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

Analyse de Données en Biosciences
Troisième Année Licence

Préparé par :
Dr. CHERFAOUI Mouloud

Université de Biskra, 2020/2021

Table des matières

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 $\Delta_2 = n_i - n_{i+1}$ 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.

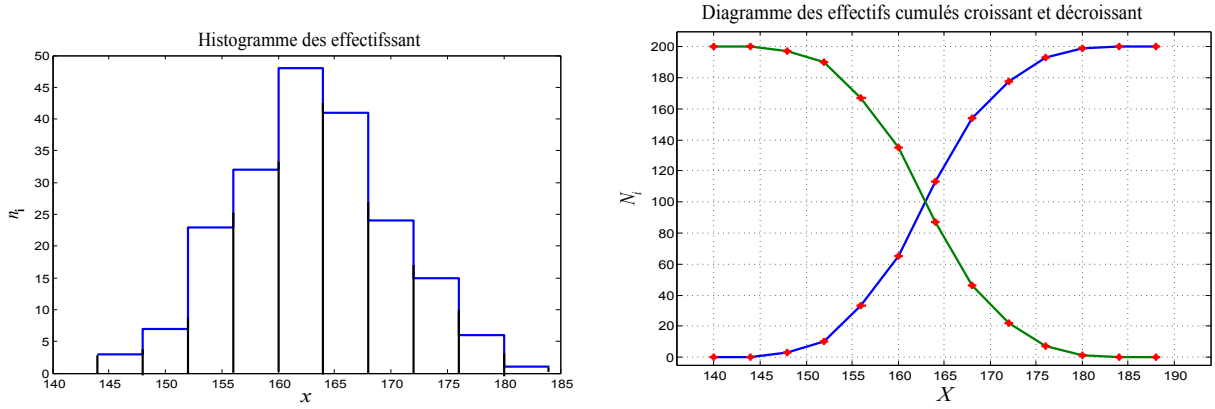

FIGURE 1.1: *Histogramme des effectifs et la courbe des effectifs cumulées*

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).

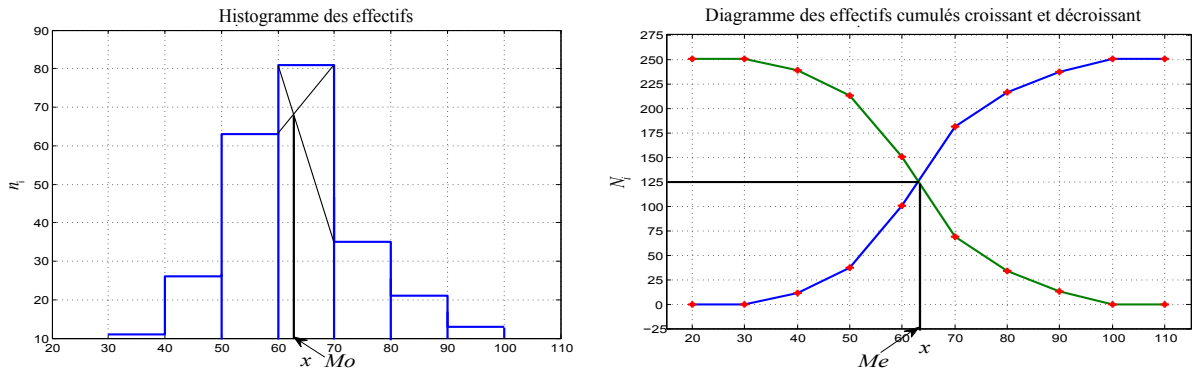

FIGURE 1.2: *Détermination graphique du mode (l'histogramme) et de la médiane (fréquences cumulées)*

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.

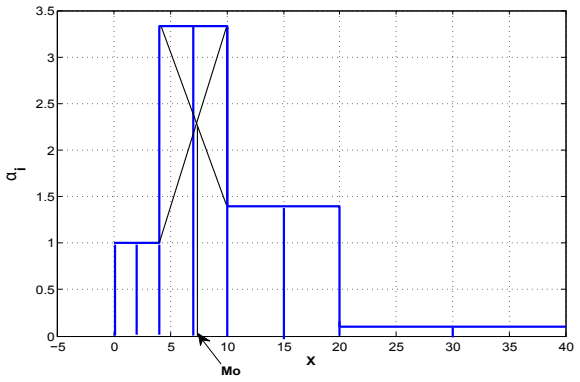

FIGURE 1.3: *Histogramme des effectifs corrigées et détermination du mode*

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 :

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

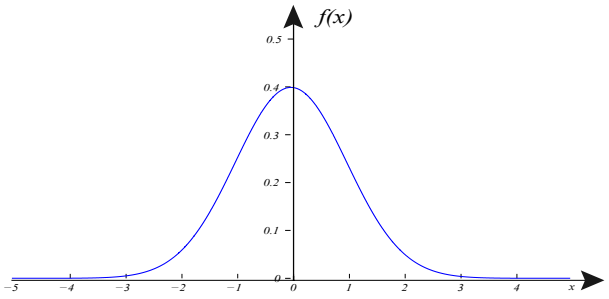

FIGURE 1.4: *Distribution de Gauss centrée et réduite*

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 de 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

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

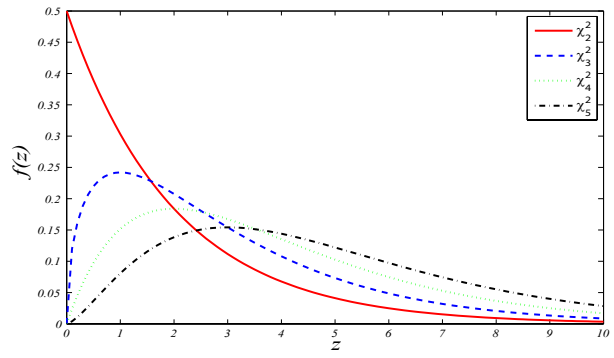

FIGURE 1.5: *Distribution de χ_2 pour différentes degré de liberté.*

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 $n > 2$.

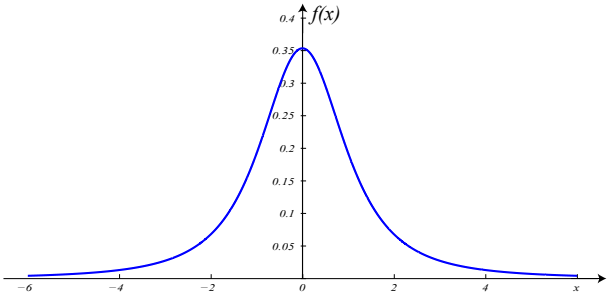

FIGURE 1.6: *Distribution de Student.*

1.3.4 La distribution de 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

2.1 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.

2.2 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.

2.2.1 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.

$$P(a \leq \theta \leq b) = 1 - \alpha, \tag{2.1}$$

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).

2.2.2 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 :

$$P(|f_n - p| < \epsilon) = 1 - \alpha, \tag{2.2}$$

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 :

$$IC_{1-\alpha}(p) = \left[f_n - q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\frac{f_n(1-f_n)}{n}}; f_n + q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \sqrt{\frac{f_n(1-f_n)}{n}} \right], \tag{2.3}$$

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.

2.2.3 Estimation de la moyenne d'une loi normale par IC

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.

2.2.3.1 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 :

$$P\left(-q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} < \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} < q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \right) = 1 - \alpha, \tag{2.4}$$

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,

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}}; \hat{\mu} + q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)} \frac{\sigma}{\sqrt{n}}\right]. \tag{2.5}$$

2.2.3.2 Cas l'écart-type σ est inconnu

La statistique utilisée dans le cas précédant

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right), \quad (\text{ avec } \hat{\mu} = \overline{X}), \tag{2.6}$$

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 (2.6) 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 :

$$P\left(-q_{\frac{\alpha}{2}}^{t_{n-1}} < \frac{\hat{\mu} - \mu}{\hat{\sigma}_c/\sqrt{n}} < q_{\frac{\alpha}{2}}^{t_{n-1}}\right) = 1 - \alpha, \tag{2.7}$$

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 :

$$IC_{1-\alpha}(\mu) = \left[\hat{\mu} - q_{\frac{\alpha}{2}}^{t_{n-1}} \frac{\hat{\sigma}_c}{\sqrt{n}}; \hat{\mu} + q_{\frac{\alpha}{2}}^{t_{n-1}} \frac{\hat{\sigma}_c}{\sqrt{n}}\right]. \tag{2.8}$$

Chapitre 3

Introduction à la théorie de test d'hypothèses ”

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

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

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 μ_0', μ_0'', ... 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 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.

3.2.1 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.

3.2.2 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.

3.2.2.1 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

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \tag{3.1}$$

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 (3.1) 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).

3.2.2.2 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 :

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}. \tag{3.2}$$

Sous l'hypothèse H_0 et l'hypothèse $\sigma_1 = \sigma_2$ la statistique du test définie dans (3.2) 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 3.2.1).

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 (3.2) 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 (3.2) 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.

3.3 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.

3.3.1 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 3.1: *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 :

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

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 3.2.2) multiples. Remarquez que l'ANOVA peut aussi être utilisée quand $p = 2$ puisque, elle retourne la même conclusion qu'un test t.

3.3.2 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 :

$$X_{ij} = \mu_i + \epsilon_{ij}, \text{ avec } i = \overline{1,n}, j = \overline{1,p} \text{ et } \epsilon_{ij} \rightsquigarrow N(0,\sigma^2), \tag{3.3}$$

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 :

$$H_0 : \text{''}\mu_1 = \mu_2 = ... = \mu_p = \mu\text{''} \text{ contre } H_1 : \text{''}\exists\, i,j \in \{1,2,...,p\} \text{ tel que } \mu_i \neq \mu_j\text{''}. \tag{3.4}$$

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.

3.3.3 Les étapes de l'$ANOVA$ 1

Afin de réaliser le test définie dans (3.4), 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 (3.4), 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 :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{p}\sum_{j=1}^{p}\sigma_i^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2, \tag{3.5}$$

ou encore :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X})^2 = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2 + \frac{1}{p}\sum_{j=1}^{p}(\overline{X}_j - \overline{X})^2. \tag{3.6}$$

On multipliant (3.6), par n on obtient :

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X})^2}_{SC_{Tot}} = \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_j)^2}_{SC_{Res}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{X}_j-\overline{X})^2}_{SC_{Fac}}, \tag{3.7}$$

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

	Somme des carrés	Degrés de libertés	Carré moyen	ratio	Ficher
source de variation	SC	ddl	CM	F_{obs}	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}	$n-1$			

3.3.4 Exemple d'application

Reprenant l'exemple présenter dans la section **??**. 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.