JG

Deuxième année de Licence MIASHS

Statistique mathématique 1[1]

Julien GREPAT[2]

Contents

[1]Reproduction et diffusion interdite sans l'accord de l'auteur

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

Part I
Éléments de probabilité

1 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

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

1.1 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

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

1.2 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

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

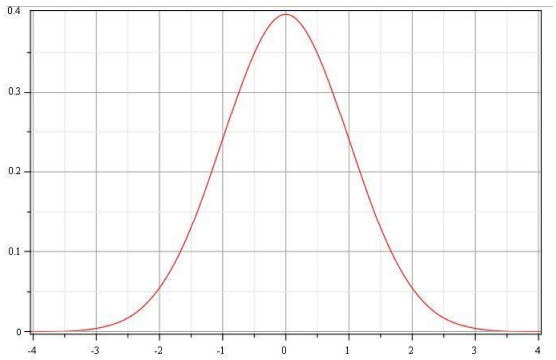
1.3 Loi normale

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. On citera la planche de Galton, renommé *The Wall* par TF1. Ce type de loi admet une densité en forme de cloche.

1.3.1 Loi normale centrée réduite

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

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



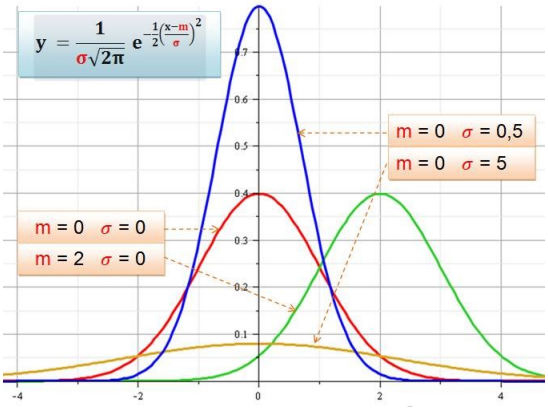
Son espérance vaut 0 et son écart-type 1.

1.3.2 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.



1.4 Combinaison linéaire de loi normales indépendanes

Theorem 1.1 *Soit X suivant une loi normale d'espérance m_X, et d'écart-type σ_X, et Y suivant une loi normale d'espérance m_Y, et d'écart-type σ_Y. En supposant les variables X et Y indépendantes, on peut affirmer que $X+Y$ suit une loi normale d'espérance $m = m_X + m_Y$, et décart-type $\sigma = \sqrt{\sigma_X^2 + \sigma_Y^2}$.*

Notons que m s'obtient par linéarité de l'espérance et que, puisque X_1 et X_2 sont indépendantes, $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$, et σ suit. Ce théorème s'étend naturellement à toute combianaison linéaire de lois normales indépendantes.

Corollary 1.2 *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.

1.5 Exemple fondamental

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

Si on veut le diamètre moyen de ces N billes, on somme les valeurs aléatoires qu'on divise par N :

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

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.

1.6 Intervalles remarquables

On aura pour ordre d'idée les intervalles remarquables :

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

2 Lois des statistiques

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

2.1 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 $\chi^2(n)$.

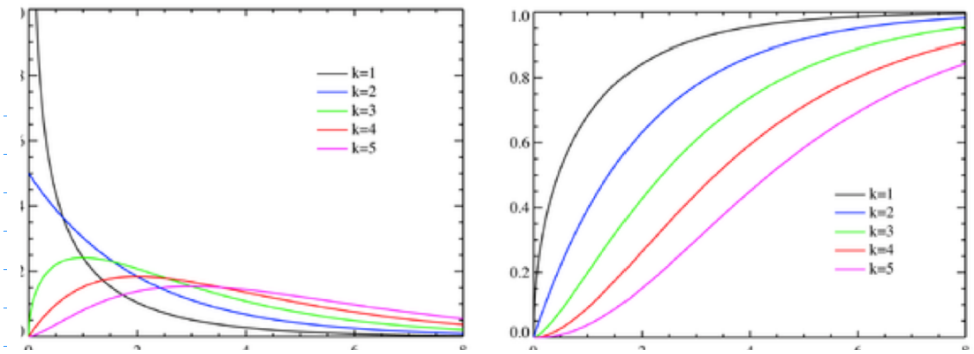
On a alors

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

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

2.2 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

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

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

On a alors

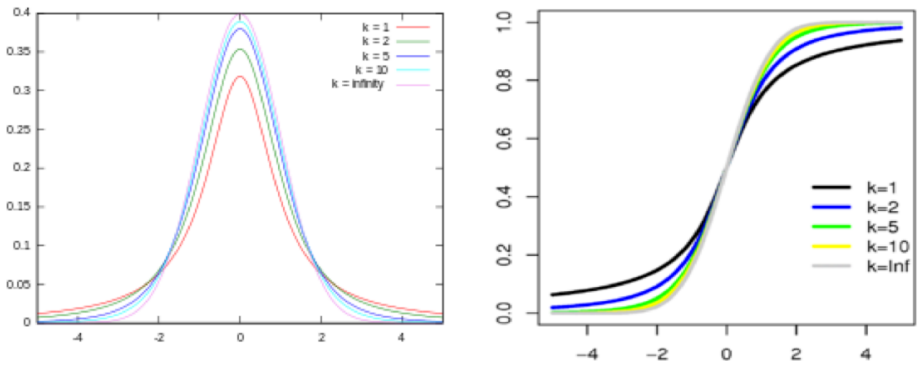
$$EX = 0, \qquad VarX = \frac{k}{k-2},.$$

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 :



2.3 Loi de Fisher

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

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

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}.

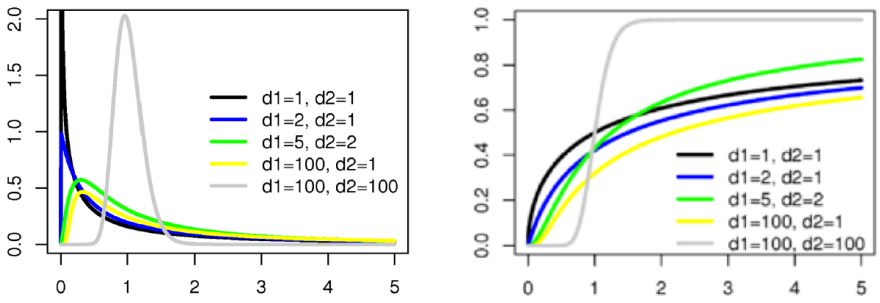
On a alors

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

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

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

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 2.2

- *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* $X^2 \sim \mathcal{F}_{1,\infty}$.

3 Théorèmes 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).

3.1 Convergence en loi

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

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

3.2 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:

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

3.3 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 3.1 *Soit (X_n) une suite de variables aléatoires i.i.d.. Supposons que pour tout $i \in \mathbb{N}$, 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 :*

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

3.4 Approximation d'une loi binomiale par une loi normale

Theorem 3.2 *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

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

3.5 Autres théorèmes limites

On a les limites suivantes.

Proposition 3.3

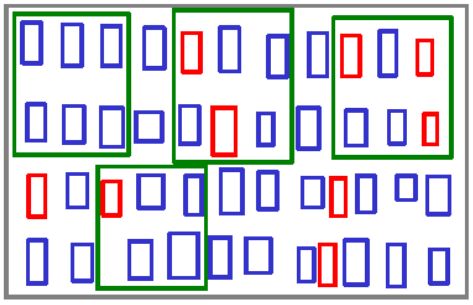
- *Si* $X \sim \mathcal{F}_{d_1;d_2}$ *alors*
$$lim_{d2\to\infty} d_1 X \to \chi^2(d_1).$$
- *La loi de Student converge vers la loi normale centrée réduite :*
$$t_k \to \mathcal{N}(0;1).$$

Part II

Méthodes de statistiques inférentielles

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.

4 Estimation



population totale

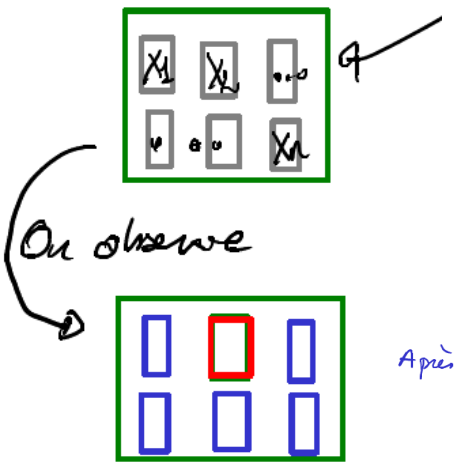
Problème : on ne peut pas tester toute une production

défectueux

Le nombre de défectueux est fluctuant en fonction du choix de l'échantillon aléatoire de taille n

4.1 Échantillon

On considère une suite de variables aléatoires $(X_i)_{i\in\mathbb{N}}$ identiques et indépendantes. Un échantillon de taille d est l'ensemble des variables $(X_1, \cdots X_d)$. Une réalisation de cet échantillon est l'observation $(x_1, \cdots x_d)$.



Echantillon aléatoire(pas encore observé) de taille n.

$(X_1, \cdots X_n)$représente le caractère à observer sur les individus.

On les suppose aléatoires, identiques et indépendantes.

Après observation, plus d'aléa. On travaillera sous l'hypothèse $x_1=0, x_2=1, x_3=0 \cdots x_n=0$.

On travaillera, en toute généralité, sous la condition $(X_1 = x_1, \cdots X_d = x_d)$.

4.2 Estimation

On cherche à connaître un paramètre θ qui dépend de la loi de X (par exemple son espérance ou sa variance). On réplique n fois X de manière indépendante $(X_1, \cdots X_n)$. On évalue alors θ par $\hat{\theta} = \hat{\theta}_n$ grâce aux réalisations possibles $(x_1, \cdots, x_n)$ de l'échantillon à n éléments. La valeur $\hat{\theta}_n$ est nommée estimé ou estimation ponctuelle.

Dans notre problème de qualité, où on compte les défectueux, la probabilité p d'être défectueux est inconnue. Les $X_i \sim \mathcal{B}(p)$ i.i.d.
p est "intuitivement" estimé par
$$P_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$ C'est l'estimateur de P.
P_n est une variable aléatoire. Sur l'échantillon choisi, on évalue P_n en remplaçant les X_i par les observations x_i
$$\hat{P}_n = \frac{\sum_{i=1}^{n} x_i}{n}.$$
C'est une estimation ponctuelle.

On prendra par exemple $\hat{\theta}_n = \sum x_i/n$ pour estimer la moyenne de X, ou $\hat{\sigma} = \sum(x_i - \hat{\theta}_n)^2/(n-1)$ pour estimer la variance.

Pour récapituler, le paramètre réel constant inconnu est θ. Il est approché par l'estimateur $\theta_n = f(X_1, \cdots X_n)$, variable aléatoire, estimé grâce au relevé statistique par $\hat{\theta} = \hat{\theta}_n = f(x_1, \cdots x_n)$.

On note parfois

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

On aura par exemple

- 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 Estimer la variance

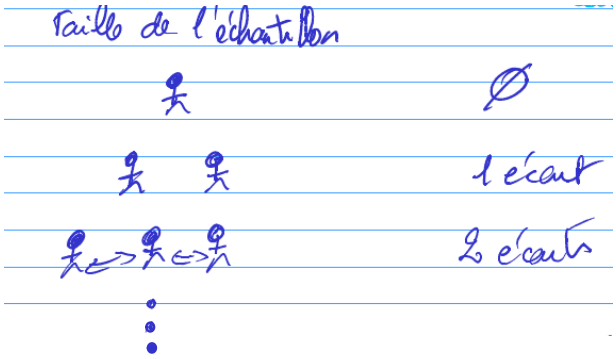
- En probabilité, nous avons défini la variance de la façon suivante,

$$Var(X) = \sum_i (x_i - EX)^2 P(X = x_i).$$

- En Introduction à la Stat (S1), nous avons posé la variance sur une **population entière** :

$$Var(x) = \frac{1}{N} \sum_i (x_i - \bar{x})^2.$$

- Lorsqu'on estime la variance sur un échantillon, on a un *problème de biais*. Cette notion sera correctement développée en S5. Mais nous pouvons comprendre le problème. Il s'agit de sommer les écarts au carré et de les diviser 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

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

5 Estimation par intervalle de confiance

Imaginons que l'on souhaite estimer la proportion de pièces défectueuses dans un grand stock (retour à notre exemple). On prend un échantillon, de taille relativement conséquente, et on calcule, sur

cet échantillon, la proportion de pièces défectueuses. On peut se demander si l'échantillon est représentatif. Et si on change d'échantillon de même taille, quelles valeurs peut-on trouver et quelle est la probabilité d'avoir une valeur complètement différente des autres ? Enfin, quelle erreur peut-on rencontrer sur notre estimation de la proportion dans le stock entier ? Toutes les réponses sont dans le problème suivant !

On souhaite évaluer la proportion de pièces défectueuses dans un stock (de grande taille) suite à un incident. On note p la probabilité pour une pièce d'être défectueuse.

On se donne un échantillon de n pièces.

- **Loi X qui compte le nombre de pièces défectueuses dans le lot de n pièces**

  On considère un échantillon de n pièces, on observe donc $X_1, \cdots, X_n$ des variables indépendantes et identiques à 2 issues : la réussite (la pièce est défecteuse) avec probabilité p et l'échec. Pour compter, on attribuera donc à X_i la valeur 1 en cas de réussite et 0 sinon.

  La variale X est la somme des X_i,

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

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

- **On pose $f_n = X/n$.**

  La variable f_n est l'estimateur de la fréquence de pièces défectueuses.

- **Approximation de f_n**

  Pour n assez grand, la variable X peut être approchée par une loi $\mathcal{N}(m, \sigma)$ en vertu du théorème central limit.

  On en déduit donc que f_n, pour n assez grand, peut être approchée par une loi $\mathcal{N}(m_n, \sigma_n)$. On prendra

  $$m_n = Ef_n = E[X/n] = \frac{EX}{n} = \frac{np}{n} = p,$$

  et

  $$\sigma_n = \sqrt{Var(f_n)} = \sqrt{Var(X/n)} = \sqrt{\frac{1}{n^2} Var(X)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}.$$

- **Risque** 5%

  - On cherche a tel que

    $$P(m_n - a\sigma_n \leq f_n \leq m_n + a\sigma_n) = 0.95$$

    On prendra $a = 1,96$.
  - Il y a donc un risque de se tromper de probabilité 0,05 si on affirme que

    $$p - 1,96\sigma_n \leq f_n \leq p + 1,96\sigma_n.$$

– En faisant l'opération $-f_n + p$, on déduit qu'au risque 0,05

$$f_n - 1,96\sigma_n \leq p \leq f_n + 1,96\sigma_n.$$

- **Estimation ponctuelle**

  On effectue un test sur 100 pièces. Il en résulte que 10 sont défectueuses.

  – l'estimation ponctuelle de la fréquence f_n de pièces défectueuses dans les 100 pièces prélevées est donnée par

  $$\hat{f}_n = 10/100 = 0,1.$$

  – On considère qu'une bonne approximation de σ_n est

  $$\sigma = \frac{\sqrt{\hat{f}_n(1-\hat{f}_n)}}{\sqrt{n}} = 0,06.$$

  Démonstration. On rappelle que puisque

  $$(\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b}) = a - b,$$

  on a, pour tout $a, b \geq \alpha > 0$,

  $$|\sqrt{a} - \sqrt{b}| \leq K_\alpha |a - b|.$$

  Il vient que

  $$\begin{aligned}
  |\sigma_n - \sigma| &= \frac{1}{\sqrt{n}} |\sqrt{p(1-p)} - \sqrt{\hat{f}_n(1-\hat{f}_n)}| \\
  &\leq \frac{K_\alpha}{\sqrt{n}} |p(1-p) - \hat{f}_n(1-\hat{f}_n)| \\
  &\leq \frac{K_\alpha}{\sqrt{n}} |p(1-p) - p(1-\hat{f}_n) + p(1-\hat{f}_n) - \hat{f}_n(1-\hat{f}_n)| \\
  &\leq \frac{K_\alpha}{\sqrt{n}} \big(|p(1-p) - p(1-\hat{f}_n)| + |p(1-\hat{f}_n) - \hat{f}_n(1-\hat{f}_n)|\big) \qquad \text{par inégalité triangualaire} \\
  &\leq \frac{K_\alpha}{\sqrt{n}} \big(p|(1-p) - (1-\hat{f}_n)| + |p - \hat{f}_n|(1-\hat{f}_n)\big) \\
  &\leq \frac{K_\alpha}{\sqrt{n}} (|\hat{f}_n - p| + |p - \hat{f}_n|) \qquad \text{car } p \leq 1, \hat{f}_n \geq 0.
  \end{aligned}$$

  On obtient le résultat en supposant que $\hat{f}_n$ est proche de p et n grand.

- **Intervalle de p avec un risque de 5%**

  On a donc, au risque 5%,

  $$\hat{f}_n - 1,96\sigma \leq p \leq \hat{f}_n + 1,96\sigma,$$

  soit

  $$0,1 - 1,96 \times 0,03 \leq p \leq 0,1 + 1,96 \times 0,03.$$

  On conclut qu'avec un risque de 5%, la proportion p de pièces défectueuses est dans l'intervalle $[0,0412; 0,1588]$.

6 Intervalle de fluctuation

Nous allons inverser la démarche précédente, supposer connue la proportion p. On peut prendre l'exemple d'une pièce jetée à pile ou face. La proportion de *Pile* est de 0,5.

- On va faire 100 lancés,
- On utilise la construction précédente pour dire que, pour 95% des tirages de 100, la proportion de *Pile* est dans l'intervalle

$$\left[0,5-2\times\sqrt{\frac{0,5\times(1-0,5)}{100}};0,5+2\times\sqrt{\frac{0,5\times(1-0,5)}{100}}\right]=[0,4;0,6].$$

- Si au terme des 100 lancers, on obtient 62 *Pile*, on en déduira qu'au risque 5%, la pièce n'est pas équilibrée.
- Un tirage de 58 *Pile* ne validera cependant pas l'équilibre de la pièce. En effet, la proportion estimée 0,58 est aussi dans l'intervalle de fluctuation d'une pièce non équilibrée de fréquence de *Pile* 0,6, à savoir

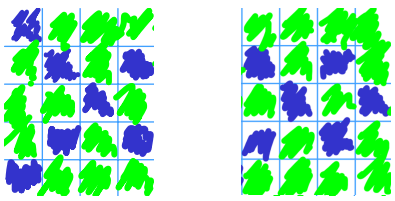
$$\left[0,6-2\times\sqrt{\frac{0,6\times(1-0,6)}{100}};0,6+2\times\sqrt{\frac{0,6\times(1-0,6)}{100}}\right]=[0,5;0,7].$$

  En fait, un tirage de 58 *Pile* ne nous dit juste que rien ne s'oppose au fait que la pièce soit équilibrée.

7 Introduction aux tests statistiques

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 $\mathcal{H}_0 =$ "$\theta \in \Theta_0$" et $\mathcal{H}_1 =$ "$\theta \in \Theta_1$" ? Ici $\Theta_0, \Theta_1 \subset \mathbb{R}$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

Exemple. On peut se poser la question "Est-ce que les deux parties de dallage suivantes sont issues du même dallage" ?



Des éléments peuvent nous aider : l'adéquation des couleurs, la proportion proche de bleus.

On pourra tester par exemple $\mathcal{H}_0$: "les morceaux sont issus du même dallage" contre $\mathcal{H}_1$: "Les deux morceaux sont issus de deux dallages différents". On notera que les hypothèses ne sont pas nécessairement le contraire l'une de l'autre.

7.1 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	OK

On notera α la probabilité d'erreur de première espèce, appelée parfois seuil. On prendra en général pour α les valeurs 0.01, 0.05 ou 0.1. Il est à noter que mécaniquement, si l'erreur de première espèce baisse, l'erreur de seconde espèce augmente.

7.2 La variable du test (la statistique du test)

Prenons la proportion de bleus dans les deux échantillons dans l'idée de vérifier si les deux morceaux ont des proportions proches.

Globalement, il y a une proportion de $p = 13/40$ bleus. C'est un paramètre de notre test que nous venons d'estimer au plus simple par une estimation ponctuelle.

Le nombre de carreaux bleus dans le tirage de 20 carreaux (non nécessairement consécutifs, mais ce n'est pas grave, ce n'est qu'un exemple illustratif) est une variable T qui suit une loi $\mathcal{B}(20, p = 13/40)$. C'est la statistique du test. On rappelle qu'en moyenne, sur un tel tirage il y a 6,5 bleus.

7.3 Zone de rejet

De notre hypothèse nulle $\mathcal{H}_0$ on peut déduire une zone de rejet W, zone où l'estimation $\hat{\theta}$ ne satisfaira plus $\mathcal{H}_0$. Dans notre exemple, on prendra $W_k = \{x \in \mathbb{R} : x \leq 6,5 - k, x \geq 6,5 + k\}$. Il reste à calculer k.

On observe avec les quantiles de $X \backsim \mathcal{B}(20, p)$ que

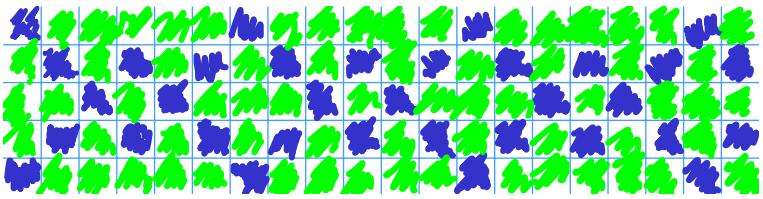
$$P(3 \leq X \leq 11) \simeq 0,95.$$

On ne peut donc pas exclure que les échantillons soient issus du même dallage s'ils sont composés d'entre 3 et 11 bleus en acceptant un risque (de se tromper), un doute de probabilité 5% (le seuil).

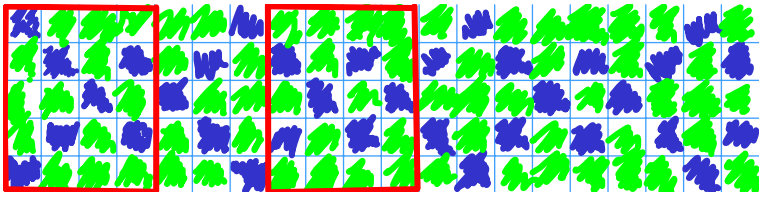
7.4 Décision

On observe $\hat{\theta}$. Si $\hat{\theta} \in W$, on rejette $\mathcal{H}_0$ est on prendra pour vraie $\mathcal{H}_1$ avec un risque de se tromper non négligeable. Si $\hat{\theta} \notin W$, rien ne s'oppose à garder $\mathcal{H}_0$ pour vraie, avec un risque de se tromper α.

Ici, $\hat{\theta}_1 = 7$ et $\hat{\theta}_2 = 6$. On ne rejette pas, au seuil 5%, leur appartenance au même dallage. En effet :



et on reconnaît



7.5 Discussion

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

- Plus l'échantillon est petit, plus le test est laxiste, mais cela peut-être l'inverse.
- 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).
- Le test est un peu pauvre, on ne regarde que le paramètre proportion de bleu, et rien sur le motif par exemple.
- 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...)

Enfin, en toute rigueur mathématique, le non rejet de l'hypothèse nulle ne vaut pas acceptation, mais nous permet juste de dire que rien ne s'oppose à $\mathcal{H}_0$. On renvoie à la discussion dans l'intervalle de fluctuation.

Part III
Régression linéaire.

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

8 Régression linéaire – Statistiques descriptives

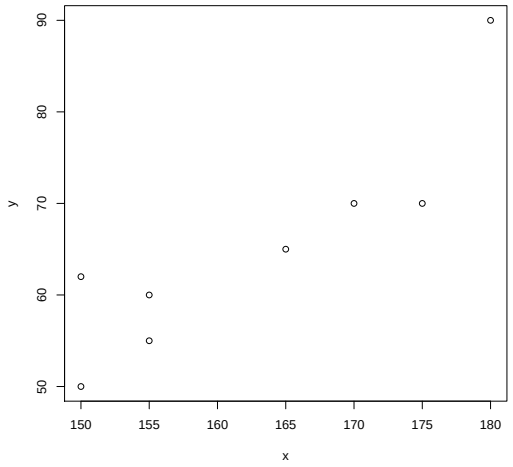
Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

8.1 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 8.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*



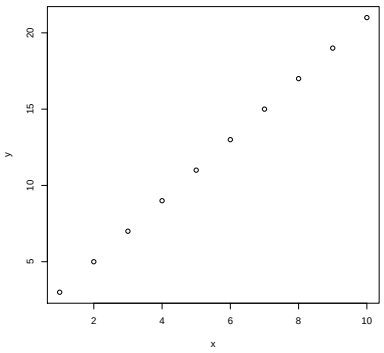
Definition 8.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

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

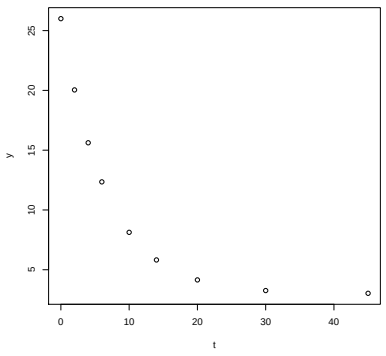
8.2 Forme du nuage de points

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

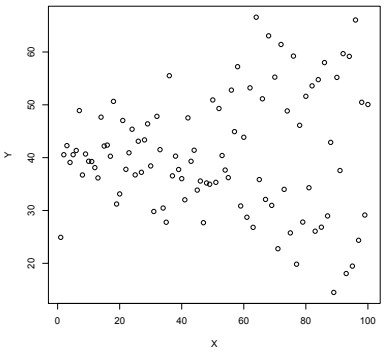
(i) forme allongée et rectiligne : les points sont plus ou moins alignés



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



(iii) forme quelconque



8.3 Ajustement affine (droite de régression linéaire)

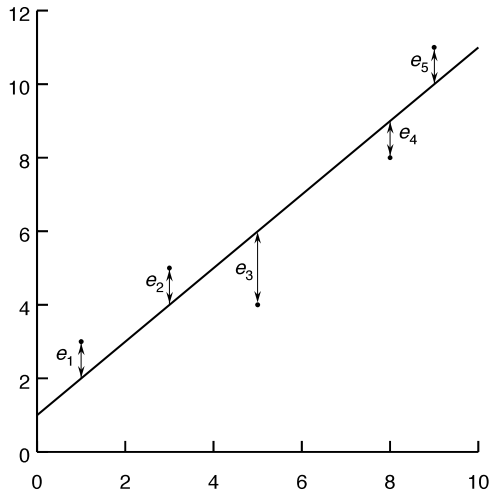
On s'intéresse plus particulièrement au premier cas 8.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

$$y = ax + b$$

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.

8.3.1 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 $D : y = ax + b$ 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 et on aura

$$y_i = ax_i + b + e_i,$$

en faisant appel au résidu e_i. On note

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

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 8.2.3.

On a

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

Preuve. On suppose qu'il n'y a pas d'erreur systématique, sinon on pourrait souhaiter translater ou modifier la pente de la droite. On suppose donc que l'erreur moyenne sur nos observations est nulle

$$\frac{1}{N}\sum_{i=1}^{N} e_i = 0$$

Il suit que

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N}(ax_i + b + e_i) = a\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N} b + \frac{1}{N}\sum_{i=1}^{N} e_i = a\bar{x} + b$$

Il est donc naturel de supposer que toute droite de régression passe par le centre de gravité.

Puisque

$$b = \bar{y} - a\bar{x} \qquad \Longleftrightarrow \qquad y_i = ax_i + \bar{y} - a\bar{x} \qquad \Longleftrightarrow \qquad y_i = a(x_i - \bar{x}) + \bar{y}.$$

On pose la somme des carrés des résidus

$$M(a) = \sum_{i=1}^{N}(y_i - (a(x_i - \bar{x}) + \bar{y}))^2 = \sum_{i=1}^{N}(y_i - \bar{y} - a(x_i - \bar{x}))^2.$$

Nous cherchons donc la valeur de a qui minimise cette quantité. Dérivons la en a

$$M'(a) = \sum_{i=1}^{N} -2(x_i - \bar{x})(y_i - \bar{y} - a(x_i - \bar{x})),$$

et résolvons

$$\begin{aligned}
M'(a) &= 0 \\
\Longleftrightarrow \qquad \sum_{i=1}^{N} -2(x_i - \bar{x})(y_i - \bar{y} - a(x_i - \bar{x})) &= 0 \\
\Longleftrightarrow \qquad \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y} - a(x_i - \bar{x})) &= 0 \\
\Longleftrightarrow \qquad \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) - a\sum_{i=1}^{N}(x_i - \bar{x})^2 &= 0.
\end{aligned}$$

Ce qui équivaut à

$$a = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

Ainsi la quantité $M(a)$ atteint un extremum en

$$a = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}.$$

Il reste idéalement à vérifier que c'est un minimum. On peut voir que la fonction est convexe en calculant la dérivée seconde de M

$$M''(a) = 2\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

qui est positive.

□

8.3.2 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 8.2.3. 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 linéaire r :

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

Proposition 8.3 *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 $r < -0.9$.

9 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 $Y = aX + b + \epsilon$ ou, avec des notations plus fréquemment utilisées,

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

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 les valeurs prédites

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

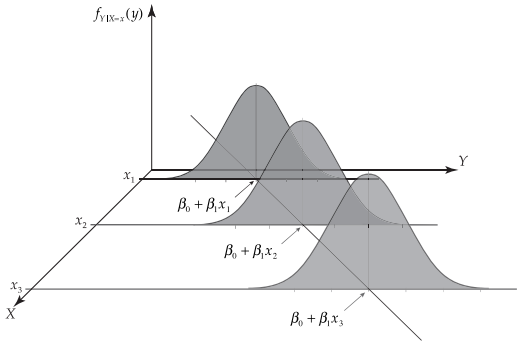
et $\hat{Y}_i$ est la valeur associée à X_i par la droite de régression (dite empirique). Il vient que

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

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

9.1 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



Note. Les hypothèses du modèle montrent que

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

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

9.2 Équation de la variance

D'après les hypothèses précédentes, il vient que

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

grâce à l'éxogénéité. Il suit

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

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 $\sum_i (\hat{Y}_i - \bar{Y})^2$ 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[3].

On a donc

$$SCT = SCM + SCE. \tag{9.1}$$

9.3 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.

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

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

9.4 Distribution de β_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 9.1 *Sous les hypothèses du modèle de régression linéaire simple, $\hat{\beta}_1$ suit une loi normale d'espérance β_1 et de variance*

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

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

Lemma 9.2 *On pose*

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

Nous avons

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

[3]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

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

$$\begin{aligned}\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} Y_i - \bar{Y}\sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (9.2)\end{aligned}$$

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 (9.2), 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 9.1 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 9.1

□

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 9.3 *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 Tests sur la pente de la droite

Pour faire simple, les tests F de Fisher 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).

10.1 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 9.3 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 $\hat{\beta}_1/s_{\beta_1}$. On rejettera $\mathcal{H}_0$ au seuil α si

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

10.2 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 (9.1). 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.

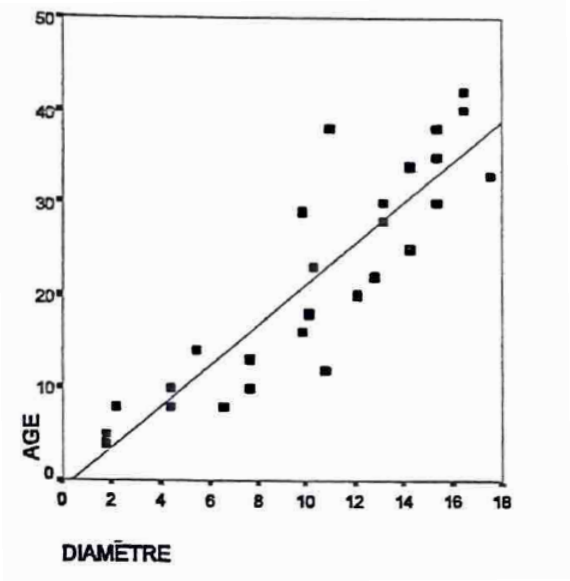
Dans le cadre d'une régression multiple (sur plusieurs variables explicatives), le test de Fischer teste l'effet global des variables sur la variable Y, les tests de Student testent l'effet de chaque variable explicative sur Y.

11 Exemple : le cas de la régression linéaire simple

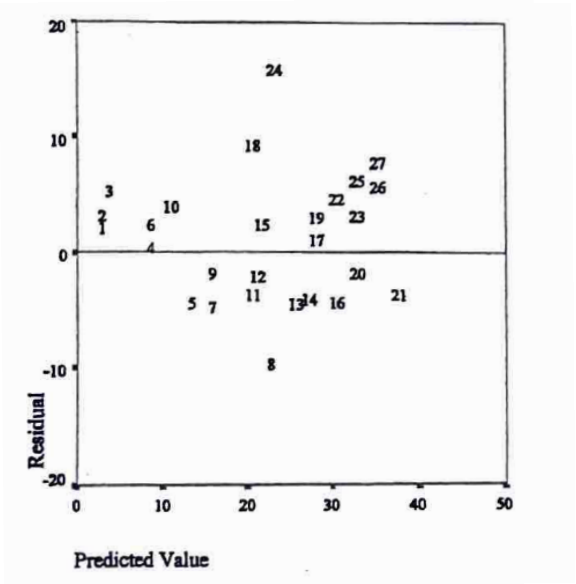
On sait que l'on mesure l'âge d'un arbre en comptant les anneaux sur une section transversale du tronc, mais cela nécessite de l'avoir abattu auparavant. Peut-on connaître l'âge à partir de la mesure de sa circonférence ?

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

Dans le cas de la régression linéaire simple, le test de Student et de Fisher sont totalement équivalents. Par conséquent, il n'est pas surprenant d'observer que $\sqrt{F} = t_{diametre}$.

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.

Part IV
ANOVA

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.

11.1 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.

11.2 Facteur à a modalités

11.2.1 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.

11.2.2 Le test de Fisher

L'hypothèse à tester est

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

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.

11.2.3 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.

$$SCT = SCM + SCE.$$

En effet,

$$\begin{aligned}
SCT &= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left[(X_{ij}-\bar{X}_i)^2+(\bar{X}_i-\bar{X})^2+2(X_{ij}-\bar{X}_i)(\bar{X}_i-\bar{X})\right] \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2+2\sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)(\bar{X}_i-\bar{X}) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{a}n_i(\bar{X}_i-\bar{X})^2+2\sum_{i=1}^{a}\left[(\bar{X}_i-\bar{X})\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)\right] \\
&= SCE+SCM+0,
\end{aligned}$$

en remarquant que

$$\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)=0.$$

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 :

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

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

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

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

11.2.4 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.

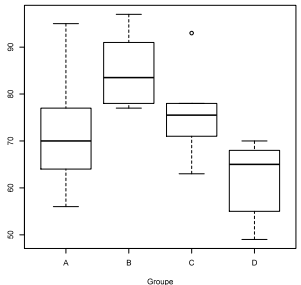
Résumé des données.

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

Avant de commencer l'analyse, notons que les quatre absorptions totales de M.G. diffèrent de façon visible : de 372 pour la 4e à 510 pour la 2e.

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



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.

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

$$\text{Pondéré } s^2 = 2018/20 = 100{,}9$$
$$s_{\bar{D}} = \sqrt{2s^2/n} = \sqrt{2 \times 100{,}9/6} = 5{,}80$$

On rappelle que le test de Fisher de l'ANOVA a pour hypothèse nulle

$$\mathcal{H}_0 = \text{“les moyennes sur chaque groupe sont égales”}.$$

On peut calculer la statistique du test

$$F = \frac{SCM/(4-1)}{SCE/(24-4)},$$

avec

$$SCM = \sum_{i=1}^{a} n_i \bar{X}_i^2 - n\bar{X}^2 = 132174 - 24 \times \left(\frac{1770}{24}\right)^2 = 1636,5,$$

et

$$SCE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{a} \bar{X}_i^2 = 2018.$$

Il vient que

$$F = \frac{1636,5/3}{2018/20} = 5,41,$$

à comparer avec le quantile théorique d'ordre 0.95 de la loi $\mathcal{F}(3;20)$ qui est 3,10. Au seuil 5%, on rejette $\mathcal{H}_0$, on en conclut que le type de matière grasse influe sur la quantité de matière grasse absorbée.

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				

Plus simplement, vu la p-value, on rejette significativement l'hypothèse nulle. De même, on en conclut que le type de matière grasse influe sur la quantité de matière grasse absorbée.