Troisième année de Licence MIASHS

Statistique mathématique 3[1]

Julien GREPAT[2]

Contents

[2]Contact : julien.grepat@univ-grenoble-alpes.fr

Part I

Introduction aux tests statistiques

Au croisement des statistiques et des probabilités, la démarche des statistiques inférentielles est de considérer l'observation $(x_1, \cdots, x_d)$ comme la réalisation de d répétitions $(X_1, \cdots, X_d)$ d'une expérience aléatoire X. La réalité est souvent entachée d'erreur. Cette démarche permet d'en tenir compte, d'imaginer, d'inférer le problème pur, et de travailler pour mieux comprendre l'erreur.

1 Le test statistique

Posons θ un paramètre lié à la problématique, par exemple, une moyenne, un écart-type, une quantification de l'indépendance ou un coefficient de corrélation linéaire...

L'objectif d'un test d'hypothèse paramétrique est une aide à la décision à propos de la question que l'on forme de la manière suivante :

Est-ce qu'au vu de l'observation d'un échantillon, on peut décider entre les deux possibilités

- $\mathcal{H}_0 : \theta \in \Theta_0$,
- $\mathcal{H}_1 : \theta \in \Theta_1$?

Ici $\Theta_0, \Theta_1 \subset \mathbb{R}$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

1.1 Erreur

Il est d'usage de souhaiter statuer sur la vraisemblance de $\mathcal{H}_0$. On a affaire à deux types d'erreur :

	$\mathcal{H}_0$ est vraie	$\mathcal{H}_1$ est vraie
Accepter $\mathcal{H}_0$	OK	Erreur de deuxième espèce
Rejeter $\mathcal{H}_0$	Erreur de première espèce	Puissance du test

On notera α la probabilité d'erreur de première espèce, appelée parfois seuil. On prendra en général pour α les valeurs 0.01 ou 0.05. Il est à noter que mécaniquement, si l'erreur de première espèce baisse, l'erreur de seconde espèce augmente.

1.2 La variable du test (la statistique du test)

Il s'agit d'un estimateur $\theta_n = f(X_1, \cdots, X_n)$ de θ. C'est donc une variable aléatoire dont la loi est connue sous l'hypotèse $\mathcal{H}_0$.

On observera la valeur de θ_n, que l'on notera $\hat{\theta}_n$, sous l'observation $(X_1 = x_1, \cdots X_n = x_n)$, que l'on calculera

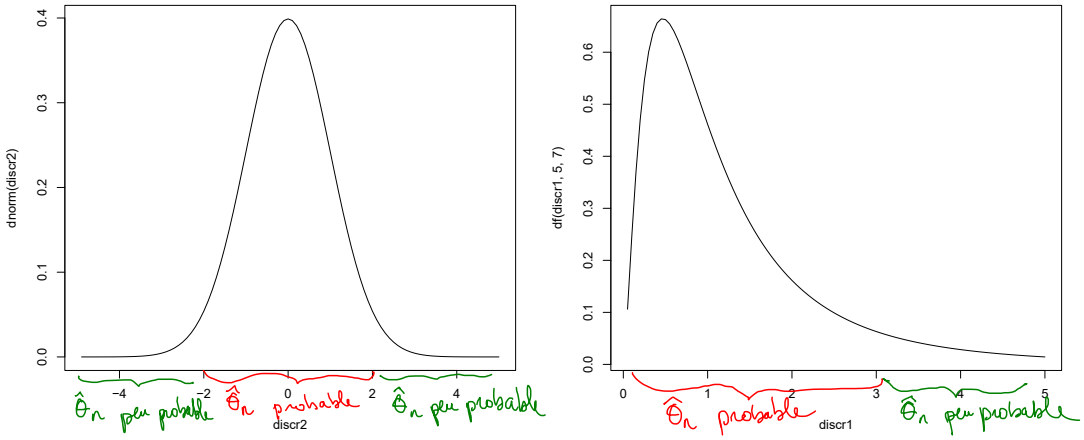
$$\hat{\theta}_n = f(x_1, \cdots, x_n).$$

On notera que

$$\hat{\theta}_n = E\big[\theta_n | X_1 = x_1, \cdots X_n = x_n\big].$$

1.3 Zone de rejet

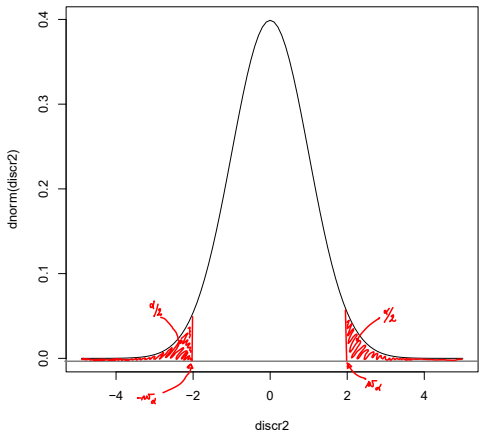
Il est indispensable de connaître la densité de la loi de notre statistique du test pour savoir où il serait le plus probable de voir les valeurs de θ observées. Considérons les distributions suivantes.



De note hypothèse nulle $\mathcal{H}_0$ on peut déduire une zone de rejet W_α, pour un seuil α. C'est la zone où il n'est pas probable de trouver l'estimation $\hat{\theta}$ sous $\mathcal{H}_0$. Pour un test *bilatéral* :

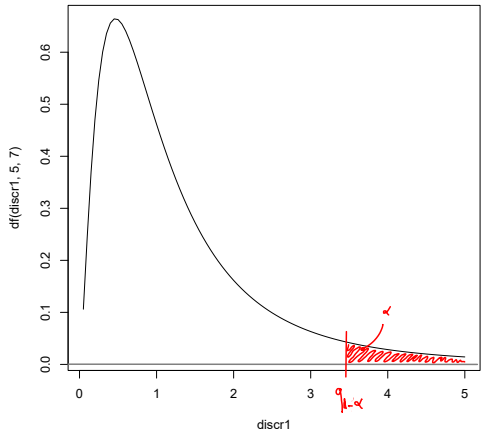
$$W_\alpha =]-\infty, -w_\alpha] \cup [w_\alpha, +\infty[,$$

où $w_\alpha = q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$.



Pour un test *unilatéral* :

$$W_\alpha = [q_\alpha, +\infty[.$$



Dans cette zone, l'hypothèse $\mathcal{H}_0$ est probablement contredite, avec un risque de se tromper α. On rejette $\mathcal{H}_0$ au seuil α.

1.4 Décision

La décision est donc binaire. Si $\hat{\theta} \in W_\alpha$, on rejette $\mathcal{H}_0$ est on prendra pour vraie $\mathcal{H}_1$ avec un risque de se tromper α non négligeable. Si $\hat{\theta} \notin W$, on rejette $\mathcal{H}_1$ est on prendra pour vraie $\mathcal{H}_0$ en espérant que le test est puissant.

1.5 $p-$value

Un paragraphe précédent fait apparaître le seuil comme la probabilité α, fixée a priori, que le test rejette l'hypothèse $\mathcal{H}_0$ à tort.

$$\mathbb{P}_{\mathcal{H}_0}[\text{ Rejet de } \mathcal{H}_0] = \alpha \ .$$

Une fois les données recueillies, la valeur prise par la statistique de test sera calculée, et la réponse sera binaire : rejet ou non de $\mathcal{H}_0$. On préfère souvent garder l'information contenue dans la valeur de la statistique de test, en retournant le seuil limite auquel $\mathcal{H}_0$ aurait été rejetée, compte tenu de l'observation.

Prenons l'exemple (fréquent) d'une hypothèse $\mathcal{H}_0$ sous laquelle la statistique de test T suit la loi normale $\mathcal{N}(0,1)$. La règle de rejet pour le test bilatéral de seuil 0.05 est :

$$\text{Rejet de } \mathcal{H}_0 \iff T \notin [-1.96\,,\, +1.96] \ .$$

Supposons que la valeur prise par T soit 2.72. L'hypothèse $\mathcal{H}_0$ sera donc rejetée. Mais elle serait également rejetée au seuil 0.01. En fait elle serait rejetée pour n'importe quel seuil supérieur à 0.00653, ce qui est un renseignement plus précis qu'une simple réponse binaire.

Definition 1.1 *Soit $\mathcal{H}_0$ l'hypothèse nulle, T la statistique de test et F_0 sa fonction de répartition sous l'hypothèse $\mathcal{H}_0$. On suppose que F_0 est continue.*

(i) Pour un test bilatéral (rejet des valeurs trop écartées) la p-valeur d'une valeur t prise par T est :

$$p(t) = \begin{cases} 2F_0(t) & si \quad F_0(t) < 0.5 \\ 2(1-F_0(t)) & si \quad F_0(t) \geq 0.5 \ . \end{cases}$$

(ii) Pour un test unilatéral à droite (rejet des valeurs trop grandes) la p-valeur d'une valeur t prise par T est :

$$p(t) = 1 - F_0(t) \ .$$

(iii) Pour un test unilatéral à gauche (rejet des valeurs trop petites) la p-valeur d'une valeur t prise par T est :

$$p(t) = F_0(t) \ .$$

Cependant calculer une p-valeur pour un test bilatéral est assez artificiel. Au vu de la valeur prise par T, on aura tendance à effectuer plutôt un test unilatéral visant à décider si la valeur observée est trop grande ou trop petite. Pour une statistique de test suivant la loi $\mathcal{N}(0,1)$, la valeur 2.72 est clairement à droite de la distribution. Le problème ne se pose plus de savoir si elle est trop petite, mais plutôt si elle est significativement trop grande. En pratique, pour une statistique de test de fonction de répartition F_0 sous $\mathcal{H}_0$, on définira souvent la p-valeur de la valeur t par :

$$p(t) = \min\{F_0(t)\ ,\ 1-F_0(t)\} \ .$$

La connaissance de la p-valeur rend inutile le calcul préalable de la région de rejet : si $p(t)$ est la p-valeur d'une observation t sous l'hypothèse $\mathcal{H}_0$, on obtient un test de seuil α par la règle de rejet :

$$\text{Rejet de } \mathcal{H}_0 \iff p(T) < \alpha \ .$$

En pratique. On retiendra qu'on rejettera l'hypothèse $\mathcal{H}_0$ si la p-value est inférieure au seuil (5/100 par exemple).

1.6 Discussion

On peut faire plusieurs remarques sur ce que l'on ressent à propos du test :

- En général, plus l'échantillon est petit, plus le test est laxiste ou inversement.
- La démarche semble plus robuste si l'on cherche à rejeter l'hypothèse $\mathcal{H}_0$ par analogie avec un raisonnement par l'absurde. (Test au sens de Fisher).
- On est tributaire du hasard dans le tirage de l'échantillon. En effet, les tests ont été créés dans le contexte de la qualité d'une production, où l'échantillonnage est reproductible et où on acceptera de jeter une production à tort si on peut sauver d'avantage de séries produites.

Pour toutes ces raisons, les tests ne restent qu'une aide à la décision, même s'ils sont largement utilisés dans tous les domaines scientifiques (sciences humaines, technologiques, économiques...)

2 Rappels : variables aléatoires élémentaires

On rappelle que l'on peut définir les notions d'espérance de variable aléatoire comme la moyenne statistique ainsi que celles d'écart-type et de variance de manière analogue aux notions vues dans le cadre des statistiques descriptives. On rappelle les propriétés suivantes.

- La linéarité de l'espérance $E[aX + bY + c] = aEX + bEY + c$, avec X, Y des variables aléatoires et a, b, c des réels.
- La formule fondamentale : $Var(aX + b) = a^2 VarX$.

- Si X et Y sont indépendantes, alors $Var(X+Y) = Var(X) + Var(Y)$.

On rappelle qu'on définit la fonction de répartition d'une variable aléatoire par

$$F_X(t) = P(X \leq t).$$

2.1 Loi de Bernouilli

Soit une expérience à deux issues complémentaires : succès ou réussite. Par exemple, on pourra considérer le succès comme *face* au lancer de pièce. La variable X_0 prendra la valeur 1 en cas de succès, et la valeur 0 sinon. La probabilité de succès est notée p, X_0 suit une loi $\mathcal{B}(p)$:

$$P(X_0 = 1) = p, \qquad P(X_0 = 0) = 1 - p.$$

Cette loi a pour valeurs caractéristiques

$$EX_0 = p, \qquad Var\ X_0 = p(1-p).$$

2.2 Loi binomiale

Soit X la loi qui compte le nombre de succès à n expériences aléatoires identiques et indépendantes à deux issues possibles :

- le succès (avec probabilité p),
- l'échec (avec probabilité $1-p$).

La variable X suit une loi binômiale $\mathcal{B}(n,p)$ avec

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Cette loi peut être vue comme la somme de n variables de loi $\mathcal{B}(p)$ indépendantes. Ainsi, cette loi a pour valeurs caractéristiques

$$EX = np, \qquad Var\ X = np(1-p).$$

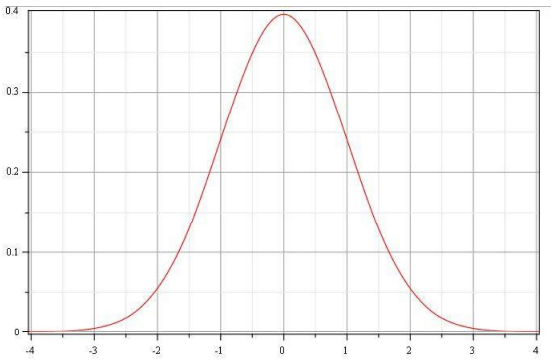
2.3 Loi normale

On parle de loi normale (ou Gaussienne) lorsqu'on a affaire à une variable continue qui dépend d'un grand nombre de causes indépendantes, dont les effets s'additionnent et dont aucune n'est prépondérante. C'est le cas de la plupart des phénomènes naturels, industriels, humain. Ce type de loi admet une densité en forme de cloche.

2.3.1 Loi normale centrée réduite

La loi normale centrée réduite, notée $\mathcal{N}(0,1)$ est la loi de référence. Elle a pour densité

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



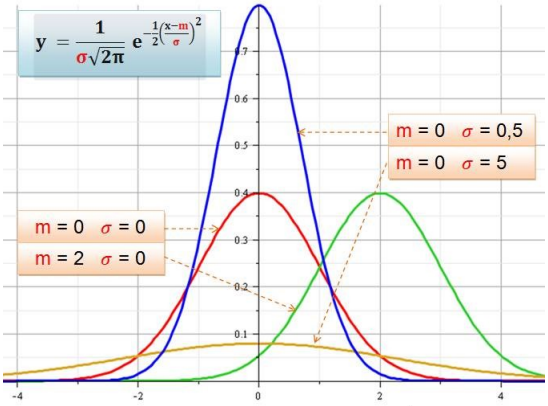
Son espérance vaut 0 et son écart-type 1.

2.3.2 Loi normale d'espérance m, d'écart-type σ

Une telle loi, notée $\mathcal{N}(m, \sigma)$ ou plus rarement $\mathcal{N}(m, \sigma^2)$ est obtenue par translation et dilatation de la loi normale centrée réduite :

$$\mathcal{N}(m, \sigma) = \sigma\mathcal{N}(0, 1) + m.$$

Une loi normale est donc entièrement déterminée par son espérance m et son écart-type σ. On observera les densités suivantes.



Theorem 2.1 *Soit $X_1, \cdots, X_N$ des lois normales (non nécessairement identiques) indépendantes. Toute combinaison linéaire des $X_1, \cdots X_N$ suit une loi normale.*

On utilisera les formules usuelles de linéarité de l'espérance, ainsi que les formules usuelles sur la variance pour déterminer les paramètre de la loi résultante.

2.3.3 Exemple fondamental

On considère une expérience aléatoire normale de paramètres m et σ répétée N fois de manière indépendante. Il s'agit des hypothèses standard de l'échantillonnage. C'est, par exemple, le cas du relevé des diamètres de N billes produites par une même machine.

Si on veut estimer le diamètre moyen de ces N billes, on somme les valeurs aléatoires qu'on divise par Nr. C'est l'estimateur de la moyenne :

$$\bar{X} = \frac{X_1 + \cdots + X_N}{N} = \frac{1}{N}X_1 + \cdots + \frac{1}{N}X_N.$$

Notons que $\bar{X}$ est une variable aléatoire, et que c'est une variable normale de paramètres $m_{\bar{X}}$ et $\sigma_{\bar{X}}$ en tant que combinaison linéaire de lois normales indépendantes. Il nous reste à connaître ses paramètres...

$$m_{\bar{X}} = E\bar{X} = E\left[\frac{1}{N}X_1 + \cdots + \frac{1}{N}X_N\right] = \frac{1}{N}EX_1 + \cdots + \frac{1}{N}EX_N = \frac{N}{N}m = m,$$

par linéarité de l'espérance. Rappelons la formule $Var(aX+b) = a^2 VarX$. On a

$$\sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \sqrt{Var\left(\frac{X_1 + \cdots + X_N}{N}\right)} = \sqrt{\frac{1}{N^2}Var(X_1 + \cdots + X_N)}$$
$$= \sqrt{\frac{Var(X_1) + \cdots + Var(X_N)}{N^2}} = \sqrt{\frac{N\sigma^2}{N^2}} = \frac{\sigma}{\sqrt{N}},$$

car les $X_1, \cdots X_N$ sont indépendantes.

2.3.4 Intervalles remarquables

On aura pour ordre d'idée les intervalles remarquables :

$$\begin{aligned} P(X \in [m-\sigma, m+\sigma]) &\approx 0.68, \\ P(X \in [m-1,96\sigma, m+1,96\sigma]) &= 0.95 \approx P(X \in [m-2\sigma, m+2\sigma]), \\ P(X \in [m-3\sigma, m+3\sigma]) &\approx 0.997. \end{aligned}$$

3 Lois des statistiques

On propose, dans cette section, une sélection de lois très importantes en statistique, définies à partir de sommes de lois normales centrées réduites.

3.1 Loi du χ^2

Si $X_1; \cdots ; X_n$ sont des variables aléatoires indépendantes de même loi $\mathcal{N}(0;1)$, alors

$$X_1^2 + \cdots + X_n^2$$

suit une loi du χ^2 à n degrés de libertés, notée $\chi^2(n)$.

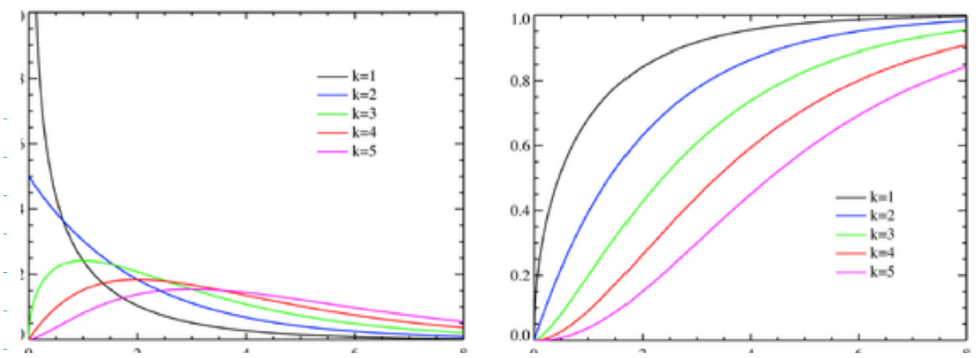
On a alors

$$EX = n, \qquad VarX = 2n.$$

Pour information, la densité de probabilité de la loi $\chi^2(n)$ est donnée par

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-t/2}, \qquad \forall x \geq 0,$$

où Γ est la fonction gamma d'Euler. Les densités et fonctions de répartition de certaines lois sont tracées sur les graphes suivants :

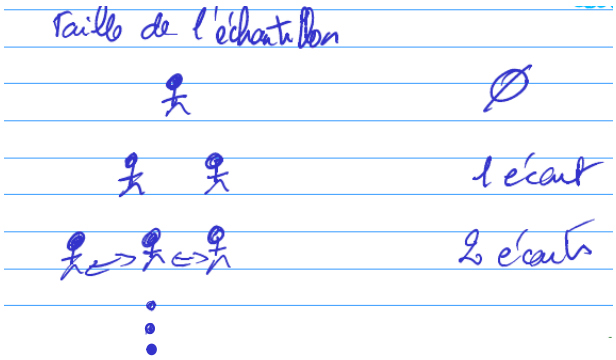


Proposition 3.1 *Si $X_1 \sim \chi^2(n)$ et $X_2 \sim \chi^2(m)$ sont deux variables aléatoires indépendantes, alors $X_1 + X_2 \sim \chi^2(n+m)$.*

Exemple fondamental

On considère une expérience aléatoire normale de paramètres μ et σ répétée n fois de manière indépendante. Il s'agit des hypothèses standard de l'échantillonnage.

On veut estimer la variance du diamètre de ces n billes, on somme les écarts au carré des valeurs aléatoires à la moyenne connue qu'on divise par $n-1$. C'est l'estimateur de la variance. Notons qu'on préfère diviser par $n-1$ pour avoir un estimateur sans biais. L'explication intuitive est qu'il est nécessaire de diviser la somme des écarts au carré par leur nombre dans un échantillon, et ici, ce n'est pas si simple. En effet



il y a $n-1$ écarts dans une population de taille n. Il suit qu'on estimera la variance par l'estimateur

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2.$$

On peut donner la loi de

$$\frac{\hat{S}_n^2}{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2.$$

Notons que si $X \sim \mathcal{N}(\mu, \sigma)$, alors

$$\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

parmi lesquelles $(N-1)$ sont indépendantes (en effet, en connaissant μ et les valeurs de $X_1, \cdots, X_{N-1}$, on peut déduire la valeur de X_N). On en déduit que

$$\frac{\hat{S}_n^2}{\sigma^2} \sim \chi^2(n-1).$$

3.2 Loi de Student

Soient U et V deux variables aléatoires indépendantes, telles que $U \sim \mathcal{N}(0;1)$ et $V \sim \chi^2(k)$. La variable

$$X = \frac{U}{\sqrt{V/k}}$$

est une loi de Student à k degrés de libertés, notée $t(k)$.

On a alors

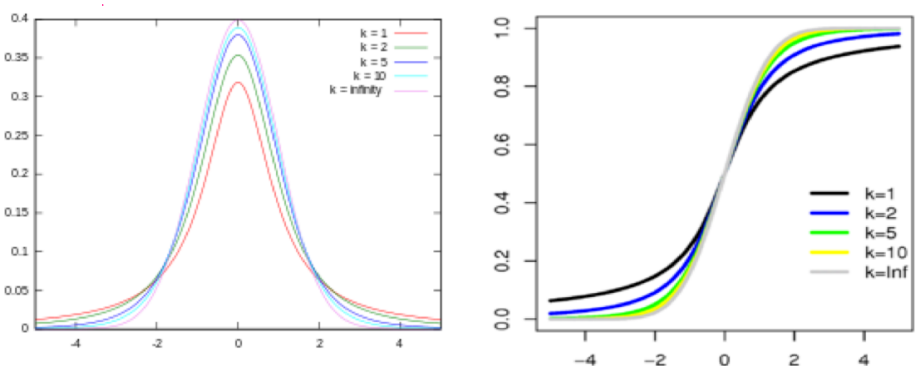
$$EX = 0, \qquad VarX = \frac{k}{k-2},.$$

pour $k > 2$.

Pour information, la densité de probabilité de la loi $t(k)$ est donnée par

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

où Γ est la fonction gamma d'Euler. Les densités et fonctions de répartition de certaines lois sont tracées sur les graphes suivants :



3.3 Loi de Fisher

Soient U et V deux deux variables aléatoires indépendantes, telles que $U \sim \chi^2(d_1)$ et $V \sim \chi^2(d_2)$. La variable

$$F = \frac{U/d_1}{V/d_2}$$

est suit une loi de Fisher–Snedecor de paramètres d_1 et d_2, notée $\mathcal{F}(d_1, d_2)$ ou F_{d_1,d_2}.

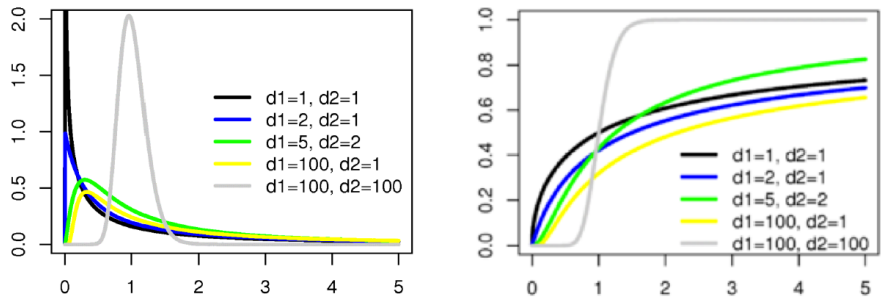
On a alors

$$EX = \frac{d_2}{d_2 - 1}, \quad d_2 > 2; \qquad VarX = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)2(d_2 - 4)}, \quad d_2 > 4.$$

Pour information, la densité de probabilité de la loi $\mathcal{F}_{d_1,d_2}$ est donnée par

$$f(x) = \frac{\left(\frac{d_1 x}{d_1 x + d_2}\right)^{(d_1/2)} \left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{(d_2/2)}}{x B(d_1/2, d_2/2)},$$

pour $x > 0$, où B est la fonction Beta d'Euler. Les densités et fonctions de répartition de certaines lois sont tracées sur les graphes suivants :



Proposition 3.2

- *Si* $X \sim \mathcal{F}_{d_1,d_2}$ *alors* $1/X \sim \mathcal{F}_{d_2,d_1}$.
- *Si* $X \sim t_k$ *alors* $X^2 \sim \mathcal{F}_{1,k}$.
- *Si* $X \sim \mathcal{N}(0;1)$ *alors* $X^2 \sim \mathcal{F}_{1,\infty}$.

4 Théorème limites

Il y a plusieurs types de convergences pour les suites de variables aléatoires. On rappelle la définition de la convergence en loi (suffisante pour tout ce dont nous auront besoin).

4.1 Convergence en loi

Soit (X_n) une suite de variables aléatoires. On dit que (X_n) converge en loi vers X si

$$\lim_{x\to+\infty} F_{X_n}(x) = F_X(x), \qquad \forall x \in \mathbb{R}.$$

4.2 Loi des grands nombres

Soit (X_n) une suite de variables aléatoires identiques et indépendantes et qui admettent la même espérance μ et le même écart-type. Alors:

$$\bar{X}_n = \frac{1}{n}\sum_{i=0}^{n} X_i \to \mu.$$

On connaît les estimateurs usuels suivants :

- estimateur de proportion p : avec $X_1, \cdots, X_n \sim \mathcal{B}(p)$ *i.i.d*,

$$\hat{p}_n = \sum x_i/n,$$

- estimateur de la moyenne μ :

$$\hat{\mu}_n = \sum x_i/n$$

- estimateur de la variance... À suivre...

Notons que, par la loi des grands nombres, on a les convergences en loi suivantes

$$\hat{p}_n \to p, \qquad \hat{\mu}_n \to \mu.$$

On dit que ce sont des estimateurs convergents.

4.3 Le théorème central limit (TCL)

Le théorème central limite établit la convergence en loi de la somme d'une suite de variables aléatoires vers la loi normale. Intuitivement, ce résultat affirme que toute somme de variables aléatoires indépendantes et identiquement distribuées (**i.i.d.** dans la suite) tend vers une variable aléatoire gaussienne.

Theorem 4.1 *Soit (X_n) une suite de variables aléatoires i.i.d.. Supposons que pour tout $n \in \mathbb{N}$, les variables X_i admettent une espérance μ et un écart-type σ. Alors la suite de variables aléatoires centrées réduites suivante converge en loi :*

$$\frac{\sum_{i=0}^{n} X_n - n\mu}{\sqrt{n}\sigma} \to \mathcal{N}(0,1).$$

4.4 Approximation d'une loi binomiale par une loi normale

Theorem 4.2 *Soit X une loi binômiale $\mathcal{B}(n,p)$. Si n est grand, on peut approximer X par une loi normale dont l'espérance est celle de X, et d'écart-type celui de X.*

En pratique, on aura les critères $n \geq 50$ et :

- $0,4 \leq p \leq 0,6$;
- ou $npq \geq 18$;
- ou $np > 5$ et $nq > 5$.

Soit Y la loi normale approximante. On a alors

$$P(a \leq X \leq b) \approx P(a \leq Y \leq b).$$

Si on souhaite faire l'approximation de $P(X = k)$, Y étant continue, on effectuera la correction de continuité

$$P(X = k) \approx P(k - 0,5 \leq Y \leq k + 0,5).$$

Exemple fondamental

Reprenons l'estimateur de proportion p : avec $X_1, \cdots, X_n \sim \mathcal{B}(p)$ *i.i.d*,

$$\hat{p}_n = \sum x_i/n \sim \mathcal{B}(n,p)/n,$$

d'espérance $np/n = p$, et de variance

$$\frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Supposons n grand, p petit, $\hat{p}_n$ s'approche par une loi normale $\mathcal{N}(p, \sqrt{p(1-p)}/n)$. En prenant en compte l'idée que p est petit, et que donc, $(1-p) \approx 1$, on peut en déduire la convergence en loi

$$\sqrt{n}\frac{p_n - p}{\sqrt{p}} \to \mathcal{N}(0,1).$$

Cet exemple nous sera très utile dans les tests du χ^2.

4.5 Autres théorèmes limites

On a les limites suivantes.

Proposition 4.3

- *Si* $X \sim \mathcal{F}_{d_1;d_2}$ *alors*
$$lim_{d2\to\infty} d_1 X \to \chi^2(d_1).$$
- *La loi de Student converge vers la loi normale centrée réduite :*
$$t_k \to \mathcal{N}(0;1).$$

Part II

Tests du χ^2

5 Test du χ^2 d'indépendance

Dans cette section, on s'intéresse aux relations entre deux variables notées X et Y. Supposons que l'on observe ces deux variables sur n unités statistiques. A chaque individu i, on peut associer un couple d'observations $(x_i; y_i)$. Chaque variable peut-être quantitative ou qualitative. Nous proposons ici un test d'indépendance.

5.1 Distributions conjointe et marginales

Notons $m_1^X, \cdots, m_J^X$ les J modalités de X et $m_1^Y, \cdots, m_K^Y$ les K modalités de Y. Si l'une des deux variables (ou les deux) est quantitative continue, les m_j^X ou les m_k^Y sont des classes modales. Introduisons les quantités suivantes :

- n_{jk} est le nombre de fois où le couple (X, Y) prend la modalité (m_j^X, m_k^Y),
- $n_{\bullet k}$ est le nombre de fois où la variable Y prend la valeur m_k^Y,
- $n_{j\bullet}$ est le nombre de fois où la variable X prend la valeur m_j^X.

On a

$$\sum_{j=1}^{J} n_{jk} = n_{\bullet k} \quad \text{et} \quad \sum_{k=1}^{K} n_{jk} = n_{j\bullet}$$

$$\sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk} = \sum_{j=1}^{J} n_{j\bullet} = \sum_{k=1}^{K} n_{\bullet k} = n$$

Les données peuvent être représentées dans un tableau à double entrée appelé **Tableau de contingence**.

	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
m_1^X	n_{11}	$\dots$	n_{1k}	$\dots$	n_{1K}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_j^X	n_{j1}	$\dots$	n_{jk}	$\dots$	n_{jK}	$n_{j\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_J^X	n_{J1}	$\dots$	n_{Jk}	$\dots$	n_{JK}	$n_{J\bullet}$
	$n_{\bullet 1}$	$\dots$	$n_{\bullet k}$	$\dots$	$n_{\bullet K}$	n

Le **tableau des fréquences** s'obtient en divisant les effectifs par le nombre d'unités statistiques n (effectif total). Comme précédemment on obtient

$$f_{jk} = \frac{n_{jk}}{n}, \quad f_{\bullet k} = \frac{n_{\bullet k}}{n} \quad f_{j\bullet} = \frac{n_{j\bullet}}{n}$$

	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
m_1^X	f_{11}	$\dots$	f_{1k}	$\dots$	f_{1K}	$f_{1\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_j^X	f_{j1}	$\dots$	f_{jk}	$\dots$	f_{jK}	$f_{j\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_J^X	f_{J1}	$\dots$	f_{Jk}	$\dots$	f_{JK}	$f_{J\bullet}$
	$f_{\bullet 1}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet K}$	1

5.2 Distributions marginales

A partir du tableau de contingence, on peut retrouver la distribution de chacune des variables séparément :

Modalité de Y	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
Fréquence empirique	$f_{\bullet 1}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet K}$	1

Modalité de X	m_1^X	$\dots$	m_j^X	$\dots$	m_J^X	total
Fréquence empirique	$f_{1\bullet}$	$\dots$	$f_{j\bullet}$	$\dots$	$f_{J\bullet}$	1

Les distributions de X et de Y sont appelées distributions marginales. Sur chaque variable, on peut calculer les indicateurs habituels (moyenne, variance, écart type si la variable est quantitative...). Ces paramètres sont qualifiés d'indicateurs marginaux.

5.3 Statistique du χ^2

En présence de deux variables, l'un des enjeux principaux est d'étudier (c'est à dire quantifier voire expliquer) la dépendance entre les deux caractères.

Si on était dans le cadre des probabilités, ce qui n'est pas le cas, alors deux caractères sont indépendants si la valeur de l'un n'a aucune influence sur la distribution de l'autre. Si tel était le cas, alors les distributions conditionnelles

$$f_{j|k} = \frac{f_{jk}}{f_{\bullet k}} \qquad \text{et} \qquad f_{k|j} = \frac{f_{jk}}{f_{j\bullet}}$$

seraient toutes semblables à la distribution marginale. Pour tout (j,k), on devrait avoir

$$f_{j|k} = f_{j\bullet} \qquad \text{et} \qquad f_{k|j} = f_{\bullet k}.$$

Ainsi, on aurait :

$$f_{kj} = f_{j|k} f_{\bullet k} = f_{j\bullet} f_{\bullet k}.$$

D'où, si les deux variables étaient indépendantes, on aurait

$$n_{jk} = \frac{n_{j\bullet} n_{\bullet k}}{n}.$$

En statistiques, on ne peut que "quantifier la distance à l'indépendance" par la statistique du χ^2,

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}}.$$

On peut remarquer que

$$D_{\chi^2} = n\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{n_{jk}^2}{n_{j\bullet}n_{\bullet k}} - 1\right),$$

ou de façon équivalente

$$D_{\chi^2} = \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{jk} - \frac{n_{j\bullet}n_{\bullet k}}{n}\right)^2}{\frac{n_{j\bullet}n_{\bullet k}}{n}},$$

où J et K sont le nombre de modalités de chacune des deux variables considérées.
Le cas d'indépendance probabiliste serait alors équivalent à $D_{\chi^2} = 0$.

5.4 Test du χ^2

Posons l'hypothèse nulle

$$\mathcal{H}_0 = \{\text{Les deux variables sont indépendantes}\}.$$

Theorem 5.1 *La variable du test D_{χ^2} suit une loi du χ^2 à $(K-1)(L-1)$ degrés de liberté sous l'hypothèse $\mathcal{H}_0$.*

Éléments de preuve On rappelle le théorème central limit dans le cas d'une fréquence. Si $T \backsim \mathcal{B}(n,p)$ alors, pour n grand,

$$\frac{T-np}{\sqrt{np(1-p)}} \approx Y \backsim N(0,1).$$

De plus, si p est petit, $(1-p)$ est proche de 1, et on factorise par n

$$\frac{n(T/n-p)}{\sqrt{np}} = \sqrt{n}\frac{(T/n-p)}{\sqrt{p}} \approx Y \backsim N(0,1).$$

On reconnaît $F = T/n$ comme une fréquence d'espérance p. On alors un résultat asymptotique simple :

$$\sqrt{n}\frac{(F-p)}{\sqrt{p}} \approx Y \backsim N(0,1),$$

pour n grand.
On pose $F = f_{jk}$ dont l'espérance p est $f_{j\bullet}f_{\bullet k}$ sous l'hypothèse $\mathcal{H}_0$. On en déduit que les quotients

$$\sqrt{n}\frac{f_{jk} - f_{j\bullet}f_{\bullet k}}{\sqrt{f_{j\bullet}f_{\bullet k}}}$$

forment asymptotiquement des lois normales centrées réduites. La somme de leur carrés est une loi du χ^2. Il reste à discuter du nombre de degrés de libertés. Il suffit de compter le nombre d'aléas (variables gaussiennes indépendantes). Puisqu'on suppose qu'il n'y a pas d'aléa sur les lois marginales et que pour chaque ligne, la somme des féquences jointes est égale à la fréquence marginale, on en déduit qu'une ligne contient $(K-1)$ aléas. De même, une colonne contient $(L-1)$ aléas. Il vient que la loi du χ^2 a $(K-1)(L-1)$ degrés de liberté.
□

5.5 interprétation

Au seuil $\alpha\%$ (le plus souvent $\alpha = 5$), il faut comparer D_{χ^2} au quantile d'ordre $1-\alpha\%$ à savoir $q_{1-\alpha}$ ($q_{0,95}$ le plus souvent) d'une loi du χ^2_d, où

$$d = (J-1)(K-1)$$

est le degré de liberté de la loi (c'est à dire le paramètre de la loi du χ^2).

L'interprétation est la suivante :

- si $D_{\chi^2} \geq q_{1-\alpha}$, on conclut que les deux variables sont dépendantes,
- sinon, on conclut qu'elles sont indépendantes.

On rejettera l'hypothèse d'indépendance si $p \leq \alpha/100$ (le plus souvent si $p \leq 0,05$.)

En pratique, on évite d'utiliser le test du χ^2 si un effectif du tableau est inférieur ou égal à 5 car l'approximation par le Théorème Central Limit est alors trop grossière.

5.6 Exemple

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

(i) On Donne le tableau des fréquences.
Pour calculer les fréquences, on divise chaque effectif par l'effectif total (ici 200) :

	Célibataire	En couple	Total
Dossier accepté	0.17	0.29	0.46
Dossier refusé	0.33	0.21	0.54
Total	0.5	0.5	1

(ii) On Calcule la statistique du Chi-deux.
La statistique du Chi-deux est donnée par :

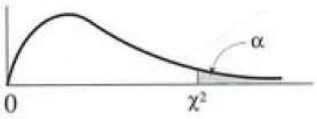
$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{\bullet j} f_{k\bullet})^2}{f_{\bullet j} f_{k\bullet}}$$

Ici on a donc :

$$\begin{aligned} D_{\chi^2} =& 200\left(\frac{(0.17 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.29 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.33 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \right. \\ & \left. + \frac{(0.21 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \right) \\ =& 200\Big(0.016 + 0.016 + 0.013 + 0.013 \Big) \\ =& 11.6 \end{aligned}$$

(iii) On fait le test du Chi-deux pour conclure.
On compare cette statistique à la valeur de la table du Chi-deux à 1 degré de liberté (2 modalités pour chaque variable).

Table χ^2 : points de pourcentage supérieurs de la distribution χ^2



dl	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	3.56	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15

On trouve 3.84.
On a donc $D_{\chi^2} > q_{0.95}$ et on conclue que les variables sont dépendantes : la situation maritale influence l'acceptation ou le refus du dossier.

6 Test du χ^2 pour l'ajustement d'une série à une loi de probabilité

Lorsqu'une loi statistique a une distribution qui ressemble celle d'une loi de probabilité connue, on peut se poser la question de leur adéquation. Des méthodes descriptives (qq-plot, droite de Henry...à suivre...) peuvent permettre une première approche. Les statistiques inférentielles, et en particulier le test du χ^2, peut donner un outil de choix supplémentaire.

On considère la série suivante détaillant le poids de 500 sacs de ciment.

Poids (kg)	effectif
[0,45]	35
]45,47]	53
]47,49]	76
]49,51]	100
]51,53]	88
]53,55]	78
]55,57]	42
]57,∞]	28

On souhaite savoir si cette série peut être ajustée par une loi Normale (m, σ). On peut faire une estimation ponctuelle des paramètres m et σ sur la série,

$$m = \frac{44 \times 35 + \cdots + 46 \times 53 + \cdots 58 \times 28}{500} \approx 50,78, \qquad \sigma \approx 3,74.$$

On souhaite donc comparer la série observée à une loi normale $\mathcal{N}(50,78;3,74)$.

On note O_i les effectifs observés. On pose t_i les extrémités des classes. On complète le tableau avec les effectifs trouvés avec la loi gaussienne,

$$T_i = 500 \times (F(t_i) - F(t_{i-1})),$$

où F est la fonction de répartition de le loi normale $\mathcal{N}(50,78;3,74)$. On obtient

poids	O_i	T_i
[0,45]	35	30,5
]45,47]	53	47,5
]47,49]	76	80
]49,51]	100	104
]51,53]	88	99
]53,55]	78	74,5
]55,57]	42	40,5
]57,∞]	28	24

6.1 Hypothèse $\mathcal{H}_0$

On considère l'hypothèse

$$\mathcal{H}_0 = \text{“La série observée est distribuée selon une loi normale } \mathcal{N}(50,78;3,74)''.$$

6.2 Variable du test

On étudie la distance à l'adéquation des effectifs de la série observée à la série théorique

$$D_{\chi^2} = \sum_{i=1}^{l} \frac{(O_i - T_i)^2}{T_i}.$$

Theorem 6.1 *La variable du test D_{χ^2} suit une loi du χ^2 à $l-s-1$ degrés de liberté, où l est le nombre de modalités observées, s est le nombre de paramètres estimés (m,σ,...).*

Éléments de preuve On rappelle le théorème central limit dans le cas d'une fréquence. Si $O \backsim \mathcal{B}(n,p)$ alors, pour n grand,

$$\frac{O - np}{\sqrt{np(1-p)}} \approx Y \backsim N(0,1).$$

De plus, si p est petit, $(1-p)$ est proche de 1, et en notant que l'effectif moyen est $T = np$ sous $\mathcal{H}_0$,

$$\frac{(O - T)}{\sqrt{T}} \approx Y \backsim N(0,1).$$

On en conclut que D_{χ^2} suit bien une loi du χ^2. Il reste à discuter des degrés de liberté, i.e. du nombre d'aléas indépendants. En connaissant l'effectif total n, on peut déduire le dernier T_i des précédents puisque leur somme vaut n. Chaque paramètre des s connus pose une équation et permet de connaître un T_i de plus par déduction. Par exemple, la moyenne connue donne une équation de plus

$$m = \sum_{i=1}^{l} \frac{T_i \times n_i}{n}.$$

On connaît donc un autre T_i à partir des $l-2$ premiers T_i et ainsi de suite. □

Dans notre cas, la loi du χ^2 a 5 d.d.l., et $D_{\chi^2} = 3,76$.

6.3 Interprétation

Au seuil $\alpha\%$ (le plus souvent $\alpha = 5$), il faut comparer D_{χ^2} au quantile d'ordre $1 - \alpha\%$ à savoir $q_{1-\alpha}$ ($q_{0,95}$ le plus souvent) de la loi du χ^2_d

L'interprétation est la suivante :

- si $D_{\chi^2} \geq q_{1-\alpha}$, on conclut que les deux distributions ne peuvent pas être identiques,
- sinon, on ne rejette pas cette hypothèse.

On rejettera l'hypothèse d'adéquation si $p \leq \alpha/100$ (le plus souvent si $p \leq 0,05$.)

En pratique, dans ce cas également, on évite d'utiliser le test du χ^2 si un effectif du tableau est inférieur ou égal à 5 à cause de l'approximation avec le Théorème Central Limit.

Dans notre exemple, $q_{0,95} = 11,07$. Puisque $D_{\chi^2} \leq q_{0,95}$, on ne rejette pas l'adéquation des lois.

Part III

Les tests de normalité

7 Les différents tests

Les lois normales (ou de Gauss) sont des lois de probabilités aux propriétés remarquables. Nous aurons l'occasion, dans cette unité, de voir leur omniprésence dans les hypothèse (au sens mathématiques) des modèles statistiques.

On va donc se demander si la série est normalement distribuée. Pour valider ou infirmer ce caractère normal d'une distribution, on pourra utiliser un test de normalité.

7.1 Hypothèse nulle et puissance du test

Tous les tests de normalité classiques testent l'hypothèse nulle $\mathcal{H}_0 =$"la distribution est normale".

Pour rappel, le seuil d'un test, communément 5% est la probabilité de rejeter à tort $\mathcal{H}_0$. Bien souvent dans ces tests, nous souhaiterions pouvoir opter pour la normalité de la série. Pour un seuil donné, la probabilité de ne pas se tromper en choisissant l'hypothèse alternative est appelée puissance du test. La problémtique est donc de trouver un test puissant.

Il est à noter que, dans bien des cas, la gêne occasionnée par une série non normale, n'est pas ce qui se passe mal dans la partie centrale de la distribution, la cloche, mais le plus souvent dans les queues, les événements rares, ce qui rend fragile le découpage en classe d'une variable continue.

7.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribuées.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Rien ne s'oppose au fait que la série soit normale (Pour autant, rien ne l'assure non plus ! D'où la nécessité de la puissance élévée pour le test).

7.3 Les différents tests de normalité

Il existe différents tests de normalité. Certains testeront l'adéquation des fonctions de répartitions quand d'autres s'intéresseront à d'autres propriétés de la loi normale.

7.3.1 Le test du χ^2 d'adéquation à une loi normale

Le reproche fait à l'utilisation de ce test est la nécessité de regrouper les observations en classes. Ainsi, l'intervention de ce choix plus ou moins arbitraire lui fait perdre de la puissance.

7.3.2 Tests de Lilliefors et d'Anderson–Darling,

Ces test sont des mises en application dans le cadre normal du test de Komogorov–Smirnov qui, de manière générale s'intéresse à l'adéquation des fonctions de répartition de la loi statistique et de celle de la loi théorique d'intérêt. Ces tests ont donc pour statistique de test la distance verticale maximale entre les deux fonctions de répartition.

Le test de Lilliefors est l'application directe du test de Komogorov–Smirnov au cadre Gaussien. Ainsi, il capte peu les différences dans les queues, c'est-à-dire, l'occurrence d'événement rare.

Le test d'Anderson–Darling tente de capter ces événements rares en pénalisant la différence qu'ils engendrent sur les fonctions de répartition. Ce test semble assez puissant, néanmoins, il reste moins populaire que le test de Shapiro–Wilk (à venir).

7.4 Tests d'Agostino et de Jarque–Bera

Ces test s'intéressent à l'aplatissement des queues et à la symmétrie des distribution, les principaux attribus des lois normales. Leur statistiques de test pénalisent donc les écarts à l'aplatissement des queues, ainsi que les écarts à la symétrie des observations. Ces tests sont relativement puissants et, ce qui est notable, le restent pour des effectifs n grands. On leur reprochera de ne pas exhiber le défaut de distribution amenant à la non normalité. Ils restent moins populaires que le test de Shapiro–Wilk.

8 Le test de Shapiro–Wilk

Le test de Shapiro–Wilk reste le plus populaire des tests de normalité, par sa puissance, mais peut-être pour le fait qu'il s'appuie sur un outil graphique de normalité qui va nous permettre de comprendre ce qui rend notre variable non normale. Il s'agit du QQ-plot, graphe quantile-quantile.

8.1 Le QQ-plot

Soit $x_1, \cdots, x_n$ une série statistique. On peut chercher à savoir si la distribution des données est gaussienne ou Poisson etc... Notons F_0 la fonction de répartition de cette loi de probabilité d'intérêt.

Le QQ-plot est un outil graphique permettant de visualiser rapidement l'adéquation de la distribution d'une série numérique à une distribution de référence. Dans notre contexte, on considérera une loi normale dont les paramètres seront estimés sur la série statistique observée.

Dans ce graphe, on reporte sur l'axe des ordonnées les fractiles q_i correspondant à la distribution observée et sur l'axe des abscisses ceux correspondant à la distribution théorique q_i^*. On reporte dans un graphique le nuage de points $(q_i^*; q_i)_i$.

8.1.1 Données brutes

Il est primordial de classer dans l'ordre croissant les observations statistiques :

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)},$$

tri obtenu sur `R` avec la fonction `sort()`. Clairement, l'observation $x_{(i)}$ est le $i$ème quantile d'ordre $1/n$, plus précisément, $q_i = x_{(i)}$ est le quantile d'ordre i/n. Il reste alors à calculer la série de quantiles théoriques

$$q_i^* = F_0^{-1}\left(\frac{i}{n}\right),$$

qu'on obtient sur `R`, dans le cas d'une loi normale, par `qnorm(i/n,mu,sigma)`. Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.

Note : Il se peut que la statistique d'ordre $x_{(i)}$ soit considérée comme le quantile d'ordre corrigé :

- $i/(n+1)$ si la population est divisée en $n+1$ tranche,
- $(2i-1)/(2n)$ si on souhaite prendre le milieu de la tranche,
- $(i-0.375)/(n+0.25)$ par certains auteurs (Saporta, 2006, p. 361).

Cela ne devrait pas changer fondamentalement les choses.

8.1.2 Données discrètes ordonnées

Dans le cas d'une variable quantitative dont les valeurs sont regroupées par modalités. Soient $m_1, \cdots, m_J$ les modalitées de la série $x_1, \cdots, x_n$ que l'on appellera quantiles observés ($q_i = m_i$). On a alors

Modalités ordonnées, quantiles observés	$q_1 = m_1$	$\cdots$	$q_J = m_J$
Fréquences cumulées	F_1	$\cdots$	F_J
Quantiles théoriques	$q_1^* = F_0^{-1}(F_1)$	$\cdots$	$q_J^* = F_0^{-1}(F_J)$

Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.

8.1.3 Données continues regroupées par classes

Dans le cas d'une variable quantitative dont les valeurs sont regroupées en classes de modalité :

Classe	$[b_0, b_1[$	$[b_1, b_2[$	$\ldots$	$[b_{J-1}, b_J[$
quantiles observés(centre de la classe)	q_1	q_2	$\ldots$	q_J
Fréq. cumulées	F_1	F_2	$\ldots$	F_J
Quantiles théoriques	$q_1^* = F_0^{-1}(F_1)$	$q_2^* = F_0^{-1}(F_2)$	$\ldots$	$q_J^* = F_0^{-1}(F_J)$

Il reste à tracer le nuage de points $(q_i^*; q_i)_i$.

8.1.4 Interprétation

- Si les points sont alignées sur la diagonale du carré de côté 1 (première bissectrice), alors la loi théorique proposée (de fonction de répartition F_0) est adaptée aux observations.
- Si les points sont alignés sur une droite parallèle à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de position de la loi théorique.
- Si les points sont alignés sur une droite passant par l'origine mais inclinée par rapport à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de dispersion de la loi théorique.
- Si les points sont alignés sur une droite ne passant pas par l'origine et inclinée par rapport à la diagonale du carré de côté 1, on soupçonnera une erreur sur les paramètres de dispersion et de position de la loi théorique.
- Si les points ne sont pas alignés sur une droite, la loi théorique n'est pas adaptée aux observations.

8.2 Le test de Shapiro Wilk

Le test de Shapiro–Wilk va tester l'alignement des points du qq-plot par rapport à une loi normale.

On a vu que le qq-plot est une représentation descriptive qui, si les points sont alignés, nous permet d'affirmer que le répartition est bien normale. C'est en cela que le le test tient sa puissance, en particulier pour des petits effectifs (inférieurs à 50).

Sous l'hypothèse nulle $\mathcal{H}_0$ ="la série statistique est normalement distribuée", la statistique du test W est un coefficient de détermination corrigé[3]du qq-plot. Ainsi, $0 \leq W \leq 1$ et plus W est élevé, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

$$W < W_{crit}.$$

Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk :

[3]les quantiles sont corrélés

n \ α	0,05	0,01
3	0,767	0,753
4	0,748	0,687
5	0,762	0,686
6	0,788	0,713
7	0,803	0,730
8	0,818	0,749
9	0,829	0,764
10	0,842	0,781
11	0,850	0,792
12	0,859	0,805
13	0,856	0,814
14	0,874	0,825
15	0,881	0,835
16	0,837	0,844
17	0,892	0,851
18	0,897	0,858
19	0,901	0,863
20	0,905	0,868
21	0,908	0,873
22	0,911	0,878
23	0,914	0,881
24	0,916	0,884
25	0,918	0,888
26	0,920	0,891

n \ α	0,05	0,01
27	0,923	0,894
28	0,924	0,896
29	0,926	0,898
30	0,927	0,900
31	0,929	0,902
32	0,930	0,904
33	0,931	0,906
34	0,933	0,908
35	0,934	0,910
36	0,935	0,912
37	0,936	0,914
38	0,938	0,916
39	0,939	0,917
40	0,940	0,919
41	0,941	0,920
42	0,942	0,922
43	0,943	0,923
44	0,944	0,924
45	0,945	0,926
46	0,945	0,927
47	0,946	0,928
48	0,947	0,929
49	0,947	0,929
50	0,947	0,930

8.3 Exemple

On observe la richesse des régions françaises en 2019.

	PIB/Hab.
Auvergne Rhônes Alpes	31639
Bourgogne Franche Comté	26218
Bretagne	27838
Centre Val de Loire	27274
Corse	26954
Grand Est	27378
Hauts de France	26095
Ile de France	55227
Normandie	27465
Nouvelle Aquitaine	27657
Occitanie	27449
Pays de la Loire	29424
Provence Alpes Côte d'Azur	30864

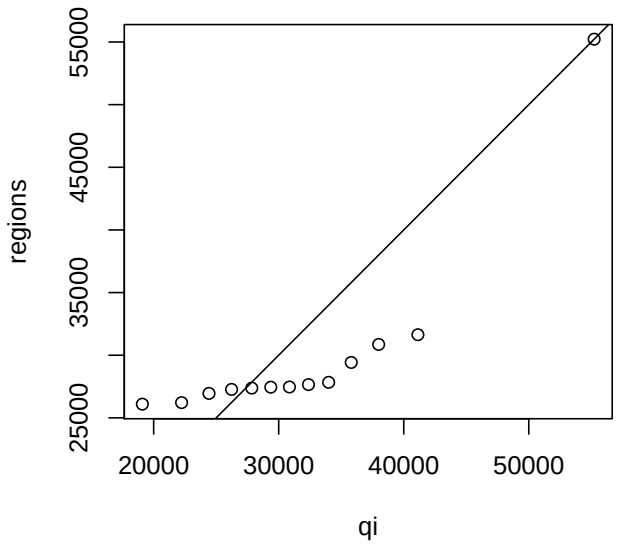
On souhaite observer l'ajustement à une loi normale. On trie les données dans l'ordre croissant, on donne les fréquences cumulées :

x	26095	26218	26954	27274	27378	27449	27465	27657	27838	29424	30864	31639	55227
$i/13$	0.0769	0.1538	0.2308	0.3077	0.3846	0.4615	0.5385	0.6154	0.6923	0.7692	0.8462	0.9231	1

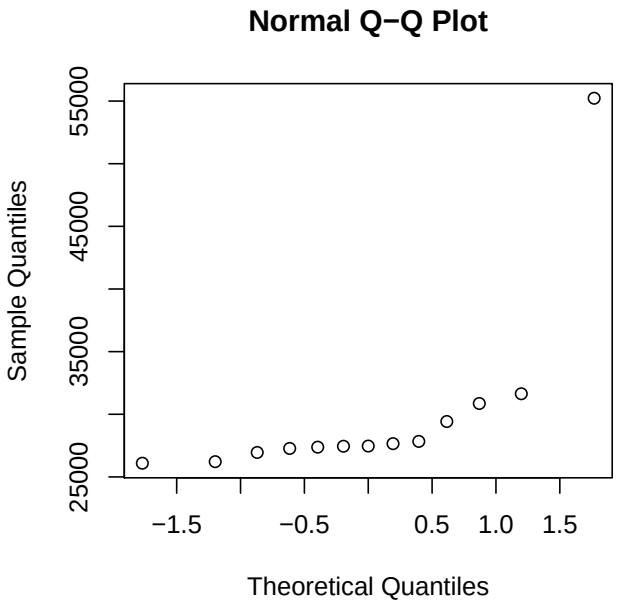
On calcule les quantiles théoriques d'après une loi normale $\mathcal{N}(30114; 7726)$ avec la fonction `qnorm(Fi, 30114,7726)`:

x	26095	26218	26954	27274	27378	27449	27465	27657	27838	29424	30864	31639	55227
$i/13$	0.0769	0.1538	0.2308	0.3077	0.3846	0.4615	0.5385	0.6154	0.6923	0.7692	0.8462	0.9231	1
q_i^*	19096	22232	24425	26232	27847	29367	30860	32380	33995	35802	37995	41131	Inf

Naturellement, le dernier quantile théorique est infini. On pourrait oublier le dernier point, ou diviser les fréquences cumulées par 14 au lieu de 13, ou, ce que nous allons faire, remplacer le dernier quantile théorique par la valeur observée 55227. On trace le nuage de points (x_i, q_i^*)



On s'aperçoit que le nuage n'est pas rectiligne. On peut supposer que les richesses ne sont pas normalement distribuées. On peut obtenir directement le qq-plot avec `R` par la commande `qqnorm(x)`. On obtient quelque-chose de similaire.



On conclut avec le test de Shapiro–Wilk.

```
        Shapiro-Wilk normality test

data:  regions
W = 0.4926, p-value = 8.91e-06
```

On rejette la normalité.

8.4 Mise en œuvre du test de Shapiro–Wilk sur R

Simulons des observations normales et appliquons le test de Shapiro–Wilk.

```
> x=rnorm(50,11,2)
> x
 [1] 10.214512  9.081032  6.620966  9.362181  8.986913 11.673092 13.729038
 [8]  9.677063 11.311007 12.167166 15.143072 10.128579  9.741033  9.510930
[15] 11.741705 11.539940  9.512374 13.226112 11.564665 13.886239  7.412722
[22] 11.397585 11.934305 12.439496  7.633424 11.339367 11.585285  5.934555
[29]  9.705902 10.930072  7.706895 11.698624 13.606455 12.948415 12.573321
[36] 10.926673  7.010748  9.134820 10.628741 10.815647  8.714866  7.589055
[43] 12.306357  9.384695 10.401019 13.733044 11.899843  9.970912 14.845173
[50] 14.696913
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.98553, p-value = 0.7942
```

La $p-$value est supérieure à 5%, on ne rejette pas l'hypothèse nulle. On peut supposer que les observations sont normales.

Simulons des observations uniformes et appliquons le test de Shapiro–Wilk.

```
> y=runif(50,-5,15)
> y
 [1] -4.0879249 -3.1266726 -0.7170668 -2.2280176 -1.7824772 14.2489004
 [7]  2.1059444 -3.3355710 -2.7841404  1.3467159  1.0099143  7.3365456
[13]  7.8388273 -0.3247354  1.2960432  8.6903540 10.2471285  1.0189554
[19]  3.9594622  7.2452501 12.5912116  3.7068249  1.8631227  8.8284095
[25]  0.7586813  0.9331361 12.0389597 -2.1943549  1.0193574  9.9336526
[31] -3.5596714  8.6926976  9.7736992  1.4294532 -4.8870104 11.9398889
[37] 10.3700827  3.0211665  0.5607261 -2.7961990 -2.4244256  3.3730912
[43]  0.4950279  9.9933305  3.8206514  1.2576399  7.2472824 -1.5863573
[49]  3.6153996  4.0258651
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.94087, p-value = 0.01459
```

La $p-$value est inférieure à 5%, on rejette l'hypothèse nulle. On peut affirmer que les observations ne sont pas normales.

Part IV

Régression linéaire.

On rappelle la régression linéaire entre deux variables d'un point de vue descriptif.

9 Régression linéaire – Statistiques descriptives

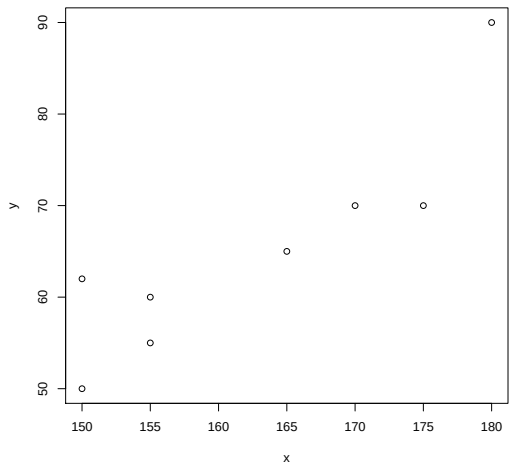
Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

9.1 Nuage de points

On relève le couple (taille, poids) de 8 individus. On résume les données dans le tableau suivant.

taille	x	150	155	155	150	165	175	170	180
poids	y	50	55	60	62	65	70	70	90

Definition 9.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*



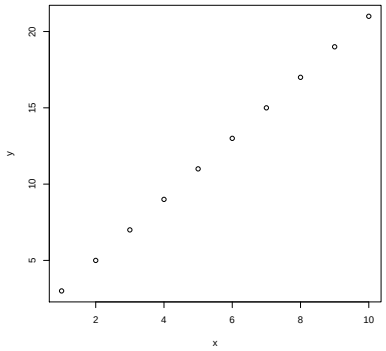
Definition 9.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

Il s'agit du centre de gravité du nuage. On rencontrera parfois cette dénomination. Dans notre exemple, le point moyen est $(65.2, 162.5)$.

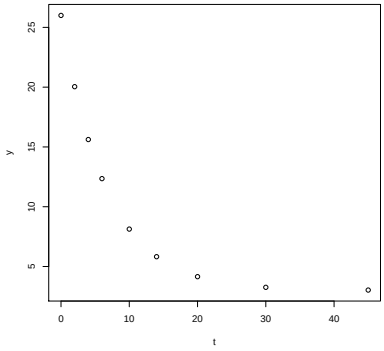
9.2 Forme du nuage de points

D'une manière générale, trois cas peuvent se présenter en ce qui concerne le profil du nuage :

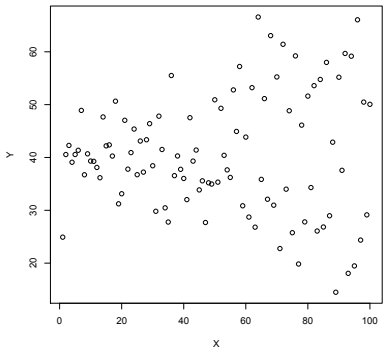
(i) forme allongée et rectiligne : les points sont plus ou moins alignés



(ii) forme allongée mais non rectiligne : les points ne sont pas alignés mais ont un profil ordonné



(iii) forme quelconque



9.3 Ajustement affine (droite de régression linéaire)

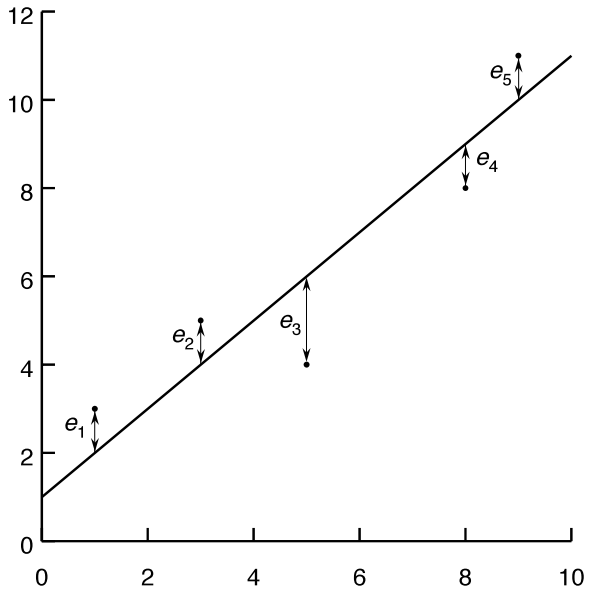
On s'intéresse plus particièrement au premier cas 9.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

$$y = \beta_1 x + \beta_0$$

qui passe au plus proche des points du nuage de points. Cette droite nous servira donc d'approximation. Bien évidemment, suivant la méthode utilisée pour la construire, on peut obtenir différentes droites. La méthode la plus utilisée car donnant la meilleure approximation est la méthode des moindres carrés.

9.3.1 La méthode des moindres carrés

L'idée de cette méthode est de chercher la droite qui minimise la somme des carrés des écarts verticaux entre la droite et les points du nuage, les *résidus*.



En pratique, on détermine les coefficients de la droite $D : y = \beta_1 x + \beta_0$ à l'aide de `R` ou d'un tableur. La droite ainsi obtenue est unique. Cette droite s'appelle la droite de régression linéaire de y en x par la méthode des moindres carrés. On note

$$\sigma_{xy} = Cov(x, y) = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

Cette quantité est nommée covariance de x et y. Si la quatité σ_x est une distance entre les valeurs de x est $\bar{x}$, on peut considérer la covariance comme un produit scalaire entre les variables x et y. Ainsi, si la covariance est proche de 0, on peut penser que les variables ont une dynamique qui n'ont rien de commun (penser à l'orthogonalité), c'est à dire le nuage 9.2.3.

On a

$$\begin{aligned} \beta_1 &= cov(x, y)/\sigma_x^2, \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}. \end{aligned}$$

Preuve. On pose la somme des carrés des résidus :

$$M(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2.$$

Le minimum de $M(\beta_1, \beta_0)$ s'obtient en annulant les dérivées partielles par rapport à β_1 et β_0.

$$\begin{cases} \frac{\partial M(\beta_1,\beta_0)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i - \beta_0) &= 0 \\ \frac{\partial M(\beta_1,\beta_0)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0) &= 0 \end{cases}.$$

On obtient donc un système à deux équations à résoudre en β_1 et β_0. En divisant chaque ligne par -2, le système est équivalent à

$$\begin{cases} \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i - \beta_0) &= 0 \\ \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0) &= 0 \end{cases} \iff \begin{cases} \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2 - \beta_0 \sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \beta_0 &= 0 \end{cases}.$$

On divisera par n, par commodité, le système est équivalent à

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{\beta_1}{n} \sum_{i=1}^{n} x_i^2 - \frac{\beta_0}{n} \sum_{i=1}^{n} x_i &= 0 \\ \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta_1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \beta_0 &= 0 \end{cases}.$$

Il vient que

$$\begin{cases} \overline{xy} - \beta_1 \overline{x^2} - \beta_0 \bar{x} &= 0 \\ \bar{y} - \beta_1 \bar{x} - \beta_0 &= 0 \end{cases}.$$

La deuxième ligne nous donne

$$\beta_0 = \bar{y} - \beta_1 \bar{x},$$

et en l'injectant dans la première équation

$$\begin{array}{rrcl} & \overline{xy} - \beta_1 \overline{x^2} - (\bar{y} - \beta_1 \bar{x})\bar{x} &=& 0 \\ \iff & \overline{xy} - \beta_1 \overline{x^2} - \bar{x}\bar{y} + \beta_1 \bar{x}^2 &=& 0 \\ \iff & \overline{xy} - \bar{x}\bar{y} - \beta_1 (\overline{x^2} - \bar{x}^2) &=& 0 \\ \iff & cov(x,y) - \beta_1 Var(x) &=& 0. \end{array}$$

On en conclut que

$$\beta_1 = \frac{Cov(x,y)}{Var(x)}.$$

Il reste à montrer que c'est un minimum, c'est à dire que le déterminant de la matrice Hessienne est strictement positif. Rappelons les dérivées partielles premières,

$$\begin{cases} \frac{\partial M(\beta_1,\beta_0)}{\partial \beta_1} &= -2\sum_{i=1}^n x_i(y_i - \beta_1 x_i - \beta_0) \\ \frac{\partial M(\beta_1,\beta_0)}{\partial \beta_0} &= -2\sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0) \end{cases}.$$

On dérive à nouveau,

$$\begin{cases} \frac{\partial^2 M(\beta_1,\beta_0)}{\partial^2 \beta_1} &= 2\sum_{i=1}^n x_i^2 &=& 2n\overline{x^2} \\ \frac{\partial^2 M(\beta_1,\beta_0)}{\partial^2 \beta_0} &= 2\sum_{i=1}^n 1 &=& 2n \\ \frac{\partial^2 M(\beta_1,\beta_0)}{\partial \beta_0 \partial \beta_1} &= 2\sum_{i=1}^n x_i &=& 2n\bar{x} \end{cases}.$$

Le déterminant de la matrice Hessienne est donc

$$det(Hess) = \begin{vmatrix} 2n\overline{x^2} & 2n\bar{x} \\ 2n\bar{x} & 2n \end{vmatrix} = 4n^2\overline{x^2} - 4n^2(\bar{x})^2 = 4n^2 Var(x) > 0.$$

Il s'agit bien d'un minimum, ce qui termine la preuve.□

9.3.2 Coefficient de corrélation linéaire

Notons que la méthode des moindres carrés peut être utilisée pour n'importe quelle série double. On peut tout à fait obtenir une droite de régression dans le cas 9.2.3. Pour s'assurer de façon objective (et non purement visuelle) que l'ajustement est valide, on considère un autre paramètre de la série : le coefficient de corrélation r

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Proposition 9.3 *On a les propriétés suivantes :*

(i) on a toujours $-1 \leq r \leq 1$ *;*

(ii) le coefficient directeur de la droite de régression et le coefficient de corrélation sont de même signe ;

(iii) le degré de corrélation est d'autant plus fort que r *est proche de* 1 *ou* -1*.*

C'est l'assertion 3.*iii* qui nous permet de dire si la droite de régression est proche des points. En pratique, une régression linéaire est légitime si $r > 0.9$ ou si $r < -0.9$.

10 Point de vue inférentiel

On peut supposer que x et y sont les observations d'un échantillon des variables X et Y. On écrit donc le modèle

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

Les valeurs β_1 et β_0 calculées ci-dessus sont en réalité les estimations $\hat{\beta}_1$ et $\hat{\beta}_0$ par la méthode des moindres carrés, *i.e.* minimisant la somme des carrés des écarts (par rapport à la droite)

$$SCE = \sum_i \varepsilon_i^2.$$

On a alors

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2},$$

et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

On notera alors les valeurs prédites

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

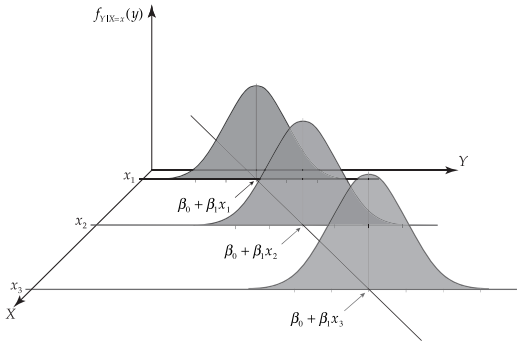
et $\hat{Y}_i$ est la valeur associée à X_i par la droite de régression (dite empirique). Il vient que

$$SCE = \sum_i (\hat{Y}_i - Y_i)^2.$$

On formule les hypothèses suivantes sur les termes d'erreurs.

10.1 Hypothèses sur les termes d'erreur ε

- Indépendance des erreurs : les $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ sont indépendants.
- Éxogénéité : les variables explicatives $(X_1, \cdots X_n)$ ne sont pas corrélées au terme d'erreur. De plus, les erreurs sont centrées $E(\varepsilon_i) = 0$
- Homoscédasticité : les termes d'erreurs sont supposés de variance constante.
- Normalité des termes d'erreur : les termes d'erreurs suivent une loi normale, centrées, de variance σ^2



Lemma 10.1 *Les hypothèses du modèle montrent que*

$$Y_i = \beta_1 X + \beta_0 + \varepsilon_i$$

suit une loi normale $\mathcal{N}(\beta_1 X_i + \beta_0, \sigma^2)$. *De plus les* Y_i *sont indépendants.*

10.2 Équation de la variance

D'après les hypothèse précédentes, il vient que

$$Var(Y) = Var(\hat{Y}) + Var(\varepsilon),$$

grâce à l'éxongénéité. Il vient que

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2.$$

Cette équation a une signification intéssante.

- Le terme $\sum_i (Y_i - \bar{Y})^2$ représente la variation totale des valeurs des Y_i par rapport à leur moyenne $\bar{Y}$. On notera cette quantité SCT : *Somme des Carrés Totale.*
- Puisque le terme $\sum_i (\hat{Y} - \bar{Y})^2$ est l'écart de la valeur prédite par rapport à la moyenne, nous dirons que ce terme est la *Somme des Carrés due au Modèle*, notée SCM[4].

On a donc

$$SCT = SCM + SCE. \tag{10.1}$$

10.3 Coefficient de détermination

Le coefficient de détermination est le rapport de variance de Y expliquée par la régression :

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)}.$$

Le coefficient R^2 est donc la proportion de variance de Y expliquée par le modèle. Dès lors que $\bar{\hat{Y}} = \bar{Y}$, on peut formuler

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{\frac{1}{n-1}\sum_i (\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2} = \frac{SCM}{SCT}.$$

Il se trouve que R^2 est le carré du coefficient de corrélation linéaire. En effet,

$$(r_{X,Y})^2 = \frac{cov^2(X,Y)}{VarXVarY} = \frac{cov^2(X,\hat{Y}+\varepsilon)}{VarXVarY} = \frac{cov^2(X,\hat{Y})}{VarXVarY} = \frac{\hat{\beta}_1^2 Var^2 X}{VarXVarY} = \frac{\hat{\beta}_1^2 VarX}{VarY} = \frac{Var\hat{Y}}{VarY}.$$

10.4 Distribution de $\hat{\beta}_1$

La méthode des moindres carrés prend le parti de ne pas considérer d'erreur sur les valeurs x_i prises par X. Il vient qu'on peut considérer l'ensemble de valeurs $X_i = x_i$ qui seront non aléatoires et le modèle équivalent :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Les termes ε_i sont des variables aléatoires normales identiques et indépendantes de moyenne nulle et variance σ_ε^2, quelque soit la valeur X_i. On en déduit le théorème suivant qui donne la distribution de $\hat{\beta}_1$.

Theorem 10.2 *Sous les hypothèses du modèle de régression linéaire simple, $\hat{\beta}_1$ suit une loi normale d'espérance β_1 et de variance*

$$\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

[4]Certains auteurs noteront cette quantité SCR, pour somme des carrés dûs à la régression. Nous éviterons cette notation qui peut prêter à confusion avec SCE qui renvoie à la somme des carrés des résidus

Pour prouver ce résultat, on aura besoin du lemme suivant.

Lemma 10.3 *On pose*

$$a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Nous avons

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i,$$

et

$$\sum_{i=1}^n a_i = 0, \qquad \sum_{i=1}^n a_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}, \qquad \sum_{i=1}^n a_i X_i = 1.$$

Proof. Tout d'abord,

$$\sum_{i=1}^n a_i = \sum_{i=1}^n \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i - n\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{n\bar{X} - n\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0.$$

Notons que

$$\begin{aligned}\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i - \bar{Y} \sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \sum_{i=1}^n \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} Y_i - \bar{Y} \sum_{i=1}^n \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (10.2)\end{aligned}$$

Ce qui revient à

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i - \bar{Y} \sum_{i=1}^n a_i = \sum_{i=1}^n a_i Y_i$$

Calculons maintenant

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

Enfin, en remplaçant Y_i par X_i et $\bar{Y}$ par $\bar{Y}$ dans le calcul (10.2), on peut observer que

$$\sum_{i=1}^n a_i X_i = \sum_{i=1}^n a_i (X_i - \bar{X}) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = 1.$$

Ce qui prouve le Lemme. □

Nous avons tous les éléments pour prouver le Théorême 10.2 de manière élégante.

Proof. Notons que $\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i$ suit une loi normale en tant que combinaison linéaire de $Y_1, \cdots Y_n$, variables normales indépendantes. Calculons son espérance.

$$\begin{aligned}E\hat{\beta}_1 &= \sum_{i=1}^n a_i EY_i = \sum_{i=1}^n a_i E(\beta_1 X_i + \beta_0 + \varepsilon_i) \\ &= \sum_{i=1}^n a_i (\beta_1 X_i + \beta_0 + E\varepsilon_i) \\ &= \sum_{i=1}^n a_i (\beta_1 X_i + \beta_0) \\ &= \beta_1 \sum_{i=1}^n a_i X_i + \beta_0 \sum_{i=1}^n a_i \\ &= \beta_1.\end{aligned}$$

En utilisant l'indépendance des Y_i, la variance est

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} \left(a_i^2 Var(Y_i)\right) = \sum_{i=1}^{n} a_i^2 \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Ce qui prouve le Théorême 10.2

On fait les remarques suivantes.

- Ce théorème montre que $\hat{\beta}_1$ est sans biais pour β_1. Cela signifie que d'un échantillon à l'autre, la valeur de $\hat{\beta}_1$ oscille autour de la valeur théorique β_1.
- Ces écarts par rapport à la moyenne β_1 sont distribués selon une loi normale dont la variance est

  $$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

  On note, donc, que la variance de $\hat{\beta}_1$ croît avec σ_ε^2, mais qu'elle décroît lorsque $\sum_{i=1}^{n}(X_i - \bar{X})^2$ croît. Ainsi, plus les X_i sont nombreux et dispersés, plus notre estimation sera fiable.
- On acceptera que la distribution de $\hat{\beta}_0$ est normale et suit la loi

  $$\mathcal{N}\left(\beta_0, \sigma_\varepsilon^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right).$$

En pratique nous ne connaissons pas σ_ε^2, la variance de la variable ε, qui est nécessaire dans le calcul de $\sigma_{\hat{\beta}_1}$. Cependant, nous disposons d'une estimation de celle-ci, à savoir :

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{SCE}{n-2}.$$

Il vient que

$$(n-2)\frac{s_\varepsilon^2}{\sigma_\varepsilon^2} \backsim \chi_{n-2}^2.$$

On en conclut que le rapport $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$ n'est pas distribué selon une loi $\mathcal{N}(0,1)$, mais plutôt selon une loi t de Student ayant $n-2$ degrés de liberté. On a donc les distributions suivantes.

Theorem 10.4 *En estimant σ_ε par s_ε, on obtient les distributions des estimateurs suivantes*

$$(\hat{\beta}_0 - \beta_0)/s_{\hat{\beta}_0} \backsim t_{n-2}, \qquad (\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1} \backsim t_{n-2},$$

avec

$$s_{\hat{\beta}_1}^2 = \frac{SCE}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

et

$$s_{\hat{\beta}_0}^2 = \frac{SCE}{(n-2)}\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right].$$

Si le nombre de degrés de liberté est assez élevé (plus de trente), on peut faire une approximation de la loi t de Student par une loi $\mathcal{N}(0,1)$.

10.5 Intervalles de confiance

10.5.1 Intervalles de confiances des coefficients de la régression

Le dernier théorême de la section précédente, donne les distributions suivantes.

$$(\hat{\beta}_0 - \beta_0)/s_{\hat{\beta}_0} \backsim t_{n-2}, \qquad (\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1} \backsim t_{n-2}, \qquad (n-2)\frac{s_\varepsilon^2}{\sigma_\varepsilon^2} \backsim \chi^2_{n-2}.$$

On en déduit les intervalles de confiance à $100(1-\alpha\%)$ suivants.

$$\frac{(n-2)s_\varepsilon^2}{q_{\alpha/2}^{\chi^2(n-2)}} \geq \sigma_\varepsilon^2 \geq \frac{(n-2)s_\varepsilon^2}{q_{1-\alpha/2}^{\chi^2(n-2)}},$$

$$\hat{\beta}_0 - q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_0},$$

$$\hat{\beta}_1 - q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_1}$$

où la valeur $q_{1-\alpha/2}^{t_{n-2}}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté (obtenu de la table de la loi de Student).

10.5.2 Intervalles pour les prévisions

Lorsque nous substituons dans l'équation de la droite de régression une valeur donnée de X, soit X_0, nous obtenons une certaine valeur que nous notons :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

Cette valeur de X_0 peut être utilisée à deux fins, car elle estime deux choses :

- $E[Y_0] = \beta_0 + \beta_1 X_0$, c'est-à-dire la moyenne de la variable Y en $X = X_0$;
- $E[Y_0] + \varepsilon = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon$, c'est-à-dire une observation de Y pour un individu ayant en $X = X_0$.

Lorsque l'on fait de telles prévisions, on préfère accompagner celles-ci de limites de confiance.

(i) Dans le premier cas, lorsque nous voulons estimer la moyenne de la variable Y lorsque la valeur de X demeure fixée à X_0, nous utilisons l'intervalle à $100\alpha\%$ de confiance suivant :

$$\hat{Y}_0 \pm q_{1-\alpha/2}^{t_{n-2}} \sqrt{s_\varepsilon^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)}.$$

où la valeur $q_{1-\alpha/2}^{t_{n-2}}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté.

Proof. On utilisera les notations et les résultat du Lemme 10.3. Pour rappel,

$$a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Nous avons

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i,$$

et

$$\sum_{i=1}^n a_i = 0, \qquad \sum_{i=1}^n a_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}, \qquad \sum_{i=1}^n a_i X_i = 1.$$

Dès lors que

$$\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0 \qquad \Longleftrightarrow \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

nous obtenons,

$$\hat{Y}_0 = \bar{Y} + \hat{\beta}_1(X_0 - \bar{X}).$$

Puisque

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i,$$

nous avons,

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i(\beta_1 X_i + \beta_0 + \varepsilon_i).$$

Il suit des égalités sur les sommes des a_i que

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} a_i \varepsilon_i.$$

Par conséquent,

$$\begin{aligned} Var(\hat{Y}_0) &= Var\big[\bar{Y} + \hat{\beta}_1(X_0 - \bar{X})\big] \\ &= Var\left[\beta_0 + \beta_1\bar{X} + \bar{\varepsilon} + (X_0 - \bar{X})\left(\beta_1 + \sum_{i=1}^{n} a_i\varepsilon_i\right)\right] \\ &= Var\left[\bar{\varepsilon} + (X_0 - \bar{X})\sum_{i=1}^{n} a_i\varepsilon_i\right]. \end{aligned}$$

On voit alors que le coefficient de ε_i est

$$\frac{1}{n} + (X_0 - \bar{X})a_i.$$

Il suit de l'indépendance des ε_i,

$$\begin{aligned} Var(\hat{Y}_0) &= \sigma_\varepsilon^2 \sum_{i=1}^{n}\left(\frac{1}{n} + (X_0 - \bar{X})a_i\right)^2 \\ &= \sigma_\varepsilon^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{2}{n}(X_0 - \bar{X})a_i + (X_0 - \bar{X})^2 a_i^2\right) \\ &= \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right). \end{aligned}$$

Cette variance est estimée par

$$Var(\hat{Y}_0) = s_\varepsilon^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

□

(ii) Dans le second cas, il s'agit de prévoir, pour un individu donné, la valeur de Y qui lui est propre, sachant que sa valeur en X est X_0. L'intervalle est

$$\hat{Y}_0 \pm q_{1-\alpha/2}^{t_{n-2}}\sqrt{s_\varepsilon^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}.$$

Proof. Il s'agit maintenant de trouver la variance de $\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon$. On note que ε est une réplication du terme d'erreur indépendante des autres. Il suit

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon) = Var(\hat{\beta}_0 + \hat{\beta}_1 X_0) + Var(\varepsilon).$$

De manière analogue à la preuve précédente,

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon) = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right),$$

qui peut être estimée par

$$s_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right).$$

□

11 Tests sur la pente de la droite

Pour faire simple, les tests F de Fischer et t de Student testent l'hypothèse $\mathcal{H}_0$ sous laquelle le coefficient β_1 est nul, contre β_1 est non nul (ce qui permet d'affirmer que X explique Y, au moins en partie).

11.1 Test de Student

Notons l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0\text{”},$$

autrement formulée, $\mathcal{H}_0$ est équivalente à “X n'explique pas Y”. L'estimation de β_1 dans le théorème 10.4 montre que

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \backsim t_{n-2},$$

et sous $\mathcal{H}_0$, nous garderons

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \backsim t_{n-2}.$$

Rappelons l'estimation

$$s_{\hat{\beta}_1} = \sqrt{\frac{SCE}{(n-2)\sum (X_i - \bar{X})^2}}.$$

Nous ferons donc un test (bilatéral) de Student sur la statistique de test $\hat{\beta}_1 / s_{\hat{\beta}_1}$. On rejettera $\mathcal{H}_0$ au seuil α si

$$\frac{|\hat{\beta}_1|}{s_{\hat{\beta}_1}} \geq q_{1-\alpha/2}^{t_{n-2}}.$$

11.2 Table d'ANOVA

L'analyse de la variance, souvent présentée sous forme d'un tableau, permet d'éclairer sur l'influence de la variable X sur la variable Y grâce à l'étude de la décomposition de la variance (10.1). Notons, encore, l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0\text{”}.$$

On note que $SCM = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2$. On a vu que

$$s_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2 = \frac{SCE}{n-2}.$$

et on rappelle que sous $\mathcal{H}_0$ ($\beta_1 = 0$), $\hat{\beta}_1 / s_{\hat{\beta}_1} \backsim t_{n-2}$, avec

$$s_{\hat{\beta}_1}^2 = \frac{SCE}{(n-2)\sum (X_i - \bar{X})^2}.$$

Il vient que

$$\left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)^2 = \frac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{\frac{SCE}{n-2}} = \frac{SCM}{SCE/(n-2)}$$

suit une loi de Fisher $\mathcal{F}_{1,n-2}$ en tant que carré de la loi de Student. La variable du test est donc

$$\frac{SCM}{SCE/(n-2)},$$

et on observe son éloignement (à droite) de zéro.

Ainsi, si la p-value dans la table d'ANOVA est proche de zéro (ou en dessous du seuil fixé), on rejettera la nullité de $\hat{\beta}_1$.

Dans ce cadre, on voit qu'on peut utiliser le test de t de Student pour le rapport $\hat{\beta}_1/s_{\hat{\beta}_1}$ ou F de Fisher pour le carré de ce rapport, sans distinction. Il est totalement équivalent en cas de régression simple (ce n'est pas le cas sur une régression multiple). On note d'ailleurs que la statistique F est le carré de la statistique t.

Dans le cadre d'une régression multiple (sur plusieurs variables explicatives), le test de Fischer teste l'effet global des variables sur la variable Y, les tests de Student testent l'effet de chaque variable explicative sur Y.

12 Tests sur régression linéaire multiple

La question qu'on se pose est de savoir si la variable réponse est expliquée par les variables explicatives dans leur globalité, ou par telle ou telle variable explicative. Cela se traduit, mathématiquement, par la non nullité des coefficients de la régression. En effet, si le coefficient d'une des variable explicative est nul ou presque nul, cette variable explicative fait peu varier la régression linéaire, elle n'influence donc pas la variable réponse. Plaçons-nous dans le cadre de la régression linéaire multiple.

12.1 Les tests t de Student

Les tests de Student testent la nullité de chaque coefficient de la régression linéaire. Ainsi, on saura quelles variables explicatives ont un effet sur la variable expliquée.

12.1.1 Hypothèse nulle

L'hypothèse nulle de chaque test est $\mathcal{H}_0$ ="La variable X_i n'a pas d'effet sur la variable réponse"="$\beta_i = 0$".

12.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que la variable X_i a un effet sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. La variable X_i n'a pas d'effet sur la variable réponse.

12.3 Les tests F de Fisher - ANOVA

Le test de Fisher teste l'effet de l'ensemble des variables explicatives sur la variable réponse. Ainsi, on saura si la variable réponse est expliquée par les variables explicatives. On appelle cela une ANalyse de la (Of) VAriance.

12.3.1 Hypothèse nulle

L'hypothèse nulle du test est $\mathcal{H}_0$ ="Les variables X_i n'ont pas d'effet, dans leur globalité, sur la variable réponse"="la variance de l'erreur est très forte face à la variance expliquée par le modèle".

12.3.2 Décision

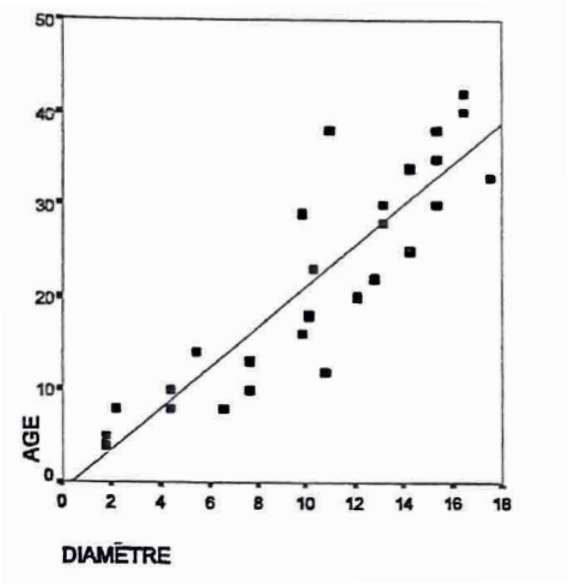
- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que les variable X_i ont un effet global sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Les variables X_i n'ont pas d'effet sur la variable réponse.

12.4 Exemple : le cas de la régression linéaire simple

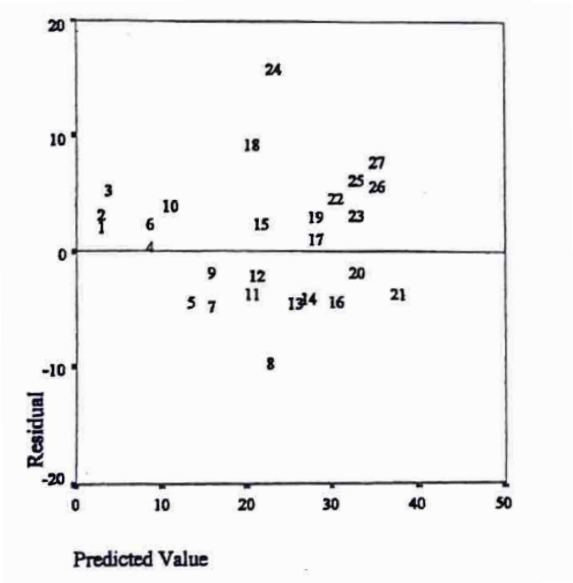
On sait que l'on mesure l'âge d'un arbre en comptant les anneaux sur une section transversale du tronc, mais cela nécessite de l'avoir abattu auparavant. Peut-on connaître l'âge à partir de la mesure de sa circonférence ?

Afin de répondre à cette question, on a effectué les mesures sur un échantillon de 27 arbres de la même espèce. À partir de ces données, on a effectué une régression de l'âge en fonction du diamètre . Les résultats ont été traités à l'aide du logiciel SPSS.

Nuage de points



Résidus



	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	2905,549	1	2905,55	93,44	,000
Résidu	777,414	25	31,097		
Total	3682,963	26			

	Coefficients non standardisés		Coefficients standardisés		
	B	Erreur standard	Bêta	t	Signification
(constante)	-,974	2,604		-,374	,711
DIAMETRE	2,206	,228	,888	9,67	,000

Très logiquement, puisqu'il n'y a qu'une variable explicative, l'effet de toutes les variables explicatives est équivalent à l'effet de chaque (l'unique) variable explicative. Par conséquent, il n'est pas surprenant d'observer que $\sqrt{F} = t_{diametre}$. C'est à dire que, dans le cas de la régression linéaire simple, le test de Student et de Fisher sont totalement équivalents.

Dans notre exemple, nous avons, tant pour F que pour t, des $p-$values inférieures à 5%. Nous en concluons que le diamètre explique significativement l'âge des arbres.

12.5 Exemple : le cas de la régression linéaire multiple

On cherche à modéliser la relation entre poids des bébés à la naissance et l'âge, le poids et le statut tabagique de la mère durant la grossesse. (Exemple fictif) On pose

- y = poids de naissance en grammes (bwt),
- x_1 = âge de la mère (age),
- x_2 = poids de la mère en kilos (weight),
- x_3 = statut tabagique de la mère pendant la grossesse (smoke) codé par un score à une échelle de 1 à 20.

```
> modele=lm(bwt~age+weight+smoke)
> summary(modele)

Call:
lm(formula = bwt ~ age + weight + smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-385.81  -65.83   -0.70   68.17  290.66

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3005.65519   22.99506  130.71   <2e-16 ***
age            0.02645    0.53148    0.05     0.96
weight         8.44845    0.30499   27.70   <2e-16 ***
smoke        -26.53764    1.82009  -14.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.58 on 996 degrees of freedom
Multiple R-squared:  0.4919,    Adjusted R-squared:  0.4904
F-statistic: 321.4 on 3 and 996 DF,  p-value: < 2.2e-16
```

En regardant la $p-$value du test F de Fisher, puisqu'elle est inférieure à 5%, on en déduit que les variables $(age, weight, smoke)$ ont globalement un effet sur la variable bwt. Dans le détail, en observant les $p-$values des test t de Student, on remarque que l'age n'a pas d'effet significatif sur bwt ($p \geq 0.05$) mais que les variables $weight$ et $smoke$ ont un effet significatif sur bwt ($p \sim 0$).

Part V
ANOVA

13 ANOVA à un facteur

Nous avons utilisé l'ANOVA dans l'étude du modèle linéaire

$$Y = \beta_1 \xi + \beta_0 + \epsilon,$$

avec certaines hypothèses sur ξ, Y, ϵ de normalité et de non corrélation des termes d'erreur. Nous avions utilisé les test de Student et de Fisher afin de vérifier la non nullité de β_1, ce qui entrainerait l'absence d'effet de ξ sur Y, par l'étude des moyennes ou des variances.

La variable explicative ξ était alors quantitative. Il n'est cependant pas rare de rencontrer une variable explicative qualitative. Le passage par une régression linéaire n'a plus de sens dès que la multiplication $\beta_1 \xi$ n'en a plus. Prenons par exemple la variable ξ à deux modalités :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",

ou plus :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",
- ξ_3 : "traitement expérimental à forte dose".

La variable ξ s'appelle le facteur. On pourra chercher à expliquer une variable réponse X, par exemple le taux d'une hormone. Pour chaque valeur ξ_i, on obtient un échantillon indépendant X_i. Dans le premier cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$

,

ou dans le deuxième cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$
$\xi = \xi_3$: "traitement expérimental à forte dose"	$X_{3,1}, X_{3,2}, \cdots X_{3,n_3}$

.

On considère le modèle

$$X_i = EX_i + \epsilon_i.$$

La question est de savoir si les $\mu_i = EX_i$ sont identiques (ξ n'as pas d'effet sur X) ou différents selon les valeurs ξ_i. Dans ce cas, ξ influence X.

13.1 Facteur à deux valeurs - t de Student

On considère deux échantillons indépendants de tailles n_1 et n_2, respectivement :

$$X_{1,1}, X_{1,2}, \cdots, X_{1,n_1};$$

$$X_{2,1}, X_{2,2}, \cdots, X_{2,n_2},$$

où $X_{i,j}$ représente la jè observation du iè échantillon, ($i = 1, 2$ et $j = 1, \cdots, n_i$). Nous notons $\bar{X}_i$ la moyenne estimées des deux groupes i, $i = 1, 2$. Nous supposons que ces échantillons sont issus de deux populations normales de moyennes μ_1 et μ_2 et de variance commune σ^2, estimée par

$$s^2 = \frac{\sum_{j=1}^{n_1}(X_{1,j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2,j} - \bar{X}_2)^2}{n_1 + n_2 - 2}.$$

On note que $(n_1 + n_2 - 2)s^2/\sigma^2$ suit une loi $\chi^2(n_1 + n_2 - 2)$.

Notre but est de tester l'hypothèse

$$\mathcal{H}_0 = \text{“}\mu_1 = \mu_2\text{”},$$

équivalente à “Le facteur n'a pas d'effet sur la variable X”, ou encore “les deux échantillons sont issus de la même population”. Nous allons donc étudier l'estimateur de la différence des moyennes $\bar{X}_1 - \bar{X}_2$, de moyenne nulle par hypothèse nulle, et de variance

$$Var(\bar{X}_1 - \bar{X}_2) = Var\left(\frac{1}{n_1}\sum_{j=1}^{n_1} X_{1,j} - \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2,j}\right) = \frac{1}{n_1^2}Var\left(\sum_{j=1}^{n_1} X_{1,j}\right) + \frac{1}{n_2^2}Var\left(\sum_{j=1}^{n_2} X_{2,j}\right),$$

par indépendance des échantillons. Il en suit

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{n_1}{n_1^2}\sigma^2 + \frac{n_2}{n_2^2}\sigma^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Cette quantité sera estimée par

$$s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Le test est basé sur la variable

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

On rejettera donc $\mathcal{H}_0$ au seuil α si $|T| \geq q_{1-\alpha/2}^{t(n_1+n_2-2)}$.

Le numérateur de la statistique T est une mesure de l'écart entre les moyennes échantillonnales, alors qu'au dénominateur figure l'écart type s qui est une mesure de la dispersion à l'intérieur des échantillons. Nous rejetons $\mathcal{H}_0$ lorsque $|T|$ prend une valeur trop grande, c'est-à-dire lorsque l'écart entre les échantillons est trop grand comparé à la dispersion à l'intérieur des échantillons. Nous utiliserons le même principe maintenant dans le cas de plus de deux échantillons.

Si ici, avec la variable T nous utiliserions un test de Student, nous pourrions également utiliser un test unilatéral de Fisher avec la variable T^2. C'est cette approche qui permet de généraliser la démarche aux facteurs à plusieurs valeurs.

13.2 Facteur à a modalités

13.2.1 Le modèle

Supposons donc qu'on prélève a échantillons indépendants :

$$\begin{array}{lll} X_{1,1}, X_{1,2}, & \cdots & , X_{1,n_1}; \\ X_{2,1}, X_{2,2}, & \cdots & , X_{2,n_2}; \\ & \vdots & \\ X_{a,1}, X_{a,2}, & \cdots & , X_{a,n_a}; \end{array}$$

où $X_{i,j}$ représente la $j$ème observation du $i$ème échantillon, ($i = 1, \cdots, a$ et $j = 1, \cdots, n_i$). Les échantillons indépendants sont issus des populations normales de moyenne $\mu_1, \cdots, \mu_a$ et de variance commune σ^2. On pose donc le modèle

$$X_{i,j} = \mu_i + \varepsilon_{ij},$$

où les ε_{ij} sont des lois normales $\mathcal{N}(0, \sigma)$ indépendantes.

13.2.2 Le test de Fisher

L'hypothèse à tester est

$$\mathcal{H}_0 : \text{“}\mu_1 = \mu_2 = \cdots = \mu_a = \mu''.$$

Chacune des espérances μ_i des échantillons sont estimées par les moyennes

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j},$$

et par la moyenne totale

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^{a} n_i \bar{X}_i$$

pour μ. Une bonne façon de tester l'égalité de toutes les moyennes est de les comparer à la moyenne commune $\bar{X}$:

$$\sum_{i=1}^{a} (\bar{X}_i - \bar{X})^2,$$

ou mieux, en faisant apparaître le rapport de force de chaque moyenne grâce à l'effectif de chaque échantillon

$$SCM = \sum_{i=1}^{a} n_i (\bar{X}_i - \bar{X})^2,$$

où SCM signifie somme des carrés due au modèle.

On observe que

$$\bar{X}_i - \bar{X} = \sum_{j=1}^{n_i} \frac{1}{n_i} (X_{ij} - \bar{X}) = \sum_{j=1}^{n_i} \frac{1}{n_i} \varepsilon_{ij} = \bar{\varepsilon}_i,$$

dont la variance est σ^2/n_i. Par conséquent,

$$\sum_{i=1}^{a} \frac{n_i (\bar{X}_i - \bar{X})^2}{\sigma^2} = \frac{SCM}{\sigma^2} \sim \chi^2(a-1).$$

On pose

$$SCE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2.$$

Puisque

$$\sum_{i=1}^{a} (n_i - 1) = n - a,$$

on estime σ^2 par $SCE/(n-a)$ et

$$\frac{SCE}{\sigma^2} \sim \chi^2(n-a).$$

La variable du test est donc

$$F = \frac{SCM/(a-1)}{SCE/n-a} \sim \mathcal{F}_{a-1,n-a}.$$

Nous rejetons $\mathcal{H}_0$ au seuil α si

$$F = \frac{CMM}{CME} = \frac{SCM/(a-1)}{SCE/n-a} \geq q_{1-\alpha}^{\mathcal{F}_{a-1,n-a}},$$

où q est le quantile d'ordre $1-\alpha$ de la dite loi.

Remarquons que nous rejetons $\mathcal{H}_0$ seulement si F est trop grand et non si F est trop petit car un F grand signifie que les $\bar{X}_i$ sont trop dispersés, et donc que les μ_i ne semblent pas être tous égaux.

13.2.3 Équation de la variance

Posons de plus

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

pour la dispersion totale. On peut aisément établir l'équation de la variance suivante.

$$SCT = SCM + SCE. \tag{13.3}$$

Cette décomposition met en évidence le fait que la dispersion totale des données (SCT) est formée d'une partie (SCM) expliquée par le fait que les populations sont différentes, et d'une autre partie (SCE) qu'on attribue au hasard. Autrement dit, SCE représente les différences individuelles alors que SCM représente les différences entre les groupes. On rejette l'hypothèse que les populations d'origine des groupes sont de même moyenne si les différences entre les groupes sont trop grandes par rapport aux différences individuelles. Cette analyse est appelée analyse de variance. Les calculs se font plus aisément à l'aide des formules suivantes.

Proposition 13.1

$$SCM = \sum_{i=1}^{a} n_i\bar{X}_i^2 - n\bar{X}^2,$$

$$SCT = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2,$$

$$SCE = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{a} n_i\bar{X}_i^2.$$

Ces formules s'obtiennent en développant les carrés et en regroupant les moyennes.

Preuve l'équation de la variance (13.3)

$$\begin{aligned} SCT &= \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{a} n_i\bar{X}_i^2 + \sum_{i=1}^{a} n_i\bar{X}_i^2 - n\bar{X}^2 \\ &= SCE + SCM \end{aligned}$$

Les résultats d'une analyse de variance sont habituellement présentés sous la forme d'un tableau comme le suivant :

Source	Somme des carrés	d.l.	Moyenne des carrés	F
Modèle	$\mathrm{SCM} = \sum_{i=1}^{a} n_i \bar{X}_i^2 - n\bar{X}^2$	$a-1$	$\frac{\mathrm{SCM}}{a-1}$	$F = \frac{\mathrm{CMM}}{\mathrm{CME}}$
Erreur	$\mathrm{SCE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2$	$n-a$	$\frac{\mathrm{SCE}}{n-a}$	
Total	$\mathrm{SCT} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2$	$n-1$	$\frac{\mathrm{SCT}}{n-1}$	

13.2.4 En cas de rejet de l'hypothèse d'égalités des moyennes

La table d'analyse de variance nous permet de tester l'hypothèse que les moyennes des populations sont toutes égales. Dans la plupart des cas, le rejet de l'hypothèse soulève de nouvelles questions : si les moyennes ne sont pas toutes égales, où sont les différences ? Nous étudions ici le cas où l'expérimentateur a formulé certaines questions (formulé certaines hypothèses) à priori.

Supposons, par exemple, qu'un expérimentateur veuille comparer trois traitements pour la culture des betteraves :

(i) Un engrais minéral appliqué en avril avant l'ensemencement ;

(ii) Le même engrais appliqué en décembre avant le labourage ;

(iii) Pas de minéraux.

Les données portent sur la récolte obtenue dans chacune de ces trois conditions. En supposant que l'hypothèse $\mu_1 = \mu_2 = \mu_3$ sera rejetée, l'expérimentateur sait qu'il voudra ensuite tester l'hypothèse :

$$\frac{\mu_1 + \mu_2}{2} = \mu_3,$$

c'est l'hypothèse qu'en moyenne, les minéraux n'ont pas d'effet. Plus généralement, supposons qu'on veuille tester une hypothèse de la forme

$$\mathcal{H}_0 = \text{“}\varphi = \sum_{i=1}^{a} \lambda_i \mu_i = 0\text{”},$$

où λ_i sont des constantes donnés. la fonction linéaire φ sera estimée par

$$\hat{\varphi} = \sum_{i=1}^{n} \lambda_i \bar{X}_i.$$

Puisque $\hat{\varphi}$ est une combinaison liéaire de variables normales indépendantes, elle suit une loi normale de variance

$$Var(\hat{\varphi}) = \sum_{i=1}^{n} \lambda_i^2 Var(\bar{X}_i) = \sum_{i=1}^{n} \lambda_i^2 \frac{\sigma^2}{n_i} = \sigma^2 \sum_{i=1}^{n} \frac{\lambda_i^2}{n_i}.$$

La variance σ^2 sera toujours estimée par

$$s^2 = \frac{SCE}{n-a},$$

et

$$(n-a)\frac{\frac{SCE}{n-a}}{\sigma^2} = \frac{SCE}{\sigma^2} \sim \chi^2(n-a).$$

Étant donné que $\hat{\phi}$ est indépendante de s^2, on en déduit que lorsque $\mathcal{H}_0$ est vraie, le rapport

$$T = \frac{\hat{\varphi}}{s\sqrt{\sum_{i=1}^{a} \frac{\lambda_i^2}{n_i}}} \sim t_{n-a}.$$

La région critique pour tester l'hypothèse

$$\mathcal{H}_0 = \text{“}\varphi = \sum_{i=1}^{a} \lambda_i \mu_i = 0\text{”}$$

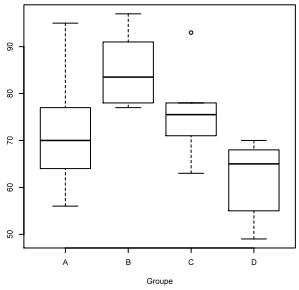
contre une alternative bilatérale est donc

$$|T| > q_{1-\alpha/2}^{t_{n-a}}.$$

13.2.5 Exemple

Nous allons reprendre ici un exemple du livre de Snedecor and Cochran (1989). Pendant leur cuisson les beignets absorbent de la matière grasse en quantité variable. On peut se demander si la quantité absorbée dépend de la matière grasse utilisée ? Pour chacune des quatre matières grasses, on a constitué six fournées de 24 beignets chacune. La mesure est la quantité, en grammes, de matière grasse absorbée, par fournée. On a simplifié les calculs en leur soustrayant 100 g. Les données de ce genre constituent une classification à une seule entrée, ou à une seule voie ou classification simple; on dit aussi à un seul facteur, chaque matière grasse représentant une classe, ou niveau du facteur.

Boxplot. Quantité de matière grasse absorbée, par fournée, en grammes.



Données et ANOVA. Poids de matière grasse absorbée par fournée (diminuée de 100 g)

	Matière grasse (indice i)				
j	**1**	**2**	**3**	**4**	**Tous**
1	64	78	75	55	
2	72	91	93	66	
3	68	97	78	49	
4	77	82	71	64	
5	56	85	63	70	
6	95	77	76	68	
$\sum_j X_{ij}$	432	510	456	372	1 770
$\bar{X}_i$	72	85	76	62	295
$\sum_j X_{ij}^2$	31 994	43 652	35 144	23 402	134 192
$n_i \bar{X}_i^2$	31 104	43 350	34 656	23 064	132 174
$\sum_j X_{ij}^2 - n_i \bar{X}_i^2$	890	302	488	338	2018
d.l.	5	5	5	5	20

RAPPORT DÉTAILLÉ

Groupes	n_i	*Somme*	*Moyenne*	*Variance*
1	6	432	72	178
2	6	510	85	60,4
3	6	456	76	97,6
4	6	372	62	67,6

ANALYSE DE VARIANCE

Source	*S. C.*	*d. l.*	*C.M.*	*F*	*Prob.*	$F_{3,20;0,05}$
Inter groupes	1636,5	3	545,5	5,41	0,0069	3,10
Intra groupes	2018,0	20	100,9			
Total	3654,5	23				

Avant de commencer l'analyse, notons que les quatre totaux de M.G. diffèrent de façon visible : de 372 pour la 4e à 510 pour la 2e. II y a en effet une séparation assez net entre les résultats individuels des matières grasses (4) et (2), 70 est la plus haute valeur donnée par la M.G. (4) tandis que 77 est la plus basse pour la M.G. (2). Pour les autres paires d'échantillons, on observe un certain chevauchement des résultats.

Peut-on penser que les beignets absorbent une quantité de matière grasse qui ne dépendrait pas du type de matière grasse ? Selon le tableau de l'analyse de la variance, puisque $F = 5,41$ et que la valeur critique est 3,10, nous rejetons $\mathcal{H}0$: "$\mu_1 = \mu_2 = \mu_3 = \mu_4''$. La quantité moyenne de graisse absorbée dépend du type de graisse.

Les deux premières matières grasses sont d'origine animale ; les deux dernières sont d'origine végétale. Nous aimerions tester l'hypothèse que les deux types de matière grasse sont absorbées en moyenne de la même façon.

Le contraste s'écrit de la manière suivante :

$$\frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2} \iff \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0.$$

Ainsi, $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = -1, \lambda_4 = -1$. Ce contraste est estimé par

$$\hat{\varphi} = \bar{X}_1 + \bar{X}_2 - \bar{X}_3 - \bar{X}_4 = 19.$$

Sous $\mathcal{H}_0$, $E\hat{\varphi} = \varphi = 0$ et la variance est :

$$Var[\varphi] = \sigma^2 \sum_{i=1}^{a} \frac{\lambda_i^2}{n_i} = \frac{2}{3}\sigma$$

que nous estimons par $\frac{2}{3}\sigma^2 = 67,26667 = 8,201626^2$. Le rapport t est donc

$$\frac{|\hat{\varphi}|}{\sqrt{\widehat{Var[\hat{\varphi}]}}} = 2,316614.$$

Si nous faisons un test à 5%, la valeur critique est 2,0860 et on rejette $\mathcal{H}0$: "$\varphi = 0''$. Il semblerait que plus de matière grasse soit absorbée quand on utilise une graisse animale.

14 ANOVA à deux voies

La généralisation de l'ANOVA à une voie à des plans d'expérience plus complexes est assez intuitive. L'exemple suivant utilise deux facteurs.

Exemple. Le loyer moyen dans une grande ville française, en fonction de deux facteurs : *Date de construction* A et *Nombre de pièces* B est donné par

		$A=1$	2	3	4
		<1981	1981-1970	1991-2001	>2001
$B=1$	1 pièce	509	503	521	795
2	2 pièces	596	661	814	1138
3	3 pièces	684	791	1071	1503
4	4 pièces	808	960	1259	1741
5	5 pièces	1075	1216	1604	2131

L'analyse d'un tel jeu de données à pour objectif d'expliquer et de quantifier l'influence des deux facteurs sur la variable réponse (le loyer). Le plan d'expérience employé est complet, en ce sens que, pour chaque combinaison des deux facteurs, on dispose d'une observation.

14.1 Position du problème

On veut mesurer maintenant le rôle conjoint de deux facteurs A et B sur la variable dépendante (réponse). Trois effets sont à mesurer :

- effet de A,
- effet de B,
- interaction entre A et B.

Les deux premiers seront les effets principaux.

14.1.1 Description des données

La population est notée P, X est la variable d'intérêt de moyenne globale μ. On étudie le rôle de deux facteurs A et B, le facteur A ayant p modalités $(A_1, \cdots A_p)$, le facteur B ayant q modalités $(B_1, \cdots B_q)$.

- Les facteurs A et B définissent $p \times q$ sous population P_{ij}, sur laquelle la sous-variable X_{ij} de X prend les observations $x_{i,j,k}$ (pour $k \leq n_{ij}$) et a pour moyenne μ_{ij}.
- On note $P_{i.}$ les individus correspondants à $A = A_i$, la sous variable $X_{i.}$ de X est observée par la concaténation sur j des $x_{i,j,k}$. La variable $X_{i.}$ a pour moyenne $\mu_{i.}$ et effectif $n_{i.} = \sum_{j=1}^{q} n_{ij}$.
- On note $P_{.j}$ les individus correspondants à $B = B_j$, la sous variable $X_{.j}$ de X est observée par la concaténation sur j des $x_{i,j,k}$. La variable X_j a pour moyenne $\mu_{.j}$ et effectif $n_{j.} = \sum_{i=1}^{p} n_{ij}$.

On suppose que dans chaque sous-population P_{ij}, les observations $x_{i,j,k}$ forme un échantillon E_{ij}.

Note. Pour simplifier l'exposé, dans tout ce qui suit, on considère que le plan d'expériences est équilibré, $card(E_{ij}) = n_{ij} = n$. Cela n'est pas gênant, en effet, en passant par un plan équilibré, on améliore la robustesse du test. D'autre part, le traitement par R saura s'occuper des différences d'effectif s'il y en a.

On définit les moyennes croisées et marginales

$$\bar{X}_{ij} = \frac{1}{n}\sum_{k=1}^{n} x_{i,j,k}, \qquad \bar{X}_{i.} = \frac{1}{q}\sum_{j=1}^{q} \bar{X}_{ij}, \qquad \bar{X}_{.j} = \frac{1}{p}\sum_{i=1}^{p} \bar{X}_{ij}.$$

14.2 Tableau de données

Facile à lire mais encombrant, pratique pour les calculs manuels.

carburant \ aspiration	atmo	turbo
diesel	52 56 (...) 58	68 65 (...) 67
essence	48 49 (...) 67	102 145 (...) 130

$$x_{2,2,2} = puissance_{essence,turbo,2}$$

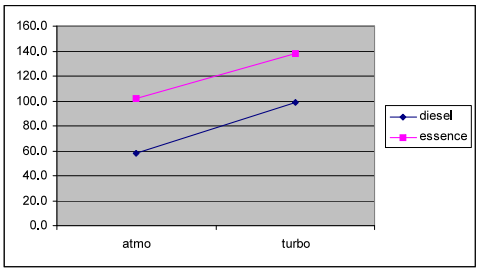
Dans cet exemple, que nous allons filer dans la suite du cours, le tableau n'est pas exhaustif, nous n'avons mis que les trois premières valeurs de chaque groupe.

14.3 Tableau des moyennes

Moyenne puissance	Aspiration atmo	Aspiration turbo	Total
diesel	$\bar{X}_{11} = 58.1$	$\bar{X}_{12} = 98.6$	$\bar{X}_{1.} = 78.35$
essence	$\bar{X}_{21} = 101.6$	$\bar{X}_{22} = 138.4$	$\bar{X}_{2.} = 120$
Total	$\bar{X}_{.1} = 79.85$	$\bar{X}_{.2} = 118.5$	$\bar{X} = 112.1$

14.4 Graphe des interactions

Ce graphe permet de distinguer les interactions lorsque les lignes se croisent.



14.5 Hypothèses statistiques

Ce sont les mêmes que pour l'ANOVA à 1 facteur : normalité de la variable dépendante, indépendance des observations inter et intra groupes, variance homogène dans les groupes.

14.6 Hypothèses soumises au test

Il y en a trois :

$$\begin{cases} H_0 : \mu_{i.} = \mu, \quad \forall i & \text{absence d'effet de } A \\ H_0 : \mu_{.j} = \mu, \quad \forall j & \text{absence d'effet de } B \\ H_0 : \mu_{ij} = \mu, \quad \forall i,j & \text{absence d'effet de l'interaction.} \end{cases}$$

14.7 Équation d'ANOVA

On a la décomposition de la moyenne suivante

$$x_{i,j,k} - \bar{X} = \underbrace{(\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X})}_{\substack{\text{Effet des facteurs} \\ \text{principaux}}} + \underbrace{(\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})}_{\substack{\text{Effet de} \\ \text{l'interacton}}} + \underbrace{(x_{i,j,k} - \bar{X}_{ij})}_{\substack{\text{Erreur} \\ \text{résiduelle}}}.$$

À partir de laquelle on extrait l'équation d'ANOVA

$$\underbrace{SCT}_{\substack{\text{Variablité} \\ \text{totale}}} = \underbrace{SCM_A + SCM_B + SCM_{AB}}_{\substack{\text{Variabilité} \\ \text{expliquée}}} + \underbrace{SCE}_{\substack{\text{Variabilité} \\ \text{résiduelle}}}.$$

Les sommes de carrés sont donc

$$\begin{aligned} SCT &= \sum_{i,j,k} (x_{ijk} - \bar{X})^2, \\ SCM_A &= nq \sum_{i=0}^{p} (\bar{X}_{i.} - \bar{X})^2, \\ SCM_B &= np \sum_{j=0}^{q} (\bar{X}_{.j} - \bar{X})^2, \\ SCE &= \sum_{i,j} \sum_{k} (x_{ijk} - \bar{X}_{ij})^2. \end{aligned}$$

Et pour faire simple

$$SCM_{AB} = SCT - SCM_A - SCM_B - SCE.$$

On calcule les carrés moyens

$$CMM_A = \frac{SCM_A}{p-1}, \qquad \begin{array}{c} CMT = \frac{SCT}{pqn-1}, \\ CMM_B = \frac{SCM_B}{q-1}, \\ CME = \frac{SCE}{pq(n-1)}. \end{array} \qquad CMM_{AB} = \frac{SCM_{AB}}{(p-1)(q-1)},$$

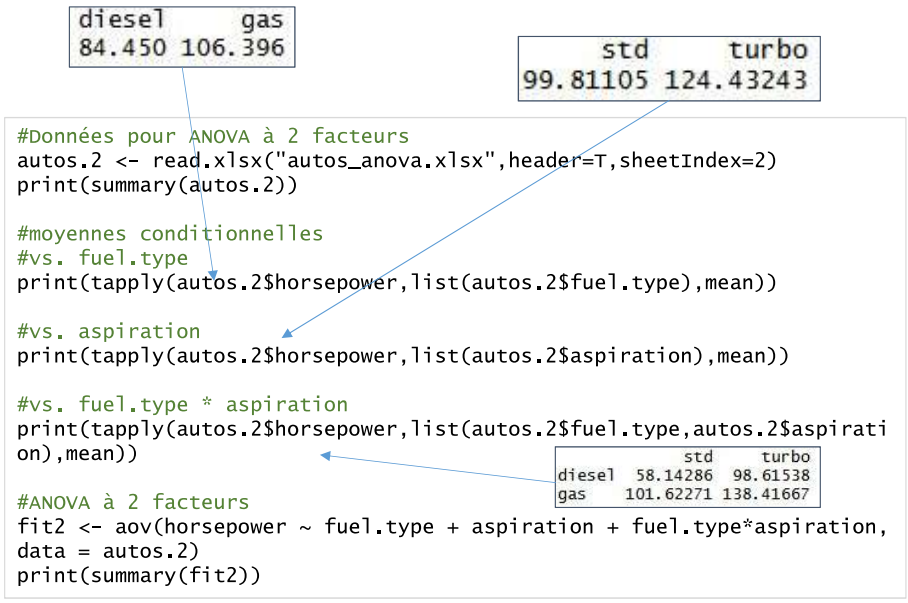
Voici les rapports de carrés moyens pour construire les statistiques de Fisher utilisées pour mettre à jour les effets (principaux et interactions).

- Effet de A :
$$F_A = \frac{CMM_A}{CME}.$$
- Effet de B :
$$F_B = \frac{CMM_B}{CME}.$$
- Effet de l'interaction de A et de B :
$$F_{AB} = \frac{CMM_{AB}}{CME}.$$

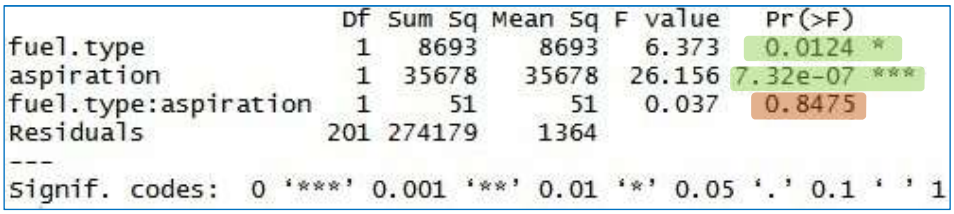
Ces quantités suivent une loi de Fischer, les degrés de libertés sont lus dans les dénominateurs des carrés moyens associés.

14.8 ANOVA à deux facteurs sous R

Puissances des véhicules en fonction du type de carburant (`Fuel-type`) et le mode d'alimentation (`aspiration`).



On obtient le tableau d'ANOVA à deux facteurs suivant

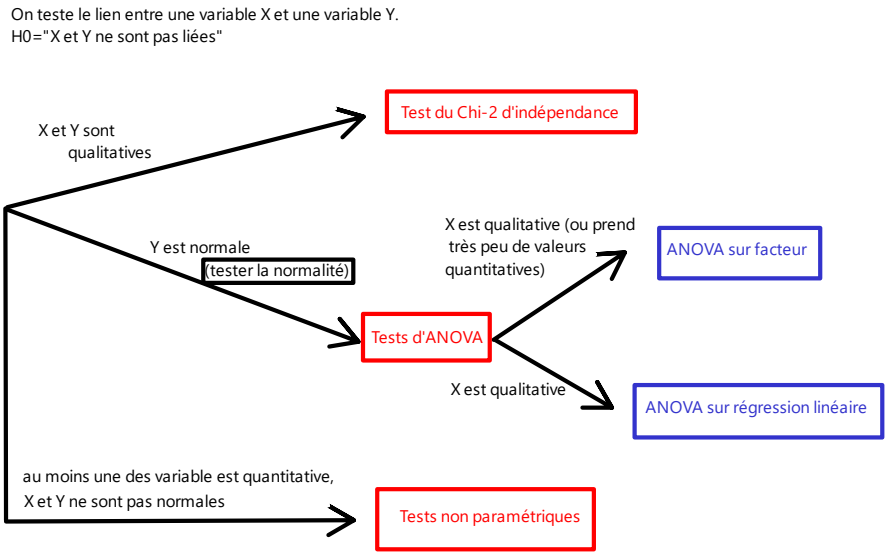


On observe que `Fuel-type` et `aspiration` influent sur la puissance mais pas leur interaction.

Part VI

Tests non paramétriques

Un test est dit non-paramétrique s'il peut être appliqué quelle que soit la distribution des observations. Puisque la distribution (normale ? exponentielle ? uniforme ? autre ?) n'est pas spécifiée, on n'a aucun paramètre à estimer. Il existe une multitude de méthodes et de tests non paramétriques, certains plutôt grossiers, d'autres beaucoup plus raffinés, qui permettent de tester plusieurs types d'hypothèses (équidistribution de deux échantillons, symétrie d'une distribution, indépendance, etc.). Pour se repérer dans l'ensemble des tests que nous avons détaillé, on pourrait utiliser le diagramme suivant.



15 Intervalle pour les quantiles – Exemple introductif

Rappelons la notion de quantile. Soit X une variable alétaoire de type continu dont la densité est $f(x)$ et dont la fonction de répartition est $F(X)$. Soit p une proportion strictement entre 0 et 1. Supposons que $F(x) = p$ admette une unique solution en x. Cette solution est dite : "quantile d'ordre p". Cette racine (ou solution) est notée ξ_p. Ainsi,

$$P(X \leq \xi_p) = F(\xi_p) = p.$$

Par exemple, le quantile d'ordre $1/2$ est la médiane de la distribution et

$$P(X \leq \xi_{1/2}) = F(\xi_{1/2}) = \frac{1}{2}.$$

Nous voudrions obtenir un intervalle de confiance pour ξ_p, le quantile d'ordre p, d'une distribution de type continu ; nous voudrions une approche qui ne dépende pas de la distribution ; comme toujours, nous prenons un échantillon de taille n, $X_1, \cdots, X_n$. Ces observations sont alors triées pour donner une série $Y_1 < Y_2 < \cdots < Y_n$. Cette série est l'ensemble des statistiques d'ordre.

Prenons $Yi < Yj$ et considérons l'événement

$$Y_i < \xi_p < Y_j.$$

Pour que la $i$ème statistique d'ordre soit inférieure à ξ_p il faut qu'il y ait au moins i valeurs observées inférieures à ξ_p. De plus, pour que la $j$ème statistique d'ordre soit supérieure à ξ_p, il faut qu'il y ait moins de j valeurs inférieures à ξ_p. Ainsi, si nous considérons qu'une valeur inférieure à ξ_p est un succès, alors, parmi les n essais indépendants, il doit y avoir entre i et $j-1$ succès pour que l'événement qui nous intéresse se réalise. Donc

$$P(Y_i < \xi_p < Y_j) = \sum_{k=i}^{j-1} \frac{n!}{k!(n-p)!} p^k (1-p)^{n-k}.$$

Ainsi pour des valeurs de i, j et n, cette probabilité peut être calculée. Si

$$P(Y_i < \xi_p < Y_j) = \gamma,$$

cela signifie que la probabilité que l'intervalle déterminé par Y_i et Y_j recouvre la valeur ξ_p est γ. Ainsi,$]Y_i; Y_j[$ est un intervalle de confiance pour ξ_p dont le niveau de confiance est $100\gamma\%$.

Exemple.

On veut un intervalle de confiance pour le troisième quartile en observant un échantillon de taille 10. On a

$$P(Y_4 < \xi_{3/4} < Y_9) = \sum_{k=4}^{9-1} \frac{10!}{k!(10-k)!} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{n-k} \approx 0.9240.$$

Ainsi, en utilisant Y_4 et Y_9, on obtient un intervalle de confiance à $92,40\%$. Les données sont dans le tableau suivant :

Un échantillon de taille 10 d'une population continue

i	1	2	3	4	5	6	7	8	9	10
X_i	10,48	14,43	0,82	4,03	4,74	4,70	4,21	0,53	14,73	0,91
Y_i	0,53	0,82	0,91	4,03	4,21	4,70	4,74	10,48	14,43	14,73

Un intervalle de confiance pour le 3e quartile est donc $]4,03; 14,73[$. Pour cet exemple, les auteurs ont échantillonné une population de loi $\mathcal{E}(5)$. On montre facilement que la valeur théorique de ce quartile est

$$Q3 = 10ln(2) = 6,93$$

16 Test d'équidistribution de deux échantillons

Comme souvent en statistiques, un problème ouvre une bibliothèque de tests. Nous proposons d'en détailler deux, le lecteur étant libre de se cultiver à volonté.

Nous avons beaucoup parlé du test du χ^2 en début de semestre ainsi qu'en début de formation tant pour tester l'adéquation d'une série observée à une série théorique que pour tester l'indépendance (qui revient à comparer une loi jointe observée à une loi jointe produit des marges). Puisque dans ce cas la distribution de la variable observée n'est pas supposée connue, nous pouvons parler de test non paramétrique.

16.1 Test du χ^2... Encore !

Dire que deux variables X et Y sont équidistribuées est équivalent à dire que la distribution d'une variable est indépendante du fait qu'il s'agisse d'un X ou d'un Y.

En découpant l'intervalle $]-\infty, +\infty[$ en k tronçons $I_1, I_2, \cdots I_k$, on peut dénombrer

$$\begin{aligned} N_{1j} &= Card\{i : X_i \in I_j\} = \text{“nombre de } X \text{ dans } I_j \text{ ”}, \\ N_{2j} &= Card\{i : Y_i \in I_j\} = \text{“nombre de } Y \text{ dans } I_j \text{ ”}. \end{aligned}$$

On obtient donc un tableau $2 \times k$ et la suite se fait exactement comme s'il s'agissait d'un test d'indépendance.

16.2 Le test de Wilcoxon

Considérons deux échantillons $X_1, X_2, \cdots, X_{n_X}$ et $Y_1, Y_2, \cdots, Y_{n_Y}$. On veut tester l'hypothèse

$$\mathcal{H}0 : \text{“}F_X = F_Y''.$$

16.2.1 Le tri

Pour appliquer le test de Wilcoxon, on ordonne (disons, de la plus petite à la plus grande) l'ensemble des $n = n_X + n_Y$ observations. On obtient alors un “mot” formé de $n_X + n_Y$ lettres (n_X fois la lettre X et n_Y fois la lettre Y).

Exemple.

Si $X = (17.1, 14.5, 20.3, 8.3)$ et $Y = (5.2, 10.3, 12.4)$, on obtient le mot

$$YXYYXXX$$

(la plus petite des 7 observations est un Y, la deuxième est un X, $\cdots$, les 3 plus grandes sont des X).

16.2.2 La variable

La variable du test de Wilcoxon est

$$W = \sum_{i=1}^{n_X} R_i,$$

où R_i désigne le rang de l'observation X_i parmi les $n = n_X + n_Y$ observations en ordre croissant.

$\mathcal{H}_0$ sera rejetée si W est significativement grand ou significativement petit.

Déterminons $\mu_W = E[W|\mathcal{H}0]$ et $\sigma_W^2 = Var[W|\mathcal{H}0]$. On voit que, sous $\mathcal{H}_0$, chaque R_i est de loi uniforme $\mathcal{U}(1, n)$. On a donc

$$E[R_i|\mathcal{H}_0] = \frac{n+1}{2}, \qquad \text{et} \qquad Var[Ri|\mathcal{H}0] = \frac{n^2-1}{12}.$$

Il suit que

$$\mu_W = \mathrm{E}[W|H_0] = \sum_{i=1}^{n_X} \mathrm{E}[R_i|H_0] = \frac{n_X(n+1)}{2}.$$

Aussi,

$$\sigma_W^2 = \mathrm{Var}\left[\sum_{i=1}^{n_X} R_i|H_0\right] = \sum_{i=1}^{n_X}\sum_{j=1}^{n_X} \mathrm{cov}\big[R_i,R_j|H_0\big]$$

$$\sigma_W^2 = \sum_{i=1}^{n_X} \mathrm{Var}[R_i|H_0] + \sum_{i=1}^{n_X}\sum_{\substack{j=1\\ i\neq j}}^{n_X} \mathrm{cov}\big[R_i,R_j|H_0\big]$$

$$\sigma_W^2 = \frac{n_X(n^2-1)}{12} + (n_X^2 - n_X)c.$$

où c désigne la covariance entre R_i et R_j quand $i \neq j$. Une astuce classique pour déterminer c est d'imaginer qu'on n'a que des X (alors $n_Y = 0$). La valeur de W sera alors nécessairement

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$

et la variance de W sera 0. En appliquant cette formule dans la dernière formule de la variance de W, on trouve

$$0 = \frac{n(n^2-1)}{12} + (n^2-n)c \qquad \Longleftrightarrow \qquad n(n-1)\left(\frac{n+1}{12}+c\right) = 0 \qquad \Longleftrightarrow \qquad c = -\frac{n+1}{12}.$$

Réintroduisant, en toute généralité, cette valeur de c dans dans la formule de la variance de W, on trouve

$$\begin{aligned}\sigma_W^2 &= \frac{n_X(n^2-1)}{12} - \frac{n_X(n_X-1)(n+1)}{12} \\ &= \frac{n_X(n+1)}{12}((n-1)-(n_X-1)) \\ &= \frac{n_X n_Y(n+1)}{12}.\end{aligned}$$

En vertue du Théorème Central Limit, si n_X et n_Y sont tous deux grands, la statistique

$$W = \sum_{i=1}^{n_x} Ri$$

suivra asymptotiquement une loi normale

$$\mathcal{N}\left(\frac{n_X(n+1)}{2}; \sqrt{\frac{n_X n_Y(n+1)}{12}}\right).$$

16.2.3 Conclusion du test

L'hypothèse $\mathcal{H}_0$ sera donc rejetée au seuil α si

$$\left|W - \frac{n_X(n+1)}{2}\right| > q_{\alpha/2}\sqrt{\frac{n_X n_Y(n+1)}{12}},$$

où $q_{\alpha/2}$ est le quantile théorique de la loi normale $\mathcal{N}(0;1)$:

$$P\big(\mathcal{N}(0;1) > q_{\alpha/2}\big) = \frac{\alpha}{2}.$$

Remarque :

Le test de Wilcoxon est excellent pour détecter les différences de position (médiane, moyenne). Pour détecter les différences de dispersion, il ne vaut pas grand chose. Si, par exemple, on observe le mot

$$XXXXXXXYYYYYYYYYXXXXXXX,$$

l'hypothèse d'équidistribution est évidemment fausse (tous les Y sont au centre et les X sont aux deux bouts) ; le test de Wilcoxon, pourtant, donnera une valeur de W tout-à-fait compatible avec l'hypothèse d'équidistribution.

17 Test d'indépendance–Test de Spearman

Nous avons déjà été amené à tester l'indépendance de deux variables avec le test du χ^2, en particulier pour les couples de variables qualitatives. Ce test reste le plus usité. On pourra également parler du test de Spearman basé sur la corrélation des rangs.

On posera pour hypothèse de test

$$\mathcal{H}_0 = \text{"les variables ne sont pas liées"} = \text{"les variables sont non corrélées"}.$$

On pourrait considérer un test basé sur le coefficient de corrélation échantillonnal r obtenu des n couples (X_i, Y_i) d'un couple de variables (X, Y). On a pu observer que la nullité de ce coefficient de corrélation n'entrainait pas à coup sûr l'indépendance des variables. En effet, le coefficient de corrélation r est pleinement aveugle face aux relations non linéaires du types $Y = X^2$.

Le coefficient de corrélation de rangs (noté R^*) est celui qu'on obtient en remplaçant simplement les observations X_i et Y_i par leurs rangs $R_{X(i)}$ et $R_{Y(i)}$. $R_{X(i)}$ est le rang obtenu par X_i dans l'échantillon X **ordonné** ; de même, $R_{Y(i)}$ est le rang de Y_i parmi les n valeurs de Y observées dans l'ordre. Le coefficient de corrélation de rangs est

$$R^* = corr(R_X, R_Y),$$

et se calcule par la formule

$$R^* = \frac{\frac{1}{n}\sum_{i=1}^{n} R_X(i)R_Y(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_X(i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_Y(i)\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} R_X^2(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_X(i)\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} R_Y^2(i) - \left(\frac{1}{n}\sum_{i=1}^{n} R_Y(i)\right)^2\right)}}$$

$$= \frac{n\sum_{i=1}^{n} R_X(i)R_Y(i) - \left(\sum_{i=1}^{n} R_X(i)\right)\left(\sum_{i=1}^{n} R_Y(i)\right)}{\sqrt{\left(n\sum_{i=1}^{n} R_X^2(i) - \left(\sum_{i=1}^{n} R_X(i)\right)^2\right)\left(n\sum_{i=1}^{n} R_Y^2(i) - \left(\sum_{i=1}^{n} R_Y(i)\right)^2\right)}}.$$

On remarque toutefois que puisque les $R_X(i)$ (et les $R_Y(i)$) ne sont qu'une permutation des entiers de 1 à n, on a

$$\begin{aligned}
\sum_{i=1}^{n} R_X(i) &= \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \\
\sum_{i=1}^{n} R_X^2(i) &= \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \\
n\sum_{i=1}^{n} R_X^2(i) - \left(\sum_{i=1}^{n} R_X(i)\right)^2 &= \frac{n^2(n+1)(2n+1)}{6} - \frac{n^2(n+1)^2}{4} \\
&= n^2(n+1)\left(\frac{2n+1}{6} - \frac{n+1}{4}\right) \\
&= \frac{n^2(n^2-1)}{12}.
\end{aligned}$$

Le coefficient de corrélation de rangs devient donc

$$\begin{aligned} R^* &= \frac{n\sum_{i=1}^{n} R_X(i)R_Y(i) - \frac{n^2(n+1)^2}{4}}{\frac{n^2(n^2-1)}{12}} \\ &= \frac{12\sum_{i=1}^{n} R_X(i)R_Y(i)}{n(n^2-1)} - \frac{3(n+1)}{n-1}. \end{aligned}$$

Posant $S^* = \sum_{i=1}^{n} R_X(i)R_Y(i)$, on observe que R^* est fonction linéaire de S^*. Noter que $R^* = AS^* + b$ où

$$a = \frac{12}{n(n^2-1)} \text{ et } b = -\frac{3(n+1)}{n-1}.$$

Il suffit donc de connaître la distribution de S^* pour connaître celle de R^*. En fait, on a

$$E[R^*|\mathcal{H}_0] = aE[S^*|\mathcal{H}_0] + b$$

et

$$Var[R^*|\mathcal{H}_0] = a^2 Var[S^*|\mathcal{H}_0].$$

En cette fin de chapitre, on se passera du détail des calculs. Si $\mathcal{H}_0$ est vraie, et si n est grand, La statistique S^* suivra approximativement une loi normale

$$\mathcal{N}\left(\frac{n^2(n+1)^2}{4}; \sqrt{\frac{n^2(n+1)(n^2-1)}{144}}\right).$$

On en déduit alors que R^* suit une loi normale (bien plus jolie)

$$\mathcal{N}\left(0, \sqrt{\frac{1}{n-1}}\right)$$

L'hypothèse $\mathcal{H}_0$ sera donc rejetée au seuil α si

$$|R^*| > q_{\alpha/2}\sqrt{\frac{1}{n-1}},$$

où $q_{\alpha/2}$ est le quantile théorique de la loi normale $\mathcal{N}(0;1)$:

$$P\big(\mathcal{N}(0;1) > q_{\alpha/2}\big) = \frac{\alpha}{2}.$$