Chapitre 1

Statistique déscriptive

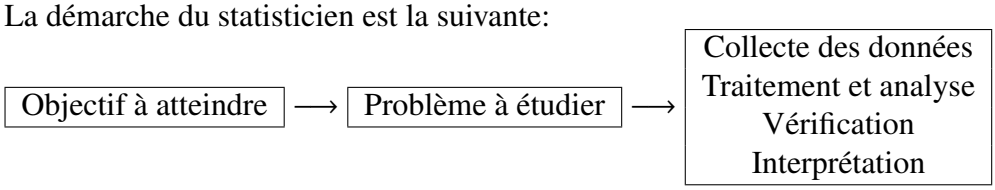
1.1 Qu'est ce que la statistique ?

La statistique est l'ensemble des méthodes scientifiques à partir desquelles on organise, on résume, on présente et on analyse des données relatives à un même phénoméne et qui permettere d'entirer des conclusions et de prendre des décisions.

Remarque 1 *Il ne faut pas confondre la statistique qui est la science qui vient d'être définie et une statistique qui est un ensemble de données chiffrées sur un sujet précis.*

1.2 Méthodologie

La démarche du statisticien est la suivante:



1.2.1 Collects des données

Le statisticien doit disposer des données sur le probléme posé (soit par une enquéte, une expérience, ou une étude historique,...). Ces données soient-elles sont qualitatives ou quantitatives.

1.2.2 Traitement et analyses des données

A partir des données, le statisticien propose un certain nombre de méthodes (en fonction du probleme posé), estimation, prévision, test,....

1.2.3 Vérification

On procède à partir des resultats obtenus à des conclusions avec les données de départ.

1.2.4 Interprétation

Le statisticien présente une signification des resultats obtenues et propose des solutions avec des évaluations de risques associés, ceci aide l'utilisateur à choisir entre les différentes décisions.

1.3 Vocabulaires statistique

1.3.1 Observations

Ce sont les données buts relatives à un phénoméne au cours d'une enquéte, une expérience, ..., il faut les regrouper, les corriger et les ordonner.

1.3.2 Population

C'est l'ensemble des éléments sur lesquels l'étude statistique sera faite, il est noté Ω.

Exemple 2 *Si l'on veut étudier la durée de vie des ampoules électriques fabriquées par une compagnie, la population considérée est l'ensemble de toutes les ampoules fabriquées par cette compagnie.*

1.3.3 Echantillon

Est tout sous-ensemble de la population.

Exemple 3 *Pour établir la durée de vie des ampoules électriques produites par une machine, on peut prélever au hasard un certain nombre d'ampoules (un échantillon) parmi toutes les celles produites par cette machine.*

1.3.4 Individu

Les éléments de Ω sont dits individus et notés ω ($\omega \in \Omega$).

Exemple 4 *Dans l'exemple précédant, chaque ampoule constitue un individu.*

Remarque 5 *Les éléments de Ω peuvent étre des personnes, des animaux ou des objets.*

1.3.5 Taille

Représente le nombre d'individus d'un échantillon ou d'une population, elle est symbolisée par $\ll n \gg$ dans le cas d'un échantillon et par $\ll N \gg$ dans le cas d'une population.

1.3.6 Caractère

C'est l'aspect particulier que l'on désire étudier.

Exemple 6 *Concernont un groupe de personnes, on peut s'intéresser à leur âge, leur sexe, leur taille, ...*

1.3.7 Modalités

Ce sont les différentes situations possibles du caractère.

Exemple 7 *Le sexe est un caractère qui présente deux modalités: féminin ou masculin.*

Exemple 8 *Quant au nombre d'enfants par famille, les modalités de ce caractère peuvent être* $0, 1, 2, ..., 10, ...$

Remarque 9 *Les modalités d'un caractère doivent être incompatibles et exhaustive, tout individu doit présenter une et une seul modalité.*

Remarque 10 *Il est d'usage de distinguer les deux types de caractère.*

Caractére qualitatif

Ses modalités ne s'expriment pas par un nombre.

Exemple 11 *La couleur du pelage, les groupes sanguins, les différents nucléotides de l'ADN, ...*

Caractére quantitatif

Ses modalités sont numériques.

Exemple 12 *Le nombre de cellules dans une culture, le taux de glycémie, le nombre de globules blancs ou rouges, ...*

1.4 Variable statistique (V.S)

Définition 13 *Une V.S: X est une application de* Ω *vers E.*

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & E \\ & \omega & \longmapsto & X(\omega). \end{array}$$

Remarque 14 *E est l'ensemble des modalités du V.S X.*

1.4.1 Variable statistique discrète

X est dite discrète si $E = \{x_1, ..., x_n, ...\}$ ensemble des valeurs isolées fini ou infini dénombrable, le plus souvent, Ces valeurs sont entieres.

1.4.2 Variable statistique continue

X est dite continue si $E = [a_0, a_1[\cup ... \cup [a_{n-1}, a_n[$ où $\forall i = 1 : n,\ a_i \in \mathbb{R}$.

1.5 Effectif et fréquences, effectif et fréquences cumulés

1.5.1 Effectif, effectif cumulés

Le nombre d'individus (notés n_i) ayant le caractère x_i s'appelle éffectif.
L'effectif cumulé du caractère x_i est $N_i = \sum_{j=1}^{i} n_j$.

1.5.2 Fréquences, fréquences cumulée

Le nombre $f_i = \frac{n_i}{N}$ ($N = \sum_i n_i$) s'appelle la fréquence du caractère x_i.

Remarque 15 $* \ 0 \leq f_i \leq 1$
$* \ \sum_i f_i = 1.$

La fréquence cumulée du caractère x_i est $F_i = \frac{N_i}{N} = \sum_{j=1}^{i} f_j$.

1.6 Courbe commulative des fréquences (fonction de répartition)

1.6.1 Cas discrète

Soient: $X : \Omega \longrightarrow \{x_1, ..., x_n\}$ un V.S et

$$\begin{array}{rccccl} F: & [x_i, x_{i+1}[& \longrightarrow & [0,1] & & \\ & x & \longmapsto & F(x) & = & \sum_{j=1}^{i} f_j \\ & & & & = & \left\{ \begin{array}{lll} 0 & \text{si} & x < x_1 \\ f_1 & \text{si} & x_1 \leq x < x_2 \\ \vdots & \vdots & \vdots \\ 1 & \text{si} & x > x_n \end{array} \right. \end{array}$$

F exprime la proportion des individus dont la valeurs du caractère est inférieur ou égale à x, c'est une fonction croissante étagée constante sur chaque intervalle $[x_i, x_{i+1}[$ et discontinue en tout point x_{i+1}. F est dite fonction de répartition de la V.S X.

1.6.2 Cas continue

Soient: $X : \Omega \longrightarrow [a_0, a_1[\cup ... \cup [a_{n-1}, a_n[$ un V.S et

$$\begin{array}{rccl} F: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & F(x) = \left\{ \begin{array}{lll} 0 & \text{si} & x < a_0 \\ \sum_{j=1}^{i} f_j + \frac{f_{i+1}}{a_{i+1}-a_i}(x - a_i) & \text{si} & x \in [a_i, a_{i+1}[\\ \vdots & \vdots & \vdots \\ 1 & \text{si} & x > a_n \end{array} \right. \end{array}$$

F exprime la proportion des individus dont la valeurs du caractère est inférieur ou égale à x, c'est une fonction croissante, continue, en segment de droite sur chaque intervalle $[a_i, a_{i+1}]$.

1.7 Présentation en tableau

En rassemblant les $1^{ère}$ données sur un phénomène quelconque, il nous est difficile de tirer profit de ces données sous cette forme là, c'est pour cela qu'on essaye de les exposer sous forme de tableau puis de graphes. Les étapes à suivre pour établir un tableau sont:

∗ Calculer l'étendue $e = X_{max} - X_{min}$.
∗ Calculer le nombre de classes $k = \sqrt{n}$.
∗ Calculer la longueur de classe $l = \frac{e}{k}$.

Exemple 16 *A fin d'étudier la structure de la population de gélinottes huppées abattues par les chasseurs, une étude du dimorphisme sexuel de cette espéce a été entreprise. Parmi les caractères mesurés figure la longueur de la rectrice centrale (plume de la queue). Les resultats observés exprimés en millimétres sur un échantillon de 50 mâles juvéniles sont notés dans la série ci-dessus:*

153	165	160	150	159	151	163	160	158	149
154	153	163	140	158	150	158	155	163	159
157	162	160	152	164	158	153	162	166	162
165	157	174	158	171	162	155	156	159	162
152	158	164	164	162	158	156	171	164	158

Les valeurs de la longueur de la rectrice peuvent être réparties de la façon suivante:
∗ *Définition de l'étendue* $e = X_{max} - X_{min} = 174 - 140 = 34$
∗ *Définition du nombre de classes* $k = \sqrt{n} \simeq 7$
∗ *Définition du longueur de classes* $l = \frac{e}{k} \simeq 5$
et par suite on a le tableau suivant:

Classes	*[140, 145[*	*[145, 150[*	*[150, 155[*	*[155, 160[*	*[160, 165[*	*[165, 170[*	*[170, 175[*
Effectif	*1*	*1*	*9*	*17*	*16*	*3*	*3*

1.8 Représentations graphiques

Les représentations graphiques ont l'avantage de renseigner immédiatement sur l'allure générale de la distribution, elles facilitent l'interprétation des données recueillies, elles repose sur la proportionnalité des longueurs, ou des aires, des graphiques, aux effectifs, ou aux fréquences des différents modalites du caractère.

1.8.1 Caractère qualitatif

Pour un caractère qualitatif, on utilise principalement deux types de représentation graphique: le tuyaux d'orgue et la représentation par secteurs. Lorsque le caractère étudié est la répartition géographique d'une population, la représentation graphique est un cartogramme.

Les tuyaux d'orgue

Les modalités de la variables sont placées sur une droite horizontale (attention: ne pas orienter cette droite car les modalités ne sont pas mesurable et il n'y a donc pas de relation d'ordre entre elles). Les effectifs (ou les frequences) sont placés sur un axe vertical, la hauteur du tuyau est proportionnelle à l'effectif.

Secteurs

L'effectif total est représenté par un disque. Chaque modalité est représentée par un secteur circulaire dont la surface (pratiquement: l'angle au centre) est proportionnelle àl'effectif correspondent.

Cartogrammes

Un cartogramme est une carte géographique dont les secteurs géographiques sont coloriés avec une couleur différente suivant l'effectif ou suivant la fréquence du caractère étudié.

Exemple 17 *On a dénombré chez un individu 1000 leucocytes et on s'intéresse à leur catégorie.*

Catégorie	*Neutrophiles*	*Eosinophiles*	*Basophiles*	*Lymphocytes*	*Monocytes*
Effectif	*600*	*20*	*10*	*110*	*260*

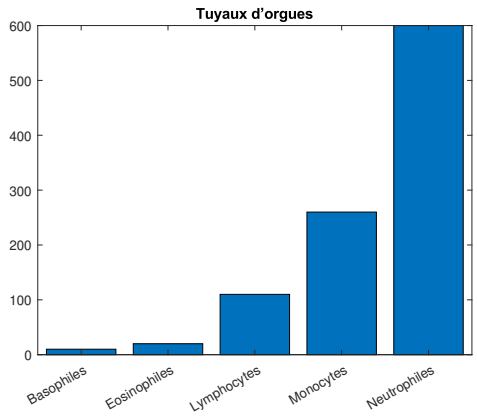


Figure 1.1: Tuyaux d'orgues.

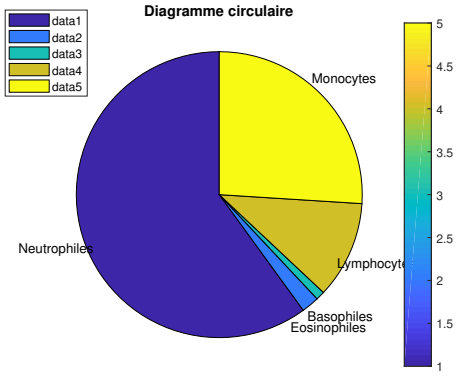


Figure 1.2: Diagramme circulaire.

1.8.2 Caractère quantitatif

Variable statistique discrète

Diagramme en bâton

On trace à partir de chaque valeur x_i de la V.S discète un segment de droite paralélle à l'axe des ordonnés telque sa longueur est égale à l'effectif n_i correspondant à la valeur x_i.

Exemple 18 *La série statistique suivante donne la taille en cm de 10 nouveaux-nés dans une maternité un jour donnée.*

x_i	*48.5*	*49.5*	*51*	*52.5*
n_i	*2*	*5*	*1*	*2*

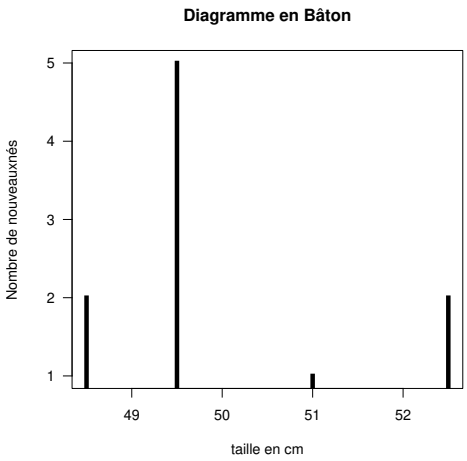


Figure 1.3: Diagramme en Bâton.

Variable statistique continue

Histogramme des effectif

Est sous forme de réctangles dont les bases sont sur l'axe des abscisses et elle représente les classes de la variable continue, tandisque les hauteurs représentent les effectifs de classes.

Polygone des effectif

Est sous forme de ligne brisée qui relie les points se situant sur le centre de chaque classe.

Exemple 19 *Dans un département, on a relevé la taille des exploitations agricoles et on a obtenu les résultats suivants:*

Taille (hectares)	*Nombre d'exploitations agricoles*
[0,10[	*30*
[10,20[	*80*
[20,30[	*60*
[30,40[	*20*
[40,50[	*10*

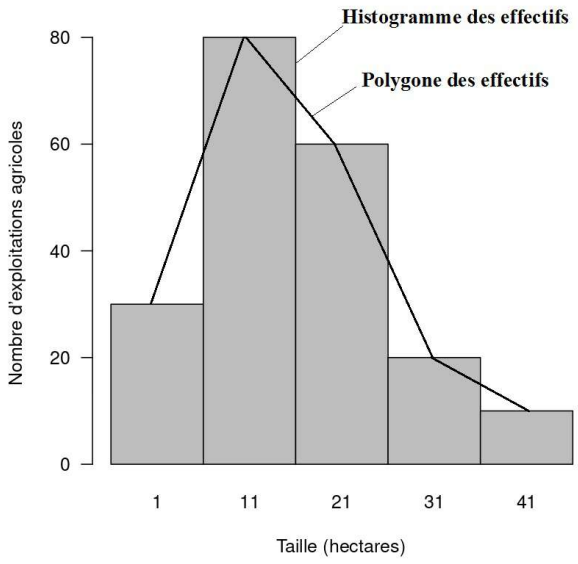


Figure 1.4: Histogramme et polygone des effectif.

1.9 Paramètres caracteristiques

1.9.1 Paramètres de position

Les paramètres de position(mode, médiane, moyenne) permettent de savoir autour de quelles valeurs se situent les valeurs d'une V.S.

Moyenne

La moyenne $\overline{X}$ ne se définit que pour une V.S quantitative. Pour une variable statistique discrète $\{(x_i, n_i), i = 1 : p\}$ à valeurs dans $\mathbb{R}$, la moyenne $\overline{X}$ est la moyenne arithmétique des modalités pondérées par les effectifs:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} x_i n_i = \sum_{i=1}^{p} f_i x_i.$$

Pour une variable statistique continue la moyenne est:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{p} x_i^* n_i = \sum_{i=1}^{p} f_i x_i^*,$$

où x_i^* présente le centre de la classe $[x_i, x_{i+1}[$

Exemple 20 *Dans le cas de l'étude du dimorphisme sexuel de la gélinotte huppée la longueur moyenne de la rectrice principale du mâle juvénile est:*
∗ *Dans le cas des données non groupées*

$$\overline{X} = \frac{153 + ... + 158}{50} = 158.9$$

∗ *Dans le cas des données groupées où les valeurs x_i correspondent aux centre des classes*

$$\overline{X} = \frac{7960}{50} = 159.2$$

Mode

On appelle mode M_o d'une série statistique la valeur de la variable qui a l'effectif le plus élevé.
Cas continu

$$M_o = e_{i-1} + a_i \frac{D_1}{D_1 + D_2}.$$

Avec:
∗ $[e_{i-1}, e_i[$: classe modale dont l'effectif est le plus élevé.

∗ a_i: longueur de la classe modale.
∗ e_{i-1}: borne inferieure de la classe modale.
∗ D_1: la différence entre les effectifs de la classe modale et la classe précédente.
∗ D_2: la différence entre les effectifs de la classe modale et la classe suivante.

Exemple 21 *Dans le cas de la distribution de la longueur de la rectrice centrale de la gélinotte huppée, la valeur du mode est:*
$M_o = 155 + 5\frac{8}{1+8} = 159.44$ *avec* $[155, 160[$ *est la classe modal,* $e_{i-1} = 155$*,* $D_1 = 17 - 9 = 8$*,* $D_2 = 17 - 16 = 1$ *et* $a_i = 160 - 155 = 5$

Médiane

Les valeurs du caractère étant rangée par ordre croissant, la médiane (M_e) est la valeur du variable statistique qui partage les individu en deux effectifs égaux. Dans le cas où les valeurs prises par le V.S ne sont pas regroupées en classe on a:
∗ Si N est impair, alors $M_e = x_k$ avec $k = \frac{N+1}{2}$
∗ Si N est pair, alors $M_e = \frac{x_k + x_{k+1}}{2}$ avec $k = \frac{N}{2}$

Cas continu

$$M_e = e_{i-1} + a_i \frac{N/2 - N^*}{n_i}.$$

Avec:
∗ $[e_{i-1}, e_i[$: classe médiane dont l'effectif cumulé depasse $N/2$.
∗ a_i: longueur de la classe médiane.
∗ e_{i-1}: borne inferieure de la classe médiane.
∗ N^*: effectif cumulé qui précède la classe médiane.
∗ n_i: effectif de la classe médiane.

Exemple 22 *Dans le cas de la distribution de la longueur de la rectrice centrale de la gélinotte hupée, la valeur de la médiane est:*
∗ *Cas des données non groupées*

$$M_e = \frac{x_{25} + x_{26}}{2} = \frac{158 + 159}{2} = 158.5$$

∗ *Cas des données groupées*

$$M_e = 155 + 5\frac{25 - 11}{17} = 159.11.$$

Avec $[155, 160[$ *est la classe médiane,* $e_{i-1} = 155$*,* $N^* = 11$*,* $n_i = 17$ *et* $a_i = 160 - 155 = 5$.

1.9.2 Paramètres de dispersion

Les paramètres de dispersion sont calculés pour les variables statistiques quantitatives.

Variance et ecart-type

Définition 23 *Soit $X = \{(x_i, n_i), i = 1 : p\}$ une V.S réelle*
$*$ *On appelle variance de X ($V(X)$ ou σ_X^2), la moyenne arithmétique des carrés des écarts de X à sa moyenne*

$$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \overline{X})^2 = \overline{X^2} - \overline{X}^2$$

$*$ *On appelle écart-type de X la racine carrée σ_X de σ_X^2*

Remarque 24 $\forall (a, b) \in \mathbb{R}^2$

$$V(a + bX) = b^2 V(X)$$

Exemple 25 *Dans le cas de l'étude du dimorphisme sexuel de la gélinotte huppée, la variance observée de la longueur de la rectrice central du mâle juvénile est*
$*$ *Cas des données non groupées $\sum_i x_i^2 = 1263647$, $\overline{X} = 158.86$ et $\sigma_X^2 = 36.44$*
$*$ *Cas des données groupées $\sum_i n_i x_i^2 = 1269012.5$, $\overline{X} = 159.2$ et $\sigma_X^2 = 35.61$*

Etendue

Soit X une V.S discrète. L'étendue e de X est la différence entre la plus grande valeur de X et la plus petite. $e = X_{max} - X_{min}$.
Ce paramètre est souvent utilisé dans les contrôles de fabrication pour lesquels on donne a priori des marges de construction. Son intérêt est limité par le fait qu'il dépend uniquement des valeurs extrêmes qui peuvent être des valeurs aberrantes.

Ecart absolu moyen

Définition 26 *Soit X une V.S, on appelle écart absolu moyen de X la moyenne arithmétique des valeurs absolues des écarts de X à sa moyenne:*

$$\xi = \frac{1}{N}\sum_i n_i |x_i - \overline{X}|$$

Quartiles et déciles

Variable statistique discrète

Pour une V.S discrète X la courbe des fréquences cumulées est une courbe en escalier. S'il existe une valeur de X pour laquelle la fréquence cumulée est 0.25, 0.5, 0.75, le quartile correspondant est cette valeur de X.

Variable statistique continue

On appelle quartiles les nombres réels Q_1, Q_2, Q_3 pour lesquels les fréquences cumulées de X sont respectivement 0.25, 0.5, 0.75. Ce sont les valeurs pour lesquelles l'ordonnée de la courbe cumulative des fréquence est respectivement égale à 0.25, 0.5, 0.75. Les quartiles partagent l'étendue en quatre intervalles qui ont le même effectif.

On peut de même définir les déciles d'une série statistique en partageant la série en dix parties de même effectif. Dans la pratique, seul le premier décile (noté D_1) et le neuvième décile (noté D_9) sont utilisés.

Moments

Soit X une V.S quantitative.

$*$ On appelle moment d'ordre r de X, la quantité:

$$m_r = \frac{1}{N}\sum_{i=1} n_i x_i^r$$

$*$ On appelle moment centré d'ordre r de X, la quantité:

$$\mu_r = \frac{1}{N}\sum_{i=1} n_i (x_i - \overline{X})^r$$

1.9.3 Paramètres de forme

Coefficient d'asymétrie

Définition 27 *Il existe plusieurs coefficients d'asymétrie, les principaux sont les suivants.*

$*$ *Le coefficient d'asymétrie de Pearson fait intervenir le mode M_o. Quand il existe, il est définie par:*

$$P = \frac{\overline{X} - M_o}{\sigma_X}$$

$*$ *Le coefficient d'asymétrie de Yule fait intervenir la médiane et les quantiles, il est définie par:*

$$Y = \frac{Q_1 + Q_3 - 2M_e}{2(Q_3 - Q_1)}$$

∗ *Le coefficient d'asymétrie de Fisher fait intervenir les moments centrés, il est définie par:*

$$F = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma_X^3}$$

∗ *Le coefficient d'asymétrie de Pearson basé sur les moments centrés, il est définie par:*

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}.$$

Lorsque le coefficient d'asymétrie est positif, la distribution est plus étalée à droite: on dit qu'il ya oblicité à gauche.
Lorsque le coefficient d'asymétrie est négatif, la distribution est plus étalée à gauche: on dit qu'il ya oblicité à droite.

Exemple 28 *Considérons la V.S X de distribution*

x_i	-1	4
n_i	4	1

$M_o = -1$, $\mu_3 = 12$, $\mu_2 = 4$, $P = 1/2$, $F = 3/2$ *et* $\beta_1 = 9/4$*: il ya oblicité à gauche.*

Exemple 29 *Considérons la V.S X de distribution*

x_i	-4	1
n_i	1	4

$M_o = 1$, $\mu_3 = -12$, $\mu_2 = 4$, $P = -1/2$, $F = -3/2$ *et* $\beta_1 = 9/4$*: il ya oblicité à droite.*

Coefficient d'aplatissement

Les principaux coefficient d'aplatissement sont les suivants:
∗ Le coefficient d'aplatissement de Pearson est

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

∗ Le coefficient d'aplatissement de Yule est

$$F_2 = \frac{\mu_4}{\mu_2^2} - 3$$

⋄ Si F_2 est égal à 0, le polygone statistique de la variable réduite a le même aplatissement qu'une courbe en cloche, on dit que la variable est mésokurtique.
⋄ Si $F_2 > 0$, le polygone statistique de la variable réduite est moins aplati qu'une courbe en cloche, on dit que la variable est leptokurtique.
⋄ Si $F_2 < 0$, le polygone statistique de la variable réduite est plus aplati qu'une courbe en cloche, on dit que la variable est platykurtique.

1.10 Exercises

Exercice 30 *On a mesuré la distance parcourue par un insecte en 30 s. Pour un lot de 10 insectes les résultats ont été les suivants (en mm) :*

78 170 173 190 90 174 166 293 149 117

− Calculer la moyenne et la variance des distances parcourues.

Exercice 31 *Le nombre d'oeufs pondus en un an par une poule a été relevé pendant 8 ans. Les résultats sont les suivants :*

Année (t)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Nombre d'oeufs pondus (n)	*160*	*140*	*122*	*112*	*96*	*88*	*72*	*62*

1− *Quelles sont les modalités* ?
2− *S'agit-il d'un caractère discret ou continu* ?
3− *Représenter graphiquement les données. Quelle relation entre n et t vous suggère cette représentation* ?
4− *Calculer la moyenne et la variance du nombre d'oeufs pondus.*
5− *Calculer les moments centré d'ordre 2, 3 et 4.*
6− *Calculer le coefficient d'asymétrie de Pearson basé sur les moments centrés.*
7− *Calculer le coefficient d'aplatissement de Yule.*

Exercice 32 *La cécidomyie du hêtre provoque sur les feuilles de cet arbre des galles dont la distribution a été observée : x est le nombre de galles par feuille, n est le nombre de feuilles portant x galles:*

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
n	*482*	*133*	*46*	*24*	*6*	*5*	*2*	*1*	*0*	*1*	*0*

1− *Quelles sont les modalités* ?
2− *S'agit-il d'un caractère discret ou continu* ?
3− *Représenter graphiquement les données.*
4− *Calculer la moyenne et la variance du nombre de galles par feuille.*
5− *Calculer les moments centré d'ordre 2, 3 et 4.*
6− *Calculer le coefficient d'asymétrie de Pearson basé sur les moments centrés.*
7− *Calculer le coefficient d'aplatissement de Yule.*

Exercice 33 *Une enquête concernant les distances entre domiciles des époux, au moment de leur mariage, a donné, dans le Finistère, les résultats suivants:*

Distance en km	*Nombre de couples*
[0; 2[	*138*
[2; 4[	*384*
[4; 6[	*210*
[6; 8[	*103*
[8; 10[	*63*
[10; 12[	*28*
[12; 14[	*20*
[14; 16[	*19*
[16; 18[	*12*
[18; 20[	*9*

1– *Quelles sont les modalités* ?
2– *S'agit-il d'un caractère discret ou continu* ?
3– *Représenter graphiquement les données.*
4– *Calculer la moyenne et la variance des distances.*
5– *Calculer les moments centré d'ordre 2, 3 et 4.*
6– *Calculer le coefficient d'asymétrie de Pearson basé sur les moments centrés.*
7– *Calculer le coefficient d'aplatissement de Yule.*

Exercice 34 *On considère la série quantitative suivante:*

7709	7710	7732	7746	7749	7750	7757	7762	7765	7767
7769	7771	7772	7772	7777	7780	7781	7783	7788	7790
7791	7792	7795	7796	7800	7801	7802	7804	7804	7805
7811	7812	7812	7817	7821	7823	7825	7826	7829	7832
7834	7839	7839	7841	7845	7850	7855	7860	7873	7889

1– *Quelle est l'étendue de la série* ?
2– *Regrouper les données en dix classes simples à manipuler.*
3– *Tracer l'histogramme de la série. En déduire le mode.*
4– *Représenter sur le même graphique le polygone des effectifs.*
5– *Tracer la courbe cumulative des effectifs. En déduire graphiquement la valeur de la médiane. Retrouver cette valeur par le calcul.*
6– *Calculer la moyenne et l'écart quadratique moyen avant et aprés le regroupement.*

Exercice 35 *Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir* 1625 *mg de bicarbonate de sodium. Afin de contrôler la*

fabrication de ces médicaments, on a prélevé un échantillon de 150 *comprimés, et on a mesuré la quantité de bicarbonate de sodium pour chacun d'eux. On a obtenu les résultats suivants :*

Classes	[1610; 1615]	]1615; 1620]	]1620; 1625]	]1625; 1630]	]1630; 1635]
Effectifs	*7*	*8*	*42*	*75*	*18*

1– *Caractériser la distribution.*
2– *Représenter graphiquement la distribution.*
3– *Déterminer les paramétres de position* (*Mode et médiane*).
4– *Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la quantité de bicarbonate de sodium.*

Exercice 36 *On s'intéresse au mélange de mangues issu de k = 4 différentes récoltes pour faire du jus. Chaque type de mangue à un taux de glucose différent et relativement imprécis, les résultats sont présentés dans la table suivants:*

Concentration (g/L)	[135; 165[	[165; 180[	[180; 195[	[195; 225[
Effectifs	*17*	*23*	*14*	*8*

1– *Représenter ces résultats graphiquement.*
2– *Déterminer les paramétres de position (moyenne, Mode et médiane).*
3– *Déterminer les paramètres de dispersion (variance et l'écart-type).*

Exercice 37 *Une coopérative laitière fabrique un fromage qui doit contenir, selon les étiquettes,* 45% *de matières grasses. Un institut de consommation dont le rôle est de vérifier que la qualité des produits est bien celle qui est affirmée par l'étiquette, fait prélever et analyser un échantillon de* 96 *fromages. Les résultats de l'analyse sont consignés dans le tableau suivant:*

Taux de matières grasses	[42, 5; 43, 5[	[43, 5; 44, 5[	[44, 5; 45, 5[	[45, 5; 46, 5[
Nombre de fromages	*12*	*24*	*38*	*22*

1– *Représenter ces résultats graphiquement.*
2– *Déterminer les paramétres de position (moyenne, Mode et médiane).*
3– *Déterminer les paramètres de dispersion (variance et l'écart-type).*

Chapitre 2

Rappels sur les probabilités

Le calcul des probabilités a pour objet l'analyse mathématique de la nature de hasard que l'on modelise par (Ω, Λ, P), où Λ: ensembles des évenements que l'on veut mesurer en termes de probabilité. Pour cela nous supposons que les éléments de Λ vérifient les propriétés suivantes:
$*$ $\Omega \in \Lambda$.
$*$ Si $A \in \Lambda$, alors $\overline{A} \in \Lambda$, $\overline{A}$ complément de A dans Ω.
$*$ Si $A_1, \ldots, A_n, \ldots$ est une suite d'éléments de Λ 2 à 2 disjoints ($A_i \cap A_j = \emptyset, i \neq j$) alors $\cup_{n\geq 1} A_n \in \Lambda$.

Définition 38 $*$ *Λ s'appelle une tribu de parties de Ω*
$*$ *(Ω, Λ) s'appelle espace probabilisable.*

Définition 39 *Soit (Ω, Λ) un espace probabilisable. On appelle probabilité toute application P de Λ vers $[0, 1]$ qui vérifie les propriétés suivantes:*
$*$ $P(\Omega) = 1$.
$*$ *Pour toute suite d'événements de Λ: $A_1, \ldots, A_n, \ldots$ 2 à 2 disjoints, on ait $P(\cup_{i\geq 1} A_i) = \sum_{i\geq 1} P(A_i)$.*

Définition 40 *(Ω, Λ, P) est un espace probabilisé.*

Exemple 41 *On lance une pièce de monnaie équilibrée une seule fois. l'ensemble des résultats possible est $\Omega = \{p, f\}$. Dans ce cas $\Lambda = \{\emptyset, \Omega, \{p\}, \{f\}\}$ et $P(\emptyset) = 0, P(\Omega) = 1, P(\{p\}) = 1/2, P(\{f\}) = 1/2$ et par suit (Ω, Λ, P) est un espace probabilisé.*

2.1 Variables aléatoires discrètes

Définition 42 *Soit (Ω, Λ, P) un espace probabilisé. On appelle variable aléatoire (v.a) discrète sur Ω toute fonction numérique X définie sur Ω et à valeur dans Ξ*

un ensemble discrèt $\Xi = \{x_1, \ldots, x_n, \ldots\}$ *telle que* $\forall x_i \in \Xi, \quad X^{-1}(\{x_i\}) \in \Lambda$ *(i.e* $\{\omega \in \Omega / X(\omega) = x_i\} \in \Lambda$*)*

2.1.1 Loi de probabilité d'une v.a discrète

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\Xi, \wp(\Xi), P_X)$ un v.a discrète.

Définition 43 *On appelle loi de probabilité de la v.a X, la probabilité* P_X *définie sur* Ξ *par:* $P_X(\{x_i\}) = P(X = x_i) = P(\omega \in \Omega, X(\omega) = x_i) = P(X^{-1}(\{x_i\}))$.

2.1.2 Moyenne et variance d'une v.a discrète

Moyenne

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\Xi, \wp(\Xi), P_X)$ un v.a discrète.
$E(X) = \sum_k x_k P_X(\{x_k\})$ E: appellée esperance mathématique de la v.a X.

Variance

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\Xi, \wp(\Xi), P_X)$ un v.a discrète.
$V(X) = \sigma_X^2 = \sum_k (x_k - E(X))^2 P_X(\{x_k\}) = E(X^2) - [E(X)]^2$

Remarque 44 $*$ $E(aX + bY) = aE(X) + bE(Y)$
$*$ $V(aX + b) = a^2 V(X)$

Exemple 45 *On lance une pièce de monnaie deux fois, soit X la variable aléatoire définie par: "nombre de piles obtenus". Dans ce cas*

$$\Omega = \{(p, p), (p, f), (f, p), (f, f)\}$$

Les valeurs possibles de X sont alors $X(\Omega) = \{0, 1, 2\}$*, alors la loi de probabilité de X est*

X	*0*	*1*	*2*
$P(X = k)$	*1/4*	*2/4*	*1/4*

En effet $X = 0$ *représente l'événement* (f, f) *alors*

$$P(X = 0) = \frac{card\{(f, f)\}}{card\Omega} = 1/4$$

De même on trouve que $P(X = 1) = 2/4$ *et* $P(X = 2) = 1/4$
$E(X) = \sum_{i=0}^{2} x_i P(X = x_i) = 1, \quad V(X) = \sum_{i=0}^{2} x_i^2 P(X = x_i) - E^2(X) = 1/2$

2.2 Quelques lois usuelles d'une v.a discrète

2.2.1 Loi de Bernoulli

Soit X une v.a à valeur dans $\{0, 1\}$ de loi P_X définie par $P_X(\{0\}) = 1 - p$ et $P_X(\{1\}) = p$, avec $p \in [0, 1]$. On dira que X suit une loi de Bernoulli de paramètre p. on notera

$$X \hookrightarrow B(p) \quad E(X) = p \quad V(X) = p(1 - p)$$

2.2.2 Loi binômiale

Soit X une v.a à valeur dans $\{0, 1, \ldots, n\}$ de loi P_X définie par $P_X(\{k\}) = C_n^k p^k q^{n-k}$ $(0 < p < 1,\ p = 1 - q)$. On dira que X suit une loi binômiale de taille n et de paramètre p. on notera

$$X \hookrightarrow b(n, p) \quad E(X) = np \quad V(X) = np(1 - p)$$

2.2.3 Loi géométrique

Soit X une v.a à valeur dans $\mathbb{N}^*$ de loi P_X définie par $P_X(\{k\}) = pq^{k-1}$ $(0 < p < 1,\ p = 1 - q)$. On dira que X suit une loi géométrique de paramètre p. on notera

$$X \hookrightarrow g(p) \quad E(X) = \frac{1}{p} \quad V(X) = \frac{1 - p}{p^2}$$

2.2.4 Loi de Poisson

Soit X une v.a à valeur dans $\mathbb{N}$ de loi P_X définie par $P_X(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda}$. On dira que X suit une loi de Poisson de paramètre λ. on notera

$$X \hookrightarrow \wp(\lambda) \quad E(X) = \lambda \quad V(X) = \lambda$$

2.2.5 Approximation d'une loi binomiale par une loi de Poisson

La loi de Poisson peut être obtenue comme la limite d'une loi binomiale lorsque le nombre n est asser grand et la probabilité p est asser petit: autrement si $X \hookrightarrow b(n, p)$ avec $n \geq 30$ et $p < 0.1$ alors pratiquement $X \hookrightarrow \wp(\lambda)$ où $\lambda = np$.

2.2.6 Fonction de répartition

Définition 46 *On appelle fonction de répartition de la v.a X la fonction* $F_X : \mathbb{R} \longrightarrow [0, 1]$ *définie par:*

$$F_X(x) = P(X \leq x)$$

2.3 Variables aléatoires réelles continues

Définition 47 *Soit* (Ω, Λ, P) *un espace probabilisé. On appelle v.a réelle sur* Ω *toute fonction numérique X définie sur* Ω *et à valeur dans* $(\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$ *telle que* $\forall A \in \mathfrak{B}_{\mathbb{R}},\ X^{-1}(A) \in \Lambda$.

2.3.1 Loi de probabilité d'une v.a réelles

Définition 48 *On appelle loi de probabilité de la v.a réelle X, la probabilité* P_X *définie sur* $\mathfrak{B}_{\mathbb{R}}$ *par:* $P_X(A) = P(X^{-1}(A)),\ \forall A \in \mathfrak{B}_{\mathbb{R}}$.

2.3.2 Fonction de répartition

Définition 49 *On appelle fonction de répartition de la v.a réelle X la fonction* $F_X : \mathbb{R} \longrightarrow [0, 1]$ *définie par:*

$$F_X(x) = P(X \leq x)$$

2.3.3 Densité de probabilité

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\mathbb{R}, \mathfrak{B}_{\mathbb{R}})$ v.a réelle, P_X sa loi et F_X sa fonction de répartition. On dit que la loi de probabilité de la v.a réelle X à densité si sa fonction de répartition F_X s'ecrit:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

où f_X est une fonction réelle positive, intégrable et $\int_{-\infty}^{\infty} f_X(x)dx = 1$
f_X est dite densité de probabilité (d.d.p) de la v.a réelle X.

Définition 50 *X est dite v.a réelle continue si sa loi de probabilité est à densité.*

2.3.4 Moyenne et variance d'une v.a réelle continue

Moyenne

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\mathbb{R}, \mathfrak{B}_{\mathbb{R}}, P_X)$ un v.a réelle continue.

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

où E: appellée esperance mathématique de la v.a réelle continue X.

Variance

Soit $X : (\Omega, \Lambda, P) \longrightarrow (\mathbb{R}, \mathfrak{B}_{\mathbb{R}}, P_X)$ un v.a réelle continue.

$$V(X) = \sigma_X^2 = E(X^2) - [E(X)]^2$$

2.4 Quelques lois usuelles d'une v.a réelle continue

2.4.1 Loi uniforme

On dira que X v.a réelle continue suit une loi Uniforme sur l'intervalle $[a, b]$ noté $X \hookrightarrow U(a, b)$ si sa d.d.p est donnée par:

$$f_X(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}(x)$$

$E(X) = \frac{a+b}{2}$, $V(X) = \frac{(b-a)^2}{12}$

2.4.2 Loi gamma

On dira que X v.a réelle continue suit une loi gamma de paramètre (α, λ), $(\alpha > 0, \lambda > 0)$ noté $X \hookrightarrow G(\alpha, \lambda)$ si sa d.d.p est donnée par:

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0$$

où

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx$$

$E(X) = \frac{\alpha}{\lambda}$, $V(X) = \frac{\alpha}{\lambda^2}$

2.4.3 Loi exponentielle

On dira que X v.a réelle continue suit une loi exponentielle de paramètre $\lambda > 0$ noté $X \hookrightarrow G(1, \lambda)$ si sa d.d.p est donnée par:

$$f_X(x) = \lambda e^{-\lambda x},\ x > 0$$

$E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$

2.4.4 Loi normale ou loi de Gauss

On dira que X v.a réelle continue suit une loi normale de paramètre $(\mu, \sigma^2), (\mu \in \mathbb{R}, \sigma^2 > 0)$ noté $X \hookrightarrow N(\mu, \sigma^2)$ si sa d.d.p est donnée par:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ x \in \mathbb{R}$$

$E(X) = \mu$, $V(X) = \sigma^2$

Représentation graphique

La fonction de densité suit une courbe en cloches appellée Gaussienne symétrique parapport à la moyenne μ, le maximum de la fonction de densité est obtenue pour $x = \mu$ avec $f_X(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$ et elle posséde deus points d'inflection pour $x = \mu \pm \sigma$.

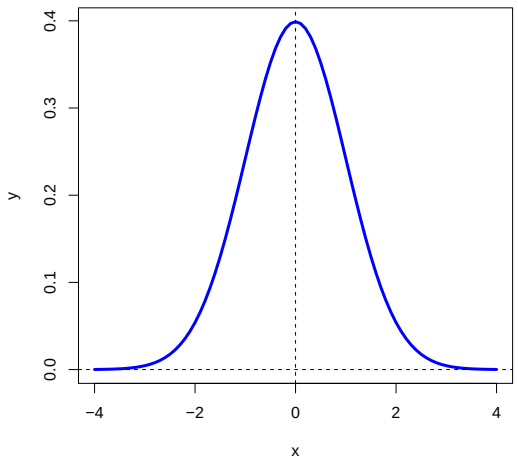


Figure 2.1: Densité de la loi normale $\mu = 0$ et $\sigma = 1$.

Utilisation des tables

Une v.a réelle continue X suit la loi normale centrée réduite si $E(X) = 0$ et $V(X) = 1$. A fin d'éviter le calcul numérique de la fonction de répartition pour la loi normale à chaque application on utilise la loi normale centrée réduite dont les valeurs sont tabulées (Tables 1 et 2).

Theorème 51 *Soit X une v.a réelle continue suivant une loi normale de paramètre (μ, σ^2). Si on applique le changement de variable $U = \frac{X-\mu}{\sigma}$ on peut écrire:*

$$F_X(x) = P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = P(U \leq u) = F_U(u)$$

et par suite la variable U suit une loi normale centrée réduite $N(0, 1)$.

2.4.5 Approximation d'une loi binomiale par une loi normale

La loi normale peut être obtenue comme la limite d'une loi binomiale lorsque le nombre n est assez grand et la probabilité p pas trop voisin de 0 et de 1: autrement si $X \hookrightarrow b(n,p)$ avec $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ alors pratiquement $X \hookrightarrow N(np, np(1-p))$.

2.4.6 La loi du Khi-Deux

Soient $Z_1, \ldots, Z_n$ une suite de v.a indépendents suivant chacune la loi normale $N(0,1)$, On dit que la variable $X = Z_1^2 + \ldots + Z_n^2$ suit la loi du Khi-Deux à n degrés de liberté noté $X \hookrightarrow \chi_n^2$ si sa d.d.p est donnée par:

$$f_X(x) = \frac{(x/2)^{n/2-1}}{2\Gamma(n/2)} e^{-x/2}, \ x > 0$$

$E(X) = n$, $V(X) = 2n$

Remarque 52 $*$ *La loi du χ_n^2 à n degrés de liberté et la loi Gamma de paramètre $\alpha = \frac{n}{2}$ et $\lambda = \frac{1}{2}$ sont identiques.*
$*$ *En pratique désque $n > 30$ on poura approximer la variable $\sqrt{2\chi_n^2} - \sqrt{2n-1}$ par la loi normal $N(0,1)$.*

Représentation graphique

On remarque que les courbes ne sont pas symetriques par rapport à leur maximum.

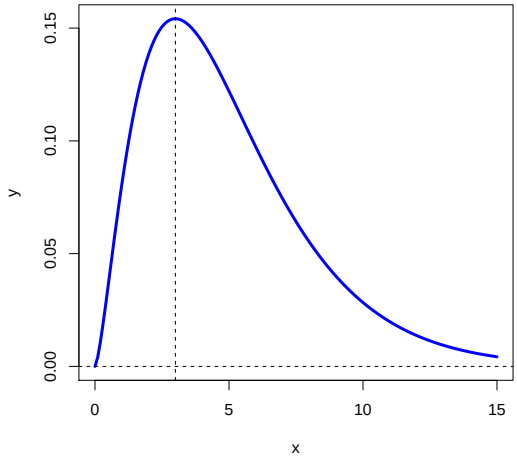


Figure 2.2: Densité de la loi du Khi-Deux $n = 5$.

Utilisation des tables

La table 3 donne l'aire sous la courbe à droite de la valeur χ^2_α lu dans la premiére ligne en fonction de degré de liberté lu dans la colone de gauche. Pour tout $\alpha \in [0,1]$ la quantité χ^2_α est définie telle que $P(X \geq \chi^2_\alpha) = \alpha$.

2.4.7 La loi de Student

Soient Z et X deux v.a indépendents où $Z \hookrightarrow N(0,1)$ et $X \hookrightarrow \chi^2_n$, On dit que la variable $T_n = \frac{Z}{\sqrt{X/n}}$ suit la loi de Student à n degrés de liberté si sa d.d.p est donnée par:

$$f_{T_n}(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma(n+1/2)}{\Gamma(n/2)} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}, \ x \in \mathbb{R}$$

$E(T_n) = 0$, $V(T_n) = \frac{n}{n-2}$, $n > 2$

Représentation graphique

Le tracé depend de n mais les courbes ont la même forme (symetriques par rapport à $x = 0$).

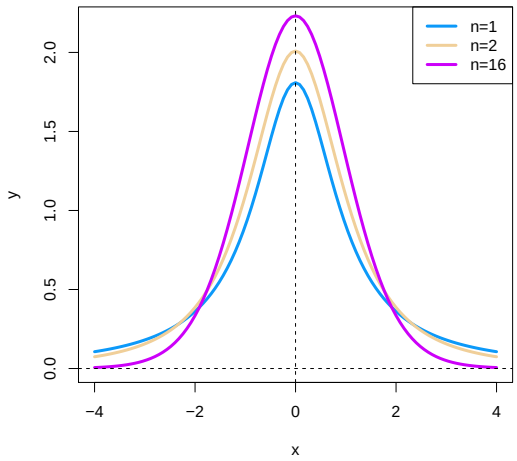


Figure 2.3: Densité de la loi de Student.

Utilisation des tables

La table 4 donne l'aire sous la courbe située à gauche de la valeur $(-t_\alpha)$ et cummulée avec celle située à droite de la valeur $(+t_\alpha)$ et le nombre de degrés de liberté n est sur la colone à gauche. Pour tout $\alpha \in [0,1]$ et n, la quantité t_α est définie telle que $P(|T_n| > t_\alpha) = \alpha$.

2.4.8 La loi de Fisher-Snédécor

Soient χ^2_n et χ^2_m deux v.a indépendents suivant chacune la loi du Khi-Deux de degrés n, m de liberté respectivement, On dit que la variable $F_{n,m} = \frac{\chi^2_n/n}{\chi^2_m/m}$ suit la loi de Fisher à n, m degrés de liberté si sa d.d.p est donnée par:

$$f_{F_{n,m}}(x) = \frac{\Gamma(\frac{m+n}{2})n^{n/2}m^{m/2}}{\Gamma(n/2)\Gamma(m/2)} \frac{x(n-2)}{2}(m+nx)^{-\frac{m+n}{2}}, \ x > 0$$

$E(F_{n,m}) = \frac{m}{m-2}, \ m > 2, V(F_{n,m}) = \frac{m^2(2m+2n-4)}{n(m-2)^2(m-4)}, \ m > 4$

Remarque 53 $F_{n,m} = \frac{1}{F_{m,n}}$

Représentation graphique

On remarque que les courbes ne sont pas symetriques par rapport à leur maximum.

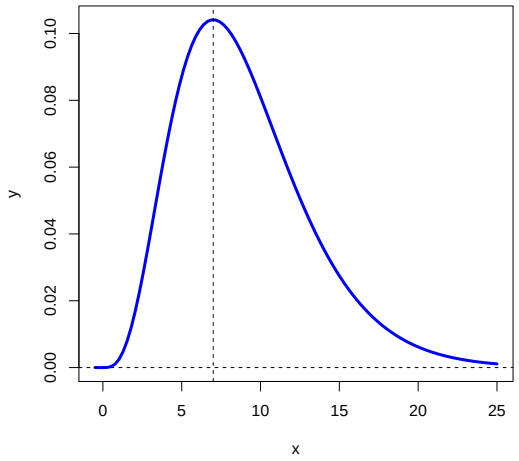


Figure 2.4: Densité de la loi de Fisher-Snédécor $n = 5, m = 9$.

Utilisation des tables

Les tables 5 et 6 dependent de deux degrés de liberté sont donc à triple entrés, pour notre cas on se contente des tables qui correspondent à $\alpha = 0.005$ et $\alpha = 0.0025$. Pour tout $\alpha \in [0, 1]$, n et m la quantité f_α est définie telle que $P(F_{n,m} \geq f_\alpha) = \alpha$.

Chapitre 3

Théorie d'estimation

Dans de nombreuses expériences aléatoires on peut déterminer parfaitement la loi de probabilité qui régit le phénoméne, cependant il existe des expériences aléatoires où la loi de probabilité est totalement inconnue ou partiellement inconnue et il s'agit de donner des précisions de cette loi aux vue de l'expérience.
∗ Dans le cas totalement inconnu les méthodes sont appellées: méthodes non-paramétriques.
∗ Dans le cas partiellement inconnu les méthodes sont appellées: méthodes paramétriques.

Définition 54 ► *Un échantillon d'une loi est une suite de v.a indépendantes identiquement distribuées (i.i.d).*
► *Un modèle statistique est la donnée de triplet* $(\mathfrak{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Theta})$ *où:* $\mathfrak{X}$ *est l'espace de réalisations,* $\mathfrak{A}$ *tribu sur* $\mathfrak{X}$*,* $P_\theta = P_X$ *loi de* X *et* Θ *l'ensemble des paramétres* θ*.*

Exemple 55 *Soit un échantillonnage de* $N(m, \sigma^2)$*, c'est à dire une suites* $X_1, \ldots, X_n$ *de v.a i.i.d avec* $\forall i$*,* $X_i \hookrightarrow N(m, \sigma^2)$*,* $\mathfrak{X} = \mathbb{R}^n$*,* $\mathfrak{A} = \mathfrak{B}_{\mathbb{R}^n}$*,* $P_\theta = N(m, \sigma^2)$ *et* $\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^{+*}$*.*

Définition 56 *Une statistique est une application* T *mesurable (v.a) de* $(\mathfrak{X}, \mathfrak{A})$ *dans un espace mesurable* $(\mathfrak{F}, \mathfrak{H})$*.*

$$\begin{array}{rccc} T: & (\mathfrak{X}, \mathfrak{A}) & \longrightarrow & (\mathfrak{F}, \mathfrak{H}) \\ & (X_1, \ldots, X_n) & \longmapsto & T(X_1, \ldots, X_n). \end{array}$$

Définition 57 *Le biais d'un estimateur est la quantité:*

$$b_\theta(T) = E_\theta(T) - \theta$$

où E_θ *espérance par rapport à* P_θ*.*
∗ *Si* $b_\theta(T) = 0$*,* T *est dit estimateur sans biais.*
∗ *Si* $b_\theta(T) \neq 0$*,* T *est dit estimateur biaisé.*

3.1 Estimation paramétrique

3.1.1 Méthode des moments

C'est une méthode naturelle dans la mesure où elle est intuitive. Supposons que l'on doive estimer le paramétre θ, la méthode des moments consiste à choisir comme estimateur $\widehat{\theta_n}$ la solution de l'équation obtenue en égalant le moment théorique d'ordre k et le moment empirique d'ordre k.

$$E(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Exemple 58 *Soit $X \hookrightarrow G(1,\theta)$, donc $E(X) = \frac{1}{\theta}$.*
$*$ Pour $k = 1$ la méthode des moments nous donne $E(X) = \overline{X}_n$, alors un estimateur de θ est:

$$\widehat{\theta_n} = \frac{1}{\overline{X}_n}$$

$*$ Pour $k = 2$ la méthode des moments nous donne $E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, or $E(X^2) = \theta\int_0^{+\infty} x^2 e^{-\theta x}dx = \frac{2}{\theta^2}$, alors un estimateur de θ est:

$$\widehat{\theta_n} = \sqrt{\frac{2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}$$

3.1.2 Méthode du maximum de vraisemblance

Définition 59 *Soient $X = (X_1, \ldots, X_n)$ une suites de v.a i.i.d, on appelle fonction de vraisemblance pour X la fonction définie par:*

$$L(X_1, \ldots, X_n, \theta) = \begin{cases} \prod_{i=1}^{n} P(X_i, \theta) & si \quad les\ X_i\ sont\ discrètes \\ \\ \prod_{i=1}^{n} f(X_i, \theta) & si \quad les\ X_i\ sont\ continues \end{cases}$$

Définition 60 *l'estimateur de θ par la méthode du maximum de vraisemblance est la valeur $\widehat{\theta_n}$ qui rend maximale la fonction de vraisemblance L.*

Les conditions requises pour assurer cette maximisation sont $\frac{dL}{d\theta} = 0$ et $\frac{d^2L}{d\theta^2} < 0$.
Il est par fois plus commode de maximiser le logarithme népérien de L par rapport à θ puisque cette fonction comporte souvent des puissances ou des formes exponentielles, les conditions deviennent alors $\frac{dlnL}{d\theta} = 0$ et $\frac{d^2lnL}{d\theta^2} < 0$.
lnL est une fonction croissante et elle aura sa valeur maximum pour la même valeur de θ qu'aurait la fonction L.

Remarque 61 *L'estimateur du maximum de vraisemblance peut ne pas exister.*

Exemple 62 *Si les X_i sont de loi $N(m, \sigma^2)$, la fonction de vraisemblance est:*

$$L(X_1, \ldots, X_n, m, \sigma^2) = \prod_{i=1}^{n} f(X_i, m, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-m)^2}{2\sigma^2}} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-m)^2}$$

D'où

$$lnL(X_1, \ldots, X_n, m, \sigma^2) = -\frac{n}{2} ln\sigma^2 - \frac{n}{2} ln2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - m)^2$$

On doit annuler les dérivées partielles de ce logarithme par rapport à m et σ^2. On a:

$$\frac{\partial}{\partial m} lnL(X_1, \ldots, X_n, m, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(X_i - m) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} X_i - nm\right),$$

qui s'annule pour

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

$$\frac{\partial}{\partial \sigma^2} lnL(X_1, \ldots, X_n, m, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - m)^2,$$

qui s'annule pour

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = S_e^2.$$

Propriétés

$*$ La moyenne empirique $\overline{X}_n$ est un estimateur sans biais pour m, en effet

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} nm = m$$

$*$ La variance empirique S_e^2 est un estimateur biaisé pour σ^2, en effet

$$\begin{aligned} E(S_e^2) &= E(X^2) - E(\overline{X}_n^2) \\ &= V(X) + E(X)^2 - V(\overline{X}_n) - E(\overline{X}_n)^2 \\ &= \frac{n-1}{n} V(X) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

En revanche, on voit que $E(\frac{n}{n-1}S_e^2) = \frac{n}{n-1}E(S_e^2) = \sigma^2$. On pose donc

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

Par conséquent S^2 (appelée variance estimée) est un estimateur sans biais pour σ^2.

3.2 Estimation par intervalles de confiance

Soit $(\mathfrak{X}, \mathfrak{A}, (P_\theta)_{\theta\in\Theta})$ un modèle statistique.
Dans les applications, on a besoin par fois d'encadrer la vraie valeur du pramêtre θ puisque les estimations ponctuelles n'apportent pas d'information sur la précision des résultats, c'est-à-dire qu'elles ne tiennent pas compte des erreurs dues aux fluctuations d'échantillonnage. Il s'agit donc de déterminer un intervalle contenant, avec une grande probabilité, la vraie valeur du paramètre θ.

Définition 63 *Soit $\alpha \in]0,1[$, on appelle intervalle de confiance (I.C) $\widehat{I} = (R(X), T(X))$ (où $X = (X_1, \ldots, X_n)$) de niveau $1-\alpha$ pour le paramètre θ un intervalle vérifiant $P(\theta \in \widehat{I}) = 1-\alpha$.*
α appellé le seuil, $R(X)$ et $T(X)$ appellés limites de confiance de θ au niveau $1-\alpha$.

Théorème central limites (TCL)
Soit $(X_n)_n$ une suite de v.a i.i.d telle que $E(X_n) = \mu$ et $V(X_n) = \sigma^2$, on pose $S_n = \sum_{i=1}^{n} X_i$, alors:

$$\frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \hookrightarrow N(0,1)$$

Remarque 64 *D'aprés le TCL on a:*

$$\overline{X}_n \hookrightarrow N(\mu, \sigma^2/n), \qquad \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \hookrightarrow N(0,1)$$

3.2.1 Intervalle de confiance pour la moyenne dans le cas d'un échantillon Gaussien

Soit $(X_1, \ldots, X_n)$ un n-échantillon de v.a de loi $N(\mu; \sigma^2)$. L'idée est de trouver une variable aléatoire U de loi connue qui serait une fonction des observations aléatoires $X_1, \ldots, X_n$ et de μ, le paramètre à estimer.

Cas σ^2 connue, n quelconque

Pour estimer μ, on utilise la moyenne empirique $\overline{X}_n$ qui a pour loi $N(\mu, \sigma^2/n)$. Il en résulte que:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \hookrightarrow N(0,1)$$

et que

$$P\left(-u_{1-\alpha/2} \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha/2}\right) = 1-\alpha$$

Ceci équivaut à

$$P\left(\overline{X}_n - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Donc un I.C pour l'espérance μ avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\overline{X}_n - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X}_n + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Cas σ^2 inconnue

► $n < 30$
Dans ce cas on approxime σ par S (variance empirique). Il faut donc considérer la quantité

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}}$$

qui suit la loi de Student à $n-1$ degrés de liberté (T_{n-1}), on en déduit donc que:

$$P\left(-t_{1-\alpha/2} \leq \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \leq t_{1-\alpha/2}\right) = 1-\alpha$$

ce qui équivaut à

$$P\left(\overline{X}_n - t_{1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

et par suite un I.C pour l'espérance μ avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\overline{X}_n - t_{1-\alpha/2}\frac{S}{\sqrt{n}}; \overline{X}_n + t_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi T_{n-1}.
► $n \geq 30$
Dans ce cas en utilise le fait que:

$$T_{n-1} \approx N(0,1)$$

et par conséquence un I.C pour l'espérance μ avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\overline{X}_n - u_{1-\alpha/2}\frac{S}{\sqrt{n}}; \overline{X}_n + u_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Exemple 65 *Pour étudier la pourriture des pommes de terre, un chercheur injecte à 13 pommes de terre des bactéries qui causent cette pourriture. Il mesure ensuite la surface pourrie (en mm^2) sur ces 13 pommes de terre. Il obtient une moyenne empirique de 7.84 mm^2 pour une variance empirique de 14.13. On modélise la surface pourrie d'une pomme de terre par une loi normale $N(\mu, \sigma^2)$.*
1. *Donner un intervalle de confiance de niveau 0.95 pour μ*
2. *Donner un intervalle de confiance de niveau 0.99 pour μ. Comparer sa largeur avec celle de l'intervalle précédent.*
3. *Si avec le même échantillon on donnait un intervalle de confiance de largeur 3.5 mm^2, quel serait son niveau de confiance* ?
4. *On souhaite maintenant estimer μ avec une précision de $\pm 1.8 mm^2$, avec un niveau de confiance de 0.95. Quelle taille minimum doit avoir l'échantillon* ?.
5. *Que seraient les intervalles de confiance de μ, si on supposait que la variance σ^2 était connue et égale à 12* ?

Réponce
1. Pour $\alpha = 0.05 \Rightarrow u_{1-\alpha/2} = 1.96$.
L'intervalle de confiance de niveau 0.95 est:

$$\left[7.84 - 1.96\frac{3.76}{\sqrt{13}}; 7.84 + 1.96\frac{3.76}{\sqrt{13}}\right] = [5.79; 9.89]$$

2. Pour $\alpha = 0.01 \Rightarrow u_{1-\alpha/2} = 2.5758$.
L'intervalle de confiance de niveau 0.99 est:

$$\left[7.84 - 2.5758\frac{3.76}{\sqrt{13}}; 7.84 + 2.5758\frac{3.76}{\sqrt{13}}\right] = [5.15; 10.53]$$

L'intervalle est plus large que le précédent. Plus la probabilité que la moyenne appartienne à l'intervalle est grande (0.99 au lieu de 0.95), plus cet intervalle doit être large. Si on veut avoir plus confiance dans l'intervalle, il faut accepter qu'il soit

moins précis.
3. La largeur de l'intervalle de confiance de niveau $1-\alpha$ est:

$$2u_{1-\alpha/2}\frac{3.76}{\sqrt{13}}$$

Si cette largeur est égale à 3.5, on obtient:

$$u_{1-\alpha/2} = \frac{3.5\sqrt{13}}{2\times 3.76} = 1.68$$

Cette valeur est le quantile d'ordre $0.9535 = 1-\alpha/2$ de la loi $N(0,1)$. Donc $\alpha = 0.093$ et $1-\alpha = 0.907$.
4. Pour un échantillon de taille n, La précision de l'intervalle de confiance de niveau 0.95 est:

$$\pm 1.96\frac{3.76}{\sqrt{n}}$$

Si elle est égale à 1.8, on obtient:

$$n = \left(\frac{1.96\times 3.76}{1.8}\right)^2 = 16.76$$

L'échantillon doit donc être de taille 17 au moins.
5.
$*$ Pour $\alpha = 0.05$, $n-1 = 12 \Rightarrow t_{1-\alpha/2} = 2.179$.
L'intervalle de confiance de niveau 0.95 est:

$$\left[7.84 - 2.179\frac{3.76}{\sqrt{13}}; 7.84 + 2.179\frac{3.76}{\sqrt{13}}\right] = [5.56; 10.11]$$

$*$ Pour $\alpha = 0.01$, $n-1 = 12 \Rightarrow t_{1-\alpha/2} = 3.055$.
L'intervalle de confiance de niveau 0.95 est:

$$\left[7.84 - 3.055\frac{3.76}{\sqrt{13}}; 7.84 + 3.055\frac{3.76}{\sqrt{13}}\right] = [4.64; 11.03]$$

3.2.2 Intervalle de confiance pour la moyenne dans le cas d'un échantillon quelconque

Si les v.a.r. $X_1, \ldots, X_n$ ne sont pas Gaussiennes mais que n est assez grand (en pratique supérieur à 30), alors le TCL nous garantit que la moyenne empirique $\overline{X}_n$ suit approximativement la loi $N(0,1)$.

Cas σ^2 connue, $n \geq 30$

Un I.C pour l'espérance μ avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\overline{X}_n - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X}_n + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Cas σ^2 inconnue, $n \geq 30$

Un I.C pour l'espérance μ avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\overline{X}_n - u_{1-\alpha/2}\frac{S}{\sqrt{n}}; \overline{X}_n + u_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Exemple 66 *On a effectué 90 mesures de concentration d'une solution de fluoresceïne. On a observé une moyenne empirique de 4.38 mg/l et un écart-type empirique de 0.08 mg/l. Donner un intervalle de confiance pour la concentration réelle de la solution, au niveaux de confiance 0.95 et 0.99.*

Réponce
$*$ Pour $\alpha = 0.05 \Rightarrow u_{1-\alpha/2} = 1.96$.
L'intervalle de confiance de niveau 0.95 est:

$$\left[4.38 - 1.96\frac{0.08}{\sqrt{90}}; 4.38 + 1.96\frac{0.08}{\sqrt{90}}\right] = [4.363; 4.397]$$

$*$ Pour $\alpha = 0.01 \Rightarrow u_{1-\alpha/2} = 2.5758$.
L'intervalle de confiance de niveau 0.99 est:

$$\left[4.38 - 2.5758\frac{0.08}{\sqrt{90}}; 4.38 + 2.5758\frac{0.08}{\sqrt{90}}\right] = [4.358; 4.402]$$

3.2.3 Intervalle de confiance pour la variance dans le cas d'un échantillon Gaussien

► $n \leq 31$
On estime la variance σ^2, par la variance empirique S^2. On sait que la v.a.r $\frac{n-1}{\sigma^2}S^2$ a pour loi χ^2_{n-1}. On obtient alors

$$P\left(a \leq \frac{n-1}{\sigma^2}S^2 \leq b\right) = 1-\alpha$$

ce qui équivaut à:

$$P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) = 1-\alpha$$

et par suite un I.C pour la variance σ^2 avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\frac{(n-1)S^2}{b}; \frac{(n-1)S^2}{a}\right]$$

avec a et b tel que $P(\chi^2_{n-1} \geq a) = 1-\frac{\alpha}{2}$ et $P(\chi^2_{n-1} \geq b) = \frac{\alpha}{2}$.
► $n > 31$
Si $n > 31 \Longrightarrow n-1 > 30$, mais les tables du χ^2_{n-1} s'arrétent habituellement au degré de liberté 30. Danc on ne peut pas utiliser les resultats présédents.

Theorème 67 *Soit $Z \hookrightarrow \chi^2_\nu$ avec $\nu > 30$ une v.a.r. On pose $U = \sqrt{2Z} - \sqrt{2\nu-1}$, alors $U \hookrightarrow N(0,1)$.*

Pour estimer σ^2, on utilise la v.a.r U avec $Z = \frac{n-1}{\sigma^2}S^2$, il en résulte que

$$P\left(-u_{1-\alpha/2} \leq \sqrt{2\frac{n-1}{\sigma^2}S^2} - \sqrt{2n-3} \leq u_{1-\alpha/2}\right) = 1-\alpha$$

Ceci équivaut à:

$$P\left(\frac{2(n-1)S^2}{\left(\sqrt{2n-3}+u_{1-\alpha/2}\right)^2} \leq \sigma^2 \leq \frac{2(n-1)S^2}{\left(\sqrt{2n-3}-u_{1-\alpha/2}\right)^2}\right) = 1-\alpha.$$

Donc un I.C pour l'espérance σ^2 avec coefficient de sécurité $1-\alpha$ est donné par:

$$\left[\frac{2(n-1)S^2}{\left(\sqrt{2n-3}+u_{1-\alpha/2}\right)^2}; \frac{2(n-1)S^2}{\left(\sqrt{2n-3}-u_{1-\alpha/2}\right)^2}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $N(0,1)$.

Exemple 68 *Reprenent les données de l'exemple 65*
1. *Donner un intervalle de confiance de niveau 0.95 pour la variance.*
2. *On supose que la taille de l'échantillon $n = 32$, donner un intervalle de confiance de niveau 0.95 pour la variance.*

Réponce
1. Le quantile d'ordre 0.025 pour la loi de khi-deux χ^2_{12} est $a = 4.404$. Le quantile

d'ordre 0.975 est $b = 23.337$. L'intervalle de confiance de niveau 0.95 pour la variance est:

$$\left[\frac{12 \times 14.13}{23.337}; \frac{12 \times 14.13}{4.404}\right] = [7.26; 38.5]$$

2. $\alpha = 0.05 \Rightarrow u_{1-\alpha/2} = 1.96$.
L'intervalle de confiance de niveau 0.95 est:

$$\left[\frac{24 \times 14.13}{\left(\sqrt{23} + 1.96\right)^2}; \frac{24 \times 14.13}{\left(\sqrt{23} - 1.96\right)^2}\right] = [7.43; 42.17]$$

3.2.4 Intervalle de confiance pour la proportion

Si la population est formée d'individus ayant ou non un caractère A, les fréquences de A sont les valeurs prises par une v.a F $(= \frac{1}{n}\sum_{i=1}^{n} X_i)$ appellée fréquence de l'echantillon de taille n $(X_1, \ldots, X_n)$. nF représente le nombre d'individus ayant le caractère A . Soit $\mathfrak{p}$ la probabilité pour qu'un individu pris au hasard présente le caractère A, donc

$$nF \hookrightarrow b(n, \mathfrak{p})$$

Pour $n \geq 30$, $n\mathfrak{p} \geq 5$ et $n\mathfrak{p}(1-\mathfrak{p}) \geq 5$ on a:

$$nF \approx N(n\mathfrak{p}, n\mathfrak{p}(1-\mathfrak{p})$$

et par conséquence

$$\frac{F - \mathfrak{p}}{\sqrt{\mathfrak{p}(1-\mathfrak{p})}/\sqrt{n}} \hookrightarrow N(0,1)$$

Pour $\alpha \in]0,1[$ donné, on peut écrire

$$P\left(-u_{1-\alpha/2} \leq \frac{F - \mathfrak{p}}{\sqrt{\mathfrak{p}(1-\mathfrak{p})}/\sqrt{n}} \leq u_{1-\alpha/2}\right) = 1 - \alpha$$

ce qui équivaut à

$$P\left(F - u_{1-\alpha/2}\sqrt{\frac{\mathfrak{p}(1-\mathfrak{p})}{n}} \leq \mathfrak{p} \leq F + u_{1-\alpha/2}\sqrt{\frac{\mathfrak{p}(1-\mathfrak{p})}{n}}\right) = 1 - \alpha$$

Remplaçant F et $\mathfrak{p}$ par f (fréquence observée sur l'echantillon), on obtient donc l'I.C pour la proportion $\mathfrak{p}$ avec coefficient de sécurité $1 - \alpha$ qui est définie par:

$$\left[f - u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}; f + u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}\right]$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $N(0,1)$.

Exemple 69 *Afin d'étudier l'influence des rayons X sur la spermatogénèse de Bombyx Mori, on a irradié des mâles au deuxième jour et au quatrième jour du stade larvaire ; ces mâles ont été accouplés avec des femelles non irradiées. On a compté le nombre d'oeufs fertiles dans la ponte des femelles, et on a obtenu 4998 oeufs fertiles pour 5646 oeufs récoltés en tout. On a aussi accouplé des mâles et des femelles non irradiés, avec un résultat de 5834 oeufs fertiles sur 6221 oeufs récoltés.*
1. *Donner un intervalle de confiance de niveau 0.95 pour la proportion d'oeufs fertiles après irradiation des mâles.*
2. *Donner un intervalle de confiance de niveau 0.95 pour la proportion d'oeufs fertiles de couples non irradiés.*
3. *Que pensez-vous de l'influence de l'irradiation sur la fertilité des oeufs* ?

Réponce
$\alpha = 0.05 \Rightarrow u_{1-\alpha/2} = 1.96$
1. La fréquence empirique des oeufs fertiles après irradiation des mâles est:

$$f = \frac{4998}{5646} = 0.885$$

L'intervalle de confiance de niveau 0.95 est:

$$\left[0.885 - 1.96\sqrt{\frac{0.885(1-0.885)}{5646}}; 0.885 + 1.96\sqrt{\frac{0.885(1-0.885)}{5646}}\right] = [0.876; 0.894]$$

2. La fréquence empirique des oeufs fertiles parmi les couples non irradiés est:

$$f = \frac{5834}{6221} = 0.938$$

L'intervalle de confiance de niveau 0.95 est:

$$\left[0.938 - 1.96\sqrt{\frac{0.938(1-0.938)}{6221}}; 0.938 + 1.96\sqrt{\frac{0.938(1-0.938)}{6221}}\right] = [0.931; 0.944]$$

3. Les deux intervalles de confiance ont une intersection vide ; la proportion d'oeufs fertiles est donc significativement plus basse pour les mâles irradiés.

3.3 Tests

Etablir un test sur un paramètre θ (d'une loi P_θ) est une opération "*inverse*" à la construction d'un intervalle de confiance de θ: au lieu de déterminer une partie de Θ (I.C) qui contient le paramètre Θ avec une grande probabilité, dans la théorie du test on se donne une partie fixée $\Theta_0 \subset \Theta$ est on cherche à controler l'exactitude de l'affirmation $\theta \in \Theta_0$, cette affirmation sera déduit au vu d'un échantillon. L'affirmation $\theta \in \Theta_0$ sera notée: $H_0 : \theta \in \Theta_0$ est appellée hypothése nulle.

Définition 70 1) *Soit $X \hookrightarrow P_\theta$, $\theta \in \Theta$, soit Θ_0 et Θ_1 une partition de Θ. Un test paramètrique entre $H_0 : \theta \in \Theta_0$ et $H_1 : \theta \in \Theta_1$ (appellée hypothése alternative) est une procédure statistique de choix entre H_0 et H_1 au vu d'un echantillon.*
2) ***Test bilatéral:*** *un test bilatéral est de la forme:*

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

3) ***Test unilatéral:*** *un test unilatéral est de la forme:*

$$\begin{cases} H_0 : \theta = \theta_0\ (\geq) \\ H_1 : \theta < \theta_0 \end{cases} \textit{ côté gauche}; \quad \begin{cases} H_0 : \theta = \theta_0\ (\leq) \\ H_1 : \theta > \theta_0 \end{cases} \textit{ côté droit}$$

4) ***Test sur la valeur d'un paramètre:*** *Soit $\lambda = \varphi(\theta)$ un nombre déterminé par θ. On peut tester: $H_0 : \lambda = \lambda_0 (= \varphi(\theta_0))$ de la façon suivante: On prand pour λ un intervalle de confiance (I.C) de seuil α et on définie le test par:*

$$\textit{Rejet } H_0 \iff \lambda_0 \notin I.C$$

3.3.1 Comparaison d'un paramètre observé à un paramètre théorique

Test sur la moyenne

▲)
∗ **Condition d'application**: population normale de variance connue, n quelconque
∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \hookrightarrow N(0, 1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\lvert U\rvert > u_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

▲▲)
∗ **Condition d'application**: population normale de variance inconnue, $n < 30$.
∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit

$$t = \frac{\overline{X} - \mu_0}{S / \sqrt{n}} \hookrightarrow T_{n-1}$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telque $P(\lvert T_{n-1}\rvert > t_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $t \leq t_\alpha$ avec t_α telque $P(\lvert T_{n-1}\rvert > t_\alpha) = 2\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $t \geq -t_\alpha$ avec t_α telque $P(\lvert T_{n-1}\rvert > t_\alpha) = 2\alpha$

▲▲▲)

∗ **Condition d'application**: population quelconque, $n \geq 30$.

∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \hookrightarrow N(0,1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\lvert U\rvert > u_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

Exemple 71 *Le taux normal de glycémie est 1g/l de sang, on dose la glycémie chez 17 sujets diabétique depuis 4 heures, la moyenne est de 1.2g/l avec un écart-type de 0.1g/l. Peut-on dire au risque de 5% que ces sujets sont hyperglycémiques, en supposant que le taux de glycémie est distribué selon une loi normal.*

Réponse

On a $n = 17 < 30$, $\overline{X} = 1.2$, $S = 0.1$, $\mu_0 = 1$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \mu = 1\ (\leq) \\ H_1 : \mu > 1 \end{cases}$$

On calcule

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{1.2 - 1}{0.1/\sqrt{17}} = 7$$

Pour $\alpha = 0.05$, on a $t_\alpha = 1.74$. Comme $t > t_\alpha$, on rejet H_0, c'est à dire le groupe d'individus est hyperglycémiques au risque 5%.

Test sur la variance

▲)
∗ **Condition d'application**: population normale, $n \leq 31$
∗ **Variable requise pour effectuer le test**: en supposant H_0 vraie et selon les conditions d'application, on a:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{nS_e^2}{\sigma_0^2} \hookrightarrow \chi^2_{n-1}$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	accépté H_0 si $\chi^2 \in [a, b]$ avec a telque $P(\chi^2_{n-1} \geq a) = 1 - \alpha/2$ et b telque $P(\chi^2_{n-1} \geq b) = \alpha/2$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	accépté H_0 si $\chi^2 \leq b$ avec b telque $P(\chi^2_{n-1} \geq b) = \alpha$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	accépté H_0 si $\chi^2 \geq -a$ avec a telque $P(\chi^2_{n-1} \geq a) = 1 - \alpha$

▲▲)
∗ **Condition d'application**: population normale, $n > 31$
∗ **Variable requise pour effectuer le test**: en supposant H_0 vraie et selon les conditions d'application, on a:

$$U = \sqrt{\frac{2nS_e^2}{\sigma_0^2}} - \sqrt{2n-3} \hookrightarrow N(0,1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\lvert U\rvert > u_\alpha) = \alpha$
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

Exemple 72 *Une méthode dit de référence pour le titrage de phosphore donne pour une confiance de 95% des résultats dont l'écart-type est de 0.44g/l, une nouvelle méthode est mise au point, on dose le phosphore dans 20 solutions, la variance des résultats obtenus est $0.06g^2/l^2$, en supposant que la $2^{ème}$ méthode suit une loi normal. Peut-on dire que l'écart-type de la $2^{ème}$ méthode est superieur à l'écart-type de la référence, au risque de 5%.*

Réponse
On a $n = 20 \leq 31$, $S_e^2 = 0.06$, $\sigma_0 = 0.44$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \sigma^2 = 0.19\ (\leq) \\ H_1 : \sigma^2 > 0.19 \end{cases}$$

On calcule

$$\chi^2 = \frac{nS_e^2}{\sigma_0^2} = \frac{20 \times 0.06}{0.19} = 6.21$$

Pour $\alpha = 0.05$, on a $b = 30.14$. Comme $\chi^2 < b$, on accépt H_0.

Test sur une proportion

$*$ **Condition d'application**: population binomiale, $n > 30$, $n\mathfrak{p} \geq 5$ et $n(1-\mathfrak{p}) \geq 5$
$*$ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{F - \mathfrak{p}_0}{\sqrt{\frac{\mathfrak{p}_0(1-\mathfrak{p}_0)}{n}}} \hookrightarrow N(0,1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mathfrak{p} = \mathfrak{p}_0$	$\mathfrak{p} \neq \mathfrak{p}_0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\|U\| > u_\alpha) = \alpha$
$\mathfrak{p} \leq \mathfrak{p}_0$	$\mathfrak{p} > \mathfrak{p}_0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\mathfrak{p} \geq \mathfrak{p}_0$	$\mathfrak{p} < \mathfrak{p}_0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

Exemple 73 *Un fabricant d'un médicament affirme qu'il est efficase à au mois 90% pour guerir une allergie en 8 heures. Dans un échantillon de 200 personnes atteintes par cette allergie 160 ont été guéries par le médicament. Peut-on dire que la proportion de guérison par le médicament est plus petite que $\mathfrak{p}_0$ au risque 5%.*

Réponse
On a $n = 200$, $\mathfrak{p}_0 = 0.9$, $F \sim f = \frac{160}{200} = 0.8$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \mathfrak{p} = 0.9\ (\geq) \\ H_1 : \mathfrak{p} < 0.9 \end{cases}$$

On calcule

$$U = \frac{F - \mathfrak{p}_0}{\sqrt{\frac{\mathfrak{p}_0(1-\mathfrak{p}_0)}{n}}} = \frac{0.8 - 0.9}{\sqrt{\frac{0.9(1-0.9)}{200}}} = -4.71$$

Pour $\alpha = 0.05$, on a $u_\alpha = 1.64$. Comme $U < u_\alpha$, on rejet H_0 c'est à dire la proportion au risque de 5% est inférieure à celle affirmée par le fabricant.

3.3.2 Comparaison de deux paramètre experimentales dans le cas d'échantillons indépendants

Test sur la moyenne

▲)
∗ **Condition d'application**: deux échantillons prélevés au hasard et indépendammment de population normales de variances connues σ_1^2 et σ_2^2.
∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \hookrightarrow N(0,1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\|U\| > u_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

Exemple 74 *On souhaite étudier l'evolution du prix d'un produit. En premier mois, àpartir de 40 points de vente pris au hasard, on a obtenu un prix moyen de 25D avec un écart-type de 2D. En deuxieme mois, à partir d'un sondage éffectué sur 35 points de vente, on a obtenu un prix moyen de 27D avec un écart-type de 4D. Ya-t-il une différence significative au risque de 5% entre les prix moyens du produit en* $1^{ère}$ *et* $2^{ème}$ *mois.*

Réponse
On a $n_1 = 40$, $n_2 = 35$, $\overline{X}_1 = 25$, $\overline{X}_2 = 27$, $\sigma_1^2 = 4$, $\sigma_2^2 = 16$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

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

On calcule

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{25 - 27}{\sqrt{\frac{4}{40} + \frac{16}{35}}} = -2.68$$

Pour $\alpha = 0.05$, on a $u_\alpha = 1.96$. Comme $-2.68 \notin]-1.96; 1.96[$, on rejet H_0 c'est à dire les prix moyens du produit en $1^{ère}$ et $2^{ème}$ mois sont significativement différent au risque de 5%.

▲▲)

∗ **Condition d'application**: deux échantillons prélevés au hasard et indépendammment dont les tailles respectives sont ≥ 30.

∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \hookrightarrow N(0,1)$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\|U\| > u_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1 - \alpha$

Exemple 75 *Un laboratoire indépendant a éffectué, pour le compte d'une revue sur la protection du consommateur, un essai de durée de vie sur un type d'ampoules éléctriques fabriquées par deux grandes entreprises, les éssais éffectués dans les mêmes conditions sur un êchantillon de 40 lampes provenant de chaque fabricant donnent les résultats suivants:* $\overline{X}_1 = 1025$ *heures,* $\overline{X}_2 = 1070$ *heures;* $S_1 = 120$ *heures et* $S_2 = 140$ *heures. Est-ce que la revue peut affirmer, qu'en moyenne, les ampoules du fabricant 1 ont une durée de vie inférieure à celles du fabricant 2* ?.

Réponse

On a $n_1 = 40$, $n_2 = 40$, $\overline{X}_1 = 1025$, $\overline{X}_2 = 1070$, $S_1^2 = 14400$, $S_2^2 = 19600$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \ (\geq) \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

On calcule

$$U = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{1025 - 1070}{\sqrt{\frac{14400}{40} + \frac{19600}{40}}} = -1.54$$

Pour $\alpha = 0.05$, on a $u_\alpha = 1.65$. Comme $-1.54 > -1.65$, on accept H_0 c'est à dire la revue peut affirmer, qu'en moyenne, les ampoules du fabricant 1 ont une durée de vie supérieure à celles du fabricant 2 au risque de 5%.

▲▲▲)

∗ **Condition d'application**: échantillons de petite taille ($n_1 < 30$ et/ou ($n_2 < 30$) prélevés au hasard et indépendammment de population normales de variances inconnues mais supposées égales à une valeur commune.

∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \hookrightarrow T_{n_1+n_2-2}$$

où

$$S_c = \sqrt{\frac{\sum_i (X_{i1} - \overline{X}_1)^2 + \sum_i (X_{i2} - \overline{X}_2)^2}{n_1 + n_2 - 2}}$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telque $P(\|t\| > t_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $t \leq t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $t \geq -t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$

Exemple 76 *Un chercheur veut étudier si l'absorption d'une certaine drogue a une influence significative sur l'éxécution d'une tâche de coordination psychomotrice, on a donc choisi au hasard vindt sujets qui ont été répartis au hasard en deux groupes: group contrôle et groupe expérimental. On a administré la drogue au group expérimental avant de leur faire subir l'épreuve, en même temps, un placebo est administré au groupe contrôle, les résultats des deux groupes sont les suivants:*

Groupe contrôle					*Groupe expérimental*				
166	167	169	170	174	167	162	165	168	162
173	172	170	166	173	160	164	158	165	169

On supposera que les résultats de l'épreuve de chaque groupe sont distribués normalement de variances inconnues mais supposées égales à une valeur commun σ^2. Tester, au seuil de signification $\alpha = 0.05$, l'hypothèse nulle selon la quelle la drogue n'a pas d'éffet significatif sur la réaction des sujets soumis à une tâche de coordination psychomotrice.

Réponse

On a $n_1 = 10$, $n_2 = 10$, $\overline{X}_1 = \frac{1700}{10} = 170$, $\overline{X}_2 = \frac{1640}{10} = 164$, $\sum_i (X_{i1} - \overline{X}_1)^2 = 80$,

$\sum_i (X_{i2} - \overline{X}_2)^2 = 112$, $S_c^2 = \frac{80+112}{18} = 10.66$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

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

On calcule

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{170 - 164}{3.26 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 4.10.$$

Pour $\alpha = 0.05$, on a: $t_\alpha = 2.10$. Comme $4.10 \notin]-2.10; 2.10[$, on rejet H_0.
▲▲▲▲)

∗ **Condition d'application**: échantillons de petite taille ($n_1 < 30$ et ($n_2 < 30$) prélevés au hasard et indépendammment de population normales de variances inconnues et différents.

∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \hookrightarrow T_\nu$$

où

$$\nu = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu = \mu_0$	$\mu \neq \mu_0$	accépté H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telque $P(\|t\| > t_\alpha) = \alpha$
$\mu \leq \mu_0$	$\mu > \mu_0$	accépté H_0 si $t \leq t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$
$\mu \geq \mu_0$	$\mu < \mu_0$	accépté H_0 si $t \geq -t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$

3.3.3 Comparaison de deux moyennes dans le cas d'échantillons apparies (dépendants)

Deux échantillons sont dits apparies l'orsqu'ils portent:

∗ Sur les même individus (Ex: 1 group de 20 personnes pour les quels on dose le cholesterol dans 2 temps t_1 et t_2), on a donc 2 séries de donnés provenant d'un même groupe d'individus.

∗ Sur des individus ayant au moins un caractére semblable (Ex: 2 groupes de personnes appries de même age pour les quels on dose le chlesterol à un temps donnée),

on a donc 2 série de données provenant de 2 groupes diférent ont un caractère en commun. Deux échantillons appries ont la même taille n.

Comparaison de deux moyenne

La méthode correcte pour comparer deux séries appariées X_{i1} et X_{i2} $i = 1 : n$ est la méthode des couples et consiste à former, pour chaque paire, la différence $d_i = X_{i2} - X_{i1}$ des deux mesures. On suppose que la différence est distribuée normalement de moyenne μ_d (inconnue) et de variance σ_d^2 (inconnue). On obtient une estimation de μ_d avec la fifférence moyenne $\overline{d} = \frac{\sum_i d_i}{n}$ et σ_d^2 avec $S_d^2 = \frac{\sum_i (d_i - \overline{d})^2}{n-1}$, $\overline{d}$ posséde les propriétes suivantes:
1/ La distribution d'échantillonnage de $\overline{d}$ est normale.
2/ La moyenne de la distribution $\overline{d}$ est $E(\overline{d}) = \mu_d$.
3/ l'écart-type de la distribution d'échantillonnage de $\overline{d}$ est:

$$S(\overline{d}) = \frac{S_d}{\sqrt{n}} \text{ où } S_d = \sqrt{\frac{\sum_i (d_i - \overline{d})^2}{n-1}}.$$

Test sur la moyenne

▲)
∗ **Condition d'application**: deux échantillons apparies distribuée normalement ($n \leq 30$).
∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$t = \frac{\overline{d}}{S_d / \sqrt{n}} \hookrightarrow T_{n-1}.$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu_d = 0$	$\mu_d \neq 0$	accépté H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telque $P(\|t\| > t_\alpha) = \alpha$
$\mu_d \leq 0$	$\mu_d > 0$	accépté H_0 si $t \leq t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$
$\mu_d \geq 0$	$\mu_d < 0$	accépté H_0 si $t \geq -t_\alpha$ avec t_α telque $P(\|t\| > t_\alpha) = 2\alpha$

Exemple 77 *La diréctrice des ressources humaines d'une entreprise veut mettre en oeuvre un programme spécial d'apprentissage pour les employés affectés au département d'assemblage, pour cela on choisi 15 employés au hasard on a observé*

le nombre de piéces assemblées durant une certaine période du temps, on a obtenu les résultats suivants:

Avant le programme X_{i1}	*Aprés le programme* X_{i2}	$d_i = X_{i2} - X_{i1}$
15	17	2
13	16	3
8	10	2
9	9	0
7	9	2
12	13	1
11	14	3
12	15	3
11	14	3
9	11	2
10	14	4
12	11	−1
11	13	2
7	10	3
12	13	1

On suppose que la différence est distribuée selon une loi normal. Est ce que le programme est éfficace ?, avec $\alpha = 0.01$.

Réponse
On a $n = 15 \leq 30$, $\overline{d} = \frac{30}{15} = 2$, $S_d = 1.30$ et $\alpha = 0.01$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \mu_d \leq 0 \\ H_1 : \mu_d > 0 \end{cases}$$

On calcule

$$t = \frac{\overline{d}}{S_d/\sqrt{n}} = \frac{2}{1.3/\sqrt{15}} = 5.91.$$

Pour $\alpha = 0.01$, on a $t_\alpha = 2.62$. Comme $5.91 > 2.62$, on rejet H_0 c'est à dire le programme est efficace au niveau 0.01.
▲▲)
∗ **Condition d'application**: deux échantillons apparies ($n > 30$).
∗ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{\overline{d}}{S_d/\sqrt{n}} \hookrightarrow N(0,1).$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mu_d = 0$	$\mu_d \neq 0$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\|U\| > u_\alpha) = \alpha$
$\mu_d \leq 0$	$\mu_d > 0$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(\|U\| > u_\alpha) = 2\alpha$
$\mu_d \geq 0$	$\mu_d < 0$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(\|U\| > u_\alpha) = 2\alpha$

Test sur la variance

∗ **Condition d'application**: échantillons de taille n_1 et n_2 prélevés au hasard et indépendammment de deux populations normales de variances σ_1^2 et σ_2^2 inconnues.
∗ **Rapport des variances et sa distribution**: en supposant H_0 vraie ($S_1^2 \geq S_2^2$ par ce que la table de la loi de Ficher est donnée pour $P(F \geq f_\alpha) = \alpha$ avec $f_\alpha \geq 1$) et selon les conditions d'application, on a:

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2}{S_2^2} \hookrightarrow F_{(n_1-1),(n_2-1)}$$

Hyp nulle	Hyp alternatives	Régles de décision
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	accépté H_0 si $F \in [F_{1-\frac{\alpha}{2},(n_1-1),(n_2-1)}; F_{\frac{\alpha}{2},(n_1-1),(n_2-1)}]$ avec $F_{\frac{\alpha}{2},(n_1-1),(n_2-1)}$ tq $P(F > F_{\alpha/2,(n_1-1),(n_2-1)}) = \frac{\alpha}{2}$ et $F_{1-\frac{\alpha}{2},(n_1-1),(n_2-1)}$ tq $P(F > F_{1-\frac{\alpha}{2},(n_1-1),(n_2-1)}) = 1-\frac{\alpha}{2}$
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	accépté H_0 si $F < F_{\alpha,(n_1-1),(n_2-1)}$ avec $F_{\alpha,(n_1-1),(n_2-1)}$ tq $P(F > F_{\alpha,(n_1-1),(n_2-1)}) = \alpha$
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	accépté H_0 si $F > F_{1-\alpha,(n_1-1),(n_2-1)}$ avec $F_{1-\alpha,(n_1-1),(n_2-1)}$ tq $P(F > F_{1-\alpha,(n_1-1),(n_2-1)}) = 1-\alpha$

Exemple 78 *On veut comparer la précision de 2 méthode de dosage de menthol dans l'essence de menth poivrée, pour cela on dose le menthol dans 16 flacons par ces 2 méthodes. Les variances des resultats obtenus sont respectivement: 0.013 g^2/l^2 (méthode 1) et 0.024 g^2/l^2 (méthode 2). Peut on dire au risque 5 % que ces 2 méthode n'ont pas la même précision* ?

Réponse
On a $n_1 = 16$, $n_2 = 16$, $S_1^2 = 0.013$, $S_2^2 = 0.024$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

On calcule

$$F = \frac{S_2^2}{S_1^2} = 1.84$$

Pour $\alpha = 0.05$, on a

$$F_{\frac{\alpha}{2},(n_2-1),(n_1-1)} = F_{\frac{0.05}{2},15,15} = 2.86$$

et

$$F_{1-\frac{\alpha}{2},(n_2-1),(n_1-1)} = \frac{1}{F_{\frac{\alpha}{2},(n_1-1),(n_2-1)}} = \frac{1}{2.86} = 0.34.$$

Comme $1.84 \in [0.34; 2.86]$, on accept H_0 c'est à dire les 2 méthode ont la même précision.

Test sur deux proportions

$*$ **Condition d'application**: deux échantillons prélevés au hasard et indépendammment ($n_1 \geq 30, n_2 \geq 30, n_1\mathfrak{p}_1 \geq 5, n_1(1-\mathfrak{p}_1) \geq 5, n_2\mathfrak{p}_2 \geq 5, n_2(1-\mathfrak{p}_2) \geq 5$).
$*$ **Ecart réduit et sa distribution**: en supposant H_0 vraie et selon les conditions d'application l'écart réduit:

$$U = \frac{F_1 - F_2}{\sqrt{\frac{\mathfrak{p}_1(1-\mathfrak{p}_1)}{n_1} + \frac{\mathfrak{p}_2(1-\mathfrak{p}_2)}{n_2}}} = \frac{F_1 - F_2}{\sqrt{\mathfrak{p}(1-\mathfrak{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \hookrightarrow N(0,1)$$

où

$$\mathfrak{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}.$$

Hypothése nulle	Hypothése alternatives	Régles de décision
$\mathfrak{p}_1 = \mathfrak{p}_2$	$\mathfrak{p}_1 \neq \mathfrak{p}_2$	accépté H_0 si $U \in]-u_\alpha, u_\alpha[$ avec u_α telque $P(\lvert U\rvert > u_\alpha) = \alpha$
$\mathfrak{p}_1 \leq \mathfrak{p}_2$	$\mathfrak{p}_1 > \mathfrak{p}_2$	accépté H_0 si $U \leq u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1-\alpha$
$\mathfrak{p}_1 \geq \mathfrak{p}_2$	$\mathfrak{p}_1 < \mathfrak{p}_2$	accépté H_0 si $U \geq -u_\alpha$ avec u_α telque $P(U \leq u_\alpha) = 1-\alpha$

Exemple 79 *A fin de tester un certain médicament, un essai clinique est réalisé en mode ambulatoire. Deux groupes de 40 sujets, chacun sont constitué par tirage au sort. Le groupe 1 reçoit le médicament, le groupe 2 reçoit un placebo. Au cour de cet essai une épidémie de grippe s'abat sur la ville et atteind 15 sujets de groupe 1 et 9 sujets de group 2. Peut on dire au risque de 5 % que le médicament administré augmente la susceptibilité à la contagion ? (c'est à dire la proportion des personnes grippés dans le groupe 1 est-elle plus grande que celle dans le groupe 2 ?).*

Réponse
On a $n_1 = 40$, $n_2 = 40$, $f_1 = \frac{15}{40} = 0.375$, $f_2 = \frac{9}{40} = 0.225$, $\mathfrak{p} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{40\times 0.375 + 40\times 0.225}{40+40} = 0.3$ et $\alpha = 0.05$. On peut effectuer le test statistique suivant:

$$\begin{cases} H_0 : \mathfrak{p}_1 \leq \mathfrak{p}_2 \\ H_1 : \mathfrak{p}_1 > \mathfrak{p}_2 \end{cases}$$

On calcule

$$U = \frac{f_1 - f_2}{\sqrt{\mathfrak{p}(1-\mathfrak{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 1.46$$

Pour $\alpha = 0.05$, on a $u_\alpha = 1.65$ Comme $1.46 < 1.65$, on accept H_0 c'est à dire que le médicament administré ne augmente pas la susceptibilité à la contagion.

3.4 Exercices

Exercice 80 *Soit X une variable aléatoire de densité:*

$$f(x,\theta) = \begin{cases} \frac{A}{x^{1+\frac{1}{\theta}}} & si\ x > 1 \\ 0 & sinon \end{cases}$$

avec $\theta > 0$, nous disposons de $(X_1, \ldots, X_n)$ un échantillon aléatoire de taille n de loi parente X.
1− *Déterminer A.*
2− *Déterminer l'estimateur du maximum de vraisemblanceb de θ.*
3− *Si $\theta < 1$, Déterminer l'estimateur de θ par la méthode des moments.*

Exercice 81 *Soit X une variable aléatoire suivant une loi de Bernoulli de paramètre λ inconnu. Nous disposons $(X_1, \ldots, X_n)$ un échantillon aléatoire de taille n de même loi de X.*
− *Estimer λ par la méthode du maximum de vraisemblance et par la méthode des moments.*

Exercice 82 *Soit $(X_1, \ldots, X_n)$ un échantillion provenant d'une population de densité:*

$$f(x_i,\theta) = \theta^2 x_i e^{-\theta x_i}, \quad x_i > 0$$

− *Déterminer un estimateur de θ par la méthode du maximum de vraisemblance.*

Exercice 83 *Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir* 1625*mg de bicarbonate de sodium. Afin de contrôler la fabrication de ces médicaments, on a prélevé un échantillon de* 150 *comprimés, et on a mesuré la quantité de bicarbonate de sodium pour chacun d'eux. On a obtenu les résultats suivants:*

Classes	[1610; 1615]	]1615; 1620]	]1620; 1625]	]1625; 1630]	]1630; 1635]
Effectifs	*7*	*8*	*42*	*75*	*18*

1− *Caractériser la distribution.*
2− *Représenter graphiquement la distribution.*

3− *Déterminer les paramétres de position (Mode et médiane).*
4− *Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la quantité de bicarbonate de sodium.*
5− *Déterminer un intervalle de confiance au seuil* 5% *de la moyenne de la quantité de bicarbonate de sodium.*
6− *Déterminer un intervalle de confiance au seuil* 2% *de la variance de la quantité de bicarbonate de sodium.*
7− *Peut-on dire aux risque* 10% *que le médicament respecte la norme souhaitée.*

Exercice 84 *On désire estimer la production d'une nouvelle espèce de pommier. On modélise la production d'un pommier de cette espèce par une loi normale d'espérance* μ *et d'écart-type* σ *inconnus. Sur un échantillon de* 15 *pommiers, on a observé une récolte moyenne de* $52kg$ *avec un écart-type de* $5kg$.
1− *Donner une estimation non biaisée de la l'espérance* μ *et de la variance* σ^2.
2− *Déterminer un intervalle de confiance pour la production moyenne des pommiers de cette espèce au risque* $\alpha = 5\%$.
3− *Déterminer un intervalle de confiance pour la variance* σ^2 *au risque* $\alpha = 5\%$.

Exercice 85 *Une clinique a proposé une nouvelle opération chirurgicale, et a connu 40 échecs, sur 200 tentatives. On note* $\mathfrak{p}$ *le pourcentage de réussite de cette nouvelle opération.*
1− *Quelle estimation de* $\mathfrak{p}$ *proposez-vous* ?
2− *En utilisant l'approximation normale, donner un intervalle de confiance pour* $\mathfrak{p}$ *de niveau de confiance 0.95.*
3− *Combien d'opérations la clinique devrait-elle réaliser pour connaître le pourcentage de réussite avec une précision de plus ou moins 1%, au niveau de confiance 0.95* ?

Exercice 86 *On a testé un nouvel antibiotique A pour traiter une infection X chez 10 sujets, qui guérissent respectivement après: 6, 13, 16, 23, 14, 19, 8, 20, 15 et 16 jours.*
1− *Donner une estimation ponctuelle non biaisée de la moyenne et de la variance de la population dont ces sujets sont extraits.*
2− *Déterminer l'intervalle de confiance de la moyenne et de la variance de la population, au risque 5%.*
Un autre antibiotique B est utilisé de nombreuses années pour traiter la même infection X avec une guérisson obtenue en moyenne après 15 jours.
3− *Peut-on dire que l'antibiotique A est meilleur que B au risque 5%.*

Exercice 87 *On a mésuré le poids de raisin par souche sur 10 souches prises au hasard. Les résultats obtenus sont les suivants (en kg):*

2.4 3.2 3.6 4.1 4.3 4.7 5.4 5.9 6.5 6.9

1− *Déterminer une estimation ponctuelle sans biais de la moyenne et de l'écart-type de la population dont ces souches sont extraites.*
2− *Donner un intervalle de confiance de la moyene de la population au risque 5%. En supposant que le poids de raisin par souche suit une loi normale. A la même époque, un grand nombre de mésure a permis d'établir que le raisin par souche avait un poids moyen de 5kg.*
3− *Peut-on dire que l'échantillon étudié est conforme à cette norme au risque 5%*?

Exercice 88 *On suppose que chez les femmes non malades, la teneur en hémoglobine du sang (en g pour 100 ml) est une variable aléatoire de loi normale de moyenne 14,5 et d'écart-type 1,1. Sur un échantillon de 20 femmes, on trouve une teneur moyenne en hémoglobine de 13,8 et un écart-type corrigé de 1,2.*
− Au risque de 5%, peut-on conclure que la population de femmes dont est extrait cet échantillon présente une teneur en hémoglobine normale ? trop faible ?

Exercice 89 *Dans une usine du secteur de l'agroalimentaire, une machine à embouteiller est alimentée par un réservoir d'eau et par une file d'approvisionnement en bouteilles vides. Pour contrôler le bon fonctionnement de la machine, on veut construire un test d'hypothèse bilatéral qui sera mis en oeuvre toutes les heures. Pour une production d'une heure, on suppose que la variable aléatoire X qui à toute bouteille, prise au hasard dans cette production, associe le volume d'eau (en litres) qu'elle contient, est une variable aléatoire d'espérance μ et d'écart-type σ inconnus. On considère que la machine est bien réglée lorsque le volume d'eau moyen dans une bouteille est σ=1.5 l. On a prélevé un échantillon de 100 bouteilles, et on a obtenu un volume d'eau moyen de 1.495l et un écart-type corrigé de 0.01. Peut-on conclure, au risque 5%, que la machine est bien réglée ?*

Exercice 90 *Afin de tester une solution toxique, on fait des injections à un groupe de 80 souris. On admet que l'injection est mortelle dans 80% des cas. Le fait que 22 souris ne soient pas mortes est-il compatible au seuil 5% avec cette hypothèse ?*

Exercice 91 *La durée de gestation humaine est en moyenne de 40,5 semaines.*
1. *Dans une maternité, on a noté l'âge gestationnel de 100 nouveaux-nés successifs. On a observé une moyenne de 38,5 semaines et un écart-type de 5 semaines. On pense que cette maternité est spécialisée dans les accouchements prématurés.*
− Tester cette hypothèse au risque de 5%.
2. *Dans cette même maternité, les mères des 100 nouveaux-nés suivants ont reçu un traitement inhibant les contractions utérines. Pour ces nouveaux-nés, on a observé une moyenne de 39,5 semaines et un écart-type de 4 semaines.*
− Tester l'égalité des moyennes des durées de gestation des 2 groupes au risque 2%.

Exercice 92 *Pour comparer deux médicaments en fonction de leur activité tachycardisante (accélératrice du rythme cardiaque), on effectue un sondage dans une population de malades de façon à grouper* $n_1 = 40$ *malades utilisant le premier médicament et* $n_2 = 30$ *malades utilisant le second médicament. On évalue ensuite pour chaque malade la différence y entre la fréquence cardiaque une heure après la prise et la fréquence juste avant la prise. On obtient, pour des fréquences exprimées en battements par minute, les résultats suivants pour la moyenne et la variance corrigée:*

Groupe	$\overline{X}$	S^2
1	*7.7*	*2*
2	*8.6*	*6*

− Peut-on conclure avec une confiance de 98% qu'il y a une différence entre les médicaments ?

Exercice 93 *Le myélome se traduit par une prolifération des plasmocytes, ce qui provoque un accroissement des globulines (protéines du sang ayant fonction d'anticorps). On veut tester cet accroissement chez des malades atteints de myélome à un stade précoce. A cet effet, on mesure en g/l les concentrations des globulines dans le sang dans un échantillon de tels malades, puis dans un échantillon de personnes en bonne santé. On suppose enfin que dans la population toute entière, le taux de globuline suit une loi normale.*
∗ Echantillon de malades: taille 20, moyenne 50 g/l, écart-type 14 g/l.
∗ Echantillon de personnes saines: taille 10, moyenne 40 g/l, écart-type 12 g/l.
− Faire un test d'hypothèse bilatéral au risque 5% pour la moyenne.

Exercice 94 *Chez un groupe de 10 malades, on expérimente les effets d'un traitement destiné à diminuer la pression artérielle. On observe les résultats suivants (valeur de la tension artérielle systolique en cm Hg):*

avant traitement	*15*	*18*	*17*	*20*	*21*	*18*	*17*	*15*	*19*	*16*
après traitement	*12*	*26*	*17*	*18*	*17*	*15*	*18*	*14*	*16*	*18*

− On se demande si le traitement à une action significative au risque de 5%.

Exercice 95 *Au cours d'une étude destinée à comparer diverses méthodes d'échantillonnage de sols forestiers, on a mesuré les teneurs en* K_2O*, d'une part pour 20 échantillons de terre prélevés individuellement, et d'autre part pour 10 échantillons mélangés obtenus chacun à partir de 25 terres différentes. On a obtenu pour les échantillons individuels:*

$$\sum_i x_i = 259.2 \quad \sum_i x_i^2 = 3662.08$$

et pour les échantillons mélangés:

$$\sum_i y_i = 109.2 \quad \sum_i y_i^2 = 1200.8$$

− On s'attend à ce que les deux méthodes d'échantillonnage donnent des variances très différentes. Justifier cela intuitivement et vérifier le par le test de Fisher.

Exercice 96 *On désire savoir si, chez les individus qui consomment régulièrement de l'huile d'olive, le risque cardio-vasculaire est diminué. On utilise pour cela le logarithme du dosage en d-dimères, modélisé par une loi normale. Sur un échantillon de 9 individus consommant de l'huile d'arachide, on a observé une moyenne de -0.78, avec un écarttype de 0.27. Sur un échantillon de 13 individus consommant de l'huile d'olive, on a observé une moyenne de -0.97, avec un écart-type de 0.32.*
1− *Tester l'hypothèse d'égalité des variances au seuil 0.05.*
2− *Au seuil de 0.05, quel test proposez-vous pour décider si l'huile d'olive abaisse significativement le risque cardio-vasculaire* ?
3− *On effectue des dosages sur 110 individus consommant de l'huile d'arachide, pour lesquels on observe une moyenne de -0.82, avec un écart-type de 0.29, et sur 130 individus consommant de l'huile d'olive, pour lesquels on observe une moyenne de -0.93, avec un écart-type de 0.31. On se demande si l'amélioration est significative au risque de 5%.*

Exercice 97 *On compare les effets d'un même traitement dans deux hopitaux différents. Dans le premier hopital, 70 des 100 malades traités montrent des signes de guérison. Dans le deuxième hopital, c'est le cas pour 100 des 150 malades traités.*
− Quelle conclusion peut-on en tirer au risque de 5% ?

Exercice 98 *Pour traiter un certain type de tumeur, on a utilisé deux schémas thérapeutiques:*
∗ *sur 40 malades traités avec le schéma A, on a observé une mortalité à 5 ans de 15%;*
∗ *sur 60 malades traités avec le schéma B, on a observé une mortalité à 5 ans de 25%.*
Si l'on considère la mortalité à 5 ans, peut-on dire que les schémas A et B diffèrent significativement au risque 10% ? *au risque 5%* ?

Chapitre 4

Corrélation et régression linéaire simple

4.1 Nuage de points

L'existence d'une corrélation entre deux variables peut être décelée graphiquement. Il s'agit de reporter les couples d'observations (X_i, Y_i) sur un graphique en prenant pour abscisse la variable X_i et pour ordonnée la variable Y_i. Le graphique résultant constitue un nuage de points appelé diagramme de dispersion. La forme de ce nuage de points nous permettra de constater si les variables concernées sont en corrélation. Nous traitons ici que la forme linéaire.

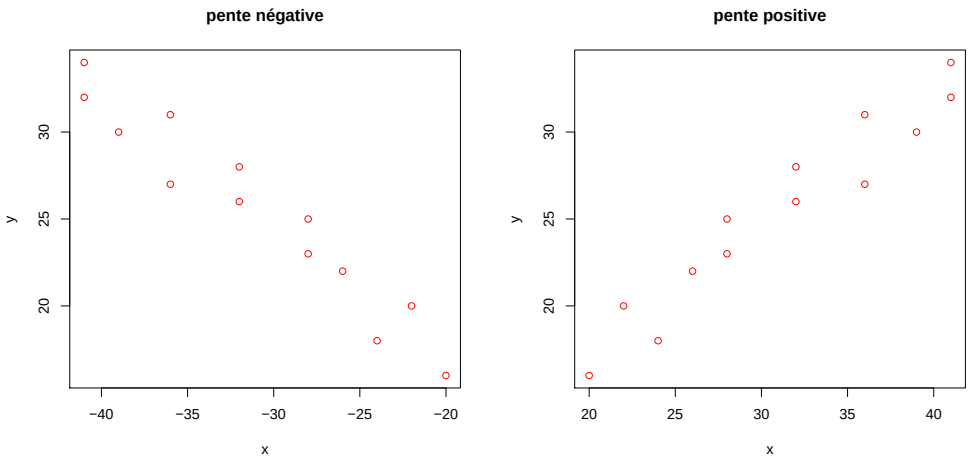


Figure 4.1: Nuage de points.

4.2 Coefficient de corrélation

Dans le cas où le nuage de points prend une forme allongé telle que les points le constituant semblent se répartir autour d'une droite (de pente positive ou négative), on peut calculer un indice qui mesure l'intensite de la liaison linéaire entre les deux variables. Nous définissons le coefficient de corrélation linéaire comme suit:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}},$$

où

$S_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ $\quad$ $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = (n-1)S^2$

$S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ $\quad$ $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad$ $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Remarque 99 $* -1 \leq r_{XY} \leq 1$
$*$ *Si* $r = -1$ *ou* $r = 1$ *alors il y a corrélation parfaite entre X et Y et les points* $(X_i; Y_i)$ *sont tous sur la droite de régression.*
$*$ *Si* $r = 0$ *alors il n'y a pas de corrélation entre X et Y et les points* $(X_i; Y_i)$ *sont dispersés au hasard.*
$*$ *Si* $0 < r < 1$ *alors il y a corrélation positive faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une augmentation de Y.*
$*$ *Si* $-1 < r < 0$ *alors il y a corrélation négative faible, moyenne ou forte entre X et Y. Dans ce cas, une augmentation de X entraîne une diminution de Y.*

4.3 Modèle de régression linéaire simple

Un probléme de régression consiste à étudier les changements de la valeur moyenne d'une variable quand une autre variable ou plusieurs autres variables prennent différentes valeurs fixer. La premiére variable est appelée variable dépendante ou variable expliquée, les autres variables sont appelées variables indépendantes ou variables explicatives. Le cas le plus élémentaire de modèle de régression est défini par une équation de la forme

$$Y_i = aX_i + b + \xi_i \quad i = 1 : n$$

où a et b sont deux paramètres inconnue à estimer et $\xi_i \sim N(0, \sigma^2)$ (avec σ^2 inconnue à estimer) est l'erreur aléatoire du modèle.

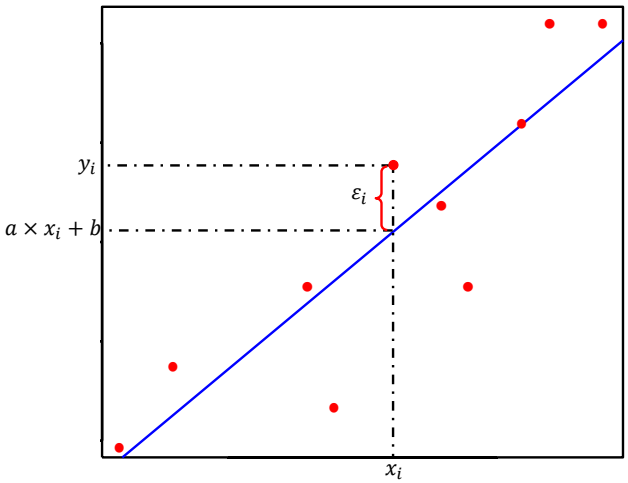


Figure 4.2: Modèle de régression linéaire simple.

4.4 Estimation des paramètres a, b et σ^2

En régression linéaire simple, l'objectif est d'obtenir une droite qui s'ajuste le mieux possible aux points du diagramme de dispersion. plusieurs droites peuvent s'ajustes à un nuage de points mais parmi toutes ces droites, on veut retenire celle qui permet de rendre minimum la somme des carrés des écarts des valeurs observées Y_i à la droite, c'est à dire il s'agit de determiner les expressions de a et b de telle sorte que $\sum_{i=1}^{n}(Y_i - aX_i - b)^2 \doteq f(a,b)$ soit la plus petite possible. Pour minimiser l'expression $f(a,b)$ par rapport à a et b, on a recours aux dérivées partielles. Le premier critere que l'on doit satisfaire pour minimiser cette somme de carrées est l'annulation des dérivées premiéres parrapport à a et b. i.e.

$$\begin{cases} \frac{\partial f}{\partial a} = 0 \\ \\ \frac{\partial f}{\partial b} = 0 \end{cases} \Longrightarrow \begin{cases} \widehat{a} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{S_{XY}}{S_{XX}} \\ \\ \widehat{b} = \overline{Y} - \widehat{a}\overline{X} \end{cases}$$

Un estimateur pour σ^2 est défini par:

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

où $\widehat{Y_i} = \widehat{a}X_i + \widehat{b}$ est la valeur prédite (estimée)

4.5 Intervalles de confiance pour a et b

$\widehat{a}$ et $\widehat{b}$ sont des estimateurs sans biais, leure variances est définie réspectivement par $V(\widehat{a}) = \frac{\sigma^2}{S_{XX}}$ et $V(\widehat{b}) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]$.
Les statistiques $\frac{\widehat{a}-a}{\sqrt{\widehat{\sigma}^2/S_{XX}}}$ et $\frac{\widehat{b}-b}{\sqrt{\widehat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{X}^2}{S_{XX}}\right]}}$ suitent la loi de Student à $n-2$ degrés de liberté.
L'intervalle de confiance de niveau $1-\alpha$ ($\alpha \in]0,1[$) pour le paramétre a est définie par:

$$\left[\widehat{a} - t_{\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}; \widehat{a} + t_{\alpha/2}\sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}\right].$$

De même un intervalle de confiance de niveau $1-\alpha$ ($\alpha \in]0,1[$) pour le paramétre b est définie par:

$$\left[\widehat{b} - t_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]}; \widehat{b} + t_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]}\right]$$

où $t_{\alpha/2}$ est la fractile d'ordre $1-\frac{\alpha}{2}$ pour la loi de Student a $(n-2)$ degrés de liberté.

4.5.1 Intervalles de confiance pour la droite de régression

Il s'agit d'un intervalle de confiance pour $E(Y_0|X_0)$, la réponse moyenne à la valeur X_0.
Pour X_0 donné, $\widehat{Y_0} = \widehat{a}X_0 + \widehat{b}$ est l'estimateur de $E(Y_0|X_0)$. Donc un intervalle de confiance de niveau $1-\alpha$ ($\alpha \in]0,1[$) pour le paramétre $E(Y_0|X_0)$ est donné par:

$$\left[\widehat{Y_0} - t_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(\overline{X}-X_0)^2}{S_{XX}}\right]}; \widehat{Y_0} + t_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(\overline{X}-X_0)^2}{S_{XX}}\right]}\right]$$

où $t_{\alpha/2}$ est la fractile d'ordre $1-\frac{\alpha}{2}$ pour la loi de Student a $(n-2)$ degrés de liberté.

4.6 Tests d'hypothèses

4.6.1 Tests sur le paramètre a

L'écart réduit:

$$t = \frac{\widehat{a} - a_0}{\sqrt{\widehat{\sigma}^2/S_{XX}}} \sim T_{n-2}$$

nous permet d'établir un test paramétrique de la forme

$$\begin{cases} H_0 : a = a_0 \\ H_1 : a \neq a_0, \end{cases}$$

et on a comme régles de décision: on accepte H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telle que $p(|t| > t_\alpha) = \alpha$ et $\alpha \in]0, 1[$.

4.6.2 Tests sur le paramètre b

L'écart réduit:

$$t = \frac{\widehat{b} - b_0}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]}} \sim T_{n-2}$$

nous permet d'établir un test paramétrique de la forme:

$$\begin{cases} H_0 : b = b_0 \\ H_1 : b \neq b_0, \end{cases}$$

et on a comme régles de décision: on accepte H_0 si $t \in]-t_\alpha, t_\alpha[$ avec t_α telle que $p(|t| > t_\alpha) = \alpha$ et $\alpha \in]0, 1[$.

4.7 Tests de significativité

Il s'agit d'un teste qui mesure l'impact de X dans l'explication de Y via le modèle, qui ce traduit par tester la nullité de la pente a (absence de liaison linéaire entre X et Y)

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0, \end{cases}$$

Accepter H_0 implique que l'on conclut qu'il n'y a pas de relation linéaire entre X et Y. Ceci peut signifier que:
► la relation entre X et Y n'est pas linéaire.
► la variation de X influe peu ou pas sur la variation de Y.
Rejeter H_0 implique que l'on conclut que la variation de X influe sur la variation de Y .

Exemple 100 *On a calculer le rendement de maïs Y (en quintal) à partir de la quantité d'engrais utilisé X (en kilo) sur des parcelles de terrain similaires. On a obtenu les résultats suivants:*

X_i	20	24	28	22	32	28	32	36	41	41
Y_i	16	18	23	24	28	29	26	31	32	34

On veut tester la réalité d'une relation linéaire entre Y et X, soit:

$$Y = aX + b + \xi$$

Réponse
$n = 10$

$\overline{X} = 30.4$	$S_{XY} = 351.6$	$S_{YY} = 314.9$
$\overline{Y} = 26.1$	$S_{XX} = 492.4$	$\sum_i (Y_i - \widehat{Y_i})_2 = 63.83$

► Coefficient de corrélation

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{351.6}{393.77} = 0.89$$

► Estimation des paramètres a, b et σ^2

$$\widehat{a} = \frac{S_{XY}}{S_{XX}} = \frac{351.6}{492.4} = 0.71; \quad \widehat{b} = \overline{Y} - \widehat{a}\overline{X} = 4.51; \quad \widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = 7.97$$

► Intervalles de confiance pour a et b
$\alpha = 0.05$, $t_{\alpha/2} = 2.306$

$$I.C_a = \left[\widehat{a} - t_{\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}; \ \widehat{a} + t_{\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}} \right] = [0.42; 1]$$

$$I.C_b = \left[\widehat{b} - t_{\alpha/2} \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}} \right]}; \ \widehat{b} + t_{\alpha/2} \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}} \right]} \right] = [-4.7; 13.72]$$

► Tests de significativité
$\alpha = 0.05$, $t_{\alpha/2} = 2.306$.
On test les deux hyphothèse:

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0, \end{cases}$$

pour ce la on utilise la statistique

$$t = \frac{\widehat{a}}{\sqrt{\widehat{\sigma}^2 / S_{XX}}} = 5.91$$

et on applique la régle de décision: on accepte H_0 si $t \in]-t_{\alpha/2}, t_{\alpha/2}[$.
Comme $5.91 \notin]-2.306, 2.306[$, donc on rejet H_0 i.e le coefficient a est très significativement différent de 0.

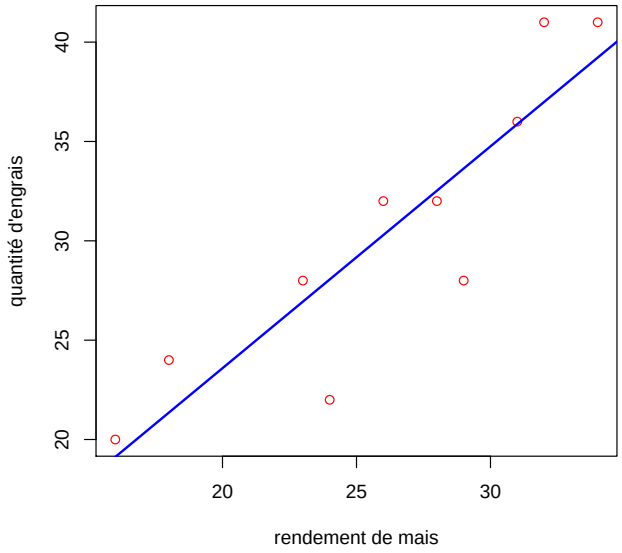


Figure 4.3: Droite de régression du rendement de maïs.

4.8 Exercices

Exercice 101 *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 X_i et le nombre de pulsations Y_i par seconde est bien approchée par une droite de régression. 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– *Tracer le nuage de points.*
2– *Calculer le coefficient de corrélation linéaire entre Y et X.*
3– *Déterminer la droite de régression de Y % X.*
4– *Si la température augmente de* 3 *degré, de combien augmentera le nombre de pulsations.*

Exercice 102 *Un étudiant en techniques forestières veut utiliser la régression linéaire pour estimer le volume X_i en bois utilisable d'un arbre debout en fonction de l'aire Y_i du tronc mesuré à* 25*cm du sol. Il a choisi au hasard* 10 *arbres et a mesuré, à la base, l'aire correspondante (en cm^2). Il a par la suite enregistré, une fois l'arbre*

coupé, le volume correspondant en m^3.

X_i	0.152	0.284	0.187	0.350	0.416	0.230	0.242	0.276	0.383	0.140
Y_i	297	595	372	687	790	520	473	585	762	232

1– *Calculer le coefficient de corrélation linéaire de volume et l'aire.*
2– *Donner l'équation de la droite de régression de volume par rapport à l'aire.*
3– *Tester la nullité de la pente a de la droite de régression de volume par rapport à l'aire au risque 5 %.*

Exercice 103 *Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O_3 dans l'air (en microgrammes par millilitre). En particulier, on cherche à savoir s'il est possible d'expliquer le taux maximal d'ozone Y_i de la journée par la température T_{12} X_i à midi. Les données sont:*

X_i	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
Y_i	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

1– *Tracer le nuage de points.*
2– *Calculer le coefficient de corrélation linéaire entre Y et X.*
3– *Déterminer la droite de régression de Y % X.*
4– *Déterminer un intervalles de confiance pour la droite de régression au risque 2%.*

Exercice 104 *On sélectionne* 10 *personnes inscrites à un stage de formation. Avant le début de la formation, ces stagiaires subissent une épreuve A notée de* 0 *à* 20*. A l'issue du stage, une épreuve B identique à la première est aussi notée de* 0 *à* 20*. Considérant les deux variables X note de A et Y note de B, on a obtenu les résultats suivants:*

Stagiaire	1	2	3	4	5	6	7	8	9	10
X_i	3	4	6	7	9	10	9	11	12	13
Y_i	8	9	10	13	15	14	13	16	13	19

1– *Représenter ces résultats par un nuage de points.*
2– *Calculer le coefficient de corrélation linéaire entre X et Y.*
3– *Déterminer une équation de la droite de régression de Y en X.*
4– *Déterminer un intervalles de confiance pour les paramètres a et b de la droite de régression au risque 5%.*

Exercice 105 *Lors d'une période de sécheresse, un agriculteur relève la quantité totale (en m^3) utilisée par son exploitation depuis le premier jour et donne le résultat suivant:*

Nombre de jours écoulés x_i	1	3	5	8	10
Volume utilisé (en m^3) y_i	2.25	4.3	8	17.5	27

1– *Représenter graphiquement la série* (x_i, y_i).
2– *Calculer le coefficient de corrélation de y en fonction de x.*
3– *Déterminer la droite de régression de y par rapport à x.*

Exercice 106 *Dans la série statistique suivante, X représente le nombre de jours d'exposition au soleil d'une feuille et Y le nombre de stomates aérifères au millimètre carré:*

X	*2*	*4*	*8*	*10*	*24*	*40*	*52*
Y	*6*	*11*	*15*	*20*	*39*	*62*	*85*

1– *Tracer le nuage des points.*
2– *Calculer le coefficient de corrélation linéaire entre X et Y.*
3– *Déterminer l'équation de la droite de régression de Y en fonction de X.*
4– *Si on expose au soleil une feuille 15 jours; quel est le nombre de stomates aérifères peut-on prévoir* ?

Exercice 107 *On répartit dans 10 tubes des volumes égaux de culture additionnées d'une quantité X d'antibiotique, et on mesure, après incubation, la densité optique D. La densité optique permet de déterminer la concentration en bactérie du milieu de culture.*

X	*0.2*	*0.2*	*0.4*	*0.4*	*0.6*	*0.6*	*0.8*	*0.8*	*1*	*1*
D	*19*	*21*	*35*	*38*	*64*	*66*	*115*	*130*	*200*	*210*

1– *Construire le nuage des points M de coordonnées* (X_i, D_i) *représentant la densité optique en fonction de la concentration d'antibiotique.*
2– *Un ajustement linéaire semble-t-il justifié* ?

Exercice 108 *L'analyse de la température de fonctionnement d'un procédé chimique sur le rendement du produit a donné les valeurs suivantes pour la température* X_i *et le rendement correspondant* Y_i*:*

X_i	*100*	*150*	*110*	*160*	*120*	*170*	*130*	*180*	*140*	*190*
Y_i	*45*	*70*	*51*	*74*	*54*	*78*	*61*	*85*	*66*	*89*

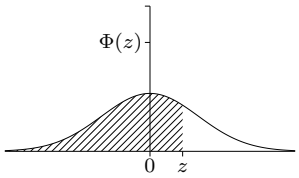
1– *Représenter ces résultats par un nuage de points.*
2– *Calculer le coefficient de corrélation linéaire entre X et Y.*
3– *Déterminer une équation de la droite de régression de Y en X.*
4– *Déterminer un intervalles de confiance pour les paramètres a et b de la droite de régression au risque 2%.*

Chapitre 5

Tables usuelles

Table 1. Loi Normale N(0,1)

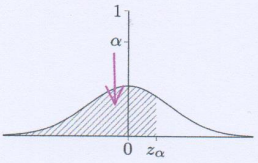
- Fonction de répartition de la loi Normale. — La fonction de répartition Φ de la loi Normale N $(0,1)$ est définie par $\Phi(z) = \int_{-\infty}^{z} \mathrm{e}^{-u^2/2}\,\mathrm{d}u/\sqrt{2\pi}$, $z \in \mathbb{R}$. Pour tout $z \in \mathbb{R}$, on a $\Phi(z) = 1 - \Phi(-z)$.



z	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

Exemples. — $\Phi(0{,}25) \approx 0{,}5987$, $\Phi(-0{,}32) = 1 - \Phi(0{,}32) \approx 1 - 0{,}6255 = 0{,}3745$.

Table 2



1. *Quantiles de la loi Normale.* — Pour $\alpha \in]0,1[$, le quantile d'ordre α de la loi Normale est $z_\alpha = \Phi^{-1}(\alpha)$. Pour tout $\alpha \in]0,1[$, on a $\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$.

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,5	0,0000	0,0251	0,0502	0,0753	0,1004	0,1257	0,1510	0,1764	0,2019	0,2275
0,6	0,2533	0,2793	0,3055	0,3319	0,3585	0,3853	0,4125	0,4399	0,4677	0,4959
0,7	0,5244	0,5534	0,5828	0,6128	0,6433	0,6745	0,7063	0,7388	0,7722	0,8064
0,8	0,8416	0,8779	0,9154	0,9542	0,9945	1,0364	1,0803	1,1264	1,1750	1,2265
0,9	1,2816	1,3408	1,4051	1,4758	1,5548	1,6449	1,7507	1,8808	2,0537	2,3263

α	0,990	0,991	0,992	0,993	0,994	0,995	0,996	0,997	0,998	0,999
$\Phi^{-1}(\alpha)$	2,3263	2,3656	2,4089	2,4573	2,5121	2,5758	2,6521	2,7478	2,8782	3,0902

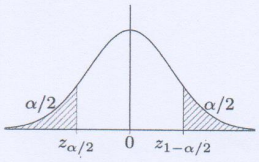
α	0,9990	0,9991	0,9992	0,9993	0,9994	0,9995	0,9996	0,9997	0,9998	0,9999
$\Phi^{-1}(\alpha)$	3,0902	3,1214	3,1559	3,1947	3,2389	3,2905	3,3528	3,4316	3,5401	3,7190

Exemples. — On a $\Phi^{-1}(0{,}75) \approx 0{,}6745$, $\Phi^{-1}(0{,}995) \approx 2{,}5758$, $\Phi^{-1}(0{,}9995) \approx 3{,}2905$; ainsi que $\Phi^{-1}(0{,}25) \approx -0{,}6745$, $\Phi^{-1}(0{,}005) \approx -2{,}5758$, $\Phi^{-1}(0{,}0005) \approx -3{,}2905$.

2. *Quantiles de la loi Normale (bis).* — Si Z est une variable aléatoire suivant la loi normale $\mathcal{N}(0,1)$, la table donne, pour α fixé, la valeur $z_{1-\alpha/2}$ telle que

$$\mathbb{P}\{|Z| \geqslant z_{1-\alpha/2}\} = \alpha.$$

Ainsi, $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	∞	2,5758	2,3263	2,1701	2,0537	1,9600	1,8808	1,8119	1,7507	1,6954
0,1	1,6449	1,5982	1,5548	1,5141	1,4758	1,4395	1,4051	1,3722	1,3408	1,3106
0,2	1,2816	1,2536	1,2265	1,2004	1,1750	1,1503	1,1264	1,1031	1,0803	1,0581
0,3	1,0364	1,0152	0,9945	0,9741	0,9542	0,9346	0,9154	0,8965	0,8779	0,8596
0,4	0,8416	0,8239	0,8064	0,7892	0,7722	0,7554	0,7388	0,7225	0,7063	0,6903
0,5	0,6745	0,6588	0,6433	0,6280	0,6128	0,5978	0,5828	0,5681	0,5534	0,5388
0,6	0,5244	0,5101	0,4959	0,4817	0,4677	0,4538	0,4399	0,4261	0,4125	0,3989
0,7	0,3853	0,3719	0,3585	0,3451	0,3319	0,3186	0,3055	0,2924	0,2793	0,2663
0,8	0,2533	0,2404	0,2275	0,2147	0,2019	0,1891	0,1764	0,1637	0,1510	0,1383
0,9	0,1257	0,1130	0,1004	0,0878	0,0753	0,0627	0,0502	0,0376	0,0251	0,0125

α	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
$z_{1-\alpha/2}$	3,2905	3,8906	4,4172	4,8916	5,3267	5,7307	6,1094

Exemples. — Pour $\alpha = 0{,}5$, on trouve $z \approx 0{,}6745$; pour $\alpha = 0{,}25$, on trouve $z \approx 1{,}1503$; pour $\alpha = 10^{-6}$, on trouve $z \approx 4{,}8916$.