Florent ARNAL

VADE MECUM
de
STATISTIQUE et PROBABILITÉS

Université de Bordeaux

Adresse électronique : florent.arnal@u-bordeaux.fr
Site internet : http ://flarnal.e-monsite.com
Septembre 2018

Table des matières

I Notion de variable

Les données statistiques se présentent sous la forme d'individus auxquels on associe des caractères (appelés également "variables statistiques"). L'ensemble des individus constitue un échantillon (ou encore une série statistique), formant ainsi un sous-ensemble d'un groupe appelé "population".
Les caractères statistiques peuvent être de plusieurs natures :

• Les variables qualitatives :
Les valeurs possibles d'une variable qualitative sont appelées les modalités.
Exemples : couleur de cheveux, sexe.
Remarque : On qualifie d'ordinale une variable qualitative pour laquelle la valeur mesurée sur chaque individu est numérique (par exemple, une appréciation où l'on qualifie le produit de Passable (1) à très bon (5)).

• Les variables quantitatives (que l'on peut mesurer) ;
— Les variables quantitatives discrètes prenant des valeurs (isolées) dans un ensemble fini ou dénombrable (en général $\mathbb{N}$).
Exemple : Notes ;
— Les variables quantitatives continues prenant leurs valeurs dans un intervalle de $\mathbb{R}$ (voire $\mathbb{R}$).
Exemples : taille d'un individu, pH, masse.

II Statistique descriptive à 1 variable

II.1 Paramètres de position

1. Médiane d'une série

DÉFINITION 1 : La médiane notée M_e est un nombre réel tel que la moitié des observations lui sont inférieures (ou égales) et la moitié supérieures (ou égales).

REMARQUE 1 : La médiane partage la série statistique $(x_i)_{1\leq i\leq N}$ en deux groupes de "même effectif" (les valeurs du caractère étant rangées par ordre croissant).
Cas d'une série discrète d'effectif total N(les observations étant rangées par ordre croissant).
On convient que :

- Si N impair, la médiane est telle que $M_e = x_{\frac{N+1}{2}}$.
- Si N est pair, on convient que la médiane est $M_e = \frac{x_{\frac{N}{2}}+x_{\frac{N}{2}+1}}{2}$.

Exemple 1 *La série* $(1;3;4;6;8;8)$ *admet pour médiane* $M_e = 5$.

2. Moyenne d'une série

DÉFINITION 2 : Soit $(x_i)_{1\leq i\leq N}$ une série statistique.
La moyenne de cette série statistique est le réel noté $\bar{x}$ défini par :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

REMARQUE 2 : La moyenne (pondérée) d'une série statistique $(x_i;n_i)$ telle que $\sum_{i=1}^{p} n_i = N$ est le réel noté $\bar{x}$ défini par :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + ... + n_px_p}{N} = \frac{1}{N}\sum_{i=1}^{p} n_ix_i$$

Remarque 3 : La moyenne et la médiane d'une distribution statistique s'expriment dans la même unité que les valeurs prises par le caractère étudié.
Il est à noter que la médiane présente l'avantage par rapport à la moyenne de ne pas dépendre des valeurs extrêmes qui peuvent être suspectes voire aberrantes.

Exemple 2 *La série* $(1;3;4;6;8;8)$ *admet pour moyenne* $\bar{x} = \frac{1}{6} \times 30 = 5$.

Propriété 1 : Soit $(x_i)_{1 \leq i \leq N}$ une série statistique. On a :

$$\sum_{1 \leq i \leq N} (x_i - \bar{x}) = 0$$

II.2 Paramètres de dispersion

1. L'étendue

Définition 3 : L'étendue R d'une série statistique est la différence entre la plus grande et la plus petite valeur observées. On la note souvent R.

Exemple 3 *Les séries* $(1;3;4;6;8;8)$ *et* $(1;7;7;8;8;8)$ *ont la même étendue* $R = 7$.

Remarque 4 : Comme l'étendue ne permet pas de différencier les séries, elle n'est pas ici une très bonne mesure de la dispersion. D'une façon générale, quand il existe des valeurs extrêmes, l'étendue est une mesure médiocre de la dispersion. Il est donc utile de faire appel à de nouveaux paramètres comme la variance et l'écart-type.

2. La variance et l'écart-type

Définition 4 : La variance de la série $(x_i)_{1 \leq i \leq N}$ est le réel noté σ_X^2 voire s_X^2 défini par :

$$\sigma_X^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Le réel σ_X est appelé écart-type de la série (x_i).
Le réel $\sum_{i=1}^{N} (x_i - \bar{x})^2$ est appelé Somme des Carrés des écarts (à la moyenne). On note :

$$SCE_X = \sum_{i=1}^{N} (x_i - \bar{x})^2$$

On a donc :

$$\sigma_X^2 = \frac{SCE_X}{N}$$

Remarque 5 : $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left({x_i}^2 + \bar{x}^2 - 2\bar{x}x_i\right) = \sum_{i=1}^{n} {x_i}^2 + n\bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} {x_i}^2 + n\bar{x}^2 - 2\bar{x}n\bar{x}$
On a donc : $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$ ce qui implique

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} {x_i}^2 - \bar{x}^2 \quad (1)$$

REMARQUE 6 : La variance de la série $(x_i; n_i)_{1\leq i\leq p}$ telle que $\sum_{i=1}^{p} n_i = N$ est définie par :

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{p} n_i \left(x_i - \bar{x}\right)^2$$

REMARQUE 7 : L'écart-type permet de mesurer la dispersion autour de la moyenne d'une série statistique. Plus l'écart-type est faible, plus la série est homogène.
Il s'exprime dans la même unité que les valeurs observées.

Exemple 4 *Considérons la série* $(5; 7; 9; 11; 13)$.
On a : $\bar{x} = 9$ *et* $SCE = (5-\bar{x})^2 + \cdots + (13-\bar{x})^2 = 40$.
Ainsi : $\sigma_X^2 = \dfrac{SCE}{N} = 8$ *et* $\sigma_X = 2\sqrt{2}$.

II.3 Quartiles d'une série statistique

Les valeurs étant rangées par ordre croissant, on rappelle que la médiane partage la population en deux populations (basse et haute).

DÉFINITION 5 : On appelle premier quartile Q_1 la médiane de la partie basse de la population, troisième quartile Q_3 la médiane de la partie haute de la population.

REMARQUE 8 : Les quartiles correspondent aux valeurs prises par le caractère qui partagent la série en quatre groupes de même effectif.
Le premier quartile peut être considéré comme la plus petite valeur du caractère telle qu' au moins 25 % des valeurs lui sont inférieures ou égales.

DÉFINITION 6 : L'intervalle $[Q_1; Q_3]$ est appelé intervalle interquartile.
La longueur de cet intervalle $L = Q_3 - Q_1$ est appelée écart interquartile.

REMARQUE 9 : L'intervalle interquartile est un paramètre de dispersion qui élimine les valeurs extrêmes qui peuvent être douteuses, ce qui est un avantage par rapport à l'étendue.
Cependant, il ne tient compte que de 50 % de la population ce qui engendre une perte parfois conséquente d'information.

Les quartiles ne sont qu'un cas particulier de la notion de quantiles (https ://fr.wikipedia.org/wiki/Quantile).

Bilan : Résumés d'une série statistique par ses paramètres

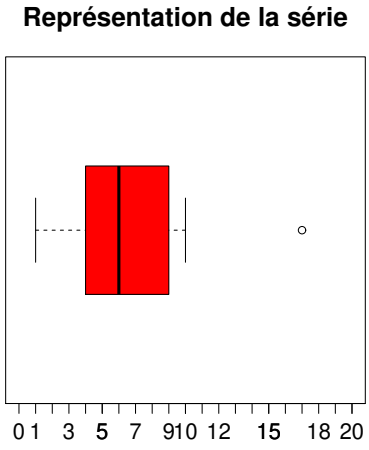
- Le couple (médiane ; étendue) est le plus "simple" à déterminer mais ne permet pas de situer les valeurs extrêmes par rapport à la moyenne.
- Le couple (médiane ; intervalle ou écart interquartile) est plus précis que le précédent car il est insensible aux valeurs extrêmes.
- Le couple (moyenne ; écart-type) est le couple le plus fréquemment utilisé, notamment pour décrire des distributions "normales" (courbe en cloche).
  Il s'avère que, dans ce cas, environ 68 % des valeurs sont comprises entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$, environ 95 % des valeurs sont comprises entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ et plus de 99 % des valeurs sont comprises entre $\bar{x} - 3\sigma$ et $\bar{x} + 3\sigma$ où σ correspond à l'écart-type de la population.

II.4 Représentation graphique d'une série à l'aide d'un boxplot

Pour représenter une série (variable) quantitative, il est souvent utile d'utiliser un boxplot (boîte à moustaches). On repère sur la boîte à moustaches d'une variable :

- l'échelle des valeurs de la variable, située sur l'axe vertical ;
- la valeur du premier quartile Q_1, correspondant au trait inférieur de la boîte ;
- la valeur de la médiane M_e, représentée par un trait horizontal à l'intérieur de la boîte ;
- la valeur du troisième quartile Q_3, correspondant au trait supérieur de la boîte ;
- les 2 «moustaches» inférieure et supérieure délimitent les valeurs dites adjacentes qui sont déterminées à partir de l'écart interquartile $(Q_3 - Q_1)$. La valeur extrême de la moustache inférieure correspond à la plus petite des valeurs supérieures ou égales à $Q_1 - 1,5 * (Q_3 - Q_1)$. La valeur extrême de la moustache supérieure correspond à la plus grande des valeurs inférieures ou égales à $Q_3 + 1,5 * (Q_3 - Q_1)$;
- d'éventuelles valeurs suspectes situées au-delà des valeurs extrêmes définies ci-dessus peuvent apparaître et sont représentées par des marqueurs (rond, étoile, etc.). Dans le cas où il n'existe pas de valeur suspecte, les valeurs extrêmes des moustaches correspondent aux valeurs minimales et maximales.

Exemple 5 *Considérons la série suivante :* $(1; 3; 4; 5; 6; 7; 9; 10; 17)$.
On a : $M_e = 6$; $Q_1 = 4$; $Q_3 = 9$. *Ainsi,* $Q_3 - Q_1 = 5$ *ce qui induit que :*
$Q_1 - 1,5 * (Q_3 - Q_1) = -3,5$ *et* $Q_3 + 1,5 * (Q_3 - Q_1) = 16,5$. *Il y a donc une valeur suspecte (*17*). Les valeurs extrêmes des «moustaches» étant égales à* 1 *et* 10.
On obtient la représentation graphique suivante :

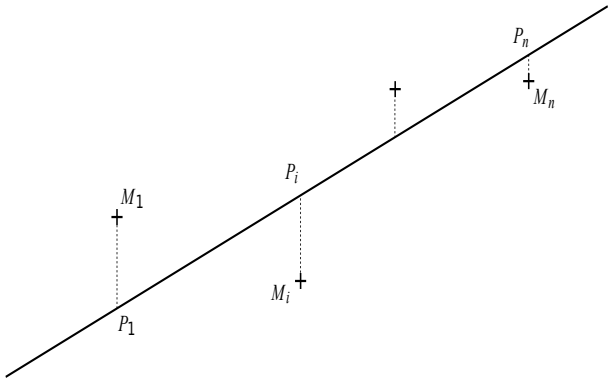


REMARQUE 10 : Dans la boîte à moustaches définie par TUKEY, la boîte a pour longueur la distance interquartile $(Q_3 - Q_1)$, et les moustaches sont basées usuellement sur 1,5 fois la longueur de la boîte. Dans ce cas, une valeur est suspecte (atypique) si elle dépasse de $1,5$ fois l'écart interquartile au dessous du premier quartile ou au dessus du troisième quartile.
Le choix de la valeur $1,5$ par TUKEY a une justification probabiliste.
En effet, si une variable suit une distribution normale, alors la zone délimitée par la boîte et les moustaches devrait contenir 99,3 % des observations. On ne devrait donc trouver que 0,7% d'observations suspectes (outliers).

III Statistique descriptive à 2 variables

III.1 Méthode des moindres carrés

On considère une série statistique double représentée par un nuage de points $M_i\,(x_i; y_i)_{1\leq i\leq n}$.
Soit (D) une droite d'ajustement. Pour tout entier naturel i tel que $1 \leq i \leq n$, on note P_i le projeté de M_i sur la droite (D) parallèlement à l'axe des ordonnées.
Ajuster ce nuage de point par la méthode des moindres carrés, c'est déterminer la droite (D) pour que la somme $\sum_{i=1}^{n} {M_iP_i}^2$ soit minimale.



III.2 Equations des droites de régression

${M_iP_i}^2 = (y_i - [ax_i + b])^2$ donc $\sum_{i=1}^{n} {M_iP_i}^2 = \sum_{i=1}^{n} [(ax_i + b) - y_i]^2$.
Ainsi, minimiser $\sum_{i=1}^{n} {M_iP_i}^2$ revient à déterminer le minimum de la fonction φ définie sur $\mathbb{R}^2$ par

$$\varphi(a, b) = \sum_{i=1}^{n} [(ax_i + b) - y_i]^2$$

On admet que si la fonction φ admet un extremum en $(x_0; y_0)$ alors ses dérivées partielles s'y annulent.
$\frac{d\varphi}{da}(a, b) = 2\sum_{i=1}^{n} x_i[(ax_i + b) - y_i]$
$\frac{d\varphi}{db}(a, b) = 2\sum_{i=1}^{n} [(ax_i + b) - y_i]$
$\frac{d\varphi}{da}(a, b) = 0$ implique $\sum_{i=1}^{n} \left(a{x_i}^2 + bx_i - x_iy_i\right) = 0$ soit $a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_iy_i = 0$.
$\frac{d\varphi}{db}(a, b) = 0$ implique $\sum_{i=1}^{n} [(ax_i + b) - y_i] = 0$ soit $a\sum_{i=1}^{n} x_i + nb - \sum_{i=1}^{n} y_i = 0$.
On a donc :

$$\sum_{i=1}^{n} x_iy_i = a\sum_{i=1}^{n} {x_i}^2 + b\sum_{i=1}^{n} x_i \quad \text{et} \quad \sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + nb$$

En posant $\alpha = \sum_{i=1}^{n} {x_i}^2$ et $\beta = \sum_{i=1}^{n} x_i$, on a :

$$\begin{pmatrix} \alpha & \beta \\ \beta & n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_iy_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$$

En notant M la matrice définie par $M = \begin{pmatrix} \alpha & \beta \\ \beta & n \end{pmatrix}$, on a : $M \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_iy_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}$.

On rappelle que :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{et} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left({x_i}^2 + \bar{x}^2 - 2\bar{x}x_i\right) = \sum_{i=1}^{n} {x_i}^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} {x_i}^2 + n\bar{x}^2 - 2\bar{x}n\bar{x}$$

On a donc : $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$ ce qui implique

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2 - \bar{x}^2 \tag{2}$$

La matrice M est inversible si et seulement si $n\alpha - \beta^2 \neq 0$ soit $n\alpha - (n\bar{x})^2 \neq 0$.
M inversible équivaut donc à $\frac{1}{n}\alpha - \bar{x}^2 \neq 0$. D'après ce qui précède, on en déduit que :
M inversible ssi $\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \neq 0$.
Or : $\sum_{i=1}^{n} (x_i - \bar{x})^2 = 0 \Longleftrightarrow \forall i \in \{1; 2; ...; n\},\ x_i - \bar{x} = 0 \Longleftrightarrow \forall i \in \{1; 2; ...; n\},\ x_i = \bar{x}$.
Ainsi : M inversible si et seulement si les x_i ne sont pas tous égaux.
Dans ce cas, on a : $M^{-1} = \frac{1}{n\alpha - \beta^2}\begin{pmatrix} n & -\beta \\ -\beta & \alpha \end{pmatrix}$

$$M\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} \text{ donc } \begin{pmatrix} a \\ b \end{pmatrix} = M^{-1}\begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}.$$

Ainsi : $a = \frac{1}{n\alpha - \beta^2}\left(n\sum_{i=1}^{n} x_i y_i - \beta\sum_{i=1}^{n} y_i\right) = \frac{n\sum_{i=1}^{n} x_i y_i - n^2\bar{x}\bar{y}}{n\sum_{i=1}^{n} {x_i}^2 - (n\bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}$

Quant à b, on a :

$$b = \frac{1}{n\alpha - \beta^2}\left(-\beta\sum_{i=1}^{n} x_i y_i + \alpha\sum_{i=1}^{n} y_i\right) = \frac{-n\bar{x}\sum_{i=1}^{n} x_i y_i + n\bar{y}\sum_{i=1}^{n} {x_i}^2}{n\sum_{i=1}^{n} {x_i}^2 - (n\bar{x})^2} = \frac{-\bar{x}\sum_{i=1}^{n} x_i y_i + \bar{y}\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}.$$

Cette relation peut s'écrire : $b = \frac{\bar{y}\left(\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2\right) - \bar{x}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = \bar{y} - a\bar{x}.$

En conclusion, φ peut admettre un extremum uniquement en $(a; b)$ tel que

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \qquad \text{et} \qquad b = \bar{y} - a\bar{x}$$

En considérant $b = \bar{y} - a\bar{x}$, on obtient :
$\varphi(a, b) = \sum_{i=1}^{n} (a(x_i - \bar{x}) + (\bar{y} - y_i))^2 = \sum_{i=1}^{n} \left(a^2(x_i - \bar{x})^2 + (\bar{y} - y_i)^2 + 2a(x_i - \bar{x})(\bar{y} - y_i)\right)^2$.
$\varphi(a, b)$ peut donc s'écrire sous la forme :

$$\varphi(a, b) = a^2\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2a\sum_{i=1}^{n}(x_i - \bar{x})(\bar{y} - y_i) + \sum_{i=1}^{n}(\bar{y} - y_i)^2$$

On constate que, si $b = \bar{y} - a\bar{x}$ alors $\varphi(a, b)$ est un polynôme du second degré avec un coefficient du terme dominant positif. Cette fonction admet donc un minimum ce qui permet de conclure que les réels a et b obtenus précédemment permettent de minimiser $\sum_{i=1}^{n} M_i P_i{}^2$.

A noter que, tout comme la formule de la variance (2) , $\sum_{i=1}^{n}(x_i-\bar{x})(\bar{y}-y_i)$ peut s'écrire autrement en développant :

$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) = \sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}$ On a donc :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y} \tag{3}$$

Une équation de la droite d'ajustement de Y en X est donc $Y = aX + b$ avec

$$a = \frac{\sum_{i=1}^{n} x_iy_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{et} \qquad b = \bar{y} - a\bar{x}$$

Théorème-Définition 1 :
On considère deux variables X et Y prenant respectivement les valeurs $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$.
La **covariance** de X et Y est le réel noté $\mathrm{Cov}(X,Y)$ ou σ_{XY} défini par :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y})$$

ou encore

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y} = \overline{xy} - \bar{x}\bar{y}$$

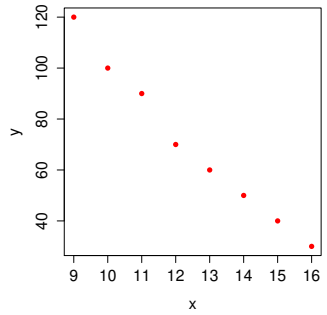
PROPRIÉTÉ 2 : La droite de régression de Y en X a pour équation $Y = aX + b$ avec

$$a = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} \text{ et } b = \bar{y} - a\bar{x}$$

Exemple 6 *Considérons le nombre d'acheteurs potentiels d'un produit en fonction de son prix de vente.*

Prix x_i en euros	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
Nombre y_i d'acheteurs éventuels	*120*	*100*	*90*	*70*	*60*	*50*	*40*	*30*

Le nuage de points associé à la série $(x_i;y_i)$ laisse apparaître un relation linéaire entre le nombre Y d'acheteurs potentiels et le prix X.



Calculons les coefficients de la droite de régression de Y en X :
$\bar{x} = 12,5$ *;* $\bar{y} = 70$ *;* $Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\,(y_i-\overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_iy_i - \bar{x}\bar{y} = -66,25$ *et* $\sigma_X^2 = 5,25$.
Ainsi : $a = \frac{Cov(X,Y)}{\sigma_X^2} = \frac{-66,25}{5,25} \simeq -12,62$ *et* $b = \bar{y} - a\bar{x} \simeq 227,74$.
La droite de régression de Y en X a pour équation

$$Y = -12,62X + 227,74$$

III.3 Résidus de la régression

DÉFINITION 7 : On appelle résidus de Y par rapport à X les écarts $e_i = y_i - \hat{y}_i$ entre les valeurs observées de la variable dépendante y_i et les valeurs correspondantes $\hat{y}_i = ax_i + b$ calculées à l'aide de l'équation de la droite de régression de Y en X .

REMARQUE 11 : Les valeurs $\hat{y}_i$ sont aussi appelées valeurs estimées de la variable dépendante Y.

PROPRIÉTÉ 3 : Les résidus sont de somme et de moyenne nulle.

En effet : $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - (ax_i + b)) = n\bar{y} - an\bar{x} - nb = n\left(\bar{y} - a\bar{x} - b\right) = 0$

REMARQUE 12 : Les séries $(y_i)_{1\leq i\leq n}$ et les estimations $(\hat{y}_i)_{1\leq i\leq n}$ ont donc la même moyenne.

Exemple 7 *Calculons les résidus associés à l'exemple précédent.*
Dans un premier temps, nous déterminons les estimations en utilisant l'équation $\hat{y}_i = -12,62x_i + 227,74$.
On obtient : $114,16$; $101,54$; $88,92$; $76,30$; $63,68$; $51,06$; $38,44$; $25,82$.
Ainsi, les résidus étant définis par $e_i = y_i - \hat{y}_i$*, on obtient :*

$$5,84; \quad ; \quad -1,54 \quad ; \quad 1,08 \quad ; \quad -6,30 \quad ; \quad -3,68 \quad ; \quad -1,06 \quad ; \quad 1,56 \quad ; \quad 4,18$$

III.4 Coefficients de corrélation et de détermination

Ces coefficients permettent d'apprécier (partiellement) la pertinence et la qualité d'un ajustement.

DÉFINITION 8 : On appelle coefficient de corrélation linéaire de la série $(x_i, y_i)_{1\leq i\leq n}$ le nombre réel

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

PROPRIÉTÉ 4 : Comme $\sigma_X$et σ_Y sont positifs alors $\rho(X,Y)$, a et $\mathrm{Cov}(X,Y)$sont de même signe.

DÉFINITION 9 : (Les différentes SCE)

On considère $(x_i; y_i)_{1\leq i\leq n}$ une série statistique à deux variables constituée de n couples.

- La variabilité totale (des y_i) est définie par :

$$SCE_{\text{totale}} = \sum_i (y_i - \bar{y})^2$$

- La variabilité expliquée SCE_{exp} (par l'ajustement affine) est définie par :

$$SCE_{\text{exp}} = \sum_i (\hat{y}_i - \bar{y})^2$$

  Elle correspond à la SCE des estimations $\hat{y}_i$.
- La variabilité résiduelle SCE_{res} est définie par :

$$SCE_{\text{res}} = \sum_i {e_i}^2$$

  Elle correspond à la SCE des résidus.

$SCE_{\text{totale}} = \sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$.
En développant, on a :
$SCE_{\text{totale}} = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)\ (\hat{y}_i - \bar{y})$.
Or, $\sum_i (y_i - \hat{y}_i)\ (\hat{y}_i - \bar{y}) = \sum_i (y_i - \bar{y} - a(x_i - \bar{x}))\ (a(x_i - \bar{x})) = a\sum_i (y_i - \bar{y} - a(x_i - \bar{x}))\ (x_i - \bar{x}) = 0$. On en déduit la relation explicitée sur la page suivante.

PROPRIÉTÉ 5 :
$$SCE_{\text{totale}} = SCE_{\text{exp}} \ + \ SCE_{\text{res}}$$

DÉFINITION 10 : Le coefficient de détermination d'une série statistique à deux variables X et Y est $\rho^2(X,Y)$.

Il s'avère que : $\dfrac{SCE_{\text{exp}}}{SCE_{\text{totale}}} = \dfrac{\sum_i (y_i - \bar{y})^2}{n\sigma_Y^2} = \dfrac{a^2\sum_i (x_i - \bar{x})^2}{n\sigma_Y^2} = a^2\dfrac{\sigma_X^2}{\sigma_Y^2} = \left(\dfrac{\sigma_{XY}}{\sigma_X^2}\right)^2 \dfrac{\sigma_X^2}{\sigma_Y^2} = \left(\dfrac{\sigma_{XY}}{\sigma_X\sigma_Y}\right)^2$.
La définition du coefficient de détermination conduit à la propriété suivante :

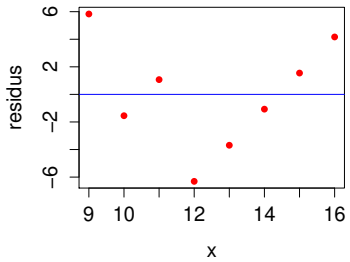
PROPRIÉTÉ 6 : Le coefficient de détermination $\rho^2(X,Y)$ correspond à la part de variabilité de Y expliquée par la régression. On a :
$$\rho^2(X,Y) = \frac{SCE_{\text{exp}}}{SCE_{\text{totale}}}$$

La relation $SCE_{\text{totale}} = SCE_{\text{exp}} \ + \ SCE_{\text{res}}$ induit la propriété suivante :

PROPRIÉTÉ 7 :
1. $\rho^2(X,Y) \leq 1$
2. $-1 \leq \rho(X,Y) \leq 1$

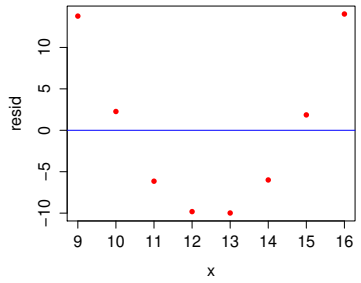
À retenir :
L'ajustement sera d'autant meilleur que $\rho^2(X,Y)$ est proche de 1.
En pratique : on considère que si $0,8 \leq |\rho(X,Y)| \leq 1$, il y a une forte corrélation.
Si $\rho^2(X,Y)$ est proche de 1, on peut dire qu'il existe une corrélation importante entre les deux variables.
Mais ceci n'implique pas nécessairement l'existence d'une relation directe de cause à effet entre les deux variables.
La pertinence d'un ajustement affine peut être vérifiée par l'étude des résidus dont la représentation graphique ne doit faire apparaître aucune tendance.
Sinon, malgré un coefficient de corrélation linéaire (ou de détermination) élevé, un autre ajustement sera souvent plus cohérent ...

Exemple 8 *Voici ci-dessous une représentation graphique des résidus calculés précédemment. Il ne laisse apparaître aucune tendance.*



Le coefficient de détermination $\rho^2(X,Y)$ est environ égal à $0,98$. L'ajustement affine est donc pertinent.

Exemple 9 *Le graphique ci-dessous représente les résidus associés à un ajustement affine avec un coefficient de détermination $\rho^2(X,Y)$ est environ égal à $0,99$.*



On voit clairement que les résidus sont positifs aux extrêmités et négatifs au centre.
Le modèle linéaire n'est pas pertinent (un ajustement quadratique le serait).

À retenir (Covariance et indépendance) :

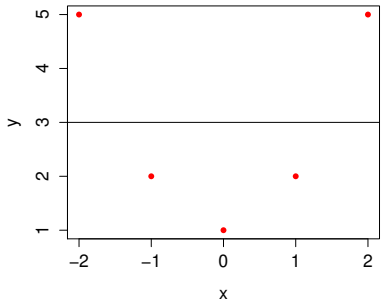
1. Si X et Y sont indépendantes alors $\text{Cov}(X,Y)=0$ (on a également $\rho(X,Y)=0$).
2. Par contre, la réciproque n'est pas vraie comme le montre l'exemple ci-dessous.
   Si $\text{Cov}(X,Y)=0$, cela n'implique pas que X et Y sont indépendantes.

Exemple 10

Considérons deux variables X et Y définies par le tableau ci-dessous :

x_i	-2	-1	0	1	2
y_i	4	1	0	1	4

Le nuage de points associé à la série $(x_i;y_i)$, représenté ci-dessous, ne laisse apparaître aucune liaison linéaire entre X et Y.



On a d'ailleurs :

$$\text{Cov}(X,Y)=0$$

Cependant, X et Y ne sont pas indépendantes car

$$Y=X^2+1$$

IV Probabilités

IV.1 Généralités

Définition 11 : Une expérience dont on ne peut prévoir le résultat à l'avance est appelée expérience aléatoire.
Lors d'une expérience aléatoire, chaque résultat possible est appelé une issue.
L'ensemble de tous les résultats possibles de cette expérience aléatoire est appelé univers noté Ω.
Toute **partie** de l'univers est appelée un événement. Ω est appelé " événement certain " ; $\emptyset$ est appelé " événement impossible ".

Remarque 13 : Un événement qui ne contient qu'une éventualité est un évènement élémentaire.

Définition 12 : Deux événements A et B qui n'ont aucune éventualité commune sont *incompatibles* (les ensembles étant disjoints). On a donc : $A \cap B = \emptyset$.

Définition 13 : Soit A un événement de l'univers Ω.
L'événement constitué de toutes les éventualités de Ω qui n'appartiennent pas à A est appelé l'événement contraire de A, noté $\overline{A}$.

IV.2 Probabilités d'événements

Définition 14 : Soit Ω un ensemble fini et $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω.
L'application $\mathbb{P} : \mathcal{P}(\Omega) \to [0\,;\,1]$ est une probabilité sur Ω si :

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ pour tous A et B tels que $A \cap B = \emptyset$.
- $\mathbb{P}(\Omega) = 1$.

Propriété 8 : La probabilité d'un événement A est la somme des probabilités des événements élémentaires de A.

Propriété 9 : Soient A et B deux évènements de l'univers Ω, on a :

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}(A)$.

Définition 15 : On dit qu'il y a équiprobabilité lorsque les probabilités de tous les événements élémentaires sont égales.

Propriété 10 : Sur un univers fini, en situation d'équiprobabilité, pour tout évènement A, on a :

$$\mathbb{P}(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} = \frac{\text{"nombre de cas favorables"}}{\text{"nombre de cas possibles"}}$$

IV.3 Conditionnement et indépendance

Définition 16 : (Probabilité conditionnelle)
Soit Ω un univers associé à une expérience aléatoire et B un événement de probabilité non nulle.
On appelle probabilité de A sachant B le nombre, noté $\mathbb{P}_B(A)$, défini par :

$$\mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Propriété 11 : (Formule des probabilités totales)
Soit Ω un univers associé à une expérience aléatoire et A,B deux événements de probabilité non nulle.
On a : $\mathbb{P}(A) = P(A \cap B) + \mathbb{P}\left(A \cap \overline{B}\right)$ soit $\mathbb{P}(A) = \mathbb{P}_B(A)\, P(B) + \mathbb{P}_{\overline{B}}(A)\, P\left(\overline{B}\right)$

Définition 17 : (Evénements indépendants)
Soient Ω un univers associé à une expérience aléatoire, Aet B étant deux événements.
Aet B sont indépendants si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\, P(B)$.

IV.4 Généralités sur les variables aléatoires

Si X est une application d'un espace Ω dans $\mathbb{R}$ et si $A' \subset \mathbb{R}$, l'ensemble $X^{-1}(A')$ est le sous-ensemble de Ω, appelé image réciproque de A' par X, défini par $X^{-1}(A') = \{\omega \in \Omega; X(\omega) \in A'\}$ que l'on note usuellement $[X \in A']$.
En d'autres termes, il s'agit de l'ensemble des éléments de Ω qui ont leur image par X dans A'.

Définition 18 : (Variable aléatoire réelle)
On appelle variable aléatoire réelle (en abrégé v.a.r.) toute application $X : \Omega \to \mathbb{R}$ telle que, pour tout intervalle I de $\mathbb{R}$, on a : $X^{-1}(I) \in A$.

Définition 19 : (Fonction de répartition)
On appelle fonction de répartition de la variable aléatoire $X : \Omega \to \mathbb{R}$ l'application F_X définie sur $\mathbb{R}$ par

$$F_X(x) = \mathbb{P}(X \leqslant x)$$

Propriété 12 :

1. Pour tout $x \in \mathbb{R}$, $F_X(x) \in [0,1]$;
2. F_X est croissante ;
3. $\lim\limits_{x \to -\infty} F_X(x) = 0$ et $\lim\limits_{x \to +\infty} F_X(x) = 1$.

IV.5 Variables aléatoires discrètes

Définition 20 : Une variable aléatoire discrète X est une application de $\Omega \to \mathbb{R}$ qui prend les valeurs isolées $x_1, x_2, \cdots$.

Remarque 14 : $X(\Omega)$ est fini ou dénombrable.

Définition 21 : Notons p_{i} la probabilité que X prenne la valeur x_i pour $i = 1, 2, \cdots$.
L'affectation des p_{i} aux valeurs x_i permet de définir une probabilité (appelée aussi loi de probabilité) $\mathbb{P}_X$ sur $X(\Omega)$.
$\mathbb{P}_X(x_i) = \mathbb{P}(\omega \in \Omega / X(\omega) = x_i)$ se note $\mathbb{P}(X = x_i)$.

Remarque 15 :

1. $\sum_i \mathbb{P}(X = x_i) = 1$.
2. La loi de probabilité d'une variable aléatoire discrète est donnée soit par la liste des probabilités (présentées dans un tableau), soit par une formule.

Propriété 13 : Soit X une variable aléatoire discrète admettant pour fonction de répartition F_X. Pour tout réel $a \in X(\Omega)$, on a :

$$\mathbb{P}(X < a) = F_X(a-1)$$

Définition 22 : Soit X une variable aléatoire discrète prenant les valeurs $x_1, x_2, \cdots$.
L'espérance de X est le réel, noté $\mathbb{E}(X)$, défini par :

$$\mathbb{E}(X) = \sum_i x_i \mathbb{P}(X = x_i)$$

Remarque 16 :

- Lorsque X prend une infinité de valeurs, l'espérance d'une variable aléatoire est définie sous réserve de convergence de la série $\sum_i x_i \mathbb{P}(X = x_i)$.
- $\mathbb{E}(X)$ est donc la **moyenne** des valeurs $x_i$pondérées par les probabilités $\mathbb{P}(X = x_i)$.

Théorème 1 : (Théorème de transfert)
Soit X une variable aléatoire discrète prenant les valeurs x_1, x_2, x_3 , ... et φ une fonction définie sur $\varphi(X(\Omega))$.
On a :

$$\mathbb{E}(\varphi(X)) = \sum_i \varphi(x_i)\, \mathbb{P}(X = x_i)$$

Propriété 14 : (Linéarité de l'espérance)
Soient X une variable aléatoire discrète et a, b deux réels. On a :

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Définition 23 : Soit Xune variable aléatoire discrète prenant les valeurs $x_1, x_2, \cdots$.

- La variance de X est le réel, noté $\mathbb{V}(X)$, défini par :

$$\mathbb{V}(X) = \sum_i [x_i - \mathbb{E}(X)]^2 \mathbb{P}(X = x_i)$$

- L'écart-type de X est le réel, noté $\sigma(X)$, défini par :

$$\sigma(X) = \sqrt{V(X)}$$

Remarque 17 :

- Lorsque X prend une infinité de valeurs, l'espérance d'une variable aléatoire est définie sous réserve de convergence de la série $\sum_i x_i \mathbb{P}(X = x_i)$.
- Pour le calcul de la variance, on utilise parfois la formule suivante (Formule de Koenig-Huygens) :

$$\mathbb{V}(X) = \sum_i x_i^2 \mathbb{P}(X = x_i) - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

Propriété 15 : Soient X une variable aléatoire discrète et a, b deux réels.

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

et

$$\sigma(aX + b) = |a| \sigma(X)$$

Propriété 16 : Considérons un couple de variables aléatoires $(X; Y)$ admettant une espérance et une variance. On a :

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \quad \text{et} \quad \mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y)$$

En outre, si X et Y sont indépendantes alors

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$$

IV.6 Lois usuelles discrètes

Définition 24 : On dit que X est distribuée selon la loi de Bernoulli de paramètre p si :

- X ne peut prendre que les valeurs 0 et 1.
- $\mathbb{P}(X = 1) = p$ et $\mathbb{P}(X = 0) = q = 1 - p$.
  On note : $X \sim \mathcal{B}(1, p)$.

Propriété 17 : On suppose que $X \sim \mathcal{B}(1, p)$. On a :

$$\mathbb{E}(X) = p \quad \text{et} \quad \mathbb{V}(X) = p(1 - p) = pq$$

Définition 25 : Supposons que l'on répète n fois, dans des conditions identiques et indépendantes, une expérience aléatoire dont l'issue est :

- soit un succès (noté S) avec la probabilité p ;
- soit un échec (noté E) avec la probabilité $q = 1 - p$.

On note X la variable aléatoire égale au nombre de succès obtenus lors de ces n expériences.
On dit que X est distribuée suivant la loi binomiale de paramètres n et p notée $\mathcal{B}(n, p)$.

Remarque 18 :

1. X prend les valeurs de $[\![0; n]\!]$.
2. Lorsque les n tirages s'effectuent dans une population contenant un grand nombre N d'individus tel que $N > 10n$ (c'est-à-dire dès que la population est 10 fois plus grande que l'échantillon), les tirages pourront être assimilés à des tirages avec remise.

Théorème 2 : X suivant la loi binomiale $\mathcal{B}(n,p)$ est la somme de n variables de Bernoulli indépendantes et de même paramètre p.

PROPRIÉTÉ 18 : Si X suit la loi binomiale $\mathcal{B}(n,p)$ alors

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \quad \forall k \in [\![0;n]\!]$$

PROPRIÉTÉ 19 : (Espérance et variance)
Soit X une variable aléatoire distribuée selon la loi binomiale $\mathcal{B}(n,p)$. On a :

$$\mathbb{E}(X)=np \quad \text{et} \quad \mathbb{V}(X)=np(1-p)=npq$$

DÉFINITION 26 : On dit qu'une variable aléatoire X est distribuée selon la loi de Poisson de paramètre $\lambda>0$ notée $\mathcal{P}(\lambda)$ si :

- X prend ses valeurs dans $\mathbb{N}$
- $P(X=k)=\mathrm{e}^{-\lambda}\dfrac{\lambda^k}{k!}$ pour tout entier naturel k.

REMARQUE 19 : Une loi de Poisson ne correspond pas à une situation-type ; elle apparaît comme modèle décrivant certaines observations statistiques (pannes de machine, nombre de personnes passant à un guichet dans un temps donné,....) ou comme loi limite.

PROPRIÉTÉ 20 : (Espérance et variance)
Soit X une variable distribuée suivant la loi de Poisson de paramètre λ. On a :

$$E(X)=\lambda \quad \text{et} \quad V(X)=\lambda$$

IV.7 Généralités sur les lois continues

Nous avons étudié précédemment des lois de probabilité de variables aléatoires discrètes (ne prenant que des valeurs isolées).
Dans cette partie, nous allons considérer des variables aléatoires pouvant prendre toutes les valeurs de $\mathbb{R}$ (voire d'un intervalle de $\mathbb{R}$).

DÉFINITION 27 : On appelle fonction densité de probabilité, toute fonction f définie sur $\mathbb{R}$, telle que :

- f est continue sur $\mathbb{R}$ (sauf éventuellement en quelques valeurs) ;
- f est positive sur $\mathbb{R}$;
- $\displaystyle\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t=1.$

DÉFINITION 28 : On dit qu'une variable aléatoire X est continue à fonction de densité f si :
pour tout intervalle I de $\bar{\mathbb{R}}=\mathbb{R}\cup\{-\infty,+\infty\}$, on a :

$$\mathbb{P}(X\in I)=\int_I f(t)\,\mathrm{d}t$$

Remarque 20 : On dit également que la loi de X admet f comme densité de probabilité. Pour tout réel a, on a :

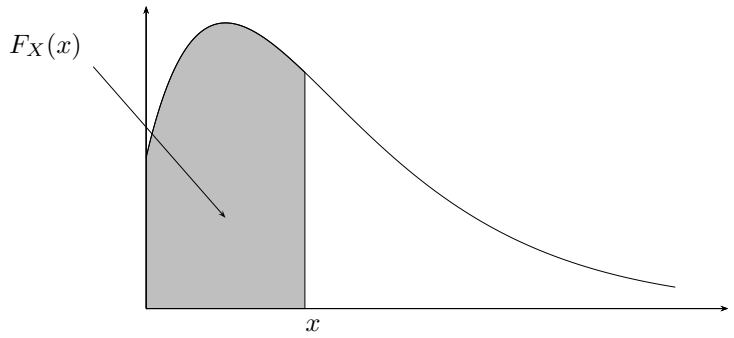
$$\mathbb{P}(X = a) = \mathbb{P}(a \leq X \leq a) = 0$$

Ainsi, la probabilité qu'une variable aléatoire continue X prenne une valeur isolée est nulle.
On a donc : $\mathbb{P}(X < a) = \mathbb{P}(X \leq a)$ pour tout réel a.
On utilise donc **indifféremment des inégalités larges ou strictes** avec des variables aléatoires continues.

Théorème-Définition 2 : Soit X une variable aléatoire continue de densité de probabilité f.

- La fonction de répartition de X est la fonction F définie sur $\mathbb{R}$ par $F_X(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \mathbb{P}(X \leq x)$.
- F_X est une primitive de f presque partout.

Illustration graphique :



Propriété 21 : Soit X une variable aléatoire continue admettant F_X comme fonction de répartition.

- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$ et $\mathbb{P}(X > b) = 1 - F_X(b)$.
- La fonction F_X est croisante sur $\mathbb{R}$ et $0 \leq F_X \leq 1$.

Définition 29 : Soit X une variable aléatoire continue de densité de probabilité f.

- L'espérance de X est le réel, noté $\mathbb{E}(X)$, défini par

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t\, f(t)\,\mathrm{d}t$$

- La variance de X est le réel, noté $\mathbb{V}(X)$, défini par

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} (t - \mathbb{E}(X))^2\, f(t)\,\mathrm{d}t$$

Remarque 21 :

- L'espérance de X n'est définie que si $\int_{-\infty}^{+\infty} t\, f(t)\, dt$ est absolument convergente.
- La variance peut, elle aussi, ne pas exister. Si elle existe, on a également : $\mathbb{V}(X) = \int_{-\infty}^{+\infty} t^2\, f(t)\, \mathrm{d}t - [\mathbb{E}(X)]^2$ soit $\mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2$.

IV.8 Lois usuelles continues

- **Lois uniformes** $\mathcal{U}(a,b)$
  $X(\Omega) = [a;b]$
  Densité : $f(x) = \dfrac{1}{b-a}$ si $x \in [a;b]$, 0 sinon
  $\mathbb{E}(X) = \dfrac{a+b}{2}$ et $\mathbb{V}(X) = \dfrac{(b-a)^2}{12}$

- **Lois normales** $\mathcal{N}(\mu,\sigma)$
  $X(\Omega) = \mathbb{R}$
  Densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  $\mathbb{E}(X) = \mu$ et $\sigma(X) = \sigma$
  La fonction de répartition de la loi $\mathcal{N}(0;1)$ est notée usuellement Φ

Propriété 22 : Si X est distribuée suivant la loi normale $\mathcal{N}(\mu,\sigma)$ alors :
$X^\star = \dfrac{X-\mu}{\sigma}$ est distribuée suivant la loi normale $\mathcal{N}(0;1)$ appelée loi normale centrée réduite.

Propriété 23 : Pour tout réel x positif, on a :

$$\Phi(-x) + \Phi(x) = 1$$

Propriété 24 : Pour tout réel x positif, on a :

$$\mathbb{P}(-x \leq U \leq x) = 2\Phi(x) - 1$$

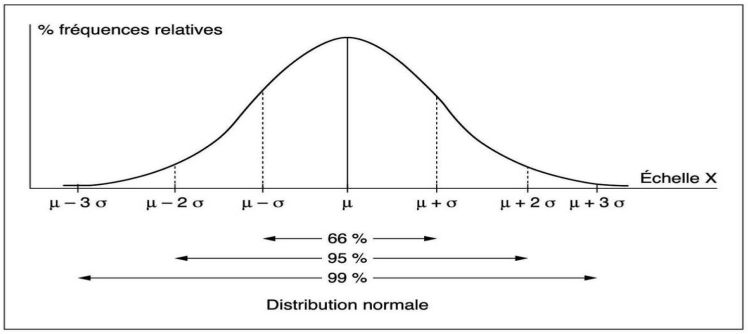


Figure 1 – Pourcentages classiques liés aux lois normales.

Propriété 25 : La somme de variables aléatoires indépendantes et identiquement distribuées suivant une loi normale suit une loi normale

- **Lois du Khi-deux**
  $X(\Omega) = \mathbb{R}^+$
  Si $X \sim \chi^2(n)$ alors $\mathbb{E}(X) = n$ et $\mathbb{V}(X) = 2n$

- **Lois de Student**
  $X(\Omega) = \mathbb{R}$
  Si $X \sim \mathcal{T}(n)$ avec $n \geqslant 2$ alors $\mathbb{E}(X) = 0$ et $\mathbb{V}(X) = \dfrac{n}{n-2}$

- **Liens entre les lois normale, du Khi-deux et de Student**

PROPRIÉTÉ 26 : Lois normales et Khi-deux
Si $X_1, ..., X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0;1)$ alors

$$X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$$

PROPRIÉTÉ 27 : Lois normales, Khi-deux et Student
Si X et Y sont deux variables indépendantes telles que $X \sim \mathcal{N}(0;1)$ et $Y \sim \chi^2(n)$ alors

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim \mathcal{T}(n)$$

- **Lois de Fisher**

PROPRIÉTÉ 28 : Si X_1 et X_2 sont deux variables indépendantes telles que $X \sim \chi^2(d_1)$ et $Y \sim \chi^2(d_2)$ alors la variable $\dfrac{\frac{X_1}{d_1}}{\frac{X_2}{d_2}} \sim \mathcal{F}(d_1, d_2)$.

V Loi faible des grands nombres, théorème central-limite

PROPRIÉTÉ 29 : Inégalité de Markov
Soit X une variable aléatoire réelle positive. Pour tout $a \geqslant 0$, on a

$$aP(X \geqslant a) \leqslant \mathbb{E}(X)$$

L'inégalité de Markov appliquée à $Y = (X - E(X))^2$ conduit à l'inégalité ci-dessous.

PROPRIÉTÉ 30 : Inégalité de Bienaymé-Tchebychev
Soit X une variable aléatoire réelle . Pour tout $\varepsilon > 0$, on a

$$P(|X - \mathbb{E}(X)| \geqslant \varepsilon) \leqslant \frac{\mathbb{V}(X)}{\epsilon^2}$$

PROPRIÉTÉ 31 : Loi faible des grands nombres
Soient $X_1, \cdots, X_n$ des variables indépendantes et identiquement distribuées de même loi d'espérance μ.
En considérant $S_n = \sum_{i=1}^{n} X_i$, on a

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geqslant \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$

Ce résultat découle de l'inégalité précédente en considérant $X = \dfrac{S_n}{n}$ d'espérance

$\mathbb{E}(X) = \mathbb{E}\left(\dfrac{S_n}{n}\right) = \dfrac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$ et $\mathbb{V}\left(\dfrac{S_n}{n}\right) = \dfrac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(X_i) = \dfrac{n\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$.

On a donc $P\left(\left|\dfrac{S_n}{n} - \mu\right| \geqslant \varepsilon\right) \leqslant \dfrac{\sigma^2}{n\epsilon^2} \underset{n\to+\infty}{\longrightarrow} 0$

PROPRIÉTÉ 32 : Théorème Central Limite (TCL)
Soient $X_1, \cdots, X_n$ des variables indépendantes et identiquement distribuées de même loi d'espérance μ et d'écart-type σ. On a

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0;1)$$

REMARQUE 22 : La somme ou la moyenne de variables de même loi d'espérance μ et d'écart-type σ **centrée réduite** convergent en loi vers une loi normale.

REMARQUE 23 : Approximation d'une loi binomiale par une loi normale
Nous avons vu précédemment qu'une variable X de loi $\mathcal{B}(n,p)$ peut s'écrire $X = \sum_{i=1}^{n} X_i$ où les X_i sont iid de loi de Bernoulli $\mathcal{B}(1,p)$.
En utilisant que $\mathbb{E}(X_i) = np$ et $\mathbb{V}(X_i) = p(1-p)$, d'après le TCL, on a

$$\frac{X - np}{\sqrt{np(1-p)}} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0;1)$$

Ainsi, pour n suffisamment grand, la loi $\mathcal{B}(n,p)$ peut être approchée par la loi $\mathcal{N}\left(np; \sqrt{np(1-p)}\right)$.

VI Statistique inférentielle

VI.1 Variables aléatoires d'échantillonnage et lois

• Moyenne : $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

• Variance : $\hat{S}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \dfrac{SCE}{n-1}$

• proportion : $F = \dfrac{X}{n}$

PROPRIÉTÉ 33 : Distribution de $\bar{X}$ (Cas de distributions normales, écart-type connu)
Si les variables $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu;\sigma)$ alors

$$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

PROPRIÉTÉ 34 : Distribution de $\bar{X}$ (Application du TCL dans de distributions quelconques, écart-type connu)
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées avec $n \geqslant 30$, $\mathbb{E}(X_i) = \mu$ et $\mathbb{V}(X_i) = \sigma^2$ alors
la loi de $\bar{X}$ converge vers $\mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$.

Ce propriété découle du TCL qui affirme que $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0;1)$. En divisant par n, il vient $\dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ ce qui permet d'affirmer que

$$\boxed{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0;1)}$$

PROPRIÉTÉ 35 : Distribution associée à $\bar{X}$ (Écart-type de la population inconnu)
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu;\sigma)$ alors

$$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$$

PROPRIÉTÉ 36 : Distribution associée à la variance
Si $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu;\sigma)$ alors

$$\frac{SCE}{\sigma^2} \sim \chi^2(n-1)$$

PROPRIÉTÉ 37 : Considérons une certaine population dans laquelle la proportion d'individus ayant une propriété donnée est égale à p.
• $X \sim \mathcal{B}(n,p)$;
• Pour n suffisamment grand, la loi de X peut être approchée par la loi $\mathcal{N}\left(np; \sqrt{np(1-p)}\right)$;
• Pour n suffisamment grand, la loi de $F = \dfrac{X}{n}$ peut être approchée par la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$.

VI.2 Estimation ponctuelle

Définition 30 : Soit θ un paramètre d'un modèle.
Une variable Y est un estimateur sans biais de θ si $\mathbb{E}(Y) = \theta$.
Une estimation ponctuelle de ce paramètre θ, notée $\widehat{\theta}$, est donnée par la réalisation y de la variable Y sur un échantillon donné. On a donc :

$$\widehat{\theta} = y$$

Propriété 38 : Estimation ponctuelle d'une moyenne
- $\bar{X}$ est un estimateur non biaisé de μ.
- Une estimation ponctuelle $\widehat{\mu}$ de la moyenne μ d'une population est donnée par

$$\widehat{\mu} = \bar{x}$$

En effet : $\mathbb{E}(\bar{X}) = \mu$.

Propriété 39 : Estimation ponctuelle d'une variance
- $\dfrac{SCE}{n-1}$ est un estimateur non biaisé de σ^2.
- Une estimation ponctuelle $\widehat{\sigma}^2$ de la variance σ^2 d'une population est donnée par

$$\widehat{\sigma}^2 = \frac{SCE}{n-1} = \hat{s}^2$$

En effet : $\dfrac{SCE}{\sigma^2} \sim \chi^2(n-1)$ et $\mathbb{E}(\chi^2(n-1)) = n-1$ donc $\mathbb{E}\left(\dfrac{SCE}{\sigma^2}\right) = n-1$.
Ainsi : $\mathbb{E}\left(\dfrac{SCE}{n-1}\right) = \sigma^2$.

Remarque 24 : La variance empirique corrigée $\hat{S}^2 = \dfrac{SCE}{n-1}$ est un estimateur non biaisé de σ^2 alors que la variance empirique $S^2 = \dfrac{SCE}{n}$ est un estimateur biaisé de σ^2.

Remarque 25 : $\hat{s}$ est une estimation biaisée de σ.

Propriété 40 : Estimation ponctuelle d'une proportion
Considérons une certaine population dans laquelle la proportion d'individus ayant une propriété donnée est égale à p.
- $F = \dfrac{X}{n}$ est un estimateur sans biais de p.
- Une estimation ponctuelle de p est donnée par

$$\widehat{p} = f$$

En effet, le variable X dénombrant les individus d'un échantillon de taille n possédant cette propriété suit la loi binomiale $\mathcal{B}(n,p)$.
On a donc : $\mathbb{E}(F) = \mathbb{E}\left(\dfrac{X}{n}\right) = \dfrac{1}{n}\mathbb{E}(X) = \dfrac{1}{n} \times np = p$.

VI.3 Estimation par intervalle de confiance

DÉFINITION 31 : Estimer un paramètre θ par intervalle (symétrique en probabilité), au niveau de confiance $1-\alpha \in]0;1[$, à partir d'un estimateur Y revient à déterminer un réel positif η tel que

$$\mathbb{P}(Y-\eta \leq \theta \leq Y+\eta) = 1-\alpha$$

L'intervalle $[y-\eta; y+\eta]$ est une estimation par intervalle de confiance de θ au niveau de confiance $1-\alpha$.

REMARQUE 26 : Dans le cas d'une distribution normale d'écart-type connu σ, l'intervalle $\left[\overline{x}-1,96\frac{\sigma}{\sqrt{n}}; \overline{x}+1,96\frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de la moyenne μ au niveau de confiance 0,95.
Ce la signifie qu'il est très vraisemblable que la moyenne μ de la population appartienne à l'intervalle $\left[\overline{x}-1,96\frac{\sigma}{\sqrt{n}}; \overline{x}+1,96\frac{\sigma}{\sqrt{n}}\right]$.

VII Généralités sur les tests statistiques

VII.1 Un peu d'histoire : Neyman, Pearson et Fisher

Les tests selon Fisher : 1 hypothèse et 1 probabilité

Dans la théorie des tests de Fisher, une seule hypothèse est testée. Elle est appelée "hypothèse nulle" dans le sens "to be nullified" c'est-à-dire à réfuter et notée $\mathcal{H}_0$.
L'hypothèse $\mathcal{H}_0$ n'est jamais "prouvée" mais peut être rejetée.
La conclusion est basée sur le calcul d'une probabilité conditionnelle (p-valeur sous $\mathcal{H}_0$) correspondant à la probabilité d'avoir une observation égale ou pire à celle que l'on a obtenue. Si cette probabilité est jugée suffisamment faible, on considère qu'on a réussi à montrer la fausseté de l'hypothèse nulle, on la rejette et le résultat est déclaré significatif. Si p est trop élevé, on suspend le jugement et le résultat est déclaré non significatif.
Il n'y a qu'un seul risque d'erreur (α) correspondant à la probabilité de rejeter $\mathcal{H}_0$ à tort. Pour juger de p, on raconte que le seuil de 5 % avait sa préférence car il percevait 5 % de royalties sur ses publications. La p-valeur est centrale. Plus elle est petite, plus on a de preuves contre $\mathcal{H}_0$.
• $p = 0,049$: l'observation est "limite"
• $p = 0,001$: l'observation est "probante"
• $p = 0,049$ et $p = 0,051$: les résultats sont voisins.

Critique :
La p-valeur correspond à la probabilité d'avoir une observation pire (ou égale) à celle que l'on a obtenue (sous $\mathcal{H}_0$).
$\mathcal{H}_0$ peut être rejetée (alors qu'elle est vraie) parce qu'elle est "en contradiction" avec des résultats qui n'ont pas été observés.

Les tests selon Neyman-Pearson : 2 hypothèses et 1 région critique

En 1928 et 1933, Neyman et Pearson introduisent la notion d'hypothèse alternative $\mathcal{H}_0$, induisant l'apparition d'un second risque (de deuxième espèce β). Basée sur la règle de décision par rapport à α (du ressort du chercheur), p non présentée, une seule hypothèse
Si le seuil est fixé à 5% alors
• $p = 0,049$ et $p = 0,001$ conduisent à la même conclusion : rejet de $\mathcal{H}_0$
• $p = 0,049$ et $p = 0,051$ conduisent à des conclusions différentes.

Critique :
Cette méthode, basée sur la notion de région critique (zone de rejet $\mathcal{H}_0$) ne tient pas compte du "degré" de preuve (lié à p).

Les tests des nos jours : l'approche N-P-F

De nos jours, on utilise une méthode hybride basée sur les construction de Neyman, Pearson et Fisher. Les différentes étapes d'un test sont les suivantes :

1. **Hypothèses d'un test**
   La première étape d'un test consiste à déterminer l'hypothèse à tester.
   Cette hypothèse, notée H_0, est appelée hypothèse nulle. Il s'agit, en général, d'une égalité.
   On définit ensuite l'hypothèse qui sera retenue si on rejette $\mathcal{H}_0$.
   Cette hypothèse est appelée l'hypothèse alternative $\mathcal{H}_1$ et se présente souvent sous la forme :
   soit " $\ldots.\neq.\ldots$ " ; on dit que le test est bilatéral.
   soit " $\ldots.>.\ldots$ " ou " $\ldots.<.\ldots..$ " ; on dit alors que le test est unilatéral.

2. **Seuil de signification et choix du modèle**
   Le niveau de signification d'un test est fixé fixé a priori, est fréquemment égal à 5 %.
   Quant à la statistique de test (variable de décision) et la loi associée, elle est définie suivant le paramètre considéré.
   La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse $\mathcal{H}_0$ est

vraie.
Par exemple, dans le cas d'un test de conformité d'une moyenne ($H_0 : \mu = \mu_0$), on prendra, sous réserve de validité : $\dfrac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \hookrightarrow \mathcal{N}(0;1)$.

3. **Calcul de la p-valeur (p-value)**
Une règle de rejet, utilisée par le logiciel R, consiste à calculer la probabilité que la statistique de test soit égale à la valeur observée ou encore plus extrême, tout en supposant que l'hypothèse nulle $\mathcal{H}_0$ est vraie :
on appelle cette probabilité la p-valeur (p-value en anglais) voire probabilité critique.
Elle correspond à la **probabilité d'avoir une observation égale ou "pire" que celle que l'on a obtenue. Dans le cas d'un test bilatéral, la p-value correspond au double de celle qui aurait été obtenue lors de la mise en place d'un test unilatéral.** Nous voyons donc que la p-valeur est une probabilité calculée a posteriori, en fonction des données.

4. **Conclusion**
Si p-valeur $\leqslant 0,05$ *alors il y avait moins de* 5% *de chance que la statistique de test prenne une valeur "pire" que la valeur observée ce qui conduit à remettre en question l'hypothèse initiale. On rejette donc* $\mathcal{H}_0$. Plus généralement :
on rejette $\mathcal{H}_0$ (au profit de $\mathcal{H}_1$) lorsque p-value $\leqslant \alpha$, et le test est effectué au risque α.
Dans le cas contraire, on ne rejette pas $\mathcal{H}_0$.
Si l'hypothèse $\mathcal{H}_0$ est rejetée alors que le seuil de signification considéré était égal à **5 %**, on dit qu'on rejette $\mathcal{H}_0$ de manière significative.

VII.2 Risques de première et de deuxième espèce

Tous les règles de décision acceptent un risque α correspondant au risque de rejeter à tort l'hypothèse $\mathcal{H}_0$, c'est-à-dire le risque de rejeter l'hypothèse $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie. Ce risque s'appelle aussi le *risque de première espèce*.
La règle de décision du test comporte également un deuxième risque, à savoir de celui de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ alors que c'est l'hypothèse $\mathcal{H}_1$ qui est vraie. C'est le *risque de deuxième espèce*.

Les deux risques peuvent se définir ainsi :

$$\begin{aligned} \alpha &= \mathbb{P}(\text{rejeter}\mathcal{H}_0 \,|\, \mathcal{H}_0\text{vraie}) = \text{probabilité de commettre une erreur de première espèce.} \\ \beta &= \mathbb{P}(\text{ne pas rejeter}\mathcal{H}_0 \,|\, \mathcal{H}_1\text{vraie}) = \text{probabilité de commettre une erreur de deuxième espèce.} \end{aligned}$$

Le risque de première espèce α est choisi a priori. Toutefois le risque de deuxième espèce β dépend de l'hypothèse alternative $\mathcal{H}_1$ et on ne peut le calculer que si on spécifie des valeurs particulières du paramètre dans l'hypothèse $\mathcal{H}_1$ que l'on suppose vraie.
Les risques liés aux tests d'hypothèses peuvent se résumer ainsi :

	$\mathcal{H}_0$ est vraie	$\mathcal{H}_0$ est fausse
$\mathcal{H}_0$ n'est pas rejetée	$1-\alpha$	β
$\mathcal{H}_0$ est rejetée	α	$1-\beta$

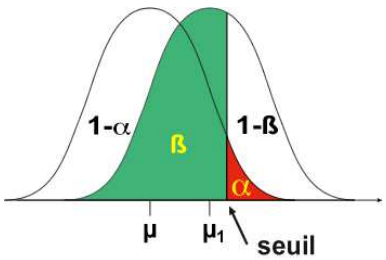
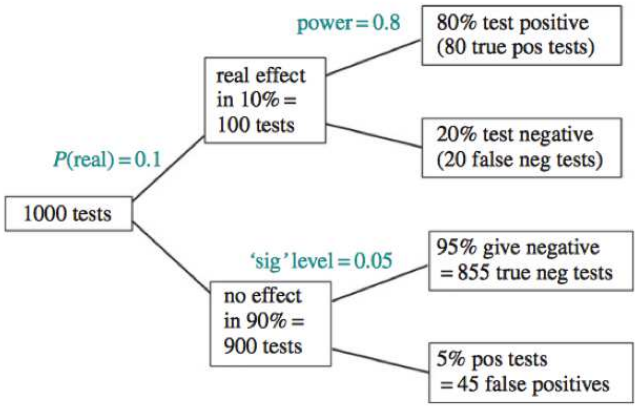


FIGURE 2 – Cas d'un test de conformité d'une moyenne

DÉFINITION 32 : Puissance d'un test
$1-\beta$ définit la *puissance du test* à l'égard de la valeur du paramètre dans l'hypothèse alternative $\mathcal{H}_1$.
La puissance du test représente la probabilité de rejeter l'hypothèse nulle $\mathcal{H}_0$ lorsque l'hypothèse vraie est $\mathcal{H}_1$.
Plus β est petit, plus le test est puissant.

VII.3 Comment utiliser les tests ?

Se limiter à un simple calcul de p-value pour décider d'un éventuel effet est trop réducteur et source de beaucoup d'erreurs.
Par exemple, supposons que sur 1000 tests, seuls 10 % aient un effet réel. On fixe $\alpha = 5\%$ et $\beta = 20\%$ soit $1-\beta = 80\%$.



Ainsi, $80 + 45 = 125$ tests s'avèreront positifs (rejet de $\mathcal{H}_0$) et seulement 80 seront de "vrais" positifs. La probabilité, lorsqu'on a déclaré un effet positif (rejet de $\mathcal{H}_0$) qu'il n'y ait pas d'effet est

$$\frac{45}{125} \simeq 0,36$$

Près de $\frac{1}{3}$ des effets annoncés comme significatifs ne le sont pas ... Il est donc important d'accompagner un test de mesures ou analyses telles que :

- Calcul de la puissance d'un test (ajustement éventuel de la taille d'échantillon)
- Détermination d'intervalles de confiance
- Calculs de la taille de l'effet.

Puissance d'un test

La puissance d'un test correspond à la probabilité de mettre en évidence un effet lorsque celui existe.
Il s'agit donc de la capacité d'un test à prendre la bonne décision lorsqu'il existe un effet.

1. Pour un même risque α et une même taille d'échantillon, on constate que, si *l'écart* Δ entre la valeur du paramètre posée en $\mathcal{H}_0$ et celle supposée dans l'hypothèse vraie $\mathcal{H}_1$ augmente, le risque β diminue ce qui induit une augmentation de $1-\beta$.
2. Une diminution de la *variabilité* (cf. écart-type) peut également induire une augmentation de la puissance du test.
3. Enfin, une *augmentation de la taille du ou des échantillons* aura pour effet de donner une meilleure précision.
   Le test est alors plus puissant.

Il est recommandé de choisir une taille d'échantillon conduisant, a priori, à une **puissance de test au moins égale à 80 %**. Avec R, on peut utiliser des fonctions implémentées de base ('power.t.test' ou 'power.anova.test') ou accessibles via le package 'pwr'.

Taille de l'effet

Pour un grand échantillon, un effet minime (et sans réelle signification) suffira à faire rejeter l'hypothèse $\mathcal{H}_0$.
On peut donc introduire la notion de *taille de l'effet* désignant à quel degré un phénomène donné est présent dans la population (*Revue des sciences de l'éducation, volume 31, 2009*).
Ainsi, **ne pas rejeter l'hypothèse nulle revient à considérer que la taille de l'effet est nulle**.

Soit Δ l'écart entre la moyenne de la population et une valeur cible, ou entre les moyennes de deux populations.
La taille de l'effet est déterminé par

$$\text{Taille de l'effet} = \frac{\Delta}{\sigma}$$

où σ correspond à l'écart-type des populations.
On peut utiliser des niveaux définis par Cohen pour qualifier un effet.
- $d = 0.2$: Effet faible
- $d = 0.5$: Effet moyen
- $d = 0.8$: Effet fort

VII.4 Tests paramétriques de conformité

On considère des échantillons de taille n.

Paramètre	**Conditions**	**Statistique de test et loi**
Moyenne μ	Distribution normale et écart-type σ connu	$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$
	Distribution normale et écart-type inconnu	$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$
	Distribution quelconque et grand échantillon ($n \geqslant 30$)	$\bar{X} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$
Variance σ^2	Distribution normale	$\frac{SCE}{\sigma^2} \sim \chi^2(n-1)$
Proportion p		$X \sim \mathcal{B}(n,p)$
	Grand échantillon	$F \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$

VIII Tests paramétriques de comparaison

On considère deux échantillons de tailles respectives n_1 et n_2.

Paramètres	**Conditions**	**Statistique de test et loi**
Variances	Distributions normales et indépendantes	$\frac{\hat{S}_1^2}{\hat{S}_2^2} \sim \mathcal{F}(n_1-1, n_2-1)$
Moyennes	Distributions normales, indépendantes, variances connues	$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(0; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
	Distributions normales, indépendantes et homoscédastiques	$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{CM}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{T}(n_1 + n_2 - 2)$ $*$
	Distributions normales, indépendantes et homoscédastiques	$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathcal{T}(\nu)$ $**$
	Grands échantillons indépendants	$\bar{X}_1 - \bar{X}_2 \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
	Échantillons appariés, distribution des différences normale	$\frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim \mathcal{T}(n-1)$
Proportions	Grands échantillons indépendants	$F_1 - F_2 \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\left(0; \sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$ $***$

$*$ $CM = \dfrac{SCE_1 + SCE_2}{n_1 + n_2 - 2}$.
Sa réalisation est une "bonne" estimation de la variance commune des deux populations.

$**$ Test t de Welch

$***$ $\hat{p}$ représente une estimation de la proportion commune d'individus pésentant le caractère étudié dans les deux populations.

VIII.1 Tests du Khi-deux, test exact de Fisher et test sur la médiane

- **Test d'indépendance**
  Ce test est utilisé pour tester l'indépendance de deux variables (caractères) X (prenant p modalités) et Y (prenant q modalités) étudiés sur une population à partir d'observations réalisées sur un échantillon.

- **Test d'homogénéité de plusieurs populations**
  Ce test est utilisé pour comparer la distribution d'une variable qualitative (à p modalités) sur q populations indépendantes (à partir d'observations réalisées sur un échantillon).
  Il peut donc être utilisé pour comparer plusieurs proportions sur des échantillons indépendants.

  La statistique de test et la loi utilisée sont, sous H_0,

$$\sum_{i,j} \frac{(N_{ij} - N_{ij}^{th})^2}{N_{ij}^{th}} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\left((p-1)(q-1)\right)$$

- **Test d'ajustement à une loi théorique**
  Ce test est utilisé pour tester si un échantillon dont on a la distribution des effectifs (observés avec répartition en p classes) peut provenir d'une loi donnée (théorique).
  La statistique de test et la loi utilisée sont, sous H_0,

$$\sum_{i} \frac{(N_i - N_i^{th})^2}{N_i^{th}} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2\left(p-1\right)$$

  où $N_i^{th} = Np_i$, p_i correspondant à la probabilité d'avoir l'évènement i et $N = \sum_i N_i$.

- **Test exact de Fisher**
  Le test exact de Fisher est un test d'indépendance (comme le test du Khi-deux) pouvant être utilisé pour tester l'indépendance de deux caractères A et B prenant chacun 2 modalités. Il présente l'avantage d'utiliser une loi exacte (d'où son nom). Les données peuvent se présenter par un tableau de contingence comme ci-dessous :

	$A1$	$A2$	$Total$
$B1$	a	b	$a+b$
$B2$	c	d	$c+d$
$Total$	$a+c$	$b+d$	$a+b+c+d=n$

  La probabilité de cette configuration est donnée, sous l'hypothèse nulle d'absence d'association, par la loi hypergéométrique :

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

  La p-value s'obtient en ajoutant les probabilités de chacun des tableaux aussi éloignés ou plus éloignés de l'indépendance que la table observée.
  Il s'agit alors d'un test bilatéral.

- **Test sur la médiane**
  Ce test permet de tester l'hypothèse selon laquelle deux populations ont la même médiane.
  On note M_e la médiane lorsqu'on regroupe les deux populations.
  Le test est basé sur un test d'indépendance lorsqu'on regroupe les données ainsi :

	$Valeurs > M_e$	$Valeurs \leq M_e$	$Total$
$Echantillon1$	a	b	n_1
$Echantillon2$	c	d	n_2
$Total$	$a+c$	$b+d$	$n_1+n_2=n$

  On peut alors utiliser le test **exact de Fisher** voire un test du **Khi-deux**.

VIII.2 ANOVA à 1 facteur

On rappelle que

$$SCE_{fact} = SCE_{inter} = \sum_{i,j} (\bar{x}_i - \bar{x})^2 \quad SCE_{res} = SCE_{intra} = \sum_{i,j} (x_{ij} - \bar{x}_i)^2 \quad SCE_{totale} = \sum_{i,j} (x_{ij} - \bar{x})^2$$

Sous H_0 : "Il n'y a pas d'effet du facteur étudié", on a

$$F_{obs} = \frac{CM_{fact}}{CM_{res}} \sim \mathcal{F}(p-1; N-p)$$

Arbre de choix autour de la recherche d'un effet factoriel :