Polytech'Lille
Département Informatique et Statistique

Statistique Inférentielle

(Notes de Cours)

Sergueï Dachian

version 2019/2020

Plan du cours

Ce cours démarre par des rappels (et des compléments) de la théorie des probabilités et de la statistique descriptive, avant de passer à la statistique inférentielle à proprement parler.

Dans le premier chapitre "*Rappels et compléments de la théorie des probabilités*", page 2, en plus de revoir les notions de base de la théorie des probabilités, on introduit les quantiles et les intervalles de fluctuation, et on rappelle les convergences de variables aléatoires et les théorèmes limites.

Dans le deuxième chapitre "*Rappels et compléments de la statistique descriptive*", page 16, on revoit les notions de base de la statistique descriptive. Rappelons ici qu'en statistique descriptive on cherche à résumer (synthétiser) ou représenter graphiquement une (voire plusieurs) variable (dite également caractère) observée sur une population entière.

Dans le troisième chapitre "*Statistique inférentielle — généralités*", page 30, on pose le cadre de la statistique inférentielle. En statistique inférentielle, la variable n'est observée que sur une "petite" partie (tirée au hasard) d'une "grande" population, et on cherche à inférer (déduire) des informations sur le comportement global de cette variable à partir de ces observations. La clé pour le faire est la théorie des probabilités : les observations forment une suite de variables aléatoires i.i.d. d'une loi inconnue (totalement ou partiellement).

Le quatrième chapitre "*Estimation et intervalles de confiance*", page 39, est consacré à la théorie de l'estimation statistique. Ici on cherche à estimer (approcher) la loi inconnue des observations (ou les paramètres dont elle dépend).

Finalement, le cinquième et dernier chapitre "*Tests d'hypothèses*", page 62, est consacré à la théorie des tests statistiques (dits également tests d'hypothèses). Ici on cherche à confirmer ou infirmer une hypothèse (une affirmation) concernant la loi inconnue des observations (ou les paramètres dont elle dépend).

Pour celles et ceux qui voudrons approfondir leurs connaissances en statistique inférentielle, je conseille le livre de G. Saporta [2] et, pour des aspects plus théoriques, celui de A. Borovkov [1], ainsi que le polycopié de cours de J. Jacques [3], dans lequel on trouve notamment une très grande collection de tests d'hypothèses.

1 Rappels et compléments de la théorie des probabilités

1.1 Notions et objets de base

Dans cette section, nous passons rapidement en revue les principaux notions et objets de la théorie des probabilités et fixons les notations qui seront utilisées dans la suite.

Espace probabilisé

L'objet fondamental de la théorie des probabilités est l'**espace probabilisé** $(\Omega, \mathscr{F}, \mathbf{P})$, où Ω est l'**ensemble des issus de l'expérience** (dit également **univers**), $\mathscr{F}$ est la tribu des **événements** observables, et $\mathbf{P}$ est la (**mesure de**) **probabilité** $\big($sur $(\Omega, \mathscr{F})\big)$, qui à chaque événement $A \in \mathscr{F}$ fait correspondre sa probabilité $\mathbf{P}(A)$.

Variable aléatoire et sa loi

Un des objets principaux d'intérêt de la théorie des probabilités est la **variable aléatoire** (**v.a.**), définie comme une application ($\mathscr{F}/\operatorname{Bor}(\mathbb{R})$ mesurable) $X : \Omega \longrightarrow \mathbb{R}$.

La **loi** ou **distribution** (**de probabilité**) d'une variable aléatoire X est la mesure de probabilité $\mathbf{P}_X$ sur $\big(\mathbb{R}, \operatorname{Bor}(\mathbb{R})\big)$ définie, pour tout $A \in \operatorname{Bor}(\mathbb{R})$, par

$$\mathbf{P}_X(A) = \mathbf{P}(X \in A) = \mathbf{P}\big(\{\omega \,:\, X(\omega) \in A\}\big).$$

On notera $\mathcal{L}_X = \mathbf{P}_X$ la loi de X, et on écrira $X \backsim \mathcal{L}$ pour dire que X suit une loi $\mathcal{L}$ donnée.

Notons que la plupart des notions introduites ci-après pour des variables aléatoires pourra également être utilisée pour leurs lois (par exemple, on pourra parler d'une fonction de répartitions d'une variable aléatoire, comme d'une fonction de répartition d'une loi).

Fonction de répartition d'une variable aléatoire

La loi d'une variable aléatoire X est entièrement décrite par sa **fonction de répartition** (**FdR**) $F = F_X$ définie, pour tout $x \in \mathbb{R}$, par

$$F(t) = \mathbf{P}(X \leqslant t) = \mathbf{P}_X\big(]-\infty, t]\big).$$

Notons que les trois propriétés suivantes sont caractéristiques d'une fonction de répartition (c'est-à-dire toute FdR les possède, et toute fonction F les satisfaisant est une FdR) :

- F est croissante $\big(s \geqslant t$ implique $F(s) \geqslant F(t)\big)$;
- $F(-\infty) = \lim\limits_{t\to-\infty} F(t) = 0$ et $F(+\infty) = \lim\limits_{t\to+\infty} F(t) = 1$;
- F est une fonction **càdlàg**, c'est à dire continue à droite et admettant une limite à gauche en tout point de $\mathbb{R}$.

Variables aléatoires continues

Une variable aléatoire X est dite **continue** (ou **à densité**) si sa fonction de répartition F peut être représentée comme

$$F(t) = \int_{-\infty}^{t} f(x)\,\mathrm{d}x,$$

où $f = f_X$ est une fonction positive ($f(x) \geqslant 0$ pour tout $x \in \mathbb{R}$) dite **densité** de X.

Pour des exemples de lois continues usuelles, nous renvoyons à la brochure **Lois de Probabilité et Tables Statistiques**.

Remarques.

1. Dans ce cas, pour tout $x \in \mathbb{R}$, on a

$$\mathbf{P}(X = x) = F(x) - F(x-) = F(x) - \lim_{t \to x-} F(t) = 0.$$

   On dit que la variable aléatoire X est **diffuse**.

2. On a

$$\mathbf{P}\big(X \in (a,b)\big) = \int_{a}^{b} f(x)\,\mathrm{d}x.$$

   Ici et dans la suite, (a,b) dénote un intervalle ouvert ou fermé (de chaque côté), fini ou infini (c'est-à-dire on peut avoir $a = -\infty$ et/ou $b = +\infty$).

3. La densité f est défini **à presque partout près**. Cela signifie, *grosso modo*, que si on modifie les valeurs de f en un nombre fini ou dénombrable de points, la nouvelle fonction obtenue est aussi une densité de X (en effet, cela ne changera pas les intégrales, et donc pas non plus la fonction de répartition). Par exemple, pour la loi uniforme $\mathscr{U}(a,b)$, on peut prendre comme densité une des fonctions $f(x) = \frac{1}{b-a}\,\mathbb{1}_{[a,b]}(x)$ et $f(x) = \frac{1}{b-a}\,\mathbb{1}_{]a,b[}(x)$ (représentées respectivement dans les Figures 1.1 (a) et (b) ci-dessous), mais aussi $f(x) = \frac{1}{b-a}\,\mathbb{1}_{[a,b[}(x)$, $f(x) = \frac{1}{b-a}\,\mathbb{1}_{]a,b]}(x)$, ou même chacune des fonctions représentées dans les Figures 1.1 (c) et (d).

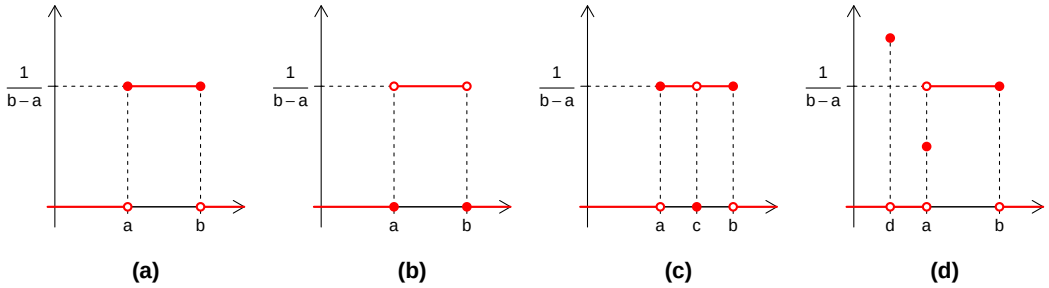

FIGURE 1.1 – Différentes densités de la loi $\mathscr{U}(a,b)$

Le bon sens est d'utiliser les versions "les plus continues possibles" (continues au moins à droite où à gauche en tout point où c'est possible) de la densité. Par exemple, pour la loi $\mathscr{U}(a,b)$, on va éviter les versions représentées dans les Figures 1.1 (c) et (d). En effet, elles admettent des discontinuités "évitables" respectivement aux points c et d. De plus, la version de la Figure 1.1 (d), n'est continue ni à droite ni à gauche au point a, ce qui est également évitable.

4. On a $f(x) = F'(x)$ pour **presque tout** $x \in \mathbb{R}$ (*grosso modo*, sauf un nombre fini ou dénombrable de points).

Notons finalement que les deux propriétés suivantes sont caractéristiques d'une densité (c'est-à-dire toute densité les possède, et toute fonction f les satisfaisant est une densité) :

— f est positive ($f(x) \geqslant 0$ pour tout $x \in \mathbb{R}$) ;

— $\displaystyle\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1.$

Variables aléatoires discrètes

Une variable aléatoire X est dite **discrète** si sa fonction de répartition F est une fonction en escalier. Comme on a $\mathbf{P}(X = x) = F(x) - F(x-)$, les abscisses des sauts de la fonction F donnent les valeurs possibles de X (il y en a un nombre fini ou dénombrable), et les hauteurs des sauts donnent les probabilités de ces valeurs.

On définit dans le cas discret une fonction $f = f_X$ par $f(x) = \mathbf{P}(X = x)$ pour tout $x \in \mathbb{R}$, et on l'appelle "**densité**" **discrète.** Ceci peut être vu comme une simple convention qui nous permettra de traiter les cas continue et discret simultanément, même si, en réalité, cela a un sens plus profond (que l'on ne détaillera ici) dans le cadre de la théorie de la mesure.

Pour des exemples de lois discrètes usuelles, nous renvoyons à la brochure **Lois de Probabilité et Tables Statistiques**.

Notons que les variables aléatoires discrètes et continues sont des types particuliers très importants de variables aléatoires, mais il existe également des variables aléatoires qui sont ni continues, ni discrètes.

Support d'une variable aléatoire

Le **support** d'une variable aléatoire X est défini comme

$$\mathscr{S}(X) = \big\{x \in \mathbb{R} \,:\, f(x) > 0\big\},$$

où f_X est la densité usuelle (si X est une v.a. continue) ou la "densité" discrète (si X est une v.a. discrète) de X. Notons que dans le cas discret, $\mathscr{S}(X)$ est l'ensemble (fini ou dénombrable) des valeurs possibles de X, et que dans le cas continue, $\mathscr{S}(X)$ n'est pas unique

(puisque f_X ne l'est pas). Cependant, même dans ce cas, $\mathscr{S}(X)$ peut être interprété comme l'ensemble des valeurs possibles de X (rappelons que les variables aléatoires continues sont diffuses, et on peut donc *grosso modo* ajouter ou enlever un nombre fini ou dénombrables de valeurs à l'ensemble des valeurs possibles).

Notons finalement que pour les variables aléatoires discrètes, $\mathscr{S}(X)$ est souvent un sous-ensemble (fini ou dénombrable) de $\mathbb{N}$, et que pour les variables aléatoires continues, en choisissant une des versions "les plus continues possibles" de la densité, $\mathscr{S}(X)$ est la plupart du temps un intervalle $(a, b) \subset \mathbb{R}$. Notons également que dans ce dernier cas, on utilise parfois la convention de prendre toujours un intervalle fermé (et choisir donc la version de la densité en conséquence).

Espérance d'une variable aléatoire

L'**espérance** (dite également **moyenne**) d'une variable aléatoire X représente la valeur que X prends "en moyenne". Elle est définie par

$$\mathbf{E}\, X = \mathbf{E}(X) = \int_{\mathscr{S}(X)} x\, f_X(x)\, \mathrm{d}x = \begin{cases} \displaystyle\int_{-\infty}^{+\infty} x\, f_X(x)\, \mathrm{d}x, & \text{si la v.a. } X \text{ est continue,} \\ \displaystyle\sum_{x \in \mathscr{S}(X)} x\, f_X(x), & \text{si la v.a. } X \text{ est discrète,} \end{cases}$$

où f_X est la densité (usuelle ou discrète, selon le cas) de X, et la notation $\int_A \cdots \mathrm{d}x$ représente (ici et dans la suite) une somme sur $x \in A$ si l'ensemble A est discret (c'est-à-dire fini ou dénombrable), et un intégral usuel sur A sinon (ici, comme $f_X(x) = 0$ pour tout $x \notin \mathscr{S}(X)$, on peut prendre l'intégral sur tout le $\mathbb{R}$). Une fois de plus, ceci peut être vu comme une simple convention qui nous permet de traiter les cas continue et discret simultanément, même si, en réalité, cela a un sens plus profond dans le cadre de la théorie de la mesure. Notons également que $\mathbf{E}\, X$ peut ne pas exister (car la série ou la somme peut ne pas converger).

Une transformation de X par une fonction φ est une nouvelle variable aléatoire $Y = \varphi(X)$. On peut calculer son espérance en utilisant la densité de X (plutôt que sa propre densité) grâce à la **formule de transfert**

$$\mathbf{E}\, \varphi(X) = \int_{\mathscr{S}(X)} \varphi(x)\, f_X(x)\, \mathrm{d}x.$$

Rappelons enfin que l'espérance est linéaire, c'est-à-dire si X et Y sont des variables aléatoires, on a $\mathbf{E}(aX + bY) = a\, \mathbf{E}\, X + b\, \mathbf{E}\, Y$ pour tout $a, b \in \mathbb{R}$, et que $\mathbf{E}(XY) = \mathbf{E}(X)\, \mathbf{E}(Y)$ si les variables aléatoires X et Y sont indépendantes (pour la notion d'indépendance, ainsi que pour celles de loi et d'espérance conditionnelles, nous renvoyons au cours de probabilité).

Variance d'une variable aléatoire

Dans le but de quantifier la dispersion de X autour de sa moyenne $\mathbf{E}\,X$, on cherche à introduire une quantité qui mesure la grandeur "moyenne" de l'écart $X - \mathbf{E}\,X$ entre X et $\mathbf{E}\,X$. Une première idée est de considérer l'espérance de cet écart. Or, $\mathbf{E}(X - \mathbf{E}\,X) = 0$ (les écarts positifs et négatifs se compensent), et on ne peut donc pas utiliser cette quantité. Une solution consisterait à utiliser la quantité $\mathbf{E}|X - \mathbf{E}\,X|$, dite **écart-moyen**. Mais la valeur absolue n'étant pas très pratique à manipuler, on lui préfère la plupart du temps la moyenne quadratique. On introduit donc la **variance** et l'**écart-type** de X (lorsque ceux-ci existent) par

$$\mathbf{Var}(X) = \mathbf{E}\,(X - \mathbf{E}\,X)^2 \quad \text{et} \quad \boldsymbol{\sigma}(X) = \sqrt{\mathbf{Var}(X)}.$$

On a également

$$\mathbf{Var}(X) = \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2.$$

Cette égalité, connue sous le nom de la **formule de König**, est un cas particulier d'une formule plus générale (qu'on appellera **formule de König généralisée**)

$$\mathbf{Var}(X) = \mathbf{E}\,(X - a)^2 - (\mathbf{E}\,X - a)^2,$$

valable pour $a \in \mathbb{R}$ quelconque. Notons que pour $a = \mathbf{E}\,X$, on retrouve la définition de la variance, et pour $a = 0$, la formule de König, et que pour la démontrer, il suffit de développer le côté droit, qui après simplifications donnera $\mathbf{E}\,X^2 - (\mathbf{E}\,X)^2$ (et, comme ce résultat ne dépend pas de a, on peut substituer $a = \mathbf{E}\,X$ et voir ainsi qu'il est égal aussi à la variance de X).

Rappelons enfin que $\mathbf{Var}(X + a) = \mathbf{Var}(X)$ et $\mathbf{Var}(aX) = a^2\,\mathbf{Var}(X)$ pour tout $a \in \mathbb{R}$, et que $\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y)$ si les variables aléatoires X et Y sont indépendantes.

Moments d'une variable aléatoire

Plus généralement, pour $p \in \mathbb{N}^*$, on introduit les différents **moments** d'ordre p d'une variable aléatoire X.

- Le **moment** (**ordinaire**) d'ordre p de X est $m_p = \mathbf{E}\,X^p$.
- Le **moment centré** d'ordre p de X est $\mu_p = \mathbf{E}\,(X - \mathbf{E}\,X)^p$.
- Le **moment absolu** d'ordre p de X est $\mathbf{E}\,|X|^p$.
- Le **moment centré absolu** d'ordre p de X est $\mathbf{E}\,|X - \mathbf{E}\,X|^p$.

Notons que les différents moments d'un même ordre p existent ou pas tous en même temps, et que l'existence des moments d'ordre supérieur implique l'existence des moments d'ordre inférieur.

Notons également que $m_1 = \mathbf{E}\, X$, et que $\mu_1 = 0$ et $\mu_2 = \mathbf{Var}(X)$. Autrement dit, la moyenne et la variance sont respectivement le premier moment et le premier moment centré non-triviaux.

Notons finalement qu'on peut parfois (et lorsqu'ils existent) utiliser également des moments d'ordre fractionnaire ou même négatif (comme, par exemple $p = 1/6$ ou $p = -2$).

Quelques autres caractéristiques d'une variable aléatoire

— Pour une variable aléatoire X possédant des moments d'ordre 2 telle que $\mathbf{E}\, X \neq 0$, on introduit le **coefficient de variation** (dit également **écart-type relatif**) de X par

$$c_{\mathrm{v}} = \frac{\boldsymbol{\sigma}(X)}{\mathbf{E}\, X}\,.$$

— Pour une variable aléatoire non-déterministe X possédant des moments d'ordre 3, on introduit le **coefficient d'asymétrie** (**skewness** en anglais) de X par

$$\gamma_1 = \mathbf{E}\left(\frac{X - \mathbf{E}\, X}{\sigma(X)}\right)^3 = \frac{\mu_3}{\sigma^3(X)}\,.$$

Notons que le signe de γ_1 permet de savoir si la densité de la loi de X est symétrique, "penchée" à droite ou "penchée" à gauche (les trois cas de figure sont représentés dans la Figure 1.2 ci-dessous).

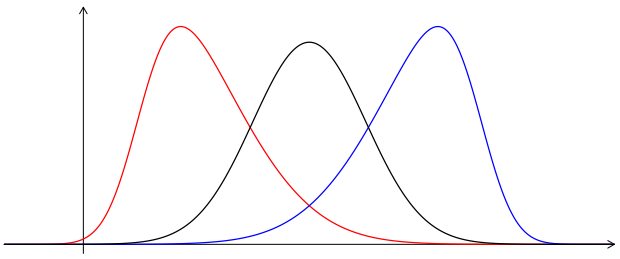

FIGURE 1.2 – Coefficients d'asymétrie nul, positif et négatif

— Pour une variable aléatoire non-déterministe X possédant des moments d'ordre 4, on introduit le **kurtosis** (dit également **coefficient d'aplatissement**) de X par

$$\beta_2 = \mathbf{E}\left(\frac{X - \mathbf{E}\, X}{\sigma(X)}\right)^4 = \frac{\mu_4}{\sigma^4(X)}\,.$$

Notons que pour les lois normales, le kurtosis vaut 3. On dit que c'est des lois **mésokurtiques**. Les lois dont le kurtosis est supérieur à 3 (lois **leptokurtiques**) ont des densités plus "pointues" que celle de la loi normale, et les lois dont le kurtosis

est inférieur à 3 (lois **platikurtiques**) ont des densités plus "plates" (les trois cas de figure sont représentés dans la Figure 1.3 ci-dessous). Notons également que le terme "coefficient d'aplatissement" pourrait laisser plutôt croire le contraire.

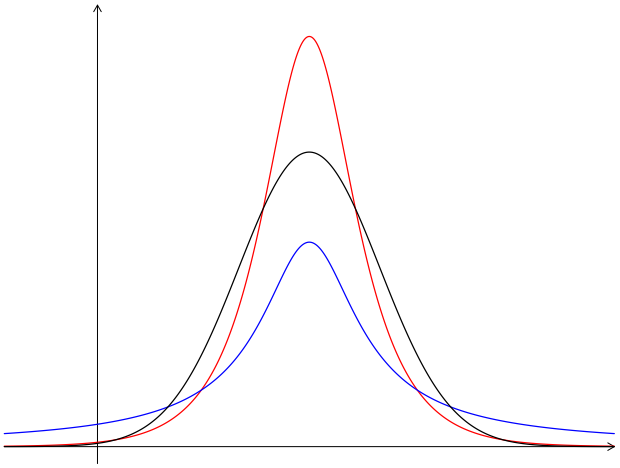

FIGURE 1.3 – Lois mésokurtique, leptokurtique et platikurtique

Notons finalement que certains logiciels et/ou auteurs utilisent plutôt le **kurtosis normalisé** (dit également **excès d'aplatissement**) $\gamma_2 = \beta_2 - 3$, qui est égal à 0 pour les lois mésokurtiques (et, en particulier, pour les lois normales).

Covariance et coefficient de corrélation

La **covariance** de (ou entre) deux variables aléatoires X et Y est définie comme

$$\mathbf{Cov}(X,Y) = \mathbf{E}\big((X - \mathbf{E}\,X)(Y - \mathbf{E}\,Y)\big) = \mathbf{E}(XY) - \mathbf{E}(X)\,\mathbf{E}(Y).$$

Comme pour la variance, une formule plus générale

$$\mathbf{Cov}(X,Y) = \mathbf{E}\big((X - a)(Y - b)\big) - (\mathbf{E}\,X - a)(\mathbf{E}\,Y - b)$$

est valable pour $a, b \in \mathbb{R}$ quelconques.

Notons que si les variables aléatoires X et Y sont indépendantes, on a $\mathbf{Cov}(X,Y) = 0$. L'implication inverse n'est pas vraie en général (elle est, cependant, vraie pour les couples gaussiens de variables aléatoires).

On introduit également le **coefficient de corrélation** de X et Y $\big($appartenant à l'intervalle $[-1,1]\big)$ par

$$\varrho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\boldsymbol{\sigma}(X)\;\boldsymbol{\sigma}(Y)}\,.$$

Finalement, pour un vecteur aléatoire $\vec{X} = (X_1, \ldots, X_n)^t$, on introduit sa **matrice de variance-covariance** comme étant $\Sigma = (\sigma_{ij})_{i,j\in\{1,\ldots,n\}}$, où $\sigma_{ij} = \mathbf{Cov}(X_i, X_j)$.

1.2 Quantiles et intervalles de fluctuation

Nous introduisons maintenant les notions de quantile (qui, bien qu'étant un objet probabiliste, est très important en statistique) et d'intervalle de fluctuation d'une variable aléatoire.

Quantiles

Nous définissons d'abord les quantiles dans le cas particulier d'une variable aléatoire X ayant une fonction de répartition F_X continue et strictement croissante, sauf éventuellement des paliers à hauteur 0 et/ou 1. Dans ce cas, on appelle **quantile d'ordre** $\alpha \in]0,1[$ de X la valeur $q_\alpha = Q_X(\alpha)$ telle que $\mathbf{P}(X \leqslant q_\alpha) = \alpha$. On complète cette définition en posant $q_0 = \inf\big(\mathscr{S}(X)\big)$ et $q_1 = \sup\big(\mathscr{S}(X)\big)$ (bords gauche et droit du support de X, qui peuvent éventuellement être infinis).

Notons que dans ce cas, (la restriction de) la fonction F_X sur l'intervalle $]q_0, q_1[$ est une bijection entre $]q_0, q_1[$ et $]0,1[$. Elle est donc inversible, et comme $F_X(q_\alpha) = \mathbf{P}(X \leqslant q_\alpha) = \alpha$, on a $q_\alpha = F_X^{-1}(\alpha)$ pour tout $\alpha \in]0,1[$ (*cf.* la Figure 1.4 (a) ci-dessous).

Notons également que si f_X est la densité de X, comme $\int_{-\infty}^{q_\alpha} f_x(x)\,\mathrm{d}x = F_x(q_\alpha) = \alpha$, l'aire sous la courbe de f_X à gauche de q_α est égale à α (*cf.* la Figure 1.4 (b) ci-dessous).

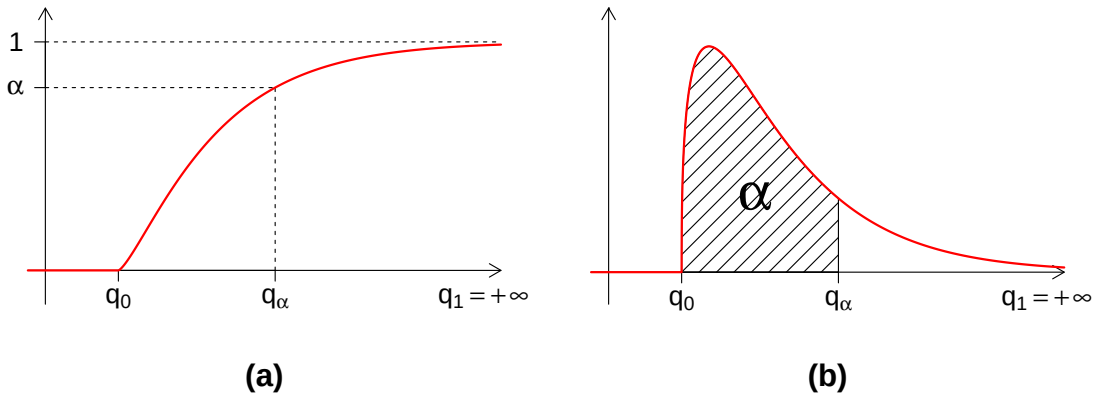

FIGURE 1.4 – Quantiles lorsque la fonction de répartition est inversible

Si F_X n'est pas continue et/ou strictement croissante (notamment pour les lois discrètes), la situation devient plus compliquée.

Comme on peut le voir dans la Figure 1.5 ci-dessous, deux types de problèmes peuvent se poser. Pour des valeurs comme α_1, l'inverse peut ne pas exister, et pour les valeurs comme α_2, il peut ne pas être unique. Dans le premier cas, il serait logique de définir le quantile d'ordre α_1 par $q_{\alpha_1} = c$. Tandis que dans le deuxième cas, on pourrait envisager deux solutions : dire que le quantile d'ordre α_2 n'est pas unique (toute valeur appartenant à

l'intervalle $[a, b]$ en est un), ou choisir une valeur particulière (par exemple, le bord gauche ou le bord droit de l'intervalle $[a, b]$, ou encore son milieu).

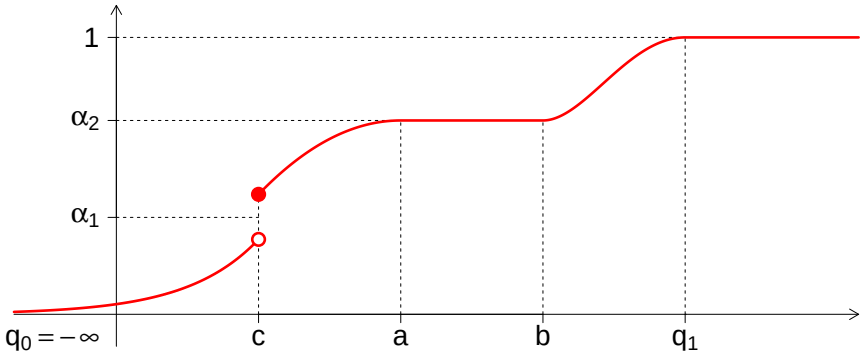

FIGURE 1.5 – Quantiles dans le cas général

En théorie des probabilités, on définit habituellement le quantile d'ordre $\alpha \in]0, 1[$ de X par $q_\alpha = Q_X(\alpha) = \inf\{x \,:\, F(x) \geqslant x\}$. Dans notre exemple, pour α_1, cela donne bien $q_{\alpha_1} = c$, et pour α_2, on obtient $q_{\alpha_2} = a$. L'intérêt de cette définition est que la fonction Q_X est bien définie, et qu'elle est ce qu'on appelle l'**inverse généralisé** de F_X.

En statistique, on préfère la plupart du temps définir le quantile d'ordre $\alpha \in]0, 1[$ de X comme toute valeur x telle que

$$\mathbf{P}(X \leqslant x) \geqslant \alpha \qquad \big(\iff F_X(x) \geqslant \alpha\big)$$

et

$$\mathbf{P}(X \geqslant x) \geqslant 1 - \alpha \qquad \big(\iff \mathbf{P}(X < x) \leqslant \alpha \iff F_X(x-) \leqslant \alpha\big).$$

Dans notre exemple, pour α_1, cela donne bien $q_{\alpha_1} = c$, et pour α_2, toute valeur appartenant à l'intervalle $[a, b]$ est bien un quantile d'ordre α_2.

Parfois (et c'est, notamment, ce que fait la plupart de logiciels de statistique), lorsque la valeur vérifiant les conditions ci-dessus n'est pas unique, on défini le quantile q_α comme le milieu de l'intervalle formé par de telles valeurs $\big($ce qui donne $q_{\alpha_2} = \frac{a+b}{2}$ dans notre exemple$\big)$.

Notons finalement que lorsque $\alpha = {}^k\!/_n$ (avec k et n entiers), on utilise parfois le terme **fractile** pour désigner le quantile q_α, et que certains fractiles particuliers ont des noms spécifiques. Notamment :

- le quantile d'ordre $1/2$ est dit **médiane**,
- les quantiles d'ordre $1/4$ et $3/4$ sont dits (premier et troisième) **quartiles**,
- les quantiles d'ordre ${}^i\!/_{10}$, $i = 1, \ldots, 10$, sont dits **déciles**,
- les quantiles d'ordre ${}^i\!/_{100}$, $i = 1, \ldots, 100$, sont dits **centiles**.

Pour les quantiles de la loi normale, de la loi de Student, de la loi de chi-deux et de la loi de Fisher-Snedecor, nous renvoyons à la brochure **Lois de Probabilité et Tables Statistiques**.

Intervalles de fluctuation

Soit une variable aléatoire X donnée, et soit $\beta \in [0,1]$. Tout intervalle $(a,b) \subset \mathbb{R}$ tel que

$$\mathbf{P}\big(X \in (a,b)\big) = \beta$$

est dit **intervalle de fluctuation** (ou **de prévision**, ou **de prédiction**) de niveau β pour X. On appelle β **niveau** (**de confiance**) et $\alpha = 1-\beta$ (**niveau de**) **risque**. Bien-sûr, en pratique β est choisi proche de 1 (et α proche de 0).

Notons que comme on considérera les intervalles de fluctuation uniquement pour des variables aléatoire continues, on pourra toujours choisir de prendre un intervalle fermé $[a,b]$.

L'intervalle de fluctuation n'est pas unique. Il peut être :

- **unilatéral à gauche** : $\left]-\infty,\, q_{1-\alpha}\right]$ $\big($on peut prendre plutôt $\left[q_0,\, q_{1-\alpha}\right]$ si $q_0 > -\infty\big)$,
- **unilatéral à droite** : $\left[q_\alpha,\, +\infty\right]$ $\big($on peut prendre plutôt $\left[q_\alpha,\, q_1\right]$ si $q_1 < +\infty\big)$,
- **bilatéral symétrique** : $\left[q_{\frac{\alpha}{2}},\, q_{1-\frac{\alpha}{2}}\right]$,
- **bilatéral asymétrique** : $\left[q_{(1-r)\alpha},\, q_{1-r\alpha}\right]$ (ici $r \in [0,1]$, mais pour $r = \frac{1}{2}$ on retombe sur l'intervalle bilatéral symétrique, et pour $r = 0$ et $r = 1$ sur les intervalles unilatéraux).

Ces différents intervalles de fluctuation sont illustrés (en bleu) dans la Figure 1.6 ci-dessous.

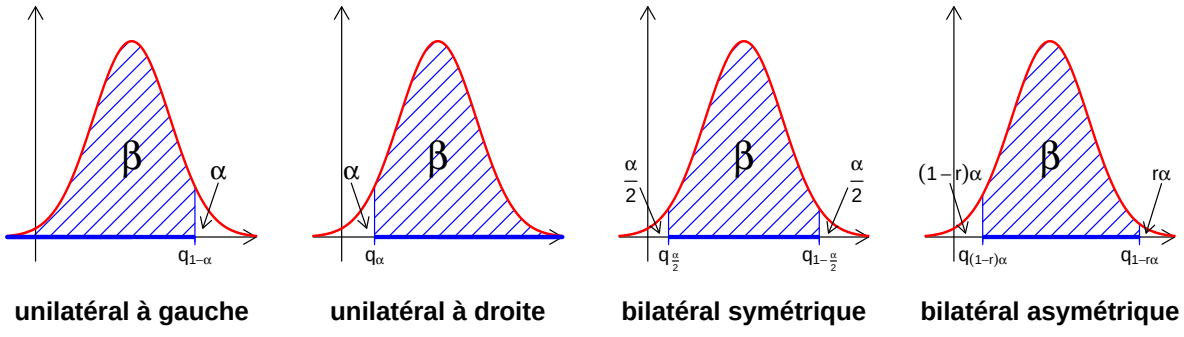

FIGURE 1.6 – Différentes formes de l'intervalle de fluctuation

Le choix de la forme de l'intervalle de fluctuation relève du bon sens : on prend toujours un intervalle de fluctuation inclus dans le support de X, et on essaie d'exclure la (ou les) zone(s) où la densité est la plus faible. Par exemple, parmi les intervalles de fluctuation

représentés dans la Figure 1.7 ci-dessous, on choisira l'intervalle (a), car les intervalles (b) et (c) excluent des zones où la densité est forte, et l'intervalle (d) dépasse du support.

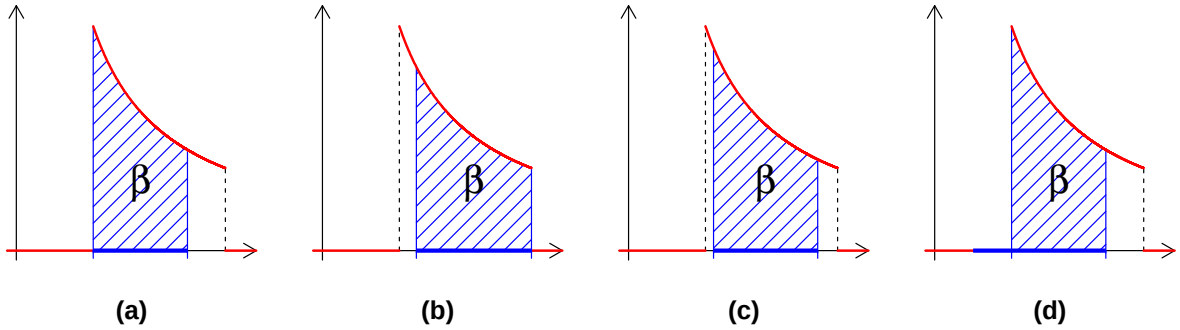

FIGURE 1.7 – Bon (a) et mauvais (b)–(d) choix de l'intervalle de fluctuation

Pour les lois ayant une densité "en cloche" (qui est croissante, puis décroissante), on utilise souvent l'intervalle de fluctuation bilatéral symétrique $\left[q_{\frac{\alpha}{2}},\, q_{1-\frac{\alpha}{2}}\right]$. Ce choix est optimal si la loi de X est symétrique (c'est-à-dire si sa densité f_X admet un axe de symétrie vertical). Mais en général, l'intervalle de fluctuation le plus court de risque α est l'intervalle bilatéral asymétrique $[q_{(1-r)\alpha},\, q_{1-r\alpha}]$ qui égalise les valeurs de la densité sur les deux bords, c'est-à-dire avec r solution de l'équation $f_X(q_{(1-r)\alpha}) = f_X(q_{1-r\alpha})$. En effet, comme on peut le constater dans la Figure 1.8 ci-dessous, cet intervalle (en rouge) est bien plus cours que l'intervalle bilatéral symétrique (en bleu), car les zones hachurées en rouge uniquement et en bleu uniquement ont la même aire, et donc la première, étant plus "haute", est forcément moins large. De plus, ce même raisonnement montre que l'intervalle en rouge est plus court que n'importe quel autre intervalle de fluctuation de risque α.

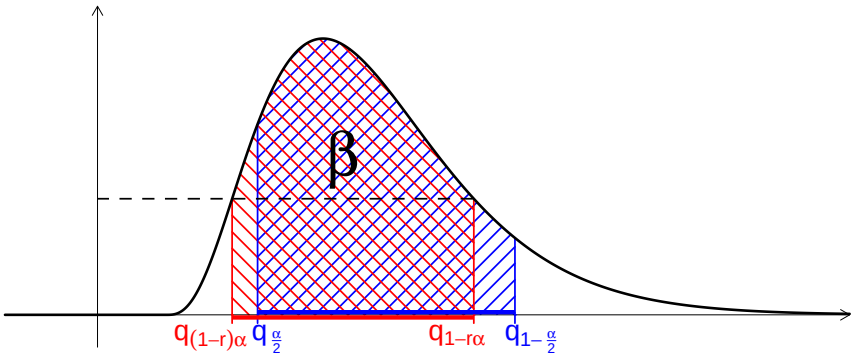

FIGURE 1.8 – Intervalle de fluctuation optimal

Notons cependant que (sauf, bien évidemment, pour les lois symétriques) il est rarement possible de trouver cet intervalle de fluctuation optimal analytiquement. Ceci explique

probablement l'usage quasi systématique des intervalles bilatérales symétriques pour les lois ayant une densité "en cloche" (malgré le fait que les intervalles optimaux pourrait sans trop de difficulté être calculés numériquement à l'aide des logiciels !).

Notons également que si $\mathbf{E}\, X = m$ et $\mathbf{Var}(X) = \sigma^2$, alors (grâce à l'inégalité de Bienaymé-Tchebychev) l'intervalle $[m - 3\sigma,\, m + 3\sigma]$ est un intervalle de fluctuation d'un niveau au moins égal à $\frac{8}{9} \approx 89\%$. D'ailleurs, pour la plupart des lois usuelles, ce niveau est beaucoup plus proche de 1 (pour les lois normales, par exemple, il est à peu près égal à 99,73%). Ceci est connu sous le nom de **règle des trois sigmas**.

Notons finalement que les intervalles de fluctuation sont déterministes et ne doivent surtout pas être confondus avec les intervalles de confiance (qu'on verra plus tard) qui sont aléatoires et dont la philosophie est complètement différente.

1.3 Convergences de variables aléatoires et théorèmes limites

Nous rappelons finalement les différents types de convergence des suites de variables aléatoires.

Soit $X_1, X_2, \ldots$ et X des variables aléatoires.

— On dit que X_n converge **presque sûrement** vers X si

$$\mathbf{P}(X_n \to X) = \mathbf{P}\big(\{\omega \,:\, X_n(\omega) \to X(\omega)\}\big) = 1.$$

On le note $X_n \xrightarrow{\text{p.s.}} X$.

— On dit que X_n converge **en probabilité** vers X si

$$\lim_{n\to+\infty} \mathbf{P}\big(|X_n - X| \geqslant \delta\big) = 0 \text{ pour tout } \delta > 0.$$

On le note $X_n \xrightarrow{\mathbf{P}} X$.

— Soit $p \in \mathbb{N}^*$ (ou, plus généralement, $p \in \mathbb{R}_+^*$). On dit que X_n converge **en** L^p (ou également **en moyenne d'ordre** p) vers X si

$$\mathbf{E}\,|X_n - X|^p \to 0.$$

On le note $X_n \xrightarrow{L^p} X$.

Le cas le plus couramment utilisé est $p = 2$. Dans ce cas, on parle également de convergence **en moyenne quadratique**.

— On dit que X_n converge **en loi** vers X si

$$\mathbf{E}\,\varphi(X_n) \to \mathbf{E}\,\varphi(X) \text{ pour toute fonction continue bornée } \varphi.$$

De manière équivalente, X_n converge en loi vers X si et seulement si $F_{X_n}(x) \to F_X(x)$ pour tout $x \in \mathbb{R}$ tel que F_X est continue au point x.

On le note $X_n \xrightarrow{\mathcal{L}} X$.

Notons que pour que la définition de la convergence presque sûre ait un sens, il faut supposer que les variables aléatoires X_i et X sont toutes définies sûr le même espace probabilisé. Il en est de même pour les convergences en probabilité et en L^p, sauf si la variable aléatoire limite X est déterministe. Par contre, ce n'est pas nécessaire pour la convergence en loi. En quelque sorte, la convergence en loi de X_n vers X est plutôt la convergence de la loi de X_n vers celle de X. D'ailleurs, par abus de notation, on écrira également $X_n \xrightarrow{\mathcal{L}} \mathcal{L}_X$.

On a la hiérarchie suivante entre différents types de convergence :

$$\begin{array}{l} X_n \xrightarrow{\text{p.s.}} X \\ \\ X_n \xrightarrow{L^p} X \end{array} \begin{array}{c} \searrow \\ \\ \nearrow \end{array} X_n \xrightarrow{\mathbf{P}} X \quad \Rightarrow \quad X_n \xrightarrow{\mathcal{L}} X.$$

Notons que les implications non présentes dans ce diagramme ne sont pas vraies en général. Cependant, dans le cas particulier où la variable aléatoire limite X est déterministe, disons $X = a$ ($\in \mathbb{R}$), on a

$$X_n \xrightarrow{\mathbf{P}} a \quad \Longleftrightarrow \quad X_n \xrightarrow{\mathcal{L}} a.$$

Notons également que, comme la fonction $x \longmapsto x^p$ n'est pas bornée, la convergence en loi ($X_n \xrightarrow{\mathcal{L}} X$) n'implique pas en général la convergence des moments ($\mathbf{E}\, X_n^p \to \mathbf{E}\, X^p$). Cependant, cette implication est vraie sous certaines conditions supplémentaires (qu'on ne détaillera pas ici) qui sont souvent vérifiées en pratique.

Pour conclure, rappelons deux théorèmes limites importants de la théorie de probabilité qui, comme on le verra plus tard, sont fondamentaux pour la statistique inférentielle. Ici et dans la suite, nous utiliserons l'abréviation **i.i.d.** pour dire que des variables aléatoires sont **indépendantes et identiquement distribuées** (c'est-à-dire suivent toutes la même loi).

Théorème (Loi forte des grands nombres ou LFGN)**.** *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. possédant des moments d'ordre 1. Alors*

$$\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{p.s.}} \mathbf{E}\, X_1.$$

Théorème (Théorème central limite ou TCL)**.** *Soit $X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d. possédant des moments d'ordre 2. Alors*

$$\frac{\sum_{i=1}^{n} X_i - n\, \mathbf{E}\, X_1}{\sqrt{n\, \mathbf{Var}(X_1)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Notons qu'on a $\mathbf{E}\sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{E}\,X_i = n\,\mathbf{E}\,X_1$ et, en utilisant l'indépendance des X_i, on a $\mathbf{Var}\big(\sum_{i=1}^n X_i\big) = \sum_{i=1}^n \mathbf{Var}(X_i) = n\,\mathbf{Var}(X_1)$. Par conséquent, $\frac{\sum_{i=1}^n X_i - n\,\mathbf{E}\,X_1}{\sqrt{n\,\mathbf{Var}(X_1)}}$ est la version centrée et réduite de la variable aléatoire $\sum_{i=1}^n X_i$. Le fait que la loi limite est centrée et réduite n'est donc pas une surprise ; la "magie" de ce théorème est plutôt dans le fait que cette loi limite est normale (et ceci, quelle que soit la loi de départ des variables aléatoires X_i !). On peut également dire que, pour n grand, on a

$$\frac{\sum_{i=1}^n X_i - n\,\mathbf{E}\,X_1}{\sqrt{n\,\mathbf{Var}(X_1)}} \hookrightarrow \approx \mathcal{N}(0,1),$$

et donc

$$\sum_{i=1}^n X_i - n\,\mathbf{E}\,X_1 \hookrightarrow \approx \mathcal{N}\big(0,\, n\,\mathbf{Var}(X_1)\big),$$

ou encore

$$\sum_{i=1}^n X_i \hookrightarrow \approx \mathcal{N}\big(n\,\mathbf{E}\,X_1,\, n\,\mathbf{Var}(X_1)\big).$$

Autrement dit, le TCL montre que la somme de beaucoup d'effets aléatoires similaires indépendants se comporte de manière universelle : elle suit, à peu près, une loi normale (avec la bonne moyenne et la bonne variance) quelle que soit la nature des effets de départ.

2 Rappels et compléments de la statistique descriptive

2.1 Notions et objets de base

En statistique descriptive, on considère une **population** constituée de n **individus** (dits parfois également **unités statistiques**). Notons qu'un "individu" n'est pas nécessairement un être humain ; la population peut être, par exemple, l'ensemble des étudiants de Polytech'Lille ou des électeurs français, tout comme l'ensemble des ménages français ou des voitures assurées par une certaine compagnie d'assurance.

Pour chaque individu, on observe la valeur d'un (ou de plusieurs) **caractère(s)** (qu'on appelle également **variable(s)**). Ça peut être, par exemple, le poids, la taille, la couleur des yeux ou des cheveux pour un étudiant ; l'âge, la situation professionnelle ou l'intention de vote aux prochaines élections pour un électeur ; le nombre de personnes le constituant ou le revenu global pour un ménage ; la puissance fiscale ou le nombre d'accidents pour une voiture, *etc.*

Autrement dit, pour un caractère x, les observations sont $x_1, \ldots, x_n$, où x_k est la valeur de ce caractère pour le $k^{\text{ème}}$ individu.

Ces observations sont résumées dans ce qu'on appelle un **tableau brut de données** contenant une ligne par individu. Un exemple de tableau brut de données (pour une petite classe de 10 élèves sur laquelle on observe quatre caractères : sexe, note, âge et taille) est donné ci-dessous.

	Sexe	Note	Age	Taille
1	M	AB	14	175,5
2	F	P	13	166
3	M	AB	14	161,5
4	F	P	12	152
5	M	P	13	159
6	F	TB	12	140,5
7	F	P	12	143
8	M	B	16	178,5
9	M	B	15	170
10	M	AB	15	169

Un caractère est dit **quantitatif** si ces valeurs sont numériques. Sinon, il est dit **qualitatif**. Par exemple, pour notre classe d'élèves, les caractères âge et poids sont quantitatifs, tandis que les caractères sexe et note sont qualitatifs. Notons que parfois, les valeurs d'un caractère qualitatif peuvent être codées par des nombres (par exemple, les réponses sur une **échelle de Likert** : 1 = « pas du tout d'accord », 2 = « pas d'accord », 3 = « ni en désaccord ni d'accord », 4 = « d'accord » et 5 = « tout à fait d'accord » à une question de sondage) ; ce caractère ne devient pas quantitatif pour autant.

Un caractère qualitatif est dit **ordinal** si ces valeurs admettent un ordre naturel. Sinon, il est dit **nominal**. Par exemple, le sexe et la couleur des yeux sont des caractères nominaux, tandis que la note (dans notre exemple d'une classe d'élèves) et la réponse sur une échelle de Likert à une question d'un sondage sont des caractères ordinaux.

Quant à un caractère quantitatif, il peut être **discret** ou **continu**. Notons que cette distinction peut parfois être assez subtile. Par exemple, l'âge d'une personne, vu comme la durée du temps écoulé depuis ça naissance, serait plutôt continu par nature, mais comme on le donne en général sous forme d'un nombre (entier) d'années, il est considéré la plupart du temps comme discret. Inversement, le revenu global d'un ménage, même s'il est donné à un euro près, est considéré la plupart du temps comme continu, car il y aurait trop de valeurs différentes si on le considérait comme discret. Pour revenir à notre exemple d'une classe d'élèves, il est naturel de considérer l'âge comme un caractère discret et la taille comme un caractère continu.

Soit $\{c_1, \dots, c_p\}$ l'ensemble des valeurs possibles (dites **modalités**) d'un certain caractère x. Pour tout $i \in \{1, \dots, p\}$, on pose $n_i = \mathrm{Card}\{k \ : \ x_k = c_i\}$ dit **effectif** de la modalité c_i. On a, évidemment, $n_1 + \cdots + n_p = n$ qu'on appelle **effectif total**.

Ainsi, on résume le caractère x par la **série** (ou **table**) **statistique** ci-dessous.

Modalité	c_1	$\cdots$	c_i	$\cdots$	c_p	Total
Effectif	n_1	$\cdots$	n_i	$\cdots$	n_p	n

Par exemple, pour le caractère âge de notre classe d'élèves, on obtient la série statistique suivante.

Modalité	12	13	14	15	16	Total
Effectif	3	2	2	2	1	10

Avant de construire la série statistique d'un caractère quantitatif continu, souvent on le **discrétise**, c'est-à-dire on découpe l'ensemble de ces modalités en intervalles (en **classes**), en faisant attention pour que ceux-ci soient bien disjoints. Par exemple, pour le caractère taille de notre classe d'élèves, on pourrait obtenir la série statistique suivante.

Modalité	$[140, 150[$	$[150, 160[$	$[160, 170[$	$[170, 180]$	Total
Effectif	2	2	3	3	10

Lorsqu'un caractère continu discrétisé a besoin d'être traité comme quantitatif (mais discret), on utilise les centres des classes. Pour la série statistique précédente, cela revient à considérer la table statistique ci-dessous.

Modalité	145	155	165	175	Total
Effectif	2	2	3	3	10

Pour tout $i \in \{1, \dots, p\}$, on pose également $f_i = {}^{n_i}\!/_{n}$ dit **fréquence** de la modalité i. Les fréquences vérifient, évidemment, $f_1 + \cdots + f_p = 1$. Par exemple, pour le caractère âge de notre classe d'élèves, on obtient la table des fréquences suivante.

Modalité	12	13	14	15	16	Total
Fréquence	0,3	0,2	0,2	0,2	0,1	1

Finalement, pour tout $i \in \{1, \dots, p\}$, on pose $N_i = \sum_{j=1}^{i} n_j$ et $F_i = {}^{N_i}\!/_{N} = \sum_{j=1}^{i} f_j$ dits respectivement **effectif cumulé** (**croissant**) et **fréquence cumulée** (**croissante**) de la modalité i. À titre d'exemple, voici les tables des effectifs et des fréquences cumulés pour le caractère âge de notre classe d'élèves.

Modalité	12	13	14	15	16
Eff. cummulé	3	5	7	9	10

Modalité	12	13	14	15	16
Fréq. cummulée	0,3	0,5	0,7	0,9	1

Notons enfin que le mot "effectif" se traduit en anglais par "frequency", et le mot "fréquence" par "percentage", ce qui peut parfois prêter à confusion.

2.2 Résumés numériques d'un caractère

Un des objectifs de la statistique descriptive est de résumer (synthétiser) un caractère par quelques valeurs numériques dites **indicateurs** (ou **caractéristiques**).

Les deux principaux types d'indicateurs sont les indicateurs **de position** (ou **de tendance centrale**), qui permettent de situer les valeurs du caractère, et les indicateurs **de dispersion**, qui permettent de mesurer leur variabilité, mais il existe également des indicateurs qui ne rentrent dans aucune de ces deux catégories.

Principaux indicateurs de position

1. **Moyenne**

   On appelle **moyenne** (**empirique**) d'un caractère quantitatif x la valeur

   $$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{1}{n} \sum_{i=1}^{p} c_i n_i = \sum_{i=1}^{p} c_i f_i.$$

   C'est l'indicateur de tendance centrale le plus naturel. À titre d'exemple, pour le caractère âge de notre classe d'élèves, la moyenne est $\overline{x} = 13{,}6$.

   *Remarque (**très importante** — lien avec la théorie des probabilités).*

   Considérons l'expérience qui consiste à tirer au hasard un individu dans la population et à regarder la modalité du caractère x pour cet individu. Cela revient

à considérer la population initiale comme un univers Ω muni de la mesure de probabilité équirépartie (un poids $1/n$ pour chaque individu), et d'observer une variable aléatoire X, définie comme la fonction sur Ω qui a l'individu k associe la valeur $X(k) = x_k$. Le support de X est $\{c_1, \dots, c_p\}$ et, pour tout $i \in \{1, \dots, p\}$, on a

$$\mathbf{P}(X = c_i) = \frac{\mathrm{Card}\{k \,:\, X(k) = c_i\}}{n} = \frac{\mathrm{Card}\{k \,:\, x_k = c_i\}}{n} = \frac{n_i}{n} = f_i.$$

Ainsi, X suit la loi discrète donnée par la table suivante.

Valeur	c_1	$\cdots$	c_i	$\cdots$	c_p
Probabilité	f_1	$\cdots$	f_i	$\cdots$	f_p

Cette loi est dite la **loi empirique** (du caractère x), et la moyenne empirique $\overline{x}$ n'est rien d'autre que l'espérance (la moyenne) de cette loi ($\overline{x} = \mathbf{E}\, X$).

On pourra donc définir les analogues empiriques des autres caractéristiques (de variables aléatoires) existantes en théorie des probabilités en les appliquant simplement à la loi empirique.

2. **Médiane**

On appelle **médiane** (**empirique**) d'un caractère x la médiane (ou, autrement dit, le quantile d'ordre $1/2$) de sa loi empirique, c'est-à-dire tout m tel que

$$\frac{\mathrm{Card}\{k \,:\, x_k \leqslant m\}}{n} \geqslant \frac{1}{2} \qquad \left(\iff \mathrm{Card}\{k \,:\, x_k \leqslant m\} \geqslant n/2\right)$$

et

$$\frac{\mathrm{Card}\{k \,:\, x_k \geqslant m\}}{n} \geqslant \frac{1}{2} \qquad \left(\iff \mathrm{Card}\{k \,:\, x_k < m\} \leqslant n/2\right).$$

Pour trouver la médiane à partir de la série statistique, il faut donc utiliser la table des effectifs (ou des fréquences) cumulé(e)s. La première modalité dont l'effectif cumulé (ou la fréquence cumulée) dépasse la valeur $n/2$ (ou $1/2$ pour les fréquences) est une médiane. Notons que si pour la modalité précédente on tombe pile sur cette valeur, la médiane n'est pas unique : toute valeur comprise entre ces deux modalités est une médiane. Par exemple, pour le caractère âge de notre classe d'élèves, on a une fréquence cumulée de pile $1/2$ pour la valeur 13, et donc toute valeur $m \in [13, 14]$ est une médiane (on prend parfois le milieu de l'intervalle, c'est-à-dire 13,5 dans cet exemple).

Pour trouver la médiane à partir du tableau brut, on réordonne d'abord la suite $x_1, \dots, x_n$ dans l'ordre croissant. On note, ici et dans la suite, $x_{(1)}, \dots, x_{(n)}$ la permutation obtenue qu'on appelle **série variationnelle**. Par exemple, la série variationnelle

du caractère âge de notre classe d'élèves est $x_{(1)} = x_{(2)} = x_{(3)} = 12$; $x_{(4)} = x_{(5)} = 13$; $x_{(6)} = x_{(7)} = 14$; $x_{(8)} = x_{(9)} = 15$ et $x_{(10)} = 16$. Ensuite :

- si n est impair, on a $n = 2\ell + 1$, et la médiane est $m = x_{(\ell+1)}$;
- si n est pair, on a $n = 2\ell$, et tout $m \in [x_{(\ell)}, x_{(\ell+1)}]$ est une médiane.

Notons que, contrairement à la moyenne, la médiane peut être défini égaelement pour des caractères qualitatifs ordinaux, et qu'elle est moins sensible aux valeurs extrêmes et/ou aberrantes.

3. **Mode**

   Pour les caractères qualitatifs (ainsi que quantitatifs discrets), on utilise parfois le **mode**, qui est la modalité la plus fréquente (ayant le plus grand effectif et/ou la plus grande fréquence). Notons que ce n'est pas toujours un indicateur de tendance centrale très pertinent. Par exemple, pour le caractère âge de notre classe d'élèves, le mode est égal à 12.

Principaux indicateurs de dispersion

1. **Intervalle des valeurs**

   L'**intervalle des valeurs** d'un caractère quantitatif (ou qualitatif ordinal) x est l'intervalle $[x_{(1)}, x_{(n)}]$. Par exemple, pour le caractère âge de notre classe d'élèves, l'intervalle des valeurs est $[12, 16]$.

2. **Étendue**

   L'**étendue** d'un caractère quantitatif x est la longueur $x_{(n)} - x_{(1)}$ de son intervalle des valeurs. Par exemple, pour le caractère âge de notre classe d'élèves, l'étendue est égale à 4.

3. **Intervalle interquartile**

   L'**intervalle interquartile** d'un caractère quantitatif (ou qualitatif ordinal) x est l'intervalle $[Q_1, Q_3]$, où Q_1 est Q_3 sont le 1^{er} et le $3^{\text{ème}}$ quartiles de x (c'est-à-dire de sa loi empirique). Par exemple, pour le caractère âge de notre classe d'élèves, l'intervalle interquartile est $[12, 15]$.

4. **Écart interquartile**

   L'**écart interquartile** d'un caractère quantitatif x est la longueur $Q_3 - Q_1$ de son intervalle interquartile. Par exemple, pour le caractère âge de notre classe d'élèves, l'écart interquartile est égale à 3.

5. **Variance**

On appelle **variance** (**empirique**) d'un caractère quantitatif x la variance de sa loi empirique, c'est-à-dire

$$s_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p}(c_i - \overline{x})^2 f_i.$$

Notons que, grâce à la formule de König (généralisée), on a aussi

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \overline{x}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - a)^2 - (\overline{x} - a)^2 \\ &= \frac{1}{n}\sum_{i=1}^{p} c_i^2\, f_i - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{p}(c_i - a)^2 f_i - (\overline{x} - a)^2, \end{aligned}$$

où $a \in \mathbb{R}$ est quelconque.

Notons également que parfois on utilise plutôt

$$\overline{s}_x^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2 = \frac{n}{n-1}\, s_x^2.$$

On verra plus tard, en statistique inférentielle, la raison d'être de cette modification.

À titre d'exemple, pour le caractère âge de notre classe d'élèves, on a $s_x^2 = 1{,}84$ et $\overline{s}_x^2 \approx 2{,}0444$.

6. **Écart-type**

On appelle **écart-type** (**empirique**) d'un caractère quantitatif x la racine carrée de sa variance empirique, c'est-à-dire $s_x = \sqrt{s_x^2}$ ou $\overline{s}_x = \sqrt{\overline{s}_x^2}$.

À titre d'exemple, pour le caractère âge de notre classe d'élèves, on a $s_x \approx 1{,}3565$ et $\overline{s}_x \approx 1{,}4298$.

Autres indicateurs

De la même manière, on peut définir, pour un caractère quantitatif x, le **coefficient de variation** (**empirique**), le **coefficient d'asymétrie** (**empirique**), le **kurtosis** (**empirique**), *etc.*

2.3 Cas de deux caractères

On considère deux caractères observés sur une même population de n individus : un caractère x ayant les modalités $c_1, \ldots, c_p$ et un caractère y ayant les modalités $d_1, \ldots, d_q$.

Tableau de contingence

Pour tout $i \in \{1, \dots, p\}$ et $j \in \{1, \dots, q\}$, on pose $n_{ij} = \text{Card}\{k \ : \ x_k = c_i \text{ et } y_k = d_j\}$ dit **effectif conjoint** des modalités c_i et d_j.

Le **tableau de contingence** (**des effectifs**) des caractères x et y se présente comme suit.

x \ y	d_1	$\cdots$	d_j	$\cdots$	d_q	Somme
c_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\bullet}$
Somme	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet q}$	$n_{\bullet\bullet} = n$

Notons que les sommes $n_{1\bullet}, \dots, n_{p\bullet}$ sont les effectifs des modalités $c_1, \dots, c_p$ de x, tandis que les sommes $n_{\bullet 1}, \dots, n_{\bullet q}$ sont ceux des modalités $d_1, \dots, d_q$ de y. Ces effectifs sont dits **effectifs marginaux**.

Par exemple, pour les caractères âge et sexe de notre classe d'élèves, on obtient le tableau de contingence suivant.

Âge \ Sexe	12	13	14	15	16	Somme
M	0	1	2	2	1	6
F	3	1	0	0	0	4
Somme	3	2	2	2	1	10

De la même manière, on défini les **fréquences conjointes** $f_{ij} = {}^{n_{ij}}\!/_{n}$, et on construit le **tableau de contingence des fréquences** ci-dessous.

x \ y	d_1	$\cdots$	d_j	$\cdots$	d_q	Somme
c_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1q}	$f_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{iq}	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
c_p	f_{p1}	$\cdots$	f_{pj}	$\cdots$	f_{pq}	$f_{p\bullet}$
Somme	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet q}$	$f_{\bullet\bullet} = 1$

Notons que ce tableau n'est rien d'autre que la loi empirique (jointe) des caractères x et y, et que les **fréquences marginales** $f_{1\bullet}, \ldots, f_{p\bullet}$ et $f_{\bullet 1}, \ldots, f_{\bullet q}$ correspondent respectivement aux lois empiriques (marginales) de x et de y.

Profils lignes

Pour tout $i \in \{1, \ldots, p\}$, on appelle **profil ligne** de la modalité c_i le vecteur ligne

$$\left\{ \frac{n_{i1}}{n_{i\bullet}}, \ldots, \frac{n_{iq}}{n_{i\bullet}} \right\}.$$

Le profil ligne de la modalité c_i est constitué des fréquences des modalités de y parmi les individus pour lesquels $x = c_i$ ou, autrement dit, sachant que $x = c_i$. C'est la loi empirique de y conditionnelle à $x = c_i$. Quant au vecteur ligne

$$\left\{ \frac{n_{\bullet 1}}{n}, \ldots, \frac{n_{\bullet q}}{n} \right\} = \{f_{\bullet 1}, \ldots, f_{\bullet q}\},$$

qui correspond à la loi empirique (marginale) de y, on l'appelle **profil ligne global**.

Profils colonnes

De la même manière, pour tout $j \in \{1, \ldots, q\}$, on appelle **profil colonne** de la modalité d_j le vecteur colonne

$$\left\{ \frac{n_{1j}}{n_{\bullet j}}, \ldots, \frac{n_{pj}}{n_{\bullet j}} \right\}^t.$$

Il est constitué des fréquences des modalités de x parmi les individus pour lesquels $y = d_j$ ou, autrement dit, sachant que $y = d_j$. C'est la loi empirique de x conditionnelle à $y = d_j$. Quant au vecteur colonne

$$\left\{ \frac{n_{1\bullet}}{n}, \ldots, \frac{n_{p\bullet}}{n} \right\}^t = \{f_{1\bullet}, \ldots, f_{p\bullet}\}^t,$$

qui correspond à la loi empirique (marginale) de x, on l'appelle **profil colonne global**.

Covariance

On appelle **covariance** (**empirique**) de (ou entre) deux caractères quantitatifs x et y la covariance de la loi empirique jointe, c'est-à-dire

$$c_{xy} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) = \sum_{i=1}^{p} \sum_{j=1}^{q} (c_i - \overline{x})(d_j - \overline{y}) f_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_k y_k - \overline{x}\,\overline{y} = \cdots$$

À titre d'exemple, pour les caractères âge et taille de notre classe d'élèves, on a $c_{xy} = 14{,}6$.

Coefficient de corrélation

On appelle **coefficient de corrélation** (**linéaire**) de deux caractères quantitatifs x et y le coefficient de corrélation de la loi empirique jointe, c'est-à-dire

$$\varrho_{xy} = \frac{c_{xy}}{s_x s_y}\,.$$

À titre d'exemple, pour les caractères âge et taille de notre classe d'élèves, on a $\varrho_{xy} \approx 0{,}8749$.

2.4 Représentations graphiques

Un autre objectif de la statistique descriptive est de représenter graphiquement (visualiser) les données. Dans cette section, nous passons en revue les représentations graphiques les plus couramment utilisées.

Un caractère qualitatif

Pour représenter un caractère qualitatif, on utilise souvent un **diagramme en barres** (ou **à bandes**), dit **bar chart** en anglais.

Chaque modalité est représentée par une barre verticale (ou horizontale) de largeur fixe et dont la hauteur (ou la longueur) est égale (ou, plus généralement, proportionnelle) à l'effectif de la modalité. Des exemples de diagramme en barres pour le caractère sexe de notre classe d'élèves sont donnés dans la Figure 2.1 ci-dessous.

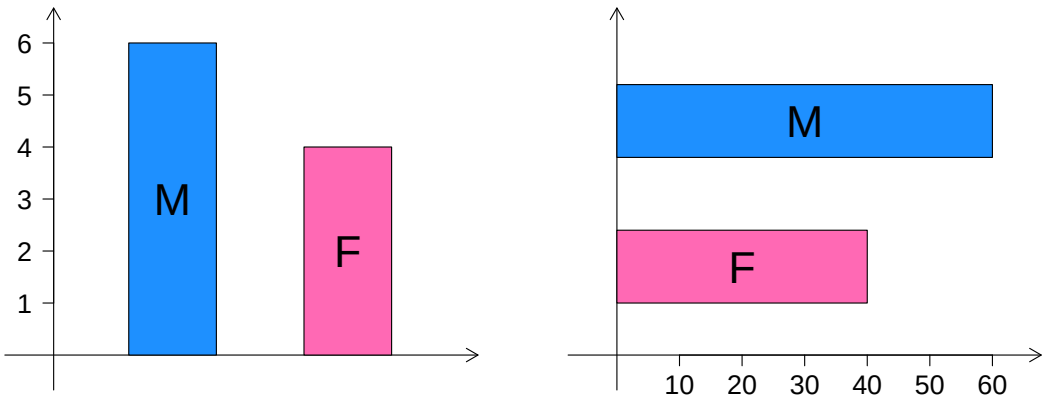

FIGURE 2.1 – Diagrammes en barres du caractère sexe

Il existe également une variante de ce diagramme utilisant une seule barre découpée en morceaux (proportionnellement aux effectifs). Un exemple (toujours pour le caractère sexe

de notre classe d'élèves) est donné dans la Figure 2.2 ci-dessous.

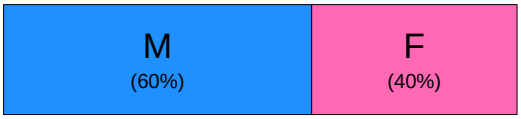

FIGURE 2.2 – Diagramme à une seule barre du caractère sexe

Un autre diagramme couramment utilisé pour représenter un caractère qualitatif est le **diagramme circulaire** (ou **en camembert**), dit **pie chart** en anglais.

Chaque modalité est représentée par un secteur circulaire dont l'angle est proportionnel à l'effectif de la modalité, 2π correspondant à l'effectif total. Un exemple de diagramme circulaire pour le caractère sexe de notre classe d'élèves est donné dans la Figure 2.3 ci-dessous.

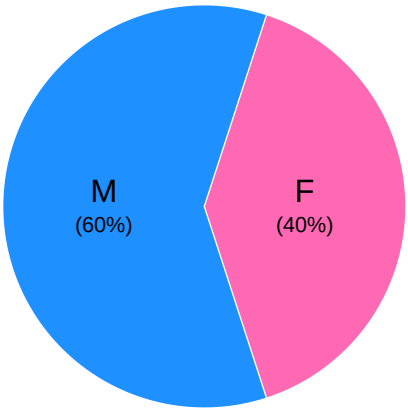

FIGURE 2.3 – Diagramme circulaire du caractère sexe

Parfois on utilise une variante où l'effectif total correspond à π (plutôt qu'à 2π) qu'on appelle **diagramme semi-circulaire** (*cf.* la Figure 2.4 ci-dessous).

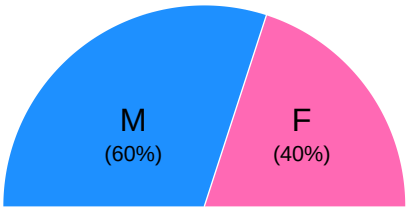

FIGURE 2.4 – Diagramme semi-circulaire du caractère sexe

Un caractère quantitatif

Pour représenter un caractère quantitatif discret, on utilise la plupart du temps un **diagramme en bâtons**.

Il est construit de la même manière que le diagramme en barres, mais au lieu d'utiliser des barres, on utilise des bâtons qu'on place correctement sur l'axe des abscisses suivant les valeurs numériques du caractère. Un exemple de diagramme en bâtons pour le caractère âge de notre classe d'élèves est donné dans la Figure 2.5 ci-dessous.

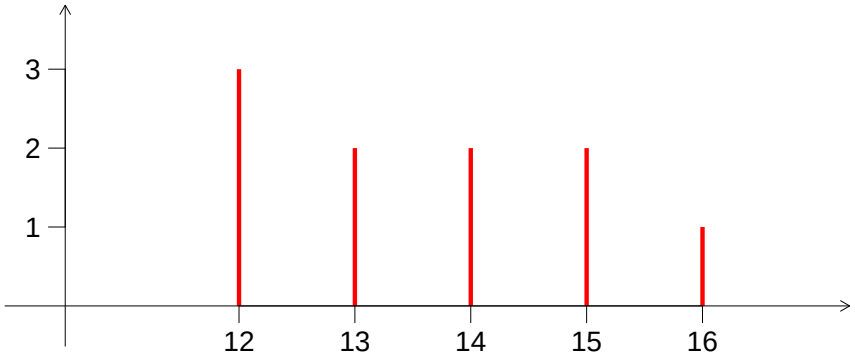

FIGURE 2.5 – Diagramme en bâtons du caractère âge

Pour représenter un caractère quantitatif continu, on utilise la plupart du temps un **histogramme**.

Pour construire un histogramme, on reporte d'abord les classes sur l'axe des abscisses. Ensuite, chaque classe est représenté par un rectangle posé sur l'axe des abscisse dont l'aire est proportionnelle à l'effectif de la classe. Un exemple d'histogramme pour le caractère taille de notre classe d'élèves est donné dans la Figure 2.6 ci-dessous.

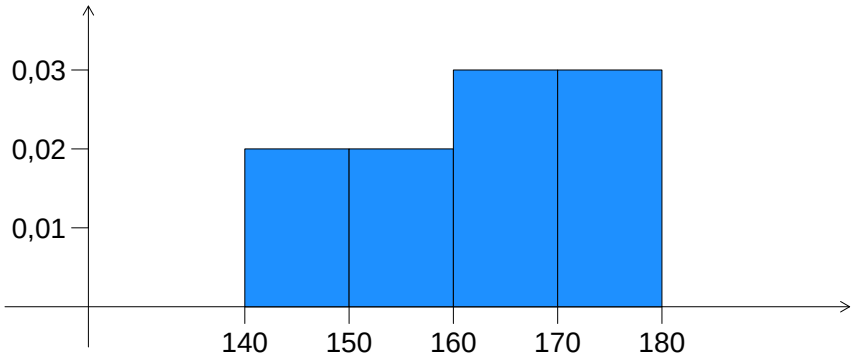

FIGURE 2.6 – Histogramme du caractère taille

Notons que souvent (comme dans l'exemple ci-dessus) on choisit l'échelle sur l'axe des ordonnées de manière que l'aire totale de l'histogramme soit égale à 1 (on peut noter l'analogie avec la densité de probabilité). Pour cela, on utilise des rectangles dont les aires sont égales aux fréquences des classes.

Notons également qu'il est très important que ça soient les aires (et non pas les hauteurs) qui soient proportionnelles aux effectifs (bien-sûr, si les classes sont de la même largeur, les hauteurs le seront aussi, mais ce n'est pas le cas en général). Par exemple, si on regroupe les deux classes du milieu du caractère taille, on obtient l'histogramme représenté dans la Figure 2.7 ci-dessous, qui est logique (tandis qu'avoir la classe du milieu deux fois plut haute ne le serait certainement pas).

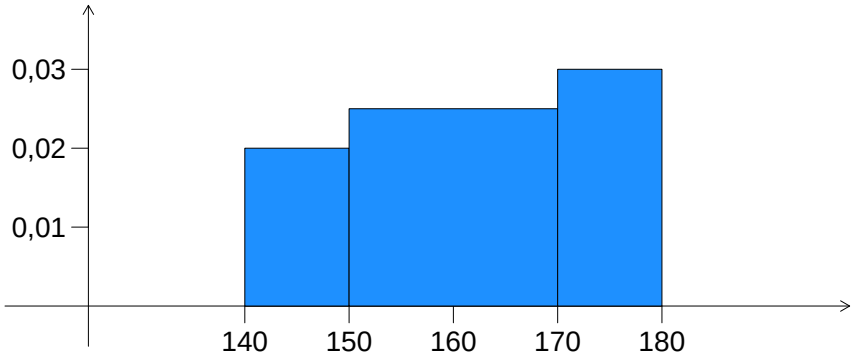

FIGURE 2.7 – Histogramme du caractère taille avec deux classes regroupées

Un autre diagramme couramment utilisé pour représenter d'une manière synthétique un caractère quantitatif (discret ou continue) est la **boîte à moustaches**, dit **box plot** (ou **box-and-whisker plot**) en anglais. Un exemple de boîte à moustaches pour le caractère taille de notre classe d'élèves est donné dans la Figure 2.8 ci-dessous.

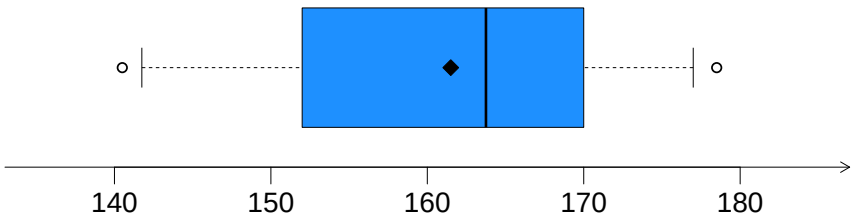

FIGURE 2.8 – Boîte à moustaches du caractère taille

Les abscisses des bords de la boîte sont le 1^{er} et le $3^{\text{ème}}$ quartile, et celle du trait la coupant en deux est la médiane. L'abscisse du petit carré (qui n'est pas toujours présent) est égale à la moyenne, et les petits cercles représentent les valeurs qui sont en dehors des "moustaches",

considérés comme “extrêmes”. Pour les moustaches, il existe plusieurs options. On utilise souvent les moustaches définies pour un caractère x par $\min\{x_k \,:\, x_k \geqslant Q_1 - 1{,}5\,(Q_3 - Q_1)\}$ et $\max\{x_k \,:\, x_k \leqslant Q_3 + 1{,}5\,(Q_3 - Q_1)\}$. C'est la version de la boîte à moustache la plus courante introduite par Tukey, mais on peu également utiliser :

— le minimum $x_{(1)}$ et le maximum $x_{(n)}$ des données (dans ce cas, il n'y a, évidemment, pas de valeurs extrêmes) ;

— 1^{er} et $9^{\text{ème}}$ décile ;

— $5^{\text{ème}}$ et $95^{\text{ème}}$ centile (c'est l'option utilisée dans le graphique ci-dessus) ;

— $\overline{x} - 2s_x$ et $\overline{x} + 2s_x$;

— *etc.*

Finalement, pour représenter un caractère quantitatif, on peut utiliser un **diagramme des effectifs** (ou **des fréquences**) **cumulé(e)s**.

Pour un caractère x, on trace la fonction

$$N(x) = \text{Card}\{k \,:\, x_k \leqslant x\} = \sum_{i\,:\,c_i \leqslant x} n_i,$$

ou la fonction

$$F(x) = \frac{N(x)}{n} = \sum_{i\,:\,c_i \leqslant x} f_i.$$

Notons que la différence entre ces deux diagrammes est juste l'échelle sur l'axe des ordonnées, et que F n'est rien d'autre que la fonction de répartition empirique (la fonction de répartition de la loi empirique). Un exemple de diagramme des effectifs cumulés pour le caractère âge de notre classe d'élèves est donné dans la Figure 2.9 ci-dessous.

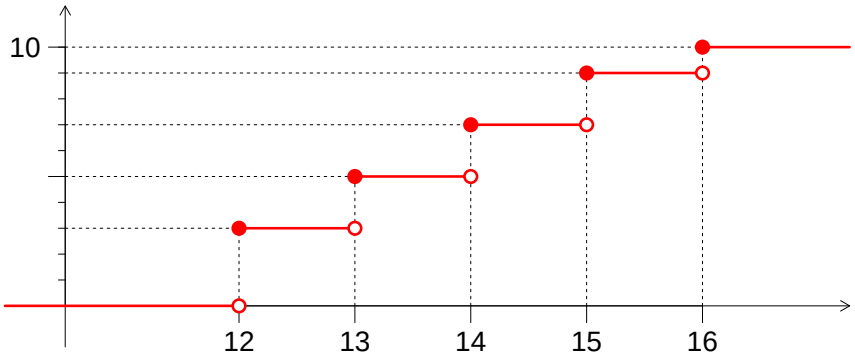

FIGURE 2.9 – Diagramme des effectifs cumulés du caractère âge

Notons également que pour des caractères quantitatifs continus discrétisés, plutôt que d'avoir une fonction en escalier, il est plus logique d'interpoler linéairement à l'intérieur

des classes. Un exemple de diagramme des fréquences cumulées pour notre caractère taille habituel découpé en quatre classes est donné dans la Figure 2.10 ci-dessous.

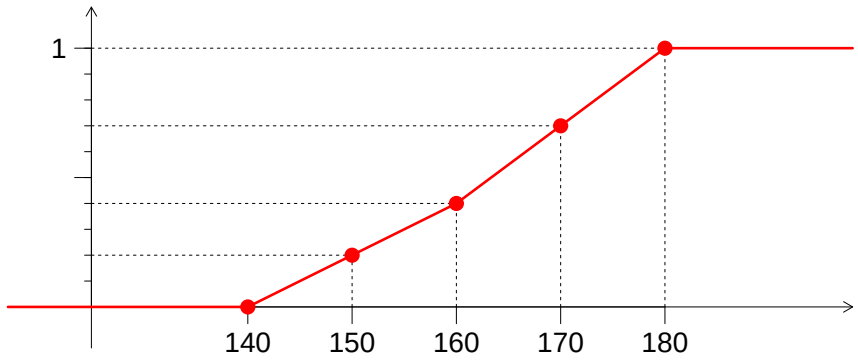

FIGURE 2.10 – Diagramme des fréquences cumulées du caractère taille

Deux caractères

Pour représenter deux caractères quantitatifs x et y, on utilise fréquemment un **nuage de points**, dit **scatter plot** en anglais.

On trace tout simplement l'ensemble des points $\{(x_k, y_k)\}_{k=1,\dots,n}$. Un exemple de nuage de points pour les caractères âge et taille de notre classe d'élèves est donné dans la Figure 2.11 ci-dessous.

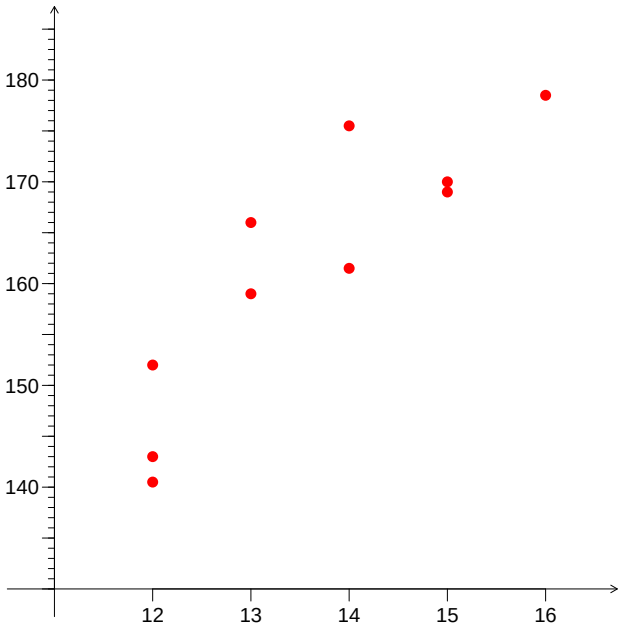

FIGURE 2.11 – Nuage de points des caractères âge et taille

3 Statistique inférentielle — généralités

3.1 Échantillonnage

Soit une grande population (par exemple, l'ensemble des électeurs français) sur laquelle est défini un caractère (par exemple, l'intention de vote aux prochaines élections comportant deux candidats, avec 1 signifiant que l'on va voter pour le candidat A, et 0 que l'on va voter pour le candidat B).

On ne peut pas connaître la valeur du caractère sur l'ensemble de la population (à moins d'attendre le jour des élections !) mais seulement sur une petite partie tirée au hasard (par exemple, en faisant un sondage d'opinion sur n personnes).

Comme nous avons déjà vu, tirer un individu au hasard dans une population revient à considérer la valeur de son caractère comme une variable aléatoire (de la loi empirique liée à la population entière et que l'on ne connaît donc pas en occurrence).

Si les n tirages sont effectués de manière indépendante et avec remise (ou dans une population suffisamment grande, voire infinie, pour que l'on puisse négliger le fait d'avoir déjà enlevé un individu), alors on obtient une suite de variables aléatoires i.i.d. d'une loi inconnue (la loi empirique liée à toute la population). Ceci s'appelle **échantillonnage** et la suite $X_1, \dots, X_n$ de variables aléatoires i.i.d. obtenue s'appelle un **n-échantillon** (ou **échantillon de taille** n).

Le but de la statistique inférentielle est de déduire (inférer) des informations sur la loi commune (inconnue) des X_i à partir des **observations**, c'est-à-dire des réalisations $x_1, \dots, x_n$ des variables aléatoires $X_1, \dots, X_n$. Notons que les termes "échantillon" et "observations" sont parfois utilisés pour désigner indifféremment les variables aléatoires $X_1, \dots, X_n$ et leurs réalisations $x_1, \dots, x_n$.

Pour revenir à l'exemple des intentions de vote, si on échantillonne correctement (par exemple, en faisant des tirages équirépartis et avec remise), on obtient une suite de variables aléatoires i.i.d. de la loi de Bernoulli de paramètre p, où p est la vraie proportion d'intentions de vote pour le candidat A dans l'ensemble de tous les électeurs. On ne connaît, bien-sûr, pas la valeur de p, et on veut l'approcher en interrogeant n personnes, c'est-à-dire en observant un n-échantillon $X_1, \dots, X_n$ de la loi $\mathscr{B}(p)$, $p \in [0, 1]$.

On peut se demander légitimement si c'est possible. La réponse est "oui", du moins pour n suffisamment grand.

En effet, d'après la loi forte des grands nombres (LFGN), on a

$$\overline{X} = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{\text{p.s.}} \mathbf{E}(X_1) = p,$$

et donc, pour n suffisamment grand, on peut approcher p par la réalisation $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ de $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

De plus, on peut quantifier la qualité de l'approximation obtenue (car une approximation dont on ne précise pas la qualité n'est pas très utile : si on dit, par exemple, que $p \approx 0{,}52$ sans préciser si c'est plutôt $0{,}52 \pm 0{,}01$ ou $0{,}52 \pm 0{,}1$, on ne pourra pas tirer des conclusions intéressantes). En effet, d'après le théorème central limite (TCL), on a

$$\begin{aligned} \frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad &\iff \quad \frac{n\big(\overline{X}_n - p\big)}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \\ &\iff \quad \sqrt{\frac{n}{p(1-p)}}\big(\overline{X}_n - p\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \end{aligned}$$

et donc, pour n suffisamment grand,

$$\overline{X}_n - p \rightsquigarrow \approx \sqrt{\frac{p(1-p)}{n}}\,\mathcal{N}(0,1) = \mathcal{N}\left(0,\, \frac{p(1-p)}{n}\right).$$

Or, $\overline{X}_n - p$ est précisément l'erreur d'approximation de p par $\overline{X}_n$, et ainsi on a trouvé sa loi (approchée pour n grand). Pour être plus honnête, il faut noter que dans cet exemple, on ne connaît pas tout à fait cette loi, car sa variance $\frac{p(1-p)}{n}$ dépend du paramètre inconnu p, mais on peut, par exemple, la majorer par $\frac{1}{4n}$, car $p(1-p) \leqslant 1/4$ pour tout $p \in [0,1]$. Finalement, la connaissance de la loi de l'erreur d'approximation permet de contrôler cette dernière. Par exemple, en utilisant la règle des trois sigmas, on peut dire que (avec une très grande probabilité) l'erreur d'approximation est inférieure à $3\sqrt{\frac{p(1-p)}{n}} \leqslant \frac{3}{2\sqrt{n}}$, ou obtenir des affirmations plus précises en utilisant les quantiles de la loi normale.

Que vient-on de faire sur cet exemple des intentions de vote ? En fait, on vient d'**estimer** le paramètre inconnu p $\big($de la loi $\mathscr{B}(p)\big)$. Les problèmes d'**estimation** seront plus généralement traités dans le chapitre 4. On pourrait aussi essayer de répondre à des questions à propos de ce paramètre inconnu (plutôt que de l'estimer). Par exemple, « a-t-on $p > 0,5$? » ou, autrement dit, « le candidat A va-t-il gagner ? ». C'est un problème de **test d'hypothèses**, et on traitera ce type de problèmes dans le chapitre 5.

3.2 Modèles Statistiques

Les considérations de la section précédente nous amènent à la notion suivante du **modèle statistique** (**i.i.d.**) : $X_1, \ldots, X_n$ i.i.d., $X_i \rightsquigarrow \mathbf{P} \in \mathscr{P}$, où $\mathscr{P}$ est une certaine famille de lois de probabilité sur $\mathbb{R}$.

Remarques.

1. Notons qu'en réalité, on a ici une suite de modèles statistiques (un modèle statistique pour chaque $n \in \mathbb{N}^*$). Ceci permet de considérer des propriétés **asymptotiques** (c'est-à-dire lorsque $n \to \infty$), qui pourront être considérées comme valable pour un (seul) modèle avec n suffisamment grand.

2. Dans ce cours on considère uniquement des modèles statistiques i.i.d. Il faut cependant noter qu'il existe des modèles statistiques plus généraux (sortant du cadre i.i.d.), où l'on observe, par exemple, des suites de variables aléatoires dépendantes et/ou non identiquement distribués, ou même des valeurs aléatoires variant avec le temps (dites **processus stochastiques**), plutôt que des suites de variables aléatoires.

On distingue deux types de modèles statistiques : paramétriques et non-paramétriques.

Définition. *Un modèle statistique $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P} \in \mathscr{P}$ est dit* ***paramétrique*** *si $\mathscr{P}$ peut s'écrire sous la forme $\mathscr{P} = \{\mathbf{P}_\theta,\ \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}^*$. Dans ce cas, on écrira le modèle comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta$, et on appellera θ* ***paramètre****.*

Dans le cas contraire, le modèle est dit ***non-paramétrique****.*

Exemples.

— On observe n réalisations indépendantes d'une loi de Bernoulli de paramètre $p \in [0,1]$ inconnu. C'est un modèle paramétrique de dimension 1 (qu'on a déjà aperçu dans la section précédente), qui s'écrit comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathscr{B}(p)$, $p \in [0,1]$ ($\subset \mathbb{R}$). Le paramètre du modèle est $\theta = p$.

— On observe n réalisations indépendantes d'une loi normale de moyenne m et de variance σ^2 inconnues. C'est un modèle paramétrique de dimension 2, qui s'écrit comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathcal{N}(m, \sigma^2)$, $\binom{m}{\sigma^2} \in \mathbb{R} \times \mathbb{R}_+^*$ ($\subset \mathbb{R}^2$). Le paramètre du modèle est $\theta = \binom{m}{\sigma^2}$.

— On observe n réalisations indépendantes d'une loi entièrement inconnue. C'est un modèle non-paramétrique, qui s'écrit comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P} \in \mathscr{P}$, où $\mathscr{P} =$ « ensemble de toutes les lois ».

— On observe n réalisations indépendantes d'une loi continue inconnue (quelconque). C'est un modèle non-paramétrique, qui s'écrit comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P} \in \mathscr{P}$, où $\mathscr{P} =$ « ensemble de toutes les lois continues ».

— On observe n réalisations indépendantes d'une loi inconnue, dont on sait (uniquement) qu'elle possède une espérance. C'est un modèle non-paramétrique, qui s'écrit comme $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P} \in \mathscr{P}$, où $\mathscr{P} =$ « ensemble de toutes les lois ayant une espérance ».

Remarques.

1. Il faut noter ici qu'on peut toujours écrire la famille $\mathscr{P}$ comme $\mathscr{P} = \{\mathbf{P}_\theta,\ \theta \in \Theta\}$. Par exemple, si $\mathscr{P} =$ « ensemble de toutes les lois », on peut écrire $\mathscr{P} = \{\mathbf{P}_F,\ F \in \mathscr{G}\}$, où $\mathscr{G} = \left\{F : F \nearrow,\ F \text{ càdlàg},\ F(-\infty) = 0 \text{ et } F(+\infty) = 1\right\}$ est l'ensemble de toutes les fonctions de répartitions, et $\mathbf{P}_F$ dénote la loi ayant F pour fonction de répartition. Cependant, l'ensemble $\mathscr{G}$ est un ensemble (assez riche) des fonctions qui n'est pas fini-dimensionnel (n'est pas inclus dans $\mathbb{R}^k$ pour aucun $k \in \mathbb{N}^*$). Par abus de langage, on peut appeler tout de même $\theta = F$ "paramètre" infini-dimensionnel dans ce cas.

Ainsi, le point important dans la définition du modèle paramétrique est le fait que le paramètre soit de dimension finie.

2. Dans le cas paramétrique, lorsque $k \geqslant 2$, on peut parler indifféremment d'un (seul) paramètre k-dimensionnel $(\theta_1, \dots, \theta_k)^t$, comme de k paramètres réels (unidimensionnels) $\theta_1, \dots, \theta_k$.
3. Dans le cas paramétrique, il faut faire attention à la "véritable" dimension du paramètre. Si, par exemple, on nous dit que l'on observe n réalisations indépendantes d'une loi de Bernoulli de moyenne m et de variance σ^2 inconnues, on pourrait croire que c'est un modèle paramétrique de dimension 2 (avec deux paramètres m et σ^2). Cependant, comme $m = p$ et $\sigma^2 = p(1-p)$, ces deux paramètres sont liés. Autrement dit, le vecteur $\theta = \binom{m}{\sigma^2}$ appartient à un sous-ensemble $\Theta = \left\{ \binom{p}{p(1-p)},\ p \in [0,1] \right\}$ de $\mathbb{R}^2$ qui est une courbe (et donc un objet unidimensionnel !), et la véritable dimension du modèles est bien égale à 1 (comme déjà vu plus haut, lorsqu'on utilisait un seul paramètre p).

Finalement, on introduit la notion suivante d'identifiabilité.

Définition. *Soit un modèle statistique $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta$. On dit que le modèle est **identifiable** si l'application $\theta \longmapsto \mathbf{P}_\theta$ est injective, c'est-à-dire si $\theta_1 \neq \theta_2$ implique $\mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2}$.*

Notons que cette condition est, en quelque sorte, nécessaire pour que les problèmes statistiques puissent recevoir des réponses, et elle est donc parfaitement naturelle. En effet, pour un modèle non identifiable, même si on arrivait à reconstruire entièrement la loi $\mathbf{P}_\theta$ à partir des observations (ce qui est, d'ailleurs, normalement impossible), on aurait aucun moyen de trancher entre les différentes valeurs de θ correspondant à cette loi. Par exemple, pour le modèle $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathscr{B}\big(|\theta|\big)$, $\theta \in [-1, 1]$, même si on arrivait à déterminer d'après les observations que leur loi n'est autre que $\mathscr{B}(0{,}5)$, on aurait aucun moyen de savoir si $\theta = 0{,}5$ ou $\theta = -0{,}5$.

3.3 La notion d'une statistique

Les réponses aux problèmes statistiques (que ça soit des problèmes d'estimation ou de test d'hypothèses) doivent être données en fonction des observations. On arrive donc à la notion suivante d'une "statistique".

Définition. *Toute fonction $S = S_n = S(X_1, \dots, X_n) = S_n(X_1, \dots, X_n)$ des observations (à valeurs réelles, vectorielles ou même fonctionnelles) est dite **statistique (réelle, vectorielle** ou **à valeurs fonctionnelles)**.*

Exemples.

— $S(X_1, \dots, X_n) = a$, avec un $a \in \mathbb{R}$ fixé, est une statistique (réelle) triviale (une constante).

— $S(X_1, \dots, X_n) = (X_1, \dots, X_n)^t$ (c'est-à-dire S est la fonction identité sur $\mathbb{R}^n$) est une statistique (vectorielle de dimension n) triviale (l'échantillon lui-même).

— $S(X_1, \dots, X_n) = \overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$ est une statistique (réelle) couramment utilisée (la moyenne empirique).

— $S(X_1, \dots, X_n) = S_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ est également une statistique (réelle) couramment utilisée (la variance empirique).

— $S(X_1, \dots, X_n) = \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix}$ est une statistique (vectorielle de dimension 2) qui combine la moyenne et la variance empiriques.

— $S(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)})^t$ est une statistique (vectorielle de dimension n) un peu moins triviale que l'échantillon lui-même (la série variationnelle de l'échantillon).

— $S(X_1, \dots, X_n) = X_{(k)}$ est une statistique (réelle), dite $k^{\text{ème}}$ **statistique d'ordre**. Notamment, on utilise couramment la première et la dernière statistiques d'ordre $X_{(1)} = \min\limits_{1 \leqslant i \leqslant n} (X_i)$ et $X_{(n)} = \max\limits_{1 \leqslant i \leqslant n} (X_i)$.

— La fonction de répartition empirique est un exemple d'une statistique à valeurs fonctionnelles. Ici, $S : (X_1, \dots, X_n)^t \longmapsto \widehat{F}_n$, où $\widehat{F}_n$ est la fonction de répartition empirique définie par

$$\widehat{F}_n(x) = \frac{\text{Card}\{i \ : \ X_i \leqslant x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}, \qquad x \in \mathbb{R}.$$

Remarque. Il n'y a pas qu'une seule manière de voir une statistique !

— Premièrement, par sa définition même, une statistique S est une fonction (sur $\mathbb{R}^n$). Mais on substitue dans S les observations, c'est-à-dire soit les variables aléatoires $X_1, \dots, X_n$ (lorsqu'on l'étudie), soit leurs réalisations $x_1, \dots, x_n$ (lorsqu'on l'applique aux données réelles).

— Une fois que l'on substitue les variables aléatoires $X_1, \dots, X_n$ dans la statistique S, elle peut elle-même être vue comme une variable aléatoire, et on peut donc écrire, par exemple, $\mathbf{E}\, S$ ou $\mathbf{Var}(S)$. Autrement dit, S sous-entend dans ces écritures la variable aléatoire $S(X_1, \dots, X_n)$.

Attention toutefois, la loi de cette variable aléatoire dépend de la loi inconnue des observations (ou, dans le cas paramétrique, du paramètre inconnu). Dans le cas paramétrique, on écrira donc plutôt $\mathbf{E}_\theta\, S$ ou $\mathbf{Var}_\theta(S)$ (en utilisant l'indice θ pour souligner le fait que la loi de S dépend de θ, et l'espérance est donc calculée "sous θ"), ou encore, par exemple, $S_n \xrightarrow{\text{p.s.}} g(\theta)$ (car la loi de la statistique S_n étant dépendante de θ, sa limite presque sûr peut, elle aussi, dépendre de θ).

— Pour ce qui est de l'application aux données réelles, S sous-entendra ici $S(x_1, \ldots, x_n)$, et sera donc un simple résultat numérique (un réel, un vecteur ou même une fonction).

3.4 La Δ-méthode

Beaucoup de statistiques usuelles (comme, par exemple, la moyenne et la variance empiriques) peuvent être représentées sous la forme $S_n = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$, en utilisant des fonctions $g : \mathbb{R}^n \longrightarrow \mathbb{R}^\ell$ et $h : \mathbb{R}^\ell \longrightarrow \mathbb{R}^k$. On peut étudier le comportement asymptotique de telles statistiques grâce aux deux théorèmes suivants.

Le premier théorème montre qu'une telle statistique converge presque sûrement vers une limite déterministe.

Théorème 1. *Soit $S_n = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$, et supposons que $g(X_1)$ possède des moments d'ordre 1, et que la fonction h est continue au point $a = \mathbf{E}\, g(X_1)$. Alors*

$$S_n \xrightarrow{\text{p.s.}} h(a).$$

Preuve. D'après la LFGN appliquée aux variables aléatoires i.i.d. $g(X_1), \ldots, g(X_n)$, on a

$$\frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow{\text{p.s.}} \mathbf{E}\, g(X_1) = a$$

et, comme h est continue en a, on en déduit

$$S_n = h\left(\frac{1}{n}\sum_{i=1}^n g(X_i)\right) \xrightarrow{\text{p.s.}} h(a),$$

ce qui termine la preuve. □

Le deuxième théorème permet de préciser la qualité de l'approximation de $h(a)$ par S_n.

Théorème 2 (Δ-méthode)**.** *Soit $S_n = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$ avec $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ et $h : \mathbb{R} \longrightarrow \mathbb{R}$, et supposons que $g(X_1)$ possède des moments d'ordre 2, et que la fonction h est dérivable au point $a = \mathbf{E}\, g(X_1)$. Alors*

$$\sqrt{n}\,\big(S_n - h(a)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

avec $V = [h'(a)]^2\, \mathbf{Var}\big(g(X_1)\big)$.

Idée (non rigoureuse) de la preuve. On a

$$\begin{aligned}\sqrt{n}\,\big(S_n - h(a)\big) &= \sqrt{n}\left[h\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right) - h(a)\right] \\ &= \sqrt{n}\left[h\left(a + \frac{\sum_{i=1}^{n} g(X_i) - na}{n}\right) - h(a)\right] \\ &= \frac{h\Big(a + \frac{1}{\sqrt{n}}\,\frac{\sum_{i=1}^{n} g(X_i) - na}{\sqrt{n}}\Big) - h(a)}{\frac{1}{\sqrt{n}}} = \frac{h\big(a + \frac{\xi_n}{\sqrt{n}}\big) - h(a)}{\frac{\xi_n}{\sqrt{n}}}\,\xi_n,\end{aligned}$$

où nous avons noté $\xi_n = \frac{\sum_{i=1}^{n} g(X_i) - na}{\sqrt{n}}$. Comme $a = \mathbf{E}\, g(X_1)$, d'après la TCL, on obtient

$$\xi_n = \frac{\sum_{i=1}^{n} g(X_i) - na}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0,\, \mathbf{Var}\big(g(X_1)\big)\Big).$$

De plus, cela implique $\frac{\xi_n}{\sqrt{n}} \xrightarrow{\mathcal{L}} 0$, et donc (en admettant qu'on ait le droit de remplacer la suite déterministe tendant vers 0 par une suite aléatoire tendant en loi vers 0 dans la définition de la dérivée) on a

$$\frac{h\big(a + \frac{\xi_n}{\sqrt{n}}\big) - h(a)}{\frac{\xi_n}{\sqrt{n}}} \xrightarrow{\mathcal{L}} h'(a).$$

Par conséquent,

$$\sqrt{n}\,\big(S_n - h(a)\big) = \frac{h\big(a + \frac{\xi_n}{\sqrt{n}}\big) - h(a)}{\frac{\xi_n}{\sqrt{n}}}\,\xi_n \xrightarrow{\mathcal{L}} h'(a)\,\mathcal{N}\Big(0, \mathbf{Var}\big(g(X_1)\big)\Big) = \mathcal{N}(0, V)$$

avec $V = [h'(a)]^2\, \mathbf{Var}\big(g(X_1)\big)$. □

Notons que le Théorème 2 est valable également dans un cadre multidimensionnel (lorsque $g : \mathbb{R}^n \longrightarrow \mathbb{R}^\ell$ et $h : \mathbb{R}^\ell \longrightarrow \mathbb{R}^k$), à condition de remplacer dans l'énoncé la dérivabilité par la différentiabilité et l'expression de V (qui est maintenant une matrice de variance-covariance de dimension k) par $V = J_h(a)\,\Sigma\,\big(J_h(a)\big)^t$, où Σ est la matrice de variance-covariance du vecteur aléatoire (ℓ-dimensionnel) $g(X_1)$, et $J_h(a)$ est la matrice jacobienne de la fonction h au point $a = \mathbf{E}\, g(X_1) \in \mathbb{R}^\ell$.

Exemples.

1. **La moyenne empirique $\overline{X}_n$**

   On a bien $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = h\big(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\big)$ avec $h(x) = x$ et $g(x) = x$.

   D'après le Théorème 1, on obtient

   $$\overline{X}_n \xrightarrow{\text{p.s.}} h\big(\mathbf{E}\, g(X_1)\big) = \mathbf{E}\, X_1.$$

Autrement dit, la moyenne empirique converge vers la moyenne théorique.

D'après le Théorème 2, on obtient

$$\sqrt{n}\,\big(\overline{X}_n - \mathbf{E}\,X_1\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \mathbf{Var}(X_1)\big).$$

En effet, comme $g(X_1) = X_1$ et $h'(x) = 1$, on a $V = [h'(a)]^2\,\mathbf{Var}\big(g(X_1)\big) = \mathbf{Var}(X_1)$.

2. **La variance empirique S_n^2**

Pour pouvoir représenter la variance empirique S_n^2 sous la forme $h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$, nous sommes obligés de nous placer dans un cadre multidimensionnel. On a

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

On prend alors $g(x) = \binom{x}{x^2}$, d'où

$$\frac{1}{n}\sum_{i=1}^n g(X_i) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix},$$

et on a donc bien $S_n^2 = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$ en choisissant $h(x, y) = y - x^2$.

D'après le Théorème 1, on obtient

$$S_n^2 \xrightarrow{\text{p.s.}} h\big(\mathbf{E}\,g(X_1)\big) = h\big(\mathbf{E}\,X_1,\, \mathbf{E}\,X_1^2\big) = \mathbf{E}\,X_1^2 - (\mathbf{E}\,X_1)^2 = \mathbf{Var}(X_1) = \sigma^2.$$

Autrement dit, la variance empirique converge vers la variance théorique.

Pour appliquer la version multidimensionnelle du Théorème 2, il est préférable de modifier légèrement les fonctions g et h dans la représentation $S_n^2 = h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$. En notant $m = \mathbf{E}\,X_1$ et en utilisant la formule de König généralisée, on obtient

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - m)^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i - m\right)^2,$$

et on peut donc prendre également $g(x) = \binom{x}{(x-m)^2}$ et $h(x, y) = y - (x - m)^2$. On a alors $\mathbf{E}\,g(X_1) = \mathbf{E}\binom{X_1}{(X_1-m)^2} = \binom{m}{\sigma^2}$. Comme $J_h(x, y) = \big(-2\,(x - m),\, 1\big)$, on trouve $J_h(a) = (0, 1)$. En notant $\Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$ la matrice de variance-covariance de $g(X_1) = \binom{X_1}{(X_1-m)^2}$ et en utilisant la version multidimensionnelle du Théorème 2, on obtient

$$\sqrt{n}\,\big(S_n^2 - \sigma^2\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

avec

$$V = J_h(a)\,\Sigma\,\big(J_h(a)\big)^t = (0, 1)\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix} = (0, 1)\begin{pmatrix}\sigma_{12}\\ \sigma_{22}\end{pmatrix}$$
$$= \sigma_{22} = \mathbf{Var}\big((X_1 - m)^2\big) = \mathbf{E}\,(X_1 - m)^4 - \big(\mathbf{E}\,(X_1 - m)^2\big)^2 = \mu_4 - \sigma^4,$$

où on a noté, comme d'habitude, $\mu_4 = \mathbf{E}(X_1 - \mathbf{E}\, X_1)^4$ le moment centré d'ordre 4 de X_1.

Pour conclure, notons qu'il existe des statistiques qui ne peuvent pas être représentés sous la forme $h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$. C'est le cas, par exemple, des statistiques d'ordre $X_{(k)}$. En effet, pour les calculer, il faut comparer les X_i entre eux. Or, dans l'expression $h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$, les X_i interviennent un à un (via la fonction g, avant d'être moyennisé et transformée par la fonction h).

L'étude du comportement asymptotique de telles statiques demande donc des considérations spécifiques. On traitera, notamment, la première et la dernière statistiques d'ordre $X_{(1)}$ et $X_{(n)}$ en TD.

4 Estimation et intervalles de confiance

4.1 Estimateur et ses propriétés

Soit un modèle paramétrique $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$. On cherche à estimer (approcher) le paramètre θ.

Définition. *Toute statistique (notée généralement $\overline{\theta}$, $\widehat{\theta}$, θ^*, ... ou encore $\overline{\theta}_n$, $\widehat{\theta}_n$, θ^*_n, ...) à valeurs dans $\mathbb{R}^k$ est dite* ***estimateur*** *de θ.*

Notons que le fait que l'estimateur doit être proche de θ n'a pas été mis dans sa définition mais sera dans ses propriétés. Ceci est fait dans l'idée de considérer tous les estimateurs (les bons comme mauvais) pour pouvoir les comparer et faire le tri après. Un estimateur peut donc être complètement fantaisiste.

Remarques.

1. Parfois on cherche à estimer non pas le paramètre θ lui-même, mais une fonction $g(\theta)$ de celui-ci, avec $g : \Theta \longrightarrow \mathbb{R}^\ell$. Dans ce cas, un estimateur sera évidemment une statistique à valeurs dans $\mathbb{R}^\ell$. C'est le cas, par exemple, si on observe un échantillon de la loi normale $\mathcal{N}(m, \sigma^2)$ dont la moyenne m et la variance σ^2 sont inconnues, mais on cherche à estimer seulement la moyenne m. Ici, $\theta = \binom{m}{\sigma^2}$ et $g(x, y) = x$.
2. Ceci est possible même dans les modèles non-paramétriques. Par exemple, on peut vouloir estimer la moyenne d'une loi dont on sait seulement qu'elle en possède une. Ici la fonction g fait correspondre à chaque loi (ayant une moyenne) la moyenne de cette loi. Dans de tels cas, on parle parfois d'estimation **semi-paramétrique**.
3. Finalement, dans un modèle non paramétrique, on peut vouloir estimer le "paramètre" infini-dimensionnel lui-même. Par exemple, estimer la fonction de répartition d'une loi inconnue, ou la densité d'une loi continue inconnue. On parle alors d'estimation **non-paramétrique**.

Introduisons maintenant quelques quantités servant à mesurer la "qualité" d'un estimateur, ainsi que quelques propriétés que celui-ci pourrait avoir (pour des raisons de simplicité on se limite au cas unidimensionnel $k = 1$).

Définition. *Soit θ^*_n un estimateur de θ.*

— *On appelle* ***biais*** *de θ^*_n la fonction*

$$\mathbf{b}(\theta) = \mathbf{b}_{\theta^*_n}(\theta) = \mathbf{E}_\theta\, \theta^*_n - \theta, \qquad \theta \in \Theta.$$

— *Si $\mathbf{b}_{\theta^*_n}(\theta) = 0$ pour tout $\theta \in \Theta$, on dit que l'estimateur θ^*_n est* ***sans biais****, ou encore que θ^*_n est un* ***estimateur sans biais (ESB)****.*

— *Si $\lim\limits_{n\to+\infty} \mathbf{b}_{\theta^*_n}(\theta) = 0$, on dit que l'estimateur θ^*_n est* ***asymptotiquement sans biais****.*

Notons que $\mathbf{b}_{\theta_n^*}(\theta) = \mathbf{E}_\theta(\theta_n^* - \theta)$, et donc le biais représente l'erreur moyenne de l'estimateur θ_n^*. Cependant ce n'est pas forcément une bonne manière de mesurer la "qualité" d'un estimateur, car les erreurs positives et négatives (même grandes) peuvent se compenser lorsqu'on fait la moyenne et donner un biais petit. Ceci nous amène à la définition suivante.

Définition. *Soit θ_n^* un estimateur de θ. On appelle* ***risque quadratique*** *(ou également* ***erreur moyenne quadratique*** *(**EMQ**) ou* ***mean squared error*** *en anglais) de θ_n^* la fonction*

$$R_{\theta_n^*}(\theta) = \mathbf{E}_\theta(\theta_n^* - \theta)^2, \qquad \theta \in \Theta.$$

Plus le risque quadratique de θ_n^* est petit, plus les valeurs de cet estimateur sont proches de θ (en moyenne quadratique).

Comme montre la proposition suivante, il y deux sources d'erreur dans l'EMQ : le biais de l'estimateur et sa variance.

Proposition (Compromis biais-variance)**.** *Soit θ_n^* un estimateur de θ. Alors*

$$R_{\theta_n^*}(\theta) = \mathbf{Var}_\theta(\theta_n^*) + \big(\mathbf{b}_{\theta_n^*}(\theta)\big)^2.$$

Preuve. En utilisant la formule de König généralisée, on obtient

$$\mathbf{Var}_\theta(\theta_n^*) = \mathbf{E}_\theta(\theta_n^* - \theta)^2 - (\mathbf{E}_\theta\, \theta_n^* - \theta)^2 = R_{\theta_n^*}(\theta) - \big(\mathbf{b}_{\theta_n^*}(\theta)\big)^2,$$

et il reste juste à passer le terme $\big(\mathbf{b}_{\theta_n^*}(\theta)\big)^2$ de l'autre côté de l'égalité. □

On voit donc que le fait d'être sans biais n'est pas forcément la panacée : si en introduisant un petit biais on peut gagner beaucoup sur la variance, on peut au final avoir une EMQ plus petite.

Passons maintenant aux propriétés asymptotiques de l'estimateur.

Définition. *Soit θ_n^* un estimateur de θ.*

- *Si $\theta_n^* \xrightarrow{\mathbf{P}} \theta$, on dit que θ_n^* est* ***consistant*** *(parfois, en français, on dit juste* ***convergeant****).*
- *Si, de plus, $\theta_n^* \xrightarrow{\text{p.s.}} \theta$, on dit que θ_n^* est* ***fortement consistant****.*
- *De même, si $\theta_n^* \xrightarrow{L^p} \theta$, on dit que θ_n^* est* ***consistant en L^p*** *(ou* ***en moyenne d'ordre*** *p).*

Notons que, comme θ est déterministe, la consistance est équivalente à $\theta_n^* \xrightarrow{\mathcal{L}} \theta$, et que c'est la propriété "minimum" qu'un estimateur raisonnable devrait avoir. Notons également que les deux dernières notions sont plus fortes que la première.

Vu que, par la définition même de la convergence en L^2, la consistance en moyenne quadratique équivaut à $\mathbf{E}_\theta(\theta_n^* - \theta)^2 \to 0$, et que $\mathbf{E}_\theta(\theta_n^* - \theta)^2 = R_{\theta_n^*}(\theta)$, nous avons la proposition suivante.

Proposition. *Un estimateur θ_n^* de θ est consistant en moyenne quadratique si et seulement si $R_{\theta_n^*}(\theta) \to 0$.*

Nous en déduisons le corollaire suivant.

Corollaire. *Un estimateur θ_n^* de θ est consistant en moyenne quadratique si et seulement si il est asymptotiquement sans biais et $\mathbf{Var}_\theta(\theta_n^*) \to 0$.*

Preuve. On a

$$R_{\theta_n^*}(\theta) = \mathbf{Var}_\theta(\theta_n^*) + \big(\mathbf{b}_{\theta_n^*}(\theta)\big)^2,$$

et, comme les deux termes de droite sont positifs, le terme de gauche converge vers zéro si et seulement si les deux termes de droite le font. □

Finalement, pour quantifier la vitesse à laquelle un estimateur θ_n^* converge vers θ, nous introduisons la définition suivante.

Définition. *Soit θ_n^* un estimateur de θ.*

— *S'il existe une suite décroissante $\varphi_n \searrow 0$ et une loi $\mathcal{L} = \mathcal{L}_\theta$ non dégénérée (c'est-à-dire non concentrée en 0) telles que*

$$\varphi_n^{-1}(\theta_n^* - \theta) \to \mathcal{L}_\theta,$$

*alors φ_n est dite **vitesse de convergence** de θ_n^* et $\mathcal{L}_\theta$ est dite **loi limite** de θ_n^*.*

— *Dans le cas particulier où $\varphi_n = 1/\sqrt{n}$ et $\mathcal{L}_\theta = \mathcal{N}\big(0, V(\theta)\big)$, c'est-à-dire lorsque*

$$\sqrt{n}(\theta_n^* - \theta) \to \mathcal{N}\big(0, V(\theta)\big),$$

*on dit que θ_n^* est **asymptotiquement normal** de **variance limite** $V(\theta)$.*

L'idée de cette définition est que si θ_n^* est un estimateur consistant (et donc $\theta_n^* \xrightarrow{\mathcal{L}} \theta$ en particulier), $\varphi_n^{-1}(\theta_n^* - \theta)$ est une forme indéterminée $(+\infty \times 0)$ qui divergera si θ_n^* converge vers θ moins vite que φ_n ne le fait vers 0 (et donc φ_n^{-1} vers $+\infty$), et convergera vers 0 si θ_n^* converge vers θ plus vite que φ_n vers 0. On a donc une limite non dégénérée seulement si φ_n correspond exactement à la vitesse à laquelle θ_n^* converge vers θ.

Remarques.

1. Parfois c'est la suite $\varphi_n^{-1} \nearrow +\infty$ qui est dite vitesse de convergence (plutôt que la suite $\varphi_n \searrow 0$).

2. La vitesse et la loi limite sont définies à une constante près. Si, par exemple, φ_n et $\mathcal{L}_\theta$ conviennent, alors $2\varphi_n$ et $\mathcal{L}_\theta/2$ le font aussi. C'est donc "l'ordre" de la vitesse de convergence qui est important.

3. On peut comparer deux estimateurs de point de vue asymptotique et conclure que l'un converge plus vite que l'autre si l'ordre de la vitesse de convergence du premier est supérieur à celui du second (par exemple, $1/n$ et $1/\sqrt{n}$).

4. Si l'ordre de vitesse de convergence des deux estimateurs θ_n^* et $\overline{\theta}_n$ est le même, alors il faut tout d'abord se ramener à la même vitesse :

$$\varphi_n^{-1}(\theta_n^* - \theta) \to \xi_\theta \ (\backsim \mathcal{L}_\theta) \quad \text{et} \quad \varphi_n^{-1}(\overline{\theta}_n - \theta) \to \zeta_\theta \ (\backsim \mathcal{L}'_\theta).$$

   En supposant qu'il y a convergence des moments d'ordre 2 dans ces convergences en loi (ce qui est souvent le cas), on aura

$$\varphi_n^{-2}\,\mathbf{E}_\theta(\theta_n^* - \theta)^2 = \varphi_n^{-2} R_{\theta_n^*}(\theta) \to \mathbf{E}\,\xi_\theta^2 \quad \text{et} \quad \varphi_n^{-2}\,\mathbf{E}(\overline{\theta}_n - \theta)^2 = \varphi_n^{-2} R_{\overline{\theta}_n}(\theta) \to \mathbf{E}\,\zeta_\theta^2.$$

   On a donc, pour n grand, $R_{\theta_n^*}(\theta) \approx \varphi_n^2\,\mathbf{E}\,\xi_\theta^2$ et $R_{\overline{\theta}_n}(\theta) \approx \varphi_n^2\,\mathbf{E}\,\zeta_\theta^2$, et il est naturel de conclure que θ_n^* converge plus vite que $\overline{\theta}_n$ si $\mathbf{E}\,\xi_\theta^2 < \mathbf{E}\,\zeta_\theta^2$ (et vice versa).

5. La quantité $\mathbf{E}\,\mathcal{L}_\theta^2$ est dite **erreur moyenne quadratique limite** (**EMQL**) de θ_n^*. Comparer de point de vue asymptotique deux estimateurs ayant la même vitesse de convergence revient donc à comparer leurs EMQL. Notons également que dans le cas (particulier) des estimateurs asymptotiquement normaux, cela revient à comparer les variances limites, car $\mathbf{E}\big[\mathcal{N}\big(0, V(\theta)\big)\big]^2 = V(\theta)$.

Notons finalement que les différentes notions introduites ci-dessus peuvent se généraliser au cas où le paramètre θ est de dimension $k \neq 1$ (cette généralisation est directe pour certaines, et un peu plus subtile pour d'autres).

Exemples.

1. **Estimation de la moyenne** m

   — La moyenne empirique $\overline{X}_n$ est un estimateur sans biais de m. En effet,

$$\mathbf{E}\,\overline{X}_n = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\,X_i = \frac{1}{n}\,nm = m.$$

   — Le risque quadratique de $\overline{X}_n$ est égale à sa variance (car c'est un ESB) et est donné par

$$\begin{aligned}\mathbf{Var}\big(\overline{X}_n\big) &= \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\,\mathbf{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{Var}(X_i) = \frac{1}{n^2}\,n\sigma^2 = \frac{\sigma^2}{n}\,,\end{aligned}$$

où nous avons supposé que $\sigma^2 = \mathbf{Var}(X_1)$ existe et, dans la troisième égalité, utilisé l'indépendance des X_i.

Notons que sans hypothèses supplémentaires sur la loi des X_i (cadre semi-paramétrique), la variance σ^2 est également inconnue, et donc le risque quadratique n'est pas déterminé par m (dépend aussi de σ^2).

— Comme on l'a déjà vu, $\overline{X}_n$ est fortement consistant (et donc, en particulier, consistant) et asymptotiquement normal de variance limite σ^2, c'est-à-dire on a $\sqrt{n}\big(\overline{X}_n - m\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$. En particulier, pour n grand, on peut approcher la loi de $\overline{X}_n$ comme $\overline{X}_n \rightsquigarrow \approx \mathcal{N}(m,\, \sigma^2/n)$.

Notons également que l'estimateur $\overline{X}_n$ est consistant en moyenne quadratique, car c'est un ESB et $\mathbf{Var}\big(\overline{X}_n\big) = \frac{\sigma^2}{n} \to 0$.

— Dans le cas d'un modèle gaussien $\big(X_i \rightsquigarrow \mathcal{N}(m, \sigma^2)\big)$, la moyenne empirique $\overline{X}_n$ étant une combinaison linéaire des variables gaussiennes indépendantes, elle est elle-même gaussienne. On connaît donc sa loi exacte : $\overline{X}_n \rightsquigarrow \mathcal{N}(m,\, \sigma^2/n)$.

2. **Estimation de la variance σ^2, la moyenne m étant inconnue également**

— La variance empirique $S^2 = S_n^2 = \frac{1}{n}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$ est un estimateur biaisé de σ^2. En effet,

$$\begin{aligned}\mathbf{E}(S_n^2) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\, X_i^2 - \mathbf{E}\,\overline{X}_n^2 = \frac{1}{n}\, n\, \mathbf{E}\, X_1^2 - \mathbf{E}\,\overline{X}_n^2 \\ &= \mathbf{Var}(X_1) + (\mathbf{E}\, X_1)^2 - \Big(\mathbf{Var}\big(\overline{X}_n\big) + \big(\mathbf{E}\,\overline{X}_n\big)^2\Big) \\ &= \sigma^2 + m^2 - \left(\frac{\sigma^2}{n} + m^2\right) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\,\sigma^2,\end{aligned}$$

d'où, $\mathbf{b}_{S_n^2}(\sigma^2) = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \neq 0$.

Une autre conséquence est que $\mathbf{E}\big(\frac{n}{n-1}\, S_n^2\big) = \frac{n}{n-1}\, \mathbf{E}(S_n^2) = \frac{n}{n-1}\,\frac{n-1}{n}\,\sigma^2 = \sigma^2$, et donc $\overline{S}^2 = \overline{S}_n^2 = \frac{n}{n-1}\, S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2$ est un ESB de σ^2. On appelle $\overline{S}_n^2$ **variance empirique corrigée** (ou **dé-biaisée**).

— En supposant que les moments d'ordre 4 de X_1 existent et en notant, comme d'habitude, $\mu_4 = \mathbf{E}(X_1 - m)^4$ le moment centré d'ordre 4, on peut montrer que

$$\mathbf{Var}(S_n^2) = \frac{n-1}{n^3}\big[(n-1)\mu_4 - (n-3)\sigma^4\big].$$

Notons que $\mathbf{Var}\big(\overline{S}_n^2\big) = \mathbf{Var}\big(\frac{n}{n-1}\, S_n^2\big) = \big(\frac{n}{n-1}\big)^2\, \mathbf{Var}(S_n^2) > \mathbf{Var}(S_n^2)$. Ainsi, même si l'estimateur $\overline{S}_n^2$ gagne sur le biais, il perd sur la variance. Il n'est donc pas clair s'il est meilleur ou moins bon (au sens du risque quadratique) que $\overline{S}_n^2$.

Cette question sera discutée un peu plus en détail en TD, même si souvent les gens ont tendance à utiliser plutôt $\overline{S}_n^2$ (pour sa propriété d'être sans biais).

— Comme on l'a déjà vu, S_n^2 est un estimateur fortement consistant (et donc, en particulier, consistant) et asymptotiquement normal de variance limite $\mu_4 - \sigma^4$, c'est-à-dire on a $\sqrt{n}\big(S_n^2 - \sigma^2\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\, \mu_4 - \sigma^4)$. Il en est de même pour l'estimateur $\overline{S}_n^2$. En effet, on a

$$\overline{S}_n^2 = \frac{n}{n-1}\, S_n^2 \xrightarrow{\text{p.s.}} \sigma^2,$$

car $\frac{n}{n-1} \to 1$, et

$$\begin{aligned}\sqrt{n}\big(\overline{S}_n^2 - \sigma^2\big) &= \sqrt{n}\Big(\frac{n}{n-1}\, S_n^2 - \sigma^2\Big) = \sqrt{n}\Big(S_n^2 + \frac{1}{n-1} S_n^2 - \sigma^2\Big) \\ &= \sqrt{n}(S_n^2 - \sigma^2) + \frac{\sqrt{n}}{n-1}\, S_n^2 \xrightarrow{\mathcal{L}} \mathcal{N}(0,\, \mu_4 - \sigma^4),\end{aligned}$$

car $\frac{\sqrt{n}}{n-1} \to 0$ et $S_n^2 \xrightarrow{\text{p.s.}} \sigma^2$.

Notons également que ces deux estimateurs sont consistants en moyenne quadratique. En effet, la variance empirique S_n^2 est un estimateur asymptotiquement sans biais (car $\mathbf{b}_{S_n^2}(\sigma^2) = -\frac{\sigma^2}{n} \to 0$) et $\mathbf{Var}(S_n^2) \to 0$. De même, la variance empirique corrigée $\overline{S}_n^2$ est un ESB et $\mathbf{Var}\big(\overline{S}_n^2\big) \to 0$

— Dans le cas d'un modèle gaussien $\big(X_i \backsim \mathcal{N}(m, \sigma^2)\big)$, on peut donner la loi exacte de S_n^2 et de $\overline{S}_n^2$ par

$$\frac{nS_n^2}{\sigma^2} = \frac{(n-1)\overline{S}_n^2}{\sigma^2} = \frac{\sum_{i=1}^n \big(X_i - \overline{X}_n\big)^2}{\sigma^2} \backsim \chi^2_{n-1}.$$

Ceci est une des affirmations du **théorème de Cochran** (admise) qui, de plus, dit que $\overline{S}_n^2$ (de même que $\overline{S}_n^2$) est indépendant de $\overline{X}_n$.

3. **Estimation de la variance σ^2, la moyenne m étant connue**

On peut, bien-sûr, utiliser les estimateurs S_n^2 et $\overline{S}_n^2$ considérés précédemment (ignorant ainsi le fait qu'on connaît la moyenne m), mais on peut également utiliser l'estimateur $\frac{1}{n}\sum_{i=1}^n (X_i - m)^2$ qui sera étudié (et comparé aux deux autres) en TD.

4. **Estimation de la proportion p $(X_i \backsim \mathscr{B}(p))$**

Comme $p = \mathbf{E}\, X_1$, on peut utiliser l'estimateur $\overline{X}_n$ pour estimer p. On a donc les mêmes propriétés que dans l'exemple 1 $\big($en précisant que $\sigma^2 = p(1-p)\big)$: le fait d'être sans biais, l'expression pour le risque quadratique, diverses consistances, la normalité asymptotique, ainsi que l'approximation de la loi qui s'ensuit.

Pour la loi exacte de $\overline{X}_n$, on a dans ce cas $n\overline{X}_n = \sum_{i=1}^n X_i \backsim \mathscr{B}(n,p)$. Notons qu'en approchant, pour n grand, cette loi binomiale par une loi normale, on obtient $n\overline{X}_n \backsim \approx \mathcal{N}\big(np, np(1-p)\big)$, et on retrouve l'approximation $\overline{X}_n \backsim \approx \mathcal{N}\big(p, \frac{p(1-p)}{n}\big)$.

4.2 Méthodes de construction d'estimateurs

Méthode des moments

Soit un modèle paramétrique unidimensionnel $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}$. Pour tout $\theta \in \Theta$, on pose $m_p(\theta) = \mathbf{E}_\theta\, X_1^p$ (le moment d'ordre p des observations sous θ). On choisit p de manière que la fonction $m_p : \Theta \longrightarrow m_p(\Theta) \subset \mathbb{R}$ soit inversible. En notant m_p^{-1} la fonction réciproque, on a donc $\theta = m_p^{-1}\big(\mathbf{E}_\theta\, X_1^p\big)$. Comme on sait que $\mathbf{E}_\theta\, X_1^p$ peut être approché par $\frac{1}{n}\sum_{i=1}^n X_i^p$ (le moment empirique d'ordre p), on propose comme estimateur de θ la statistique

$$\theta_n^* = m_p^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^p\right),$$

dite **estimateur par la méthode des moments** de θ.

Si le paramètre θ est de dimension $k \neq 1$, le principe est le même, mais il faut utiliser k moments (d'ordres différents), pour que la fonction $m : \Theta \longrightarrow m(\Theta) \subset \mathbb{R}^k$ ait une chance d'être inversible.

Exemples.

1. Soit $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathscr{E}(\lambda)$, $\lambda \in \mathbb{R}_+^*$. On a

$$m_1(\lambda) = \mathbf{E}_\lambda\, X_1 = \frac{1}{\lambda} = y \quad \Longleftrightarrow \quad \lambda = \frac{1}{y} = m_1^{-1}(y).$$

   Donc la fonction réciproque de m_1 est $m_1^{-1}(y) = 1/y$, $y \in \mathbb{R}_+^*$, et l'estimateur par la méthode des moments de λ est

$$\lambda_n^* = m_1^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^1\right) = \frac{1}{\frac{1}{n}\sum_{i=1}^n X_i} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X}_n}\,.$$

   Notons que c'est un estimateur parfaitement naturel : comme $\overline{X}_n$ approche $\mathbf{E}\, X_1 = \frac{1}{\lambda}$, il est logique d'approcher λ par $\lambda_n^* = 1/\overline{X}_n$. C'est l'essence même de la méthode des moments.

2. Soit $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathcal{N}(m, \sigma^2)$, $\theta = \binom{m}{\sigma^2} \in \mathbb{R} \times \mathbb{R}_+^* \subset \mathbb{R}^2$. On a

$$m_{1,2}(\theta) = \begin{pmatrix} m_1(\theta) \\ m_2(\theta) \end{pmatrix} = \begin{pmatrix} \mathbf{E}_\theta\, X_1 \\ \mathbf{E}_\theta\, X_1^2 \end{pmatrix} = \begin{pmatrix} m \\ \sigma^2 + m^2 \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \Longleftrightarrow \begin{cases} m = x \\ \sigma^2 = y - m^2 = y - x^2 \end{cases}.$$

Donc $m_{1,2}^{-1}(x,y) = \binom{x}{y-x^2}$, et l'estimateur par la méthode des moments est

$$\theta_n^* = m_{1,2}^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^1\,,\ \frac{1}{n}\sum_{i=1}^n X_i^2\right) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \end{pmatrix} = \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix}.$$

Une fois de plus, nous trouvons un estimateur parfaitement naturel : on estime la moyenne et la variance inconnues par la moyenne et la variance empiriques.

3. Soit $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathscr{U}(-\theta, \theta)$, $\theta \in \mathbb{R}_+^*$. On a

$$m_1(\theta) = \mathbf{E}_\theta\, X_1 = \frac{\theta + (-\theta)}{2} = 0.$$

La fonction m_1, n'est donc pas inversible, et il faut utiliser un moment d'un ordre différent. On a

$$m_2(\theta) = \mathbf{E}_\theta\, X_1^2 = \mathbf{Var}_\theta(X_1) = \frac{\big(\theta - (-\theta)\big)^2}{12} = \frac{\theta^2}{3} = y \quad \Longleftrightarrow \quad \theta = \sqrt{3y} = m_2^{-1}(y).$$

Donc la fonction réciproque de m_2 est $m_2^{-1}(y) = \sqrt{3y}$, $y \in \mathbb{R}_+^*$, et l'estimateur par la méthode des moments de θ est

$$\theta_n^* = m_2^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \sqrt{\frac{3}{n}\sum_{i=1}^n X_i^2}.$$

Notons aussi que si dans le modèle initial on avait $\theta \in [0,1]$ (au lieu de $\theta \in \mathbb{R}_+^*$), la fonction m_2^{-1} serait définie sur $[0, \frac{1}{3}]$ (et non pas sur tout $\mathbb{R}_+^*$). Par contre, rien ne garantit que le moment empirique $\frac{1}{n}\sum_{i=1}^n X_i^2 \in [0, \frac{1}{3}]$. Donc, lorsqu'on propose l'estimateur θ_n^* dans ce cas, on utilise, en fait, l'extension "naturelle" $m_2^{-1}(y) = \sqrt{3y}$, $y \in \mathbb{R}_+^*$, de la fonction $m_2^{-1}(y) = \sqrt{3y}$, $y \in [0,1]$.

L'avantage de la méthode des moments est qu'elle est assez simple à mettre en place. De plus, elle produit toujours des estimateurs de la forme $h\big(\frac{1}{n}\sum_{i=1}^n g(X_i)\big)$ (avec $h = m_p^{-1}$ et $g(x) = x^p$), ce qui garantit le comportement asymptotique suivant.

— En supposant que la fonction m_p^{-1} est continue et en utilisant le Théorème 1 de la Δ-méthode, on a

$$\theta_n^* \xrightarrow{\text{p.s.}} h\big(\mathbf{E}_\theta\, g(X_1)\big) = m_p^{-1}\big(\mathbf{E}_\theta\, X_1^p\big) = m_p^{-1}\big(m_p(\theta)\big) = \theta,$$

c'est-à-dire θ_n^* est fortement consistant.

— En supposant que la fonction m_p^{-1} est dérivable (et que les moments d'ordre $2p$ de X_1 existent), d'après le Théorème 2 de la Δ-méthode, θ_n^* est asymptotiquement normal de variance limite

$$\left[(m_p^{-1})'\big(m_p(\theta)\big)\right]^2 \mathbf{Var}\big(X_1^p\big) = \frac{\mathbf{Var}\big(X_1^p\big)}{\left[m'(\theta)\right]^2},$$

où la dernière égalité est due à un théorème d'analyse sur la dérivée de la fonction réciproque.

Notons que ce comportement asymptotique reste valable (avec une expression adaptée pour la matrice de variance-covariance limite) lorsque le paramètre θ est de dimension $k \neq 1$.

Par contre, la méthode ne dit pas comment choisir p (est-ce que pour certains choix de p, on a de meilleurs estimateurs que pour d'autres ?). De plus, les estimateurs par la méthode des moments convergent toujours à la vitesse $1/\sqrt{n}$ (car ils sont forcément asymptotiquement normaux). Or, comme on le verra en TD, dans certains modèles il existe des estimateurs qui convergent plus vite (et qui ne peuvent donc pas être trouvés par cette méthode).

Méthode du maximum de vraisemblance

Pour bien comprendre l'idée sous-jacente à la méthode du maximum de vraisemblance, supposons d'abord que l'on observe une (seule) réalisation d'une variable aléatoire X continue, dont la densité est soit f_{θ_1}, soit f_{θ_2}, ces deux densités étant représentées respectivement en rouge et en bleu dans la Figure 4.1 ci-dessous.

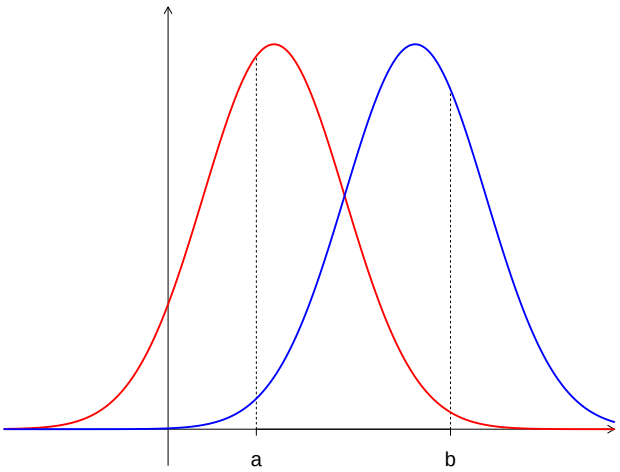

FIGURE 4.1 – Deux densités et la vraisemblance

Si la valeur observée de X est a, même si elle peut bien-sûr provenir de chacune de ces deux densités, la densité f_{θ_1} est bien plus à même d'avoir produit cette valeur que la densité f_{θ_2}.

Autrement dit, θ_1 est plus "vraisemblable" que θ_2. De même, si la valeur observée de X est b, on peut dire que θ_2 est plus "vraisemblable" que θ_1. Ainsi, plus la densité évaluée en l'observation X est grande, plus il est vraisemblable que cette densité ait produit cette observation.

Soit maintenant un modèle paramétrique $X_1, \ldots, X_n$ i.i.d., $X_i \rightsquigarrow \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$. On suppose que pour tout $\theta \in \Theta$, la loi $\mathbf{P}_\theta$ des X_i possède une densité f_θ.

Notons que les f_θ peuvent être des densités usuelles (si les X_i sont des v.a. continues), comme des "densités" discrètes (si les X_i sont des v.a. discrètes), mais elles doivent être de même nature pour toutes les valeurs de θ.

Notons également que dans le cas continu, il faut, comme d'habitude, choisir les versions "les plus continues possibles" des densités, en essayant, en plus, d'utiliser la même convention pour le choix du support.

Ainsi, la densité du vecteur d'observations $\vec{X} = (X_1, \ldots, X_n)^t$ est

$$\boldsymbol{f}_\theta(\vec{x}) = \boldsymbol{f}_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

et l'idée est que la valeur de θ pour laquelle la densité au point $\vec{X} = (X_1, \ldots, X_n)^t$ est la plus grande est la plus à même d'avoir produit ces observations. Elle est donc la plus "vraisemblable".

Définition. *On appelle* ***(fonction de) vraisemblance*** *(**likelihood (function)** en anglais) la fonction de $t \in \Theta$ définie par*

$$L(t) = \boldsymbol{f}_t\big(\vec{X}\big) = \prod_{i=1}^{n} f_t(X_i).$$

On appelle ***estimateur du maximum de vraisemblance (EMV)*** *de θ l'estimateur*

$$\widehat{\theta}_n = \operatorname*{argsup}_{t \in \Theta} L(t).$$

La notation argsup utilisée dans cette définition est une généralisation de argmax. On rappelle que si une fonction $f(t)$, $t \in T$, admet un maximum, alors toute valeur t^* telle que $f(t^*) = \max_{t \in T} f(t)$ est dite $\operatorname{argmax}_{t \in T} f(t)$. Plus généralement, toute valeur t^* telle qu'au voisinage de t^* il existe des valeurs de t avec $f(t)$ aussi proche de $\sup_{t \in T} f(t)$ qu'on veut (dans le cas des fonctions càdlàg sur $\mathbb{R}$, cela équivaut à $\max\{f(t^*-), f(t^*+)\} = \sup_{t \in T} f(t)$) est dite $\operatorname{argsup}_{t \in T} f(t)$. Notons que l'argsup (comme, d'ailleurs, l'argmax) peut ne pas être unique.

Pour mieux comprendre, on a tracé quelques exemples dans la Figure 4.2 ci-dessous (la fonction f est définie sur $T = [\alpha, \beta]$ à chaque fois). Dans la Figure 4.2 (a), on a $\max_{t \in T} f(t) = m$ et $\operatorname{argsup}_{t \in T} f(t) = \operatorname{argmax}_{t \in T} f(t) = a$. Dans la Figure 4.2 (b), on a $\max_{t \in T} f(t) = m$, mais

$\operatorname{argmax}_{t\in T} f(t)$ n'est pas unique : toute valeur entre a et b est un $\operatorname{argmax}_{t\in T} f(t)$ (également un $\operatorname{argsup}_{t\in T} f(t)$, d'ailleurs). Dans la Figure 4.2 (c), $\max_{t\in T} f(t)$ et $\operatorname{argmax}_{t\in T} f(t)$ n'existent pas, mais on a $\sup_{t\in T} f(t) = m$ et $\operatorname{argsup}_{t\in T} f(t) = a$. Finalement, dans la Figure 4.2 (d), même si $\sup_{t\in T} f(t) = +\infty$, on a $\operatorname{argsup}_{t\in T} f(t) = a$.

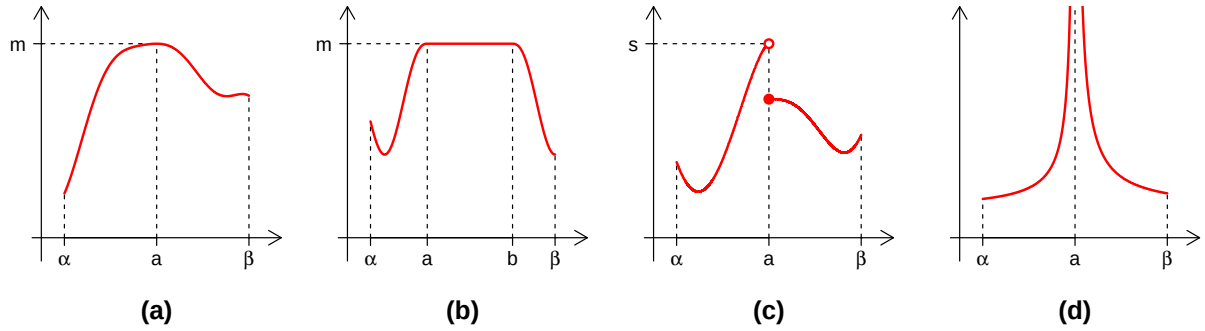

FIGURE 4.2 – Différents cas d'argmax et d'argsup

Remarques.

1. Si le support de $\mathbf{P}_\theta$ ne dépend pas de θ, alors $f_t(X_i) > 0$ pour tout $t \in \Theta$, et donc $L(t) > 0$. Dans ce cas, on peut plutôt maximiser

$$\ell(t) = \ln\big(L(t)\big) = \ln\left(\prod_{i=1}^{n} f_t(X_i)\right) = \sum_{i=1}^{n} \ln\big(f_t(X_i)\big),$$

   car on a bien $\widehat{\theta}_n = \operatorname{argsup}_{t\in\Theta} L(t) = \operatorname{argsup}_{t\in\Theta} \ell(t)$.

2. Souvent, on ne change pas la variable θ en t dans la définition de la vraisemblance, et on écrit $\widehat{\theta}_n = \operatorname{argsup}_{\theta\in\Theta} L(\theta)$. Dans ce cas, il faut garder en tête que le θ dans cette écriture est une variable muette (c'est-à-dire le résultat ne dépend pas de cette variable), et non pas la vraie valeur du paramètre (comme on verra, c'est surtout important lorsque le support de la loi des observations dépend de θ).

Exemples.

1. Soit $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathscr{E}(\lambda)$, $\lambda \in \mathbb{R}_+^*$. On a $f_\lambda(x) = \lambda\, e^{-\lambda x}\, \mathbb{1}_{\{x>0\}}$, d'où

$$L(t) = \prod_{i=1}^{n} f_t(X_i) = \prod_{i=1}^{n} \big[t\, e^{-tX_i}\, \mathbb{1}_{\{X_i>0\}}\big] = t^n\, e^{-t\sum_{i=1}^n X_i},$$

   où on a pu enlever l'indicatrice car $X_i > 0$ est toujours vrai (vu que $\mathscr{S}(X_i) = \mathbb{R}_+^*$ pour toutes les valeurs de λ). Par conséquent,

$$\ell(t) = \ln\big(L(t)\big) = \ln\big(t^n\, e^{-t\sum_{i=1}^n X_i}\big) = n\ln(t) - t\sum_{i=1}^{n} X_i.$$

Notons, qu'on aurait pu directement écrire

$$\ell(t) = \sum_{i=1}^{n} \ln\big(f_t(X_i)\big) = \sum_{i=1}^{n} \ln\big(t\, e^{-tX_i}\, \mathbb{1}_{\{X_i>0\}}\big) = \sum_{i=1}^{n} \big[\ln(t) - tX_i\big] = n\ln(t) - t\sum_{i=1}^{n} X_i.$$

Maximisons maintenant la fonction ℓ. On a

$$\ell'(t) = \frac{n}{t} - \sum_{i=1}^{n} X_i \gtreqless 0 \quad \Longleftrightarrow \quad \frac{n}{t} \gtreqless \sum_{i=1}^{n} X_i \quad \Longleftrightarrow \quad t \lesseqgtr \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}_n}\,.$$

Ainsi la fonction ℓ est (strictement) croissante sur l'intervalle $\big]0,\, 1/\overline{X}_n\big[$ et est (strictement) décroissante sur l'intervalle $\big]1/\overline{X}_n,\, +\infty\big[$, et donc l'estimateur du maximum de vraisemblance de λ est $\widehat{\lambda}_n = \operatorname{argsup}_{t\in\mathbb{R}_+^*} \ell(t) = 1/\overline{X}_n$ (le même que par la méthode des moments).

2. Soit $X_1, \ldots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \mathbb{R}_+^*$, où $\mathbf{P}_\theta$ est la loi de probabilité ayant pour densité $f_\theta(x) = \frac{2x}{\theta^2}\, \mathbb{1}_{\{0<x<\theta\}}$ (*cf.* la Figure 4.3 ci-dessous).

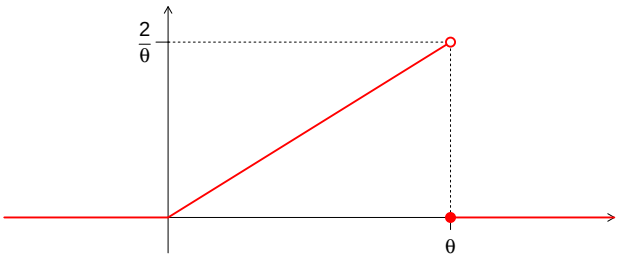

FIGURE 4.3 – Densité de l'exemple 2

On a

$$L(t) = \prod_{i=1}^{n} f_t(X_i) = \prod_{i=1}^{n}\left[\frac{2X_i}{t^2}\, \mathbb{1}_{\{0<X_i<t\}}\right] = \prod_{i=1}^{n}\left[\frac{2X_i}{t^2}\, \mathbb{1}_{\{X_i<t\}}\right] = \frac{2^n \prod_{i=1}^{n} X_i}{t^{2n}} \prod_{i=1}^{n} \mathbb{1}_{\{X_i<t\}}$$
$$= \frac{C}{t^{2n}}\, \mathbb{1}_{\{\bigcap_{i=1}^{n}\{X_i<t\}\}} = \frac{C}{t^{2n}}\, \mathbb{1}_{\{X_{(n)}<t\}} = \frac{C}{t^{2n}}\, \mathbb{1}_{\{t>X_{(n)}\}},$$

où nous avons noté $C = 2^n \prod_{i=1}^{n} X_i$.

On voit donc que la fonction L est nulle sur $\big]0, X_{(n)}\big]$, et qu'elle est strictement positive et décroissante sur $\big]X_{(n)}, +\infty\big[$. Par conséquent, l'estimateur du maximum

de vraisemblance de θ est $\widehat{\theta}_n = \operatorname{argsup}_{t\in\mathbb{R}_+^*} L(t) = X_{(n)}$ (*cf.* la Figure 4.4 ci-dessous).

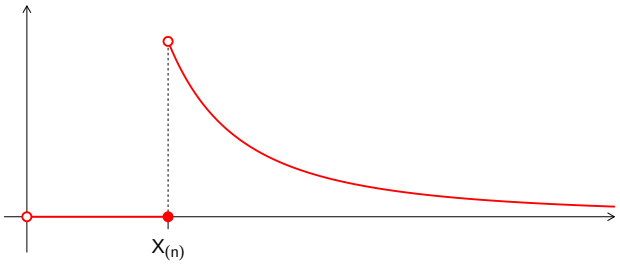

FIGURE 4.4 – Vraisemblance de l'exemple 2

Notons que si on n'avait pas changé la variable θ en t dans la définition de la vraisemblance, on aurait pu faire le calcul erroné

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \left[\frac{2X_i}{\theta^2}\,\mathbb{1}_{\{0<X_i<\theta\}}\right] = \prod_{i=1}^{n}\left[\frac{2X_i}{\theta^2}\right] = \frac{2^n \prod_{i=1}^n X_i}{\theta^{2n}} = \frac{C}{\theta^{2n}}$$

en croyant que dans les indicatrices $X_i < \theta$ est toujours vrai (ce qui n'est pas le cas, car ici θ est juste une variable muette qu'on aurait pu, et même dû, appeler autrement ; et comme $\mathscr{S}(X_i) =]0,\theta[$, on sait bien que $X_i < \theta$, mais cette fois-ci θ est la vraie valeur du paramètre), et conclure faussement que $\widehat{\theta}_n = \operatorname{argsup}_{\theta\in\mathbb{R}_+^*} L(\theta) = 0$.

Méthode de Bayes

Soit un modèle paramétrique $X_1,\dots,X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$. Imaginons que le parmètre inconnu θ soit une variable aléatoire suivant une certaine loi (dite **loi *a priori***) ayant pour support Θ et pour densité une fonction q (dite **densité *a priori***) strictement positive et continue (sur Θ). Notons qu'une telle densité peut être vue comme une manière de coder quelles valeurs de θ sont *a priori* plus ou moins réalistes (en attribuant des valeurs plus ou moins élevées à la densité). Dans ce cas, la fonction

$$\boldsymbol{f}_t(\vec{x}) = \boldsymbol{f}_t(x_1,\dots,x_n) = \prod_{i=1}^{n} f_t(x_i)$$

doit être vu comme la densité conditionnelle de $\vec{X} = (X_1,\dots,X_n)^t$ sachant que $\theta = t$. Par conséquent, la densité jointe de θ et de $\vec{X}$ est $f(t,\vec{x}) = \boldsymbol{f}_t(\vec{x})\,q(t)$, la densité marginale de $\vec{X}$ est $\int_\Theta \boldsymbol{f}_t(\vec{x})\,q(t)\,\mathrm{d}t$, et la densité conditionnelle de θ sachant que $\vec{X} = \vec{x}$ est

$$q_\theta\big(t \,\big|\, \vec{X} = \vec{x}\big) = \frac{\boldsymbol{f}_t(\vec{x})\,q(t)}{\int_\Theta \boldsymbol{f}_t(\vec{x})\,q(t)\,\mathrm{d}t}\,.$$

Donc, la loi conditionnelle de θ sachant $\vec{X}$ (dite **loi *a posteriori***) a pour densité

$$q_\theta\big(t \,\big|\, \vec{X}\big) = \frac{\boldsymbol{f}_t\big(\vec{X}\big)}{\int_\Theta \boldsymbol{f}_t\big(\vec{X}\big)\, q(t)\, \mathrm{d}t} = \frac{L(t)\, q(t)}{\int_\Theta L(t)\, q(t)\, \mathrm{d}t}$$

(dite **densité *a posteriori***), et l'espérance de cette loi est

$$\mathbf{E}\big(\theta \,\big|\, \vec{X}\big) = \int_\Theta t\, q_\theta\big(t \,\big|\, \vec{X}\big)\, \mathrm{d}t = \frac{\int_\Theta t\, L(t)\, q(t)\, \mathrm{d}t}{\int_\Theta L(t)\, q(t)\, \mathrm{d}t}\,.$$

Définition. *On appelle* ***estimateur bayésien (avec la densité a priori*** *q) de θ la statistique*

$$\widetilde{\theta}_n = \frac{\int_\Theta t\, L(t)\, q(t)\, \mathrm{d}t}{\int_\Theta L(t)\, q(t)\, \mathrm{d}t}\,.$$

Remarques.

1. On sait que $\widetilde{\theta}_n = \mathbf{E}\big(\theta \,\big|\, \vec{X}\big) = \varphi_*\big(\vec{X}\big)$, où φ_* est la fonction qui minimise $\mathbf{E}\big(\varphi\big(\vec{X}\big) - \theta\big)^2$ parmi toutes les fonctions $\varphi = \varphi\big(\vec{X}\big)$ ou, autrement dit, parmi tous les estimateurs (pour simplifier, on suppose ici que le paramètre θ est de dimension $k = 1$). Or,

$$\begin{aligned}\mathbf{E}\big(\varphi\big(\vec{X}\big) - \theta\big)^2 &= \mathbf{E}\Big[\mathbf{E}\Big(\big(\varphi\big(\vec{X}\big) - \theta\big)^2 \,\Big|\, \theta\Big)\Big] = \int_\Theta \mathbf{E}\Big(\big(\varphi\big(\vec{X}\big) - \theta\big)^2 \,\Big|\, \theta = t\Big)\, q(t)\, \mathrm{d}t \\ &= \int_\Theta \mathbf{E}_t\big(\varphi\big(\vec{X}\big) - \theta\big)^2\, q(t)\, \mathrm{d}t = \int_\Theta R_{\varphi(\vec{X})}(t)\, q(t)\, \mathrm{d}t.\end{aligned}$$

   Ainsi, l'estimateur bayésien a le plus petit risque quadratique "moyennisé" (par rapport à θ en utilisant une densité q sur Θ) possible (parmi tous les estimateurs).

2. On peut, plus généralement, utiliser l'estimateur

$$\widetilde{\theta}_n = \frac{\int_\Theta t\, L(t)\, q(t)\, \mathrm{d}t}{\int_\Theta L(t)\, q(t)\, \mathrm{d}t}$$

   quand q est une fonction strictement positive et continue quelconque (qui n'est pas nécessairement une densité de probabilité). Si $\int_\Theta q(t)\, \mathrm{d}t = A < +\infty$, en divisant le numérateur et le dénominateur par A, on voit que $\widetilde{\theta}_n$ est l'estimateur bayésien avec la densité *a priori* $q(t)/A$, $t \in \Theta$. Si, par contre, $\int_\Theta q(t)\, \mathrm{d}t = +\infty$, on appelle q **pseudo-densité *a priori*** et $\widetilde{\theta}_n$ **estimateur pseudo-bayésien**.

Exemple.

Soit $X_1, \ldots, X_n$ i.i.d., $X_i \hookrightarrow \mathscr{E}(\lambda)$, $\lambda \in \mathbb{R}_+^*$. Comme on l'a déjà vu, la vraisemblance est $L(t) = t^n\, e^{-t\sum_{i=1}^n X_i}$, et donc les estimateurs (pseudo-)bayésiens sont donnés par

$$\widetilde{\lambda}_n = \frac{\int_0^{+\infty} t^{n+1}\, e^{-t\sum_{i=1}^n X_i}\, q(t)\, \mathrm{d}t}{\int_0^{+\infty} t^n\, e^{-t\sum_{i=1}^n X_i}\, q(t)\, \mathrm{d}t}\,,$$

ou q est la (pseudo-)densité *a priori*.

Si, par exemple, il n'y a pas de préférence entre différentes valeurs de λ, il est naturel de choisir la pseudo-densité $q(t) = 1$, $t \in \mathbb{R}_+^*$. En utilisant le changement de variable $t = y \,/\, \sum_{i=1}^n X_i$, on obtient alors

$$\begin{aligned}
\widetilde{\lambda}_n &= \frac{\int_0^{+\infty} t^{n+1}\, e^{-t\sum_{i=1}^n X_i}\, \mathrm{d}t}{\int_0^{+\infty} t^{n}\, e^{-t\sum_{i=1}^n X_i}\, \mathrm{d}t} = \frac{\int_0^{+\infty} \big(\frac{y}{\sum_{i=1}^n X_i}\big)^{n+1}\, e^{-y}\, \cancel{\frac{1}{\sum_{i=1}^n X_i}}\, \mathrm{d}y}{\int_0^{+\infty} \big(\frac{y}{\sum_{i=1}^n X_i}\big)^{n}\, e^{-y}\, \cancel{\frac{1}{\sum_{i=1}^n X_i}}\, \mathrm{d}y} \\
&= \frac{1}{\sum_{i=1}^n X_i}\, \frac{\int_0^{+\infty} y^{n+1}\, e^{-y}\, \mathrm{d}y}{\int_0^{+\infty} y^{n}\, e^{-y}\, \mathrm{d}y} = \frac{1}{\sum_{i=1}^n X_i}\, \frac{\Gamma(n+2)}{\Gamma(n+1)} \\
&= \frac{1}{\sum_{i=1}^n X_i}\, \frac{(n+1)!}{n!} = \frac{n+1}{\sum_{i=1}^n X_i}\,.
\end{aligned}$$

Notons finalement, qu'on adopte ici le point de vue dit **fréquentiste**, et on voit l'approche bayésienne juste comme une méthode de construction d'estimateurs. Mais il y a toute une branche de statistique (dite **bayésienne**), dans laquelle on considère que θ est "vraiment" aléatoire, et que les observations ne font qu'apporter des précisions sur sa loi (en passant de la loi *a priori* à la loi *a posteriori*).

4.3 Quelques notions de comparaison d'estimateurs

Soit un modèle paramétrique unidimensionnel $X_1, \ldots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}$. Nous avons déjà discuté de la comparaison d'estimateurs par leurs vitesses de convergence, c'est-à-dire de manière asymptotique. Pour les comparer d'une manière non-asymptotique (pour n fixe), on peut utiliser leurs risques quadratiques.

Définition. *Soit θ_n^* et $\overline{\theta}_n$ deux estimateurs de θ. On dit que θ_n^* est **meilleur** que $\overline{\theta}_n$ si*

$$R_{\theta_n^*}(\theta) \leqslant R_{\overline{\theta}_n}(\theta), \qquad \forall \theta \in \Theta.$$

*Si, de plus, il existe un $\theta \in \Theta$ tel que cette inégalité est stricte, on dit que θ_n^* est **strictement meilleur** (on dit même parfois, dans ce cas, que $\overline{\theta}_n$ est **inadmissible**).*

Notons que les risques quadratiques étant des fonctions (de θ), parmi deux estimateurs il n'y a pas forcément un qui est meilleur que l'autre. Il se peut, en effet, très bien que pour certaines valeurs de θ ça soit une de ces fonctions qui est plus petite, et pour d'autres valeurs de θ ça soit l'autre. Et comme justement on ne connaît pas θ...

Pire encore, il n'est pas difficile de voir que **le meilleur** estimateur (parmi tous) n'existe pas, sauf le cas très particulier (et inintéressant !) des modèles statistiques **dégénérés**, c'est-à-dire des modèles où l'on peut estimer le paramètre inconnu sans erreur (autrement dit, le déterminer entièrement à partir des observations). Voici un exemple d'un modèle

statistique dégénérée : $X \hookrightarrow \mathscr{U}(\theta, \theta + 1/2)$, $\theta \in \mathbb{Z}$. Une seule réalisation x de X suffit pour déterminer θ (par exemple, si on observe $x = 3{,}14\ldots$, on est sûr que $\theta = 3$).

En effet, supposons que θ_n^* est meilleur que tout autre estimateur. Alors, pour tout $a \in \Theta$ arbitrairement fixé, il doit être meilleur que l'estimateur "stupide" déterministe $\overline{\theta}_n = a$. Or, $R_{\overline{\theta}_n}(\theta) = (a - \theta)^2$ et donc, en particulier, $R_{\overline{\theta}_n}(a) = 0$. D'où, comme $R_{\theta_n^*}(a) \leqslant R_{\overline{\theta}_n}(a)$, on obtient $R_{\theta_n^*}(a) = 0$. Mais comme a était fixé arbitrairement, ceci est vrai pour tout $a \in \Theta$. En particulier, pour $a = \theta$ (la vraie valeur inconnue), cela s'écrit

$$R_{\theta_n^*}(\theta) = \mathbf{E}_\theta(\theta_n^* - \theta)^2 = 0,$$

d'où $\theta_n^* \overset{\text{p.s.}}{=} \theta$ (égalité presque sûre sous la loi $\mathbf{P}_\theta$). Autrement dit, θ_n^* estime θ sans erreur, ce qui veut dire que le modèle statistique est dégénéré.

Ainsi, nous venons de voir que (à part dans des modèles statistiques dégénérés), il est impossible de minimiser (sur l'ensemble de tous les estimateurs) le risque quadratique (pour tous les $\theta \in \Theta$ à la fois). Pour palier à cela il existe plusieurs approches.

Dans le cadre bayésien, par exemple, on peut chercher à minimiser (toujours sur l'ensemble de tous les estimateurs) le risque quadratique "moyennisé" (par rapport à θ en utilisant une densité q sur Θ), défini pour un estimateur $\overline{\theta}_n$ par

$$\int_\Theta R_{\overline{\theta}_n}(t)\, q(t)\, \mathrm{d}t.$$

On sait déjà que l'estimateur que réalise ce minimum est l'estimateur bayésien $\widetilde{\theta}_n$ vu dans la section précédente.

Nous considérons ici une autre approche qui consiste à essayer de minimiser le risque quadratique pour tous les $\theta \in \Theta$ à la fois, mais uniquement parmi les estimateurs dans une certaine famille des estimateurs "raisonnables" (excluant notamment les estimateurs déterministes), et non pas sur l'ensemble de tous les estimateurs. On peut notamment essayer de chercher le meilleur estimateur parmi les ESB. Le risque quadratique étant égale à la variance pour les ESB, un tel estimateur (s'il existe) est dit **estimateur sans biais uniformément de variance minimal** (**ESBUVM**). Comme on verra, l'idée n'est pas absurde, du moins pour certains modèles statistiques dits réguliers.

Définition. *Un modèle statistique $X_1, \ldots, X_n$ i.i.d., $X_i \hookrightarrow \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}$, ayant une vraisemblance $L(t)$, $t \in \Theta$, est dit* ***régulier*** *si :*

- *Le support de $\mathbf{P}_\theta$ ne dépend pas de θ.*
- *La vraisemblance L est dérivable (et donc la log-vraisemblance ℓ l'est aussi, avec $\ell'(t) = L'(t)/L(t)$).*
- *La quantité*

  $$I_n(\theta) = \mathbf{Var}_\theta\big(\ell'(\theta)\big) = \mathbf{E}_\theta\big(\ell'(\theta)\big)^2$$

  (on admet ici que $\mathbf{E}_\theta\, \ell'(\theta) = 0$) dite ***information de Fisher (apportée par $X_1, \ldots, X_n$ sur θ)*** *existe, est strictement positive et est continue par rapport à θ.*

Remarques.

1. La vraisemblance $L(t) = \prod_{i=1}^n f_t(X_i)$ est dérivable si et seulement si $f_t(x)$ est dérivable par rapport à t. Cette dérivée sera notée $\dot{f}_t(t)$ (pour se différencier de $f'_t(x)$, qui est la dérivée par rapport à x).
2. L'information de Fisher peut également être calculée comme $I_n(\theta) = nI_1(\theta)$ avec

$$I_1(\theta) = \mathbf{Var}_\theta\left(\frac{\dot{f}_\theta(X_1)}{f_\theta(X_1)}\right) = \mathbf{E}_\theta\left(\frac{\dot{f}_\theta(X_1)}{f_\theta(X_1)}\right)^2 = \int\limits_{\mathscr{S}(X_1)} \frac{\big(\dot{f}_\theta(x)\big)^2}{f_\theta(x)}\,\mathrm{d}x.$$

   En effet,

$$\begin{aligned}\mathbf{Var}_\theta\big(\ell'(\theta)\big) &= \mathbf{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{i=1}^n \ln f_\theta(X_i)\right) = \mathbf{Var}_\theta\left(\sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_\theta(X_i)\right)\\ &= \sum_{i=1}^n \mathbf{Var}_\theta\left(\frac{\dot{f}_\theta(X_i)}{f_\theta(X_i)}\right) = nI_1(\theta).\end{aligned}$$

3. Si la vraisemblance L est deux fois dérivable, on peut également calculer l'information de Fisher comme $I_n(\theta) = -\,\mathbf{E}_\theta\big(\ell''(\theta)\big)$ (admis).
4. On peut définir la notion du modèle régulier en imposant des conditions légèrement plus faibles, mais nous avons préféré ici d'imposer des conditions plus simples (plutôt que plus générales).

On admet le théorème suivant.

Théorème. *Soit un modèle statistique régulier. Alors la borne inférieure suivante (dite* ***borne de Cramér-Rao*** *ou, parfois dans les sources françaises,* ***borne de Fréchet-Darmois-Cramér-Rao*** *ou* ***borne FDCR****) a lieu : pour tout ESB $\overline{\theta}_n$, on a*

$$\mathbf{Var}_\theta\big(\overline{\theta}_n\big) \geqslant \frac{1}{I_n(\theta)} = \frac{1}{nI_1(\theta)}\,, \qquad \forall\theta\in\Theta.$$

Cette borne montre que (au moins) dans les modèles réguliers le minimum de la variance (et donc du risque quadratique) sur l'ensemble des ESB n'est pas nul, et donc il n'est peut-être pas impossible dans ce cas de trouver un ESB atteignant ce minimum.

Définition. *Soit un modèle statistique régulier. Un ESB θ_n^* est dit* ***efficace*** *(****au sens de Cramér-Rao****) s'il atteigne la borne de Cramér-Rao, c'est-à-dire s'il vérifie*

$$\mathbf{Var}_\theta\big(\theta_n^*\big) = \frac{1}{I_n(\theta)} = \frac{1}{nI_1(\theta)}\,, \qquad \forall\theta\in\Theta.$$

Notons qu'un estimateur efficace est évidemment un ESBUVM, mais que l'implication inverse n'est pas vraie : un ESBUVM peut ne pas être efficace au sens de Cramér-Rao (même dans un modèle statistique régulier). Dans ce cas, ce n'est pas un défaut de l'estimateur, mais plutôt celui de la borne, cette dernière donnant une minoration trop basse de la variance.

Le théorème suivant (admis) montre que (dans un modèle régulier) si un estimateur efficace existe, on peut le trouver par la méthode du maximum de vraisemblance.

Théorème. *Soit un modèle statistique régulier.*

1. *S'il existe un estimateur efficace, c'est l'EMV.*
2. *Dans tous les cas, si on suppose de plus que le modèle est identifiable, l'EMV $\widehat{\theta}_n$ est consistant, est asymptotiquement normal de variance limite $1/I_1(\theta)$:*

$$\sqrt{n}\big(\widehat{\theta}_n - \theta\big) \xrightarrow{\mathcal{L}} \xi \backsim \mathcal{N}\big(0,\, 1/I_1(\theta)\big),$$

et il y a convergence des moments de tout ordre $p > 0$ dans cette dernière convergence en loi :

$$n^{p/2}\ \mathbf{E}_\theta\big(\widehat{\theta}_n - \theta\big)^p \to \mathbf{E}\,\xi^p.$$

De plus, ces propriétés asymptotiques sont également valables pour les estimateurs bayésiens (avec des densités a priori strictement positives et continues quelconques).

En plus de décrire le comportement asymptotique de l'EMV et des estimateurs bayésiens (dans les modèles réguliers), la deuxième partie du théorème dit que même dans les cas où l'estimateur efficace n'existe pas, ces estimateurs sont proches de l'être pour n grand. En effet, pour l'EMV, la convergence du moment d'ordre 1 nous dit que $\sqrt{n}\big(\mathbf{E}_\theta\,\widehat{\theta}_n - \theta\big) \to 0$, d'où $\mathbf{E}_\theta\,\widehat{\theta}_n \to \theta$, c'est-à-dire l'EMV $\widehat{\theta}_n$ est asymptotiquement sans biais (et donc il est "presque" sans biais pour n grand). De plus, d'après la convergence du moment d'ordre 2, nous avons $n\,\mathbf{Var}_\theta\big(\widehat{\theta}_n\big) \to 1/I_1(\theta)$, d'où, pour n grand, $\mathbf{Var}_\theta\big(\widehat{\theta}_n\big) \approx \frac{1}{nI_1(\theta)} = \frac{1}{I_n(\theta)}$ (et donc l'EMV $\widehat{\theta}_n$ atteigne "presque" la borne de Cramér-Rao). Ainsi, pour n grand, l'EMV $\widehat{\theta}_n$ est "presque" efficace. Les mêmes considérations sont, bien-sûr, valables pour les estimateurs bayésiens.

Notons également que, comme on verra en TD, dans les modèles non-réguliers l'EMV et les estimateurs bayésiens pourront converger plus vite qu'à la vitesse $1/\sqrt{n}$.

Notons finalement que, comme on a pu le voir, la question de la comparaison d'estimateurs est un problème très complexe, et il n'y a pas une seule "bonne" manière de comparer les estimateurs. Citons ici rapidement une autre manière de comparer les estimateurs, qui consiste à comparer les valeurs maximales (sur Θ) des fonctions représentant leurs risques quadratiques. L'idée sous-jacente est que, comme on ne connaît pas θ, un estimateur doit être "performant" pour toutes les valeurs possibles de θ.

4.4 Intervalles de confiance

Définition. *Soit un modèle paramétrique unidimensionnel $X_1, \dots, X_n$ i.i.d., $X_i \backsim \mathbf{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}$, et soit deux statistiques $A = A(\vec{X})$ et $B = B(\vec{X})$ (où $\vec{X} = (X_1, \dots, X_n)^t$) telles que $A(\vec{X}) \leqslant B(\vec{X})$. L'intervalle $I(\vec{X}) = \big[A(\vec{X}), B(\vec{X})\big]$ est dit* ***intervalle de confiance (IC)*** *de* ***niveau (de confiance)*** *β (ou de (**niveau de**) **risque** $\alpha = 1 - \beta$) pour θ si*

$$\mathbf{P}_\theta\Big(\theta \in \big[A(\vec{X}), B(\vec{X})\big]\Big) = \beta.$$

Cette définition peut paraître simple et similaire à celle de l'intervalle de fluctuation, mais conceptuellement c'est très différent. D'ailleurs, le concept de l'IC est très souvent mal compris (et c'est pire encore pour les tests d'hypothèses qu'on verra plus tard). En 2015, par exemple, un grand scandale à éclaté car une revue de psychologie assez réputée est allée jusqu'à bannir l'utilisation des IC et des tests d'hypothèses dans les articles qu'elle publie, en prétendant que ces méthodes statistiques était fausses ! Il est donc très important de bien comprendre et de correctement interpréter le concept de l'IC, qu'on va essayer d'expliquer un peu plus dans les remarques suivantes.

Remarques.

1. Il est tout d'abord très important de comprendre que θ est déterministe (bien qu'inconnu), et c'est l'intervalle $\big[A(\vec{X}), B(\vec{X})\big]$ qui est aléatoire. Pour souligner ça, on aurait peut-être dû plutôt écrire

$$\mathbf{P}_\theta\Big(\big[A(\vec{X}), B(\vec{X})\big] \ni \theta\Big) = \beta.$$

2. Une fois qu'on remplace $\vec{X}$ par sa réalisation $\vec{x} = (x_1, \dots, x_n)^t$ (les valeurs numériques observées), il est FAUX de dire que l'intervalle (déterministe) $I(\vec{x})$ contient la vraie valeur (également déterministe) de θ avec probabilité β ! C'est précisément cette mauvaise interprétation de l'IC qui a conduit au scandale mentionné ci-dessus. En effet, une fois qu'on remplace $\vec{X}$ par $\vec{x}$, il n'y a plus rien d'aléatoire, et la probabilité que θ appartient à $I(\vec{x})$ ne peut donc pas valoir β (elle vaut, d'ailleurs, 1 ou 0, car c'est juste vrai ou faux).
3. La seule bonne interprétation est donc la suivante. Lors de l'application d'un intervalle de confiance $I(\vec{X})$ à un (seul) jeu de données, on ne peut pas vraiment conclure quant à l'appartenance ou non de θ à l'intervalle $I(\vec{x})$ obtenu. Par contre, si on utilise l'intervalle $I(\vec{X})$ régulièrement (sur des jeux de données différents), l'intervalle obtenu contiendra θ dans $(100\beta)\%$ des cas environ (ce qui explique que β est le niveau de confiance). Autrement dit, β est la "fiabilité" de la "recette" qu'on utilise, par contre lors d'une utilisation particulière ça peut marcher comme ne pas marcher.

La méthode générale de construction des IC est la suivante. On considère un estimateur θ_n^* de θ (où, plus généralement, une fonction quelconque des observations ET de θ), dont on

connaît la loi (où, du moins, une approximation de la loi pour n grand). On construit un intervalle de fluctuation $[a, b]$ de niveau β pour cette loi. Comme cette loi dépend de θ, on a, en fait, $a = a(\theta)$ et $b = b(\theta)$. On obtient donc

$$\mathbf{P}_\theta\big(a(\theta) \leqslant \theta_n^* \leqslant b(\theta)\big) = \beta.$$

Il reste donc à résoudre ces inégalités par rapport à θ pour transformer $a(\theta) \leqslant \theta_n^* \leqslant b(\theta)$ en $A(\theta_n^*) \leqslant \theta \leqslant B(\theta_n^*)$, et conclure que $\mathbf{P}_\theta\big(A(\theta_n^*) \leqslant \theta \leqslant B(\theta_n^*)\big) = \beta$ ou, autrement dit, que l'intervalle $\big[A(\theta_n^*),\, B(\theta_n^*)\big]$ est un IC de niveau β pour θ. Bien sûr, si on utilise une loi approchée (pour n grand) pour θ_n^*, le niveau de l'IC ne sera pas égale à β exactement, mais d'une manière approximative (pour n grand). On dit parfois dans ce cas que l'IC est de **niveau asymptotique** β.

Exemples.

1. **IC pour la moyenne m, la variance σ^2 étant connue**

   Dans le modèle gaussien $\big(X_i \backsim \mathcal{N}(m, \sigma^2)\big)$, on a $\overline{X}_n \backsim \mathcal{N}\big(m, \frac{\sigma^2}{n}\big)$, ce qui donne $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$, d'où $\mathbf{P}\Big(-u_{1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leqslant u_{1-\frac{\alpha}{2}}\Big) = \beta$. Ici et dans la suite, on utilise les notations de la brochure **Lois de Probabilité et Tables Statistiques** ; en particulier, u_ε désigne le quantile d'ordre ε de la loi normale centrée réduite. Finalement, comme

   $$\begin{aligned} -u_{1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leqslant u_{1-\frac{\alpha}{2}} \quad &\Longleftrightarrow \quad -\frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}} \leqslant \overline{X}_n - m \leqslant \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}} \\ &\Longleftrightarrow \quad \overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}} \leqslant m \leqslant \overline{X}_n + \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}, \end{aligned}$$

   on obtient que $\mathbf{P}\Big(\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}} \leqslant m \leqslant \overline{X}_n + \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\Big) = \beta$ ou, autrement dit, que l'intervalle $\Big[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\,,\ \overline{X}_n + \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\Big]$ est un IC de niveau β pour m.

   Si le modèle n'est pas gaussien, on peut utiliser le même IC pour n grand (typiquement pour $n \geqslant 30$), car $\sqrt{n}\big(\overline{X}_n - m\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, et donc $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

2. **IC pour la moyenne m, la variance σ^2 étant inconnue également**

   Dans le cas gaussien, on remplace la variance inconnue σ^2 par son estimateur $\overline{S}_n^2$ dans la construction précédente. On a :

   $$T_n = \frac{\overline{X}_n - m}{\sqrt{\overline{S}_n^2/n}} = \frac{\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\overline{S}_n^2}{\sigma^2}\Big/(n-1)}} \backsim \mathcal{T}_{n-1}$$

   car, d'après le théorème de Cochran, $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0,1)$ et $\frac{(n-1)\overline{S}_n^2}{\sigma^2} \backsim \chi^2_{n-1}$, ces deux

v.a. étant indépendantes. Par conséquent, $\mathbf{P}\Big(-t_{n-1,1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n - m}{\sqrt{\overline{S}_n^2/n}} \leqslant t_{n-1,1-\frac{\alpha}{2}}\Big) = \beta$, et donc $\Big[\overline{X}_n - \sqrt{\frac{\overline{S}_n^2}{n}}\, t_{n-1,1-\frac{\alpha}{2}}\ ,\ \overline{X}_n + \sqrt{\frac{\overline{S}_n^2}{n}}\, t_{n-1,1-\frac{\alpha}{2}}\Big]$ est un IC de niveau β pour m.

Si le modèle n'est pas gaussien, l'IC $\Big[\overline{X}_n - \sqrt{\frac{\overline{S}_n^2}{n}}\, u_{1-\frac{\alpha}{2}}\ ,\ \overline{X}_n + \sqrt{\frac{\overline{S}_n^2}{n}}\, u_{1-\frac{\alpha}{2}}\Big]$ peut être utilisé pour n grand. En effet,

$$T_n = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{\overline{S}_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

car $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ (puisque $\overline{X}_n$ est un estimateur de m qui est asymptotiquement normal de variance limite σ^2) et $\sqrt{\frac{\sigma^2}{\overline{S}_n^2}} \xrightarrow{\text{p.s.}} 1$ (puisque $\overline{S}_n^2$ est un estimateur fortement consistant de σ^2).

3. **IC pour la variance σ^2 dans le cas gaussien**

D'après le théorème de Cochran, nous avons $\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{(n-1)\overline{S}_n^2}{\sigma^2} \backsim \chi^2_{n-1}$, d'où $\mathbf{P}\Big(u_{n-1,\frac{\alpha}{2}} \leqslant \frac{nS_n^2}{\sigma^2} \leqslant u_{n-1,1-\frac{\alpha}{2}}\Big) = \beta$, et donc $\Big[\frac{nS_n^2}{u_{n-1,1-\frac{\alpha}{2}}},\ \frac{nS_n^2}{u_{n-1,\frac{\alpha}{2}}}\Big]$ est un IC de niveau β pour σ^2.

Notons que si la moyenne m est connue, on peut construire un autre IC à partir de la statistique $\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} \backsim \chi^2_n$.

4. **IC pour la proportion p ($X_i \backsim \mathscr{B}(p)$)**

Nous avons $\sqrt{n}\big(\overline{X}_n - p\big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, p(1-p)\big)$, ou encore $\frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Par conséquent,

$$\mathbf{P}_p\left(-u_{1-\frac{\alpha}{2}} \leqslant \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leqslant u_{1-\frac{\alpha}{2}}\right) \approx \beta$$

pour n grand. Pour construire un IC, on pourrait, bien sûr, résoudre ces inégalités par rapport à p, mais les expressions ainsi obtenues sont très complexes et ne seront pas présentées ici. On peut également utiliser un IC plus simple (mais peut-être moins précis) $\Big[\overline{X}_n - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\ ,\ \overline{X}_n + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\Big]$ obtenu en remplaçant p par son estimateur $\overline{X}_n$ dans le dénominateur de S_n, ce qui est justifié par le fait que $\overline{X}_n \xrightarrow{\text{p.s.}} p$, et par conséquent $\frac{\overline{X}_n - p}{\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}} = \frac{\overline{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sqrt{\frac{p(1-p)}{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

4.5 Quelques notions d'estimation non-paramétrique

Estimation de la fonction de répartition

La fonction de répartition F inconnue des observations X_i peut être estimée à l'aide de la **fonction de répartition empirique**

$$\widehat{F}_n(x) = \frac{\text{Card}\{i \,:\, X_i \leqslant x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant x\}}, \qquad x \in \mathbb{R}.$$

En effet, soit $x \in \mathbb{R}$ quelconque. Comme $\mathbf{P}\{X_i \leqslant x\} = F(x)$, on a $\mathbb{1}_{\{X_i \leqslant x\}} \backsim \mathscr{B}\big(F(x)\big)$. Donc, en utilisant les Théorèmes 1 et 2 de la Δ-méthode (avec $g(t) = \mathbb{1}_{\{t \leqslant x\}}$ et $h(y) = y$), on obtient

$$\widehat{F}_n(x) \xrightarrow{\text{p.s.}} \mathbf{E}\,\mathbb{1}_{\{X_1 \leqslant x\}} = F(x)$$

et

$$\sqrt{n}\big(\widehat{F}_n(x) - F(x)\big) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V),$$

où $V = \mathbf{Var}(\mathbb{1}_{\{X_1 \leqslant x\}}) = F(x)\big(1 - F(x)\big)$. Autrement dit, $\widehat{F}_n(x)$ est un estimateur de $F(x)$ fortement consistant et asymptotiquement normal de variance limite $F(x)\big(1 - F(x)\big)$.

Cependant, même si les considérations précédentes sont valables pour tout $x \in \mathbb{R}$, elles ne montrent pas que la fonction $\widehat{F}_n$ approche "globalement" la fonction F. La réponse est apportée par le théorème suivant (admis).

Théorème (Glivenko-Cantelli)**.** *On a :*

$$\sup_{x \in \mathbb{R}} \big|\widehat{F}_n(x) - F(x)\big| \xrightarrow{\text{p.s.}} 0.$$

Estimation de la densité

La densité f inconnue des observations X_i (que l'on suppose donc d'être des v.a. continues) peut être estimée à l'aide de l'**histogramme** dont l'air total est égale à 1, c'est-à-dire celui pour lequel le rectangle correspondant à une classe d'effectif n_j a l'aire égale à $\frac{n_j}{n}$, et donc la hauteur égale à $\frac{n_j}{nh}$, où h est la largeur des classes (pour simplifier, on considère que toutes les classes ont la même largeur). On peut montrer que si $h = h_n \to 0$ de telle sorte que $nh_n \to +\infty$, l'histogramme est un estimateur consistant de f. Autrement dit, si h est suffisamment (mais pas trop !) petit, l'histogramme approchera bien la densité inconnue.

Une autre possibilité est de passer par l'estimateur $\widehat{F}_n$ de la fonction de répartition F des X_i. Cependant, on ne peut pas directement dériver la fonction de répartition empirique $\widehat{F}_n$, car celle-ci est discontinue (la loi empirique étant discrète). En s'inspirant du fait que

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h},$$

on peut proposer l'estimateur

$$f_n^*(x) = \frac{\widehat{F}(x+h) - \widehat{F}(x-h)}{2h} = \frac{1}{2h}\,\frac{\mathrm{Card}\{i\,:\,x-h<X_i\leqslant x+h\}}{n}$$
$$= \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}_{\{x-h<X_i\leqslant x+h\}} = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\,\mathbb{1}_{\left\{-1<\frac{X_i-x}{h}\leqslant 1\right\}},$$

dit estimateur **par fenêtre mobile**. Notons que la deuxième (ou la troisième) représentation de $f_n^*(x)$ montre qu'il est assez proche de l'histogramme, sauf qu'ici les classes fixes (de largeur h) sont remplacées par une fenêtre (mobile) centrée en x (de largeur $2h$).

La dernière représentation de $f_n^*(x)$ suggère un estimateur plus général, dit estimateur **par la méthode de noyau** (ou **à noyau**, ou **de Parzen**), défini par

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

où le **noyau** K est une fonction positive vérifiant $\int_{-\infty}^{\infty} K(x)\,\mathrm{d}x = 1$. La fenêtre mobile correspond au noyau uniforme $K(x) = \frac{1}{2}\,\mathbb{1}_{\{-1<x\leqslant 1\}}$, mais pour les densités lisses il est préférable d'utiliser d'autres noyaux, comme le noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}$ ou le noyau d'Epanechnikov $K(x) = \frac{3}{4}\,(1-x^2)\,\mathbb{1}_{\{-1<x\leqslant 1\}}$.

Tout comme pour l'histogramme, on peut montrer que si $h = h_n \to 0$ de telle sorte que $nh_n \to +\infty$, l'estimateur $\widehat{f}_n$ (et donc, en particulier, l'estimateur f_n^*) est un estimateur consistant de f.

Notons finalement que pour tous ces estimateurs, le choix de h est crucial. On recommande généralement l'utilisation de $h = \sqrt{S_n^2}\,n^{-1/5}$, ce qui permet de garantir une erreur moyenne quadratique d'ordre $n^{-4/5}$, c'est-à-dire d'avoir $\mathbf{E}\int_{-\infty}^{+\infty}\big(\widehat{f}_n(x) - f(x)\big)^2\,\mathrm{d}x \leqslant \frac{C}{n^{4/5}}$.

5 Tests d'hypothèses

5.1 Principes généraux de tests d'hypothèses

Un **test** (**statistique**) est un procédé permettant de trancher, en se basant sur un échantillon, entre deux hypothèses (contradictoires) sur la loi dont est issue cet échantillon.

Les rôles des deux hypothèses ne sont pas symétriques : on teste une hypothèse dite **nulle** (notée, en général, H_0) contre une hypothèse dite **alternative** (notée, en général H_1). L'hypothèse nulle H_0 est en général **conservative** (admise par défaut). Par exemple, si un laboratoire pharmaceutique veux prouver l'efficacité d'un nouveau médicament qu'elle a développé, alors H_0 : « *le médicament n'a pas d'effet* » et H_1 : « *le médicament a l'effet escompté* ». Quelque part, on cherche des "preuves" (en faveur) de H_1 (et tant qu'on n'en a pas trouvé, on continue à considérer que c'est H_0 qui est vraie), comme c'est un peu le cas dans un tribunal, où un suspect est présumé *innocent* (H_0), tant qu'on n'a pas trouvé des preuves montrant qu'il est *coupable* (H_1).

L'hypothèse est dite **simple** si la loi des observations est entièrement déterminée sous cette hypothèse, sinon elle est dite **composée** (ou **composite**) et peut être **paramétrique** ou **non-paramétrique**. Reprenons l'exemple du laboratoire pharmaceutique. Imaginons, par exemple, que le médicament est censé augmenter le nombre moyen m de globules rouges dans le sang, qu'on suppose distribué d'après une loi normale ($X_i \hookrightarrow \mathcal{N}(m, \sigma^2)$ avec σ^2 connu). Supposons également que sans le traitement, ce nombre moyen est égale à un certain $m_0 \in \mathbb{R}$ connu. On a donc l'hypothèse H_0 : « $m = m_0$ », qui est une hypothèse simple, et l'hypothèse H_1 : « $m > m_0$ », qui est une hypothèse composée paramétrique. Notons que si σ^2 était inconnu également, les deux hypothèses seraient composées paramétriques.

Un test est généralement basé sur une statistique T dite **statistique de test** ou **variable de décision**. La loi de cette statistique doit être entièrement déterminée sous H_0 (notons qu'elle l'est forcément si H_0 est simple, ce qui est souvent le cas). On peut donc construire un intervalle de fluctuation I de niveau $1 - \alpha$ pour T, où $\alpha \in [0, 1]$ est un niveau de risque (généralement plutôt petit, 0,05 par exemple) fixée d'avance.

Si la statistique T (calculée sur les observations) prends une valeur en dehors de cet intervalle de fluctuation I (c'est-à-dire si $T \in W = \mathbb{R} \setminus I$), alors de deux choses l'une :

- soit un évènement peu probable (de probabilité α) s'est produit,
- soit l'hypothèse H_0 n'est pas vraie.

En quelque sorte, dans ce cas, ce qu'on a observé "contredit" l'hypothèse H_0, et il est donc naturel de conclure en faveur de l'hypothèse H_1 (on dit parfois que le test est **significatif**).

Par contre, si $T \in I$, rien ne "contredit" l'hypothèse H_0, et donc on la garde (par défaut).

Ainsi, nous avons abouti à la règle de décision suivante :

- si $T \in W$, on **rejette** (H_0 **en faveur de** H_1),
- si $T \in I$, on **accepte** (H_0).

Notons que le terme “accepter” (largement repandu) est assez mal choisi, car on n’a rien trouvé qui prouve H_0 (on n’a juste rien pu trouver qui prouve le contre). Il crée donc plus confusion qu’autre chose. On devrait plutôt dire (comme certains auteurs, malheureusement, peu nombreux) “on **conserve** (H_0)” ou “on **ne peut pas rejeter** (H_0 **en faveur de** H_1)”.

Notons également que W est dit **zone** (ou **région**) **de rejet** (ou **critique**), et qui I est parfois dit **zone** (ou **région**) **d’acceptation**.

Notons finalement qu’on peut parfois entendre parler de rejeter ou d’accepter H_1 (n’oubliez pas que par défaut on parle de H_0), ce qui ne fait que rajouter à la confusion.

On va voir sur l’exemple suivant que lors de la construction d’un test, il faut faire très attention à la forme de l’intervalle de fluctuation I. Celle-ci ne doit pas être choisie (comme on le fait d’habitude) en fonction de la forme de la loi sous H_0, mais de telle sorte que les valeurs qui sont plus “plausibles” sous H_1 que sous H_0 (car si on rejette, cela doit être en faveur de H_1 !) se retrouvent en dehors de l’intervalle de fluctuation I (c’est-à-dire dans la zone de rejet W).

Exemple.

Soit $X_1, \dots, X_n$ i.i.d., $X_i \hookrightarrow \mathcal{N}(m, \sigma^2)$ avec σ^2 connu et $m \in \mathbb{R}$ inconnu, et soit $m_0 \in \mathbb{R}$ une valeur fixe.

Comme statistique de test on choisit $T = \overline{X}_n$. On sait déjà que la loi de $\overline{X}_n$ est $\mathcal{N}(m, \sigma^2/n)$. En particulier, sous H_0, on a $\overline{X}_n \hookrightarrow \mathcal{N}(m_0, \sigma^2/n)$.

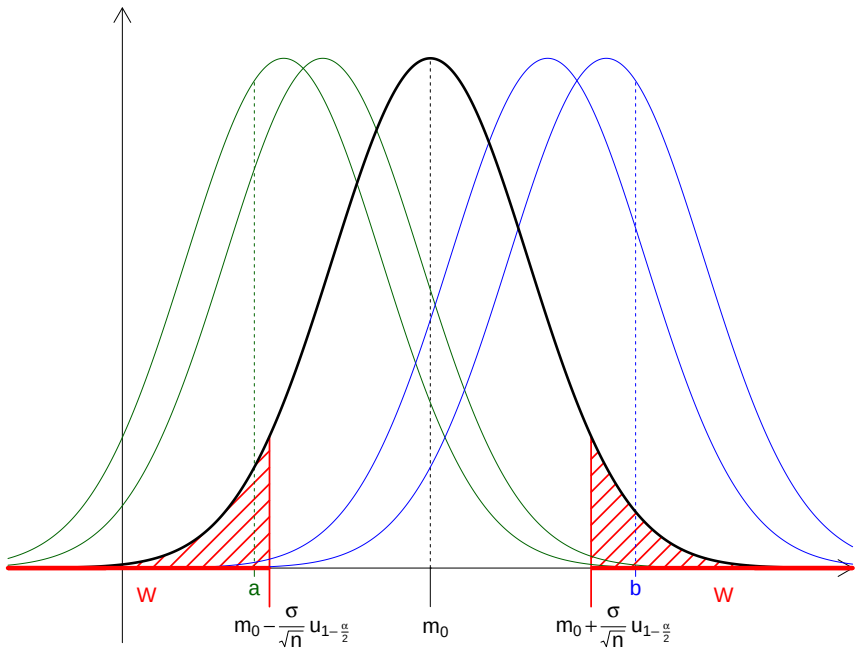

FIGURE 5.1 – Test bilatéral symétrique

On veut d'abord tester H_0 : « $m = m_0$ » contre H_1 : « $m \neq m_0$ » (test paramétrique bilatéral). Dans ce cas, l'intervalle de fluctuation qu'il faut utiliser est l'intervalle de fluctuation bilatéral symétrique $I = \left[m_0 - \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\,,\ m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\right]$. En effet, comme on peut le voir dans la Figure 5.1 ci-dessus, les valeurs dans $W = \mathbb{R} \setminus I$ (en rouge) sont peu "plausibles" (trop "extrêmes") pour H_0 (densité noire épaisse) et, surtout, elles le sont plus sous H_1 : les valeurs trop à gauche (comme a) sont plus "plausibles" sous les densités vertes (qui font bien partie de H_1), et celles trop à droite (comme b), sous les densités bleues (qui font aussi partie de H_1). Il est donc tout à fait logique de rejeter H_0 en faveur de H_1 si la valeur de $\overline{X}_n$ tombe dans la région de rejet $W = \left]-\infty\,,\ m_0 - \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\right[\cup \left]m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\frac{\alpha}{2}}\,,\ +\infty\right[$.

Par contre, si on veut tester H_0 : « $m = m_0$ » contre H_1 : « $m > m_0$ » (test paramétrique unilatéral à droite), l'intervalle de fluctuation précédent ne convient pas. En effet, les valeurs trop à gauche (comme a) sont toujours peu "plausibles" sous H_0, mais ils le sont encore moins sous H_1, car celui-ci ne contient plus les densités vertes, mais seulement les densités bleues. Par contre, les valeurs qui sont trop à droite (comme b), sont toujours plus "plausibles" sous H_1 que sous H_0. Il faut donc utiliser l'intervalle de fluctuation unilatéral à gauche $I = \left]-\infty\,,\ m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}\right]$ et rejeter si $\overline{X}_n > m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}$ (*cf.* la Figure 5.2 ci-dessous). Une autre manière de l'expliquer est de remarquer que sous H_1, la statistique de test $\overline{X}_n$ a tendance de prendre des valeurs plus grandes que sous H_0 (puisque les X_i le font), et il est donc logique de rejeter les grandes valeurs de $\overline{X}_n$ (et pas les petites).

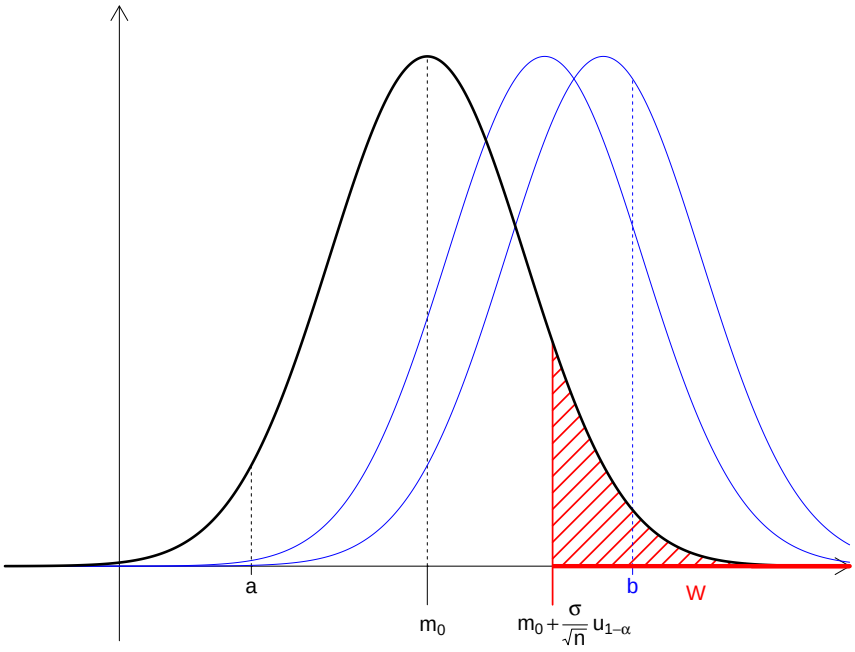

FIGURE 5.2 – Test unilatéral à droite

De même, pour tester H_0 : « $m = m_0$ » contre H_1 : « $m < m_0$ » (test paramétrique unilatéral à gauche), il faut prendre l'intervalle de fluctuation unilatéral à droite $I = \left]m_0 - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}\,,\ +\infty\right[$ et rejeter si $\overline{X}_n < m_0 - \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}$.

5.2 Deux espèces d'erreur

On considère un test basé sur la statistique T et dont la région de rejet est W. Ce test peut se tromper de deux manières :

- **erreur de 1ère espèce** : on rejette H_0, tandis qu'en réalité elle est vraie ;
- **erreur de 2nde espèce** : on conserve H_0, tandis qu'en réalité c'est H_1 qui est vraie.

La probabilité d'erreur de 1ère espèce est la probabilité sous H_0 de décider de rejeter, c'est-à-dire $\mathbf{P}_{H_0}(T \in W)$. Cette probabilité est bien définie, car la loi de T est entièrement déterminée sous H_0. Elle est également dite **risque** (**de 1ère espèce**), **niveau** (**de risque**), **seuil** (**de risque**), ou encore **taille** du test.

Notons que nous contrôlons cette probabilité d'erreur de 1ère espèce : en choisissant pour W le complémentaire d'un intervalle de fluctuation de niveau $1-\alpha$, nous garantissons qu'elle est égale à α (le test T est de niveau α).

Il faut bien comprendre ce que représente le niveau α du test. On ne peut pas le voir sur une (seule) utilisation du test, par contre α correspond bien à la proportion de "vraies" H_0 qu'on rejette lors de son utilisation régulière. Par exemple, si dans un laboratoire pharmaceutique le département "Statistique" utilise systématiquement un test de niveau 0,05 pour tester l'efficacité des médicaments élaborés par le département "Recherche et Développement" avant de les mettre sur le marché, environ 5% des médicaments inefficaces seront mis sur le marché quand-même.

La probabilité d'erreur de 2nde espèce est en général plus compliquée. L'hypothèse H_1 étant la plupart de temps composée, la loi de T sous H_1 n'est pas forcément déterminée, et donc $\mathbf{P}_{H_1}(T \notin W)$ n'est pas bien définie. Il faudrait préciser "où" on se trouve exactement "dans H_1" pour pouvoir la calculer.

Pour mieux le comprendre, reprenons, par exemple, le test de H_0 : « $m = m_0$ » contre H_1 : « $m > m_0$ » dans le cas gaussien ($X_i \hookrightarrow \mathcal{N}(m, \sigma^2)$ avec σ^2 connu) ayant par variable de décision $T = \overline{X}_n$ et pour zone de rejet $W = \left]m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}\,,\ +\infty\right[$ (autrement dit, on rejette si $\overline{X}_n > m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}$). On ne peut pas calculer la probabilité de $\{T \notin W\}$ sous H_1 (c'est-à-dire si on sait uniquement que $m > m_0$), mais seulement si on connaît la valeur de m ou, autrement dit, "sous m" $\big($auquel cas, on sait que $T = \overline{X}_n \hookrightarrow \mathcal{N}(m,\, \sigma^2/n)\big)$:

$$\begin{aligned}\mathbf{P}_m(T \notin W) &= \mathbf{P}_m\left(T < m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha}\right) = \mathbf{P}_m\left(\frac{T-m}{\sigma/\sqrt{n}} < \frac{m_0 + \frac{\sigma}{\sqrt{n}}\, u_{1-\alpha} - m}{\sigma/\sqrt{n}}\right) \\ &= \mathbf{P}\left(\mathcal{N}(0,1) < u_{1-\alpha} - \frac{m-m_0}{\sigma/\sqrt{n}}\right) = \Phi\left(u_{1-\alpha} - \frac{m-m_0}{\sigma/\sqrt{n}}\right) = \alpha_2(m),\end{aligned}$$

où Φ est la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$. Ainsi, le risque de 2$^{\text{nde}}$ espèce α_2 est une fonction $\big($sur $]m_0,+\infty[\big)$ de m.

Notons que dans cet exemple (et, en fait, c'est typique pour un test) $\lim\limits_{m\searrow m_0} \alpha_2(m) = 1-\alpha$ (ce qui est beaucoup trop grand !) et $\lim\limits_{m\to+\infty} \alpha_2(m) = 0$.

Notons aussi que la fonction $\beta = 1-\alpha_2$ $\big($toujours sur $]m_0,+\infty[\big)$ est dite **puissance** du test, et que $\lim\limits_{m\searrow m_0} \beta(m) = \alpha$ et $\lim\limits_{m\to+\infty} \beta(m) = 1$.

Notons finalement que pour $m > m_0$ fixé, dans l'exemple ci-dessus on a $\lim\limits_{n\to+\infty} \alpha_2(m) = 0$ $\Big($ou encore $\lim\limits_{n\to+\infty} \beta(m) = 1\Big)$. On dit dans ce cas que le test est **consistant**.

Pour illustrer toutes ces convergences, nous avons tracé dans la Figure 5.3 ci-dessous les puissances du test pour différentes valeurs de n.

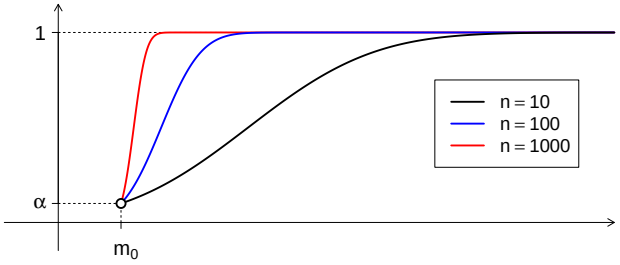

FIGURE 5.3 – Puissances pour différentes valeurs de n

On voit bien sur ce graphique que le test n'arrive pas à bien détecter les valeurs de m "trop proches" de m_0 $\Big(\lim\limits_{m\searrow m_0} \beta(m) = \alpha\Big)$, et que ça marche beaucoup mieux pour les valeurs de m qui sont plus "éloignées" $\Big(\lim\limits_{m\to+\infty} \beta(m) = 1\Big)$. De plus, plus n est grand, moins il y a des valeurs qui sont "trop proches" $\Big(\lim\limits_{n\to+\infty} \beta(m) = 1\Big)$.

Quand on a le choix entre plusieurs tests de niveau α, il est évidemment préférable de choisir celui qui est plus puissant (autrement dit, celui dont la puissance est plus grande ou, d'une manière équivalente, celui dont le risque de 2$^{\text{nde}}$ espèce est plus petit). Malheureusement, un peu comme dans les problèmes d'estimation, le test le plus puissant n'existe que rarement (par exemple, dans les problèmes de test avec deux hypothèses simples et dans certains problèmes de test paramétriques unilatéraux), d'où la multitude des tests proposés dans certains problèmes de test (surtout dans les problèmes non-paramétriques).

Finalement, en supposant que la région de rejet $W = W_\alpha$ varie continûment avec α, pour une valeur observée t de la statistique de test T on peut calculer le **niveau critique** α_*, au dessus duquel t bascule dans W_α (pour $\alpha = \alpha_*$, la valeur t se trouvera ainsi pile sur le bord de la région W_{α_*}). Ce niveau critique, dit également p**-value** (et parfois p**-valeur** en

français), est ce que calcule la plupart des logiciels de statistique. Quand on connaît α_*, on sait qu'on rejettera H_0 si $\alpha > \alpha_*$, et qu'on le conservera si $\alpha < \alpha_*$.

En quelque sorte, la p-value est la probabilité sous H_0 d'observer la valeur t de T qu'on a effectivement observée (ou, pour être un peu plus précis, cette valeur ou une valeur encore plus "extrême").

Ainsi, si la p-value est petite, soit quelque chose de peu probable s'est produit, soit H_0 n'est pas vraie. Ce que l'on a observé contredit donc, d'une certaine manière, l'hypothèse H_0, ce qui nous amène à la rejeter (et conclure H_1).

Par contre, si la p-value n'est pas trop petite, on n'a rien vu qui contredit l'hypothèse H_0, et donc on la conserve (on la garde par défaut).

Notons également que la p-value est une statistique (car elle ne dépend que de la valeur observée t de la statistique T). De plus, en supposant toujours que la région de rejet varie continûment avec α, cette statistique suit sous H_0 la loi uniforme sur $[0, 1]$. En effet, nous avons $\mathbf{P}_{H_0}(\alpha_* < \alpha) = \mathbf{P}_{H_0}(T \in W_\alpha) = \alpha$ pour tout $\alpha \in [0, 1]$.

Notons enfin qu'en aucun cas, la p-value est la probabilité de H_0 d'être vraie ni, d'ailleurs, celle de H_1 d'être fausse : ces hypothèses étant simplement soit vraies, soit fausses, il n'y a pas lieu de parler de probabilités ici (un peu comme pour les intervalles de confiance). Cette mauvaise interprétation (malheureusement bien repandue) est à l'origine de l'interdiction (et du scandale qui s'en est suivi, dont on a déjà parlé dans le chapitre précédent) des intervalles de confiance et des tests statistiques par une revue de psychologie réputée.

5.3 Quelques tests usuels

Test sur la moyenne m, la variance σ^2 étant connue

On considère d'abord le cas du modèle gaussien $\big(X_i \backsim \mathcal{N}(m, \sigma^2)\big)$ qu'on a déjà traité plus haut, mais qu'on présente ici d'une manière légèrement différente (bien qu'équivalente).

L'hypothèse nulle est :

$$H_0 : \text{« } m = m_0 \text{ »},$$

où $m_0 \in \mathbb{R}$ est une valeur fixée.

La statistique de test est :

$$T = \frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}}\,.$$

Sa loi sous H_0, comme on l'a déjà vu, est :

$$\frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}} \backsim \mathcal{N}(0, 1).$$

La région critique est donc :

— $W = \left]-\infty, -u_{1-\frac{\alpha}{2}}\right[\cup \left]u_{1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $m \neq m_0$ » ;

— $W = \left]u_{1-\alpha}, +\infty\right[$, si H_1 : « $m > m_0$ » (car T augmente sous H_1) ;

— $W = \left]-\infty, -u_{1-\alpha}\right[$, si H_1 : « $m < m_0$ » (car T diminue sous H_1).

Si le modèle n'est pas gaussien, le même test peut être utilisé pour n grand car, comme on l'a déjà vu, on a $\frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}} \xrightarrow[\text{sous } H_0]{\mathcal{L}} \mathcal{N}(0,1)$.

Test sur la moyenne m, la variance σ^2 étant inconnue

On considère d'abord le cas du modèle gaussien $\left(X_i \hookrightarrow \mathcal{N}(m, \sigma^2)\right)$.

L'hypothèse nulle est :

$$H_0 : \text{« } m = m_0 \text{ »},$$

où $m_0 \in \mathbb{R}$ est une valeur fixée.

La statistique de test est :

$$T = \frac{\overline{X}_n - m_0}{\sqrt{\overline{S}_n^2/n}}.$$

Sa loi sous H_0, comme on l'a déjà vu (en utilisant le théorème de Cochran), est :

$$\frac{\overline{X}_n - m_0}{\sqrt{\overline{S}_n^2/n}} = \frac{\frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)\overline{S}_n^2}{\sigma^2}\Big/(n-1)}} \hookrightarrow \mathcal{T}_{n-1}.$$

La région critique est donc :

— $W = \left]-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right[\cup \left]t_{n-1,1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $m \neq m_0$ » ;

— $W = \left]t_{n-1,1-\alpha}, +\infty\right[$, si H_1 : « $m > m_0$ » (car T augmente sous H_1) ;

— $W = \left]-\infty, -t_{n-1,1-\alpha}\right[$, si H_1 : « $m < m_0$ » (car T diminue sous H_1).

Si le modèle n'est pas gaussien, comme on l'a déjà vu, on a :

$$\frac{\overline{X}_n - m_0}{\sqrt{\overline{S}_n^2/n}} = \frac{\overline{X}_n - m_0}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{\overline{S}_n^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Par conséquent, la même statistique de test peut être utilisé pour n grand, mais, cette fois-ci, avec la région critique :

- $W = \left]-\infty, -u_{1-\frac{\alpha}{2}}\right[\cup \left]u_{1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $m \neq m_0$ » ;
- $W = \left]u_{1-\alpha}, +\infty\right[$, si H_1 : « $m > m_0$ » (car T augmente sous H_1) ;
- $W = \left]-\infty, -u_{1-\alpha}\right[$, si H_1 : « $m < m_0$ » (car T diminue sous H_1).

Test sur la variance σ^2 dans le cas gaussien $\left(X_i \hookrightarrow \mathcal{N}(m, \sigma^2)\right)$

L'hypothèse nulle est :

$$H_0 : \text{« } \sigma^2 = \sigma_0^2 \text{ »},$$

où $\sigma_0^2 \in \mathbb{R}_+^*$ est une valeur fixée.

La statistique de test est :

$$T = \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma_0^2} = \frac{nS_n^2}{\sigma_0^2} = \frac{(n-1)\overline{S}_n^2}{\sigma_0^2}.$$

Sa loi sous H_0, d'après le théorème de Cochran, est :

$$\frac{(n-1)\overline{S}_n^2}{\sigma_0^2} \hookrightarrow \chi^2_{n-1}.$$

La région critique est donc :

- $W = \left[0, u_{n-1,\frac{\alpha}{2}}\right[\cup \left]u_{n-1,1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $\sigma^2 \neq \sigma_0^2$ » ;
- $W = \left]u_{n-1,1-\alpha}, +\infty\right[$, si H_1 : « $\sigma^2 > \sigma_0^2$ » $\left(\text{car } T > \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \hookrightarrow \chi^2_{n-1} \text{ sous } H_1\right)$;
- $W = \left[0, u_{n-1,\alpha}\right[$, si H_1 : « $\sigma^2 < \sigma_0^2$ » $\left(\text{car } T < \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \hookrightarrow \chi^2_{n-1} \text{ sous } H_1\right)$.

Notons que si la moyenne m est connue, on peut construire un autre test basé sur la statistique $T' = \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma_0^2} \underset{\text{sous } H_0}{\hookrightarrow} \chi^2_n$.

Test sur la proportion p $\left(X_i \hookrightarrow \mathscr{B}(p)\right)$

L'hypothèse nulle est :

$$H_0 : \text{« } p = p_0 \text{ »},$$

où $p_0 \in \left]0, 1\right[$ est une valeur fixée.

La statistique de test est :

$$T = \frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Sa loi approchée (pour n grand) sous H_0, comme on l'a déjà vu, est :

$$\frac{\overline{X}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \rightsquigarrow \approx \mathcal{N}(0,1).$$

Notons qu'ici, on n'utilise pas la loi exacte de T sous H_0, mais son approximation. Le niveau du test construit ne sera donc pas exactement égal à α, mais sera environ α pour n grand. On dit aussi que le test est de **niveau asymptotique** α.

La région critique est donc :

— $W = \left]-\infty, -u_{1-\frac{\alpha}{2}}\right[\cup \left]u_{1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $p \neq p_0$ » ;

— $W = \left]u_{1-\alpha}, +\infty\right[$, si H_1 : « $p > p_0$ » (car T augmente sous H_1) ;

— $W = \left]-\infty, -u_{1-\alpha}\right[$, si H_1 : « $p < p_0$ » (car T diminue sous H_1).

Tests sur 2 populations indépendantes gaussiennes

On suppose que l'on observe un n_1-échantillon $X_{1,1}, \dots, X_{1,n_1}$ de la loi $\mathcal{N}(m_1, \sigma_1^2)$ et un n_2-échantillon $X_{2,1}, \dots, X_{2,n_2}$ de la loi $\mathcal{N}(m_2, \sigma_2^2)$, et que ces deux échantillons sont indépendants (par exemple, les taux de globules rouges d'un groupe de patients sous traitement, et ceux d'un autre groupe de patients sous placebo).

Notons que l'indépendance des échantillons est très importante ici. Lorsqu'on observe deux mesures sur la même population (par exemple, les taux de globules rouges des patients avant et après le traitement), on parle de deux échantillons **appariés**. Dans ce cas, il faut plutôt raisonner sur les différences $Y_i = X_{2,i} - X_{1,i}$.

1. **Test d'égalité des variances**

   L'hypothèse nulle est :

   $$H_0 : \text{« } \sigma_1^2 = \sigma_2^2 \text{ »}.$$

   En notant $\overline{S}^2_{1,n_1}$ et $\overline{S}^2_{2,n_2}$ les variances empiriques corrigées des deux échantillons, la statistique de test est :

   $$T = \frac{\overline{S}^2_{1,n_1}}{\overline{S}^2_{2,n_2}}.$$

   Sa loi sous H_0 est :

   $$\frac{\overline{S}^2_{1,n_1}}{\overline{S}^2_{2,n_2}} = \frac{\overline{S}^2_{1,n_1}}{\overline{S}^2_{2,n_2}} \frac{\sigma_2^2}{\sigma_1^2} = \frac{\overline{S}^2_{1,n_1} \Big/ \sigma_1^2}{\overline{S}^2_{2,n_2} \Big/ \sigma_2^2} = \frac{\frac{(n_1-1)\overline{S}^2_{1,n_1}}{\sigma_1^2} \Big/ (n_1-1)}{\frac{(n_2-1)\overline{S}^2_{2,n_2}}{\sigma_2^2} \Big/ (n_2-1)} \rightsquigarrow \mathcal{F}_{n_1-1,\, n_2-1},$$

où nous avons utilisé (deux fois) le théorème de Cochran, l'indépendance des échantillons et la définition de la loi de Fisher-Snedecor.

La région critique est donc :

— $W = \left[0,\ f_{n_1-1,\,n_2-1,\,\frac{\alpha}{2}}\right[\cup \left]f_{n_1-1,\,n_2-1,\,1-\frac{\alpha}{2}},\ +\infty\right[$, si H_1 : « $\sigma_1^2 \neq \sigma_2^2$ » ;

— $W = \left]f_{n_1-1,\,n_2-1,\,1-\alpha},\ +\infty\right[$, si H_1 : « $\sigma_1^2 > \sigma_2^2$ » $\left(\text{car } T > \frac{\overline{S}_{1,n_1}^2}{\overline{S}_{2,n_2}^2}\frac{\sigma_2^2}{\sigma_1^2} \text{ sous } H_1\right)$;

— $W = \left[0,\ f_{n_1-1,\,n_2-1,\,\alpha}\right[$, si H_1 : « $\sigma_1^2 < \sigma_2^2$ » $\left(\text{car } T < \frac{\overline{S}_{1,n_1}^2}{\overline{S}_{2,n_2}^2}\frac{\sigma_2^2}{\sigma_1^2} \text{ sous } H_1\right)$.

2. **Test d'égalité des moyennes, sachant que les variances (inconnues) sont égales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)**

L'hypothèse nulle est :

$$H_0 : \text{« } m_1 = m_2 \text{ ».}$$

La statistique de test est :

$$T = \frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

où nous avons noté $\overline{X}_{1,n_1}$ et $\overline{X}_{2,n_2}$ les moyennes empiriques des deux échantillons, et $S^2 = \frac{n_1 S_{1,n_1}^2 + n_2 S_{2,n_2}^2}{n_1+n_2-2}$ est un estimateur de la variance (commune) σ^2 construit à partir de l'ensemble des deux populations (ici, S_{1,n_1}^2 et S_{2,n_2}^2 sont les variances empiriques des échantillons).

On peut montrer que la loi de cette statistique sous H_0 est :

$$\frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \hookrightarrow \mathcal{T}_{n_1+n_2-2}.$$

La région critique est donc :

— $W = \left]-\infty,\ -t_{n_1+n_2-2,\,1-\frac{\alpha}{2}}\right[\cup \left]t_{n_1+n_2-2,\,1-\frac{\alpha}{2}},\ +\infty\right[$, si H_1 : « $m_1 \neq m_2$ » ;

— $W = \left]t_{n_1+n_2-2,\,1-\alpha},\ +\infty\right[$, si H_1 : « $m_1 > m_2$ » (car T augmente sous H_1) ;

— $W = \left]-\infty,\ -t_{n_1+n_2-2,\,1-\alpha}\right[$, si H_1 : « $m_1 < m_2$ » (car T diminue sous H_1).

3. **Test d'égalité des moyennes sans supposer l'égalité des variances**

Si on n'assume pas l'égalité des variances, on peut utiliser le test **de Welch** dont la statistique est légèrement différente :

$$T = \frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{\frac{\overline{S}_{1,n_1}^2}{n_1} + \frac{\overline{S}_{2,n_2}^2}{n_2}}},$$

où $\overline{S}^2_{1,n_1}$ et $\overline{S}^2_{2,n_2}$ sont les variances empiriques corrigées des échantillons.

On peut montrer que cette statistique suit, sous H_0, une loi de Student avec un nombre de degrés de liberté ν compliqué (et non nécessairement entier !) qui peut être approché par l'**équation de Welch-Satterthwaite** :

$$\nu = \frac{\left(\frac{S^2_{1,n_1}}{n_1} + \frac{S^2_{2,n_2}}{n_2}\right)^2}{\frac{S^4_{1,n_1}}{n_1^2(n_1-1)} + \frac{S^4_{2,n_2}}{n_2^2(n_2-1)}}\,.$$

La région critique est donc :

— $W = \left]-\infty,\, -t_{\nu,\,1-\frac{\alpha}{2}}\right[\cup \left]t_{\nu,\,1-\frac{\alpha}{2}},\, +\infty\right[$, si H_1 : « $m_1 \neq m_2$ » ;

— $W = \left]t_{\nu,\,1-\alpha},\, +\infty\right[$, si H_1 : « $m_1 > m_2$ » (car T augmente sous H_1) ;

— $W = \left]-\infty,\, -t_{\nu,\,1-\alpha}\right[$, si H_1 : « $m_1 < m_2$ » (car T diminue sous H_1).

Comparaison de 2 proportions

On suppose maintenant que l'on observe un n_1-échantillon $X_{1,1}, \ldots, X_{1,n_1}$ de la loi $\mathscr{B}(p_1)$ et un n_2-échantillon $X_{2,1}, \ldots, X_{2,n_2}$ de la loi $\mathscr{B}(p_2)$, et que ces échantillons sont indépendants.

L'hypothèse nulle est :

$$H_0 : \text{« } p_1 = p_2 \text{ »}.$$

La statistique de test est :

$$T = \frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\,,$$

où nous avons, comme avant, noté $\overline{X}_{1,n_1}$ et $\overline{X}_{2,n_2}$ les moyennes empiriques des deux échantillons, et $\widehat{p} = \frac{n_1\overline{X}_{1,n_1} + n_2\overline{X}_{1,n_1}}{n_1+n_2}$ est un estimateur (sous H_0) du paramètre $p_1 = p_2 = p$ construit à partir de l'ensemble des deux populations.

On peut montrer que la loi approchée (lorsque n_1 et n_2 sont grands) de cette statistique sous H_0 est :

$$\frac{\overline{X}_{1,n_1} - \overline{X}_{2,n_2}}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \hookrightarrow \approx \mathcal{N}(0,1).$$

Notons qu'ici, la loi de T sous H_0 n'est pas entièrement déterminée : elle varie un peu en fonction de la valeur de p. Mais si n est grand, les variations de cette loi sont infimes, car pour tout p elle peut être approchée par la même loi $\big(\mathcal{N}(0,1)\big)$. Par conséquent, on peut construire un test de niveau asymptotique α de manière habituelle.

La région critique est donc :

— $W = \left]-\infty, -u_{1-\frac{\alpha}{2}}\right[\cup \left]u_{1-\frac{\alpha}{2}}, +\infty\right[$, si H_1 : « $p_1 \neq p_0$ » ;

— $W = \left]u_{1-\alpha}, +\infty\right[$, si H_1 : « $p_1 > p_2$ » (car T augmente sous H_1) ;

— $W = \left]-\infty, -u_{1-\alpha}\right[$, si H_1 : « $p_1 < p_2$ » (car T diminue sous H_1).

Tests d'ajustement

Les tests **d'ajustement** ou **d'adéquation** (**goodness-of-fit tests** en anglais) permettent de tester si les observations suivent une loi donnée. Attention, un test d'ajustement vérifie juste que rien ne "contredit" cette hypothèse (mais, en aucun cas, il ne la "prouve" pas ; si on la conserve, c'est par défaut). Voici trois tests d'adéquation les plus utilisés.

1. **Test du χ^2**

   On suppose que l'on observe un n-échantillon $X_1, \dots, X_n$ d'une loi discrète à k valeurs, dont la table de répartition est :

v_1	$\cdots$	v_k
p_1	$\cdots$	p_k

   Soient $n_1, \dots, n_k$ les effectifs des modalités $v_1, \dots, v_k$ dans l'échantillon (avec donc $n_1 + \cdots + n_k = n$). On veut tester l'hypothèse

   $$H_0 : \text{« } p_1 = p_1^0, \, p_2 = p_2^0, \dots, p_k = p_k^0 \text{ » (adéquation),}$$

   où $p_1^0, \dots, p_k^0$ sont des valeurs fixées vérifiant $p_i^0 > 0$ et $p_1^0 + \cdots + p_k^0 = 1$, contre l'alternative

   $$H_1 = \overline{H_0} : \text{« } \exists\, i \in \{1, \dots, k\} \text{ tel que } p_i \neq p_i^0 \text{ » (pas d'adéquation).}$$

   Notons que sous H_0, les n_i (également dits **effectifs observés**) doivent être proches des np_i^0 (dits **effectifs espérés**).

   La statistique de test est :

   $$T = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}.$$

   On peut montrer que sa loi approchée (pour n grand) sous H_0 est :

   $$\sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0} \underset{\mathcal{L}}{\approx} \chi^2_{k-1}.$$

   Comme sous H_1 la statistique T augmente (car les effectifs observés diffèrent plus des effectifs espérés que sous H_0), la région critique est :

   $$W = \left]u_{k-1,\,1-\alpha}, +\infty\right[.$$

Remarques.

1. Si $k = +\infty$ (comme, par exemple, pour une loi de Poisson), on groupe les valeurs à faible probabilité (et donc faible effectif, typiquement $n_i < 5$) en une seule classe (pour la loi de Poisson, ça pourrait donner, par exemple, les classes 0, 1, 2, 3 et $\{4, 5, \ldots\}$), et on effectue le test comme décrit précédemment.
2. On peut utiliser le test du χ^2 pour tester l'adéquation d'un caractère quantitatif continu à une loi de probabilité continue en les discrétisant (tous les deux), mais il y a d'autres tests plus adaptés pour ce cas de figure.

2. **Test de Kolmogorov-Smirnov**

Le test **de Kolmogorov-Smirnov** est utilisé pour tester l'hypothèse

$$H_0 : \text{« } F = F_0 \text{ » (adéquation)},$$

où F est la fonction de répartition des X_i et F_0 est la fonction de répartition d'une loi de probabilité continue fixée (à laquelle on veut tester l'adéquation), contre l'alternative

$$H_1 = \overline{H_0} : \text{« } F \neq F_0 \text{ » (pas d'adéquation)}.$$

La statistique de test est :

$$T = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|.$$

La loi de la statistique T sous H_0 est connue et, comme sous H_1 cette statistique augmente (car $\widehat{F}_n$ diffère plus de F_0), le test consiste à rejeter dès que T dépasse le quantile d'ordre $1 - \alpha$ de cette loi.

3. **Test de Cramér-Von Mises**

Le test **de Cramér-Von Mises** est similaire à celui de Kolmogorov-Smirnov, mais la statistique de test est différente :

$$T = \int_{-\infty}^{+\infty} \left(\widehat{F}_n(x) - F_0(x) \right)^2 f_0(x) \, \mathrm{d}x,$$

où $f_0(x) = F_0'(x)$ est la densité de la loi de probabilité à laquelle on est en train de tester l'adéquation.

Notons finalement qu'il existe également beaucoup de tests permettant de tester l'adéquation à des lois particulières. Citons, à titre d'exemple, le test de Shapiro-Wilk (qu'on ne détaillera pas ici) qui permet de tester l'adéquation à une loi normale.

Test du χ^2 d'indépendance

On suppose que l'on observe deux n-échantillons appariés : $X_1, \dots, X_n$ d'une loi discrète ayant k valeurs $v_1, \dots, v_k$ et $Y_1, \dots, Y_n$ d'une loi discrète ayant r valeurs $u_1, \dots, u_r$. En notant n_{ij} le nombre d'observations k telles que $X_k = v_i$ et $Y_k = u_j$, nous obtenons le tableau de contingence suivant.

X \ Y	u_1	$\cdots$	u_j	$\cdots$	u_r	
v_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1r}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
v_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ir}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
v_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kr}	$n_{k\bullet}$
	$n_{\bullet 1}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet r}$	$n_{\bullet\bullet} = n$

On veut tester l'hypothèse

$$H_0 : \text{« } X_1 \text{ et } Y_1 \text{ sont indépendantes »}$$

contre l'alternative

$$H_1 = \overline{H_0} : \text{« } X_1 \text{ et } Y_1 \text{ sont dépendantes »}.$$

Notons que sous H_0, on peut s'attendre que les **effectifs observés** n_{ij} soient proches des $n\,\mathbf{P}(X_1 = v_i)\,\mathbf{P}(Y_1 = u_j) \approx n\,\frac{n_{i\bullet}}{n}\,\frac{n_{\bullet j}}{n} = \frac{n_{i\bullet}\,n_{\bullet j}}{n}$ (dits **effectifs espérés** dans ce cas).

La statistique de test est :

$$T = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i\bullet}\,n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\,n_{\bullet j}}{n}}\,.$$

On peut montrer que sa loi approchée (pour n grand) sous H_0 est :

$$\sum_{i=1}^{k}\sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i\bullet}\,n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\,n_{\bullet j}}{n}} \;\overset{\mathcal{L}}{\approx}\; \chi^2_{(k-1)(r-1)}.$$

Comme sous H_1 la statistique T augmente (car les effectifs observés diffèrent plus des effectifs espérés que sous H_0), la région critique est :

$$W = \left]u_{(k-1)(r-1),\,1-\alpha},\, +\infty\right[.$$

Test du χ^2 d'homogénéité

On suppose que l'on observe k échantillons, chacun prélevé dans sa propre population. Chaque échantillon peut provenir de sa propre loi, mais on suppose que toutes ces lois sont discrètes et ont un support commun $\{v_1, \dots, v_r\}$. On note n_{ij} l'effectif de la $j^{\text{ème}}$ valeur dans le $i^{\text{émé}}$ échantillon, et p_{ij} la probabilité de cette valeur dans la $i^{\text{émé}}$ population. Notons que les n_{ij} forment une tableau de contingence similaire à celui de la page précédente, et que $n_{i\bullet}$ est égal ici à la taille du $i^{\text{ème}}$ échantillon.

On veut savoir si la loi est la même dans les k populations, c'est-à-dire tester l'hypothèse

$$H_0 : \text{« } p_{1j} = p_{2j} = \cdots = p_{kj} \text{ pour tout } j \in \{1, \dots, r\} \text{ » (homogénéité)}$$

contre l'alternative

$$H_1 = \overline{H_0} : \text{« } \exists\, i, i' \in \{1, \dots, k\} \text{ et } j \in \{1, \dots, r\} \text{ tels que } p_{ij} \neq p_{i'j} \text{ ».}$$

L'hypothèse H_0 d'homogénéité revient à dire que la variable observée est indépendante du numéro de la population. On peut donc procéder de la même manière que pour le test du χ^2 d'indépendance.

La statistique de test est :

$$T = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i\bullet}\, n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\, n_{\bullet j}}{n}}.$$

On peut montrer que sa loi approchée (lorsque les tailles de tous les k échantillons sont grandes) sous H_0 est :

$$\sum_{i=1}^{k} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i\bullet}\, n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\, n_{\bullet j}}{n}} \overset{\mathcal{L}}{\approx} \chi^2_{(k-1)(r-1)}.$$

Comme sous H_1 la statistique T augmente (car les effectifs observés diffèrent plus des effectifs espérés que sous H_0), la région critique est :

$$W = \left]u_{(k-1)(r-1),\,1-\alpha},\, +\infty\right[.$$

Introduction à ANOVA (ANalysis Of VAriance)

On suppose que l'on observe k échantillons gaussiens indépendants (qu'on appellera également **classes**). On note $X_{i1}, \dots, X_{in_i}$ le $i^{\text{éme}}$ échantillon, et $\mathcal{N}(m_i, \sigma^2)$ la loi dont il est issu (la variance est supposée être la même pour les k échantillons : cas **homoscédastique**). Notons que le numéro d'échantillon peut être vu comme un caractère qualitatif ayant les modalités $1, \dots, k$ (dit **facteur**), qu'on notera A.

On veut savoir si la moyenne de la loi normale est la même dans les k classes, c'est-à-dire tester l'hypothèse

$$H_0 : \text{« } m_1 = m_2 = \cdots = m_k \text{ »}$$

contre l'alternative

$$H_1 = \overline{H_0} : \text{« } \exists\, i, i' \in \{1, \ldots, k\} \text{ tels que } m_i \neq m_{i'} \text{ ».}$$

On note

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad \text{et} \quad S_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$$

la moyenne et la variance empiriques calculées sur la $i^{\text{ème}}$ classe, et

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} \quad \text{et} \quad S^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2$$

la moyenne et la variance empiriques calculées sur l'ensemble des observations.

Notons que

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \, \overline{X}_i = \sum_{i=1}^{k} \overline{X}_i \, \frac{n_i}{n} \, .$$

C'est la formule de l'espérance totale $\mathbf{E}\, X = \mathbf{E}\big(\mathbf{E}(X \,|\, A)\big)$ appliquée à la loi empirique. De même, la formule de la variance totale $\mathbf{Var}(X) = \mathbf{Var}\big(\mathbf{E}(X \,|\, A)\big) + \mathbf{E}\big(\mathbf{Var}(X \,|\, A)\big)$ donne

$$S^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2 + \frac{1}{n} \sum_{i=1}^{k} n_i \, S_i^2.$$

Cette formule montre que la variance empirique se décompose en deux termes : la variance **expliquée par le** (ou **due au**) **facteur** A

$$S_A^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \left(\overline{X}_i - \overline{X}\right)^2,$$

dite également variance **inter-classe**, et la variance **résiduelle**

$$S_{\mathrm{R}}^2 = \frac{1}{n} \sum_{i=1}^{k} n_i \, S_i^2,$$

dite également variance **intra-classe**.

D'après le théorème de Cochran (appliqué au $i^{\text{ème}}$ échantillon), on a $\frac{n_i S_i^2}{\sigma^2} \rightsquigarrow \chi^2_{n_i - 1}$. Vu que les échantillons sont indépendants, et en remarquant que $\sum_{i=1}^{k} (n_i - 1) = n - k$, on en déduit $\frac{n S_{\mathrm{R}}^2}{\sigma^2} = \sum_{i=1}^{k} \frac{n_i S_i^2}{\sigma^2} \rightsquigarrow \chi^2_{n-k}$.

Sous H_0, on peut également montrer que S_{R}^2 et S_A^2 sont indépendants et que $\frac{nS_A^2}{\sigma^2} \backsim \chi_{k-1}^2$. À titre d'illustration, vérifions la dernière affirmation dans le cas particulier où les échantillons sont tous de la même taille ($n = k\,n_1$). Dans ce cas (et sous H_0), $\overline{X}_1, \ldots, \overline{X}_k$ est un k-échantillon de la loi $\mathcal{N}(m_1^2,\, \sigma^2/n_1)$, sa moyenne est $\frac{1}{k}\sum_{i=1}^k \overline{X}_i = \frac{1}{n}\sum_{i=1}^k n_i\,\overline{X}_i = \overline{X}$ et, d'après le théorème de Cochran (appliqué à ce k-échantillon), on a

$$\frac{\sum_{i=1}^k \big(\overline{X}_i - \overline{X}\big)^2}{\sigma^2/n_1} = \frac{\sum_{i=1}^k n_i\,\big(\overline{X}_i - \overline{X}\big)^2}{\sigma^2} = \frac{nS_A^2}{\sigma^2} \backsim \chi_{k-1}^2.$$

Ainsi on aboutit au test suivant, dit test **de Fisher**.

La statistique de test est :

$$T = \frac{S_A^2 \,\big/\, (k-1)}{S_{\mathrm{R}}^2 \,\big/\, (n-k)} = \frac{\frac{nS_A^2}{\sigma^2} \,\Big/\, (k-1)}{\frac{nS_{\mathrm{R}}^2}{\sigma^2} \,\Big/\, (n-k)} \underset{\text{sous } H_0}{\backsim} \mathcal{F}_{k-1,\,n-k}.$$

Comme sous H_1 la statistique T augmente (car la part de la variance expliquée par le facteur A devient plus grande), la région critique est : et la région critique est :

$$W = \big]\, f_{k-1,\,n-k,\,1-\alpha},\, +\infty \big[.$$

Bibliographie

[1] Alexander Borovkov, "*Statistique Mathématique*", Editions Mir, 1987 (éditions de la version russe : 2010, 2007, 1997 et 1984).

[2] Gilbert Saporta, "*Probabilités, Analyse des Données et Statistique*", Editions Technip, 2011 (éditions précédentes : 2006 et 1990).

[3] Julien Jacques, "*Statistiques Inférentielles*", Polycopié de cours téléchargeable sur `http://eric.univ-lyon2.fr/~jjacques/enseignement.html`