République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA
FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE
DÉPARTEMENT DE MATHÉMATIQUE

Polycopié du Cours

BIOSTATISTIQUES

Statistiques Appliquées à l'Expérimentation En Sciences Biologique

Préparé par :
Dr. CHERFAOUI Mouloud

Université de Biskra, 2016/2017

Avant-Propos

On n'a pas besoin de grandes enquêtes d'opinions pour se rendre compte que les biologistes sont globalement frileux à se frotter aux statistiques. L'étape de l'analyse des résultats est souvent vécue comme une contrainte, un passage obligé mais désagréable, voire même parfois un calvaire. Pourtant, le premier objectif des statistiques est bien de révéler ce que les données ont à nous dire. Passer à côté d'une bonne analyse par manque de temps, de motivation ou de compétence, c'est surtout prendre le risque de rater un phénomène intéressant qui était pourtant là, sous nos yeux.

Ce polycopié de cours est destiné principalement aux étudiants de la Licence Biologie mais peut-être utile à toute personne souhaitant connaître et surtout utiliser les principales méthodes de la statistique.

Le niveau mathématique requis est celui de la première année et deuxième année Licence avec quelques notions : Statistiques descriptives, probabilités unidimensionnelles, suites, intégrales,... (souvent enseignées en première et en deuxième année Licence).

L'enseignement de la Biostatistique en premier cycle (département biologie de l'université de Biskra) est subdivisé en trois parties essentielles : "*Mathématique Statistique et Informatique*", "*Biostatistique*" et "*Analyse de données en Bioscience*". Ces trois parties complémentaires ont pour objectif de permettre aux étudiants de développer des compétences qui leur permettront :

- D'acquérir et de parfaire la connaissance des principales notions relatives à l'utilisation des méthodes statistiques.
- De résoudre des questions empiriques par l'utilisation des tests statistiques.
- De maîtriser et de compléter les notions de bases des statistiques en vue de les appliquer à des exemples spécifiques aux sciences biologiques, prises dans leur sens général (biologie, médecine, pharmacie, écologie,...).
- D'appliquer ces notions et méthodes sur des données biologiques à partir de logiciels simples (Excelstat, SPSS, R,...).
- D'utiliser des logiciels de statistique et d'apprendre la lecture de leurs résultats.

Le présent polycopié se conclut par des exercices, orientés dans le sens de la biologie, avec un corrigé suffisamment détaillé permettant ainsi de contrôler l'acquisition des notions essentielles qui ont été introduites.

Table des matières

Introduction générale

La statistique a une origine très ancienne, se réduisant initialement à une collecte de données (recensement). Le terme "Statistique", vient du latin "statisticus", relatif à l'état (status), il est employé alors dans un sens purement descriptif de recueil ou de collection de données chiffrées, "les statistiques". Le mot employé au singulier, avec l'article défini "la statistique" évoque la méthode utilisée pour étendre les résultats et dégager des lois (l'inférence) c'est donc une méthode scientifique d'analyse des données recueillies.

Les statistiques constituent, en biologie, l'outil permettant de répondre à de nombreuses questions qui se posent souvent aux biologistes, à titre d'exemples :

- Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
- Quelle est la fiabilité d'une mesure ou d'une observation ?
- Quel est le risque ou l'avantage d'un traitement ?
- Les conditions expérimentales "A" sont-elles plus efficaces que celles des conditions de "B" ?
- Les effets de la variable "A" sont-ils les mêmes ou diffèrent-ils des effets de la variable "B" ?

Les statistiques, dans le sens populaire du terme, traitent des populations. Leur objectif consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. On peut alors chercher à extrapoler une information obtenue à partir de l'échantillon. En effet, l'objectif de *statistique inférentielle* en générale et *Biostatistique* en particulier, est de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons ou de comparer statistiquement et de façon significative si des échantillons sont identiques ou non selon un ou plusieurs paramètres ou tests (indépendance, hypothèses, estimation,...).

Le présent polycopié reprend les éléments de bases des statistiques descriptives en y introduisant une approche plus probabiliste. Les méthodes statistiques orientées vers des études classiques d'estimation et tests d'hypothèse, de manière à satisfaire les conditions d'application des méthodes de l'inférence (approche déductiviste). Il fournit également, des outils statistiques qui permettent d'étendre ou de généraliser, dans certaines conditions, les conclusions obtenues par la statistique descriptive à partir de la fraction des individus (échantillon) que l'on a observé ou étudié expérimentalement, à l'ensemble des individus constituant la population.

En plus de la présente introduction, le document contient 6 chapitres, une bibliographie et une annexe. La suite de cette introduction décrit brièvement le contenu de chaque chapitre.

Chapitre 1 : La statistique inférentielle a pour objet de généraliser les résultats observés sur un échantillon. Toutefois, elle demande d'abord une description précise de cet échantillon par des

méthodes graphiques ou des résumés statistiques. C'est pourquoi nous présenterons, dans ce chapitre, un rappel sur ces méthodes descriptives qui décrit quelques exemples typiques de problèmes statistiques descriptives ainsi que les bases de probabilités nécessaires à la compréhension des méthodes d'analyse statistique.

Chapitre 2 : Nous introduisons dans ce chapitre les techniques de base pour l'estimation de paramètres. Ces techniques sont regroupées en deux catégories : estimation ponctuelle et estimation par intervalle de confiance.

Chapitre 3 : L'approche inférentielle choisie ici est particulière car on n'a présenté que les tests d'hypothèses classiques. En effet, le chapitre concerne les tests de conformité d'un échantillon, d'homogénéité de deux échantillons, d'homogénéité de plus de deux échantillons (ANOVA 1 et ANOVA 2), d'indépendance de deux variables et d'ajustement de la distribution d'un échantillon.

Chapitre 4 : Nous introduisons dans ce chapitre la notions de régression d'une variable réelle Y par rapport à une variable X par une droite (régression linéaire simple). Ensuite, nous présentons une généralisation de la notion de droite, en remplaçant X par plusieurs variables X_1, X_2, ..., X_p (régression linéaire multiple), chargées de permettre la prévision linéaire de Y.

Chapitre 5 : Une présentation très élémentaire d'une Analyse en Composantes Principales (ACP) proposée sur un exemple est présenté dans ce chapitre.

Chapitre 6 : Dans ce chapitre, le polycopié est enrichi d'une trentaine d'exercices soigneusement sélectionnées et tous intégralement corrigés qui permettront, aux lecteurs, d'assimiler les connaissances développées dans les chapitres précédents.

Bibliographie : Une liste de références est présentée pour les lecteurs désireux enrichir leurs connaissances dans le sujet.

Annexe : Cette annexe contient cinq tables statistiques pour la lecture des valeurs critiques, à savoir : La table de la loi normale, de la loi de Student, de la loi de $Khi - Deux$, de la loi de Fisher-Snédécor et celle des valeurs critiques du test d'justement Kolmogorov-Smirnov.

Il est à noter que certains démonstrations et notions purement statistiques (Biais, convergence,...) ont été omises volontairement afin d'éviter toute ambiguïté de compréhension des étudiants et de ne pas s'éloigner de l'objectif fixé qui consiste en mises en œuvre des techniques statistiques présenter et n'ont pas leurs développement.

Chapitre 1

Rappels : Statistiques descriptives & Probabilités

Introduction

L'objectif du présent chapitre est de rappeler quelques notions de base des statistiques descriptives ainsi que de la théorie des probabilités. Plus précisément, dans un premier temps nous allons présenter à travers des exemples numérique les principales caractéristiques descriptives d'une série statistique quantitative (nous se limitons au cas de variables continues) à savoir : les présentations graphiques (Histogramme, polygone des fréquences cumulées,...), les paramètres de position (moyenne, médiane, les quartiles, le mode,...) et les paramètres de dispersions (variance, écart-type,...). Dans un deuxième temps, nous focalisons sur les notions de bases de la théorie des probabilités (variable aléatoire, densité, fonction de répartition, espérance mathématique, etc.). On conclut le chapitre par la présentation de quelques lois usuelles.

1.1 Rappels sur les statistiques descriptives

Supposons qu'on dispose d'une série statistique d'une taille n résumée comme suit :

X	n_i	X_i	$N_i \nearrow$	$F_i \nearrow$	$N_i \searrow$
$[a_0\ ,\ a_1]$	n_1	$\frac{a_0+a_1}{2}$	$N_1 = 0$	$F_1 = N_1/n$	n
$[a_1\ ,\ a_2]$	n_2	$\frac{a_1+a_2}{2}$	$N_2 = 0 + n_1$	$F_2 = N_2/n$	$n - n_1$
$[a_2\ ,\ a_3]$	n_3	$\frac{a_2+a_3}{2}$	$N_3 = 0 + n_1 + n_2$	$F_3 = N_3/n$	$n - n_1 - n_2$
$\vdots$					
$[a_{i-1}\ ,\ a_i]$	n_i	$\frac{a_{i-1}+a_i}{2}$	$N_i = 0 + n_1 + ... + n_i$	$F_i = N_i/n$	$n - n_1 - ... - n_i$
$\vdots$					
$[a_{m-1}\ ,\ a_m]$	n_m	$\frac{a_{m-1}+a_m}{2}$	$N_m = 0 + n_1 + ... + n_{m-1}$	$F_m = N_m/n$	$n - n_1 - ... - n_{m-1}$
Σ	n	$-$	n	1	0

alors, la moyenne, la variance et l'écart-type de l'échantillon sont définis respectivement par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{m} n_i X_i; \tag{1.1}$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{m} n_i \left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{m} n_i X_i^2 - \left(\overline{X}\right)^2 = \overline{X^2} - \left(\overline{X}\right)^2. \tag{1.2}$$

$$\textit{Écart-type}(X) = \sigma(X) = \sqrt{Var(X)}. \tag{1.3}$$

Les différents quartiles de la série en question peuvent être quantifiés à l'aide des formules suivantes :

$$Q_1 = a_i + (a_{i+1} - a_i)\left(\frac{n/4 - N_i}{N_{i+1} - N_i}\right) = a_i + (a_{i+1} - a_i)\left(\frac{1/4 - F_i}{F_{i+1} - F_i}\right); \tag{1.4}$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient l'élément $X_{n/4}$.

$$Q_2 = Me = a_i + (a_{i+1} - a_i)\left(\frac{n/2 - N_i}{N_{i+1} - N_i}\right) = a_i + (a_{i+1} - a_i)\left(\frac{1/2 - F_i}{F_{i+1} - F_i}\right); \tag{1.5}$$

tel que a_i et a_{i+1} sont les bornes de la classe médiane.

$$Q_3 = a_i + (a_{i+1} - a_i)\left(\frac{n * 3/4 - N_i}{N_{i+1} - N_i}\right) = a_i + (a_{i+1} - a_i)\left(\frac{3/4 - F_i}{F_{i+1} - F_i}\right); \tag{1.6}$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient $X_{n*3/4}$.
Une notions plus générale est bien que le fractal d'ordre p $(0 \leq p \leq 1)$, définis par :

$$Q_p = a_i + (a_{i+1} - a_i)\left(\frac{n * p - N_i}{N_{i+1} - N_i}\right) = a_i + (a_{i+1} - a_i)\left(\frac{p - F_i}{F_{i+1} - F_i}\right); \tag{1.7}$$

tel que a_i et a_{i+1} sont les bornes de la classe qui contient X_{n*p}.
Le mode de la série peut être quantifié en utilisant la formule suivante.

$$M_o = a_i + (a_{i+1} - a_i)\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right); \tag{1.8}$$

tel que : $\Delta_1 = n_i - n_{i-1}$ et $\Delta_2 = n_{i+1} - n_i$ et a_i et a_{i+1} sont les bornes de la classe modale.

Afin de mettre en évidence les différentes notions introduites ci-dessus, nous allons présenter trois exemples numériques.

Exemple 1 *Soit le tableau statistique suivant :*

X	n_i	X_i	$N_i \nearrow$	$N_i \searrow$	$n_i * X_i$	X_i^2	$n_i * X_i^2$
[144 , 148]	*3*	*146*	*0*	*200*	*438*	*21316*	*63948*
[148 , 152]	*7*	*150*	*3*	*197*	*1050*	*22500*	*157500*
[152 , 156]	*23*	*154*	*10*	*190*	*3542*	*23716*	*545468*
[156 , 160]	*32*	*158*	*33*	*167*	*5056*	*24964*	*798848*
[160 , 164]	*48*	*162*	*65*	*135*	*7776*	*26244*	*1259712*
[164 , 168]	*41*	*166*	*113*	*87*	*6806*	*27556*	*1129796*
[168 , 172]	*24*	*170*	*154*	*46*	*4080*	*28900*	*693600*
[172 , 176]	*15*	*174*	*178*	*22*	*2610*	*30276*	*454140*
[176 , 180]	*6*	*178*	*193*	*7*	*1068*	*31684*	*190104*
[180 , 184]	*1*	*182*	*199*	*1*	*182*	*33124*	*33124*
Σ	*200*	-	*200*	*0*	*32608*	-	*5326240*

Les caractéristiques de cette série sont :

1. La moyenne de la série statistique :

$$\begin{aligned}\overline{X} &= \frac{1}{n}\sum_{i=1}^{10} n_i X_i = \frac{1}{200}\left(3*146+7*150+...+1*182\right)\\ &= \frac{1}{200}*32608 = 163.0400\end{aligned}$$

2. La variance et l'écart-type de la série statistique :

$$\begin{aligned}Var(X) &= \overline{X^2}-(\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{10} n_i X_i^2 - (\overline{X})^2\\ &= \left(\frac{1}{200}*5326240\right) - (163.04)^2 = 49.1584.\end{aligned}$$

   et $\sigma(X) = \sqrt{Var(X)} = \sqrt{49.1584} = 7.0113$.

3. Le premier quartile, la médiane et le troisième quartile sont :
   On a la classe qui contient $X_{n/4}$ est $[156\ ,\ 160[$, alors :
   $Q_1 = a_4 + (a_5 - a_4)\left(\frac{200/4-N_4}{N_5-N_4}\right) = 156 + (160-156)\left(\frac{200/4-33}{65-33}\right) = 158.1250.$
   On a la classe médiane est $[160\ ,\ 164[$, alors :
   $Me = a_5 + (a_6 - a_5)\left(\frac{200/2-N_5}{N_6-N_5}\right) = 162.9167.$
   On a la classe qui contient $X_{n*3/4}$ est $[164\ ,\ 168[$, alors :

   $Q_3 = a_6 + (a_7 - a_6)\left(\frac{200*3/4-N_6}{N_7-N_6}\right) = 167.6098.$

4. Le mode : on constate que la classe modale est $[160\ ,\ 164[$, alors :
   $M_o = a_4 + (a_5 - a_4)\left(\frac{n_5-n_4}{(n_5-n_4)+(n_5-n_6)}\right) = 162.7826.$

5. L'histogramme et la courbe des effectifs cumulés croissant et décroissant sont présentés dans la figure 1.1.

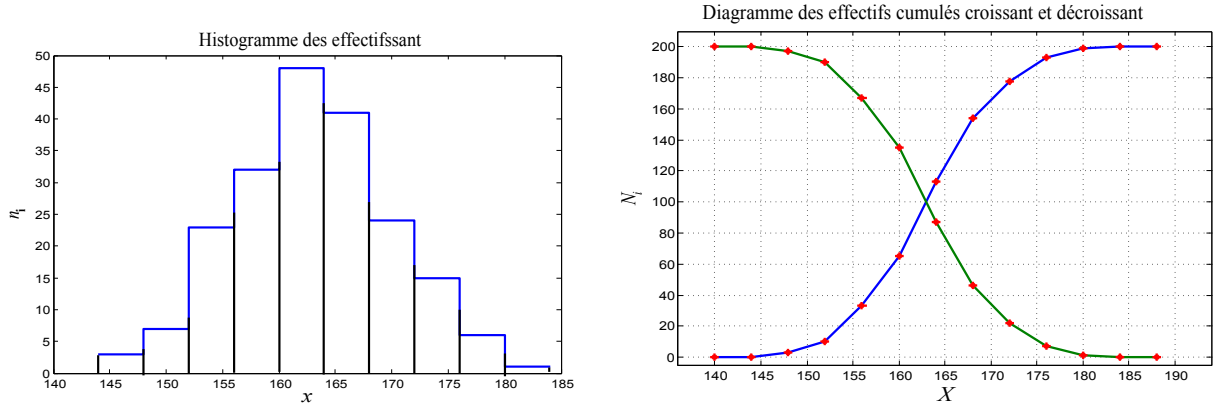


FIGURE 1.1: *Histogramme des effectifs et la courbe des effectifs cumulées*

Exemple 2 *Soit le tableau statistiques suivant :*

X	n_i	X_i	$N_i \nearrow$	$N_i \searrow$	$n_i * X_i$	X_i^2	$n_i * X_i^2$
[30 , 40]	11	35	0	250	385	1225	13475
[40 , 50]	26	45	11	239	1170	2025	52650
[50 , 60]	63	55	37	213	3465	3025	190575
[60 , 70]	81	65	100	150	5265	4225	342225
[70 , 80]	35	75	181	69	2625	5625	196875
[80 , 90]	21	85	216	34	1785	7225	151725
[90 , 100]	13	95	237	13	1235	9025	117325
$\sum$	250	-	250	0	15930	-	1064850

1. La moyenne de cette série est :
   $\overline{X} = \frac{1}{n}\sum_{i=1}^{7} n_i X_i = \frac{1}{250} * 15930 = 63.7200,$
2. La variance de cette série est :
   $Var(X) = \overline{X^2} - (\overline{X})^2 = \left(\frac{1}{250} * 1064850\right) - (63.7200)^2 = 199.1616,$
   et son écart-type est :
   $\sigma(X) = \sqrt{Var(X)} = \sqrt{199.1616} = 14.1125.$
3. La médiane, premier quartile et le troisième quartile :
   On a la classe qui contient $X_{n/4}$ est [50 , 60[, alors :
   $Q_1 = a_3 + (a_4 - a_3)\left(\frac{250/4 - N_3}{N_4 - N_3}\right) = 50 + (60 - 50)\left(\frac{250/4 - 37}{100 - 37}\right) = 54.0476.$
   On a la classe médiane est [60 , 70[, alors :
   $Me = a_4 + (a_5 - a_4)\left(\frac{250/2 - N_4}{N_5 - N_4}\right) = 63.0864.$
   On a la classe qui contient $X_{n*3/4}$ est [70 , 80[, alors :
   $Q_3 = a_5 + (a_6 - a_5)\left(\frac{250*3/4 - N_5}{N_6 - N_5}\right) = 71.8571.$
4. Le mode : on a la classe modale est [60 , 70[, alors :
   $Mo = a_4 + (a_6 - a_5)\left(\frac{n_4 - n_3}{(n_4 - n_3) + (n_4 - n_5)}\right) = 62.8125.$

On peut également déterminer les différents quartiles et le mode graphiquement. En effet, à partir du polygone des effectifs cumulés on détermine les différents quartiles et à partir de l'histogramme on peut déterminer le mode (voir l'exemple illustratif présenté dans la figure 1.2).

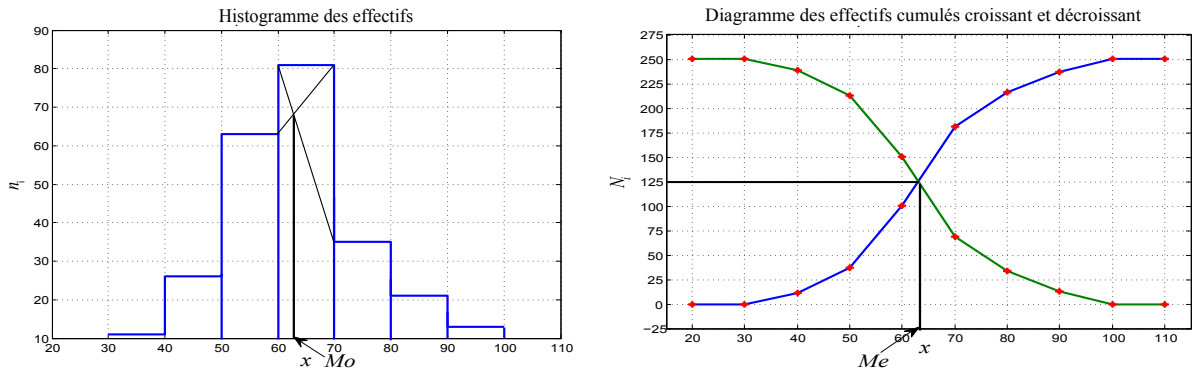


FIGURE 1.2: *Détermination graphique du mode (l'histogramme) et de la médiane (fréquences cumulées)*

Exemple 3 *(Correction de la classe)Dans les deux exemples précédents on a constaté que la taille des classes est une constante (c'est-à-dire $c_i = a_{i+1} - a_i = c, \forall i = \overline{0, m-1}$). Cependant, dans la pratique ce n'est pas toujours le cas, dans cette situation certaines caractéristique ne peuvent être déterminé qu'après la correction des classes(l'exemple du mode). Le présent exemple illustre la situation en question ainsi que la correction des classes.*

X	n_i	u_i	$\alpha = n_i/u_i$
$[0\ ,\ 4[$	*4*	*4*	*1.00*
$[4\ ,\ 10[$	*20*	*6*	*3.33*
$[10\ ,\ 20[$	*14*	*10*	*1.40*
$[20\ ,\ 40[$	*2*	*20*	*0.10*

Les caractéristiques de cette série sont résumées dans le tableau suivant :

Caractéristiques	$\overline{X}$	$Var(X)$	Mo
Valeur	10.450	39.448	7.281
Caractéristiques	Q_1	Me	Q_3
Valeur	5.800	8.800	14.286

La présentation graphique (histogramme des effectifs) de cette série ne se fait qu'après la correction des classes, ainsi on aura la figure 1.3.

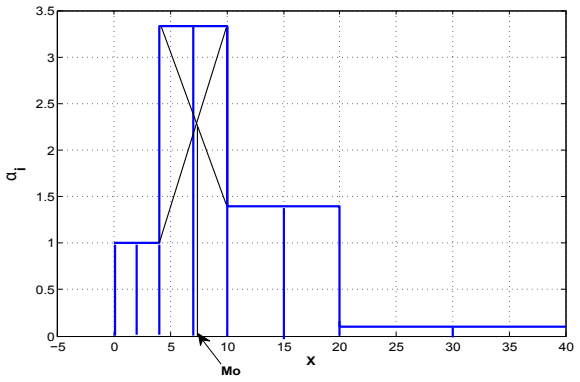


FIGURE 1.3: *Histogramme des effectifs corrigées et détermination du mode*

1.2 Caractérisation d'une variable aléatoire

Le calcul des probabilités s'occupe d'épreuves aléatoires et de phénomènes aléatoires c'est-à-dire d'expériences ou de phénomènes naturels qui, dans des conditions déterminées et stables, ne mènent pas toujours à la même issue. On observe, cependant, une certaine régularité statistique. L'étude de cette régularité fait l'objet d'une théorie mathématique. Dans cette section nous nous limitons à la présentation des outils nécessaires aux applications statistiques qui seront traitées dans les chapitres ultérieurs.

1.2.1 Le concept de variables aléatoires

Une variable aléatoire ($v.a.$) est un nombre réel associe au résultat d'une épreuve, donc un nombre aléatoire. Si l'épreuve est répétée, ce nombre change en général.

Exemple 4 *La taille d'un individu extrait au hasard d'une population ou encore le nombre de "faces" dans une série de 10 jets d'une monnaie.*

1.2.2 La distribution d'une variable aléatoire

1.2.2.1 Cas d'une variable aléatoire discrète

D'une manière générale, si une variable aléatoire discrète X peut prendre les valeurs $x_1,\ x_2, ..., x_n$ avec des probabilités respectives $p_1,\ p_2, ..., p_n$, nous dirons que X a pour *distribution de probabilité* l'ensemble des couples :

$$(x_1, p_1),\ (x_2, p_2), ...(x_n, p_n)$$

Notons qu'une distribution de probabilité a les propriétés suivantes :

- $p_i \geq 0$ pour tout i ;
- $\sum_i p_i = 1$.

On peut représenter graphiquement une distribution de probabilité d'une variable aléatoire discrète à l'aide d'un diagramme en bâtons ou en colonnes (comme si c'était une distribution de fréquences dans les statistiques descriptives).

1.2.2.2 Cas d'une variable aléatoire continue

Pour calculer les probabilités afférentes à une variable continue on utilise sa *fonction de densité*, c'est-à-dire une fonction qui permet de calculer la probabilité que X soit dans un intervalle $[a, b]$. Plus précisément, la densité de probabilité (ou densité) de X est une fonction f_X telle que :

1. $f_X(x) \geq 0$ pour tout x,
2. $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
3. $P(a \leq X \leq b) = \int_a^b f_X(x)dx$.

S'il n'y a pas de possibilité de confusion, on utilisera le symbole f à la place de f_X.

1.2.3 La fonction de répartition (distribution cumulative)

En général, pour exprimer toutes les probabilités associées à une variable aléatoire discrète ou continue il suffit de déterminer les probabilités des intervalles de la forme $I =]\infty, x]$, où x est un nombre réel. L'outil fondamental qui exprime ces probabilités est *la fonction de répartition*. La *fonction de répartition* d'une variable aléatoire X est la fonction

$$\begin{aligned} F_X(x) &= P(X < x) \\ &= \text{probabilité de l'événement "que } X \text{ soit plus petit à } x\text{"}. \end{aligned}$$

Cette fonction est définie pour tout x réel et prend des valeurs entre 0 et 1 ($F_X(x) \in [0, 1]$). S'il n'y a pas de confusion, on utilise le symbole F à la place de F_X.

En général, une fonction de distribution cumulative quelconque a les propriétés suivantes :

- elle est non décroissante ;
- elle prend des valeurs entre 0 et 1 ;
- elle tend vers 0 si x tend vers $-\infty$ et vers 1 si x tend vers $+\infty$.

- La fonction de répartition d'une variable continue est une fonction continue tandis que la fonction de distribution cumulative d'une variable discrète est une fonction discontinue en escalier.
- Entre la distribution de probabilité (x_1, p_1), $(x_2, p_2), ...$ et la fonction de répartition F d'une variable discrète il y a la relation suivante :

$$F(x) = \sum_{x_i \leq x} p_i.$$

- Entre la fonction de densité f et la fonction de répartition F d'une variable continue, y a les relations suivantes :

$$F(x) = \int_{-\infty}^{x} f(t)dt; \tag{1.9}$$

$$f(x) = \frac{d}{dx}F(x) \text{ si } F \text{ est dérivable en } x. \tag{1.10}$$

Remarque 1.1 *Pour une variable aléatoire continue* $P(X \leq x) = P(X < x)$.

1.2.4 Espérance (moyenne) d'une variable aléatoire

Souvent, il suffit d'avoir quelques nombres caractérisant la distribution au lieu de la distribution complète. Les mesures les plus fréquemment utilisées sont la moyenne (ou espérance), la variance, l'écart-type et les fractals.

Soit X une variable aléatoire discrète avec une distribution (x_i, p_i), $i = 1, 2, ..., n$. Alors, l'espérance mathématique (ou espérance) de X est définie par :

$$\mu(x) = x_1p_1 + x_2p_2 + ... = \sum_i x_ip_i. \tag{1.11}$$

On utilise aussi le symbole $E(X)$ à la place de $\mu(X)$.

Si X une variable aléatoire continue ayant une densité f alors l'espérance de X est :

$$\mu(x) = \int_{-\infty}^{\infty} xf(x)dx. \tag{1.12}$$

On utilise aussi le symbole $E(X)$ à la place de $\mu(X)$.

Propriété 1.1

1. *Une propriété de grande importance est que l'espérance est une application linéaire. Soient* X *et* Y *deux variables aléatoires et* a *et* b *deux constantes, alors*

$$E(aX + bY) = aE(X) + bE(Y); \tag{1.13}$$

2. *Une autre propriété très utile, concerne l'espérance d'une transformation d'une variable aléatoire. Soit* g *une fonction réelle quelconque et* $Y = g(X)$ *une transformation de* X. *Alors l'espérance de la variable* Y *est donnée par :*

$$E(Y) = E(g(X)) = \sum_i g(x_i)P(X = x_i), \quad \textit{dans le cas discret} \tag{1.14}$$

$$E(Y) = E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx, \quad \textit{dans le cas continue.} \tag{1.15}$$

1.2.5 La variance et l'écart-type d'une variable aléatoire

L'espérance d'une variable donne une idée de la valeur moyenne de cette variable mais ne prend pas en compte d'autres aspects importants. Par exemple, les variables

- X_1 avec distribution uniforme dans l'intervalle $[-1, 1]$,
- X_2 avec distribution uniforme dans l'intervalle $[-1000, 1000]$,

ont toutes les deux une espérance égale à 0 mais avec une variabilité très différente. Pour mesurer cet aspect on utilise la variance ou l'écart-type.
Soit X une variable aléatoire, la *variance* (de population) de X est définie par :

$$\sigma^2(X) = E\left([X - E(X)]^2\right), \tag{1.16}$$

c'est-à-dire,

$$\sigma^2(X) = \sum_i [X_i - E(X)]^2 P(X = x_i), \quad \text{Si } X \text{ discret}; \tag{1.17}$$

$$\sigma^2(X) = \int_{-\infty}^{\infty} [X_i - E(X)]^2 f(x)dx, \quad \text{Si } X \text{ continue}. \tag{1.18}$$

L'écart-type (de population) de X est défini par

$$\sigma(X) = \sqrt{\sigma^2(X)}. \tag{1.19}$$

1.2.6 Fractals d'une variable aléatoire

Le *quantile* (d'ordre) α $(0 < \alpha < 1)$ d'une variable aléatoire continue X ayant une fonction de répartition F est le nombre q_α tel que :

$$F(q_\alpha) = \alpha, \qquad \text{c'est-à-dire,} \qquad q_\alpha = F^{-1}(\alpha). \tag{1.20}$$

Ainsi, on définit des percentiles, des quartiles et des déciles de population. Pour une variable discrète on procède comme dans le cas d'une distribution des fréquences cumulées en statistiques descriptives (voir chapitre 1).

1.3 Quelques lois de probabilités usuelles

Cette partie définit brièvement les modèles de distributions uni-variées les plus fréquemment utilisés comme descriptions approximatives de distributions réelles(en statistique). Comme ces modèles dépendent de paramètres qui doivent être déterminés à l'aide des données que l'on souhaite décrire on les appels des modèles paramétriques.

1.3.1 La distribution de Gauss (Normale)

On dit que la variable aléatoire X à (ou suit) une distribution normale centrée et réduite ou une distribution de Gauss centrée et réduite qu'on note $X \rightsquigarrow N(0, 1)$ si elle a pour densité

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{1.21}$$

Le graphique de f est une courbe "en cloche" (voir figure 1.4).

Si X a une distribution $N(0,1)$ on obtient $E(X) = 0$ et $var(X) = 1$.
La fonction de répartition de X est :

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt. \tag{1.22}$$

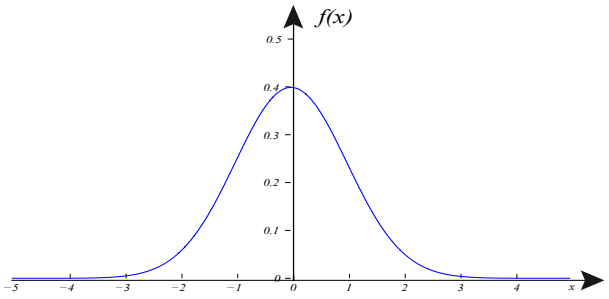


FIGURE 1.4: *Distribution de Gauss centrée et réduite*

Pour déterminer les valeurs de $F(x)$, on se réfère à des tables numériques (voir Tables de la distribution de Gauss) ou on utilise des programmes d'intégration numérique.

Si $X \rightsquigarrow N(0,1)$ alors, la variable aléatoire $Y = \sigma X + \mu$ a une distribution de Gauss de moyenne μ et de variance σ^2, notée $N(\mu, \sigma^2)$ et sa densité est donnée par :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/2\sigma^2}. \tag{1.23}$$

La transformation précédente ($Y = \sigma X + \mu$) effectuée en sens inverse permet de passer d'une variable aléatoire Y de distribution $N(\mu, \sigma^2)$ à la variable centrée et réduite

$$X = \frac{Y - \mu}{\sigma}, \tag{1.24}$$

qui suit une distribution $N(0,1)$. Cette transformation permet de calculer des probabilités relatives à la variable Y à l'aide de la fonction de répartition et des tables de $N(0,1)$.

L'un des principaux résultats obtenus sur la distribution Normale est le théorème central limite résumé comme suit :

Théorème 1.1 *(Théorème Limite Centrale (TCL))*
Supposons que $X_1, ..., X_n$ soient i.i.d. (indépendantes et identiquement distribuées) selon une distribution F_X inconnue, telle que que $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$. Alors, si $n \to \infty$,

$$\frac{\left(\sum_{i=1}^{n} X_i/n\right) - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1). \tag{1.25}$$

Ce théorème peut être interpréter comme suit : La distribution de la moyenne arithmétique centrée et réduite est donc approximativement Gaussienne $N(0,1)$, indépendamment de la distribution F_X, pourvu que n soit suffisamment élevé. La distribution de la moyenne arithmétique est approximativement normale de moyenne μ et variance σ^2/n et ces paramètres peuvent être estimés. Malheureusement, il n'y a pas en général une règle simple pour déterminer la valeur minimale de n pour que l'approximation soit bonne. Cette valeur dépend de la forme de F_X. Mais généralement, dans la pratique, on se contente de $n \geq 30$.

1.3.2 La distribution χ^2 $(Khi - Deux)$

Soient $X_1, ..., X_n$, n variables aléatoires indépendantes et identiquement distribuées $(i.i.d)$ selon une distribution normale standard. On dit que la variable aléatoire

$$Z = X_1^2 + X_2^2 + ... + X_n^2,$$

à une *distribution* χ^2 à n *degrés de liberté* notée χ_n^2. La densité de cette distribution est

$$f(z) = \frac{z^{(\frac{n}{2}-1)}}{2^{n/2}\Gamma(n/2)} e^{-z/2}, \qquad z \geq 0, \tag{1.26}$$

avec $\Gamma(.)$ indique la fonction Γ (Gamma), définie par

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx, \qquad p > 0, \tag{1.27}$$

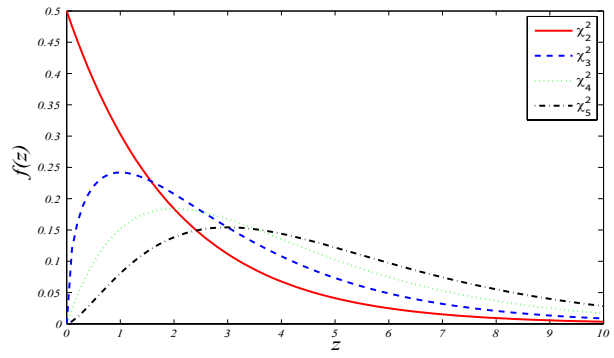


FIGURE 1.5: *Distribution de χ_2 pour différentes degré de liberté.*

La fonction de répartition est généralement calculée à l'aide d'un programme informatique ou de "tables de la distribution χ^2" (voir Tables). La moyenne et la variance de la distribution χ^2 sont $E(Z) = n$ et $\sigma^2(Z) = 2n$.

Remarque 1.2 *Soit Z_1 et Z_2 deux variables aléatoires de distribution χ^2 de degré liberté n et m respectivement, alors la variable aléatoire $Z = Z_1 + Z_2$ est aussi une variable aléatoire d'une distribution de χ^2 de degré liberté $n + m$ $(Z \rightsquigarrow \chi^2_{(n+m)})$.*

1.3.3 La distribution Student (t)

Supposons que X_0, X_1..., X_n, n variables aléatoires indépendantes et identiquement distribuées selon une distribution normale standard. On dit que la variable aléatoire

$$T = \frac{X_0}{\sqrt{\frac{1}{n}(X_1^2 + ... + X_n^2)}} = \frac{X_0}{\sqrt{Z/n}} \quad (avec\ Z \rightsquigarrow \chi_n^2) \tag{1.28}$$

à une distribution t (ou distribution de Student) à n degrés de liberté notée t_n. La densité de cette distribution est

$$f(t) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{n\pi}} (1 + t^2/n)^{-(n+1)/2}. \tag{1.29}$$

La fonction de distribution cumulative est généralement calculée à l'aide d'un programme informatique ou de "tables de la distribution t" (voir Tables). La moyenne et la variance de la distribution t sont $E(T) = 0$ et $\sigma^2(T) = n/(n-2)$, pour $n > 2$.

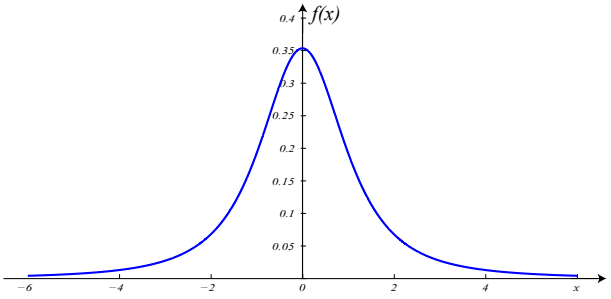


FIGURE 1.6: *Distribution de Student.*

1.3.4 La distribution Fisher (Fisher-Snedecor) F

Soit $X_1, ..., X_{n+m}$, $n+m$ variables aléatoires indépendantes qui suivent une distribution normale centrée et réduite. On dit que la variable aléatoire

$$Y = \frac{\frac{1}{n}(X_1^2 + ... + X_n^2)}{\frac{1}{m}(X_{n+1}^2 + ... + X_{n+m}^2))} = \frac{Z_1/n}{Z_2/m} \quad (Z_1 \rightsquigarrow \chi_n^2 \text{ et } Z_2 \rightsquigarrow \chi_m^2) \tag{1.30}$$

a une distribution F avec n degrés de liberté au numérateur et m degrés de liberté au dénominateur notée $F_{(n,m)}$ ou de degrés de libertés (n, m). La densité de cette distribution est

$$f(y) = \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} n^{n/2} m^{m/2} y^{(n/2)-1} (m + ny)^{-(n+m)/2}, \quad y \geq 0. \tag{1.31}$$

Les calculs concernant la distribution F sont généralement effectués à l'aide d'un programme d'ordinateur ou de "tables de la distribution F" (voir Tables). La moyenne de la distribution F est $E(Y) = \frac{m}{m-2}$ pour $m > 2$ et sa variance $\sigma^2(Y) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$, pour $m > 4$.

Chapitre 2

Théorie statistique de l'estimation : Estimation ponctuelle & par intervalle

Introduction

L'un des problèmes les plus importants en statistique, d'une manière générale, et en Biostatistique, en particulier, est le suivant : On désire étudier un caractère d'une population qui suit une loi $f(x, \theta)$ (de paramètre inconnu θ) dont on connaît sa forme sachant qu'il suffit de déterminer une valeur de θ pour que f soit entièrement déterminée. Cela, se fait par l'estimation qui consiste à donner des valeurs approchées au paramètre θ de la population à l'aide d'un échantillon de n observations issues de cette population.

Dans ce chapitre, pour bien comprendre la démarche à suivre et la mise en œuvre des techniques existantes dans la littérature pour l'estimation du paramètre θ, nous allons introduire d'abord les notions et les définitions liées à ce domaine. Ensuite, nous présentons les techniques d'estimation paramétrique les plus usitées dans la pratique, qu'on peut regrouper en deux familles, à savoir : Estimation ponctuelle (la méthode empirique, la méthode des moments et la méthode du maximum de vraisemblance) et estimation par intervalle de confiance.

2.1 Définitions et notions de base

Définition 2.1 *(Échantillon i.i.d)*
Un échantillon de taille n (ou n-échantillon) indépendant identiquement distribué (i.i.d.) de loi F est constitué de n variables aléatoire v.a indépendantes qui suivent une même loi F. Aussi un échantillon peut être définir comme un sous-ensemble de la population.

Définition 2.2 *(Paramètre)*
Un paramètre est une caractéristique de la population, c'est-à-dire c'est la valeur inconnue d'une population à quantifier à partir d'un échantillon, par exemple : la moyenne (μ), pourcentage (p), écart-type (σ), la variance (σ^2),...

Définition 2.3 *(Modèle statistique)*
On appelle " modèle statistique " ou " la structure statistique " la donnée $(E, \mathcal{A}, \{P_\theta; \theta \in \Theta\})$, où $(E, \mathcal{A})$ est l'espace des réalisations de la v.a, et Θ est l'ensemble des paramètres, où $\theta \in \Theta$, P_θ est la loi de probabilité de la (v.a) lorsque le paramètre vaut θ.

Définition 2.4 *(Statistique)*

1. *On appelle statistique T toute fonction mesurable d'une v.a $(X_1, ..., X_n)$.*
2. *Une statistique est une caractéristique de l'échantillon $\left(\overline{X}, S_n^2, f_n, etc\right)$.*

Exemple 5 *Soit*

$$E = \mathbb{R}^+,\ \mathcal{A} = \mathcal{B}\left(\mathbb{R}^+\right),\ \Theta = \mathbb{R}^+,\ P_\theta = \exp(\lambda).$$

qui est un modèle exponentielle. Soit aussi, $X_1, ..., X_n$ un n-échantillon avec X_i sont i.i.d. alors $T = \sum_{i=1}^{i=n} X_i$ est une statistique qui correspond à la moyenne de l'échantillon.

Rappelons que le but est de construire à partir des données observées sur un échantillon une nouvelle variable notée $\hat{\theta}$ qui est une valeur approchée de θ, cette variable aléatoire appelée estimateur de θ, dont la définition précise est la suivante :

Définition 2.5 *(Estimateur)*
On appelle estimateur d'un paramètre inconnue θ de la population, toute fonction statistique $T(x_1, ..., x_n)$ de l'échantillon utilisée pour trouver une valeur estimative (approcher) de θ. C'est aussi une variable aléatoire noté θ_n, T_n, T ou encore $\hat{\theta}$.

Remarque 2.1 *Vue que l'estimateur est une variable aléatoire, dans certaines situations, on s'intéresse au calcul de ses caractéristique tels l'espérance et la variance et voir même à déterminer sa distribution.*

2.2 Estimateur empirique

A partir d'un échantillon $(X_1, ..., X_n)$ de X, nous définirons la loi de probabilité empirique P_n tel que : $P_n = \frac{1}{n}\sum_{i=1}^{i=n}\delta_{X_i}$ où δ_{X_i} est la masse de Dirac (fonction indicatrice) au point X_i. Cette loi de probabilité admet une fonction de répartition empirique, notée F_n :

$$F_n(x) = \frac{\text{nombre de } x_i \text{ inférieur ou égale à } x}{n} = P_n(]-\infty, x]). \tag{2.1}$$

Définition 2.6 *(moyenne empirique)*
La moyenne empirique d'un n-échantillon (i.i.d) est la moyenne de la loi empirique notée :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i,$$

cette v.a. admet espérance et variance.

Définition 2.7 *(variance empirique) La variance empirique de la loi empirique est :*

$$\begin{cases} S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, & \text{si } \mu \text{ est connue ;} \\ S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2, & \text{si } \mu \text{ est inconnue.} \end{cases}$$

Remarque 2.2 *La valeur moyenne de la variance empirique n'est pas exactement égale à la variance théorique σ^2 (estimateur avec biais), c'est pourquoi on introduit la variance empirique corrigée définie par :*

$$S_{n,c}^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(X_i - \overline{X}_n\right)^2.$$

2.3 Estimateur ponctuelle : méthode des moments

Comme son nom l'indique, cette technique se base sur les moments, pour cela il est important de rappeler la définition des moments qui sont répartie en deux familles : moments théoriques et moments empiriques.

Définition 2.8 *(Les moments théoriques)*
Soit X une variable aléatoire de densité $f(x,\theta)$, $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Le moment théorique d'ordre r se calcule comme suit

$$\mu_r = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x,\theta) dx. \tag{2.2}$$

Définition 2.9 *(Les moments empiriques)*
Soit $X_1, X_2, ..., X_n$ un n-échantillon issu de la variable aléatoire X ayant une densité $f(x,\theta)$, $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Le moment empirique de d'ordre r de la v.a. X se calcule comme suit

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r. \tag{2.3}$$

La méthode des moments est un outil d'estimation intuitif qui a commencé à la naissance des statistiques. Elle consiste à estimer les paramètres recherchés en égalisant entre les moments théoriques (qui dépendent de ces paramètres) et les moments empiriques. L'égalisation se justifie par la loi des grands nombres qui implique que l'on peut "approcher" une espérance mathématique par une moyenne empirique. On est donc amené à résoudre un système d'équations, pour cela il faut alors prendre garde à pouvoir identifier tous les paramètres c'est-à-dire que le nombre d'équation dans le système doit être égale aux nombres de paramètres à estimer.

Les moments empiriques servent à estimer un ou plusieurs paramètres de la loi d'intérêt. En effet, estimer θ, consiste à résoudre le système d'équations.

$$\mu_r = m_r, \quad r = 1, 2, ..., k. \tag{2.4}$$

Exemple 6 *Pour mettre en évidence la technique des moments ci-dessus soit les deux exemples suivant.*

1. *Estimation du paramètres d'une loi de Poisson $\mathcal{P}(\lambda)$.*

   On observe un échantillon $(X_1, ..., X_n)$ de variables aléatoires i.i.d issu d'une loi de Poisson (voir annexe B) de paramètre λ. Supposons qu'on désire estimer le paramètre λ par la méthode des moments. Dans ce cas, on se contente de l'utilisation du moment d'ordre 1, le fait qu'il n'y a qu'un seul paramètre à estimer (la dimension de θ est d'ordre 1). Par conséquent, l'estimateur de λ par la méthode des moments sera $\hat{\lambda}$ qu'on obtiendra par l'égalisation de μ_1 avec m_1 :

   $$\mu_1 = m_1 \Rightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

   le fait que $\mu_1 = E(X) = \lambda$ et $m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$.

   On remarque sur cet exemple que la statistique T_n n'est autre que la moyenne empirique de l'échantillon.

2. *Estimation des paramètres d'une loi normale $N(\mu, \sigma^2)$*
   La loi normale contient deux paramètres à savoir : le paramètre μ et le paramètre σ (σ^2). Donc, pour estimer ces paramètres, par la méthode des moments, il nous faut deux équations. Pour cela, il faut réaliser une égalisation entre les deux premiers moments théoriques et empiriques :

$$\begin{cases} \mu_1 &= m_1, \\ \mu_2 &= m_2, \end{cases}$$

On a d'une part $\mu_1 = \mu$ et $\mu_2 = \sigma^2 + \mu^2$ et d'autre part $m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, alors la détermination des estimateurs de μ et σ^2 consiste à résoudre le système suivant :

$$\begin{cases} \mu &= \frac{1}{n}\sum_{i=1}^{n} X_i, \\ \sigma^2 + \mu^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{cases}$$

La résolution de ce dernier système nous fournit les estimateurs suivants :
$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *et* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.

2.4 Estimateur ponctuelle : Estimateur du Maximum de Vraisemblance (EMV)

La méthode du maximum de vraisemblance est une méthode statistique d'estimation est couramment utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné, elle a été développée entre 1912 et 1922 par le statisticien *Ronald Fisher*.

Soit X une variable aléatoire de loi quelconque dépendant d'un paramètre θ que l'on veut estimer, et soit $x_1, x_2, ..., x_n$ une réalisation de l'échantillon théorique $X_1, X_2, ..., X_n$.

Le maximum de vraisemblance consiste à déterminer la valeur de θ qui rend l'échantillon observé le plus vraisemblable. On va donc chercher la valeur de θ qui donne à l'échantillon qui a été observé la plus grande probabilité possible de l'avoir obtenu si X est une variable aléatoire discrète, ou la plus grande densité si X est une variable aléatoire continue, c'est-à-dire la plus grande vraisemblance, et à choisir cette valeur pour estimer θ. Cette méthode permet de construire des estimateurs qui ont, en général de bonnes propriétés, dès que la taille de l'échantillon n est grande.

2.4.1 Principe de la méthode du maximum de vraisemblance

Avant d'énoncer la forme ou le procédé à suivre pour l'obtention d'un estimateur par la méthode de la vraisemblance, nous allons définir la fonction vraisemblance.

Définition 2.10 *(fonction de vraisemblance) Soit $x_1, x_2, ..., x_n$ la réalisation de l'échantillon théorique $X_1, X_2, ..., X_n$ issu de la variable aléatoire X qui admet $f(x, \theta)$ comme densité de probabilité.*

On appelle la vraisemblance notée $L(x_1, x_2, ..., x_n; \theta)$ la variable aléatoire (ou une fonction de la variable θ) définie comme suit

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta). \tag{2.5}$$

Remarque 2.3 *Si X est une variable aléatoire discrète qui admet comme loi de probabilité P_θ alors, la formule (2.5) sera réécrite comme suit :*

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} P_\theta(X = x_i). \tag{2.6}$$

Exemple 7 *Soit x_1, x_2,...,x_n, la réalisation d'un échantillon issu d'une loi P_θ.*

1. *Cas P_θ est une loi de poisson (cas discret) :*

$$L(X) = P_\lambda(X).$$
$$On\ a : P_\lambda(X = x_i) = \left(\frac{\lambda^{x_i}}{x_i!}\right) e^{-\lambda}; \quad i = 1, ..., n.$$
$$alors : L(x_1, ..., x_n, \lambda) = \prod_{i=1}^{i=n} \left(\frac{\lambda^{x_i}}{x_i!}\right) e^{-\lambda},$$
$$= \left(\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}\right) e^{-n\lambda};$$

2. *Cas P_θ est une loi normale (cas continue) :*

$$L(X) = N(m, \sigma).$$
$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2}; \quad i = 1, ..., n,$$

cette loi dépend de m et σ, sa fonction de vraisemblance est :

$$L(x_1, ..., x_n, m, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2}$$
$$= (2\pi)^{\frac{-n}{2}} \sigma^{-n} e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2}$$

Définition 2.11 *(Estimateur du Maximum de Vraisemblance (EMV))*
On appelle estimateur du maximum de vraisemblance de θ, tout élément $\hat{\theta}$ de Θ maximisant la vraisemblance, c'est-à-dire

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(x_1, ..., x_n, \theta).$$

La recherche de l'EMV peut se faire, sous certaines conditions, d'une manière directe et cela par la recherche de l'extremum max de L, c'est-à-dire quand la fonction L est deux fois dérivable par rapport à θ, il suffit de vérifier les points suivants :

$$\begin{cases} 1) & \frac{\partial L}{\partial \theta}(x_1, ..., x_n, \theta) = 0, \quad \text{la solution de cette équation fournie } \hat{\theta}. \\ 2) & \frac{\partial^2 L}{\partial \theta^2}(x_1, ..., x_n, \theta) < 0, \quad \text{pour assurer que } \hat{\theta} \text{ est un extremum maximal.} \end{cases} \tag{2.7}$$

Par définition la vraisemblance se calcule à partir d'un produit de n éléments, cependant, on préfère remplacer le problème définie dans (2.7) par un problème équivalent moins complexe. Puisque, la fonction logarithme Népérien (ln) est strictement croissante et $L(x_1, ..., x_n, \theta)$ et $\ln L(x_1, ..., x_n, \theta)$ atteignent leurs maximums pour la même valeur de θ, il serait souvent plus aisé de résoudre le système d'équations équivalent suivant :

$$\begin{cases} 1) & \frac{\partial \ln L(x_1,..,x_n,\theta)}{\partial \theta} = 0, \\ 2) & \frac{\partial^2 \ln L(x_1,..,x_n,\theta)}{\partial \theta^2} < 0, \end{cases} \tag{2.8}$$

que le système (2.7).

Il est à noter que dans le cas où θ est un paramètre de dimension k (c'est-à-dire $\theta = (\theta_1, ..., \theta_k)$), on résout un système de k équations dont ces dernières sont obtenues en dérivant $L(x_1, .., x_n, \theta)$ par rapport à chacune des composantes de θ.

2.4.2 Quelques exemples d'application

Dans cette section nous allons présenter quelques exemples d'application de la méthode du maximum de vraisemblance.

2.4.2.1 Modèle de Bernoulli

Supposons que les observations $x_1, x_2, ..., x_n$ soit indépendantes et identiquement distribuées suivant une loi de Bernoulli de paramètre $\theta(\theta \in]0, 1[)$ c'est-à-dire $P(X = x) = \theta^x(1-\theta)^{1-x}(x \in \{0,\ 1\})$. La vraisemblance du modèle est définie par

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} P(X = x_i),\ \theta \in]0, 1[\tag{2.9}$$

$$= \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i},\ \theta \in]0, 1[\tag{2.10}$$

$$= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i},\ \theta \in]0, 1[. \tag{2.11}$$

La fonction log-vraisemblance est

$$\log L(x_1, x_2, ..., x_n; \theta) = \sum_{i=1}^{n} x_i (\log \theta) + n - \sum_{i=1}^{n} x_i (\log(1-\theta)), \tag{2.12}$$

on a alors,

$$\frac{\partial \log L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0 \Longleftrightarrow \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta} = 0 \Longrightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} (\text{ extremum }),$$

et

$$\frac{\partial^2 \log L(x_1, x_2, ..., x_n; \theta)}{\partial \theta^2} = \frac{-\sum_{i=1}^{n} x_i}{\theta^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-\theta)^2} < 0 (\text{ maximum }).$$

Donc l'estimateur du maximum de vraisemblance de θ est donné par $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$.

2.4.2.2 Modèle Gaussien

Supposons que les observations $x_1, x_2, ..., x_n$ soit indépendantes et identiquement distribuées suivant une loi normale de moyenne θ inconnue et de variance σ^2 connue.
La vraisemblance du modèle est définie par

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\right). \tag{2.13}$$

La fonction log-vraisemblance est

$$\log L(x_1, x_2, ..., x_n; \theta) = -n \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2. \tag{2.14}$$

L'estimateur du maximum de vraisemblance de θ est donné par

$$\frac{\partial \log L(x_1, x_2, ..., x_n; \theta)}{\partial \theta} = 0 \Longleftrightarrow \sum_{i=1}^{n} (x_i - \theta) = 0 \Longrightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n},$$

et

$$\frac{\partial^2 \log L(x_1, x_2, ..., x_n; \theta)}{\partial \theta^2} = \frac{-n}{\sigma^2} < 0.$$

2.4.3 Comparaison : EMV ou méthode des moments ?

La méthode des moments est conceptuellement plus simple que la méthode du maximum de vraisemblance, mais les estimateurs ainsi produits n'ont pas les bonnes propriétés asymptotiques.

Pour la plupart des lois de probabilité usuelles, l'estimateur du maximum de vraisemblance est défini de façon unique, et se calcule explicitement. Sur le plan théorique, il présente de nombreux avantages. Sous des hypothèses vérifiées par de nombreux modèles courants, on démontre qu'il est asymptotiquement sans biais et convergent. On démontre de plus que sa variance est minimale. La méthode du maximum de vraisemblance est donc théoriquement la meilleure des méthodes d'estimation.

Mais toutefois, dans certains cas comme celui de la loi Gamma, le calcul de la fonction vraisemblance peut poser des problèmes tandis que l'estimation des moments est très facilement accessible. Finalement, la méthode des moments est une méthode applicable juste dans le cas ou on ne peut pas calculer la fonction de vraisemblance.

2.5 Estimation ponctuelle : Méthode des Moindres Carrées

Un modèle très général pour des situations où l'on mesure une quantité inconnue μ est

$$X_i = \mu + \epsilon_i, \qquad i = 1, ..., n, \tag{2.15}$$

où X_i est la i-ème mesure, μ la quantité inconnue et ϵ_i est l'erreur de mesure qui a influencé la i-ème observation. Souvent, on assume que les erreurs sont *i.i.d.* selon une certaine distribution F mais on ne souhaite pas décrire cette distribution de façon plus précise à l'aide d'un modèle de distribution. Dans une modélisation de ce type, la méthode MC (Moindres Carrés) définit un estimateur $\hat{\mu}$ de μ de la façon suivante : On définit la quantité

$$Q(\mu) = \sum_{i=1}^{n} \epsilon_i^2 = (x_1 - \mu)^2 + ... + (x_n - \mu)^2, \tag{2.16}$$

qui représente la somme des erreurs quadratique et on cherche la valeur $\hat{\mu}$ de μ telle que la somme $Q(\mu)$ soit minimale. Cette valeur vérifie la relation

$$\frac{dQ(\mu)}{\mu} = -2(x_1 - \mu) - ... - 2(x_n - \mu) = 0, \tag{2.17}$$

et donc

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.18}$$

La relation (2.15) est l'exemple le plus simple d'une multitude de situations. Aux prochains chapitres nous allons considérer des relations plus complexes (les modèles de régression).

2.6 Distribution d'un estimateur d'une moyenne et d'une variance

En général, une simple estimation ne suffit pas : il est nécessaire de connaître son degré d'imprécision. L'outil fondamental pour évaluer un estimateur et le comparer à d'autres, est bien que sa distribution d'échantillonnage. Par exemple, à égalité entre différents aspects, on préférera l'estimateur avec la plus petite variance. Cette section s'occupe du calcul de la distribution de quelques estimateurs usuels (moyenne, variance). Si on suppose que la distribution des données peut être décrite par un modèle paramétrique, on aura une approche paramétrique au calcul de la distribution de l'estimateur ; autrement on parlera d'une approche non-paramétrique.

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ, de variance σ^2, et un échantillon $X_1,\ X_2, ...,\ X_n$ de X de taille n.

1. Pour chaque échantillonnage on peut calculer la moyenne observée du caractère

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

   On démontre que $E(\overline{X}) = \mu$ (un estimateur sans biais de μ) et $Var(\overline{X}) = \frac{\sigma^2}{n}$.

2. Si la moyenne μ est **connue**, alors on considère la variance d'échantillon

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \mu^2.$$

3. Si la moyenne μ est **inconnue** alors dans ce cas, l'estimateur sans biais de la variance de l'échantillon est définie comme suit :

$$\hat{\sigma_c}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Les lois de probabilité de l'estimateur d'une moyenne et d'une variance pour certaine situations peuvent être résumées dans ce qui suit :

1. Cas d'un petit échantillon gaussien $n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$
   - Si σ est connu alors, $U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ suit la loi normale $N(0,1)$.
   - Si σ est inconnu, alors, $T = \frac{\overline{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté.
2. Cas d'un grand échantillon ($n > 30$) et X de loi quelconque :
   - Dans ce cas, $U = \frac{\overline{X}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ suit approximativement la loi normale $N(0,1)$.
3. Cas d'un échantillon gaussien (X de loi normale $N(\mu, \sigma^2)$) :
   - Si μ est connue alors la variable $Y^2 = n\frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi-Deux$ à n degrés de liberté.
   - Si μ est inconnue alors la variable $Y^2 = (n-1)\frac{\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi-Deux$ à $n-1$ degrés de liberté.

2.7 Estimation par intervalle : Intervalle de confiance

L'estimation est un utile très important pour avoir des informations sur certaines paramètres que l'on cherche à estimer. Cependant, les méthodes d'estimation qui nous avons exposé jusqu'à maintenant ne nous donnent aucune information concernant la précision des estimateurs construits.

Donc on a besoin de déterminer d'autres techniques d'estimation d'un paramètre qui nous permettent de déterminer un ensemble de valeurs, sous forme d'un ensemble de valeurs fort probable qu'elles soient la vraie valeur du paramètre à estimer. Dans cette section nous allons présenter une autre technique d'estimation d'un paramètre unidimensionnelle qui nous permis de déterminer un ensemble de valeurs, sous forme d'un intervalle, fort probable qu'elle soit la vraie valeur du paramètre à estimer. Cette nouvelle approche est souvent préférée dans la pratique car elle introduit la notion d'incertitude.

2.7.1 Principe général

Dans cette approche on cherche à déterminer l'intervalle $[a, b]$ (généralement, centré sur la valeur numérique estimée du paramètre inconnu) contenant la vraie valeur du paramètre inconnu θ ($\theta \in \Theta$ tel que $\Theta \subseteq \mathbb{R}$) avec une probabilité $1 - \alpha$ fixée a priori.

$$P\left(a \leq \theta \leq b\right) = 1 - \alpha, \tag{2.19}$$

où la probabilité $1 - \alpha$ permet de s'adapter aux exigences de l'application.

L'intervalle $[a, b]$ est appelé *intervalle de confiance* et $1 - \alpha$ est le niveau de confiance. Une estimation par intervalle de confiance sera d'autant meilleure que l'intervalle sera petit pour un niveau de confiance grand.

En générale, pour construire un intervalle de confiance, on dispose deux cas :

1. Le cas des grands échantillons : Les intervalles de confiance peuvent être obtenus, par le TCL.
2. Le cas des petits échantillons : les intervalles de confiance peuvent être obtenus, par calcule de la loi exact.

D'après ce qui précède, on constate que la donnée de départ, outre l'échantillon, sera la connaissance de la loi de probabilité du paramètre à estimer. Comme il n'existe pas de résolution générale de ce problème, nous nous abordons que les cas les plus fréquents dans la pratique (estimation d'une proportion, d'une moyenne et d'une variance).

2.7.2 Estimation d'une proportion par IC

Soit une population dont les individus possèdent un caractère A avec une probabilité p (loi de Bernoulli 0/1). On cherche à déterminer cette probabilité inconnue en prélevant un échantillon de taille n dans cette population.

A partir de l'échantillon prélevé, on constate que x parmi les n individus possèdent le caractère A. Que peut-on en déduire ? C'est-à-dire, la proportion $f_n = X/n$ est une approximation (estimation) de la vraie valeur de p, mais avec quelle confiance (précision) ?

f_n peuvent être obtenue par l'une des techniques présenter dans les sections précédentes (MV, moments,...).

L'approximation $f_n = X/n$, f_n est une $v.a$ construite par la somme de n $v.a$ X tel que $X \in \{0, 1\}$ $i.i.d$. Donc, d'après le **TCL**, f_n est une $v.a$ dont la loi de tend vers une loi normale

de moyenne p et d'écart-type $\sqrt{\frac{p(1-p)}{n}}$. Bien évidement, que cette approximation est valable uniquement si la taille de l'échantillon est suffisamment grande (c'est-à-dire $n > 30$ en pratique). Construisons l'intervalle de confiance autour de p sous la forme :

$$P(|f_n - p| < \epsilon) = 1 - \alpha, \tag{2.20}$$

où α est le risque (a priori, on construit un intervalle symétrique) et f_n est une réalisation d'une $v.a$ de la loi $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$. Donc, on peut par la normalisation et le centrage obtenir une nouvelle $v.a$ U tel que :

$$U = \frac{f_n - p}{\sqrt{\frac{p(1-p)}{n}}} \rightsquigarrow \mathcal{N}(0,1).$$

On montre qu'une bonne approximation de l'IC de niveau $1-\alpha$ de p, fondé sur la valeur observée f_n, est donnée par l'intervalle ci-dessous :

$$IC_{1-\alpha}(p) = \left[f_n - q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\frac{f_n(1-f_n)}{n}}; f_n + q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\frac{f_n(1-f_n)}{n}} \right], \tag{2.21}$$

où $q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)}$ est le fractile (quantile) d'ordre $1-\alpha/2$ de la loi normale centré et réduite.

2.7.3 IC pour la moyenne d'une loi normale

De même, qu'une proportion, il existe une expression approchée pour l'IC de niveau $1-\alpha$ d'une moyenne μ, où l'intervalle est fondé sur la valeur observée $\hat{\mu}$ obtenue après une expérience portant sur n individus. Mais cet intervalle dépend de l'écart-type σ à savoir : σ est connu ou σ est inconnu. Dans ce qui suit nous allons traiter ces deux situations.

2.7.3.1 Cas l'écart-type σ est connu

A partir de l'estimateur $\hat{\mu}$, qui distribué selon une loi normale $\mathcal{N}(\mu, \sigma^2/n)$, on détermine la valeur $q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)}$ du fractile d'ordre $1-\alpha/2$ de la loi normale centré et réduite, tel que :

$$P\left(-q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \right) = 1 - \alpha, \tag{2.22}$$

ce qui conduit à la construction d'un intervalle symétrique centré sur $\hat{\mu}$ de forme :

$$\hat{\mu} - q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}},$$

c'est-à-dire,

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}}; \hat{\mu} + q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}} \right]. \tag{2.23}$$

2.7.3.2 Cas l'écart-type σ est inconnu

La statistique utilisée dans le cas précédant

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \left(\frac{\hat{\mu} - \mu}{\sigma} \right), \quad (\text{ avec } \hat{\mu} = \overline{X}), \tag{2.24}$$

ne peut pas convenir dans ce cas (σ est inconnu), le fait qu'elle intervienne le paramètre inconnu σ. A cet effet, on est contraint a remplacé σ par son estimateur ponctuelle, basé sur la variance modifiée :

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

On substituant cette dernière dans la statistique (2.24) on aura :

$$T = \frac{\hat{\mu} - \mu}{\hat{\sigma}_c/\sqrt{n}}.$$

Par définition (d'une loi de Student), on déduit que la statistique T suit une loi de Student à $n-1$ degré de liberté. A cet effet, la détermination d'un IC de la moyenne, dans cette situation, consiste à déterminer la valeur du fractile d'ordre $1-\alpha/2$ d'une loi de Student de degré de liberté $n-1$ $\left(q_{\frac{\alpha}{2}}^{t_{n-1}}\right)$, c'est-à-dire :

$$P\left(-q_{\frac{\alpha}{2}}^{t_{n-1}} < \frac{\hat{\mu} - \mu}{\hat{\sigma}_c/\sqrt{n}} < q_{\frac{\alpha}{2}}^{t_{n-1}}\right) = 1-\alpha, \tag{2.25}$$

d'où la déduction de l'intervalle bilatéral symétrique, au tour de $\hat{\mu}$, de forme :

$$\hat{\mu} - q_{\frac{\alpha}{2}}^{t_{n-1}}\frac{\hat{\sigma}_c}{\sqrt{n}} < \mu < \hat{\mu} + q_{\frac{\alpha}{2}}^{t_{n-1}}\frac{\hat{\sigma}_c}{\sqrt{n}},$$

ou encore :

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - q_{\frac{\alpha}{2}}^{t_{n-1}}\frac{\hat{\sigma}_c}{\sqrt{n}}; \hat{\mu} + q_{\frac{\alpha}{2}}^{t_{n-1}}\frac{\hat{\sigma}_c}{\sqrt{n}}\right]. \tag{2.26}$$

2.7.4 IC pour la variance d'une loi normale

Comme dans le cas de la moyenne, l'intervalle de confiance de la variance σ^2 nécessite une information préalable sur le paramètre $\hat{\mu}$. En effet, deux constructions d'un intervalle de confiance de σ^2 peut être envisagé, et cela selon μ est connu ou non.

2.7.4.1 Cas la moyenne μ est connue

L'estimateur sans biais, convergent et efficace, d'une variance dans le cas où la vraie moyenne est connue est donné par :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Donc pour déterminer un intervalle de confiance de la variance dans ce cas il suffit de quantifier les deux paramètres a_α et b_α, telle que :

$$P\left(a_\alpha < n\frac{\hat{\sigma}^2}{\sigma^2} < b_\alpha\right) = 1-\alpha, \tag{2.27}$$

ou encore,

$$n\frac{\hat{\sigma}^2}{b_\alpha} < \sigma^2 < n\frac{\hat{\sigma}^2}{a_\alpha}. \tag{2.28}$$

Rappelons que

$$n\frac{\hat{\sigma}^2}{\sigma^2} \rightsquigarrow \chi_n^2,$$

alors, les valeurs de a_α et b_α serons déterminées de la manière suivante :

$$P\left(\chi_n^2 < a_\alpha\right) = \alpha_1 \text{ et } P\left(\chi_n^2 > b_\alpha\right) = \alpha_2, \tag{2.29}$$

avec la seule contrainte $\alpha_1 + \alpha_2 = \alpha$.

2.7.4.2 Cas de la moyenne μ est inconnue

L'estimateur sans biais, convergent de σ^2, qu'il faut retenir dans cette situation est bien que :

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2,$$

donc la loi est connue

$$(n-1)\frac{\hat{\sigma}_c^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2.$$

On peut donc déterminer les valeurs de a_α et b_α telle sort :

$$P\left(a_\alpha < (n-1)\frac{\hat{\sigma}_c^2}{\sigma^2} < b_\alpha\right) = 1 - \alpha, \tag{2.30}$$

ce qui conduit à l'IC suivant :

$$(n-1)\frac{\hat{\sigma}_c^2}{b_\alpha} < \sigma^2 < (n-1)\frac{\hat{\sigma}_c^2}{a_\alpha}. \tag{2.31}$$

Les valeurs de a_α et b_α sont en fait déterminer par :

$$P\left\{\chi_{n-1}^2 < a_\alpha\right\} = \alpha_1 \text{ et } P\left\{\chi_{n-1}^2 > b_\alpha\right\} = \alpha_2, \tag{2.32}$$

avec la seule contrainte $\alpha_1 + \alpha_2 = \alpha$.

Remarque 2.4 *Dans la pratique la majorité des cas, l'hors de la construction d'un intervalle de confiance de la variance on pose* $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

Chapitre 3

Introduction à la théorie de test d'hypothèses ”

Introduction

Les tests statistiques sont des méthodes de la statistique inférentielle qui, comme l'estimation, permettent d'analyser des données obtenues par tirages au hasard. Ils consistent à généraliser les propriétés constatées sur des observations à la population d'où ces dernières sont extraites, et à répondre à des questions concernant par exemple la nature d'une loi de probabilité, la valeur d'un paramètre ou l'indépendance de deux variables aléatoires.

Il serait important de chercher à présenter en détail l'ensemble des tests statistiques, mais la littérature est très abondante sur le sujet. Pour cela, dans ce chapitre nous allons se limiter aux tests classiques les plus simples et les plus usuels dans la pratique. En effet, Les tests présentés, sont concernés les tests à un seul échantillon, d'adéquation d'une loi, de comparaison de deux échantillons et enfin l'analyse de la variance à un seule facteur et analyse de la variance à deux facteurs.

3.1 Tests de conformité pour une moyenne

Considérons un caractère quantitatif représenté par une variable aléatoire X d'espérance mathématique μ, d'écart-type σ, et un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n de X. La moyenne et la variance corrigée d'échantillon sont données respectivement par :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } \hat{\sigma}_c^2 = \frac{n}{n-1}\hat{\sigma}^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

3.1.1 Cas d'un petit échantillon gaussien ($n \leq 30$ et X de loi normale $N(\mu, \sigma^2)$)

Dans ce test deux cas sont envisageable. En effet, on peut distinguer le cas où l'écart-type est une quantité bien connue et le cas où l'écart-type n'est connue qu'approximativement à travers son estimateur.

3.1.1.1 Cas σ connu

Il s'agit de faire un choix entre plusieurs hypothèses possibles sur μ sans disposer d'informations suffisantes pour que ce choix soit sûr. On met en avant deux hypothèses privilégiées : l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Par exemple, on testera

$$H_0 : \text{"}\mu = \mu_0''\text{ contre } H_1 : \text{"}\mu \neq \mu_0'',$$

avec μ_0 fixé arbitrairement. On veut savoir si l'on doit rejeter H_0 ou pas.
La résolution du présent problème consiste, en résumé, à réaliser les étapes suivantes :

1. Utilise une variable aléatoire dont on connaît la loi de probabilité lorsque H_0 est vraie. Par exemple, on prend $U = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, en raison que lorsque H_0 est vraie, $U = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ suit la loi $N(0,1)$, et cela le fait que l'échantillon est issue d'une variable aléatoire d'une loi normale $X \rightsquigarrow N(\mu, \sigma^2)$.
2. Fixe une valeur $\alpha \in]0,1[$. En général, on prend α (le risque) petit, le plus souvent

$$\alpha \in \{0.10, 0.05,\ 0.01,\ 0.01,\ 0.001\}.$$

3. Quantifier un réel u_α, tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$. Ce réel u_α peut être extrait de la table de la loi normale centrée et réduite (voir annexe A).
4. Comparer la moyenne empirique $\overline{X}$ de l'échantillon à la moyenne théorique $\mu = \mu_0$, sachant que l'hypothèse H_0 signifiera que les différences observées sont seulement dues aux fluctuations d'échantillonnage (i.e. ne sont pas significatives). En fin, on decide ce qui suit :
   - On ne rejettera pas H_0 si les différences observées ne sont pas significatives, c'est-à-dire si U est "petite", ce que l'on peut formuler par $-u_\alpha < U < u_\alpha$, ou encore $|U| < u_\alpha$.
   - On rejettera H_0 si les différences observées sont significatives, ce que l'on peut formuler par $U < -u_\alpha$ ou $U > u_\alpha$, c'est-à-dire $|U| > u_\alpha$. Par construction de u_α, on a $P(U > u_\alpha) = P(U < -u_\alpha) = \frac{\alpha}{2}$, soit encore $P(|U| > u_\alpha) = \alpha$ i.e. $P(U \notin]-u_\alpha, u_\alpha[) = \alpha$.

   En pratique, on calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$ et on décide
   - de rejeter H_0 si $u \notin]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité forte de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ n'est pas conforme à la valeur théorique μ_0 mais on ne pourra pas donner de valeurs acceptable de μ ;
   - de ne pas rejeter H_0 si $u \in]-\mu_\alpha, \mu_\alpha[$, car si H_0 était vraie, l'événement $U \notin]-\mu_\alpha, \mu_\alpha[$ aurait une probabilité faible de se réaliser ; on pourra dire que la valeur observée $\overline{X}$ est conforme à la valeur théorique μ_0 et que la valeur μ_0 ne peut être rejeter.
     Attention : d'autres valeurs $\mu_0',\ \mu_0'', \ldots$ peuvent également convenir.

Erreurs de décision Il est à noter que, l'aspect aléatoire de l'échantillon (observations) peut nous faussé la décision finale (rejeter ou non l'hypothèse H_0). On effet, lorsque on rejette H_0 alors que H_0 est vraie, on commet une erreur. On a donc une probabilité α (car lorsque H_0 est vraie, on a $P(U \notin]-\mu_\alpha, \mu_\alpha[) = \alpha$) de se tromper : α est appelée ***erreur de première espèce***.

Une autre situation où on peut commettre une erreur de décision est bien que celle lorsque on ne rejette pas H_0 alors que H_0 est fausse. Dans ce cas, on a une probabilité β de se tromper : β est appelée ***erreur de deuxième espèce***. Cette probabilité est difficilement calculable car dans la plupart des temps, on ne connaît pas la loi de U lorsque H_0 est fausse. La valeur $1-\beta$ est appelée la ***puissance du test***.

Finalement, ces déférentes situations peuvent être résumées par le schéma suivant :

		Réalité	
		H_0	H_1
Décision	H_0	$1-\alpha$	α
	H_1	β	$1-\beta$

Les différents tests usuels (formulation et décision) correspondant à la présente situation peuvent être résumer comme suit :

Test (bilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu \neq \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :
- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu > \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U \geq u_\alpha) = 1-\alpha$ et on décide que :
- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu < \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$. On détermine u_α, à partir de la table de la loi normale, tel que $P(U < u_\alpha) = 1-\alpha$ et on décide que :
- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

3.1.1.2 Cas σ inconnu

Par définition, on sait que $T = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit la loi de Student à $n-1$ degrés de liberté (voir section 1.3.3). Alors, les différents tests précédents (bilateral et unilatéral) se font comme suit :

Test (bilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu \neq \mu_0''$,
On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α sur la table de Student pour un degré de liberté $n-1$ tel que $P(-t_\alpha < T < t_\alpha) = 1-\alpha$ et on décide que :
- Si $t \in]-t_\alpha, t_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $t \notin]-t_\alpha, t_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu > \mu_0''$,
On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T \geq t_\alpha) = 1-\alpha$ et on décide que :
- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu < \mu_0''$,
On calcule $t = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine t_α tel que $P(T < t_\alpha) = 1-\alpha$ et on décide que :
- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

3.1.2 Cas d'un grand échantillon : $n > 30$

Dans cette situation ($n > 30$), on se basons sur le TCL, on sait que la variable aléatoire $U = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$ suit approximativement une loi normale centrée et réduite ($U \rightsquigarrow N(0,1)$).

Test (bilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu \neq \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$, et on décide que :
- Si $u \in]-u_\alpha, u_\alpha[$, alors on ne peut rejeter H_0 ;
- Si $u \notin]-u_\alpha, u_\alpha[$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu > \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U \geq u_\alpha) = 1-\alpha$ et on décide que :

- Si $u < u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \geq u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\mu = \mu_0''$ **contre** H_1 : $''\mu < \mu_0''$,
On calcule $u = \frac{\overline{X}-\mu_0}{\frac{\hat{\sigma}_c}{\sqrt{n}}}$. On détermine u_α tel que $P(U < u_\alpha) = 1-\alpha$ et on décide que :

- Si $u > -u_\alpha$, alors on ne peut rejeter H_0 ;
- Si $u \leq -u_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

3.2 Test de conformité : pour une variance σ^2

Considérons un caractère quantitatif représenté par une variable aléatoire X de loi normale $N(\mu, \sigma^2)$ et un échantillon $X_1,\ X_2, ...,\ X_n$ de taille n de X. La moyenne de l'échantillon est $\overline{X}$ et sa variance corrigée est $\hat{\sigma}_c^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Par définition la variable aléatoire définie par $Y^2 = \frac{(n-1)\hat{\sigma}_c^2}{\sigma^2}$ suit la loi de $Khi - Deux$ à $n-1$ degrés de liberté (voir section 1.3.2).

Les différents tests simples de conformité de la variance (formulations et décisions) sont résumés dans ce qui suit :

Test (bilatéral) H_0 : $''\sigma^2 = \sigma_0^2''$ **contre** H_1 : $''\sigma^2 \neq \sigma_0^2''$
On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine a_α et b_α tel que $P(Y^2 \geq a_\alpha) = 1 - \frac{\alpha}{2}$ et $P(Y^2 \geq b_\alpha) = \frac{\alpha}{2}$ et ensuite on décide de :

- ne peut rejeter H_0 si $y^2 \in]a_\alpha, b_\alpha[$;
- rejette H_0 avec une probabilité α de se tromper si $y^2 \notin]a_\alpha, b_\alpha[$.

Test (unilatéral) H_0 : $''\sigma^2 = \sigma_0^2''$ **contre** H_1 : $''\sigma^2 > \sigma_0^2''$
On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine b_α tel que $P(Y^2 \geq b_\alpha) = \alpha$ et on décide :

- Si $y^2 < b_\alpha$, de ne peut rejeter H_0 ;
- Si $y^2 \geq b_\alpha$, de rejette H_0 avec une probabilité α de se tromper.

Test (unilatéral) H_0 : $''\sigma^2 = \sigma_0^2''$ **contre** H_1 : $''\sigma^2 < \sigma_0^2''$
On calcule $y^2 = \frac{n-1}{\sigma^2}\hat{\sigma}_c^2$, on détermine a_α tel que $P(Y^2 \geq a_\alpha) = 1-\alpha$ et on décide :

- Si $y^2 > a_\alpha$, de ne peut rejeter H_0 ;
- Si $y^2 \leq a_\alpha$, de rejette H_0 avec une probabilité α de se tromper.

3.3 Tests de conformité d'une distribution : Tests d'ajustement

Dans les paragraphes précédent, nous avons construit des tests portant sur un paramètre réel. Nous souhaitons désormais, à partir de l'observation d'un échantillon de taille n issu d'une loi de fonction de répartition F, tester si $F = F_0$ ou non (où F_0 est une fonction de répartition que l'on se donne). Il s'agit donc de tester que la loi des X_i est une loi donnée F_0. La formulation statistique de ce test peut être présenter comme suite :

$$H_0 : ''F_n(x) = F_0(x)'' \text{ contre } H_1 : ''F_n(x) \neq F_0(x)''.$$

où,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i<x\}}, \tag{3.1}$$

est la fonction de répartition empirique de la variable échantillonnée.

Dans ce document, nous allons présenté les tests les plus usuels qui nous permis de réaliser ce test, à savoir : le test de $Khi-Deux$ et le test $Kolmogorov-Smirnov$.

3.3.1 Test de $Khi-Deux$

Le test du $Khi-Deux$ utilise des propriétés de la loi multi-nomiale. Il permet de juger si une hypothèse concernant la loi de probabilité d'une variable aléatoire est compatible avec la réalisation d'un échantillon ou non.

Dans ce test deux cas sont à distinguer :

- La fonction de répartition F_0 est entièrement spécifiée, et ses paramètres sont connus.
- On connaît seulement la forme de la loi de distribution, et ses paramètres sont estimés à partir d'un échantillon.

Soit $X_1, X_2, ..., X_n$ un n-échantillon issu d'une variable aléatoire X dont on désir tester si la loi de l'échantillon est une loi entièrement spécifiée F_0. Le test de $Khi-Deux$ se base principalement sur les étapes suivantes :

- On partage le domaine D de la variable X en r classes $c_1, c_2, ..., c_r$ (avec le même principe que dans les statistiques descriptives voir Chapitre 1).
- Quantifier les N_i (pratique) : l'effectif de la classe c_i, $i = \overline{1:r}$.
- Pour $i = \overline{1:r}$, calculer p_i (théorique) : la probabilité de se trouver dans la classe c_i. Elle est déduite à partir de la loi de probabilité F_0.
- Pour $i = \overline{1:r}$, calculer $n_i = np_i$ (théorique) : effectif théorique de la classe c_i, $i = \overline{1:r}$.
- Déterminer la valeur de k_n^2 la réalisation de variable aléatoire K_n^2 définie par :

$$K_n^2 = \sum_{i=1}^{r} \frac{(N_i - n_i)^2}{n_i}. \tag{3.2}$$

Sachant que **Pearson** a démontré que la variable aléatoire K_n^2 suit asymptotiquement un $Khi-Deux$ à $(r-1)$ degré de liberté, alors la valeur critique du test sera déterminer à partir de le table de $Khi-Deux$, $\chi^2_{(r-1,\alpha)}$ et la decision sera prise comme suite :

- Si $k_n^2 < \chi^2_{(r-1,\alpha)}$, alors on accepte l'ajustement de la distribution de la variable aléatoire X par la loi choisie F_0.

- Si $k_n^2 \geq \chi^2_{(r-1,\alpha)}$, alors on rejette l'ajustement de la distribution de la variable aléatoire X par la loi choisie F_0.

Lorsque les paramètres de la loi à valider sont estimés à partir de l'échantillon, le degré de liberté de la distribution de $Khi-Deux$ est alors égal à $(r-q-1)$, q étant le nombre de paramètres estimés c'est-à-dire valeur critique du test est égale au fractile $\chi^2_{(r-q-1,\alpha)}$.

L'application du test $Khi-Deux$ doit satisfaire les conditions suivantes :

- Le nombre de classes doit être supérieur ou égal à 7.
- L'effectif théorique de chaque classe doit être supérieur ou égal à 5.
- Les effectifs théoriques des k classes doivent être sensiblement égaux.

3.3.2 Test de Kolmogorov-Smirnov

Dans la section précédente nous avons exposé le test d'ajustement de χ^2, même si ce test est facile à implémenter, nous avons constaté sa mise en œuvre exige certaines conditions à titre d'exemple l'effectifs théoriques doit être au moins égales à 5 dans tous les classes. Une alternative à ce test est l'utilisation du test de Kolmogorov-Smirnov. En effet, le test de Kolmogorov-Smirnov est un test non paramétrique qui peut être utilisé même dans le cas où les effectifs théoriques sont inférieurs à 5, ce qui fait ce test est plus puissant que le précédent.

Contrairement au test d'ajustement de χ^2, qui consiste à vérifier si les effectifs empiriques étaient conformes à ceux d'une distribution théorique, le test de Kolmogorov-Smirnov se base sur les différences en valeur absolue entre les fréquences théoriques et les fréquences empiriques cumulées.

La statistique, notée D_n, utilisée par le test de Kolmogorov-Smirnov est la plus grande différence en valeur absolue entre les fréquences empiriques cumulées (F_n) et les fréquences théoriques cumulées (F_0) :

$$D_n = max|F_n(x) - F_0(x)|, \tag{3.3}$$

où : $F_n(x)$=(nombre d'observations $\leq x$/ la taille de l'échantillon) définie dans (3.1) et $F_0(x)$ la fonction de répartition théorique. Supposons qu'on dispose de n réalisations $x_1,\ x_2, ...,\ x_n$ d'une variable aléatoire X et on désir tester à partir de ces réalisation si la loi de X est une loi F_0 qu'on a fixé préalablement on utilisons le test de Kolmogorov-Smirnov. Alors, pour réaliser ce test on suit les étapes suivantes :

- Classer les observations selon un ordre croissant ;
- Déterminer la fonction des fréquences empiriques cumulées croissantes F_n(x) à partir de ces n observations ;
- Comparer la fonction de répartition empirique $F_n(x)$ avec la fonction de répartition théorique $F_0(x)$ et cela on calculons

$$D_n(x_i) = |F_n(x_i) - F_0(x_i)|, \quad i = \overline{1:n}. \tag{3.4}$$

- Chercher $D_n = \max D_n(x_i)$;
- Fixer un seuil de signification (de risque) α pour déterminer la valeur de $d(\alpha)$ à partir de la table de Kolmogorov-Smirnov (voir annexe).
- Et en fin, la décision sera prise de la manière suivante :
  - ✓ Si $D_n < d(\alpha) \Rightarrow$ on ne rejette pas H_0, c'est-à-dire on admet que la distribution de X est $F_0(x)$.
  - ✓ Si $D_n \geq d(\alpha) \Rightarrow$ on rejette H_0, c'est-à-dire on admet que la distribution de X n'est pas $F_0(x)$.

Remarque 3.1 *Pour des échantillons de petits taille ($n \leq 30$), il existe une table qui donne la valeur critique de D_n. Tandis que pour des échantillons supérieurs à 30, les valeurs critiques du test seront calculées en utilisant l'expression $a/\sqrt{n}$ où $a = 1.63$ pour un seuil de signification de 1%, $a = 1.36$ pour un seuil de signification de 5% et $a = 1.22$ pour un seuil de signification de 10% (n est la taille de l'échantillon).*

3.4 Tests d'homogénéité

Dans les différents tests présenté dans les sections précédentes on n'a considéré qu'un seule échantillon, pour lequel on s'intéresse si l'un de ses caractères (moyenne, variance, distribution) est conforme à une quantité fixée arbitrairement (cette dernière quantité représente généralement une

norme du phénomène étudié). Cependant, dans la pratique, dans certaines situation on dispose de deux populations P_1 et P_2 ou voir même plus de deux populations, dont on étudie un même caractère et on désir comparer les populations quant à ce caractère, et donc à savoir si elles sont homogènes ou non. Dans cette section, nous se limitons au cas de test d'homogénéité de variance et de moyennes de deux populations indépendantes.

3.4.1 Comparaison de deux variances

Soient X et Y deux variables aléatoires indépendantes représentant le même caractère quantitative dans chacune des populations P_1 et P_2. On suppose que X et Y suivent des lois normales respectivement, $N(\mu_1;\sigma_1^2)$ et $N(\mu_2;\sigma_2^2)$.

De P_1, on extrait un échantillon $X_1,\ X_2,...,\ X_{n_1}$ de taille n_1 de X et de P_2, on extrait un échantillon $Y_1,\ Y_2,...,\ Y_{n_2}$ de taille n_2 de Y.

Les moyennes empiriques des deux échantillons sont alors

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i;$$

et leurs variances corrigées sont :

$$\hat{\sigma}_{c,1}^2 = \frac{n_1}{n_1-1}\hat{\sigma}_1^2 \text{ avec } \hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2 - \overline{X}^2,$$

$$\hat{\sigma}_{c,2}^2 = \frac{n_2}{n_2-1}\hat{\sigma}_2^2 \text{ avec } \hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2 - \overline{Y}^2.$$

On veut réaliser le test bilatérale suivant :

$$H_0 :\ "\sigma_1^2 = \sigma_2^2" \text{ contre } H_1 :\ "\sigma_1^2 \neq \sigma_2^2".$$

Les étapes de la réalisation de ce test peuvent être résumées comme suit :

1. On calcule la réalisation $f_c = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$. Si nécessaire, on permute les échantillons de sorte que $f_c \geq 1$ $\left(\text{c'est-à-dire } f_c = \frac{\max\{\sigma_{c,1}^2,\sigma_{c,2}^2\}}{\min\{\sigma_{c,1}^2,\sigma_{c,2}^2\}}\right)$.
2. Sachant que sous l'hypothèse H_0, la statistique (variable aléatoire) $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2}$ suit une loi de Fisher à $(n_1-1;n_2-1)$ degrés de liberté, alors à partir de la table de Fisher on détermine f_α tel que : $P(F \geq f_\alpha) = \frac{\alpha}{2}$(ou encore $P(F \leq f_\alpha) = 1-\frac{\alpha}{2}$).
3. La règle de décision se fait comme suite :
   - si $f_c < f_\alpha$, alors on ne peut rejeter H_0 (H_0 est vraie).
   - si $f_c \geq f_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper.

Avec le même raisonnement on va trouver la zone de non rejet de l'hypothèse nulle dans les tests unilatéral. Les résultats des différents tests sont résumés dans le tableau suivant :

Hypothèse	Zone de non-rejet H_0
$H_0 :\ "\sigma_1^2 = \sigma_2^2\ "$ contre $H_1 :\ "\sigma_1^2 \neq \sigma_2^2\ "$	$[1; f(n_1-1, n_2-1, 1-\frac{\alpha}{2})]$
$H_0 :\ "\sigma_1^2 = \sigma_2^2\ "$ contre $H_1 :\ "\sigma_1^2 > \sigma_2^2\ "$	$[1; f(n_1-1, n_2-1, 1-\alpha)]$
$H_0 :\ "\sigma_1^2 = \sigma_2^2\ "$ contre $H_1 :\ "\sigma_1^2 < \sigma_2^2\ "$	$[1; f(n_2-1, n_1-1, 1-\alpha)]$, avec $f_c = \frac{\hat{\sigma}_{c,2}^2}{\hat{\sigma}_{c,1}^2}$

tel que $f(n,m,1-\alpha)$ est lu dans la table de loi Fisher-Snedecor $(1-\alpha)$ à colonne n, ligne m, de plus on ne rejettera pas H_0 si f_c appartient à la zone de non-rejet de H_0 et on rejettera H_0 sinon.

3.4.2 Comparaison de deux moyennes

Dans cette section, nous allons intéresser à l'homogénéité de deux populations par rapport à la moyenne. Notons que, le test de comparaison de deux moyennes dépend de la distribution des échantillons dont on dispose. Dans le cadre de ce document, nous allons se focalisé sur le cas où les deux échantillons sont de grand taille issues d'une loi quelconque et le cas où les deux échantillons sont gaussien et de petite taille.

3.4.2.1 Cas des grands échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère qualitative étudié dans chaque population. On suppose que X et Y suivent une loi quelconque de moyennes respectives μ_1 et μ_2 et d'écart-types respectifs σ_1 et σ_2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 > 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 > 30$ de Y.

Soit la statistique

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{3.5}$$

et u sa réalisation.

Test (bilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 \neq \mu_2''$,

Sous l'hypothèse H_0, la statistique U définie par (3.5) suit approximativement la loi normale centrée réduite $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1}+\frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que :

$$P(-u_\alpha < U < u_\alpha) = 1 - \alpha,$$

c'est-à-dire

$$P(U < u_\alpha) = 1 - \frac{\alpha}{2},$$

et on décide :

- de ne pas rejeter H_0 si $u \in]-u_\alpha, u_\alpha[$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \notin]-u_\alpha, u_\alpha[$.

Test (unilatéral) de $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1}+\frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que : $P(U \geq u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u < u_\alpha$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \geq u_\alpha$.

Test (unilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, la statistique U suit approximativement la loi normale $N(0, 1)$.

On calcule $u = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{\hat{\sigma}_{c,1}^2}{n_1}+\frac{\hat{\sigma}_{c,2}^2}{n_2}}}$, et on détermine u_α, sur la table de la loi normale, tel que $P(U < u_\alpha) = 1 - \alpha$ et on décide :

- de ne pas rejeter H_0 si $u > -u_\alpha$;
- de rejeter H_0, avec une probabilité α de se tromper, si $u \leq -u_\alpha$.

La démarche et les résultats des trois tests ci-dessus restent valable si on remplace σ_1^2 ou σ_2^2 par leurs estimations $\hat{\sigma}_{c,1}^2$, le fait que $U = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}}$ suit aussi une loi normale centrée réduite (on peut le justifier par le TCL).

3.4.2.2 Cas de petits échantillons

Soient X et Y des variables aléatoires indépendantes représentant le caractère dans chaque population. On suppose que X et Y suivent une loi normal de moyennes respectives μ_1 et μ_2, de variance respectives σ_1^2 et σ_2^2. On extrait un échantillon $X_1,\ X_2, ...,\ X_{n_1}$ de taille $n_1 \leq 30$ de X et un échantillon $Y_1,\ Y_2, ...,\ Y_{n_2}$ de taille $n_2 \leq 30$ de Y.

Test (bilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 \neq \mu_2''$,

Afin de réaliser ce test, nous définissons la statistique suivante :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}. \tag{3.6}$$

Sous l'hypothèse H_0 et l'hypothèse $\sigma_1 = \sigma_2$ la statistique du test définie dans (3.6) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Cependant, dans la pratique on ne sait pas si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ou non. A cet effet, on doit d'abord tester l'égalité des deux variances, $\sigma_1^2 = \sigma_2^2$ (Voir section 3.4.1).

Si cette dernière hypothèse est retenue, alors la valeur commune σ^2 peut être estimer par $\hat{\sigma}_c^2 = \frac{(n_1-1)\sigma_{c,1}^2+(n_2-1)\sigma_{c,2}^2}{n_1+n_2-2}$. Ensuite, on calcule la réalisation de la statistique T, c'est-à-dire $t = \frac{\overline{x}-\overline{y}}{\hat{\sigma}_c \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ et on détermine sur la table de la loi de Student la valeurs critique, t_α, du test tel que : $P(-t_\alpha < T < t_\alpha) = 1 - \alpha$. Finalement, on décide que :

- On ne peut rejeter H_0 si $t \in\ ]-t_\alpha, t_\alpha[$;
- On rejette H_0 si $\notin\ ]-t_\alpha, t_\alpha[$, avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 > \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (3.6) suit approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Ainsi, on détermine t_α sur la table de la loi de Student pour un $n = n_1 + n_2 - 2$ et qui vérifié l'égalité $P(T \geq t_\alpha) = 1 - \alpha$ et on décide :

- Si $t < t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \geq t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la décision.

Test (unilatéral) $H_0 :\ ''\mu_1 = \mu_2''$ **contre** $H_1 :\ ''\mu_1 < \mu_2''$,

Sous l'hypothèse H_0, si $\sigma_1 = \sigma_2$ alors la statistique, T, du test définie dans (3.6) suit encore approximativement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Pour prendre la décision sur le rejet de l'hypothèse H_0, il suffit de déterminer sur la table de Student pour un *ddl* $n = n_1 + n_2 - 2$ la valeur critique t_α tel que $P(T < t_\alpha) = 1 - \alpha$ et on décide :

- Si $t > -t_\alpha$, alors on ne peut rejeter H_0 ;
- Si $t \leq -t_\alpha$, alors on rejette H_0 avec une probabilité α de se tromper dans la decision.

3.5 Test d'indépendance : Test de $Khi - Deux$

La mise en évidence de l'existence d'une liaison entre deux caractères aléatoires a beaucoup d'importance dans toutes les études biologique. Les techniques employées sont différentes suivant

que les variables étudiées sont discrètes ou continues ; elles sont différentes aussi suivant que le type de loi des variables est connu ou non. Nous distinguerons trois cas fondamentaux qui donnent lieu chacun à diverses méthodes : les variables sont toutes les deux discrètes, une seule est continue et les deux le sont. Partant de là, nous allons introduire une méthode, plus générale, qui peut être appliqué dans les trois situations en question qui nous permet de mettre en évidence l'indépendance entre deux caractères (facteurs) aléatoires.

3.5.1 Position du problème

On veut savoir si le temps écoulé depuis la vaccination contre la petite vérole a ou non une influence sur le degré de gravité de la maladie lorsqu'elle apparaît. Les patients sont divisés en trois catégories selon la gravité de leur maladie : légère (L), moyenne (M), ou grave (G) et en trois autres quant à la durée écoulée depuis la vaccination : moins de 10 ans (A), entre 10 et 25 ans (B), plus de 25 ans (C).

Les résultats d'une observation portant sur $n = 1574$ malades sont les suivants :

Degré de gravité Y de la maladie	Durée X écoulée depuis la vaccination			Total
	A	B	C	
G	1	42	230	273
M	6	114	347	467
L	23	301	510	834
Total	30	457	1087	1574

Pour mettre en évidence, l'existence d'une liaison entre la durée écoulée depuis la vaccination et le degré de gravité de la maladie, on choisit de tester les hypothèses nulle et alternative :

H_0 **:** la durée écoulée depuis la vaccination et le degré de gravité de la maladie sont indépendantes",

H_1 **:** la durée écoulée depuis la vaccination et le degré de gravité de la maladie sont liées".

C'est dans ce genre de situations, que le test d'indépendance de $Khi - Deux$ peut intervenir.

3.5.2 Principe du test

De manière générale, le problème du test d'indépendance est posé de la manière suivante : soient X et Y deux variables discrètes (respectivement continues), X à r modalités (respectivement classes) et Y à k modalités (respectivement classes), notées respectivement $i = 1, ..., r$ et $j = 1, ..., k$ et n_{ij} l'effectif observé, dans le tableau croisé, des individus pour lesquels X vaut A_i et Y vaut B_j (voir le tableau 3.2).

X \ Y	B_1	B_2	$\cdots$	B_j	$\cdots$	B_k	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rk}	$n_{r\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet k}$	n

TABLE 3.2: *Présentation des effectifs observés sous forme tableau croisé*

On note $n_{\bullet j}$ le nombre total de ceux pour lesquels Y vaut B_j et qui figure au bas de la $j^{\text{ème}}$ colonne, et $n_{i\bullet}$ le nombre total de ceux pour lesquels X vaut A_i et qui figure à droite de la ligne i.

Pour mettre en évidence ou nie une liaison entre X et Y, on choisit de tester les hypothèses nulle et alternative :

H_0 **:** "X et Y sont indépendantes",

H_1 **:** "X et Y sont liées".

Sous l'hypothèse H_0 d'indépendance de X et Y :

$$P(X = i, Y = j) = P(X = i) \times P(Y = j) \tag{3.7}$$
$$p_{ij} = p_{i\bullet} \times p_{\bullet j}. \tag{3.8}$$

Comme les estimateurs de chacune de ces probabilités à partir du tableau des effectifs du tableau des observations, sont

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \quad \hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} \quad \text{et} \quad \hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}. \tag{3.9}$$

Alors, d'après les formules (3.8) et (3.9) on aura :

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n} \Longrightarrow n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}.$$

Si H_0 est vraie alors les écarts

$$e_{ij} = \hat{p}_{ij} - \hat{p}_{i\bullet} \times \hat{p}_{\bullet j} \quad \left(\text{ou encore } E_{ij} = n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n} \right),$$

ne doivent être dus qu'aux fluctuations d'échantillonnage.

Afin de répondre à notre objectif nous allons définir la statistique des erreurs suivante :

$$K_n^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet} \times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}}.$$

On peut démontrer que la variable K_n^2 suit une loi proche de celle d'un χ^2 à $(r-1)(k-1)$ degrés de liberté, pourvu que les dénominateurs $n_{i\bullet}p_{\bullet j}$ soient tous supérieurs à 5 (si ce n'est pas le cas, on regroupe plusieurs classes de tel sorte que cette condition soit vérifié), alors la règle de décision du test d'indépendance de $Khi - Deux$ sera sous la forme suivante :

- Si $k_n^2 < \chi^2_{((r-1)(k-1),\alpha)}$, alors les deux variables en question sont indépendantes.
- Si $k_n^2 > \chi^2_{((r-1)(k-1),\alpha)}$, alors les deux variables en question sont liées.

où k_n^2 est la réalisation de la statistique K_n^2, et $\chi^2_{((r-1)(k-1),\alpha)}$ le fractile d'ordre $1-\alpha$ d'une loi de $Khi - Deux$ à $(r-1)(k-1)$ ddl

3.5.3 Exemple d'application

Reprenant l'exemple exposé dans la section 3.5.1. Sous l'hypothèse H_0 (les deux variables sont indépendantes) les effectifs n_{ij} devrai être comme suit :

$n_{ij} =$

Y	X		
	A	B	C
G	5.20	79.26	188.53
M	8.90	135.59	322.51
L	15.90	242.15	575.96

Ainsi, les écarts E_{ij} entre les effectifs théoriques et observées sont :

$E_{ij} =$

Y	X		
	A	B	C
G	-4.20	-37.26	41.47
M	-2.90	-21.59	24.49
L	7.10	58.86	-65.96

$\Rightarrow \frac{E_{ij}^2}{\frac{n_{i\bullet} \times n_{\bullet j}}{n}} =$

Y	X		
	A	B	C
G	3.3981	17.5193	9.1206
M	0.9461	3.4379	1.8598
L	3.1742	14.3043	7.5534

On a d'une part, d'après les résultats ci-dessous, la réalisation de la variable K_n^2 vaut $k_n^2 = 61.3137$. D'autre part, sous l'hypothèse H_0, K_n^2 suit une loi du χ^2 à (r -1) (k - 1) = 4 degrés de liberté, ainsi la valeur critique du test pour un risque $\alpha = 5\%$ vaut 11.143 (voir table de la loi de $Khi - Deux$ en annexe).

On constate que la valeur critique du test est inférieur à la valeur observée k_n^2, alors on rejette l'hypothèse d'indépendance de la gravité de la maladie et du délai écoulé depuis la vaccination.

3.6 Analyse de la variance à un facteur (ANOVA 1)

Dans cette section, nous allons intéressé à un cas plus générale pour la comparaison de moyennes et cela lorsque le nombre d'échantillon est supérieur strictement à deux. Plus précisément nous allons intéressé à la technique d'analyse de la variance à un seul facteur qui est la plus adéquate avec la situation.

3.6.1 Position du problème

Supposons que nous ayons 3 forêts contenant un type d'arbre bien déterminé où nous désirons savoir si ces forêts ont une influence sur la hauteur des arbres ou non. À cet effet, nous avons réalisés un recueil de hauteur de six (06) arbres dans chaque forêt, dont les mesures sont rangées dans le tableau suivant.

N°	forêt 1	forêt 2	forêt 3
1	23.3	18.9	22.5
2	24.4	21.1	22.9
3	24.6	21.1	23.7
4	24.9	22.1	24.0
5	25.0	22.5	24.0
6	26.2	23.5	24.5

TABLE 3.3: *Tailles des arbres selon la forêt*

Soit les notions et les notations suivantes :

- Les forêts : Variable qualitative contenant trois modalités, appelée facteur.
- Hauteur des arbres : Réponse, notée X, et μ_i la hauteur moyenne des arbres de la $i^{\text{ème}}$ forêt ($i = \overline{1,3}$).

Répondre à notre objectif consiste à la réalisation du test suivant :

$$H_0 : \text{''}\mu_1 = \mu_2 = \mu_3 = \mu\text{''} \text{ contre } H_1 : \text{''}\exists\ i, j \in \{1, 2, 3\} \text{ tel que } \mu_i \neq \mu_j;\text{''}.$$

Pour réaliser ce test nous pourrons le décomposer en trois sous-tests où nous comparons la hauteur moyenne des arbres deux à deux selon les forêts. Mais afin de contourner le problème d'erreur α gonflé, le fait elle ne réalise qu'une seule comparaison à la fois, nous utilisons la technique statistique connue sous le nom d'analyse de variance (en anglais : Analyse Of Variance (ANOVA)) plutôt que des tests de Student t (voir section 3.4.2) multiples. Remarquez que l'ANOVA peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t.

3.6.2 Analyse de la variance à un seul facteur

L'identification de l'*ANOVA* 1 au sens littéraire peut être résumée dans la définition suivante :

Définition 3.1 *(ANOVA 1)*
L'analyse de la variance à un facteur teste l'effet d'un facteur contrôlé A ayant p modalités (groupes) sur les moyennes d'une variable quantitative X.

Les problèmes concernés par la technique *ANOVA* 1 s'écrivent en générale de la manière suivante :

N°	*groupe* 1	*groupe* 2		*groupe p*
1	$X_{1,1}$	$X_{1,2}$	$\cdots$	$X_{1,p}$
2	$X_{2,1}$	$X_{2,2}$	$\cdots$	$X_{2,p}$
3	$X_{3,1}$	$X_{3,2}$	$\cdots$	$X_{3,p}$
4	$X_{4,1}$	$X_{4,2}$	$\cdots$	$X_{4,p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_j	$X_{n_1,1}$	$X_{n_2,2}$	$\cdots$	$X_{n_p,p}$

et le modèle mathématique leurs associés est donné par :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1, n}, j = \overline{1, p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0, \sigma^2), \tag{3.10}$$

où X_{ij} est la $j^{\text{ième}}$ réalisation de la variable quantitative X dans le $i^{\text{ième}}$ échantillon et ϵ_{ij} sont les erreurs de mesure.

Si on retient ce modèle alors le test à réaliser est défini par :

$$H_0 : \text{''}\mu_1 = \mu_2 = ... = \mu_p = \mu\text{''} \text{ contre } H_1 : \text{''}\exists\ i, j \in \{1, 2, ..., p\} \text{ tel que } \mu_i \neq \mu_j''. \tag{3.11}$$

Dans ce qui suit, nous allons énumérer les étapes de la mise en oeuvre de l'ANOVA 1 qui nous permet de réaliser ce test.

3.6.3 Les étapes de l'*ANOVA* 1

Afin de réaliser le test définie dans (3.11), trois conditions doit être vérifiées préalablement, à savoir :

- Les p échantillons comparés sont indépendants.
- La variable quantitative étudiée suit une loi normale dans les p populations comparées.
- Les p populations comparées ont même variance : *Homogénéité* des variances ou *homoscédasticité*.

Si ces dernières conditions sont vérifiées alors, on peut utiliser la technique *ANOVA* 1 pour réaliser le test (3.11), et pour ce faire nous avons besoin des quantités (statistiques) suivantes :

- La moyenne de toutes les observations : $\overline{X} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} X_{ij}$ avec $n = \sum_{j=1}^{p} n_j$;
- Moyenne de chaque échantillon : $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$, pour $j = \overline{1,p}$;
- Variance de chaque échantillon : $\hat{\sigma}_i^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$, pour $j = \overline{1,p}$;
- La variance de toutes les observations : $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2$ avec $n = \sum_{j=1}^{p} n_j$.

On peut démonter facilement que la variance de toutes les observations est la somme de la variance des moyennes et de la moyenne des variances des p échantillons, c'est-à-dire :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2, \tag{3.12}$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2. \tag{3.13}$$

On multipliant (3.13), par n on obtient :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j - \overline{X})^2}_{SC_{Fac}}, \tag{3.14}$$

où,

SC_{Tot} : est la variation totale qui représente la dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur (variation inter-groupes) qui représente la dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle (variation intra-groupes) qui représente la dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

L'idée la plus naturelle est que le facteur n'a pas d'impact sur le caractère étudié si la variation totale n'est engendrée que par la variation intra-groupes (résiduelle) associée au caractère, c'est-à-dire,

- Si H_0 est vraie, alors la variation SC_{Fac} due au facteur doit être petite par rapport à la variation résiduelle SC_{Res}.
- Par contre, si H_1 est vraie alors la variation SC_{Fac} due au facteur doit être grande par rapport à la quantité SC_{Res}.

Pour comparer ces quantités, Fisher a considéré le rapport des carrés moyens associés au facteur CM_{Fac} et les carrés moyens résiduels CM_{Res}, où

le carré moyen associé au facteur est : $CM_{Fac} = \frac{SC_{Fac}}{p-1}$.

le carré moyen résiduel est : $CM_{Res} = \frac{SC_{Res}}{n-p}$.

Si les 3 conditions d'application d'ANOVA (Indépendance, Normalité et Homogénéité) sont vérifiées et H_0 est vraie, alors

$$F_{obs} = \frac{CM_{Fac}}{SC_{Res}} \rightsquigarrow f_{(p-1,n-p)}.$$

Décision : Pour un seuil de risque donné α les tables de Fisher nous fournissent une valeur critique f_α telle que :

$$P\left(\frac{CM_{Fac}}{SC_{Res}} < f_\alpha\right) = 1 - \alpha,$$

- si $f_{obs} < f_\alpha \Longrightarrow$ on ne peut pas rejeter H_0 (le facteur n'a aucune influence sur le caractère étudier),
- si $f_{obs} \geq f_\alpha \Longrightarrow$ on rejette H_0 (le facteur influe sur le caractère étudié),

avec f_{obs} est la réalisation de la variable (statistique) F_{obs}.

Les résultats d'une ANOVA 1 sont souvent présentés dans un tableau sous la forme suivante :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe (Fac)	SC_{Fac}	$p-1$	CM_{Fac}	$\frac{CM_{Fac}}{CM_{Res}}$	c
Intra-groupe (Rés)	SC_{Res}	$n-p$	CM_{Res}		
Total	SC_{Tot}	$n-1$			

3.6.4 Exemple d'application

Reprenant l'exemple présenter dans la section 3.5.1. Les étapes qu'on doit suivre pour réaliser le test

$$H_0 : \text{ " } \mu_1 = \mu_2 = \mu_3 = \mu \text{ " contre } H_1 : \text{ " } \exists\, i,j \in \{1,2,3\} \text{ tel que } \mu_i \neq \mu_j \text{ ",}$$

à l'aide de la technique ANOVA 1, sont les suivantes :

- Calculer les moyennes des différents échantillons : $\overline{X}_1 = 24.73$, $\overline{X}_2 = 21.53$ et $\overline{X}_3 = 23.60$.
- Calculer la moyenne globale de toutes les observations : $\overline{X} = \frac{1}{n}(n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3) = 23.2889$.
- Compléter le tableau de l'ANOVA à un seul facteur :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}	Ficher c
Inter-groupe	31.5911	2	15.7956	12.02	3.6823
Intra-groupe	19.7067	15	1.3138		
Total	51.2978	17			

- Décision : on constate que $f_{obs} = 12.02 > f_\alpha = 3.6823$ (pour un risque de $\alpha = 5\%$), donc les hauteurs moyennes des arbres sont significativement différentes d'une forêt à une autre. Cela signifié que le facteur forêt influe sur la hauteur des arbres.

3.7 Analyse de la variance à deux facteurs (ANOVA 2)

Cette section est consacrée à l'étude des situations expérimentales dans lesquelles l'effet de deux facteurs (variables qualitatives) est étudié simultanément, c'est-à-dire dans le même protocole expérimental. En cela, elle constitue une extension à la situation précédente dans laquelle on n'étudiait qu'un seul facteur à la fois (ANOVA d'ordre 1).

3.7.1 Position du problème

Nous avons réalisés un recueil de rendement de trois variétés du blé selon le type d'engrais utilisé, les mesures obtenus sont rangées dans la table 3.4.

	Variété 1	Variété 2	Variété 3
Engrais 1	46	41	35
	35	26	21
	19	11	31
Engrais 2	37	49	45
	18	37	66
	18	35	61
Engrais 3	32	65	34
	43	67	66
	32	58	58

TABLE 3.4: *Variation de rendement du blé selon la variété et le type d'engrais*

Si l'on s'intéresse, à l'effet des trois type d'engrais et les trois variétés du blé sur les rendements séparément, on pourrait réaliser deux expériences dans lesquelles on manipulerait chacun des deux facteurs, et analyser les résultats à l'aide d'ANOVA 1 s'il y a plus de 2 modalités ou niveaux pour chaque facteur : on saurait ainsi si le type d'engrais affecte sensiblement les rendements mesurées, et également si la variété affecte les rendements. Mais, on ne saurait pas si l'effet du type d'engrais est le même quelque soit la variété du blé ; en d'autres termes, on perd l'information concernant l'interaction entre ces deux facteurs.

Les modèles d'ANOVA d'ordre 2 sont comparables sur le fond au modèle précédent de l'ANOVA d'ordre 1, mais elle incluent en plus de l'étude des effets principaux des deux facteurs, celle du l'effet d'interaction des deux facteurs.

3.7.2 Analyse de variance à deux facteurs

L'identification de l'ANOVA d'ordre 2 (ANOVA 2) au sens littéraire peut être résumée dans la définition suivante :

Définition 3.2 *(ANOVA 2)*
L'analyse de la variance à deux facteurs teste l'effet de deux facteurs contrôlés A et B (variables qualitatives) ayant respectivement I et J modalités sur les moyennes d'une variable quantitative X.

Les problèmes concernés par la technique ANOVA 2 se présente en générale de la manière suivante :

N°	B_1	B_2	$\cdots$	B_J
A_1	$X_{1,1,1}$	$X_{1,2,1}$	$\cdots$	$X_{1,J,1}$
	$X_{1,1,2}$	$X_{1,2,2}$	$\cdots$	$X_{1,J,2}$
	$\vdots$			$\vdots$
	$X_{1,1,K}$	$X_{1,2,K}$	$\cdots$	$X_{1,p,K}$
A_2	$X_{2,1,1}$	$X_{2,2,1}$	$\cdots$	$X_{2,J,1}$
	$X_{2,1,2}$	$X_{2,2,2}$	$\cdots$	$X_{2,J,2}$
	$\vdots$			$\vdots$
	$X_{2,1,K}$	$X_{2,2,K}$	$\cdots$	$X_{2,J,K}$
$\vdots$		$\vdots$		
A_I	$X_{I,1,1}$	$X_{I,2,1}$	$\cdots$	$X_{I,J,1}$
	$X_{I,1,2}$	$X_{I,2,2}$	$\cdots$	$X_{I,J,2}$
	$\vdots$			$\vdots$
	$X_{I,1,K}$	$X_{I,2,K}$	$\cdots$	$X_{I,J,K}$

et sont modèle mathématique est donné par :

$$X_{ijk} = \mu_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1,I}, j = \overline{1,J} \text{ et } k = \overline{1,K}, \tag{3.15}$$

où X_{ijk} est la $k^{\text{ième}}$ réalisation de la variable quantitative X, lorsque on fixe le premier facteur à la $i^{\text{ième}}$ modalité et le deuxième facteur à la $j^{\text{ième}}$ modalité et ϵ_{ijk} sont les erreurs de mesure (inconnues) de plus $E(\epsilon_{ijk}) = 0$.

Le modèle (3.15) peut être réécrit sous sa forme détaillée comme suit :

$$X_{ijk} = \mu + a_i + b_j + c_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1,I}, j = \overline{1,J} \text{ et } k = \overline{1,K}, \tag{3.16}$$

ce qui s'explique que la réalisation de la variable X est un cumule d'une constate μ (indépendante des deux facteurs), de l'effet du premier facteur a, de l'effet du deuxième facteur b, l'effet d'interaction des deux facteurs c et de l'erreur de mesure ϵ.

Si le modèle de référence retenu est le modèle (3.15), alors le test pour lequel nous nous intéressons à réaliser sera formulé comme suit :

$$\begin{aligned} & H_0 :'' \forall \left\{ \begin{array}{l} i \in \{1,2,\cdots,I\} \\ j \in \{1,2,\cdots,J\} \end{array} \right\}, \ \mu_{ij} = \mu'' \\ \text{contre} & \\ & H_1 :'' \exists \left\{ \begin{array}{l} i_1, i_2 \in \{1,2,\cdots,I\} \\ j_1, j_2 \in \{1,2,\cdots,J\} \end{array} \right\} \text{ tel que } \mu_{i_1 j_1} \neq \mu_{i_2 j_2}.'' \end{aligned} \tag{3.17}$$

Par contre, si le modèle de référence retenu est le modèle (1), alors l'analyse de la variance à deux facteurs avec repetitions consiste en réalisation de trois tests de Fisher à la fois, dont la formulation est :

1. **Effet du premier facteur** :
   H_0 : "*les paramètres a_i sont tous nuls*" contre H_1 : "*les paramètres a_i ne sont pas tous nuls*"
2. **Effet du second facteur** :
   H_0 : "*les paramètres b_j sont tous nuls* " contre H_1 : "*les paramètres b_j ne sont pas tous nuls*"
3. **Effet de l'interaction des deux facteurs** :
   H_0 : "*les paramètres c_{ij} sont tous nuls* " contre H_1 : "*les paramètres c_{ij} ne sont pas tous nuls*"

3.7.3 Les étapes de l'ANOVA 2

La lise en oeuvre d'une ANOVA 2, se fait principalement en 4 étapes. Les détails de ces étape sont comme suit :

Étape 0 : (Conditions)

Afin de réaliser une analyse de la variance à deux facteurs, les conditions suivantes doivent être vérifiées préalablement :

- Les $I * J$ échantillons comparés sont mutuellement indépendants.
- La variable quantitative étudiée suit une loi normale dans les $I * J$ populations comparées.
- Les $I * J$ populations comparées ont même variance : *Homogénéité* des variances (*homoscédasticité*).

Étape 1 : (Moyennes et variances)

Quantifier les différentes statistiques intervenant dans l'ANOVA à 2 facteurs et qui sont :

- La moyenne globale de toutes les observations :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} \text{ avec } n = I * J * K;$$

- Moyenne de chaque échantillon :

$$\overline{X}_{ij\bullet} = \frac{1}{K}\sum_{k=1}^{K} X_{ijk} \text{ pour } i = \overline{1, I} \text{ et } j = \overline{1, J};$$

- Moyenne de chaque modalité du premier facteur :

$$\overline{X}_{i\bullet\bullet} = \frac{1}{J * K}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} \text{ pour } i = \overline{1, I};$$

- Moyenne de chaque modalité du deuxième facteur :

$$\overline{X}_{\bullet j\bullet} = \frac{1}{I * K}\sum_{i=1}^{I}\sum_{k=1}^{K} X_{ijk} \text{ pour } j = \overline{1, J};$$

- La somme des carrés des erreurs totale :

$$SC_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \overline{X})^2;$$

- La somme des carrés des erreurs résiduelles :

$$SC_{res} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(X_{ijk} - \overline{X}_{ij\bullet})^2;$$

- La somme des carrés des erreurs du premier facteur :

$$SC_{a} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{i\bullet\bullet} - \overline{X})^2;$$

- La somme des carrés des erreurs du deuxième facteur :

$$SC_b = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{\bullet j\bullet} - \overline{X})^2;$$

- La somme des carrés des erreurs des deux facteurs facteur :

$$SC_c = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{X}_{ij\bullet} - \overline{X}_{i\bullet\bullet} - \overline{X}_{\bullet j\bullet} + \overline{X})^2.$$

Avec le même raisonnement que dans l'ANOVA 1, on peut démontrer que la variation quadratique totale des observations autour de la moyenne $\overline{X}$ peut être décomposé comme suit :

$$SC_{Tot} = SC_{Res} + SC_a + SC_b + SC_c. \tag{3.18}$$

Étape 2 : (Les Carrés moyens)
A partir de la décomposition (3.18), l'idée la plus naturelle est que le facteur ou l'interaction des facteurs n'a pas d'impact sur le caractère étudié si la variation inter-groupes (engendrée par les deux facteur ou/et leurs interaction) associée au caractère est négligeable par rapport aux fluctuations individuelles. Pour comparer ces quantités, on considère les carrés moyens suivants :

Carré moyen due aux fluctuations individuelles : $CM_{Res} = \frac{SC_{Res}}{I*J*(K-1)}$.

Carré moyen de mesure de l'effet du premier facteur :$CM_a = \frac{SC_a}{(I-1)}$.

Carré moyen de mesure de l'effet du second facteur : $CM_b = \frac{SC_b}{(J-1)}$.

Carré moyen de mesure de l'effet de l'interaction entre les deux facteurs : $CM_c = \frac{SC_c}{(I-1)*(J-1)}$.

Notons que, si les trois conditions citées précédemment (Indépendance, Normalité et Homogénéité) sont vérifiées alors sous l'hypothèse H_0

$$\frac{CM_a}{CM_{Res}} \rightsquigarrow F_{((I-1),I*J(K-1))}, \tag{3.19}$$

$$\frac{CM_b}{CM_{Res}} \rightsquigarrow F_{((J-1),I*J(K-1))}, \tag{3.20}$$

$$\frac{CM_c}{CM_{Res}} \rightsquigarrow F_{((I-1)*(J-1),I*J(K-1))}, \tag{3.21}$$

où F la loi de Fisher.

Étape 3 : (Décision)
Pour un seuil de risque donné α, nous quantifions les valeurs critiques c_α, c_β et c_γ(par la lecture sur la table de Fisher), telle que :

$$P\left(\frac{CM_a}{SC_{Res}} < f_a\right) = 1-\alpha,$$

$$P\left(\frac{CM_b}{SC_{Res}} < f_b\right) = 1-\alpha,$$

$$P\left(\frac{CM_c}{SC_{Res}} < f_c\right) = 1-\alpha,$$

Ainsi les décisions des test se font comme suit :

Décision sur le premier facteur :

- Si $\frac{CM_a}{SC_{Res}} < f_a$, alors le premier facteur n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_a}{SC_{Res}} \geq f_a$, alors le premier facteur a une influence significative sur le caractère étudié.

Décision sur le deuxième facteur :

- Si $\frac{CM_b}{SC_{Res}} < f_b$, alors le deuxième facteur n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_b}{SC_{Res}} \geq f_b$, alors le deuxième facteur a une influence significative sur le caractère étudié.

Décision sur l'interaction des deux facteurs :

- Si $\frac{CM_c}{SC_{Res}} < f_c$, alors l'interaction des deux facteurs n'a pas une influence significative sur le caractère étudié.
- Si $\frac{CM_c}{SC_{Res}} \geq f_c$, alors l'interaction des deux facteurs a une influence significative sur le caractère étudié.

Remarque 3.2 *Les résultats d'une ANOVA 2 sont souvent présentés dans un tableau de la forme suivante :*

source de variation	*Somme des carrés* SC	*Degrés de libertés* ddl	*Carré moyen* CM	*ratio* F_{obs}	*Ficher* F_c
due à F_A	SC_a	$I-1$	CM_a	CM_a/CM_{Res}	f_a
due à F_B	SC_b	$J-1$	CM_b	CM_b/CM_{Res}	f_b
due à $F_A \times F_B$	SC_c	$(I-1)*(J-1)$	CM_c	CM_c/CM_{Res}	f_c
Résiduelle	SC_{Res}	$I*J*(K-1)$	CM_{Res}		
Totale	SC_{Tot}	$n-1$			

TABLE *3.5: Tableau de l'ANOVA d'ordre 2*

3.7.4 Exemple d'application

Afin de concrétiser les différentes étapes citées auparavant, reprenant l'exemple présenté dans la section 3.7.1. Supposons que les trois conditions d'application de la technique d'ANOVA 2 sont vérifiées et on désir prendre nos décisions pour un risque $\alpha = 5\%$ de se tromper, alors le reste des étapes se fait comme suit :

- Calculer les moyennes des différents échantillons :

	Variété 1	Variété 2	Variété 3	$\overline{X}_{i\bullet\bullet}$
Engrais 1	33.3333	26.0000	29.0000	29.4444
Engrais 2	24.3333	40.3333	57.3333	40.6667
Engrais 3	35.6667	63.3333	52.6667	50.5556
$\overline{X}_{\bullet j\bullet}$	31.1111	43.2222	46.3333	40.2222

TABLE 3.6: *Moyennes des différents échantillons*

- Calculer les variances des différents échantillons :

	Variété 1	Variété 2	Variété 3	$\sigma^2_{i\bullet\bullet}$
Engrais 1	184.3333	225.0000	52.0000	125.5278
Engrais 2	120.3333	57.3333	120.3333	278.7500
Engrais 3	40.3333	22.3333	277.3333	231.0278
$\sigma^2_{\bullet j\bullet}$	113.1111	342.1944	285.5000	272.7179

TABLE 3.7: *Variances des différents échantillons*

- Compléter la table 3.5 :

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio f_{obs}	Ficher f_α
due à F_a	1164.222	2	582.111	4.766	$f_a = 3.55$
due à F_b	2008.222	2	1004.111	8.220	$f_b = 3.55$
due à $F_a \times F_b$	1719.556	4	429.889	3.519	$f_c = 2.93$
Résiduelle	2198.667	18	122.148		
Totale	50772.000	27			

- Décision : Dans les trois situations on constate que $F_{obs} > f_\alpha$ ($4.766 > 3.55$, $8.220 > 3.55$ et $3.519 > 2.93$), cela signifié qu'on doit rejeter l'hypothèse H_0 dans les trois tests. L'interprétation de ces résultats vis-à-vis le problème étudié est que le facteur variété et le facteur type d'engrais ainsi que l'interaction entre eux influent significativement sur le rendement du blé.
  Rappelons que cette décision est prise pour un risque de 5% mais si on souhaite diminuer ce risque à 1% alors la décision sera différente (pour cet exemple seulement).
  En effet, pour un seuil de risque 2%, on constate que le type d'engrais qui influe significativement sur le rendement le fait que $8.220 > f_b = 6.01$. Par contre la variété (respectivement l'interaction des deux facteurs) n'a pas une influence significative sur le rendement car $4.766 < f_a = 6.01$ (respectivement $3.519 < f_c = 4.58$).

Remarque 3.3

- *Dans la littérature on distingue trois types d'ANOVA (I, II et III) et cela selon la nature du/des facteur(s) étudié :*

  type I : *modèle à effet fixe lorsque les modalités des facteurs sont choisies délibérément par l'expérimentateur. C'est le cas dans la plupart des protocoles expérimentaux, et c'est le type que nous avons développé dans ce document.*

  type II : *modèle à effet aléatoire lorsque les modalités des facteurs sont issues d'un processus d'échantillonnage.*

  type III : *modèle à effets mixtes, lorsque on dispose des facteurs à effet fixe et des facteurs à effet aléatoire simultanément.*

- *Une autre classification de l'ANOVA existe également et cela la présentation des observations :*

  Plan avec répétitions *: chaque cellule contient plusieurs observations.*

  Plan sans répétitions *: chaque cellule contient une seule observation (voir chapitre exercice).*

  Plan équilibré *: chaque cellule contient le même nombre d'observations.*

  Plan non équilibré *: les cellules ne contient pas le même nombre d'observations (voir chapitre exercice).*

  Plan à mesures répétées *: les mêmes sujets ont été utilisés pour les observations au sein de différentes cellules (appariement).*

Conclusion

A partir les différentes notions et différents tests exposés dans ce chapitre on peut conclure que :

Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validité d'hypothèses relatives à une ou plusieurs populations.

Les méthodes de l'inférence statistique nous permettent de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard, ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

Les tests d'hypothèses font appel à un certain nombre d'hypothèses concernant la nature de la population dont provient l'échantillon étudié (normalité de la variable, égalité des variances, indépendance, etc.) et qui doivent être vérifier préalablement.

Chapitre 4

Régression linéaire simple et multiple

Introduction et problématique

La régression est l'une des méthodes les plus connues et les plus appliquées en statistiques pour l'analyse de données quantitatives sous forme d'un modèle. Si on s'intéresse à la relation entre deux variables, on parlera de *régression simple* en exprimant l'une des deux variables en fonction de l'autre. Tandis que, si la relation porte entre une variable et plusieurs autres variables (≥ 2), on parlera de *régression multiple*.

La mise en œuvre d'une régression impose l'existence d'une relation de cause à effet entre les variables prises en compte dans le modèle. Cette méthode peut être mise en place sur des données quantitatives observées sur n individus et présentées sous la forme :

$$y = f(x) + \epsilon, \tag{4.1}$$

où

- y est une variable quantitative prenant la valeur y_i pour l'individu i $(i = 1, ..., n)$, appelée variable à expliquer ou variable réponse.
- $x_1, x_2, ..., x_p$ sont p variables quantitatives prenant respectivement les valeurs $x_{1i}, x_{2i}, ...x_{pi}$ pour le $i^{\text{ième}}$ individu, appelées variables explicatives ou prédicteurs.
- ϵ est une variable aléatoire (résidus).

Considérons un couple de variables quantitatives (X, Y). S'il existe une liaison entre ces deux variables, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y. Si l'on admet qu'il existe une relation de cause à effet entre X et Y, le phénomène aléatoire représenté par X peut donc servir à prédire celui représenté par Y et la liaison s'écrit sous la forme (4.1) et on dit que l'on fait de la régression de y sur x (dans le cas d'une régression multiple de y sur $x_1, x_2, ..., x_p$ la liaison peuvent être écrite sous la forme $y = f(x_1, x_2, ..., x_p)$).

Dans les cas les plus fréquents, on choisit l'ensemble des fonctions affinées du type :

Cas de régression linéaire simple :

$$f(x) = ax + b. \tag{4.2}$$

Cas de régression linéaire multiple :

$$f(x) = f(x_1, x_2, ..., x_p) = a_0 + a_1x_1 + a_2x_2 + ... + a_px_p. \tag{4.3}$$

4.1 Le modèle de régression linéaire simple

Soit un échantillon de n individus. Pour un individu i $(i = 1, ..., n)$, on a observé y_i la valeur de la réalisation de la variable quantitative Y et x_i la valeur de la variable quantitative x.

On veut étudier la relation entre ces deux variables, et en particulier, l'effet de x (variable explicative) sur y (variable réponse).

Dans un premier temps, on peut représenter graphiquement cette relation en traçant le nuage des n points de coordonnées (x_i, y_i) et on constate que la relation entre y_i et x_i s'écrit sous la forme d'un modèle de régression linéaire

La relation entre y et x est supposée n'être qu'approximative : elle est perturbée par un "terme d'erreur" additif, noté ϵ_i avec $E(\epsilon_i),\ i = \overline{1:n}$.

L'équation de la régression linéaire simple (ou le "modèle de régression") s'écrit donc de la façon suivante :

$$Y = a + b + \epsilon, \tag{4.4}$$

$$E(Y) = a + bE(x), \tag{4.5}$$

ou encore,

$$E(Y/x) = a + bx, \tag{4.6}$$

où a et b sont les paramètres du modèle, et ϵ est le terme d'erreur qui est une variable aléatoire.

Remarque 4.1

1. *a représente le point d'intersection de la droite de régression avec l'ordonnée ("intercept", "constante").*
2. *b représente la pente de la droite de régression.*
3. *La valeur de b donne le nombre d'unités supplémentaires de Y associées à une augmentation par une unité de x.*
4. *$E(Y/x)$ est la moyenne de Y pour une valeur de x donnée.*

Exemple 8

(a) : *$Y_i = a + bx_i + cx_i^2 + \epsilon_i$, est un modèle linéaire tandis que la relation entre x et y n'est pas linéaire mais de type polynomial.*

(b) : *$Y_i = a + b\ cos(x_i) + \epsilon_i$, est un modèle linéaire.*

(c) : *$Y_i = ae^{bx_i} + \epsilon_i$, n'est pas un modèle linéaire.*

(d) : *$Y_i = ab + cx_i + \epsilon_i$, n'est pas un modèle linéaire.*

Enfin, la linéarité est relier aux paramètres du modèle et non pas aux variables explicatives.

4.2 Analyse du modèle de régression linéaire simple

Soit le couple (X, Y) de variable aléatoire où X est une variable indépendante et Y la variable dépendante. On cherche une relation du type

$$Y = a + bx + \epsilon.$$

Notons que la mise en oeuvre et l'exploitation de ce modèle nécessite une quantification préalable des paramètres inconnus a et b.

4.2.1 Estimation des paramètres du modèle

On suppose que la variable X est contrôlée par l'expérimentateur où il réalise n expériences y_1 , y_2 , $y_3, \ldots, y_n$ aux points x_1 , x_2 , x_3 , $\ldots$, x_n fixés. De plus, on suppose que les Y_i sont mutuellement indépendants.

Le modèle s'écrit

$$y_i = a + bx_i$$

pour $i = \overline{1:n}$, tel que :

- $E(\epsilon_i) = 0$
- $Cov(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$
- $Var(\epsilon_i) = \sigma^2 \quad \forall i = \overline{1:n}$

Supposons qu'on opte pour la méthode des moindre carrées pour quantifier a et b, alors les estimateurs des paramètres a et b sont $\hat{a}$ et $\hat{b}$ qui minimise la fonction $Q(a, b)$, définie par :

$$Q(a,b) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - a - bx)^2. \tag{4.7}$$

Cela revient à la détermination d'un optimum minimal de la fonction des erreurs quadratique $Q(a, b)$, qui consiste à résoudre le système des équations suivant :

$$\begin{cases} \frac{\partial Q(a,b)}{\partial a} &= 0, \\ \frac{\partial Q(a,b)}{\partial b} &= 0, \end{cases} \tag{4.8}$$

c'est-à-dire,

$$\begin{cases} -2\sum_{i=1}^{n}(Y_i - a - bx) &= 0 \\ -2\sum_{i=1}^{n} x_i(Y_i - a - bx) &= 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\ \sum_{i=1}^{n} Y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 &= 0 \end{cases}. \tag{4.9}$$

Finalement, le système à résoudre, pour estimer les coefficients de régression a et b, ni rien d'autre qu'un système linéaire à deux équations et à deux inconnus, qui est donné par :

$$\begin{cases} a\left(\sum_{i=1}^{n} 1\right) + b\left(\sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{n} Y_i \\ a\left(\sum_{i=1}^{n} x_i\right) + b\left(\sum_{i=1}^{n} x_i^2\right) &= \sum_{i=1}^{n} Y_i x_i \end{cases} \tag{4.10}$$

La résolution du système (4.10), nous fournis la solution suivante :

$$\boxed{\hat{b} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2}} \quad \text{et} \quad \boxed{\hat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i}, \tag{4.11}$$

ou encore :

$$\boxed{\hat{b} = \frac{Cov(x,y)}{Var(x)}} \text{ et } \boxed{\hat{a} = \overline{Y} - \hat{b}\overline{X},} \tag{4.12}$$

où :

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] - \overline{X}\ \overline{Y}. \tag{4.13}$$

4.2.2 Estimation de σ^2

En plus de l'estimation des paramètres du modèle (a et b), l'une des caractéristique statistique importante liée au modèle est bien que la variance inconnue σ^2. Pour cela, nous allons estimer σ^2, où nous proposons d'utiliser la méthode MLE (voir chapitre 2). A cet effet, on suppose que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, alors

$$Y_i \rightsquigarrow \mathcal{N}(a + bx_i, \sigma^2).$$

Dans ce cas, la fonction de vraisemblance correspondante au modèle est donnée par :

$$\mathcal{L}(Y_1, Y_2, ..., Y_n, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2\right],$$

L'expression de la variance qui maximise cette fonction est donnée par :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2. \tag{4.14}$$

Mais, cet estimateur est un estimateur avec Biais, alors il doit être corriger. Ainsi, après sa correction on aura l'estimateur sans Biais de σ^2 suivant :

$$\hat{\sigma}_c^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \rightsquigarrow \chi^2_{n-2}. \tag{4.15}$$

4.2.3 Qualité et validation du modèle :

Dans cette section, nous allons présenter deux manière du jugé la qualité et l'adéquation du modèle linéaire :

$$Y_i = a + bx_i + \epsilon_i \quad , i = 1, ..., n.$$

pour l'explication de la variable Y à l'aide de la variable x.

4.2.3.1 Coefficients de correlation et de détermination

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoire, c'est étudier l'intensité de la liaison qui peut être existée entre ces variables.

Une mesure de cette corrélation dans le cadre linéaire est obtenue par le calcul du coefficient appelé coefficient de corrélation. Ce coefficient est égal au rapport de leurs covariances et du produit non nul de leurs écarts types :

$$\rho = Cor(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = r(x, y). \tag{4.16}$$

avec,

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2, \tag{4.17}$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2, \tag{4.18}$$

et $Cov(x, y)$ est donnée dans (4.13).

Le coefficient de corrélation est toujours compris entre -1 et +1. De plus, son signe donne le sens de la corrélation où le signe positif indique que les deux variables sont proportionnelles

dans le même sens, tandis que le signe négatif indique que les deux variables sont inversement proportionnelles.

Plus $|\rho|$ est près de 1, plus la corrélation est grande donc le modèle linéaire décrit bien le phénomène étudié. Par contre, si $|\rho|$ est près de zéro le modèle linéaire n'est pas adéquat pour la modélisation du problème étudié.

Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire entre les deux variables considérées. A cet effet, un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux variables mais seulement l'absence d'une relation linéaire. Pour cela, il ne faut pas confondre la corrélation et la relation causale. Une bonne corrélation entre deux variables peut révéler une relation de cause à effet entre elle, mais pas nécessairement.

Pour mieux juger la qualité d'une régression linéaire, on définit un autre indicateur compris entre 0 et 1, nommé : *coefficient de détermination*, noté R^2 :

$$R^2 = \rho^2.$$

Ce nombre mesure l'adéquation entre le modèle et les données observées où plus, R^2 est près de 1, plus le modèle est adéquat et le contraire est vrai.

4.2.3.2 Le test de Fisher

Une autre technique, plus puissante que le calcule de coefficient de corrélation, pour mesuré la pertinence et l'adéquation d'un modèle est l'utilisation du test de Ficher qui se base sur l'analyse de la variance.

On peut démontrer que la variation totale de Y se décompose comme suit :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{SCE}$$

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Dans la logique des choses, le modèle est validé, si la variation totale du modèle n'est engendré que par la variation des résidus et non pas par la variation de la régression, autrement dit la variation moyenne des résidus doit être supérieur à la variation moyenne de la régression pour valider le modèle,

$$\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} > k > 1,$$

donc, il nous reste à savoir comment déterminer la valeur critique k.

Sachant que :

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \rightsquigarrow \chi^2_{n-2}$$

et

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2 \rightsquigarrow \chi^2_1,$$

alors,

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} \rightsquigarrow f_{(1,n-2)}$$

où la notation $f_{(1,n-2)}$ désigne est une loi de Fisher de degrés de liberté $n_1 = 1$ et $n_2 = n-2$, cela signifie que, pour un risque α, la valeur critique k n'est rien d'autre que le fractale d'ordre $1-\alpha$ d'une loi de Fisher de degrés de liberté 1 et $n-1$ ($k = f_{(1,n-2,1-\alpha)}$) ainsi on décide que :

- Si $f_c > f_{(1,n-2,1-\alpha)}$ alors le modèle est valide.
- Si $f_c \leq f_{(1,n-2,1-\alpha)}$ le modèle n'est pas valide.

où f_c est la réalisation de la statistique F.

4.3 Régression linéaire multiple

Dans la pratique les principales étapes d'une analyse de régression Multiple sont :

1. Définir la variable dépendante et les variables explicatives.
2. Spécifier la nature de la relation entre la variable dépendante et les variables explicatives.
3. Estimer les paramètres du modèle, en suite, quantifier sa qualité et vérifier sa validité.
4. Dans le cas où le modèle est retenu, interpréter sa signification par rapport au problème posé.

Dans cette section, nous nous intéresserons à la régression multiple dans le cadre du modèle linéaire. La régression linéaire multiple est la généralisation de la régression linéaire simple qui ne considère qu'une seule variable explicative. Considérons le modèle linéaire multiple dont la forme est la suivante :

$$Y = b_1 x_1 + b_2 x_2 + ... + b_k x_k + \epsilon,$$

pour la $i^{\text{ième}}$ observation le modèle peut être représenté de la manière suivante :

$$Y_i = b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \epsilon_i \ , \qquad i = 1, ..., n,$$

ou encore, sous sa forme Matricielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & ... & x_{k1} \\ x_{12} & x_{22} & ... & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

$$Y \quad = \quad X \quad b \quad + \quad \epsilon$$

A partir des étapes de l'analyse de régression multiple, cité précédemment, on est au niveau de l'étape (3), c'est-à-dire on doit estimer les paramètres (coefficients) du modèle.

4.3.1 Estimation des paramètres du modèle

Supposons qu'on a :

$$Y = Xb + \epsilon,$$

avec,

$$X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & & & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad et \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

de plus,

- $E(\epsilon) = 0$,
- $Var(\epsilon) = \sigma^2 I_n$, où I_n est une matrice d'identités d'ordre n.

On utilisant la méthode des moindres carrés, pour estimer les coefficients du modèle, on aura un système linéaire à k équations et k variables. Ce système s'écrit sous sa forme matricielle comme suit :

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \dots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{1i}x_{2i} & \sum_{i=1}^{n} x_{1i}^2 & \dots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{1i}x_{ki} & \dots & \dots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}}_{M} \times \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}}_{b} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n} y_i x_{1i} \\ \sum_{i=1}^{n} y_i x_{2i} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ki} \end{bmatrix}}_{m}$$

où $M = X^t\ X$ et $m = X^t\ Y$.

Finalement, l'estimation des coefficients du modèle sont données par le calcul matriciel suivant :

$$\hat{b} = (X^t.X)^{-1}.X^t.Y.$$

4.3.2 Test sur la validité du modèle

Avec le même raisonnement abordé dans le cas de la regression linéaire simple on peut construire le test de validation du modèle. En effet, la variation totale de Y se décompose comme suit :

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{Y})^2}_{SCT} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SCR} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}_{SCE}$$

où

SCT : Variation de Y ou variation totale.

SCR : Variation des résidus.

SCE : Variation de la régression ou variation expliqué par la régression.

Pour valider le modèle, on test

$$H_0 \text{ "}b_1 = b_2 = ... = b_k = 0\text{" contre } H_1 \text{ "}\exists j \in \{1,\ 2, ...,\ k\}/\ b_j \neq 0\text{"},$$

avec le même raisonnement que dans le cas de régression linéaire on obtient la statistique du test suivante :

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2/(k-1)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-k)} \rightsquigarrow f_{(k-1,n-k)},$$

où la notation $f_{(k-1,n-k)}$ désigne la loi de Fisher à $k-1$ et $n-k$ degrés de liberté.

Ainsi, on décide :

- Si $f_c > f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de valider le modèle.
- Si $f_c \leq f_{(k-1,n-k,1-\alpha)} \Rightarrow$ de ne pas valider le modèle.

avec f_c est la réalisation de la statistique F.

Chapitre 5

Analyse en Composantes Principales (ACP)

La description des liaisons entre deux variables par des techniques statistiques bidimensionnelle conduisent à se poser la question de la représentation simultanée de données en dimension plus grande que 2.

Quelle graphique permettrait de "*généraliser*" le nuage de points tracé dans le cas de deux variables permettant d'aborder la structure de corrélation présente entre plus de 2 variables ? L'outil utilisé est alors l'analyse en composantes principales (ACP).

L'idée et le principe de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales.

5.1 Exemple d'une ACP

Une présentation très élémentaire de l'Analyse en Composantes Principales est proposée sur un exemple jouet de données. Considérons le tableau, de données, suivant :

$$Y = \begin{pmatrix} 2 & 2 & 3 \\ 3 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 4 \\ 2 & 1 & 3 \end{pmatrix}, \tag{5.1}$$

correspondant à des mesures effectuées sur 5 individus de poids statistiques égaux pour les trois variables Y_1, Y_2 et Y_3.

Réaliser une ACP centrée-réduite sur ce dernier tableau consiste en générale à répondre au questions suivantes :

1. Calculer :
   - l'individu moyen $\overline{Y}$,
   - le vecteur $\sigma = (\sigma_1,\ \sigma_2, \sigma_3)$ des écarts types des variables,
   - la matrice X des données centrées-réduites.
2. Calculer la matrice des corrélations ρ.
3. Effectuer la décomposition aux valeurs propres de ρ.
4. Calculer les vecteurs propres v_1, v_2 et v_3 de ρ.
5. Calculer les composantes principales c_1, c_2 et c_3 dont on vérifiera les propriétés statistiques.
6. Représenter les individus dans le plan factorial (v_1, v_2).
7. Donner une interprétation de cette ACP.

5.2 Solution

Avant de répondre à notre objectif, introduisons la notation suivante qui nous sera utile par la suite :

$$Y = \begin{pmatrix} y_{11} & y_{21} & y_{31} \\ y_{12} & y_{22} & y_{32} \\ y_{13} & y_{23} & y_{33} \\ y_{14} & y_{24} & y_{34} \\ y_{15} & y_{25} & y_{35} \end{pmatrix} \tag{5.2}$$

1. L'individu moyen est obtenu en faisant la moyenne de chacune des colonnes du tableau Y, soit

$$\overline{X} = \left(\frac{1}{5}\sum_{i=1}^{5} Y_{1i}, \frac{1}{5}\sum_{i=1}^{5} Y_{2i}, \frac{1}{5}\sum_{i=1}^{5} Y_{3i} \right) = (2, 1, 3).$$

Le vecteur des écarts types est obtenu en calculant les écarts types de chaque colonnes de Y. Soit Y_c la matrice des données centrées, $Y_c = Y - \overline{X}$, c'est-à-dire :

$$Y_c = \begin{pmatrix} 2-2 & 2-1 & 3-3 \\ 3-2 & 1-1 & 2-3 \\ 1-2 & 0-1 & 3-3 \\ 2-2 & 1-1 & 4-3 \\ 2-2 & 1-1 & 3-3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.3}$$

Le vecteur σ contient les termes en racine carrée des éléments diagonaux de la matrice des variances (Var-Cov) définie par : $S^2 = \frac{1}{n} Y_c^t Y_c$.

$$S^2 = \frac{1}{n} Y_c^t Y_c = \frac{1}{5} \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \frac{1}{5} \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}, \tag{5.4}$$

alors, $\sigma^2 = \frac{1}{5}(2, 2, 2)$, d'où $\sigma = \left(\sqrt{\frac{2}{5}}, \sqrt{\frac{2}{5}}, \sqrt{\frac{2}{5}} \right)$.

2. Le calcul de la matrice X, des données centrées réduites, revient à diviser chaque colonne de Y_c sur l'écart-type de la variable correspondante $\left(x_{ij} = \frac{Y_{ij} - \overline{Y}_i}{\sigma_i} \right)$, ainsi on aura la matrice suivante :

$$X = \frac{1}{\sqrt{2/5}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \sqrt{\frac{5}{2}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.5}$$

3. La matrice des corrélations ρ est définie par :

$$\rho = \frac{1}{n} X^t X = \frac{1}{5}\sqrt{\frac{5}{2}} \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \end{pmatrix} \times \sqrt{\frac{5}{2}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.5 & -0.5 \\ 0.5 & 1.0 & 0 \\ -0.5 & 0 & 1.0 \end{pmatrix} \tag{5.6}$$

4. L'ACP, centrée-réduite, de Y nécessite le calcul préalable des vecteurs propres de ρ. A cet effet, on résout d'abord l'equation

$$\det(\rho - \lambda I) = 0,$$

après le calcul du déterminant on obtiens ce qui suit :

$$\det(\rho - \lambda I) = -\lambda^3 + 3\lambda^2 - (5/2)\lambda + 1/2,$$

soit,

$$P_3(\lambda) = -\lambda^3 + 3\lambda^2 - (5/2)\lambda + 1/2.$$

Le polynôme $P_3(\lambda)$ s'appel le polynôme caractéristique de ρ, dont ses racine sont les valeurs propres correspondantes à la matrice ρ.
On remarque que $\lambda = 1$ est une racine du polynôme $P_3(\lambda)$. Pour cela, $P_3(\lambda)$ peut s'écrire sous la forme :

$$P_3(\lambda) = (\lambda - 1)P_2(\lambda) = (\lambda - 1)(-\lambda^2 + a\lambda + b) = (\lambda - 1)(-\lambda^2 + 2\lambda - 1/2),$$

Il est à noter que l'utilité de cette décomposition n'est rien d'autre que la facilité de résolution de l'équation $P_3(\lambda) = 0$. La résolution de l'équation $P_3(\lambda) = 0$, nous fournit les 3 valeurs propres suivantes :

$$\lambda_1 = 1 + \frac{1}{\sqrt{2}}, \quad \lambda_2 = 1 \text{ et } \lambda_3 = 1 - \frac{1}{\sqrt{2}}.$$

Rappelons que, le vecteur propre v associé à une valeur propre λ est la solution du système linéaire $\rho v = \lambda v$. Ainsi, les vecteur propres associe à ces valeurs sont :

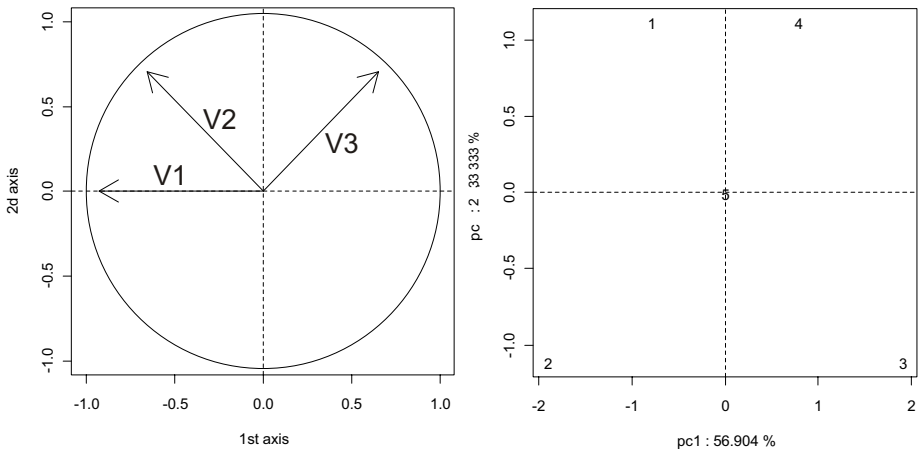
$$v_1 = \begin{pmatrix} -\sqrt{2}/2 \\ -1/2 \\ 1/2 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ \sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix} \text{ et } v_3 = \begin{pmatrix} \sqrt{2}/2 \\ -1/2 \\ 1/2 \end{pmatrix}.$$

5. Le calcul des composantes principales est donné par

$$c^i = Xv_i, \quad i = 1,\ 2,\ 3. \tag{5.7}$$

et on obtiens : $c^1 = \begin{pmatrix} -0.7906 \\ -1.9086 \\ 1.9086 \\ 0.7906 \\ 0.0000 \end{pmatrix}$, $c^2 = \begin{pmatrix} 1.1180 \\ -1.1180 \\ -1.1180 \\ 1.1180 \\ 0.0000 \end{pmatrix}$ et $c^3 = \begin{pmatrix} -0.7906 \\ 0.3275 \\ -0.3275 \\ 0.7906 \\ 0.0000 \end{pmatrix}$.

6. Représentation des individus dans le plan factoriel (v_1, v_2).



Le premier axe oppose les variations de V_1, V_2 avec V_3. Le second est un axe de taille. Les individus 2 et 3 présentent de faibles valeurs de V_2 et V_3, l'individu 2 étant caractérisé par une forte valeur de V_1. Les individus 1 et 4 sont attachés aux variables V_2 et V_3 respectivement. L'individu 5 est le plus consensual puisqu'il est confondu avec le centre de gravité de l'ACP.

Chapitre 6

Exercices corrigés

Introduction

Dans ce chapitre nous proposons 38 exercices avec des solutions détaillées qui sont ordonnés selon les notions introduites dans les chapitres précédents afin de permettre aux utilisateurs de ce polycopié de contrôler l'acquisition des notions essentielles qui ont été introduites.

6.1 Énoncés des exercices

Exercice 1 *(Densité de probabilité)*

a) Dans chacun des cas, dites si la fonction f définit une densité de probabilité

$$1) f(x) = \begin{cases} \frac{-3}{x^4}, & \text{si } x \geq 1\,; \\ 0, & \text{sinon} \end{cases} \qquad 2) f(x) = \begin{cases} \frac{2}{x^4}, & \text{si } x \geq 1\,; \\ 0, & \text{sinon} \end{cases} \qquad 3) f(x) = \begin{cases} xe^{-x}, & \text{si } x \geq 0\,; \\ 0, & \text{sinon} \end{cases}$$

b) Soit I l'intervalle $[1; 10]$ et f_λ la fonction définie sur I par : $f_\lambda(x) = \lambda x^{-2}$.
- Déterminer le réel λ pour lequel f est une densité de probabilité.
- Même question avec $I = [1, \infty[$.

Exercice 2 *(Fonction de Répartition, Espérance, Variance et Écart-type)*

a) Reprendre les fonctions f de l'exercice précèdent et déterminer leurs : fonctions de répartition, espérances, variances et écart-type.

b) Trouver la loi de probabilité de la $v.a.$ X dont la fonction de répartition est

$$F(x) = \begin{cases} 1 - \frac{a^3}{x^3}, & \text{si } x \geq a\,; \\ 0, & \text{si } x < a. \end{cases}$$

où a est une constante positive.

Exercice 3 *(Lecture sur la table de la loi Normale)*

a) Soit X une $v.a.$ de loi $N(0, 1)$.

1. Calculer :

1)	$P(X \leq 2.41)$	6)	$P(1.34 \leq X \leq 2.41)$
2)	$P(X \geq 1.34)$	7)	$P(-1.53 \leq X \leq 2.41)$
3)	$P(X \leq -1.72)$	8)	$P(-2.74 \leq X \leq -1.45)$
4)	$P(X \leq -1.45)$	9)	$P(\lvert X\rvert \leq 1.45)$
5)	$P(X \geq -1.53)$	10)	$P(\lvert X\rvert \geq 1.96)$

2. Déterminer x tel que :

1)	$P(X \leq x) = 0.95$	3)	$P(X \leq x) = 0.486$
2)	$P(X \geq x) = 0.239$	4)	$P(X \geq x) = 0.812$

b) Soit X une $v.a.$ de loi $N(5, 4)$. Calculer :

1)	$P(X \leq 6)$	4)	$P(X \geq -2)$
2)	$P(X \leq -1)$	5)	$P(3 \leq X \leq 7)$
3)	$P(X \geq 7)$		

c) Soit X une $v.a.$ de loi $N(\mu, \sigma^2)$ avec $\mu = 3$ et $\sigma^2 = 4$. Déterminer x tel que :

1) $P(X \leq x) = 0.95$, 2) $P(X \geq x) = 0.015$, 3) $P(X \geq x) = 0.812$.

Exercice 4 *(Lecture sur les tables statistique)*

1. Soit T une $v.a.$ d'une loi de Student de degré de liberté n ($T \rightsquigarrow t_n$). Determiner la valeur de t si :

1)	$n = 18$ et $P(T \leq t) = 0.95$	4)	$n = 25$ et $P(T \geq t) = 0.25$
2)	$n = 10$ et $P(T \leq t) = 0.80$	5)	$n = 25$ et $P(T \geq t) = 0.975$
3)	$n = 40$ et $P(T \leq t) = 0.95$		

2. Soit Y une $v.a.$ d'une loi de $Khi - Deux$ de degré de liberté m ($Y \rightsquigarrow \chi^2_m$). Determiner la valeur de y si :

1)	$m = 15$ et $P(Y \geq y) = 0.900$	3)	$m = 20$ et $P(Y \leq y) = 0.975$
2)	$m = 15$ et $P(Y \geq y) = 0.975$	4)	$m = 50$ et $P(Y \geq y) = 0.975$

3. Soit f une $v.a.$ d'une loi de $Fisher$ de degrés de libertés n, m ($f \rightsquigarrow F_{n,m}$). Determiner la valeur de f si :

1) $n = 6$, $m = 2$ et $P(F \leq f) = 0.99$ 2) $n = 20$, $m = 15$ et $P(F \geq f) = 0.05$.

Exercice 5 *(Estimation ponctuelle et intervalle de confiance)*

On admet que le taux de cholestérol chez une femme suit une loi normale $N(\mu, \sigma^2)$. Sur un échantillon de 10 femmes, on a obtenu les taux de cholestérol (en g/l) suivants :

3.0	1.8	2.1	2.7	1.4	1.9	2.2	2.5	1.7	2.0

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type du taux.
2. Déterminer un intervalle de confiance pour la moyenne du taux au seuil 1%.
3. Déterminer un intervalle de confiance pour l'écart-type du taux au seuil 5%.

Exercice 6 *(Estimation ponctuelle et intervalle de confiance)*

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. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la quantité de bicarbonate de sodium.
2. Déterminer un intervalle de confiance au seuil 5% de la moyenne de la quantité de bicarbonate de sodium.

3. Quelle devrait-être la taille n de l'échantillon pour connaître la quantité moyenne de bicarbonate de sodium à 1 mg près ?

Exercice 7 *(Estimation ponctuelle et intervalle de confiance)*

La statistique des pannes d'un distributeur de billets pendant 200 jours donne les valeurs suivantes :

Nombre de pannes x (x_i)	0	1	2	3	4	5	6
Jour avec x pannes (n_i)	32	50	52	34	19	10	3

1. Déterminer une estimation ponctuelle du nombre moyen des pannes ainsi que leurs variation.
2. Déterminer un intervalle de confiance de nombre moyen des pannes pour un risque $\alpha = 5\%$.
3. Déterminer une estimation par intervalles de confiance de l'écart-type de X pour un risque $\alpha = 5\%$.
4. Si la variance de l'échantillons σ^2 val 2 alors, quelle doit être la taille (minimale) de l'échantillons pour que la longueur de ce dernier est inférieur à 0.1.

Exercice 8 *(Test de conformité de moyenne)*

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 9 *(Test de conformité de variance)*

Le volume d'une pipette d'un type donné suit une loi normale $N(\mu, \sigma^2)$. Le fabriquant annonce un écart-type $\sigma = 0.2$ ml. Pour le vérifier, on pipette 20 fois un liquide. On observe une moyenne de 10 ml et un écart-type de 0.4 ml. Tester l'affirmation du fabricant ($\alpha = 5\% et \alpha = 1\%$).

Exercice 10 *(Test de conformité de moyenne et de variance)*

On a mesuré, après une course de 400 mètres, le pouls (en battements par minute) de 7 étudiants suivants un cours d'éducation physique :

X	83	96	99	110	130	95	74

Supposons que l'accroissement du pouls est une variable aléatoire de loi normale $N(\mu, \sigma^2)$, alors à un risque $\alpha = 5\%$, peut-ont considérer que :

1. Le nombre des pouls est inférieur à 100 battements en moyenne.
2. Le variation des pouls est différente de 300.

Exercice 11 *(Estimation par intervalle de confiance et Test de conformité de moyenne)*

On admet que le PH d'une certaine boisson alimentaire suit une loi normale $N(\mu, \sigma^2)$. Sur un échantillon de 10 bouteilles, on a obtenu les PH suivants :

8.0	6.8	7.3	7.7	6.4	6.9	8.2	7.7	6.7	7.3

On donne : $\sum_{i=1}^{n} x_i = 73$.

1. Déterminer un intervalle de confiance pour la moyenne du PH du produit au seuil $\alpha = 2\%$.
2. Déterminer un intervalle de confiance pour la variance du PH du produit au seuil $\alpha = 2\%$.

3. On désir connaître la nature de la boisson acide ($PH < 7$) , neutre ($PH = 7$) ou base ($PH > 7$).
   a) Donner la forme du test à réalisé dans ce cas.
   b) Donner la statistique du test ainsi que sa réalisation.
   c) Donner la valeur critique du test pour un seuil de risque $\alpha = 10\%$.
   d) Que peut-on conclure sur la nature de la boisson (acide, neutre ou base).

Exercice 12 *(Estimation et Test de conformité de moyenne)* Un échantillon de 15 enfants d'une ville donnée à fourni les tailles suivants (en cm) :

70	85	93	99	101	105	110	121	138	166	74	85	93	99	102

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de la tailles des enfants.
2. Déterminer une estimation par intervalles de confiance de la moyenne et de l'écart-type de la tailles des enfants pour un risque $\alpha = 10\%$.
3. Au vu de l'échantillon, peut-on considérer, au seuil de signification 2%, que la taille moyenne des enfants est de 110 *cm* ?

Exercice 13 *(Estimation et Test de conformité de moyenne)*
Afin d'analyser un certain types d'arbre, nous avons réalisés un recueil de hauteur de quelques arbres, dont les mesures sont rangées dans le tableau suivant.

X	21.1	21.1	22.1	22.4	23.3	23.3	24.0	24.3	24.5	25.0	25.9

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type de X.
2. Déterminer une estimation par intervalles de confiance de la moyenne et de l'écart-type de X pour un risque $\alpha = 10\%$.
3. Un biologiste indique que la taille moyenne des arbres en question est égale à 25 unités. Au vu de l'échantillon précédent, peut-on confirmer, au seuil de signification 2%, que ce biologiste à raison ?

Exercice 14 *(Test de Student : conformité et d'homogénéité)*
On dispose de deux échantillons d'étudiants de sexe masculin et féminin, dont on a relevé la taille. On se demande si les tailles observées peuvent être considérées comme différentes entre les deux groupes. Le tableau suivant résume les principaux caractéristiques des deux échantillons en question :

	Masculin	*Féminin*
Effectif du groupe	*11*	*10*
Moyenne	*182.43*	*168.80*
$\hat{\sigma}_c^2$	*54.95*	*26.20*

1. Peut-on dire, au seuil de signification 5%, que la taille moyenne des garçon est de 180cm ?
2. Peut-on dire, au seuil de signification 5%, que les tailles moyennes observées peuvent être considérées différentes entre les deux groupes ?
3. Au seuil $\alpha = 5\%$, peut-on dire que la taille moyenne des garçons est supérieur a la taille moyenne des filles ?

Exercice 15 *(Test de Student : conformité et d'homogénéité)*

Une laiterie produit deux types du camemberts. La masse d'un camembert tiré au hasard dans la production, par la contrôle, est distribuée selon une loi normale de moyenne $\mu = 250$ et de variance σ^2. L'agent de contrôle a tiré un échantillon simple de chaque type, dont le tableau suivant fournit les masses mesurées en g :

X	*257*	*241*	*253*	*251*	*245*	*248*	*251*	*264*	*261*	$\times$	$\times$
Y	*235*	*252*	*243*	*240*	*243*	*239*	*240*	*246*	*246*	*246*	*243*

1. L'agent de contrôle indique que, les deux types des camemberts n'ont pas la même masse moyenne. Peut-on conclure, au seuil $\alpha = 5\%$, que l'agent de contrôle a raison ?
2. L'agent indique, aussi, que la masse moyenne des camemberts de la deuxième production (Y) est inférieure à la norme. Au vu de l'échantillon précédent, au seuil de signification 5%, l'agent de contrôle aurait-il le droit de pénaliser l'entreprise ?
3. Le responsable de production réclame et dit que si l'agent prend un seuil de risque 2%, alors il constatera que la masse moyenne des camemberts de la deuxième production (Y) respecte la norme. Dans ce cas, est-ce-que l'agent de contrôle aura le droit de pénaliser l'entreprise ?

Exercice 16 *(Test d'homogénéité de Student)*

Afin de comparer deux types d'arbre, nous avons réalisés un recueil de hauteur de quelques arbres, dont les mesures sont rangées dans le tableau suivant.

							Somme
Arbre 1	23.3	24.0	24.3	24.5	25.0	25.9	147
Arbre 2	21.1	21.1	22.1	22.4	23.3		110

1. Déterminer une estimation ponctuelle de la moyenne et de la variance de chaque échantillons.
2. Supposons qu'on désir savoir si les deux types d'arbres ont la même hauteur en moyenne.
   a) Donner la forme du test à réalisé dans ce cas.
   b) Vérifier si les conditions du test sont satisfaites pour un seuil de risque $\alpha = 2\%$.
3. Si les conditions de 2.b) sont vérifiées alors :
   a) Donner la statistique du test donner dans 2.a) ainsi que sa réalisation.
   b) Donner la valeur critique associe à ce test, pour un seuil de risque $\alpha = 2\%$.
   c) Que peut-on conclure sur la hauteur moyenne des deux types d'arbres ?

Exercice 17 *(Test d'ajustement : $Khi - Deux$ et Kolmogorov-Smirnov)*

On effectue le croisement entre des pois à fleurs blanches et des pois à fleurs rouges. On obtient en deuxième génération sur 600 plantes les effectifs suivants :

Phénotype	*Rouge*	*Rose*	*Blanc*
Effectif	*141*	*325*	*134*

1. Donner les proportions théoriques de la répartition *Mendélienne* pour les trois couleurs. Calculer la statistique de test pour le test du *Khi-Deux*. Quelle est votre conclusion ?
2. Confirmer vos résultats en utilisons le test de Kolmogorov-Smirnov.

Exercice 18 *(Test d'ajustement : Khi − Deux et Kolmogorov-Smirnov)*

Un chercheur s'intéresse aux facteurs qui déterminent le choix des cours des étudiants. Il pose la question suivante à un échantillon de 50 étudiants : Parmi les 4 facteurs suivants, lequel est le plus important lorsque vous sélectionnez un cours ?. Les étudiants doivent choisir 1 des 4 propositions suivantes

1. l'intérêt pour le contenu du cours ;
2. le degré de complexité de l'examen ;
3. le professeur ;
4. l'heure à laquelle le cours se donne.

Voici les résultats que le chercheur obtient :

	Contenu cours	*Examen*	*Professeur*	*Horaire*
Effectif	*18*	*17*	*7*	*8*

Sur base de ces données, le chercheur peut-il conclure qu'un facteur (ou plusieurs facteurs) est (sont) plus important(s) que les autres ?

Exercice 19 *(Estimation ponctuelle et Test d'ajustement de Kolmogorov-Smirnov)* On désir verifier si le taux de cholestérol chez une femme suit une loi normale $N(\mu, \sigma^2)$. Un prélèvement du taux de cholestérol (en g/l) sur un échantillon de 10 femmes, nous a fourni les résultats suivants :

3.0	*1.8*	*2.1*	*2.7*	*1.4*	*1.9*	*2.2*	*2.5*	*1.7*	*2.0*

1. Estimer les paramètres μ et σ^2 de la loi normale $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
2. Poser les hypothèses à tester (hypothèse nulle et hypothèse alternative).
3. La comparaison de la distribution observée à la distribution théorique s'effectue par un test de Kolmogorov-Smirnov. Que peut-on en conclure ?

Exercice 20 *(Test d'ajustement Khi − Deux et Kolmogorov-Smirnov)*

La statistique des pannes d'un distributeur de billets pendant 200 jours donne les valeurs suivantes :

Nombre de pannes x (x_i)	0	1	2	3	4	5	6
Jours avec x pannes (n_i)	32	50	52	34	19	10	3

Tester, à un seuil de risque $\alpha = 5\%$, si les pannes suivent une distribution de poisson en utilisant :

1. Le test d'ajustement de $Khi - Deux$?
2. le test d'ajustement de Kolmogorov-Smirnov ?

N.B. Rappelons qu'on dit que X suit une loi de Poisson de paramètre λ si :

a) Sa densité est définie par :

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ pour } x \in \mathbb{N}$$

.

b) Sa fonction de répartition est définie par :

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} f(k) = \sum_{k=0}^{x} \frac{\lambda^k}{k!} e^{-\lambda} \text{ pour } x \in \mathbb{N}$$

.

c) la moyenne d'une distribution de Poisson est λ, son estimateur est définie par

$$\hat{\lambda} = \overline{X} = \frac{1}{n}\sum_{i=1}^{m} n_i X_i.$$

Exercice 21 *(Analyse de la variance à un seul facteur)*

Lors d'une expérimentation pédagogique, on désire comparer l'efficacité de quatre méthodes d'enseignement. On dispose des notes obtenues à un examen par quatre groupes d'étudiants (chaque groupe contient 25 étudiants) ayant chacun reçu un des 4 types d'enseignement a, b, c ou d. Pour répondre à l'objectif la technique statistique la plus adéquate est bien que l'ANOVA à seul facteur. L'application de cette dernière nous a fournis ce qui suit :

	SC	ddl	MC	F
Inter-groupes (Fac)			31.82	6.64
Intra-groupes (Rés)				
Total				

1. Completer la table d'analyse de variance ci-dessous.
2. A un seuil de risque $\alpha = 5\%$, que-peut-on conclure sur l'efficacité moyens des 4 méthodes.

Exercice 22 *(ANOVA 1 à un plan équilibré)*

Nous souhaitons comparer quatre traitements, notés A, B , C et D. Nous répartissons par tirage au sort les patients, et nous leur affectons l'un des quatre traitements. Nous mesurons sur chaque patient la durée, en jours, séparant de la prochaine crise d'asthme. Les mesures sont reportées dans le tableau ci-dessous :

Traitement A	*Traitement B*	*Traitement C*	*Traitement D*
36 ; 37 ; 35 ; 38 ; 41	*42 ; 38 ; 39 ; 42 ; 44*	*26 ; 26 ; 30 38 ; 34*	*42 ; 45 ; 50 ; 56 ; 58*

Pouvons-nous conclure, à un seuil de risque 1%, que les facteur traitement a une influence sur le critère retenue ? (On donne $SCT = 1324.55$.)

Exercice 23 *(ANOVA 1 à un plan non équilibré)*

On s'intéresse au rendement d'orge pour quatre variétés différentes. On dispose de quatre parcelles avec une variété d'orge pour chacune. On répète cette expérience à des endroits différents. On a obtenu :

	variété 1	variété 2	variété 3	variété 4
	46	57	50	39
	43	53	41	51
	48	43	47	45
		54	42	43
		48		
Somme	137	255	180	178

1. Calculer les estimations des moyennes μ_1, μ_2, μ_3, μ_4 et m.
2. Considérons l'hypothèse (H_0) : les rendements moyens de chaque variété sont égaux.
   a) Donner la table d'analyse de variance du problème posé, sachant que la somme des carrés résiduels $SC_{Res} = 263.667$.
   b) Que peut-on conclure sur le rendement de chaque variété (l'hypothèse H_0) à un seuil de risque $\alpha = 1\%$.

Exercice 24 *(ANOVA 2 à un plan équilibré)*

Les données sont issues d'une expérience dans laquelle la concentration de calcium dans le plasma a été mesurée chez 20 sujets des deux sexes ayant subi ou non l'administration d'un traitement hormonal.
Les données individuelles en fonction des deux traitements (h_1 et h_2) et les deux sexes (S_1 et S_2) sont résumées dans le tableau suivant

h_1		h_2	
S_1	S_2	S_1	S_2
16.5	14.5	39.1	32.0
18.4	11.0	26.2	23.8
12.7	10.8	21.3	28.8
14.0	14.3	35.8	25.0
12.8	10.0	40.2	29.3

Supposons qu'on désir tester l'influence du traitement, du sexe et les deux simultanément sur concentration de calcium.

1. Donner les hypothèses à tester.
2. Quelle techniques statistiques qui nous permet de réaliser le test citer en (1).
3. Appliquer la techniques citer en (2), pour répondre à notre objectif.

Exercice 25 *(ANOVA 2 à un plan non équilibré)*

Pour tester la possible différence entre divers laborantins dans un examen de microscopie en laboratoire, on dispose de 3 préparations différentes. Chaque laborantin compte le nombre d'un certain type de cellules dans chaque préparation. On obtient le tableau suivant :

	Laborantins		
Préparations	1	2	3
A	48 51	44	75 63
B	62	58 57 69	75
C	57 54	55	83

Y a-t-il une différence significative entre les laborantins ?

Exercice 26 *(ANOVA 2 à un plan sans répétitions)*

Supposons que lors d'une étude statistique d'un certain phénomène, nous nous somme intéressés à l'influence de deux facteurs F_1 (ayant *4 modalités*) et F_2 (ayant *5 modalités*) sur un caractère quantitatif X. Pour cela nous avons utilisé l'ANOVA 2 dont certains résultats, fournis par cette méthode, sont donnés comme suit :
$SC_{tot} = 1350, CM_{F_1} = 140$ et le ratio du premier facteur $f_c = 3.5$.

Si l'expérience réaliser pour répondre à notre objectif est *sans répétitions* alors :

1. Donner et compléter la table d'ANOVA correspondante au problème.
2. Que peut-on conclure sur l'effet des facteurs sur la variable X, à un seuil de risque $\alpha = 5\%$?

Exercice 27 *(ANOVA 2 à un plan sans répétitions)*

Trois équipes (matin, midi, soir) se relaient sur une chaîne de montage. Elles occupent quatre post de travail A, B, C et D. Sur la production d'un mois, on note le nombre total de pièces défectueuses ventilées par équipe et par post de travail :

	post				
équipe	A	B	C	D	moyenne
équipe du matin	26	13	35	6	20
équipe du soir	18	17	31	2	17
équipe de nuit	31	24	33	4	23
moyenne	25	18	33	4	20

On souhaite interpréter ce tableau à l'aide d'une ANOVA.

1. Dans un premier temps, on analyse les deux facteurs équipe et post séparément.
   (a) Peut-on affirmer l'existence d'une différence entre les performances globales des équipes ?
   (b) Peut-on conclure que les post présentent des difficultés de montage inégales ?
2. On désire maintenant tenir compte simultanément des deux facteurs. On modélise le tableau par un modèle additif d'analyse de la variance à deux facteurs :

$$Y_{ij} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij},$$

   où les variables aléatoires, ϵ_{ij} sont indépendantes, normales centrées de même variance σ^2
   (a) Serait-il pertinent de modéliser le tableau par un modèle complet à deux facteurs ?
   (b) Y-a-t-il un effet "post de travail" ?
   (c) Y-a-t-il un effet "équipe" ?
   (d) Y-a-t-il un effet "équipe" et "post de travail" ? que peut-on conclure ?
   (e) Comparer les résultats obtenus dans 1 et 2, que peut-on conclure ?

Exercice 28 *(ANOVA 2 à un plan sans répétitions)*

Un botaniste veut déterminer si la présence d'insectes a un effet sur la fécondité des plantes dans un champ. Afin d'empêcher les insectes d'attaquer ses plantes, le botaniste a l'idée d'installer des cages au-dessus des quadrants expérimentaux. Il propose d'utiliser trois traitements : contrôle (plantes non couvertes), plantes recouvertes de cages laissant les insectes entrer et plantes recouvertes de cages empêchant les insectes d'atteindre les plantes. Pour s'assurer que les différences qu'il observe à la fin de l'étude sont dues aux manipulations et non à un effet quelconque des propriétés des cages utilisées, il utilise 3 quadrants par traitement, et il échantillonne 6 plantes par quadrant. La variable mesurée est la fécondité moyenne (le nombre moyen de graines produit par les 6 plantes). Les données se trouvent ci-dessous.

		Contrôle	Cages fermées	Cages ouvertes
Quadrants	1	78.4	71.2	76.2
	2	75.2	61.8	77.3
	3	83.5	68.5	77.7

Question : A l'aide d'une technique statistique adéquate, indiquer si la présence d'insectes ainsi que les cages ont un effet significatif sur la fécondité moyenne des plantes ou non.

Exercice 29 *(Test d'indépendance de Khi − Deux)*

En vue de comparer deux traitements T_1 et T_2 d'une affection bénigne, on répartit entre ces deux traitements 250 malades par tirage au sort. Les résultats, sur l'état du malade après 5 jours de traitement, sont indiqués dans le tableau ci-dessous.

	État du malade après 5 jours		
Traitement	Stationnaire	Amélioré	Guéri
T_1	15	70	35
T_2	25	85	20

Les deux variables Traitement et l'état du malade après 5 jours de traitement sont ils indépendantes ?

Exercice 30 *(Test d'indépendance de Khi − Deux)*

On s'intéresse à l'association entre le mode de vie, "seul" ou "en famille", et la présence ou l'absence d'une névrose. Dans un échantillon aléatoire d'individus d'une certaine population on a trouve les fréquences ci-dessous :

	Névrose		
Mode de vie	Présente	Absente	Total
En famille	40	60	100
Seul	100	60	160
Total	140	120	260

Question : Peut-on rejeter, au seuil de 1%, l'hypothèse de non association entre le mode de vie et la présence d'une névrose ?

Exercice 31 *(Test d'indépendance de Khi − Deux)*

Lors de l'interrogation du module "*Analyse de Données en Biosciences*", des troisièmes années Licence option Biologie et physiologie végétale qui contiens trois groupes ($G01$, $G02$ et $G03$), dans la grande salle numéro 1 qui contiens quatre (04) rangées, le responsable de module s'est posé la question suivante :
Est-ce-que les étudiants choisissent le rang pour s'asseoir selon leurs groupes ou non ?
Après l'analyse de la situation l'enseignant a conclu que le choix du rang est indépendant du groupe. Indiquer, pour un seuil de risque $\alpha = 5\%$, si l'enseignant à raison ou non. Sachant que la répartition des étudiants dans la salle selon leurs groupes est comme suite :

Groupe	Rang 1	Rang 2	Rang 3	Rang 4	$\sum$
Groupe 1	4	6	8	6	24
Groupe 2	4	3	5	5	17
Groupe 3	6	5	3	2	16
$\sum$	14	14	16	13	57

Exercice 32 *(Test d'indépendance de Khi − Deux)*

Afin de comparer l'action de deux types de levures (A et B) sur une pâte à gâteaux, on prélève, pour chacune des levures, un échantillon aléatoire de gâteaux. L'aptitude des pâtes à lever est définie par les critères suivants : Moyenne, Bonne et Très bonne. Les résultats constatés sont rassemblés dans le tableau suivant :

Aptitude à lever	Moyenne	Bonne	Très bonne
A	41	16	63
B	22	27	51

Quel est le test statistique adéquat pour déterminer s'il y a une différence d'activité des deux levures ? À l'aide de ce test, au risque de 5%, peut-on conclure à une différence d'activité des deux levures ?

Exercice 33 *(ANOVA 1 et test d'indépendance de Khi − Deux)*

Trente sept étudiants d'une promotion ont été répartis, en début d'année académique, de manière strictement aléatoire dans trois séries de travaux pratiques de statistique dirigés par trois assistants différents $A1$, $A2$ et $A3$. Les résultats obtenus par les étudiants de chaque série sont notés sur 10 et regroupés dans le tableau suivant.

Assistant	*Note des étudiants (sur 10)*												
$A1$	9	5.5	6	3	3	2	7	5	0	8	4.5	7	×
$A2$	4	3	6	8	2	3	5	7	7	4.5	3.5	0	×
$A3$	8	6	4	8	10	4	4.5	5	7	8	10	9	6

Y-a-t-il un effet d'appartenance sur le niveau des étudiants ?

Afin d'étudier l'indépendance des résultats par rapport à la série d'appartenance de l'étudiant, un chercheur fait un décompte en termes de nombre de réussites et d'échecs par série de travaux pratiques.
Quel test que le chercheur doit utiliser dans ce cas ? Réaliser ce test sur les présentes données.

Exercice 34 *(Régression linéaire simple)*

Dans le cadre de travaux de recherche sur la *Biomasse* (mg), d'un certain type de plante, en fonction de la concentration de l'Azote NH_4^+ (μmol), nous avons réalisé des expériences dont la biomasse moyenne (Y) ainsi que la concertation du l'Azote (X) en question sont données dans le tableau ci-dessus :

Concentration μmol	0	100	200	400	600
Biomasse mg	305	378	458	540	565

On donne : $\sum x_i = 1300$; $\sum y_i = 2246$; $\sum x_i^2 = 570000$; $\sum y_i^2 = 1056498$; $\sum x_i y_i = 684400$;

Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y \ = \ a \ + \ bx.$$

1. Présenter graphiquement le nuage des points (X_i, Y_i). Que peut-on conclure sur le modèle proposer ?
2. Calculer les estimations des paramètres a et b et donner la droite de régression.
3. Calculer le coefficient de corrélation linéaire. Que peut-on conclure ?
4. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?
5. Quelle Biomasse prévoyez-vous à une concentration 500 μmol ?

Exercice 35 *(Régression linéaire simple)*

Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne (variable Y en degrés Celsius) ainsi que l'altitude (variable X en mètres) de chaque station données dans le tableau ci-dessous :

altitude	*1040*	*1230*	*1500*	*1600*	*1740*	*1950*	*2200*	*2530*	*2800*	*3100*
température	*7.4*	*6*	*4.5*	*3.8*	*2.9*	*1.9*	*1*	*-1.2*	*-1.5*	*-4.5*

On donne : $\sum x_i = 19690$; $\sum y_i = 20.3$; $\sum x_i^2 = 42925500$; $\sum y_i^2 = 162.41$; $\sum x_i y_i = 17671$;

1. Calculer le coefficient de corrélation linéaire.
2. Calculer les estimations des paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Quelle température moyenne prévoyez-vous à 1100 m ? à 2300 m ?

Exercice 36 *(Régression linéaire simple)*

Dans le cadre d'une enquête visant à comparer, selon certains critères, différents *Sandwich* vendus dans les fast-foods, nous avons retenu les informations se trouvant dans le tableau ci-dessous.

Sandwich	S_1	S_2	S_3	S_4	S_5	S_6
Poids (g)	150	92	193	90	135	169
Prix (DA)	190	140	270	90	180	130

On donne :
$\sum Poids = 829$; $\sum Poids^2 = 123099$; $\sum Prix = 1000$; $\sum Prix^2 = 186000$; et
$\sum Poids * Prix = 147860$;

1. Y-a-t-il une relation linéaire entre les variables Poids et Prix ? Pour répondre à cette question, faire un graphique, calculer le coefficient de corrélation des deux variables et l'équation de la droite de régression.
2. Supposons qu'on désir d'augmenter le poids du Sandwich S_6 à 180 g, alors quelle sera son nouveau prix ?
3. Le modèle linéaire proposé est-il adéquat pour la description de la relation entre les variables Poids et Prix ?

Exercice 37 *(Régression linéaire simple et transformation des variables)*

On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $X = \ln(D)$ et $Y = \ln(H)$. Voici les mesures faites sur 5 arbres :

D	*0.1999*	*0.3012*	*0.3791*	*0.6005*	*0.6570*
H	*9.2073*	*9.6794*	*10.8049*	*13.4637*	*14.1540*

1. Donner le coefficient de corrélation linéaire entre X et$Y$.
2. Donner l'équation de la droite de régression de Y par rapport à X.
3. Tester la pertinence de la régression au seuil de 5%.
4. Donner la hauteur prévue d'un arbre de diamètre 0.7.

Exercice 38 *(Régression linéaire simple et changement des variables)*

Dans le cadre de travaux de recherche sur l'absorbance, d'un produit en fonction de sa concentration, par une certaine plante, nous avons réalisé des expériences dont l'absorbance moyenne (Y) ainsi que la concertation du produit (x) en question sont données dans le tableau ci-dessus :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.205	0.331	0.515	0.584	0.671	2.3060

a) Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y = a_1\, x + b_1.$$

1. Calculer les estimations des paramètres a_1 et b_1 et donner la droite de régression.
2. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$? Que peut-on conclure ?
3. Calculer le coefficient de corrélation linéaire, ce résultat confirme t-il les résultats obtenue en 3) ?
4. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?

b) Vue les doutes qu'on a sur le modèle précèdent, nous avons proposé le modèle suivant :

$$Z \;=\; e^{Y} \;=\; a_2\, x \;+\; b_2.$$

1. Complétez le tableau suivant :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Z	1.0000						

2. Calculer les estimations des paramètres a_2 et b_2 pour la régression linéaire de Z sur X.
3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$. Que peut-on conclure par rapport au premier modèle ?
4. Calculer le coefficient de corrélation linéaire de ce nouveau modèle.
5. Indiquer quel est le meilleur modèle parmi les deux proposés (avec justification).

6.2 Solution des exercices

Solution 1 *(Densité de probabilité)*

a) $f(x)$ est une densité si et seulement si :

$$\begin{cases} f(x) \geq 0, & \forall x \in \mathbb{R}\,; \\ \int\limits_{-\infty}^{+\infty} f(x)dx = 1. \end{cases}$$

alors,

1. $f(x) = \frac{-3}{x^4}$ n'est pas une densité car : $\exists\, x \in [1, +\infty[\ f(x) < 0$.
2. $f(x) = \frac{2}{x^4}$ n'est pas une densité car : $\int_1^{+\infty} \frac{2}{x^4}dx = 2/3 \neq 1$.
3. $f(x) = xe^{-x}$ est une densité le fait que :
   $\forall x \in [0, +\infty[\ \ f(x) \geq 0$ et $\int\limits_0^{+\infty} xe^{-x}dx = [-xe^{-x}]_0^{+\infty} - \int\limits_0^{+\infty} e^{-x}dx = 1$ (Intégration par partie).

b) 1. $f(x) \geq 0$ pour toute x alors pour quelle soit une densité il suffit que :
   $\int\limits_1^{10} \lambda x^{-2}dx = [-\lambda x^{-1}]_1^{10} = 1$ c'est-à-dire $9\lambda/10 = 1 \Rightarrow \lambda = 10/9$.

2. De la même manière, pour $I = [1, \infty[$ on obtient $\lambda = 1$ car $\int\limits_1^{+\infty} \lambda x^{-2}dx = \lambda$.

Solution 2 *(Fonction de Répartition, Espérance, Variance et Écart-type)*

Fonction de répartition : Par définition $F(x) = \int_{-\infty}^{x} f(t)dt$.

1. Pour $x \in [1, +\infty]$, on a : $\int_0^x te^{-t}dt = 1 - (x+1)e^{-x}$ alors,

$$F(x) = \begin{cases} 0, & \text{si } x < 0\,; \\ 1 - (x+1)e^{-x}, & \text{si } x \geq 0. \end{cases}$$

2. Pour $x \in [1, 10]$, on a : $\lambda = 10/9$ et $\int_1^x (10/9)t^{-2}dt = \frac{10(x-1)}{9x}$ alors,

$$F(x) = \begin{cases} 0, & \text{si } x < 1\,; \\ \frac{10(x-1)}{9x}, & \text{si } 1 \leq x \leq 10\,; \\ 1, & \text{si } x \geq 10. \end{cases}$$

3. Pour $x \in [1, +\infty]$, on a : $\lambda = 1$ et $\int_1^x t^{-2}dt = \frac{x-1}{x}$ alors,

$$F(x) = \begin{cases} 0, & \text{si } x < 1\,; \\ \frac{x-1}{x}, & \text{si } x \geq 1. \end{cases}$$

Espérance : Par définition $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$.

1. $E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{0} x \times 0 \; dx + \int_0^{+\infty} x \times (xe^{-x})dx = \int_0^{+\infty} x^2e^{-x}dx = [-x^2e^x]_0^{\infty} + 2\int_0^{+\infty} xe^{-x}dx = 2.$
2. $E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{1} x \times 0 \; dx + \int_1^{10} x \times (10/9x^{-2})dx + \int_{10}^{+\infty} x \times 0 \; dx = \int_1^{10} (10/9)x^{-1} \; dx = [(10/9)\log(x)]_1^{10} = \frac{10\log(10)}{9}$.
3. $E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{1} x \times 0 \; dx + \int_1^{+\infty} x \times (x^{-2})dx = \int_1^{+\infty} x^{-1} \; dx = [\log(x)]_1^{+\infty} = +\infty$.

Variance : Par définition $Var(X) = E(X^2) - E(X)^2 = \int_{-\infty}^{+\infty} x^2f(x)dx - E(X)^2$.

1. $E(X^2) = \int_{-\infty}^{+\infty} x^2f(x)dx = \int_{-\infty}^{0} x^2 \times 0 \; dx + \int_0^{+\infty} x^2 \times (xe^{-x})dx = \int_0^{+\infty} x^3e^{-x}dx = [-3x^2e^x]_0^{\infty} + 3\int_0^{+\infty} x^2e^{-x}dx$
   $E(X^2) = 3 \times 2 = 6$ alors $Var(X) = 6 - 2^2 = 2$.
2. $E(X^2) = \int_{-\infty}^{+\infty} x^2f(x)dx = \int_{-\infty}^{1} x^2 \times 0 \; dx + \int_1^{10} x^2 \times (10/9x^{-2})dx + \int_{10}^{+\infty} x^2 \times 0 \; dx = \int_1^{10} (10/9) \; dx$
   $E(X^2) = [(10/9)x]_1^{10} = 10.$ alors $Var(X) = 10 - \left(\frac{10\log(10)}{9}\right)^2$.
3. $E(X^2) = \int_{-\infty}^{+\infty} x^2f(x)dx = \int_{-\infty}^{1} x^2 \times 0 \; dx + \int_1^{+\infty} x^2 \times (x^{-2})dx = \int_1^{+\infty} 1 \; dx = [x]_1^{+\infty} = +\infty.$
   alors $Var(X) = \infty - (\infty)^2 \Rightarrow Var(x) \; \nexists$.

b) Par définition on a : $f(x) = \frac{\partial F(x)}{\partial x}$, alors $\frac{\partial\left(1-\frac{a^3}{x^3}\right)}{\partial x} = \frac{3a^3}{x^2} \Longrightarrow f(x) = \begin{cases} 0, & \text{si } x < a\,; \\ \frac{3a^3}{x^2}, & \text{si } x \geq a. \end{cases}$

Solution 3 *(Lecture sur la table de la loi Normale)*

a) La résolution de cet exercice nécessite la table de la loi normale centrée et réduite.

1. La quantification des probabilités dans ces situations se fait par la lecture sur la table de la loi normale centrée et réduite, et cela soit par une lecture directe sur la table de la loi normale, soit après une transformations adéquate en se basant sur les propriétés de la loi normale.
   - $P(X \leq 2.41) = 0.99202$ (obtenue directement sur la table).
   - $P(X \geq 1.34) = 1 - P(X \leq 1.34) = 1 - 0.9099 = 0.0901$
   - Pour déterminer la valeur de $P(X \leq -1.72)$ sur la table de la loi normale centrée et réduite, on doit réaliser d'abord la transformation de l'écriture. On se basant sur la propriété de symétrie de la loi normale et la propriété de probabilité totale, on obtient ce qui suit :
   $$P(X \leq -1.72) = P(X \geq 1.72) = 1 - P(X \leq 1.72),$$
   et par une lecture directe sur la table de la loi normale centrée et réduite on aura $P(X \leq 1.72) = 0.95728$ ainsi $P(X \leq -1.72) = 1 - 0.95728 = 0.0427$.
   - Avec le même raisonnement que dans l'exemple precedent on obtient : $P(X \leq -1.45) = 1 - P(X \leq 1.45) = 1 - 0.9265 = 0.0735$.
   - On utilisons la propriété de la symétrie de la loi normale on aura : $P(X \geq -1.53) = P(X \leq 1.53) = 0.9370$.
   - $P(1.34 \leq X \leq 2.41) = P(X \leq 2.41) - P(X \leq 1.34) = 0.99202 - 0.9099 = 0.0821$.
   - $P(-1.53 \leq X \leq 2.41) = P(X \leq 2.41) - P(X \leq -1.53) = P(X \leq 2.41) - [1 - P(X \leq 1.53)] = P(X \leq 2.41) + P(X \leq 1.53) - 1 = 0.99202 + 0.9370 - 1 = 0.9290$.
   - $P(-2.74 \leq X \leq -1.45) = P(1.45 \leq X \leq 2.74) = P(X \leq 2.74) - P(X \leq 1.45) = 0.9969 - 0.9265 = 0.0705$.
   - $P(|X| \leq 1.45) = P(-1.45 \leq X \leq 1.45) = P(X \leq 1.45) - P(X \leq -1.45) = P(X \leq 1.45) - [1 - P(X \leq 1.45)] = 2 * P(X \leq 1.45) - 1 = 2 * 0.9265 - 1 = 0.8529$.
   - $P(|X| \geq 1.96) = P(X \leq -1.96) + P(X > 1.96) = [1 - P(X \leq 1.96)] + [1 - P(X \leq 1.96)] = 2 - 2P(X \leq 1.96) = 2 - 2 * 0.975 = 0.05$.
2. La détermination de la valeur de x dans ces situation se fait par la lecture inverse sur la table de la loi normale centrée et réduite.
   - $P(X \leq x) = 0.95 \Rightarrow x = 1.64$ (obtenue directement sur la table).
   - $P(X \geq x) = 0.239 \Rightarrow 1 - P(X \leq x) = 0.239$ c'est-à-dire $P(X \leq x) = 1 - 0.239 = 0.7610 \Rightarrow x = 0.71$
   - $P(X \leq x) = 0.486$ on a $P(X \leq x) \leq 0.5 \Rightarrow x \leq 0$ et $P(X \leq x) = 1 - P(X \leq -x) \Rightarrow 1 - P(X \leq -x) = 0.486 \Rightarrow P(X \leq -x) = 0.5140 \Rightarrow -x = 0.04 \Rightarrow x = -0.04$
   - $P(X \geq x) = 0.812 \Rightarrow P(X \leq -x) = 0.812 \Rightarrow -x = 0.89 \Rightarrow x = -0.89$.

b) Le calcul des probabilités dans ce cas nécessite la transformation des variables. En effet, avant de lire sur la table de la loi normale centrée et réduite il faut posé $Z = \frac{X-\mu}{\sigma}$, tel que $\mu = 5$ et $\sigma = 2$.

1. $P(X \leq 6) = P(\frac{X-\mu}{\sigma} \leq \frac{6-\mu}{\sigma}) = P(\frac{X-5}{2} \leq \frac{6-5}{2}) = P(Z \leq 0.5) = 0.69$.
2. $P(X \leq -1) = P(\frac{X-\mu}{\sigma} \leq \frac{-1-\mu}{\sigma}) = P(\frac{X-5}{2} \leq \frac{-1-5}{2}) = P(Z \leq -3) = 1 - P(Z \leq 3) = 0.0013$.

3. $P(X \geq 7) = P(\frac{X-\mu}{\sigma} \geq \frac{7-\mu}{\sigma}) = P(\frac{X-5}{2} \geq \frac{7-5}{2}) = P(Z \geq 1) = 1 - P(Z \leq 1) = 0.1587$.
4. $P(X \geq -2) = P(\frac{X-\mu}{\sigma} \geq \frac{-2-\mu}{\sigma}) = P(\frac{X-5}{2} \geq \frac{-2-5}{2}) = P(Z \geq -3.5) = P(Z \leq 3.5) = 0.9998$.
5. $P(3 \leq X \leq 7) = P(\frac{3-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{7-\mu}{\sigma}) = P(\frac{3-5}{2} \leq \frac{X-5}{2} \leq \frac{7-5}{2}) \Rightarrow P(3 \leq X \leq 7) = P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = 2P(Z \leq 1) - 1 = 2 * 0.8413 - 1 = 0.6827$.

c) Soit X une *v.a.* de loi $N(\mu, \sigma^2)$ tel que $\mu = 3$ et $\sigma^2 = 4$.
Pour déterminer les valeurs de x tel que : $P(X \leq x) = 0.95$, $P(X \geq x) = 0.015$ et $P(X \geq x) = 0.812$, on ne peut pas exploiter d'une manière directe la table de la loi normale centrée et réduite et cela le fait que X suit une loi normale qui n'est ni centrée et ni réduite. A cet effet, pour répondre à la question on doit effectué les trois étapes suivante :

1. Réaliser une transformation de la *v.a.* X vers une *v.a.* Y qui suit une loi $N(0,1)$, en posant $y = \frac{X-\mu}{\sigma}$.
2. Par une lecture inverse sur la table de la loi normale centrée et réduite on détermine la valeur de y.
3. Par une transformation inverse, on détermine la valeur de x, c'est-à-dire $x = \sigma y + \mu$.

$\underline{P(X \leq x) = 0.95 :}$
• $P(X \leq x) = 0.95 \Leftrightarrow P\left(\frac{X-3}{2} \leq \frac{x-3}{2}\right) = 0.95 \Leftrightarrow P(Y \leq y) = 0.95$.
• A partir de la table de la loi $N(0,1)$ par une lecture inverse on obtiens 1.645
• Alors $x = 2 * y + 3 = 2 * 1.645 + 3 = 6.2897$.
$\underline{P(X \geq x) = 0.015 :}$
•$P(X \geq x) = 0.015 \Leftrightarrow P(\frac{X-\mu}{\sigma} \geq \frac{x-3}{2}) = 0.015 \Leftrightarrow P(Y \geq y) = 0.015 \Leftrightarrow P(Y \leq y) = 0.985$
• à partir de la table de la loi $N(0,1)$ par une lecture inverse on obtiens $y = 2.17$
• Alors $x = 2 * 2.17 + 3 = 7.34$
$\underline{P(X \geq x) = 0.812 :}$
• avec le même raisonnement que dans le premier cas on obtiens : $P(Y \leq y) = 0.188 \leq 0.5$ cela signifié que y est négatif alors pour réaliser une lecture sur la table $N(0,1)$ la transformation suivante est nécessaire $P(Y \leq y) = 0.188 \Leftrightarrow P(Y \leq -y) = 0.812$.
• à partir de la table de la loi $N(0,1)$ par une lecture inverse on obtiens $-y = 0.89 \Rightarrow y = -0.89$
• Alors $x = 2 * (-0.89) + 3 = 1.22$.

Solution 4 *(Lecture sur les tables statistique)*

1. T une *v.a.* d'une loi de Student de degré de liberté n ($T \rightsquigarrow t_n$).
   - On a $n = 18$ et $P(T \leq t) = 0.95 = 1 - \frac{0.1}{2}$, alors a partir de la table de la loi de Student dans l'intersection de la ligne 18 et la colonne $P = 0.10$ on aura $t = t(18, 1 - \frac{0.1}{2}) = 1.734$.
   - On a $n = 10$ et $P(T \leq t) = 0.80 \Rightarrow t = t_{(10,0.80)} = t_{(10,1-0.2)}$, alors a partir de la table de la loi de Student dans l'intersection de la ligne 10 et la colonne $P = 0.40$ on aura $t = t_{(10,1-0.2)} = 0.879$.
   - On a $n = 40$ $P(T \leq t) = 0.95 = 1 - \frac{0.1}{2}$ alors, a partir de la table de la loi de Student dans l'intersection de la ligne $n = \infty$ et la colonne $P = 0.10$ on aura $t = 1.645$.
   - On a $P(T \geq t) = 0.25$, dans cette situation, avant de lire sur la table de la loi de Student on réalise la transformation suivante : $P(T \geq t) = 0.25 \Rightarrow P(T \leq t) = 1 - 0.25$. A partir de la table de la loi de Student dans l'intersection de la ligne 25 et la colonne $P = 0.50$ on aura $t = t_{(10,0.75)} = t_{(10,1-0.25)} = 0.684$.
   - On a $n = 25$ et $P(T \geq t) = 0.975$, le fait que $P \leq 0.5$ cela signifier que t est négative ($t < 0$), alors $P(T \geq t) = 0.975 \Rightarrow P(T \leq -t) = 0.975$(propriété de symétrie de

la loi de Student) c'est-à-dire $P(T \leq -t) = 1 - \frac{0.05}{2}$. Ainsi, à partir de la table de la loi de Student dans l'intersection de la ligne 25 et la colonne $P = 0.05$ on aura $-t = t(25, 1 - \frac{0.05}{2}) = 2.060 \Rightarrow t = -2.060$.

2. Y une $v.a.$ d'une loi de $Khi - Deux$ de degré de liberté m ($Y \rightsquigarrow \chi^2_m$).

   - Pour determiner la valeur de y pour le cas $m = 15$ et $P(Y \geq y) = 0.90$ il suffit de lire directement la valeur qui se trouve dans l'intersection de la ligne $m = 15$ et la colonne $p = 0.90$ où on aura $y = 8.547$.
   - On a $m = 15$ et $P(Y \geq y) = 0.975$, avec la même démarche que le cas précédent on aura $y = 6.262$.
   - Pour determiner la valeur de y pour le cas $m = 20$ et $P(Y \leq y) = 0.975$, il faut réaliser d'abord la transformation suivante : $P(Y \leq y) = 0.90 \Rightarrow P(Y \geq y) = 1 - 0.975 = 0.025$. Maintenant, pour obtenir la valeur de y, il suffit de lire la valeur située dans l'intersection de la ligne 20 et la colonne $p = 0.025$ où on obtient $y = 34.170$.
   - On a $m = 50$ et $P(Y \geq y) = 0.975$ dans cette situation on ne peut pas lire sur la table de la loi de $Khi - Deux$ mais plutôt c'est à partir de la table de la loi normale que y sera déterminer. En effet, afin de déterminer la valeur de y, il faut faire recoure à l'approximation de loi de $Khi - Deux$ par une loi normale centrée réduite et cela toute en utilisant la transformation suivante, $z = \sqrt{2 * Y} - \sqrt{2 * m - 1} \leq \sqrt{2 * y} - \sqrt{2 * m - 1}$ (z suit une loi normale centrée réduite), avec $P(Z \leq z) = 0.975$. A partir de la table de la loi normale on aura $z = 1.96 \Rightarrow \sqrt{2 * y} - \sqrt{2 * m - 1} = 1.96$ d'où $y = 70.9226$.

3. Soit f une $v.a.$ d'une loi de $Fisher$ de degrés de libertés n, m ($f \rightsquigarrow F_{n,m}$).

   - A partir de la table de la loi de Fisher $P = 0.99$ (troisième table), dans l'intersection de la colonne $n = 6$ et de la ligne $m = 2$, on a $f = F_{(6,2,0.99)} = 99.3$.
   - Avant de lire sur la table de la loi de Fisher on doit réaliser la transformation suivante : $P(F \geq f) = 0.05 \Rightarrow P(T \leq t) = 0.95$. A partir de la table de la loi de Fisher $P = 0.95$ (première table), dans l'intersection de la ligne $m = 15$ et la colonne $n = 20$, on a $f = F_{(20,15,0.95)} = 2.33$.

Solution 5 *(Estimation ponctuelle et intervalle de confiance)*

On admet que le taux de cholestérol chez une femme suit une loi normale $N(\mu, \sigma^2)$. Sur un échantillon de 10 femmes, on a obtenu les taux de cholestérol (en g/l) suivants :

3.0	1.8	2.1	2.7	1.4	1.9	2.2	2.5	1.7	2.0

1. Déterminer une estimation ponctuelle de la moyenne et de l'écart-type du taux.
   $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 2.13$ et $\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = 0.2357 \Rightarrow \hat{\sigma}_c = \sqrt{0.2357} = 0.4855$.
2. Déterminer un intervalle de confiance pour la moyenne du taux au seuil 1%.
   On remarque que la taille de l'échantillon est inférieur à 30, et X suit une loi normale donc l'intervalle de confiance de la moyenne à un seuil de confiance $1 - \alpha$=99% est défini comme suite :

$$IC_{1-\alpha} = \left[\hat{m} - t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}; \hat{m} + t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}\right]$$

$$\Rightarrow IC_{1-\alpha} = \left[2.13 - t_{(9,1-0.01/2)}\sqrt{\frac{0.2357}{10}}; 2.13 + t_{(9,1-0.01/2)}\sqrt{\frac{0.2357}{10}}\right].$$

   Sur la table de la loi de Student, on obtient $t_{(9,1-0.005)} = 3.25$ alors l'intervalle de confiance de la moyenne à un seuil de risque $\alpha = 1\%$ est donné par : [1.6310 ; 2.6290].

3. Déterminer un intervalle de confiance pour l'écart-type du taux au seuil 5%.
   Dans ce cas, on a affaire à la determination d'un intervalle de confiance d'une variance σ^2 d'une variable aléatoire issue d'une loi normale d'espérance μ inconnue, donc cet intervalle est défini comme suite :
$$(n-1)\frac{\sigma_c^2}{b_\alpha} < \sigma^2 < (n-1)\frac{\sigma_c^2}{a_\alpha} \tag{6.1}$$
   où a_α et b_α sont déterminer par :
$$\alpha_1 = P\left\{\chi^2_{n-1} < a_\alpha\right\} \text{ et } \alpha_2 = P\left\{\chi^2_{n-1} > b_\alpha\right\} \tag{6.2}$$
   avec $\alpha_1 + \alpha_2 = \alpha$.
   Supposons qu'on fixe $\alpha_1 = \alpha_2 = \alpha/2 = 0.025$, alors
   par la lecture sur la table de $Khi - Deux$ à un $ddl = 9$ au seuil 0.975, la valeur de $a_\alpha = 2.7$.
   par la lecture sur la table de $Khi - Deux$ à un $ddl = 9$ au seuil 0.025, la valeur de $b_\alpha = 19.023$.
   donc,
$$9\ \frac{0.2357}{19.023} < \sigma^2 < 9\ \frac{0.2357}{2.7} \Rightarrow 0.1115 < \sigma^2 < 0.7857$$

Solution 6 *(Estimation ponctuelle et intervalle de confiance)*

1. L'estimation ponctuelle de la moyenne et de l'écart-type de la quantité de bicarbonate de sodium sont donnés par :
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 1625.4667,$$
   et le fait que la vraie moyenne est connue ($\mu = 1625$) alors l'estimateur de la variance est donnée par : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - 1625)^2$ c'est-à-dire la variance de l'échantillon est $\hat{\sigma}^2 = 21.9167$ d'où
$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - 1625)^2} = \sqrt{21.9167} = 4.6815.$$

2. On à la taille de l'échantillon supérieur à 30, alors l'intervalle de confiance de la moyenne à un seuil de confiance 95% est défini comme suite :
$$IC_{1-\alpha} = \left[\hat{m} - u_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}; \hat{m} + u_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right]$$
$$\Rightarrow IC_{95\%} = \left[1625.5 - 1.96\sqrt{\frac{21.9167}{150}}; 1625.5 + 1.96\sqrt{\frac{21.9167}{150}}\right];$$
   d'où l'intervalle de confiance de la moyenne à un seuil de risque $\alpha = 5\%$ est donné par :
$$[1624.75\ ;\ 1626.25].$$

3. Pour connaître la quantité moyenne de bicarbonate de sodium à 1 mg près, il faut que $2u_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq 1$, c'est-à-dire,
$$u_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq 1/2 \Rightarrow u^2_{\alpha/2}\frac{\sigma^2}{n} \leq 1/4 \Rightarrow n \geq 4u^2_{\alpha/2}\sigma^2,$$
   alors pour α fixé à 5% on aura $n \geq 4 * 1.96^2 * 22 = 338.0608$, et le fait que $n \in \mathbb{N}$ alors on prend $n_0 = 339$.

Solution 7 *(Estimation ponctuelle et intervalle de confiance)*

1. Calcule de la moyenne et de la variance :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{7} n_i X_i = \frac{1}{200}\left(32\times 0 + 50\times 1 + ... + 3\times 6\right) = 2. \tag{6.3}$$

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{7} n_i (X_i - \overline{X})^2 = \frac{1}{199}\sum_{i=1}^{7} n_i (X_i - 2)^2 = 2.1407 \tag{6.4}$$

$$\hat{\sigma}_c = \sqrt{\hat{\sigma}_c^2} = 1.4631. \tag{6.5}$$

2. Intervalle de confiance de la moyenne :
Sachant que la taille de l'échantillon est assez grand $n = 200 > 30$ et on ne connaît pas la distribution de l'échantillon alors l'IC est définis comme suit :

$$IC_{1-\alpha}(\mu) = \left[\overline{X} - U_\alpha \frac{\hat{\sigma}_c}{\sqrt{n}} \ ; \ \overline{X} + U_\alpha \frac{\hat{\sigma}_c}{\sqrt{n}}\right].$$

où U_α est le fractile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée et réduite. Sur la table de la loi normale on obtient $U_\alpha = 1.96$ alors :

$$IC_{95\%}(\mu) = [1.7972 \ ; \ 2.2028].$$

3. Intervalle de confiance de l'écart-type :
on L'IC d'une variance est définie comme suit :

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)\sigma_c^2}{b_\alpha} \ ; \ \frac{(n-1)\sigma_c^2}{a_\alpha}\right],$$

avec a_α et b_α sont des fractiles d'une loi de $Khi - Deux$ de de gré de liberté $n-1$, respectivement, d'ordre $\alpha/2$ et $1-\alpha/2$.
Sur la table de $Khi - Deux$ on obtient :
$a_\alpha = \chi_{(199,1-0.05/2)} = 161.826$ et $b_\alpha = \chi_{(199,0.05/2)} = 239.960$, donc

$$IC_{1-\alpha}(\sigma^2) = [1.7753 \ ; \ 2.6325]$$

d'où

$$IC_{1-\alpha}(\sigma) = [\sqrt{1.7753} \ ; \ \sqrt{2.6325}] = [1.3324 \ ; \ 1.6225]$$

4. La longueur de l'IC de la moyenne est inférieur à 0.1 signifier que : $2U_\alpha \frac{\hat{\sigma}_c}{\sqrt{n}} \leq 0.1$ d'où

$$\frac{1}{\sqrt{n}} \leq \frac{0.1}{2U_\alpha \hat{\sigma}_c} \Rightarrow \sqrt{n} \leq \frac{2U_\alpha \hat{\sigma}_c}{0.1} \Rightarrow n \geq \left(\frac{2U_\alpha \hat{\sigma}_c}{0.1}\right)^2 \Rightarrow n \geq 400\ U_\alpha^2\ \sigma^2.$$

Sachant que $\sigma^2 = 2$ et que pour n assez grand U_α sera déterminer sur la table de la loi normale. A cet effet, pour $\alpha = 5\%$ alors $U_\alpha = 1.96$, ainsi on aura :

$$n \geq 3073.28 \Rightarrow n_{minimal} = 3074 \text{ vu que } n \in \mathbb{N}.$$

Solution 8 *(Test de conformité de moyenne)*

1. Dans cet exercice, l'objectif, de la première question, est de réaliser le test de conformité bilatéral, suivant :
$$H_0 : \ ''\mu = \mu_0'' \text{ contre } H_1 : \ ''\mu \neq \mu_0'',$$
c'est-à-dire,
$$H_0 : \ ''\mu = 14.5'' \text{ contre } H_1 : \ ''\mu \neq 14.5'',$$
Le fait qu'il n'est y a pas une précision sur la population de l'échantillon prélevé alors la vraie variance (σ^2) est inconnue, dans ce cas l'écart-type σ est remplacé par son estimateur corrigé, de ce fait la statistique du test sera
$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}_c/\sqrt{n}},$$
cette statistique suit une loi de Student de degré de liberté $n-1$.
On a, d'une part, la réalisation de cette statistique est
$$t = \frac{13.8 - 14.5}{1.2/\sqrt{20}} = -2.60874,$$
et d'autre part, sur la table de Student,
$$t_{(n-1,1-\alpha/2)} = t_{(19,1-0.05/2)} = 2.0930.$$
On constate que $|t| > t_{(n-1,1-\alpha/2)}$ alors on rejette H_0, cela signifier que la population dont l'échantillon en question est extrait ne présente pas une teneur en hémoglobine normale.
2. L'objectif, de la deuxième question, est de réaliser le test de conformité unilatéral, suivant :
$$H_0 : \ ''\mu = \mu_0'' \text{ contre } H_1 : \ ''\mu < \mu_0'',$$
c'est-à-dire,
$$H_0 : \ ''\mu = 14.5'' \text{ contre } H_1 : \ ''\mu < 14.5'',$$
La statistique de décision est la même que la première question, c'est-à-dire
$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}_c/\sqrt{n}},$$
qui suit une loi de Student de degré de liberté $n-1$, dont sa réalisation est
$$t = \frac{13.8 - 14.5}{\frac{1.2}{\sqrt{20}}} = -2.60874.$$
Le fractile qu'on doit cherché cette fois-ci sur la table de la loi de Student est bien que $t_{(n-1,1-\alpha)}$, et dans ce cas on aura $t_{(19,1-0.05)} = 1.7291$.
On constate que $|t| > t_{(n-1,1-\alpha/2)}$ alors on rejette H_0, cela signifier que la population dont l'échantillon en question est extrait ne présente pas une teneur en hémoglobine normale, plus précisément elle est significativement trop faible.

Solution 9 *(Test de conformité de variance)*
Cet exercice est classé dans le cadre de test de conformité, d'une variance d'un échantillon normale, dans le cas où la moyenne est inconnue (observée), qui peut s'écrire sous la forme suivante :
$$H_0 : \ ''\sigma^2 = \sigma_0^2 \ '' \text{ contre } H_1 : \ '' \ \sigma^2 \neq \sigma_0^2 \ '',$$

plus précisément,

$$H_0 : \ ''\sigma^2 = 0.04'' \text{ contre } H_1 : \ ''\sigma^2 \neq 0.04''.$$

Alors, la statistique du test est :

$$Y^2 = (n-1)\frac{\hat{\sigma}_c^2}{\sigma^2},$$

qui suit une loi de χ^2 de degré de liberté $n-1$.
La réalisation de cette statistique égale à : $y^2 = (n-1)\frac{\hat{\sigma}_c^2}{\sigma^2} = (20-1)\frac{0.16}{0.04} = 76$.
Le domaine d'acceptation de H_0 est définie par $[a_\alpha, b_\alpha]$ tel que :

$$P(Y^2 \geq a_\alpha) = 1 - \frac{\alpha}{2} \qquad \text{et} \qquad P(Y^2 \geq b_\alpha) = \frac{\alpha}{2}.$$

A cet effet,

1. Pour $\alpha = 5\%$ et un *ddl* $n = 19$ l'intervalle de non rejet de H_0 est donné par :

   $$[8,907 \ , \ 32.852],$$

   où on constate que la réalisation $y^2 = 76 \notin [8,907 \ , \ 32.852]$ donc on rejette H_0, c'est-à-dire le fabricant à tort.
2. Pour $\alpha = 1\%$ et un *ddl* $n = 19$ l'intervalle d'acceptation de H_0 est donné par :

   $$[6.844 \ , \ 38.582],$$

   où on constate que la réalisation $y^2 = 76 \notin [6,844 \ , \ 38.582]$ donc on rejette H_0, c'est-à-dire le fabricant à tort.

Solution 10 *(Test de conformité de moyenne et de variance)*
Afin de répondre aux questions de l'exercice on aura besoin des quantités suivantes :
La moyenne :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(83+96+99+110+130+95+74) = 98.1429,$$

et la variance :

$$\begin{aligned}
\hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \\
&= \frac{1}{6}\left((83-98.1429)^2 + (96-98.1429)^2 + ... + (95-98.1429)^2 + (74-98.1429)^2\right) \\
&= 330.4762.
\end{aligned}$$

1. La formulation du test à réaliser dans ce cas est :

   $$H_0 ''\mu = \mu_0'' \ \textit{contre} \ H_1 ''\mu < \mu_0'', \tag{6.6}$$

   plus précisément :

   $$H_0 ''\mu = 100'' \ \textit{contre} \ H_1 ''\mu < 100''. \tag{6.7}$$

   (a) La statistique du test est : $T = \frac{\overline{X}-\mu_0}{\sqrt{\hat{\sigma}_c^2/n}} = \frac{98.1429-100}{\sqrt{330.4762/7}} = -0.2703$,

   (b) La valeur critique du test est $t_\alpha = t_{(n_1-1,1-\alpha)} = t_{(7-1,1-0.05)} = 1.943$ (de la table de la loi de Student),

(c) On remarque que $T \in]-t_\alpha, t_\alpha[$, alors on ne rejette pas H_0, c'est-à-dire le nombre des pouls est égale à 100 battements en moyenne, avec un risque 5% de se tromper.

2. Le test à réaliser dans ce cas est :

$$H_0\text{"}\sigma^2 = \sigma_0^2\text{"} \ \textit{contre} \ H_1\text{"}\sigma^2 \neq \sigma_0^2\text{"}, \tag{6.8}$$

plus précisément :

$$H_0\text{"}\sigma^2 = 300\text{"} \ \textit{contre} \ H_1\text{"}\sigma^2 \neq 300\text{"}. \tag{6.9}$$

(a) La statistique du test est : $Y = \frac{(n-1)\hat{\sigma}_c^2}{\sigma^2} = \frac{6\times 330.4762}{300} = 6.6095$,

(b) Les valeurs critiques du test sont :
$P\left(\chi^2_{(n-1)} > a_\alpha\right) = 1-\alpha/2 \Rightarrow P\left(\chi^2_{(6)} > a_\alpha\right) = 0.975 \Rightarrow a_\alpha = 1.237$
$P\left(\chi^2_{(n-1)} > b_\alpha\right) = \alpha/2 \Rightarrow P\left(\chi^2_{(6)} > b_\alpha\right) = 0.025 \Rightarrow b_\alpha = 14.449.$
D'où l'intervalle du test de la variance σ^2 est $[1.237\,;\,14.449]$.

(c) On remarque que $Y \in]a_\alpha, b_\alpha[$, alors on ne rejette pas H_0, c'est-à-dire la variation des pouls est égale à 300, avec un risque 5% de se tromper.

Solution 11 *(Estimation par intervalle de confiance et Test de conformité de moyenne)*

1. On a, la vraie valeur de σ est inconnue, de plus la taille de l'échantillon $n < 30$ alors l'intervalle de confiance de la moyenne dans ce cas est donné par :

$$IC_{1-\alpha} = \left[\overline{X} - t_{(n-1,1-\alpha/2)}\sqrt{\frac{\hat{\sigma}_c^2}{n}}\ ;\ \overline{X} + t_{(n-1,1-\alpha/2)}\sqrt{\frac{\hat{\sigma}_c^2}{n}}\right].$$

On a :

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(73) = 7.3,\\
\hat{\sigma}_c^2 &= Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{9}(3.2) = 0.3556,\\
t_{(n-1,1-\alpha/2)} &= t_{(9,1-0.02/2)} = 2.821 \ \text{(par la lecture sur la table de Student)},
\end{aligned}$$

alors,

$$\begin{aligned}
IC_{98\%} &= \left[7.3 - 2.821\sqrt{\frac{0.3556}{10}}\ ;\ 7.3 + 2.821\sqrt{\frac{0.3556}{10}}\right].\\
&= [6.7680\ ;\ 7.8320].
\end{aligned}$$

2. Dans ce cas la vraie valeur de la moyenne μ est inconnue, alors l'intervalle de confiance de la variance sera définie par :

$$IC_{98\%} = \left[(n-1)\frac{\hat{\sigma}_c^2}{b_\alpha}\ ;\ (n-1)\frac{\hat{\sigma}_c^2}{a_\alpha}\right];$$

avec

$$P\left\{\chi^2_{n-1} < a_\alpha\right\} = \alpha/2 \text{ et } P\left\{\chi^2_{n-1} > b_\alpha\right\} = \alpha/2 \tag{6.10}$$

donc,

$$P\left\{\chi^2_{n-1} < a_\alpha\right\} = \alpha/2 \Rightarrow P\left\{\chi^2_9 > a_\alpha\right\} = 0.99$$

$$P\left\{\chi^2_{n-1} > b_\alpha\right\} = \alpha/2 \Rightarrow P\left\{\chi^2_9 > b\right\} = 0.01.$$

Par la lecture sur la table de $Khi - Deux$ on obtient :

$$a_\alpha = 2.088 \quad \text{et} \quad b_\alpha = 21.666.$$

Alors,

$$\begin{aligned} IC_{98\%} &= \left[(10-1)\,\frac{0.3556}{21.666}\;;\;(10-1)\,\frac{0.3556}{2.088}\right]; \\ &= [0.1477\;;\;1.5328]. \end{aligned}$$

3. On désir connaître la nature de la boisson : acide ($PH < 7$) , neutre ($PH = 7$) ou base ($PH > 7$).
   a) On a $\overline{X} = 7.3$ cela signifier que la moyenne ne peut être que soit égale à 7 ou elle est supérieur à 7 (la boisson est "neutre" ou une "base"), alors le test adéquat est :
   $$H_0 : \ ''\mu = 7'' \text{ contre } H_1 : \ ''\mu > 7'',$$
   b) La statistique correspondante à ce test est :
   $$T = \frac{\overline{X} - 7}{\hat{\sigma}_c/\sqrt{n}},$$
   qui suit une loi de Student de degré de liberté $n - 1 = 9$, et la réalisation de cette statistique est : $t = \frac{\overline{X}-7}{\hat{\sigma}_c/\sqrt{n}} = \frac{7.3-7}{0.5963/\sqrt{10}} = 1.5909$.
   c) Donner la valeur critique du test pour un seuil de risque $\alpha = 10\%$. La valeur critique du test correspond au fractile $1 - \alpha = 1 - 0.1$ d'une loi de Student de degré de liberté 9, c'est-à-dire $t_\alpha = 1.383$.
   d) Que peut-on conclure sur la nature de la boisson (acide, neutre ou base). On comparant t et t_α on constate que $t > t_\alpha$ cela signifier que H_0 est fausse elle doit être rejeté, c'est-à-dire la moyenne du PH est significativement supérieur à 7 avec un risque $\alpha = 10\%$. On conclusion, la boisson est une "base" avec un seuil de confiance 90%.

Solution 12 *(Estimation et Test de conformité de moyenne)*

1. L'estimation ponctuelle de la moyenne et de l'écart-type de X sont données par :
$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 102.7333 \\ \hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = 598.9238(\text{ le fait que la vraie moyenne, } \mu, \text{ est inconnue}) \\ \hat{\sigma}_c &= \sqrt{\hat{\sigma}_c^2} = \sqrt{598.9238} = 24.4729 \end{aligned}$$
2. Afin de déterminer l'intervalle de confiance de l'écart-type il faut déterminer d'abord celui de la variance. Rappelons que, la taille de l'échantillon est inférieure à 30, la variable aléatoire X suit une loi normale de paramètre (μ, σ^2) et on ne connaît pas la vraie valeur de la variance σ^2 alors, dans ce cas l'intervalle de confiance de la moyenne pour un risque $\alpha = 10\%$ est définie par :
$$\begin{aligned} IC_{1-\alpha}(\mu) &= \left[\overline{X} - t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}; \overline{X} + t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}\right] \\ &= \left[102.7333 - t_{(15-1,1-0.10/2)}\frac{24.4729}{\sqrt{15}}; 102.7333 + t_{(15-1,1-0.10/2)}\frac{24.4729}{\sqrt{15}}\right] \\ &= [91.6058; 113.8608]\,, \end{aligned}$$

sachant que la valeur $t_{(15-1,1-0.10/2)}$ est lue à partir de la table de la loi de Student et qui égale à : $t_{(15-1,1-0.10/2)} = t_{(14,1-0.10/2)} = 1.761$

3. Dans cet exercice, la variable aléatoire X suit une loi normale de paramètre (μ, σ^2) et on ne connaît pas la vraie valeur de la moyenne μ alors, dans ce cas l'intervalle de confiance de la variance pour un risque $\alpha = 10\%$ est définie par :

$$IC_{1-\alpha}(\sigma^2) = \left[(n-1)\frac{\hat{\sigma}_c^2}{b_\alpha}; (n-1)\frac{\hat{\sigma}_c^2}{a_\alpha}\right].$$

sachant que à partir de la table de la loi de $Khi - Deux$ pour un degré de liberté $n-1$ et $\alpha = 10\%$, on a :
$P\left(\chi^2_{(n-1)} > a_\alpha\right) = 1 - \alpha/2 \Rightarrow P\left(\chi^2_{(14)} > a_\alpha\right) = 0.95 \Rightarrow a_\alpha = 6.571$
$P\left(\chi^2_{(n-1)} > b_\alpha\right) = \alpha/2 \Rightarrow P\left(\chi^2_{(14)} > b_\alpha\right) = 0.05 \Rightarrow b_\alpha = 23.685.$
D'où l'intervalle de confiance de la variance σ^2 est donné comme suit :

$$\begin{aligned} IC_{1-\alpha}(\sigma^2) &= \left[(15-1)\frac{598.9238}{23.685}; (15-1)\frac{598.9238}{6.571}\right] \\ &= [1.3364; 6.2094]. \end{aligned}$$

Ainsi, on déduit aisément l'intervalle de confiance de l'écart-type définie par :

$$\begin{aligned} IC_{1-\alpha}(\sigma) &= \left[\sqrt{1.3364}; \sqrt{6.2094}\right] \\ &= [354.0187; 1276.0513]. \end{aligned}$$

4. Dans cette question, la formulation du test à réaliser est la suivante :

$$H_0 : \text{"}\mu = \mu_0''\text{ contre } H_1 : \text{"}\mu \neq \mu_0'.$$

Au vu de l'échantillon précédent (la taille de l'échantillon < 30, X suit une loi normale et la vraie variance est inconnue) c'est le test de Student qu'il faut réaliser. On a :
• d'une part la statistique du test $t = \frac{\overline{X}-\mu_0}{\hat{\sigma}_c/\sqrt{n}} = \frac{102.7333-110}{24.4729/\sqrt{15}} = -1.1500$
• et d'autre part sur la table de la loi de Student $t_\alpha = t_{(n-1,1-\alpha/2)} = t_{(14,1-.02/2)} = 2.625$.
A cet effet, on constate que $t \in [-t_\alpha, t_\alpha](-1.1500 \in [-2.625, 2.625]) \Rightarrow H_0$ est vraie, c'est-à-dire à un seuil de risque 2% la taille moyenne des enfants est égale de 110cm.

Solution 13 *(Estimation et Test de conformité de moyenne)*

1. L'estimation ponctuelle de la moyenne et de l'écart-type de X sont données par :

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = 23.3636 \\ \hat{\sigma}_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 = 2.4465(\text{ le fait que la vraie moyenne, } \mu, \text{ est inconnue}) \\ \hat{\sigma}_c &= \sqrt{\hat{\sigma}_c^2} = \sqrt{2.4465} = 1.56414 \end{aligned}$$

2. Afin de déterminer l'intervalle de confiance de l'écart-type il faut déterminer d'abord celui de la variance.
Dans cet exercice, la taille de l'échantillon est inférieure à 30, la variable aléatoire X suit une loi normale de paramètre (μ, σ^2) et on ne connaît pas la vraie valeur de la variance σ^2

alors, dans ce cas l'intervalle de confiance de la moyenne pour un risque $\alpha = 10\%$ est définie par :

$$\begin{aligned} IC_{1-\alpha}(\mu) &= \left[\overline{X} - t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}; \overline{X} + t_{(n-1,1-\alpha/2)}\frac{\hat{\sigma}_c}{\sqrt{n}}\right] \\ &= \left[23.3636 - t_{(11-1,1-0.10/2)}\frac{1.56414}{\sqrt{11}}; 23.3636 + t_{(11-1,1-0.10/2)}\frac{1.56414}{\sqrt{11}}\right] \\ &= [22.508; 24.218]\,, \end{aligned}$$

sachant que la valeur $t_{(11-1,1-0.10/2)}$ est lue à partir de la table de la loi de Student et qui égale à : $t_{(11-1,1-0.10/2)} = t_{(10,1-0.10/2)} = 1.8125$

3. Dans cet exercice, la variable aléatoire X suit une loi normale de paramètre (μ, σ^2) et on ne connaît pas la vraie valeur de la moyenne μ alors, dans ce cas l'intervalle de confiance de la variance pour un risque $\alpha = 10\%$ est définie par :

$$IC_{1-\alpha}(\sigma^2) = \left[(n-1)\frac{\hat{\sigma}_c^2}{b_\alpha}; (n-1)\frac{\hat{\sigma}_c^2}{a_\alpha}\right].$$

sachant que à partir de la table de la loi de $Khi - Deux$ pour un degré de liberté $n-1$ et $\alpha = 2\%$, on a :
$P\left(\chi^2_{(n-1)} > a_\alpha\right) = 1 - \alpha/2 \Rightarrow P\left(\chi^2_{(10)} > a_\alpha\right) = 0.95 \Rightarrow a_\alpha = 3.9403$
$P\left(\chi^2_{(n-1)} > b_\alpha\right) = \alpha/2 \Rightarrow P\left(\chi^2_{(10)} > b_\alpha\right) = 0.05 \Rightarrow b_\alpha = 18.3070.$
D'où l'intervalle de confiance de la variance σ^2 est donné comme suit :

$$\begin{aligned} IC_{1-\alpha}(\sigma^2) &= \left[(11-1)\frac{2.4465}{18.3070}; (11-1)\frac{2.4465}{3.9403}\right] \\ &= [1.3364; 6.2089]\,. \end{aligned}$$

Ainsi, on déduit aisément l'intervalle de confiance de l'écart-type définie par :

$$\begin{aligned} IC_{1-\alpha}(\sigma) &= \left[\sqrt{1.3364}; \sqrt{6.2089}\right] \\ &= [1.1560; 2.4918]\,. \end{aligned}$$

4. Dans cette question, la formulation du test à réaliser est la suivante :

$$H_0 : \text{"}\mu = \mu_0''\text{ contre } H_1 : \text{"}\mu \neq \mu_0'.$$

Au vu de l'échantillon précédent (la taille de l'échantillon < 30, X suit une loi normale et la vraie variance est inconnue) c'est le test de Student qu'il faut réaliser. On a :
• d'une part la statistique du test $t = \frac{\overline{X}-\mu_0}{\hat{\sigma}_c/\sqrt{n}} = \frac{23.3636-25}{1.56414/\sqrt{11}} = -3.4698$
• et d'autre part sur la table de la loi de Student $t_\alpha = t_{(n-1,1-\alpha/2)} = t_{(10,1-.02/2)} = 2.764$.
A cet effet, on constate que $t \notin [-t_\alpha, t_\alpha](-3.4698 \notin [-2.764, 2.764]) \Rightarrow H_0$ est fausse, c'est-à-dire à un seuil de risque 2% la taille moyenne des arbres est différente de 25.

Solution 14 *(Test de Student : conformité et d'homogénéité)*
Soit les notation suivantes :
μ_1 est la vraie taille moyenne des garçons.
μ_2 est la vraie taille moyenne des filles.

1. Le test à réaliser dans ce cas est :

$$H_0\text{"}\mu_1 = \mu_0\text{"} \ \textit{contre} \ H_1\text{"}\mu_1 \neq \mu_0\text{"}, \tag{6.11}$$

plus précisément :

$$H_0\text{"}\mu_1 = 180\text{"} \ \textit{contre} \ H_1\text{"}\mu_1 \neq 180\text{"}. \tag{6.12}$$

   (a) La statistique du test est : $T_1 = \frac{\overline{X}-\mu_0}{\sqrt{\hat{\sigma}_c^2/n_1}} = \frac{182.43-180}{\sqrt{54.95/11}} = 1.0872$,

   (b) La valeur critique du test est $t_\alpha = t_{(n_1-1,1-\alpha/2)} = t_{(10,1-0.05/2)} = 2.228$ (de la table de la loi de Student),

   (c) On remarque que $T_1 \in]-t_\alpha, t_\alpha[$ alors on rejette pas H_0, c'est-à-dire on admet que la taille moyenne des garçon est de 180cm avec un risque 5% de se tromper.

2. Le test à réaliser dans ce cas est :

$$H_0\text{"}\mu_1 = \mu_2\text{"} \ \textit{contre} \ H_1\text{"}\mu_1 \neq \mu_2\text{"}$$

mais pour réaliser ce test on est contraint à réaliser d'abord le test suivant :

$$H_0\text{"}\hat{\sigma}_{c,1}^2 = \hat{\sigma}_{c,2}^2\text{"} \ \textit{contre} \ H_1\text{"}\hat{\sigma}_{c,1}^2 \neq \hat{\sigma}_{c,2}^2\text{"}.$$

   (a) Sachant que $\hat{\sigma}_{c,1}^2 > \hat{\sigma}_{c,2}^2$ alors, la statistique du test est : $F = \frac{\hat{\sigma}_{c,1}^2}{\hat{\sigma}_{c,2}^2} = \frac{54.95}{26.20} = 2.0973$,

   (b) La valeur critique du test est $f_\alpha = f_{(n_1-1,n_2-1,1-\alpha/2)} = f_{(10,9,0.975)} = 3.96$ (de la table de la loi de Fisher),

   (c) On remarque que $F \in [1, f_\alpha[$ alors on admet que les deux échantillons ont la même variance avec un risque 5% de se tromper.

   (d) Le fait que les deux échantillons ont la même variance, alors on calcule la variance commune : $\hat{\sigma}_c^2 = \frac{(n_1-1)\hat{\sigma}_{c,1}^2+(n_2-1)\hat{\sigma}_{c,2}^2}{n_1+n_2-2} = \frac{10*54.95+9*26.20}{11+10-2} = 41.3316$.

   (e) Revenant maintenant au test (6.11), la statistique de ce dernier est :

$$T_2 = \frac{\overline{X}-\overline{Y}}{\sqrt{\hat{\sigma}_c^2(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{182.43-168.80}{\sqrt{41.3316(\frac{1}{11}+\frac{1}{10})}} = 4.8522,$$

   (f) La valeur critique du test est $t_\alpha = t_{(n_1+n_2-2,1-\alpha/2)} = t_{(19,1-0.05/2)} = 2.093$ (de la table de la loi de Student),

   (g) On remarque que $T_2 \notin]-t_\alpha, t_\alpha[$ alors on rejette H_0, c'est-à-dire on admet que la taille moyenne des garçon est significativement différente de celle des filles.

3. Le test à réaliser dans ce cas est :

$$H_0\text{"}\mu_1 = \mu_2\text{"} \ \textit{contre} \ H_1\text{"}\mu_1 > \mu_2\text{"}$$

   (a) Les conditions de ce test sont vérifier dans la deuxième question (homogénéité de variance et calcul de la variance commune), de plus la statistique correspondante à ce test est la même que celle du test (6.11), c'est-à-dire $T_3 = T_2 = 4.8522$.

   (b) la valeur critique du test est définie cette fois-ci par :
   $t_\alpha = t_{(n_1+n_2-2,1-\alpha)} = t_{(19,1-0.05)} = 1.729$ (de la table de la loi de Student),

   (c) On remarque que $T_3 > t_\alpha$ alors on rejette H_0, c'est-à-dire on admet que la taille moyenne des garçon est supérieur à celle des filles.

Solution 15 *(Test de Student : conformité et d'homogénéité)*

1. Afin de confirmer ou de démentir ce que l'agent indique, nous devons réaliser le test d'homogénéité de moyenne des deux échantillons, dont la formulation du test est donné par :

$$H_0 : \text{ " } \mu_X = \mu_Y \text{ " contre } H_1 : \text{ " } \mu_X < \mu_Y \text{ ",} \tag{6.13}$$

mais on doit vérifier d'abord l'homogénéité de leurs variances, et cela en réalisant le test suivant :

$$H_0 : \text{ " } \sigma_X^2 = \sigma_Y^2 \text{ " contre } H_1 : \text{ " } \sigma_X^2 \neq \sigma_Y^2 \text{ ",}$$

Notons que les moyennes est les variances des deux échantillons sont données par :
$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{9} X_i = \frac{1}{9}\sum_{i=1}^{9} X_i = 252.3333$ et $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{11} Y_i = \frac{1}{11}\sum_{i=1}^{11} Y_i = 243$
Sachant que, dans cet exercice, la vraie moyenne est connue où $\mu = 250$ alors
$\hat{\sigma}_X^2 = \frac{1}{n_1}\sum_{i=1}^{9}(X_i - \mu)^2 = \frac{1}{9}\sum_{i=1}^{9}(X_i - 250)^2 = 48.6667.$
$\hat{\sigma}_Y^2 = \frac{1}{n_2}\sum_{i=1}^{11}(Y_i - \mu)^2 = \frac{1}{11}\sum_{i=1}^{11}(Y_i - 250)^2 = 18.7273.$

   a) La statistique du test d'homogénéité de variance des deux échantillons est donnée par : $F = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2}$ (le fait que $\hat{\sigma}_X^2 > \hat{\sigma}_Y^2$) est sa réalisation est $f = 2.5987$.

   b) La valeur critique du test, pour $\alpha = 5\%$, est : $f_\alpha = f_{(n_1,n_2,1-\alpha/2)} = f_{(9,11,0.975)} = 3.59$.

   c) On constate que : $f < f_\alpha (2.5987 < 3.59)$, cela signifier que les deux échantillons ont la même variance.

   d) Le fait que les deux échantillons ont la même variance donc on calcule la variance commune définie par :

$$\hat{\sigma}^2 = \frac{n_1\hat{\sigma}_X^2 + n_2\hat{\sigma}_Y^2}{n_1 + n_2} = 32.2000$$

   e) Ainsi, la statistique du test d'homogénéité de moyenne (6.13) est donnée par :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}^2(\frac{1}{n_1} + \frac{1}{n_2})}} \quad \text{et sa réalisation est} \quad t = 3.6594.$$

   f) La valeur critique du test est : $t_\alpha = t(n_1 + n_2, 1 - \alpha/2) = t(20, 1 - 0.05/2) = 2.086$.

   g) On constate que, $t \notin [-t_\alpha, t_\alpha]$, cela signifier que l'agent à raison une fois de plus, les deux types du camembert n'ont pas la même masse moyenne.

2. Le test à réaliser dans ce cas est bien que :

$$H_0 : \text{ "}\mu_Y = 250\text{" contre } H_1 : \text{ "}\mu_Y < 250\text{".}$$

   a) La réalisation de statistique du test est : $t = \frac{\overline{Y} - 250}{\sqrt{\hat{\sigma}_Y^2/n_2}} = -5.3648$

   b) La valeur critique du test est : $t_\alpha = t_{(n_2,1-\alpha)} = t_{(11,1-0.05)} = 1.796$.

   c) On constate que $t < -t_\alpha (-5.3648 < -1.796)$, donc l'agent a le droit de pénaliser l'entreprise.

3. Dans cette situation, la statistique du test est la même $t = -5.3648$ et la valeur critique est $t_\alpha = t_{(n_2,1-\alpha)} = t_{(11,1-0.02)} = 2.3281$. On constate également que $t < -t_\alpha$ $(-5.3648 < -2.3281)$, ce qui confirme que l'agent à le droit de pénalisé l'entreprise le fait que la masse moyenne du deuxième type de camembert est significativement inférieur à la norme (¡250g).

Solution 16 *(Test d'homogénéité de Student)*

1. Déterminer une estimation ponctuelle de la moyenne et de la variance de chaque échantillons. On a :

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = 24.5 \quad \text{et} \quad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i = 22$$

$$\hat{\sigma}^2_{c,1} = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\overline{X})^2 = 0.7880 \text{ et } \hat{\sigma}^2_{c,2} = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(y_i-\overline{Y})^2 = 0.8700$$

2. Supposons qu'on désir savoir si les deux types d'arbres ont la même hauteur en moyenne.

   a) Le test à réaliser est bien que le test d'homogénéité de moyennes, qu'on peut formulé comme suit :

$$H_0 : \text{"}\mu_1 = \mu_2\text{"} \text{ contre } H_1 : \text{"}\mu_1 \neq \mu_2\text{"}, \qquad (*)$$

   b) La condition nécessaire pour la réalisation de ce test est que les deux échantillons ont la même variance, c'est-à-dire il faut réaliser d'abord le test suivant :

$$H_0 : \text{" } \sigma_1^2 = \sigma_2^2 \text{ " } \textit{ contre } \text{ } H_1 : \text{" } \sigma_1^2 \neq \sigma_2^2 \text{ ".}$$

   La statistique de ce test (homogénéité de variance) est :
   $F = \frac{\hat{\sigma}^2_{c,2}}{\hat{\sigma}^2_{c,1}}$, car la deuxième variance est plus grande que la première.
   La réalisation de cette statistique est : $f = \frac{0.8700}{0.7880} = 1.1041$.
   La valeur critique du test, f_α, correspond au fractile d'ordre $1-\alpha/2 = 1-0.02/2 = 0.99$ d'une loi de Fisher de degrés de liberté $(n_2-1, n_1-1) = (4,5)$ d'où par la lecture sur la table de la loi de Fisher on obtient $f_\alpha = 11.39$. On constate que $f < f_\alpha$, alors les deux échantillons ont les mêmes variances.

3. On a les deux échantillons sont issus d'une loi normale, mutuellement indépendants de plus ils ont la même variance (voir réponse 2.*b*)), donc le test $(*)$ sera réaliser à l'aide du test d'homogénéité de moyenne de Student.

   a) La statistique du test $(*)$ est :

$$T = \frac{\overline{X}-\overline{Y}}{\hat{\sigma}_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},$$

   où $\hat{\sigma}_c^2$ est la variance globale des deux échantillons définie par :

$$\hat{\sigma}_c^2 = \frac{(n_1-1)\hat{\sigma}^2_{c,1}+(n_2-1)\hat{\sigma}^2_{c,2}}{n_1+n_2-2} = \frac{(6-1)0.7880+(5-1)0.8700}{6+5-2} = 0.8244.$$

   Donc, la réalisation de la statistique T est :

$$t = \frac{24.5-22}{0.9080\sqrt{\frac{1}{6}+\frac{1}{5}}} = 4.5469.$$

   b) La valeur critique du test est le fractile d'ordre $1-\alpha/2 = 1-.02/2$ d'une loi de Student de degré de liberté $n_1+n_2-2 = 9$, d'où $t_\alpha = 2.821$.

   c) On à $t \notin [-t_\alpha, t_\alpha]$, donc H_0 est fausse, cela signifier que c'est l'hypothèse H_1 qui est vraie c'est-à-dire les deux type d'arbres ont des hauteurs moyennes significativement différentes.

Solution 17 *(Test d'ajustement : $Khi-Deux$ et Kolmogorov-Smirnov)*

1. **Test de *Khi-Deux*** :

   a) Les proportions théorique de la répartition des couleurs selon le modèle de Mendel doit être comme suite :

Phénotype	Rouge	Rose	Blanc
p_i	1/4	2/4	1/4
np_i	150	300	150

   b) La statistique du $Khi-Deux$ est définie par :

$$K_n^2 = \sum_{i=1}^{r} \frac{(N_i - np_i)^2}{np_i}. \tag{6.14}$$

   La réalisation de cette statistique est donnée par :

$$k_n^2 = \sum_{i=1}^{3} \frac{(n_i - np_i)^2}{np_i} = \frac{(141-150)^2}{150} + \frac{(325-300)^2}{300} + \frac{(134-150)^2}{150} = 4.33. \tag{6.15}$$

   c) Pour savoir (conclure) si l'expérience du croisement obéit au modèle de Mendel, il suffit de comparé la réalisation de la statistique de *Khi-Deux* avec la valeur tabulée de *Khi-Deux* pour $r-1$ degré de liberté et un risque α qu'on doit fixer préalablement. Supposons que $\alpha = 5\%$, alors sur la table du *Khi-Deux* pour un degré de liberté $r-1=2$, $\chi^2 = 5.991 > 4.33 = k_n^2$, donc on peut conclure que le croisement en question est du modèle de Mendel avec un risque $\alpha = 5\%$.

2. **test de *Kolmogorov-Smirnov* (K.S.)** :

   La statistique du test de *Kolmogorov-Smirnov* est donnée par :

$$D_n = max|F_n(x) - F_0(x)| = \max D(x_u) \tag{6.16}$$

   Le test de *K.S.*, se base sur les fréquences cumulées, alors on doit d'abord calculer les fréquences empiriques.

Phénotype	Rouge	Rose	Blanc
f_i (empirique)	0.2350	0.5417	0.2233
p_i (théorique)	0.2500	0.500	0.2500

   ainsi, on obtient les fréquence cumulées théoriques et empiriques suivantes :

Phénotype	Rouge	Rose	Blanc
F_i (empirique)	0.2350	0.7767	1
F_0 (théorique)	0.2500	0.7500	1

   On a d'une part,

$$D(x_u) = \{|0.235 - 0.250|; |0.7767 - 0.7500|; |1-1|\} = \{0.0150;\ 0.0267,\ 0\},$$

   d'où le $\max D(x_u) = 0.0267$.
   D'autre part, sur la table de *K.S.* pour $n = 600$ on aura $d_\alpha = \frac{1.358}{\sqrt{n}} = \frac{1.358}{\sqrt{600}} = 0.0554$; de plus on remarque que $\max D(x_u) < d_\alpha$, cela signifie que l'échantillon obtenue du croisement est distribués selon le modèle de Mendel.

Solution 18 *(Test d'ajustement : $Khi-Deux$ et Kolmogorov-Smirnov)* Si aucun facteur n'est plus important que les autres alors la distribution de la répartition des 4 facteurs doit être uniforme. Pour vérifier cette dernière hypothèse nous allons utiliser un test d'ajustement.

Test de *Khi-Deux* Les proportions ainsi que les effectifs théorique de la répartition des 4 facteurs doivent être comme suite :

	Effectif observé	Proportions théorique	Effectif théorique	Résidu
Contenu cours	7	1/4	12.5	-5.5
Examen	8	1/4	12.5	-4.5
Professeur	17	1/4	12.5	4.5
Horaire	18	1/4	12.5	5.5
Total	50			

Pour conclure qu'au moins un facteur est plus important que les autres, il suffit de comparé la réalisation de la statistique de *Khi-Deux* donnée par :

$$K_n^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{6.17}$$

avec la valeur tabulée de *Khi-Deux* pour $r-1$ (r est le nombre de modalité) degré de liberté et un risque α qu'on doit fixer préalablement.
La réalisation de cette statistique est donnée par :

$$k_n^2 = \sum_{i=1}^{4} \frac{(n_i - np_i)^2}{np_i} = 8.080.$$

Supposons qu'on fixe $\alpha = 5\%$ alors on aura les résultats résumés dans le tableau suivant.

k_n^2	ddl	$\chi^2_{(3,0.05)}$
8.080	$r-1=3$	7.815

A partir de ces résultats numérique on peut conclure que : aucun facteur n'est plus important que les autres, c'est-à-dire le choix du cours se fait d'une manière uniforme sans favoriser un critère de choix à un autre.

test de *Kolmogorov-Smirnov* (K.S.) :
Rappelons que le test de *Kolmogorov-Smirnov* se base sur les fréquences cumulées croissantes. Les différentes quantités nécessaires pour le test sont résumées dans le tableau suivant :

	Contenu cours	Examen	Professeur	Horaire
Effectif	18	17	7	8
fréquence observées f_n	0.36	0.34	0.14	0.16
fréquence théorique f_t	0.25	0.25	0.25	0.25
F_n	0.36	0.70	0.84	1.00
F_t	0.25	0.50	0.75	1.00
Écart absolu	0.11	0.20	0.09	0

On a d'une part, le $d_c = max(|F_n - F_t|) = 0.20$ et d'autre part $d_\alpha = \frac{1.358}{\sqrt{50}} = 0.1921(\alpha = 5\%)$. On constate que $d_c < d_\alpha$ alors on conclu que : aucun facteur n'est plus important que les autres ce qui confirme les résultats du test de $Khi - Deux$.

Solution 19 *(Estimation ponctuelle et Test d'ajustement de Kolmogorov-Smirnov)*

1. L'estimation de la moyenne μ et la variance σ^2 de la loi normale à partir de l'échantillon est :
$\hat{\mu} = \overline{X} = \frac{1}{10}\sum_{i}^{10} x_i = 2.13$ et $\hat{\sigma}^2 = \hat{\sigma}_c^2 = \frac{1}{10-1}\sum_{i}^{10}(x_i - \overline{X})^2 = 0.2357 \Rightarrow \hat{\sigma}_c = 0.4855$

2. Afin de réaliser le test de Kolmogorov-Smirnov sur l'échantillon :

   a) La première étape est bien que d'ordonner les observation à l'ordre croissant.

   a) Le fait que le tirage de l'échantillon est aléatoire est simple alors la probabilité de chaque observation x_i $(i = \overline{1,10})$ égale à 1/10, de ce fait on obtiens la distribution empirique F_n décrite dans la deuxième ligne du tableau ci-dessous.

   a) Calculer la distribution théorique $F(x_i) = P(x < x_i)$, dans cet exercice après transformation de la variable aléatoire X qu'on supposé $X \rightsquigarrow N(2.13, 0.4855)$ vers une loi normale $N(0,1)$, la lecture sur la table de la loi normale nous fournie ce qui suit :
   $P(X < x_1) = P(\frac{X-\mu}{\sigma} < \frac{1.4-2.13}{0.4855}) = P(Y < -1.5036) = 1 - P(Y < 1.5036) = 0.0337$
   $P(X < x_2) = P(\frac{X-\mu}{\sigma} < \frac{1.7-2.13}{0.4855}) = P(Y < -0.8857) = 1 - P(Y < 0.8857) = 0.0121$
   $P(X < x_3) = P(\frac{X-\mu}{\sigma} < \frac{1.8-2.13}{0.4855}) = P(Y < -0.6797) = 1 - P(Y < 0.6797) = 0.0517$
   $P(X < x_4) = P(\frac{X-\mu}{\sigma} < \frac{1.9-2.13}{0.4855}) = P(Y < -0.4737) = 1 - P(Y < 0.4737) = 0.0822$
   $P(X < x_5) = P(\frac{X-\mu}{\sigma} < \frac{2.0-2.13}{0.4855}) = P(Y < -0.2678) = 1 - P(Y < 0.2678) = 0.1056$
   $P(X < x_6) = P(\frac{X-\mu}{\sigma} < \frac{2.1-2.13}{0.4855}) = P(Y < -0.0618) = 1 - P(Y < 0.0618) = 0.1246$
   $P(X < x_7) = P(\frac{X-\mu}{\sigma} < \frac{2.2-2.13}{0.4855}) = P(Y < 0.1442) = 0.1427$
   $P(X < x_8) = P(\frac{X-\mu}{\sigma} < \frac{2.5-2.13}{0.4855}) = P(Y < 0.7621) = 0.0230$
   $P(X < x_9) = P(\frac{X-\mu}{\sigma} < \frac{2.7-2.13}{0.4855}) = P(Y < 1.1740) = 0.0202$
   $P(X < x_10) = P(\frac{X-\mu}{\sigma} < \frac{3.0-2.13}{0.4855}) = P(Y < 1.7920) = 0.0366$

X	1.4000	1.7000	1.8000	1.9000	2.0000	2.1000	2.2000	2.5000	2.7000	3.0000
F_n	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
F	0.0663	0.1879	0.2483	0.3178	0.3944	0.4754	0.5573	0.7770	0.8798	0.9634
D	0.0337	0.0121	0.0517	0.0822	0.1056	0.1246	0.1427	0.0230	0.0202	0.0366

   A partir des résultats rangés dans le tableau ci-dessus on obtiens :

$$D = max\{|F_n(x_i) - F(x_i)|\} = 0.1427,$$

   et à partir de la table de Kolmogorov-Smirnov pour $\alpha = 5\%$ on a :

$$d_{(10,1-\alpha)} = 0.409.$$

   On constate que $D < d_{(10,1-\alpha)}$ cela signifie que H_0 est varie c'est-à-dire on admet que la distribution de l'échantillon est bien une distribution normale de paramètre $\mu = 2.13$ et $\sigma^2 = 0.2357$.

Solution 20 *(Test d'ajustement Khi − Deux et Kolmogorov-Smirnov)*

1. Avant de répondre à la question on doit d'abord estimer le paramètre, λ, de la loi de Poisson. d'après la remarque (c) :

$$\hat{\lambda} = \frac{1}{n}\sum_{i=0}^{6} n_i X_i = \frac{1}{200} \times 400 = 2.$$

   Alors la distribution de Poisson correspondante à l'échantillon est définie comme suit :

$$P(X = x) = \frac{2^x}{x!}e^{-2}, \quad \text{pour } i = 0,\ 1,\ 2, ..., 6.$$

Alors P_i (théorique) sont calculées comme suit (voir remarque (b)) :

$$P_0 = P(X=0) = \frac{2^0}{0!}e^{-2} = 0.1353 \qquad P_4 = P(X=4) = \frac{2^4}{4!}e^{-2} = 0.0902$$

$$P_1 = P(X=1) = \frac{2^1}{1!}e^{-2} = 0.2707 \qquad P_5 = P(X=5) = \frac{2^5}{5!}e^{-2} = 0.0361$$

$$P_2 = P(X=2) = \frac{2^2}{2!}e^{-2} = 0.2707 \qquad P_6 = P(X=6) = \frac{2^6}{6!}e^{-2} = 0.0120$$

$$P_3 = P(X=3) = \frac{2^3}{3!}e^{-2} = 0.1804$$

Nombre de pannes x (x_i)	0	1	2	3	4	5	6
Jours avec x pannes (n_i)	32	50	52	34	19	10	3
P_i	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120
nP_i	27.0671	54.1341	54.1341	36.0894	18.0447	7.2179	2.4060
$e_i = N_i - nP_i$	4.9329	-4.1341	-2.1341	-2.0894	0.9553	2.7821	0.5940

La réalisation de la statistique du test est calculée comme suit :

$k_n^2 = \sum_{i=0}^{7} \frac{e_i^2}{nP_i} = \frac{4.9329^2}{27.0671} + \frac{(-4.1341)^2}{54.1341} + \frac{(-2.1341)^2}{54.1341} + \frac{(-2.0894)^2}{36.0894} + \frac{0.9553^2}{18.0447} + \frac{2.7821^2}{7.2179} + \frac{0.5940^2}{2.4060} = 2.6894.$

Il ne nous reste qu'à déterminer la valeur critique du test qui est le fractile d'une loi de $Khi - Deux$ d'ordre $1-\alpha$ à un $r-1-q$ degré de liberté $(\chi^2_{(r-1-q,\alpha)})$, avec q est le nombre de paramètres estimé. Dans notre cas, on n'a estimé que le paramètre λ ainsi d'après la table de la loi de $Khi - Deux$. Ainsi d'après la table de la loi de $Khi - Deux$: $\chi^2_{(r-1-q,\alpha)} = \chi^2_{(7-1-1,0.05)} = 11.0705$.

On remarque que $k_n^2 < \chi^2_{(r-1-q,\alpha)}$ donc on accepte l'ajustement de la distribution de l'échantillon par une loi de Poisson de paramètre $\lambda = 2$.

2. Dans le test K.S. nous avons besoin des fréquences cumulées croissantes (observées) et la fonction de répartition de la loi de Poisson (théorique).

$F(0) = P_0 = 0.1353$
$F(1) = P_0 + P_1 = 0.4060$
$F(2) = P_0 + P_1 + P_2 = 0.6767$
$F(3) = P_0 + P_1 + P_2 + P_3 = 0.8571$

$F(4) = P_0 + P_1 + P_2 + P_3 + P_4 = 0.9473$
$F(5) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5 = 0.9834$
$F(6) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5 + P_6 = 0.9955$

Nombre de pannes x (x_i)	0	1	2	3	4	5	6
Jours avec x pannes (n_i)	32	50	52	34	19	10	3
f_i	0.1600	0.2500	0.2600	0.1700	0.0950	0.0500	0.0150
F_i	0.1600	0.4100	0.6700	0.8400	0.9350	0.9850	1.0000
$F(x_i)$	0.1353	0.4060	0.6767	0.8571	0.9473	0.9834	0.9955
D_i	0.0247	0.0040	-0.0067	-0.0171	-0.0123	0.0016	0.0045

$d_c = max(|D_i|) = \max\{0.0247\ ,\ 0.0040\ ,\ 0.0067\ ,\ 0.0171\ ,\ 0.0123\ ,\ 0.0016\ ,\ 0.0045\}$ donc $d_c = 0.0247$. La valeur critique du test est définie par : $d_\alpha = d(n, 1-\alpha)$ et sur la table de K.S. on aura $d_\alpha = d(n, 1-\alpha) = \frac{1.358}{\sqrt{n}} = \frac{1.358}{\sqrt{200}} = 0.0960$. On constate que $d_c < d_\alpha$, ce qui signifier qu'on peut admettre que la distribution de l'échantillon est une loi de Poisson de paramètre $\lambda = 2$.

Solution 21 *(Analyse de la variance à un seul facteur)*

À partir de l'énoncé on a $p = 4$ et $n = p * 25 = 100$, ces données nous permet de déterminer facilement les différents degrés de liberté ($d.d.l$ de la troisième colonne de la table d'ANOVA 1.

1. $p - 1 = \mathbf{3}$.

2. $n - p = \mathbf{96}$.
3. $n - 1 = \mathbf{99}$.

Pour le reste des quantités on a :

- $CM_{F_1} = SC_{F_1}/(I-1) = 31.82$, alors $SC_{F_1} = CM_{F_1} * (I-1) = 31.82 * 3 = \mathbf{95.46}$.
- $F_c = CM_{F_1}/CM_{Res} = 6.64$, alors $CM_{Res} = CM_{F_1}/F_c = 31.82/6.64 = \mathbf{4.7922}$.
- $CM_{Res} = SC_{Res}/(n-p)$, alors $SC_{Res} = CM_{Res} * (n-p) = 4.7922 * 96 = \mathbf{460.0512}$.
- $SC_{Tot} = CM_{Res} + SC_{F_1}$, alors $SC_{Tot} = 95.46 + 464.8434 = \mathbf{555.5112}$.

Ainsi, on aura la table suivante :

	Somme des carrés	*ddl*	Moyenne des carrés	*F*
Inter-groupes	95.46	3	31.82	6.64
Intra-groupes	460.0512	96	4.7922	
Total	555.5112	99		

TABLE 6.9: *Table de l'ANOVA associée au problème*

Sur la table de la loi de Ficher pour un seuil de confiance 0.95 ($\alpha = 5\%$) on obtient : $f_\alpha = f_{(3,96,0.95)} \approx f_{(3,100,0.95)} = 2.70 < 6.64 \Rightarrow$ le facteur "Méthode d'enseignement" a une influence significative sur le niveau des étudiants.

Solution 22 *(ANOVA 1 à un plan équilibré)*

Les différentes moyennes (de chaque échantillon et globale) sont données par :

$$\overline{X}_1 = 37.40,\ \overline{X}_2 = 41.00,\ \overline{X}_3 = 30.80,\ \overline{X}_4 = 50.20 \text{ et } \overline{X} = 39.85.$$

En exploitant ces dernières quantités pour le calcul des différentes variations on obtient :

	SC	*ddl*	*CM*	*f*	f_α
Inter-groupes	961.2667	3	320.4222	14.1123	5.29
Intra-groupes	363.2833	16	22.7052		
Total	1324.55	19			

On constate que $f > f_\alpha$ cela signifie qu'on doit rejeter H_0. C'est-à-dire le facteur traitement a une influence significative sur les durées séparant deux crise d'asthme.

Solution 23 *(ANOVA 1 à un plan non équilibré)*

1. Les estimations des différentes moyennes μ_1, μ_2, μ_3, μ_4 et m sont données respectivement par :

   ▷ $\overline{X}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} = \frac{1}{3}(137) = 45.67$, ▷ $\overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} = \frac{1}{5}(255) = 51.00$,

   ▷ $\overline{X}_3 = \frac{1}{n_3}\sum_{j=1}^{n_3} x_{3j} = \frac{1}{4}(180) = 45.00$, ▷ $\overline{X}_4 = \frac{1}{n_4}\sum_{j=1}^{n_4} x_{4j} = \frac{1}{4}(178) = 44.50$,

   et $\overline{X} = \frac{1}{n}\sum_{i=1}^{4} n_i \overline{X}_i = \frac{1}{16}(3 * 45.67 + 5 * 51.00 + 4 * 45.00 + 4 * 44.50) = 46.8750$.

2. Considérons l'hypothèse (H_0) : les rendements moyens de chaque variété sont égaux.

   a) Dans ce cas, le problème est le même que celui de l'exercice 22. Contrairement à ce dernier, on remarque que les tailles des échantillons ne sont pas les mêmes, mais le raisonnement de l'ANOVA ne changera pas . En effet, la décomposition de la variation

totale dans cette situation sera :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2}_{SC_{Tot}}=\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2}_{SC_{Res}}+\underbrace{\sum_{j=1}^{p}n_j(\overline{X}_j-\overline{X})^2}_{SC_{Fac}}, \tag{6.18}$$

où, n_j : est la taille du $j^{\text{ième}}$ échantillon (groupe).

SC_{Tot} : est la variation totale qui représente dispersion des données autour de la moyenne générale.

SC_{Fac} : est la variation due au facteur qui représente dispersion des moyennes autour de la moyenne générale.

SC_{Res} : est la variation résiduelle qui représente dispersion des données à l'intérieur de chaque échantillon autour de sa moyenne.

source de variation	Somme des carrés SC	Degrés de libertés ddl	Carré moyen CM	ratio F_{obs}
Inter-groupe	126.0833	3	42.0278	1.9128
Intra-groupe	263.6667	12	21.9722	
Total	389.7500	15		

b) On a d'une part $F_{obs} = 1.9128$ et d'autre part $f_\alpha = f(3, 12, 1-0.01) = 5.9525$, alors on rejette pas H_0 car $F_{obs} < f_\alpha$, cela signifier qu'il y a pas une différence significatives entre les rendement des différentes variétés d'orge.

Solution 24 *(ANOVA 2 à un plan équilibré)*

Dans cet exercice le problème posé est l'analyse de l'influence d'un traitement hormonal, du sexe et leurs interactions sur la concentration de calcium dans le plasma d'un être humain.

1. Sachant que le modèle correspondant au problème est :

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \text{ avec } i = \overline{1, I}, j = \overline{1, J} \text{ et } k = \overline{1, K},$$

alors les hypothèses à tester sont :

Effet du premier facteur (le traitement hormonal) :

H_0 : "*les paramètres α_i sont tous nuls*" contre H_1 : "*les paramètres α_i ne sont pas tous nuls*"

Effet du second facteur (le sexe) :

H_0 : "*les paramètres β_j sont tous nuls* " contre H_1 : "*les paramètres β_j ne sont pas tous nuls*"

Effet de l'interaction des deux facteurs (le traitement et le sexe simultanément) :

H_0 : "*les paramètres γ_{ij} sont tous nuls* " contre H_1 : "*les paramètres γ_{ij} ne sont pas tous nuls*"

2. La technique adéquate pour l'analyse du problème est bien que l'ANOVA à deux facteur.
3. L'application de l'ANOVA 2 sur les données nous fournies ce qui suit :

(*a*) Les caractéristiques descriptives des données qui sont résumées dans la table suivante :

Hormone	Sexe	Moyenne	Variance
Hormone 1	Homme	14.880	4.9736
	Femme	12.120	3.5816
	Total	13.500	6.1820
Hormone 2	Homme	32.520	55.7736
	Femme	27.780	8.9456
	Total	30.150	37.9765
Total	Homme	23.700	108.1660
	Femme	19.950	67.5725
	Total	21.825	91.3849

TABLE 6.13: *Analyse descriptive de la variable dépendante : Concentration de Calcium dans le plasma.*

(*b*) La table de l'ANOVA 2 suivante :

Source	Somme des carrés	ddl	Moyenne des carrés	f_c
Hormone	1386.113	1	1386.113	60.534
Sexe	70.312	1	70.312	3.071
Hormone * Sexe	4.900	1	4.900	.214
Résiduelles	366.372	16	22.898	
Total	1827.698	19		

TABLE 6.14: *Décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs avec réplications.*

Supposons qu'on desire prendre la décision avec risque $\alpha = 5\%$, de se tromper. Alors d'après les résultats obtenus (voir table 6.14), on constate que :
La présence hormonal lors du traitement, à un effet significatif sur la concentration de calcium dans le plasma (le fait que $f_c > f_{(1,16,1-0.05)} = 4.49$), par contre y a aucune différence de concentration du calcium chez la femme ou l'homme, ce qui reste vraie pour l'interaction Sexe et Hormone qui n'a aucun effet significatif sur la concentration du calcium dans le plasma (le fait que $f_c < 4.49$ pour la variable Sexe et l'interaction des deux variables).

Solution 25 *(ANOVA 2 à un plan non équilibré)*

La réponse à cet exercice consiste à réaliser une ANOVA à deux facteurs lorsque le plan d'expérience est avec répétions de mesures non équilibrées.

D'après les résultats numérique obtenus (voir tableaux ci-dessous), on constate que :

Le choix du laborantin d'analyse, à un effet significatif sur le nombre de cellule (le fait que $f_c > f_{(2,5,1-0.05)} = 5.7861$), par contre y a aucune différence significative du nombre de cellule si on change la préparation (le fait que $f_c < 5.7861$), ce qui reste vraie pour l'interaction des deux facteurs, préparation et laborantin , qui n'a aucun effet significatif sur le nombre de cellule aussi(le fait que $f_c < 5.1922$).

Le tableau suivant résumes les caractéristiques des différents échantillons du problème posé :

Équipe	Laborantins	Moyenne	Variance	n_{ij}
A	1	49.50	2.25	2
	2	44.00	0	1
	3	69.00	36	2
	Total	56.20	128.56	5
B	1	62.00	0	1
	2	61.33	29.56	3
	3	75.00	0	1
	Total	64.20	46.96	5
C	1	55.50	2.25	2
	2	55.00	0	1
	3	83.00	0	1
	Total	62.25	144.69	4
Total	1	54.40	23.44	5
	2	56.60	63.44	5
	3	74.00	51	4
	Total	60.7857	116.31	14

Le tableau de l'analyse de la variance correspondant au problème pour $\alpha = 5\%$ est :

Source	Somme des carrés	d.d.l	Moyenne des carrés	f_c	$f_{(n_1,n_2,1-\alpha)}$
Préparation	172.007	2	86.004	2.534	5.7861
Laborantins	1189.65	2	594.825	17.529	5.7861
Préparation * Labo	97.033	4	24.258	.715	5.1922
Erreur	169.667	5	33.933		
Total	1628.357	13			

Solution 26 *(ANOVA 2 à un plan sans répétitions)*

Avant de répondre à l'exercice introduisons quelques informations utiles pour la réalisation d'une ANOVA à deux facteurs lorsque l'expérience n'est réalisée qu'une seule fois c'est-à-dire y a une seule observation par combinaison des deux facteurs.

Rappelons que la Table de décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs avec réplications ce présente comme suit :

Variation	*SC*	*ddl*	*CM*	F_{obs}
Facteur A	$KJ\sum_{i=1}^{I}\left(\overline{X}_{i\bullet\bullet}-\overline{X}\right)^2$	$I-1$	$CM_\alpha = SC_\alpha/dll$	CM_α/CM_{Res}
Facteur B	$KI\sum_{j=1}^{J}\left(\overline{X}_{\bullet j\bullet}-\overline{X}\right)^2$	$J-1$	$CM_\beta = SC_\beta/dll$	CM_β/CM_{Res}
Inter. $A\times B$	$K\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{ij\bullet}-\overline{X}_{i\bullet\bullet}-\overline{X}_{\bullet j\bullet}+\overline{X}\right)^2$	$(I-1)*(J-1)$	$CM_\gamma = SC_\gamma/dll$	CM_γ/CM_{Res}
Résiduelle	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(X_{ijk}-\overline{X}_{ij\bullet}\right)^2$	$I*J*(K-1)$	$CM_{Res} = SC_{Res}/dll$	
Totale	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(X_{ijk}-\overline{X}\right)^2$	$n-1$		

TABLE *6.17: Table de l'ANOVA d'ordre 2 cas : avec réplications équilibrées.*

avec I et J représentent le nombre de modalités ou niveaux des facteurs A et B, K l'effectif pour chaque combinaison des facteurs, où le nombre de réplications pour un plan est équilibrée, SC représente la somme des carrés, *ddl* le nombre de degrés de libertés et CM le carré moyen.

Cependant, lorsque le protocole expérimental ne comporte pas de réplications, c'est-à-dire qu'on ne dispose que d'une observation par combinaison de facteurs, il est toujours possible d'effectuer une ANOVA, mais l'analyse est limitée à la seule étude des effets principaux, avec comme hypothèse implicite supplémentaire l'absence d'interaction entre les facteurs. En effet, puisqu'il y n'y a qu'une seule observation pour chaque combinaison de chaque niveau des différents facteurs, il n'est plus possible d'estimer la variabilité intra pour cette combinaison particulière et l'on ne peut plus estimer la variabilité résiduelle à partir de ces intra-variabilités. Celle-ci doit donc être estimée à partir du CM de l'interaction (en fait, les composantes d'interaction et résiduelles sont confondues).

La décomposition de la variance suit le même principe que dans le cas des plans avec réplications (exception faite de l'indice de réplications) et est illustrée dans la table 6.18.

Variation	*SC*	*ddl*	*CM*	F_{obs}
Facteur A	$J\sum_{i=1}^{I}\left(\overline{X}_{i\bullet}-\overline{X}\right)^2$	$I-1$	$CM_A = SC_A/dll$	CM_A/CM_{Res}
Facteur B	$I\sum_{j=1}^{J}\left(\overline{X}_{\bullet j}-\overline{X}\right)^2$	$J-1$	$CM_B = SC_B/dll$	CM_B/CM_{Res}
Résiduelle	$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij}-\overline{X}_{i\bullet}-\overline{X}_{\bullet j}+\overline{X}\right)^2$	$(I-1)*(J-1)$	$CM_{Res} = SC_{Res}/dll$	
Totale	$\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij}-\overline{X}\right)^2$	$n-1$		

TABLE *6.18: Table de l'ANOVA d'ordre 2 cas sans réplications*

La table 6.18 représente (résume) la décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs sans réplications (SC représente la somme des carrés, ddl le nombre de degrés de libertés et CM le carré moyen.

1. La table de décomposition de la variance en fonction des sources de variabilité, pour un plan factoriel à 2 facteurs sans réplications correspondante au problème est donnée par 6.17. À partir de l'énoncé on a $I = 4$ et $J = 5$. De plus, le fait qu'il n'y a pas de réplications (sans répétions) alors $n = I * J = 20$, ces données nous permet de déterminer facilement les différents degrés de liberté (*d.d.l* de la troisième colonne de la table 6.18).
   (a) $I - 1 = \mathbf{3}$.
   (b) $J - 1 = \mathbf{4}$.
   (c) $(I - 1) * (J - 1) = \mathbf{12}$.
   (d) $n - 1 = \mathbf{19}$.

   Pour le reste des quantités on a :
   - $CM_{F_1} = SC_{F_1}/(I-1) = 140$, alors $SC_{F_1} = CM_{F_1} * (I-1) = 140 * 3 = \mathbf{420}$.
   - $F_c = CM_{F_1}/CM_{Res} = 3.5$, alors $CM_{Res} = CM_{F_1}/F_c = 140/3.5 = \mathbf{40}$.
   - $CM_{Res} = SC_{Res}/((I-1)(J-1))$, alors $SC_{Res} = CM_{Res} * ((I-1)(J-1)) = \mathbf{480}$.
   - $SC_{Tot} = CM_{Res} + SC_{F_1} + SC_{F_2}$, alors $SC_{F_2} = 1350 - 480 - 420 = \mathbf{450}$.
   - $CM_{F_2} = SC_{F_2}/(J-1)$, alors $CM_{F_2} = \mathbf{112.5}$
   - $F_{c_{F_2}} = CM_{F_2}/CM_{Res}$, alors $F_{c_{F_2}} = 112.5/40 = \mathbf{2.8125}$

   Ainsi, on aura la table suivante :

Variation	*SC*	*ddl*	*CM*	F_c
Facteur F_1	420	3	140	3.5
Facteur F_2	450	4	112.5	2.8125
Résiduelle	480	12	40	
Totale	1350	19		

TABLE 6.19: *Table de l'ANOVA associée au problème*

2. Sur la table de la loi de Ficher pour un seuil de confiance 0.95 ($\alpha = 5\%$) on obtient :

$f_\alpha = f_{(3,12,0.95)} = 3.49 < 3.5 \Rightarrow$ le premier facteur a une influence significative sur la variable X.
$f_\alpha = f_{(4,12,0.95)} = 3.26 > 2.8125 \Rightarrow$ le deuxième facteur n'a pas une influence significative sur X.

Solution 27 *(ANOVA 2 à un plan sans répétitions)*

La réponse à cet exercice consiste à réaliser une ANOVA à deux facteurs lorsque le plan d'expérience est sans répétitions de mesures. L'objectif du présent exercice est de mettre en évidence le lien de l'ANOVA 1 (facteur par facteur) avec l'ANOVA 2 lorsque n'y ont pas de répétitions. Le tableau suivant résume les caractéristiques des différents échantillons du problème posé :

	post					
équipe	A	B	C	D	Moyenne	Variance
équipe du matin	26	13	35	6	20	126.5
équipe du soir	18	17	31	2	17	105.5
équipe de nuit	31	24	33	4	23	131.5
Moyenne	25	18	33	4	20	113.5
Variance	28.6667	20.6667	2.6667	2.6667	6	127.1667

ANOVA 1 : Le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur post*, pour $\alpha = 5\%$, est :

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	1362	3	454	22.146	4.0662
Intra-groupes	164	8	20.5		
Total	1526	11			

Le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur équipe*, pour $\alpha = 5\%$, est :

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	72	2	36	0.223	4.2565
Intra-groupes	1454	9	161.556		
Total	1526	11			

D'après le tableau d'ANOVA du facteur *post* on constate que le facteur post a un effet significatif sur le nombre de pièces défectueuses et cela le fait que $f_c > f_\alpha$. D'après le tableau d'ANOVA du facteur *équipe* on constate que le facteur équipe n'a pas un effet significatif sur le nombre de pièces défectueuses et cela le fait que $f_c < f_\alpha$.

ANOVA 2 : Certes, il est pertinent de modéliser le tableau par un modèle à deux facteurs, mais au lieu du modèle complet :

$$Y_{ij} = m + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij},$$

on utilise le modèle suivant :

$$Y_{ij} = m + \alpha_i + \beta_j + \epsilon_{ij}.$$

A cet effet, pour $\alpha = 5\%$, on aura le tableau de l'analyse de la variance des pièces défectueuses selon le *facteur équipe* et le *facteur post* simultanément suivant :

Variation	SC	ddl	CM	F_{obs}	$f_{(n_1,n_2,1-\alpha)}$
post	1362	3	454	29.6093	4.7571
Équipe	72	2	36	2.3479	5.1433
Résiduelle	92	6	15.333		
Total	1526	11			

L'analyse de ce dernier tableau nous permet de faire les mêmes conclusions que dans la première partie de l'exercice (ANOVA 1).
Remarque : La réponse à cet exercice peut ce faire également par l'étude d'indépendance entre la variable équipe et la variable post on utilisons le test de $Khi - Deux$.

Solution 28 *(ANOVA 2 à un plan sans répétitions)*
Deux solution sont possible dans cette exercice, et cela selon l'hypothèse posée. En effet, si nous considérons *les Quadrants* comme facteur donc on réalise ANOVA 2 sans répétions pour répondre au problème, si nous considérons *les Quadrants* comme réplication donc on réalise une ANOVA 1 pour répondre au problème. Les résultats correspondant aux deux situations sont résumés, respectivement, dans les deux tables suivantes.

Source	*SC*	*ddl*	*MC*	f_c	$f_{(n_1,n_2,1-\alpha)}$
Quadrants	*42.736*	*2*	*21.368*	*2.117*	*6.9443*
Cages	*242.696*	*2*	*121.348*	*12.025*	*6.9443*
Erreur	*40.364*	*4*	*10.091*		
Total	*325.796*	*8*			

TABLE *6.24: Table d'ANOVA 2 sans répétions*

Source	*SC*	*ddl*	*MC*	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	*242.696*	*2*	*121.348*	*8.762*	*5.1433*
Intra-groupes	*83.100*	*6*	*13.850*		
Total	*325.796*	*8*			

TABLE *6.25: Table d'ANOVA 1*

- A partir les résultats rangés dans la table 6.24, on constate que pour un risque de 5%, les *Quadrants* n'ont pas d'influence sur la fécondité moyenne tandis que le facteur *cage* a une influence significative sur la fécondité. On conclue que la présence des insectes ont une influence sur fécondité moyenne des plantes en question.
- A partir les résultats rangés dans la table 6.25, on constate que pour un risque de 5%, le facteur *cage* a une influence significative sur la fécondité. On conclue que la présence des insectes ont une influence sur fécondité moyenne des plantes en question.

Solution 29 *(Test d'indépendance de $Khi - Deux$)*
Afin de vérifier si les deux variables *Traitement* et l'*état du malade après 5 jours de traitement* sont indépendantes, nous utilisons le test d'indépendance de $Khi - Deux$.
Les effectifs observés, attendus et ainsi que l'écart entre eux sont résumés dans le tableau suivant :

			État			Total
			Stationnaire	Amélioré	Guéri	
Traitement	T_1	Effectif	15	70	35	120
		Effectif théorique	19.2	74.4	26.4	120
		Résidu	-4.2	-4.4	8.6	
	T_2	Effectif	25	85	20	130
		Effectif théorique	20.8	80.6	28.6	130
		Résidu	4.2	4.4	-8.6	
Total	Effectif		40	155	55	250
	Effectif théorique		40.0	155.0	55.0	250

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique de *Khi-Deux* avec la valeur tabulée de *Khi-Deux* pour $r-1$ degré de liberté et un risque α qu'on doit fixer préalablement. Supposons que ce dernier est fixé à $\alpha = 5\%$.

On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 7.6548.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(3-1),0.05)} = 5.991$. A cet effet, on conclut que le traitement et l'état du malade après 5 jours sont dépendants (liés) le fait que $7.6548 > 5.991$.

Solution 30 *(Test d'indépendance de Khi − Deux)*

Le test statistique adéquat est bien que le test d'indépendance de χ^2. Les résultats fournis par ce test, appliqué sur nos données, sont résumés dans le tableau suivant :

			Névrose		Total
			Présente	Absente	
Mode de vie	En famille	Effectif	40	60	100
		Effectif théorique	53.8	46.2	100
		Résidu	-13.8	13.8	
	Seul	Effectif	100	60	160
		Effectif théorique	86.2	73.8	160
		Résidu	13.8	-13.8	
Total		Effectif	140	120	260

on a :

$$K_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 12.536. \tag{6.19}$$

$$et \quad \chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(2-1),0.01)} = \chi^2_{(1,0.01)} = 6.635. \tag{6.20}$$

De (6.19) et (6.20), on conclut qu'il y a un lien entre le mode de vie et la névrose le fait que $K_n^2 > \chi^2_{(1,0.01)}$.

Solution 31 *(Test d'indépendance de Khi − Deux)*

Le calcul des $n_{ij} = \frac{n_{i\bullet}\times n_{\bullet j}}{n}, i = 1, 2, 3$ *et* $j = 1, 2, 3, 4$, nous fournis ce qui suit :

			Rang				Total
			Rang 1	Rang 2	Rang 3	Rang 4	
Groupe	Groupe 1	Effectif	4	6	8	6	24
		Effectif théorique	5.9	5.9	6.7	5.5	24.0
		Résidu	-1.9	0.1	1.3	0.5	
	Groupe 2	Effectif	4	3	5	5	17
		Effectif théorique	4.2	4.2	4.8	3.9	17.0
		Résidu	-0.2	-1.2	0.2	1.1	
	Groupe 3	Effectif	6	5	3	2	16
		Effectif théorique	3.9	3.9	4.5	3.6	16.0
		Résidu	2.1	1.1	-1.5	-1.6	
Total		Effectif	14	14	16	13	57
		Effectif théorique	14.0	14.0	16.0	13.0	57.0

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique de *Khi-Deux* avec la valeur tabulée de *Khi-Deux* pour $r - 1$ degré de liberté et un risque α qu'on doit fixer préalablement.

On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 4.195.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((4-1)(3-1),0.05)}$ est de 12.592. A cet effet, on

conclu que le choix de rang est indépendant du groupe le fait que $k_n^2 < \chi^2_{(6,0.05)}$, cela signifié que les étudiants choisissent leurs places arbitrairement sans prendre en considération leurs appartenance à un groupe bien déterminé.

Solution 32 *(Test d'indépendance de $Khi - Deux$)*
Le test statistique adéquat pour déterminer s'il y a une différence d'activité des deux levures ou non est bien que le test d'indépendance ce χ^2. Les résultats fournis par ce test, appliqué sur nos données, sont résumés dans le tableau suivant :

			Aptitude à lever			Total
			Moyenne	Bonne	Très Bonne	
Levure	*A*	Effectif	41	16	63	120
		Effectif théorique	34.4	23.5	62.2	120
		Résidu	6.6	-7.5	0.8	
	B	Effectif	22	27	51	100
		Effectif théorique	28.6	19.5	51.8	100
		Résidu	-6.6	7.5	-0.8	
Total	Effectif		43	63	114	220

Pour répondre à notre objectif, il suffit de comparé la réalisation de la statistique du test avec la valeur critique qui est le fractile d'ordre $\alpha = 5\%$ d'une loi de *Khi-Deux* à $(r-1)*(k-1)$ degré de liberté, avec $r = 2$ (nombre de levures) et $k = 3$ (nombre de critères).

$$K_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(N_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} \approx 8.1, \tag{6.21}$$

et la valeur critique du test est donnée par :

$$\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((2-1)(3-1),0.05)} = 5.991 \tag{6.22}$$

De (6.21) et (6.22), on conclu qu'il y a une différence entre l'activité des deux levures.

Solution 33 *(Test d'indépendance de $Khi - Deux$)*
La réponse à la première question de cet exercices consiste à réalisé une analyse de la variance à un seul facteur sur les notes des étudiants selon l'assistant, tandis que la deuxième concerne l'indépendance des deux variables note de l'étudiant et l'assistant dont la réponse se fait par le test d'indépendance du $Khi - Deux$.

1. Le tableau suivant résume les caractéristiques descriptives (moyenne et variance) des différents échantillons du problème posé :

	n	Moyenne	Variance
Assistant 1	12	5.0000	6.3750
Assistant 2	12	4.4167	4.9514
Assistant 3	13	6.8846	4.2367
Total	37	5.4730	6.2966

La table 6.31 résume les résultats de l'analyse de la variance des notes des étudiants selon le facteur Assistant.

Source	SC	d.d.l	MC	f_c	$f_{(n_1,n_2,1-\alpha)}$
Inter-groupes	41.979	2	20.990	3.737	3.2759
Intra-groupes	190.994	34	5.617		
Total	232.973	36			

TABLE 6.31: *Table de l'ANOVA associée au problème des notes selon l'Assistant*

A partir de cette dernière table, on constate que $f_c > f_\alpha$ cela signifie que le choix de l'Assistant influe sur les notes moyenne des groupes.

2. On admet que l'étudiant a échoué s'il a une note inférieur strictement à 5 ($note < 5$) et il a réussit si sa note est supérieure ou égale à 5 ($note \geq 5$), ansi le tableau croisé "Assistant" × "réussite et échec" est donné comme suit :

	État	
Assistant	Échec	réussite
A_1	5	7
A_2	7	5
A_3	3	10

Les effectifs observés, attendus et ainsi que l'écart entre eux sont résumés dans le tableau suivant :

			État		Total
			Échec	Réussite	
Assistant	A_1	Effectif	7	5	12
		Effectif théorique	4.9	7.1	12.0
		Résidu	2.1	-2.1	
	A_2	Effectif	5	7	12
		Effectif théorique	4.9	7.1	12.0
		Résidu	.1	-.1	
	A_3	Effectif	3	10	13
		Effectif théorique	5.3	7.7	13.0
		Résidu	-2.3	2.3	
Total		Effectif	15	22	37
		Effectif théorique	15.0	22.0	37.0

On a d'une part, la réalisation de la statistique du test est donnée par :

$$k_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet}\times n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\times n_{\bullet j}}{n}} = \sum_{i=1}^{r}\sum_{j=1}^{k} e_{ij}^2/n_{ij} = 3.8027.$$

et d'autre par la valeur tabulée de $\chi^2_{((r-1)(k-1),\alpha)} = \chi^2_{((3-1)(2-1),0.05)}$ est 5.991. A cet effet, on conclu que les résultats des étudiants(échec ou réussite) sont indépendant du l'assistant qui assure le TP.

Solution 34 *(Régression linéaire simple)*

1. À partir de la présentation graphique (voir figure 6.1), on constate que le nuage des points est distribué sous une forme linéaire, à priori le modèle proposé est adéquat pour l'explication de Y en fonction de x.

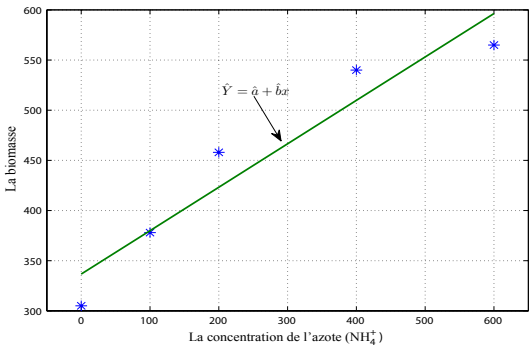


FIGURE 6.1: *Présentation graphique du nuage des points* (X_i, Y_i)

2. On a d'une part :

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} \qquad \text{et} \qquad \hat{a} = \overline{Y} - \hat{b}\overline{X}. \tag{6.23}$$

et d'autre part :

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}1300 = 260, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}2246 = 449.2, \\ Cov(x,y) &= \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\ \overline{Y} = \frac{1}{5}(684400) - (260)\ (449.2) = 20088, \\ Var(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{5}(570000) - (260)^2 = 46400, \end{aligned}$$

ainsi,

$$\hat{b} = 0.4329, \qquad \text{et} \qquad \hat{a} = 336.6460,$$

de ce fait, la droite de régression de la biomasse (Y) en fonction de la concentration (x) est :

$$\hat{Y} = 0.4329\ x\ +\ 336.6460.$$

3. On a d'une part,

$$r = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \tag{6.24}$$

et d'autre part : $Cov(x,y) = 20088$, Écart-type$(x) = \sqrt{46400} = 215.4066$ et Écart-type$(Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{\frac{1}{5}(1056498) - (449.2)^2} = \sqrt{9518.36} = \mathbf{97.5621}$, alors

$$\rho = \rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.9559 = 95.59\%, \tag{6.25}$$

Le fait que la valeur de $\rho \approx 1$, on déduit qu'il y a une forte liaison linéaire entre x et Y.

4. Afin de valider le modèle nous aurons besoin des $\hat{Y}_i = 0.4329\ x_i\ +\ 336.6460$ dont leurs valeurs sont rangées dans le tableau suivant :

Concentration (μmol)	0	100	200	400	600
Biomasse (mg)	305	378	458	540	565
$\hat{y}_i$ (mg)	336.646	379.936	423.226	509.806	596.386
$e_i = y_i - \hat{y}_i$	-31.646	-1.936	34.774	30.194	-31.386

On a d'une part

$$f_c = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n} e_i^2/(n-2)} = \frac{43477.3591/1}{4111.2071/(5-2)} = 31.7260,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1,3,0.95) = 10.1.$$

On constate que $f_c > f_\alpha$, alors le modèle est valide (pertinent), c'est-à-dire on admet qu'on peut expliquer la Biomasse de la plante en fonction de la concentration de l'azote par la droite

$$\hat{Y} = 0.4329\ x\ +\ 336.6460.$$

5. On a : $\hat{Y} = 0.4329\ x\ +\ 336.6460$ alors la Biomasse qu'on peut prévoir à une concentration 500 μmol est $\hat{Y} = 0.4329 * 500\ +\ 336.6460 = 553.0960mg$.

Solution 35 *(Régression linéaire simple)*

1. Par définition le coefficient de corrélation linéaire est donné par :

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

on a : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{19690}{10} = 1969, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{20.3}{10} = 2.03,$

$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = \sqrt{\frac{1}{10}42925500 - 1969^2} = 679.5333,$

$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{\frac{1}{10}162.4100 - 2.03^2} = 3.6697,$

$Cov(x,y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\ \overline{Y} = \left(\frac{1}{10}17671\right) - 1969 \times 2.03 = -2.22997,$

alors le coefficient de corrélation est :

$$r = -0.9936$$

2. Calculer les estimations des paramètres a, b et σ^2 pour la régression linéaire de Y sur X.
3. Le modèle linéaire de Y sur X est donné par :

$$Y = aX + b + \epsilon,$$

en utilisant la méthode des moindres carrées les estimateurs de a et b sont définis comme suite :

$$\boxed{\hat{a} = \frac{Cov(x,y)}{Var(x)} = -0.0054.} \text{ et } \boxed{\hat{b} = \overline{Y} - \hat{a}\overline{X} = 12.5953} \tag{6.26}$$

c'est-à-dire, la droite de régression est :

$$\hat{Y} = -0.0054X + 12.5953.$$

on a,

$$\hat{\sigma}_c^2 = var(\epsilon) = var(y - \hat{a} - \hat{b}x) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2, \tag{6.27}$$

donc

$$\hat{\sigma}_c^2 = \frac{1}{10-2}\sum_{i=1}^{10}(y_i - \hat{a} - \hat{b}x_i)^2 = 0.1931.$$

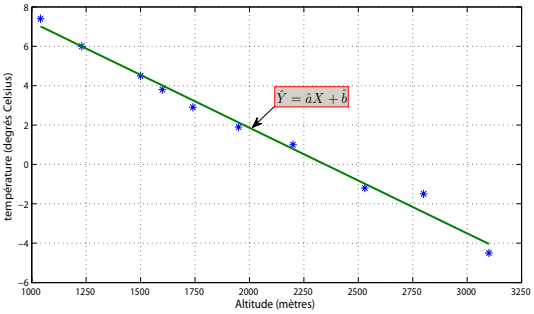


FIGURE 6.2: *Nuage des points observés et la droite de régression*

4. Les températures moyennes correspondantes à aux altitudes 1100 m et 2300 m.
   (a) $1100m$ est : $y = -0.0054 * 1100 + 12.5953 = 6.6929$.
   (b) $2300m$ est : $y = -0.0054 * 2300 + 12.5953 = 0.2539$.

Solution 36 On note X est le poids, Y est le Prix et le modèle de régression est $Y = aX + b$.

1. A partir des données on a :

variable	Moyenne	Variance	Carrée Moyenne
X	138.1667	1426.4722	20516.50
Y	166.6667	3222.2222	31000
$X * Y$	24643.33		

   d'où : $Cov(X, Y) = 1615.54$, $\rho = 0.754$, $\hat{a} = 1.133$, $\hat{b} = 10.186$ et $\hat{Y} = 1.133X + 10.186$.

2. Si on augmente le poids du Sandwich S_6 à 180 g, alors son nouveau prix sera :
   $\hat{Y} = 1.133(180) + 10.186 = 214.126DA$.

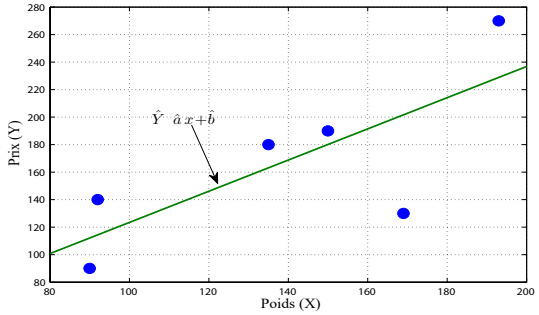


FIGURE 6.3: *Nuage de variation du Prix des Sandwichs en fonction de leurs Poids.*

3. La table d'analyse de la variance, du modèle, est donnée comme suite :

Source	SC	ddl	MC	f_c	Sig.
Régression	10978.215	1	10978.215	5.256	0.084
Résidu	8355.118	4	2088.780		
Total	19333.333	5			

TABLE 6.34: *Table d'ANOVA du modèle*

   A partir de ces résultats, on constate que pour un risque de 5%, le modèle linéaire n'est pas adéquat pour la description de la relation entre les variables Poids et Prix. Mais on peut conclure que ce modèle est adéquat pour un risque de 10%.

Solution 37 *(Régression linéaire simple et transformation des variables)*
Pour faire une régression linéaire, on effectue un changement de variable en posant $X = \ln(D)$ et $Y = \ln(H)$. Après le calcul des valeurs des variable X et Y on aura les résultats suivants :

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65
$\hat{Y}$	2.1690	2.3255	2.4133	2.5889	2.6232
ϵ	0.0510	-0.0555	-0.0333	0.0111	0.0268

1. Le calcul du coefficient de corrélation linéaire entre X et Y nécessite les quantités suivantes :
   $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = -0.9420$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.4240$,

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2} = \sqrt{0.1945} = 0.4410,$$
$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2} = \sqrt{0.0299} = 0.1728,$$
$$Cov(x,y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\ \overline{Y} = 0.0742,$$

ainsi on aura le coefficient de corrélation :

$$r = 0.9737.$$

2. On a,

$$\hat{a} = \frac{Cov(x,y)}{Var(x)} = 0.3817 et \hat{b} = \overline{Y} - \hat{a}\overline{X} = 2.7836.$$

alors,

$$Y = 0.38172X + 2.78358, \tag{6.28}$$

3. Le test de validation du modèle se base sur la statistique :

$$F = \frac{\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} \rightsquigarrow f_{(1,n-2)},$$

or on a,

$$\sum_{i=1}^{n}(\hat{y} - \overline{Y})^2/1 = 0.1417$$

et

$$\sum_{i=1}^{n}(y_i - \hat{y})^2/(n-2) = 0.0025$$

alors la réalisation, f, de la statistique F est égale à : 55.7266.
A partir de la table de Fisher pour un seuil de risque $\alpha = 5\%$, on obtient $f_{(1,n-2,1-\alpha)} = f_{(1,3,0.95)} = 10.1$, on constate que la valeurs de la réalisation de la statistique F est supérieur à la valeurs tabulée de fisher, cela signifier que le modèle est valide c'est-à-dire le modèle linéaire définie dans (6.28) est adéquat pour l'explication de la variable Y en fonction de la variable X.

4. Donner la hauteur prévue d'un arbre de diamètre 0.7. on a,

$$\hat{Y} = 0.38172X + 2.78358 \Rightarrow \ln(\hat{H}) = 0.38172\ln(D) + 2.78358 \Rightarrow \hat{H} = e^{0.38172\ln(D)+2.78358}.$$

Alors, pour un diamètre D=0.7, on prévoit une hauteur $H = e^{0.38172\ln(0.7)+2.78358} = 14.1177$.

Solution 38 *(Régression linéaire simple et changement des variables)*

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Y	0	0.205	0.331	0.515	0.584	0.671	2.3060
X^2	0	400	1600	3600	6400	10000	22000
Y^2	0	0.0420	0.1096	0.2652	0.3411	0.4502	1.2081
$X * Y$	0	4.10	13.24	30.90	46.72	67.10	162.06

a) Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y \;=\; a_1\, x \;+\; b_1.$$

1. Calcul des estimateurs des paramètres a_1 et b_1. On a :

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}300 = 50.\\
\overline{Y} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}3002.3060 = 0.3843.\\
Cov(x,y) &= \frac{1}{n}\sum_{i=n}^{n} x_i y_i - \overline{X}\;\overline{Y} = \frac{1}{6}(162.06) - (50)\;(0.3843) = 7.7950\\
Var(x) &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{1}{6}(22000) - (50)^2 = 1166.6667
\end{aligned}$$

alors,

$$\hat{a}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n}\sum\limits_{i=n}^{n} x_i y_i - \overline{X}\;\overline{Y}}{\frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 - \overline{X}^2} = 0.0067.$$

$$\hat{b}_1 = \overline{Y} - \hat{a}_1\overline{X} = 0.0503,$$

de ce fait la droite de régression de l'absorbance (Y) en fonction de la concentration (x) est donnée par :

$$\hat{Y} = 0.0067\; x \;+\; 0.0503.$$

2. Quelle absorbance prévoyez-vous à une concentration 50 $\mu g/\mu l$?

$$\hat{Y} = 0.0067\;(50) \;+\; 0.0503 \;=\; 0.3853.$$

3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$? Que peut-on conclure ?

$$\hat{Y} = 0.0067\;(40) \;+\; 0.0503 = 0.3183.$$

On constate que la valeur de régression est très proche de la vraie valeur (0.331), donc à priori le modèle retenu est adéquate pour la représentation des données du tableau.

4. Calcul du coefficient de corrélation linéaire.

$$r = r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y} = 0.9851,$$

avec $\sigma_y = \sqrt{var(Y)} = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n} y_i^2 - \overline{Y}} = \sqrt{0.0536} = 0.2316$; La valeur du coefficient de corrélation est très proche de 1, i.e. X et Y sont fortement linéairement liés donc le modèle est efficace ce qui confirme les résultats de la question 3).

5. Pour un seuil de risque $\alpha = 5\%$, le modèle proposé est-il pertinent ?
Pour répondre à cette question on utilise le test de validation du modèle (Fisher). On d'une part

$$f_c = \frac{\sum\limits_{i=1}^{n}(\hat{y} - \overline{Y})^2/1}{\sum\limits_{i=1}^{n}(y_i - \hat{y})^2/(n-2)} = \frac{0.3124/1}{0.0095/(6-2)} = 131.5368,$$

et d'autre par

$$f_\alpha = f(1, n-2, 1-\alpha) = f(1, 4, 0.95) = 7.71.$$

On constate que $f_c > f_\alpha$, alors on accepte le modèle proposé, c'est-à-dire le modèle est valide (pertinent)

b) Vue les doutes qu'on a sur le modèle précèdent, nous avons proposé le modèle suivant :

$$Z \;=\; e^Y \;=\; a_2\, x \;+\; b_2.$$

1. Complétez le tableau suivant :

							Somme
X $\mu g/\mu l$	0	20	40	60	80	100	300
Z	1.0000	1.2275	1.3924	1.6736	1.7932	1.9562	9.0429
Z^2	1.0000	1.5068	1.9387	2.8011	3.2156	3.8267	14.2888
$X * Z$	0	24.5505	55.6944	100.4183	143.4558	195.6193	519.7382

2. Calculer les estimations des paramètres a_2 et b_2 pour la régression linéaire de Z sur X.

$$\begin{aligned}
\overline{X} &= 50. \\
\overline{Z} &= \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{6}(9.0429) = 1.5072. \\
Cov(x,z) &= \frac{1}{n}\sum_{i=n}^{n} x_i z_i - \overline{X}\;\overline{Z} = \frac{1}{6}(162.06) - (50)\;(0.3843) = 7.7950 \\
Var(x) &= 1166.6667
\end{aligned}$$

alors,

$$\hat{a}_2 = \frac{Cov(x,z)}{Var(x)} = 0.0097.$$

$$\hat{b}_2 = \overline{Z} - \hat{a}_2\overline{X} = 1.0243,$$

de ce fait la droite de régression de (Z) en fonction de (x) est donnée par :

$$\hat{Z} = 0.0097\; x \;+\; 1.0243.$$

3. Quelle absorbance prévoyez-vous à une concentration 40 $\mu g/\mu l$. Que peut-on conclure par rapport au premier modèle ?
On $Z = 0.0097\;(40) + 1.0243 = 1.4123$ donc l'absorbance $y = log(z) = log(1.4123) = 0.3452$.
On constate que se modèle nous fournit une valeur proche à la vraie valeur mais c'est le premier modèle qui nous fournis une valeur plus proche d se fait il se peut que c'est le premier modèle qui est meilleur.
4. Calculer le coefficient de corrélation linéaire de ce nouveau modèle.

$$r = r(x,y) = \frac{Cov(x,z)}{\sigma_x \sigma_z} = 0.9946,$$

5. On constate que le coefficient de corrélation est plus grand pour le deuxième modèle donc le meilleur modèle est le deuxième.

Bibliographie

[1] J. Bass, *Eléménts de calcul de probabilités*. Masson, 1974.

[2] N. Ben Righi, M. Cherfaoui *Estimation paramétrique : Intervalle et région de confiance.* Mémoire Master en Mathématique Option Statistique, Université de Biskra, 2016.

[3] G. Calot, *Cours de calcul des probabilités*. Dunod, 1967.

[4] D. Foudrinier, *Statistique inférentielle : Cours et exercices*. Dunod, Paris 2002.

[5] H. Gudeida, A. Roubi *Tests de comparaison.* Mémoire Master en Mathématique Option Statistique, Université de Biskra, 2016.

[6] J. Guégand, J. P. Gavini, *Probabilités.* 1998.

[7] K. Khaldi, *Méthodes statistique et Probabilités.* Casbah, 2000.

[8] A. Krief, S. Levy, *Calcul des probabilités.* Hermann, 1972.

[9] M. Laviéville, *Statistique et Probabilités : Rapepels de cours et exercuces résolus.* Dunod, 1996.

[10] J.P. Lecoutre, S. Legait, P. Tassi, *Statistique : Exercices corrigés et rappels de cours.* Masson, 1987.

[11] J.P. Lecoutre, *Statistique et probabilité, manuel et exercices corrigés.* quatrième édition. Masson, 2009.

[12] M. Sheldon, M. Ross, *Initiation aux probabilités*. Presses polytechniques et universitaires normandes, 1994.

[13] G. Saporta, *Probabilité, analyse des données et statistique.* Editions Technip, 1990.

[14] P. Tassi, *Méthodes statistiques.* Edition Economica, 2004.

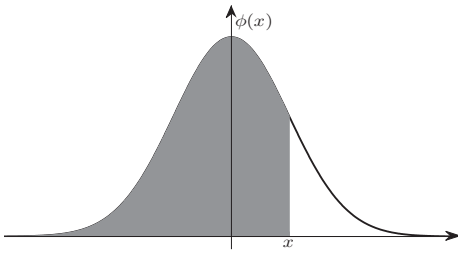
Annexe : Tables des lois statistique

Dans cette section on présente les tables les plus représentatives des lois statistiques les plus usuelles. En effet, cette annexe contient les cinq tables suivantes :

1. Table de la loi normale
2. Table de la loi de Student
3. Table de la loi du $Khi - Deux$
4. Table de la loi de Fisher-Snédécor
5. Table des valeurs critiques du test d'justement Kolmogorov-Smirnov

Fonction de répartition de la loi normale centrée réduite

(probabilité $\phi(x)$ de trouver une valeur inférieure à x)

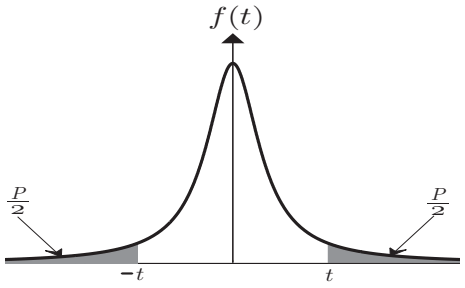


x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.60	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.70	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.80	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.90	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

Table de la loi de Student

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

fournit les quantiles t tels que
$P(|T| \geq t) = p$ pour $t \rightsquigarrow t_n$



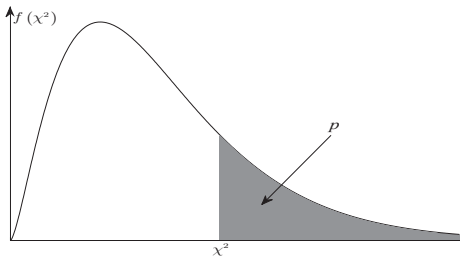
n \ p	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.128	0.260	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

- n est le nombre de degrés de liberté.
- Le quantile d'ordre $1 - \frac{\alpha}{2}$ se lit dans la colonne $P = \alpha$.
- Le quantile d'ordre $1 - \alpha$ se lit dans la colonne $P = 2\alpha$.

Table de la loi du $Khi - Deux$

Valeurs de χ^2 ayant la probabilité P d'être dépassées

fournit les quantiles χ^2 tels que
$P(X \geq \chi^2) = p$ pour $X \rightsquigarrow \chi^2_n$

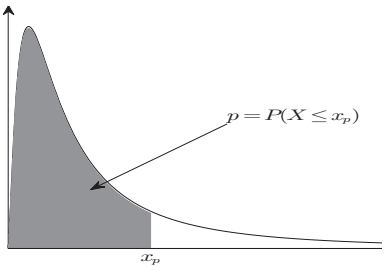


n \ p	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00004	0.0002	0.001	0.0039	0.0158	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672

- n est le nombre de degrés de liberté.
- Pour $n > 30$, on peut admettre que la quantité $\sqrt{2\chi^2} - \sqrt{2n-1}$ suit la loi normale centrée réduite.

Tables des quantiles de la $v.a.$ de Fisher-Snédécor

fournit les quantiles x_p tels que
$P(X \leq x_p) = p$ pour $X \rightsquigarrow F_{(n_1;n_2)}$



$$P(X \leq x_p) = 0.95$$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	252	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.44	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.75	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.32	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.02	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.80	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.64	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.51	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.40	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.31	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.24	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.18	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.12	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.08	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.04	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.97	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.92	1.84	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.76	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	1.96	1.88	1.79	1.70	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.66	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.89	1.81	1.71	1.63	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.60	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.56	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62	1.53	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70	1.60	1.51	1.32
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69	1.59	1.49	1.30
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.48	1.28
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73	1.64	1.54	1.44	1.22
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.52	1.41	1.19
∞	3.75	2.97	2.59	2.36	2.21	2.09	2.01	1.94	1.88	1.83	1.67	1.57	1.46	1.35	1.00

Tables des quantiles de la $v.a.$ de Fisher (suite)

$$P(X \leq x_p) = 0.975$$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	648	800	864	900	922	937	948	957	963	969	985	993	1001	1008	1018
2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5
3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.2	14.1	14.0	13.9
4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.38	8.26
5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.14	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	4.98	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.28	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.81	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.47	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.22	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.12	3.03	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	2.96	2.87	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.84	2.74	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.73	2.64	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.55	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68	2.57	2.47	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62	2.50	2.41	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56	2.44	2.35	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51	2.39	2.30	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.25	2.09
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.30	2.18	2.08	1.91
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	1.97	1.79
35	5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.24	2.12	2.00	1.89	1.70
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.83	1.64
45	5.38	4.01	3.42	3.09	2.86	2.70	2.58	2.49	2.41	2.35	2.14	2.03	1.90	1.79	1.59
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.99	1.87	1.75	1.55
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.70	1.48
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24	2.03	1.91	1.78	1.66	1.44
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21	2.00	1.88	1.75	1.63	1.40
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19	1.98	1.86	1.73	1.61	1.37
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.85	1.71	1.59	1.35
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92	1.80	1.67	1.54	1.27
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90	1.78	1.64	1.51	1.23
∞	4.93	3.67	3.11	2.78	2.56	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.43	1.00

$$P(X \leq x_p) = 0.99$$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	15	20	30	50	∞
1	4052	5000	5403	5625	5764	5859	5982	6022	6056	6157	6209	6261	6303	6334	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	26.9	26.7	26.5	26.4	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.2	14.0	13.8	13.7	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.72	9.55	9.38	9.24	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.09	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.86	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.07	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.52	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.12	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.81	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.57	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.38	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.22	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.08	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	2.97	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.00	2.87	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.78	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.84	2.71	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.64	2.42
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.54	2.40	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.25	2.01
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.60	2.44	2.28	2.14	1.89
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.06	1.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.46	2.31	2.14	2.00	1.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70	2.42	2.27	2.10	1.95	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.88	1.60
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.31	2.15	1.98	1.83	1.54
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.12	1.94	1.79	1.49
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52	2.24	2.09	1.92	1.76	1.46
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.74	1.43
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16	2.00	1.83	1.66	1.33
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13	1.97	1.79	1.63	1.28
∞	6.59	4.61	3.79	3.33	3.02	2.81	2.64	2.52	2.41	2.32	2.04	1.88	1.70	1.52	1.00

Table des valeurs critiques du test d'justement Kolmogorov-Smirnov

n \ $1-\alpha$	0.90	0.95	0.99
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.510	0.563	0.669
6	0.468	0.520	0.617
7	0.436	0.483	0.576
8	0.410	0.454	0.542
9	0.387	0.430	0.513
10	0.369	0.409	0.489
11	0.352	0.391	0.468
12	0.338	0.375	0.450
13	0.325	0.361	0.432
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.327	0.392
17	0.286	0.318	0.381
18	0.279	0.309	0.371
19	0.271	0.301	0.361
20	0.265	0.294	0.352

n \ $1-\alpha$	0.90	0.95	0.99
21	0.259	0.287	0.344
22	0.253	0.281	0.337
23	0.247	0.275	0.330
24	0.242	0.269	0.323
25	0.238	0.264	0.317
26	0.233	0.259	0.311
27	0.229	0.254	0.305
28	0.225	0.250	0.300
29	0.221	0.246	0.295
30	0.218	0.242	0.290
31	0.214	0.238	0.285
32	0.211	0.234	0.281
33	0.208	0.231	0.277
34	0.205	0.227	0.273
35	0.202	0.224	0.269
> 35	$\frac{1.224}{\sqrt{n}}$	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$