TMM1 - Cours de Statistiques

said.qasmi@cerfacs.fr

1 Introduction

Un *série statistique* est une suite de mesures comme, par exemple, la suite des températures maximales Tx mesurées (en °C) à la station de Toulouse-Blagnac lors de l'année 2003 :

01-01-2003 : 10,3 °C 02-01-2003 : 19,2 °C … 31-12-2003 : 6 °C

Par la suite on désignera par x_k ces valeurs, l'indice *k* prenant les valeurs entières de 1 à N (N étant la taille de la *population* ou de l'*échantillon*, ici N = 365).

$x_1 = 10,3$ $x_2 = 19,2$ ... $x_{365} = 6$

Lorsque le nombre de données est important, on peut regrouper celles-ci par classes :

Tx (°C)	[-4;0[	[0;4[	[4;8[	[8;12[	[12;16[	[16;20[	[20;24[	[24;28[	[28;32[	[32;36[	[36;40[	[40;44[
effectifs	3	6	22	49	58	64	38	42	43	23	13	4

Pour comprendre une telle série la première idée est de la représenter graphiquement. Dans notre exemple, un histogramme fournit un bon résumé graphique des données, toujours utile pour commencer l'analyse d'une série.

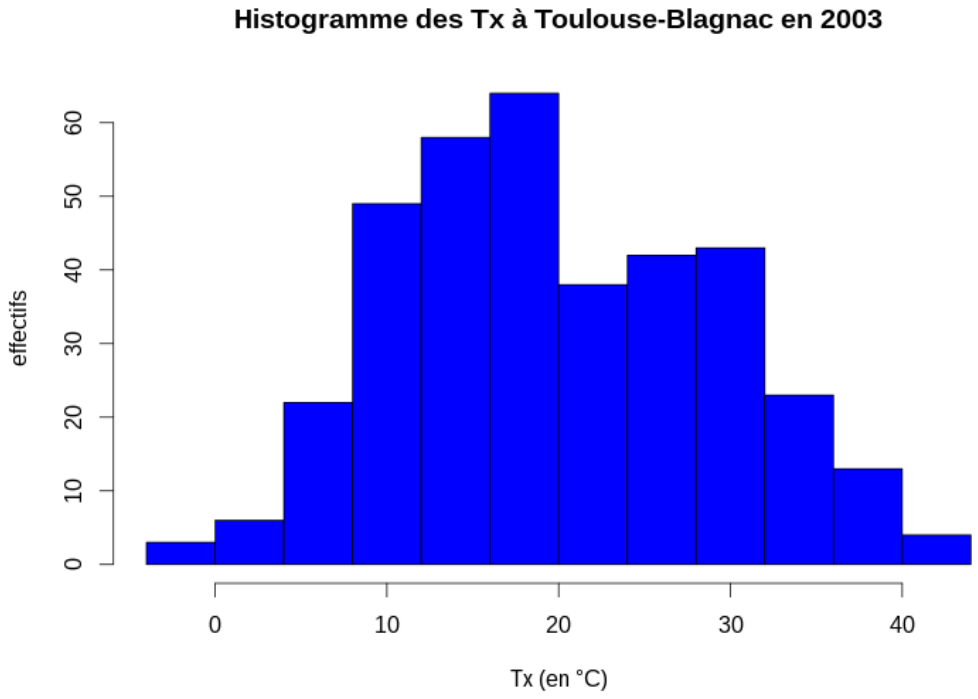


fig.1 : Histogramme des Tx

Cependant, il faudra le plus souvent l'accompagner d'autres résumés, quantitatifs cette fois, dont les plus utilisés sont la moyenne, la médiane, la variance et l'écart-type.

Dans notre exemple, la moyenne des Tx est de 19,93 °C, la médiane 18,5 °C, la variance 85,03 °C² et donc l'écart-type type est de 9,22 °C. La majorité des Tx sont donc comprises entre 10 et 28 °C (moyenne +/- écart-type), la moitié étant inférieure à 18,5 °C (la médiane), et l'autre moitié supérieure.

2 Statistiques descriptives

2.1 Représentation graphique des données

2.1.1 Variables qualitatives

On parle de *variable qualitative* lorsque celles-ci ne prend pas de valeur numérique (par exemple la direction du vent, les candidats à une élection). On représente généralement une variable qualitative à l'aide d'un *diagramme en barre* (ou éventuellement par un *diagramme circulaire*).

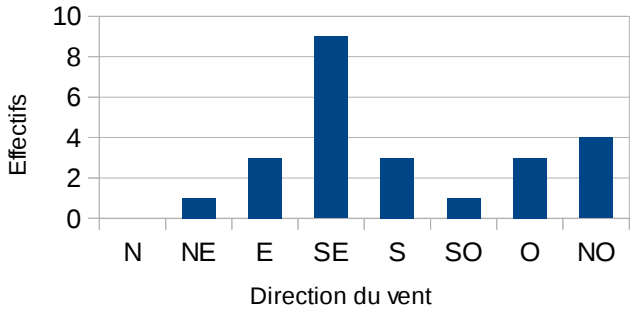


fig.2 : diagramme en barre

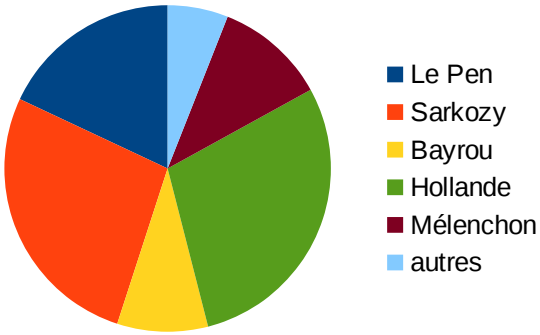


fig.3 : diagramme circulaire

2.1.6 Variables quantitatives

On parle de *variable quantitative* lorsque celles-ci prend des valeurs numériques (par exemple la température, le nombre d'enfants).
On représente généralement une variable quantitative à l'aide d'un *diagramme en bâtons* pour les variables *discrètes* (par ex. le nombre d'enfants), et d'un *histogramme* pour les variables réparties en classes (ce que l'on fera pour les variables *continues*).

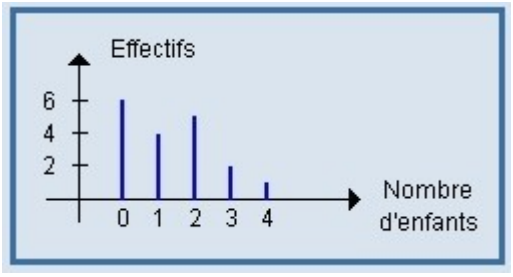


fig.4 : diagramme en bâton

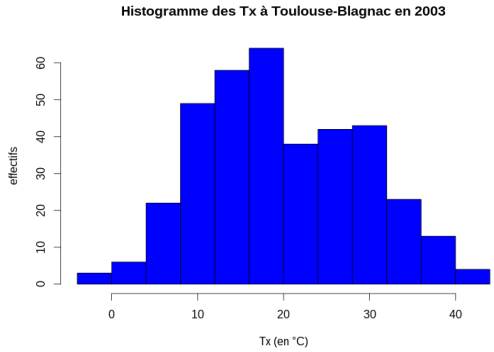


fig.5 : histogramme

2.2 Paramètres descriptifs d'un échantillon (ou d'une population)

En statistique on étudie une *population*, ou, si celle-ci est trop importante, un *échantillon*, c'est-à-dire un sous-ensemble beacoup plus petit, de cette population. L'échantillon doit être bien choisi pour représenter correctement la population.
Par exemple, lorsqu'on désire connaître les intentions de vote des Français, on réalise un sondage auprès d'un échantillon représentatif de la population.

2.2.1 Paramètres de position

Mode : Pour une variable discrète le *mode* est la *modalité* (les valeurs que peuvent prendre les x_k) qui représente le plus grand effectif (exemple : sur la fig.2 le mode est le vent de secteur SE). Pour une variable quantitative continue on parle de *classe modale* : c'est la classe dont l'effectif est maximum (exemple : sur la fig.1 la classe modale est [16 ; 20[).

Moyenne : On considère un ensemble de données $X = \{x_1, x_2, x_3, \dots, x_N\}$.
La *moyenne* (arithmétique) des valeurs de X, notée $\bar{x}$ est :

$$\bar{x} = \frac{x_1+x_2+\dots+x_N}{N} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

Cas d'un tableau d'effectif : lorsque les valeurs sont affectées de coefficients (ici d'effectifs), on parle de *moyenne pondérée*. On utilise la formule ci-dessous :

$$\bar{x} = \frac{n_1 x_1+n_2 x_2+\dots+n_p x_p}{N} = \frac{1}{N}\sum_{k=1}^{p} n_k x_k \qquad \text{avec N l'effectif total : } \quad N = \sum_{k=1}^{p} n_k$$

où $n_1, n_2, \dots, n_N$ sont les effectifs correspondants aux modalités $x_1, x_2, \dots, x_N$, dans le cas discret, ou aux centres des classes dans le cas continue.

Exemple 1 : Calculons la moyenne de la série ci-dessous où N = 20 + 16 + 10 +5 + 1 = 52

xk	0	1	2	3	4
effectifs	20	16	10	5	1

$$\bar{x} = \frac{20\cdot 0+16\cdot 1+10\cdot 2+5\cdot 3+1\cdot 4}{52} \approx 1{,}058$$

Médiane : La *médiane* des valeurs de X, notée *Md*, est la valeur qui partage l'ensemble des valeurs de X en deux parties de même taille. La moitié (50 %) des valeurs de X seront donc plus petites que *Md* alors que l'autre moitié seront plus grande.

Exemple 2: a) X_1 = { 5 ; 8 ; 3 ; 4 ; 7 }

On ordonne les valeurs de X_1 : X_1 = { 3 ; 4 ; **5** ; 7 ; 8 }

La médiane est : $Md_1 = 5$

b) X_2 = { 30 ; 15 ; 25 ; 40 ; 20 ; 30 }

On ordonne les valeurs de X_2 : X_2 = { 15 ; 20 ; **25** ; **30** ; 30; 40 }

La médiane est : $Md_2 = \frac{25+30}{2} = 27{,}5$

Cas d'un tableau d'effectif : Pour déterminer la médiane lorsque l'on a un tableau d'effectif, il suffit de calculer les *effectifs cumulés croissants*.

- Si l'effectif total est *impair*, N = 2k+1, la médiane est x_{k+1} : $Md = x_{k+1}$

- Si l'effectif total est *pair*, N = 2k, la médiane est la moyenne entre x_k et x_{k+1} : $Md = \frac{x_k + x_{k+1}}{2}$

Exemple 3 : On étudie le nombre d'enfants d'un échantillon de 52 individus : N = 52 = 2×26

Nbr d'enfants (x_k)	0	1	2	3	4
effectifs	20	16	10	5	1
eff. cumulés croissants	20	36	46	51	52

de x_1 à x_{20} de x_{21} à x_{36}

La médiane est la modalité «1 enfant» qui correspond à la moyenne des 26ème et 27ème valeurs : (x_{26}) et (x_{27}) .

2.2.2 Paramètres de dispersion

Quartiles : Si on ordonne et partage une série en 4 parties de même taille on obtient les *quartiles.*

- Le *premier quartile*, noté Q1, est la plus petite valeur des termes de la série pour laquelle au moins un quart (25 %) des données sont inférieures ou égales à Q1.
- Le *troisième quartile*, noté Q3, est la plus petite valeur des termes de la série pour laquelle au moins trois quart (75 %) des données sont inférieures ou égales à Q3.
- L'*intervalle interquartile* est l'intervalle [Q1 ; Q3].
- L'*écart interquartile* est le nombre Q3 - Q1.

remarque : La médiane est en fait le deuxième quartile (50 % des valeurs de la série sont inférieures ou égales à Md).

Déciles : Si on ordonne et partage une série en 10 parties de même taille on obtient les *déciles.*

- Le *premier décile*, noté D1, est la plus petite valeur des termes de la série pour laquelle au moins un dixième (10 %) des données sont inférieures ou égales à D1.
- Le *neuvième décile*, noté D9, est la plus petite valeur des termes de la série pour laquelle au moins neuf dixièmes (90 %) des données sont inférieures ou égales à D9.
- L'*intervalle interdécile* est l'intervalle [D1 ; D9].
- L'*écart interdécile* est le nombre D9 - D1.

Boîte à moustaches (ou diagramme de Tukey) : Le *diagramme de Tukey*, ou plus communément appelé, *boîte à moustaches*, est un résumé graphique d'une série. On y fait apparaître la médiane, les premier et

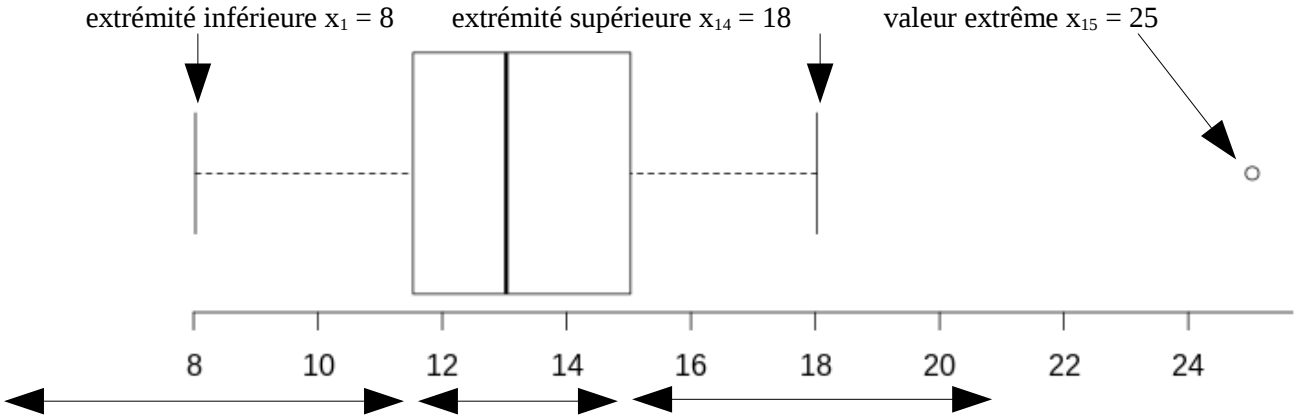
troisième quartiles (Q1 et Q3), et éventuellement les valeurs extrêmes.

Exemple 4 : on étudie la série ordonnée de 15 nombres (N = 15) ci-dessous.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
8	8	9	11	12	12	13	13	13	14	15	15	16	18	25

- *calcul des quartiles :* Q1 : $0{,}25 \times 15 = 3{,}75 \leq 4$ Q3 : $0{,}75 \times 15 = 11{,}25 \leq 12$
donc Q1 = $x_4 = 11$ donc Q3 = $x_{12} = 15$

Délimitation des longueurs des moustaches :
On définit les frontières basse et haute par :
frontière basse FB = Q1 - 1,5*(Q3-Q1) frontière haute : FH = Q3 + 1,5*(Q3-Q1)
- l'extrémité de la moustache inférieure est la plus petite valeur supérieure à la frontière basse.
- l'extrémité de la moustache supérieure est la plus grande valeur inférieure à la frontière haute.



Variance : La *variance* de *X*, notée *V(x)*, est *la moyenne des écarts quadratique à la moyenne* :

$$Var(X) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2$$

Elle mesure la dispersion d'une série autour de sa moyenne.
Pour la calculer, on utilise de préférence la *formule de Huygens* :

$$Var(X) = \frac{1}{n}[\sum_{k=1}^{n} x_k^2] - \bar{x}^2$$

Cas d'un tableau d'effectif : $Var(X) = \frac{1}{N}\sum_{k=1}^{p} n_k (x_k - \bar{x})^2 \overset{Huygens}{=} \frac{1}{N}[\sum_{k=1}^{p} n_k {x_k}^2] - \bar{x}^2$ où $N = \sum_{k=1}^{p} n_k$

Exemple 5 : En réutilisant les données de l'exemple 1, on trouve :

$$Var(X) = \frac{20 \cdot 0^2 + 16 \cdot 1^2 + 10 \cdot 2^2 + 5 \cdot 3^2 + 1 \cdot 4^2}{52} - 1{,}058^2 \approx 1{,}192$$

Ecart-Type : L'*écart-type* des valeurs de X, notée σ_X (*sigma* de X), est la racine carrée de la variance :

$$\sigma_X = \sqrt{(Var(X))} \quad \text{(ou bien } \sigma_X^2 = Var(X) \text{)}$$

L'écart-type est plus commode que la variance pour étudier la dispersion des valeurs la série autour de sa moyenne. En effet, il est exprimé dans la même unité que les x_k.
Plus l'écart-type est faible, plus les valeurs sont regroupées autour de la moyenne.

Exemple 5 (suite) : $\sigma_X \approx \sqrt{(1{,}192)} \approx 1{,}09$ enfants. L'écart-type est important compte tenu des ordres de grandeurs considérés (le nombre d'enfants). La moyenne étant de 1,06 enfants par personnes.

Conclusion : la majorité des personnes ont entre 0 et 2 enfants !

3 Estimation

Si le hasard conduit à un résultat non prévisible, l'observation de plusieurs résultats d'une même expérience aléatoire permettra cependant de choisir judicieusement le modèle aléatoire à retenir. En jetant un dé plusieurs fois consécutivement et, en notant le résultat à l'issue de chaque jet, nous obtenons une suite de nombres entiers compris entre un et six, que nous pourrons appeler *échantillon* de la loi associée à un jet de dé (loi uniforme sur l'ensemble Ω = {1,2,3,4,5,6}). Si on calcule la fréquence observée de chacun de ces chiffres, pour un nombre suffisament élevé de lancers, on obtiendra une distribution empirique (obtenue par l'observation) de valeurs proches les unes des autres et proches de la valeur 16,7 %. Ceci nous oriente donc vers la loi théorique qui est la loi uniforme attribuant la même probabilité 1/6 à chaque chiffres 1,2,3,4,5,6. C'est donc à partir d'observations qui vont constituer ce qu'on appelle un échantillon qu'on pourra déterminer la distribution empirique nous permettant de retenir la distribution théorique qui lui ressemble le plus.

3.1 Échantillonnage

On appelle échantillon de taille *n* d'une variable aléatoire X, une suite $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes et de même loi de probabilité que X (loi que l'on ne connait pas !).

On note μ l'espérance de X et σ son écart-type : $E(X) = \mu$ et $V(X) = \sigma^2$

On cherche donc à estimer ces deux quantités à partir de l'échantillon $(X_1, \ldots, X_n)$.

3.2 Estimateurs de la moyenne et de la variance d'une population

A partir des *n* variables aléatoires X_i, on en construit de nouvelles, appelées *estimateurs*. On va notamment s'intéresser aux estimateurs de moyenne et de variance.

Estimateur de la moyenne : $$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Estimateur de la variance : $$V_n = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M_n)^2$$

Remarque 1 : La variance empirique (c'est-à-dire la variance de l'échantillon) n'est pas exactement égale à la variance théorique, c'est pourquoi on introduit la variance empirique corrigée V_n en divisant par n-1 au lieu de n.

*Remarque 2 : Les estimateurs M_n et V_n sont **sans biais**, en effet $E(M_n) = \mu$ et $E(V_n) = \sigma^2$ (si on moyennise les estimations sur tous les échantillons de taille n on retrouve le paramètre à estimer).*

Remarque 3 : Les estimateurs M_n et V_n sont les meilleurs estimateurs de μ et de σ², dans la mesure où ils tendent vers μ et σ² lorsque n tend vers l'infinie, qu'ils sont sans biais, et que leur écart-type est minimal.

Remarque 4 : En pratique, on parle d'estimations ponctuelles que l'on note avec un ˆ : $\hat{\mu}$ et $\hat{\sigma}$.

3.3 Lois des estimateurs – Intervalles de confiance

3.3.1 Grands échantillons

A partir de la v.a. M_n, on construit la variable centrée réduite (i.e. de moyenne nulle et d'écart-type 1) :

$$Z_n = \frac{M_n - \mu}{\sigma / \sqrt{n}}$$

Rappelons que la loi de probabilité de chaque X_i et donc de M_n et de Z_n est inconnue. Néanmoins on a :

Théorème central limite : ***La loi de probabilité de Z_n tend vers celle de la loi normale centrée réduite N(0,1) lorsque n tend vers l'infini.***

Application : estimation par intervalle de confiance

On ne cherche plus à donner une valeur estimée la meilleur possible de la moyenne ou de l'écart-type mais un intervalle de valeurs dans lequel la vraie valeur (μ ou σ) se trouve avec une probabilité donnée (le coefficient de confiance, dans la pratique 95 % ou 99 %). Si on écrit le coefficient de confiance sour la forme 1-α, α est appelé le « risque » (5 % ou 1 %).

Estimation d'une moyenne μ par intervalle de confiance :
Ainsi, lorsque l'on dispose d'un grand échantillon, on sait d'après le théorème central limite que $T = \frac{M_n - \mu}{\sigma/\sqrt{n}}$ suit approximativement une loi normale centrée réduite. On peut donc trouver, à l'aide par exemple de la table de *N(0,1)*, le nombre t_α tel que : $P(|T| < t_\alpha) = 1-\alpha$.
Les valeurs à retenir : - si α = 5 % alors $t_\alpha = 1{,}96$
- si α = 1 % alors $t_\alpha = 2{,}576$
- si α = 0,1 % alors $t_\alpha = 3{,}29$

On a : $P(|T| < t_\alpha) = P\left(-t_\alpha < \frac{M_n - \mu}{\sigma/\sqrt{n}} < t_\alpha\right) = P\left(M_n - t_\alpha \frac{\sigma}{\sqrt{n}} < \mu < Mn + t_\alpha \frac{\sigma}{\sqrt{n}}\right)$
d'où l'intervalle de confiance à 1-α :

$$\boxed{I_{\mu,1-\alpha} = \left[\ M_n - t_\alpha \frac{\sigma}{\sqrt{n}}\ ;\ M_n + t_\alpha \frac{\sigma}{\sqrt{n}}\ \right]} \qquad \textbf{(I)}$$

exemple : le caractère X d'une grande population suit une loi d'écart-type σ = 2,5 et de moyenne μ inconnue. On considère un échantillon de 100 individus sur lequel la moyenne des valeurs (moyenne empirique) de X est $M_n = 4{,}3$. Un intervalle de confiance pour μ avec un coefficient de confiance de 0,95 (ou un risque d'erreur α = 0,05 = 5 %) est :

$$I_{\mu,0,95} = \left[\ 4{,}3 - 1{,}96 \cdot \frac{2{,}5}{\sqrt{100}}\ ;\ 4{,}3 + 1{,}96 \cdot \frac{2{,}5}{\sqrt{100}}\ \right] = [\ 3{,}81\ ;\ 4{,}79\ ]$$

Remarque : en pratique, on ne connait pas l'écart-type σ, on utilise alors l'estimation $\hat{\sigma}$.

Estimation d'un écart-type par intervalle de confiance.
De la même manière, on peut donner un intervalle de confiance pour estimer l'écart-type σ. Ainsi, lorsque l'on dispose d'un grand échantillon de taille n et de l'écart-type estimé $\hat{\sigma}$ sur celui-ci, l'intervalle de confiance est donné par la formule :

$$\boxed{I_{\sigma,1-\alpha} = \left[\ \hat{\sigma} - t_\alpha \frac{\hat{\sigma}}{\sqrt{2n}}\ ;\ \hat{\sigma} + t_\alpha \frac{\hat{\sigma}}{\sqrt{2n}}\ \right]} \ \textbf{(II)}$$

où les valeurs de t_α sont déterminées comme ci-dessus à partir de la loi normale centrée réduite *N(0,1)*.

Estimation d'une proportion par intervalle de confiance.
Dans une population P, un caractère X ne peut prendre que deux valeurs (1,0) et la proportion (inconnue) de la population vérifiant X = 1 est p. Celle de l'événement contraire est q = 1 - p. On veut donner un intervalle de confiance pour p à partir de son estimation $\hat{p}$. On considère l'estimateur. P_n définie sur l'ensemble des échantillons de taille n par : $P_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (proportion de 1 dans l'échantillon).

Son espérance est p et son écart-type $\sqrt{\frac{pq}{n}}$. On utilise à nouveau le théorème central limite :

Si n est suffisamment grand ($n \geq \frac{3}{pq}$ *), alors la va..* $T = (P_n - p)/\sqrt{\frac{pq}{n}}$ *suit la loi N(0,1).*

En notant les estimations $\hat{p}$ et $\hat{q} = 1 - \hat{p}$, l'intervalle de confiance est donné par la formule :

$$\boxed{I_{p,1-\alpha} = \left[\ \hat{p} - t_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}\ ;\ \hat{p} + t_\alpha \sqrt{\frac{\hat{p}\hat{q}}{n}}\ \right]} \qquad \textbf{(III)}$$

où les valeurs de t_α sont déterminées comme ci-dessus à partir de la loi normale centrée réduite *N(0,1)*.

3.3.2 Petits échantillons

Dans la cadre de petits échantillons, on ne peut plus utiliser de résultats asymptotiques (lorsque n →∞), on est donc obliger d'émettre une hypothèse supplémentaire : la v.a. X étudiée suit une loi normale.

Estimation de la moyenne μ par intervalle de confiance

Lorsque la taille de l'échantillon est petite (n ≤ 30) et que l'on ne connait pas l'écart-type σ de la population, on n'est plus dans le cadre du théorème central limite : $T = \frac{M_n - \mu}{\sigma/\sqrt{n}}$ ne suit plus la loi normale. On montre cependant que, dans ce cas, T suit une loi de Student à ν = n - 1 degrés de liberté (T ~ T_{n-1}). On trouve t_α à l'aide des tables de cette loi.

Par exemple, si n = 10 (ν = 9), on a : - si α = 5 % alors t_α = 2,262
- si α = 1 % alors t_α = 3,25
- si α = 0,1 % alors t_α = 4,781

L'intervalle de confiance est ensuite donné par le même raisonnement et donc les mêmes formules, où μ et σ sont la moyenne et l'écart-type de la population et $\hat{\mu}$ et $\hat{\sigma}$ les estimateurs associés :

1er cas : **on connait l'écart-type σ** de la population, on utilise alors l'intervalle de confiance pour les grands échantillons :

$$I_{\mu,1-\alpha} = \left[\hat{\mu} - t_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \hat{\mu} + t_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

où on lit la valeur de t_α sur la table de la loi normale N(0,1).

2ème cas : **on ne connait pas l'écart-type σ** de la population, mais on l'estimation $\hat{\sigma}$ de celui-ci sur un échantillon.
On peut montrer que T suit une loi de Student à ν = n-1 degré de liberté.
L'intervalle de confiance à 1-α est donc :

$$I_{\mu,1-\alpha} = \left[\hat{\mu} - t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} \; ; \; \hat{\mu} + t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

où on lit la valeur de t_α sur la table de Student à n-1 degré de liberté.

Exemple : on considère un échantillon de n = 10 valeurs dont la moyenne et l'écart-type sont
$\mu_e = 19$ *et* $\sigma_e = 0,5$
Les estimations de la moyenne et de l'écart-type sont :

$$\hat{\mu} = 19 \quad et \quad \hat{\sigma} = \sqrt{\frac{10}{9}} \cdot 0,5 \approx 0,527$$

On ne connait pas l'écart-type σ de la population, ainsi T suit une loi de Student à 9 degrés de liberté. L'intervalle de confiance à 95 % ($t_{0,05}$ = 2,262 avec ν = 9) est :

$$I_{\mu;0,95} = \left[19 - \frac{2,262 \cdot 0,527}{\sqrt{10}} \; ; \; 19 + \frac{2,262 \cdot 0,527}{\sqrt{10}} \right] = [18,623 \, ; 19,377]$$

on peut noter aussi : μ = 19 +/- 0,377

Estimation de l'écart-type σ par intervalle de confiance

Sous cette hypothèse, les intervalles de confiance à 1- α de la variance σ² et de l'écart-type σ sont :

variance : $I_{\sigma^2,1-\alpha} = \left[\frac{\nu \hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \; ; \; \frac{\nu \hat{\sigma}^2}{\chi^2_{\alpha/2}} \right]$ écart-type : $I_{\sigma,1-\alpha} = \left[\sqrt{\frac{\nu}{\chi^2_{1-\alpha/2}}} \, \hat{\sigma} \; ; \; \sqrt{\frac{\nu}{\chi^2_{\alpha/2}}} \, \hat{\sigma} \right]$

où on lit les valeurs $\chi^2_{1-\alpha/2}$ et $\chi^2_{\alpha/2}$ sur la table du χ² *(khi deux)* avec ν = n-1 degré de liberté.

Exemple : on considère le même échantillon qu'à l'exemple précédant en supposant que celui-ci est représentatif d'une v.a. X suivant une loi normale de paramètres μ et σ inconnus.
Ainsi, en lisant sur la table du χ² les valeurs $\chi^2_{1-\alpha/2}$ et $\chi^2_{\alpha/2}$ pour α = 0,05 et ν = 9, on a :
$\chi^2_{0,975} = 19,02$ $\chi^2_{0,025} = 2,7$
et donc l'intervalle de confiance à 95 % pour l'écart-type est :

$$I_{\sigma;0,95} = \left[\sqrt{\frac{9}{19,02}} \cdot 0,527 \; ; \; \sqrt{\frac{9}{2,7}} \cdot 0,527 \right] = [0,36 \, ; 0,96]$$

4 Tests d'hypothèses

Exemple : on veut tester la résistance d'un nouveau modèle de sonde dont le constructeur certifie qu'elle est de $m_0 = 1,5$ Ω avec une précision de +/- 0,01 Ω. Pour cela, on effectue $n = 100$ mesures et on calcule la moyenne m de l'échantillon (qui est une estimation de la moyenne μ de la population).
Question : cette moyenne m est-elle sensiblement différente ou pas de la moyenne théorique m_0 ?
C'est pour répondre à cette question que l'on va effectuer un *test d'hypothèse*.

4.1 Principe d'un test d'hypothèse

• On étudie une population dont les éléments possèdent un caractère (quantitatif ou qualitatif) et dont la valeur du paramètre (moyenne, écart-type, ...) relative au caractère étudié est inconnue.
• Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.
• On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.

Il est bien évident que la statistique (c'est-à-dire la variable d'échantillonnage) servant d'estimateur au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse du fait des fluctuations d'échantillonnage.
Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

La construction d'un test d'hypothèse consiste en fait à déterminer entre quelles valeurs peut varier la variable aléatoire, en supposant l'hypothèse vraie, sur la seule considération du hasard de l'échantillonnage.

4.2 Hypothèses et erreurs

Une hypothèse statistique est une affirmation concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une population.
Le test d'hypothèse a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques : l'hypothèse nulle H_0 et l'hypothèse alternative H_1.
L'hypothèse nulle H_0 est l'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière. N'importe quelle autre hypothèse qui diffère de l'hypothèse H_0 s'appelle l'hypothèse alternative et est notée H_1.

C'est l'hypothèse nulle qui est soumise au test et toute la démarche du test s'effectue en considérant cette hypothèse comme vraie.

Dans notre exemple, on a : H_0 : « $\mu = 1,5\ \Omega$ »
H_1 : « $\mu \neq 1,5\ \Omega$ »

Nous allons établir des règles de décision qui vont nous conduire à l'acceptation ou au rejet de H_0.
Toutefois cette décision est fondée sur une information partielle, les résultats d'un échantillon. Il est donc statistiquement impossible de prendre la bonne décision à coup sûr. La conclusion qui sera déduite des résultats de l'échantillon aura un caractère probabiliste : on ne pourra prendre une décision qu'en ayant conscience qu'il y a un certain risque qu'elle soit erronée. Ce risque nous est donné par le seuil de signification du test (ou risque de première espèce)

Le risque (de première espèce), consenti à l'avance et que nous notons α de rejeter à tort l'hypothèse nulle H_0 alors qu'elle est vraie, s'appelle le seuil de signification du test et s'énonce en probabilité ainsi :

$$\alpha = P\ (\text{ rejeter } H_0 \mid H_0 \text{ vraie })$$

4.3 Statistique de test

La statistique (ou fonction discriminante) du test est une variable aléatoire dont on connaît la loi sous H_0 et dont la valeur observée sera utilisée pour décider du « rejet » ou du « non-rejet » de H_0. *La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse H_0 est vraie.*

4.4 Région critique et région d'acceptation

Au seuil de signification α, on fait correspondre sur la distribution d'échantillonnage de la statistique une région de rejet de l'hypothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité α . Si par exemple , on choisit α = 0,05 , cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de rejet de H_0, bien que H_0 soit vraie et ceci uniquement d'après le hasard de l'échantillonnage.
Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite région d'acceptation de H_0 (ou région de non-rejet) de probabilité $1-\alpha$.

4.5 Mise en œuvre d'un test

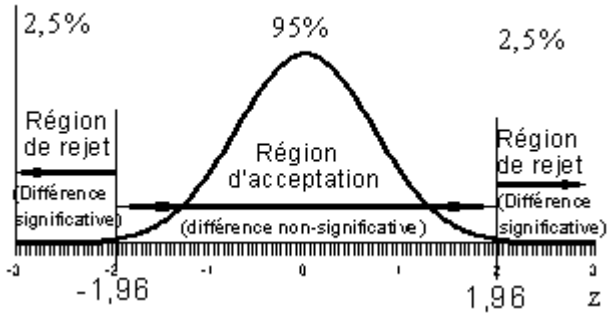
1ère étape : formulation des hypothèses
Dans notre exemple, les hypothèses sont : H_0 : « $\mu = m_0 = 1{,}5\ \Omega$ »
H_1 : « $\mu \neq m_0 = 1{,}5\ \Omega$ »
où μ est la moyenne de la population alors que l'on ne dispose que d'une estimation de celle-ci : la moyenne *m* de l'échantillon.
A partir de là, on effectue le test en considérant H_0 vraie.

2ème étape : détermination de la statistique T (fonction discriminante) du test
On a effectué *n = 100* mesures. On dispose donc d'un grand échantillon. On utilise comme statistique de test la v.a. centrée réduite T ci-dessous :

$$T = \frac{m - m_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

qui, d'après le *théorème central limite*, suit une loi normale centrée réduite N(0,1).

3ème étape : détermination de la région d'acceptation au risque α
On fixe α = 0,05 = 5 %. On cherche donc la valeur tx telle que $P(|T|<t_\alpha) = 1-\alpha \Leftrightarrow P(|T|>t_\alpha) = \alpha$
T ~ *N(0,1)* on a donc t_α = 1,96 pour α = 0,05.



La région d'acceptation au risque α, notée A_α, est donc l'intervalle A_α = [-1,96 ; 1,96].
La région critique au risque α, notée W_α, est ainsi l'intervalle W_α =] -∞ ; -1,96 [∪] 1,96 ; +∞ [.

4ème étape : calcul de la valeur de T prise dans l'échantillon et conclusion
Sur notre échantillon, on a calculé m = 1,495 Ω et l'écart-type, que l'on connait, est σ = 0,01 Ω.
Ainsi la statistique T prend comme valeur sur l'échantillon :

$$u = \frac{m_0 - m}{\sigma/\sqrt{n}} = \frac{1{,}495 - 1{,}5}{0{,}01\sqrt{100}} = -5$$

Conclusion : *u* = -5 n'appartient pas à A_α = [-1,96 ; 1,96] mais à W_α, on rejete donc l'hypothèse H_0 !

$$u \notin A_\alpha \Rightarrow \textit{rejet de } H_0$$

Avec un risque de 5 %, on rejete l'hypothèse H_0, c'est-à-dire que la résistance de notre nouveau modèle de sondes n'est pas égale à 1,5 Ω (elle fluctue beaucoup trop autour de 1,5).

4.6 Exemples de tests

4.6.1 Comparaison d'une moyenne m observée à une moyenne théorique μ_0 , n>30

Hypothèses : H_0 : $\mu = \mu_0$ où μ est la moyenne de la population dont *m* est l'estimation
H_1 : $\mu \neq \mu_0$

Statistique : $T = \frac{m - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$ où *N(0,1)* est la loi normale centrée réduite.

Région d'acceptation : - si α = 5 % alors t_α = 1,96 et A_α = [-1,96 ; 1,96]
- si α = 1 % alors t_α = 2,576 et A_α = [-2,576 ; 2,576]
- si α = 0,1 % alors t_α = 3,29 et A_α = [-3,29 ; 3,29]

Remarque : si on ne connait pas l'écart-type σ, on utilise alors l'estimation $\hat{\sigma}$.

4.6.2 Comparaison d'une moyenne m observée à une moyenne théorique μ_0 , n≤30

On suppose, dans le cas d'un petit échantillon, que ***la population est distribuée selon une loi normale***.

Hypothèses : H_0 : $\mu = \mu_0$ où μ est la moyenne de la population dont *m* est l'estimation
H_1 : $\mu \neq \mu_0$

Statistique : - 1er cas : **on connaît *σ*** l'écart-type de la population : $T = \frac{m - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$

- 2ème cas : **on ne connaît pas *σ*** l'écart-type de la population : $T = \frac{m - \mu_0}{\hat{\sigma} / \sqrt{n}} \sim T_{n-1}$
où T_{n-1} est la *loi de Student à n-1 degrés de liberté*.

Région d'acceptation : - 1er cas : voir 4.6.1 ci-dessus
- 2ème cas : par exemple, pour α = 5 % et n=20 on aura A_α = [-2,093 ; 2,093]

4.6.3 Comparaison d'une variance s^2 observée à une variance théorique σ_0^2

Hypothèses : H_0 : $\sigma^2 = \sigma_0^2$
H_1 : $\sigma^2 \neq \sigma_0^2$
où *σ* est l'écart-type de la population dont l'estimation est $\hat{\sigma} = \sqrt{\frac{n}{n-1}}\, s \Leftrightarrow \hat{\sigma}^2 = \frac{n}{n-1} s^2$

Statistique : - 1er cas : grand effectif (n>30) : $T = \frac{\hat{\sigma} - \sigma_0}{\hat{\sigma} / \sqrt{2n}} \sim N(0,1)$

- 2ème cas : petit effectif (n≤30), on suppose ici que ***la population est distribuée selon une loi normale.***

$$T = \frac{(n-1)\hat{\sigma}^2}{{\sigma_0}^2} \sim {\chi^2}_{n-1}$$

où χ^2_{n-1} est la *loi du Khi deux à n-1 degrés de liberté*.

Remarque : Attention, dans le premier cas (grand échantillon) on raisonne sur l'écart-type σ alors que dans le deuxième cas (petit échantillon) on raisonne avec la variance σ² !

Région d'acceptation : - 1er cas : voir 4.6.1 ci-dessus
- 2ème cas : par exemple, pour α = 5 % et n=20 on aura A_α = [8,91 ; 32,85]
On a lu sur la table du χ^2 les valeurs $t_{\alpha/2}$ et $t_{1-\alpha/2}$ avec 19 degrés de liberté : $t_{0,025}$ = 8,91 et $t_{0,975}$ = 32,85.

5 Comparaison d'échantillons

On veut comparer ici non plus un échantillon avec une population théorique, comme nous l'avons fait dans la partie précédente, mais comparer deux échantillons entre eux. Plus exactement, c'est leurs moyennes et leurs écarts-types que nous comparerons dans les deux cas habituels : grands et petits échantillons.
Il s'agira ainsi de faire à chaque fois un test d'hypothèse sur l'égalité des moyennes ou des variances.

On dispose ainsi de deux échantillons <u>indépendants</u> Ω_1 et Ω_2 de tailles n_1 et n_2, de moyennes μ_1 et μ_2 et d'écart-type σ_1 et σ_2.

5.1 Grands échantillons ($n_1 > 30$ et $n_2 > 30$)

On utilisera ici la loi normale centrée réduite *N(0,1)* pour déterminer la région d'acceptation.

5.1.1 Comparaison de deux moyennes

On effectue un <u>test bilatéral</u> d'égalité des moyennes. Dans ce cas, les hypothèses, la statistique de test et la région d'acceptation sont :

<u>hypothèse :</u> H_0 : $\mu_1 = \mu_2$
H_1 : $\mu_1 \neq \mu_2$

<u>statistique :</u> $T = \dfrac{\mu_1 - \mu_2}{\sqrt{\dfrac{\hat{\sigma_1}^2}{n_1} + \dfrac{\hat{\sigma_2}^2}{n_2}}} \sim N(0,1)$ où *N(0,1)* est la loi normale centrée réduite

<u>région d'acceptation :</u>
- si $\alpha = 5\ \%$ alors $t_\alpha = 1{,}96$ et $A_\alpha = [\ -1{,}96\ ;\ 1{,}96\ ]$
- si $\alpha = 1\ \%$ alors $t_\alpha = 2{,}576$ et $A_\alpha = [\ -2{,}576\ ;\ 2{,}576\ ]$
- si $\alpha = 0{,}1\ \%$ alors $t_\alpha = 3{,}29$ et $A_\alpha = [\ -3{,}29\ ;\ 3{,}29\ ]$

5.1.2 Comparaison de deux variances

<u>hypothèse :</u> H_0 : $\sigma_1^2 = \sigma_2^2$
H_1 : $\sigma_1^2 \neq \sigma_2^2$

<u>statistique :</u> $T = \dfrac{\hat{\sigma}_1^{\,2} - \hat{\sigma}_2^{\,2}}{\sqrt{\dfrac{2\hat{\sigma}_1^{\,4}}{n_1} + \dfrac{2\hat{\sigma}_2^{\,4}}{n_2}}} \sim N(0,1)$

<u>région d'acceptation :</u>
- si $\alpha = 5\ \%$ alors $t_\alpha = 1{,}96$ et $A_\alpha = [\ -1{,}96\ ;\ 1{,}96\ ]$
- si $\alpha = 1\ \%$ alors $t_\alpha = 2{,}576$ et $A_\alpha = [\ -2{,}576\ ;\ 2{,}576\ ]$
- si $\alpha = 0{,}1\ \%$ alors $t_\alpha = 3{,}29$ et $A_\alpha = [\ -3{,}29\ ;\ 3{,}29\ ]$

5.2 Petits échantillons ($\min(n_1,n_2) \leq 30$)

On suppose, dans le cas de petits échantillons, que la <u>population étudiée suit une loi normale</u>.

5.2.1 Comparaison de deux moyennes

On suppose ici que **les variances sont égales : $\sigma_1^2 = \sigma_2^2$**

Ainsi, pour effectuer le test de comparaison des moyennes, il est nécessaire, en cas de petits échantillons, de tester l'égalité des variances en premier (voir test ci-dessous).

<u>hypothèse :</u> H_0 : $\mu_1 = \mu_2$
H_1 : $\mu_1 \neq \mu_2$

statistique : $T = \dfrac{\mu_1 - \mu_2}{\hat{\sigma}\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim T_v$ avec $v = n_1 + n_2 - 2$ et $\hat{\sigma}^2 = \dfrac{(n_1-1)\hat{\sigma}_1{}^2 + (n_2-1)\hat{\sigma}_2{}^2}{n_1+n_2-2}$

où T_v est la *loi de Student* à v degrés de liberté.

région d'acceptation : $A_\alpha = [\, -T_{v,1-\alpha/2} \,;\, T_{v,1-\alpha/2}]$

5.2.2 Comparaison de deux variances

hypothèse : $H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$

statistique : $F = \dfrac{\hat{\sigma}_2{}^2}{\hat{\sigma}_1{}^2} \sim F(v_1, v_2)$ avec $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$

où $F(\nu_1,\nu_2)$ est la loi de Fisher-Snédécor à (ν_1, ν_2) degrés de liberté

région d'acceptation : $A_\alpha = [\, F_{\alpha/2}(v_1, v_2) \,;\, F_{1-\alpha/2}(v_1, v_2)\,]$

Remarque importante : Pour calculer le quantile d'ordre α/2 de la loi de Fisher-Snédécor, on utilise la propriété suivante : $F_{\alpha/2}(v_1, v_2) = \dfrac{1}{F_{1-\alpha/2}(v_1, v_2)}$

5.3 Test unilatéral

Dans certains problèmes, il est plus pertinent de considérer une hypothèse alternative unilatérale du type :

i) $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 < \theta_2$

ou

ii) $H_0 : \theta_1 = \theta_2$ contre $H_1 : \theta_1 > \theta_2$ avec θ un paramètre statistique quelconque

La définition de la région de rejet (et donc de celle d'acceptation) du test dépend de la forme de l'hypothèse alternative :

i) $A = \{t \leq t_\alpha\}$: H_0 accepté et $W = \{t > t_\alpha\}$: H_0 rejeté et H_1 ($\theta_1 < \theta_2$) acceptée

ii) $A = \{t \geq t_\alpha\}$: H_0 accepté et $W = \{t_\alpha < t\}$: H_0 rejeté et H_1 ($\theta_1 > \theta_2$) acceptée

La statistique de test reste la même que dans le cas bilatéral, seul les hypothèses et la région d'acceptation vont changer.

Exemples : *1) comparaison de deux moyennes*

hypothèse : $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 < \mu_2$
région d'acceptation :
- si α = 5 % alors $t_\alpha = 1{,}6449$ et $A_\alpha = [\, -\infty ; 1{,}6449\,]$
- si α = 1 % alors $t_\alpha = 2{,}3263$ et $A_\alpha = [\, -\infty ; 2{,}3263\,]$
- si α = 0,1 % alors $t_\alpha = 3{,}0902$ et $A_\alpha = [\, -\infty ; 3{,}0902\,]$

2) comparaison de deux variances

hypothèse : $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 < \sigma_2^2$
région d'acceptation : si α = 5 % , $\nu_1 = 20$ et $\nu_2 = 20$ alors $t_{1-\alpha} = 2{,}12$
et $A_\alpha = [\, -\infty ; 2{,}12\,]$

Remarque : Attention, si l'on choisit comme hypothèse alternative $H_1 : \sigma_1^2 > \sigma_2^2$ alors on prendra comme statistique de test $F = \dfrac{\hat{\sigma}_1{}^2}{\hat{\sigma}_2{}^2}$ (pour avoir F > 1) *et on lit t_α sur la table de Fisher-Snédécor à (v_2, v_1) degrés de liberté.*