République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université de Saida-Dr. Moulay Tahar.
Faculté des Sciences.
Département de Biologie.

Polycopié de Travaux Dirigés

Biostatistiques pour les 2ième années (LMD)

Résumé de Cours
&
Exercices corrigés .

Dr. Yahiaoui LAHCENE
Courriels :lahceneya8@gmail.com / lahcen.yahiaoui@univ-saida.dz

Année 2020/2021

Table des matières

1 Statistique descriptive univariée **5**

1.1 Généralités 5

1.1.1 Vocabulaire 5

1.2 Représentation graphique d'une variable 7

1.3 Paramètres de position 11

1.3.1 Le mode 11

1.3.2 La médiane 11

1.3.3 Les quartiles 12

1.3.4 Les moyennes 13

1.4 La boite à moustaches (box-and-Wiskersplot) 14

1.5 Paramètres de dispersion : 15

1.5.1 L'étendue 15

1.5.2 L'écart interquartile 15

1.5.3 L'écart type 15

1.5.4 Le coefficient de variation 16

1.6 Exercices corrigés 17

2 Statistique à deux variables **23**

2.1 Présentation des données à deux variables qualitatives 23

2.1.1 Tableau de contingence 23

2.1.2 Tableau des fréquences 24

2.1.3 Profils lignes et profils colonnes 25

2.1.4 Effectifs théoriques et khi-deux 26

2.2 Présentation des données à deux variables quantitatives 26

2.2.1 Covariance 27

2.2.2 Corrélation 27

2.2.3 Driote de régression linéaire 28

2.3 Exercices corrigés 32

3 Quelque lois usuelles **37**
3.1 Variable aleatoire . . . 37
3.1.1 Loi de probabilité . . . 37
3.1.2 Fonction de répartition . . . 38
3.1.3 Espérance mathématique, Moments, Variance mathématique . . . 38
3.2 Quelque lois usuelles discrètes . . . 39
3.3 Quelque lois usuelles continues . . . 40
3.4 Exercices corrigés . . . 44

4 Échantiannage et Estimation **48**
4.1 Le plan d'échantillonnage . . . 49
4.2 Loi d'échantillonage . . . 50
4.3 Estimation ponctuelle . . . 50
4.4 Estimation par intervalle de confiance . . . 52
4.5 Exercices corrigés . . . 57

5 Tests d'hypothèses **62**
5.1 Principe . . . 62
5.2 Tests paramétriques et non paramétriques . . . 68
5.3 Test paramétrique . . . 69
5.3.1 Tests de conformité . . . 69
5.3.2 Tests d'homogénéité . . . 71
5.4 Test non paramétrique "Tests du χ^2" . . . 75
5.4.1 Test d'indépendance . . . 75
5.4.2 Test d'adéquation d'une loi à une loi donnée . . . 76
5.5 Exercices corrigés . . . 78

Préambule

Ce polycopié de Biostatistiques a pour objectif d'initier des étudiants des troncs communs des sciences de la nature et de la vie aux traitements des données liées á leurs thématiques de travail via les biostatistiques, sachant que la statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. C'est un outil essentiel pour la compréhension et la gestion des phénomènes complexes. Les données étudiées peuvent être de toute nature, ce qui rend la statistique utile dans tous les champs disciplinaires et explique pourquoi elle est enseignée dans toutes les filières universitaires, de l'économie à la biologie en passant par la psychologie et bien sûr les sciences de l'ingénieur. La statistique consiste à :

- Recueillir des données.
- Présenter et résumer ces données.
- Tirer des conclusions sur la population étudiée et d'aider à la prise de décision.
- En présence de données dépendant du temps, nous essayons de faire de la prévision.

L'analyse statistique se subdivise en deux parties

✓ **Statistique descriptive** : a pour but de décrire, de résumer ou représenter les données.
Questions typiques

* Représentation graphique

* Paramètres de position, de dispersion, de relation.

✓**Statistique inférentielle** : l'ensemble des méthodes permettant de formuler un jugement. Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).
Questions typiques

* Estimation des paramètres

* Intervalle de confiance

* Tests d'hypothèses

Chapitre 1

Statistique descriptive univariée

1.1 Généralités

Les statistiques se sont développées dans la deuxième moitié du XIXe siècle dans le domaine des sciences humaines (sociologie, économie, anthropologie, ...). Elles se sont dotées d'un vocabulaire particulier.

1.1.1 Vocabulaire

Épreuve statistique
L'épreuve statistique est une expérience que l'on provoque.
Population
On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω.

Individu (unité statistique)
On appelle individu tout élément de la population , il est noté ω (ω dans Ω).
Échantillon
C'est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la taille de l'échantillon.
Caractère (variable statistique)
On appelle caractère (ou variable statistique, dénotée V.S) toute application

$$X : \Omega \longrightarrow C.$$

L'ensemble C est dit : ensemble des valeurs du caractère X (c'est ce qui est mesuré, nombre ou description sur les individus).

Modalités

Les modalités d'une variable statistique sont les différentes valeurs que peut prendre celle-ci.

Types des caractères

Nous distinguons deux catégories de caractères : les caractères **qualitatifs** et les caractères **quantitatifs**.

Caractère qualitatif

lorsque les modalités (ou les valeurs) qu'elle prend sont désignées par des noms.

On distingue deux types de variables qualitatives : les variables qualitatives **ordinales** et les variables qualitatives **nominales**.

Caractère quantitatif

L'ensemble des valeurs est représenté par des valeurs réelles.

On distingue 2 types de variables quantitatives : les variables quantitatives **discrètes** et les variables quantitatives **continues**.

Une variable discrète (discontinue) qui ne prend que des valeurs entières est dite discrète (exemple : nombre d'enfants d'une famille). Elle est dite continue lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini (exemple : diamètre de pièces, salaires...).

Série statistique

On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation. Le nombre d'unités d'observation est noté n. Les valeurs de la variable X sont notées $x_1, ..., x_i, ..., x_n$.

Tableau statistique :

On regroupe toutes les données de la série statistique dans un tableau indiquant la répartition des individus selon le caractère étudié.

- Si le caractère est qualitatif ou discontinu, un groupe contient tous les individus ayant la même modalité (nombre ou catégorie) du caractère.
- Si le caractère est continu, un groupe contient tous les individus ayant les modalités dans un intervalle, cet intervalle s'appelle une classe.

Pour construire ces intervalles, on respecte les règles suivantes :

1. Le nombre de classes est compris entre 5 et 20 (de préférence entre 6 et 12)
2. Chaque fois que cela est possible, les amplitudes des classes sont égales.
3. Chaque classe (sauf la dernière) contient sa borne inférieure mais pas sa borne supérieure..

Dans les calculs, une classe sera représentée par son centre, qui est le milieu de l'intervalle. Une fois la classe constituée, on considère les individus répartis uniformément entre les

deux bornes ce qui entraîne une perte d'informations par rapport aux données brutes. La répartition en classes des données nécessite de définir à priori le nombre de classes J et donc l'amplitude de chaque classe. En règle générale, on choisit au moins cinq classes de même amplitude. Cependant, il existe des formules qui nous permettent d'établir le nombre de classes et l'intervalle de classe (l'amplitude) pour une série statistique de N observations.

- La règle de Sturge :$J = 1 + (3.3 \log_{10}(N))$.
- La règlede Yule :$J = 2.5\sqrt[4]{n}$.

L'intervalle de classe est obtenue ensuite de la manière suivante : longueur de l'intervalle $L = (x_{max} - x_{min})/J$, où x_{max} (resp. x_{min}) désigne la plus grande (resp. la plus petite) valeur observée.

Remarque : Il faut arrondir le nombre de classe J à l'entier le plus proche. Par commodité, on peut aussi arrondir la valeur obtenue de l'intervalle de classe.

A partir de la plus petite valeur observée, on obtient les bornes de classes en additionnant successivement l'intervalle de classe (l'amplitude).

Effectifs, fréquences

On appelle **effectif** d'une modalité (ou une classe) le nombre des individus du groupe qui correspond à cette modalité (ou cette classe) , noté par n_i où i le i-ième modalité (ou classe) .

La fréquence d'une modalité (ou d'une classe) est le rapport de l'effectif cette modalité (ou cette classe) par le nombre total d'observation (l'effectif total $\sum n_i = n$) , noté par f_i où i le i-ième modalité (ou classe), C-a-d : $f_i = \frac{n_i}{n}$.

Effectifs cumulées, fréquences cumulées

Effectif cumulé est l'effectif de la modalité (ou la classe) augmenté de ceux effectifs des modalitées (ou des classes) précédentes (lorsque la variable statistique est quantitative), noté par N_i.

La fréquence cumulée d'une modalité est la fréquence de la modalité (ou d'une classe) augmenté de ceux fréquences des modalités (ou des classes) précédentes (lorsque la variable statistique est quantitative), noté par F_i.

1.2 Représentation graphique d'une variable

Ils servent á visualiser la répartition des individus

✠ Une variable statistique **qualitative** :

On utilise des diagrammes à **secteurs circulaires** (voir 1.1), des **diagrammes en bandes (ou en tuyaux d'orgue)** (voir 1.2). Le principe est de représenter des aires proportionnelles aux fréquences de la variable statistique .

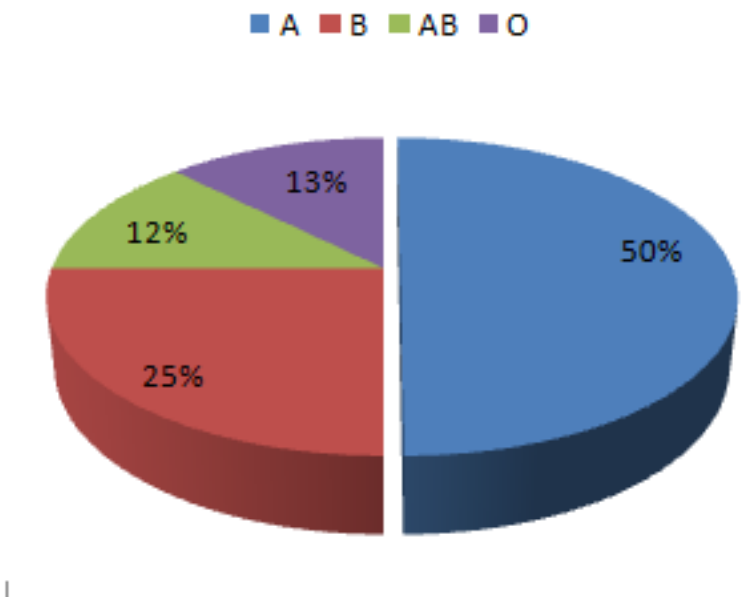


FIGURE 1.1 – Diagramme en secteurs

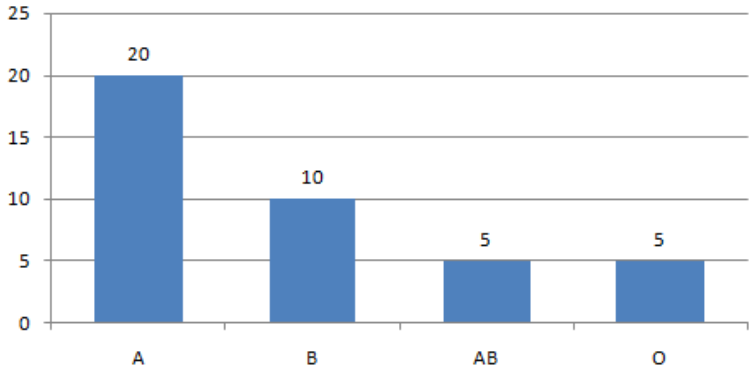


FIGURE 1.2 – Diagramme en tuyaux d'orgue

✠ Une variable statistique **discrète** : On utilise un **diagramme différentiel en bâtons** (voir 1.3), complété du diagramme des fréquences cumulées appelé **diagramme cumulatif** (voir 1.10). Le diagramme cumulatif est la représentation graphique d'une fonction F ($F(x_i) = F_i.$), appelée fonction de répartition de la variable statistique .

✠ Une variable statistique **continue** :

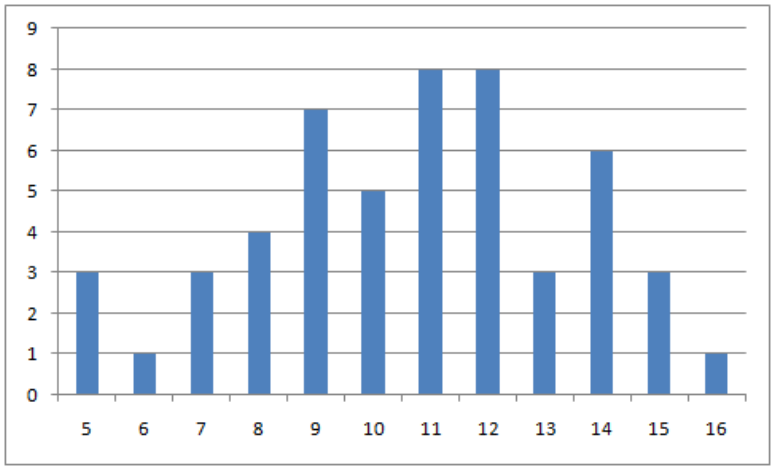


FIGURE 1.3 – Diagramme en bâtons

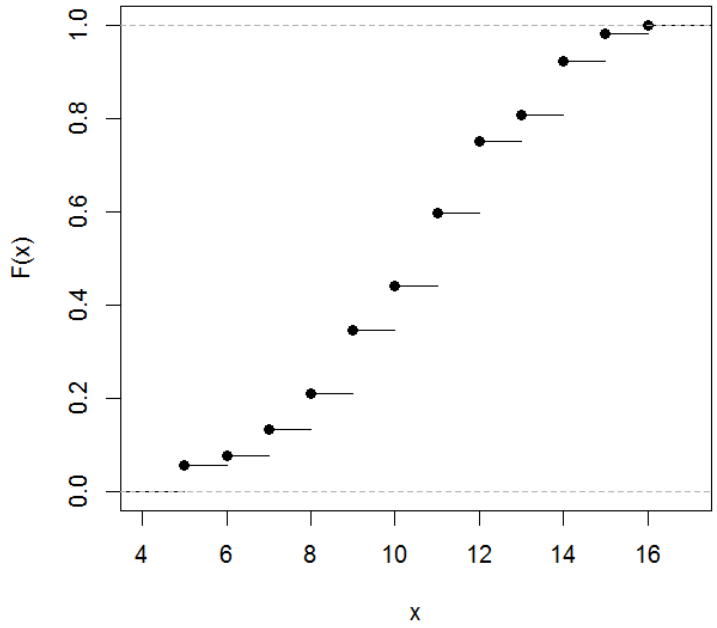


FIGURE 1.4 – Diagramme cumulatif ou la courbe cumulative.

- Le diagramme représentant la série est un **histogramme** : ce sont des rectangles juxta posés dont chacune des bases est égale à l'intervalle de chaque classe et dont la hauteur est telle que l'aire de chaque rectangle soit proportionnelle aux effectifs (**histogramme des effectifs** (voir 1.10))) ou aux fréquences de la classe correspondante (**histogramme des fréquences** .
- On obtient **le polygone des effectifs** (ou **des fréquences** (voir 1.6)) en reliant les milieux des bases supérieures des rectangles.

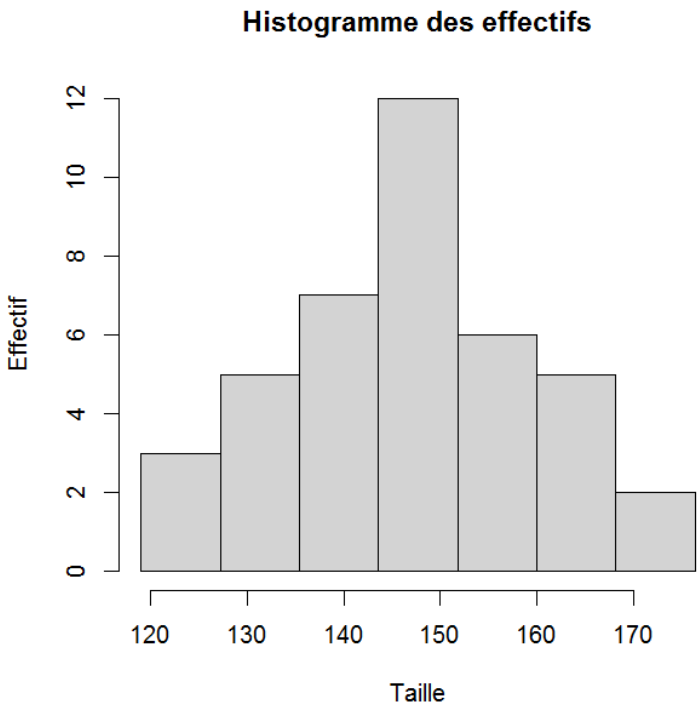


FIGURE 1.5 – Histogramme des effectifs

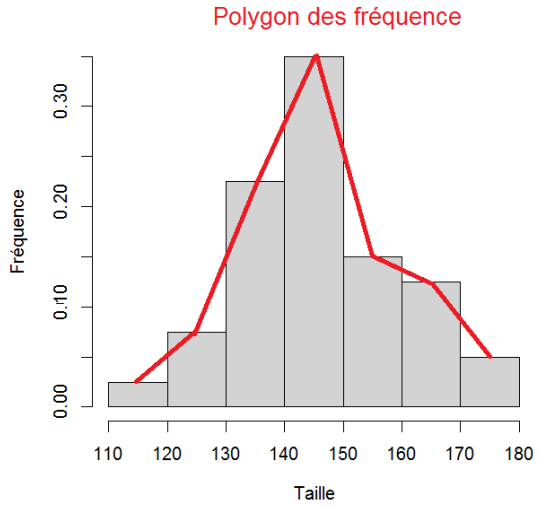


FIGURE 1.6 – Le polygone des fréquences

– **La courbe cumulative** (ou **polygone des fréquences cumulées**) est obtenue en portant les points dont les abscisses représentent la borne supérieure de chaque classe et les ordonnées les fréquences cumulées correspondantes, puis en reliant ces points par des segments de droite, c-a-d la courbe de la fonction de répartition

F (voir 1.7) tel que $F(x) = F_i + \frac{f_{i+1}}{h}(x - a_i), x \in [a_i; a_{i+1}[$.

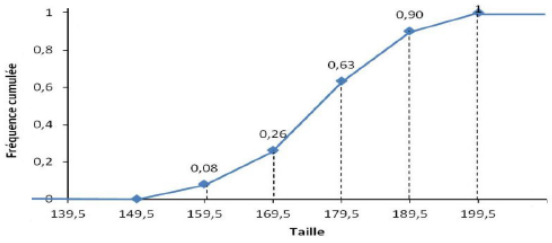


FIGURE 1.7 – La fonction de repartition

1.3 Paramètres de position

1.3.1 Le mode

Le mode noté par M_o correspond à la modalité la plus fréquente. Pour un caractère continu pour le quel les données sont groupées en classes, la classe modale correspond à celle associée à l'effectif (corrigé) le plus élevé, dans ce cas le mode est calculé à partir la méthode suivante : $M_o \in [a; b[$

$$M_o = a + \frac{\Delta_1}{\Delta_1 + \Delta_2}(b - a),$$

tel que :

- Δ_1= Différence entre l'effectif de la classe modale et l'effectif de la classe précédente.
- Δ_2= Différence entre l'effectif de la classe modale et l'effectif de la classe qui suit.

1.3.2 La médiane

La médiane est la modalité qui divise la série des données statistiques en deux parties égales après avoir ranger ces données en ordre croissant (ou décroissant), noté par M_e.
Cas d'un caractère discret : Lorsqu'on possède la série des données brutes et distribution (non groupée), on doit ranger les N observations en ordre croissant.
Si N est impair, la médiane est la $\frac{N+1}{2}$-ième observation. Si N est pair, la médiane est habituellement définie comme étant le point milieu entre la $\frac{N}{2}$-ième et la $\frac{N}{2} + 1$-ième observation.
Cas d'un caractère continue : la médiane est la modalité x qu'elle vérifie

$$F(x) = 0.5$$

Pour calculer la médiane on doit déterminer la classe médiane à partir des fréquences cumulées croissant, puis on calcule la valeur ponctuelle de la médiane selon l'hypothèse de l'uniformité de la répartition des individus à l'intérieur de la classe médiane. voila la formule : si $M_e \in [a; b[$

$$M_e = a + \frac{0.5 - F_{-1}}{f}(b - a),$$

tel que

- F_{-1} : La fréquence cumulée de la classe qui précède la classe médiane,
- f : La fréquence de la classe médiane.

Remarque : La médiane se caractérise par le fait que sa valeur n'est pas influencée par les observations aberrantes ou les observations extrêmes.

1.3.3 Les quartiles

Les quartiles sont des indicateurs qui divisent la distribution en quatre parties égales. Le premier quartile est indicateur noté Q_1 tel que

$$F(Q_1) = 0.25$$

, dans le cas continue calculé par cette méthode : si $Q_1 \in [a; b[$

$$Q_1 = a + \frac{0.25 - F_{-1}}{f}(b - a)$$

tel que

- F_{-1} : La fréquence cumulée de la classe qui précède la classe de premier quartile,
- f : La fréquence de la classe de premier quartile.

Le troisième quartile est noté Q_3 tel que

$$F(Q_3) = 0.75,$$

dans le cas continue calculé par cette méthode : si $Q_3 \in [a; b[$

$$Q_3 = a + \frac{0.75 - F_{-1}}{f}(b - a)$$

tel que

- F_{-1} : La fréquence cummilée de la classe qui pécède la classe de troisième quartile,
- f : La fréquence de la classe de troisième quartile.

1.3.4 Les moyennes

La moyenne est un indicateur de tendance centrale qui permet de déterminer le centre de la distribution, la moyenne arithmétique est la moyenne est la plus utilisée, mais il existe d'autres types de moyennes utilisées dans le calcul de la tendance centrale de distributions statistiques telles que la moyenne géométrique et la moyenne quadratique.

La moyenne arithmétique :

La moyenne arithmétique est la somme de toutes les données observées divisées par le nombre des individus de l'échantillon, c-a-d

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

Si les données sont présentées dans un tableau statistique dans le quel chaque modalité est associée à fréquence absolue ou relative alors on calcule la moyenne arithmétique pondérée ainsi :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i.$$

la moyenne géométrique :

La moyenne géométrique d'une série statistique brute est définie ainsi :

$$\bar{X_g} = \sqrt[N]{\prod_{i=1}^{N} x_i}.$$

Pour des données groupées la moyenne géométrique pondérée est calculée ainsi :

$$\bar{X_g} = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}}$$

Ce type de moyenne est surtout utilisé pour calculer des pourcentages moyens.

La moyenne harmonique :

La moyenne harmonique est la moyenne de l'inverse de la variable x, ou bien l'inverse de la moyenne arithmétique, elle est calculée ainsi pour des données brutes :

$$\bar{X_h} = N \left(\sum_{i=1}^{N} \frac{1}{x_i} \right)^{-1}.$$

Pour des données groupées la moyenne harmonique est égale à :

$$\bar{X_h} = N \left(\sum_{i=1}^{k} \frac{n_i}{x_i} \right)^{-1}.$$

La moyenne harmonique permet de calculer la moyenne des grandeurs obtenues à partir d'un rapport de deux variables tels que le taux de change, l'indice du prix le taux de chômage ...

la moyenne quadratique :

La moyenne quadratique permet de calculer la moyenne des carrés des caractères, pour une série de données brute elle est calculée ainsi :

$$\bar{X_q} = \frac{1}{N} \sum_{i=1}^{N} x_i^2.$$

Lorsque les données sont présentées dans un tableau statistique alors :

$$\bar{X_q} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i^2.$$

Remarque : Dans le cas d'un tableau d'un caractère continu on remplace x_i par le centre de la classe.

L'ensemble des moyennes calculées pour un caractère doivent vérifier l'inégalité suivante :

$$\min x_i \leq \bar{X_h} \leq \bar{X_g} \leq \bar{X} \leq \bar{X_q} \leq \max x_i.$$

1.4 La boite à moustaches (box-and-Wiskersplot)

Ce type de graphique représente de façon simplifiée la dispersion des données contenues dans un vecteur. Il permet donc d'avoir un aperçu de la distribution et de la variabilité des données, et d'identifier les observations aberrantes.

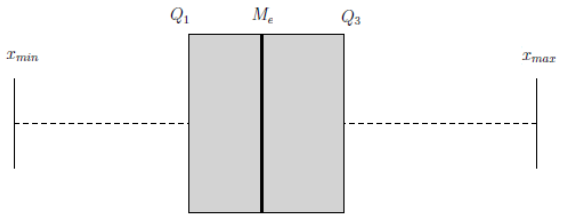
Objectif :

- Résume la série à partir de ses valeurs extrêmes, ses quartiles et sa médiane.
- Permet une comparaison visuelle immédiate de plusieurs séries.

Construction :

- Sur un axe horizontal, on place les valeurs extrêmes et les quartiles.
- on trace un rectangle de longueur l'interquartile $(Q_3 - Q_1)$ et la largeur proportionnelle à la racine carrée de la taille de la série.

– on partage le rectangle par un segment vertical au niveau de la médiane.



1.5 Paramètres de dispersion :

Pour analyser une distribution on peut utiliser en plus des indicateurs de tendance centrale, telles que la médiane ou la moyenne, d'autres indicateurs qui permettent de mesurer la dispersion ou l'éparpillement de la série dans le but de bien décrire la distribution d'une variable.

1.5.1 L'étendue

Définition 1.5.1.1. *L'étendue noté par E est un paramètre qui mesure l'écart entre la valeur la plus élevée et la valeur la plus faible de la distribution :*

$$E = x_{max} - x_{min}.$$

1.5.2 L'écart interquartile

Définition 1.5.2.1. *l'intervalle interquartile est l'intervalle $[Q_1; Q_3[$, cet intervalle contient 50% des observations. L'écart interquartile est l'amplitude de l'intervalle interquartile :*

$$EIQ = Q_3 - Q_1.$$

L'écart interquartile est un indicateur qui a l'avantage d'écarter les observations extrêmes.

1.5.3 L'écart type

Définition 1.5.3.1. *L'écart type noté par σ_x est l'indicateur de dispersion le plus utilisé et le plus simple à interpréter. Il permet de comparer les distributions dont la tendance centrale est identique. Il donne la variation moyenne de la distribution autour de la moyenne arithmétique. Pour calculer l'écart type on doit d'abord calculer la variance de X qui est*

égale à la somme des carrés des écarts à la moyenne divisée par l'effectif n, par la suite l'écart-type est égal à la racine de la variance.
La variance de X est calculée ainsi :
Pour des données brutes la variance est égale à :

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{X}^2.$$

Lorsque les données sont groupées alors :

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \bar{X}^2.$$

1.5.4 Le coefficient de variation

Définition 1.5.4.1. *Lorsqu'on veut comparer la dispersion ou l'étalement de deux séries d'observations qui n'ont pas le même ordre de grandeur ou qui portent sur des variables différentes, on ne peut pas utiliser directement les écarts types. Le coefficient de variation se définit comme le rapport de l'écart type divisé par la moyenne, exprimé en pourcentage, c-a-d*

$$CV = \frac{\sigma_x}{\bar{X}}.$$

1.6 Exercices corrigés

Exercice 01

Parmi ces assertions, préciser celles qui sont vraies, celles qui sont fausses.

1. On appelle variable, une caractéristique que l'on étudie.
2. La tâche de la statistique descriptive est de recueillir des données.
3. La tâche de la statistique descriptive est de présenter les données sous forme de tableaux, de graphiques et d'indicateurs statistiques.
4. En Statistique, on classe les variables selon différents types.
5. Les valeurs des variables sont aussi appelées modalités.
6. Pour une variable qualitative, chaque individu statistique ne peut avoir qu'une seule modalité.
7. Pour faire des traitements statistiques, il arrive qu'on transforme une variable quantitative en variable qualitative.
8. La variable quantitative poids d'automobile peut être reclassée en compacte, intermédiaire et grosse.
9. En pratique, lorsqu'une variable quantitative discrète prend un grand nombre de valeurs distinctes, on la traite comme continue.

Solution. le corrigé en ordre est donné par

1. VRAI
2. FAUX
3. VRAI
4. VRAI
5. VRAI
6. VRAI
7. VRAI
8. VRAI
9. VRAI

Exercice 02

Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population,

Groupes sanguins	A	B	AB	O
L'effectif	20	10	n_3	5

1. Déterminer la variable statistique et son type.
2. Déterminer l'effectif des personnes ayant un groupe sanguin AB.
3. Donner toutes les représentations graphiques possibles de cette distribution.

Solution.

1. La population dans cette étude est les 40 personnes. Donc $N = 40$. La variable statistique est le groupe sanguin des individus et elle est qualitative.
2. L'effectif total est égal à 40. Par conséquent, $N = \sum_{i=1}^{4} n_i$. Alors $40 = 20+10+n_3+5$. Ce qui implique que $n_3 = 5$.
3. Nous avons deux représentations possibles "Tyaux d'orgue" et "Diagramme en secteur".

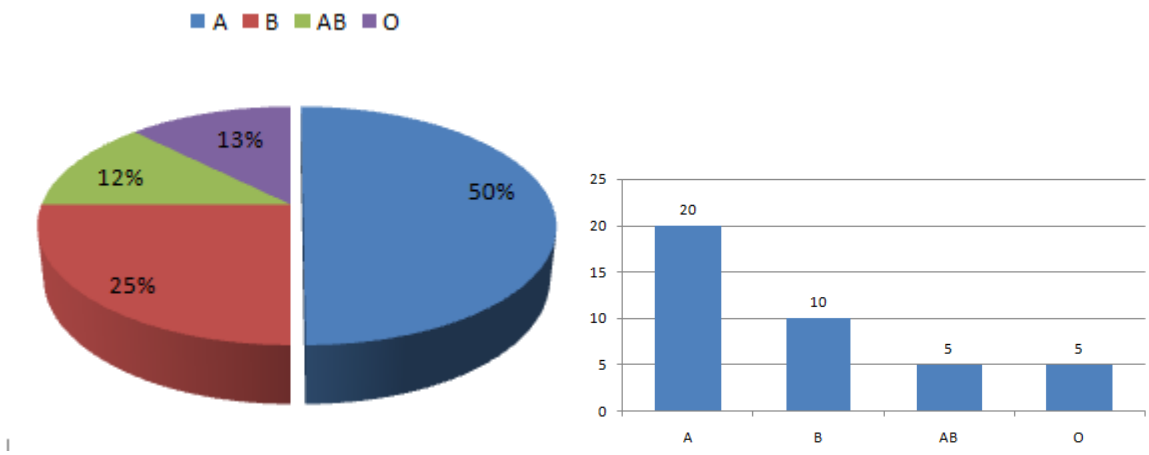


FIGURE 1.8 – A gauche "Diagramme en secteur" et à droite "Tuyaux d'orgue"

Exercice 03

Le gérant d'un magasin vendant des articles de consommation courante a relevé pour un article particulier qui semble connaître une très forte popularité, le nombre d'articles vendus par jour. Son relevé a porté sur les ventes des mois de Mars et Avril, ce qui correspond à 52 jours de vente. Le relevé des observations se présente comme suit : 7; 13; 8; 10; 9; 12; 10; 8; 9; 10; 6; 14; 7; 15; 9; 11; 12; 11; 12; 5; 14; 11; 8; 10; 14; 12; 8; 5; 7; 13; 12 ; 16; 11; 9; 11; ; 11; 12; 12; 15; 14; 5; 14; 9; 9; 14; 13; 11; 10; 11; 12; 9; 15.

1. Quel type est la variable statistique étudiée.
2. Déterminer le tableau statistique en fonction des effectifs, des fréquences, des effectifs cumulés et des fréquences cumulés.
3. Tracer le diagramme des bâtonnés associé à la variable X.

4. Soit F_x la fonction de répartition. Déterminer F_x.
5. Calculer le mode M_o et la moyenne arithmétique $\bar{X}$.
6. Déterminer à partir du tableau puis à partir du graphe, la valeur de la médiane M_e.
7. Calculer la variance et l'écart-type.

Solution.

1. La population est les 52 jours et la variable statistique étudiée est le nombre d'articles vendus par jour. Son type est bien évidement quantitatif discret (nombre).
2. Le tableau statistique est donné par

x_i	5	6	7	8	9	10	11	12	13	14	15	16	$\sum$
n_i	3	1	3	4	7	5	8	8	3	6	3	1	52
f_i	$\frac{3}{52}$	$\frac{1}{52}$	$\frac{3}{52}$	$\frac{4}{52}$	$\frac{7}{52}$	$\frac{5}{52}$	$\frac{8}{52}$	$\frac{8}{52}$	$\frac{3}{52}$	$\frac{6}{52}$	$\frac{3}{52}$	$\frac{1}{52}$	1
N_i	3	4	7	11	18	23	31	39	42	48	51	52	
F_i	$\frac{3}{52}$	$\frac{4}{52}$	$\frac{7}{52}$	$\frac{11}{52}$	$\frac{18}{52}$	$\frac{23}{52}$	$\frac{31}{52}$	$\frac{39}{52}$	$\frac{42}{52}$	$\frac{48}{52}$	$\frac{51}{52}$	1	

3. L'élaboration du diagramme des bâtonnets de X

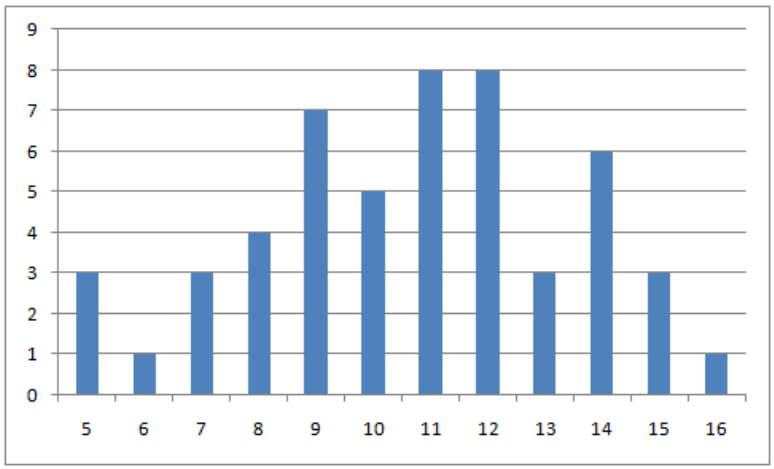


FIGURE 1.9 – Diagramme à bâtons

4. La fonction de répartition est donnée par

$$F_X(x) = \begin{cases} 0 & ;\text{si } x < 5 \\ \frac{3}{52} & ;\text{si } 5 \leq x < 6 \\ \frac{4}{52} & ;\text{si } 6 \leq x < 7 \\ \frac{7}{52} & ;\text{si } 7 \leq x < 8 \\ \ddots & \ddots \\ 1 & ;\text{si } 16 \leq x \end{cases}$$

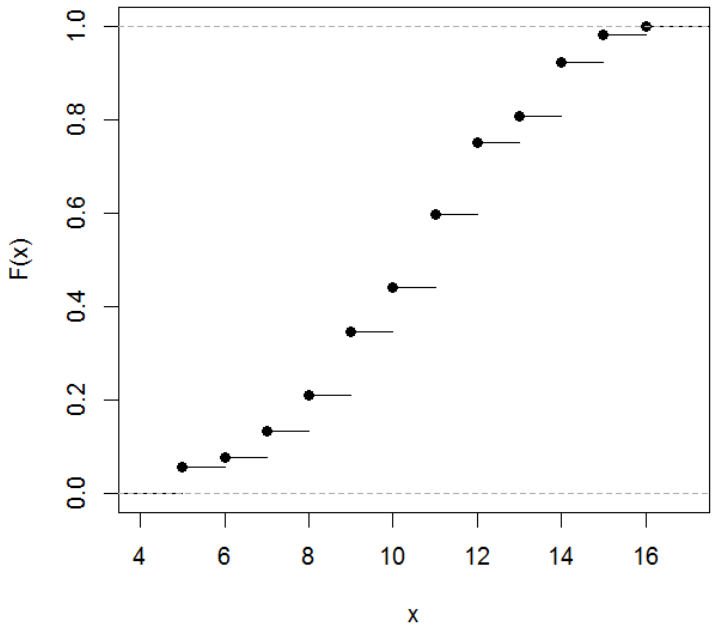


FIGURE 1.10 – Diagramme cumulatif ou la courbe cumulative.

5. Le mode est la valeur de la variable qui a le plus grand effectif, c'est à dire, $n_i = 8$. Donc,

$$M_o = 11 \text{ et } M_o = 12.$$

La moyenne arithmétique est donnée par ;

$$\begin{aligned} \bar{X} = \frac{1}{N}\sum_{i=1}^{12} n_i x_i &= \sum_{i=1}^{12} f_i x_i \\ &= \frac{555}{52} \\ &\simeq 10.67 \end{aligned}$$

6. La médiane est la valeur de la variable qui divise la population de la série statistique en deux parties égales. Nous avons, $N = 52$ est un nombre pair donc

$$M_e = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} = 11.$$

7. Nous commençons par la variance,

$$var(X) = \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{12} n_i (x_i - \bar{X})^2.$$

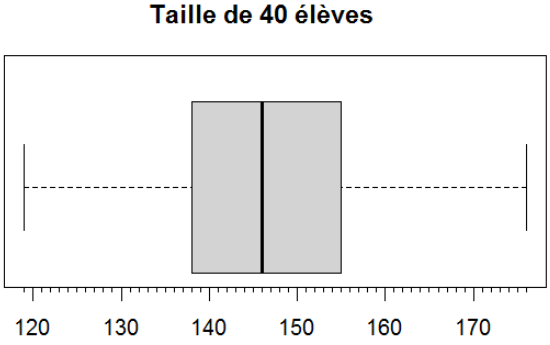
Après calcule, on trouve

$$var(X) = \sigma_X^2 = 7.64.$$

Par conséquent, l'écart type est calculé à partir de

$$\sigma_X = \sqrt{var(X)} = 2.76.$$

Exercice 04



1. Que représente ce graphique ?
2. A partir du graphique, donner la population étudiée, la taille de l'échantillon, la variable ou le caractère, l'étendue de la série statistique, les quartiles.

Solution. :

1. Ce graphique représente **La boite à moustache**
2. – la population étudiée : Les élèves
   – la taille de l'échantillon : 40
   – la variable ou le caractère : La taialle de élève
   – l'étendue de la série statistique : $E = 176 - 119 = 57$
   – les quartiles :

$$Q_1 = 138, M_e = 146, Q_3 = 155$$

Exercice 05

On a mesuré la taille (en cm) de 40 élèves d'une classe et on a obtenu les résultats suivants :

138 164 150 132 144 125 149 157 146 158 140 147 136 148 152 144 168 126 138 176
163 119 154 165 146 173 142 147 135 153 140 135 161 145 135 142 150 156 145 128

1. Regrouper les données en des classes.
2. Représenter graphiquement les données obtenues l'aide d'un histogramme.

Solution.

1. Regroupement les données en des classes :
   On va calculer le nombre des classes par le règle de Yule

$$Nc = 2.5\sqrt[4]{40} = 6.29,$$

   on prend NC=7.
   La longeur de chaque classe :

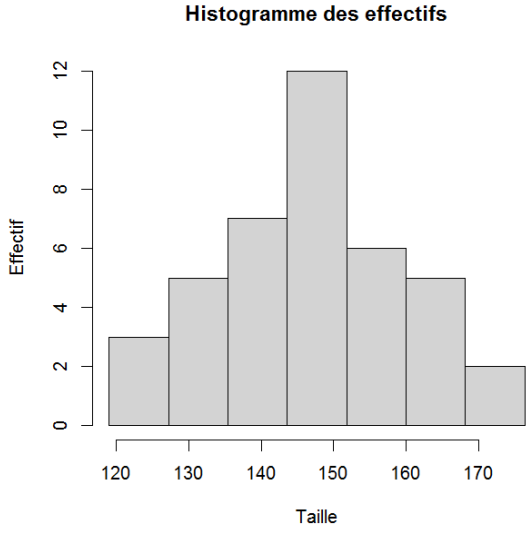
$$L = \frac{X_{max} - X_{min}}{Nc} = \frac{176 - 119}{7} = 8.17,$$

   on prend L=8.2

Classe	[119 ;127.2[	[127.2 ;135.4[	[135.4 ;143.6[	[143.6 ;151.8[	[151.8 ;160[
Effectife	3	5	7	12	6
Fréquence	$\frac{3}{40}$	$\frac{5}{40}$	$\frac{7}{40}$	$\frac{12}{40}$	$\frac{6}{40}$

Classe	[160 ;168.2[	[168.2 ;176.4]
Effectife	5	2
Fréquence	$\frac{5}{40}$	$\frac{2}{40}$

2. Représentation graphiquement l'aide d'un histogramme.



Chapitre 2

Statistique à deux variables

Pour approfondir l'analyse, il est souvent utile de croiser certaines variables entre elles : croiser le niveau de satisfaction avec le sexe (les femmes sont-elles plus satisfaites que les hommes par rapport à ce produit ?), croiser l'âge avec le sexe (quelle est la moyenne d'âge chez les hommes ? Chez les femmes ?), croiser l'âge avec le poids (l'âge est-il corrélé au poids ?).
Les représentations statistiques diffèrent en fonction du type de variables croisées : qualitative/qualitative, qualitative/quantitative, quantitative/quantitative.
L'analyse descriptive bivariée prépare l'inférence statistique : liaison entre variables, corrélation entre variables.

2.1 Présentation des données à deux variables qualitatives

Considérons $X = x_1, x_2, ..., x_l, Y = y_1, y_2, ..., y_m$ deux variables qualitatives ayant respectivement l et m modalités. Les valeurs de ces variables ont été observées sur une population de n individus.

2.1.1 Tableau de contingence

Définition 2.1.1.1. *La répartition des effectifs suivant les modalités de X et de Y, se présente sous forme d'un tableau à double entrée, appelé tableau de contingence ou encore tableau croisé :*

$X\backslash Y$	y_1	$\dots$	y_j	$\dots$	y_m	*Total*
x_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1m}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{im}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_l	n_{l1}	$\dots$	n_{lj}	$\dots$	n_{lm}	$n_{l\bullet}$
Total	$n_{\bullet 1}$	$\dots$	$n_{\bullet j}$	$\dots$	$n_{\bullet m}$	n

- n_{ij} : ***effectif joint****(ou* ***les effectifs croisés observés****) qui se trouve sur la i-ème ligne et la j-ème colonne du tableau de contingence, est le nombre d'individus qui possèdent à la fois la modalité x_i et de la modalité y_j.*
- $n_{i.}$: ***effectif marginal*** *qui se trouve sur la i-ème ligne et la colonne Total est le nombre d'individus qui possèdent de la modalité x_i ; on a donc $n_{i\bullet} = n_{i1} + \ldots + n_{im}$*
- $n_{.j}$: ***effectif marginal*** *qui se trouve sur la j-ème colonne et la ligne Total est le nombre d'individus qui possèdent de la modalité y_j ; on a donc $n_{\bullet j} = n_{1j} + \ldots + n_{lj}$*
- *n : taille de l'échantillon qui se trouve sur la ligne Total et la colonne Total est le nombre d'individus de la population étudiée ; on a donc : $n = n_{1\bullet} + \ldots + n_{l\bullet} = n_{\bullet 1} + \ldots + n_{\bullet m}$*

2.1.2 Tableau des fréquences

Définition 2.1.2.1. *Le tableau de fréquences s'obtient en divisant tous les effectifs par la taille de l'échantillon :*

$$f_{ij} = \frac{n_{ij}}{n}, f_{i\bullet} = \frac{n_{i\bullet}}{n} and f_{\bullet j} = \frac{n_{\bullet j}}{n}, i = 1, ..., l, j = 1, ..., m,$$

c-a-d on a

$X\backslash Y$	y_1	$\dots$	y_j	$\dots$	y_m	*Total*
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1m}	$f_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{im}	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_l	f_{l1}	$\dots$	f_{lj}	$\dots$	f_{lm}	$f_{l\bullet}$
Total	$f_{\bullet 1}$	$\dots$	$f_{\bullet j}$	$\dots$	$f_{\bullet m}$	*1*

La donnée des modalités x_i de la variable X et des fréquences correspondantes $f_{i\bullet}$ (ou encore des effectifs correspondant $n_{i\bullet}$) est appelée distribution marginale de la variable X. La donnée des modalités Y_j de la variable Y et des fréquences correspondantes $f_{\bullet j}$ (ou encore des effectifs correspondant $n_{\bullet j}$) est appelée distribution marginale de la variable Y.

2.1.3 Profils lignes et profils colonnes

Définition 2.1.3.1. *Un tableau de contingence s'interprété toujours en comparant des fréquences en lignes ou des fréquences en colonnes (appelés aussi profils lignes et profils-colonnes).*
Les profils lignes sont définis par

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} \text{ tel que } f_{1|j} + \ldots + f_{l|j} = 1, i = 1, ..., l, j = 1, ..., m,$$

et les profils colonnes par

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} \text{ tel que } f_{1|i} + \ldots + f_{m|i} = 1, i = 1, ..., l, j = 1, ..., m,$$

Le tableau suivant est appelé tableau des profils colonnes

$X \backslash Y$	y_1	$\ldots$	y_j	$\ldots$	y_m	*Total*
x_1	$f_{1\|1}$	$\ldots$	$f_{1\|j}$	$\ldots$	$f_{1\|m}$	$f_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_i	$f_{i\|1}$	$\ldots$	$f_{i\|j}$	$\ldots$	$f_{i\|m}$	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_l	$f_{l\|1}$	$\ldots$	$f_{l\|j}$	$\ldots$	$f_{l\|m}$	$f_{l\bullet}$
Total	*1*	$\ldots$	*1*	$\ldots$	*1*	*1*

ce tableau permet de comparer les profils colonnes (les colonnes) au profil marginal colonne (dernière colonne) etde les comparer entre eux.
Le tableau suivant est appelé tableau des profils lignes

$X \backslash Y$	y_1	$\ldots$	y_j	$\ldots$	y_m	*Total*
x_1	$f_{1\|1}$	$\ldots$	$f_{j\|1}$	$\ldots$	$f_{m\|1}$	*1*
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_i	$f_{1\|i}$	$\ldots$	$f_{j\|i}$	$\ldots$	$f_{m\|i}$	*1*
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
x_l	$f_{1\|l}$	$\ldots$	$f_{j\|l}$	$\ldots$	$f_{m\|l}$	*1*
Total	$f_{\bullet 1}$	$\ldots$	$f_{\bullet j}$	$\ldots$	$f_{\bullet m}$	*1*

ce tableau permet de comparer les profils lignes (les lignes) au profil marginal ligne (dernière ligne) et de les comparer entre eux.

2.1.4 Effectifs théoriques et khi-deux

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien :

Définition 2.1.4.1. *on construit un tableau d'effectifs théoriques qui représente la situation oú les variables ne sont pas liées (indépendance). Ces effectifs théoriques sont construits de la manière suivante :*

$$n_{ij}^* = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Les effectifs observés n_{ij} ont les mêmes marges que les effectifs théoriques n_{ij}^*.
La dépendance du tableau se mesure au moyen du khi-carré défini par

$$\chi^2_{obs} = \sum_{j=1}^{m} \sum_{i=1}^{l} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}.$$

Le khi-carré peut être normalisè pour ne plus dépendre du nombre d'observations. On définit le phi-deux par :

$$\phi^2 = \frac{\chi^2_{obs}}{n}.$$

Le ϕ^2 ne dépend plus du nombre d'observations. Il est possible de montrer que

$$\phi^2 \leq \min(l-1; m-1)$$

Le V de Cramer est définit par

$$V = \sqrt{\frac{\phi^2}{\min(l-1; m-1)}}.$$

Le V de Cramer est compris entre 0 et 1. Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si $V \approx 0$, les deux variables sont indépendantes. Si $V = 1$, il existe une relation fonctionnelle entre les variables, ce qui signifie que chaque ligne et chaque colonne du tableau de contingence ne contiennent qu'un seul effectif différent de 0.

2.2 Présentation des données à deux variables quantitatives

Dans ce cas, chaque couple est composé de deux valeurs numériques. Un couple de nombres (entiers ou réels) peut toujours être représenté comme un point dans un plan $(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)$ que s'appelle **nuage des points**.

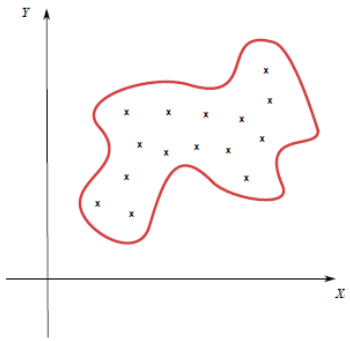


FIGURE 2.1 – Représentation sous forme de nuage de points.

Les variables x et y peuvent être analysées séparément. On peut calculer tous les paramètres dont les moyennes et les variances :

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$$

Ces paramètres sont appelés paramètres marginaux : variances marginales, moyennes marginales, écarts-types marginaux, quantiles marginaux, etc.. . .

2.2.1 Covariance

La covariance est définie

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{X}\bar{Y}$$

Remarque

- La covariance peut prendre des valeurs positives, négatives ou nulles.
- Quand $x_i = y_i$, pour tout $i = 1, ..., n$, la covariance est égale à la variance.

2.2.2 Corrélation

Le coefficient de corrélation est la covariance divisée par les deux écart-types marginaux :

$$r = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

Le coefficient de détermination est le carré du coefficient de corrélation :

$$r^2 = \frac{cov(X,Y)^2}{\sigma_x^2 \sigma_y^2}$$

Remarque

- Le coefficient de corrélation mesure la dépendance linéaire entre deux variables.
- $-1 \leq r \leq 1$
- $0 \leq r^2 \leq 1$
- Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante.
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante.
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.

2.2.3 Driote de régression linéaire

Le problème de l'ajustement d'un ensemble de points représentés dans un système d'axes par une droite, ou plus généralement par une courbe, est essentiel dans le développement de la statistique. Au 18ème siècle, Leonhard Euler et Tobias Mayer développent, indépendamment l'un de l'autre, la méthode des moyennes permettant d'ajuster des points par une droite.
Le premier texte paru faisant mention de la méthode des moindres carrés est dû à Adrien-Marie Legendre dans un article sur ses " nouvelles méthodes pour la détermination des orbites des comètes ", publié en 1805. Un an plus tard, Gauss fait aussi allusion à cette méthode. C'est avec l'apparition de la loi normale que cette méthode va trouver sa justification et va devenir pour longtemps la méthode d'ajustement.
La paternité de la corrélation a donné lieu à une littérature abondante. Signalons simplement que Galton exprime le désir de construire un coefficient de réversion qui se mutera en régression et qu'en 1888 il utilise les termes de "partial correlation " annonçant déjà la corrélation multiple. En 1896, Karl Pearson reprend les concepts de Galton pour leur donner leur forme actuelle. Au 20ème siècle, d'autres mesures d'association allaient naître comme, en 1904, le coefficient de corrélation de rang avec Spearman et la même année la statistique " classique " du chi-deux par Pearson.
Exemple. On désire savoir comment le taux de cholestérol sérique dépend de l'âge chez l'homme. Pour cela on a pris 5 échantillons d'hommes adultes d'ages bien déterminés 25,

35, 45, 55 et 65 ans. On a obtenu les données suivantes :

Ages	25	25	25	35	35	35	45	45	55	65
Taux	1.8	2.3	2.4	2.6	2.9	2.7	3.7	3.3	2.9	2.7

– Que peut-on conclure de ces données ?

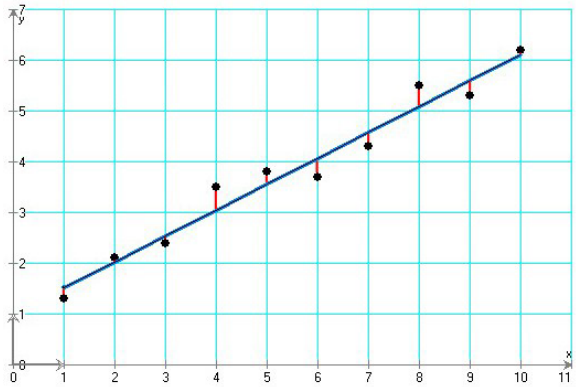
En pratique nous sommes souvent amenés à rechercher une relation entre deux variables x et y. Pour cela, dans un premier temps, nous collectons des données $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Ensuite nous représentons graphiquement ces données.

Ajustement linéaire par la méthode des moindres carrés

La méthode des moindres carrés consiste à déterminer la droite (que l'on appelle aussi droite de régression) telle que la somme des carrés des n valeurs $y_i - \hat{y}_i$ soit minimale (ce qui explique le nom de la méthode).

NB :$\hat{y}_i$ est la coordonnée verticale du point de la droite d'abscisse x_i. Donc $\hat{y}_i = ax_i + b$.

Sur le dessin, chaque trait vertical rouge représente la valeur $y_i - \hat{y}_i$.



On veut donc minimiser la quantité $q = \sum[y_i - (a\ x_i + b)]^2$. Rappelons que la valeur minimale d'une fonction se calcule en posant sa dérivée égale à 0. Pour trouver a et b, calculons cette dérivée. Calculons d'abord la dérivée de q par rapport à a.

$$\begin{aligned} \frac{dq}{da} &= -2\sum((y_i - a\ x_i - b)\ x_i) = 0 \\ \sum x_i\ y_i &= a\sum x_i^2 + b\sum x_i \end{aligned} \tag{2.1}$$

Calculons maintenant la dérivée de q par rapport à b.

$$\begin{aligned} \frac{dq}{db} &= -2\sum(y_i - a\ x_i - b) = 0 \\ \sum y_i &= \sum a\ x_i + \sum b \\ \sum y_i &= \sum a\ x_i + n\ b \\ b &= \bar{y} - a\bar{x} \end{aligned} \tag{2.2}$$

Ce résultat indique que la droite passe par le point moyen $(\bar{x}; \bar{y})$. Introduisons le résultat de (2.2) dans (2.1) pour trouver a :

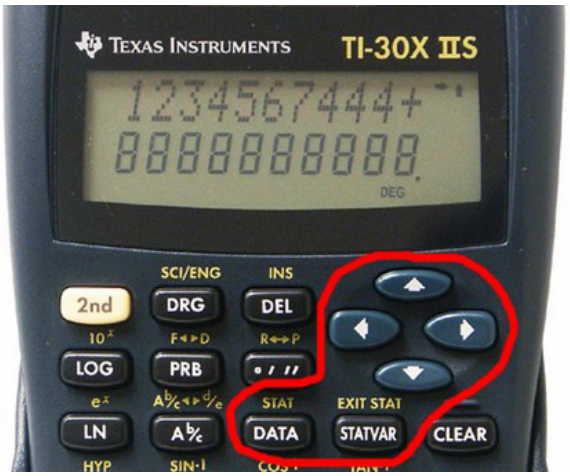
$$\begin{aligned} \sum x_i y_i &= a\sum x_i^2 + (\bar{y} - a\bar{x})\sum x_i \\ \sum x_i y_i &= a\sum x_i^2 + \bar{y}\sum x_i - a\bar{x}\sum x_i \\ a\sum x_i^2 - a\bar{x}\sum x_i &= \sum x_i y_i - \bar{y}\sum x_i \\ a &= \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_i^2 - \bar{x}\sum x_i} \\ a &= \frac{\frac{1}{n}\sum x_i y_i - \bar{y}\bar{x}}{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \frac{cov(x,y)}{\sigma_x^2} \end{aligned} \tag{2.3}$$

L'équation de la droite des moindres carrés

La droite des moindres carrés y = ax + b a pour coefficients :

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

On utilise la calculatrice qui va donner l'équation de la droite cherchée. Avec TI : - Appuyer sur " STAT " puis " Edite " et saisir les valeurs de xi dans L1 et les valeurs de yi dans L2. - Appuyer à nouveau sur " STAT " puis " CALC " et " RegLin(ax+b) " Avec CASIO : - Aller dans le menu " STAT ". - Saisir les valeurs de xi dans List1 et les valeurs de yi dans List2. - Sélectionner " CALC " puis " SET ". - Choisir List1 pour 2Var XList et List2 pour 2Var YList puis " EXE ". - Sélectionner " REG " puis " X " et " aX+b ".



Remarque. Certaines calculatrices ont des fonctions statistiques qui fournissent ces valeurs très rapidement. Consultez le mode d'emploi de votre machine .

Analyse de la variance et qualité d'ajustement

Dans le but d'analyser la qualité d'ajustement et déterminer le pouvoir explicatif de la droite de régression à ajuster la relation entre les variables X et Y, on doit décomposer la variance totale ou bien la variance expliquées par la variable endogène Y en deux types de variabilité ou variances, une expliquées par la droite de régression ou bien variance explicative et un résiduelle expliquée par le terme d'erreur ou l'écart e_i.
La décomposition de la variance est présentée dans l'équation de l'analyse de la variance : La somme des carrés totale= somme des carrés expliquée+ somme des carrés résiduelle.

$$SCT = SCE + SCR$$

- $SCT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}$
- $SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $SCR = \sum_{i=1}^{n} e_i^2$

L'équation de l'analyse de la variance permet de calculer un indicateur d'évaluation de la qualité de l'ajustement appelé **le coefficient de détermination** :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

donc $0 < R^2 < 1$

- Si R^2 tend vers zéro alors l'ajustement est de mauvaise qualité
- Si R^2 tend vers un alors l'ajustement est de bonne qualité

2.3 Exercices corrigés

Exercice 01

On s'intéresse à une éventuelle relation entre le sexe de 200 personnes et la couleur des yeux. Le Tableau suivant reprend le tableau de contingence.

	Bleu	Vert	Marron	Total
Homme	10	50	20	80
Famme	20	60	40	120
Total	30	110	60	200

1. Déterminer tableau des fréquences
2. Déterminer tableau des Profils lignes, des profils colonnes et des effectifs théoriques
3. Déterminer khi-deux, phi-deux, le V de Cramer. conclusion.

Solution.

1. Tableau des fréquences

	Bleu	Vert	Marron	Total
Homme	0.05	0.25	0.10	0.40
Famme	0.10	0.30	0.20	0.60
Total	0.15	0.55	0.30	1.00

2. Tableau des Profils lignes, des profils colonnes et des effectifs théoriques

	Bleu	Vert	Marron	Total
Homme	0.13	0.63	0.25	1.00
Famme	0.17	0.50	0.33	1.00
Total	0.15	0.55	0.30	1.00

	Bleu	Vert	Marron	Total
Homme	0.33	0.45	0.33	0.40
Famme	0.67	0.55	0.67	0.60
Total	1.00	1.00	1.00	1.00

	Bleu	Vert	Marron	Total
Homme	12	44	24	80
Famme	18	66	36	120
Total	30	110	60	200

3. khi-deux, phi-deux, le V de Cramer.
   - Le khi-deux observé vaut $\chi^2_{obs} = 3.03$.
   - Le phi-deux vaut $\phi^2 = 0.01515$.

– Comme le tableau a deux lignes $\min(I-1, J-1) = \min(2-1, 3-1) = 1$, Le V de Cramer est égal à $\sqrt{\phi^2} = 0.123$.

conclusion :

La dépendance entre les deux variables est très faible.

Exercice 02

Les criquets ont un organe spécial sur leurs ailes avant qui produit un son lorsqu'ils frottent leurs ailes les unes contre les autres. En règle générale, plus la température de l'air est élevée, plus ils frottent leurs ailes rapidement. La relation entre la température (notée Y et mesurée en degré Celsius) et le nombre de pulsations par seconde (notée X) est bien approchée par une droite de régression (chaque espèce a sa droite propre). On a relevé les mesures suivantes :

x_i	15	17	20	21	23	24	27	28	30	32	34
y_i	13.5	14.1	14.5	14.4	16.3	15.5	17.1	17.8	18.2	20.2	20.1

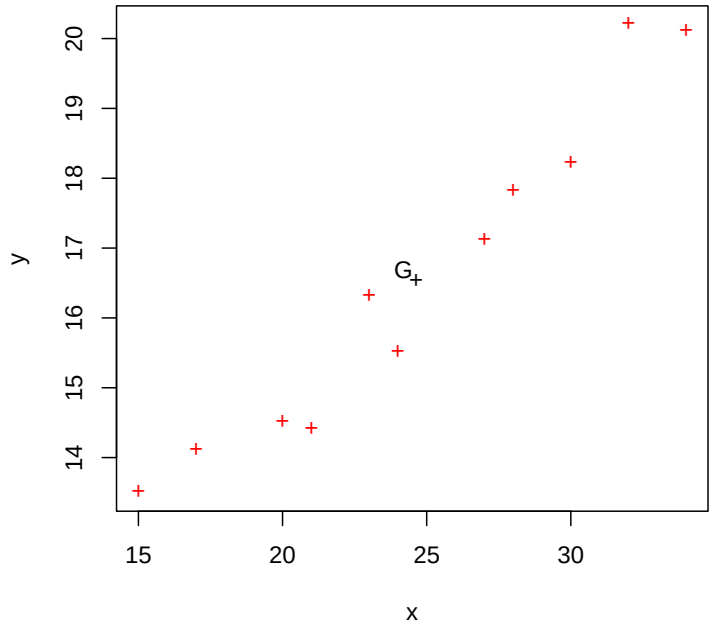
1. Calculer les coordonnées du point moyen $G(\overline{X}; \overline{Y})$.
2. Dans un repère, représenter le nuage de points $(x_i; y_i)$ et le points G.
3. Calculer le coefficient de corrélation r.
4. Déterminer une équation de la droite d'ajustement par la méthode des moindres carrés, puis représenter la droite d'ajustement de y en x.

Solution.

1. les coordonnées du point moyen $G(\overline{X}; \overline{Y})$:

$$\overline{X} = \frac{1}{n}\sum x_i = 24.64,\ \overline{Y} = \frac{1}{n}\sum y_i = 16.52$$

2. La représentation du nuage de points $(x_i; y_i)$ et le points G.



3. Le coefficient de corrélation r

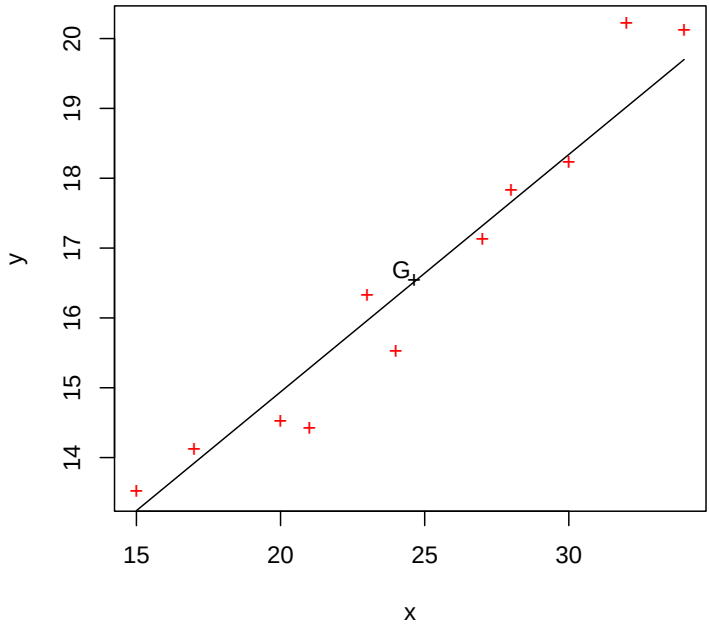
   - $Var(X) = \frac{1}{n}\sum x_i^2 - \overline{X}^2 = 37.70,\ \sigma_X = \sqrt{Var(X)} = 6.14,$
   - $Var(Y) = \frac{1}{n}\sum y_i^2 - \overline{Y}^2 = 4.97,\ \sigma_Y = \sqrt{Var(Y)} = 2.23,$
   - $COV(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{XY} = 12.82,$
   - $r = \frac{COV(X,Y)}{\sigma_X * \sigma_Y} = 0.88$

4. Léquation de la droite d'ajustement $y = ax + b$ par la méthode des moindres carrés :

$$a = \frac{COV(X,Y)}{Var(X)} = 0.34 \text{ et } \ b = \overline{Y} - a\overline{X} = 8.14$$

$$y = 0.34x + 8.14$$

   - Représentation de la droite d'ajustement :



Exercice 03

On désire avoir s'il existe une relation entre le poids de naissance d'un enfant et âge de sa mère à l'accouchement. Dans ce but, on prélève 100 dossiers médicaux dans le fichier des naissances d'une maternité. On calcule les quantités suivantes :

	Poids de naissance de l'enfant	L'age de la mère à l'accouchement
Moyenne observée	3100	25
Variance observée σ^2	10000	25

La covariance observée entre le poids de naissance de l'enfant et l'âge de la mère à l'accouchement :

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = 450$$

1. Donner la valeur numérique du coefficient de corrélation entre le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement.
2. Calculer la droite de régression linéaire entre le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement.
3. La qualité du modèle est jugée par le coefficient de détermination de la régression, exprimer leur valeur.

Solution. :

1. la valeur numérique du coefficient de corrélation entre le poids de naissance de l'enfant et l'âge de sa mère à l'accouchement est :

$$r = \frac{COV(X,Y)}{\sigma_X \times \sigma_Y} = \frac{480}{100 \times 5} = 0.96$$

2. la droite de régression linéaire $y = a \times x + b$ entre le poids de naissance de l'enfant (y) et l'âge de sa mère à l'accouchement (x) est :

$$a = \frac{COV(X,Y)}{\sigma_X^2} = \frac{480}{25} = 19.2 \text{ et } b = \bar{Y} - a \times \bar{X} = 3100 - 19.2 \times 25 = 2620,$$

d'ou

$$y = 19.2 \cdot x + 2620$$

3. le coefficient de détermination de la régression est :
on a

$$R^2 = \frac{SCE}{SCT} = 0.92$$

tellque

- $SCT = \sum_{i=1}^{n} y_i^2 - n\bar{y} = n \times \sigma_Y^2 = 1000000$
- $SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = a^2 \sum_{i=1}^{n} (x_i - \bar{X})^2 = 19.2^2 \times 100 \times 25 = 921600$

R^2 tend vers un alors l'ajustement est de bonne qualité

*

Chapitre 3

Quelque lois usuelles

3.1 Variable aleatoire

Il est courant que la valeur soit associée à tout résultat d'une expérience aléatoire. Le concept de variables aléatoires est une formalisation mathématique de cette événement.

Définition 3.1.1. *Soit Ω un ensemble fondamental associe à une épreuve. On appelle variable aléatoire toute application X définie sur Ω à valeurs numériques.*

Notation
Si X est une variable aléatoire définie sur l'ensemble fondamental Ω relatif à une épreuve , et si x est un nombre réel , alors on pose :

$$\{X \leq x\} = \{\omega \in \Omega / X(\omega) \leq x\}$$

Si I est une partie de $\mathbb{R}$, on pose :

$$\{X \in I\} = \{\omega \in \Omega / X(\omega) \in I\}$$

3.1.1 Loi de probabilité

Définition 3.1.1.1. *on appelle la loi de probabilité de la variable aléatoire X définie sur Ω, la donnée des probabilités $\mathbb{P}(X \in I)$ de tout partie I de $\mathbb{R}$.*

1. **Variable aléatoire discrète**

   Soit $X(\Omega) = \{x_1, x_2, ..., x_n\}$. La loi de probabilité de X est entièrement déterminée par la donnée des $p_i = \mathbb{P}(X = x_i)$ pour $i = 1, ..., n$. On a
   – $p_i \geq 0, \forall i \in \{1, ..., n\}$

– $\sum_{i=1}^{n} p_i = 1$

Remarque. Si $X(\Omega)$ dénombrable et infini, on prend n tend vers infini

2. **Variable aléatoire continue**
Soit $X(\Omega)$ une reunion d'intervalles de $\mathbb{R}$. La loi de probabilité de X est définie par la donnée des probabilités $\mathbb{P}(X \leq x)$, pour tout $x \in \mathbb{R}$.

3.1.2 Fonction de répartition

Définition 3.1.2.1. *Si X est une variable aléatoire définie sur Ω, alors la fonction de repartition de X est l'application F définie sur $\mathbb{R}$ de la manière suivante :*

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x)$$

Définition 3.1.2.2. *Si F la fonction de répartition de la variable aléatoire continue X est dérivable en tout point $x \in \mathbb{R}$, de dérivée $f(x)$, sauf éventuellement en un nombre limité de points, et :*

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

on dit que X est une variable aléatoire absolument continue, et f est appelée la densité de probabilité de X.

3.1.3 Espérance mathématique, Moments, Variance mathématique

Espérance mathématique

1. **Variable aléatoire discret**
Soit X une variable aléatoire discret sur $X(\Omega) = \{x_1, x_2, ..., x_n\}$. appelle espérance mathématique, ou moyenne, de X le nombre

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_1 \mathbb{P}(X = x_i)$$

2. **Variable aléatoire continue**
Si X est une variable aléatoire absolument continue de densité de probabilité f, on appelle espérance mathématique de X le nombre :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

lorsque l'intégrale est convergente.

Moments d'une variable aléatoire
On appelle moment d'ordre k, $k \in \mathbb{N}$, d'une variable aléatoire X, le nombre m^k défini par :

$$m^k = \mathbb{E}(X^k)$$

Variance mathématique On appelle la variance mathématique, d'une une variable aléatoire X, le nombre $V(X)$ défini par

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

3.2 Quelque lois usuelles discrètes

1. **Loi de Bernoulli.**
   Une v.a. X suit une loi de Bernoulli de paramètre $p \in [0; 1]$ si elle ne prend que les deux valeurs 0 et 1 avec

   $$\mathbb{P}(X = 1) = p;\ \mathbb{P}(X = 0) = 1 - p = q.$$

   Son espérance est $\mathbb{E}[X] = p$. Sa variance est $Var(X) = \sigma_X^2 = p \cdot q$.

2. **Loi binomiale**
   Considérons l'expérience qui consiste à répéter n fois une expérience aléatoire (expérience de Bernoulli) de façon indépendante telle que le résultat de chaque expérience est un succès ou un échec avec une probabilité de succès p. Posons X la variable aléatoire qui donne le nombre total de succès sur les n tentatives. La variable aléatoire X suit une loi Binomiale de paramètres n et $p \in [0; 1]$, notée $\mathcal{B}(n, p)$.
   Le support de cette variable aléatoire est $X(\Omega) = \{0, 1, 2, \ldots, n\}$, et la loi de probabilité est donnée par

   $$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{n-k}$$

   où

   $$C_n^k = \frac{n!}{k!(n-k)!}$$

   Son espérance est $\mathbb{E}[X] = n \cdot p$. Sa variance est $Var(X) = \sigma_X^2 = n \cdot p \cdot q$.

3. **Loi de Poisson**
   Considérons X la v.a. qui donne le nombre d'événements observés dans une unité de temps. Ces événements sont vérifient les suivants :

(a) un seul événement arrive à la fois.

(b) le nombre d'événements se produisant ne dépend que du temps de l'observation.

(c) les événements sont indépendants.

On a alors la variable aléatoire X suit une loi de Poisson, notée $X \sim \mathcal{P}(\lambda)$, où λ est le nombre moyen d'événements par unité de temps.

Les valeurs possibles de la variable aléatoire sont $X(\Omega) = \{0, 1, 2, \ldots\}$, et la loi de probabilité est donnée par

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda), \text{ pour } k = 0, 1, 2, \ldots$$

Son espérance est $\mathbb{E}[X] = \lambda$. Sa variance est $Var(X) = \sigma_X^2 = \lambda$.

Remarque. Lorsque le nombre d'épreuves n est grand et très petit (proche de 0) la loi Binomiale $\mathcal{B}(n, p)$ tend vers une loi de Poisson $\mathcal{P}(\lambda)$ de seul paramètre λ (espérance et variance de la loi binomiale approchée par la loi de Poisson). La loi de Poisson est une distribution discrète. Elle est tabulée

3.3 Quelque lois usuelles continues

1. **Loi Normale**
   On dit que la v.a. X suit une loi Normale $\mathcal{N}(\mu; \sigma)$ si elle a une fonction densité comme suit :

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

   Son espérance est $\mathbb{E}[X] = \mu$. Sa variance est $Var(X) = \sigma^2$.

   – **La distribution Normale centrée réduite**
   On dit que la distribution est centrée si son espérance μ est nulle ; elle est dite réduite si sa variance σ^2 (et son écart-type σ) est égale à 1.

   Remarque.

   – Soit Z suit une loi Normale $\mathcal{N}(0; 1)$ et Φ la fonction de répartition, comme la fonction Φ est symétrique par rapport à l'axe $(x = 0)$ alors

$$\Phi(-z) = 1 - \Phi(z).$$

   – Les probabilités $\Phi(a) = P(X \leq a)$ ont été calculées et regroupées dans une table numérique.

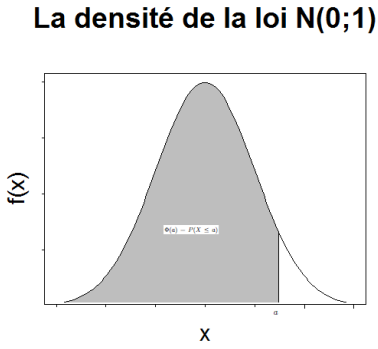


FIGURE 3.1 – La densité de loi Normale $\mathcal{N}(0;1)$

Remarque. Lorsque le nombre d'épreuves n est grand la loi Binomiale $\mathcal{B}(n,p)$ approchée par loi Normale $\mathcal{N}(n\cdot p;\sqrt{n\cdot p(1-p)})$ sous la condition $\min(n\cdot p, n\cdot (1-p))>5$.

– **Transformation d'une loi Normale quelconque en loi Normale centrée réduite**
Soit F la fonction de répartition de loi normal $\mathcal{N}(\mu;\sigma)$, pour calcul $F(a)=P(X\leq a)$, on pose$X=\mu+\sigma Z$ alors

$$Z=\frac{X-\mu}{\sigma}$$

où Z suit une loi normal $\mathcal{N}(0;1)$. on a

$$F(a)=P(X\leq a)=\Phi(\frac{a-\mu}{\sigma})$$

2. **Loi du khi-deux**
Soient $Z_1,\ldots,Z_n$ des v.a. indépendantes de même loi $\mathcal{N}(0;1)$. Posons $\chi^2=\sum_{i=1}^{n}Z_i^2$. La v.a.$\chi^2$ suit une loi du khi-deux à n degrés de liberté (abréviation d.d.l.). On note cette loi par $\chi^2(n)$.
La densité de probabilité de χ^2 notée f_{χ^2} sera :

$$f_{\chi^2}=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\quad\text{pour tout x positif.}$$

où Γ est la fonction gamma. On a

$$\mathbb{E}(\chi^2)=n\text{ et }var(\chi^2)=2n$$

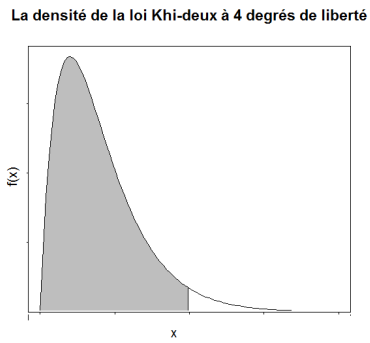


FIGURE 3.2 – La densité de probabilité de χ^2 à 4 degrés de liberté

3. **Loi de Student**
   Soit Z une variable aléatoire de loi Normale centrée et réduite et soit U une variable indépendante de Z et distribuée suivant la loi du χ^2 á n degrés de liberté. Par définition, la variable

$$T = \frac{Z}{\sqrt{\frac{U}{n}}}$$

   suit une loi de Student á n degrés de liberté. La densité de T, notée f_T, est donnée par :

$$f_T(t) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right), \text{ pour} n \in \mathbb{N}^*$$

   où Γ est la fonction Gamma d'Euler. La densité f_T associée à la variable T est symétrique, centrée en 0.

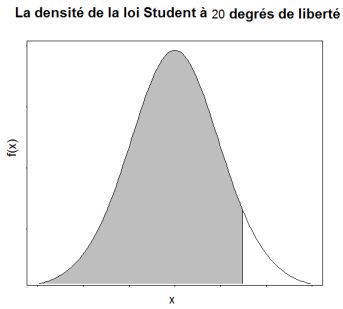


FIGURE 3.3 – La densité de probabilité de Student à 20 degrés de liberté

4. **Loi de Fisher-Snedecor**
   On dit qu'une variable aléatoire X suit la loi de Fisher-Snedecor de paramètres m

et n, On note cette loi par $\mathcal{F}_{n,m}$, si elle admet une densité qui vaut

$$f(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{x^{\frac{m}{2}-1}}{(xm+n)^{\frac{m+n}{2}}}$$

Suivant les valeurs de m et n, X admet alors une espérance et une variance qui sont

$$\mathbb{E}(X) = \frac{n}{n-2} \text{ si } n \geq 3$$

$$var(X) = \left(\frac{n}{n-2}\right)^2 \frac{2(m+n-2)}{m(n-4)} \text{ si } n \geq 5$$

La loi de Fisher-Snedecor de paramètres m et n est la loi du quotient normalisé de deux variables aléatoires Y_1 et Y_2 qui suivent une loi du Khi-deux à respectivement m et n degrés de liberté :

$$\frac{\frac{Y_1}{m}}{\frac{Y_2}{n}}$$

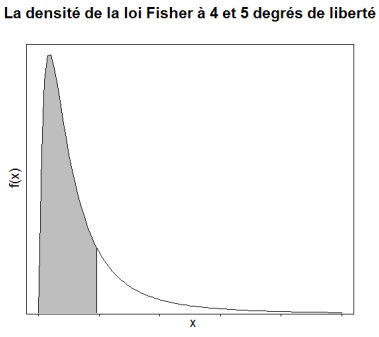


FIGURE 3.4 – Loi de Fisher-Snedecor de paramétres 4 et 5

3.4 Exercices corrigés

Exercice 1. Soit X une variable aléatoire gaussienne de moyenne $\mu = 2$ et d'écart-type $\sigma = 2$, $(X \rightsquigarrow \mathcal{N}(2,\, 2^2))$.

1. Calculer les valeurs des probabilités : $\mathbb{P}(2 \leq X \leq 3.5)$ et $\mathbb{P}(4.5 < X < 5.5)$.
2. Déterminer t et k si : $\mathbb{P}(X \geq t) = 0.33$, $\mathbb{P}(X \geq t) = 0.12098$ et $\mathbb{P}(X < k) = 0.92698$.

Solution. Si $X \sim \mathcal{N}(2,\, 4)$ alors :

$$\mathbb{P}(X \in [2,\, 3.5]) = \mathbb{P}\left(\frac{2-2}{2} \leq \frac{X-2}{2} \leq \frac{3.5-2}{2}\right) = \mathbb{P}\left(0 \leq Z \leq 0.75\right),$$

avec $Z = \dfrac{X-2}{2}$, alors

$$\mathbb{P}(X \in [2,\, 3.5]) = \mathbb{P}\left(0 \leq Z \leq 0.75\right) = \mathbb{P}\left(Z \leq 0.75\right) - \mathbb{P}\left(Z \leq 0\right) = 0.7734 - 0.5000 = 0.2734.$$

Si $X \rightsquigarrow \mathcal{N}(2,\, 4)$, alors

$$\begin{aligned}
\mathbb{P}(4.5 \leq X \leq 5.5) &= \mathbb{P}\left(\frac{4.5-2}{2} < \frac{X-2}{2} < \frac{5.5-2}{2}\right) \\
&= \mathbb{P}\left(1.25 < Z < 1.75\right) = \mathbb{P}\left(Z < 1.75\right) - \mathbb{P}\left(Z < 1.25\right), \text{ avec } Z = \frac{X-2}{2} \\
&= \mathbb{P}\left(Z \leq 1.75\right) - \mathbb{P}\left(Z \leq 1.25\right) = 0.9599 - 0.8944 = 0.0655.
\end{aligned}$$

Si $X \sim \mathcal{N}(2,\, 4)$ alors

$$\begin{aligned}
\mathbb{P}(X \geq t) = 0.33 &\iff \mathbb{P}\left(\frac{X-2}{2} \geq \frac{t-2}{2}\right) = 0.33 \\
&\iff 1 - \mathbb{P}\left(Z \leq \frac{t-2}{2}\right) = 0.33 \\
&\iff \mathbb{P}\left(Z \leq \frac{t-2}{2}\right) = 0.67 \\
&\iff \frac{t-2}{2} = 0.44,\ Z \rightsquigarrow \mathcal{N}(0,\, 1).
\end{aligned}$$

Alors $t = 2.88$
Si $Z \rightsquigarrow \mathcal{N}(2,\, 2^2)$, alors

$$\begin{aligned}
\mathbb{P}(|Z| < t) = 0.12098 &\iff \mathbb{P}(-t \leq Z \leq t) = 0.12098 \\
&\iff \mathbb{P}(Z \leq t) - \mathbb{P}(z \leq -t) = 0.12098 \\
&\iff \mathbb{P}(Z \leq t) - 1 + \mathbb{P}(Z \leq t) = 2\mathbb{P}(Z \leq t) = 1.12098 \\
&\iff \mathbb{P}(Z \leq t) = \frac{1.12098}{2} = 0.56049 \\
&\iff \mathbb{P}\left(\frac{Z-2}{2} \leq \frac{t-2}{2}\right) = 0.56049 \\
&\iff \mathbb{P}\left(K \leq \frac{t-2}{2}\right) = 0.56049,\ \ K \rightsquigarrow \mathcal{N}(0,\, 1).
\end{aligned}$$

On utilise la table de la loi $\mathcal{N}(0,\, 1)$, on trouve que : $\dfrac{t-2}{2} = 0.15$, ce qui est équivalent à dire que : $t = 2.3$.
Soit $X \sim \mathcal{N}(2,\, 4)$, et soit $\mathbb{P}(X < t) = 0.92698$, on veut calculer la valeur de t. En effet :

$$\begin{aligned}
\mathbb{P}(X < t) = 0.92698 &\iff \mathbb{P}\left(\frac{X-2}{2} < \frac{t-2}{2}\right) = 0.92698 \\
&\iff \mathbb{P}\left(Z < \frac{t-2}{2}\right) = 0.92698,\ Z = \frac{x-2}{2} \rightsquigarrow \mathcal{N}(0,\, 1).
\end{aligned}$$

De la table de la loi Normale centrée réduite $\mathcal{N}(0,\, 1)$, on trouve : $\dfrac{t-2}{2} = 1.45 \iff t = 4.9$
On peut aussi utilier la formule :$F(t^{'}) = F(t_1) + (F(t_2) - F(t_1)) \times \dfrac{t^{'} - t_1}{t_2 - t_1}$ *avec* $t_1 \leq t^{'} \leq t_2$.
Pour $t_1 = 1.45$ on a $F(t_1) = 0.9265$ et pour $t_2 = 1.46$ on a $F(t_2) = 0.9279$, par conséquent $t^{'} = 1.4534$ et $t^{'} = \dfrac{t-2}{2}$, alors $t = 4.9068$.

Exercice 2. Soit une variable aléatoire X suit une loi Normale de moyenne $\mu = 12$ et d'écart-type $\sigma = 3$.

1. Calculer $\mathbb{P}(8 < X < 16)$.
2. Déterminer un intervalle $]a,\, b[$ tel que la probabilité pour que $a < x < b$ soit 0.70 dans le cas où l'intervalle $]a,\, b[$ est centré en μ.

Solution. Soit $X \rightsquigarrow \mathcal{N}(12,\, 3^2)$, alors pour :

1. Calculons $\mathbb{P}(7 < X < 10)$:

   On pose : $T = \dfrac{X-\mu}{\sigma} = \dfrac{X-12}{3}$, d'où $X = 3t + 12$, et

$$\begin{aligned}
\mathbb{P}(7 < X < 10) &= \mathbb{P}(7 < 3T + 12 < 10) = \mathbb{P}\left(\frac{-5}{3} < T < \frac{-2}{3}\right) \\
&= F\left(\frac{-2}{5}\right) - F\left(\frac{-5}{3}\right) = 1 - F\left(\frac{2}{3}\right) - 1 + F\left(\frac{5}{3}\right) \\
&= F\left(\frac{5}{3}\right) - F\left(\frac{2}{3}\right) \\
&= F(1.66) - F(0.66) \\
&= 0.2061
\end{aligned}$$

2. Déterminons un intervalle $]a,\ b[$ tel que la probabilité pour que $a < X < b$ soit 0.70 dans le cas où l'intervalle $]a,\ b[$ est centré en 12. si l'on pose $b-a = 2\alpha$, on a, puisque 12 est la moyenne, l'intervalle : $]a = 12 - \alpha,\ \ b = 12 + \alpha[$.

   L'évènement $12 - \alpha < X < 12 + \alpha$ s'écrit : $(12 - \alpha < 3T + 12 < 12 + \alpha)$.

   Soit : $\left(\dfrac{-\alpha}{3} < T < \dfrac{\alpha}{3}\right)$, donc $\mathbb{P}(a < X < b) = 0.70$

$$\begin{aligned}
\mathbb{P}\left(a < X < b\right) = 0.70 &\iff \mathbb{P}\left(\frac{-\alpha}{3} < T < \frac{\alpha}{3}\right) = 0.70 \\
&\iff \mathbb{P}\left(0 < T < \frac{\alpha}{3}\right) = 0.35 \\
&\iff F\left(\frac{\alpha}{3}\right) - F\left(0\right) = 0.35 \\
&\iff F(\frac{\alpha}{3}) = 0.85, \\
&\iff \alpha = 3.1095,
\end{aligned}$$

   et l'intervalle $]a,\ b[$ est bien $]a,\ b[=]8,8905,\ 15.1095[$.

Exercice 3. Pour $X \rightsquigarrow N(2,\ 3^2)$, quelle est $\mathbb{P}(X > 5)$? et pour $X \rightsquigarrow N(5,\ 4^2)$, quelle est $\mathbb{P}(X \geq 10)$?

Solution. Si $X \rightsquigarrow \mathcal{N}(2,\ 3^2)$, alors

$$\mathbb{P}(X > 5) = 1 - \mathbb{P}(X \leq 5) = 1 - \mathbb{P}\left(\frac{X-2}{3} \leq \frac{5-2}{3}\right) = 1 - \mathbb{P}(Z < 1) = 1 - 0.8413 = 0.1587$$

Si $X \rightsquigarrow \mathcal{N}(5,\, 4^2)$, alors

$$\begin{aligned}
\mathbb{P}(X \geq 10) &= 1 - \mathbb{P}(X \leq 10) \\
&= 1 - \mathbb{P}\left(\frac{X-5}{4} \leq \frac{10-5}{4}\right) \\
&= 1 - \mathbb{P}(Z \leq 1.25), \;\; Z = \frac{X-5}{4} \rightsquigarrow \mathbb{N}(0,\, 1) \\
&= 1 - \mathbb{P}(Z \leq 1.25) = 1 - 0.8944 = 0.1056.
\end{aligned}$$

Et

$$\begin{aligned}
\mathbb{P}(|X| > 10) &= 1 - \mathbb{P}(|X| \leq 10) \\
&= 2 - 2\mathbb{P}(X \leq 10) \\
&= 2 - 2\mathbb{P}\left(Z \leq \frac{10-5}{4}\right) \\
&= 2 - 2\mathbb{P}(Z \leq 1.25) \\
&= 2 - 2(0.8944) \\
&= 0.2112.
\end{aligned}$$

Chapitre 4

Échantiannage et Estimation

L'objectif de l'estimation statistique est d'évaluer certaines grandeurs associées à une population à partir d'observations faites sur un échantillon. Bien souvent, ces grandeurs sont des moyennes ou des variances.

Remarque

Il faut prendre soin de distinguer ces grandeurs théoriques (inconnues et à estimer) de celles observées sur un échantillon.

Exemples de problèmes :

• Quelle est la fréquence (probabilité) de survenue d'un certain cancer chez les souris ?

• Quelle est la glycémie moyenne d'un patient ?

• Quelle est l'écart moyen de la glycémie d'un patient autour de sa glycémie moyenne ?

On apporte deux types de réponses à ces questions : à partir d'un échantillon,

♣ On calcule une valeur qui semble être la meilleure possible : on parle d'estimation ponctuelle.

♣ On calcule un intervalle de valeurs possibles : c'est la notion d'intervalle de confiance.

On se placera toujours dans la situation suivante :

• Un échantillon ω est obtenu par tirages avec remise de n individus dans la population de référence.

• Les valeurs observées $x_1, ..., x_n$ d'une grandeur (ex : poids) sur un échantillon ω ne dépendront donc pas les unes des autres (tirages sans remise).

Un échantillon est la donnée de n va. $X_1, ..., X_n$ de même loi.
Une observation correspond à une réalisation $\omega \in \Omega$ du hasard. On a alors

$$x_1 = X_1(\omega), ..., x_n = X_n(\omega).$$

Si on change d'observation, cela correspond à changer la réalisation du hasard en $\omega' \in \Omega$ et on a d'autres valeurs observées sur l'échantillon :

$$x'_1 = X_1(\omega'), ..., x'_n = X_n(\omega').$$

On modélisera donc cette situation par un ensemble fondamental

$$\Omega = \{\text{échantillons } \omega \text{ de taille n avec remise}\}$$

et des variables aléatoires $X_1, ..., X_n$ indépendantes (tirages avec remise) et de même loi (car on observe la même grandeur). On a ainsi pour un échantillon ω donné, des valeurs observées $X_1(\omega) = x_1, ..., X_n(\omega) = x_n$.

4.1 Le plan d'échantillonnage

On utilise un plan d'échantillonnage lorsque l'on réalise une étude par enquête, i.e. lorsque l'on collecte des informations sur un groupe d'individus dans leur milieu habituel, mais que tous les individus ne sont pas accessibles (par choix ou par contrainte).
Les principales méthodes d'échantillonnage peuvent être regroupées en deux ensembles :

1. **L'échantillonnage aléatoire** : tous les individus (au sens statistique) ont la même probabilité d'être choisis, et le choix de l'un n'influence pas celui des autres. Différentes méthodes d'échantillonnage aléatoire existent :
   - ✓ *L'échantillonnage aléatoire et simple* : le choix se fait parmi tous les individus de la population (au sens statistique), qui ne forme qu'un grand ensemble.
   - ✓ *L'échantillonnage stratifié* : si la population est très hétérogène, elle peut être divisée en sous - ensembles exclusifs (ou strates). Au sein de ces strates l'échantillonnage est ensuite aléatoire et simple.
   - ✓ *L'échantillonnage en grappes* : si les strates sont très nombreuses, on en choisit certaines au hasard (les grappes). Au sein de ces grappes l'échantillonnage est ensuite aléatoire et simple
   - ✓ *L'échantillonnage par degrés* : il est une généralisation de l'échantillonnage en grappes (qui est en fait un échantillonnage du premier degré). Au sein de la

population on choisit des grappes " primaires ", puis à l'intérieur de celles-ci des grappes " secondaires " (toujours au hasard), et ainsi du suite... Au dernier niveau l'échantillonnage est aléatoire et simple.

2. **L'échantillonnage systématique** : un premier individu est choisi aléatoirement, puis les autres sont choisis de façon régulière à partir du précédent (dans le temps ou l'espace). L'analyse de ce type d'échantillonnage, qui fait appel à la statistique spatiale ou à l'analyse des séries chronologiques, n'est pas abordée dans cet cours.

4.2 Loi d'échantillonage

✓Pour des moyennes

Soit une population d'effectif total N connu. On considère un échantillon d'effectif n. Un élément quelconque X de l'échantillon suit la loi d'échantillonage de taille n et de moyenne $\bar{X}_n$. Quand n devient grand ($n \geq 30$), la loi d'échantillonnage peut être approchée par la loi normale $\mathcal{N}(\bar{X}, \sigma^2/n)$ où σ^2 est supposée connue.

Example

Dans une population, l'écart-type de la taille est 5 *cm. Si sur* 200 *personnes, la taille moyenne observée est* $\bar{X} = 150$ *cm, alors la taille* X *d'un individu quelconque issu de cette population suit la loi d'échantillonnge* $\mathcal{N}(150; 0, 125)$ *(car* $\sigma^2/n = 25/200$*).*

✓Pour des fréquences

On étudie une population de taille N (connu) et un caractère X à deux éventualités (échec ou succès) avec probabilité p. On sait (loi de Bernoulli) que $\mathbb{E}[X] = p$ et $Var(X) = p(1-p)$. Si on prélève un échantillon de taille n, le nombre de succès X_n est compté par une loi binomiale $\mathcal{B}(n, p)$ avec $\mathbb{E}[X_n] = np$ et $Var(X_n) = np(1-p)$. Quand n est grand ($n > 30$), la loi de la fréquence X_n/n des succès s'approxime par $\mathcal{N}(p, \frac{p(1-p)}{n})$.

Example

Considérons une population où 10% des gens développent une certaine allergie. Dans un échantillon de 200 personnes de cette population, le nombre d'allergiques suit la loi binomiale $\mathcal{B}(200; 0, 1)$*. On l'approxime la loi de la fréquence par la loi normale* $\mathcal{N}(0, 1;)$.

4.3 Estimation ponctuelle

On cherche à estimer une valeur θ inconnue liée à un certain phénomène aléatoire, en général, la moyenne μ, la variance σ^2 ou encore l'écart-type σ de la loi du phénomène.

Pour cela, on dispose d'observations indépendantes du phénomènes, c-à-d de variables aléatoires $X_1, ..., X_n$ indépendantes et de même loi (celle du phénomène), on parle d'un échantillon. On définit à partir de l'échantillon une nouvelle variable aléatoire notée $\hat{\theta}_n$ dont les valeurs seront proches de celle de la grandeur θ à estimer. Cette nouvelle variable aléatoire $\hat{\theta}_n$ sera appelée estimateur de θ. Il peut y avoir plusieurs estimateurs pour une même grandeur θ, certains meilleurs que d'autres.

✓Estimateur

Définition 4.3.1. *Un n-échantillon aléatoire issu d'une v.a.r. X est un ensemble* $(X_1, ..., X_n)$ *de n v.a.r. indépendantes et de même loi que X*

Définition 4.3.2. *Un estimateur* $\hat{\theta}_n$ *de* θ *est une fonction qui dépend uniquement du n-échantillon* $(X_1, ..., X_n)$*. Il est dit convergent s'il est proche de* θ *au sens de la convergence en probabilité : pour tout* $\epsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to_{n\to\infty} 0.$$

Le but de la théorie de l'estimation est de choisir, parmi toutes les statistiques possibles, le "meilleur" estimateur convergent, c'est-à-dire celui qui donnera une estimation ponctuelle la plus proche possible du paramètre et ceci, quel que soit l'échantillon.

Théorème 4.3.1. *Soit* $\hat{\theta}_n$ *un estimateur de* θ*. Si l'on a :*

$$\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta \quad et \quad \lim_{n\to\infty} Var(\hat{\theta}_n) = 0,$$

alors $\hat{\theta}_n$ *est un estimateur convergent de* θ*.*

Théorème 4.3.2. *Soit* $\hat{\theta}_n$ *un estimateur convergent d'un paramètre* θ*. On appelle biais la quantité* $\mathbb{E}(\hat{\theta}_n) - \theta$*. L'estimateur* $\hat{\theta}_n$ *est dit sans biais si* $\mathbb{E}(\hat{\theta}_n) = \theta$*, et biaisé sinon.*

Remarque

On choisit, parmi les estimateurs convergents et sans biais, celui qui a la variance la plus petite. En d'autres termes, si $\hat{\theta}_n$ *est un estimateur convergent et sans biais de* θ *on a tout intérêt à ce que* θ *ne varie pas trop autour de sa moyenne. Cette propriété traduit ce que l'on appelle l'efficacité de l'estimateur.*

✓Estimation de la moyenne, de la proportion et de la variance

On considère un n-échantillon $(X_1, ..., X_n)$ issu d'une loi de moyenne μ et de variance σ^2,

toutes deux inconnues. On admet que

• Le meilleur estimateur de la moyenne $\mu = \mathbb{E}[X]$ du caractère X est

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$*$ La moyenne empirique $\bar{X_n}$ est un estimateur sans biais et convergent de μ, et par indépendance : $Var(\bar{X_n}) = \frac{\sigma^2}{n}$.

$*$ Si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, alors $\bar{X_n} \rightsquigarrow \mathcal{N}(\mu, \sigma^2/n)$.

• Dans le cas particulier où X suit une loi de Bernoulli $\mathcal{B}(p)$, comme la moyenne μ est égale à la proportion p, c'est une estimation de proportion (ou de fréquence) qu'on fait quand on estime sa moyenne $\mathbb{E}[X] = p$.

• Le meilleur de la variance $\sigma^2 = Var(X)$ du caractère X est la variance empirique corrigée

$$\begin{aligned} \acute{S}_n^2 &= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \end{aligned}$$

$*$ $\acute{S}_n^2$ est un estimateur convergent et sans biais de la variance σ^2.

4.4 Estimation par intervalle de confiance

✓Principe

Un estimateur permet de calculer une valeur sur un échantillon qui devrait être proche du paramètre θ sans pour autant savoir si cette valeur est totalement fiable. C'est pourquoi on a introduit la notion d'intervalle de confiance : c'est un intervalle dans lequel se trouve θ avec une grande probabilité $1-\alpha$ (où α est un risque qu'on se fixe, en général, petit). On peut en théorie choisir $1-\alpha$ aussi proche de 1 que l'on veut, mais alors l'intervalle de confiance grandit et devient imprécis. Il s'agit donc d'un compromis entre précision (intervalle peu étendu) et sûreté (α petit). La probabilité $1-\alpha$ est appelée niveau de confiance et α le risque (de 1ère espèce), c-à-d la probabilité que l'intervalle proposé (qu'on notera

I, pour intervalle de confiance) ne contienne pas la valeur à estimer θ.
On suppose que les observations $x_1, ..., x_n$ sont issues de n v.a. indépendantes $X_1, ..., X_n$ de même loi $\mathcal{N}(\mu, \sigma^2)$. Si la loi n'est pas gaussienne, on suppose alors que la taille de l'échantillon est grande ($n \geq 30$ en pratique), le théorème central limite (TCL) permet de faire des approximations par des lois normales, ce qui donnera des intervalles de confiance approximatifs mais suffisant en pratique. On fera donc systématiquement comme si les échantillons sont gaussiens lorsque sa taille est élevé.

✓Intervalle de confiance pour une moyenne

✓✓Lorsque σ^2 est connue
Etant donné $\bar{X}_n$, l'estimateur ponctuel de μ qui a pour loi $\mathcal{N}(\mu, \sigma^2/n)$. Alors

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \rightsquigarrow \mathcal{N}(0,1),$$

et que

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Ceci est équivaut à

$$\mathbb{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{n} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{n}\right) = 1 - \alpha$$

L'intevalle de confaince I pour l'espérance μ avec coefficient de sécurité $1 - \alpha$ (il s'agit de l'intervalle aléatoire)

$$I = \left[\bar{X}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right];$$

Ainsi, dans les calculs, l'IC est donné par

$$I = \left[\bar{x}_n - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right];$$

où $\bar{x}_n$ est l'estimation ponctuelle de μ associée à la réalisation du n-échantillon $(X_1, ..., X_n)$.

✓✓Lorsque σ^2 est inconnue
Dans cette situation l'expression précédente de l'intervalle de confiance ne peut être calculée car σ^2 n'est plus connu, remplacer σ^2 par son estimateur $\acute{S}_n^2$,

Alors

$$\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\acute{S}_n^2}\right)\rightsquigarrow \mathcal{T}(n-1),$$

où $\mathcal{T}(n-1)$ est une loi de Student à n-1 degrés de liberté. et que

$$\mathbb{P}\left(-t_{alpha,n-1}\leq\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\acute{S}_n^2}\right)\leq t_{alpha,n-1}\right)=1-\alpha.$$

Ceci est équivaut à

$$\mathbb{P}\left(\bar{X}_n-t_{alpha,n-1}\frac{\acute{S}_n^2}{n}\leq\mu\leq\bar{X}_n+t_{alpha,n-1}\frac{\acute{S}_n^2}{n}\right)=1-\alpha.$$

L'intevalle de confaince I pour l'espérance μ avec coefficient de sécurité $1-\alpha$ (il s'agit de l'intervalle aléatoire)

$$I=\left[\bar{X}-t_{alpha,n-1}\frac{\acute{S}_n^2}{\sqrt{n}},\bar{X}+t_{alpha,n-1}\frac{\acute{S}_n^2}{\sqrt{n}}\right].$$

Ainsi, dans les calculs, l'I est donné par

$$I=\left[\bar{x}_n-t_{alpha,n-1}\frac{\acute{s}_n^2}{\sqrt{n}},\bar{x}_n+t_{alpha,n-1}\frac{\acute{s}_n^2}{\sqrt{n}}\right];$$

où $\bar{x}_n$ et $\acute{s}_n^2$ sont les estimations ponctuelles respectives de la moyenne μ et de la variance σ^2.

Remarque

Quand n est grand ($n\geq 30$), *on peut considérer que la loi de Student est proche de la normale et prendre* $t_{1-\frac{\alpha}{2}}$ *dans la table de la loi normale.*

✓Intervalle de confiance pour une varaince

✓✓Lorsque μ est connue

L'intervalle de confiance de la variance σ^2 se calcule à partir de l'échantillon de taille n par

$$I=\left[\frac{nT_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n)},\frac{nT_n^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right]$$

Avec $T_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, et $\chi^2_{1-\frac{\alpha}{2}}(n)$, $\chi^2_{\frac{\alpha}{2}}(n)$ se trouvent dans la table de la loi $\chi^2(n)$ de la v.a. U. Ainsi, dans les calculs, l'I est donné par

$$I = \left[\frac{nt_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}, \frac{nt_n^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right]$$

✓✓Lorsque μ est inconnue

Comme μ est inconnue, l'idée est de la remplacer par son estimation $\bar{X}$. L'intervalle de confiance de la variance σ^2 se calcule alors à partir de l'échantillon de taille n par

$$I = \left[\frac{(n-1)\acute{S}_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)\acute{S}_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

Avec $\chi^2_{1-\frac{\alpha}{2}}(n-1)$, $\chi^2_{\frac{\alpha}{2}}(n-1)$ se trouvent dans la table de la loi $\chi^2(n-1)$ de la v.a. U. Ainsi, dans les calculs, l'I est donné par

$$I = \left[\frac{(n-1)\acute{s}_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)\acute{s}_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

✓Intervalle de confiance pour une proportion

L'intervalle de confiance pour une proportion p inconnue est donné comme suit

$$I = \left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

où $\hat{p}$ est la fréquence observée du caractère considéré sur l'échantillon étudié (c'est donc l'estimateur sur l'échantillon de l'inconnue p)

Remarque

Les conditions requises pour une bonne approximation par la loi normale sont $n \geq 30$, $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$.

✓Exemple d'application

On suppose que le taux de cholestérol X d'un individu choisi au hasard dans une population donnée suit une loi normale. Sur un échantillon ω de 100 individus, on constate la moyenne des taux observés est $\bar{x} = 1.55$ (gr pour mille). On constate aussi une variance corrigée $s'^2 = 0.25$. L'intervalle de confiance pour la moyenne μ au niveau de confiance 0.95. est donné comme suit

$$I = [\bar{x} - t_{1-\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}, \bar{x} + t_{1-\frac{\alpha}{2}} \frac{s'}{\sqrt{n}}];$$

D'où

$$I = [1.55 - 1.984 \frac{0.25}{10}, 1.55 + 1.984 \frac{0.25}{10}],$$

qui donne

$$I = [1.504, 2.046]$$

4.5 Exercices corrigés

Exercice 01

Un biologiste étudie un type d'algue qui attaque les plantes marines. La toxine contenue dans cette algue est obtenue sous forme d'une solution organique. Il mesure la quantité de toxine par gramme de solution. Il a obtenu les neuf mesures suivantes, exprimées en milligrammes : 1.2 ; 0.8 ; 0.6 ; 1.1 ; 1.2 ; 0.9 ; 1.5 ; 0.9 ; 1.0. Nous supposerons que ces mesures sont les réalisations de variables aléatoires indépendantes et identiquement distribuées suivant une loi normale d'espérance μ et d'écart-type σ.

1. Donnez une estimation ponctuelle de l'espérance μ et de l'ecart-type σ de la quantité de toxine par gramme de solution.
2. Determinez un intervalle de confiance a 95% pour l'espérance μ de la quantité de toxine par gramme de solution.
3. Le biochimiste trouve que l'intervalle obtenu n'est pas satisfaisant car trop étendu. Que doit-il faire pour obtenir une estimation plus précise ?

Solution.

1. Donnons une estimation ponctuelle de l'espérance μ de la quantité de toxine par gramme de solution. Nous avons

$$\begin{aligned} \bar{x} &= \frac{1.2+0.8+0.6+1.1+1.2+0.9+1.5+0.9+1.0}{9} \\ &= 1.02 \text{ mg.} \end{aligned}$$

   Nous donnons une estimation ponctuelle de l'écart-type σ de la quantité de toxine par gramme de solution a l'aide de l'estimateur corrige S'_n. Nous avons

$$s'_n = 0,264 \text{ mg.}$$

2. L'espérance μ et l'écart-type σ étant inconnus, l'intervalle de confiance pour μ de niveau 95% s'obtient avec la formule suivante :

$$\begin{aligned} \bar{x} - t_{0.05,8}\frac{s'_n}{\sqrt{9}} &< \mu < \bar{x} + t_{0.05,8}\frac{s'_n}{\sqrt{9}} \\ 0.82 &< \mu < 1.22 \end{aligned}$$

   où $t_{0.05,8} = 2.31$ est le quantile du risque 0,05 pour la loi de Student à huit degrés de liberté t(8).

3. Il doit augmenter la taille de l'échantillon.

Exercice 02

Le staff médical d'une grande entreprise fait ses petites statistiques sur le taux de cholestérol de ses employés ; les observations sur 100 employés tirés au sort sont les suivantes.

taux de cholestérol en cg :(centre classe)	effectif d'employés :
120	9
160	22
200	25
240	21
280	16
320	7

1. Estimer la moyenne et l'écart-type pour le taux de cholestérol dans toute l'entreprise.
2. Déterminer un intervalle de confiance á 95% pour la moyenne.
3. Déterminer la taille minimum d'échantillon pour que l'amplitude de l'intervalle de confiance soit inferieure à 10.

Solution.

1. On obtient, sur l'échantillon, la moyenne :

$$\bar{x}_e = \frac{1}{n}\sum n_i x_i = 213.6,$$

l'écart-type

$$s_e^2 = \frac{1}{n}\sum n_i x_i^2 - \bar{x}^2 = 55.77.$$

La moyenne sur l'entreprise est estimée par

$$\bar{x}_e.$$

L'écart-type est estimé par :

$$s'_n = \sqrt{\frac{n}{n-1}} \times s_e = 56.05$$

2. On en déduit, au seuil 95%, un intervalle de confiance pour la moyenne :

$$[\bar{x}_e - z_{1-\frac{\alpha}{2}}\frac{s'_n}{\sqrt{n}}, \bar{x}_e + z_{1-\frac{\alpha}{2}}\frac{s'_n}{\sqrt{n}}] = [202.61; 224.6]$$

3. la taille minimum d'échantillon pour que l'amplitude de l'intervalle de confiance soit inferieure à 10 :

$$2z_{1-\frac{\alpha}{2}}\frac{s'_e}{\sqrt{n}} \leq 10$$

alors

$$\left(z_{1-\frac{\alpha}{2}}\frac{s'_n}{5}\right)^2 \leq n$$

d'ou

$$\left(z_{1-\frac{\alpha}{2}}\frac{s'_n}{5}\right)^2 \leq n$$

alors, la taille minimum d'échantillon

$$n^* = 483.$$

Exercice 03

Une clinique a proposé une nouvelle opération chirurgicale, et a connu 40 échecs, sur 200 tentatives. On note p le pourcentage de réussite de cette nouvelle opération.

1. Quelle estimation de p proposez-vous ?
2. En utilisant l'approximation normale, donner un intervalle de confiance pour p de niveau de confiance 0.95.

Solution.

1. On tire un échantillon de la population d'intérêt et on calcule la proportion échantillonnage

$$\hat{p} = \frac{\text{nombre d'élément possédant l'attribut}}{\text{la taille de l'échantillon}} = 0.8$$

cette proportion est utilisée comme estimation ponctuelle de la proportion P de la population.

2. un intervalle de confiance pour p de niveau de confiance 0.95 est donné comme suit

$$IC = [\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}],$$

d'ou

$$IC = [0.8 - z_{0.975}\sqrt{\frac{0.8\times 0.2}{200}}, 0.8 + z_{0.975}\sqrt{\frac{0.8\times 0.2}{200}}] = [0.7446; 0.8554].$$

Les conditions requises pour une bonne approximation par la loi normale sont $n \geq 30$, $n\hat{p} \geq 10$, $n(1-\hat{p}) \geq 10$. sont satisfaits.

Exercice 04

On suppose que le poids d'un nouveau né est une variable normale d'écart-type égal à 0.5 kg. Le poids moyen des 49 enfants nés dans un hôpital a été de 3.6 kg.

1. Déterminer un intervalle de confiance à 95% le poids moyen d'un nouveau né dans cet hôpital.
2. Quel serait le niveau de confiance d'un intervalle de longueur 0.1 kg centré en 3.6 pour ce poids moyen.

$$(t_{0.05} = 1.96, \qquad \pi(0.7) = 0.758).$$

Solution.

1. L'intervalle de confiance à 95% de poids moyen est :

$$\begin{aligned} IC_{0.05} &= [\bar{x} - 1.96\frac{\sigma}{\sqrt{49}}; \bar{x} + 1,96\frac{\sigma}{\sqrt{49}}] \\ &= [3.46; 3.74]. \end{aligned}$$

2. Le niveau de confiance d'un intervalle de longueur 0.1 kg centré en 3.6 pour ce poids moyen est :

$$\begin{aligned} \mathbb{P}[\bar{x} - 0.05 \leq m \leq \bar{x} + 0.05] &= \mathbb{P}\left[\frac{-0.05}{(\sigma/7)} \leq \frac{\bar{x} - m}{(\sigma/\sqrt{n})} \leq \frac{0.05}{(\sigma/7)}\right] \\ &= 2\pi(\frac{0.05}{0.5/7}) - 1 \\ &= 2\pi(0.7) - 1 \\ &= 0.516. \end{aligned}$$

Le niveau de confiance est donc 0.516.

Exercice 05

Dans un centre avicole, des études antérieures ont montré que la masse d'un œuf choisi au hasard peut être considéré comme la réalisation d'une variable aléatoire normale X, de moyenne m et de variance . On admet que les masses des oeufs sont indépendantes les unes des autres. On prend un échantillon de n = 29 œufs que l'on pèse. Les mesures sont données dans le tableau correspondant.

50.34	52.62	53.79	54.99	55.82	57.67	51.41
53.13	53.89	55.04	55.91	57.99	51.51	53.28
54.63	55.12	55.95	58.10	52.07	53.30	54.76
55.24	57.05	59.30	57.18	53.32	54.78	60.58

1. Estimer ponctuellement la moyenne et l'écart-type de cette série statistique.
2. Donner un intervalle de confiance au niveau 95 de la masse moyenne m d'un œuf

Solution.

1. Estimation ponctuellement la moyenne et l'écart-type par cette série statistique.
   - Pour la moyenne
   $$\bar{x}_e = \frac{1}{n}\sum x_i = 53,41.$$
   - Pour l'écart-type
   $$s'_n = \sqrt{\frac{n}{n-1}} \times s_e = 3,44$$

   où $s_e^2 = \frac{1}{n}\sum n_i x_i^2 - \bar{x}^2 = 11,39.$
2. l'intervalle de confiance au niveau 95 de la masse moyenne m d'un œuf est donné comme suit :
$$\begin{aligned} IC_{0.05} &= [\bar{x} - t_{0.975,28}\frac{s'_n}{\sqrt{29}}; \bar{x} + t_{0.975,28}\frac{s'_n}{\sqrt{29}}] \\ &= [52.10; 54.72] \end{aligned}$$

   où $t_{0.975,28} = 2.048$ d'après table de loi de student $\mathcal{T}(28)$ de degré de liberté 28

Chapitre 5

Tests d'hypothèses

5.1 Principe

Définition 5.1.0.1. *Un test statistique est une procédure de décision entre deux hypothèses concernant un ou plusieurs échantillons.*

✓Hypothèses

Définition 5.1.0.2. • *L'hypothèse nulle notée H_0 est celle que l'on considère vraie à priori. Le but du test est de décider si cette hypothèse à priori est crédible.*
• *L'hypothèse alternative notée H_1 est l'hypothèse complémentaire de H_0.*

Remarque

• *Les deux hypothèses ne sont pas symétriques. H_1 est choisie uniquement par défaut si H_0 n'est pas considérée comme crédible.*

• *Le choix de H_0 et de H_1 est en général imposé par le test qu'on utilise et ne relève donc pas de l'utilisateur.*

Example
Soit μ_1 et μ_2 les moyennes de tension des deux populations correspondant à la prise de médicament ou de placebo. Une manière de démontrer que le médicament modifie la tension est de montrer que μ_2 est différent de μ_1. Les hypothèses deviennent alors

$$H_0 : \text{ les moyennes des deux populations sont égales}$$

et

$$H_1 : \text{ les moyennes des deux populations sont différentes.}$$

On l'écrit succintement sous la forme :

$$\begin{cases} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{cases}$$

Remarque

Les moyennes $\bar{x}_1$ et $\bar{x}_2$ des échantillons résultent d'échantillonnages, et ne sont donc que des estimations de μ_1 et μ_2. Ce n'est pas parce qu'elles sont différentes que μ_1 et μ_2 le sont (et vice-versa, mais c'est rare !). Comparer les moyennes des échantillons ne peut en aucun cas suffire !

Remarque

Les signes $=$, $\neq$, $>$ et $\leq$ dans l'écriture succinte des hypothèses ne correspondent pas à l'égalité ou aux inégalités au sens mathématique du terme. Il s'agit d'une façon d'écrire :

$$\begin{cases} H_0 : & \text{Il est crédible de penser que } \mu_1 = \mu_2 \\ H_1, & \mu_1 \text{ est significativement différent de } \mu_2 \end{cases}$$

✓Statistique

La statistique de test S est une fonction qui résume l'information sur l'échantillon qu'on veut tester. On la choisit de façon à pouvoir calculer sa loi sous H_0. • S est une variable aléatoire, définie indépendemment des données observées. La valeur que prend cette variable aléatoire pour les données observées sera appelée statistique observée et notée Sobs dans la suite.

• Suivant le type de statistique choisi, le test sera paramétrique ou non-paramétrique.

✓Région de rejet

Définition 5.1.0.3. *La région de rejet est le sous-ensemble $\mathcal{I}$ de $\mathbb{R}$ tel qu'on rejette H_0 si S_{obs} appartient à $\mathcal{I}$.*

Définir une procédure de test peut donc se faire en définissant

1. Une statistique
2. Une région de rejet pour cette statistique.

Définition 5.1.0.4. *La forme de la région de rejet définit la latéralité du test :*

• *test multilatéral : On veut rejetter H_0 si S_{obs} est trop grand ou trop petit, sans à priori. La région de rejet est alors de la forme $]-\infty; a] \cup [b; +\infty[$.*

• *test unilatéral à droite : On veut rejetter H_0 seulement si S_{obs} est trop grand. La région*

de rejet est alors de la forme $[a; +\infty[$.

• *test unilatéral à gauche : On veut rejetter* H_0 *seulement si* S_{obs} *est trop petit. La région de rejet est alors de la forme* $]-\infty; b]$.

Example

On considère toujours des médicaments réduisant la tension artérielle. Quelles sont les hypothèses pour répondre aux questions suivantes ?

• *Comparaison entre deux médicaments en vente*

$$\begin{cases} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 \neq \mu_2 \end{cases}$$

• *Intérêt d'un nouveau médicament plus cher que l'existant.*

$$\begin{cases} H_0 : & \mu_n \geq \mu_a \\ H_1 : & \mu_n < \mu_a \end{cases}$$

• *Intérêt d'un nouveau médicament moins cher que l'existant.*

$$\begin{cases} H_0 : & \mu_n \leq \mu_a \\ H_1 : & \mu_n > \mu_a \end{cases}$$

✓Probabilité critique

Définition 5.1.0.5. *La probabilité critique (ou p-valeur) est la probabilité, sous* H_0*, que la statistique soit au moins aussi éloignée de son espérance que la valeur observée. En d'autres termes, c'est la probabilité d'observer quelque chose d'au moins aussi surprenant que ce que l'on observe.*

Si le test est unilatéral à droite, la probabilité critique est $\mathbb{P}(S > S_{obs})$.

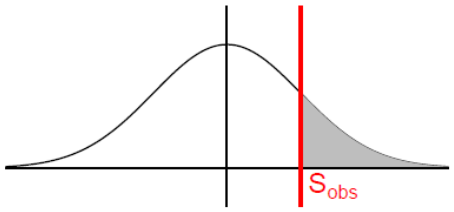


FIGURE 5.1 – Test unilatéral à droite

• Si le test est unilatéral à gauche, la probabilité critique est $\mathbb{P}(S < S_{obs})$.

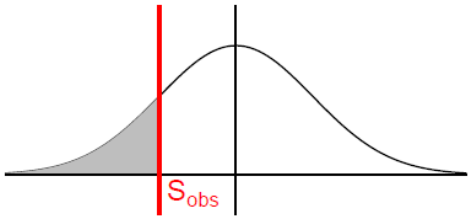


FIGURE 5.2 – Test unilatéral à gauche

• Si le test est bilatéral et que la loi de la statistique est symétrique par rapport à 0, la probabilité critique est $\mathbb{P}(|S| > |S_{obs}|)$.

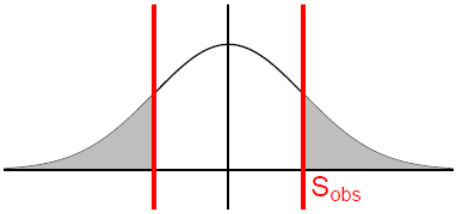


FIGURE 5.3 – Test bilatéral

✓Risque de première espèce

Définition 5.1.0.6. *Le risque de première espèce α est la probabilité sous H_0 de la région de rejet. En d'autres termes, il s'agit de la proabilité avec laquelle on accepte de décider H_1 si la vérité est H_0. $\alpha = \mathbb{P}(H_1/H_0)$ La quantité $1-\alpha$ est la confiance du test.*

En d'autres termes, une proportion α des situations dans lesquelles la vérité est H_0 verront une décision en faveur de H_1. α est la probabilité avec laquelle on accepte de se tromper quand la vérité est H_0

✓Autre manière de mener le test

On peut comparer la p-valeur à α plutôt que S_{obs} et la région de rejet.

• Si la p-valeur est supérieure à α, il n'est pas exceptionnel sous H_0 d'observer la valeur eαectivement observée. Par conséquent, H_0 est acceptée.

• Si la p-valeur est inférieure à α, la valeur observée est jugée exceptionnelle sous H_0. On décide alors de rejeter H_0 et de valider H_1.

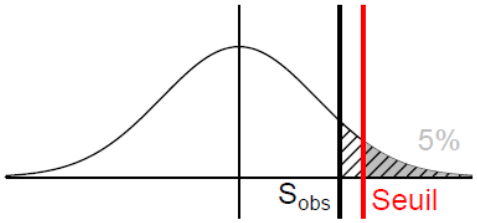


FIGURE 5.4 – Acceptation

On peut comparer la p-valeur à α plutôt que S_{obs} et la région de rejet.

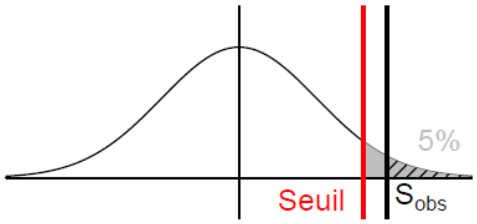


FIGURE 5.5 – Rejet

On peut comparer la p-valeur à α plutôt que S_{obs} et la région de rejet.

• si la p-valeur est supérieure à α, il n'est pas exceptionnel sous H_0 d'observer la valeur eαectivement observée. Par conséquent, H_0 est acceptée.

• si la p-valeur est inférieure à α, la valeur observée est jugée exceptionnelle sous H_0. On décide alors de rejeter H_0 et de valider H_1.

Avantage

Cette méthode permet de se rendre compte à quel point on est sur de sa décision : la position de la p-valeur par rapport à α ne dépend pas de l'échelle des données, contrairement à Sobs et au(x) seuil(s) de la région de rejet.

✓Risque de première espèce

• Hormis dans des cas de tests multiples non abordés dans ce cours, α varie généralement entre 0.01 et 0.05.

• Dans le cas de variables continues, on peut choisir une valeur arbitraire de α et obtenir une région de rejet présentant exactement le risque α.

• Dans le cas de variables discrètes, le nombre de régions de rejet, et donc de risques, possibles est fini ou dénombrable. Dans ce cas, on fixe un risque, dit risque nominal, par exemple de 5%. On cherche alors la plus grande région ne dépassant pas ce risque, qui devient la région de rejet. Le véritable risque, dit risque réel, peut alors être recalculé.

Risque de deuxième espèce

Définition 5.1.0.7. *Le risque de deuxième espèce β est la probabilité d'accepter H_0 alors que la vérité est H_1. $\beta = \mathbb{P}(H_0/H_1)$ La quantité $1-\beta$ est la puissance du test.*

		Vérité	
		H_0	H_1
Décision	H_0	1-α	β
	H_1	α	1-β

FIGURE 5.6 – Decision

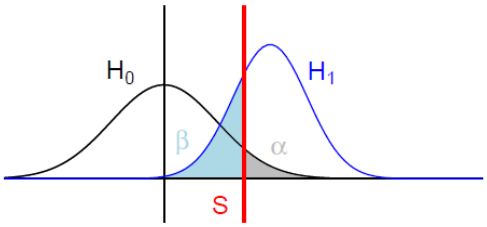


FIGURE 5.7 – Choix du α et β

Si l'échantillon reste inchangé, une diminution de α entraîne une augmentation de β et inversement. Autrement dit, si on décide de réduire le nombre de faux positifs, on augmente forcément le nombre de faux négatifs. La seule manière d'améliorer les deux critères est d'augmenter la taille de l'échantillon.

✓Principe du test

Les étapes d'un test sont toujours réalisées dans l'ordre suivant :

1. Choix du risque α
2. Choix du type de test et de sa latéralité si besoin
3. Calcul de la statistique de test
4. Calcul de la p-valeur
5. Conclusion

En pratique, l'utilisation d'un logiciel type R permet de ne pas se soucier des partie 3) et 4). Par contre, les choix liées aux étapes 1) et 2) ainsi que l'interprétation finale ne peuvent être faits par le logiciel.

5.2 Tests paramétriques et non paramétriques

Définition 5.2.0.8. ***Test paramétrique*** *Un test paramétrique est un test pour lequel on fait une hypothèse sur la forme des données sous H_0 (normale, Poisson, ...). Les hypothèses du test concernent alors les paramètres gouvernant cette loi.*

Définition 5.2.0.9. ***Test non-paramétrique*** *Un test non paramétrique est un test qui nécessite pas d'hypothèse sur la forme des données. Les données sont alors remplacées par des statistiques ne dépendant pas des moyennes/variances des données initiales (tables de contingence, statistique d'ordre ...).*

✓Choix du test

• Les tests paramétriques, quand leur utilisation est justifiée, sont en général plus puissants que les tests non-paramétriques.

• Les tests paramétriques reposent cependant sur l'hypothèse forte que l'échantillon considéré est tiré suivant une distribution appartenant à une famille donnée. Il est possible de s'en affranchir pour des échantillons suffisamment grands en utilisant des théorèmes asymptotiques tels le TCL.

Remarque

Les tests non-paramétriques sont cependant à préférer dans de nombreux cas pratiques pour lesquels les tests paramétriques ne peuvent être utilisés sans violer les postulats dont ils dépendent (notamment les échantillons trop petits).

• Les données sont parfois récupérés sous forme de rangs et non de données brutes. Seuls les tests non-paramétriques sont alors applicables.

5.3 Test paramétrique

5.3.1 Tests de conformité

✓Comparaison d'une moyenne observée et une moyenne théorique

✓✓ Variance connue.

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu;\sigma^2)$ ou la variance est connue.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, c'est le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0;\frac{\sigma^2}{n})$ et par conséquent la statistique

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

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|Z| \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha$ avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]-\infty; -z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}}; +\infty[$$

On calcule alors pour les valeurs de l'échantillon Z et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu > \mu_0$ par exemple, on obtient pour un risque d'erreur α : $\mathbb{P}(Z \leq z_{1-\alpha}) = 1-\alpha$ avec $z_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]z_{1-\alpha}; +\infty[$$

✓✓ Variance inconnue

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu;\sigma^2)$ ou la variance est inconnue.

On veut tester $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, c'est le cas bilatéral.

Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0;\frac{\sigma^2}{n})$. Comme la variance est inconnue, on l'estime par la variance empirique :

$$\acute{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X_n})^2$$

On a déjà vu qu'alors la variable

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

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|T| \leq t_{1-\frac{\alpha}{2}}) = 1-\alpha$ avec $t_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté ; et donc la région de rejet est

$$]-\infty; -t_{1-\frac{\alpha}{2}}[\cup]t_{1-\frac{\alpha}{2}}; +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

✓Comparaison d'une variance observée et une variance théorique

✓✓Moyenne connue.

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu; \sigma^2)$ ou la moyenne est connue.

On veut tester $H_0 : \sigma = \sigma 0$ contre $H_1 : \sigma \neq \sigma_0$. Sous l'hypothèse H_0 la variable aléatoire

$$V = \frac{nS_{n^2}}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \mu}{\sigma_0}\right)^2 \text{ suit une loi du } \chi^2 \text{ à n degrés de libertés.}$$

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\frac{\alpha}{2}}(n) \leq \frac{nS_{n^2}}{\sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n)\right) = 1-\alpha$$

avec$\chi^2_{\frac{\alpha}{2}}(n)$ et$\chi^2_{1-\frac{\alpha}{2}}(n)$ les quantiles d'ordre $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$ de la loi $\chi^2(n)$. Donc la région de rejet est

$$[0; \chi^2_{\frac{\alpha}{2}}(n)[\cup]\chi^2_{1-\frac{\alpha}{2}}(n); +\infty[.$$

On calcule alors pour les valeurs de l'échantillon, V , et on accepte ou on rejette au risque α H_0 suivant la valeur trouvée.

Si on a une hypothèse alternative $H_1 : \sigma > \sigma_0$ on fera un test unilatéral, et obtient au risque α

$$\mathbb{P}\left(\frac{nS_{n^2}}{\sigma_0^2} \leq \chi^2_{1-\alpha}(n)\right) = 1-\alpha$$

avec $\chi^2_{1-\alpha}(n)$ le quantile d'ordre $1-\alpha$ de la loi $\chi^2(n)$. Donc la région de rejet est

$$]\chi^2_{1-\alpha}(n); +\infty[.$$

✓✓Moyenne inconnue.

On suppose que l'on a un échantillon qui suit une loi normale $\mathcal{N}(\mu; \sigma^2)$ ou la moyenne est

connue.
On veut tester $H_0 : \sigma = \sigma 0$ contre $H_1 : \sigma \neq \sigma_0$. Sous l'hypothèse H_0 la variable aléatoire

$$V = \frac{nS_{n^2}}{\sigma_0^2} = \sum_{k=1}^{n} \left(\frac{X_k - \overline{X_n}}{\sigma_0} \right)^2 \text{ suit une loi du } \chi^2 \text{ à n-1 degrés de libertés.}$$

Pour un risque d'erreur α fixé on a donc (en choisissant un intervalle symétrique) :

$$\mathbb{P}\left(\chi^2_{\frac{\alpha}{2}}(n-1) \leq \frac{nS_{n^2}}{\sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n-1) \right) = 1 - \alpha$$

avec$\chi^2_{\frac{\alpha}{2}}(n-1)$ et$\chi^2_{1-\frac{\alpha}{2}}(n-1)$ les quantiles d'ordre $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ de la loi $\chi^2(n-1)$. Donc la région de rejet est

$$[0; \chi^2_{\frac{\alpha}{2}}(n-1)[\cup]\chi^2_{1-\frac{\alpha}{2}}(n-1); +\infty[.$$

On calcule alors pour les valeurs de l'échantillon, V , et on accepte ou on rejette au risque α H_0 suivant la valeur trouvée.

✓Comparaison d'une fréquence observée et une fréquence théorique

Le modèle mathématique est le suivant. On dispose d'une population dans laquelle chaque individu présente ou non un certain caractère, la proportion d'individus présentant le caracère étant notée p, et un échantillon aléatoire de taille n extrait de cette population. La proportion f calculée à partir de l'échantillon est considérée comme une réalisation d'une v.a. de loi binomiale $\mathcal{B}(n; p)$ qu'on peut assimiler, si n est assez grand, à une loi normale $\mathcal{N}(p; \frac{\sqrt{p(1-p)}}{\sqrt{n}})$.

On veut tester $H0 : p = p_0$ contre $H1 : p \neq p_0$, dans le cas bilatéral. On obtient la région de rejet pour un risque α

$$\left] -\infty; p_0 - z_{1-\frac{\alpha}{2}} \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} \right[\cup \left] p_0 + z_{1-\frac{\alpha}{2}} \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}; +\infty \right[$$

5.3.2 Tests d'homogénéité

✓Test de comparaison de deux moyennes

Si les deux échantillons on la même taille $n_1 = n_2 = n$. Le test se ramèène à une test à une moyenne nulle de l'échantillon $(U_1, \ldots, U_n)$, avec $U_i = X_i - Y_i$.

✓✓ Variance connue.

On suppose que l'on a deux échantillons $(X_1, \ldots, X_{n_1})$ et $(Y_1, \ldots, Y_{n_2})$qui suivent une loi

normale $\mathcal{N}(\mu_1;\sigma_1^2)$ et $\mathcal{N}(\mu_2;\sigma_2^2)$ où les variances sont connues.
On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, c'est la cas bilatéral. Sous l'hypothèse H_0 la variable aléatoire $\overline{X}_{n_1} = \frac{1}{n_1}\sum_{k=1}^{n_1} X_k$ suit une loi $\mathcal{N}(\mu_1;\frac{\sigma_1^2}{n_1})$ et $\overline{Y}_{n_2} = \frac{1}{n_2}\sum_{k=1}^{n_2} Y_k$ suit une loi $\mathcal{N}(\mu_2;\frac{\sigma_2^2}{n_2})$, par conséquent la statistique

$$Z = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \text{ suit une loi normale centrée réduite.}$$

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|Z| \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha$ avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]-\infty; -z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}}; +\infty[$$

On calcule alors pour les valeurs de l'échantillon Z et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.
Si on considère un test unilatéral et une hypothèse alternative $H_1 : \mu_1 > \mu_2$ par exemple, on obtient pour un risque d'erreur α : $\mathbb{P}(Z \leq z_{1-\alpha}) = 1-\alpha$ avec $u_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]z_{1-\alpha}; +\infty[$$

✓✓ Variance inconnue.
On suppose que l'on a deux échantillons $(X_1,\ldots,X_{n_1})$ et $(Y_1,\ldots,Y_{n_2})$qui suivent une loi normale $\mathcal{N}(\mu_1;\sigma_1^2)$ et $\mathcal{N}(\mu_2;\sigma_2^2)$ où les variances sont inconnues.
•*Cas 1 : n_1 et n_2 supérieurs à 30.*
On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, c'est la cas bilatéral. Sous l'hypothèse H_0 la variable aléatoire $\overline{X}_{n_1} - \overline{Y}_{n_1}$ suit une loi $\mathcal{N}(0; n_1\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\acute{S}_{n_1}^2 + \acute{S}_{n_2}^2 = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(X_k - \overline{X}_{n_1})^2 + \frac{1}{n_2-1}\sum_{k=1}^{n_2}(Y_k - \overline{Y}_{n_2})^2$$

Alors la variable aléatoire

$$Z = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\acute{S}_{n_1}^2}{n_1} + \frac{\acute{S}_{n_2}^2}{n_2}}} \text{ peut être approximé par loi normale centrée réduite.}$$

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|Z| \leq z_{1-\frac{\alpha}{2}}) = 1-\alpha$ avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]-\infty; -z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}}; +\infty[$$

On calcule alors pour les valeurs de l'échantillon Z et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

•*Cas 2 : n_1 et n_2 inférieur à 30 et $\sigma_1 = \sigma_2$*

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, c'est la cas bilatéral. Sous l'hypothèse H_0 la variable aléatoire $\overline{X}_{n_1} - \overline{Y}_{n_1}$ suit une loi $\mathcal{N}(0; n_1\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\acute{S}^2_{n_1,n_2} = \frac{1}{n_1+n_2-2}\left(\sum_{k=1}^{n_1}(X_k - \overline{X}_{n_1})^2 + \sum_{k=1}^{n_2}(Y_k - \overline{Y}_{n_2})^2\right)$$

Alors la variable aléatoire

$$T = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\acute{S}^2_{n_1,n_2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \text{ suit une de Student à} n_1+n_2-2 \text{ degrés de liberté.}$$

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|T| \leq t_{1-\frac{\alpha}{2}}) = 1-\alpha$ avec $t_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à n_1+n_2-2 degrés de liberté ; et donc la région de rejet est

$$]-\infty; -t_{1-\frac{\alpha}{2}}[\cup]t_{1-\frac{\alpha}{2}}; +\infty[$$

On calcule alors pour les valeurs de l'échantillon T et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

•*Cas 3 : n_1 et n_2 inférieur à 30 et $\sigma_1 \neq \sigma_2$*

On veut tester $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$, c'est la cas bilatéral. Sous l'hypothèse H_0 la variable aléatoire $\overline{X}_{n_1} - \overline{Y}_{n_1}$ suit une loi $\mathcal{N}(0; n_1\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. Comme la variance est inconnue, on l'estime par la variance empirique corrigée

$$\acute{S}^2_{n_1} + \acute{S}^2_{n_2} = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(X_k - \overline{X}_{n_1})^2 + \frac{1}{n_2-1}\sum_{k=1}^{n_2}(Y_k - \overline{Y}_{n_2})^2$$

Alors la variable aléatoire

$$T = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\acute{S}^2_{n_1}}{n_1} + \frac{\acute{S}^2_{n_2}}{n_2}}} \text{ suit une de Student à} \nu \text{ degrés de liberté.}$$

où ν est l'entier le plus proche de

$$\frac{\left(\frac{\acute{S}^2_{n_1}}{n_1}+\frac{\acute{S}^2_{n_2}}{n_2}\right)^2}{(n_1-1)\frac{\acute{S}^4_{n_1}}{n_1}+(n_2-1)\frac{\acute{S}^4_{n_2}}{n_2}}$$

Pour un risque d'erreur α fixé on a donc $\mathbb{P}(|T|\leq t_{1-\frac{\alpha}{2}})=1-\alpha$ avec $t_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student précédente ; et donc la région de rejet est

$$]-\infty;-t_{1-\frac{\alpha}{2}}[\cup]t_{1-\frac{\alpha}{2}};+\infty[$$

On calcule alors pour les valeurs de l'échantillon Z et on accepte ou on rejette H_0 suivant la valeur trouvée, au risque α.

✓Test de comparaison de deux variances

On dispose de deux échantillons d'écart-types respectifs $\acute{S}^2_{n_1}=\frac{1}{n_1-1}\sum_{k=1}^{n_1}(X_k-\overline{X}_{n_1})^2$ et $\acute{S}^2_{n_2}=\frac{1}{n_2-1}\sum_{k=1}^{n_2}(Y_k-\overline{Y}_{n_2})^2$. On se demande s'il est raisonnable de penser que les deux échantillons ont été tirés suivant des lois de même écart-type ou si ils sont signifcativement différents.

$$\begin{cases} H_0: & \sigma_1=\sigma_2 \\ H_1: & \sigma_1\neq\sigma_2 \end{cases}$$

Statistique

$$F=\frac{\acute{S}^2_{n_1}}{\acute{S}^2_{n_2}}$$

Sous l'hypothèse H_0 la statistique F suit une loi de Fisher-Snedecor $\mathcal{F}(n_1-1,n_2-1)$ à n_1-1 et n_2-1 degrés de liberté. Pour un risque d'erreur α fixé on a une région de rejet $[0;F\frac{\alpha}{2}[\cup]F\frac{\alpha}{2};+\infty[$ où les quantiles sont déterminées à l'aide de la loi précédente.

✓Test de comparaison de deux proportions

On veut comparer deux proportions p_1 et p_2 à partir de deux échantillons. Le modèle mathématique est le suivant. On considère les proportions f_1 et f_2 associés aux deux échantillons. On veut tester

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

On prend la statistique

$$Z=\frac{f_1-f_2}{\sqrt{F(1-F)\left(\frac{1}{n_1}\frac{1}{n_2}\right)}} \text{ avec } F=\frac{n_1f_1+n_2f_2}{n_1+n_2}$$

On obtient la région de rejet pour un risque α

$$]-\infty;-z_{1-\frac{\alpha}{2}}[\cup]z_{1-\frac{\alpha}{2}};+\infty[$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0;1)$

5.4 Test non paramétrique "Tests du χ^2"

Le test du χ^2 s'applique a des variables qualitatives a plusieurs modalites.

5.4.1 Test d'indépendance

Étude de la liaison entre deux caractères qualitatifs X et Y.

$$\begin{cases} H_0 & : \quad \text{Les variables X et Y sont indépendantes.} \\ H_1 & : \text{Les variables X et Y ne sont pas indépendantes.} \end{cases}$$

Nous considérons donc le tableau suivant :

X\ Y	Modalité 1	...	Modalité j	...	Modalité J	Total
Modalité 1	m_{11}	...	m_{1j}	...	m_{1J}	$m_{1\bullet}$
⋮	⋮	...	⋮	⋮	⋮	⋮
Modalité i	m_{i1}	...	m_{ij}	...	m_{iJ}	$m_{i\bullet}$
⋮	⋮	...	⋮	⋮	⋮	⋮
Modalité I	m_{I1}	...	m_{Ij}	...	m_{IJ}	$m_{I\bullet}$
Total	$m_{\bullet 1}$	...	$m_{\bullet j}$	...	$m_{\bullet J}$	$m_{\bullet\bullet}=n$

où m_{ij} correspond aux nombre d'individus observés ayant la modalité i pour X et la modalité j pour Y.
La notation $m_{i\bullet}$ correspond à $\sum_{j=1}^{J} m_{ij}$ et la notation $m_{\bullet j}$ correspond à $\sum_{i=1}^{I} m_{ij}$
Le principe du test consiste à comparer les effectifs tels que nous les avons, à la répartition que nous aurions si les variables étaient indépendantes. Dans ce cas, en considérant que les marges

$$(m_{1\bullet},\ldots,m_{i\bullet},\ldots,m_{I\bullet},m_{\bullet 1},\ldots,m_{\bullet j},\ldots,m_{\bullet J})$$

sont fixées, nous pouvons calculer cette répartition théorique dans chacun des échantillons. Nous avons alors :

$$c_{ij}=\frac{m_{i\bullet}m_{\bullet j}}{n}$$

Il s'agit donc des effectifs théoriques sous l'hypothèse d'indépendance H_0. Afin d'étudier l'écart entre ces deux répartitions, observée et théorique, nous adoptons l'indice suivant, dû à Pearson

$$\chi^2_{obs} = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(m_{ij} - c_{ij})^2}{c_{ij}}$$

Les conditions d'application du test détaillé ci-dessous sont

$$c_{ij} \geq 5 \text{ et } n \geq 50$$

Sous l'hypothèse nulle H_0 et lorsque les conditions d'application du test sont remplies, χ^2_{obs} est une réalisation d'une variable aléatoire qui suit approximativement une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté.
Pour un seuil fixe α, les tables de la loi du χ^2 à $(I-1)(J-1)$ degrés de liberté, nous fournissent une valeur critique c telle que $\mathbb{P}\left[\chi^2_{(I-1)(J-1)} \leq c\right] = 1-\alpha$. Si nous utilisons un logiciel de statistique celui-ci nous fournit une p-valeur.
Règles de décision
Alors nous décidons

$$\begin{cases} H_0 & \text{est vraie si} \quad \chi^2_{obs} < c \\ H_1 & \text{est vraie si} \quad \chi^2_{obs} \geq c \end{cases}$$

Dans le cas où nous ne pouvons pas rejeter l'hypothèse nulle H_0 et par conséquent nous l'acceptons, nous devrions calculer le risque de seconde espèce du test. Dans le cadre de ce livre, nous ne donnerons pas la formule pour calculer le risque et par conséquent la puissance du test.

1. Lorsque les conditions ne sont pas remplies, il existe des corrections, par exemple celle de Yates, ou le test exact de Fisher dans le cas de deux variables qualitatives à deux modalités.
2. S'il y a plus de deux modalités, nous pouvons essayer d'en regrouper si cela est possible, c'est-à-dire si cela a un sens.
3. Ce test, tel qu'il est exposé, ne peut pas être appliqué à des échantillons appariés.

5.4.2 Test d'adéquation d'une loi à une loi donnée

Le test suivant est un test adapte pour s'intéresser a la possibilité d'une adéquation d'une distribution a une loi de probabilité donnée. Il est adapte pour des lois de

probabilité discrètes et peut être également utilise pour une loi continue entièrement specifiée. Nous détaillons son application au cas d'une loi discrète de support fini définie par $(q_{x_k})_{x_k \in \mathcal{X}}$ avec $Card\mathcal{X} = n$.

$$\begin{cases} H_0 & : & (p_{x_k})_{x_k \in \mathcal{X}} = (q_{x_k})_{x_k \in \mathcal{X}} \\ H_1 & : & (p_{x_k})_{x_k \in \mathcal{X}} \neq (q_{x_k})_{x_k \in \mathcal{X}} \end{cases}$$

Nous mesurons ensuite la distance entre les effectifs observés m_i et les effectifs théoriques c_i de la même façon que dans le paragraphe précédent

$$\chi^2_{obs} = \sum_{i=1}^{I} \frac{(m_i - c_i)^2}{c_i}$$

Les conditions d'application du test détaillé ci-dessous sont

$$c_i \geq 5 \text{ et } n \geq 50$$

Sous l'hypothèse nulle H_0 et lorsque les conditions d'application du test sont remplies, χ^2_{obs} est une réalisation d'une variable aléatoire qui suit approximativement une loi du χ^2 a (I-1) degrés de liberté, ou I est le nombre de modalités. La règle de décision est identique au cas précédent.

5.5 Exercices corrigés

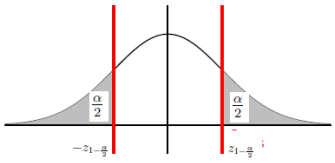
Exercice 01

Quand les gens fument, la nicotine qu'ils absorbent est transformée en cotonnine qui peut être mesurée. Un échantillon de 40 fumeurs a un niveau moyen de cotonnine de 172.5 ng/ml. Supposant que σ est connue et vaut 119.5.

– Utiliser un niveau de significativité de 0.01 pour tester l'affirmation que le niveau moyen de cotonnine pour tous les fumeurs est de 200.

Solution.

$$\begin{cases} H_0 & : & \mu = 200 \\ H_1 & : & \mu \neq 200 \end{cases}$$ c'est le cas bilatéral.



Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0; \frac{\sigma^2}{n})$ et par conséquent la statistique

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -1.46, \text{ la statistique suit une loi normale centrée réduite.}$$

Pour un risque d'erreur $\alpha = 0.01$ on a donc $\mathbb{P}(|Z| \leq z_{1-\frac{0.01}{2}}) = 1 - 0.01 = 0.99$ alors $z_{1-\frac{0.01}{2}} = 2.575$ le quantile d'ordre 0.995 de la loi $\mathcal{N}(0;1)$; et donc la région de rejet est

$$]-\infty; -2.575[\cup]2.575; +\infty[,$$

échec du rejet de H_0. Il n'y a pas suffisamment de preuves pour garantir le rejet de l'affirmation selon laquelle la moyenne est égale à 200.

Exercice 02

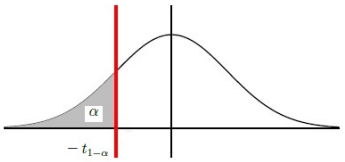
Un échantillon de paquet de céréales est sélectionné aléatoirement et le contenu en sucre est enregistré. Ces quantités sont résumées par les statistiques :

$$n = 16, \bar{x} = 0.295g, s' = 0.168.$$

– Utiliser le niveau de significativité de 0.05 pour tester l'affirmation que le contenu moyen en sucre est inférieur à 0.3 g.

Solution.

$$\begin{cases} H_0 & : & \mu = 0.3 \\ H_1 & : & \mu < 0.3 \end{cases}$$ c'est le cas unilatéral.



Sous l'hypothèse H_0 la variable aléatoire $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ suit une loi $\mathcal{N}(\mu_0; \frac{\sigma^2}{n})$. Comme la variance est inconnue, on l'estime par la variance empirique $\acute{s}_n^2 = 0.168^2$:
On a

$$t = \frac{\overline{x} - \mu_0}{\frac{\acute{s}_n}{\sqrt{n}}} = -0.119, \text{ la statistique suit une loi de Student à } n-1 \text{ degrés de liberté.}$$

On obtient pour un risque d'erreur $\alpha = 0.05$, avec $t_{1-\alpha} = 1.753$ le quantile du risque 0.05 de la loi de Student à $n - 1 = 15$ degrés de liberté ; et donc la région de rejet est

$$]-\infty; -1.753[$$

échec du rejet de H_0., il n'y a pas suffisamment de preuves pour confirmer l'affirmation selon laquelle la moyenne est inférieur à 0.3.

Exercice 03

Dans un sondage d'une revue américaine sur 1012 adultes sélectionnés aléatoirement, 9% ont dit que le clonage humain pourrait être autorisé.

- Utiliser un niveau de significativité de 0.05 pour tester l'affirmation que moins de 10% des adultes disent que le clonage humain pourrait être autorisé.

Solution.

$$\begin{cases} H_0 & : & p = 0.1 \\ H_1 & : & p < 0.1 \end{cases}$$

La proportion f calculée à partir de l'échantillon est considérée comme une réalisation d'une v.a. de loi binomiale $\mathcal{B}(n; p)$ qu'on peut assimiler, comme $n = 1012$ est assez grand, à une loi normale $\mathcal{N}(p_0; \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}})$.

$$z = \frac{f - p_0}{\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}} = -1.06,$$

On obtient la région de rejet pour un risque $\alpha = 0.05$

$$]-\infty; -1.645[$$

échec du rejet de H0. Il n'y a pas suffisamment de preuves pour confirmer l'affirmation selon laquelle moins de 10% des adultes disent que le clonage des humains devrait être autorisé.

Exercice 04

Des dosages de calcium sur deux échantillons de yaourt ont donné les résultats suivants (en mg de Ca). On suppose que la variance de population est la même dans les deux cas.

	1er échantillon	2ème échantillon
Effectif	11	9
Moyenne	3.92	4.18
Variance estimée	0.3130	0.4231

1. Par quelle formule peut-on calculer l'estimation conjointe de cette variance ?
2. Les moyennes des deux échantillons diffèrent elles significativement ?

Solution.

1. $\acute{s}^2_{n_1,n_2} = (10 \times 0.313 + 8 \times 0.4231)/18 = 0.36$.
2. $$\begin{cases} H_0 & : & \mu_1 = \mu_2 \\ H_1 & : & \mu_1 \neq \mu_2 \end{cases}$$

   $t = (3.92 - 4.18)/(0.36 \times ((1/11) + (1/9)))^0.5 = -0.964$. les valeurs critiques sont -2.101 et 2.101 (On obtient pour un risque d'erreur $\alpha = 0.05$, avec $t_\alpha = 2.101$ le quantile du risque 0.05 de la loi de Student à $n_1 + n_2 - 2 = 18$ degrés de liberté), non rejet de H_0. La différence des moyennes n'est pas significative.

Exercice 05

Des truites sont mesurées sur deux échantillons. Le premier échantillon est composé de 50 truites d'élevage et donne $\bar{x}_a = 158.86$ mm , $s^2_a = 37.18$(variance estimé). Le second échantillon est composé de 67 truites de rivière et donne $\bar{x}_b = 134.46$, $s^2_b = 25.92$(variance estimé).

– Les moyennes de ces deux échantillons diffèrent elles significativement ?

Solution.

$$\begin{cases} H_0 & : & \mu_1 = \mu_2 \\ H_1 & : & \mu_1 \neq \mu_2 \end{cases}$$

$t = (158.86 - 134.46)/((37.18/49) + (25.92/66))^0.5 = 22.74$. $t(\min(50-1, 67-1))$ qui peut être approximé par une valeur de normale , les valeurs critiques sont -1.96, 1.96 d'où le

rejet de H_0. La différence des deux moyennes est significative.

Exercice 06

En 1908, William Gosset sous le pseudonyme de " student " publie un article où il inclut les données ci dessous, concernant les rendements de deux types de graines de maïs (habituelles et séchées au four) utilisées sur deux lots de terrain adjacents. Les valeurs sont des rendements des cultures en tonne par hectare.

habituel	2.156	2.548	2.576	2.576	2.52	2.212	2.744	1.736	2.016	1.596	1.904
séché	2.81	2.688	2.688	3.136	2.52	2.184	2.492	1.792	1.932	1.764	1.932

– Utiliser un niveau de significativité de 0.05 pour tester l'affirmation qu'il n'y a pas de différence entre les rendements des deux types de graines.

Solution.

d : -0.644, -0.140 , -0.112, -0.560, 0.000, 0.028, 0.252, -0.056, 0.084, -0.168, -0.028 $\bar{d} = -0.122, s'_d = 0.3$,

$$\begin{cases} H_0 & : & \mu_d = 0 \\ H_1 & : & \mu_d \neq 0 \end{cases}$$

t= -1.32, les valeurs critiques (table de student) $\pm$ 2.228, non rejet de H_0. Il n'y a pas suffisamment de preuves pour dire qu'il y n'y a pas une différence entre les rendements des deux types de graines

Exercice 07

Dans une étude sur le daltonisme, 500 hommes et 2100 femmes ont été sélectionnés et testés aléatoirement. Parmi les hommes, 45 sont daltoniens. Parmi es femmes 6 sont daltoniennes (basé sur des données de USA Today).

– Y a -t il suffisamment de preuves pour confirmer l'affirmation que les hommes ont un taux de daltonisme plus élevé que les femmes ? Utiliser un niveau de significativité de 0.01.

Solution. :

$$\begin{cases} H_0 & : & p_1 = p_2 \\ H_1 & : & p_1 > p_2 \end{cases}$$

$\bar{p} = 0.0196$, z= 12.633, valeur critique (table normale) 2.326 rejet de H_0. Il y a suffisamment de preuves pour confirmer l'affirmation que les hommes ont un taux de daltonisme plus élevé que les femmes.

Exercice 08

On veut étudier la liaison entre les caractères : "être fumeur" (plus de 20 cigarettes par

jour, pendant 10 ans) et "avoir un cancer de la gorge", sur une population de 1000 personnes, dont 500 sont atteintes d'un cancer de la gorge. Voici les résultats observés :

Observé	cancer	non cancer	marge
fumeur	342	258	600
non fumeur	158	242	400
marge	500	500	1000

Faire un test d'indépendance pour établir la liaison entre ces caractères avec seuil de risque est 0.001 (0.1%).

$$(\mathcal{X}^2_{0.001} = 10.83)$$

Solution. :
Mise en œuvre du test :

1. Le risque : 0.1%. Pour étudier la dépendance de ces caractères faisons l'hypothèse H_0 : "lesdeux caractères sont indépendants" et voyons ce qui se passerait sous cette hypothèse.
Notons les événements :
C : "avoir un cancer dans la population observée"
F : "être fumeur dans la population observée"
Si les événements F et C sont indépendants, alors : $\mathbb{P}_(C \cap F) = \mathbb{P}(F) * \mathbb{P}(C)$ et de même pour les trois autres possibilités : $\mathbb{P}_(\bar{C} \cap \bar{F}), \mathbb{P}_(\bar{C} \cap F), \mathbb{P}_(C \cap \bar{F})$, quantités que l'on peut donc calculer sous H_0 : $\mathbb{P}(F) = 600/1000, \mathbb{P}(C) = 500/1000, \mathbb{P}(F) * \mathbb{P}(C) = 3/10$, alors l'effectif théorique correspondant à la catégorie "fumeur et cancéreux" est de 300.

2. On en déduit le tableau théorique sous H0 :

Théorique	cancer	non cancer	marge
fumeur	300	300	600
non fumeur	200	200	400
marge	500	500	1000

3. On calcule alors la valeur de $\mathcal{X}^2_c$:

$\mathcal{X}^2_c = \Sigma^4_{i=1} \frac{(O_i - T_i)^2}{T_i}$, on obtient, $\mathcal{X}^2_c = 34.73$. On a précisé le risque $\alpha = 0.001$, on lit dans la table du khi-deux à un $((nl-1)*(nc-1) = (2-1)*(2-1) = 1)$ degré de liberté : $P[\mathcal{X}^2_{0.001} > 10.83] = 0.001$ et le $\mathcal{X}^2_c$ calculé est 34.73.

4. On décide de rejeter H_0

Ainsi, en rejetant l'hypothèse de l'indépendance des caractères "être fumeur" et " avoir un cancer de la gorge", on a moins de une chance sur 1000 de se tromper, puisque moins de un tableau possible sur mille conduit à un calcul de $\mathcal{X}_c^2$ plus grand que 10.83 ; beaucoup moins sans doute, conduiraient à un calcul de $\mathcal{X}_c^2$ plus grand que 34.73.

Exercice 09

Une chercheuse a développé un modèle théorique pour prédire la couleur des yeux. Après avoir examiné un échantillon aléatoire de parents elle prédit la couleur des yeux du premier enfant, le tableau ci dessous liste la couleur des yeux des enfants. Selon sa théorie, elle prédit que 87% des enfants devraient avoir les yeux marrons, 8% devraient avoir les yeux bleus et 5% devraient avoir les yeux verts.

	Yeux marrons	yeux bleus	yeux verts
fréquences	132	17	0

– Utiliser un niveau de significativité de 0.05 pour tester l'affirmation que les fréquences actuelles correspondent aux fréquences qu'elle a prédites.

Solution.

$$\begin{cases} H_0 : & \text{les fréquences actuelles correspondent aux fréquences prédites.} \\ H_1 : & \text{les fréquences actuelles ne correspondent pas aux fréquences prédites} \end{cases}$$

(Il s'agit d'un test d'ajustement)

Effectifs	Yeux marrons	Yeux bleus	Yeux verts
No_i	132	17	0
P_i	0.87	0.08	0.05
Nt_i	129.63	11.92	7.45
$No_i - Nt_i$	2.37	5.08	-7.45
$(No_i - Nt_i)^2$	5.62	25.81	55.50
$((No_i - Nt_i)^2)/Nt_i$	0.04	2.16	7.45

$\chi_c^2 = 9.66 > 5.991$ (la valeur critique de la table de khi deux ddl=3-0-1=2, colonne 0.05) ; rejet de H_0 Il y a suffisamment de preuves pour garantir le rejet de l'affirmation selon laquelle les fréquence observées correspondent aux fréquences prédites.

Exercice 10

Le tableau suivant fournit des données sur plusieurs méthodes pour arrêter de fumer.

	Gomme de nicotine	Patch de nicotine	Inhalateur de nicotine
Fume encore	191	263	95
Arrêté de fumer	59	57	27

– Tester avec un niveau de significativité de 0.05 l'hypothèse qu'arrêter de fumer est indépendante de la méthode

Solution.

$$\begin{cases} H_0 & : \quad \text{arrêter de fumer est indépendant de la méthode.} \\ H_1 & : \quad \text{arrêter de fumer n'est pas indépendant de la méthode.} \end{cases}$$

	Gomme de nicotine	Patch de nicotine	Inhalateur de nicotine
Fume encore	198.34	253.83	96.97
Arrêté de fumer	51.66	66.13	25.21

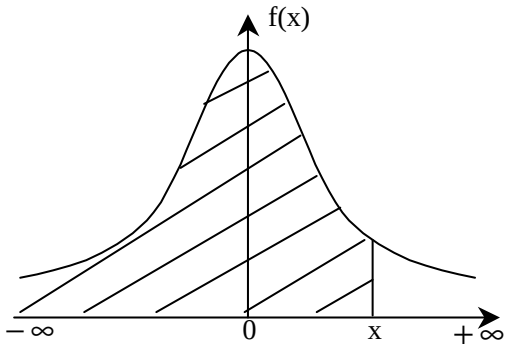
$\chi^2_c = 3.06 < 5.991$ (la valeur critique de la table de khi deux ddl=(3-1)*(2-1)=2, colonne 0.05) ; non rejet de H0. Il n'y a pas suffisamment de preuves pour garantir l'affirmation selon laquelle arrêter le tabac est indépendant de la méthode.

ANNEXE : Tables Statistiques usuelles

1. Loi Normale centrée réduite
2. Loi de Student
3. Loi du Khi-deux
4. Loi de Fisher-Snedecor

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.



$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2p}} e^{-\frac{u^2}{2}} du$$

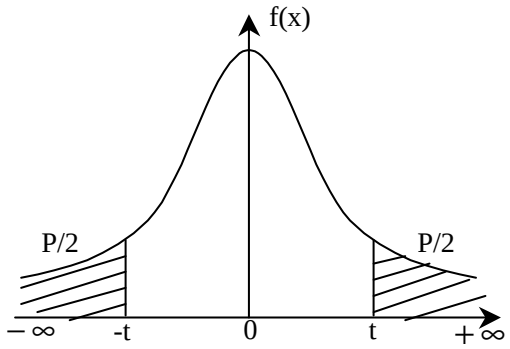
X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Table pour les grandes valeurs de x :

x	3	3,2	3,4	3,6	3,8	4	4,2	4,4	4,6	4,8
F(x)	0,99865003	0,99931280	0,99966302	0,99984085	0,99992763	0,99996831	0,99998665	0,99999458	0,99999789	0,99999921

Loi de Student

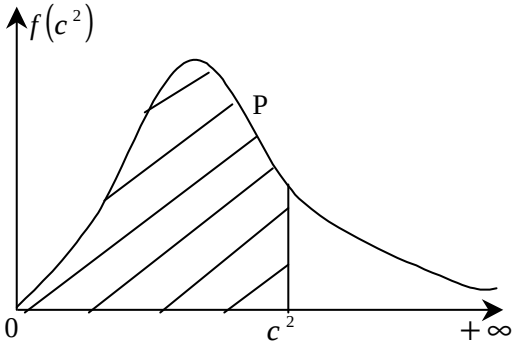
Valeurs de t ayant la probabilité P d'être dépassées en valeur absolue.



n \P	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	1%
1	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3137	12,7062	63,6559
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	9,9250
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	5,8408
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318	2,7765	4,6041
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	4,0321
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,7074
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	3,4995
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	3,3554
9	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	3,2498
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	3,1693
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	3,1058
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	3,0545
13	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	3,0123
14	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613	2,1448	2,9768
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1315	2,9467
16	0,1277	0,2576	0,3923	0,5350	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,9208
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,0690	1,3334	1,7396	2,1098	2,8982
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,8620	1,0672	1,3304	1,7341	2,1009	2,8784
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,8610	1,0655	1,3277	1,7291	2,0930	2,8609
20	0,1273	0,2567	0,3909	0,5329	0,6870	0,8600	1,0640	1,3253	1,7247	2,0860	2,8453
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,8314
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,8188
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,8073
24	0,1270	0,2562	0,3900	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,7970
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,7874
26	0,1269	0,2560	0,3896	0,5309	0,6840	0,8557	1,0575	1,3150	1,7056	2,0555	2,7787
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,7707
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,0560	1,3125	1,7011	2,0484	2,7633
29	0,1268	0,2557	0,3892	0,5302	0,6830	0,8542	1,0553	1,3114	1,6991	2,0452	2,7564
30	0,1267	0,2556	0,3890	0,5300	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,7500
40	0,1265	0,2550	0,3881	0,5286	0,6807	0,8507	1,0500	1,3031	1,6839	2,0211	2,7045
50	0,1263	0,2547	0,3875	0,5278	0,6794	0,8489	1,0473	1,2987	1,6759	2,0086	2,6778
60	0,1262	0,2545	0,3872	0,5272	0,6786	0,8477	1,0455	1,2958	1,6706	2,0003	2,6603
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,6387
100	0,1260	0,2540	0,3864	0,5261	0,6770	0,8452	1,0418	1,2901	1,6602	1,9840	2,6259
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6576	1,9799	2,6174
200	0,1258	0,2537	0,3859	0,5252	0,6757	0,8434	1,0391	1,2858	1,6525	1,9719	2,6006
∞	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,9600	2,5758

Loi du c^2

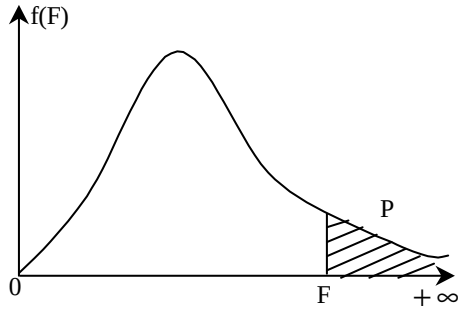
Valeur de c^2 ayant la probabilité P d'être dépassée.



ddl/P	0,5%	1,0%	2,5%	5,0%	10,0%	50,0%	90,0%	95,0%	97,5%	99,0%	99,5%
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	1,647	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,041	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	12,878	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,335
30	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	30,336	41,422	44,985	48,232	52,191	55,002
32	15,134	16,362	18,291	20,072	22,271	31,336	42,585	46,194	49,480	53,486	56,328
33	15,815	17,073	19,047	20,867	23,110	32,336	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	33,336	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	34,336	46,059	49,802	53,203	57,342	60,275

Lorsque $n > 30$ on peut admettre que la quantité $\sqrt{2c^2} - \sqrt{2n-1}$ suit une loi normale centrée réduite.

Loi de Fisher-Snedecor



Valeurs de F ayant 5% de chances d'être dépassées.

n_2 \ n_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	161,446	199,499	215,707	224,583	230,160	233,988	238,884	241,882	243,905	247,324	249,052	250,096	251,774	252,196	253,254
2	18,513	19,000	19,164	19,247	19,296	19,329	19,371	19,396	19,412	19,440	19,454	19,463	19,476	19,479	19,487
3	10,128	9,552	9,277	9,117	9,013	8,941	8,845	8,785	8,745	8,675	8,638	8,617	8,581	8,572	8,549
4	7,709	6,944	6,591	6,388	6,256	6,163	6,041	5,964	5,912	5,821	5,774	5,746	5,699	5,688	5,658
5	6,608	5,786	5,409	5,192	5,050	4,950	4,818	4,735	4,678	4,579	4,527	4,496	4,444	4,431	4,398
6	5,987	5,143	4,757	4,534	4,387	4,284	4,147	4,060	4,000	3,896	3,841	3,808	3,754	3,740	3,705
7	5,591	4,737	4,347	4,120	3,972	3,866	3,726	3,637	3,575	3,467	3,410	3,376	3,319	3,304	3,267
8	5,318	4,459	4,066	3,838	3,688	3,581	3,438	3,347	3,284	3,173	3,115	3,079	3,020	3,005	2,967
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,137	3,073	2,960	2,900	2,864	2,803	2,787	2,748
10	4,965	4,103	3,708	3,478	3,326	3,217	3,072	2,978	2,913	2,798	2,737	2,700	2,637	2,621	2,580
11	4,844	3,982	3,587	3,357	3,204	3,095	2,948	2,854	2,788	2,671	2,609	2,570	2,507	2,490	2,448
12	4,747	3,885	3,490	3,259	3,106	2,996	2,849	2,753	2,687	2,568	2,505	2,466	2,401	2,384	2,341
13	4,667	3,806	3,411	3,179	3,025	2,915	2,767	2,671	2,604	2,484	2,420	2,380	2,314	2,297	2,252
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,602	2,534	2,413	2,349	2,308	2,241	2,223	2,178
15	4,543	3,682	3,287	3,056	2,901	2,790	2,641	2,544	2,475	2,353	2,288	2,247	2,178	2,160	2,114
16	4,494	3,634	3,239	3,007	2,852	2,741	2,591	2,494	2,425	2,302	2,235	2,194	2,124	2,106	2,059
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,450	2,381	2,257	2,190	2,148	2,077	2,058	2,011
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,412	2,342	2,217	2,150	2,107	2,035	2,017	1,968
19	4,381	3,522	3,127	2,895	2,740	2,628	2,477	2,378	2,308	2,182	2,114	2,071	1,999	1,980	1,930
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,348	2,278	2,151	2,082	2,039	1,966	1,946	1,896
21	4,325	3,467	3,072	2,840	2,685	2,573	2,420	2,321	2,250	2,123	2,054	2,010	1,936	1,916	1,866
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,297	2,226	2,098	2,028	1,984	1,909	1,889	1,838
23	4,279	3,422	3,028	2,796	2,640	2,528	2,375	2,275	2,204	2,075	2,005	1,961	1,885	1,865	1,813
24	4,260	3,403	3,009	2,776	2,621	2,508	2,355	2,255	2,183	2,054	1,984	1,939	1,863	1,842	1,790
25	4,242	3,385	2,991	2,759	2,603	2,490	2,337	2,236	2,165	2,035	1,964	1,919	1,842	1,822	1,768
26	4,225	3,369	2,975	2,743	2,587	2,474	2,321	2,220	2,148	2,018	1,946	1,901	1,823	1,803	1,749
27	4,210	3,354	2,960	2,728	2,572	2,459	2,305	2,204	2,132	2,002	1,930	1,884	1,806	1,785	1,731
28	4,196	3,340	2,947	2,714	2,558	2,445	2,291	2,190	2,118	1,987	1,915	1,869	1,790	1,769	1,714
29	4,183	3,328	2,934	2,701	2,545	2,432	2,278	2,177	2,104	1,973	1,901	1,854	1,775	1,754	1,698
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,165	2,092	1,960	1,887	1,841	1,761	1,740	1,683
31	4,160	3,305	2,911	2,679	2,523	2,409	2,255	2,153	2,080	1,948	1,875	1,828	1,748	1,726	1,670
32	4,149	3,295	2,901	2,668	2,512	2,399	2,244	2,142	2,070	1,937	1,864	1,817	1,736	1,714	1,657
33	4,139	3,285	2,892	2,659	2,503	2,389	2,235	2,133	2,060	1,926	1,853	1,806	1,724	1,702	1,645
34	4,130	3,276	2,883	2,650	2,494	2,380	2,225	2,123	2,050	1,917	1,843	1,795	1,713	1,691	1,633
35	4,121	3,267	2,874	2,641	2,485	2,372	2,217	2,114	2,041	1,907	1,833	1,786	1,703	1,681	1,623
40	4,085	3,232	2,839	2,606	2,449	2,336	2,180	2,077	2,003	1,868	1,793	1,744	1,660	1,637	1,577
50	4,034	3,183	2,790	2,557	2,400	2,286	2,130	2,026	1,952	1,814	1,737	1,687	1,599	1,576	1,511
80	3,960	3,111	2,719	2,486	2,329	2,214	2,056	1,951	1,875	1,734	1,654	1,602	1,508	1,482	1,411
100	3,936	3,087	2,696	2,463	2,305	2,191	2,032	1,927	1,850	1,708	1,627	1,573	1,477	1,450	1,376
120	3,920	3,072	2,680	2,447	2,290	2,175	2,016	1,910	1,834	1,690	1,608	1,554	1,457	1,429	1,352

Valeurs de F ayant 2,5% de chances d'être dépassées.

n_2 \ n_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	647,793	799,482	864,151	899,599	921,835	937,114	956,643	968,634	976,725	990,345	997,272	1001,405	1008,098	1009,787	1014,036
2	38,506	39,000	39,166	39,248	39,298	39,331	39,373	39,398	39,415	39,442	39,457	39,465	39,478	39,481	39,489
3	17,443	16,044	15,439	15,101	14,885	14,735	14,540	14,419	14,337	14,196	14,124	14,081	14,010	13,992	13,947
4	12,218	10,649	9,979	9,604	9,364	9,197	8,980	8,844	8,751	8,592	8,511	8,461	8,381	8,360	8,309
5	10,007	8,434	7,764	7,388	7,146	6,978	6,757	6,619	6,525	6,362	6,278	6,227	6,144	6,123	6,069
6	8,813	7,260	6,599	6,227	5,988	5,820	5,600	5,461	5,366	5,202	5,117	5,065	4,980	4,959	4,904
7	8,073	6,542	5,890	5,523	5,285	5,119	4,899	4,761	4,666	4,501	4,415	4,362	4,276	4,254	4,199
8	7,571	6,059	5,416	5,053	4,817	4,652	4,433	4,295	4,200	4,034	3,947	3,894	3,807	3,784	3,728
9	7,209	5,715	5,078	4,718	4,484	4,320	4,102	3,964	3,868	3,701	3,614	3,560	3,472	3,449	3,392
10	6,937	5,456	4,826	4,468	4,236	4,072	3,855	3,717	3,621	3,453	3,365	3,311	3,221	3,198	3,140
11	6,724	5,256	4,630	4,275	4,044	3,881	3,664	3,526	3,430	3,261	3,173	3,118	3,027	3,004	2,944
12	6,554	5,096	4,474	4,121	3,891	3,728	3,512	3,374	3,277	3,108	3,019	2,963	2,871	2,848	2,787
13	6,414	4,965	4,347	3,996	3,767	3,604	3,388	3,250	3,153	2,983	2,893	2,837	2,744	2,720	2,659
14	6,298	4,857	4,242	3,892	3,663	3,501	3,285	3,147	3,050	2,879	2,789	2,732	2,638	2,614	2,552
15	6,200	4,765	4,153	3,804	3,576	3,415	3,199	3,060	2,963	2,792	2,701	2,644	2,549	2,524	2,461
16	6,115	4,687	4,077	3,729	3,502	3,341	3,125	2,986	2,889	2,717	2,625	2,568	2,472	2,447	2,383
17	6,042	4,619	4,011	3,665	3,438	3,277	3,061	2,922	2,825	2,652	2,560	2,502	2,405	2,380	2,315
18	5,978	4,560	3,954	3,608	3,382	3,221	3,005	2,866	2,769	2,596	2,503	2,445	2,347	2,321	2,256
19	5,922	4,508	3,903	3,559	3,333	3,172	2,956	2,817	2,720	2,546	2,452	2,394	2,295	2,270	2,203
20	5,871	4,461	3,859	3,515	3,289	3,128	2,913	2,774	2,676	2,501	2,408	2,349	2,249	2,223	2,156
21	5,827	4,420	3,819	3,475	3,250	3,090	2,874	2,735	2,637	2,462	2,368	2,308	2,208	2,182	2,114
22	5,786	4,383	3,783	3,440	3,215	3,055	2,839	2,700	2,602	2,426	2,332	2,272	2,171	2,145	2,076
23	5,750	4,349	3,750	3,408	3,183	3,023	2,808	2,668	2,570	2,394	2,299	2,239	2,137	2,111	2,041
24	5,717	4,319	3,721	3,379	3,155	2,995	2,779	2,640	2,541	2,365	2,269	2,209	2,107	2,080	2,010
25	5,686	4,291	3,694	3,353	3,129	2,969	2,753	2,613	2,515	2,338	2,242	2,182	2,079	2,052	1,981
26	5,659	4,265	3,670	3,329	3,105	2,945	2,729	2,590	2,491	2,314	2,217	2,157	2,053	2,026	1,954
27	5,633	4,242	3,647	3,307	3,083	2,923	2,707	2,568	2,469	2,291	2,195	2,133	2,029	2,002	1,930
28	5,610	4,221	3,626	3,286	3,063	2,903	2,687	2,547	2,448	2,270	2,174	2,112	2,007	1,980	1,907
29	5,588	4,201	3,607	3,267	3,044	2,884	2,669	2,529	2,430	2,251	2,154	2,092	1,987	1,959	1,886
30	5,568	4,182	3,589	3,250	3,026	2,867	2,651	2,511	2,412	2,233	2,136	2,074	1,968	1,940	1,866
31	5,549	4,165	3,573	3,234	3,010	2,851	2,635	2,495	2,396	2,217	2,119	2,057	1,950	1,922	1,848
32	5,531	4,149	3,557	3,218	2,995	2,836	2,620	2,480	2,381	2,201	2,103	2,041	1,934	1,905	1,831
33	5,515	4,134	3,543	3,204	2,981	2,822	2,606	2,466	2,366	2,187	2,088	2,026	1,918	1,890	1,815
34	5,499	4,120	3,529	3,191	2,968	2,808	2,593	2,453	2,353	2,173	2,075	2,012	1,904	1,875	1,799
35	5,485	4,106	3,517	3,179	2,956	2,796	2,581	2,440	2,341	2,160	2,062	1,999	1,890	1,861	1,785
40	5,424	4,051	3,463	3,126	2,904	2,744	2,529	2,388	2,288	2,107	2,007	1,943	1,832	1,803	1,724
50	5,340	3,975	3,390	3,054	2,833	2,674	2,458	2,317	2,216	2,033	1,931	1,866	1,752	1,721	1,639
80	5,218	3,864	3,284	2,950	2,730	2,571	2,355	2,213	2,111	1,925	1,820	1,752	1,632	1,599	1,508
100	5,179	3,828	3,250	2,917	2,696	2,537	2,321	2,179	2,077	1,890	1,784	1,715	1,592	1,558	1,463
120	5,152	3,805	3,227	2,894	2,674	2,515	2,299	2,157	2,055	1,866	1,760	1,690	1,565	1,530	1,433

Valeurs de F ayant 1% de chances d'être dépassées.

n_2 \ n_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	4052,185	4999,340	5403,534	5624,257	5763,955	5858,950	5980,954	6055,925	6106,682	6191,432	6234,273	6260,350	6302,260	6312,970	6339,513
2	98,502	99,000	99,164	99,251	99,302	99,331	99,375	99,397	99,419	99,444	99,455	99,466	99,477	99,484	99,491
3	34,116	30,816	29,457	28,710	28,237	27,911	27,489	27,228	27,052	26,751	26,597	26,504	26,354	26,316	26,221
4	21,198	18,000	16,694	15,977	15,522	15,207	14,799	14,546	14,374	14,079	13,929	13,838	13,690	13,652	13,558
5	16,258	13,274	12,060	11,392	10,967	10,672	10,289	10,051	9,888	9,609	9,466	9,379	9,238	9,202	9,112
6	13,745	10,925	9,780	9,148	8,746	8,466	8,102	7,874	7,718	7,451	7,313	7,229	7,091	7,057	6,969
7	12,246	9,547	8,451	7,847	7,460	7,191	6,840	6,620	6,469	6,209	6,074	5,992	5,858	5,824	5,737
8	11,259	8,649	7,591	7,006	6,632	6,371	6,029	5,814	5,667	5,412	5,279	5,198	5,065	5,032	4,946
9	10,562	8,022	6,992	6,422	6,057	5,802	5,467	5,257	5,111	4,860	4,729	4,649	4,517	4,483	4,398
10	10,044	7,559	6,552	5,994	5,636	5,386	5,057	4,849	4,706	4,457	4,327	4,247	4,115	4,082	3,996
11	9,646	7,206	6,217	5,668	5,316	5,069	4,744	4,539	4,397	4,150	4,021	3,941	3,810	3,776	3,690
12	9,330	6,927	5,953	5,412	5,064	4,821	4,499	4,296	4,155	3,910	3,780	3,701	3,569	3,535	3,449
13	9,074	6,701	5,739	5,205	4,862	4,620	4,302	4,100	3,960	3,716	3,587	3,507	3,375	3,341	3,255
14	8,862	6,515	5,564	5,035	4,695	4,456	4,140	3,939	3,800	3,556	3,427	3,348	3,215	3,181	3,094
15	8,683	6,359	5,417	4,893	4,556	4,318	4,004	3,805	3,666	3,423	3,294	3,214	3,081	3,047	2,959
16	8,531	6,226	5,292	4,773	4,437	4,202	3,890	3,691	3,553	3,310	3,181	3,101	2,967	2,933	2,845
17	8,400	6,112	5,185	4,669	4,336	4,101	3,791	3,593	3,455	3,212	3,083	3,003	2,869	2,835	2,746
18	8,285	6,013	5,092	4,579	4,248	4,015	3,705	3,508	3,371	3,128	2,999	2,919	2,784	2,749	2,660
19	8,185	5,926	5,010	4,500	4,171	3,939	3,631	3,434	3,297	3,054	2,925	2,844	2,709	2,674	2,584
20	8,096	5,849	4,938	4,431	4,103	3,871	3,564	3,368	3,231	2,989	2,859	2,778	2,643	2,608	2,517
21	8,017	5,780	4,874	4,369	4,042	3,812	3,506	3,310	3,173	2,931	2,801	2,720	2,584	2,548	2,457
22	7,945	5,719	4,817	4,313	3,988	3,758	3,453	3,258	3,121	2,879	2,749	2,667	2,531	2,495	2,403
23	7,881	5,664	4,765	4,264	3,939	3,710	3,406	3,211	3,074	2,832	2,702	2,620	2,483	2,447	2,354
24	7,823	5,614	4,718	4,218	3,895	3,667	3,363	3,168	3,032	2,789	2,659	2,577	2,440	2,403	2,310
25	7,770	5,568	4,675	4,177	3,855	3,627	3,324	3,129	2,993	2,751	2,620	2,538	2,400	2,364	2,270
26	7,721	5,526	4,637	4,140	3,818	3,591	3,288	3,094	2,958	2,715	2,585	2,503	2,364	2,327	2,233
27	7,677	5,488	4,601	4,106	3,785	3,558	3,256	3,062	2,926	2,683	2,552	2,470	2,330	2,294	2,198
28	7,636	5,453	4,568	4,074	3,754	3,528	3,226	3,032	2,896	2,653	2,522	2,440	2,300	2,263	2,167
29	7,598	5,420	4,538	4,045	3,725	3,499	3,198	3,005	2,868	2,626	2,495	2,412	2,271	2,234	2,138
30	7,562	5,390	4,510	4,018	3,699	3,473	3,173	2,979	2,843	2,600	2,469	2,386	2,245	2,208	2,111
31	7,530	5,362	4,484	3,993	3,675	3,449	3,149	2,955	2,820	2,577	2,445	2,362	2,221	2,183	2,086
32	7,499	5,336	4,459	3,969	3,652	3,427	3,127	2,934	2,798	2,555	2,423	2,340	2,198	2,160	2,062
33	7,471	5,312	4,437	3,948	3,630	3,406	3,106	2,913	2,777	2,534	2,402	2,319	2,176	2,139	2,040
34	7,444	5,289	4,416	3,927	3,611	3,386	3,087	2,894	2,758	2,515	2,383	2,299	2,156	2,118	2,019
35	7,419	5,268	4,396	3,908	3,592	3,368	3,069	2,876	2,740	2,497	2,364	2,281	2,137	2,099	2,000
40	7,314	5,178	4,313	3,828	3,514	3,291	2,993	2,801	2,665	2,421	2,288	2,203	2,058	2,019	1,917
50	7,171	5,057	4,199	3,720	3,408	3,186	2,890	2,698	2,563	2,318	2,183	2,098	1,949	1,909	1,803
80	6,963	4,881	4,036	3,563	3,255	3,036	2,742	2,551	2,415	2,169	2,032	1,944	1,788	1,746	1,630
100	6,895	4,824	3,984	3,513	3,206	2,988	2,694	2,503	2,368	2,120	1,983	1,893	1,735	1,692	1,572
120	6,851	4,787	3,949	3,480	3,174	2,956	2,663	2,472	2,336	2,089	1,950	1,860	1,700	1,656	1,533

Valeurs de F ayant 0,5% de chances d'être dépassées.

n_2 \ n_1	1	2	3	4	5	6	8	10	12	18	24	30	50	60	120
1	16212,463	19997,358	21614,134	22500,753	23055,822	23439,527	23923,814	24221,838	24426,728	24765,730	24937,093	25041,401	25212,765	25253,743	25358,051
2	198,503	199,012	199,158	199,245	199,303	199,332	199,376	199,390	199,419	199,449	199,449	199,478	199,478	199,478	199,492
3	55,552	49,800	47,468	46,195	45,391	44,838	44,125	43,685	43,387	42,881	42,623	42,466	42,211	42,150	41,990
4	31,332	26,284	24,260	23,154	22,456	21,975	21,352	20,967	20,705	20,258	20,030	19,892	19,667	19,611	19,469
5	22,785	18,314	16,530	15,556	14,939	14,513	13,961	13,618	13,385	12,985	12,780	12,656	12,454	12,402	12,274
6	18,635	14,544	12,917	12,028	11,464	11,073	10,566	10,250	10,034	9,664	9,474	9,358	9,170	9,122	9,001
7	16,235	12,404	10,883	10,050	9,522	9,155	8,678	8,380	8,176	7,826	7,645	7,534	7,354	7,309	7,193
8	14,688	11,043	9,597	8,805	8,302	7,952	7,496	7,211	7,015	6,678	6,503	6,396	6,222	6,177	6,065
9	13,614	10,107	8,717	7,956	7,471	7,134	6,693	6,417	6,227	5,899	5,729	5,625	5,454	5,410	5,300
10	12,827	9,427	8,081	7,343	6,872	6,545	6,116	5,847	5,661	5,340	5,173	5,071	4,902	4,859	4,750
11	12,226	8,912	7,600	6,881	6,422	6,102	5,682	5,418	5,236	4,921	4,756	4,654	4,488	4,445	4,337
12	11,754	8,510	7,226	6,521	6,071	5,757	5,345	5,085	4,906	4,595	4,431	4,331	4,165	4,123	4,015
13	11,374	8,186	6,926	6,233	5,791	5,482	5,076	4,820	4,643	4,334	4,173	4,073	3,908	3,866	3,758
14	11,060	7,922	6,680	5,998	5,562	5,257	4,857	4,603	4,428	4,122	3,961	3,862	3,697	3,655	3,547
15	10,798	7,701	6,476	5,803	5,372	5,071	4,674	4,424	4,250	3,946	3,786	3,687	3,523	3,480	3,372
16	10,576	7,514	6,303	5,638	5,212	4,913	4,521	4,272	4,099	3,797	3,638	3,539	3,375	3,332	3,224
17	10,384	7,354	6,156	5,497	5,075	4,779	4,389	4,142	3,971	3,670	3,511	3,412	3,248	3,206	3,097
18	10,218	7,215	6,028	5,375	4,956	4,663	4,276	4,030	3,860	3,560	3,402	3,303	3,139	3,096	2,987
19	10,073	7,093	5,916	5,268	4,853	4,561	4,177	3,933	3,763	3,464	3,306	3,208	3,043	3,000	2,891
20	9,944	6,987	5,818	5,174	4,762	4,472	4,090	3,847	3,678	3,380	3,222	3,123	2,959	2,916	2,806
21	9,829	6,891	5,730	5,091	4,681	4,393	4,013	3,771	3,602	3,305	3,147	3,049	2,884	2,841	2,730
22	9,727	6,806	5,652	5,017	4,609	4,322	3,944	3,703	3,535	3,239	3,081	2,982	2,817	2,774	2,663
23	9,635	6,730	5,582	4,950	4,544	4,259	3,882	3,642	3,474	3,179	3,021	2,922	2,756	2,713	2,602
24	9,551	6,661	5,519	4,890	4,486	4,202	3,826	3,587	3,420	3,125	2,967	2,868	2,702	2,658	2,546
25	9,475	6,598	5,462	4,835	4,433	4,150	3,776	3,537	3,370	3,075	2,918	2,819	2,652	2,609	2,496
26	9,406	6,541	5,409	4,785	4,384	4,103	3,730	3,492	3,325	3,031	2,873	2,774	2,607	2,563	2,450
27	9,342	6,489	5,361	4,740	4,340	4,059	3,687	3,450	3,284	2,990	2,832	2,733	2,565	2,522	2,408
28	9,284	6,440	5,317	4,698	4,300	4,020	3,649	3,412	3,246	2,952	2,794	2,695	2,527	2,483	2,369
29	9,230	6,396	5,276	4,659	4,262	3,983	3,613	3,376	3,211	2,917	2,759	2,660	2,492	2,448	2,333
30	9,180	6,355	5,239	4,623	4,228	3,949	3,580	3,344	3,179	2,885	2,727	2,628	2,459	2,415	2,300
31	9,133	6,316	5,204	4,590	4,195	3,918	3,549	3,314	3,149	2,855	2,697	2,598	2,429	2,385	2,269
32	9,090	6,281	5,172	4,559	4,166	3,889	3,521	3,286	3,121	2,828	2,670	2,570	2,401	2,356	2,240
33	9,049	6,248	5,141	4,531	4,138	3,861	3,495	3,260	3,095	2,802	2,644	2,544	2,374	2,330	2,213
34	9,012	6,217	5,113	4,504	4,112	3,836	3,470	3,235	3,071	2,778	2,620	2,520	2,350	2,305	2,188
35	8,976	6,188	5,086	4,479	4,088	3,812	3,447	3,212	3,048	2,755	2,597	2,497	2,327	2,282	2,164
40	8,828	6,066	4,976	4,374	3,986	3,713	3,350	3,117	2,953	2,661	2,502	2,401	2,230	2,184	2,064
50	8,626	5,902	4,826	4,232	3,849	3,579	3,219	2,988	2,825	2,533	2,373	2,272	2,097	2,050	1,925
80	8,335	5,665	4,611	4,028	3,652	3,387	3,032	2,803	2,641	2,349	2,188	2,084	1,903	1,854	1,720
100	8,241	5,589	4,542	3,963	3,589	3,325	2,972	2,744	2,583	2,290	2,128	2,024	1,840	1,790	1,652
120	8,179	5,539	4,497	3,921	3,548	3,285	2,933	2,705	2,544	2,251	2,089	1,984	1,798	1,747	1,606

Pour les grands échantillons, $\dfrac{s_1 - s_2}{s\sqrt{\dfrac{1}{2n_1} + \dfrac{1}{2n_2}}} \to \mathrm{N}(0,1)$ avec $s = \sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$.

Bibliographie et ouvrages / documents / liens recommandés

Abdennasser Chekroun. Polycopié : Statistiques descriptives et exercices. Université Abou Bekr Belkaid Tlemcen, 2017 - 2018

Antoine Ayache , Julien Hamonier. Cours en line : Cours de Statistique Descriptive. [http ://math.univ-lille1.fr/ ayache/].

Emile Amzallag, Norbert Piccioli. *INTRODUCTION A LA STATISTIQUE*. Hermann 293 rue Lecourbe, 75015 Paris, ISBN 2-7056-5889-0, 1978.

Chala Adel. Polycopié : Introduction aux Biostatistiques. UNIVERSITÉ MOHAMED KHIDER-BISKRA,2014-2015

Daniel Fredon, Myriam Maumy-Bertrand, Frédéric Bertrand. *Mathématiques Statistique et probabilités en 30 fiches*. Dunod, Paris, ISBN 978-2-10-054257-4, 2009.

Foued Ben said. Polycopié : Résumé de cours de Statistiques descriptives. UNIVERSITE DE LA MANOUBA-ECOLE SUPERIEURE DE COMMERCE DE TUNIS , 2012/2013.

Louis Houde. Polycopié : Module 7 Lois De Probabilité. Université du Québec à Trois-Riviéres. [https ://oraprdnt.uqtr.uquebec.ca/Gscdepot/paf1010/13/M7.pdf]

Jean-Christophe Breton. Statistiques "IUT Biotechnologie 2ème annèe". Université de La Rochelle, Octobre-Novembre 2008. [https ://perso.univ-rennes1.fr/jean-christophe .breton/Fichiers/stat_IUT.pdf]

Jean-Pierre. *Statistique et probabilités ; 6e édition*.Dunod, 11 rue Paul Bert, 92240 Malakoff, www.dunod.com ISBN 978-2-10-075259-1, 2016.

Maxime Herve. Aide - mémoire de statistique appliquée à la biologie. [https ://cran.r-project.org/doc/contrib/Herve-Aide-memoire-statistique.pdf]

Menaceur Amor. Polycopié : Biostatistiques 3ème Année Licence LMD. Université 8 Mai 1945-Guelma, 2017

Valentin Rousson. *Statistique appliquée aux sciences de la vie*. Springer-Verlag France, ISBN 978-2-8178-0393-7, 2013.

Yves Tillé. Polycopié : Résumé du Cours de Statistique Descriptive. Université de Neuchâtel, 2010.

Document :https ://www.apprendre-en-ligne.net/MADIMU2/STATI/STATI2.PDF