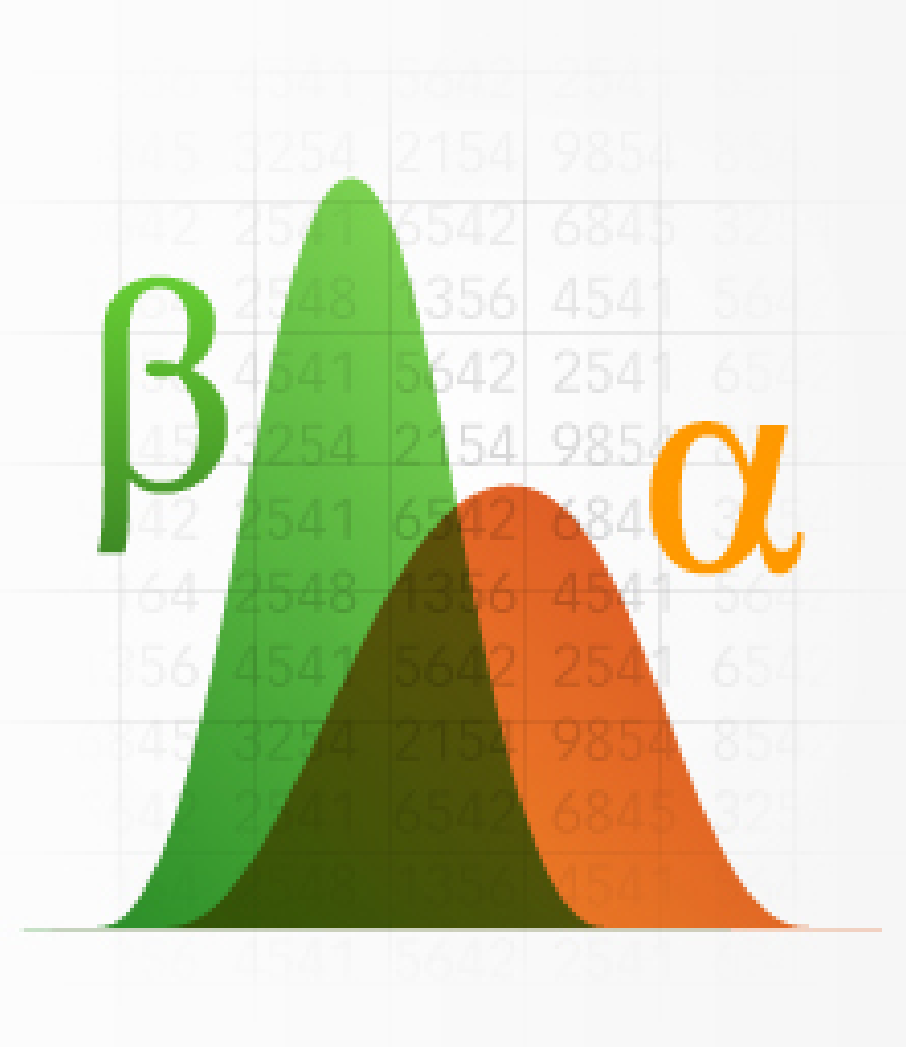
Licence 2-S4 SI-MASS
Année 2018

Cours de Statistiques inférentielles

Pierre DUSART



Chapitre 1

Lois statistiques

1.1 Introduction

Nous allons voir que si une variable aléatoire suit une certaine loi, alors ses réalisations (sous forme d'échantillons) sont encadrées avec des probabilités de réalisation. Par exemple, lorsque l'on a une énorme urne avec une proportion p de boules blanches alors le nombre de boules blanches tirées sur un échantillon de taille n est parfaitement défini. En pratique, la fréquence observée varie autour de p avec des probabilités fortes autour de p et plus faibles lorsqu'on s'éloigne de p.

Nous allons chercher à faire l'inverse : l'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

1.1.1 Fonction de répartition

La densité de probabilité $p(x)$ ou la fonction de répartition $F(x)$ définissent la loi de probabilité d'une variable aléatoire continue X. Elles donnent lieu aux représentations graphiques suivantes :

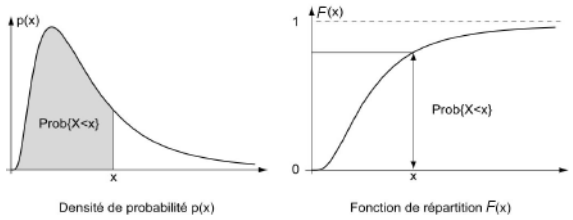


FIGURE 1.1 – fonction répartition

La fonction de distribution cumulée $F(x)$ exprime la probabilité que X n'excède pas la valeur x :

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

De même, la probabilité que X soit entre a et b $(b > a)$ vaut

$$P(a < X < b) = F(b) - F(a).$$

1.1.2 Grandeurs observées sur les échantillons

L'espérance $E(X)$ d'une variable aléatoire discrète X est donnée par la formule

$$E(X) = \sum_i x_i P(x_i).$$

L'espérance est également appelée moyenne et notée dans ce cas μ_X.

Sa variance σ_X^2 est l'espérance des carrés des écarts avec la moyenne :

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_i (x_i - \mu_X)^2 P(x_i) = \sum_i x_i^2 P(x_i) - \mu_X^2.$$

Son écart-type σ_X est la racine positive de la variance.

1.2 Lois usuelles

1.2.1 Loi normale ou loi de Gauss

Une variable aléatoire réelle X suit une loi normale (ou loi gaussienne, loi de Laplace-Gauss) d'espérance μ et d'écart type σ (nombre strictement positif, car il s'agit de la racine carrée de la variance σ^2) si cette variable aléatoire réelle X admet pour densité de probabilité la fonction $p(x)$ définie, pour tout nombre réel x, par :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Une telle variable aléatoire est alors dite variable gaussienne.

Une loi normale sera notée de la manière suivante $\mathcal{N}(\mu, \sigma)$ car elle dépend de deux paramètres μ (la moyenne) et σ(l'écart-type). Ainsi si une variable aléatoire X suit $\mathcal{N}(\mu, \sigma)$ alors

$$E(X) = \mu \quad \text{et} \quad V(X) = \sigma^2.$$

Lorsque la moyenne μ vaut 0, et l'écart-type vaut 1, la loi sera notée $\mathcal{N}(0, 1)$ et sera appelée loi normale standard. Sa fonction caractéristique vaut $e^{-t^2/2}$. Seule la loi $\mathcal{N}(0, 1)$ est tabulée car les autres lois (c'est-à-dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant : Si Y suit $\mathcal{N}(\mu, \sigma)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $\mathcal{N}(0, 1)$.

On note Φ la fonction de répartition de la loi normale centrée réduite :

$$\Phi(x) = P(Z < x)$$

avec Z une variable aléatoire suivant $\mathcal{N}(0, 1)$.

Propriétés et Exemples : $\Phi(-x) = 1 - \Phi(x)$,

$$\Phi(0) = 0.5, \quad \Phi(1.645) \approx 0.95, \quad \Phi(1.960) \approx 0.9750$$

Pour $|x| < 2$, une approximation de Φ peut être utilisée ; il s'agit de son développement de Taylor à l'ordre 5 au voisinage de 0 :

$$\Phi(x) \approx 0.5 + \frac{1}{\sqrt{2\pi}} \left(x - \frac{x^3}{6} + \frac{x^5}{40} \right).$$

Inversement, à partir d'une probabilité, on peut chercher la borne pour laquelle cette probabilité est effective.

Notation : on notera $z_{\alpha/2}$ le nombre pour lequel

$$P(Z > z_{\alpha/2}) = \alpha/2$$

lorsque la variable aléatoire suit la loi normale standard.

risque α	0.01	0.02	0.05	0.10
valeur critique $z_{\alpha/2}$	2.58	2.33	1.96	1.645
coefficient de sécurité c	99%	98%	95%	90%

A l'aide des propriétés de la loi normale standard, on remarque que le nombre $z_{\alpha/2}$ vérifie également

$$\begin{aligned} P(Z < z_{\alpha/2}) &= \\ P(Z < -z_{\alpha/2}) &= \\ P(-z_{\alpha/2} < Z < z_{\alpha/2}) &= \\ P(|Z| > z_{\alpha/2}) &= \end{aligned}$$

La somme de deux variables gaussiennes indépendantes est elle-même une variable gaussienne (stabilité) : Soient X et Y deux variables aléatoires indépendantes suivant respectivement les lois $\mathcal{N}(\mu_1, \sigma_1)$ et $\mathcal{N}(\mu_2, \sigma_2)$. Alors, la variable aléatoire $X + Y$ suit la loi normale $\mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

1.2.2 Loi du χ^2 (khi-deux)

Définition 1 *Soit $Z_1, Z_2, \ldots, Z_\nu$ une suite de variables aléatoires indépendantes de même loi $\mathcal{N}(0,1)$. Alors la variable aléatoire $\sum_{i=1}^{\nu} Z_i^2$ suit une loi appelée **loi du Khi-deux** à ν degrés de liberté, notée $\chi^2(\nu)$.*

Proposition 1.2.1 *1. Sa fonction caractéristique est $(1-2it)^{-\nu/2}$.*

2. La densité de la loi du $\chi^2(\nu)$ est

$$f_\nu(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & \text{pour } x > 0 \\ 0 \quad \text{sinon.} \end{cases}$$

où Γ est la fonction Gamma d'Euler définie par $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$.

3. L'espérance de la loi du $\chi^2(\nu)$ est égale au nombre ν de degrés de liberté et sa variance est 2ν.

4. La somme de deux variables aléatoires indépendantes suivant respectivement $\chi^2(\nu_1)$ et $\chi^2(\nu_2)$ suit aussi une loi du χ^2 avec $\nu_1 + \nu_2$ degrés de liberté.

Preuve Calculons la fonction caractéristique de Z^2 lorsque Z suit $\mathcal{N}(0,1)$.

$$\begin{aligned} \varphi(t) &= E(e^{itZ^2}) = \int_{-\infty}^{\infty} e^{itz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2it)z^2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}u^2}}{(1-2it)^{1/2}} dt \quad \text{en posant } u = (1-2it)^{1/2} z \\ \varphi(t) &= (1-2it)^{-1/2} \end{aligned}$$

Maintenant pour la somme de ν variables Z_i^2 indépendantes, on a

$$\varphi(t) = (1-2it)^{-\nu/2}.$$

Montrons maintenant que la fonction de densité est correcte. Pour cela, calculons la fonction caractéristique à partir de la densité :

$$\begin{aligned}
\varphi(t) &= E(e^{itx}) = \int_0^{+\infty} e^{itx} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^{+\infty} x^{(-1/2-it)x} dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \frac{1}{(1/2-it)(1/2-it)^{\nu/2-1}} \int_0^{+\infty} u^{\nu/2-1} e^{-u} du \quad \text{en posant } u = (1/2-it)x \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \frac{1}{(1/2-it)^{\nu/2}} \underbrace{\int_0^{+\infty} u^{\nu/2-1} e^{-u} du}_{=\Gamma(\nu/2)} \\
\varphi(t) &= \frac{1}{(1-2it)^{\nu/2}}
\end{aligned}$$

Calculons maintenant l'espérance et la variance. Selon la définition de la loi du χ^2, chaque variable Z_i suit la loi normale centrée réduite. Ainsi $E(Z_i^2) = Var(Z_i) = 1$ et $E(\sum_{i=1}^{\nu} Z_i^2) = \nu$. De même, $V(Z_i r) = E(Z_i^4) - (E(Z_i^2))^2 = \mu_4 - 1$. On sait que pour une loi normale centrée réduite $\mu_4 = 3$ donc $Var(Z_i^2) = 2$ et $Var(\sum_{i=1}^{\nu} Z_i^2) = 2\nu$.

La dernière proposition est évidente de par la définition de la loi du χ^2.

Fonction inverse : on peut trouver une tabulation de la fonction réciproque de la fonction de répartition de cette loi dans une table (en annexe) ou sur un logiciel tableur :

$$\alpha \mapsto \chi^2_{\alpha;\nu} \quad (\text{Fonction } \texttt{KHIDEUX.inverse}(\alpha;\nu)),$$

c'est-à-dire la valeur de $\chi^2_{\alpha;\nu}$ telle que $P(\chi^2(\nu) > \chi^2_{\alpha;\nu}) = \alpha$.

Exemple : Pour $\alpha = 0.990$ et $\nu = 5$, $\chi^2_\alpha = 0.554 = \chi^2_{0.99;5}$.

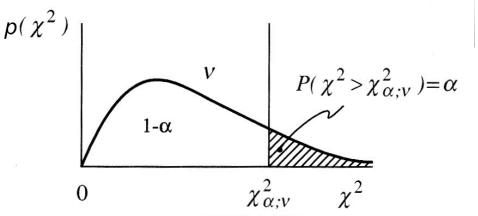


FIGURE 1.2 – fonction χ^2 inverse

1.2.3 Loi de Student

Définition 2 *Soient Z et Q deux variables aléatoires indépendantes telles que Z suit $\mathcal{N}(0,1)$ et Q suit $\chi^2(\nu)$. Alors la variable aléatoire*

$$T = \frac{Z}{\sqrt{Q/\nu}}$$

*suit une loi appelée **loi de Student** à ν degrés de liberté, notée $St(\nu)$.*

Proposition 1.2.2 *1. La densité de la loi de la loi de Student à ν degrés de liberté est*

$$f(x) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)} \frac{1}{(1+x^2/\nu)^{\frac{\nu+1}{2}}}.$$

2. L'espérance n'est pas définie pour $\nu = 1$ et vaut 0 si $\nu \geq 2$. Sa variance n'existe pas pour $\nu \leq 2$ et vaut $\nu/(\nu-2)$ pour $\nu \geq 3$.

3. La loi de Student converge en loi vers la loi normale centrée réduite.

Remarque : pour $\nu = 1$, la loi de Student s'appelle loi de Cauchy, ou loi de Lorentz.

1.2.4 Loi de Fisher-Snedecor

Définition 3 *Soient Q_1 et Q_2 deux variables aléatoires indépendantes telles que Q_1 suit $\chi^2(\nu_1)$ et Q_2 suit $\chi^2(\nu_2)$ alors la variable aléatoire*

$$F = \frac{Q_1/\nu_1}{Q_2/\nu_2}$$

suit une loi de Fisher-Snedecor à (ν_1, ν_2) degrés de liberté, notée $F(\nu_1, \nu_2)$.

Proposition 1.2.3 *La densité de la loi $F(\nu_1, \nu_2)$ est*

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{\nu_1/2-1}}{(1+\frac{\nu_1}{\nu_2}x)^{\frac{\nu_1+\nu_2}{2}}} \text{ si } x > 0 \quad (0 \text{ sinon}).$$

Son espérance n'existe que si $\nu_2 \geq 3$ et vaut $\frac{\nu_2}{\nu_2-2}$. Sa variance n'existe que si $\nu_2 \geq 5$ et vaut $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$.

Proposition 1.2.4 *1. Si F suit une loi de Fisher $F(\nu_1, \nu_2)$ alors $\frac{1}{F}$ suit une loi de Fisher $F(\nu_2, \nu_1)$.*

2. Si T suit une loi de Student à ν degrés de liberté alors T^2 suit une loi de Fisher $F(1, \nu)$.

1.2.5 Fonctions inverses et Tableur

Loi	Notation	Variable	Fct Répartition	V. critique	Fonction inverse
Gauss	$\mathcal{N}(0,1)$	Z	`loi.normale.standard`(z)	z_α	`loi.normale.standard.inverse`$(1-\alpha)$
Khi-Deux	$\chi^2(\nu)$	K^2	`khideux`$(k;\nu;1)$	$\chi^2_{\alpha;\nu}$	`khideux.inverse`$(\alpha;\nu;1)$
Student	$St(\nu)$	T	`Loi.student`$(t;\nu;1)$	$t_{\alpha;\nu}$	`Loi.student.inverse`$(\alpha;\nu)$
Fisher	$F(\nu_1,\nu_2)$	F	`Loi.f`$(f;\nu_1;\nu_2))$	$f_{\alpha;\nu_1,\nu_2}$	`inverse.Loi.f`$(\alpha;\nu_1;\nu_2))$

Chapitre 2

Convergences

2.1 Convergence en probabilité

2.1.1 Inégalités utiles

Inégalité de Markov simplifiée
Soit Y une v.a.r., g une fonction croissante et positive ou nulle sur l'ensemble des réels, vérifiant $g(a) > 0$, alors

$$\forall a > 0, P(Y \geq a) \leq \frac{E(g(Y))}{g(a)}.$$

Preuve

$$\begin{aligned} E(g(Y)) &= \int_\Omega g(y)f(y)dy = \int_{Y<a} g(y)f(y)dy + \int_{Y\geq a} g(y)f(y)dy \\ &\geq \int_{Y\geq a} g(y)f(y)dy \quad \text{car } g \text{ est positive ou nulle} \\ &\geq g(a)\int_{Y\geq a} f(y)dy \quad \text{car } g \text{ est croissante} \\ &= g(a)P(Y \geq a) \end{aligned}$$

Ainsi $E(g(Y)) \geq g(a)P(Y \geq a)$.

Rappel : Inégalité de Bienaymé-Chebyshev
Soit X une variable aléatoire admettant une espérance $E(X)$ et de variance finie σ^2 (l'hypothèse de variance finie garantit l'existence de l'espérance).

L'inégalité de Bienaymé-Chebychev s'énonce de la façon suivante : pour tout réel ε strictement positif,

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Preuve Voir Cours S3 ou prendre $Y = |X - E(X)|$, $a = \varepsilon$ et $g(t) = t^2$ dans l'inégalité de Markov.

2.1.2 Convergence en probabilité

Définition 4 (Convergence en probabilité) *On considère une suite (X_n) d'une v.a. définie sur Ω, X une autre v.a. définie sur Ω.*
On dit que la suite (X_n) converge en probabilité vers une constante réelle ℓ si

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - \ell| > \varepsilon) = 0.$$

On dit que la suite (X_n) converge en probabilité vers X si

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0.$$

Exemple de la loi binomiale : On réalise n expériences indépendantes et on suppose que lors de chacune de ces expériences, la probabilité d'un événement appelé "succès" est p. Soit S_n le nombre de succès obtenus lors de ces n expériences. La variance aléatoire S_n, somme de n variables de Bernoulli indépendantes, de même paramètre p, suit une loi binomiale : $S_n \hookrightarrow \mathcal{B}(n,p)$.

On s'intéresse alors à la variable aléatoire $\frac{S_n}{n}$, proportion de succès sur n expériences, a donc pour espérance $E(\frac{S_n}{n}) = p$ et pour variance $V(\frac{S_n}{n}) = \frac{1}{n^2}V(S_n) = \frac{p(1-p)}{n}$. Comme $p(1-p)$ atteint son maximum lorsque $p = 1/2$, on a ainsi $p(1-p) \leq 1/4$. En appliquant l'inégalité de Bienaymé-Chebyshev, il vient

$$P(|S_n/n - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}.$$

Ainsi pour tout $\varepsilon > 0$, il existe $\eta > 0$ (plus précisément $\eta > \frac{1}{4n\varepsilon^2}$) tel que $P(|S_n/n - p| \geq \varepsilon) < \eta$ ou encore $\lim_{n\to\infty} P(|S_n/n - p| \geq \varepsilon) = 0$. La variable aléatoire $\frac{S_n}{n}$ converge en probabilité vers p.

Théorème 2.1.1 *Soit (X_n) une suite de variables aléatoires sur le même espace probabilisé (Ω, P) admettant des espérances et des variances vérifiant*

$$\lim_{n\to\infty} E(X_n) = \ell \quad \text{et} \quad \lim_{n\to\infty} V(X_n) = 0,$$

alors les (X_n) convergent en probabilité vers ℓ.

Preuve Soit $\varepsilon > 0$. Posons $E(X_n) = \ell + u_n$ avec $\lim u_n = 0$. Alors il existe $N \in \mathbb{N}$ tel que :

$$n \geq N \Rightarrow |u_n| < \varepsilon/2$$

et donc à partir du rang N,

$$|X_n - E(X_n)| < \varepsilon/2 \Rightarrow |X_n - \ell| < \varepsilon, \tag{2.1}$$

car $|X_n - \ell| = |X_n - E(X_n) + E(X_n) - \ell| \leq |X_n - E(X_n)| + |E(X_n) - \ell|$.
L'implication (2.1) peut être encore écrite sous la forme

$$|X_n - \ell| \geq \varepsilon \Rightarrow |X_n - E(X_n)| \geq \varepsilon/2.$$

Par conséquent, en utilisant l'inégalité de Bienaymé-Chebyshev,

$$P(|X_n - \ell| \geq \varepsilon) \leq P(|X_n - E(X_n)| \geq \varepsilon/2) \leq \frac{V(X_n)}{(\varepsilon/2)^2},$$

qui tend vers 0 quand n tend vers l'infini.

Conséquence : Pour que (X_n) converge en probabilité vers X, il suffit que $E(X_n - X) \to 0$ et $V(X_n - X) \to 0$ lorsque $n \to \infty$ (la démonstration passe par l'inégalité de Bienaymé-Chebychev).

2.1.3 Convergence en moyenne quadratique

Définition 5 *Une suite de v.a.r. $(X_n)_{n\in\mathbb{N}}$ converge en moyenne quadratique vers une v.a.r. X si*

$$\lim_{n\to\infty} E((X_n - X)^2) = 0.$$

Propriétés :

1. La convergence en moyenne quadratique entraîne la convergence en probabilité.
2. Pour les (X_n) sont des variables aléatoires d'espérance et de variance finies, si $E(X_n) \to \mu$ et $Var(X_n) \to 0$ alors X_n converge en moyenne quadratique vers μ.

Preuve
1. On applique l'inégalité de Markov avec $Y = |X_n - X|$, $a = \varepsilon^2$ et $g(t) = t^2$. Il suffit ensuite de remarquer que $P(|X_n - X|^2 > \varepsilon^2) = P(|X_n - X| > \varepsilon)$ et utiliser l'hypothèse que $\lim E((X_n - X)^2) = 0$.
2. $\lim E((X_n - \mu)^2) = \lim E(X_n^2) - 2\mu E(X) + \mu^2 = \lim E(X_n^2) - E(X_n)^2 = \lim V(X_n) = 0$.

2.1.4 Loi faible des grands nombres

Théorème 2.1.2 *Soit (X_n) une suite de variables aléatoires indépendantes sur le même espace probabilisé (Ω, P) ayant une même espérance mathématique ℓ et des variances vérifiant $\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} \sigma_i^2 = 0$. On pose $S_n = X_1 + \cdots + X_n$ alors $\frac{S_n}{n}$ converge en probabilité vers ℓ.*

Si on considère une suite de variables aléatoires (X_n) indépendantes définies sur un même espace probabilisé, ayant même espérance et même variance finie notées respectivement $E(X)$ et $V(X)$. La loi faible des grands nombres stipule que, pour tout réel ε strictement positif, la probabilité que la moyenne empirique $\frac{S_n}{n}$ s'éloigne de l'espérance d'au moins ε, tend vers 0 quand n tend vers l'infini. La moyenne $\frac{S_n}{n}$ converge en probabilité vers l'espérance commune $E(X)$.

Preuve On a $E(S_n/n) = \ell$ et $\lim V(S_n/n) = \lim \frac{1}{n^2}\sum \sigma_i^2 = 0$ par hypothèse. Ainsi par le théorème 2.1.1, S_n/n converge en probabilité vers ℓ.

2.2 Convergence en loi

Définition 6 *Soient (X_n) et X des variables aléatoires sur un même espace probabilisé (Ω, P), de fonctions de répartition respectives F_n et F ; on dit que les (X_n) convergent vers X en loi (et on note $X_n \xrightarrow{\mathcal{L}} X$) si en tout point x où F est continue, les $F_n(x)$ convergent vers $F(x)$.*

Propriétés : (admises)

1. La convergence en probabilité entraîne la convergence en loi. $(X_n \xrightarrow{\mathcal{P}} X) \Rightarrow (X_n \xrightarrow{\mathcal{L}} X)$
2. Si les (X_n) et X sont des variables aléatoires discrètes, alors X_n converge en loi vers X si et seulement si
$$\forall x \in \mathbb{R}, \lim_{n\to\infty} P(X_n = x) = P(X = x).$$

Preuve Il s'agit de montrer que si $(X_n)_n$ converge en probabilité vers X, la suite $(F_{X_n})_n$ converge vers F_X (respectivement préalablement notées F_n et F). On utilise le lemme suivant : soient A, B des variables aléatoires réelles, c un réel et $\varepsilon > 0$. Alors on a l'inégalité

$$P(A \le c) \le B(\le c + \varepsilon) + P(|A - B| > \varepsilon),$$

car

$$\begin{aligned} P(A \leq C) &= P(A \leq c \cap B \leq c+\varepsilon) + P(A \leq c \cap B > c+\varepsilon) \\ &= P(A \leq c | B \leq c+\varepsilon) \cdot P(B \leq c+\varepsilon) + P(A \leq c \cap B - \varepsilon > c) \\ &\leq P(B \leq c+\varepsilon) + P(A - B > -\varepsilon) \quad \text{car } P(\cdot|\cdot) \leq 1 \\ &\leq P(B \leq c+\varepsilon) + P(|A-B| > \varepsilon) \\ &\quad \text{car } P(|A-B| > \varepsilon) = P(A-B > \varepsilon) + P(A-B < -\varepsilon) \geq P(A-B < -\varepsilon) \end{aligned}$$

De ce lemme, il vient respectivement pour $(A = X_n,\, c = x,\, B = X)$ puis $(A = X,\, c = x - \varepsilon,\, B = X_n)$

$$P(X_n \leq x) \leq P(X \leq x+\varepsilon) + P(|X_n - X| > \varepsilon) \tag{2.2}$$

$$P(X_n \leq x) \geq P(X \leq x-\varepsilon) + P(|X_n - X| > \varepsilon) \tag{2.3}$$

Passons à la démontration proprement dite. Soit x un point où F est continue. Soit $\eta > 0$. Par continuité de F_X en x, il existe $\varepsilon > 0$ tel que $|F_X(x+\varepsilon) - F_X(x)| < \eta/2$ et $|F_X(x-\varepsilon) - F_X(x)| < \eta/2$. Pour cet ε, de part la convergence de $(X_n)_n$ vers X, il existe n_0 tel que, pour tout $n \geq n_0$,

$$P(|X_n - X| > \varepsilon) < \eta/2.$$

Ainsi par (2.2),

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\leq F_X(x+\varepsilon) + P(|X_n - X| > \varepsilon) - F_X(x) \\ &\leq F_X(x+\varepsilon) - F_X(x) + P(|X_n - X| > \varepsilon) < \eta/2 + \eta/2 = \eta \end{aligned}$$

et par (2.3),

$$\begin{aligned} F_{X_n}(x) - F_X(x) &\geq F_X(x-\varepsilon) - F_X(x) - P(|X_n - X| > \varepsilon) \\ &\geq -\eta/2 - \eta/2 = -\eta \end{aligned}$$

Donc $\forall \eta > 0, \exists n_0$ tel que $\forall n \geq n_0, |F_{X_n}(x) - F_X(x)| < \eta$.

Proposition 2.2.1 (Convergence de la loi hypergéométrique vers la loi binomiale) *Soit (X_N) une suite de variables aléatoires sur un même espace probabilisé, de loi hypergéométrique : $X_N \hookrightarrow \mathcal{H}(N, n, p)$ où n et p sont supposés constants. Alors (X_N) convergent en loi, quand N tend vers l'infini, vers X de loi binomiale $\mathcal{B}(n, p)$ (mêmes valeurs de paramètres).*

Preuve La probabilité ponctuelle de X_N est

$$P(X_N = k) = \frac{C_{Np}^k C_{Nq}^{n-k}}{C_N^n}.$$

Lorsque N tend vers l'infini avec n constant,

$$C_N^n = \frac{N(N-1)\cdots(N-n+1)}{n!} = N^n(1 - \frac{1}{N})\cdots(1 - \frac{n-1}{N})\frac{1}{n!} \equiv \frac{N^n}{n!}$$

car $(1 - \frac{m}{N}) \equiv 1$ lorsque N tend vers l'infini. De même, lorsque N tend vers l'infini avec p et k fixes, alors

$$C_{Np}^k \equiv \frac{(Np)^k}{k!} \quad \text{et} \quad C_{N(1-p)}^{n-k} \equiv \frac{(N(1-p))^{n-k}}{(n-k)!}.$$

Finalement,

$$P(X_N = k) \equiv \frac{p^k(1-p)^{n-k}n!}{k!(n-k)!} = C_n^k p^k (1-p)^{n-k},$$

ce qui correspond à la probabilité ponctuelle d'une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n, p)$.

C'est pour cela que lorsque la population (de taille N) est très grande, on peut assimiler la loi d'une variable aléatoire comptant le nombre de réussite sur un tirage sans remise (loi hypergéométrique) à une loi binomiale (tirage avec remise).

Proposition 2.2.2 (Convergence de la loi binomiale vers une loi de Poisson) *Soit (X_n) une suite de variables aléatoires binomiales sur un même espace probabilisé : pour tout n, X_n suit $\mathcal{B}(n,p_n)$. On suppose que $\lim_{n\to+\infty} p_n = 0$ et $\lim_{n\to+\infty} np_n = \lambda$. Alors (X_n) convergent en loi, quand n tend vers l'infini, vers une loi de Poisson de paramètre λ.*

Preuve Pour k fixé,

$$\begin{aligned} P(X_n = k) &= \frac{n(n-1)\cdots(n-k+1)}{k!} p_n^k (1-p_n)^{n-k} \\ &= \frac{(np_n)^k}{k!}(1-p_n)^n (1-\frac{1}{n})\cdots(1-\frac{k-1}{n})(1-p_n)^{-k} \end{aligned}$$

On cherche la limite de $(1-p_n)^n = \exp(n\ln(1-p_n)) = \exp(n\ln(1-np_n/n))$. Comme $\lim_{n\to+\infty} np_n = \lambda$, on pose $np_n = \lambda+\varepsilon_n$ avec $\lim_{n\to+\infty}\varepsilon_n = 0$ et ainsi $\ln(1-np_n/n) \sim_\infty -\lambda/n$ donc $\lim_{n\to+\infty}(1-p_n)^n = e^{-\lambda}$. Comme k est fixé, $\lim_{n\to+\infty}(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})(1-p_n)^{-k} = 1$

Ainsi

$$\lim_{n\to+\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

ce qui correspond à la probabilité ponctuelle d'une variable aléatoire qui suit une loi de Poisson $\mathcal{P}(\lambda)$. Il s'agit donc d'une convergence en loi en appliquant le point 2 des propriétés.

Corollaire 2.2.3 (Application pratique) *On peut remplacer $\mathcal{B}(n,p)$ par $\mathcal{P}(\lambda)$ avec $\lambda = np$ pour n très grand ($n > 50$) et p très petit ($p < 0,1$).*

2.3 Convergence des fonctions caractéristiques

2.3.1 Continuité

Théorème 2.3.1 (théorème de continuité de Levy) *Soit (X_n) une suite de variables aléatoires de fonctions caractéristiques φ_{X_n} et X une variable aléatoire de fonction caractéristique φ_X, toutes sur un même espace probabilisé. Si les (X_n) convergent en loi vers X alors la suite de fonctions (φ_{X_n}) converge uniformément vers φ_X sur tout intervalle $[-a,a]$.*

Inversement si les (φ_{X_n}) convergent vers une fonction φ dont la partie réelle est continue en 0, alors φ est la fonction caractéristique d'une variable aléatoire X vers laquelle les X_n convergent en loi.

On peut le résumer ainsi :

$$\{\forall t\in\mathbb{R}; \varphi_{X_n}(t)\to\varphi_X(t)\} \Leftrightarrow \{X_n \xrightarrow{\mathcal{L}} X\}$$

2.3.2 Théorème central limite

Corollaire 2.3.2 (Théorème central limite) *Soit une suite (X_n) de variables aléatoires définies sur le même espace de probabilité, suivant la même loi D et dont l'espérance μ et l'écart-type σ communes existent et soient finis ($\sigma \neq 0$). On suppose que les (X_n) sont indépendantes. Considérons la somme $S_n = X_1 + \cdots + X_n$. Alors l'espérance de S_n est $n\mu$ et son écart-type vaut $\sigma\sqrt{n}$ et $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converge en loi vers une variable aléatoire normale centrée réduite.*

Preuve Posons $Y_i = \frac{X_i - \mu}{\sigma\sqrt{n}}$. Alors

$$\varphi_{Y_i}(t) = \varphi_{\frac{X_i-\mu}{\sigma\sqrt{n}}}(t) = \varphi_{X_i-\mu}(\frac{t}{\sigma\sqrt{n}})$$

Pour t fixé, lorsque n tend vers l'infini, $\frac{t}{\sigma\sqrt{n}}$ est infiniment petit. Ecrivons le développement limité, au voisinage de 0, de la fonction caractéristique d'une variable aléatoire W :

$$\begin{aligned}\varphi_W(u) &= \varphi_W(0) + u\ \varphi'_W(0) + \frac{u^2}{2}\varphi''_W(0) + u^2\varepsilon(u) \\ &= 1 + i\ u\ E(W) - \frac{u^2}{2}E(W^2) + u^2\varepsilon(u)\end{aligned}$$

En posant $W = X_i - \mu$, $u = t/(\sigma\sqrt{n})$, on a $E(W) = E(X_i - \mu) = 0$ et $E(W^2) = E((X_i - \mu)^2) = V(X_i) = \sigma^2$ d'où

$$\varphi_{X_i-\mu}(\frac{t}{\sigma\sqrt{n}}) = 1 - \frac{t^2}{2\sigma^2 n}\sigma^2 + \frac{1}{n}\varepsilon(t^3/\sigma^3\sqrt{n}) = 1 - \frac{t^2}{2n} + \frac{1}{n}\varepsilon_i(n)$$

avec $\lim_{n\to+\infty}\varepsilon_i(n) = 0$.

Maintenant, posons $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^n Y_i$.

L'indépendance des X_n entraîne celle des Y_i et ainsi

$$\begin{aligned}\varphi_{Z_n}(t) &= \prod_{i=1}^n \varphi_{Y_i}(t) \\ &= \exp\left(\sum_{i=1}^n \ln(1 - \frac{t^2}{2n} + \frac{1}{n}\varepsilon_i(n))\right)\end{aligned}$$

et $\lim_{n\to+\infty}\varphi_{Z_n}(t) = e^{-t^2/2}$ qui est la fonction caractéristique de $\mathcal{N}(0,1)$.

Ce théorème établit une propriété générale, qui va justifier l'importance considérable de la loi normale, à la fois comme modèle pour décrire des situations pratiques, mais aussi comme outil théorique. Il s'énonce ainsi :

« Soit $X_1, ..., X_i, ..., X_n$, une suite de n variables aléatoires indépendantes, de moyennes $\mu_1, ..., \mu_i, ..., \mu_n$, et de variances ${s_1}^2, ..., {s_i}^2, ..., {s_n}^2$, et de lois de probabilité quelconques, leur somme suit une loi qui, lorsque n augmente, tend vers une loi normale de moyenne $\mu = \sum_{i=1}^n \mu_i$ et de variance $s^2 = \sum_{i=1}^n {s_i}^2$. Il y a une seule condition restrictive, c'est que les variances soient finies et qu'aucune ne soit prépondérante devant les autres. »

La loi normale comme modèle : prenons l'exemple du fonctionnement d'un tour d'usinage du bois. Le réglage du tour a pour but d'obtenir des pièces présentant une cote bien définie ; mais on sait que de multiples causes perturbatrices agissent au cours de l'usinage d'une pièce : vibrations, usures, variations de courant ... Or si les causes perturbatrices sont nombreuses, si leurs effets interviennent de façon additive, enfin si la dispersion provoquée par chacune d'elles reste faible par rapport à la dispersion totale, alors le théorème central limite signifie qu'on doit observer une fluctuation globale très voisine de la loi normale. Et, comme ce mécanisme d'intervention de causes perturbatrices est très répandu dans la nature, il en résulte que la loi normale occupe en statistique une place privilégiée.

2.3.3 convergence de $\mathcal{P}$ vers $\mathcal{N}$

Corollaire 2.3.3 *Soit (X_n) une suite de variables aléatoires suivants des lois de Poisson de paramètres λ_n. Si $\lim_{n\to+\infty}\lambda_n = \infty$, alors $\frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$ converge en loi vers $\mathcal{N}(0,1)$.*

Preuve On utilise la fonction caractéristique de la loi de Poisson de paramètre λ :

$$\varphi_X(t) = e^{\lambda(\cos t + i\ \sin t - 1)}.$$

En utilisant les propriétés de la fonction caractéristique ($\varphi_{aX}(t) = \varphi(at)$ et $\varphi_{X+b}(t) = e^{itb}\varphi_X(t)$), il vient $\varphi_{X-\lambda}(t) = e^{-it\lambda}e^{\lambda(\cos t + i\ \sin t - 1)}$ puis $\varphi_{\frac{X-\lambda}{\sqrt{\lambda}}}(t) = e^{\lambda(\cos \frac{t}{\sqrt{\lambda}} + i\ \sin \frac{t}{\sqrt{\lambda}} - 1)}e^{i\frac{t}{\sqrt{\lambda}}(-\lambda)}$. Or, lorsque λ tend vers l'infini, $1/\lambda$ est au voisinage de 0 et

$$\begin{array}{ll} \cos(t/\sqrt{\lambda}) & \sim 1 - \frac{(t/\sqrt{\lambda})^2}{2} + \frac{1}{\lambda}\varepsilon(\lambda) \\ \sin(t/\sqrt{\lambda}) & \sim (t/\sqrt{\lambda}) + \frac{1}{\lambda}\varepsilon(\lambda) \end{array}$$

avec $\lim_{\lambda\to\infty} \varepsilon(\lambda) = 0$. Ou encore le développement de l'exposant avec $1/\lambda$ au voisinage de 0 est

$$e^{it/\sqrt{\lambda}} - 1 = \frac{it}{\sqrt{\lambda}} + \frac{(it)^2}{2\lambda} + \frac{1}{\lambda}\varepsilon(\lambda).$$

Ainsi

$$\lambda(\cos(t/\sqrt{\lambda}) + i\sin(t/\sqrt{\lambda}) - 1) - i\sqrt{\lambda}t \sim -t^2/2$$

et $\varphi_{\frac{X-\lambda}{\sqrt{\lambda}}}(t) \sim e^{-t^2/2}$, fonction caractéristique de $\mathcal{N}(0,1)$.

Application pratique : Pour λ suffisamment grand (disons $\lambda > 1000$), la distribution normale de moyenne λ et de variance λ est une excellente approximation de la distribution de Poisson de paramètre λ. Si λ est plus grand que 10, alors la distribution normale est une bonne approximation si une correction de continuité est appliquée, c'est-à-dire $P(X \leq x)$ lorsque x est un entier positif ou nul est remplacé par $P(X \leq x + 0,5)$.

2.3.4 convergence de $\mathcal{B}$ vers $\mathcal{N}$

Corollaire 2.3.4 (Théorème de Moivre-Laplace) *Soit (X_n) une suite de variables aléatoires telles que $(X_n) \in \mathcal{B}(n,p)$. Alors $\frac{X_n - np}{\sqrt{npq}}$ converge en loi vers la variable centrée réduite $Z \in \mathcal{N}(0,1)$ ou encore X_n converge en loi vers $\mathcal{N}(np, \sqrt{npq})$.*

Preuve On rappelle que l'on a défini une variable de Bernoulli comme une variable qui prend la valeur 1 avec la probabilité p, et la valeur 0 avec la probabilité $(1-p)$, et montré que sa moyenne est égale à p et sa variance à $p(1-p)$. Or on peut considérer une variable binomiale comme la somme de n variables de Bernoulli. Il résulte du théorème central limite que, si n est suffisamment grand (en pratique à partir de $n = 50$), la loi binomiale peut être approximée par une loi normale de moyenne np et de variance $np(1-p)$. C'est pourquoi les tables de la loi binomiale s'arrêtent généralement à $n = 50$.

Application pratique : on peut assimiler une loi binomiale à une loi normale dès que $np > 15$ et $nq > 15$ ou $n > 30, np > 5, nq > 5$.

2.3.5 Correction de continuité

Pour un meilleur résultat, une correction de continuité peut être appliquée, c'est-à-dire $P(X \leq x)$ lorsque x est un entier positif ou nul est remplacé par $P(X \leq x + 0,5)$. Cela permet également de différencier $P(X \leq x)$ de $P(X < x)$ lorsque l'on approche une loi discrète par une loi continue.

Chapitre 3

Echantillonnage, Estimations

3.1 Echantillonnage

Nous allons étudier comment se comporte un échantillon (éléments pris au hasard) dans une population dont on connaît les caractéristiques statistiques (lois,...) d'une variable considérée X. Dans ce cas, prendre un échantillon aléatoire de taille n consiste à considérer n réalisations de X ou encore considérer n variables aléatoires $X_1, \ldots, X_n$ indépendantes, de même loi que X.

Définition 7 *Soit X une variable aléatoire sur un référentiel Ω. Un **échantillon** de X de taille n est un n-uplet $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes de même loi que X. La loi de X sera appelée **loi mère**. Une réalisation de cet échantillon est un n-uplet de réels $(x_1, \ldots, x_n)$ où $X_i(\omega) = x_i$.*

3.1.1 Moyenne et variance empiriques

Définition 8 *On appelle **statistique** sur un n-échantillon une fonction de $(X_1, \ldots, X_n)$.*

Définition 9 *On appelle moyenne de l'échantillon ou **moyenne empirique**, la statistique notée $\overline{X}$ définie par*

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Proposition 3.1.1 *Soit X une variable aléatoire de moyenne μ et d'écart-type σ. On a :*

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}.$$

De plus, par le théorème central limite, $\overline{X}$ converge en loi vers $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ lorsque n tend vers l'infini.

Preuve

$$E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu.$$

Et, en raison de l'indépendance des X_i,

$$V\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Théorème 3.1.2 *Toute somme de variables aléatoires normales indépendantes est une variable aléatoire normale. Ainsi, si* $X \hookrightarrow \mathcal{N}(\mu, \sigma)$ *alors pour toute valeur de* n, $\overline{X} \hookrightarrow \mathcal{N}(\mu, \sigma/\sqrt{n})$.

Preuve Il suffit de démontrer le résultat avec deux variables aléatoires, l'extension se faisant de proche en proche. On suppose X_1 et X_2 indépendantes de lois respectives $\mathcal{N}(\mu_1, \sigma_1)$ et $\mathcal{N}(\mu_2, \sigma_2)$. On obtient le résultat sur la somme en utilisant les fonctions caractéristiques (voir cours S3).

Définition 10 *On appelle* ***Variance empirique****, la statistique notée* $\tilde{S}^2(X)$ *définie par*

$$\tilde{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Proposition 3.1.3 *Soit* X *une variable aléatoire d'écart-type* σ *et de moment centré d'ordre 4,* μ_4. *On a :*

$$E(\tilde{S}^2) = \frac{n-1}{n}\sigma^2, \qquad V(\tilde{S}^2) = \frac{n-1}{n^3}\left((n-1)\mu_4 - (n-3)\sigma^4\right).$$

De plus, lorsque n *tend vers l'infini,* $V(\tilde{S}^2) \sim \frac{\mu_4 - \sigma^3}{n}$.

Preuve

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{2}{n}(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + (\overline{X} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)^2 + (\overline{X} - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2
\end{aligned}$$

D'où

$$E(\tilde{S}^2) = \frac{1}{n}\sum_{i=1}^{n}V(X_i) - V(\overline{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Preuve Démontrons l'autre égalité. On rappelle les notations : les X_i suivent la loi normale $\mathcal{N}(\mu, \sigma)$ et les moments centrés d'ordre k sont définis par

$$\mu_k = E((X - \mu)^k).$$

Ainsi $\mu_1 = 0$ et $\mu_2 = \sigma^2$. On peut écrire $\tilde{S}^2$ sous la forme

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2. \tag{3.1}$$

D'autre part,

$$\begin{aligned}
\sum_{i,j}(X_i - X_j)^2 &= \sum_{i,j}(X_i^2 - 2X_i\ X_j + X_j^2) \\
&= \sum_{i,j} X_i^2 - 2\sum_i\sum_j X_i\ X_j + \sum_i\sum_j X_j^2 \\
&= 2n\sum_{i=1}^{n} X_i^2 - 2\sum_i X_i \sum_j X_j \\
&= 2n\sum_{i=1}^{n} X_i^2 - 2(n\overline{X})(n\overline{X}) \\
\sum_{i,j}(X_i - X_j)^2 &= 2n^2\tilde{S}^2 \quad \text{par (3.1).}
\end{aligned}$$

On peut donc calculer la variance de $\tilde{S}^2$ en utilisant la relation suivante :

$$Var(\tilde{S}^2) = cov(\tilde{S}^2, \tilde{S}^2) = \frac{1}{(2n^2)^2}\sum_{i,j,k,l} cov((X_i - X_j)^2, (X_k - X_l)^2).$$

On calcule alors les différentes covariances selon la forme des facteurs :
– de la forme $cov((X_i - X_j)^2, (X_k - X_l)^2)$ avec i, j, k, l tous différents,
– de la forme $cov((X_i - X_j)^2, (X_k - X_j)^2)$ avec i, j, k différents,
– de la forme $cov((X_i - X_j)^2, (X_i - X_j)^2)$ avec i, j différents.
On remarque que si $i = j$ ou $k = l$, alors on obtient une covariance avec zéro (de la forme $cov(0, (X_k - X_l)^2)$) ou $cov((X_i - X_j)^2, 0)$ qui est nulle.

Commençons par le calcul de $cov((X_i - X_j)^2, (X_i - X_j)^2)$ avec $i \neq j$.

$$cov((X_i - X_j)^2, (X_i - X_j)^2) = E((X_i - X_j)^4) - [E((X_i - X_j)^2)]^2.$$

On introduit la moyenne μ dans le calcul de l'espérance :

$$\begin{aligned}
(X_i - X_j)^4 &= [(X_i - \mu) - (X_j - \mu)]^4 \\
&= (X_i - \mu)^4 - 4(X_i - \mu)(X_j - \mu)^3 + 6(X_i - \mu)^2(X_j - \mu)^2 - 4(X_i - \mu)^3(X_j - \mu) \\
&\quad +(X_j - \mu)^4 \\
E\left((X_i - X_j)^4\right) &= 2\mu_4 - 8\mu_3\mu_1 + 6\mu_2^2 \\
&= 2\mu_4 + 6\sigma^4 \quad \text{car } \mu_1 = 0 \text{ et } \mu_2 = \sigma^2.
\end{aligned}$$

$$\begin{aligned}
(X_i - X_j)^2 &= [(X_i - \mu) - (X_j - \mu)]^2 \\
&= (X_i - \mu)^2 - 2(X_i - \mu)(X_j - \mu) + (X_j - \mu)^2 \\
E\left((X_i - X_j)^2\right) &= 2\mu_2 = 2\sigma^2.
\end{aligned}$$

Ainsi, pour $i \neq j$,

$$cov((X_i - X_j)^2, (X_i - X_j)^2) = 2\mu_4 + 2\sigma^4.$$

Continuons par le calcul de $cov((X_i - X_j)^2, (X_k - X_j)^2)$ avec i, j, k différents.

$$\begin{aligned}
cov((X_i - X_j)^2, (X_k - X_j)^2) &= E((X_i - X_j)^2(X_k - X_j)^2) - [E((X_i - X_j)^2)E((X_k - X_j)^2] \\
&= E((X_i - X_j)^2(X_k - X_j)^2) - (2\sigma^2)^2.
\end{aligned}$$

$$\begin{aligned}
&(X_i - X_j)^2(X_k - X_j)^2 \\
&= \left[(X_i-\mu)^2 - 2(X_i-\mu)(X_j-\mu) + (X_j-\mu)^2\right]\left[(X_k-\mu)^2 - 2(X_k-\mu)(X_j-\mu) + (X_j-\mu)^2\right] \\
&= (X_i-\mu)^2(X_k-\mu)^2 - 2(X_i-\mu)(X_j-\mu)(X_k-\mu)^2 + (X_j-\mu)(X_k-\mu)^2 \\
&\quad -2(X_i-\mu)^2(X_k-\mu)(X_j-\mu) + 4(X_i-\mu)(X_k-\mu)(X_j-\mu)^2 - 2(X_k-\mu)(X_j-\mu)^3 \\
&\quad +(X_i-\mu)^2(X_j-\mu)^2 - 2(X_i-\mu)(X_j-\mu)^3 + (X_j-\mu)^4 \\
&= 3(\mu_2)^2 + \mu_4
\end{aligned}$$

Ainsi, pour i, j, k différents,

$$cov((X_i - X_j)^2, (X_k - X_j)^2) = \mu_4 - \sigma^4.$$

Le dernier cas est rapidement calculé : si i, j, k, l sont différents, alors, par indépendance des X_i,

$$cov((X_i - X_j)^2, (X_k - X_l)^2) = 0.$$

Il reste à compter le nombre de termes dans chaque cas présenté.
- $cov((X_i - X_j)^2, (X_k - X_l)^2)$ est un terme de la forme $cov(X_i - X_j)^2, (X_i - X_j)^2)$ lorsque $(k = i, l = j)$ ou $(k = j, l = i)$ avec $i \neq j$, soit $2n(n-1)$ termes.
- $cov((X_i - X_j)^2, (X_k - X_l)^2)$ est un terme de la forme $cov(X_i - X_j)^2, (X_k - X_j)^2)$ lorsque ($l = j$ ou $l = i$) et k, i, j différents ou ($k = i$ ou $k = j$) et l, i, j différents, soit $(2+2)n(n-1)(n-2)$ termes.

$$\begin{aligned}
\sum_{i,j,k,l} cov &= 2n(n-1)(2\mu_4 + 2\sigma^4) + 4n(n-1)(n-2)(\mu_4 - \sigma^4) \\
&= 4n(n-1)^2\left[\mu_4 - \frac{n-3}{n-1}\sigma^4\right]
\end{aligned}$$

Corollaire 3.1.4 $\sqrt{n}\frac{\tilde{S}^2-\sigma^2}{\sqrt{\mu_4-\sigma^4}}$ *converge en loi vers* $\mathcal{N}(0,1)$ *lorsque* n *tend vers l'infini.*

3.1.2 Fréquence

Soit $(X_i)_{i=1..n}$ un échantillon aléatoire de taille n ayant une loi de Bernoulli de paramètre p comme loi mère. Alors

$$F = \frac{X_1 + \cdots + X_n}{n}$$

est la fréquence de la valeur 1 dans l'échantillon et nF suit une loi binomiale de paramètres n et p. Ainsi

$$E(F) = p \quad \text{et} \quad Var(F) = \frac{pq}{n}.$$

Donc, quand n tend vers l'infini, F converge en loi vers $\mathcal{N}(p, \sqrt{\frac{pq}{n}})$.

En effet,

$$E(F) = E\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = p.$$

$$Var(F) = Var\left(\frac{1}{n}\sum X_i\right) \overset{(ind)}{=} \frac{1}{n^2}\sum Var(X_i) = \frac{npq}{n^2} = \frac{pq}{n}.$$

On peut aussi recalculer la variance par le théorème de König :

$$\begin{aligned}
Var(F) &= E(F^2) - E(F)^2 \\
&= E\left(\left[\frac{1}{n}\sum X_i\right]^2\right) - p^2 \\
&= \frac{1}{n^2}E\left(\left[\sum_i X_i^2 + \sum_i \sum_{j\neq i} X_i\ X_j\right]\right) - p^2 \\
&= \frac{1}{n^2}\left[\sum_i E(X_i^2) + \sum_i \sum_{j\neq i} E(X_i\ X_j)\right] - p^2 \\
&\overset{(ind)}{=} \frac{1}{n^2}\left[\sum_i E(X_i^2) + \sum_i \sum_{j\neq i} E(X_i)\ E(X_j)\right] - p^2 \\
&= \frac{1}{n^2}\left[np + n(n-1)p^2\right] - p^2 \quad \text{car } E(X_i^2) = 0^2 * (1-p) + 1^2 * p = p, \\
Var(F) &= \frac{p(1-p)}{n}
\end{aligned}$$

Exercice : Montrer que $E(\ F(1-F)\ ) = pq(1-1/n)$.

3.2 Estimation paramétrique ponctuelle

Cette fois il s'agit d'estimer certaines caractéristiques statistiques de la loi (moyenne, variance, fonction de répartition) au travers d'une série d'observations $x_1, x_2, \ldots, x_n$. C'est la problématique inverse de l'échantillonnage.

À partir des caractéristiques d'un échantillon, que peut-on déduire des caractéristiques de la population dont il est issu ?

L'estimation consiste à donner des valeurs approximatives aux paramètres d'une population à l'aide d'un échantillon de n observations issues de cette population. On peut se tromper sur la valeur exacte, mais on donne la "meilleure valeur" possible que l'on peut supposer.

3.2.1 Estimateur ponctuel

On souhaite estimer un paramètre θ d'une population (cela peut être sa moyenne μ, son écart-type σ, une proportion p. Un estimateur de θ est une statistique T (donc une fonction de $(X_1, \ldots, X_n)$) dont la réalisation est envisagée comme une "bonne valeur" du paramètre θ. On parle d'estimation de θ associée à cet estimateur la valeur observée lors de l'expérience, c'est-à-dire la valeur prise par la fonction au point observé $(x_1, \ldots, x_n)$.

Exemple : pour estimer l'espérance $E(X)$ de la loi de X, un estimateur naturel est la moyenne empirique $\overline{X}$ qui produit une estimation $\overline{x}$, moyenne descriptive de la série des valeurs observées.

3.2.2 Qualité d'un estimateur

Définition 11 *On appelle **biais** de T pour θ la valeur*

$$b_\theta(T) = E(T) - \theta.$$

*Un estimateur T est dit **sans biais** si $E(T) = \theta$.*

Définition 12 *Un estimateur T est dit **convergent** si $E(T)$ tend vers θ lorsque n tend vers l'infini. Il sera dit **consistant** si T converge en probabilité vers θ lorsque n tend vers l'infini.*

Théorème 3.2.1 *Si T est convergent et de variance tendant vers 0 lorsque n tend vers l'infini alors T est consistant.*

Preuve On a, pour tous réels θ et $\alpha > 0$,

$$|T - \theta| > \alpha \Rightarrow |T - E(T)| > \alpha - |\theta - E(T)|.$$

Si $\lim E(T) = \theta$, alors à partir d'un certain rang N, on a $|\theta - E(T)| < \frac{\alpha}{2}$. Ainsi

$$\begin{aligned} P(|T-\theta|) > \alpha) &\leq P(|T - E(T)| > \alpha - |\theta - E(T)|) \\ &\leq P(|T - E(T)| > \alpha/2) \\ &\leq \frac{4}{\alpha^2}\mathrm{Var}(T) \quad \text{(par Bienaymé-Chebishev)} \end{aligned}$$

borne supérieure qui tend vers 0 lorsque n tend vers l'infini.

De façon générale, on peut écrire

$$T - \theta = (T - E(T)) + (E(T) - \theta)$$

ainsi
– la grandeur $T - E(T)$ représente les fluctuations de T autour de sa moyenne
– et $E(T) - \theta$ représente l'erreur systématique (biais).

Définition 13 *La qualité d'un estimateur se mesure également par l'**erreur quadratique moyenne** (ou risque quadratique) définie par $E((T - \theta)^2)$.*

Théorème 3.2.2 *Soit T un estimateur du paramètre θ à étudier. On a :*

$$E((T - \theta)^2) = Var(T) + [E(T) - \theta]^2.$$

Preuve

$$\begin{aligned} E([T-\theta]^2) &= E([T - E(T) + E(T) - \theta]^2) \\ &= E([T - E(T)]^2) + E([E(T) - \theta]^2) + 2E([T - E(T)][E(T) - \theta]) \\ &= Var(T) + (E(T) - \theta)^2 \quad \text{car } E(T - E(T)) = 0. \end{aligned}$$

Remarque : Entre deux estimateurs sans biais, le "meilleur" sera celui dont la variance est minimale (on parle d'**efficacité**).

Remarque : Le critère d'erreur quadratique moyenne n'est pas parfait mais il est préféré à d'autres critères qui semblent plus naturels comme l'erreur absolue moyenne $E(|T - \theta|)$ car il s'exprime en fonction de notions simples comme le biais et la variance et est relativement facile à manipuler analytiquement.

3.2.3 Quelques estimateurs classiques

1. $\overline{X}$ est un estimateur sans biais de la moyenne μ. Son estimation $\overline{x}$ est la moyenne observée dans une réalisation de l'échantillon.
2. $\tilde{S}^2$ est un estimateur consistant de σ^2 (mais biaisé).
3. $S^2 = \frac{n}{n-1}\tilde{S}^2$ est un estimateur sans biais et consistant de σ^2. Son estimation est $s^2 = \frac{n}{n-1}\sigma_e^2$ où σ_e est l'écart-type observé dans une réalisation de l'échantillon.
4. Si p est la fréquence d'un caractère, F constitue un estimateur sans biais et consistant de p. Son estimation est notée f.

Remarque : Si la moyenne μ de X est connue, $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ est un meilleur estimateur de σ^2 que S^2. (Preuve en TD)

3.2.4 Estimation par la méthode du maximum de vraisemblance

Soit X une variable aléatoire réelle de loi paramétrique (discrète ou continue), dont on veut estimer le paramètre θ. Alors on définit une fonction f telle que :

$$f(x;\theta) = \begin{cases} f_\theta(x) & \text{si } X \text{ est une v.a. continue de densité } f \\ P_\theta(X = x) & \text{si } X \text{ est une v.a. discrète de probabilité ponctuelle } P \end{cases}$$

Définition 14 *On appelle* ***fonction de vraisemblance*** *de θ pour une réalisation $(x_1, \ldots, x_n)$ d'un échantillon, la fonction de θ :*

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

Définition 15 *La méthode consistant à estimer θ par la valeur qui maximise L (vraisemblance) s'appelle méthode du maximum de vraisemblance.*

$$\hat{\theta} = \{\theta \ / \ L(\hat{\theta}) = \sup_{\theta} L(\theta)\}.$$

Ceci est un problème d'optimisation. On utilise généralement le fait que si L est dérivable et si L admet un maximum global en une valeur, alors la dérivée première s'annule en et que la dérivée seconde est négative.

Réciproquement, si la dérivée première s'annule en $\theta = \hat{\theta}$ et que la dérivée seconde est négative en $\theta = \hat{\theta}$, alors $\hat{\theta}$ est un maximum local (et non global) de $L(x_1, ..., x_i, ..., x_n; \theta)$. Il est alors nécessaire de vérifier qu'il s'agit bien d'un maximum global. La vraisemblance étant positive et le logarithme népérien une fonction croissante, il est équivalent et souvent plus simple de maximiser le logarithme népérien de la vraisemblance (le produit se transforme en somme, ce qui est plus simple à dériver).

Ainsi en pratique :

1. La condition nécessaire

$$\frac{\partial L(x_1, \cdots, x_n; \theta)}{\partial \theta} = 0 \quad \text{ou} \quad \frac{\partial \ln L(x_1, \cdots, x_n; \theta)}{\partial \theta} = 0$$

   permet de trouver la valeur $\hat{\theta}$.
2. $\theta = \hat{\theta}$ est un maximum local si la condition suffisante est remplie au point critique :

$$\frac{\partial^2 L(x_1, \cdots, x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0 \quad \text{ou} \quad \frac{\partial^2 \ln L(x_1, \cdots, x_n; \theta)}{\partial \theta^2}(\hat{\theta}) \leq 0.$$

Exemple 1 : Avec une loi discrète

On souhaite estimer le paramètre λ d'une loi de Poisson à partir d'un n-échantillon. On a $f(x;\lambda) = P_\lambda(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$. La fonction de vraisemblance s'écrit

$$L(x_1,\dots,x_n;\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-\lambda n}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}.$$

Il est plus simple d'utiliser le logarithme, la vraisemblance étant positive :

$$\ln L(x_1,\dots,x_n;\lambda) = \ln e^{-\lambda n} + \ln\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!} = -\lambda n + \sum_{i=1}^{n}\ln\frac{\lambda^{x_i}}{x_i!} = -\lambda n + \ln\lambda\cdot\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\ln(x_i!).$$

La dérivée première

$$\frac{\partial \ln L(x_1,\cdots,x_n;\theta)}{\partial\theta} = -n + \frac{\sum_{i=1}^{n}x_i}{\lambda}$$

s'annule pour $\hat\lambda = \frac{\sum_{i=1}^{n}x_i}{n}$. La dérivée seconde

$$\frac{\partial^2 \ln L(x_1,\cdots,x_n;\theta)}{\partial\theta^2} = -\frac{\sum_{i=1}^{n}x_i}{\lambda^2}$$

est toujours négative ou nulle. Ainsi l'estimation donnée par $\Lambda = \frac{\sum_{i=1}^{n}X_i}{n} = \overline{X}$ conduit à un estimateur du maximum de vraisemblance égal à $\hat\lambda = \overline{x}$. Il est normal de retrouver la moyenne empirique qui est le meilleur estimateur possible pour le paramètre λ (qui représente aussi l'espérance d'une loi de Poisson).

Exemple 2 : Avec une loi continue

On souhaite estimer les paramètres μ et σ d'une loi normale à partir d'un n-échantillon.

La loi normale $\mathcal{N}(\mu,\sigma)$ a pour fonction densité

$$f(x;\mu,\sigma) = f_{(\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Ecrivons la fonction de vraisemblance pour une réalisation d'un échantillon de n variables indépendantes :

$$f(x_1,\dots,x_n;\mu,\sigma) = \prod_{i=1}^{n} f(x_i;\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

Or (théorème de König) $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2$, où $\overline{x}$ représente la moyenne de l'échantillon. Ainsi la fonction de vraisemblance peut être écrite sous la forme

$$f(x_1,\dots,x_n;\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right).$$

$$\frac{\partial}{\partial\mu}\ln L = \frac{\partial}{\partial\mu}\left(\ln\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} - \left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right)\right) = 0 - \frac{-2n(\overline{x}-\mu)}{2\sigma^2}$$

On obtient donc l'estimateur par le maximum de vraisemblance de l'espérance :

$$\hat\mu = \overline{x} = \sum_{i=1}^{n}x_i/n.$$

Pour le second paramètre, on calcule

$$\frac{\partial}{\partial\sigma}\ln L = \frac{\partial}{\partial\sigma}\left(\frac{n}{2}\ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{\sigma^3}$$

Donc

$$\hat{\sigma}^2 = \sum_{i=1}^n (x_i-\hat{\mu})^2/n$$

que l'on peut traduire par

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2.$$

On vérifie que c'est bien des maxima locaux :

$$\frac{\partial^2 \ln L}{\partial\mu^2} = -n/\sigma^2 \leq 0$$

$$\frac{\partial^2 \ln L}{\partial\sigma^2} = n/\sigma^2 - \frac{3}{\sigma^4}(\sum (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2).$$

Au point $\hat{\sigma}$,

$$\frac{\partial^2 \ln L}{\partial\sigma^2}(\hat{\sigma}) = n/\hat{\sigma}^2 - \frac{3}{\sigma^4}(n\hat{\sigma}^2 + n(\overline{x}-\mu)^2) \leq 0.$$

La méthode fournit un estimateur non biaisé de la moyenne ($E(\hat{\mu}) = \mu$) mais par contre, l'estimateur de la variance est biaisé ($E(\hat{\sigma}^2) = \frac{n}{n-1}\sigma^2$). Néanmoins l'estimateur est asymptotiquement sans biais.

Chapitre 4

Intervalles de confiance

Au lieu de se donner une fonction (estimateur) qui donne une estimation ponctuelle d'un paramètre, on cherche un intervalle dans lequel se trouve le paramètre étudié avec une probabilité contrôlée (et généralement grande).

4.1 Estimation d'une proportion par intervalle de confiance

On considère une population telle que pour le caractère observé la proportion p d'une certaine catégorie est inconnue. On souhaite estimer cette proportion p de cette population à partir d'un échantillon de taille n dont la fréquence de la catégorie étudiée est f. Soit F la variable aléatoire qui à chaque échantillon de taille n associe la fréquence du nombre d'éléments qui appartiennent à la catégorie choisie. On sait que F suit approximativement la loi $\mathcal{N}(p;\sigma)$ avec $\sigma = \sqrt{pq/n}$, pour n suffisamment grand ($n > 30$). On dispose de

$$\sigma' = \sqrt{\frac{f(1-f)}{n}}$$

l'écart type associé à la fréquence f de l'échantillon de taille n. On se sert de l'estimation ponctuelle de σ puisque p est inconnue :

$$\sigma = \sigma' \sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n}} \cdot \sqrt{\frac{n}{n-1}} = \sqrt{\frac{f(1-f)}{n-1}}.$$

Donc la variable aléatoire Z définie par :

$$Z = \frac{F-p}{\sigma}$$

suit approximativement une loi normale centrée réduite $\mathcal{N}(0;1)$. On cherche un intervalle de confiance de la proportion p, c'est-à-dire un intervalle tel que la probabilité que la proportion p n'appartienne pas à cet intervalle soit égale à α où $\alpha \in [0;1]$. On appelle cet intervalle de confiance avec le risque α ou avec le coefficient de confiance $c = 1-\alpha$. Le risque que l'on prend à dire que p appartient à cet intervalle est donc de α ou encore la probabilité que p n'appartienne pas à cet intervalle est le risque α.

Déterminons cet intervalle de confiance : On rappelle que l'on a défini $z_{\alpha/2}$ comme étant la valeur telle que

$$P(Z > z_{\alpha/2}) = \alpha/2$$

où Z suit $\mathcal{N}(0;1)$. A l'aide des propriétés de la loi normale centrée réduite, on a $P(Z < -z_{\alpha/2}) = \alpha/2$ et $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha$.

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha \quad &\Leftrightarrow \quad P\left(-z_{\alpha/2} < \frac{F-p}{\sigma} < z_{\alpha/2}\right) = 1-\alpha \\ &\Leftrightarrow \quad P(-z_{\alpha/2}\cdot\sigma < F-p < z_{\alpha/2}\cdot\sigma) = 1-\alpha \\ &\Leftrightarrow \quad P(F-z_{\alpha/2}\cdot\sigma < p < F+z_{\alpha/2}\cdot\sigma) = 1-\alpha \\ &\Leftrightarrow \quad P\left(F-z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}} < p < F+z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}}\right) = 1-\alpha \end{aligned}$$

L'intervalle de confiance de la proportion p avec un coefficient de confiance de $1-\alpha$ est :

$$\left]f-z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}}; f+z_{\alpha/2}\sqrt{\frac{f(1-f)}{n-1}}\right[.$$

Remarque : lorsque n est grand, la différence entre n et $n-1$ devient négligeable, aussi la formule devient

$$\left]f-z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}; f+z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}\right[.$$

C'est la formule la plus couramment utilisée.

On peut encore simplifier : Avec un risque $\alpha = 5\%$, et $f \approx 0.5$, la formule peut être approchée par

$$\left]f-\frac{1}{\sqrt{n}}; f+\frac{1}{\sqrt{n}}\right[.$$

4.2 Moyenne

On considère une variable aléatoire X suivant $\mathcal{N}(\mu,\sigma)$ et $X_1, ..., X_n$, n variables indépendantes et de même loi que X. On rappelle que les définitions de la moyenne empirique et la variance empirique corrigée (ou modifiée) sont respectivement données par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Soit $z_{\alpha/2}$ le nombre réel positif tel que $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$. D'après la proposition 3.1.1, on sait que la variable aléatoire $\overline{X}$ suit la loi normale $\mathcal{N}(\mu;\sigma/\sqrt{n})$ d'où

$$\begin{aligned} 1-\alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \\ &= P\left(\overline{X}-z_{\alpha/2}\cdot\sigma/\sqrt{n} < \mu < \overline{X}+z_{\alpha/2}\cdot\sigma/\sqrt{n}\right) \end{aligned}$$

L'intervalle de confiance pour la moyenne d'une population de variance σ^2 connue est donné par

$$\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\text{soit } I = \left]\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right[.$$

Cet intervalle reste valable lorsque la variance est inconnue et l'échantillon très grand.

Proposition 4.2.1 *La variable* $\frac{(n-1)S^2}{\sigma^2}$ *suit une loi du* χ^2 *à* $\nu = n-1$ *degrés de liberté.*

Preuve (partielle) Dans la preuve de la proposition 3.1.3, on obtient

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}(X_i-\overline{X})^2 + n(\overline{X}-\mu)^2.$$

En divisant par σ^2, on obtient

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

Le premier terme est une somme de carrés de variables indépendantes suivant $\mathcal{N}(0,1)$, le deuxième vaut $\frac{(n-1)S^2}{\sigma^2}$ et le dernier est un carré d'une variable suivant $\mathcal{N}(0,1)$ d'après la proposition 3.1.1 sur la moyenne empirique. En supposant l'indépendance (admise) des variables $\overline{X}$ et S^2, on peut écrire cette égalité en terme de fonctions caractéristiques, où φ est la fonction caractéristique de $\frac{(n-1)S^2}{\sigma^2}$ et où la fonction caractéristique de la loi du χ^2 est utilisée (Proposition 1.2.1) :

$$\left(\frac{1}{1-2it}\right)^{n/2} = \varphi(t)\cdot\left(\frac{1}{1-2it}\right)^{1/2},$$

ou encore

$$\varphi(t)\cdot\left(\frac{1}{1-2it}\right)^{(n-1)/2}.$$

Ainsi, d'après la proposition 1.2.1, $\frac{(n-1)S^2}{\sigma^2}$ suit une loi du χ^2 à $\nu = n-1$ degrés de liberté.

Lorsqu'on ne dispose que de n observations d'une population de distribution normale d'écart-type inconnu, cet intervalle est modifié. En effet, on se base sur la moyenne de l'échantillon et l'écart-type estimé de la population pour donner un intervalle de confiance sur la moyenne μ de la population. On a $\frac{\overline{X}-\mu}{S/\sqrt{n}} \hookrightarrow St(n-1)$ (loi de Student à $n-1$ degrés de liberté) car cette variable peut s'écrire sous la forme d'un produit, en posant $Q = \frac{(n-1)S^2}{\sigma^2}$,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\Big/\sqrt{\frac{Q}{n-1}},$$

variable suivant une loi de Student d'après la définition 2 du Chapitre 1.

Ainsi cet intervalle est donné par :

$$\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

où $t_{\alpha/2} = t_{\alpha/2;(n-1)}$ c'est-à-dire que ce nombre sera lu dans la distribution de Student au risque $\alpha/2$ avec $\nu = n-1$ degrés de liberté.

4.3 Variance

On considère la variance empirique modifiée S^2. D'après la proposition 4.2.1, on sait que

$$\frac{(n-1)S^2}{\sigma^2} \hookrightarrow \chi^2(n-1). \quad \text{(Loi du } \chi^2 \text{ à } \nu = n-1 \text{ ddl)}$$

De plus, $P(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}) = 1-\alpha$, D'où

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{1-\alpha/2} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) \\ &= P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) \end{aligned}$$

où $\chi^2_{\alpha/2} = \chi^2_{\alpha/2;(n-1)}$ sera lu dans la table de χ^2 avec $\nu = n-1$ degrés de liberté. On cherchera donc les valeurs telles que $P\left(K^2 > \chi^2_{\alpha/2;(n-1)}\right) = \alpha/2$, et $P\left(K^2 < \chi^2_{1-\alpha/2;(n-1)}\right) = \alpha/2$.

Chapitre 5

Notion de test d'hypothèse

La description de "la" réalité en statistiques se fait à l'aide de "variables" qui sont des colonnes de valeurs numériques. On se pose souvent la question de comparer ces variables, de tester si elles sont égales ou différentes, de savoir si on peut considérer qu'elles correspondent ou non à une même population [sous-jacente], si elles correspondent à une distribution donnée, si elles sont conformes à un modèle précis etc. sachant que ces variables et leurs données ne correspondent qu'à un échantillon de valeurs.

Etant donné qu'on ne peut jamais être sûr que le résultat des calculs correspond à "la" réalité, les statisticiens et statisticiennes ont développé un cadre d'analyse qui permet de prendre de telles décisions tout en disposant d'une estimation du "risque" de ces décisions.

Les tests d'hypothèses ont pour buts de
– clarifier et définir le cadre rigoureux de ces études,
– fournir un formalisme précis pour toutes les situations,
– savoir si les différences mises en jeu sont importantes ("significatives" pour un seuil donné) ou non.

5.1 Hypothèse nulle, risques de première et deuxième espèce

Le cadre mathématique est celui des événements probabilisés où l'hypothèse, la comparaison de départ est convertie en un événement intégré à un modèle probabiliste réfutable. On distingue en général deux hypothèses seulement : la première, également nommée **hypothèse nulle**, notée H_0 est celle où justement la différence est considérée comme nulle (on dira en fait non significative, par rapport à un seuil défini plus loin comme "risque de première espèce") ; la seconde, complémentaire de la première, regroupant tous les autres cas, est nommée **hypothèse alternative** et parfois notée H_1.

Une hypothèse doit spécifier une valeur, disons θ_0 pour un paramètre θ de la population. On testera donc

$$H_0 : \theta = \theta_0.$$

Une possibilité classique pour l'hypothèse alternative est

$$H_1 : \theta \neq \theta_0,$$

qui teste chaque côté de l'égalité (on parlera de test bilatéral).

Mais on peut écrire également un autre choix d'hypothèse :

$$H_0 : \theta \geq \theta_0, \quad \text{parfois noté encore } H_0 : \theta = \theta_0$$

et l'hypothèse alternative correspondante sera

$$H_1 : \theta < \theta_0,$$

qui teste un seul côté de l'égalité (on parlera de test unilatéral).
Le dernier cas est facile à trouver : $H_0 : \theta \leq \theta_0$ et $H_1 : \theta > \theta_0$ (unilatéral également).

On peut soit rejeter l'hypothèse nulle, soit ne pas la rejeter alors qu'en fait, soit cette hypothèse est vraie soit elle ne l'est pas ce qui oblige à utiliser un tableau à 4 cases qui résume l'ensemble des couples (décisions/réalité) :

Décision / Réalité	H_0 est vraie	H_0 est fausse
ne pas rejeter H_0	Vrai Positif	Faux Positif
rejeter H_0	Faux Négatif	Vrai Négatif

Ex (Test de grossesse) : Dans le cadre d'un test de grossesse par autodiagnostic, un résultat est qualifié de « faux négatif » lorsqu'il indique que la personne n'êtes pas enceinte, bien que la fécondation ait eu lieu. A l'inverse un test positif erroné -beaucoup plus rare- indique un début de grossesse, alors qu'il n'en est rien.

Les cas VN (Rejeter H_0 quand elle est Fausse) et VP (Ne pas rejeter H_0 quand elle est Vraie) sont des "bonnes décisions". Par contre, FN (Rejeter H_0 quand elle est Vraie) est nommée **erreur de première espèce** et FP (Ne pas rejeter H_0 quand elle est Fausse) est nommée **erreur de deuxième espèce**. A chacune de ces erreurs, on associe un risque lié à la probabilité de la décision : on le nomme α pour FP, β pour FN. Il n'y a aucune raison de supposer ces risques équivalents et souvent on prend $\alpha = 5\%$ (ou 1% quand on veut être plus strict) alors qu'il est "habituel" de prendre 0.20 pour β. La probabilité de rejeter H_0 alors qu'elle est vraie vaut α et est appelé niveau du test (ou seuil). La probabilité de rejeter une fausse hypothèse nulle est $(1 - \beta)$ qui est appelée la **puissance** du test.

Il faut bien comprendre que les tests d'hypothèse ne permettent pas d'accepter H_0 mais seulement de rejeter H_0. Ne pas rejeter H_0 ne signifie pas que H_0 est vraie mais seulement que la probabilité qu'elle soit fausse est très petite. On n'est donc en fait jamais vraiment totalement sûr de rien.

Ce qui nous donne en tableau :

	H_0 est vraie	H_0 est fausse
non rejet de H_0	cohérent	Erreur type II (non rejet à tort) : risque β
rejet de H_0	Erreur type I (rejet à tort) : risque α	cohérent

Dans le cadre de tests statistiques, on doit décider si on peut considérer par exemple que 0.21 et 0.22 sont proches, si 15% et 20% peuvent être considérés comme peu éloignés etc., la loi statistique de la différence entre ces lois étant supposée connue, tabulée et consultable.

5.2 Mécanique des tests d'hypothèse

Pour réaliser un test d'hypothèse, il y a un enchaînement strict d'actions à effectuer. Cela commence par la formulation de l'hypothèse dans le domaine considéré (médical, économique, social...) et sa traduction en événements probabilistes liés à H_0. On doit ensuite considérer la statistique d'écart (la loi théorique de la différence) et choisir un seuil (alpha) de décision. On doit ensuite calculer la valeur de la statistique d'écart pour les valeurs observées puis comparer à la valeur théorique de la statistique d'écart pour le seuil choisi et en déduire si on refuse H_0 ou non. Enfin, le calcul (ou la lecture) de la "p-value" associé au dépassement de la valeur de la statistique d'écart permet de conclure de façon fine sur le fait que la différence est significative ou non.
Le fait de "Ne pas rejeter H_0" au risque α sera parfois confondu par la suite avec "On accepte H_0" par abus de langage, le risque β n'étant pas considéré pour ce cours.

Chapitre 6

Test d'indépendance

6.1 Test d'indépendance de deux variables qualitatives

Dans la plupart des tests que nous venons de présenter, on suppose toujours les valeurs de l'échantillon indépendantes. C'est une condition nécessaire. Il est donc souvent utile de vérifier cette hypothèse par un test. Ce test met en place une variable aléatoire qui suit une loi du χ^2, aussi ce test est appelé Test d'indépendance du χ^2.

Ce test permet de contrôler l'indépendance de deux caractères dans une population donnée.

On dispose de deux variables aléatoires X et Y, les valeurs possibles de X sont réparties en l modalités (ou classes) $X_1, \ldots, X_l$, celles de Y en k modalités $Y_1, \ldots, Y_k$. Pour chaque intersection de modalités X_i et Y_j, un effectif $n_{i,j}$ est observé. Ainsi

$$n = \sum_{i=1}^{l} \sum_{j=1}^{k} n_{i,j}.$$

Hypothèse testée H_0 : « Les variables X et Y sont indépendantes ».

Déroulement du test : On crée le tableau des effectifs qui est un tableau à double-entrée. A l'intersection de la i-ème ligne et de la j-ième colonne, on écrit l'effectif $n_{i,j}$. On calcule les effectifs marginaux : $S_i = \sum_j n_{i,j}$ est la somme des termes sur la i-ème ligne, $T_j = \sum_i n_{i,j}$ est la somme des termes sur la j-ième colonne.

		Y_j		
X_i	$\cdots$	$\vdots$ $n_{i,j}$ $\vdots$	$\cdots$	S_i
		T_j		n

On calcule les effectifs théoriques :

$$C_{i,j} = \frac{S_i T_j}{n}.$$

Remarque : Sous l'hypothèse H_0, on a $C_{i,j} = n_{i,j}$.

On calcule la valeur de la variable de test :

$$\chi_c^2 = \sum_{i,j} \frac{(n_{i,j} - C_{i,j})^2}{C_{i,j}}.$$

On cherche la valeur critique χ^2_α dans la table de la loi du χ^2 à $\nu = (l-1) \times (k-1)$ degrés de liberté.

Décision : si $\chi^2_c < \chi^2_\alpha$, on accepte l'hypothèse H_0, sinon on la rejette.
Vérification a posteriori des conditions d'application : il faut $C_{i,j} \geq 5$ pour tous i, j.

Exemple : Pour comparer l'efficacité de deux médicaments agissant sur la même maladie, mais aux prix très différents, la Sécurité Sociale a effectué une enquête sur les guérisons obtenues en suivant chacun des traitements. Les résultats sont consignés dans le tableau suivant :

	Médicament	Générique
Guérisons	48	158
Non Guérisons	6	44

Les effectifs marginaux sont les suivants :

	Médicament	Générique	
Guérisons	48	158	206
Aucun Effet	6	44	50
	54	202	256

Les effectifs théoriques sont :

	Médicament	Générique	
Guérisons	$\frac{206\times54}{256}$	$\frac{206\times202}{256}$	206
Non Guérisons	$\frac{50\times54}{256}$	$\frac{50\times202}{256}$	50
	54	202	256

On calcule $\chi^2_c = \frac{(48-43{,}45)^2}{43{,}45} + \frac{(158-162{,}55)^2}{162{,}55} + \frac{(6-10{,}55)^2}{10{,}55} + \frac{(44-39{,}45)^2}{39{,}45} \approx 3{,}1.$

La variable de test χ^2_c vaut approximativement 3,1, alors que la valeur critique, pour un niveau de risque de 5%, est 3,84 (on explore la table du χ^2 à un degré de liberté). On peut donc raisonnablement estimer ici que le taux de guérison ne dépend pas du prix du médicament et se poser des questions sur l'opportunité de continuer à vendre le médicament cher.

6.2 Test d'indépendance de deux variables quantitatives : test de corrélation nulle

Soit r le coefficient de corrélation de l'échantillon composé de n paires d'observations extrait de populations gaussiennes. Il s'agit de tester l'hypothèse nulle :

$$H_0 : \rho = 0 \quad \text{(corrélation nulle entre les populations)}$$

au risque α. On peut montrer sous H_0 que la variable aléatoire $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$ suit une loi de Student à $\nu = n-2$ degrés de liberté.

On calculera donc

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

puis on cherchera la valeur t_α ou $t_{\alpha/2}$ dans la table de loi t de Student à $\nu = n-2$ degrés de liberté tel que

$$P(T_{n-2} > t_{\alpha/2}) = \alpha/2$$

et on adoptera la règle de décision suivante :

- Si l'hypothèse alternative est $H_1 : \rho \neq 0$ (cas bilatéral) : rejet de H_0 au risque α si $t \notin]-t_{\alpha/2}; t_{\alpha/2}[$ avec $\nu = n-2$ degrés de liberté.
- Si l'hypothèse alternative est $H_1 : \rho > 0$ (cas unilatéral) : rejet de H_0 au risque α si $t > t_\alpha$ avec $\nu = n-2$ degrés de liberté.
- Si l'hypothèse alternative est $H_1 : \rho < 0$ (cas unilatéral) : rejet de H_0 au risque α si $t < -t_\alpha$ avec $\nu = n-2$ degrés de liberté.

Chapitre 7

Tests de conformité en loi

7.1 Cas général

7.1.1 Test d'adéquation du χ^2

Soit X une variable aléatoire de loi $\mathcal{L}$ (le plus souvent inconnue). On souhaite tester l'ajustement de cette loi à une loi connue $\mathcal{L}_0$ (Poisson, Exponentielle, normale, etc) retenue comme étant un modèle convenable.

On teste donc l'hypothèse $H_0 : \mathcal{L} = \mathcal{L}_0$ contre l'hypothèse $H_1 : \mathcal{L} \neq \mathcal{L}_0$.

Les n observations de X sont partagées en k classes. On désigne par O_i l'effectif observé de la classe i. Ainsi $\sum_i O_i = n$.

Pour chaque classe, l'effectif théorique est défini :

$$C_i = n \cdot p(X \in \text{Classe}_i / X \hookrightarrow \mathcal{L}_0).$$

Classe	1	2	$\cdots$	i	$\cdots$	k
Effectif observé	O_1	O_2	$\cdots$	O_i	$\cdots$	O_k
Effectif théorique	C_1	C_2	$\cdots$	C_i	$\cdots$	C_k

On calcule la valeur $\chi^2_c = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$. On compare cette valeur à la valeur théorique χ^2_α lue dans la table du χ^2 à $\nu = k - 1 - r$ degrés de liberté où r est le nombre de paramètres de la loi $\mathcal{L}_0$ qu'il a fallu estimer. (Exemples : $r = 0$ si la loi est connue ou imposée, $r = 1$ pour une loi de Poisson, $r = 2$ pour une loi normale sans autre précision)

On rejette H_0 lorsque $\chi^2_c > \chi^2_\alpha$.

Exemple : Un pisciculteur possède un bassin qui contient trois variétés de truites : communes, saumonées et arc-en-ciel. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela, il effectue, au hasard 399 prélèvements avec remise et obtient les résultats suivants :

Variétés	commune	saumonée	arc-en-ciel
Effectifs	145	118	136

On cherche à savoir s'il y a équirépartition des truites entre chaque espèce c'est-à-dire on suppose de $\mathcal{L}_0$ est la loi uniforme, une probabilité de $1/3$ pour chaque classe (soit $C_i = 399 \cdot \frac{1}{3} = 133$).

Variétés	commune	saumonée	arc-en-ciel
Effectifs O_i	145	118	136
Effectifs C_i	133	133	133

On obtient

$$\chi_c^2 = \frac{(145-133)^2}{133} + \frac{(118-133)^2}{133} + \frac{(136-133)^2}{133} \approx 2.84$$

La valeur théorique lue dans la table du χ^2 au risque de 5% avec $\nu = 3-1-0 = 2$ degrés de liberté vaut 5.99.

On ne peut rejeter l'hypothèse que son bassin contient autant de truites de chaque variété car $\chi_c^2 < \chi_\alpha^2$.

7.1.2 Test de Kolmogorov-Smirnov

Comme précédemment, l'objectif est d'établir la plausibilité de l'hypothèse selon laquelle l'échantillon a été prélevé dans une population ayant une distribution donnée. Le test de Kolmogorov est "non-paramétrique" : il ne place aucune contrainte sur la distribution de référence, et ne demande pas qu'elle soit connue sous forme analytique (bien que ce soit pourtant le cas le plus courant).

Etant donnés :

1. Un échantillon de taille n d'observations d'une variable,
2. Et une fonction de répartition de référence $F(x)$,

le test de Kolmogorov teste l'hypothèse H_0 selon laquelle l'échantillon a été prélevé dans une population de fonction de répartition $F(x)$.

Pour cela, il calcule sur l'échantillon une quantité D, appelée "statistique de Kolmogorov", dont la distribution est connue lorsque H_0 est vraie. La statistique de Kolmogorov-Smirnov D_n est définie par

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

où $F_n(x)$ est la proportion des observations dont la valeur est inférieure ou égale à x (fonction de répartition empirique).

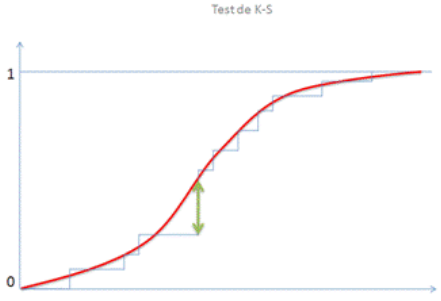


FIGURE 7.1 – Test de Kolmogorov-Smirnov

Une valeur élevée de D est une indication que la distribution de l'échantillon s'éloigne sensiblement de la distribution de référence $F(x)$, et qu'il est donc peu probable que H_0 soit correcte. Plus précisément,

$$P\left(\sup_x |F_n(x) - F(x)| > \frac{c}{\sqrt{n}}\right) \longrightarrow_{n\to\infty} \alpha(c) = 2\sum_{r=1}^{+\infty}(-1)^{r-1}\exp(-2r^2c^2)$$

pour toute constante $c > 0$. Le terme $\alpha(c)$ vaut 0.05 pour $c = 1.36$. Pour $n > 100$, la valeur critique du test est approximativement de la forme $\frac{c}{\sqrt{n}}$. Les valeurs usuelles de c en fonction de α sont :

α	0.20	0.10	0.05	0.02	0.01
c	1.073	1.224	1.358	1.517	1.628

Si $D_n > \frac{c}{\sqrt{n}}$, on rejette H_0.

Exemple : `http://www.jybaudot.fr/Inferentielle/kolmogorov.html`
Une nouvelle clientèle étrangère est attendue dans une station balnéaire. Afin de mieux connaître leurs goûts, des brasseurs ont commandé une étude de marché. En début de saison, on demande à vingt de ces nouveaux touristes de donner leur préférence parmi cinq types de bières, de la moins amère (bière 1) à la plus amère (bière 5). A l'aide d'un test de K-S, le chargé d'études décide de comparer les résultats avec une loi uniforme, c'est-à-dire une situation où chaque bière aurait eu la préférence de quatre répondants.

Les résultats de l'enquête sont les suivants :

1 3 2 5 1 2 2 4 1 2 2 1 3 3 2 4 5 1 1 2

On se fixe un risque d'erreur de 5%. L'hypothèse H_0 à tester est celle de l'égalité avec une loi uniforme.

Résumons les écarts entre observations et répartition uniforme :

Classe	Effectif	Uniforme	Cumul réel	Cumul théorique	D
1	6	4	0,30	0,20	0,10
2	7	4	0,65	0,40	0,25
3	3	4	0,80	0,60	0,20
4	2	4	0,90	0,80	0,10
5	2	4	1,00	1,00	0,00

La distance la plus élevée s'établit à $d = 0,25$.

On calcule pour $n = 20$ et $\alpha = 5\%$ la valeur de $c/\sqrt{20} = 0,303$. Bien que ces touristes semblent préférer les bières les moins amères, on ne peut pas rejeter l'hypothèse selon laquelle ils n'ont pas de préférence particulière.

7.2 Test de normalité

Les tests précédents sont des tests généraux s'appliquant sur n'importe quelle loi. Lorsque la loi à tester est la loi normale, on parle de test de normalité.

On cherche à se déterminer entre :
H_0 : les données suivent une loi normale.
H_1 : les données ne suivent pas une loi normale.

7.2.1 Méthodes graphiques : Droite de Henry

La droite de Henry est une méthode pour visualiser les chances qu'a une distribution d'être gaussienne. Elle permet de lire rapidement la moyenne et l'écart type d'une telle distribution.

Principe : On représente les quantiles théoriques en fonction des quantiles observés (Diagramme Q-Q).

Si X est une variable gaussienne de moyenne $\overline{x}$ et de variance σ^2 et si Z est une variable de loi normale centrée réduite, on a les égalités suivantes :

$$P(X < x_i) = P\left(\frac{X - \overline{x}}{\sigma} < \frac{x_i - \overline{x}}{\sigma}\right) = P(Z < y_i) = \Phi(y_i)$$

où $y_i = \frac{x_i - \overline{x}}{\sigma}$. (on note Φ la fonction de répartition de la loi normale centrée réduite).

Pour chaque valeur x_i de la variable X, on peut calculer $P(X < x_i)$ puis en déduire, à l'aide d'une table de la fonction Φ, y_i tel que $\Phi(y_i) = P(X < x_i)$.

Si la variable est gaussienne, les points de coordonnées $(x_i; y_i)$ sont alignés sur la droite d'équation $y = \frac{x - \overline{x}}{\sigma}$.

Exemple numérique : Lors d'un examen noté sur 20, on obtient les résultats suivants :
- 10% des candidats ont obtenu moins de 4
- 30% des candidats ont obtenu moins de 8
- 60% des candidats ont obtenu moins de 12
- 80% des candidats ont obtenu moins de 16

On cherche à déterminer si la distribution des notes est gaussienne, et, si oui, ce que valent son espérance et son écart type.

On connaît donc 4 valeurs x_i, et, pour ces 4 valeurs, on connaît $P(X < x_i)$.

En utilisant la table "Table de la fonction de répartition de la loi normale centrée réduite", on détermine les y_i correspondants :

x_i	$P(X < x_i) = \Phi(y_i)$	y_i
4	$0,10$	$-1,282$
8	$0,30$	$-0,524$
12	$0,60$	$0,253$
16	$0,80$	$0,842$

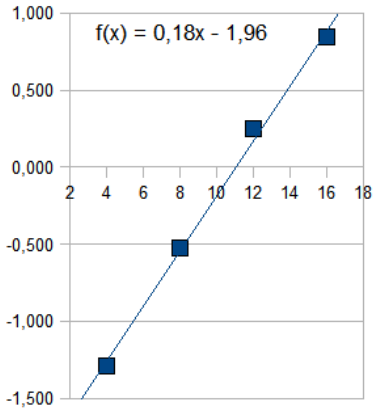


FIGURE 7.2 – Droite de Henry

Les points paraissent alignés. La droite coupe l'axe des abscisses au point d'abscisse 11 et le coefficient directeur est 0.18 environ, ce qui donnerait un écart type de $1/0.18 = 5,6$.

Cela laisse penser que la distribution est gaussienne de paramètres $\mu = 11$ et $\sigma = 5.6$.

Remarque : On peut faire de même en comparant sur un graphique les probabilités cumulées théoriques et les probabilités cumulées empiriques (comparaison des fonctions de répartition : Diagramme P-P). On est alors dans une sorte de validation type Kolmogorov-Smirnov mais graphique.

7.2.2 Test Jarque-Bera (ou test de Bowman-Shelton)

Le test de Jarque-Bera est un test de normalité. On pose

$$S = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) \quad \text{Coefficient d'asymétrie : Moment d'ordre 3 d'une variable centrée-réduite}$$

$$K = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) \quad \text{Kurtosis : Moment d'ordre 4 d'une variable centrée-réduite}$$

On rappelle qu'une loi normale a un coefficient d'asymétrie = 0 et une kurtosis = 3. On peut traduire les hypothèses sous la forme :
$H_0 : S = 0$ et $K = 3$
$H_1 : S \neq 0$ ou $K \neq 3$.

On remarque ainsi que s'il y a rejet, le test ne permet pas d'en connaître la raison principale (asymétrie ou applatissement).

On calcule

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right),$$

où n est le nombre d'observations. Il faut que n soit suffisamment grand ($n > 50$).

La statistique JB suit asymptotiquement une loi du χ^2 à 2 degrés de liberté. Si les données suivent une loi normale, le test s'approche alors de 0 et on accepte (ne rejette pas) H_0 au seuil α.

Chapitre 8

Test sur les pourcentages

8.1 Relation test et intervalles de confiance

Un test correspond à construire un intervalle de confiance autour d'une valeur à partir d'un échantillon et de regarder si sa valeur supposée sous H_0 est finalement dans cet intervalle, construit à partir d'un certain risque. La valeur intéressante pour un test est le risque pris pour rejeter H_0. Cela permet de s'assurer de la pertinence (vraisemblabilité) de H_0 ou de H_1. Les lois qui interviennent dans les calculs sont les mêmes mais au lieu de construire un intervalle de confiance pour chaque risque pris, on compare une partie fixe (calculée à partir des observations) avec une partie ne dépendant que du risque pris.

8.2 Test de conformité

Soit p_r la proportion (valeur connue) possédant le caractère considéré dans une population de référence. Il s'agit de tester si la proportion p d'une autre population, dont on a extrait un échantillon de taille n et observé une fréquence f pour ce caractère, correspond à celle d'une population de référence, soit

$$\begin{aligned} H_0 &: p = p_r \\ H_1 &: p \neq p_r \end{aligned}$$

On considère F la variable aléatoire qui suit les fréquences observées dans les échantillons. Sous H_0, la loi de F peut être approchée par $\mathcal{N}\left(p_r, \sqrt{\frac{p_r(1-p_r)}{n}}\right)$.

On se fixe α le risque que $p \neq p_r$, ce qui revient à rechercher un intervalle I centré sur p_r tel que $P(p \notin I) = 1 - \alpha$ c'est-à-dire

$$P\left(-z_{\alpha/2} < \frac{F - p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

On teste donc si la valeur calculée

$$z = \frac{f - p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}}$$

appartient à l'intervalle $]-z_{\alpha/2}; z_{\alpha/2}[$.

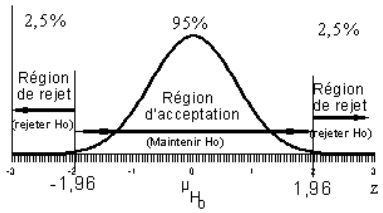


FIGURE 8.1 – test bilatéral pour $\alpha = 5\%$

Décision : on accepte H_0 si $z \in]-z_{\alpha/2}; z_{\alpha/2}[$ au risque α et on rejette H_0 sinon.

Lorsque une partie de l'hypothèse H_1 est a priori à écarter (non sens, impossibilité), alors le risque ne répartit plus de chaque côté de l'inégalité mais est réparti sur une seule partie (on parle alors de test unilatéral). On teste donc uniquement $H_0 : p = p_r$ contre $H_1 : p > p_r$, ou $H_0 : p = p_r$ contre $H_1 : p < p_r$. on rejettera H_0 lorsque p sera bien plus grand que p_r ou respectivement p sera bien plus petit que p_r.

Les hypothèses considérées sont donc dans un cas :

$$\begin{aligned} H_0 &: p = p_r \\ H_1 &: p > p_r \end{aligned}$$

ce qui revient à rechercher un intervalle I tel que

$$P\left(\frac{F - p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}} < z_\alpha\right) = 1 - \alpha.$$

On compare donc la valeur calculée

$$z = \frac{f - p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}}$$

avec une valeur z_α lue dans la table de l'écart-réduit (lire au risque 2α).

Décision : on accepte H_0 si $z \in]0; z_\alpha[$ au risque α et on rejette H_0 sinon ($z > z_\alpha$).

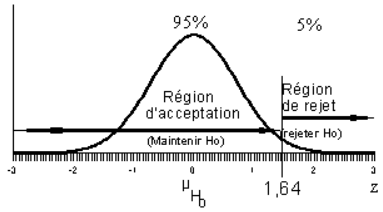


FIGURE 8.2 – test unilatéral pour $\alpha = 5\%$

Les hypothèses considérées sont donc dans un second cas :

$$\begin{aligned} H_0 &: p = p_r \\ H_1 &: p < p_r \end{aligned}$$

ce qui revient à rechercher un intervalle I tel que

$$P\left(-z_\alpha < \frac{F - p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}}\right) = 1 - \alpha.$$

On compare donc la valeur calculée $z = \frac{f-p_r}{\sqrt{\frac{p_r(1-p_r)}{n}}}$ avec une valeur z_α lue dans la table de l'écart-réduit (lire au risque 2α, et mettre un signe moins)

Décision : on accepte H_0 si $z \in]-z_\alpha; 0[$ au risque α et on rejette H_0 sinon ($z < -z_\alpha$).

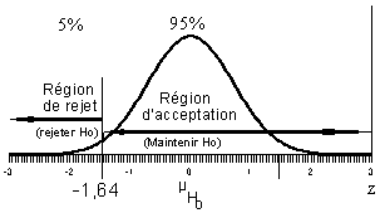


FIGURE 8.3 – test unilatéral pour $\alpha = 5\%$

Exemples :

1. test bilatéral (Un test bilatéral rejette les valeurs trop écartées) On désire tester le "chiffre" annoncé de 20% des personnes qui écoutent une certaine émission radiophonique correspond à la réalité. Une sondage de 1000 auditeurs donne une proportion de 0.1875.

   $$\begin{aligned} H_0 &: p = 0.2 \\ H_1 &: p \neq 0.2 \end{aligned}$$

   On choisit un test bilatéral car on n'a aucune idée du pourcentage réel. ($z \approx -0.99$)

2. test unilatéral à droite (Un test unilatéral à droite rejette les valeurs trop grandes de la statistique de test) Un magicien prétend qu'il peut souvent deviner à distance la couleur d'une carte tirée au hasard d'un jeu de cartes bien battu et comportant des cartes de deux couleurs différentes en nombre égal. Sur un échantillon de taille 100, la magicien a obtenu 64 succès. Quel niveau de risque prend-t-on pour déclarer que le magicien n'est pas un imposteur ?

   $$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p > 0.5 \end{aligned}$$

   ($z \approx 2.8$)

3. test unilatéral à gauche (Un test unilatéral à gauche rejette les valeurs trop petites) On sait que la grippe touche 30% d'une population lors d'une épidémie. Pour tester l'efficacité d'un vaccin antigrippal, on vaccine préalablement 300 personnes. A la fin de la saison grippale, on dénombre cinquante personnes qui ont été atteintes par la grippe parmi les vaccinés. Ce résultat permet-il d'apprécier l'efficacité du vaccin ?

   $$\begin{aligned} H_0 &: p = 0.3 \\ H_1 &: p < 0.3 \end{aligned}$$

   ($z \approx -5.04$)

8.3 Test d'homogénéité

Soit X une variable qualitative prenant deux modalités (succès $X = 1$, échec $X = 0$) observée sur deux populations et deux échantillons indépendants extraits de ces deux populations. On observe une fréquence f_1 dans la population 1 de taille n_1 et f_2 dans la population 2 de taille n_2.

On fait l'hypothèse que les deux échantillons proviennent de deux populations dans lesquelles les probabilités de succès sont identiques.

$$\begin{aligned} H_0 &: p_1 = p_2 \\ H_1 &: p_1 \neq p_2 \end{aligned}$$

La distribution d'échantillonnage de la fréquence de succès dans la population 1, F_1 converge en loi vers $\mathcal{N}(p_1; \sqrt{p_1 q_1 / n_1})$ et de même F_2 converge en loi vers $\mathcal{N}(p_2; \sqrt{p_2 q_2 / n_2})$ (On rappelle que nF suit la loi binomiale de paramètres (n, p)). Comme F_1 et F_2 sont deux variables aléatoires indépendantes, on a

$$\begin{aligned} E(F_1 - F_2) &= E(F_1) - E(F_2) = p_1 - p_2 \\ V(F_1 - F_2) &= V(F_1) + V(F_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \end{aligned}$$

Dans les conditions d'approximation de $\mathcal{B}$ par $\mathcal{N}$ ($n_1 p_1, n_1 q_1, n_2 p_2, n_2 q_2 > 5$ et $n_1, n_2 > 30$), la variable aléatoire $F_1 - F_2$ suit la loi normale $\mathcal{N}\left(p_1 - p_2; \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$ et ainsi la variable normale centrée réduite

$$Z = \frac{(F_1 - F_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

devient sous H_0,

$$Z = \frac{F_1 - F_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

La valeur p, probabilité du succès commune aux deux populations n'est en réalité pas connue. On l'estime à partir des résultats observés sur les deux échantillons :

$$\hat{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

où f_1 et f_2 représentent les fréquences observées respectivement pour l'échantillon 1 et pour l'échantillon 2.

Une valeur observée z de la variable aléatoire Z est calculée de la façon suivante :

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

avec $\hat{q} = 1 - \hat{p}$.

Cette valeur sera comparée avec la valeur seuil z_α lue sur la table de la loi normale centrée réduite $\mathcal{N}(0; 1)$ pour un risque d'erreur α fixé.

Décision :
- si $z \in]-z_{\alpha/2}; z_{\alpha/2}[$, l'hypothèse H_0 est acceptée : les deux échantillons sont extraits de deux populations ayant même probabilité de succès p.
- si $z > z_{\alpha/2}$ ou $z < -z_{\alpha/2}$ (ou encore $z \notin]-z_{\alpha/2}; z_{\alpha/2}[$) l'hypothèse H_0 est rejetée au risque d'erreur α : les deux échantillons sont extraits de deux populations ayant des probabilités de succès différentes respectivement p_1 et p_2.

Remarque : on peut aussi tester un seul côté de l'inégalité (H_0 restant $p_1 = p_2$) : on calcule de la même façon

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

puis on décide et conclut selon le cas

– Si l'hypothèse alternative est $H_1 : p_1 > p_2$ (cas unilatéral) : rejet de H_0 au risque α si $z > z_\alpha$.
– Si l'hypothèse alternative est $H_1 : p_1 < p_2$ (cas unilatéral) : rejet de H_0 au risque α si $z < -z_\alpha$.
Exemple : on veut tester l'impact de l'assiduité aux travaux dirigés dans la réussite à l'examen de statistique.

	groupe 1	groupe 2
Nbre d'heures en TD	18 h	30 h
Nbre d'étudiants	180	150
Nbre d'étudiants ayant réussi à l'examen	126	129

Qu'en concluez-vous ?

On choisit un test unilatéral car on suppose que la réussite est meilleure avec plus d'heures de TD. Ainsi on teste l'hypothèse : $H_0 : p_1 = p_2$ contre $H_1 : p_1 < p_2$.

Calculs :

$$z = \frac{f_1 - f_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -3,45 \quad \text{avec } \hat{p} = 0,773$$

Décision : avec $\alpha = 0,05$, la valeur théorique, lue dans la table de l'écart centré réduit, vaut $-z_\alpha = -1,64$ (il s'agit d'un test unilatéral). Comme $z < -z_\alpha$, H_0 est rejetée au risque d'erreur 0,05.

On peut regarder le risque critique, c'est-à-dire le risque minimal qu'il faut prendre pour rejeter H_0. La valeur $z = -3,45$ correspond à une probabilité critique $\alpha \approx 0,001$ (p-value).

Comme $\alpha < 0,001$, le risque d'erreur de type I, c'est-à-dire de rejeter H_0 alors qu'elle est vraie, est très faible. On peut donc rejeter l'hypothèse H_0 avec un risque pratiquement nul de se tromper.

Comme espéré, le taux de réussite est significativement plus grand lorsque l'assiduité aux TD est plus élevé.

Chapitre 9

Tests sur Moyennes et Variances

9.1 Test sur les moyennes

9.1.1 Test de conformité

On se donne un échantillon de n observations extrait d'une population gaussienne de moyenne μ. On souhaite tester cette moyenne vis-à-vis de la valeur μ_0. Le test de conformité d'une moyenne relatif à l'hypothèse nulle

$$H_0 : \mu = \mu_0$$

sera réalisé en utilisant la moyenne $\overline{x}$ et l'écart-type estimé s. On a

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \text{ suit une loi de Student à } \nu = n - 1 \text{ degrés de liberté.}$$

On calcule donc la valeur

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}.$$

Décision :
- Si l'hypothèse alternative est $H_1 : \mu \neq \mu_0$ (cas bilatéral) : rejet de H_0 au risque α si $t \notin]-t_{\alpha/2}; t_{\alpha/2}[$ avec $\nu = n - 1$ degrés de liberté.
- Si l'hypothèse alternative est $H_1 : \mu > \mu_0$ (cas unilatéral) : rejet de H_0 au risque α si $t > t_\alpha$ avec $\nu = n - 1$ degrés de liberté.
- Si l'hypothèse alternative est $H_1 : \mu < \mu_0$ (cas unilatéral) : rejet de H_0 au risque α si $t < -t_\alpha$ avec $\nu = n - 1$ degrés de liberté.

Pour la décision, il y a deux façons de procéder :
- Soit on définit un risque a priori : on utilise assez systématiquement un risque $\alpha = 5\%$ dans beaucoup de domaines (biologie, médecine). On l'abaisse si nécessaire après (dans le cas où une erreur de type I pourrait avoir des conséquences jugées graves)
- Soit on se décide sur le risque a posteriori : la plupart des logiciels de statistique donne le risque minimal qu'il faut prendre pour rejeter H_0. On note par valeur p (en anglais : p-value), le plus petit niveau de risque auquel on rejette l'hypothèse nulle. En d'autres termes, la valeur p est la probabilité de commettre une erreur de première espèce, c'est-à-dire de rejeter à tort l'hypothèse nulle et donc d'obtenir un faux négatif. Par exemple dans le cas d'un test bilatéral,

$$p\text{-value} = 2P\left(\left|\frac{\overline{x} - \mu_0}{s/\sqrt{n}}\right| > t_{p/2} \mid H_0 : \mu = \mu_0\right).$$

La règle de décision en sera simplifiée : H_0 sera rejetée lorsque p-value $< \alpha$.

Si la variance de la population est connue, l'écart-type estimé est remplacé par sa vraie valeur et la valeur théorique est lue dans la table de l'écart réduit au lieu de la table de Student (cela correspond à un degré de liberté infini). Dans ce cas,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \text{ suit la loi normale centrée réduite.}$$

On comparera $z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ à une valeur lue dans la table de l'écart réduit.

Décision :
– Si l'hypothèse alternative est $H_1 : \mu \neq \mu_0$ (cas bilatéral) : rejet de H_0 au risque α si $z \notin]-z_{\alpha/2}; z_{\alpha/2}[$.
– Si l'hypothèse alternative est $H_1 : \mu > \mu_0$ (cas unilatéral) : rejet de H_0 au risque α si .
– Si l'hypothèse alternative est $H_1 : \mu < \mu_0$ (cas unilatéral) : rejet de H_0 au risque α si .
Dans le dernier cas, la valeur p sera

$$p\text{-value} = P\left(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -t_p \mid H_0 : \mu = \mu_0\right)$$

Exemple :

Une compagnie de vente de licences de nouveaux logiciels e-commerce fait la publicité que les entreprises utilisant ce logiciel peuvent obtenir, en moyenne pendant la première année, un rendement de 10% sur leurs investissements initiaux. Les rendements affichés pour un échantillon aléatoire de 10 de ces franchises pour la première année de fonctionnement sont les suivants :

$$6,1 \quad 9,2 \quad 11,5 \quad 8,6 \quad 12,1 \quad 3,9 \quad 8,4 \quad 10,1 \quad 9,4 \quad 8,9$$

En supposant que les rendements de la population sont normalement distribués, tester l'affirmation de la compagnie. ($n = 10,\ \overline{x} = 8.82,\ s = 2.4,\ t = -1.55$, p-value = LOI.STUDENT(1.55 ; 9 ;2)=0.1546). On accepte H_0 au risque $\alpha = 5\%$ car p-value $\geq \alpha$. On a considéré qu'il s'agit d'un test bilatéral (ce qui peut être contestable ici). Avec le test unilatéral ($H_1 : r < 10\%$), on rejette à 10% (p-valeur$\approx$ 8%).

9.1.2 Test d'homogénéité : populations indépendantes

On s'intéresse à la différence entre les moyennes μ_1 et μ_2 au sein de deux populations au travers de deux échantillons indépendants.

On suppose que les deux échantillons, respectivement de n_1 et n_2 observations, sont extraits de populations gaussiennes qui ont une variance (inconnue) commune σ^2 c'est-à-dire $\sigma_1^2 = \sigma_2^2 = \sigma^2$. On **testera cette égalité de variance** si elle ne peut être supposée.

On considère la variable qui suit les différences entre $\overline{X_1}$ et $\overline{X_2}$. Elle suit une loi normale de moyenne $(\mu_1 - \mu_2)$ et de variance

$$V(\overline{X_1} - \overline{X_2}) = V(\overline{X_1}) + V(\overline{X_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}.$$

Ainsi

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \text{ suit la loi normale centrée réduite.}$$

Lorsque la variance commune est inconnue, on l'estime par

$$\hat{s}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Le test d'hypothèse utilisera alors une loi t de Student :

$$T = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\hat{S}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ suit une loi de Student à } \nu = n_1 + n_2 - 2 \text{ degrés de liberté.}$$

L'hypothèse nulle (l'hypothèse à tester) et l'alternative sont les suivantes :

$$H_0 : \mu_1 = \mu_2 \quad \text{ou} \quad \mu_1 - \mu_2 = 0$$

La statistique t est la suivante :

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\hat{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

Décision :
- Si l'hypothèse alternative est $H_1 : \mu_1 \neq \mu_2$ (cas bilatéral) : rejet de H_0 au risque α si $t \notin]-t_{\alpha/2}; t_{\alpha/2}[$ où $-t_{\alpha/2}$ sera lu avec $\nu = n_1 + n_2 - 2$ degrés de liberté.
- Si l'hypothèse alternative est $H_1 : \mu_1 > \mu_2$ (cas unilatéral) : rejet de H_0 au risque α si .
- Si l'hypothèse alternative est $H_1 : \mu_1 < \mu_2$ (cas unilatéral) : rejet de H_0 au risque α si .

Dans le cas où les variances sont inconnues mais supposées différentes, le test reste le t de Student avec un degré de liberté égal à

$$\nu = \frac{\left[s_1^2/n_1 + s_2^2/n_2\right]^2}{\left(\frac{s_1^2}{n_1}\right)^2/(n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2/(n_2 - 1)}.$$

On comparera

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

au risque α avec une valeur t_α ou $t_{\alpha/2}$ selon le cas (test unilatéral ou bilatéral) lue avec ν degrés de liberté.

9.1.3 Test d'homogénéité : populations appariées

On observe un échantillon de n paires d'observations que l'on notera $(x_1, y_1), \cdots, (x_n, y_n)$, extraites de populations de moyennes μ_X et μ_Y. Soit

$$\overline{D} = \overline{X} - \overline{Y}$$

et S_D les variables aléatoires respectivement de la moyenne observée et de l'écart-type estimé des différences entre les paires des échantillons.

On suppose que la distribution des différences est gaussienne.

On se ramène à tester une moyenne observée et une moyenne théorique : l'hypothèse nulle sera

$$H_0 : \mu_X - \mu_Y = D_0$$

et la variable

$$\frac{\overline{D} - D_0}{S_D/\sqrt{n}} \text{ suit une distribution } t \text{ de Student à } \nu = n - 1 \text{ degrés de liberté.}$$

On calculera

$$t = \frac{\overline{d} - D_0}{s_D/\sqrt{n}}.$$

Décision :

– Si l'hypothèse alternative est $H_1 : \mu_X - \mu_Y \neq D_0$ (cas bilatéral) : rejet de H_0 au risque α si $t \notin]-t_{\alpha/2}; t_{\alpha/2}[$ avec $\nu = n-1$ degrés de liberté.
– Si l'hypothèse alternative est $H_1 : \mu_X - \mu_Y > D_0$ (cas unilatéral) : rejet de H_0 au risque α si $t > t_\alpha$ avec $\nu = n-1$ degrés de liberté.
– Si l'hypothèse alternative est $H_1 : \mu_X - \mu_Y < D_0$ (cas unilatéral) : rejet de H_0 au risque α si $t < -t_\alpha$ avec $\nu = n-1$ degrés de liberté.

9.2 Test sur les variances

9.2.1 Test de conformité

Il s'agit d'une comparaison d'une variance expérimentale et d'une variance théorique, ou encore de l'étude de l'influence d'un facteur A sur une population $\mathcal{P}$ et sur un échantillon.

Dans la population, on connaît la variance σ_0^2 des valeurs.

Soit un échantillon E de taille n. On calcule dans cet échantillon, la moyenne $\overline{x}$ et la variance s^2 expérimentales.

Hypothèse nulle : $H_0 : \sigma^2 = \sigma_0^2$ (la variance expérimentale de l'échantillon est conforme à celle de la population)

Hypothèse alternative
• $H_1 : \sigma^2 \neq \sigma_0^2$ (test bilatéral)
• $H_1 : \sigma^2 > \sigma_0^2$ (test unilatéral)
• $H_1 : \sigma^2 < \sigma_0^2$ (test unilatéral)
Sous l'hypothèse que la distribution des données dans la population est normale, la variable aléatoire

$$Y^2 = \frac{n-1}{\sigma_0^2} S^2$$

suit une loi du χ^2 à $\nu = n-1$ degrés de liberté.

On calcule $y^2 = \frac{n-1}{\sigma^2} s^2$ et on compare cette valeur à une valeur lue dans la table du χ^2 à $\nu = n-1$ degrés de liberté.

Décision
• Dans le cas d'un test bilatéral,
▷ Si $n \leq 30$ (la table ne contient pas des degrés de liberté supérieurs à 30), on cherche a tel que $P(\chi^2 < a) = \alpha/2$ (ou $P(\chi^2 \geq a) = 1 - \alpha/2$ et b tel que $P(\chi^2 \geq b) = \alpha/2$. Ainsi

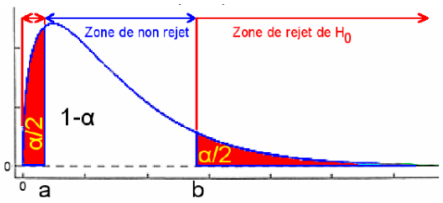


FIGURE 9.1 – Loi χ^2 : Zones de rejet de l'hypothèse nulle

– Si $y^2 \notin]a; b[$, on rejette H_0 (la variance expérimentale n'est pas conforme à la variance théorique : la variance expérimentale est différente de celle de la population).
– Sinon H_0 n'est pas rejetée. Rien ne permet de dire que la variance expérimentale n'est pas conforme à la variance de la population.

▷ Si $n > 30$, la variable aléatoire $Z = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ suit à peu près une loi normale centrée réduite. On rejettera H_0 lorsque $z = \sqrt{2y^2} - \sqrt{2n-3} \notin]-z_{\alpha/2}, z_{\alpha/2}[$.

- Si $H_1 : \sigma^2 > \sigma_0^2$, on cherche b tel que $P(\chi^2 \geq b) = \alpha$. Si $y^2 > b$, on rejette H_0 : la variance expérimentale est supérieure à celle de la population.
- Si $H_1 : \sigma^2 < \sigma_0^2$, on cherche a tel que $P(\chi^2 \leq a) = \alpha$. Si $y^2 < a$, on rejette H_0 : la variance expérimentale est inférieure à celle de la population.

Exemple : Une société produit des dispositifs électriques gérés par un contrôle thermostatique. L'écart-type de la température à laquelle ces contrôles fonctionnent ne devrait en réalité pas excéder 2.0 degrés. Pour un échantillon aléatoire de 20 de ces commandes, l'écart-type d'un échantillon aléatoire de températures d'exploitation était 2.36 degrés. Effectuer un test au seuil de 5 % de l'hypothèse nulle selon laquelle l'écart-type de population est 2.0 contre l'alternative selon laquelle cet écart est en réalité plus grand (Vous énoncerez et supposerez les hypothèses nécessaires au test)

($\chi_c^2 = 26.45$, $\chi_\alpha^2 = 30.14$; on ne peut pas rejeter H_0)

9.2.2 Test d'homogénéïté

Ce test est nécessaire pour valider l'hypothèse d'égalité des variances du paragraphe 9.1.2.

On souhaite comparer les variances de deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On dispose de deux échantillons. Soit s_1^2 la variance d'un échantillon aléatoire de n_1 observations extrait d'une population gaussienne $\mathcal{P}_1$ de variance σ_1^2. On dispose indépendamment d'un deuxième échantillon aléatoire de taille n_2 et de variance s_1^2 extrait d'une population gaussienne $\mathcal{P}_2$ de variance σ_2^2. Alors la variable aléatoire

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

suit une distribution F, construite comme le rapport de deux variables aléatoires suivant chacune une loi du χ^2, avec un degré de liberté au dénominateur égal à $(n_1 - 1)$ et un degré de liberté au dénominateur égal à $(n_2 - 1)$. On notera F_{ν_1,ν_2} avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$.

Soit H_0 l'hypothèse nulle $\sigma_1^2 = \sigma_2^2$. Sous H_0 (lorsque les variances des populations sont égales), la variable aléatoire devient

$$F = \frac{S_1^2}{S_2^2}.$$

Ainsi, on calcule le rapport

$$f = \frac{s_1^2}{s_2^2}.$$

Dans les applications pratiques, pour comparer correctement avec les valeurs théoriques limites de la table F, on s'arrange pour que ce rapport soit supérieur à 1 en échangeant le rôle des deux échantillons si nécessaire.

Décision

- Si $H_1 : \sigma_1^2 > \sigma_2^2$, on cherche f_α tel que $P(F_{(n_1-1,n_2-1)} \geq f_\alpha) = \alpha$. Si $f > f_\alpha$, on rejette H_0.
- Si $H_1 : \sigma_1^2 \neq \sigma_2^2$, on cherche $f_{\alpha/2}$ tel que $P(F_{(n_1-1,n_2-1)} \geq f_{\alpha/2}) = \alpha/2$. Si $f > f_{\alpha/2}$, on rejette H_0 (la règle semble être une règle pour un test unilatéral mais il s'agit bien d'un test bilatéral au risque α, le complémentaire étant testé avec la règle du rapport $f > 1$).

Exemple : On suppose que le total des ventes d'une société devrait varier plus dans une industrie où la concurrence des prix est active que dans un duopole avec la collusion tacite.

Dans une étude de l'industrie de production de marchandises, il a été constaté que, sur une période de quatre ans de concurrence active des prix, la variance du total des ventes d'une compagnie était de 114,09. Au cours des sept années suivantes, dans laquelle on peut supposer collusion tacite, la variance était de

16,08. Supposons que les données peuvent être considérées comme un échantillon aléatoire indépendant de deux distributions normales. Tester au seuil de 5 %, l'hypothèse nulle selon laquelle les deux variances de population sont égales contre l'hypothèse alternative que la variance du total des ventes est plus élevée dans les années de concurrence active des prix.

($f = 7.095$; $f_\alpha = 4.76$ ($\nu_1 = 3, \nu_2 = 6$) =INVERSE.LOI.F(0,05 ;3 ;6) ; H_0 rejetée)

Exercice : (ronces) La taille des feuilles de ronces ont été mesurées pour voir si il y a une différence entre la taille des feuilles qui poussent en plein soleil et celles qui poussent à l'ombre. Les résultats sont les suivants (Largeur des feuilles en cm)

Soleil	6.0	4.8	5.1	5.5	4.1	5.3	4.5	5.1
Ombre	6.5	5.5	6.3	7.2	6.8	5.5	5.9	5.5

$$\overline{x_1} = 5.05 \quad s_1 = 0.59 \quad n_1 = 8$$

$$\overline{x_2} = 6.15 \quad s_2 = 0.65 \quad n_2 = 8$$

$\hat{s} = 0.62 \quad t = 3.55 \quad t_{\alpha/2} = 2.145$

Chapitre 10

Région critique, risque β

Ref : Statistique, exercices corrigés, Tome 3, Christian Labrousse

Soit une variable aléatoire X dont la loi de probabilité $\mathcal{L}(X)$ dépend d'un paramètre θ. La densité de probabilité est $f(x_i, \theta)$. Le paramètre θ inconnu peut prendre deux valeurs θ_0 et θ_1.

On dispose d'un échantillon de la variable aléatoire X de taille n : $x_1, x_2, \cdots, x_n$. Cet échantillon peut être représenté par un point M de coordonnées $(x_1, x_2, \cdots, x_n)$.

Les hypothèses H_0 et H_1 peuvent être caractérisées par les fonctions de vraisemblance :
– Pour H_0, $L_0(M) = L(x_1, x_2, \ldots, x_n, \theta_0) = \prod_{i=1}^{n} f(x_i, \theta_0)$;
– Pour H_1, $L_1(M) = L(x_1, x_2, \ldots, x_n, \theta_1) = \prod_{i=1}^{n} f(x_i, \theta_1)$.
La région critique ω_0 est définie par α et β. Or

$$\begin{aligned} \alpha &= P(\text{décider } H_1/H_0 \text{ vraie}) = P(M \in \omega_0/H_0 \text{ vraie}) = \int_{\omega_0} L_0(M)dM; \\ \beta &= P(\text{décider } H_0/H_1 \text{ vraie}) = P(M \notin \omega_0/H_1 \text{ vraie}) = \int_{\overline{\omega_0}} L_1(M)dM; \end{aligned}$$

Principe de la méthode de Neyman et Pearson.

On fixe le risque de première espèce $\alpha = \alpha_0$. Parmi toutes les régions critiques possibles, on choisit celle qui minimise le risque de seconde espèce β, ou encore qui maximise la quantité $1 - \beta = \eta$, appelée puissance du test. Or,

$$\begin{aligned} \eta &= 1 - \beta = 1 - \int_{\overline{\omega_0}} L_1(M)dM = \int_{\omega_0} L_1(M)dM, \\ \eta &= \int_{\omega_0} \frac{L_1(M)}{L_0(M)} L_0(M)dM. \end{aligned}$$

Construction pratique de la région critique ω_0. A chaque point de $\mathbb{R}^n$ est attaché l'indicateur

$$r(M) = \frac{L_1(M)}{L_0(M)}.$$

Pour maximiser η, on recherche les points M tels que

$$r(M) \geq C,$$

soit $\frac{L_1(M)}{L_0(M)} \geq C$ ou encore $\frac{L_0(M)}{L_1(M)} \leq \frac{1}{C} = k$.

La région critique ω_0 est définie, selon un test de Neyman et Pearson, par le rapport des fonctions de vraisemblance associées aux deux hypothèses H_0 et H_1. La constante $k = 1/C$ est déterminée par

$$\int_{r(M)\geq C} L_0(M)dM = \alpha.$$

Remarquons que les risques α et β sont antagonistes car plus ω_0 est petit, plus $\overline{\omega_0}$ est grand, et réciproquement.

Exemple (Décision par test de Neyman et Pearson) : On se propose de tester la qualité d'un lot important de pièces mécaniques. Soit X une caractéristique aléatoire de ces pièces dont la loi de probabilité est une loi normale de moyenne m et d'écart-type $\sigma = 4$. A la suite d'erreurs survenant lors de la fabrication de ces pièces, on ignore si m égale 20 ou si m égale 22. On doit néanmoins prendre une décision. Pour cela on prélève dans un lot un échantillon aléatoire de 25 pièces. Quelle décision doit-on prendre ?

Solution. L'échantillon $(x_1, x_2, \ldots, x_n)$ est de taille $n = 25$. Soit $\overline{x}$ la moyenne de cet échantillon. Construisons un test selon la méthode de Neyman et Pearson. Soit ω_0 la région critique définie par :

$$\frac{L_0(M)}{L_1(M)} \leq k,$$

où $L_0(M)$ et $L_1(M)$ sont les fonctions de vraisemblance, associées respectivement aux hypothèses H_0 et H_1 :

$$\begin{aligned} H_0 &: m = m_0 = 20 \\ H_1 &: m = m_1 = 22. \end{aligned}$$

La densité de probabilité $f(x, m)$ d'une loi normale, de moyenne m et d'écart-type σ est :

$$f(x, m) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

La fonction de vraisemblance $L_0(M)$ est :

$$L_0(M) = f(x_1, m_0) \cdot f(x_2, m_0) \cdots f(x_n, m_0),$$

soit

$$L_0(M) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_0)^2\right).$$

La fonction de vraisemblance $L_1(M)$ est :

$$L_1(M) = f(x_1, m_1) \cdot f(x_2, m_1) \cdots f(x_n, m_1),$$

soit

$$L_1(M) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_1)^2\right).$$

Formons le rapport :

$$\frac{L_0(M)}{L_1(M)} = \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - m_1)^2 - \sum_{i=1}^{n}(x_i - m_0)^2\right]\right\}.$$

La région critique, étant définie par

$$\frac{L_0(M)}{L_1(M)} \leq k,$$

l'est encore par

$$\log_e \frac{L_0(M)}{L_1(M)} \leq \log_e k.$$

Il vient ici :

$$\log_e \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-m_1)^2-\sum_{i=1}^{n}(x_i-m_0)^2\right]\right\} \leq \log_e k,$$

soit

$$\sum_{i=1}^{n}(x_i-m_1)^2-\sum_{i=1}^{n}(x_i-m_0)^2 \leq 2\sigma \log_e k.$$

En développant les sommations :

$$n(m_1^2-m_0^2)+2(n\overline{x})(m_0-m_1) \leq 2\sigma \log_e k,$$

soit

$$n(m_0-m_1)[2\overline{x}-(m_0+m_1)] \leq 2\sigma \log_e k.$$

La quantité (m_0-m_1) est négative :

$$m_0-m_1 = 20-22 = -2.$$

Il est alors nécessaire de changer le sens de l'inégalité, en isolant la moyenne $\overline{x}$ de l'échantillon :

$$2\overline{x}-(m_0+m_1) \geq \frac{2\sigma^2 \log_e k}{n(m_0-m_1)},$$

d'où

$$\overline{x} \geq \frac{\sigma^2}{n(m_0-m_1)} \log_e k + \frac{m_0+m_1}{2}.$$

Désignons cette dernière quantité par π. Avec les données numériques :

$$m_0 = 20 \quad m_1 = 22 \quad \sigma = 4 \quad n = 25,$$

la région critique ω_0 est déterminée par :

$$\overline{x} \geq \pi, \text{ avec } \pi = 21 - 0.32 \log_e k.$$

La quantité π s'appelle le seuil critique.

La loi de probabilité de la moyenne $\overline{x}$ de l'échantillon aléatoire est une loi normale de moyenne m_0 ou m_1 et d'écart-type

$$\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{25}} = \frac{4}{5}$$

Désignons par $f(\overline{x})$ la densité de probabilité correspondante.

La règle de décision, du test de Neyman et Pearson est :
– décider $H_0(m = m_0 = 20)$, si $M \notin \omega_0$;
– décider $H_1(m = m_1 = 22)$, si $M \in \omega_0$.
Pour que M appartienne à ω_0, il faut que la moyenne $\overline{x}$ soit supérieure ou égale à π :

$$\overline{x} \geq 21 - 0.32 \log_e k.$$

L'erreur de première espèce α est égale à la probabilité de décider H_1, alors que l'hypothèse H_0 est vraie :

$$\begin{aligned}
\alpha &= P(\{\text{décider } H_1 \text{ / } H_0 \text{ vraie}\});\\
\alpha &= P(\{M \in \omega_0 \text{ / } H_0 \text{ vraie}\});\\
\alpha &= P(\{M \in \omega_0 \text{ / } m = m_0 = 20\});\\
\alpha &= P(\{\overline{x} \geq \pi \text{ / } m = 20\}).
\end{aligned}$$

La loi de probabilité de $\overline{X}$ étant la loi $\mathcal{N}(m; \frac{4}{5})$, faisons le changement de variable

$$Z = \frac{\overline{X} - \text{moyenne}}{\text{écart-type}},$$

soit :

$$Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - 20}{\frac{4}{5}};$$

d'où

$$\pi = \frac{4}{5}z + 20;$$

La variable aléatoire Z suit la loi $\mathcal{N}(0; 1)$.

Selon le principe de la méthode de Neyman, fixons $\alpha = \alpha_0 = 0.05$:

$$\alpha = 0,05 = \int_{z_\alpha}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-U^2/2} dU,$$

ou encore

$$0,95 = \int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-U^2/2} dU = \Phi(z_\alpha).$$

Dans la table intégrale de la loi $\mathcal{N}(0; 1)$, on trouve

$$\Phi(z_\alpha = 1.65) = 0,9505.$$

Il vient alors $z_\alpha = 1.65$. Nous déterminons le seuil critique π :

$$\begin{aligned}
\pi &= \frac{4}{5}(1.65) + 20,\\
\pi &= 21,32.
\end{aligned}$$

La région critique ω_0 est donc déterminée par :

$$\overline{x} \geq \pi = 21,32.$$

La règle de décision est ainsi :
– si $\overline{x} < 21.32$, on décide $H_0(m = m_0 = 20)$;
– si $\overline{x} \geq 21.32$, on décide $H_1(m = m_1 = 22)$.
Calculons la puissance du test η :

$$\eta = 1 - \beta,$$

où β est l'erreur de seconde espèce :

$$\beta = \int_{-\infty}^{z_\beta} \frac{1}{\sqrt{2\pi}} e^{-U^2/2} dU = \Phi(\beta).$$

La loi de probabilité correspondante étant $\mathcal{N}(22;\frac{4}{5})$, il vient :

$$\begin{aligned} z_\beta &= \frac{\pi - 22}{\frac{4}{5}} = \frac{21.32 - 22}{0.8}, \\ z_\beta &= -0.85, \end{aligned}$$

d'où la valeur de β :

$$\begin{aligned} \beta &= \Phi(-0.85) = 1 - \Phi(0.85) = 1 - 0.8023, \\ \beta &= 0.1977 \end{aligned}$$

le risque de seconde espèce est :

$$\beta \approx 0.20.$$

La puissance du test est :

$$\eta \approx 0.80.$$

A titre indicatif, déterminons la constante k :

$$\pi = 21 - 0.32 \log_e k = 21.32,$$

$$\log_e k = \frac{21 - 21.32}{0.32} = -1,$$

soit

$$k = e^{-1} = 0.368.$$

La région critique ω_0 est ainsi définie par :

$$\frac{L_0(M)}{L_1(M)} \leq 0,368.$$

Chapitre 11

Tests non paramétriques

Contrairement aux tests paramétriques qui nécessitent que les données soient issues d'une distribution paramétrée, les tests non paramétriques ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests *distribution free*. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire (test de conformité en loi). En contrepartie, ils sont moins puissants que les tests paramétriques lorsque les hypothèses sur les données peuvent être validées.

Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation tests de rangs est souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

11.1 Test de Mann-Whitney

Le test de Mann-Whitney est un test non paramétrique qui permet de tester si deux échantillons issus de populations indépendantes ont même moyenne. Les valeurs doivent être numériques (i.e. pouvoir être ordonnées). Il ne nécessite pas de connaître les valeurs des échantillons mais leurs rangs. On suppose donc toujours que l'on dispose de deux échantillons $x_1, \ldots, x_n$ et $y_1, \ldots, y_m$ issus de variables numériques ou ordinales indépendantes, de lois respectives $\mathcal{L}_X$ et $\mathcal{L}_Y$. On teste $H_0 : \mathcal{L}_X = \mathcal{L}_Y$ ou encore par rapport aux fonctions de distribution $H_0 : F_X = F_Y$.

Le test de Mann-Whitney compte le nombre de couples pour lesquels $X_i < Y_j$. L'alternance des X_i et des Y_j devrait être assez régulière sous H_0. On aura des doutes sur H_0 si les Y_j sont plutôt plus grands que les X_i, ou plus petits ou plus fréquents dans une certaine plage de valeurs.

Statistique du test de Mann-Whitney :

$$U_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} 1_{\{x<y\}}(X_i, Y_j),$$

où $1_{\{x<y\}}(X_i, Y_j)$ vaut 1 si $X_i < Y_j$, 0.5 si $X_i = Y_j$ et 0 sinon.

C'est le nombre de termes Y_j supérieurs à la médiane de $X \cup Y$.

On comptera, pour chaque valeur x_i du premier échantillon, le nombre de valeurs y_j du deuxième échantillon telles que $y_j \geq x_i$ (On comptera 0.5 pour chaque y_j est égal à x_i). On notera U_1 cette valeur obtenue à partir du premier échantillon et U_2 la valeur trouvée en échangeant les rôles des échantillons. Seule la plus petite des deux valeurs trouvées sera comparée aux tables.

On peut également calculer cette statistique en considérant la somme R_1 de tous les rangs après ordonnancement des observations de la première population. On a alors

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}.$$

On aura de même,

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

où R_2 est la somme des rangs du deuxième échantillon.

En sachant que $R_1 + R_2 = N(N+1)/2$ avec $N = n_1 + n_2$, on trouve que

$$U_1 + U_2 = n_1 n_2.$$

Cela permet de vérifier le calcul des valeurs U_1, U_2 ou de calculer l'une à partir de l'autre.

Règle de décision : dans le cas d'un test bilatéral, on prend $u = \min(u_1, u_2)$. On rejette H_0 si $u \in [0, m_\alpha]$ avec m_α donné par la table de Mann et Whitney.

En supposant l'hypothèse nulle que les positions centrales des deux populations sont les mêmes, la variable U de Mann-Whitney vérifie

$$E(U) = \frac{n_1 n_2}{2} \qquad \mathrm{Var}(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Ainsi pour des échantillons de grande taille, la distribution de la variable aléatoire

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

est approximativement la loi normale centrée réduite.

Remarques : Le test de Mann-Whitney a donc le même objectif qu'un autre test d'identité important, le "Test du Chi-2 d'identité", dans sa version pour variable numérique. Si les populations sont supposées normales et de même variance, le test T aura la préférence.

Le test de Kruskal-Wallis peut être perçu comme une extension du test de Mann-Whitney à plus de deux échantillons (de même que ANOVA univariée est une extension du test t à plus de deux échantillons).

Exemple : La taille des feuilles de ronces ont été mesurées pour voir si il y a une différence entre la taille des feuilles qui poussent en plein soleil et celles qui poussent à l'ombre. Les résultats sont les suivants (Largeur des feuilles en cm)

Soleil	6.0	4.8	5.1	5.5	4.1	5.3	4.5	5.1
Ombre	6.5	5.5	6.3	7.2	6.8	5.5	5.9	5.5

Valeurs ordonnées

E1	4.1	4.5	4.8	5.1	5.1	5.3		5.5				6.0				
E2							5.5		5.5	5.5	5.9		6.3	6.5	6.8	7
rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
rang moyen	1	2	3	4.5	4.5	6	8.5	8.5	8.5	8.5	11	12	13	14	15	16

$$\begin{aligned} U_1 &= 8+8+8+8+8+8+6.5+4 = 58.5 \\ U_2 &= 1.5+1.5+1.5+1+0+0+0+0 = 5.5 \end{aligned}$$

$$\begin{aligned} R_1 &= 1+2+3+4.5+4.5+6+8.5+12 = 41.5 \\ R_2 &= 8.5+8.5+8.5+11+13+14+15+16 = 94.5 \end{aligned}$$

Dans tous les cas, on obtient la valeur $U = \min(U_1, U_2) = 5.5$.

Ensuite on utilise la table de Mann-Whitney au risque de 5% ($n_1 = 8, n_2 = 8$), pour obtenir une valeur $m_\alpha = 13$.

On rejettera l'hypothèse nulle si U est inférieure à la valeur m_α. Dans le cas de l'exemple, comme $U < m_\alpha$, on rejette H_0. La différence entre la taille des feuilles à l'ombre et au soleil est donc significative au risque $\alpha = 5\%$.

11.2 Test de Wilcoxon (*Wilcoxon signed rank test*)

Le test de Wilcoxon est un test non paramétrique qui permet de tester si deux populations appariées ont même moyenne en se basant sur deux échantillons.

Sur les N paires observées, il reste qu'un échantillon de n différences non nulles (on enlève les éléments de différence nulle)

Soient d_i (pour $i = 1$ à n) les différences entre chaque paire d'observations. Nous classons les rangs des valeurs absolues de ces différences. La statistique de Wilcoxon tient compte uniquement des rangs des observations.

La statistique de rangs signés de Wilcoxon s'écrit :

$$W = \min(\sum_{d_i>0} r_i, \sum_{d_i<0} r_i).$$

Règle de décision : On ne peut rejeter H_0 si $W \in]W_{\alpha/2}, W_{1-\alpha/2}[$ avec $W_{1-\alpha/2} = n(n+1)/2 - W_{\alpha/2}$. Les tables ne donnent que $W_{\alpha/2}$: on rejette H_0 lorsque $W < W_{\alpha/2}$ dans le cas bilatéral.

Dans le cas des « grands » échantillons, lorsque n est supérieur à 25, il peut être démontré que la somme des rangs est pratiquement normale ; on utilise alors l'approximation normale

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim \mathcal{N}(0,1).$$

Remarque : Il existe d'autres tests. Par exemple, le test des signes n'utilise que l'information sur la direction des différences entre paires. Si nous pouvons prendre en compte, en plus, la grandeur des différences, un test plus puissant peut être utilisé : le test de Wilcoxon donne plus de poids à une paire qui montre une large différence entre les deux conditions, qu'à une paire ayant une faible différence. Cela implique que l'on puisse dire quel membre d'une paire est plus grand que l'autre (donner le signe de la différence), mais aussi que l'on puisse ranger les différences en ordre croissant.

Exemple : Un échantillon aléatoire de dix étudiants est consulté pour noter, dans un test à l'aveugle, la qualité de deux types de bière, l'une locale, l'autre importée. Les notes sont sur une échelle de 1 (pas bon) à 10 (excellent). Utiliser le test de Wilcoxon pour tester l'hypothèse nulle "la distribution des différences entre paires est centrée sur zéro" contre l'hypothèse alternative "'La population des étudiants buveurs de bières préfère la catégorie d'importation.

Etudiant	Locale	Import	Etudiant	Locale	Import
A	2	6	F	4	8
B	3	5	G	3	9
C	7	6	H	4	6
D	8	8	I	5	4
E	7	5	J	6	9

Différences : -4 ; -2 ; 1 ; 0 ; 2 ; -4 ; -6 ; -2 ; 1 ; -3

Tri	0	1	1	-2	2	-2	-3	-4	-4	-6
rang	-	1	2	3	4	5	6	7	8	9
rang moyen		1,5	1,5	4	4	4	6	7,5	7,5	9
$r_i > 0$		1,5	1,5		4					
$r_i < 0$				4		4	6	7,5	7,5	9

Ainsi $W = \min(7, 38) = 7$

On a $W_{0.05} = 8$ (test unilatéral).

11.3 Test de Corrélation de rang de Spearman

Pour valider l'existence d'un lien entre deux variables, on réalise ordinairement une régression linéaire simple, voire une régression non linéaire. La qualité du lien supposé est mesurée par le coefficient de corrélation (dit « de Pearson »). Cependant, il existe des situations pour lesquelles une mesure de la corrélation sur les valeurs est inadaptée. Si les variables sont ordinales, discrètes, ou si des valeurs extrêmes risquent de biaiser les résultats, ou encore que les valeurs en elles-mêmes n'ont que peu d'importance, ou enfin qu'elles ne suivent pas une loi normale, il nous reste un joker : les corrélations des rangs.

On n'utilise alors pas les VALEURS des observations dans les calculs mais leur RANG.

Le rang de chaque élément dans la série croissante de X et de Y sera calculé. On calcule ensuite le coefficient de corrélation r_s entre la série des rangs de X et la série des rangs de Y. On peut retrouver cette valeur ainsi :

1. calculer la différence de classement d_i pour chaque couple de valeur (r_{x_i}, r_{y_i}).
2. la valeur r_s sera donnée par
$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2.$$

La variable R_s sous l'hypothèse d'indépendance des deux variables a pour espérance $E(R_s) = 0$ et pour variance $V(R_s) = \frac{1}{n-1}$.

Si $n > 30$ alors $Z = \frac{R_s - E(R_s)}{\sqrt{V(R_s)}} = R_s\sqrt{n-1}$ suit la loi normale centrée réduite. Si $n \leq 30$, les valeurs théoriques sont données dans la table du coefficient r de Spearman.

Exemple : Placez les enfants dans une classe, par ordre ascendant en fonction de leur taille, en prenant note du rang de chaque enfant (premier, deuxième, troisième, etc.), du plus court au plus grand. Vous les placez ensuite en fonction de leur poids, puis vous prenez note de leur rang. Est-ce que chaque enfant occupe le même rang, dans chacune des mesures ? Peut-être que oui, dans l'ensemble, bien qu'un enfant court puisse également être au-dessus de son poids ou qu'un enfant grand, être, lui aussi, en-dessous de son poids, ce qui les classerait dans un rang différent pour chaque variable. La corrélation des rangs démontre le degré de correspondance entre le classement hiérarchique d'un échantillonnage d'observations sur deux variables. Les formules de Kendall ou Spearman sont les variétés communes de ce type de corrélations, car elles donnent toutes les deux une valeur de -1,0 (classement inverse parfait) à 0,0 (aucun accord) à +1,0 (classement identique des deux variables).

On ordonne la taille et le poids sur 10 enfants. On obtient les résultats suivants :

Enfant n°	1	2	3	4	5	6	7	8	9	10
Taille	1	5	3	8	10	4	2	7	6	9
Poids	5	3	9	10	2	1	6	8	7	4

On trouve $r_s = -0,07$. On accepte H_0 : indépendance des deux caractères.

A Table de Mann-Whitney

Référence : Table A5.07 : Critical Values for the Wilcoxon/Mann-Whitney Test (U)

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	0	0	0	0	1	1	1	1	1	2	2	2	2
3	–	–	–	–	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	–	–	–	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	–	–	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	–	–	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	–	–	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	–	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	–	0	2	4	7	10	12	15	17	21	23	26	28	31	34	37	39	42	45	48
10	–	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	–	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	–	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	–	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	–	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	–	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	–	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	–	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	–	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	–	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	–	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Bilatéral $\alpha = .05$ (Unilatéral $\alpha = .025$)

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0
3	–	–	–	–	–	–	–	–	0	0	0	1	1	1	2	2	2	2	3	3
4	–	–	–			0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	–	–	–		0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	–	–	–	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	–	–	–	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	–	–	–	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	–	–	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	–	–	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	–	–	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	46
12	–	–	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	–	–	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	–	–	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	–	–	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	–	–	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	–	–	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	–	–	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	–	0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	–	0	3	8	13	18	24	30	36	42	46	54	60	67	73	79	86	92	99	105

Bilatéral $\alpha = .01$ (Unilatéral $\alpha = .005$)

B Table de Wilcoxon

Critical Values of the Wilcoxon Signed Ranks

	Test bilatéral		Test unilatéral	
n	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
5	-	-	0	-
6	0	-	2	-
7	2	-	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Ref : `http://facultyweb.berry.edu/vbissonnette/tables/wilcox_t.pdf`

Calcul des valeurs sur : `http://comp9.psych.cornell.edu/Darlington/wilcoxon/wilcox0.htm`

C Table du coefficient de rang de Spearman

Valeurs critiques pour un test unilatéral utilisant ρ.

n	5%	1%	n	5%	1%
4	1.000	$*$	18	.401	.550
5	.900	1.000	19	.391	.535
6	.829	.943	20	.380	.522
7	.714	.893	21	.370	.509
8	.643	.833	22	.361	.497
9	.600	.783	23	.353	.486
10	.564	.745	24	.344	.476
11	.536	.709	25	.337	.466
12	.503	.678	26	.331	.457
13	.484	.648	27	.324	.449
14	.464	.626	28	.318	.441
15	.446	.604	29	.312	.433
16	.429	.582	30	.306	.425
17	.414	.566	40	.264	.368

Les données de la table sont les plus petites valeurs de ρ (jusqu'à 3 décimales) qui correspondent à une probabilité $\leq 5\%$ (ou 1%) sur un seul côté. La valeur observée est significative si elle est supérieure ou égale à la valeur de la table. Le niveau de signification exact ne dépasse jamais la valeur nominale (5% ou 1%). La table peut également être utilisée pour les valeurs critiques à 10% et 2% d'un test bilatéral. L'étoile indique que la signification associée au risque proposé ne peut être calculée dans ce cas.

Valeurs critiques pour un test bilatéral utilisant ρ.

n	5%	1%	n	5%	1%
4	$*$	$*$	18	.472	.600
5	1.000	$*$	19	.460	.584
6	.886	1.000	20	.447	.570
7	.786	.929	21	.436	.556
8	.738	.881	22	.425	.544
9	.700	.883	23	.416	.532
10	.648	.794	24	.407	.521
11	.618	.755	25	.398	.511
12	.587	.727	26	.390	.501
13	.560	.703	27	.383	.492
14	.538	.679	28	.375	.483
15	.521	.654	29	.368	.475
16	.503	.635	30	.362	.467
17	.488	.618	40	.313	.405

Les données de la table sont les plus petites valeurs de ρ (jusqu'à 3 décimales) qui correspondent à une probabilité $\leq 5\%$ (ou 1%) sur les deux côtés. La valeur observée est significative si elle est supérieure ou égale à la valeur de la table. Le niveau de signification exact ne dépasse jamais la valeur nominale (5% ou 1%). La table peut également être utilisée pour les valeurs critiques à 2.5% et 0.5% d'un test unilatéral. L'étoile indique que la signification associée au risque proposé ne peut être calculée dans ce cas.

`Ref : http://www.answers.com/topic/critical-values-for-spearman-s-rank-correlation-coefficient`

Annexe A

Exercice : Montrons que $\frac{n}{n-1}F(1-F)$ converge en probabilité vers $p(1-p)$.

Calculons dans un premier temps $E((\sum X_i)^2)$.

$$\begin{aligned}
E\left((\sum X_i)^2)\right) &= \sum_{i=1}^{n} E(X_i^2) + \sum_{i=1}^{n} E\left(\sum_{i\neq j} X_i X_j\right) \\
&= \sum_{i=1}^{n} E(X_i) + \sum_{i=1}^{n} E\left(\sum_{j\neq i} X_i X_j\right) \quad \text{car } X_i = 0 \text{ ou } 1 \\
&= \sum_{i=1}^{n} E(X_i) + \sum_{i=1}^{n}\sum_{j\neq i} E(X_i)E(X_j) \quad \text{car } X_i, X_j \text{ indépendantes.} \\
E\left((\sum X_i)^2)\right) &= np + n(n-1)p^2.
\end{aligned}$$

Calculons maintenant $E(F(1-F))$.

$$\begin{aligned}
E\left(\frac{1}{n}\sum X_i(1-\frac{1}{n}\sum X_i)\right) &= \frac{1}{n}E\left(\sum X_i - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j)\right) \\
&= \frac{1}{n}(np) - \frac{1}{n^2}(E\left(\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j)\right) \\
&= p - \frac{1}{n^2}(np + n(n-1)p^2) = p(1-p)\left(1-\frac{1}{n}\right)
\end{aligned}$$

Donc $\frac{n}{n-1}F(1-F)$ est un estimateur sans biais de $p(1-p)$.

Pour calculer $E\left([F(1-F) - pq(1-1/n)]^2\right)$, nous avons besoin de quelques calculs intermédiaires : on a déjà vu

$$E((\sum X_i)^2) = np + n(n-1)p^2,$$

on aura besoin de $E((\sum X_i)^3)$, de $E((\sum X_i)^4)$ et de $E([F(1-F)]^2)$.

$$\begin{aligned} E\left((\sum X_i)^3)\right) &= \sum_{i=1}^{n} E(X_i^3) + \sum_{i=1}^{n}\sum_{j\neq i} E(X_i^2)E(X_j) + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j} E(X_i)E(X_j)E(X_k) \\ &= np + n(n-1)p^2 + n(n-1)(n-2)p^3. \end{aligned}$$

$$\begin{aligned} E\left((\sum X_i)^4)\right) &= \sum_{i=1}^{n} E(X_i^4) + \sum_{i=1}^{n}\sum_{j\neq i} E(X_i^3)E(X_j) + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j} E(X_i^2)E(X_j)E(X_k) \\ &\quad + \sum_{i=1}^{n}\sum_{j\neq i} E(X_i^2)E(X_j^2) + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} E(X_i)E(X_j)E(X_k)E(X_l) \\ &= np + n(n-1)(n-2)p^3 + 2n(n-1)p^2 + n(n-1)(n-2)(n-3)p^4. \end{aligned}$$

$$\begin{aligned} E([F(1-F)]^2) &= E((\frac{1}{n}\sum_{i=1}^{n} X_i)^2(1 - \frac{1}{n}\sum_{i=1}^{n} X_i)^2) \\ &= E((\frac{1}{n}\sum_{i=1}^{n} X_i)^2(1 - \frac{2}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} X_i)^2) \\ &= E((\frac{1}{n^2}(\sum_{i=1}^{n} X_i)^2 - \frac{2}{n^3}(\sum_{i=1}^{n} X_i)^3 + \frac{1}{n^4}(\sum_{i=1}^{n} X_i)^4) \end{aligned}$$

et

$$\lim_{n\to+\infty} E([F(1-F)]^2) = p^2 - 2p^3 + p^4.$$

Maintenant

$$\begin{aligned} & E\left(\left[F(1-F) - pq(1-\frac{1}{n})\right]^2\right) \\ &= E\left([F(1-F)]^2 - 2p(1-p)(1-\frac{1}{n})F(1-F) + [p(1-p)(1-\frac{1}{n})]^2\right) \\ &= E\left([F(1-F)]^2\right) - \left[p(1-p)(1-\frac{1}{n})\right]^2 \end{aligned}$$

Ainsi

$$\lim_{n\to+\infty} E\left([F(1-F) - pq(1-1/n)]^2\right) = p^2 - 2p^3 + p^4 - [p(1-p)(1-1/n)]^2 = 0.$$

Donc $\frac{n}{n-1}F(1-F)$ converge en probabilité vers $p(1-p)$.

Table des matières

Durée : 12h de cours
12 h TD + 6h TP

TP n°1 : Tests de normalité
TP n°2 : Tests moyennes, pourcentages
TP n°3 : Test non-paramétriques.
TP n°4 : Sur le risque β