$$

c'est-à-dire que plutôt que d'utiliser la vraie valeur de la variance nous utilisons son estimateur. Ce χ^2 est appelé traditionnellement le χ^2 de Neyman. Lorsqu'il est utilisé, la méthode des moindres carrés est qualifiée de "modifiée".

La méthode des moindres carrés sur un échantillon *binné* doit être appliquée avec beaucoup de prudence lorsque la taille de l'échantillon est considérée comme une variable aléatoire (comme ci-dessus, nous noterons $\nu = \mathbb{E}[N]$). En effet, dans ce cas $\hat{\nu}_{\mathrm{MC}}$ peut être très différent de la valeur observée N. Considérons tout d'abord le cas non modifié

$$\chi^2 = \sum_{k=1}^{n} \frac{(y_k - \nu p_k)^2}{\nu p_k} = \frac{1}{\nu} \sum_{k=1}^{n} \frac{y_k^2}{p_k} + \nu - 2N$$

où nous avons utilisé $\sum_k p_k = 1$ et $\sum_k y_k = N$. Nous trouvons

$$\frac{\partial \chi^2}{\partial \nu} = -\frac{1}{\nu^2} \sum_k \frac{y_k^2}{p_k} + 1 = -\frac{1}{\nu} \left(\chi^2 + 2N - \nu \right) + 1 = \frac{-\chi^2 - 2N + 2\nu}{\nu}$$

Soit

$$\hat{\nu}_{\mathrm{MC}} = N + \frac{\chi^2}{2}$$

Un calcul similaire permet de montrer qu'avec la méthode modifiée

$$\hat{\nu}_{\mathrm{MC}} = N - \chi^2$$

Nous nous attendons à avoir typiquement $\chi^2 \approx n$. L'incertitude relative sur $\hat{\nu}_{\mathrm{MC}}$ est donc typiquement de $n/(2N)$ dans le cas non modifié et n/N dans le cas modifié. Nous pourrons retenir qu'approximativement

$$\text{incert. relative} \approx \frac{1}{\text{nombre moyen d'entrées par bin}}$$

6.4.3 Méthode des moindres carrés linéaire

La méthode des moindres carrés est particulièrement bien adaptée au cas où $m(x_i, \theta)$ dépend de manière linéaire des paramètres θ. En effet, elle fournit dans ce cas une solution analytique pour les estimateurs et ces derniers sont efficaces et non biaisés.

Notons

$$m(x_i, \theta) = \sum_{j=1}^{q} a_j(x_i)\theta_j$$

où q est le nombre de paramètres et les a_j sont des fonctions de x linéairement indépendantes. Afin d'alléger les notations, introduisons la matrice $A_{ij} = a_j(x_i)$

$$m(x_i, \theta) = A\theta$$

L'expression 6.17 s'écrit dans ce cas

$$\chi^2(\theta) = (y - A\theta)^T V^{-1} (y - A\theta) \tag{6.21}$$

et les estimateurs sont solutions du système

$$\frac{\partial \chi^2}{\partial \theta_p} = -2 \left(A_p^T V^{-1} y - A_p^T V^{-1} A\theta \right)$$

où $A_p = A_{ip}$ (i est l'indice sur lequel porte la somme dans l'expression ci-dessus). En explicitant les produits matriciels et en utilisant la convention d'Einstein (sommation sur les indices répétés)

$$\frac{\partial \chi^2}{\partial \theta_p} = -2 \left(A_{ip} \left(V^{-1}\right)_{ij} y_j - A_{ip} \left(V^{-1}\right)_{ij} A_{jk}\theta_k \right)$$

Si la matrice $A_p V^{-1} A$ est inversible, les estimateurs MC sont

$$\hat{\theta}_{\mathrm{MC}} = \left(A^T V^{-1} A\right)^{-1} A^T V^{-1} y = By$$

où nous avons défini $B = \left(A^T V^{-1} A\right)^{-1} A^T V^{-1}$. Les estimateurs MC sont donc des fonctions linéaires des observations y. La matrice de covariance U des estimateurs a pour éléments (en supprimant l'indice MC pour alléger les notations)

$$\mathrm{cov}\left(\hat{\theta}_i, \hat{\theta}_j\right) = \mathbb{E}\left[\hat{\theta}_i \hat{\theta}_j\right] - \mathbb{E}\left[\hat{\theta}_i\right] \mathbb{E}\left[\hat{\theta}_j\right] = B_{ik} B_{jl} \mathrm{cov}\left(Y_k, Y_l\right)$$

Donc

$$U = BVB^T = \left(A^T V^{-1} A\right)^{-1} \tag{6.22}$$

Nous voyons aussi à partir de 6.21 que

$$\left(U^{-1}\right)_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \tag{6.23}$$

Dans le cas où les observations sont gaussiennes, nous retrouvons ainsi la limite RCF (rappelons que dans ce cas $\chi^2 = -2 \ln \mathcal{L}$).

Comme le montrent les équations 6.22 et 6.23, $\partial^2\chi^2/(\partial\theta_i\partial\theta_j)$ est indépendant de θ. Le développement limité de χ^2 s'arrête donc à l'ordre 2 :

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + \frac{1}{2}\sum_{i,j=1}^{q}\left(\theta_i - \hat{\theta}_i\right)\left(\theta_j - \hat{\theta}_j\right)\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}$$

où, sous forme matricielle

$$\chi^2(\theta) = \chi^2(\hat{\theta}) + \left(\theta - \hat{\theta}\right)^T U^{-1}\left(\theta - \hat{\theta}\right) \tag{6.24}$$

Nous nous trouvons ici dans la même situation que celle rencontrée en 6.3.2. Dans le cas bidimensionnel, cette équation est celle d'une ellipse de centre $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2\right)$ et le tracé des tangentes de l'ellipse $\chi^2 = \chi^2(\hat{\theta}) + 1$ permet de remonter aux incertitudes sur les paramètres.

6.5 Méthode des moments

La méthode des moments consiste à prendre pour estimateur les valeurs telles que les moments soient égaux aux moments empiriques. Elle est justifiée par la loi des grands nombres qui stipule que les moments empiriques tendent vers les moments lorsque la taille de l'échantillon tend vers l'infini. Les estimateurs obtenus par cette méthode seront appelés "estimateurs MM" ou "MME". Nous les indicerons d'un MM (e.g. $\hat{\theta}_{\mathrm{MM}}$). Dans le cas à un seul paramètre nous avons donc, par définition :

$$\mathbb{E}\left[X\right]\big|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \int x f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6.25}$$

Dans le cas où le nombre de paramètres est $q > 1$, il faut résoudre le système de q équations suivant :

$$\begin{cases} \mathbb{E}\left[X\right]\big|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int x f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} x_i}{n} \\ \mathbb{E}\left[X^2\right]\big|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int x^2 f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} x_i^2}{n} \\ \qquad\vdots \qquad\qquad\qquad \vdots \\ \mathbb{E}\left[X^q\right]\big|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int x^q f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} x_i^q}{n} \end{cases} \tag{6.26}$$

Exemple 6.8: Considérons la loi binomiale

$$p(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Soit K_i ($i \in [1, l]$) un ensemble de variables aléatoires distribuées suivant cette loi. Nous savons que $\mathbb{E}[K_i] = np$ et que $\mathbb{E}\left[K_i^2\right] = \mathrm{var}[K_i] + \mathbb{E}[K_i]^2 = np(1-p) + (np)^2$. Les estimateurs MM sont donc donnés par :

$$\begin{cases} \hat{n}_{\mathrm{MM}}\hat{p}_{\mathrm{MM}} = \dfrac{\sum_{i=1}^{l} k_i}{l} \\ \hat{n}_{\mathrm{MM}}\hat{p}_{\mathrm{MM}}(1-\hat{p}_{\mathrm{MM}}) + (\hat{n}_{\mathrm{MM}}\hat{p}_{\mathrm{MM}})^2 = \dfrac{\sum_{i=1}^{l} k_i^2}{l} \end{cases}$$

Afin d'alléger les notations, introduisons $m_1 = (1/l)\sum_i k_i$ et $m_2 = (1/l)\sum_i k_i^2$:

$$\begin{cases} \hat{n}_{\mathrm{MM}}\hat{p}_{\mathrm{MM}} = m_1 \\ m_1(1-\hat{p}_{\mathrm{MM}}) + m_1^2 = m_2 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{n}_{\mathrm{MM}} = \dfrac{m_1^2}{m_1 + m_1^2 - m_2} \\ \hat{p}_{\mathrm{MM}} = \dfrac{m_1 + m_1^2 - m_2}{m_1} \end{cases}$$

Exemple 6.9: Considérons la loi uniforme

$$\begin{cases} f(x; a, b) = \dfrac{1}{b-a} \quad \text{pour } a \leq x \leq b \\ f(x; a, b) = 0 \quad \text{pour } x < a \text{ et } x > b \end{cases}$$

Soit X_i ($i \in [1, n]$) un ensemble de variables aléatoires distribuées suivant cette loi. Nous savons que $\mathbb{E}[X_i] = (a+b)/2$ et que $\mathbb{E}\left[X_i^2\right] = \mathrm{var}[X_i] + \mathbb{E}[X_i]^2 = (b-a)^2/12 + (a+b)^2/4 = (b^2 + a^2 + ab)/3$. Les estimateurs sont donc donnés par :

$$\begin{cases} \dfrac{\hat{a}_{\mathrm{MM}} + \hat{b}_{\mathrm{MM}}}{2} = \dfrac{\sum_{i=1}^{n} x_i}{n} \\ \dfrac{\hat{b}_{\mathrm{MM}}^2 + \hat{a}_{\mathrm{MM}}^2 + \hat{a}_{\mathrm{MM}}\hat{b}_{\mathrm{MM}}}{3} = \dfrac{\sum_{i=1}^{n} x_i^2}{n} \end{cases}$$

Posons $m_1 = (1/n)\sum_{i=1}^{n} x_i$ et $m_2 = (1/n)\sum_{i=1}^{n} x_i^2$:

$$\begin{cases} \hat{a}_{\mathrm{MM}} = 2m_1 - \hat{b}_{\mathrm{MM}} \\ \hat{b}_{\mathrm{MM}}^2 - 2m_1\hat{b}_{\mathrm{MM}} + 4m_1^2 - 3m_2 = 0 \end{cases}$$

Les solutions de la seconde équation sont :

$$\hat{b}_{\mathrm{MM}} = m_1 \pm \sqrt{3(m_2 - m_1^2)}$$

Comme $\hat{b}_{\mathrm{MM}} \geq m_1$, la bonne solution est celle avec le signe $+$. Finalement,

$$\begin{cases} \hat{a}_{\mathrm{MM}} = m_1 - \sqrt{3(m_2 - m_1^2)} \\ \hat{b}_{\mathrm{MM}} = m_1 + \sqrt{3(m_2 - m_1^2)} \end{cases}$$

6.5.1 Méthode des moments généralisée

La méthode des moments peut se généraliser en considérant non plus les moments de l'observable X mais un ensemble de fonctions linéairement indépendantes $a_j(X)$ ($j \in [1, q]$, q étant est le nombre de paramètres). Les estimateurs $\hat{\theta}_{\mathrm{MM}}$ sont obtenus en résolvant le système suivant :

$$\begin{cases} \mathbb{E}\left[a_1(X)\right]|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int a_1(x) f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} a_1(x_i)}{n} \\ \mathbb{E}\left[a_2(X)\right]|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int a_2(x) f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} a_2(x_i)}{n} \\ \qquad\qquad\vdots \qquad\qquad\qquad \vdots \\ \mathbb{E}\left[a_q(X)\right]|_{\theta=\hat{\theta}_{\mathrm{MM}}} = \displaystyle\int a_q(x) f(x;\hat{\theta}_{\mathrm{MM}})\mathrm{d}x = \frac{\sum_{i=1}^{n} a_q(x_i)}{n} \end{cases} \tag{6.27}$$

Ceci peut être formulé de manière équivalente en utilisant les variables centrées $c_j(X,\theta) = a_j(X) - \mathbb{E}\left[a_j(X)\right]$ (la dépendance de c_j en θ, qui provient de la dépendance de $\mathbb{E}\left[a_j(X)\right]$ en θ, est ici notée

explicitement pour plus de clarté) :

$$\begin{cases} \sum_{i=1}^{n} c_1(x_i, \hat{\theta}_{\mathrm{MM}}) = 0 \\ \sum_{i=1}^{n} c_2(x_i, \hat{\theta}_{\mathrm{MM}}) = 0 \\ \qquad\qquad \vdots \\ \sum_{i=1}^{n} c_q(x_i, \hat{\theta}_{\mathrm{MM}}) = 0 \end{cases} \tag{6.28}$$

Une généralisation encore plus grande peut être faite en considérant des fonctions $a_j(X, \theta)$ qui dépendent de X et θ et ont une espérance nulle. Dans ce cas nous avons $c_j(X, \theta) = a_j(X, \theta)$ et les estimateurs MM s'obtiennent en résolvant, comme précédemment, le système 6.28.

6.5.2 Lien avec la méthode du maximum de vraisemblance

Un cas particulier de cette méthode correspond au choix $c_j(X, \theta) = \frac{\partial \ln f(X;\theta)}{\partial \theta_j}$ (un calcul analogue à celui réalisé en 6.2.3 permet de montrer que $\mathbb{E}[c_j(X, \theta)] = 0, \forall j \in [1, q]$). Dans ce cas nous avons, pour tout $j \in [1, q]$,

$$\sum_{i=1}^{n} \frac{\partial \ln f(x_i; \hat{\theta}_{\mathrm{MM}})}{\partial \theta_j} = \frac{\partial \left(\sum_{i=1}^{n} \ln f(x_i; \hat{\theta}_{\mathrm{MM}}) \right)}{\partial \theta_j} = \left. \frac{\partial \ln \mathcal{L}}{\partial \theta_j} \right|_{\theta = \hat{\theta}_{\mathrm{MM}}} = 0$$

Pour ce choix particulier de $c_j(X, \theta)$, la méthode des moments généralisée est donc équivalente à la méthode du maximum de vraisemblance.

Chapitre 7

Test d'hypothèse

Plusieurs objectifs peuvent être poursuivis lorsqu'on réalise une expérience. Dans certains cas, nous pouvons par exemple vouloir déterminer à l'aide des données expérimentales un ou plusieurs paramètres de la théorie sous-jacente. C'est le cas que nous avons discuté dans le chapitre sur l'estimation des paramètres. Ceci suppose que nous connaissions la théorie sous-jacente. Dans d'autres cas, nous ne connaissons pas la théorie sous-jacente et l'objectif est de la déterminer. C'est de ce dernier cas que nous allons traiter dans ce chapitre.

Déterminer la théorie sous-jacente ne signifie en réalité pas grand chose. Nous ne pouvons en effet jamais arriver à une conclusion définitive sur la nature de cette théorie. Il n'est par exemple pas possible de répondre à une question du type "Le modèle standard de la physique des particules est-il le modèle qui décrit les particules et leurs interactions ?". Nous pouvons par contre chercher à répondre à la question "telle observation est-elle compatible avec le modèle standard de la physique des particules ou non ?" ou bien "les données sont-elles plus en faveur du modèle standard ou bien du modèle supersymétrique ?". Ainsi, ce que nous faisons n'est pas de déterminer la théorie sous-jacente mais d'évaluer la compatibilité entre une ou plusieurs hypothèses sur la théorie sous-jacente et les observations. L'ensemble des méthodes mises en œuvre pour réaliser ces évaluations est regroupé sous l'appelation "test d'hypothèse".

Nous ne discutons dans ce chapitre que des tests d'hypothèses classiques (ou fréquentistes). Le cas bayésien est discuté dans le chapitre 9.

7.1 Principe

La réalisation d'un test d'hypothèse classique (ou fréquentiste) procède schématiquement de la manière suivante :

1. Choix d'une variable résumant les données (dans la suite nous l'appelerons variable test)
2. Détermination de la distribution de la variable test sous l'hypothèse à tester
3. Calcul de la valeur observée de la variable test
4. Confrontation de la valeur observée à la distribution trouvée dans l'étape 2

À l'issue de la dernière étape, nous pouvons conclure sur l'accord entre les données observées et l'hypothèse considérée. Par exemple, si la valeur observée de la variable test se trouve très loin dans

la queue de la distribution alors nous pourrons conclure que l'hypothèse considérée n'est pas favorisée par les données. Si par contre elle se trouve proche du mode de la distribution, nous pourrons conclure que les données observées et l'hypothèse sont en bon accord.

La procédure décrite ci-dessus s'applique également au cas où plusieurs hypothèses sont testées. Il suffit de déterminer autant de distributions que d'hypothèses à tester et de comparer la valeur observée à chacune de ces distributions.

En guise d'exemple considérons une expérience de comptage décrite par une loi de Poisson. Soit n le nombre d'événements observés et ν le nombre d'événements moyens attendus :

$$P(n|\nu) = \frac{\nu^n}{n!} e^{-\nu}$$

Le choix de la variable test (étape 1) est dans ce cas assez immédiat. Il suffit en effet de prendre n lui même. La détermination de la distribution de cette variable (étape 2) suppose que nous fassions une hypothèse sur la théorie sous-jacente. Soit H_0 cette hypothèse (par exemple le modèle standard de la physique des particules) et soit ν_0 la valeur de ν sous H_0. Dans la suite nous prendrons $\nu_0 = 3,47$. Si H_0 est vrai, n est distribué suivant $P(n|\nu_0)$ représenté sur la figure 7.1.

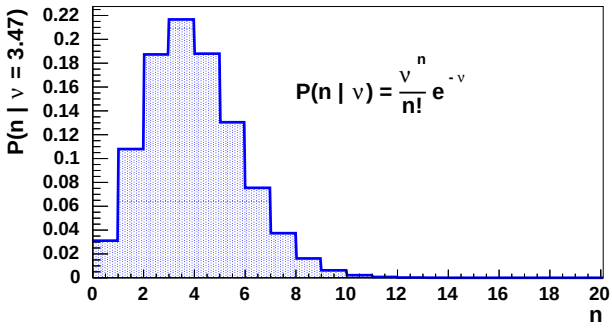


FIGURE 7.1 – Distribution de Poisson pour $\nu = \nu_0 = 3,47$.

La troisième étape du test d'hypothèse consiste à calculer la valeur observée de n (nous noterons cette valeur n_{obs}). Dans ce cas il n'y a en fait aucun calcul à faire puisque n_{obs} est directement le résultat de l'expérience. C'est la confrontation de cette valeur observée et de la distribution montrée sur la figure 7.1 qui permet de conclure quant à la validité de H_0 (étape 4). Supposons que $n_{\text{obs}} = 10$. La probabilité d'observer une telle valeur sous H_0 est très faible. H_0 doit donc être considéré comme défavorisé. Si par contre $n_{\text{obs}} = 2$, la probabilité sous H_0 est relativement grande. H_0 n'est dans ce cas pas défavorisé (on dit aussi que H_0 est corroboré).

Il arrive souvent que nous souhaitions confronter non pas une seule hypothèse aux observations mais plusieurs. Imaginons par exemple qu'en plus de H_0 nous considérions une autre hypothèse H_1 (correspondant par exemple à la supersymétrie) qui prédit une valeur différente pour ν. Une observation pourra permettre de favoriser une hypothèse plutôt que l'autre s'il lui correspond une grande probabilité dans une hypothèse et une faible dans l'autre.

Les étapes décrites ci-dessus et illustrées dans le cas de l'expérience de comptage sont précisées, approfondies et généralisées dans les sections suivantes.

7.2 Définitions

7.2.1 Variable test

Comme nous l'avons vu dans la section 7.1, il faut, pour réaliser un test d'hypothèse, disposer d'au moins une grandeur expérimentale dont la distribution puisse être prédite par la (ou les) théorie(s) testée(s). Cette grandeur (qui comme toute grandeur expérimentale est une variable aléatoire) s'appelle variable test (on parle souvent de test tout court [1]). Le test T est une fonction de l'échantillon ($T = T(X_1, \ldots, X_n)$) et sa distribution est connue si celle des observations $\{X_i\}$ ($i \in [1, n]$) est connue.

Soit H une hypothèse. La distribution du test T sous l'hypothèse H sera notée soit sous forme de probabilité conditionnelle

$$f(t|H)$$

soit sous forme abrégée

$$f_H(t) \quad \text{ou} \quad f_H$$

7.2.2 Hypothèse nulle et hypothèse alternative

Lorsque deux hypothèses sont confrontées à l'expérience, il est de coutume d'appeler l'une d'elles l'hypothèse nulle et l'autre l'hypothèse alternative. L'hypothèse nulle, notée H_0, est celle qui est considérée par défaut comme correspondant à la théorie sous-jacente. L'hypothèse alternative, notée H_1, est l'autre hypothèse. Il peut y avoir plusieurs hypothèses alternatives, auquel cas H_1 désigne non pas une seule hypothèse mais une famille d'hypothèses.

Cette catégorisation en hypothèses nulle et alternative est dans une grande mesure arbitraire. Les intervertir ne doit pas changer les conclusions. Le plus important est toujours de bien savoir de quelle hypothèse on parle.

7.2.3 *p-value*

La *p-value* (valeur p ou p-valeur en français) est la grandeur utilisée pour évaluer la compatibilité entre l'observation et une hypothèse donnée. C'est la probabilité pour que le test ait une valeur au moins aussi extrême que celle observée. Elle dépend bien entendu de l'hypothèse puisque la distribution du test en dépend.

Supposons que le test soit distribué suivant $f(t|H)$ (voir figure 7.2) et que la valeur observée du test soit t_{obs}. La *p-value* (représentée par l'aire colorée sur la figure) est

$$p\text{-}value = \int_{t_{\text{obs}}}^{\infty} f(t|H)\mathrm{d}t$$

Lorsque deux hypothèses sont considérées, la *p-value* pour une hypothèse donnée est calculée en intégrant la distribution du test pour cette hypothèse dans la direction où se trouve la distribution

1. Nous voyons ici que le mot test peut être employé dans un double sens. Il peut désigner soit la variable aléatoire qui est utilisée pour réaliser le test d'hypothèse soit le test d'hypothèse lui-même. Le sens devrait être clair en fonction du contexte.

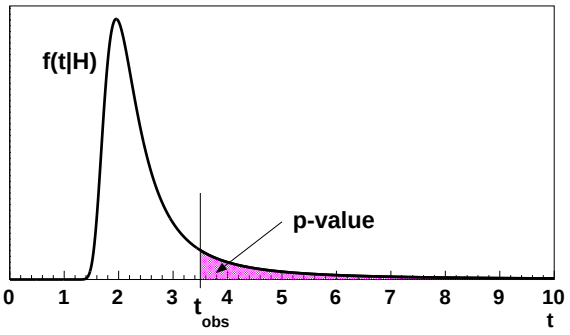


FIGURE 7.2 – Distribution du test sous l'hypothèse H et *p-value* .

pour l'autre hypothèse. Sur l'exemple de la figure 7.3, nous avons deux *p-values*

$$p_0 = \int_{t_{\text{obs}}}^{\infty} f(t|H_0)\mathrm{d}t$$

et

$$p_1 = \int_{-\infty}^{t_{\text{obs}}} f(t|H_1)\mathrm{d}t$$

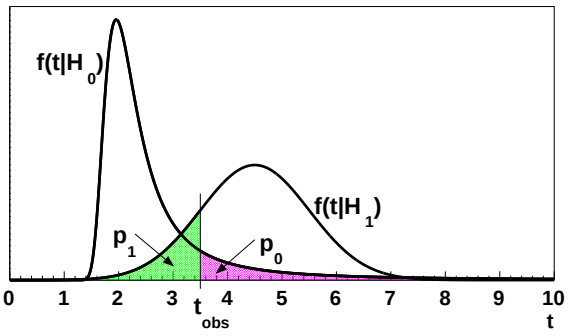


FIGURE 7.3 – Distribution du test sous H_0 et H_1 avec les *p-values* correspondant à une observation t_{obs}.

Il est important de noter que la *p-value* est une variable aléatoire (puisque t_{obs} en est une). Supposons que $t_{\text{obs}} \sim f_{H_0}$. La distribution $g(p_0|H_0)$ de p_0 sous cette hypothèse est

$$g(p_0|H_0) = f(t_{\text{obs}}|H_0)\frac{1}{\left|\mathrm{d}p_0/\mathrm{d}t_{\text{obs}}\right|} = 1$$

Nous obtenons ainsi le résultat important suivant : si l'hypothèse est la bonne alors la distribution de la *p-value* est uniforme.

7.2.4 Signification statistique

La signification statistique Z est définie par

$$Z = \Phi^{-1}(1 - p) \tag{7.1}$$

où p est la *p-value* et Φ^{-1} est le quantile (inverse de la fonction de répartition) de la loi normale centrée réduite (voir figure 7.4).

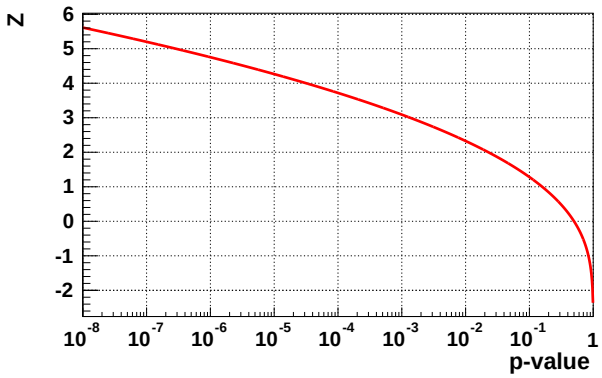


FIGURE 7.4 – Relation entre la signification statistique et la *p-value* .

La signification statistique est souvent plus pratique à manipuler que la *p-value* car elle s'étend sur beaucoup moins d'ordres de grandeur. La table 7.1 donne pour quelques *p-value* la signification correspondante.

p-value	Z
0.05	1,64
0.00135	3
$2,87 \times 10^{-7}$	5

TABLE 7.1 – Correspondance signification statistique - *p-value* .

7.2.5 Hypothèse simple et composée

Une hypothèse simple détermine entièrement la distribution du test. Une hypothèse composée ne détermine pas entièrement la distribution du test. Une hypothèse composée correspond à un ensemble d'hypothèses simples différents soit par la valeur d'un ou plusieurs paramètres soit par la famille à laquelle appartient la distribution du test. Dans la suite nous considérerons uniquement le premier cas.

Supposons par exemple une hypothèse suivant laquelle le test est distribué suivant une loi normale de moyenne et écart-type connus. Une telle hypothèse est simple. Si par contre au moins un des deux paramètres est inconnu, l'hypothèse est complexe (car pour connaître la distribution du test, il faut, en plus de spécifier l'hypothèse, choisir des valeurs pour les paramètres).

En fonction du problème traité, nous pouvons recontrer différents cas de figure : hypothèses nulle et alternative simples, hypothèse nulle simple et alternative composée (ou l'inverse) et hypothèse nulle et alternative composées.

Considérons par exemple le cas où l'hypothèse nulle est simple et l'hypothèse alternative composée. Supposons de plus que la distribution du test appartient dans les deux cas à la même famille, la différence provenant seulement de la valeur d'un ou plusieurs paramètres (θ). Notons θ_0 la valeur correspondant à l'hypothèse nulle et Ξ l'ensemble des valeurs correspondant à l'hypothèse alternative ($\theta_0 \notin \Xi$).

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \in \Xi$$

Lorsque Ξ correspond à l'ensemble des valeurs $\theta > \theta_0$ (ou $\theta < \theta_0$) le test est dit unilatéral. Lorsqu'il correspond à l'ensemble des valeurs $\theta \neq \theta_0$ le test est dit bilatéral.

7.2.6 Région critique et seuil de signification

On appelle région critique l'ensemble des valeurs du ou des tests pour lesquelles l'hypothèse nulle H_0 est rejetée. Dans le cas unidimensionnel où un seul test est calculé à partir de l'échantillon, la région critique est typiquement définie par une condition booléenne simple du type $t > t_{\text{seuil}}$ pour un test unilatéral et $t < t_{\text{seuil inf}}$ et $t > t_{\text{seuil sup}}$ pour un test bilatéral. Dans le cas multidimensionnel où plusieurs tests sont calculés, la région critique est délimitée par un contour dans l'espace des tests. Nous noterons de manière générale la région critique $\mathcal{C}$. Ainsi

- si $t_{\text{obs}} \in \mathcal{C}$ alors H_0 est rejetée
- si $t_{\text{obs}} \notin \mathcal{C}$ alors H_0 est acceptée (la région complémentaire de la région critique est parfois nommée région d'acceptance).

La région critique est en général choisit de telle sorte que la probabilité de tomber dedans sous l'hypothèse nulle soit égale à une valeur prédéfinie α :

$$\alpha = \int_{t \in \mathcal{C}} f(t|H_0)\mathrm{d}t$$

α est le seuil de signification du test.

7.2.7 Erreurs de première et seconde espèce, efficacité, puissance, réjection

Lorsqu'un test d'hypothèse est réalisé, il est évidemment possible de se tromper. Il y a deux façons de le faire. Nous pouvons soit rejeter l'hypothèse nulle alors qu'elle est vraie soit accepter cette hypothèse alors qu'elle est fausse. Nous sommes ainsi amenés à définir deux grandeurs :

- Erreur de première espèce = probabilité de rejeter l'hypothèse nulle si elle est vraie (elle est égale au seuil de signification α)
- Erreur de seconde espèce (β) = probabilité d'accepter l'hypothèse nulle si elle est fausse

$$\beta = \int_{t \notin \mathcal{C}} f(t|H_1)\mathrm{d}t$$

À partir de ces grandeurs on définit l'efficacité (notée ε et correspondant à la probabilité d'accepter H_0 si elle est vraie) et la puissance (ou réjection) (notée r et correspondant à la probabilité de rejeter H_0 si elle est fausse).

$$\varepsilon = 1 - \alpha \qquad \text{et} \qquad r = 1 - \beta$$

Toutes ces grandeurs sont résumées sur la figure 7.5 dans le cas unidimensionnel.

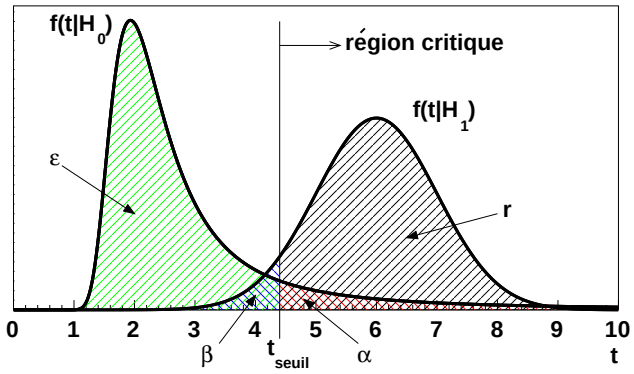


FIGURE 7.5 – Erreurs de première et seconde espèces, efficacité et puissance du test.

7.2.8 Pureté

Lorsqu'une même expérience est réalisée N fois, nous obtenons un ensemble de tests observés $\{t^1_{\text{obs}}, t^2_{\text{obs}}, \ldots, t^N_{\text{obs}}\}$. Il arrive couramment que ces différentes observations correspondent à deux hypothèses simples différentes (comme précédemment, nous les noterons H_0 et H_1). Ainsi, la distribution du test est une combinaison de $f(t|H_0)$ et $f(t|H_1)$:

$$f(t) = c \times f(t|H_0) + (1 - c) \times f(t|H_1)$$

où $c \leq 1$ est la probabilité qu'une observation corresponde à H_0 et $1 - c$ la probabilité qu'elle correspondent à H_1 :

$$c = P(H_0) \qquad \text{et} \qquad 1 - c = P(H_1)$$

Une question que nous sommes souvent amenés à nous poser est : quelle fraction des observations dans la région d'acceptance correspondent à l'hypothèse nulle ? La fraction recherchée porte le nom de pureté et correspond à $P(H_0|t \notin \mathcal{C})$. À l'aide du théorème de Bayes nous voyons que

$$\begin{aligned} P(H_0|t \notin \mathcal{C}) &= \frac{P(t \notin \mathcal{C}|H_0)P(H_0)}{P(t \notin \mathcal{C}|H_0)P(H_0) + P(t \notin \mathcal{C}|H_1)P(H_1)} \\ &= \frac{\varepsilon \times P(H_0)}{\varepsilon \times P(H_0) + \beta \times P(H_1)} \end{aligned}$$

ou de manière équivalente

$$P(H_0|t \notin \mathcal{C}) = \frac{1}{1 + \frac{\beta \times P(H_1)}{\varepsilon \times P(H_0)}} \tag{7.2}$$

La plupart du temps le coefficient c est connu et le seul paramètre sur lequel nous pouvons jouer pour augmenter la pureté est le rapport

$$\frac{\text{efficacité}}{\text{erreur de seconde espèce}}$$

L'équation 7.2 montre que pour maximiser la pureté il faut choisir la région critique de telle sorte à maximiser ce rapport.

7.3 Lemme de Neyman-Pearson

Dans la section précédente nous avons vu que pour maximiser la pureté il faut maximiser le rapport $(1-\alpha)/\beta$. Supposons que le seuil de signification α ait une valeur prédéfinie. Maximiser la pureté revient donc à maximiser la puissance. Le problème consiste maintenant à choisir la région critique de telle sorte à ce que cette puissance soit maximale. Dans le cas unidimensionnel (un seul test), ce choix est impossible car β est déterminé dès que α l'est. Dans le cas multidimensionnel (plusieurs tests), il peut y avoir plusieurs régions critiques ayant le même α.

Le lemme de Neyman-Pearson stipule que, dans le cas multidimensionnel, la région critique qui maximise la puissance (appelée région BCR [2] et notée $\mathcal{C}_{NP}$) est telle que

$$\frac{f(t|H_0)}{f(t|H_1)} \leq k_\alpha \tag{7.3}$$

où k_α est une constante qui dépend de α. Ce lemme fait apparaître pour la première fois une grandeur qui sera d'une grande importance par la suite : le rapport des *likelihoods*.

Il est important de noter que ce résultat suppose la distribution du test sous H_0 et H_1 connue. Il s'agit donc d'hypothèses simples.

Le lemme de Neyman-Pearson se démontre de la manière suivante. Considérons une autre région critique $\mathcal{C}_A$ ayant le même seuil de signification que $\mathcal{C}_{NP}$ (α) et notons C_{NP} et C_A les événements $t \in \mathcal{C}_{NP}$ et $t \in \mathcal{C}_A$ respectivement. Il faut montrer que

$$P(C_{NP}|H_1) \geq P(C_A|H_1)$$

ou, de manière équivalente, que

$$P(C_{NP} \cap \overline{C_A}|H_1) \geq P(C_A \cap \overline{C_{NP}}|H_1)$$

puisque, quelle que soit l'hypothèse,

$$P(C_{NP}) = P(C_{NP} \cap C_A) + P(C_{NP} \cap \overline{C_A}) \quad \text{et} \quad P(C_A) = P(C_A \cap C_{NP}) + P(C_A \cap \overline{C_{NP}})$$

Nous avons, si le lemme est vrai

$$\begin{aligned} P(C_{NP} \cap \overline{C_A}|H_1) &= \int_{\mathcal{C}_{NP} \cap \overline{\mathcal{C}_A}} f(t|H_1)\mathrm{d}t \\ &\geq \frac{1}{k_\alpha} \int_{\mathcal{C}_{NP} \cap \overline{\mathcal{C}_A}} f(t|H_0)\mathrm{d}t = \frac{1}{k_\alpha} P(C_{NP} \cap \overline{C_A}|H_0) \end{aligned}$$

2. *Best Critical Region*

Or

$$P(C_{NP} \cap \overline{C_A}|H_0) = P(C_A \cap \overline{C_{NP}}|H_0)$$

car

$$P(C_{NP}|H_0) = P(C_A|H_0) = \alpha$$

Donc

$$P(C_{NP} \cap \overline{C_A}|H_1) \geq \frac{1}{k_\alpha} \int_{\mathcal{C}_A \cap \overline{\mathcal{C}_{NP}}} f(t|H_0)\mathrm{d}t \geq \int_{\mathcal{C}_A \cap \overline{\mathcal{C}_{NP}}} f(t|H_1)\mathrm{d}t = P(C_A \cap \overline{C_{NP}}|H_1)$$

La figure 7.6 représente schématiquement les différentes régions mises en jeu dans cette demonstration dans le cas bi-dimensionnel.

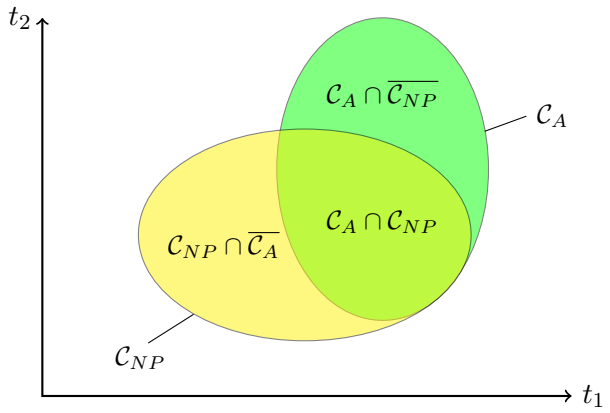


FIGURE 7.6 – Représentation des régions critiques intervenant dans la démonstration du lemme de Neyman-Pearson.

Une façon alternative de voir les choses consiste à définir un nouveau test Λ égal au rapport des *likelihoods* :

$$\Lambda = \frac{f(t|H_0)}{f(t|H_1)}$$

Λ est une variable aléatoire puisque c'est une fonction de l'échantillon. Il lui est donc associé une distribution sous l'hypothèse H_0 et une autre sous l'hypothèse H_1. Puisque les événements $t \in \mathcal{C}_{NP}$ et $\Lambda \leq k_\alpha$ sont identiques, nous avons

$$P(\Lambda \leq k_\alpha|H_0) = P(t \in \mathcal{C}_{NP}|H_0) = \alpha$$

et

$$P(\Lambda \leq k_\alpha|H_1) = P(t \in \mathcal{C}_{NP}|H_1) = 1 - \beta_{\min}$$

où $1 - \beta_{\min}$ est la puissance maximale atteignable pour le seuil de signification α. Nous sommes donc ramené d'un problème multidimensionnel à un problème unidimensionnel (voir figure 7.7).

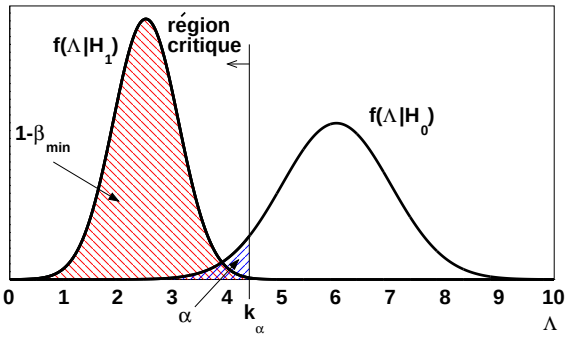


FIGURE 7.7 – Distributions du rapport des *likelihoods* et *p-values* associées.

Exemple 7.1: Soit X et Y deux variables indépendantes de même espérance μ distribuées suivant une gaussienne à deux dimensions de largeur unité. La densité de probabilité conjointe est d'après la section A.2.2

$$f(x,y;\mu) = \frac{1}{2\pi} e^{-\frac{1}{2}\left[(x-\mu)^2+(y-\mu)^2\right]} = \frac{1}{2\pi} e^{-\frac{1}{2}\left[x^2+y^2-2\mu(x+y)+2\mu^2\right]}$$

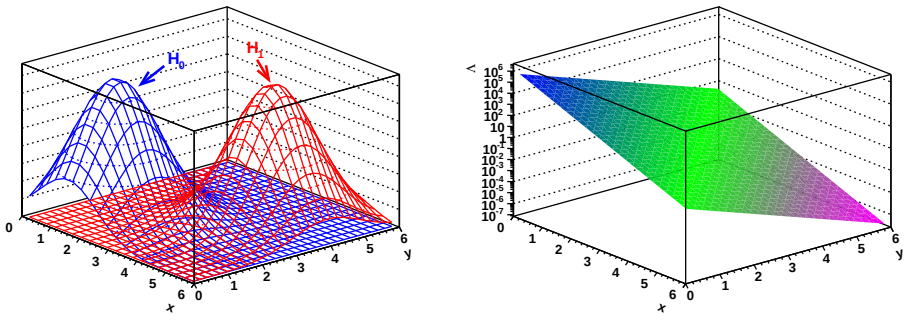
Considérons les deux hypothèses suivantes :

- $H_0 : \mu = \mu_0$
- $H_1 : \mu = \mu_1$

Nous avons

$$\Lambda = \frac{f(x,y;\mu_0)}{f(x,y;\mu_1)} = e^{-(\mu_0^2-\mu_1^2)} e^{(\mu_0-\mu_1)(x+y)}$$

La figure suivante représente $f(x,y;\mu_0)$, $f(x,y;\mu_1)$ et Λ dans le cas $(\mu_0,\mu_1) = (1.5, 4)$.



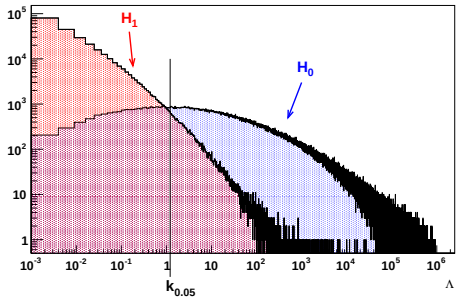
La région critique pour un seuil de signification de 0,05 est donnée par

$$\Lambda \leq k_{0,05}$$

ce qui, dans le cas $\mu_0 < \mu_1$, correspond à

$$\Leftrightarrow x + y \geq \mu_0 + \mu_1 + \frac{\ln k_{0,05}}{\mu_0 - \mu_1}$$

Elle est donc délimitée par une droite dans le plan (x, y). La figure ci-dessous montre les distributions de Λ sous H_0 et H_1 pour les mêmes valeurs de μ_0 et μ_1 que précédemment.



À partir de ces distributions nous trouvons immédiatement

$$k_{0,05} = 1,2 \quad \text{et} \quad 1 - \beta = 96,4\%$$

7.4 Test UMP

Lorsque l'hypothèse alternative H_1 est composée, nous pouvons trouver, grâce au lemme de Neyman-Pearson, une région BCR pour chaque hypothèse simple au sein de H_1. Notons comme précédemment Ξ l'ensemble des valeurs de θ correspondant à H_1 :

$$H_1 : \theta \in \Xi$$

Dans le cas le plus général, la région BCR dépend de la valeur de θ. Il existe toutefois des cas particuliers où la région BCR n'en dépend pas. Le test est, dans ces cas, qualifié d'UMP (*Uniformely Most Powerful*). Un test UMP est donc caractérisé par une région critique unique.

7.4.1 Exemple de test UMP sur un cas concret

En guise d'exemple, considérons le cas de n variables X_i ($i \in [1, n]$) i.i.d. distribuées suivant une loi normale de paramètre μ et écart-type unité ainsi que le test unilatéral

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1 > \mu_0$$

Nous avons commencé à discuter ce problème dans le cas bidimensionnel dans l'exemple 7.1. Nous voyons sans difficulté que, dans le cas général,

$$\Lambda = \frac{f(x_1, \ldots, x_n; \mu_0)}{f(x_1, \ldots, x_n; \mu_1)} = e^{-\frac{n}{2}(\mu_0^2 - \mu_1^2)} e^{(\mu_0 - \mu_1)\sum x_i} = e^{-\frac{n}{2}(\mu_0^2 - \mu_1^2)} e^{(\mu_0 - \mu_1) nm}$$

où m est la moyenne empirique. La région BCR est donc donnée par

$$m \geq \frac{1}{2}(\mu_0 + \mu_1) + \frac{\ln k_\alpha}{n(\mu_0 - \mu_1)}$$

Dans la suite nous noterons b_α le second membre de cette équation

$$m \geq b_\alpha \tag{7.4}$$

Cette expression pourrait laisser penser que la région BCR dépend de μ_1 (car le second membre de l'inégalité semble en dépendre) et que donc le test n'est pas UMP. Il n'en est rien. En fait, le second membre ne dépend pas de μ_1 car il y a une dépendance de k_α en μ_1 que nous n'avons pas explicitée. En effet, pour chaque valeur de μ_1, l'événement représenté par la condition 7.4 et l'événement $t \in$BCR sont identiques. La puissance du test est donc

$$1 - \beta = P(t \in \text{BCR}|H_1) = P(m \geq b_\alpha|H_1)$$

Dans cette expression, $P(t \in \text{BCR}|H_1)$ et $P(m \geq b_\alpha|H_1)$ sont des probabilités sous l'hypothèse H_1 pour une valeur donnée de μ_1 (la notation est ici un peu abusive puisque H_1 désigne dans ce cas une hypothèse simple correspondant à une certaine valeur de μ_1 plutôt qu'une hypothèse composite comme précédemment). m étant une grandeur empirique, sa valeur ne dépend pas des paramètres et sa distribution est connue puisque celle des X_i l'est. Nous savons en effet que

$$m \sim \mathcal{N}(\mu, 1/\sqrt{n})$$

La région BCR est donc entièrement déterminée dès lors qu'un seuil de signification α a été choisi et b_α ne dépend pas de μ_1. Nous avons

$$b_\alpha = \mu_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(1 - \alpha)$$

La figure 7.8 représente la distribution de m sous H_0 et H_1 pour différentes valeurs de μ_1.

Nous concluons de ce qui précède que le test unilatéral $\mu_1 > \mu_0$ est un test UMP. Il faut bien noter que cela signifie qu'il y a une seule et unique région BCR pour toutes les valeurs de μ_1 et non pas que la puissance est indépendante de μ_1. En effet, nous avons, pour cet exemple (cf figure 7.9)

$$1 - \beta = 1 - \Phi\left(\sqrt{n}\,(\mu_0 - \mu_1) + \Phi^{-1}\,(1 - \alpha)\right)$$

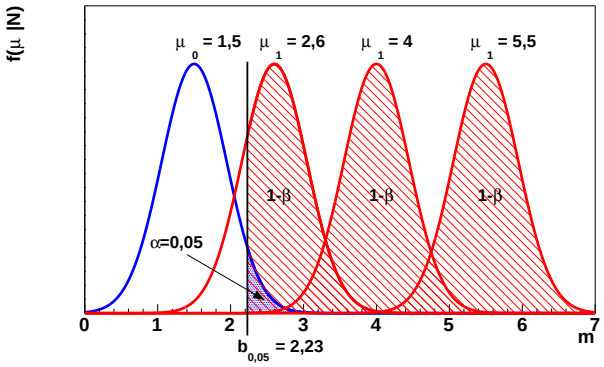


FIGURE 7.8 – Distribution de la moyenne empirique ($\mathcal{N}(\mu, 1/\sqrt{n})$) pour $n = 5$, $\mu_0 = 1,5$ et plusieurs valeurs de μ_1.

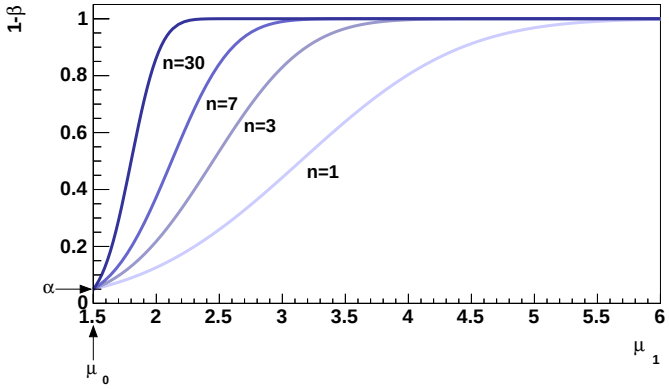


FIGURE 7.9 – Puissance du test UMP en fonction de μ_1 ($\mu_1 > \mu_0 = 1,5$) pour $\alpha = 0,05$ et plusieurs valeurs de n.

Si plutôt que de considérer le test $\mu_1 > \mu_0$ nous avions considéré le test $\mu_1 < \mu_0$ nous serions arrivé à la même conclusion, à savoir qu'il s'agit d'un test UMP (il suffit d'inverser le sens de l'inégalité 7.4). Dans le cas du test bilatéral $\mu_1 \neq \mu_0$, nous voyons par contre qu'il ne s'agit pas d'un test UMP car la région critique n'est pas la même dans les cas $\mu_1 > \mu_0$ et $\mu_1 < \mu_0$.

7.4.2 Généralisation

L'exemple particulier considéré dans la section précédente se généralise à toutes les distributions de la famille exponentielle

$$f(x;\theta) = A(\theta)B(x)e^{C(\theta)D(x)}$$

pour lesquelles $C(\theta)$ est une fonction monotone de θ. En effet, dans ce cas le *likelihood* est

$$f(x_1, \ldots, x_n; \theta) = A(\theta)^n \left[\prod_{i=1}^{n} B(x_i)\right] e^{C(\theta) \sum_{i=1}^{n} D(x_i)}$$

Considérons le test unilatéral où $H_0 : \theta = \theta_0$ est comparé à $H_1 : \theta > \theta_0$. Nous noterons, comme précédemment, θ_1 une valeur particulière de θ sous l'hypothèse H_1. Le rapport des *likelihoods* est, pour une valeur θ_1 donnée,

$$\Lambda = \left[\frac{A(\theta_0)}{A(\theta_1)}\right]^n e^{[C(\theta_0) - C(\theta_1)] \sum_{i=1}^{n} D(x_i)}$$

La région BCR pour cette valeur de μ_1 est donc donnée par

$$\sum_{i=1}^{n} D(x_i) \leq \frac{1}{C(\theta_0) - C(\theta_1)} \left[\ln k_\alpha - n \left(\ln A(\theta_0) - \ln A(\theta_1)\right)\right] \quad \text{si} \quad C(\theta_0) - C(\theta_1) > 0$$

$$\sum_{i=1}^{n} D(x_i) \geq \frac{1}{C(\theta_0) - C(\theta_1)} \left[\ln k_\alpha - n \left(\ln \ A(\theta_0) - \ln A(\theta_1)\right)\right] \quad \text{si} \quad C(\theta_0) - C(\theta_1) < 0$$

Le premier cas correspond à une fonction $C(\theta)$ décroissante et le second à une fonction croissante. Puisque $\sum D(x_i)$ n'est fonction que de l'échantillon il s'agit bien d'un test UMP. Comme dans l'exemple précédent, seule la connaissance de la distribution de cette somme sous H_0 est nécessaire pour déterminer la région BCR. Cette dernière est donc indépendante de θ_1.

Nous voyons aussi que, comme dans l'exemple gaussien, le test unilatéral où $H_0 : \theta = \theta_0$ est comparé à $H_1 : \theta < \theta_0$ est un test UMP alors que le test bilatéral où $H_0 : \theta = \theta_0$ est comparé à $H_1 : \theta \neq \theta_0$ n'en est pas un.

7.5 Test *likelihood ratio*

Les tests *likelihood ratio* sont beaucoup utilisés lorsque les hypothèses sont composites. Ils sont basés sur la variable test appelée *likelihood ratio* (d'où leurs noms). L'utilisation du terme *likelihood ratio* peut préter à confusion. En effet, tous les tests de la forme 7.3 pourraient être appelé *likelihood ratio* car ils correspondent à un rapport de *likelihood*. Dans la pratique, lorsqu'on parle de *likelihood ratio*, on fait référence le plus souvent à un rapport bien précis de *likelihoods* qui diffère de 7.3. Ce rapport est défini un peu plus loin.

Considérons le cas général où les hypothèses nulles et alternatives sont composées et notons

- θ l'ensemble (de dimension m) des paramètres,
- θ_r l'ensemble (de dimension r) des paramètres fixés par l'hypothèse nulle H_0 et
- θ_s l'ensemble (de dimension $m - r$) des paramètres libres sous H_0 [3]

De manière abrégée :

$$\theta = (\theta_1, \ldots, \theta_m) = (\theta_r, \theta_s)$$

3. Cette notation peut induire en erreur car θ_r et θ_s pourraient sembler désigner les $r^{\text{ième}}$ et $s^{\text{ième}}$ composantes de θ. Il n'en est rien, θ_r et θ_s désignerons toujours des ensembles de paramètres plutôt qu'un paramètre en particulier.

Notons θ_{r0} la valeur de θ_r sous H_0. Nous pouvons ainsi définir H_0 par

$$H_0 : \theta_r = \theta_{r0}, \theta_s$$

L'hypothèse alternative peut quant à elle dépendre du problème considéré (elle peut par exemple correspondre à une autre valeur de θ_r ou bien à un test bilatéral $\theta_r \neq \theta_{r0}$). Sa nature exacte n'est pas importante dans ce qui suit.

Afin d'alléger les notations et par soucis de cohérence avec les notations employées dans la littérature nous noterons le *likelihood*

$$\mathcal{L}(\theta_r, \theta_s)$$

7.5.1 Définition

Le test *likelihood ratio* est défini par

$$\Lambda = \frac{\mathcal{L}(\theta_{r0}, \hat{\hat{\theta}}_s)}{\mathcal{L}(\hat{\theta}_r, \hat{\theta}_s)} \tag{7.5}$$

où $\hat{\theta}_r$ et $\hat{\theta}_s$ sont les estimateurs ML (voir section 6.3) de θ_r et θ_s et $\hat{\hat{\theta}}_r$ est l'estimateur ML conditionnel de θ_s pour $\theta_r = \theta_{r0}$

$$\hat{\hat{\theta}}_s = \hat{\hat{\theta}}_s(\theta_{r0})$$

Il est évident d'après 7.5 que

$$0 \leq \Lambda \leq 1$$

Sous l'hypothèse nulle nous avons $\hat{\theta}_r \simeq \theta_{r0}$ et $\hat{\hat{\theta}}_s \simeq \hat{\theta}_s$. Le test a donc une valeur plutôt grande. La région critique correspond donc à $\Lambda \leq k_\alpha$. Dans la suite nous utiliserons plutôt le test

$$t = -2 \ln \Lambda$$

auquel correspond la région critique $t \geq -2 \ln k_\alpha$ (sous H_0, t est proche de 0).

7.5.2 Distribution dans la limite asymptotique

Comme à chaque fois qu'un test d'hypothèse est réalisé, il est nécessaire de connaître la distribution du test sous l'hypothèse nulle (et potentiellement sous l'hypothèse alternative H_1). Un des grands avantage du test *likelihood ratio* par rapport aux autres tests est que, dans la limite asymptotique, sa distribution est connue analytiquement (ce qui évite d'avoir recours à des simulations parfois lourdes pour la déterminer).

Nous démontrerons dans un premier temps la formule de la distribution asymptotique dans le cas à un paramètre puis donnerons dans un deuxième temps la formule de la distribution dans le cas général à plusieurs paramètres.

Dans le cas à un paramètre, le test est

$$t = -2 \ln \frac{\mathcal{L}(\theta_{r0})}{\mathcal{L}(\hat{\theta}_r)} = -2 \left[\ln \mathcal{L}(\theta_{r0}) - \ln \mathcal{L}(\hat{\theta}_r) \right]$$

Développons $\ln \mathcal{L}(\theta_{r0})$ autour de $\ln \mathcal{L}(\hat{\theta}_r)$:

$$\ln \mathcal{L}(\theta_{r0}) = \ln \mathcal{L}(\hat{\theta}_r) + \left(\theta_{r0} - \hat{\theta}_r\right) \left.\frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_r} + \frac{1}{2}\left(\theta_{r0} - \hat{\theta}_r\right)^2 \left.\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\right|_{\theta=\theta^\star}$$

où $\theta^\star$ est une valeur entre θ_{r0} et $\hat{\theta}_r$. Le second terme du membre de droite est nul, par définition de $\hat{\theta}_r$. Donc

$$t = -\left(\hat{\theta}_r - \theta_{r0}\right)^2 \left.\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\right|_{\theta=\theta^\star}$$

Dans la limite asymptotique et sous l'hypothèse nulle

$$\hat{\theta}_r \xrightarrow{P} \theta_{r0} \quad \Rightarrow \quad \theta^\star \xrightarrow{P} \theta_{r0}$$

et

$$\left.\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2}\right|_{\theta=\theta^\star} \xrightarrow{P} -I_n(\theta_{r0})$$

Donc

$$t \simeq I_n(\theta_{r0}) \left(\hat{\theta}_r - \theta_{r0}\right)^2$$

Or nous avons montré dans la section 6.3.1 que $\hat{\theta}_r - \theta_{r0}$ tend vers une loi normale centrée de variance $I_n(\theta_{r0})^{-1}$, donc t tend vers une loi de χ^2 à un degré de liberté. Ce résultat constitue le théorème de Wilks.

Dans le cas à m paramètres dont r sont fixés par l'hypothèse nulle, A. Wald a montré (voir [7] pour la demonstration) que le test est, dans la limite asymptotique,

$$t = \left(\hat{\theta}_r - \theta_{r0}\right)^T V_r^{-1} \left(\hat{\theta}_r - \theta_{r0}\right)$$

où V_r^{-1} est la matrice de covariance des $\hat{\theta}_r$. Puisque les $\hat{\theta}_r$ sont asymptotiquement gaussiens d'espérance θ_r, nous voyons que t est distribué, dans la limite asymptotique, suivant une loi de χ^2 non centrée avec r degrés de liberté et de paramètre de décentralisation

$$(\theta_r - \theta_{r0})^T V_r^{-1} (\theta_r - \theta_{r0})$$

Sous l'hypothèse nulle nous avons $\theta_r = \theta_{r0}$, le test est donc distribué suivant une loi de χ^2 à r degrés de libertés.

7.6 Cas poissonnien

Le cas poissonnien est particulièrement important car il s'applique à toutes les expériences de comptages réalisées en physique subatomique. Il a déjà été partiellement discuté dans la section 7.1. Il est temps maintenant de le compléter à l'aide des notions vues dans les sections 7.2, 7.3 et 7.5. Notons $p(n|H)$ la probabilité d'observer n événements sous l'hypothèse H :

$$p(n|H) = \frac{\nu_H^n}{n!} e^{-\nu_H}$$

où ν_H est le paramètre de la loi de Poisson sous l'hypothèse H.

7.6.1 *p-value*

La *p-value* associée à une observation n_{obs} est

$$p\text{-}value = \begin{cases} \displaystyle\sum_{n=n_{\text{obs}}}^{\infty} p(n|H) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} p(n|H) & \text{si } n_{\text{obs}} > \nu_H \\ \displaystyle\sum_{n=0}^{n_{\text{obs}}} p(n|H) & \text{si } n_{\text{obs}} \leq \nu_H \end{cases} \tag{7.6}$$

Les sommes ci-dessus peuvent être exprimées différemment en utilisant l'égalité

$$\sum_{n=0}^{n_{\text{obs}}} p(n|H) = \frac{\Gamma(n_{\text{obs}}+1, \nu_H)}{\Gamma(n_{\text{obs}}+1)}$$

où

$$\Gamma(n_{\text{obs}}+1, \nu_H) = \int_{\nu_H}^{\infty} x^{n_{\text{obs}}} e^{-x} \mathrm{d}x$$

est la fonction gamma incomplète et

$$\Gamma(n_{\text{obs}}+1) = \int_{0}^{\infty} x^{n_{\text{obs}}} e^{-x} \mathrm{d}x$$

est la fonction gamma. Nous pouvons écrire de manière équivalente

$$\sum_{n=0}^{n_{\text{obs}}} p(n|H) = 1 - F_{\Gamma}(n_{\text{obs}}+1, \nu_H)$$

où $F_{\Gamma}(n_{\text{obs}}+1, \nu_H)$ est la fonction de répartition de la distribution gamma.

7.6.2 Limite gaussienne

Lorsque ν_H tend vers l'infini, la loi de Poisson tend vers la loi normale de moyenne ν_H et écart-type ν_H :

$$p(n|H) \simeq \frac{1}{\sqrt{2\pi\nu_H}} e^{-\frac{(n-\nu_H)^2}{2\nu_H}}$$

Nous avons donc

$$\sum_{n=n_{\text{obs}}}^{\infty} p(n|H) = 1 - \sum_{n=0}^{n_{\text{obs}}-1} p(n|H) \simeq 1 - \Phi\left(\frac{n_{\text{obs}} - 1 - \nu_H}{\sqrt{\nu_H}}\right) \simeq 1 - \Phi\left(\frac{n_{\text{obs}} - \nu_H}{\sqrt{\nu_H}}\right)$$

où Φ est la fonction de répartition de la loi normale centrée réduite. D'après 7.1, la signification statistique est, dans le cas $n_{\text{obs}} > \nu_H$,

$$Z \simeq \frac{n_{\text{obs}} - \nu_H}{\sqrt{\nu_H}}$$

7.6.3 Cas où ν_H est connu de manière incertaine

Il arrive très souvent (quasiment tout le temps en réalité), que le paramètre de la loi de Poisson ν_H ne soit pas connu de manière exacte. Une façon de tenir compte de l'incertitude sur ce paramètre dans le calcul de *p-values* et de significations statistiques consiste à utiliser un modèle marginal plutôt que la loi de Poisson qui ne décrit que les fluctuations statistiques. Supposons que ν_H soit distribué suivant une fonction gamma f de paramètre a et b :

$$f(\nu_H; a, b) = \frac{a(a\nu_H)^{b-1}e^{-a\nu_H}}{\Gamma(b)}$$

La distribution marginale $p_m(n; a, b)$ est

$$p_m(n; a, b) = \int_0^\infty p(n|H)f(\nu_H; a, b)\mathrm{d}\nu_H$$

Nous avons vu dans l'exemple 3.3 que $p_m(n; a, b)$ est une distribution binomiale négative :

$$p_m(n; a, b) = \frac{a^b\Gamma(n+b)}{n!\Gamma(b)(a+1)^{n+b}}$$

La *p-value* se calcule dans ce cas en remplaçant $p(n|H)$ par $p_m(n|H)$ dans 7.6.

7.7 Cas d'une distribution expérimentale *binnée*

Cette section décrit deux méthodes couramment utilisées pour comparer des données expérimentales *binnées* à une distribution théorique (*binnée* ou non). Elles s'adressent donc typiquement au problème représenté sur la figure 7.10 où un histogramme est comparé à une distribution théorique.

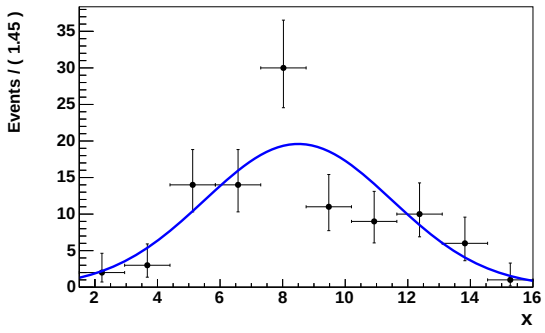


FIGURE 7.10 – Comparaison d'une distribution expérimentale *binnée* et d'une distribution théorique.

7.7.1 Méthode basée sur le test de Pearson

Le test de Pearson, que nous avons déjà rencontré en 6.4.2, est

$$\chi^2_P = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}$$

où N est le nombre de *bins*, n_i le nombre d'événements observés dans le *bin* i et ν_i le nombre d'événements attendus dans le *bin* i. Notons p_i la probabilité sous H_0 d'avoir une entrée dans le *bin* i :

$$p_i = \int_{bin\ \mathrm{i}} f(x)\mathrm{d}x$$

Dans le cas où le nombre total d'éléments dans l'échantillon $n_{\mathrm{tot}} = \sum_i n_i$ est fixe, nous avons

$$\nu_i = n_{\mathrm{tot}} p_i$$

alors que dans le cas où il est considéré comme une variable aléatoire poissonienne d'espérance $\nu_{\mathrm{tot}} = \sum_i \nu_i$ nous avons

$$\nu_i = \nu_{\mathrm{tot}} p_i$$

Nous sommes donc amené à définir les deux tests

$$\chi^2_{P1} = \sum_{i=1}^{N} \frac{(n_i - \nu_{\mathrm{tot}} p_i)^2}{\nu_{\mathrm{tot}} p_i} \quad \text{dans le cas où } n_{\mathrm{tot}} \text{ est variable}$$

et

$$\chi^2_{P2} = \sum_{i=1}^{N} \frac{(n_i - n_{\mathrm{tot}} p_i)^2}{n_{\mathrm{tot}} p_i} \quad \text{dans le cas où } n_{\mathrm{tot}} \text{ est constant}$$

L'intérêt de ces tests réside dans le fait que, dans la limite asymptotique, ils sont distribués suivant une loi connue. Il est donc facile de calculer une *p-value* et éventuellement une signification statistique suite à une observation.

Distribution asymptotique de χ^2_{P1}

Dans le cas où n_{tot} est variable, les n_i sont indépendants et distribués suivant une loi de Poisson de paramètre ν_i. Dans la limite asymptotique, ils sont donc distribués suivant une loi gaussienne d'espérance ν_i et écart-type $\sqrt{\nu_i}$. χ^2_{P1} est par conséquent distribué suivant une loi de χ^2 à N degrés de liberté.

Distribution asymptotique de χ^2_{P2}

Dans le cas où n_{tot} est constant, les n_i ne sont pas indépendants. Ils sont distribués suivant une loi multinomiale de paramètres $(p_1, \ldots, p_N)$:

$$P(n_1, \ldots, n_N) = \frac{n_{\mathrm{tot}}!}{\prod_{i=1}^{N} n_i!} \prod_{i=1}^{N} p_i^{n_i}$$

Le nombre de termes indépendants dans la somme de χ^2_{P2} est donc $N - 1$. χ^2_{P2} est donc distribué suivant une loi de χ^2 à $N - 1$ degrés de liberté.

Cas où certains paramètres sont estimés à partir des données

Dans le cas où la distribution théorique dépend de certains paramètres estimés à partir des données, χ^2_P est toujours distribué suivant une loi de χ^2 mais le nombre de degrés de libertés est inférieur à N et $N-1$ pour χ^2_{P1} et χ^2_{P2}. Considérons le cas où la distribution théorique dépend de r paramètres θ_i ($i \in [1,r]$), parmi lesquels m sont estimés à partir des données. Nous pouvons écrire

$$\chi^2_{P1} = \sum_{i=1}^{N} \frac{\left(n_i - \nu_{\text{tot}} p_i \left(\hat{\theta}_1, \ldots, \hat{\theta}_m, \ldots, \theta_r\right)\right)^2}{\nu_{\text{tot}} p_i \left(\hat{\theta}_1, \ldots, \hat{\theta}_m, \ldots, \theta_r\right)}$$

où les m paramètres $\hat{\theta}_i$ avec $i \in [1,m]$ sont des fonctions des n_i

$$\hat{\theta}_i = \hat{\theta}_i \left(n_1, \ldots, n_N\right) \tag{7.7}$$

Les équations 7.7 représentent m contraintes entres les n_i. Le nombre de termes indépendants dans χ^2_{P1} est donc $N-m$. χ^2_{P1} est donc asymptotiquement distribué suivant une loi de χ^2 à $N-m$ degrés de liberté. Un raisonnement similaire permet de montrer que χ^2_{P2} est distribué suivant une loi de χ^2 à $N-m-1$ degrés de liberté.

7.7.2 Méthode basée sur le *likelihood ratio*

Considérons le cas où les observations sont distribuées de manière multinomiale. Le *likelihood* est, en supprimant les termes ne dépendant pas des p_i,

$$\mathcal{L}(p_1, \ldots, p_N) = \prod_{i=1}^{N} p_i^{n_i}$$

Le *likelihood ratio* s'écrit

$$\Lambda = \frac{\mathcal{L}(p_1, \ldots, p_N)}{\mathcal{L}(\hat{p}_1, \ldots, \hat{p}_N)}$$

où $\hat{p}_i$ est l'estimateur ML de p_i. Ces estimateurs sont donnés par (nous laissons au lecteur le soin de le démontrer)

$$\hat{p}_i = \frac{n_i}{n_{\text{tot}}}$$

Donc

$$\Lambda = \prod_{i=1}^{N} \left(\frac{n_{\text{tot}} p_i}{n_i}\right)^{n_i}$$

où, de manière équivalente,

$$t = -2 \ln \Lambda = 2 \sum_{i=1}^{N} n_i \ln \frac{n_i}{n_{\text{tot}} p_i}$$

Asymptotiquement, t est distribué suivant une loi de χ^2 à $N-1$ degrés de libertés (voir 7.5.2).

7.7.3 Lien entre le test de Pearson et le *likelihood ratio*

Dans la limite asymptotique, les estimateurs ML convergent en probabilité vers les vraies valeurs (voir 6.3.1), donc $\hat{p}_i \xrightarrow{P} p_i$ ou, de manière équivalente, $n_i \xrightarrow{P} n_{\text{tot}} p_i$. Nous pouvons donc faire le développement limité de t autour de $n_i = n_{\text{tot}} p_i$. Posons

$$n_i = n_{\text{tot}} p_i \left(1 + \Delta_i\right)$$

avec $\Delta_i \xrightarrow{P} 0$. Nous avons donc

$$\begin{aligned} t &= 2 \sum_{i=1}^{N} n_{\text{tot}} p_i \left(1 + \Delta_i\right) \ln\left(1 + \Delta_i\right) \\ &\simeq 2 \sum_{i=1}^{N} n_{\text{tot}} p_i \left(1 + \Delta_i\right) \left(\Delta_i - \frac{\Delta_i^2}{2}\right) \\ &\simeq 2 \sum_{i=1}^{N} n_{\text{tot}} p_i \left(\Delta_i + \frac{\Delta_i^2}{2}\right) \end{aligned}$$

en se limitant à l'ordre 2 en Δ_i. Or $\sum_i p_i \Delta_i = 0$, donc

$$t \simeq \sum_{i=1}^{N} n_{\text{tot}} p_i \Delta_i^2$$

De plus, puisque $\Delta_i = \left(n_i - n_{\text{tot}} p_i\right) / \left(n_{\text{tot}} p_i\right)$, nous trouvons

$$t \simeq \chi^2_{P2}$$

Ainsi, le test *likelihood ratio* et celui de Pearson sont équivalents dans la limite asymptotique.

Chapitre 8

Intervalle de confiance

Il arrive fréquemment que nous voulions trouver un intervalle de valeurs probables pour un paramètre plutôt qu'une estimation ponctuelle comme décrit dans le chapitre 6. Quelques méthodes fréquentistes permettant de trouver de tels intervalles, appelés intervalles de confiance, sont décrites dans ce chapitre (les méthodes bayésiennes sont décrites dans le chapitre 9).

8.1 Introduction

Un intervalle de confiance est un intervalle censé contenir la vraie valeur d'un paramètre avec une grande probabilité. Ils peuvent s'écrire d'une manière générale comme ceci :

$$[\theta_{\min}(x); \theta_{\max}(x)]$$

où $\theta_{\min}(x)$ et $\theta_{\max}(x)$ sont les bornes inférieure et supérieure de l'intervalle. Toutes deux sont des fonctions des données x. Ce sont donc des variables aléatoires susceptibles de varier d'une expérience à l'autre. La probabilité que la vraie valeur θ_0 soit contenue dans l'intervalle

$$P\left(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]\right)$$

est appelée couverture. La couverture d'un intervalle de confiance est une notion centrale dans l'approche classique décrite dans ce chapitre. Les intervalles de confiance sont la plupart du temps construits avec l'objectif d'atteindre une couverture égale à une valeur prédéfinie appelée niveau de confiance. Dans la suite, nous appelerons le niveau de confiance α. Les valeurs de α les plus fréquemment utilisées sont 68%, 90% et 95%. Ainsi, un intervalle correspondant à un niveau de confiance $\alpha = 95\%$ est censé contenir la vraie valeur dans 95% des expériences.

Le résultat d'une expérience visant à construire un intervalle de confiance est typiquement formulé de la manière suivante : $\theta \in [\theta_{\min}; \theta_{\max}]$ à 95% CL. Ceci signifie que le niveau de confiance α est de 95% (CL signifie *Confidence Level*).

Il est important de voir que la couverture peut être différente du niveau de confiance. Ce dernier correspond à l'objectif visé mais il est parfois difficile de l'atteindre. Trois cas sont à distinguer :

(a) $P\left(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]\right) = \alpha$: c'est le cas idéal

(b) $P\left(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]\right) > \alpha$: ce cas n'est pas idéal mais acceptable puisque le niveau de confiance est sous-estimé (on dit qu'il y a *overcoverage*)

(c) $P(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]) < \alpha$: c'est le pire des cas car le niveau de confiance est sur-estimé (on dit qu'il y a *undercoverage*)

Dans la section 8.2 nous décrirons des méthodes permettant de construire des intervalles de confiance pouvant correspondre aux trois cas cités ci-dessus. Le fait que ces méthodes puissent conduire au cas (c) (sauf dans des cas simples) est leur principal défaut. Dans les sections 8.3, 8.4 et 8.5 nous décrirons des méthodes offrant la garantie d'avoir une couverture supérieure ou égale au niveau de confiance (cas (a) et (b)). Celles-ci doivent être préférées, dans la mesure du possible, à celles décrites dans la section 8.2. Il faut tout de même noter qu'elles peuvent être assez complexes à mettre en œuvre et très coûteuses en temps de calcul. Si un résultat est attendu rapidement alors les méthodes simples de la section 8.2 peuvent être utilisées (mais il faut toujours avoir en tête qu'elle peuvent parfois sur-estimer assez largement le niveau de confiance).

8.2 Méthodes approximatives

Supposons que nous voulions déterminer un intervalle de confiance pour un paramètre θ inconnu d'une certaine loi de probabilité. Soit $\{X_i\}$ ($i \in [1, n]$) l'échantillon de données. L'intervalle de confiance le plus simple est construit à partir d'un estimateur $\hat{\theta}(\{X_i\})$:

$$\theta \in \left[\hat{\theta} - d\sqrt{\text{var}\left[\hat{\theta}\right]}; \hat{\theta} + d\sqrt{\text{var}\left[\hat{\theta}\right]}\right] \tag{8.1}$$

où d est un réel permettant d'ajuster la taille de l'intervalle (et donc la couverture). Nous reviendrons un peu plus loin sur le choix de d. Les différentes méthodes décrites dans le chapitre 6 peuvent être utilisées pour construire l'estimateur $\hat{\theta}$. La principale difficulté réside dans le calcul de var $\left[\hat{\theta}\right]$. Ce n'est que dans des cas simples que la variance est calculable analytiquement. Dès que le problème devient un peu complexe, il n'est souvent pas possible de faire autrement que de l'estimer. Dans ce cas, l'intervalle s'écrira plutôt

$$\theta \in \left[\hat{\theta} - d\sqrt{\widehat{\text{var}}\left[\hat{\theta}\right]}; \hat{\theta} + d\sqrt{\widehat{\text{var}}\left[\hat{\theta}\right]}\right] \tag{8.2}$$

pour bien faire ressortir le fait que c'est un estimateur de la variance qui est utilisé. Une méthode qui peut être utiliser pour estimer la variance est la méthode Monte Carlo. Dans le cas où l'estimateur est calculé par la méthode du maximum de vraisemblance (c'est le cas le plus fréquent), la limite RCF ou la méthode graphique décrites dans la section 6.3.2 peuvent être utilisées si l'approximation asymptotique est supposée valide.

Ces intervalles ne sont pas construit dans le but d'atteindre un niveau de confiance prédéterminé (c'est en cela qu'ils sont qualifiés d'approximatifs). Ce n'est que dans des cas simples que leur couverture est connue. Nous savons par exemple que dans le cas gaussien que la relation entre la couverture et d pour un intervalle du type 8.1 est

$$d = \Phi^{-1}((1-\alpha)/2 + \alpha) \tag{8.3}$$

où Φ est la fonction de répartition de la loi normale centrée réduite (la couverture est par exemple de 68% pour $d = 1$, 90% pour $d = 1,64$ et 95% pour $d = 1,96$). Dans des cas complexes, la couverture n'est à priori pas connue. Une solution souvent adoptée est de supposer que l'estimateur est gaussien

(c'est une bonne approximation dans la limite asymptotique). Dans ce cas, la relation 8.3 est utilisée. Sinon, il est toujours possible de calculer la couverture a posteriori et d'ajuster la variable d de telle sorte à l'accroître ou la décroître mais cela peut être fastidieux et ne jamais aboutir. En effet, il est observé, pour ce type d'intervalle, que la couverture peut varier énormément en fonction de la vraie valeur du paramètre θ (nous verrons un exemple plus loin). Il est parfois impossible de trouver une valeur de d de telle sorte à ce que la couverture soit supérieure à une valeur acceptable (par exemple 90%) pour toutes les valeurs vraies du paramètre. C'est pour ces raisons que les intervalles décrit dans les sections suivantes sont préférables.

Comme il vient d'être dit, la couverture pour un intervalle du type 8.1 n'est pas connue précisément (sauf dans des cas simples tel que le cas gaussien). Il est toutefois possible de la borner inférieurement grâce à l'inégalité de Bienaymé-Tchebichef. D'après l'équation 2.3, nous pouvons en effet écrire, si $\hat{\theta}$ est un estimateur non biaisé de θ,

$$P\left(\left|\hat{\theta}-\theta\right| \geq d\sqrt{\text{var}\left[\hat{\theta}\right]}\right) \leq \frac{1}{d^2}$$

ou, de manière équivalente,

$$P\left(\left|\hat{\theta}-\theta\right| \leq d\sqrt{\text{var}\left[\hat{\theta}\right]}\right) \geq 1-\frac{1}{d^2}$$

ou encore

$$P\left(\hat{\theta}-d\sqrt{\text{var}\left[\hat{\theta}\right]} \leq \theta \leq \hat{\theta}+d\sqrt{\text{var}\left[\hat{\theta}\right]}\right) \geq 1-\frac{1}{d^2}$$

Nous voyons donc que la couverture de l'intervalle 8.1 est bornée inférieurement par $1-1/d^2$. La table 8.1 donne quelques valeurs de cette borne pour différentes valeur de d.

d	1	2	3	4	5
couverture $\geq$	0	0,75	0,88	0,9375	0,96

TABLE 8.1 – Borne inférieure sur la couverture pour différentes valeur de d pour des intervalles de la forme 8.1.

8.2.1 Intervalle de confiance pour l'espérance d'une distribution de variance connue

La détermination d'un intervalle de confiance pour l'espérance d'une distribution de variance σ^2 connue est un des cas les plus simples. En effet, nous avons vu dans les sections 4.1 et 6.2.1 que la moyenne empirique $M = \frac{1}{n}\sum_i X_i$ est un estimateur non biaisé de l'espérance et que sa variance est $\text{var}[M] = \sigma^2/n$. Ainsi, l'intervalle de confiance pour l'espérance est

$$\left[M - d\frac{\sigma}{\sqrt{n}}; M + d\frac{\sigma}{\sqrt{n}}\right] \tag{8.4}$$

Cet intervalle est de la forme 8.1 car c'est la vraie variance de l'estimateur qui est utilisée et non pas une estimation de la variance.

Exemple 8.1: Considérons le cas d'une distribution gaussienne de moyenne μ inconnue et variance σ^2 connue :

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

La moyenne empirique est aussi distribuée de manière gaussienne :

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

À partir de la fonction de répartition de la loi normale, nous savons que

$$P\left(M < \mu + \frac{\sigma}{\sqrt{n}}\right) = 0,841 \Rightarrow P\left(\mu > M - \frac{\sigma}{\sqrt{n}}\right) = 0,84$$

et

$$P\left(M > \mu - \frac{\sigma}{\sqrt{n}}\right) = 0,841 \Rightarrow P\left(\mu < M + \frac{\sigma}{\sqrt{n}}\right) = 0,84$$

Nous avons donc

$$P\left(M - \frac{\sigma}{\sqrt{n}} < \mu < M + \frac{\sigma}{\sqrt{n}}\right) = 2 \times 0,84 - 1 = 0,68$$

L'intervalle de confiance donné dans l'équation 8.4 pour $d = 1$ a donc une couverture de 68%. De même, nous trouvons que pour $d = 1,64$ la couverture est de 90% et que pour $d = 1,96$ elle est de 95%. Le cas gaussien est l'un des rares cas pour lequel la couverture est calculable analytiquement. Pour d'autres distributions, il est nécessaire en général d'avoir recours à des méthodes Monte Carlo.

8.2.2 Intervalle de confiance lorsque la variance est inconnue

Le cas où la variance est inconnue est à priori plus compliqué que celui considéré dans la section précédente où elle est connue. Dans ce cas il n'est en général pas possible de connaître la vraie variance de l'estimateur. Ce sont donc plutôt des intervalles de la forme 8.2 qui sont déterminés. La variété des méthodes utilisées en pratique est telle qu'il est impossible d'en rendre compte dans cette section. Nous nous contenterons d'illustrer, sur l'exemple de la loi binomiale, la construction d'un intervalle qui montre bien les difficultés susceptibles d'être rencontrées dans ce cas.

Exemple 8.2: Considérons le cas d'une variable aléatoire k distribuée suivant une loi binomiale de paramètres N et p :

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

Cherchons à construire un intervalle de confiance pour p en supposant que

N est connu. Nous savons (voir exemple 6.6) que l'estimateur ML de p est

$$\hat{p} = \frac{k}{N}$$

De plus, la variance de k est $Np(1-p)$. Celle de $\hat{p}$ est donc

$$\text{var}[\hat{p}] = \frac{p(1-p)}{N}$$

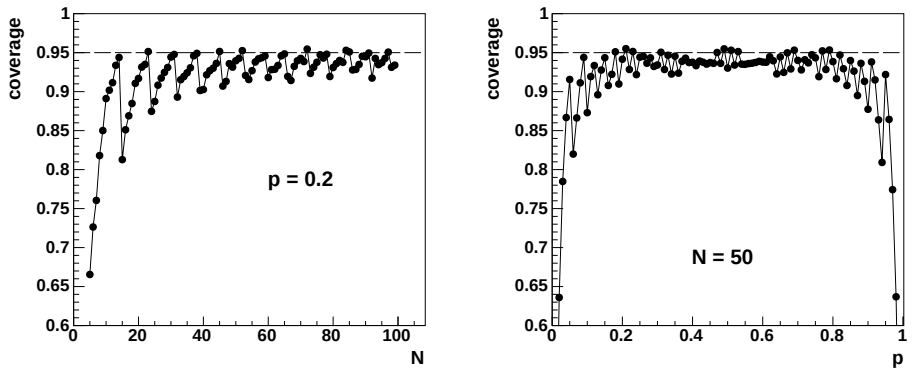
La variance de l'estimateur dépend de la grandeur inconnue p. Il est donc impossible de construire un intervalle de la forme 8.1. Une solution consiste à construire un intervalle de la forme 8.2 en remplaçant dans la variance p par $\hat{p} = k/N$:

$$\widehat{\text{var}}[\hat{p}] = \frac{\hat{p}(1-\hat{p})}{N}$$

Ainsi, l'intervalle de confiance pour p est

$$p \in \left[\hat{p} - d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}; \hat{p} + d\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right]$$

Cet intervalle est appelé intervalle de Wald. Bien que très utilisé, il est en fait très mauvais (nous donnons dans l'appendice B des intervalles meilleurs). Premièrement, nous voyons que l'intervalle tend vers l'ensemble vide lorsque $\hat{p}$ tend vers 0 ou 1. Ceci n'est physiquement pas acceptable (par exemple, il n'est pas raisonnable de dire que $p = 0 \pm 0$ lorsque $(N, k) = (2, 0)$ (ou que $p = 1 \pm 0$ lorsque $(N, k) = (2, 2)$). Deuxièmement, la couverture montre une forte dépendance avec les paramètres N et p. Ceci est illustré sur les figures suivantes pour $d = 1,96$, ce qui correspond dans l'approximation gaussienne à un niveau de confiance de 95%.



Ces figures montrent que la couverture est quasiment systématiquement inférieure au niveau de confiance attendu dans le cas gaussien, même lorsque N est grand (où l'on pourrait s'attendre à ce que l'approximation gaussienne soit valable). Plus grave, la couverture dépend énormément de N et p. Et comme p est inconnu, nous ne pouvons pas savoir quelle est la couverture de l'intervalle.

8.3 Construction de Neyman

8.3.1 Description de la méthode

La construction de Neyman est une méthode pour déterminer un intervalle de confiance sur un paramètre ou un ensemble de paramètres θ qui offre la garantie d'avoir une couverture égale (supérieure ou égale) au niveau de confiance dans le cas continu (discret). La première étape de cette méthode consiste à déterminer, pour chaque valeur de θ, un ensemble de valeurs de x correspondant à la probabilité α. Cette méthode est illustrée sur la figure 8.1 où les densités de probabilité $f(x;\theta)$ pour sept valeurs de θ sont représentées (sur cette figure nous avons considéré le cas gaussien avec $\alpha = 90\%$ mais la méthode s'applique de manière similaire si la distribution et α sont différents). Notons

$$[x_{\min}(\theta); x_{\max}(\theta)]$$

l'ensemble de valeurs de x correspondant à la probabilité α :

$$P\left(x \in [x_{\min}(\theta); x_{\max}(\theta)]; \theta\right) = \alpha$$

Il existe plusieurs façons de déterminer cet ensemble. Un ensemble tel que $x_{\min}(\theta) \neq -\infty$ et $x_{\max}(\theta) \neq +\infty$ correspond à un intervalle dit bilatéral. Si $x_{\min}(\theta) = -\infty$ ou $x_{\max}(\theta) = +\infty$ l'intervalle est dit unilatéral. Un intervalle bilatéral tel que

$$P\left(x < x_{\min}(\theta); \theta\right) = P\left(x > x_{\max}(\theta); \theta\right) = (1-\alpha)/2 \tag{8.5}$$

conduit à un intervalle de confiance dit central. Un intervalle unilatéral tel que

$$P\left(x < x_{\min}(\theta); \theta\right) = 1 - \alpha$$

conduit à une limite supérieure sur θ (dans ce cas, $x_{\max}(\theta) = +\infty$ pour tout θ). La figure 8.1 montre les ensembles correspondant aux intervalles centraux (Eq. 8.5).

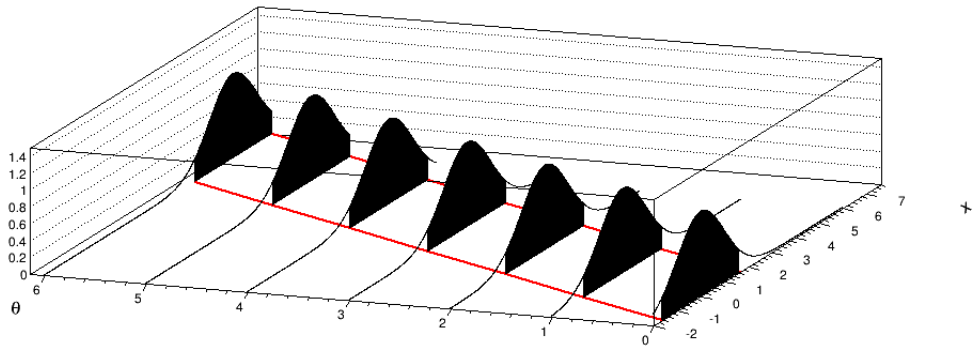


FIGURE 8.1 – Illustration de la construction de Neyman.

La deuxième étape consiste à construire, à partir des ensembles obtenus dans la première étape, la bande de confiance (*confidence belt*) dans le plan (θ, x) et à déduire de cette bande l'intervalle

$[\theta_{\min}\left(x^{\text{obs}}\right);\theta_{\max}\left(x^{\text{obs}}\right)]$ correspondant à la valeur observée de x. La bande de confiance ainsi que l'intervalle de confiance correspondant aux intervalles centraux de la figure 8.1 et à une observation x^{obs} sont représentés sur la figure 8.2.

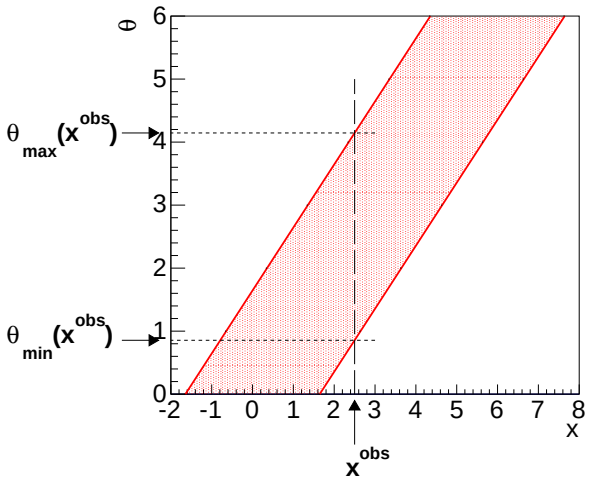


FIGURE 8.2 – Bande de confiance obtenue à partir de la figure 8.1.

La figure 8.3 représente la bande de confiance dans le cas unilatéral et pour la même valeur de α que précédemment (90%).

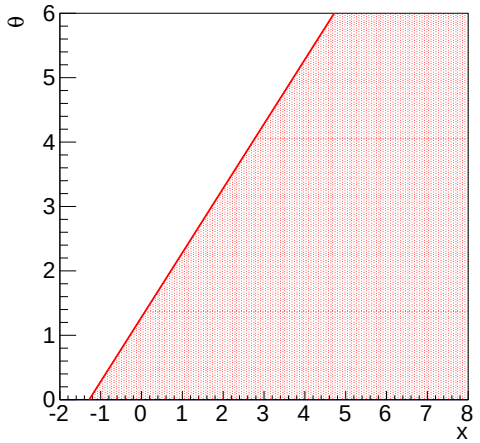


FIGURE 8.3 – Bande de confiance dans le cas unilatéral pour $\alpha = 90\%$.

La méthode décrite ci-dessus a une couverture qui, par construction, est égale exactement au niveau de confiance dans le cas continu. En effet, la probabilité pour que l'intervalle contienne la vraie valeur

est

$$P\left(\theta_0 \in [\theta_{\min}(x); \theta_{\max}(x)]; \theta_0\right) = P\left(x \in [x_{\min}\left(\theta_0\right); x_{\max}\left(\theta_0\right)]; \theta_0\right) = \alpha \tag{8.6}$$

8.3.2 *Flip-flopping*

La construction d'intervalle de confiance suivant la méthode décrite dans la section précédente suppose qu'une décision ait été prise à priori quant à la nature de cette intervalle (bilatéral ou unilatéral). Or il n'est souvent pas souhaitable de prendre une telle décision à priori mais plutôt de la baser sur le résultat de l'expérience (c'est ce que l'on appelle le *flip-flopping*).

Considérons par exemple le cas gaussien de la section précédente et le cas de *flip-flopping* suivant :

- si $x^{\text{obs}} < 3$ l'observation n'est pas significativement différente de 0, nous décidons alors de déterminer une limite supérieure sur θ
- si $x^{\text{obs}} >= 3$ l'observation est significativement différente de 0, nous décidons alors de déterminer un intervalle bilatéral.

Dans ce cas la bande de confiance devient celle représentée sur la figure 8.4 correspondant à la combinaison de la figure 8.3 jusqu'à $x = 3$ et de la figure 8.2 pour $x \geq 3$.

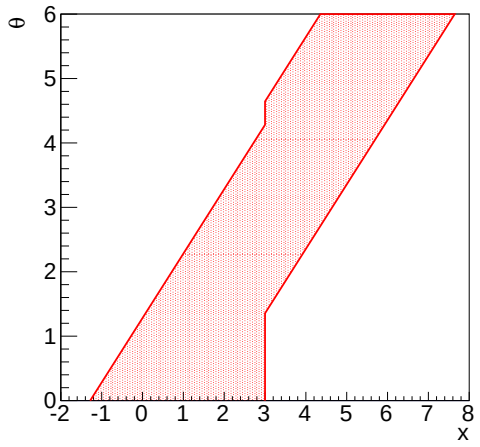


FIGURE 8.4 – Bande de confiance dans le cas de *flip-flopping* pour $\alpha = 90\%$.

Le problème avec ce type de constructions est que la couverture peut-être inférieure au niveau de confiance pour certaines valeurs de θ (*undercoverage*). Dans l'intervalle $1,36 < \theta < 4,28$, nous avons, pour $\alpha = 90\%$,

$$P\left(x < x_{\min}(\theta); \theta\right) = 10\%$$

et

$$P\left(x > x_{\max}(\theta); \theta\right) = 5\%$$

Donc

$$P\left(\theta \in [\theta_{\min}(x); \theta_{\max}(x)]; \theta\right) = P\left(x \in [x_{\min}\left(\theta\right); x_{\max}\left(\theta\right)]; \theta\right) = 85\%$$

Tout ce qui a été dit jusqu'ici sur la construction de Neyman et le problème du *flip-flopping* s'applique également aux distributions discrètes, à la différence que l'équation 8.6 n'est pas vérifiée et ceci même sans *flip-flopping*. En effet, les valeurs de x étant discrètes, il n'est pas possible d'avoir une couverture exactement égale au niveau de confiance. Dans ce cas nous choisissons, pour être conservatif, l'intervalle $[x_{\min}(\theta); x_{\max}(\theta)]$ de telle sorte à ce qu'il y ait *overcoverage*.

8.4 Intervalle de Feldman-Cousins

G. Feldman et R. Cousins ont développé une méthode pour construire des intervalles de confiance basée sur la construction de Neyman mais ne souffrant pas du problème de *flip-flopping* et se comportant mieux que ceux présentés précédemment proche des limites physiques sur les paramètres. Cette méthode consiste à baser l'inclusion d'une valeur de x dans l'intervalle $[x_{\min}(\theta); x_{\max}(\theta)]$ sur le rapport de *likelihoods*

$$R = \frac{\mathcal{L}(\theta; x)}{\mathcal{L}\left(\hat{\theta}; x\right)}$$

où $\hat{\theta}$ est l'estimateur ML de θ pour x. Les valeurs de x sont incluses dans $[x_{\min}(\theta); x_{\max}(\theta)]$ par ordre décroissant de R jusqu'à ce que la couverture soit supérieure ou égale au niveau de confiance (dans le cas continu il est en principe possible d'atteindre l'égalité parfaite alors que dans le cas discret nous faisons en sorte, comme mentionné précédemment, d'avoir *overcoverage*).

Exemple 8.3: Considérons une expérience de Poisson en présence d'un signal s et d'un bruit de fond b. Dans la suite nous prendrons $b = 3$. Cet exemple est extrait de l'article original de G. Feldman et R. Cousins [8]. Le *likelihood* est

$$P(n; s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

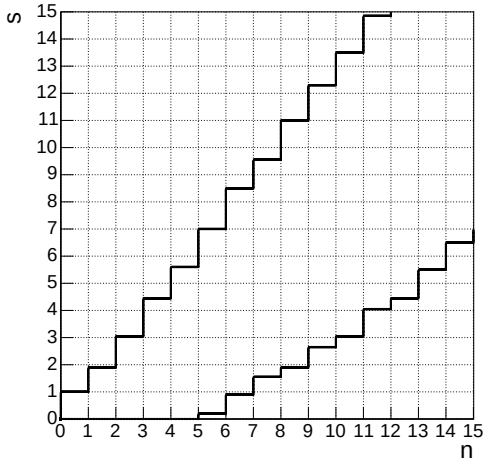
Pour chaque s, les valeurs de n incluses dans la bande de confiance sont obtenues en examinant le rapport

$$R = \frac{P(n; s)}{P(n; \hat{s})}$$

où $\hat{s}$ est l'estimateur ML de s. Nous ne considérons ici que le cas où s est positif. Nous contraignons donc son estimateur à être positif. Ainsi, $\hat{s} = \max(0, n - b)$. En guise d'exemple, prenons $s = 0,5$ et $\alpha = 90\%$. Les valeurs de $P(n; s)$, $\hat{s}$, $P(n; \hat{s})$ et R pour plusieurs valeurs de n sont données dans la table suivante.

n	$P(n\|s=0,5)$	$\hat{s}$	$P(n\|\hat{s})$	R	rang
0	0.030	0.	0.050	0.607	6
1	0.106	0.	0.149	0.708	5
2	0.185	0.	0.224	0.826	3
3	0.216	0.	0.224	0.963	2
4	0.189	1.	0.195	0.966	1
5	0.132	2.	0.175	0.753	4
6	0.077	3.	0.161	0.480	7
7	0.039	4.	0.149	0.259	
8	0.017	5.	0.140	0.121	
9	0.007	6.	0.132	0.050	
10	0.002	7.	0.125	0.018	
11	0.001	8.	0.119	0.006	

La valeur de n pour laquelle R est le plus grand est $n = 4$. C'est la première valeur incluse dans la bande de confiance. La deuxième valeur incluse est $n = 3$. La troisième $n = 2$ et ainsi de suite. L'ordre dans lequel les valeurs sont incluses est donné dans la dernière colonne de la table. Après l'inclusion de $n = 6$ la probabilité est 85,8%. L'ajout de $n = 7$ fait passer la probabilité à 93,5%, ce qui est supérieur à α. La bande de confiance pour $s = 0,5$ est donc formée par l'ensemble des valeurs $n \in [0, 6]$. Si nous répétons cette procédure pour chaque valeur de s, nous obtenons la bande de confiance représentée sur la figure suivante



Nous voyons que, même pour $n = 0$, l'intervalle de confiance n'est pas vide (ce qui serait le cas si nous avions construit une bande de confiance unilatérale ou centrée) et qu'il n'y a pas de *flip-flopping* (et donc pas *d'undercoverage*).

8.5 Construction par inversion d'un test d'hypothèse

Une autre méthode permettant de construire des intervalles de confiance est l'inversion d'un test d'hypothèse. Cette méthode fait appel aux notions décrites dans le chapitre 7. Elle permet de construire des intervalles unilatéraux et bilatéraux. Nous verrons par la suite que ce type de construction est dans certaines situations totalement identique à la construction de Neyman. L'intérêt de cette nouvelle méthode est qu'elle permet de traiter des cas complexes avec de nombreux paramètres de nuisance de façon relativement simple. Elle permet aussi dans des cas simples de fournir des formules analytiques qui peuvent se révéler d'un grand intérêt pratique.

8.5.1 Principe

Réaliser un test d'hypothèse consiste à regarder si la valeur observée lors d'une expérience se trouve dans la région critique ou non. Si elle s'y trouve, l'hypothèse est rejetée. Sinon, elle est acceptée. Inverser un test d'hypothèse revient à réaliser un test d'hypothèse pour chaque valeur du paramètre et à inclure dans l'intervalle de confiance toutes les valeurs du paramètre non exclues. Nous noterons ici le seuil de signification du test $1-\alpha$ afin d'être cohérent avec la définition de α utilisée précédemment dans ce chapitre. Cette notation est différente de celle utilisée au chapitre 7.

Considérons par exemple la construction d'un intervalle bilatéral centré pour la moyenne μ de la loi normale de variance σ^2 connue. La densité de probabilité est

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8.7}$$

Dans la suite nous noterons x_{obs} la valeur de x obtenue lors d'une expérience et nous prendrons $\sigma = 1$, $\alpha = 90\%$ et $x_{\text{obs}} = 0,9$. La densité de probabilité 8.7 ainsi que les régions critiques sont représentées sur la figure 8.5 pour différentes valeurs de μ. L'intervalle étant bilatéral la région critique est l'union de deux sous-régions de part et d'autre de la valeur moyenne. À chaque sous-région correspond une probabilité $(1-\alpha)/2$. D'après cette figure, nous voyons que $\mu = -1$ et $\mu = 3$ sont en dehors de l'intervalle de confiance car, pour ces deux valeurs, l'observation se trouve dans la région critique. La valeur $\mu = 0$ est en revanche incluse dans l'intervalle de confiance.

Les valeurs de μ qui composent l'intervalle de confiance sont telles que

$$x_{\text{obs}} \geq \Phi^{-1}\left(\frac{1-\alpha}{2}\right) + \mu \quad \text{et} \quad x_{\text{obs}} \leq -\Phi^{-1}\left(\frac{1-\alpha}{2}\right) + \mu$$

où Φ est la fonction de répartition de la loi normale centrée réduite. L'intervalle de confiance est donc

$$x_{\text{obs}} + \Phi^{-1}\left(\frac{1-\alpha}{2}\right) \leq \mu \leq x_{\text{obs}} - \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Avec les valeurs considérées ci-dessous, cela donne

$$-0,74 \leq \mu \leq 2,54$$

La description qui vient d'être faite dans le cas d'un intervalle bilatéral pour la moyenne d'une loi normale se généralise sans difficulté aux intervalles unilatéraux et aux autres densités de probabilité.

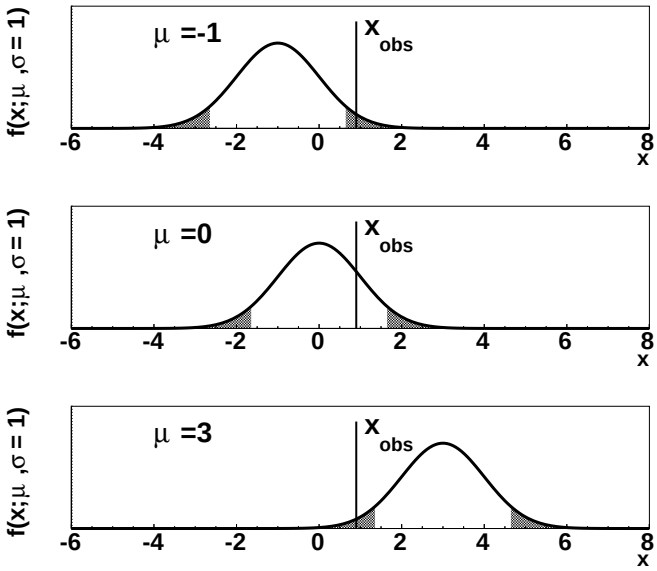


FIGURE 8.5 – Illustration de la construction d'un intervalle de confiance par inversion d'un test d'hypothèse. Les aires dans les queues de distribution correspondent aux régions critiques pour un seuil de signification de 10%.

8.5.2 Lien avec la construction de Neyman

La description qui a été faite dans la section précédente montre que la construction par inversion d'un test d'hypothèse correspond à la construction de Neyman, mais formulée dans un langage différent. Afin d'apporter plus de lumière sur la relation entre les deux méthodes, considérons un cas différent de celui considéré ci-dessus qui est celui de la loi de Poisson. La différence principale avec la loi normale est que la loi de Poisson est discrète.

Reprenons le problème considéré dans l'exemple 8.3 :

$$P(n;s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \tag{8.8}$$

et construisons des intervalles de confiance "classiques" (c'est-à-dire sans la méthode de Feldman-Cousins) unilatéraux et bilatéraux. Dans le cas unilatéral (établissement d'une limite supérieure), la valeur maximale de s non exclue s_{up} est donnée par

$$\sum_{n=0}^{n_{\mathrm{obs}}} \frac{(s_{\mathrm{up}}+b)^n}{n!} e^{-(s_{\mathrm{up}}+b)} = 1-\alpha \tag{8.9}$$

Le membre de gauche de cette expression est la fonction de répartition de la distribution de Poisson. Nous pouvons l'écrire en fonction de la fonction de répartition F_{χ^2} de la loi de χ^2 :

$$\sum_{n=0}^{n_{\mathrm{obs}}} \frac{(s_{\mathrm{up}}+b)^n}{n!} e^{-(s_{\mathrm{up}}+b)} = 1 - F_{\chi^2}\left(2\left(s_{\mathrm{up}}+b\right); 2\left(n_{\mathrm{obs}}+1\right)\right)$$

L'intérêt d'une telle opération est de faire passer s_{up} de paramètre à variable et donc de pouvoir l'exprimer en inversant F_{χ^2}. Nous trouvons

$$s_{\text{up}} = -b + \frac{1}{2} F_{\chi^2}^{-1}\left(\alpha; 2\left(n_{\text{obs}} + 1\right)\right) \tag{8.10}$$

La *p-value* apparaissant dans l'équation 8.9, à partir de laquelle le résultat 8.10 a été obtenu, est parfois appelée CL_{s+b}. La méthode employée pour en arriver à 8.10 est par conséquent parfois appelée "méthode CL_{s+b}". Un calcul similaire dans le cas d'un intervalle bilatéral conduit à

$$-b + \frac{1}{2} F_{\chi^2}^{-1}\left(\frac{1-\alpha}{2}; n_{\text{obs}}\right) \leq s \leq -b + \frac{1}{2} F_{\chi^2}^{-1}\left(\frac{\alpha}{2}; 2\left(n_{\text{obs}} + 1\right)\right) \tag{8.11}$$

Les intervalles donnés par les équations 8.10 et 8.11 sont représentés pour $b = 3$ sur la figure 8.6 avec les intervalles obtenus par la construction de Neyman. Nous voyons que, même dans le cas discret, les deux méthodes sont identiques. La différence dans la représentation (la construction par inversion d'un test d'hypothèse fournit une représentation continue de s en fonction de n alors que la construction de Neyman produit des marches) n'a pas d'incidence sur le résultat car n est discret et, pour chaque valeur possible, les courbes noires et rouges coïncident.

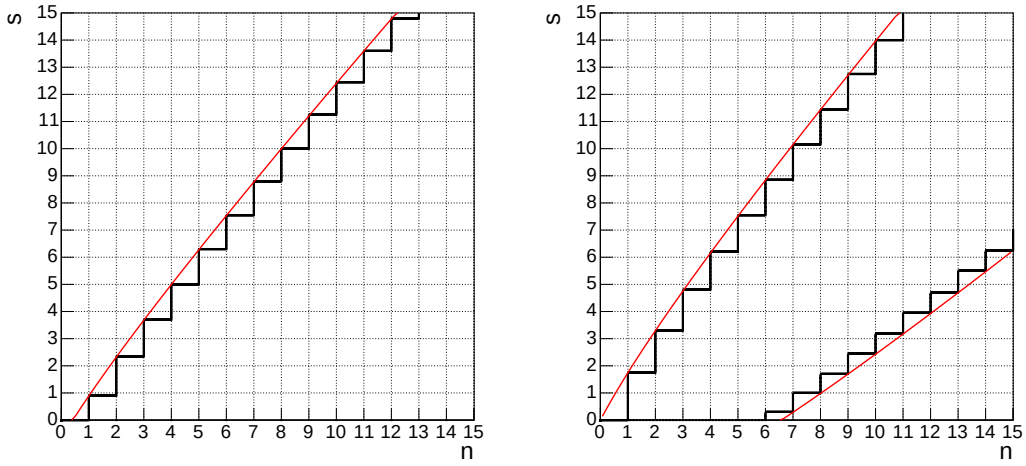


FIGURE 8.6 – Intervalle de confiance unilatéral (à gauche) et bilatéral (à droite) pour le paramètre s de la loi de Poisson (Eq. 8.8) obtenus par construction de Neyman (en noir) et inversion de test d'hypothèse (en rouge) pour $\alpha = 0,9$ et $b = 3$.

La figure 8.6 montre que, pour $n = 0$, l'intervalle de confiance unilatéral est vide. Ceci est souvent considéré comme indésirable et il serait préférable de construire des intervalles de confiance unilatéraux non vides pour $n = 0$. Nous avons vu dans la section 8.4 une méthode permettant de le faire. Dans la section suivante, nous allons décrire une autre méthode particulièrement à la mode en physique des particules.

8.5.3 Méthode CL_s pour le calcul de limite

Comme nous l'avons vu dans la section précédente, l'intervalle unilatéral pour le paramètre s de la loi de Poisson est vide pour $n = 0$. Une méthode permettant de construire des intervalles de confiance non vide est la méthode CL_s. Celle-ci consiste à remplacer, dans 8.9, CL_{s+b} par $CL_s = CL_{s+b}/CL_b$, où CL_b est donné par

$$CL_b = \sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!} e^{-b}$$

c'est-à-dire par CL_{s+b} avec $s = 0$. L'équation à résoudre est donc

$$\frac{\sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}}+b)^n}{n!} e^{-(s_{\text{up}}+b)}}{\sum_{n=0}^{n_{\text{obs}}} \frac{b^n}{n!} e^{-b}} = 1 - \alpha \tag{8.12}$$

CL_b ne dépend pas de s_{up}. Résoudre 8.12 revient donc à résoudre 8.9 en remplaçant $1 - \alpha$ par $(1 - \alpha) \times CL_b$. La méthode CL_s est donc identique à la méthode CL_{s+b} avec un seuil de signification réduit. La relation 8.10 devient donc

$$s_{\text{up}} = -b + \frac{1}{2} F_{\chi^2}^{-1} \left(1 - (1 - \alpha) \left[1 - F_{\chi^2}\left(2b; 2(n_{\text{obs}} + 1)\right)\right]; 2\left(n_{\text{obs}} + 1\right)\right)$$

La figure 8.7 montre les intervalles de confiance unilatéraux trouvés par la construction de Neyman, la méthode CL_{s+b} (comme sur la figure 8.6) et la méthode CL_s(la bande de confiance bilatérale de Feldman-Cousins trouvée dans l'exemple 8.3 est également rappelée). Les trois méthodes sont équivalentes pour les grandes valeurs de n. La méthode CL_s se distingue des deux autres à petit n, où elle produit des intervalles de confiance plus grands. Nous voyons aussi que, à petit n, la limite supérieure obtenue par la méthode CL_s est plus conservative que celle obtenue par la méthode de Feldman-Cousins. Ceci n'est pas surprenant, cette dernière ayant une converture égale au niveau de confiance alors que la méthode CL_s présente un *overcoverage*.

La figure 8.8 montre une comparaison des résultats obtenus par les méthodes CL_s et CL_{s+b} pour différentes valeurs de b et de n_{obs} et pour un niveau de confiance de 95%. Nous voyons que CL_s ne produit jamais d'intervalle vide. Même pour $n_{\text{obs}} = 0$ et $b > 0$, l'intervalle ne l'est pas : pour toutes les valeurs de b nous avons $s_{\text{up}} = 3$. Pour $n_{\text{obs}} > 0$, la limite supérieure tend vers 3 lorsque b tend vers l'infini car

$$CL_{s+b} \xrightarrow[b\to\infty]{} e^{-s} CL_b$$

L'équation 8.12 devient donc

$$e^{-s_{\text{up}}} = 1 - \alpha$$

ce qui donne, à 95% CL,

$$s_{\text{up}} = \ln\left(\frac{1}{1 - 0,95}\right) = 3$$

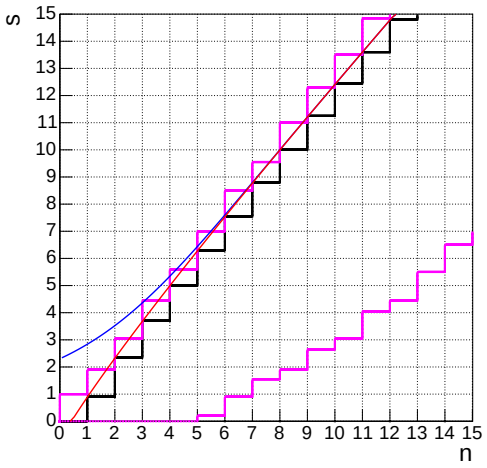


FIGURE 8.7 – Intervalle de confiance unilatéral pour le paramètre s de la loi de Poisson (Eq. 8.8) obtenus par la construction de Neyman (en noir), la méthode CL_b (en rouge) et la méthode CL_s (en bleu) pour $\alpha = 0,9$ et $b = 3$. La bande de confiance trouvée par la méthode Feldman-Cousins est également montrée en magenta.

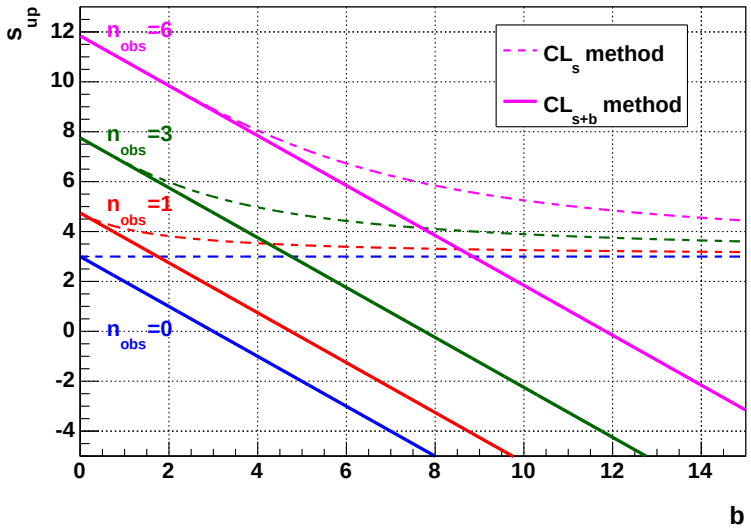


FIGURE 8.8 – Limite supérieure à 95% CL sur s en fonction de b pour différentes valeurs de n_{obs}. Les courbes pleines (pointillées) ont été obtenues avec la méthode CL_{s+b} (CL_s).

8.5.4 Combinaison de plusieurs mesures

Méthode fréquentiste classique

Utilisation du *likelihood ratio*

8.5.5 Prise en compte des paramètres de nuisance

Chapitre 9

Inférence bayésienne

L'inférence bayésienne repose sur la construction, à partir du théorème de Bayes, de la distribution *a posteriori* donnant la probabilité qu'une théorie soit vraie ou qu'un paramètre ou un ensemble de paramètres de la théorie prenne telle ou telle valeur. Cette distribution s'écrit (nous utilisons ici la notation dans le cas discret mais cette expression reste vraie dans le cas continu, il suffit de remplacer les probabilités par les densités de probabilité)

$$P(\text{théorie}|\text{données}) = \frac{P(\text{données}|\text{théorie})P(\text{théorie})}{P(\text{données})}$$

Le terme $P(\text{données}|\text{théorie})$ correspond au *likelihood* et le terme $P(\text{théorie})$ à la distribution *a priori*. Le terme $P(\text{données})$ est une constante de normalisation. Il est courant de la supprimer de la notation et d'écrire

$$P(\text{théorie}|\text{données}) \propto P(\text{données}|\text{théorie}) \times P(\text{théorie})$$

9.1 Inférence à partir de la distribution *a posteriori*

Toute l'inférence sur la théorie se fait à partir de la distribution *a posteriori*. Afin d'illustrer ceci, considérons le problème de l'estimation d'un paramètre θ. Notons, comme dans les chapitres précédents, x l'échantillon de données (qui peut être multidimensionnel) et $\mathcal{L}$ le *likelihood*. Nous avons

$$f(\theta|x) \propto \mathcal{L} \times g(\theta)$$

où $g(\theta)$ est la distribution *a priori*.

Tous les types d'estimations (ponctuelle, calcul d'intervalle de confiance et calcul de limite) se font à partir de $f(\theta|x)$. Alors que ces différents types d'estimations se font souvent par des techniques différentes dans le cas fréquentiste, nous voyons que dans le cas bayésien elles découlent toutes de la même fonction.

9.1.1 Estimation ponctuelle

L'estimation ponctuelle de θ peut se faire à partir des différentes méthodes de mesure de localisation décrites en 2.1 appliquées à la distribution *a posteriori*. Nous pouvons par exemple choisir l'espérance

comme estimation ponctuelle de θ :

$$\hat{\theta} = \int \theta f(\theta|x) \mathrm{d}\theta$$

ou bien le mode :

$$\hat{\theta} = \max_{\theta} f(\theta|x)$$

ou encore la médiane :

$$F_\theta(\hat{\theta}) = \int_{-\infty}^{\hat{\theta}} f(\theta|x) \mathrm{d}\theta = 1/2$$

où F_θ est la fonction de répartition de la distribution *a posteriori*.

Il est intéressant de noter que, dans le cas d'une distribution *a priori* uniforme, le mode correspond à l'estimateur par maximum de vraisemblance rencontré en 6.3.

9.1.2 Intervalle de confiance

La construction d'un intervalle de confiance nécessite, comme dans le cas fréquentiste, le choix d'un niveau de confiance. Le terme "niveau de confiance" ne recouvre toutefois pas la même signification que dans le cas fréquentiste et nous préférerons utiliser le terme "niveau de crédibilité" dans le contexte bayésien. En effet, en analyse fréquentiste la notion de niveau de confiance sous-entend celle de couverture. Dans le cas bayésien, l'inférence ne repose que sur les données réellement obtenues lors de l'expérience et pas sur des pseudo-données comme dans le cas fréquentiste. La notion de couverture n'a donc pas de signification.

Un intervalle de confiance $[\theta_1; \theta_2]$ avec un niveau de crédibilité α est tel que

$$\int_{\theta_1}^{\theta_2} f(\theta|x) \mathrm{d}\theta = \alpha$$

Plusieurs choix peuvent être fait pour θ_1 et θ_2. Un intervalle centré est tel que

$$\int_{-\infty}^{\theta_1} f(\theta|x) \mathrm{d}\theta = \int_{\theta_2}^{\infty} f(\theta|x) \mathrm{d}\theta = (1-\alpha)/2$$

L'intervalle le plus étroit est tel que $f(\theta_1|x) = f(\theta_2|x)$ (pour tout $\theta \in [\theta_1; \theta_2]$ et $\theta_c \notin [\theta_1; \theta_2]$, $f(\theta|x) > f(\theta_c|x)$). L'intervalle centré sur la moyenne est tel que

$$\int_{\theta_1}^{\mathbb{E}[\theta]} f(\theta|x) \mathrm{d}\theta = \int_{\mathbb{E}[\theta]}^{\theta_2} f(\theta|x) \mathrm{d}\theta = \alpha/2$$

Il est aussi possible de définir des intervalles unilatéraux ($\theta_1 = -\infty$ ou $\theta_2 = +\infty$). Un intervalle unilatéral avec $\theta_1 = -\infty$ conduit à une limite supérieure sur θ égale à θ_2.

9.2 Accroissement de la connaissance par l'approche bayésienne

L'approche bayésienne offre l'avantage d'expliquer de manière simple et naturelle la façon dont la connaissance progresse grâce à l'expérience. La connaissance que nous avons a priori (c'est-à-dire avant que l'expérience ne soit réalisée) est augmentée grâce aux données récoltées dans l'expérience. Cette augmentation se mesure en comparant les distributions *a priori* et *a posteriori*. Si la distribution *a posteriori* est très différente de la distribution *a priori*, l'expérience a apporté beaucoup d'informations. Si en revanche les deux distributions sont similaires alors l'expérience n'a servi à rien. Dans le cas où l'expérience contient énormément d'informations sur le ou les paramètres à estimer la distribution *a posteriori* dépend principalement des données recueillies lors de celle-ci et relativement peu de la distribution *a priori*.

Ce processus d'accroissement de la connaissance est représenté sur le schéma suivant :

$$\text{connaissance a priori} \xrightarrow[\textbf{expérience}]{} \text{connaissance a posteriori}$$

En utilisant les grandeurs mathématiques :

$$g(\theta) \xrightarrow[x]{} f(\theta|x)$$

Cette interprétation de l'approche bayésienne est intéressante car elle permet de comprendre comment la connaissance s'accroît d'expérience en expérience. Supposons par exemple que deux expériences soient réalisées à plusieurs années d'intervalle pour contrainte le même paramètre théorique. Notons $\mathcal{L}_1$ et $\mathcal{L}_2$ (x_1 et x_2) les *likelihoods* (échantillons de données) pour la première et la deuxième expérience respectivement. La distribution *a posteriori* obtenue à l'issue de la première expérience est

$$f_1(\theta|x_1) \propto \mathcal{L}_1 \times g(\theta)$$

Lorsque la deuxième expérience est réalisée, il serait dommage de prendre

$$f_2(\theta|x_2) \propto \mathcal{L}_2 \times g(\theta)$$

comme distribution *a posteriori* car les données obtenues lors de la première seraient totalement ignorées. Il est plus avantageux d'en tenir compte et d'utiliser la distribution *a posteriori* de la première expérience comme distribution *a priori* de la deuxième :

$$f_2(\theta|x_2, x_1) \propto \mathcal{L}_2 \times f_1(\theta|x_1) \propto \mathcal{L}_2 \times \mathcal{L}_1 \times g(\theta)$$

L'application séquentielle du théorème de Bayes que nous venons de décrire est équivalente à l'application unique tenant compte des données recueillies dans les deux expériences en même temps (à la condition qu'elles soient indépendantes). En effet, nous avons, en notant $f_3(\theta|x_1, x_2)$ la distribution *a posteriori* obtenue lors d'une application unique,

$$f_3(\theta|x_1, x_2) = f(x_1, x_2|\theta)g(\theta)$$

où $f(x_1, x_2|\theta)$ est la densité de probabilité conjointe des deux expériences. Si elles sont indépendantes, $f(x_1, x_2|\theta) = \mathcal{L}_1\mathcal{L}_2$. $f_3(\theta|x_1, x_2)$ est donc identique à $f_2(\theta|x_2, x_1)$.

Dans le cas où n expériences sont réalisées, nous avons

$$f_n(\theta|x_n,\ldots,x_1) \propto \mathcal{L}_n \times \mathcal{L}_{n-1} \times \ldots \times \mathcal{L}_1 \times g(\theta)$$

Notons que les conclusions auxquelles nous sommes arrivés restent valables si les expériences ne sont pas indépendantes (c'est-à-dire si x_k dépend de x_{k-1}). Dans ce cas, la densité de probabilité conjointe des n expériences, au lieu de s'écrire comme le produit de *likelihoods* indépendants, s'écrit

$$f(x_1,\ldots,x_n|\theta) = f(x_1|\theta)f(x_2|x_1,\theta)\ldots f(x_n|x_1,x_2,\ldots,x_{n-1},\theta)$$

9.3 Choix de la distribution *a priori*

L'inférence bayésienne nécessite, comme nous l'avons vu, le choix d'une distribution *a priori*. Ce choix est dans une grande mesure arbitraire. Ce fait est vu par plusieurs personnes comme un argument contre l'approche bayésienne. A contrario, les partisants de l'approche bayésienne ne comprennent pas comment les fréquentistes peuvent dire quelque chose sur la théorie sans cet à priori.

Les bayésiens eux-mêmes ne sont souvent pas d'accord sur les règles à suivre pour choisir la distribution *a priori*. Certains pensent qu'il faut choisir les distributions les moins informatives possibles pour réduire leur impact sur l'inférence (ils qualifient ces distributions d'objectives). D'autres au contraire acceptent que l'inférence puisse dépendre parfois dans une grande mesure de la distribution *a priori* et cherchent à construire des distributions qui reflètent le mieux la connaissance à priori. Ces derniers ne voient pas l'utilisation d'une distribution *a priori* comme une limitation de la méthode bayésienne mais comme quelque chose d'inhérent à la méthode et d'inévitable.

Dans cette section nous donnerons quelques éléments permettant de comprendre comment les distributions *a priori* objectives sont construites.

Il est important de se souvenir que le choix de la distribution *a priori* n'est important que lorsque la taille de l'échantillon est relativement petite. Lorsque celle-ci augmente, le *likelihood* prend de plus en plus d'importance par rapport à la distribution *a priori*. La distribution *a posteriori* devient donc de plus en plus indépendante de la distribution *a priori*.

9.3.1 Information

9.3.2 Distribution *a priori* objective

9.4 Exemple élémentaire

Considérons une expérience de comptage réalisée dans le but de déterminer le paramètre μ de la loi de Poisson. Le *likelihood* est

$$P(N|\mu) = \frac{\mu^N}{N!}e^{-\mu}$$

Prenons une distribution *a priori* uniforme (le choix de cette distribution sera discuté plus en détail dans la section 9.3). La distribution *a posteriori* est donc

$$f(\mu|N) = \frac{\frac{\mu^N}{N!}e^{-\mu}}{\int_0^\infty \frac{\mu^N}{N!}e^{-\mu}\mathrm{d}\mu}$$

Le dénominateur de l'expression précédente est imposé par la condition de normalisation. En simplifiant, nous obtenons

$$f(\mu|N) = \frac{\mu^N e^{-\mu}}{\int_0^\infty \mu^N e^{-\mu} \mathrm{d}\mu}$$

Nous reconnaissons la fonction gamma au dénominateur, qui est égale à $N!$. Donc

$$f(\mu|N) = \frac{\mu^N}{N!} e^{-\mu}$$

La distribution *a posteriori* est donc identique au *likelihood*. Il s'agit de la distribution gamma avec $a = 1$ et $b = N + 1$ (voir A.9). Elle est représenté sur la figure 9.1 pour quatre valeurs de N.

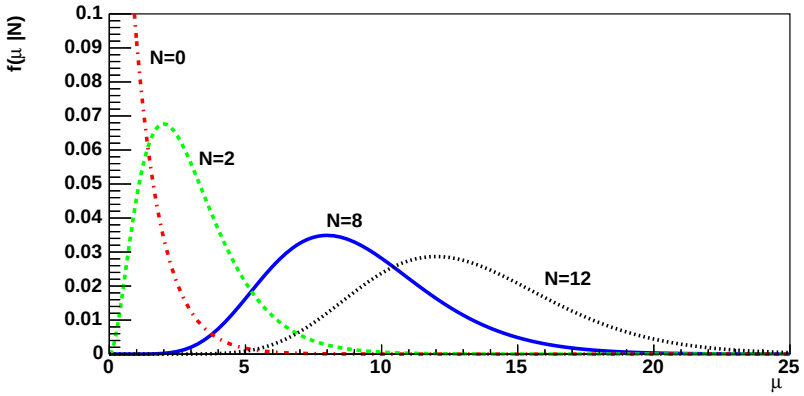


FIGURE 9.1 – Distribution *a posteriori* pour différentes valeurs de N.

Dans le cas $N = 0$, la distribution *a posteriori* est $e^{-\mu}$ et sa fonction de répartition est $1 - e^{-\mu}$. Il est naturel dans ce cas de déterminer une limite supérieure sur μ. La limite μ_{up} à 95% CI est donnée par

$$1 - e^{-\mu_{\mathrm{up}}} = 0,95$$

soit

$$\mu_{\mathrm{up}} = \ln 20 \simeq 3$$

Nous trouvons ici un résultat identique à celui trouvé en inférence fréquentiste par la méthode CL_s (voir section 8.5.3). Ce n'est pas un hasard. Nous verrons dans la section 9.7 que la méthode bayésienne avec une distribution *a priori* uniforme est identique à la méthode CL_s pour le calcul de limite.

Dans le cas $N = 8$ il est par contre plus naturel de reporter un intervalle bilatéral. Les intervalles centré et centré sur la moyenne à 68% CI sont representés sur la figure 9.2.

Considérons maintenant le cas où n expériences de comptages sont réalisées. Soit N_i le nombre d'événements mesurés dans la $i^{\mathrm{ème}}$ expérience ($i \in [1, n]$). La distribution *a posteriori* est

$$f(\mu|\{N_i\}) \propto \prod_{i=1}^{n} \frac{\mu^{N_i}}{N_i!} e^{-\mu} \times g(\mu) \propto e^{-n\mu} \mu^{N_{\mathrm{tot}}} \times g(\mu)$$

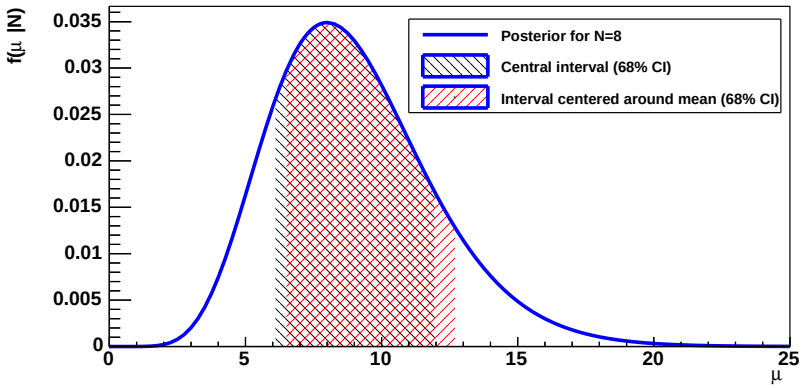


FIGURE 9.2 – Distribution *a posteriori* pour $N = 8$ avec deux intervalles à 68% CI.

où $g(\mu)$ est la distribution *a priori* et $N_{\text{tot}} = \sum_i N_i$. Il a été dit précédemment que, lorsque la taille de l'échantillon augmente, la distribution *a posteriori* (et donc l'inférence) dépend de moins en moins de la distribution *a priori*. Afin d'illustrer ceci, nous avons représenté les distributions *a priori* et *a posteriori* sur la figure 9.3 pour $(n, N_{\text{tot}}) = (1, 3),\ (5, 15)$ et $(10, 30)$ avec :

- $g(\mu) =$ constante (distribution *a priori* uniforme)
- $g(\mu) = e^{\mu}$ (distribution *a priori* exponentielle)

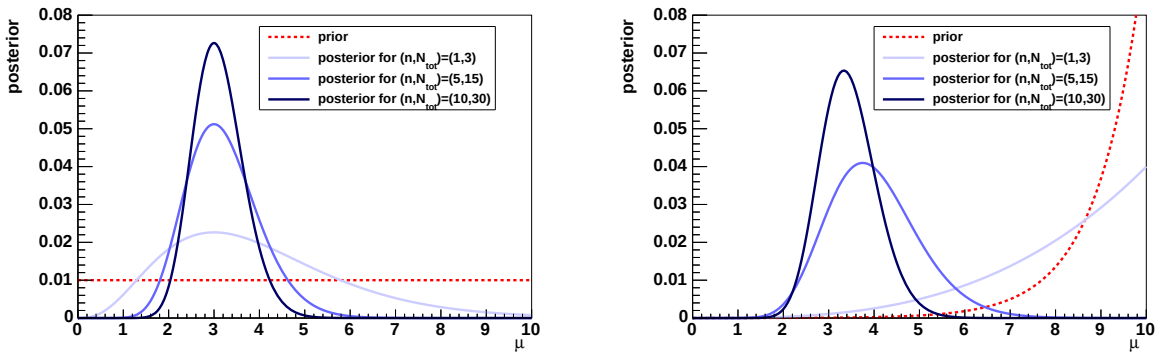


FIGURE 9.3 – Distribution *a posteriori* pour différentes tailles d'échantillon avec une distribution *a priori* uniforme (à gauche) et exponentielle (à droite).

Pour $(n, N_{\text{tot}}) = (1, 3)$ les distributions *a posteriori* dans le cas d'une distribution *a priori* uniforme et exponentielle sont très différentes l'une de l'autre. L'inférence dépend fortement de la connaissance à priori. Lorsque $(n, N_{\text{tot}}) = (10, 30)$ les distributions *a posteriori* sont par contre assez similaires. Elles tendent l'une vers l'autre dans la limite asymptotique. Le cas d'une distribution *a priori* exponentielle est particulièrement intéressant car il montre que la distribution *a posteriori* est localisée, lorsque

l'échantillon est grand, là où la distribution *a priori* a très peu de poids. Les données tendent à favoriser dans ce cas des valeurs de μ très différentes de celles supposées a priori.

9.5 Le *likelihood principle*

Le *likelihood principle* est le principe suivant lequel toute l'information nécessaire à l'inférence est contenue dans le *likelihood* et les données obtenues lors de l'expérience (et pas dans des données autres que celle obtenues). De plus, deux *likelihoods* identiques à une fonction des données près contiennent la même information sur θ. L'approche bayésienne de l'inférence satisfait à ce principe (comme nous le verrons dans un instant) alors que l'approche fréquentiste non.

Ce principe fait l'objet de nombreux débats depuis de nombreuses années. Certains pensent qu'il est fondamental dans le problème de l'inférence et qu'il est donc absolument nécessaire de le satisfaire. Ces personnes rejettent donc l'approche fréquentiste. D'autres personnes pensent qu'il n'est pas si fondamental que cela et accepte donc l'inférence fréquentiste.

Il est immédiat de voir que l'approche bayésienne satisfait au *likelihood principle* à partir de

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)\mathrm{d}\theta}$$

En effet, seules les données obtenues lors de l'expérience x sont utilisées et multiplier $f(x|\theta)$ par une fonction de x (indépendante de θ) conduit à la même distribution *a posteriori* et donc à la même inférence sur θ.

L'approche classique ne satisfait pas au *likelihood principle* car l'inférence ne se base pas uniquement sur les données recueillies lors de l'expérience mais aussi sur les données que l'on suppose recueillir si l'expérience est répétée plusieurs fois. Considérons par exemple le problème de l'estimation d'un paramètre. Nous avons vu au chapitre 6 que, dans l'approche classique, un bon estimateur doit vérifier un certain nombre de propriétés. Une de ces propriétés est l'absence de biais. Rappelons qu'un estimateur $\hat{\theta}(x)$ est non biaisé si

$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta \tag{9.1}$$

Si par conséquent l'inférence consiste en la recherche d'un estimateur non biaisé, elle se base non pas uniquement sur les données recueillies mais aussi sur les données non recueillies (que l'on recueillerait si l'expérience était répétée) qui interviennent dans le calcul de l'espérance dans 9.1. Le non respect du *likelihood principle* dans l'approche classique ne se limite pas au problème de l'estimation de paramètre mais se produit aussi dans les tests d'hypothèse. La description que nous avons faite au chapitre 7 des tests d'hypothèses fréquentiste montre que, pour construire un test d'hypothèse, il faut construire la distribution du test statistique et donc utiliser la aussi les résultats d'expériences autres que celle réalisée.

Exemple 9.1: Dans cet exemple nous décrivons une situation illustrant particulièrement bien comment la méthode fréquentiste peut violer le *likelihood principle* et donc comment elle se distingue de la méthode bayésienne. Considérons les deux expériences suivantes :

- Expérience 1 : n expériences de Bernouilli de paramètre θ sont réalisées au cours desquelles k succès sont obtenus. Cette expérience est décrite par la loi binomiale :

$$P_{\text{Bin}}(k;\theta,n) = \binom{n}{k}\theta^k\,(1-\theta)^{n-k}$$

- Expérience 2 : des expériences de Bernouilli de paramètre θ sont réalisées jusqu'à ce que k succès soient obtenus. Soit n le nombre total d'expériences. n est décrit par la loi binomiale négative :

$$P_{\text{NegBin}}(n;\theta,k) = \binom{n-1}{k-1}\theta^k\,(1-\theta)^{n-k}$$

Dans les deux cas, l'inférence porte sur le paramètre θ. La méthode bayésienne conduit à des résultats identiques car les *likelihoods* sont identiques à une fonction indépendante de θ près. La distribution *a posteriori* est, quelque soit l'expérience,

$$f(\theta|n,k) = \frac{\theta^k\,(1-\theta)^{n-k}\,g(\theta)}{\displaystyle\int \theta^k\,(1-\theta)^{n-k}\,g(\theta)\mathrm{d}\theta}$$

L'estimation de θ par la méthode fréquentiste nécessite le choix de certains critères pour définir l'estimateur. Si nous décidons que l'estimateur doit être non biaisé, alors les seuls estimateurs valables $\hat{\theta}$ sont ceux pour lesquels

$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta$$

Dans le cas de l'expérience 1, l'estimateur non biaisé est l'estimateur ML

$$\hat{\theta}_1 = \frac{k}{n}$$

Dans le cas de l'expérience 2 l'estimateur ML (qui est le même que dans l'expérience 1) est biaisé. L'estimateur non biaisé est (nous laissons au lecteur le soin de le démontrer)

$$\hat{\theta}_2 = \frac{k-1}{n-1}$$

Nous voyons donc que, dans le cas fréquentiste, le résultat de l'inférence dépend de la nature de l'expérience. Dans le cas bayésien elle n'en dépend pas. Deux expériences produisant les mêmes résultats conduisent à la même inférence, même si elles sont de nature différente.
Considérons maintenant le cas d'un test d'hypothèse réalisé sur les deux expériences. Les hypothèses testées sont $H_0 : \theta = 0,5$ et $H_1 : \theta > 0,5$. Comme pour l'estimateur du paramètre θ, le traitement bayésien conduit aux mêmes conclusions quel que soit l'expérience (la distribution *a posteriori* étant la même) alors que le traitement fréquentiste non. Prenons par exemple $k = 3$ et $n = 12$. Les *p-values* pour les deux expériences sont :

- Expérience 1 : $p_1 = \sum_{k'=0}^{3} \binom{12}{k'} 0,5^{k'} (1-0,5)^{12-k'} = 0,073$
- Expérience 2 : $p_2 = \sum_{n'=12}^{\infty} \binom{n'-1}{3-1} 0,5^3 (1-0,5)^{n'-3} = 0,033$

Si nous utilisons la valeur standard de 0,05 pour le seuil de signification du test, nous voyons que dans le cas de l'expérience 1 l'hypothèse nulle est rejetée alors que dans le cas de l'hypothèse 2 elle est acceptée.

9.6 Marginalisation

Nous n'avons considéré précédemment que le cas où le *likelihood* dépend uniquement du paramètre d'intérêt (c'est-à-dire du paramètre sur lequel l'inférence porte). Il arrive souvent (en fait tout le temps dans les cas réalistes) qu'il dépende en plus d'autres paramètres. Ces derniers portent le nom de paramètres de nuisance. En statistique bayésienne, ils sont traités par marginalisation (voir 2.9). À l'issue de la marginalisation, le *likelihood* et la distribution *a posteriori* ne dépendent plus que du paramètre d'intérêt.

Considérons un *likelihood* dépendant d'un paramètre d'intérêt θ et d'un paramètre de nuisance ν : $\mathcal{L}(\theta, \nu)$. Soit $g(\nu)$ la distribution *a priori* de ν. Le *likelihood* marginalisé est

$$\mathcal{L}(\theta) = \int \mathcal{L}(\theta, \nu) g(\nu) \mathrm{d}\nu$$

La distribution *a posteriori* est donc (en notant $g(\theta)$ la distribution *a priori* de θ)

$$f(\theta|x) \propto \mathcal{L}(\theta) g(\theta) \propto \int \mathcal{L}(\theta, \nu) g(\nu) g(\theta) \mathrm{d}\nu$$

Plutôt que de construire d'abord le *likelihood* marginalisé puis la distribution *a posteriori*, nous pouvons de manière équivalente construire d'abord la distribution *a posteriori* multidimensionnelle puis l'intégrer sur le paramètre de nuisance :

$$f(\theta, \nu|x) \propto \mathcal{L}(\theta, \nu) g(\theta) g(\nu) \xrightarrow[\text{marginalisation}]{} f(\theta|x) = \int f(\theta, \nu|x) \mathrm{d}\nu$$

La procédure que nous venons de décrire dans le cas à un paramètre de nuisance se généralise directement au cas où le *likelihood* dépend de plus d'un paramètre.

Exemple 9.2: Reprenons le cas de l'expérience de Poisson considérée dans la section 9.4 mais en supposant maintenant que deux types de processus contribuent au comptage : un signal et un bruit de fond. Notons s (b) le nombre d'événements de signal (bruit de fond) attendu. Le *likelihood* est

$$P(N|s, b) = \frac{(s+b)^N}{N!} e^{-(s+b)}$$

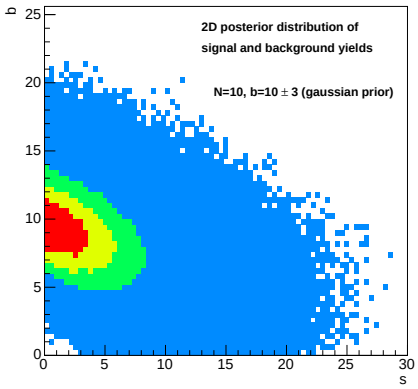
Supposons de plus que le bruit de fond ne soit pas connu de manière certaine. Soit b_0 la valeur nominale du bruit de fond attendu et σ son incertitude. Un choix possible dans ce cas est de prendre pour b une distribution *a priori* normale de moyenne b_0 et variance σ^2 (ce choix n'est pas forcément le meilleur mais nous le faisons pour des raisons de simplicité et laissons de côté les difficultés qui lui sont associées) :

$$g(b|b_0,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b-b_0)^2}{2\sigma^2}}$$

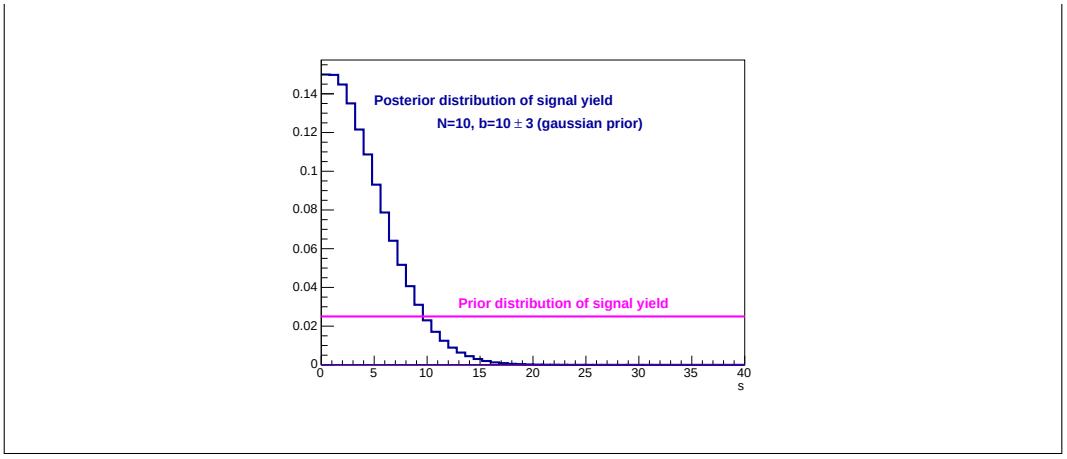
En prenant une distribution *a priori* uniforme pour s nous obtenons la distribution *a posteriori* suivante :

$$\begin{aligned} f(s|N) = \int f(s,b|N)\mathrm{d}b &\propto \int P(N|s,b)g(b|b_0,\sigma)\mathrm{d}b \\ &\propto \int (s+b)^N e^{-(s+b)} e^{-\frac{(b-b_0)^2}{2\sigma^2}} \mathrm{d}b \end{aligned}$$

L'intégrale apparaissant dans cette expression n'est pas calculable analytiquement, il faut avoir recours à des outils numériques. Le calcul à l'aide d'une chaîne de Markov conduit à la distribution *a posteriori* représentée sur la figure suivante.



Cette distribution bi-dimensionnelle montre un comportement général qui est que les variables dont dépend le *likelihood* sont correlées après l'expérience bien qu'elles soient considérées comme décorrélées avant. Dans notre cas s et b sont anti-corrélées, ce qui est attendu puisque le *likelihood* ne dépend que de leur somme. L'inférence ne portant que sur s, il faut intégrer la distribution *a posteriori* bi-dimensionnelle sur b pour construire la distribution *a posteriori* unidimensionnelle de s. Celle-ci est montrée sur la figure ci-dessous. Les valeurs de s favorisées par les données sont proche de 0. Il est naturel dans ce cas de reporter une limite supérieure. Nous trouvons, avec un niveau de crédibilité de 95%, que $s < 9.95$.



9.7 Equivalence entre la méthode CL_s et la méthode bayesienne pour le calcul de limite dans le cas poissonnien

En inférence bayésienne, l'obtention d'une limite supérieure sur un paramètre θ s'obtient, comme nous l'avons vu dans la section 9.1.2, à partir de la fonction de répartition de la distribution *a posteriori*. En notant θ_{up} la limite supérieur correspondant à un niveau de crédibilité α, nous avons

$$\int_{-\infty}^{\theta_{\text{up}}} f(\theta|x)\mathrm{d}\theta = \alpha$$

Dans le cas d'une expérience poissonnienne en présence d'un signal et d'un bruit de fond (supposé parfaitement connu), cela donne, avec une distribution *a priori* uniforme pour $s \geq 0$ et nulle pour $s < 0$,

$$\frac{\int_0^{s_{\text{up}}} (s+b)^N e^{-(s+b)}\mathrm{d}s}{\int_0^{\infty} (s+b)^N e^{-(s+b)}\mathrm{d}s} = \alpha$$

Les numérateur et dénominateur de cette expression peuvent s'exprimer avec la fonction gamma incomplète $\Gamma(n+1;\nu) = \int_{\nu}^{\infty} x^n e^{-x}\mathrm{d}x$:

$$\frac{\Gamma(N+1;b) - \Gamma(N+1;s_{\text{up}}+b)}{\Gamma(N+1;b)} = \alpha$$

soit

$$\frac{\Gamma(N+1;s_{\text{up}}+b)}{\Gamma(N+1;b)} = 1-\alpha \tag{9.2}$$

De plus, nous savons que

$$\sum_{n=0}^{N} \frac{\nu^n}{n!} e^{-\nu} = \frac{\Gamma(N+1;\nu)}{\Gamma(N+1)}$$

Ainsi, 9.2 peut s'écrire

$$\frac{CL_{s+b}\left(s_{\mathrm{up}}\right)}{CL_b} = \alpha$$

Cette dernière expression correspond à la définition de la limite supérieure par la méthode CL_s. Ceci prouve, que, pour une expérience poissonnienne, la méthode CL_s et la méthode bayésienne avec une distribution *a priori* uniforme sont équivalentes.

Annexe A

Quelques distributions fréquentes

Cette annexe donne quelques propriétés des distributions rencontrées lors des chapitres précédents. Pour plus d'information, le lecteur peut consulter [9].

A.1 Uniforme

La densité de probabilité uniforme entre deux valeurs a et b est donnée par :

$$\begin{cases} f(x;a,b) = \dfrac{1}{b-a} & \text{pour } a \leq x \leq b \\ f(x;a,b) = 0 & \text{pour } x < a \text{ et } x > b \end{cases}$$

Nous vérifions sans difficultés que $\int f(x)\mathrm{d}x = 1$. Son espérance et son écart-type sont

- $\mathbb{E}[X] = \frac{a+b}{2}$
- $\sigma[X] = \sqrt{\int \frac{x^2}{b-a}\mathrm{d}x - \left(\frac{a+b}{2}\right)^2} = \sqrt{\frac{b^3-a^3+3a^2b-3b^2a}{12(b-a)}} = \sqrt{\frac{(b-a)^3}{12(b-a)}} = \frac{b-a}{\sqrt{12}}$

La distribution uniforme pour $a = -2$ et $b = 6$ est montrée sur la figure A.1.

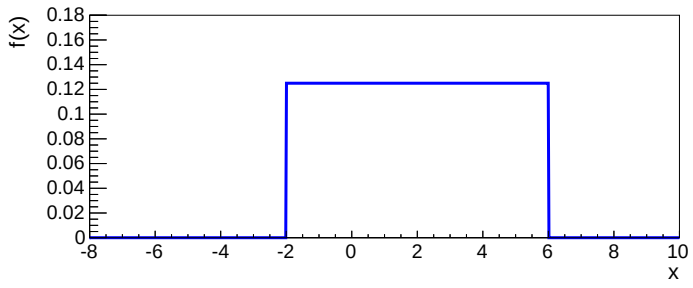


FIGURE A.1 – Distribution uniforme pour $a = -2$ et$b = 6$.

A.2 Gaussienne (Normale)

A.2.1 Cas unidimensionnel

La densité de probabilité gaussienne unidimensionnelle (ou loi normale) est donnée par :

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

Son espérance et son écart-type sont

- $\mathbb{E}[X] = \mu$
- $\sigma[X] = \sigma$

La distribution gaussienne pour $\mu = 5$ et différentes valeurs de σ est représentée sur la figure A.2.

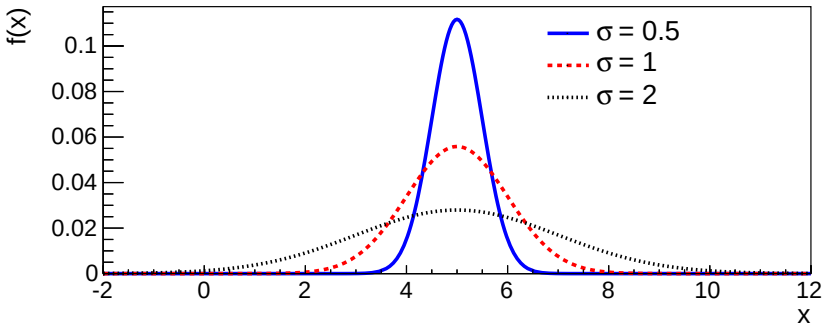


FIGURE A.2 – Distribution normale pour $\mu = 5$ et trois valeurs de σ.

La somme de deux variables aléatoires gaussiennes indépendantes est aussi gaussienne. L'espérance de la somme est égale à la somme des espérances et la variance de la somme est égale à la somme des variances.

A.2.2 Cas multidimensionnel

La densité de probabilité gaussienne multidimensionnelle est donnée par :

$$f(x;\mu,\Sigma) = \frac{1}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

où N est le nombre de variables et Σ est la matrice de covariance.

En guise d'exemple, considérons le cas de deux variables aléatoires. Soit μ_1, μ_2, σ_1 et σ_2 leurs moyennes et écarts-types et ρ leur fonction de corrélation. La matrice de covariance est

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

et la densité de probabilité conjointe

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)}$$

A.2.3 Théorème central limite

La loi normale tire son importance du théorème central limite. Suivant ce théorème, la somme de n variables aléatoires indépendantes suivant la même loi de probabilité est distribuée suivant une loi normale dans la limite où n tend vers l'infini. Notons S_n la somme de n variables X_i ($i \in [1, n]$) d'espérance μ et écart-type σ :

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

Nous avons :

$$\mathbb{E}[S_n] = n\mu \quad \text{et} \quad \text{var}[S_n] = n\sigma^2$$

Définissons les variables centrées réduites associées aux X_i et à S_n :

$$Y_i = \frac{X_i - \mu}{\sigma} \quad \text{et} \quad Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Nous voyons que :

$$Z_n = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$$

La fonction caractéristique de Z_n est :

$$\Phi_{Z_n}(t) = \mathbb{E}\left[e^{itZ_n}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{it\frac{Y_i}{\sqrt{n}}}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{it\frac{Y_i}{\sqrt{n}}}\right] = \prod_{i=1}^{n} \Phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left(\Phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

où nous avons utilisé le fait que les Y_i sont indépendants et identiquement distribués.

La fonction caractéristique d'une variable Y centrée réduite s'exprime, à l'ordre 2, comme ceci :

$$\Phi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$$

Nous avons finalement :

$$\Phi_{Z_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \xrightarrow[n\to\infty]{} e^{-\frac{t^2}{2}}$$

Ce qui correspond à la fonction caractéristique de la loi normale standard. Le type de convergence dont il s'agit ici est la convergence en loi, que nous notons $\xrightarrow{L}$ et que nous définirons en 4.4.1. Nous avons donc :

$$Z_n \xrightarrow{L} \mathcal{N}(0,1) \text{ et } S_n \xrightarrow{L} \mathcal{N}(n\mu, n\sigma^2)$$

A.3 Student

La densité de probabilité de Student est

$$f(t;n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}$$

C'est la densité de probabilité de la variable

$$t = \frac{\sqrt{n-1}\,(M - \mathbb{E}[X])}{s}$$

où $M = \frac{1}{n}\sum_i X_i$ est la moyenne empirique, $s^2 = \frac{\sum_i (X_i - M)^2}{n}$ est la variance empirique (biaisée) et les X_i sont des variables idépendantes et distribuées suivant une loi normale d'espérance $\mathbb{E}[X]$.

La loi de Student tend vers la loi normale lorsque n tend vers l'infini. Son espérance et sa variance sont

- $\mathbb{E}[T] = 0$ pour $n > 1$ (non définie pour $k \leq 1$)
- $\mathrm{var}[T] = \frac{k}{k-2}$ pour $k > 2$ (infinie pour $k \leq 2$)

Loi loi de Student est représentée sur la figure A.3 pour différentes valeurs de n.

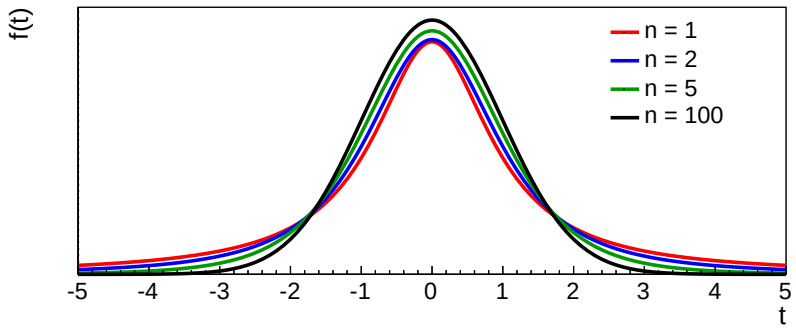


FIGURE A.3 – Distribution de Student pour différentes valeurs de n.

A.4 Lognormale

Une variable aléatoire X est distribuée suivant la distribution lognormale si son logarithme $Y = \ln X$ est distribué suivant une distribution normale :

$$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$$

La distribution lognormale est donc donnée par :

$$f_X(x;\mu,\sigma) = f_Y(y;\mu,\sigma)\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = \frac{1}{x\sqrt{2\pi}\sigma}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

La distribution lognormale n'est définie que sur $\mathbb{R}^+$. Son espérance et son écart-type sont

- $\mathbb{E}[X] = e^{\mu+\frac{\sigma^2}{2}}$
- $\sigma[X] = \sqrt{\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}}$

A.5 Binomiale

Soit B_i ($i \in [1,n]$) un ensemble de n variables aléatoires suivant toute la même loi de Bernouilli de paramètre p. Soit K la somme des B_i : $K = B_1 + B_2 + \ldots + B_n$. K est distribuée suivant la loi binomiale donnée par :

$$p(k;p,n) = \binom{n}{k}p^k(1-p)^{n-k}$$

où $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Son espérance et son écart-type sont

- $\mathbb{E}[K] = np$
- $\sigma[K] = \sqrt{np(1-p)}$

La distribution binomiale est représentée pour $n = 10$ et différentes valeurs de p sur la figure A.4.

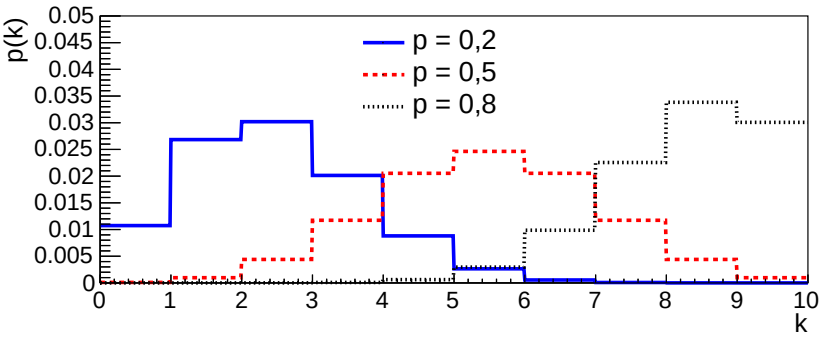


FIGURE A.4 – Distribution binomiale pour $n = 10$ et trois valeurs de p.

La loi binomiale peut, de manière équivalente, être définie comme la probabilité conjointe de deux variables aléatoires N_0 et N_1 correspondant aux nombres d'échec et de succès au cours de n de processus de Bernouilli ($N_0 + N_1 = n$ est fixe). Renommons $p_1 = p$ la probabilité de succès et $p_0 = 1 - p$ la probabilité d'échec. La loi binomiale peut être réécrite comme ceci :

$$p(n_0,n_1) = \frac{n!}{n_0!n_1!}p_0^{n_0}p_1^{n_1}$$

La loi binomiale tire son nom du lien qu'elle entretient avec la formule du binôme de Newton. Rappelons en effet que

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Nous voyons que la probabilité $p(k;p,n)$ a la même structure que les termes dans la formule du binôme. L'application de la formule du binôme à $p(k;p,n)$ montre immédiatement que la condition de normalisation est bien vérifiée

$$\sum_{k=0}^{\infty} p(k;p,n) = 1$$

A.6 Multinomiale

La loi multinomiale est une généralisation de la loi binomiale au cas où le nombre de résultats m est supérieur à deux :

$$p(n_1,\ldots,n_m) = \frac{n!}{\prod_{i=1}^{m} n_i!} \prod_{i=1}^{m} p_i^{n_i}$$

avec $\sum_{i=1}^{m} p_i = 1$ et $\sum_{i=1}^{m} N_i = n$ (fixe). Chaque variable N_i est une variable binomiale d'espérance et écart-type :

$$\mathbb{E}[N_i] = np_i \quad \text{et} \quad \sigma[N_i] = \sqrt{np_i(1-p_i)}$$

et la covariance est :

$$\operatorname{cov}(N_i, N_j) = -np_ip_j$$

La formule de l'espérance se démontre comme suit (les formules de l'écart-type et de la covariance sont similaires).

$$\begin{aligned}
\mathbb{E}[N_i] &= \sum_{n_0=0}^{\infty} \cdots \sum_{n_m=0}^{\infty} n_i p(n_0,n_1,\ldots,n_m) = \sum_{n_0} \cdots \sum_{n_m} n_i \frac{n!}{\prod_{j=1}^{m} n_j!} \prod_{j=1}^{m} p_j^{n_j} \\
&= \sum_{n_i} \frac{n_i n!}{n_i!} p_i^{n_i} \sum_{n_0} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_m} \frac{\prod_{j=1,j\neq i}^{m} p_j^{n_j}}{\prod_{j=1,j\neq i}^{m} n_j!} \\
&= \sum_{n_i} \frac{n_i n!}{(n-n_i)!n_i!} p_i^{n_i} \underbrace{\sum_{n_0} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_m} \frac{(n-n_i)!}{\prod_{j=1,j\neq i}^{m} n_j!} \prod_{j=1,j\neq i}^{m} p_j^{n_j}}_{\left(\sum_{j=1,j\neq i}^{m} p_j\right)^{n-n_i} = (1-p_i)^{n-n_i}} \\
&= \sum_{n_i} n_i \frac{n!}{(n-n_i)!n_i!} p_i^{n_i} (1-p_i)^{n-n_i}
\end{aligned}$$

où nous avons utilisé la formule du multinôme de Newton

$$(x_1 + \ldots + x_m)^n = \sum_{k_1,\ldots,k_m;k_1+\ldots+k_m=n} \binom{n}{k_1,\ldots,k_m} x_1^{k_1} \ldots x_m^{k_m}$$

Nous retrouvons ainsi l'espérance de la loi binomiale de paramètres (p_i, n) qui comme nous le savons est égale à np_i.

A.7 Binomiale négative

La loi binomiale négative est la loi qui régit le nombre d'expériences de Bernouilli à réaliser avant d'obtenir un nombre donné de succès. Soit k le nombre de succès, p le paramètre de la loi de Bernouilli et n le nombre d'expériences de Bernouilli ($k, n \in \mathbb{N}$ et $0 \leq p \leq 1$). Nous avons

$$p(n; k, p) = \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Son espérance et sa variance sont

- $\mathbb{E}[n] = \frac{k}{p}$
- $\text{var}[n] = \frac{k(1-p)}{p^2}$

A.8 Poisson

Soit $N \in \mathbb{N}$ une variable aléatoire discrète. N suit une loi de Poisson de paramètre μ si la probabilité d'obtenir n lors d'un processus aléatoire est

$$p(n; \mu) = \frac{\mu^n}{n!} e^{-\mu}$$

La loi de Poisson est représentée pour trois valeurs de μ sur la figure A.5.

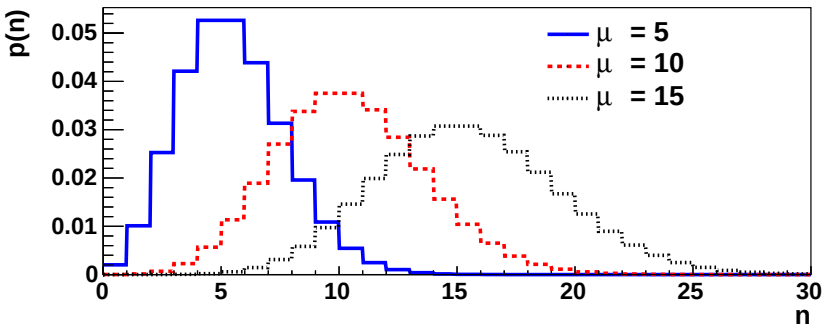


FIGURE A.5 – Distribution de Poisson pour μ =5, 10 et 15.

Son espérance et sa variance sont

- $\mathbb{E}[N] = \mu$. En effet,

$$\mathbb{E}[N] = \sum_{n=0}^{\infty} np(n;\mu) = e^{-\mu}\sum_{n=0}^{\infty}\frac{n\mu^n}{n!} = e^{-\mu}\sum_{n=1}^{\infty}\frac{n\mu^n}{n!} = \mu e^{-\mu}\sum_{n=1}^{\infty}\frac{\mu^{n-1}}{(n-1)!} = \mu e^{-\mu}\underbrace{\sum_{n=0}^{\infty}\frac{\mu^n}{n!}}_{e^{\mu}}$$

- $\mathrm{var}[N] = \mu^2$. En effet,

$$\mathrm{var}[N] = \mathbb{E}\left[N^2\right] - \mathbb{E}[N]^2 = e^{-\mu}\sum_{n=0}^{\infty}\frac{n^2\mu^n}{n!} - \mu^2$$

En utilisant $n^2 = n + n(n-1)$, nous trouvons que la somme dans cette dernière expression est

$$\sum_{n=0}^{\infty}\frac{n\mu^n}{n!} + \sum_{n=0}^{\infty}\frac{n(n-1)\mu^n}{n!} = \mu e^{\mu} + \mu^2\sum_{n=2}^{\infty}\frac{\mu^{n-2}}{(n-2)!} = \mu e^{\mu} + \mu^2\sum_{n=0}^{\infty}\frac{\mu^n}{n!} = \mu e^{\mu} + \mu^2 e^{\mu}$$

A.8.1 Limite de la loi binomiale

La loi de Poisson est un cas limite de la loi binomiale lorsque $n \to \infty$, $p \to 0$ et lorsque l'espérance np reste fini. Dans la suite nous noterons $\lambda = np$. Puisque n tend vers l'infini, nous pouvons utiliser la formule de Stirling :

$$n! \simeq \left(\frac{n}{e}\right)^n\sqrt{2\pi n} \quad \text{et} \quad (n-k)! \simeq \left(\frac{n-k}{e}\right)^{n-k}\sqrt{2\pi(n-k)}$$

La loi binomiale s'écrit donc :

$$p(k;p,n) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \simeq \frac{n^n e^{-k}}{k!(n-k)^{n-k}}\sqrt{\frac{n}{n-k}}p^k(1-p)^{n-k} \simeq \frac{n^n e^{-k}}{k!(n-k)^{n-k}}p^k(1-p)^{n-k}$$

En introduisant $\lambda = pn$,

$$\begin{aligned} p(k;p,n) &\simeq \frac{n^n e^{-k}}{k!(n-k)^{n-k}}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &\simeq \frac{n^n e^{-k}}{k!(n-k)^n(n-k)^{-k}}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^n \\ &\simeq \frac{e^{-k}}{k!(1-\frac{k}{n})^n(1-\frac{k}{n})^{-k}}\lambda^k e^{-\lambda} \\ &\simeq \frac{e^{-k}}{k!e^{-k}}\lambda^k e^{-\lambda} \simeq \frac{\lambda^k}{k!}e^{-\lambda} \end{aligned}$$

Notons que la probabilité ne dépend maintenant que d'un seul paramètre λ : $p(k;p,n) = p(k;\lambda)$

A.8.2 Limite gaussienne

Lorsque μ devient très grand, la loi de Poisson tend vers la loi normale de moyenne λ et écart-type $\sqrt{\lambda}$. En effet, les fluctuations autour de la moyenne deviennent petites devant cette dernière et nous pouvons écrire

$$n = \mu(1+\varepsilon)$$

où $\varepsilon \ll 1$. En utilisant la formule de Stirling

$$\begin{aligned} p(n;\mu) &\simeq \frac{\mu^{\mu(1+\varepsilon)} e^{-\mu}}{\sqrt{2\pi\mu(1+\varepsilon)} \times (\mu(1+\varepsilon))^{\mu(1+\varepsilon)} \times e^{-\mu(1+\varepsilon)}} \\ &\simeq \frac{e^{-\mu}}{\sqrt{2\pi\mu} \times (1+\varepsilon)^{\mu(1+\varepsilon)+1/2} \times e^{-\mu(1+\varepsilon)}} \end{aligned}$$

La limite de $(1+\varepsilon)^{\mu(1+\varepsilon)+1/2}$ s'obtient en prenant le logarithme :

$$(\mu(1+\varepsilon)+1/2)\ln(1+\varepsilon) \xrightarrow[\varepsilon\to 0]{} (\mu(1+\varepsilon)+1/2)\left(\varepsilon - \frac{\varepsilon^2}{2}\right) = \left(\mu\left(\varepsilon + \frac{\varepsilon^2}{2}\right) - \frac{1}{2}\left(\varepsilon - \frac{\varepsilon^2}{2}\right)\right) \simeq \mu\left(\varepsilon + \frac{\varepsilon^2}{2}\right)$$

Donc

$$p(n;\mu) \simeq \frac{e^{\mu\varepsilon} e^{-\mu\left(\varepsilon+\frac{\varepsilon^2}{2}\right)}}{\sqrt{2\pi\mu}} = \frac{e^{-\mu\frac{\varepsilon^2}{2}}}{\sqrt{2\pi\mu}} = \frac{1}{\sqrt{2\pi\mu}} e^{-\frac{(n-\mu)^2}{2\mu}}$$

Il s'agit bien d'une gaussienne de moyenne μ et écart-type $\sqrt{\mu}$.

A.9 Gamma

La densité de probabilité gamma est

$$f(x;a,b) = \frac{a(ax)^{b-1}e^{-ax}}{\Gamma(b)} \tag{A.1}$$

où $\Gamma(b) = \int x^{b-1}e^{-x}\mathrm{d}x$ est la fonction gamma. b est le paramètre de forme et a le paramètre d'échelle. Lorsque b est un entier positif ($\Gamma(b) = (b-1)!$), on la nomme parfois distribution d'Erlang.

Son espérance et sa variance sont

- $\mathbb{E}[X] = b/a$
- $\mathrm{var}[X] = b/a^2$

Remarque : lorsque $a = 1$, la distribution d'Erlang

$$f(x;a=1,b) = \frac{x^{b-1}}{(b-1)!}e^{-x}$$

a la même forme que la loi de Poisson (mais les rôles de variable et paramètre sont inversés)

La distribution gamma est représentée pour différentes valeurs de a et b sur la figure A.6.

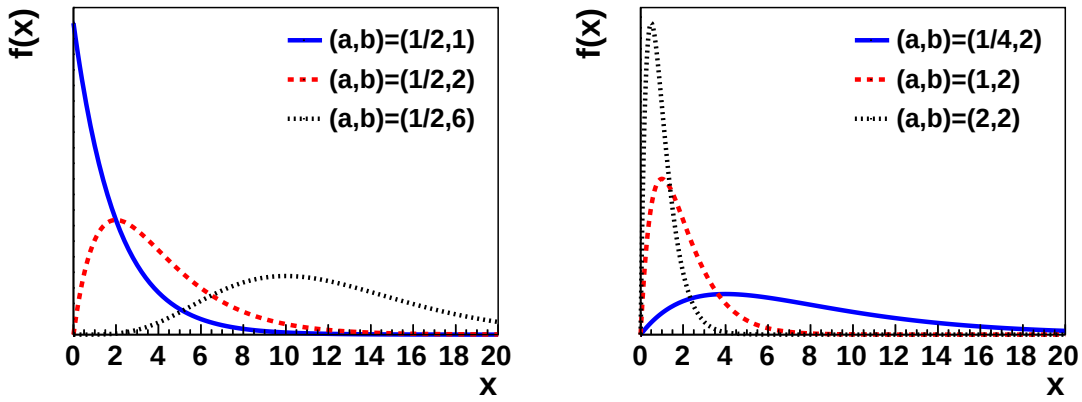


FIGURE A.6 – Distribution gamma pour différentes valeurs de a et b.

A.10 Bêta

La densité de probabilité bêta, définie entre 0 et 1, est

$$f(x;a,b) = \frac{x^{a-1}\,(1-x)^{b-1}}{B(a,b)}$$

où $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ est la fonction beta d'Euler. Son espérance et sa variance sont

- $\mathbb{E}[X] = \frac{a}{a+b}$
- $\mathrm{var}[X] = \frac{ab}{(a+b)^2(a+b+1)}$

La distribution bêta est représentée pour différentes valeurs de a et b sur la figure A.7.

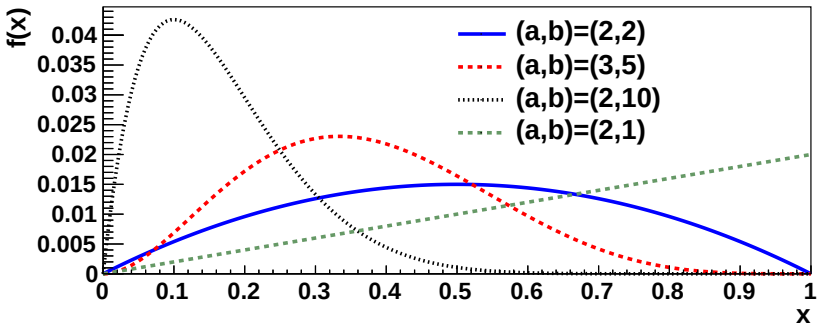


FIGURE A.7 – Distribution bêta pour différentes valeurs de a et b.

A.11 Khi carré

La distribution du khi carré est

$$f(x;n) = \frac{\left(\frac{x}{2}\right)^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2\Gamma\left(\frac{n}{2}\right)}$$

où n est un paramètre entier appelé nombre de degrés de libertés. Son espérance et sa variance sont

- $\mathbb{E}[X] = n$
- $\text{var}[X] = 2n$

La distribution du Khi carré est un cas particulier de distribution gamma (voir équation A.1) pour $a = 1/2$ et $b = n/2$. Elle est représentée sur la figure A.8 pour différentes valeurs de n.

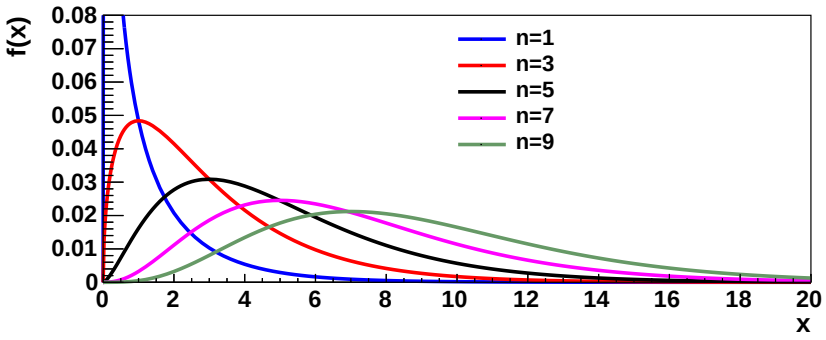


FIGURE A.8 – Distribution du khi carré pour différentes valeurs de n.

Annexe B

Intervalle de confiance pour une proportion binomiale

Dans cette annexe nous décrivons quelques méthodes populaires pour estimer un intervalle de confiance pour une proportion binomiale. Pour plus de détails, le lecteur pourra consulter [10, 11]. Nous utiliserons les notations suivantes :

- p : proportion à estimer
- N : nombre total d'événements
- k : nombre d'événements après sélection ($k \leq N$)
- $\hat{p} = k/N$: proportion observée (estimateur ML de p)
- $\hat{q} = 1 - \hat{p}$
- $z_\alpha = \Phi^{-1}\left((1+\alpha)/2\right)$: nombre de déviations standards correspondant au niveau de confiance α (Φ^{-1} est l'inverse de la fonction de repartition de la loi normale centrée réduite)

La loi qui régit k est la loi binomiale :

$$p(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k}$$

Nous rappelons que, pour cette loi,

$$\mathbb{E}[k] = Np$$

$$\text{var}[k] = Np(1-p)$$

B.1 Intervalle standard (ou intervalle de Wald)

L'intervalle standard CI_s, au niveau de confiance α, est

$$CI_s = \hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}\hat{q}}{N}}$$

Démonstration 1 : le centre de l'intervalle est l'estimateur de la proportion $\hat{p} = k/N$ et sa largeur est évaluée de la manière suivante :

$$\sigma_{\hat{p}} = \frac{\partial \hat{p}}{\partial k}\sigma_k = \frac{\sigma_k}{N} = \frac{\sqrt{Np(1-p)}}{N} = \sqrt{\frac{p(1-p)}{N}}$$

Hypothèse 1 : nous supposons que k, et donc $\hat{p}$, est gaussien (i.e. nous faisons l'approximation du théorème central limite). Nous pouvons donc utiliser le quantile de la loi normale z_α

$$CI_s = \hat{p} \pm z_\alpha \sqrt{\frac{p(1-p)}{N}}$$

Hypothèse 2 : nous supposons que N est suffisamment grand pour que $\hat{p} \simeq p$ (i.e. var $[\hat{p}]$ négligeable). Donc

$$CI_s = \hat{p} \pm z_\alpha \sqrt{\frac{\hat{p}\hat{q}}{N}}$$

Démonstration 2 : c'est la même démonstration que ci-dessus, mais formulée différemment. Ici, l'intervalle est obtenu en inversant la région d'acceptance du test de Wald dans la limite des grands échantillon :

$$\left| \frac{\hat{p} - p}{\hat{se}\,(\hat{p})} \right| \leq z_\alpha$$

où $\hat{p}$ est l'estimateur ML de p et $\hat{se}\,(\hat{p})$ est son écart-type estimé. Dans le cas binomiale, nous avons $\hat{p} = k/N$ et $\hat{se}\,(\hat{p}) = \sqrt{\hat{p}\hat{q}/N}$. La formulation en terme de test de Wald conduit donc au même résultat que la première démonstration.

Commentaires : cet intervalle est le pire possible : (a) la couverture peut être assez différente du niveau de confiance α (ceci est dû au fait que la distribution de $\hat{p}$ est discrète, asymétrique et que N n'est pas suffisamment grand pour dire que $\hat{p} \simeq p$), (b) la largeur de l'intervalle tend vers 0 quand $\hat{p}$ est proche de 0 ou 1, ce qui n'est clairement pas physique. Les autres intervalles décrits ci-dessous corrigent au moins en partie ces défauts.

B.2 Intervalle de Wilson

L'intervalle de Wilson CI_W, au niveau de confiance α, est donné par :

$$CI_W = \frac{\hat{p} + z_\alpha^2/(2N)}{1 + z_\alpha^2/N} \pm \frac{z_\alpha \sqrt{N}}{N + z_\alpha^2} \sqrt{\hat{p}\hat{q} + \frac{z_\alpha^2}{4N}}$$

Démonstration : la démonstration est similaire à la seconde démonstration de l'intervalle standard (voir ci-dessus) mais plutôt que d'utiliser l'estimation de l'écart-type la vrai valeur est utilisée

$$\left| \frac{\hat{p} - p}{se\,(\hat{p})} \right| \leq z_\alpha$$

L'écart-type est $se\,(\hat{p}) = \sqrt{p(1-p)/N}$. L'intervalle est donc

$$\left|\frac{\hat{p}-p}{\sqrt{p(1-p)/N}}\right| \leq z_\alpha$$

$$\Leftrightarrow (\hat{p}-p)^2 \leq z_\alpha^2 \frac{p(1-p)}{N}$$

$$\Leftrightarrow N(\hat{p}^2+p^2-2p\hat{p}) - z_\alpha^2 p + z_\alpha^2 p^2 \leq 0$$

$$\Leftrightarrow (N+z_\alpha^2)p^2 - (2N\hat{p}+z_\alpha^2)p + N\hat{p}^2 \leq 0$$

C'est une équation quadratique dont les solutions sont

$$p_\pm = \frac{\hat{p}+z_\alpha^2/(2N)}{1+z_\alpha^2/N} \pm \frac{z_\alpha\sqrt{N}}{N+z_\alpha^2}\sqrt{\hat{p}\hat{q}+\frac{z_\alpha^2}{4N}}$$

L'intervalle de Wilson est $CI_W = [p_-, p_+]$.

Commentaires : Le centre de cet intervalle n'est pas l'estimateur ML. Cependant, la couverture est bien meilleure que dans le cas de l'intervalle standard.

B.3 Intervalle d'Agresti-Coull

L'intervalle d'Agresti-Coull CI_{AC}, au niveau de confiance α, est

$$CI_{AC} = \tilde{p} \pm z_\alpha \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{N}}}$$

avec

- $\tilde{k} = k + z_\alpha^2/2$
- $\tilde{N} = N + z_\alpha^2$
- $\tilde{p} = \tilde{k}/\tilde{N}$ et $\tilde{q} = 1-\tilde{p}$

Démonstration : Cet intervalle n'est pas le résultat d'une démonstration comme CI_s et CI_W mais il se trouve qu'il a des bonnes propriétés et qu'il se comporte bien mieux que l'intervalle standard. L'idée est de construire un intervalle qui a la même forme que l'intervalle standard, mais avec une couverture plus proche de α. Ceci est réalisable en définissant $\tilde{k}$, $\tilde{N}$, $\tilde{p}$ et $\tilde{q}$ (voir ci-dessus pour leurs définitions) tel que le centre de l'intervalle coïncide avec le centre de l'intervalle de Wilson. $\tilde{p}$ est le centre de l'intervalle de Wilson et est donc aussi utilisé comme centre dans l'intervalle d'Agresti-Coull. L'intervalle d'Agresti-Coull, $\tilde{p} \pm z_\alpha\sqrt{\tilde{p}\tilde{q}/\tilde{N}}$, est, de manière détaillée,

$$CI_{AC} = \frac{\hat{p}+z_\alpha^2/(2N)}{1+z_\alpha^2/N} \pm \frac{z_\alpha}{\sqrt{N+z_\alpha^2}}\sqrt{\frac{k+z_\alpha^2/2}{N+z_\alpha^2}\left(1-\frac{k+z_\alpha^2/2}{N+z_\alpha^2}\right)}$$

Commentaires : La couverture de l'intervalle d'Agresti-Coull est toujours plus grande que celle de l'intervalle de Wilson.

B.4 Intervalle de Jeffrey

L'intervalle de Jeffrey CI_J, au niveau de confiance α, est

$$CI_J = [L(k), U(k)]$$

avec $L(0) = 0$, $U(N) = 0$ et

- $L(k) = b_{(1-\alpha)/2}(k + 1/2, N - k + 1/2)$
- $U(k) = b_{(1+\alpha)/2}(k + 1/2, N - k + 1/2)$

où $b_{(1-\alpha)/2}(a, b) = F^{-1}((1 + \alpha)/2; a, b)$ est le $100((1 + \alpha)/2)^{\text{ème}}$ percentile de la distribution bêta $Be(x; a, b)$ ($F^{-1}(x; a, b)$ est l'inverse de la fonction de répartition de la distribution bêta).

Démonstration : Cet intervalle est dérivé suivant une approche bayésienne. Soit $p(k|N, p) = \binom{N}{k} p^k (1-p)^{N-k}$. La distribution *a posteriori* de p est

$$p(p|N, k) = \frac{p(k|N, p)p(p|N)}{\mathcal{Z}}$$

où $\mathcal{Z}$ est une constante de normalisation et $p(p|N)$ est la distribution *a priori*. Cette dernière est choisie dans la famille des distributions bêtas (voir section A.10) avec $a = b = 1/2$:

$$p(p|N) = \frac{p^{1/2}(1-p)^{1/2}}{B(1/2, 1/2)}$$

La distribution *a posteriori* est donc une distribution bêta de paramètres $k + 1/2$ and $N - k + 1/2$

$$P(p|N, k) = \frac{p^{k+1/2}(1-p)^{N-k+1/2}}{B(k + 1/2, N - k + 1/2)}$$

L'intervalle est formé en prenant l'intervalle centré (voir section 9.1.2).

Commentaires : L'intervalle de Jeffrey a une bonne couverture (sauf lorsque p est proche de 0 ou 1). Ses bornes inférieures et supérieures ne sont pas calculables analytiquement. Il faut avoir recours à des méthodes numériques. Des formules approchées peuvent être utilisées si une grande précision n'est pas nécessaire [10]. De plus, il est important de remarquer que ni l'espérance ni le mode de la distribution *a posteriori* ne correspondent à l'estimateur ML $\hat{p} = k/N$. En effet, l'espérance et le mode de la distribution beta de paramètres a et b sont

$$\mathbb{E}[X] = \frac{a}{a + b}$$

$$\text{mod}[X] = \frac{a - 1}{a + b - 2}$$

Dans notre cas $a = k + 1/2$, $b = N - k + 1/2$. Donc

$$\mathbb{E}[p] = \frac{k + 1/2}{N + 1}$$

$$\text{mod}[p] = \frac{k - 1/2}{N - 1}$$

B.5 Comparaison des couvertures

Les figures B.1 et B.2 montrent les couvertures en fonction de N et p pour les différentes intervalles décrits ci-dessus.

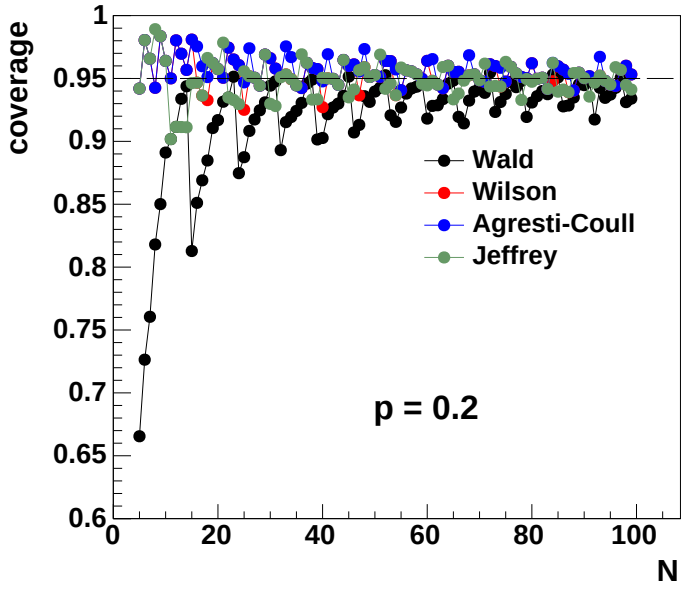


FIGURE B.1 – Couverture en fonction de N pour $p = 0,2$ pour les intervalles de Wald, Wilson, Agresti-Coull et Jeffrey.

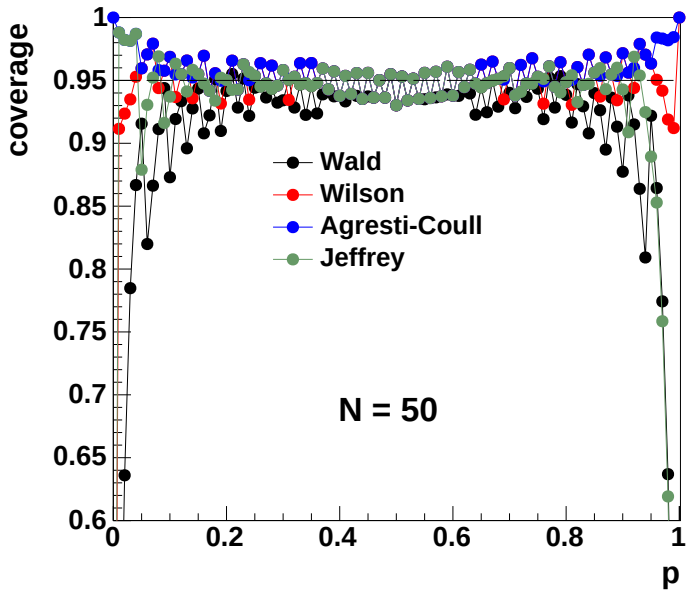


FIGURE B.2 – Couverture en fonction de p pour $N = 50$ pour les intervalles de Wald, Wilson, Agresti-Coull et Jeffrey.

Bibliographie

[1] Alan Stuart et al. Kendall's advanced theory of statistics ; 6th ed. of the 3-vol. ed.. Chichester : Wiley, 1994.

[2] William T Eadie et al. Statistical methods in experimental physics. Amsterdam : North-Holland, 1971.

[3] Glen D Cowan. Statistical data analysis. Oxford : Oxford Univ. Press, 1998.

[4] Louis Lyons. Statistics for Nuclear and Particle Physicists. Cambridge : Cambridge Univ. Press, 1986.

[5] A. N. Kolmogorov. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin, 1933.

[6] Andrey N. Kolmogorov. Foundations of the Theory of Probability. 2e éd. Chelsea Pub Co, juin 1960. URL : http://www.clrc.rhul.ac.uk/resources/fop/Theory%20of%20Probability%20(small).pdf.

[7] Abraham Wald. "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large". Dans : Transactions of the American Mathematical Society 54.3 (nov. 1943), p. 426–482. ISSN : 00029947. DOI : 10.2307/1990256. URL : http://dx.doi.org/10.2307/1990256.

[8] G. J. Feldman et R. D. Cousins. "Unified approach to the classical statistical analysis of small signals". Dans : Phys.Rev.D 57 (avr. 1998), p. 3873–3889. DOI : 10.1103/PhysRevD.57.3873. eprint : physics/9711021.

[9] Christian Walck. Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists. Déc. 1996.

[10] Lawrence D. Brown, T. Tony Cai et Anirban DasGupta. "Interval Estimation for a Binomial Proportion". Dans : Statistical Science 16.2 (mai 2001), p. 101–133. DOI : 10.1214/ss/1009213286. URL : http://dx.doi.org/10.1214/ss/1009213286.

[11] D. Casadei. "Estimating the selection efficiency". Dans : Journal of Instrumentation 7 (août 2012), p. 8021. DOI : 10.1088/1748-0221/7/08/P08021. eprint : 0908.0130.