Introduction à la statistique inférentielle

Didier Concordet
Unité de Biostatistique
Ecole Nationale Vétérinaire de Toulouse

Sommaire

Chapitre 1

Statistiques descriptives

L'objet de ce chapitre est de présenter brièvement la première étape de l'analyse des données : la description. L'objectif poursuivi dans une telle analyse est de 3 ordres :
tout d'abord, obtenir un contrôle des données et éliminer les données aberrantes ensuite, résumer les données (opération de réduction) sous forme graphique ou numérique, enfin, étudier les particularités de ces données ce qui permettra éventuellement de choisir des méthodes plus complexes. Les méthodes descriptives se classent en deux catégories qui souvent sont complémentaires : la description numérique et la description graphique.

1.1 Description numérique

Avant de donner des définitions formelles de tous les indices, nous les calculerons sur la série de données suivante (GMQ de porcs exprimés en g):

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
737	630	573	615	718	620	820	763	786	529

Nous noterons n la taille de la série de données, ici $n = 10$

1.1.1 Paramètres de position

Les paramètres de position, aussi appelés valeurs centrales, servent à caractériser l'ordre de grandeur des données.

• **moyenne arithmétique :**

Elle est plus souvent appelée moyenne, et est en général notée $\bar{x}$, elle est calculée en utilisant la formule:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dans notre exemple,$\bar{x} = 679$.

• **moyenne géométrique**

La moyenne géométrique ($\bar{x}_g$) est toujours inférieure (ou égale) à la moyenne arithmétique. Elle est donnée par:

$$\bar{x}_g = \left[\prod_{i=1}^{n} x_i\right]^{1/n}$$

Dans notre exemple, $\bar{x}_g = 672.6$

On peut remarquer que

$$\log(\bar{x}_g) = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

en d'autres termes, le log de la moyenne géométrique est la moyenne arithmétique du log des données. Elle est très souvent utilisée pour les données distribuées suivant une loi log normale (par exemple les comptages cellulaires du lait).

• **moyenne harmonique**

La moyenne harmonique ($\bar{x}_h$) est toujours inférieure (ou égale) à la moyenne géométrique, elle est en général utilisée pour calculer des moyennes sur des intervalles de temps qui séparent des événements. Elle est donnée par:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Dans notre exemple,$\bar{x}_h = 666.05$
On peut remarquer que

$$\frac{1}{\bar{x}_h} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}.$$

• médiane

La médiane $\tilde{x}$ est la valeur telle que la moitié des observations lui sont supérieures (ou égales) et la moitié inférieures (ou égales). Il est clair que la médiane existe pour toutes les distributions (ce qui n'est pas le cas de la moyenne) de plus, elle est peu sensible aux valeurs extrêmes.
Lorsque le nombre d'observations est pair, la médiane n'est pas définie de façon unique. La valeur usuellement retenue est la moyenne des observations de rang $\frac{n}{2}$ et de rang $\frac{n}{2}+1$ Dans notre exemple $\tilde{x} = 674$.

• les quartiles

Les quartiles sont au nombre de trois. La médiane est le deuxième.
Le premier quartile q_1 est la valeur telle que 75% des observations lui sont supérieures (ou égales) et 25% inférieures (ou égales).
Lorsqu'il n'est pas défini de façon unique, on utilise généralement la moyenne des observations qui l'encadrent pour le calculer. Dans notre exemple, $q_1 = 615$.
Le troisième quartile q_3 est la valeur telle que 25% des observations lui sont supérieures (ou égales) et 75% inférieures (ou égales).
Lorsqu'il n'est pas défini de façon unique, on utilise la moyenne des observations qui l'encadrent pour le calculer. Dans notre exemple, $q_3 = 763$.

• le mode

est la (ou les) valeur(s) pour laquelle les effectifs sont maximums, il est en général assez difficile de l'évaluer (quand il existe) sur des échantillons de petite taille.

• les extrêmes

Ce sont les minimum et maximum de l'échantillon qui ici valent respectivement 529 et 820.

> *La moyenne n'est pas toujours le meilleur indice pour d'écrire la position des données, tout dépend de la forme de la distribution.*

En effet, pour des distributions non symétriques ou multimodales, il est souvent préférables de donner les percentiles qui sont plus facile à interpréter.

1.1.2 Paramètres de dispersion

Ces paramètres (comme leur nom l'indique) mesurent la dispersion des données.

• **la variance**

Elle est définie comme la moyenne des carrés des écarts à la moyenne, soit:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Il est aussi possible d'en donner la définition suivante:

$$\hat{\sigma}_n^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$$

On voit donc, que la variance est proportionnelle à la somme des carrés de toutes les différences possibles entre les observations.

Cette définition de la variance n'est pas utilisée en pratique pour une raison que nous verrons au chapitre suivant. En fait, on utilise la définition suivante

$$\hat{\sigma}_{n-1}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

La variance s'exprime dans l'unité au carré des données ; dans notre exemple, la variance vaut :$\hat{\sigma}_{n-1}^2 = 9664.989g^2$

• **l'écart type**

est la racine carrée de la variance. il vaut ici:$\hat{\sigma}_{n-1} = 93.26g$ Utilisez le à bon escient (cf TD)

• **l'étendue ou amplitude**

est définie comme la différence entre la maximum et le minimum, soit ici :$820 - 529 = 291g$

• **la distance inter-quartile**

est définie comme la différence entre q_3 et q_1, soit:$763 - 615 = 148$

• **le coefficient de variation**

est définie comme le rapport entre l'écart type et la moyenne.

$$CV = \sqrt{\frac{S^2}{\bar{x}}}$$

1.1.3 Paramètres de forme

Les logiciels de statistiques fournissent généralement les paramètres Skewness et Kurtosis construits à partir des moments centrés d'ordre 2,3 et 4 qui mesurent respectivement la symétrie et l'aplatissement de la distribution dont l'échantillon est issu.

Pour une loi normale centrée réduite, ces coefficients sont nuls.

Les moments centrés d'ordre 3 et 4 sont définis par:

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

A partir de ces définitions, les paramètres Skewness et Kurtosis sont respectivement définis par:

$$\gamma_1 = \frac{m_3}{s^3}$$

$$\gamma_2 = \frac{m_4}{s^4} - 3$$

Dans notre exemple,$\gamma_1 = -0.037$ et $\gamma_2 = -1.339$

Le paramètre γ_1 est nul pour une distribution symétrique. Le graphique suivant montre un exemple de distribution avec un γ_1 positif et négatif. Le paramètre γ_2 est nul pour une loi normale. Le graphique suivant montre un exemple de distribution avec un γ_1 positif et négatif.

1.2 Description graphique

Les graphiques présentés dans ce paragraphe décrivent d'une part la densité de la distribution et d'autre part la fonction de répartition de la distribution.

1.2.1 Description de la densité

Histogramme (cf fig 1.1)

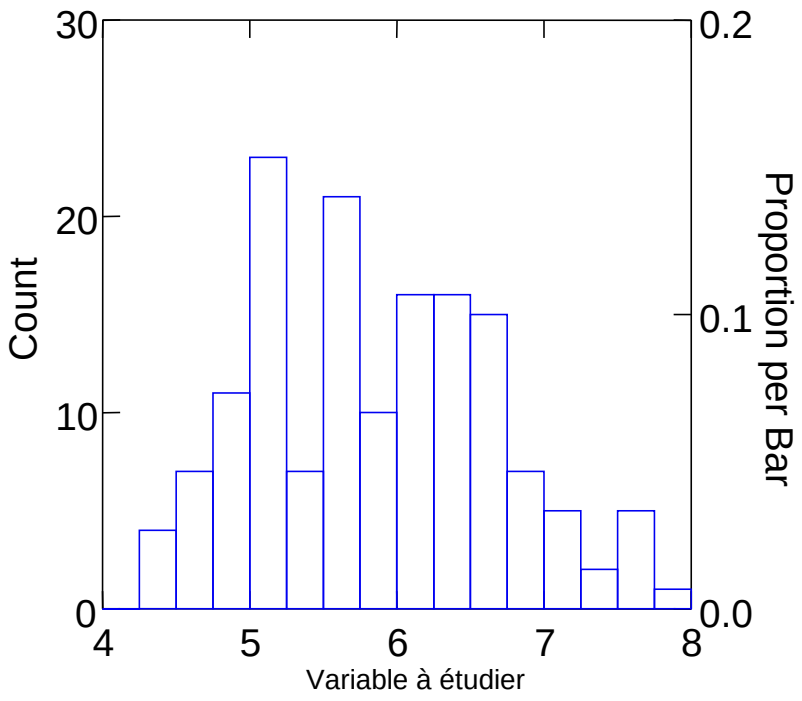

Figure 1.1: Histogramme d'une variable quantitative. La variable quantitative est découpée en classes représentées en abscisse. Le pourcentage (et/ou le nombre) de données de l'échantillon appartenant à chaque classe est représenté en ordonnée. L'inconvénient majeur de cette représentation graphique est l'arbitraire dans le choix des classes.

Stem and leaf

4		3
4		4445
4		666677
4		88888999999
5	H	0000000000111111111
5		22223
5		4444445555555
5		66666677777777
5	M	8888888999
6		000000111111
6		2222333333333
6	H	444444455555
6		6677777777
6		8889999
7	01	
7		2223
7		4
7		67777
7		9

C'est un de mes graphiques préférés. Il s'agit d'un histogramme fait avec des chiffres. Les données sont classées par ordre croissant. Le minimum de l'échantillon est 4.3 (première ligne du stem). La deuxième ligne nous indique que l'échantillon contient 3 valeurs qui après arrondi valent 4.4 et une valeur égale (après arrondi) à 4.5. Le maximum vaut 7.9. Les H nous indiquent les classes qui contiennent respectivement les premier et troisième quartiles tandis que le M nous donne la classe qui contient la médiane. On en déduit que 25% des données sont inférieures à 5.0 ou 5.1, 50 % sont inférieures à 5.8 ou 5.9 et 25% sont supérieures à 6.4 ou 6.5.

1.2.2 Description de la fonction de répartition

Qplot (Quantile plot) ou encore fonction de répartition empirique (cf fig 1.2)

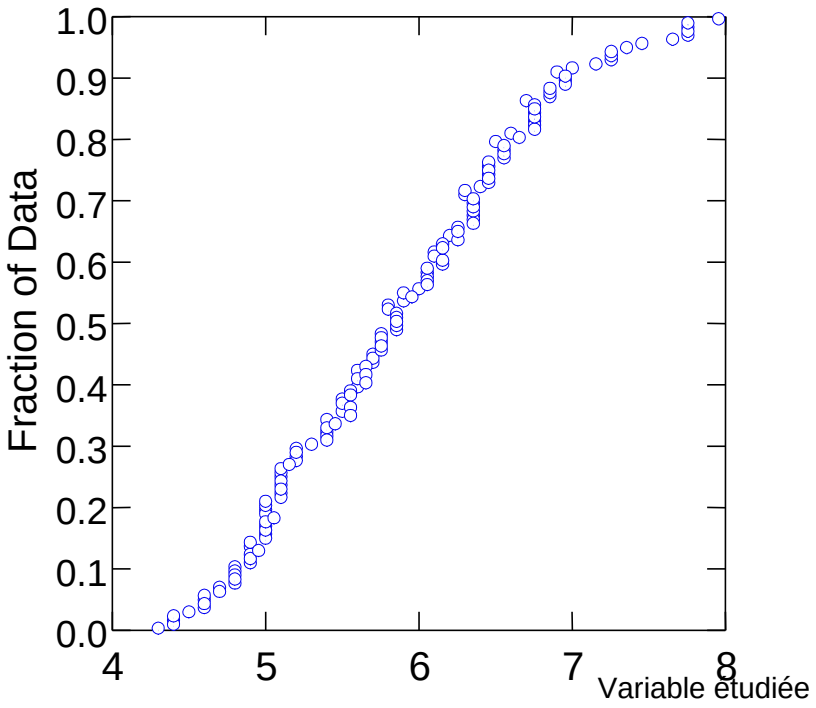

Figure 1.2: Ce graphique est homogène au graphique des fréquences cumulées pour une variable qualitative. La variable étudiée est représentée sur l'axe des abscisses. L'axe des ordonnées donne le pourcentage de données de l'échantillon inférieures ou égales à l'abscisse.

Pplot (Probability plot) aussi appelé dans le cas de la loi normale droite de Henry. (cf fig 1.3). Toutes les fonctions de répartition se ressemble, ce sont des courbes croissantes en général sigmoïdale. En bref, elles ne permettent pas facilement d'identifier une loi. L'idée des Pplot est de déformer l'axe des ordonnées de telle façon que si la loi empirique est proche de la loi que l'on cherche à identifier alors les points sont à peu prés alignés. Le Pplot le plus courant est la droite de Henry qui permet de reconnaître la loi normale. Formellement voilà comment cela marche. Notons $\hat{F}(x)$ la fonction de répartition empirique construite avec notre échantillon. On pense que cette fonction de répartition est proche de la fonction de répartition de la loi

normale $N(m, \sigma^2)$ (cf paragraphe refgauss0 pour plus de détails). On pense donc que $\hat{F}(x) \simeq \Phi\left(\frac{x-m}{\sigma}\right)$ où Φ est la fonction de répartition de la la loi normale $N(0,1)$. Si $\hat{F}(x) \simeq \Phi\left(\frac{x-m}{\sigma}\right)$ alors $\Phi^{-1}\left(\hat{F}(x)\right) \simeq \frac{x-m}{\sigma}$. En d'autres termes, si $\hat{F}(x)$ est proche de la fonction de répartition de la loi normale alors le graphique de $\Phi^{-1}\left(\hat{F}(x)\right)$ contre x devrait nous donner une droite d'équation $\frac{x-m}{\sigma}$. Les points devraient donc se situer autour de cette droite si la distribution est gaussienne (aux effets de bords prés).

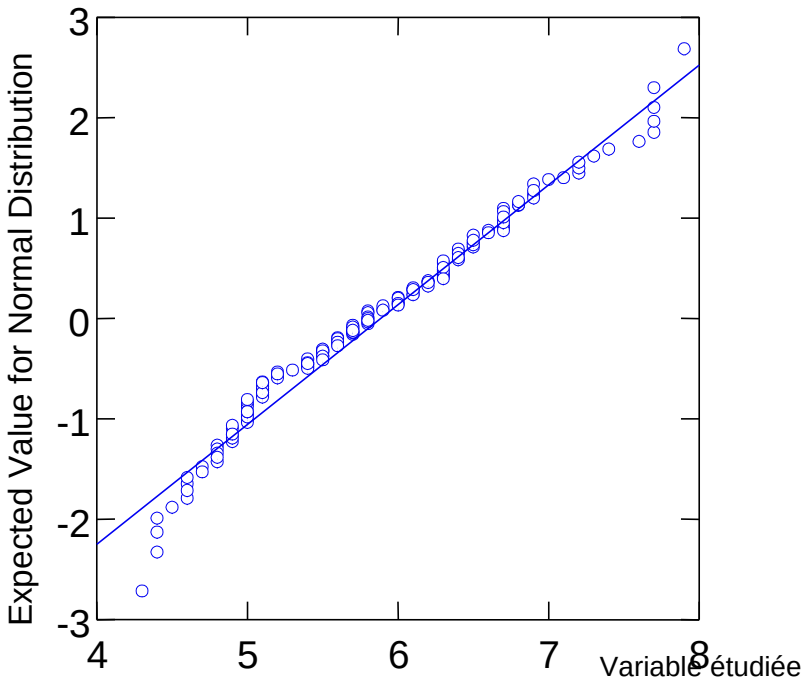

Figure 1.3: Ce graphique nous montre clairement que cette distribution ne peut pas être considérée comme gaussienne, il y a trop de courbure.

Chapitre 2

Le zoo des lois de probabilité

Une des notions fondamentales des statistiques est celle de **variable aléatoire**. On considère un ensemble d'individus qui sera appelé Ω. Un individu de cet ensemble sera noté ω. On note $X(\omega)$ une caractéristique de l'individu ω. Par exemple, Ω est l'ensemble des bactéries que l'on trouve dans du lait de mammites, ω est une bactérie particulière et $X(\omega)$ est type de la bactérie ω. La quantité $X(.)$ est appelée variable aléatoire (en général on note v.a.). Les valeurs possibles que peut prendre $X(\omega)$ quand $\omega \in \Omega$ détermine la nature de la variable aléatoire. Ainsi, si $X(\omega)$ [1]prend ses valeurs dans $\mathbb{R}$, on parlera de variable aléatoire continue, si $X(.)$ prend ses valeurs dans un ensemble fini ou dénombrable, $X(.)$ sera alors appelée v.a. discrète.
En résumé,

$$X : \Omega \longrightarrow \mathbf{E}$$
$$\omega \longrightarrow X(\omega)$$

Quelques exemples de variables aléatoires :
1) le nombre d'étudiants présents au cours de stat ;
2) le nombre de vaches qui ont une mammite dans un élevage ;
3) le pourcentage de réussite aux examens ;
4) le temps pendant lequel un animal est porteur d'une maladie ;

[1]Pour simplifier les notations, on note généralement X au lieu de $X(\omega)$. Par la suite, cet abus de notation sera abondamment utilisé

5) la température d'un chien;
6) les concentrations en fer et en cuivre dans le sang d'un animal sain.

Les trois premières v.a. sont discrètes, et ne peuvent prendre que des valeurs qu'il est possible d'énumérer d'avance. En revanche, les v.a. 4), 5), 6) sont continues. La variable aléatoire 6) est une va à deux dimensions. Nous adopterons dorénavant la convention suivante : les lettres majuscules désigneront les variables aléatoires, les lettres minuscules désigneront les valeurs que peuvent prendre les variables aléatoires.
L'étude des lois de probabilité usuelles est en fait l'étude de la distribution des valeurs que peut prendre une variable aléatoire.

2.1 Lois de probabilité discrètes

Pour complètement définir une loi de probabilité d'une va discrète X, il suffit de définir la probabilité d'occurrence de chaque valeur k que peut prendre cette va. En d'autres termes, la donnée des quantités $P(X = k)$ et ceci pour toutes les valeurs k possibles déterminent une loi de proba particulière. De façon équivalente, pour complètement caractériser une loi de proba, il suffit de définir sa **fonction de répartition** , définie par :

$$F(n) = \sum_{k \leq n} P(X \leq k).$$

Cette fonction s'interprète comme la probabilité que la va X soit au plus égale à n. C'est évidemment une fonction positive et croissante (on ajoute des probabilités qui sont des quantités positives ou nulles). Pour illustrer ce qu'elle représente, prenons un petit exemple. Supposons que X est le nombre de clients d'un vétérinaire le mardi matin. La va X est discrète et ne peut prendre que les valeurs $k = 0, 1, \ldots, 10$. Supposons de plus que la distribution de X est donnée par

k	0	1	2	3	4	5	6	7	8	9	10
$P(X = k)$	0.01	0.03	0.09	0.14	0.17	0.17	0.15	0.11	0.07	0.04	0.02

alors la fonction de répartition est donnée par

n	0	1	2	3	4	5	6	7	8	9	10
$F(n)$	0.01	0.04	0.13	0.27	0.45	0.62	0.77	0.88	0.94	0.98	1.00

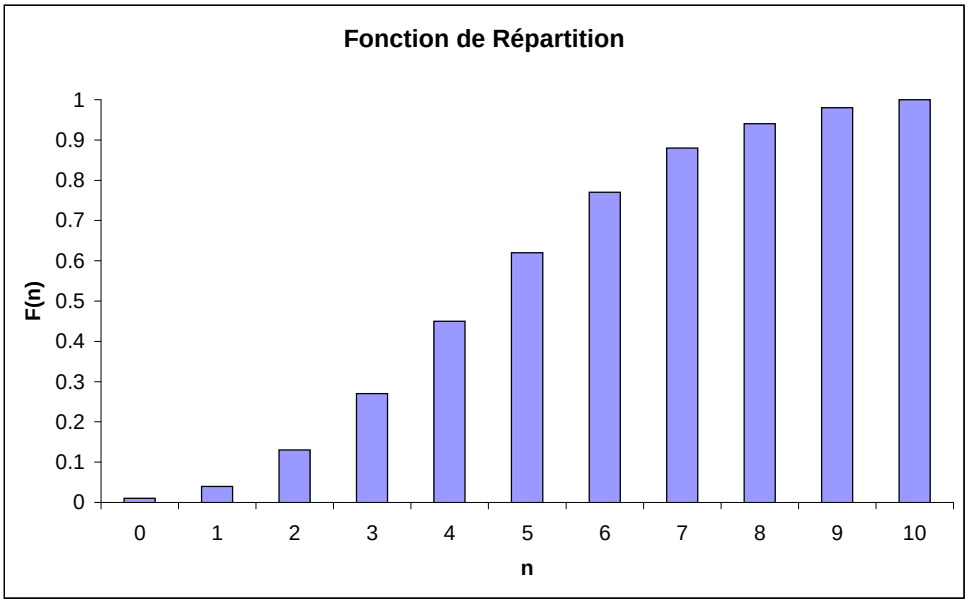

Figure 2.1: Fonction de répartition du nombre de clients d'un vétérinaire le mardi matin

Il est bien évident que si le nombre de valeurs que peut prendre la variable aléatoire est très élevé, il peut être très fastidieux (voire impossible) de donner toutes ces probabilités. Or, comme nous allons le voir, les lois de proba usuelles sont en fait définies par un petit nombre de paramètres : les moments de la loi de proba. Pour définir les moments, nous avons besoin d'un opérateur appelé espérance mathématique qui est noté $\mathbb{E}$. Cet

opérateur placé devant une variable aléatoire, fournit la moyenne de cette variable, ainsi la quantité $\mathbb{E}(X)$ est définie par

$$\mathbb{E}(X) = \sum_k kP(X = k)$$

Dans notre exemple, le nombre de clients moyen du vétérinaire le mardi matin est donné par

$$\begin{aligned}\mathbb{E}(X) = \quad & 0 \times 0.01 + 1 \times 0.03 + 2 \times 0.09 + 3 \times 0.14 + 4 \times 0.17 + 5 \times 0.17 + \\ & 6 \times 0.15 + 7 \times 0.11 + 8 \times 0.07 + 9 \times 0.04 + 10 \times 0.02 = 4.95\end{aligned}$$

Plus généralement, on peut définir l'espérance mathématique de n'importe quelle fonction Φ (ayant de bonnes propriétés) de la va X ainsi,

$$\mathbb{E}(\Phi(X)) = \sum_k \Phi(k)P(X = k)$$

On peut maintenant définir le moment d'ordre p par :

$$\mathbb{E}(X^p) = \sum_k k^p P(X = k).$$

Le moment centré d'ordre p est défini par

$$m_p = \mathbb{E}((X - \mathbb{E}(X))^p) = \sum_k (k - \mathbb{E}(X))^p P(X = k).$$

Vous connaissez déjà le moment centré d'ordre 2 qui est aussi appelé **variance**. Nous reviendrons un peu plus loin sur l'interprétation pratique de cet indice ainsi que sur celle des moments centrés d'ordre 3 et 4. Dans l'exemple précédent, la variance du nombre de clients du mardi matin est donnée par

$$\begin{aligned}\mathbb{E}((X - \mathbb{E}(X))^2) = \quad & (0 - 4.95)^2 \times 0.01 + (1 - 4.95)^2 \times 0.03 + (2 - 4.95)^2 \times 0.09 + \\ & (3 - 4.95)^2 \times 0.14 + (4 - 4.95)^2 \times 0.17 + (5 - 4.95)^2 \times 0.17 + \\ & (6 - 4.95)^2 \times 0.15 + (7 - 4.95)^2 \times 0.11 + (8 - 4.95)^2 \times 0.07 + \\ & (9 - 4.95)^2 \times 0.04 + (10 - 4.95)^2 \times 0.02 = 4.6275\end{aligned}$$

Nous pouvons maintenant passer à l'inventaire des lois de probabilités les plus courantes.

2.1.1 Loi de Bernoulli

C'est la loi de probabilité la plus simple: l'individu ω peut se trouver dans deux états (en général notés 0 et 1).
Exemple : Ω est l'ensemble des bactéries dans du lait de mammite, ω est une bactérie particulière, $X(\omega) = 0$ si la bactérie ω est gram (-) et, $X(\omega) = 1$ si la bactérie ω est gram (+). La loi de probabilité de X est entièrement déterminée par la seule donnée du nombre $P(X(\omega) = 0) = p$ qui permet de déduire que $P(X(w) = 1) = 1 - p$. On dit alors que la v.a. X suit une loi de BERNOULLI de paramètre p. On peut interpréter p dans notre exemple comme la probabilité qu'une bactérie donnée soit gram (-). La loi de BERNOULLI nous sera essentiellement utile pour définir d'autres lois de probabilité.

2.1.2 Loi binomiale

Une v.a. qui suit une loi binomiale ne peut prendre qu'un nombre fini de valeurs que nous noterons N. Pour illustrer l'utilisation de la loi binomiale, prenons l' exemple suivant : supposons que la prévalence de la dysplasie de la hanche chez le CN est de p (la proportion de CN non porteur de cette anomalie est donc de $1 - p$). A l'école vétérinaire, il passe par an N CN, on note X le nombre de CN porteurs de la dysplasie de la hanche parmi les N traités à l'école. On suppose que l'école a une chance égale d'être choisie comme centre de traitement par les propriétaires de CN à dysplasie de la hanche. Alors,

$$P(X = k) = C_N^k p^k (1-p)^{N-k} \text{ et ceci pour } k = 0, 1...N.$$

$C_N^k = \dfrac{N!}{k!(N-k)!}$ est le nombre de "paquets de k que l'on peut faire parmi N".
Une propriété élémentaire de C_N^k est

$$C_N^k = C_N^{N-k}.$$

Le nombre moyen de CN porteur de la dysplasie que l'on peut trouver au cours d'une année à l'école véto est donné par $\mathbb{E}(X) = Np$. En d'autres termes si la prévalence de la dysplasie de la hanche est de $p = 0.1$, et s'il passe dans les cliniques de l'école $N = 500$ CN par an, on trouvera en moyenne $Np = 500\ 0.1 = 50$ CN porteurs de cette anomalie. Il est bien évident que le nombre de CN porteurs trouvés sur les 500 examinés par an ne sera pas toujours égal à 50. Il y a donc des variations de CN porteurs qui seront observés à l'école. Un indice mesure ces variations c'est la variance. La variance d'une loi binomiale est donnée par

$$Var(X) = Np(1-p).$$

Très souvent la quantité $1-p$ est notée q ; ceci explique le fait que $Var(X) = Npq$.Quand X suit une loi binomiale de paramètre N et p on note

$$X \sim \mathcal{B}(N, p).$$

Le graphique 2.2 montre les formes caractéristiques d'une loi binomiale en fonction des valeurs du paramètre p.

Remarque Il existe une autre façon de construire la loi binomiale. Voyons sur l'exemple des bactéries comment procéder.
On considère N bactéries. Chaque bactérie a une probabilité p d'être gram (-), à chaque bactérie on fait correspondre une v.a. de Bernoulli de paramètre p qui prend la valeur 0 si elle est gram (-) et 1 si elle est gram (+). On appelle X_i la variable aléatoire attachée à la $i^{ième}$ bactérie. En supposant que les bactéries sont indépendantes on a:

$$X = \sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p).$$

X représente ici le nombre total de bactéries gram (+) parmi les N considérées.

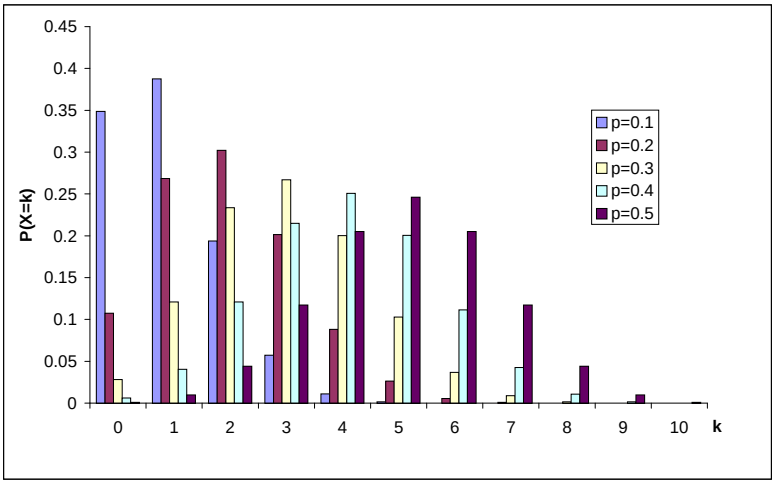

Figure 2.2: Forme de la loi binomiale pour différentes valeurs du paramètre p.

2.1.3 Loi hypergéométrique

Pour bien faire comprendre la loi hypergéométrique prenons un petit exemple. Supposons que vous ayez à évaluer la prévalence des mammites de la vache en Midi-Pyrénées. On sait que dans cette région il y a N vaches. Parmi ces vaches N_1 sont atteintes et N_2 sont saines (on a évidemment $N_1 + N_2 = N$.) Vous ne pouvez pas contrôler toutes les vaches de Midi-Pyrénées, vous êtes donc obligé de prendre un échantillon de taille $n < N$. On appelle X le nombre de vaches à mammite que vous avez trouvé dans votre échantillon. X [2] est une quantité aléatoire, en effet, si vous faites plusieurs fois des échantillons de taille n, vous ne retrouvez pas à chaque fois le même nombre de vaches atteintes. On s'interesse aux probabilités suivantes $P(X = k)$ k varie entre 0 et $N_1 \wedge n$. Il y a C_N^n façons de tirer un échantillon de taille n parmi les N vaches de M.P.

[2] X est ici mis pour $X(\omega)$. ω représente un tirage de n vaches

$C_{N_1}^k$ est le nombre de façons de tirer k vaches à mammites parmi les N_1 présentes en M.P. et enfin $C_{N_2}^{n-k}$ est le nombre de façons de tirer $n-k$ vaches saines parmi N_2 présentes en M.P.
On en déduit que

$$P(X=k) = \frac{\sharp\text{cas probables}}{\sharp\text{cas possibles}} = \frac{C_{N_1}^k C_{N_2}^{n-k}}{C_N^n} \text{si } k \leq N_1 \text{ et } n-k \leq N_2$$
$$= 0 \text{ sinon}$$

La variable aléatoire X suit une loi hypergéométrique. Quand X suit une loi hypergéométrique de paramètres N, n, N_1 on note,

$$X \sim \mathcal{H}(N, n, \frac{N_1}{N}).$$

Sa moyenne est donnée par

$$\mathbb{E}(X) = n\frac{N_1}{N}$$

et sa variance par

$$Var(X) = n\frac{N_1}{N}\frac{N_2}{N}\frac{N-n}{N-1}$$

On peut noter que lorsque $N \longrightarrow \infty$, si $\frac{N_1}{N} \longrightarrow p$ (p est le pourcentage vache atteintes présentes parmi les N à contrôler) alors

$$\mathcal{H}(N, n, \frac{N_1}{N}) \longrightarrow \mathcal{B}(n, p).$$

En d'autres termes, si le nombre total de vaches en MP est très élevé, on peut utiliser la loi binomiale (plus simple) à la place de la loi hypergéométrique.

2.1.4 Loi de Poisson ou loi des événements rares

Une va qui suit une loi de poisson peut prendre une infinité de valeurs.
On dit que la va X suit une loi de poisson de paramètre λ, et on note $X \sim \mathcal{P}(\lambda)$, si

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k = 0, 1, ...$$

La moyenne d'une va qui suit une loi de poisson est égale à $\mathbb{E}(X) = \lambda$, sa variance est $Var(X) = \lambda$.
Le graphique ci-dessous montre les différentes formes de distribution d'une loi de poisson en fonction de la valeur du paramètre

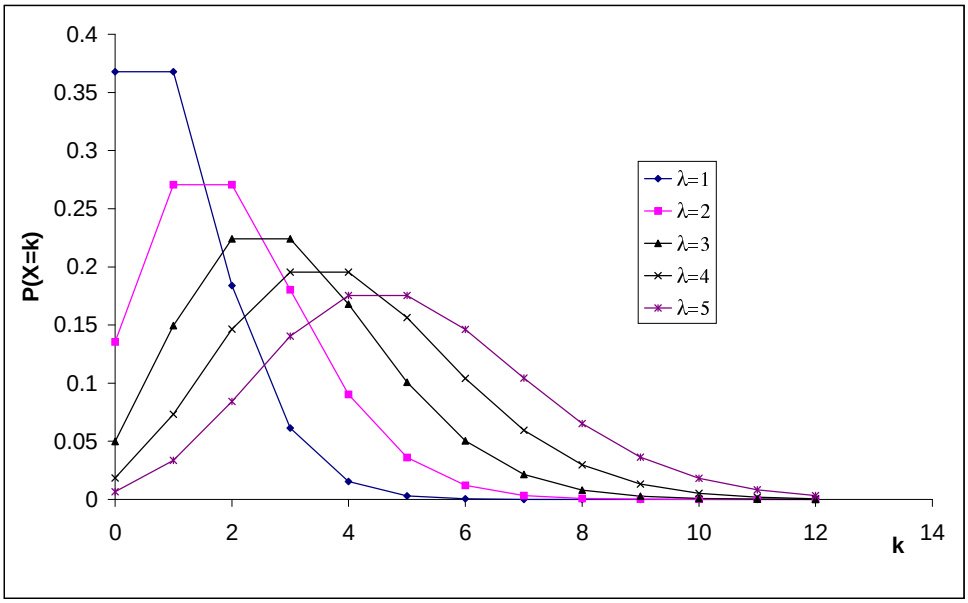

Figure 2.3: Loi de poisson pour différentes valeurs de λ

La loi de poisson est souvent utilisée pour approximer certaines lois discrètes. On l'appelle aussi loi des événements rares. En effet, si X est le nombre de fois où apparaît un événement de probabilité très petite (p), alors la loi de X peut être approximée par une loi de poisson. Prenons un exemple pour illustrer ce phénomène. Soit une maladie dont la prévalence est très petite ($p = 0.01$) On tire un échantillon de taille 100 et on s'interesse à la distribution du nombre

de sujets atteints trouvés dans l'échantillon (noté X). En d'autres termes, on veut calculer

$$(Bi) \qquad P(X = k) = C_{100}^{k}(0.01)^{k}(1 - 0.01)^{100-k}.$$

Il est bien évident que le calcul d'une telle probabilité n'est pas si facile à cause du terme C_{100}^{k} (pour vous en convaincre essayez de calculer avec votre calculette C_{100}^{50}). L'idée est alors d'**approximer** la quantité (Bi) par une quantité plus facilement calculable:

$$P(X = k) = C_{100}^{k}(0.01)^{k}(1 - 0.01)^{100-k} \simeq e^{-100\times 0.01}\frac{(100 \times 0.01)^{k}}{k!}$$

Plus généralement, si $X \sim B(N, p)$, si N est grand, si p est petit et si Np est raisonnable on peut approximer la loi $B(N, P)$ par une loi de poisson de paramètre $\lambda = Np$. Ces conditions sont évidemment très vagues. Les conditions usuelles sous lesquelles on considère que la qualité de l'approximation est "raisonnable" sont les suivantes : $N > 30$, et $Np > 5$. D'autres valeurs de ces paramètres peuvent être tout à fait acceptables pour peu que vous ne soyez pas trop regardant sur la qualité d'approximation de certaines probabilités.
La loi de poisson est souvent utilisée pour modéliser des quantités dont la variance est à peu prés égale à la moyenne. Lorsque la variance est supérieure à la moyenne, on utilise dans certains cas la loi Binomiale négative.

2.1.5 Loi binomiale négative

Une va qui suit une loi binomiale négative peut prendre un nombre infini de valeurs. On dit que la va X suit une loi binomiale négative de paramètre N et p si

$$P(X = k) = C_{N+k-1}^{k}\frac{p^{k}}{(1+p)^{n+k}},\ k = 0..$$

Sa moyenne est égale à $\mathbb{E}(X) = Np$ et sa variance $Var(X) = Np(1+p)$. On peut remarquer que ces distributions sont d'autant plus *surdispersées* que p est grand. Le graphique suivant montre comment varie les distributions binomiales négatives quand p varie.

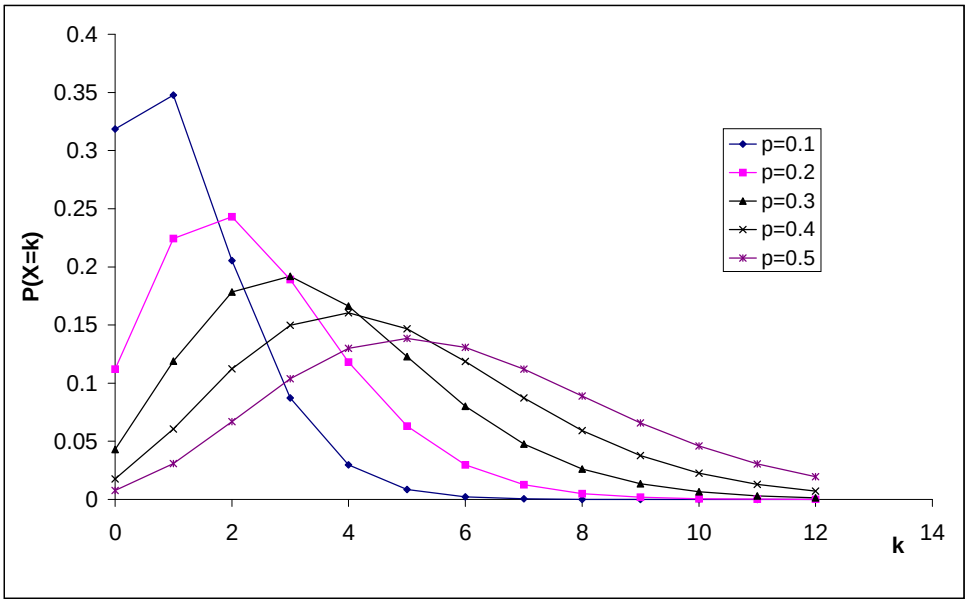

Figure 2.4: Loi binomiale négative pour différentes valeurs de p. Plus p augmente plus la loi est surdispersée

2.1.6 Loi de Pascal

Une va qui suit une loi de pascal peut prendre une infinité de valeurs. On dit que la va X suit une loi de Pascal de paramètre p si

$$P(X = k) = p\ (1-p)^{k-1},\ k = 1, 2, ...$$

Pour illustrer son utilisation, reprenons l'exemple de la dysplasie de la hanche chez le CN. Supposons que l'école a une chance égale d'être choisie comme centre de traitement par les propriétaires de CN à dysplasie de la hanche. Notons p la prévalence de cette anomalie et X le nombre de CN à examiner

avant d'en trouver un atteint, alors si on pose $q = 1 - p$, on a:
$P(X = 1) = p$, $P(X = 2) = pq...$, $P(X = k) = pq^{k-1}$.
Le nombre moyen de CN à examiner avant d'en trouver un atteint est

$$\mathbb{E}(X) = \frac{1}{p},$$

la variance de ce nombre est

$$Var(X) = \frac{q}{p^2}.$$

2.2 Quelques lois de probabilité continues

2.2.1 Quelques définitions préliminaires

Dans l'étude des lois de proba continues, il apparaît une nouvelle quantité : la **densité de probabilité**.
Pour bien comprendre ce dont il s'agit, imaginons que l'on s'interesse à l'étude de la distribution de la taille des Français. Pour étudier cette distribution, on fait des classes de tailles, et on compte le pourcentage d'individus qui appartiennent à cette classe. Une représentation graphique de cette distribution est donnée par l'histogramme qui sera revu au chapitre suivant.Supposons maintenant que le nombre d'individus de la population d'intérêt (ici les Français) est infini. Un histogramme avec un nombre fini de classes nous donne une piètre information sur la distribution de la taille. Pour être plus précis on augmente le nombre de classes et on diminue la taille de chaque classe. On obtient ainsi un histogramme plus précis. Que se passe t-il quand le nombre de classes tend vers l'infini et que la taille de chaque classe tend vers zéro ? On obtient une courbe limite, cette courbe limite est en fait une représentation graphique d'une fonction (notée f) que nous appellerons densité de probabilité.
Il est clair que par construction, cette fonction possède un certain nombre de propriétés:
- elle est positive ou nulle (en effet la valeur de cette fonction en un point x

représente en quelque sorte le pourcentage d'individus qui mesure x)
- la surface totale sous cette courbe est égale à 1 ; la surface sous la courbe représente le pourcentage cumulé de tous les individus (par définition il vaut 1).
La fonction de répartition F est définie à partir de la densité de proba de la façon suivante :

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

La quantité $F(x)$ représente donc le cumul des pourcentages d'individus dont la taille est inférieure à x. Ce constat nous permet de définir la fonction de répartition par

$$F(x) = P(X \leq x).$$

Par définition $F(x)$ est donc toujours un nombre compris entre zéro et un, et la fonction $x \longrightarrow F(x)$ est une fonction croissante (c'est un cumul de pourcentages). De plus on a $F(+\infty) = 1$ (on l'a déjà dit) et $F(-\infty) = 0$.
Soit Δx un accroissement infinitésimal de la taille, alors la quantité

$$\frac{F(x + \Delta x) - F(x)}{\Delta x}$$

représente en quelque sorte le pourcentage d'individus dont la taille est comprise entre x et $x + \Delta x$, et en faisant tendre $\Delta x \longrightarrow 0$ on obtient

$$\lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = f(x).$$

En d'autres termes, la dérivée de la fonction de répartition est la densité de probabilité.Tout comme dans le cas discret, il est possible de définir les moments d'une loi de probabilité. Ce sont en général ces quantités dont nous nous servirons en statistique pour travailler. Le moment d'ordre 1 d'une loi de probabilité est défini quand il existe [3] par

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx$$

[3] Il existe certaines lois de proba dont les moments sont infinis par exemple la loi de Cauchy

On reconnaît ici l'analogue continu de la définition donnée dans le paragraphe précédent. Il suffit en effet de changer le signe $\int$ par le signe $\sum$ pour retrouver la même formule. De même, le moment centré d'ordre p est défini par

$$m_p = \mathbb{E}((X - \mathbb{E}(X))^p) = \int_{\mathbb{R}} (x - \mathbb{E}(X))^p f(x) dx$$

Le moment centré d'ordre 2 est aussi appelé variance, les moments centrés d'ordre 3 et 4 sont respectivement appelés kurtosis et skewness.

2.2.2 Loi normale ou de Laplace Gauss

La loi normale joue un rôle particulièrement important dans la théorie des probabilités et dans les applications pratiques. La particularité fondamentale de la loi normale la distinguant des autres lois est que c'est une loi *limite* vers laquelle tendent les autres lois pour des conditions se rencontrant fréquemment en pratique.On peut montrer que la somme d'un nombre suffisamment grand de va indépendantes (ou faiblement liées) suivant des lois quelconques (ou presque), tend vers une loi normale et ceci avec d'autant plus de précision que le nombre de termes de cette somme est important. La majorité des va que l'on rencontre en pratique, comme par exemple des erreurs de mesures, peuvent souvent être considérées comme des sommes d'un nombre important de termes, erreurs élémentaires, dues chacune à une cause différente indépendante des autres. Quelque soit la loi des erreurs élémentaires, les particularités de ces répartitions n'apparaissent pas dans la somme d'un grand nombre de celles-ci, la somme suivant une loi voisine de la loi normale.
La loi normale est caractérisée par sa **densité** de probabilité. Pour une loi normale de moyenne m et de variance σ^2, elle est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

La courbe représentative de la densité a la forme d'une courbe en cloche symétrique. Le graphique 2.5 montre comment varie la densité d'une loi normale, quand la variance est fixée, en fonction de sa moyenne (ici $m_1 < m_2$.)

Le graphique 2.6 montre comment varie la densité d'une loi normale (à moyenne fixée) quand la variance augmente : Les variances des lois **I, II, III** sont de plus en plus élevées.

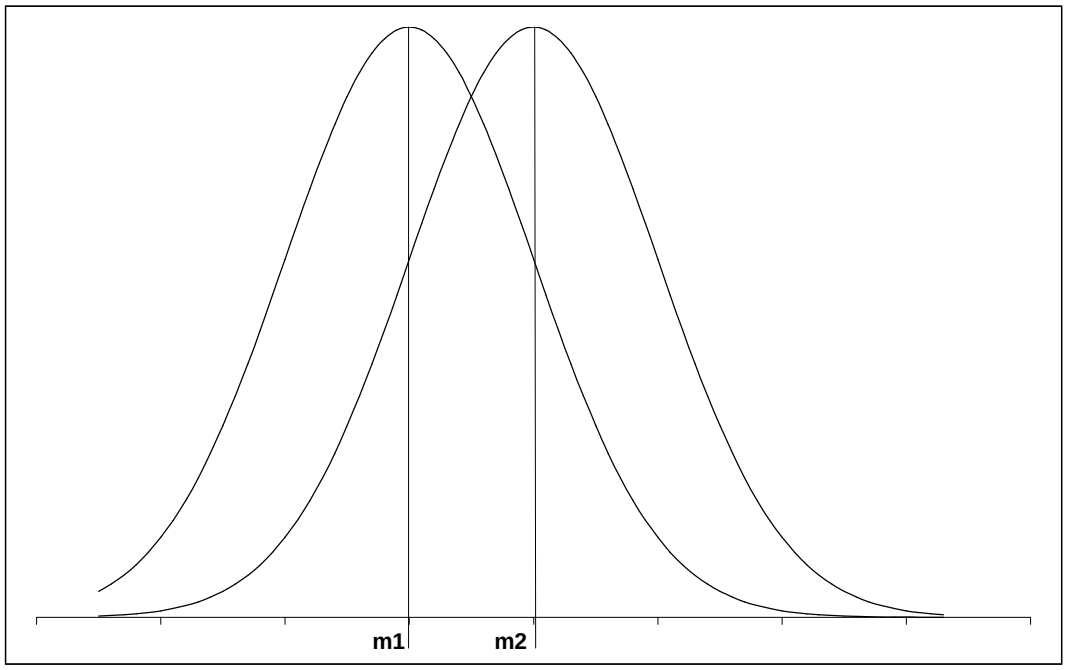

Figure 2.5: Un exemple de deux lois normales. Les deux lois ont la même variance. La moyenne m_1 de la première loi est inférieure à celle m_2 de la seconde

La fonction de répartition de la loi normale est définie à partir de la densité par :

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

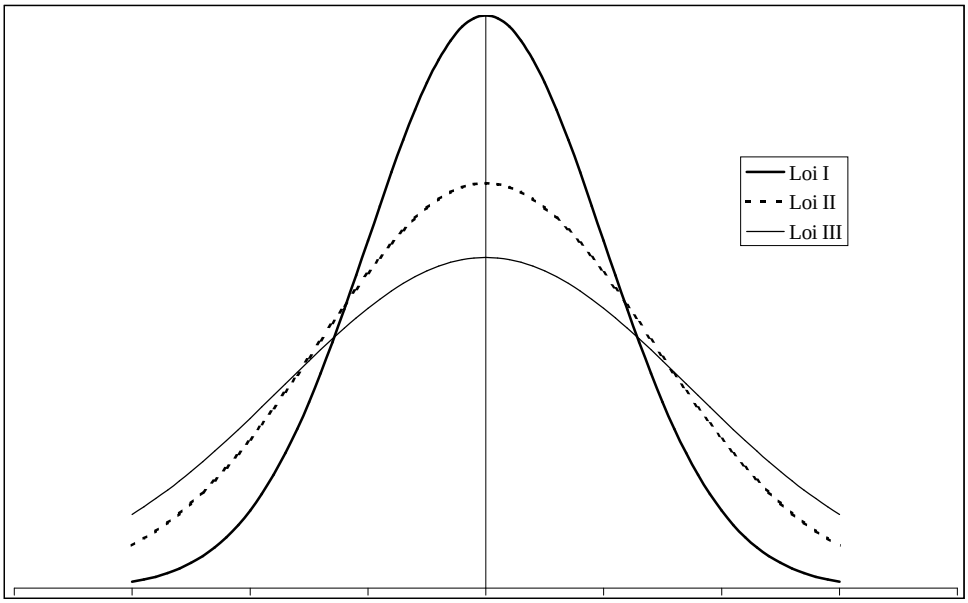

Figure 2.6: Les trois lois ont la même moyenne. Les variances des lois **I, II, III** sont de plus en plus élevées.

Cette dernière propriété traduit géométriquement le fait qu'une probabilité peut s'interpréter comme la surface sous la courbe densité comme l'indique le graphique 2.7:

Il n'existe pas d'expression algébrique donnant l'aire sous la courbe en fonction de x. Il faut donc utiliser des valeurs tabulées. Comme il est impossible d'avoir autant de tables que de valeurs possibles de m et de σ^2, on a recours a l'astuce suivante :
supposons que X est une va suivant une loi normale de moyenne m et de variance σ^2 (on note $X \sim N(m, \sigma^2)$, alors la quantité $\dfrac{X-m}{\sigma}$ suit une loi $N(0,1)$. On en déduit que si F représente la fonction de répartition de la

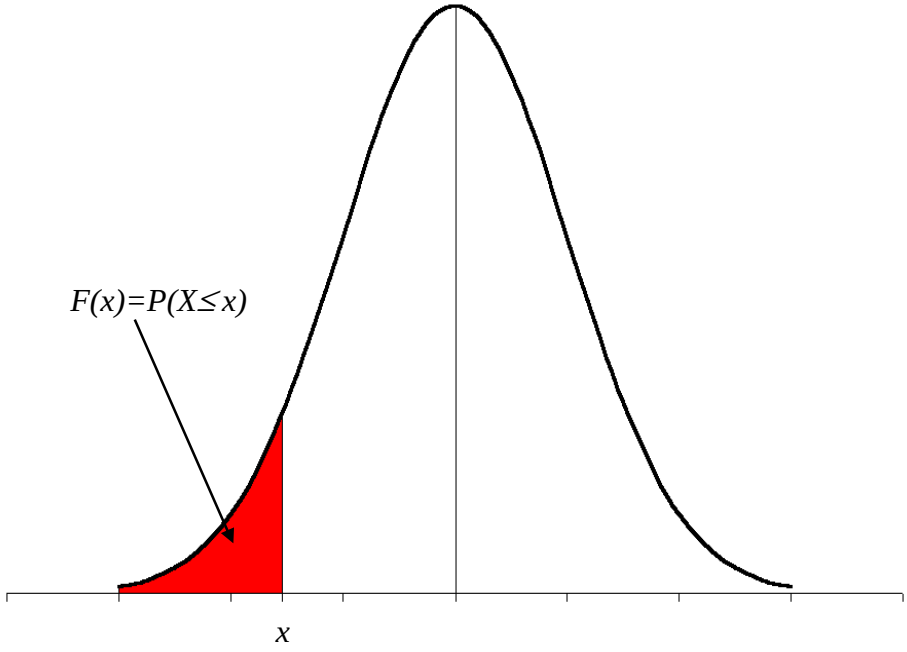

Figure 2.7: Une probabilité s'interprète comme la surface sous la courbe représentant la densité

$N(m, \sigma^2)$ et Φ la fonction de répartition de la $N(0,1)$ alors :

$$P(a < X < b) = F(b) - F(a) = P(a - m < X - m < b - m)$$
$$= P(\tfrac{a-m}{\sigma} < \tfrac{X-m}{\sigma} < \tfrac{b-m}{\sigma}) = \Phi(\tfrac{b-m}{\sigma}) - \Phi(\tfrac{a-m}{\sigma}).$$

remarque : Par définition Φ est une fonction croissante et on a $\Phi(+\infty) = 1$ et $\Phi(-\infty) = 0$.

2.2.3 Loi du χ^2

Cette loi nous sera très utile pour étudier la distribution des variances. Elle est construite à partir de la loi normale de la façon suivante : Soient

$X_1, X_2, \ldots, X_n$ n va indépendantes de même loi N(0,1), et soit

$$K = X_1^2 + X_2^2 + \ldots + X_n^2 = \sum_{i=1}^{n} X_i^2$$

alors, K suit une loi du Khi 2 à n degrés de liberté ($K \sim \chi_n^2$). On peut remarquer qu'une va qui suit une loi du χ^2 est par construction toujours positive ou nulle (c'est une somme de carrés). La densité de probabilité d'une loi du χ^2 est asymétrique (reportez vous aux tables que je vous ai données pour en avoir une idée).

2.2.4 Loi de Student

La loi de Student est construite à partir de la loi normale et de la loi du Khi 2. Nous l'utiliserons intensivement pour faire des tests d'hypothèses.
Soient X une va de loi N(0,1), et K une va qui suit une loi du χ_n^2 (Khi 2 à n degrés de liberté). On suppose de plus que K et X sont indépendantes. Soit

$$T_n = \frac{X}{\sqrt{\frac{K}{n}}},$$

alors T_n suit une loi de student à n degrés de liberté.

2.2.5 Loi de Fisher

Tout comme la loi de student, la loi de Fisher sera très utilisée par la suite. Voyons en rapidement sa construction.
Soient K_1 et K_2 deux variables aléatoires indépendantes de loi respectives χ_n^2 et χ_p^2, alors la quantité

$$F_{n,p} = \frac{K_1/n}{K_2/p}$$

suit une loi de Fisher à n et p degrés de liberté. Il faut faire très attention à l'ordre des degrés de liberté. Le premier degré de liberté (ici n) est le degré de liberté du numérateur, alors que le second (p) est celui du dénominateur.

2.3 Quelques remarques sur l'opérateur $\mathbb{E}$

L'opérateur $\mathbb{E}$ est un opérateur linéaire en d'autres termes, si X et Y sont des va avec de "bonnes propriétés", et si $\alpha,\ \beta$ et γ sont des réels, alors

$$\mathbb{E}(\alpha X + \beta Y + \gamma) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y) + \gamma$$

et ceci que les variables aléatoires X et Y soient indépendantes ou pas. En revanche, l'opérateur variance (noté Var) construit avec l'opérateur $\mathbb{E}$ de la façon suivante

$$Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

n'est pas un opérateur linéaire. On peut constater que par définition, c'est un opérateur positif. La condition nécessaire et suffisante pour que $Var(X)$ soit nulle, est que X soit déterministe c'est à dire non aléatoire. On a de plus des propriétés suivantes: si $\alpha \in \mathbb{R}$, alors

$$Var(\alpha X) = \alpha^2 Var(X)$$

Si X et Y sont deux variables aléatoires **indépendantes**, alors

$$Var(X + Y) = Var(X) + Var(Y)$$

et par conséquent

$$\begin{aligned} Var(\alpha X + \beta Y + \gamma) &= \alpha^2 Var(X) + \beta^2 Var(Y) + Var(\gamma) \\ &= \alpha^2 Var(X) + \beta^2 Var(Y) + 0. \end{aligned}$$

Si les variables aléatoires X et Y ne sont **pas indépendantes**, alors

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

où $Cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$ est la covariance entre X et Y. On voit donc que lorsque les variables aléatoires ne sont pas indépendantes, il apparaît un terme supplémentaire dans le calcul de la variance. On pourrait être tenté de prendre la covariance comme une mesure d'indépendance. Ceci

est en général faux sauf dans le cas où les va X et Y sont normalement distribuées. En résumé :
si X et Y sont indépendantes alors $Cov(X,Y) = 0$,
si $Cov(X,Y) = 0$ et si X et Y sont des va gaussiennes alors X et Y sont indépendantes.

La quantité

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

est un nombre sans dimension appelé coefficient de corrélation linéaire de Pearson. Nous voyons que si X et Y sont gaussiennes et si $\rho(X,Y) = 0$, alors les variables aléatoires X et Y sont indépendantes. Nous l'utiliserons dans le paragraphe suivant consacré à la loi normale à 2 dimensions.

2.4 Lois à deux dimensions

2.4.1 Généralités

Tout comme dans le cas unidimensionnel, les lois à plusieurs dimensions sont caractérisées par leur
- fonction de répartition,
- densité,
- moments.

On appelle **fonction de répartition** du couple de va (X,Y) la probabilité de vérification simultanée des deux inégalités $(X < x)$ et $(Y < y)$:

$$F(x,y) = P((X < x)(Y < y)).$$

En interprétant le couple (X,Y) comme un point aléatoire dans le plan, on voit que la fonction de répartition $F(x,y)$ n'est rien d'autre que la probabilité pour que le point aléatoire (X,Y) appartienne au quadrant de sommet le point (x,y), situé à gauche et en bas de celui-ci (cf fig 2.8).

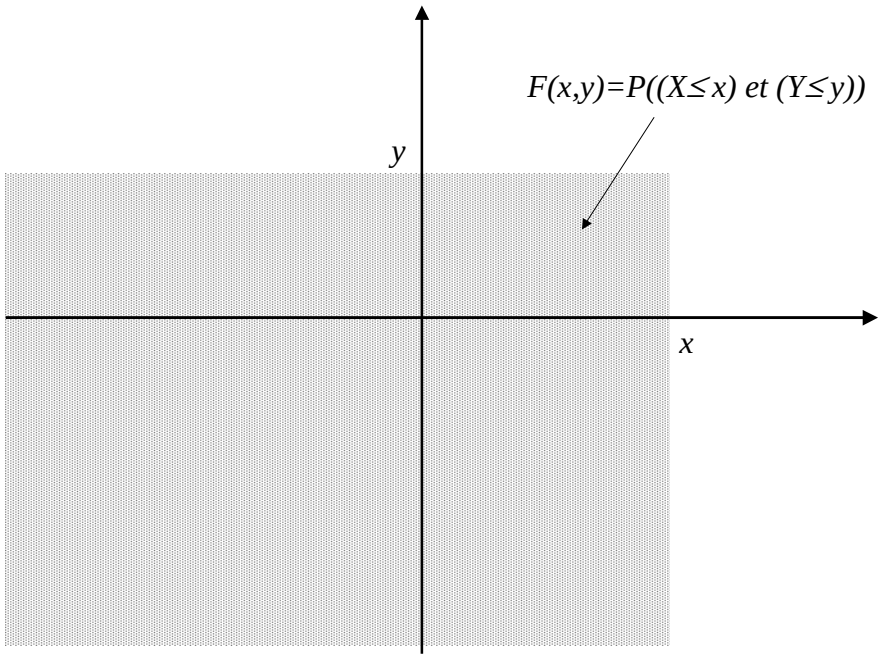

Figure 2.8: La probabilité $F(x, y)$ s'interprète comme la probabilité pour que le point aléatoire (X, Y) appartienne au quadrant de sommet le point (x, y), situé à gauche et en bas de celui-ci

1) Cette interprétation géométrique, permet de voir que si x augmente, ou si y augmente, la fonction $F(x, y)$ augmente aussi.
2) Partout en $-\infty$ la fonction de répartition est égale à zéro :

$$F(x, -\infty) = F(-\infty, y) = F(-\infty, -\infty) = 0.$$

Pour avoir cette propriété, il suffit de déplacer indéfiniment la limite supérieure (ou la limite droite) du quadrant de la figure précédente vers $-\infty$; la probabilité de tomber dans ce quadrant tend alors vers 0.
3) Lorsque un des arguments vaut $+\infty$, la fonction de répartition du couple de va devient alors une fonction de répartition correspondant à l'autre

argument :

$$F(x, +\infty) = F_1(x),\ F(+\infty, y) = F_2(y),$$

où $F_1(x),\ F_2(y)$ sont respectivement les fonctions de répartition des variables aléatoires X et Y. On peut facilement s'en rendre compte en faisant $x \longrightarrow +\infty$, ou $y \longrightarrow +\infty$; à la limite le quadrant devient un demi-plan, la probabilité de tomber dans ce demi-plan est donnée par la fonction de répartition de la variable respective.
4) Si les deux arguments sont égaux à $+\infty$, la fonction de répartition du couple de va est égale à 1 :

$$F(+\infty, +\infty) = 1.$$

En effet, on obtient alors le plan tout entier et le point (X, Y) s'y trouve certainement. De façon analogue, le point (X, Y) peut se trouver dans un domaine quelconque D dans le plan. La probabilité $P((X, Y) \in D)$ ne s'exprime alors pas simplement à partir de la fonction de répartition F sauf dans quelques cas très particuliers sur lesquels nous reviendrons.**Densité de probabilité**
Soit un couple de va continues (X, Y) interprété comme un point aléatoire de ce plan. Considérons dans ce plan un petit rectangle R_Δ dont les cotés sont Δx et Δy avec un sommet au point x, y.

La proba de tomber dans ce rectangle est

$$P((X, Y) \in R_\Delta)$$

$$= F(x + \Delta x, y + \Delta y) - F(x + \Delta x, y) - F(x, y + \Delta y) + F(x, y)$$

En divisant la proba de tomber dans le rectangle R_Δ par l'aire de ce rectangle, on obtient

$$\lim_{\substack{\Delta x \longrightarrow 0 \\ \Delta y \longrightarrow 0}} \frac{P((X, Y) \in R_\Delta)}{\Delta x \Delta y}$$

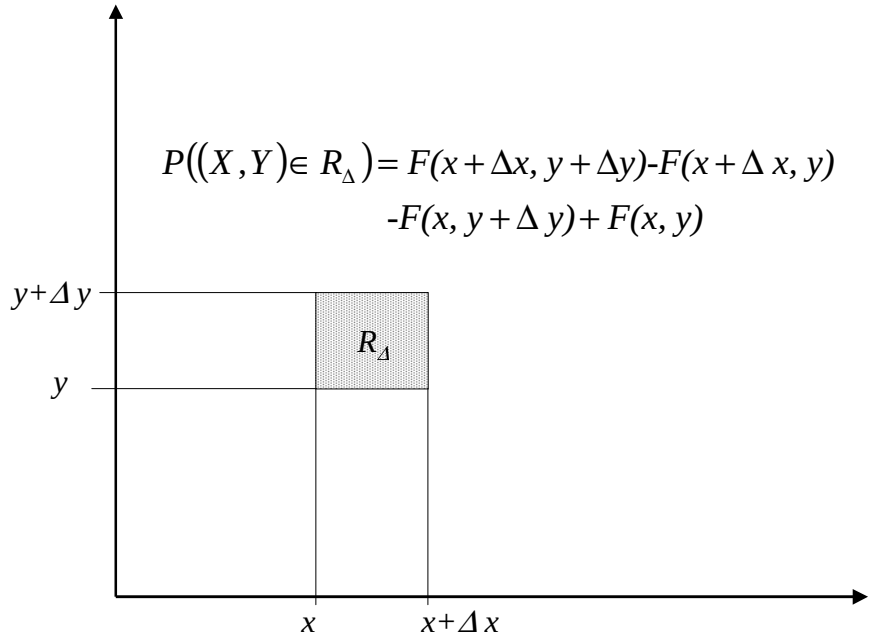

Figure 2.9: La densité s'obtient en faisant des accroissements infinitésimaux de la fonction de répartition

$$= \lim_{\substack{\Delta x \longrightarrow 0 \\ \Delta y \longrightarrow 0}} \frac{F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y) - F(x, y+\Delta y) + F(x, y)}{\Delta x \Delta y}$$

Si on suppose que la fonction F est dérivable, le second membre de la précédente inégalité est alors la dérivée partielle seconde mixte de F. Désignons cette dérivée par $f(x, y)$:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = F''_{xy}(x, y)$$

La fonction f est la densité de proba du couple (X, Y), en d'autres termes,

$$P((X, Y) \in D) = \int_{(x,y)\in D} f(x, y) dx dy$$

De toutes les distributions de couple de va, la plus fréquemment utilisée est la loi normale aussi nous contenterons nous d'étudier la loi normale.

2.4.2 Loi normale à deux dimensions

Dans la suite, nous supposons que le couple (X, Y) *suit une loi normale à deux dimensions.* La loi normale à deux dimensions est définies par 5 paramètres :
sa moyenne (m_x, m_y) et sa matrice de variance-covariance :

$$V = \begin{pmatrix} \sigma_x^2 & Cov(X,Y) \\ Cov(X,Y) & \sigma_y^2 \end{pmatrix}$$

avec $m_x = \mathbb{E}(X),\ m_y = \mathbb{E}(Y)$ et $\sigma_x^2 = Var(X),\ \sigma_y^2 = Var(Y)$.
On voit donc que si les va X et Y sont indépendantes, la matrice de variance-covariance est diagonale.
Si on note ρ le coefficient de correlation entre X et Y, la densité de la loi normale à deux dimensions s'exprime par la formule :

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\quad \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}\right]\right) \end{aligned}$$

Le graphe de cette fonction est représenté à la figure 2.10.

En coupant la surface de répartition par un plan parallèle au plan xOy, on obtient une courbe sur laquelle la densité est constante en chaque point. En reprenant l'équation de la densité, on voit que la densité est constante si et seulement si :

$$\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2} = C^2$$

où C est une constante. Vous reconnaissez l'équation d'une ellipse de centre (m_x, m_y).

Si les va sont indépendantes (donc si $\rho = 0$*), l'équation de l'ellipse devient*

$$\frac{(x-m_x)^2}{\sigma_x^2} + \frac{(y-m_y)^2}{\sigma_y^2} = C^2$$

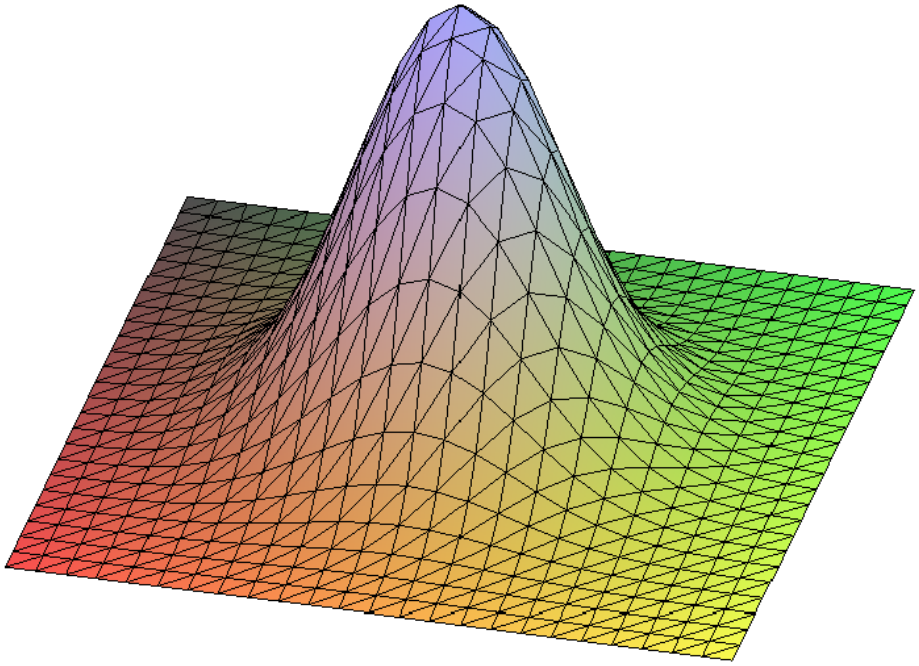

Figure 2.10: Densité de la loi normale à 2 dimensions

Ceci est l'équation d'une ellipse dont les axes sont parallèles aux axes (x, y). Si de plus $\sigma_x^2 = \sigma_y^2$ on obtient alors l'équation d'un cercle de centre (m_x, m_y) et de rayon $C\sigma_x^2$.
Dans le cas général où $\rho \neq 0$, les axes de symétrie de l'ellipse forme un angle θ avec l'axe Ox donné par

$$tg(2\theta) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}.$$

En statistique, on s'interesse très souvent à des domaines dans lesquels on a un certain nombre de chances de trouver un point aléatoire donné. On recherche par exemple des domaines D vérifiant

$$P((X, Y) \in D) = 1 - \alpha$$

où α est un nombre fixé. Quand la loi du couple (X, Y) est gaussienne, le plus simple est de rechercher le domaine D sous la forme d'une ellipse. On recherche donc D tel que

$$\begin{aligned} P((X,Y) \in D) &= 1 - \alpha = \int_{(x,y)\in D} f(x,y)dxdy \\ &= \int_{(x,y)\in D} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\exp(-\frac{1}{2(1-\rho^2)}[\frac{(x-m_x)^2}{\sigma_x^2} - 2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y} + \frac{(y-m_y)^2}{\sigma_y^2}])dxdy \end{aligned}$$

La recherche d'un tel domaine dans ce système de coordonnées est difficile aussi allons nous faire une rotation d'angle

$$\theta = \frac{1}{2}Arctg(\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2})$$

on obtient

$$P((X,Y) \in D) = \int_{D'} \frac{1}{2\pi\tilde{\sigma}_x\tilde{\sigma}_y} \exp(-\frac{1}{2}[\frac{(x-m_x)^2}{\tilde{\sigma}_x^2} + \frac{(y-m_y)^2}{\tilde{\sigma}_y^2}])dxdy$$

avec

$$\begin{aligned} \tilde{\sigma}_x &= \sigma_x cos^2\theta + \rho\sigma_x\sigma_y sin2\theta + \sigma_y^2 sin^2\theta \\ \tilde{\sigma}_y &= \sigma_x sin^2\theta - \rho\sigma_x\sigma_y sin2\theta + \sigma_y^2 cos^2\theta \end{aligned}$$

après un changement de variables trivial, en passant en coordonnées polaires, on en déduit que :

$$P((X,Y) \in D) = \frac{1}{2\pi}\int_{-\pi}^{+\pi}\int_0^{r_0} e^{\frac{-r^2}{2}} r dr d\theta$$

En conclusion il faut que $\alpha = e^{-r_0^2/2}$ soit $r_0 = \sqrt{-2\ln\alpha}$.
L'ellipse ainsi obtenue est de centre (m_x, m_y) et fait un angle θ avec Ox et la longueur des demi-axes est donnée par $r_0\tilde{\sigma}_x$ et $r_0\tilde{\sigma}_y$.

Chapitre 3

Estimation

L'objet de ce chapitre n'est pas de donner une méthode générale d'estimation, mais plutôt d'exposer quelques propriétés et définitions qui seront reprises par la suite.

3.1 Généralités

L'estimation consiste à rechercher la valeur numérique d'un ou plusieurs paramètres inconnus d'une loi de probabilité à partir d'observations (valeurs prises par la v.a. qui suit cette loi de probabilité). On utilise pour cela un estimateur fonction de la v.a. étudiée: quand la v.a. prend comme valeur l'observation, la valeur de l'estimateur est appelée estimation. L'exemple suivant illustre ces définitions. On s'interesse au GMQ des porcs . Supposons que ce GMQ que nous noterons X est distribué normalement, en d'autres termes que X suit une loi $\mathbf{N}(m, \sigma^2)$, où m représente le GMQ moyen de toute la population de porcs et σ^2 la variance de la distribution des GMQ. Les paramètres m et σ^2 sont inconnus, l'objet de l'estimation est de trouver une valeur "raisonnable" pour ces paramètres. Deux possibilités s'offrent à nous:- soit on peut mesurer le GMQ de tous les porcs de la population et, dans ce cas, les paramètres m et σ^2 seront parfaitement connus,- soit la population est trop grande, et, on est obligé de travailler sur un échantillon.Cet

échantillon va nous donner des informations sur les vraies valeurs (celles de la population) de m et σ^2. Supposons que l'on ait étudié le GMQ (en grammes) sur un échantillon de taille n=10. Notons $X_1, X_2...X_{10}$, le GMQ des porcs $N°1, N°2...N°10$ de cet échantillon.
La moyenne de l'échantillon (notée $\bar{X}$) est une "approximation" de la moyenne m de la population. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur de m.

Num porc	1	2	3	4	5	6	7	8	9	10
GMQ (g)	500	530	560	510	620	560	540	610	600	580

Table 3.1: Table des Gains Moyens Quotidiens observés sur un échantillon de 10 porcs

Le mot **estimateur** se réfère au procédé de calcul utilisé pour approximer $m.\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 561$ est une estimation de m.

Le mot **estimation** se réfère à la valeur numérique utilisée pour approximer.

En général un estimateur est une variable aléatoire, en d'autres termes l'estimation du paramètre dépend des individus présents dans l'échantillon. Si un autre échantillon avait été considéré, une autre estimation du paramètre aurait été obtenue. Le choix de l'estimateur se fait selon des critères qui mesurent sa proximité au paramètre inconnu. Nous allons dans ce qui suit présenter la liste des critères les plus souvent utilisés pour définir les "qualités " d'un estimateur.

3.2 Estimateur convergent

Une des propriétés élémentaires que doit remplir un estimateur est d'être convergent. En d'autres termes, lorsque la taille de l'échantillon tend vers l'infini, il faut que l'estimateur se "rapproche" du paramètre qu'il estime. Il existe plusieurs façons de mesurer cette proximité qui donnent lieu à la définition de plusieurs types de convergence. Notre objectif n'étant pas ici de faire un cours de statistiques fondamentales, nous nous bornerons à citer

les principaux types de convergence et à les illustrer à l'aide des deux exemples suivants :

exemple 1 :

Soient $X_1, \ldots, X_n$, n variables aléatoires de même loi $\mathcal{N}(m, \sigma^2)$. On s'interesse à la convergence de la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ vers m.

exemple 2 :

Soit X une variable aléatoire distribuée selon une loi $\mathcal{B}(n, p)$. On s'interesse à la convergence de $\hat{p}_n = X/n$ vers p.

Dans un cadre plus général, nous noterons T_n un estimateur du paramètre θ obtenu à partir d'un échantillon de taille n qui vérifie pour tout n, $\mathbb{E}(T_n) = \theta$ (cf paragraphe suivant).

Définition : *L'estimateur T_n est convergent en* **moyenne quadratique** *si :*

$$Var(T_n) \longrightarrow 0$$

quand $n \longrightarrow \infty$.

Rappelons que la variance d'une variable aléatoire est définie par $Var(T_n) = \mathbb{E}(T_n - \mathbb{E}(T_n))^2 = \mathbb{E}(T_n - \theta)^2$. Dire que T_n converge en moyenne quadratique signifie en fait que lorsque n tend vers l'infini la distance moyenne qui sépare T_n de θ tend vers 0.

Il est facile d'établir que $Var(\bar{X}_n) = \frac{\sigma^2}{n}$. Par conséquent lorsque $n \longrightarrow \infty$, $Var(\bar{X}_n) \longrightarrow 0$.

De même $Var(\hat{p}_n) = \frac{p(1-p)}{n}$ tend vers 0 quand n tend vers ∞.

Définition : *L'estimateur T_n est convergent en* **probabilité** *si : pour tout $\varepsilon > 0$ fixé la quantité*

$$P(\|T_n - \theta\| > \varepsilon)$$

tend vers 0 quand n tend vers ∞

Ce type de convergence peut s'interpréter de la façon suivante : Supposons que l'on se fixe un intervalle de largeur 2ε centré sur θ. Supposons de plus que nous disposons d'un grand nombre de réalisations de T_n (obtenu avec un grand nombre d'échantillons de taille n). On s'interesse au pourcentage de ces réalisations qui "tombent" dans en dehors de cet intervalle. Alors, l'estimateur T_n converge en probabilité vers θ si ce pourcentage tend vers 0

quand n tend vers l'infini. Il faut noter que ceci ne présume en rien de la distance qui sépare les réalisations de T_n en dehors de l'intervalle, de la valeur de θ. En revanche, si T_n converge en moyenne quadratique alors il converge en probabilité.
Vous avez déjà montré en prépa que la moyenne empirique (resp. $\hat{p}$) converge en probabilité vers m (resp. p). La preuve est une simple application de l'inégalité de Tchébychev.
Définition :*L'estimateur T_n est presque sûrement convergent si :*

$$P(\lim_{n\to\infty} T_n \neq \theta) = 0$$

On voit à travers cette définition que la convergence presque sure est une convergence beaucoup plus "forte" que la convergence en probabilité : elle implique la convergence en probabilité. Pour obtenir une convergence presque sure, il est nécessaire que la convergence en proba soit suffisamment rapide pour que n assez grand un très faible pourcentage de réalisations de T_n ne tombent en dehors de l'intervalle que nous avons défini précédemment. En réfléchissant un peu, on peut voir que si T_n converge en probabilité alors, il est possible de trouver une sous suite de $(T_n)_n$ qui converge presque surement. La preuve de la convergence presque sure de la moyenne empirique et de $\hat{p}$ sort des objectifs de ce cours.

3.3 Estimateur sans biais

Un estimateur peut être **sans biais.** Un estimateur est sans biais si, à taille d'łéchantillon finie et fixée, les différentes estimations d'un même paramètre sur différents échantillons admettent le paramètre à estimer comme barycentre; ou plus simplement, si T est un estimateur de θ , $\mathbb{E}(T) = \theta$.
L'opérateur $\mathbb{E}(.)$ est utilisé pour symboliser la moyenne de population de la variable aléatoire sur laquelle il opère. Revenons à notre exemple des GMQ et supposons que 1000 échantillons aient été faits. Ces 1000 échantillons ont fournis 1000 estimations du GMQ moyen (celui de la population). Dire que

$\bar{X}$ est un estimateur sans biais de m équivaut à dire que sur un grand nombre d'échantillons, m est la moyenne des $\bar{X}_i$. On pourrait croire à tort que tous les estimateurs usuels sont sans biais, c'est faux, les exemples suivants sont les plus connus.
Un estimateur classiquement utilisé pour la variance est:

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

c'est un estimateur biaisé de la variance, il sous-estime en moyenne la variance de population, en effet

$$\mathbb{E}(\hat{\sigma}_n^2) = (1 - \frac{1}{n})\sigma^2.$$

On voit à partir de la formule précédente qu'un estimateur sans biais de la variance est donné par

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

Si la moyenne de population m est connue, il est facile de montrer qu'un estimateur sans biais de la variance est donné par

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

Plus généralement, si g est une fonction non linéaire, et si T est un estimateur sans biais de θ, alors

$$I\!E(g(T)) \neq g(\theta).$$

Ainsi, en prenant $g(x) = \sqrt{x}$ un obtient

$$I\!E(\sqrt{\hat{\sigma}_{n-1}^2}) \neq \sigma$$

la quantité $\sqrt{\hat{\sigma}_{n-1}^2}$ n'est donc pas un estimateur sans biais de l'écart type σ.

3.4 Estimateur de variance minimum

Un estimateur peut être de **variance minimum.** Comme le montre le schéma ci-dessus, $\bar{X}$ est aléatoire, en d'autres termes pour différents échantillons, on obtient différentes estimations de m. En général on utilise comme indice de dispersion de l'estimateur sans biais T de θ, la quantité $\mathbb{E}[(T-\theta)^2]$ c'est-à-dire la moyenne des carrés des écarts de T au paramètre estimé θ. Cette quantité n'est autre que la variance (théorique càd calculée avec les paramètres de population) de l'estimateur quand il est sans biais.
Un critère de choix des estimateurs est que sa dispersion ne soit pas trop grande. Une technique d'estimation (le maximum de vraisemblance) permet de construire des estimateurs qui asymptotiquement sont de variance minimum.

> *La plupart des estimateurs que vous utilisez classiquement sont des estimateurs de variance minimum, en d'autres termes, il n'existe pas d'estimateurs plus "précis" permettant d'estimer la quantité que vous étudiez.*

Définition : Soit $x = (x_1, \ldots, x_n)$ une observation d'un échantillon. $(X_1, \ldots, X_n)$ de taille n dont la densité $f_\theta(x)$ dépend d'un paramètre θ (à estimer).

On définit la vraisemblance de l'échantillon par :

$$L(x_1, \ldots, x_n, \theta) = f(x_1, \theta) \ldots f(x_n, \theta)$$

Les n observations étant indépendantes, la **vraisemblance** apparaît comme la probabilité d'obtention de l'échantillon dans le cas discret et comme la densité de cette probabilité dans le cas continu.

Sous certaines conditions de régularité de la vraisemblance, on a l'inégalité suivante (Cramer-Rao) : Soit T un estimateur d'une fonction $g(\theta)$ alors

$$var(T) \geq \frac{[g'(\theta)]^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln L(x_1, \ldots, x_n, \theta)\right)^2\right]}$$

avec

$$\ln L(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ln f(x_i, \theta)$$

On voit donc que si T est un estimateur sans biais de θ alors $g(\theta) = \theta$ et $g'(\theta) = 1$. De plus, si f vérifie certaines conditions de régularité alors :

$$Var(T) \geq \frac{-1}{\mathbb{E}(\frac{\partial^2 \ln f_\theta}{\partial \theta^2})}$$

Cette inégalité montre qu'à taille d'échantillon finie, la variance d'un estimateur sans biais ne peut être inférieure à une certaine limite. Il est donc illusoire de penser qu'il est possible d'accéder aux paramètres de population sur un échantillon de taille finie). Un estimateur est **efficace** *si sa variance atteint la borne inférieure de Cramer-Rao en d'autres termes si:*

$$Var(T) = \frac{-1}{I\!E(\frac{\partial^2 \ln f_\theta}{\partial \theta^2})} = \text{borne inf de cramer Rao.}$$

Exemple :
On veut estimer le GMQ d'une population de porc. A cet effet deux échantillons indépendants sont tirés. Sur la premier échantillon de taille 10, une moyenne de $\bar{x} = 580g$ est observée, sur le second échantillon de taille 30 on observe une moyenne de 620 g.

Pour estimer la moyenne de population, on vous propose deux procédés de calcul

(1) $$z_1 = \frac{\bar{x} + \bar{y}}{2} = \frac{580 + 620}{2} = 600g$$

(2) $$z_2 = \frac{10\bar{x} + 30\bar{y}}{10 + 30} = 610g$$

A votre avis, y a t-il une estimation meilleure que l'autre ?
Pour répondre à cette question simple, nous allons examiner deux propriétés de ces estimateurs. Tout d'abord, nous allons regarder si ces estimateurs sont biaisés, nous examinerons ensuite la "précision" de chacun de ces estimateurs. Nous noterons par la suite

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i,\ \bar{Y} = \frac{1}{30}\sum_{i=1}^{30} Y_i,$$

et nous supposerons que les va X_i sont indépendantes, que les va Y_i sont indépendantes et que les X_i et les Y_i sont indépendantes.
Pour examiner le biais éventuel de chacun des estimateurs Z_1 et Z_2, il suffit de calculer leur espérance:

$$\mathbb{E}(Z_1) = \mathbb{E}(\frac{\bar{X}+\bar{Y}}{2}) = \frac{1}{2}\mathbb{E}(\bar{X}+\bar{Y}) = \frac{1}{2}\big(\mathbb{E}(\bar{X})+\mathbb{E}(\bar{Y})\big)$$

Or nous savons que les porcs proviennent de la même population et que $\bar{X}$ et $\bar{Y}$ sont des estimateurs non biaisés de m. On en déduit que

$$\mathbb{E}(Z_1) = \mathbb{E}(\frac{\bar{X}+\bar{Y}}{2}) = \frac{1}{2}(m+m) = m.$$

Z_1 est donc un estimateur non biaisé de m. Faisons le même travail pour Z_2

$$\mathbb{E}(Z_2) = \mathbb{E}(\frac{10\bar{X}+30\bar{Y}}{10+30}) = \frac{10}{10+30}\mathbb{E}(\bar{X})+\frac{30}{10+30}\mathbb{E}(\bar{Y}) = \frac{10m}{10+30}+\frac{30m}{10+30} = m$$

Z_2 est aussi un estimateur non biaisé de m : ce critère ne suffit donc pas pour faire un choix.
Comme ces estimateurs sont non biaisés, un indice de mesure de leur dispersion est donné par leur variance :

$$Var(Z_1) = Var(\frac{\bar{X}+\bar{Y}}{2}) = \frac{1}{4}(Var(\bar{X})+Var(\bar{Y})) = \frac{1}{4}(\frac{\sigma^2}{10}+\frac{\sigma^2}{30}) = \frac{\sigma^2}{30}$$

et

$$\begin{aligned} Var(Z_2) &= Var\big((\tfrac{10}{10+30})\bar{X}+(\tfrac{30}{10+30})\bar{Y}\big) = \big(\tfrac{10}{10+30}\big)^2 Var(\bar{X}) + \big(\tfrac{30}{10+30}\big)^2 Var(\bar{Y}) \\ &= \big(\tfrac{10}{10+30}\big)^2 \tfrac{\sigma^2}{10} + \big(\tfrac{30}{10+30}\big)^2 \tfrac{\sigma^2}{30} = \tfrac{\sigma^2}{40} \end{aligned}$$

L'estimateur Z_2 possède donc une variance plus petite que l'estimateur Z_1.

3.5 Une méthode générale d'estimation : le maximum de vraisemblance

Fisher a proposé une méthode basée sur la remarque suivante : les meilleures valeurs du paramètre inconnu θ sont celles qui donnent à l'événement observé $(x_1, \ldots, x_n)$ la plus grande probabilité.

On a vu que cette probabilité peut être "représentée" par la vraisemblance

$$L(x, \theta) = f(x_1, \theta) \dots f(x_n, \theta).$$

L'estimation "maximum de vraisemblance" de θ sera une fonction des observations qui rend $L(x, \theta)$ maximum.

Remarque : il est équivalent de rendre maximum

$$\ln L(x, \theta) = \sum_{i=1}^{n} \ln f(xi, \theta).$$

Un exemple d'application

Estimation de la moyenne et de la variance d'un échantillon gaussien.

Soit $(x_1, \dots, x_n)$ une observation d'un échantillon $(X_1, \dots, X_n)$ de taille n. Les v.a. X_i sont indépendantes et de loi $\mathcal{N}(m, \sigma^2)$ avec m et σ^2 inconnus.

Ecrivons la vraisemblance.

$$L(x_1, \dots, x_n, m, \sigma^2) = f(x_1, m, \sigma^2) \times f(x_2, m, \sigma^2) \times \dots \times f(x_n, m, \sigma^2)$$

on en déduit que

Or

$$\ln f(x_i, m, \sigma^2) = -\tfrac{1}{2} ln(2\pi\sigma) - \tfrac{(x_i - m)^2}{2\sigma^2}$$
$$\Longrightarrow \quad \textstyle\sum_{i=1}^{n} \ln f(x_i, m, \sigma^2) = -n\tfrac{1}{2} ln(2\pi\sigma) - \sum_{i=1}^{n} \tfrac{(x_i - m)^2}{2\sigma^2}$$

On cherche d'abord la valeur σ^2 qui maximise $\ln L$. C'est la valeur qui annule la dérivée par rapport à σ.

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^3} = 0$$

De même, on cherche la valeur de m qui annule la dérivée partielle de la log vraisemblance par rapport à m et on trouve :

$$\frac{\partial \ln L}{\partial m} = \sum_{i=1}^{n} \frac{(x_i - m)}{\sigma^2} = 0$$

On arrive finalement à

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i \; et \; \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{m})^2$$

Remarque : Si on calcule $\mathbb{E}(\hat{\sigma}_n^2)$ on a :

$$\mathbb{E}(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2 = (1 - \frac{1}{n})\sigma^2$$

L'estimateur $\hat{\sigma}_n^2$ n'est donc pas sans biais (il sous estime la variance), en revanche l'estimateur :

$\hat{\sigma}_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{m})^2$ est sans biais.

3.6 Une bricole sur le théorème central limit

Un théorème important sera souvent évoqué dans ce cours, **le théorème "central limit"**.

En voici un énoncé un peu formel:

Soient $X_1, X_2, ..., X_n$ n variables aléatoires indépendantes identiquement distribuées de moyenne m et de variance σ^2 alors:

$$\lim_{n\to\infty} \mathcal{L}\big(\sqrt{n}\frac{\bar{X} - m}{\sigma}\big) = \mathcal{N}(0,1)$$

ou encore : $\forall a, b \in \mathbb{R}$

$$\lim_{n\to\infty} P(a \leq \sqrt{n}\frac{\bar{X} - m}{\sigma} \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(b) - \Phi(a)$$

où Φ est la fonction de répartition d'une loi normale $N(0,1)$. Ce théorème, signifie, que si un grand nombre de quantités aléatoires indépendantes, de même variance sont ajoutées, alors la distribution de la somme est une loi normale. C'est une des raisons qui justifie l'utilisation de la loi normale pour les opérations sur les moyennes, même quand la population n'est pas normalement distribuée (cf le jeu de dés vu en cours).

3.7 Applications

L'objet de ce paragraphe est de montrer l'utilisation de certains estimateurs couramment rencontrés en statistiques. Le mot estimation recouvre en fait deux types de technique :
- l'**estimation ponctuelle** $\longrightarrow$ une valeur du paramètre à estimer,
- l'**estimation par intervalle** $\longrightarrow$ un intervalle dans lequel il est vraisemblable de trouver avec une probabilité donnée $(1-\alpha)$ le paramètre à estimer (on parle alors d'intervalle de confiance de sécurité $1-\alpha$).

3.7.1 Estimation des paramètres d'une loi normale

Soient $X_1, \ldots, X_n$ n va indépendantes de même loi $\mathcal{N}(m, \sigma^2)$. Nous commençons par estimer la variance puis nous estimons la moyenne. Afin d'effectuer des estimations par intervalle, nous avons besoin de la proposition suivante :
Proposition :
Soit $\bar{X} = \frac{1}{n}\sum X_i$ *et* $\hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum (X_i - \bar{X})^2$ *alors :*

1 $\bar{X} \sim \mathcal{N}(m, \dfrac{\sigma^2}{n})$

2 $\dfrac{(n-1)\hat{\sigma}^2_{n-1}}{\sigma^2} \sim \chi^2_{n-1}$

3 $\bar{X} \text{et} \hat{\sigma}^2_{n-1}$

sont indépendantes Pour illustrer l'emploi des formules, nous reprendrons les données de l'exemple des GMQ de la page 40 nous supposons donc que la normalité des GMQ est déjà démontrée).

$Numporc$	1	2	3	4	5	6	7	8	9	10
$GMQ(g)$	500	530	560	510	620	560	540	610	600	580

Estimation de la variance

Un estimateur sans biais de la variance est donné par

$$\hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

D'après l'affirmation (2) de la proposition précédente,

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

d'où

$$P(c^2_{\alpha/2} \leq \frac{(n-1)\hat{\sigma}^2_{n-1}}{\sigma^2} \leq c^2_{1-\alpha/2} = 1-\alpha$$

où $c^2_{\alpha/2}$ est la valeur limite au seuil $\alpha/2$ d'une loi du χ^2 à $n-1$ degrés de liberté.

Un intervalle confiance de sécurité $1-\alpha$ de σ^2 est donc donné par

$$(n-1)\frac{\hat{\sigma}^2_{n-1}}{c^2_{1-\alpha/2}} \leq \sigma^2 \leq (n-1)\frac{\hat{\sigma}^2_{n-1}}{c^2_{\alpha/2}}$$

Application :

Dans cet exemple $n = 10$ et une estimation de la variance est donnée par $\hat{\sigma}^2_{n-1} = 1721.11$ Un intervalle de sécurité 0.95 peut alors facilement être construit : la table du χ^2 nous donne pour $10-1 = 9$ degrés de liberté $c^2_{0.05/2} = 2.700$ et $c^2_{1-0.05/2} = 19.023$ nous en déduisons donc que nous avons 95 chances sur 100 de trouver la variance dans l'intervalle

$$[(10-1)\frac{1721.11}{19.023}; (10-1)\frac{1721.11}{2.700}] \text{ soit}$$

$$814.277 \leq \sigma^2 \leq 5737.03$$

Les logiciels de stat (presque tous américains) fournissent en général deux quantités supplémentaires : la *standard deviation* (notée SD) qui ici vaut 41.486 et le *standard error* (noté se) dont la valeur est 13.119 Ces deux quantités n'estiment pas la même chose : SD est définie comme la racine carrée de la variance et peut être assimilée à une estimation (biaisée) de

l'écart-type. SD nous donne donc une idée de la dispersion des GMQ dans la population des porcs. Quand la taille de l'échantillon augmente, il est donc tout à fait naturel de voir SD se stabiliser autour d'une valeur.
La quantité se est définie par $SD/\sqrt{n}$ et elle peut être utilisée comme uns estimation (biaisée elle aussi) de l'**écart-type de la moyenne**. se nous donne donc une idée de la "précision" de l'estimation de la moyenne que nous obtenons avec un échantillon de taille n. Quand la taille de l'échantillon augmente il faut donc s'attendre à une diminution de se (plus on a de données plus on est précis).

Estimation de la moyenne

Un estimateur sans biais de la moyenne est donné par

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

En utilisant l'affirmation 1 de la proposition, il vient

$$\sqrt{n}\frac{\bar{X} - m}{\sigma} \sim \mathcal{N}(\prime, \infty)$$

et d'après la deuxième affirmation

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

. Comme $\bar{X}$ et σ^2_{n-1} sont indépendantes, nous en déduisons que la statistique

$$T = \frac{\bar{X} - m}{\frac{\hat{\sigma}^2_{n-1}}{\sqrt{n}}} \sim \text{Student}_{n-1}$$

Un intervalle confiance de sécurité $1-\alpha$ de m est donc donné par

$$(MOY) \qquad \bar{X} - t^{1-\alpha/2}_{n-1}\sqrt{\frac{\hat{\sigma}^2_{n-1}}{n}} \leq m \leq \bar{X} + t^{1-\alpha/2}_{n-1}\sqrt{\frac{\hat{\sigma}^2_{n-1}}{n}}$$

où encore

$$\bar{X} - t^{1-\alpha/2}_{n-1} se \leq m \leq \bar{X} + t^{1-\alpha/2}_{n-1} se$$

avec $t_{n-1}^{1-\alpha/2}$ est la valeur limite au seuil $1-\alpha/2$ d'une loi du student à $n-1$ degrés de liberté.
Application :
Dans notre exemple $n = 10$ et une estimation de la moyenne est donnée par $\bar{X} = 561$ Un intervalle de sécurité 0.95 peut alors facilement être construit : la table de Student nous donne pour $10-1 = 9$ degrés de liberté $t_9^{1-0.05/2} = 2.262$ nous en déduisons donc que nous avons 95 chances sur 100 de trouver la moyenne de population dans l'intervalle

$$[561 - 2.262\sqrt{\frac{1721.11}{10}}; 561 + 2.262\sqrt{\frac{1721.11}{10}}] \text{ soit}$$

$$526.6 \leq m \leq 595.36$$

Attention : Il y a souvent confusion entre l'intervalle de confiance de la moyenne défini par (MOY) et l'intervalle dans lequel se trouve une certaine fraction de la population défini comme suit :

$$(POP) \qquad [\bar{X} - t_{n-1}^{1-\alpha/2}\sqrt{\frac{n+1}{n}\hat{\sigma}_{n-1}^2}; \bar{X} + t_{n-1}^{1-\alpha/2}\sqrt{\frac{n+1}{n}\hat{\sigma}_{n-1}^2}]$$

Cette confusion est souvent renforcée par des présentations de résultats de la forme $\bar{x} \pm et$ où et est une quantité qui est soit SD soit se. Il est clair que pour être interprétable il est nécessaire de savoir ce que et représente.
Pour obtenir (POP), considérons une va X indépendante des $(X_i)_i$ et de loi $\mathcal{N}(m, \sigma^2)$. Alors $X - \bar{X} \sim \mathcal{N}(0, \sigma^2 \frac{n+1}{n})$ et en reprenant le même raisonnement que celui que nous venons de faire pour la construction de (MOY), il est facile d'obtenir le résultat.
Dans notre exemple, l'intervalle dans lequel se trouvent 95 % de la population vaut

$$[561-2.262\sqrt{\frac{10+1}{10}1721.11}; 561+2.262\sqrt{\frac{10+1}{10}1721.11}] \text{ soit } [447.00; 674.99]$$

En utilisant le théorème central limit il est facile de voir que l'intervalle de confiance de la moyenne (MOY) ne dépend pas tellement de la distribution des données si la taille de l'échantillon

est suffisante. En d'autres termes, l'hypothèse de normalité de la distribution peut être relaxée pour des échantillons de taille assez grande. En revanche, il est clair que la forme de la distribution est très importante pour les intervalles dans lesquels se trouvent une certaine portion de la population (POP).

3.7.2 Estimation d'un pourcentage

L'objet de ce paragraphe est de montrer les techniques de construction des intervalles de confiance des pourcentages. Pour construire un intervalle de confiance, nous avons besoin d'identifier les lois de probabilités sous-jacentes. A cet effet prenons des notations. Soit X une variable aléatoire distribuée selon une loi Binomiale de paramètre N et p. X est donc le nombre d'individus qui satisfait une certaine condition de la forme $(0, 1)$ avec une probabilité p. La quantité N est déterministe et connue et on cherche une valeur raisonnable de p. Il est clair qu'un estimateur sans biais de p est donné par $\hat{p} = \frac{X}{N}$. En revanche, la recherche d'un intervalle de confiance de p pose quelques problèmes : les seuls intervalles faciles de construire ne sont qu'approximatifs et ils ne deviennent vraiment fiables que lorsque n est assez grand.

méthode 1 (exacte)

Cette méthode de construction d'intervalle de confiance est exacte. Par conséquent aucune hypothèse concernant la taille de l'échantillon n'est requise. Il est difficile de l'utiliser directement sans faire appel à des techniques d'analyse numérique ; aussi on a souvent recours à des tables ou à des logiciels spécialisé. Notons $\hat{P}_{sup}$ la solution de

$$\sum_{i=0}^{x} C_N^i p^i (1-p)^{N-i} = \alpha/2$$

et $\hat{P}_{inf}$ la solution de

$$\sum_{i=x}^{N} C_N^i p^i (1-p)^{N-i} = \alpha/2$$

alors un intervalle de sécurité $1-\alpha$ est donné par $[\hat{P}_{inf}; \hat{P}_{sup}]$.

méthode 2

Cette méthode repose sur le même principe que la méthode exacte. On approxime la loi Binomiale (de paramètres N et p par la loi de Poisson de paramètre Np. Il faut donc que les conditions.requises pour cette approximation soient vérifiées (N grand p petit, Np raisonnable).

méthode 3

Grace au théorème central limit et à la loi des grand nombres, nous savons que pour N assez grand, la quantité

$$U = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}.$$

est approximativement distribuée selon une loi $\mathcal{N}(0,1)$. (Il faut que les conditions.requises pour cette approximation soient vérifiées) Un intervalle de sécurité $1-\alpha$ est donc donné par

$$\hat{p} - u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \leq p \leq \hat{p} + u_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

où $u_{1-\alpha/2}$ est la valeur limite au seuil $\alpha/2$ d'une loi $N(0,1)$ (Si $\alpha = 0.05$ alors $u_{1-\alpha/2} = 1.96$).

Application :

On s'intéresse au pourcentage d'animaux porteur d'une anomalie. Supposons que sur un échantillon de taille $N = 100$ on a observé $x = 10$ animaux porteurs de cette anomalie alors $\hat{p} = 0.1 = 10/100$. Notre objectif est de construire l'intervalle de confiance de sécurité $1-\alpha$.

En utilisant la méthode 1 nous devons résoudre :

$$\sum_{i=0}^{10} C_{100}^{i} \hat{p}_{sup}^{i} (1-\hat{p}_{sup})^{100-i} = 0.025$$

et

$$\sum_{i=10}^{100} C_{100}^{i} \hat{p}_{inf}^{i} (1-\hat{p}_{inf})^{100-i} = 0.025$$

Un calcul avec un logiciel spécialisé nous donne $\hat{p}^i_{sup} = 0.1762$ et $\hat{p}^i_{inf} = 0.0491$
L'intervalle de confiance de sécurité 0.95 de p est donc : [0.0491 ; 0.1762].

Enfin, la construction d'un intervalle de confiance de sécurité 95% avec la méthode 3 nous conduit à

$$[0.1 - 1.96\sqrt{\frac{0.1 \times 0.9}{100}}; 0.1 + 1.96\sqrt{\frac{0.1 \times 0.9}{100}}] = [0.0412; 0.1588].$$

Ces résultats sont proches de ceux que l'on obtient avec la méthode exacte et sont obtenus grace à un calcul direct.

Chapitre 4

Tests d'hypotheses

4.1 Généralités

Un test d'hypothèses sert à répondre à une question.Répondre à une question suppose que soient déjà définis: la question (des hypothèses) et, une façon d'y répondre (une règle de décision). L'objet de ce chapitre est d'examiner plus précisément les questions (les hypothèses) et les règles de décision ; en d'autres termes les tests d'hypothèses. Pour situer le problème, commençons par un exemple.

Exemple : Comparaison de 2 insulines (A et B) sur la diminution de la concentration en glucose dans le sang chez des chiens diabétiques. Une expérience est réalisée sur 20 chiens sur lesquels un prélèvement de sang est effectué 15 minutes après l'administration de l'insuline. 10 chiens ont reçu l'insuline A, et 10 chiens ont reçu l'insuline B. L'objectif de l'expérience est de comparer les diminutions moyennes de glucose des chiens. Pour simplifier, nous supposerons que :

- la diminution de la concentration en glucose est normalement distribuée,
- pour les deux insulines, l'écart-type de diminution de concentration en glucose est connue et vaut 59 $mg/100ml$
- les deux moyennes m_A et m_B des diminutions sont inconnues.

Des exemples de questions:

1) La diminution moyenne (de la concentration en glucose) des animaux

traités avec A est elle égale à la diminution moyenne des animaux traités avec B ou la diminution moyenne des animaux traités avec A est elle différente de la diminution moyenne des animaux traités avec B ? Ce qui peut encore s'écrire : $m_A = m_B$ ou $m_A \neq m_B$.
2) La diminution moyenne (de la concentration en glucose) des animaux traités avec A est elle égale à la diminution moyenne des animaux traités avec B ou la diminution moyenne des animaux traités avec A est elle supérieure à la diminution moyenne des animaux traités avec B ? Ce qui peut encore s' écrire : $m_A = m_B$ ou $m_A \geq m_B$.
3) La diminution moyenne (de la concentration en glucose) des animaux traités avec A est elle égale à la diminution moyenne des animaux traités avec B ou la diminution moyenne des animaux traités avec A est elle inférieure d'au moins $20mg/100ml$ à la diminution moyenne des animaux traités avec B ? Ce qui peut encore s' écrire : $m_A = m_B$ ou $m_A \leq m_B - 20$. Pour répondre à ces questions, il faut avoir des informations sur m_A et m_B. Deux possibilités se présentent :
- soit on connaît déjà m_A et m_B, auquel cas on peut répondre à la question posée,
- soit m_A et m_B sont inconnues, et dans ce cas il faut faire une expérience pour avoir des informations sur ces paramètres.
Supposons que m_A et m_B sont inconnues et donc que l'on fasse une expérience. Il existe à nouveau 2 cas de figures:
- soit l'essai est mené sur toute la population des animaux pouvant recevoir les insulines A et B, et, dans ce cas les valeurs de m_A et m_B seront connues avec certitude, et l'on peut répondre à la question posée,
- soit il est impossible de mener l'essai sur tous les animaux pouvant recevoir ces traitements et dans ce cas, il faut se contenter d'échantillons des populations concernées.
Par la suite nous nous placerons toujours dans ce cas de figure où m_A et m_B sont inconnues et estimées à partir d'échantillons. Comme ces moyennes sont estimées à partir d'échantillons, on ne dispose pas des vraies valeurs de m_A et m_B (celles de la population), les seules valeurs dont nous disposons sont

$\hat{m}_A$ et $\hat{m}_B$, qui (sauf extraordinaire coup de chance) sont différentes de m_A et m_B. La règle de décision qui nous permettra de répondre à la question posée sera donc construite à partir de valeurs "approximatives" de m_A et m_B, valeurs obtenues sur les échantillons. Des exemples de règles de décision:
1) On dira que la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec A est différente de la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec B si $\hat{m}_A$ est très différente de $\hat{m}_B$, par exemple si $|\hat{m}_A - \hat{m}_B| > 30mg/100ml$.
2) On dira que la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec A est supérieure à la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec B si par exemple $\hat{m}_A \geq \hat{m}_B + 30mg/100ml$.

Passons à des définitions un peu plus formelles des hypothèses et des règles de décisions associées.

4.2 Hypothèse

Une hypothèse est un ensemble de valeurs des paramètres inconnus (paramètres de population).
Par exemple l'hypothèse: "la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec A est égale à la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec B" peut encore s'écrire : $\{(m_A, m_B)$ tels que $m_A - m_B = 0\}$.
Une hypothèse peut être simple ou composée.
Une hypothèse est dite simple si elle contient une unique valeur des paramètres inconnus, elle est composée dans le cas contraire.
Un exemple d'hypothèse simple: la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec A est égale à 80 mg/100ml, ou encore, $\{m_A = 80\}$. Il faut noter que si la variance de la réponse était inconnue, cette hypothèse ne serait pas simple.
Un exemple d'hypothèse composée: "la diminution moyenne de la concentration en glucose dans le sang des animaux traités avec A est égale à la

diminution moyenne de la concentration en glucose dans le sang des animaux traités avec B" : $\{(m_A, m_B)$ tels que $m_A - m_B = 0\}$. En effet, si nous supposons que la variance de la réponse est connue, l'ensemble $\{(m_A, m_B)$ tels que $m_A - m_B = 0\}$ contient une infinité de couple (m_A, m_B). En revanche, si les mêmes chiens reçoivent successivement les deux insulines, et si nous supposons que le variance de la différence est connue, la paramètre inconnu est alors $\delta = m_A - m_B$; l'hypothèse H_0 s'exprime alors sous la forme $\{\delta = 0\}$ et l'hypothèse H_0 est simple. Nous verrons un peu plus loin dans ce chapitre le rôle fondamental que joue cette propriété.
Dans une question, il y a deux hypothèses: une hypothèse d'équivalence que nous appellerons **hypothèse nulle**, notée H_0 une **hypothèse alternative**, en général de non équivalence, qui sera notée H_1.
On appellera **test**, la donnée d'un jeu d'hypothèses et d'une règle de décision.
Un test est **unilatéral** si l'hypothèse H_1 s'exprime sous forme d'inégalités.
Il est **bilatéral** si l'hypothèse H_1 est exprimée avec des symboles "$\neq$".

4.3 Définition des risques

Supposons que l'on se soit fixé une règle de décision pour répondre à la question $N°1$:La diminution moyenne (de la concentration en glucose) des animaux traités avec A est elle égale à la diminution moyenne des animaux traités avec B ou la diminution moyenne des animaux traités avec A est elle différente de la diminution moyenne des animaux traités avec B soit $H_0 : m_A = m_B$, $H_1 : m_A \neq m_B$. Comme nous l'avons déjà vu, cette règle de décision est fondée sur des valeurs estimées de m_A et m_B, elle peut donc conduire à des erreurs. Ces erreurs sont habituellement classées en 2 catégories: l'erreur de première espèce et évidemment l'erreur de seconde espèce.
A chacune de ces erreurs correspond un (ou des) risque(s).
Ainsi le risque de commettre une erreur de première espèce s'appelle **risque de première espèce** (il est noté α), et , le risque de commettre une erreur

de seconde espèce s'appelle **risque de seconde espèce** (il est noté β).[1]
Le risque de première espèce est le risque de rejeter (avec la règle de décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est vraie.
Le risque de seconde espèce est le risque d'accepter (avec la règle de décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est fausse.
En général on présente ces risques dans le tableau suivant La quantité $1-\beta$

		DECISION
REALITE	H_0 vraie	H_1 vraie
H_0 vraie	$1-\alpha$	α
H_1 vraie	β	$1-\beta$

est une probabilité de bonne décision appelée **puissance du test.**
Revenons à notre exemple, supposons que la règle de décision choisie pour répondre à la question $N°1$ soit la suivante:
On dira que les insulines A et B sont différentes si $|\hat{m}_A - \hat{m}_B| > 50$.
Le risque α peut s'interpréter dans ce problème comme le risque de décider que les insulines A et B sont différentes alors qu'en réalité elles sont équivalentes.
En d'autres termes, α est le risque d'observer sur les échantillons des valeurs $\hat{m}_A$ et $\hat{m}_B$ telles que $|\hat{m}_A - \hat{m}_B| > 50$ alors qu'en réalité $m_A = m_B$.
Le risque β s'interprète comme le risque de décider que les insulines sont équivalentes alors qu'en réalité elles sont différentes.
β est donc le risque d'observer sur les échantillons des valeurs $\hat{m}_A$ et $\hat{m}_B$ telles que $\hat{m}_B$ telles que $|\hat{m}_A - \hat{m}_B| < 50$ alors qu'en réalité $m_A \neq m_B$.

Supposons que nous ayons utilisé la règle de décision suivante:

On dira que les insulines A et B sont différentes si $|\hat{m}_A - \hat{m}_B| > 60$.
Cette nouvelle règle est d'une part plus "contraignante" que la précédente

[1]D.SCHWARTZ a défini pour des hypothèses unilatérales un troisième risque noté γ. Ce risque permet de définir ce qu'il appelle *l'attitude pragmatique*. Bien que conceptuellement intéressante, cette approche n'est pas utilisée en dehors de nos frontières

pour rejeter l'hypothèse H_0 ; il faut que la différence entre $\hat{m}_A$ et $\hat{m}_B$ soit "grande" pour dire que m_A et m_B sont différents; et d'autre part moins "exigeante" que la précédente pour accepter l'hypothèse H_0 (même une différence de l'ordre de 55 entre $\hat{m}_A$ et $\hat{m}_B$ ne permet pas de conclure à la différence entre m_A et m_B).

Il apparaît donc que cette nouvelle règle de décision possède un risque de première espèce inférieur à la règle 1), et, un risque de seconde espèce supérieur.

Ce petit exemple illustre bien le fait que: **les risques α et β sont liés et varient en sens inverse.**

Quand on réalise un test, la démarche est inversée: les hypothèses H_0 et H_1 et le risque de première espèce α sont fixés [2] ; une règle de décision dont le risque de première espèce correspond à celui que l'on s'est fixé est alors recherchée.

A taille d'échantillon donnée, se fixer un risque α équivaut à se fixer un risque β.

Voyons sur un exemple les conséquences (souvent désastreuses) de cette propriété:

Exemple:

On veut tester $H_0 : m_A = m_B$ contre $H_1 : m_A \neq m_B$ (m_A et m_B ont le même sens que précédemment).

A cet effet un essai a été effectué sur des échantillons de taille 10. Les résultats sont les suivants : $\hat{m}_A = 150$, $\hat{m}_B = 100$. On suppose (pour simplifier le problème) que les variances sont connues de façon déterministes : $\sigma_A = \sigma_B = 59$

Si on se fixe un risque $\alpha = 5\%$, la règle de décision est la suivante: on rejette l'hypothèse H_0 si $|\hat{m}_A - \hat{m}_B| > 55.4$. Avec les résultats de l' essai, l'hypothèse H_0 n'est pas rejetée.

Le prince de la formule conclura que $m_A = m_B$ avec un risque de 5% "de se tromper" ? Analysons l'erreur que commet ce prince si souvent rencontré: le "risque de 5% de se tromper" correspond à un risque de première espèce

[2] Le risque α est classiquement fixé à 5%. Je ne connais pas l'argument scientifique qui milite en faveur de cette valeur. Toute explication sera la bienvenue

que nous avons fixé a priori à 5%.
Ce risque s'interprète comme le risque de décider à tort que les effets des insulines A et B sont différents. Or, notre règle de décision n'a pas rejeté l'hypothèse H_0 d'équivalence des effets.
Le risque α n'est donc d'aucune utilité dans cette décision, le risque qui garde un sens est le risque de seconde espèce β qui est ici voisin de 70%.
On a donc presque 70% de chance avec cette règle de décision et cette taille d'échantillon de conclure à l'égalité des effets des insulines alors qu'en réalité ces effets sont différents.
Pour éviter ce gag classique, il existe une solution: calculer le nombre de sujets nécessaires.
Un test statistique est par nature négatif.
Accepter H_0 ne signifie pas que cette hypothèse est vraie mais seulement que les observations disponibles ne sont pas incompatibles avec cette hypothèse et que l'on n'a pas de raison suffisante de lui préférer l'hypothèse H_1 compte tenu des résultats expérimentaux.

4.4 Ce qu'il ne faudrait pas croire

Quand on écrit les hypothèses à tester, on utilise un certain formalisme qui est souvent trompeur. Par exemple, l'hypothèse que nous écrivons H_0 : $m_A = m_B$ est un moyen pratique pour écrire que nous voulons voir si m_A et m_B ne sont pas trop différentes, en d'autres termes si $|m_A - m_B| < \Delta$.
Δ est le seuil à partir duquel on estime que les moyennes sont "biologiquement" différentes.Lorsque Δ n'est pas fixé *a priori*, ce sont les riques α et β adoptés et la taille d'échantillon qui le fixe à votre place. Ceci explique le comportement courant de certains biologistes qui devant des résultats de tests "très significatifs" proclament que cette différence statistique n'a aucun sens biologique. Il est clair que dans ce cas, le nombre d'unités statistiques qui a été utilisé est trop important compte-tenu des objectifs fixés. La différence minimale que le test est alors capable de mettre en évidence devient alors sans intérêt biologique. Un test est un peu comme un microscope dont le

grossissement est réglé par la taille de l'échantillon.

Il faut noter que les hypothèses formulées sous la forme

$$H_0 \ : \ |m_A - m_B| \leq \Delta$$

ne sont pas simples et que par conséquent les risques α et β ne sont pas uniquement définis.

4.5 Tests paramétriques et non paramétriques

Un test paramétrique est un test pour lequel des hypothèses sur la distribution des populations sont requises. La plupart des tests paramétriques qui seront abordés dans ce cours sont construits en faisant l'hypothèse de normalité des distributions.
On qualifie de non paramétriques les méthodes statistiques qui sont applicables dans les conditions générales quant aux distributions des populations.
Les anglo-saxons utilisent l'expression "distribution free", qui bien mieux que "non paramétriques", décrit ce dont il s'agit.

4.6 Quelques remarques

Le paragraphe suivant contient une batterie de tests qui devraient vous permettre de "faire face" à la plupart des situations rencontrées en pratique. Un certain nombre de remarques doivent être faites concernant l'utilisation et l'interprétation des tests.
La plupart des logiciels de statistiques et des publications fournissent une valeur de probabilité P : comment s'interprète t-elle ?
Lorsque nous réalisons "à la main" un test, nous calculons une statistique que nous comparons (pour un risque α fixé) à une valeur théorique. Dans l'exo précédent, nous avons calculé $u = \frac{50}{59\sqrt{\frac{2}{10}}}$ que nous avons comparé à la valeur limite d'une loi $N(0,1)$ (i.e. 1.96 pour un risque α de 5%.) La règle de décision que nous avons utilisé est la suivante : si $u > 1.96$ alors on rejette H_0. On peut noter que 1.96 est la valeur pour laquelle $P(X > 1.96) = 0.05$

(où X est une va N(0,1)).
La valeur P annoncée correspond à la définition suivante : soient X une va de même loi que la statistique de test quand l'hypothèse nulle est vraie et u la valeur observée sur l'échantillon de cette statistique de test, alors $P = P(X > u)$. Par conséquent, si $P < 5\%$, l'hypothèse H_0 est rejetée avec un risque $\alpha = 5\%$. De même, si $P < 1\%$, l'hypothèse H_0 est rejetée avec un risque $\alpha = 1\%$. C'est une démarche légèrement différente de celle que nous avons utilisée dans le paragraphe précédent dans lequel toutes les règles de décisions annoncées sont construites en supposant que le risque de première espèce α est fixé a priori.
Les logiciels fonctionnent différemment: la valeur P est le risque de première espèce maximal, calculé à partir de l'échantillon. Ainsi, dans un test de Student de comparaison de moyennes, une valeur $P = 0.02$ signifie que l'on prend un risque de 2% de dire que les moyennes sont différentes alors qu'en réalité elles sont égales. Ces quantités (P) sont des variables aléatoires (elles dépendent des observations) qui mesurent un risque observé. Il n'est donc pas conseillé de les interpréter telles quelles, mais plutôt de les comparer á des risques fixés *a priori*. Les valeurs "P" ne mesurent pas nécessairement l'importance (biologique) d'une variable.
Une variable (biologiquement) importante peut avoir (dans un test) une valeur P élevée (non significative) si l'échantillon est petit ou si cette variable est mesurée avec beaucoup d'erreur.
De même, une variable qui n'est pas (biologiquement) importante peut avoir une valeur P très petite dans un échantillon de grande taille.
Calculer un intervalle de confiance d'un paramètre, donnera souvent une information plus pertinente que la simple valeur de P.
De plus, et en guise de conclusion sur ce sujet, **les valeurs de P annoncées par les logiciels sont des approximations. Les hypothèses requises pour calculer la valeur exacte de P ne sont jamais satisfaites en pratique.**

Chapitre 5

Tests classiques

5.1 Comparaisons portant sur les variances

La comparaison de variances est un outil essentiel des statistiques, nous l'utiliserons intensivement en régression multiple et en analyse de la variance.
Supposons que nous disposons de p échantillons gaussiens indépendants de tailles respectives $n_1, \ldots, n_p$. On peut pour chaque échantillon, calculer un estimateur sans biais de la variance de la population. Par exemple, pour le k^{ieme} échantillon, un estimateur sans biais de la variance de population σ_k^2 est donné par:

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (X_i^k - \bar{X}^k)^2$$

où (X_i^k est la i^{ieme} donnée de l'échantillon k, et, $\bar{X}^k$ est la moyenne de l'échantillon k.
Maintenant que nous disposons de notations, passons aux tests.

5.1.1 Comparaison d'une variance à une valeur déterministe

On veut ici comparer la variance obtenue à partir d'un échantillon, que nous noterons $\hat{\sigma}_1^2$ à une valeur donnée (fixée) *a priori* notée σ_0^2
Test de $H_0 : \sigma_1^2 = \sigma_0^2$ contre $H_1 : \sigma_1^2 \neq \sigma_0^2$

La règle de décision est la suivante:
on rejette H_0 avec un risque de première espèce α si :

$$(n_1 - 1)\frac{\hat{\sigma}_1^2}{\sigma_0^2} > \chi^2_{1-\alpha/2} \quad ou \quad si\ (n_1 - 1)\frac{\hat{\sigma}_1^2}{\sigma_0^2} < \chi^2_{\alpha/2}$$

où $\chi^2_{\alpha/2}$ est la valeur limite au seuil $\alpha/2$ d'une loi du χ^2 à $n_1 - 1$ degrés de liberté.

5.1.2 Comparaison de deux variances

a) Test bilatéral
On veut tester l'hypothèse: $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$
On ne sait pas à priori si une des variances est supérieure à l'autre.
Sans perte de généralités, on peut supposer que $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$
La règle de décision est alors la suivante: si $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} > f^{1-\alpha/2}_{n_1-1,n_2-1}$ alors on rejette l'hypothèse nulle.
où $f^{1-\alpha/2}_{n_1-1,n_2-1}$ est la valeur limite au seuil $1 - \alpha/2$ d'une loi de FISHER à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.
Le **premier** degré de liberté $n_1 - 1$ est celui du **numérateur**, le **second** degré de liberté est celui du **dénominateur**.
b) Test unilatéral
On veut tester l'hypothèse: $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 > \sigma_2^2$
La règle de décision est alors la suivante: si $F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} > f^{1-\alpha}_{n_1-1,n_2-1}$ alors on rejette l'hypothèse nulle.
où $f^{1-\alpha}_{n_1-1,n_2-1}$ est la valeur limite au seuil $1 - \alpha$ d'une loi de FISHER à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.
Le **premier** degré de liberté $n_1 - 1$ est celui du **numérateur**, le **second** degré de liberté est celui du **dénominateur**.

5.1.3 Comparaison de plusieurs variances

On veut tester l'hypothèse: $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$
Il existe plusieurs méthodes pour tester ces hypothèses, la plus couramment utilisée est le test de Bartlett.

Test de Bartlett

On dispose des estimations de ces p variances à comparer
Notons $n = \sum_{i=1}^{p} n_i$, $SCE = \sum_{i=1}^{p}(n_i - 1)\hat{\sigma}_i^2$ et enfin,$\hat{\sigma}^2 = \frac{SCE}{n-p}$.
Si l'hypothèse H_0 est vraie, alors $\hat{\sigma}^2$ est une estimateur sans biais de $\hat{\sigma}_1^2$
Le principe du test de Bartlett est, en quelque sorte, de comparer cette valeur aux $\hat{\sigma}_i^2$
La règle de décision est la suivante:
si

$$\chi^2_{obs} = \frac{(n-p)ln(\hat{\sigma}^2) - \sum_{i=1}^{p}(n_i - 1)ln(\hat{\sigma}_i^2)}{1 + \frac{1}{3(p-1)}(\sum_{i=1}^{p}\frac{1}{n_i - 1} - \frac{1}{n-p})} > \chi^2_{1-\alpha}$$

où $\chi^2_{1-\alpha}$ est la valeur limite au seuil $1-\alpha$ d'une loi du χ^2 à $p-1$ degrés de liberté, alors on rejette l'hypothèse nulle.
Ce test est très utilisé, car il permet de comparer des variances calculées sur des effectifs différents.

Test de Hartley

On note n_{min} la taille du plus petit échantillon dont nous disposons, et n_{max} la taille du plus grand échantillon.
Notons de plus SCE_{max} la plus grande de toutes les valeurs $(n_1 - 1)\hat{\sigma}_1^2, (n_2 - 1)\hat{\sigma}_2^2, ..., (n_p - 1)\hat{\sigma}_p^2$, et, SCE_{min} la plus petite de toutes les valeurs $(n_1 - 1)\hat{\sigma}_1^2, (n_2 - 1)\hat{\sigma}_2^2, ..., (n_p - 1)\hat{\sigma}_p^2$.
Le test de Hartley repose sur la statistique :$H = \frac{SCE_{max}}{SCE_{min}}$ et la règle de décision est la suivante:
on rejette H_0 si $H > H_{p,n_{min}-1}$ et on accepte H_0 si $H < H_{p,n_{max}-1}$.
Les quantités $H_{p,n_{max}-1}$ et $H > H_{p,n_{min}-1}$ se trouvent dans les tables de Hartley.

Test de Cochran

Le test de Cochran ne peut être utilisé que si les effectifs de chaque échantillon sont égaux. Il est basé sur la statistique $C = \frac{\hat{\sigma}^2_{max}}{\sum_{i=1}^{p}\hat{\sigma}_i^2}$

où $\hat{\sigma}^2_{max}$ est le plus grand des $(\hat{\sigma}^2_i)$.
On rejette l'hypothèse nulle si: $C > C^{1-\alpha}_{p,n_1-1}$ où $C^{1-\alpha}_{p,n_1-1}$ est lue dans la table de Cochran.

5.2 Comparaisons portant sur les moyennes

La plupart des techniques permettant de comparer deux moyennes ne peuvent être utilisées que si un certain nombre d'hypothèses sont vérifiées.
Dans un premier temps, donnons nous des notations et précisons ces hypothèses. Supposons que nous disposons de deux échantillons de taille respective n et p que nous noterons $X_1, X_2, ..., X_n$ et $Y_1, Y_2, ...Y_p$.
Les $(X_i)_{i=1..n}$ suivent une loi $\mathbf{N}(m_X, \sigma^2_X)$ et sont indépendantes.
De même les $(Y_i)_{i=1..p}$ suivent une loi $\mathbf{N}(m_Y, \sigma^2_Y)$, elles sont indépendantes et elles sont indépendantes des $(X_i)_{i=1..n}$.
Le fait de supposer que toutes les variables aléatoires ($(X_i)_{i=1..n}$ par exemple) suivent une même loi de probabilité, signifie simplement que toutes les observations dont nous pouvons disposer doivent provenir d'une même population et que, pour cette population, la variable étudiée (X par exemple) ait une moyenne m_X et une variance σ^2_X.
L'indépendance, signifie que la valeur que va prendre X_2 par exemple ne doit pas être "influencée" par les autres valeurs (pas de phénomène de contagion).
Comme nous disposons d'échantillons, nous ne pouvons avoir accès aux valeurs de populations de la moyenne et de la variance ; les seules informations dont nous disposons sont des estimations de ces valeurs.
Donnons donc un nom à ces estimations.
Nous noterons $\bar{x}$ et $\bar{y}$ les moyennes respectives des (x_i) et des (y_i) soit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$

Les variances de population sont estimées sans biais par: $\hat{\sigma}^2_X = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, et $\hat{\sigma}^2_Y = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \bar{y})^2$.

Rappelons enfin que la moyenne $\bar{X}$ est aléatoire (la valeur qu'elle prend dépend de
l'échantillon), elle a une variance $\frac{\sigma_X^2}{n}$, de même $\bar{Y}$ a une variance égale à $\frac{\sigma_Y^2}{p}$.
Nous pouvons maintenant passer aux tests.

5.2.1 Comparaison d'une moyenne à une valeur donnée $\mathbf{m_0}$

Il existe deux possibilités de tests suivant la connaissance que l'on a, *a priori*, du phénomène étudié.

a) La variance de population est connue σ_0^2

• test bilatéral:ł$H_0 : m_X = m_0$ contre $H_1 : m_x \neq m_0$
la règle de décision est la suivante:
rejet de H_0 si

$$\frac{|\bar{X} - m_0|}{\sqrt{\frac{\sigma_0^2}{n}}} \geq u_{1-\alpha/2}$$

• test unilatéral:ł$H_0 : m_X = m_0$ contre $H_1 : m_X > m_0$
la règle de décision est la suivante:
rejet de H_0 si

$$\frac{\bar{X} - m_0}{\sqrt{\frac{\sigma_0^2}{n}}} \geq u_{1-\alpha}$$

b) La variance de population est inconnue

Elle est donc estimée à partir de l'échantillon par $\hat{\sigma}_X^2$
• test bilatéral:
$H_0 : m_X = m_0$ contre $H_1 : m_X \neq m_0$ la règle de décision est la suivante:
rejet de H_0 si

$$\frac{|\bar{X} - m_0|}{\sqrt{\frac{\hat{\sigma}_X^2}{n}}} \geq t_{1-\alpha/2}^{n-1}$$

• test unilatéral:ł$H_0 : m_X = m_0$ contre $H_1 : m_X > m_0$
la règle de décision est la suivante

rejet de H_0 si

$$\frac{\bar{X} - m_0}{\sqrt{\frac{\hat{\sigma}_X^2}{n}}} \geq t_{n-1}^{1-\alpha}$$

5.2.2 Comparaison de deux moyennes

Deux cas de figures se présentent, soit les échantillons sont appariés, en d'autres termes les observations des deux échantillons sont réalisées sur les mêmes individus, soit les échantillons sont indépendants.
Si les échantillons sont appariés, il faut calculer la moyenne des différences et on est alors ramené au cas précédent de comparaison d'une moyenne à une valeur donnée.
Si les échantillons sont indépendants, il existe à nouveau deux possibilités:
- soit les variances des deux des populations dont proviennent les échantillons peuvent être considérées comme égales (résultat issu d'un test)
- soit les variances des deux populations ne sont pas égales.

a) Premier cas: les variances sont égales

Si les variances des deux populations sont égales, alors un estimateur sans biais de la variance de population est donnée par:

$$\hat{\sigma}^2 = \frac{(n-1)\hat{\sigma}_X^2 + (p-1)\hat{\sigma}_Y^2}{n+p-2}$$

Test de comparaison de la différence de deux moyennes à une valeur donnée D_0

• test bilatéral:
$H_0 : m_X - m_Y = D_0$ contre $H_1 : m_X - m_Y \neq D_0$
la règle de décision est la suivante:rejet de H_0 si:

$$\frac{|\bar{X} - \bar{Y} - D_0|}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{1}{p})}} \geq t_{n+p-2}^{1-\alpha/2}$$

Il faut noter que le fait de ne pas rejeter l'hypothèse nulle n'implique nullement que cette hypothèse est vraie. Il est tout à fait possible que l'hypothèse H_1 soit vraie, mais que compte tenu de la taille des échantillons, la puissance

de ce test soit epsilonesque. Supposons que $D_0 = 0$ (cette hypothèse n'est pas nécessaire, mais elle permet de simplifier les notations). Les hypothèses testées sont donc $H_0 : m_X = m_Y$ contre $H_1 : m_X \neq m_Y$
Notons que pour montrer l'égalité stricte entre les moyennes, il faudrait toute la population. En général, on se fixe un nombre Δ au delà de laquelle la différence $|m_X - m_Y|$ a un sens biologique. Supposons ce nombre Δ fixé alors, sous l'hypothèse H_1, la quantité

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{1}{p})}}$$

est distribuée selon une loi de Student décentrée à $n+p-2$ degrés de liberté et avec un paramètre de décentrage δ avec

$$\delta = \frac{\Delta}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{1}{p})}}$$

Supposons que $T_{n+p-2}(\delta)$ est une variable aléatoire qui suit une loi de Student décentrée à $n + p - 2$ degrés de liberté et avec un paramètre de décentrage δ, alors la puissance $1 - \beta$ est donnée par

$$P(T_{n+p-2}(\delta) > t_{n+p-2}^{1-\alpha/2}) = 1 - \beta.$$

Cette probabilité peut être trouvée dans les tables de la loi de Student décentrée. Si vous ne disposez pas de telles tables, vous pouvez utiliser l'approximation suivante : Soit Z une va $\mathcal{N}(0, 1)$, alors

$$P(T_{n+p-2}(\delta) > t_{n+p-2}^{1-\alpha/2}) = P(Z > z_\beta)$$

avec

$$z_\beta = \frac{t_{n+p-2}^{1-\alpha/2} - \delta}{\sqrt{1 + \frac{\left(t_{n+p-2}^{1-\alpha/2}\right)^2}{2(n+p-2)}}}$$

Si les effectifs par groupe sont assez grands et sont égaux, on peut utiliser l'approximation suivante :

$$n = 2(u_{1-\alpha/2} + u_{1-\beta})^2 \frac{\sigma^2}{\Delta^2}$$

n est l'effectif par groupe, et $u_{1-\alpha/2}$ est la valeur limite au seuil $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$.
Enfin, il existe des abaques ou des programmes qui permettent le calcul de la puissance.
• test unilatéral:ł$H_0 : m_X - m_Y = D_0$ contre $H_1 : m_X - m_Y > D_0$
la règle de décision est la suivante:rejet de H_0 si:

$$\frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{1}{p})}} \geq t_{n+p-2}^{1-\alpha}$$

Dans le cas unilatéral, la puissance est calculée en utilisant les formules du cas bilatéral après avoir substitué α par 2α. Ainsi, quand les effectifs sont assez grand on a:

$$n = 2(u_{1-\alpha} + u_{1-\beta})^2 \frac{\sigma^2}{\Delta^2}$$

b) Second cas: les variances ne sont pas égales
Si les variances des deux populations sont différentes, on peut utiliser le test d'Aspin-Welch
Ce test est basé sur la statistique

$$\frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{p}}}$$

Ce test possède exactement les mêmes règles de décision que lorsque les variances sont égales, seul le nombre de degrés de liberté de la loi de Student utilisée doit être changé.
Il est calculé en utilisant la formule:

$$ddl = \frac{\left(\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{p}\right)^2}{\left(\frac{\hat{\sigma}_X^2}{n}\right)^2/(n-1) + \left(\frac{\hat{\sigma}_Y^2}{p}\right)^2/(p-1)}$$

Ce degré de liberté est toujours inférieur ou égale à n+p-2. Il est d'autant plus petit que les variances sont hétérogènes (l'égalité a lieu lorsque les variances observées sont égales). Le fait de diminuer le degré de liberté implique une augmentation des valeurs limites auxquelles la statistique de test est comparée et par conséquent l'utilisation d'un test plus conservatif (qui maîtrise mieux le risque α en le surestimant).

5.3 Comparaisons portant sur les proportions

5.3.1 Comparaison d'une proportion à une valeur donnée

Considérons une population infinie d'individus possédant l'un ou l'autre de deux caractères opposés de laquelle on prélève un échantillon aléatoire d'effectif n. On note X le nombre d'individus qui possèdent le premier caractère, $\hat{p} = \frac{X}{n}$ est alors un estimateur sans biais de la proportion p d'individus de la population qui possèdent ce caractère.
On peut se poser un certain nombre de questions sur p: par exemple savoir si cette proportion est égale à une proportion donnée p_0 (fixée *a priori*). Pour répondre à cette question, deux tests d'hypothèses peuvent être réalisés selon que l'hypothèse alternative est unilatérale ou bilatérale.
Ces deux tests ne sont à utiliser que si x et $n-x$ sont assez grands (la valeur 5 est généralement la valeur minimale tolérée par les biologistes).

a) Test bilatéral

$H_0 : p = p_0$ contre $H_1 : p \neq p_0$.
Deux règles de décision sont usuellement utilisées:
1) on rejette H_0 si

$$u_{obs} = \frac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > u_{1-\alpha/2}$$

2) on rejette H_0 si

$$u_{obs} = 2\sqrt{n}|arcsin\sqrt{\frac{x}{n}} - arcsin\sqrt{p_0}| > u_{1-\alpha/2}$$

$u_{1-\alpha/2}$ est la valeur limite au seuil $1-\alpha/2$ d'une loi $\mathbf{N}(0,1)$ et $arcsin$ est la fonction réciproque de la fonction sinus.

ATTENTION

Si vous utilisez la seconde règle de décision, il faut qu'au moment du calcul de $arcsin$, les angles soient exprimés en radians, pas en degrés.

b) Test unilatéral

$H_0 : p = p_0$ contre $H_1 : p > p_0$.
Deux règles de décision sont usuellement utilisées:

1) on rejette H_0 si

$$u_{obs} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} > u_{1-\alpha}$$

2) on rejette H_0 si

$$u_{obs} = 2\sqrt{n}(arcsin\sqrt{\frac{x}{n}} - arcsin\sqrt{p_0}) > u_{1-\alpha}$$

5.4 Comparaison de deux proportions

Souvent, on veut comparer la proportion d'individus d'une population à une autre proportion d'individus, ou encore comparer p_1 et p_2.
Les données dont nous disposons sont, d'une part les effectifs n_1 et n_2 des deux échantillons, d'autre part la répartition de ces n_1 et n_2 individus en fonction du caractère étudié.
Les données peuvent être présentées dans une table de contingence qui a la forme suivante :

	échantillon 1	échantillon 2	Totaux
caractère 1	a	b	a+b
caractère 2	c	d	c+d
Totaux	a+c	b+d	a+b+c+d
	ou n_1	ou n_2	ou $n_1 + n_2$

Les symboles a, b, c, d représentent les effectifs observés correspondants aux quatre cellules de ce tableau.
Test des hypothèses: $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.
a) Test exact
Les tests usuellement utilisés, sont des tests asymptotiques tout à fait acceptables pour des effectifs assez élevés. Dans certains cas, les effectifs sont trop faibles pour faire raisonnablement confiance au risque annoncé par les logiciels, il reste alors une solution: utiliser un test exact. La loi Hypergéométrique permet de déterminer la probabilité de rencontrer, lorsque

H_0 est vraie, une hypothèse aussi anormale que celle réellement observée.
On obtient:

$$P(a) = \frac{C_{a+c}^{a} C_{b+d}^{b}}{C_{a+b+c+d}^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!(a+b+c+d)!}$$

Si la probabilité d'observer un effectif égal à a ou un effectif plus anormal (quand
l'hypothèse H_0 est vraie) est faible, c'est à dire si la répartition observée n'est pas compatible avec l'hypothèse H_0 alors, on rejette cette hypothèse.
Prenons un exemple.
On veut comparer la sensibilité de deux races bovines à la trypanosomiase. Cinquante bovins, appartenant à deux races différentes, ont été observés dans le but de comparer la sensibilité de ces deux races à la trypanosomiase. Les résultats sont consignés dans le tableau suivant: [h] Les marges du tableau

	Race 1	Race 2	Total
non infestés	14	0	14
infestés	5	31	36
Total	19	31	50

étant fixées (nombre de bêtes infestées et non infestées, et nombres de bêtes de race 1 et 2) le tableau suivant donne la probabilité d'observer les effectifs a, b, c, d quand H_0 est vraie: En additionnant ces probabilités à partir des deux extrémités de la distribution, on constate que l'hypothèse d'égalité des taux d'infestation des deux races doit être rejetée au niveau 0.05 lorsque a est soit inférieur ou égal à 1, soit supérieur ou égal à 9. C'est en effet entre 1 et 2 d'une part et entre 8 et 9 d'autre part que la probabilité cumulée dépasse la valeur $0.0250 = \frac{0.05}{2}$.
Il en résulte que la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie, est $0.0045 + 0.0202 = 0.0247$, c'est à dire moins que le risque initialement fixé.

b) Méthodes asymptotiques

• Test bilatéral

a	b	c	d	$P(a)$	$\sum P(a)$
0	14	19	17	0.0003	0.0003
1	13	18	18	0.0042	0.0045
2	12	17	19	0.0257	0.0302
3	11	16	20	0.0875	0.1177
4	10	15	21	0.1833	0.3010
5	9	14	22	0.2500	.
6	8	13	23	0.2282	.
7	7	12	24	0.1413	0.2208
8	6	11	25	0.0593	0.0795
9	5	10	26	0.0167	0.0202
10	4	9	27	0.0031	0.0035
11	3	8	28	0.0004	0.0004
12	2	7	29	0.0000	0.0000
13	1	6	30	0.0000	0.0000
14	0	5	31	0.0000	0.0000

Quand les effectifs des échantillons sont suffisamment élevés, on peut utiliser pour tester les hypothèses $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$ les approximations suivantes:

$$u_{obs} = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{p_0(1-p_0)(1/n_1 + 1/n_2)}}$$

avec $p_0 = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$ et on rejette H_0 si $u_{obs} \geq u_{1-\alpha/2}$
où $u_{1-\alpha/2}$ est la valeur limite au seuil $1 - \alpha/2$ d'une loi $\mathcal{N}(0, 1)$.
Ce test est équivalent au test du χ^2. La valeur du χ^2 observé se déduit de u_{obs} par la relation :$\chi^2_{obs} = u^2_{obs}$.
La formule suivante relie l'erreur de première espèce (α), l'erreur de seconde espèce (β), l'effectif par groupe n et les pourcentages p_1 et p_2

$$n = \frac{(u_{1-\alpha/2} + u_{1-\beta})^2}{2(arcsin\sqrt{p_1} - arcsin\sqrt{p_2})^2}.$$

- Test unilatéral

Pour tester les hypothèses $H_0 : p_1 = p_2$ contre $H_1 : p_1 > p_2$ on peut utiliser

les approximations suivantes:
si

$$u_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)(1/n_1 + 1/n_2)}} > u_{1-\alpha}$$

alors on rejette l'hypothèse nulle. La relation entre les risques, l'effectif par groupe n et les pourcentages p_1 et p_2 devient alors

$$n = \frac{(u_{1-\alpha} + u_{1-\beta})^2}{2(arcsin\sqrt{p_1} - arcsin\sqrt{p_2})^2}.$$

5.5 Test de conformité a une loi de proba

Une loi de probabilité est définie par "la probabilité" qu'elle donne à chaque point.
Pour les variables continues (poids, tailles) une fonction appelée densité [1] caractérise complètement la loi de probabilité.
La densité n'est, en fait, que l'histogramme des fréquences construit sur la totalité de la population quand les classes sont réduites à un point.
A partir de la densité, on peut construire d'autres fonctions, comme par exemple, la fonction de répartition F. Cette dernière peut s'interpréter comme la fonction des fréquences cumulées. Comme la densité, cette fonction définit complètement la loi de probabilité.
Un histogramme est un estimateur de la densité, la fonction des fréquences cumulées $\hat{F}$ [2] est un estimateur de la fonction de répartition.
La plupart des tests de conformité à une loi de probabilité, sont construits en comparant soit la fonction de répartition empirique à la fonction de répartition, soit, l'histogramme à la densité.

5.5.1 Test de Kolmogorov-Smirnov (KS)

Il permet de comparer la fonction de répartition empirique (construite à partir de l'échantillon) à la fonction de répartition théorique F d'une loi

[1] pour la loi normale, la densité est représentée par une courbe en cloche
[2] On dit aussi fonction de répartition empirique

normale. De façon plus précise, pour un échantillon $z_1, z_2, ..., z_n$ de taille n, $\hat{F}(z)$ est définie comme le pourcentage d'observations inférieures ou égale à z, ou encore

$$\hat{F}(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{[z_i \leq z]}$$

avec

$$\begin{aligned} 1_{[z_i \leq z]} &= 1 \quad si \quad z_i \leq z \\ &= 0 \quad sinon \end{aligned}$$

Le test de KS permet de tester les hypothèses:
H_0 : *La distribution de la population dont est issu l'échantillon est normale*, contre
H_1 : *La distribution de la population dont est issu l'échantillon n'est pas normale.*
Ce test est basé sur la statistique:

$$K = \sqrt{n}[\max_i |F(z_i) - \frac{i - 0.5}{n}| + \frac{1}{2n}]$$

qui mesure l'éloignement de la fonction de répartition empirique et de la fonction de répartition théorique.
La règle de décision est la suivante:
pour $\alpha = 0.05$, on rejette H_0 si $K \geq 1.36$
pour $\alpha = 0.01$, on rejette H_0 si $K \geq 1.63$

5.5.2 Test du χ^2 pour une loi normale

Il permet de comparer la densité d'une loi normale à l'histogramme construit à partir des observations. Le problème avec l'utilisation de l'histogramme, est le choix toujours arbitraire des classes, supposons néanmoins que p classes sont choisies.
Le principe du test du χ^2 est de comparer le pourcentage d'observations observé dans la classe numéro i, que nous noterons $\hat{P}_i$, au pourcentage

d'observation que contiendrait cette même classe,que nous noterons P_i, si la distribution de la population était normale.
Le test du χ^2 repose donc sur le calcul d'une distance entre P_i et $\hat{P}_i$, et ceci pour chaque classe, ou, pour être plus précis,

$$\chi^2_{obs} = n \sum_{i=1}^{n} \frac{(\hat{P}_i - P_i)^2}{P_i}$$

ce qui peut aussi s'exprimer avec les effectifs de chaque classe n_i:

$$\chi^2_{obs} = \sum_{i=1}^{n} \frac{(n_i - nP_i)^2}{nP_i}$$

Pour tester les hypothèses:
H_0 : *La distribution de la population dont est issu l'échantillon est normale*, contre
H_1 : *La distribution de la population dont est issu l'échantillon n'est pas normale.*
pour un risque de première espèce α, la règle de décision est la suivante:
on rejette H_0 si:$\chi^2_{obs} \geq \chi^2_{1-\alpha}$ où $\chi^2_{1-\alpha}$ est la valeur limite au seuil $1-\alpha$ d'une loi du χ^2 à $p-3$ degrés de liberté. Ce test peut être utilisé si pour tout i les quantités nP_i sont assez grandes (en général on impose à ces quantités d'être au moins supérieures à 5). Dans le cas contraire, il faut faire des regroupements des classes jusqu'à ce que cette condition soit vérifiée.

5.6 Comparaisons multiples

Nous allons examiner dans ce paragraphe les propriétés de l'analyse de variance à un facteur ainsi que les comparaisons multiples réalisables après cette analyse. Notre objectif n'est pas ici d'étudier les techniques de modélisation dans toutes leurs généralités, mais plutôt de présenter un outil particulier que nous utiliserons pour comparer plusieurs moyennes. L'exemple suivant illustre bien le type de question auquel nous allons essayer d'apporter une réponse.

5.6.1 Exemple

Une expérience a été réalisée pour comparer 5 traitements. Les résultats sont consignés dans le tableau suivant : Nous voulons savoir si :

T_1	T_2	T_3	T_4	T_5
92	112	118	124	123
100	113	112	117	121
106	109	116	118	130
97	113	116	121	120
104	110	113	122	121
100	112	121	115	122
100	113	118	119	120
97	107	115	126	122
95	111	112	122	123
103	109	109	111	124

- tous les traitements sont en moyenne équivalents.
- le traitement 1 étant un témoin, les autres traitements lui sont ils en moyenne supérieurs ?
- les traitements 2,3,4,5 sont-ils en moyennes équivalents ?
- peut -on ordonner les traitements ?

Pour répondre à ces questions, nous allons tout d'abord nous donner des notations et des hypothèses, puis une analyse de variance à un facteur sera réalisée, les hypothèses seront vérifiées, enfin les résultats de cette analyse nous permettront de répondre aux questions. ß5.1 Notations et hypothèses Les notations suivantes sont adoptées

$Y_{i,j}$ la réponse de l'unité expérimentale $N°j$ soumis au traitement $N°i$,

μ_i est l'effet moyen du traitement (i.e. la moyenne de la réponse de toute la population)

μ l'effet moyen général (il ne dépend pas du traitement)

τ_i l'effet différentiel du niveau i du facteur traitement ,

$\varepsilon_{i,j}$ l'erreur du modèle pour l'unité expérimentale $N°j$ soumis au traitement $N°i$.

Avec ces notations, nous pouvons maintenant écrire le modèle

$$Y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}.$$

ou de façon équivalente

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}.$$

Dans notre exemple, i varie de 1 à 5, et j varie de 1 à 10. Nous supposerons que les $(Y_{i,j})$ sont des variables aléatoires
- de même variance
- indépendantes
- normalement distribuées.

Ces hypothèses sur la réponse Y sont équivalents aux mêmes hypothèses sur les ε. La première hypothèse signifie que l'erreur faite sur chacune des unités expérimentales doit être à peu près constante. Les paramètres μ, τ_i et les paramètres de dispersions sont inconnus et doivent être estimés à partir des observations. C'est l'objet de l'analyse de variance.

5.6.2 Analyse de la variance

Les résultats de l'analyse de variance sont donnés ci-dessous:
DEP VAR: Y N: 50 MULTIPLE R: 0.922 SQUARED MULTIPLE R: 0.851

ESTIMATES OF EFFECTS

		Y
CONSTANT	113.48	
T	1	-14.08
T	2	-2.58
T	3	1.52
T	4	6.02

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
T	.326628E+04	4	816.5700000	64.2181929	0.0000000
ERROR	572.2000000	45	12.7155556		

5.6.3 Estimation des paramètres

[3] Dans un premier temps, les paramètres μ et τ_i sont estimés à partir des observations. Les estimateurs obtenus sont les estimateurs de maximum de vraisemblance qui, comme la variance est constante (hypothèse 1), sont aussi les estimateurs des moindres carrés. Ils sont donc obtenus en minimisant la quantité

$$\sum_{i=1}^{5}\sum_{j=1}^{10}\big(Y_{i,j}-(\mu+\tau_i)\big)^2=\sum_{i=1}^{5}\sum_{j=1}^{10}\varepsilon_{i,j}^2$$

On trouve ainsi :

$$\hat{\mu}=\frac{1}{10\times 5}\sum_{i=1}^{5}\sum_{j=1}^{10}Y_{i,j}$$

$\hat{\mu}$ est donc la moyenne générale et

$$\hat{\tau}_i=\frac{1}{10}\sum_{j=1}^{10}Y_{i,j}-\hat{\mu}$$

en d'autres termes, les $\hat{\tau}_i$ sont obtenus en calculant la différence entre la moyenne du traitement $N°i$ et la moyenne générale. On peut noter que par construction

$$\sum_i \hat{\tau}_i=0$$

Dans notre exemple, $\hat{\mu}=113.48$
$\hat{\tau}_1=-14.08$ $\hat{\tau}_2=-2.58$
$\hat{\tau}_3=1.52$

[3]Les formules qui suivent sont vraies lorsque le plan d'expérience est équilibré, en d'autres termes lorsque le même nombre d'unités expérimentales est utilisé pour chaque traitement. Lorsque le plan est déséquilibré, il faut tenir compte de certains facteurs de pondérations

$\hat{\tau}_4 = 6.02$
On en déduit donc que

$$\hat{\tau}_5 = -\hat{\tau}_4 - \hat{\tau}_3 - \hat{\tau}_2 - \hat{\tau}_1 = 9.12$$

Remarque :
On peut retrouver les moyennes par traitements. Par construction elles sont données par

$$\bar{Y}_i = \hat{\mu} + \hat{\tau}_i$$

Par exemple, pour le traitement $N°1$ on a:

$$\bar{Y}_1 = \hat{\mu} + \hat{\tau}_1 = 113.48 - 14.08 = 99.40$$

Il reste à calculer la variance expliquée par le facteur traitement, et la variance expliquée par la différence entre les unités expérimentales. Pour obtenir ces variances, calculons d'abord les sommes des carrés des écarts (SCE) associées. Notons tout d'abord que la SCE totale (que l'on peut interpréter comme la quantité d'information contenue dans les données) est donnée par

$$SCE_{totale} = \sum_{i=1}^{5} \sum_{j=1}^{10} \big(Y_{i,j} - \hat{\mu})\big)^2 = 3838.48$$

La variance totale est donc donnée par

$$\hat{\sigma}^2_{totale} = \frac{SCE_{totale}}{5 * 10 - 1} = 78.336$$

La SCE expliquée par la différence entre les unités expérimentales (c'est à dire non expliquée par le facteur traitement) est celle que nous avons minimisée soit :

$$SCE_{erreur} = \sum_{i=1}^{5} \sum_{j=1}^{10} \big(Y_{i,j} - (\hat{\mu} + \hat{\tau}_i)\big)^2$$

Elle est estimée avec $50 - 5 = 45$ degrés de liberté. Pour comprendre l'origine de ce nombre de degrés de liberté détaillons un petit peu. Cette SCE est en

fait la somme de SCE par traitement[4] que l'on calcule comme d'habitude

$$\begin{aligned} SCE_{erreur} &= SCE_{err,trt1} + SCE_{err,trt2} + SCE_{err,trt3} + SCE_{err,trt4} + SCE_{err,trt5} \\ &= \sum_{j=1}^{10} \left(Y_{1,j} - (\hat{\mu} + \hat{\tau}_1)\right)^2 + \sum_{j=1}^{10} \left(Y_{2,j} - (\hat{\mu} + \hat{\tau}_2)\right)^2 + \\ &\quad \sum_{j=1}^{10} \left(Y_{3,j} - (\hat{\mu} + \hat{\tau}_3)\right)^2 + \sum_{j=1}^{10} \left(Y_{4,j} - (\hat{\mu} + \hat{\tau}_4)\right)^2 + \sum_{j=1}^{10} \left(Y_{5,j} - (\hat{\mu} + \hat{\tau}_5)\right)^2 \end{aligned}$$

Or chacune de ces SCE est estimée avec $10 - 1$ degrés de libertés, le degré de liberté de la somme est ici la somme des degrés de liberté soit $5 \times (10 - 1) = 50 - 5 = 45$. On en déduit que la variance non expliquée par le modèle est

$$\hat{\sigma}^2_{erreur} = \frac{SCE_{erreur}}{45} = \frac{572.2}{45} = 17.715$$

Il reste maintenant à calculer la SCE expliquée par le facteur traitement. Comme rien ne se crée, rien ne se perd et tout se transforme, On obtient cette SCE par différence entre la SCE totale et la SCE résiduelle. On fait de même pour les degrés de liberté. On obtient ainsi

$$SCE_T = \sum_{i=1}^{5} \hat{\tau}_i^2$$

On voit que cette quantité ne peut être nulle que si tous les $\hat{\tau}_i$ sont nuls (ce qui est équivalent à dire que tous les $\hat{\mu}_i$ sont égaux). Le degré de liberté avec lequel est estimée cette SCE est $49 - 45 = 4$. La variance expliquée par le facteur traitement est la somme des carrés des écarts divisée par le degré de liberté soit

$$\hat{\sigma}^2_T = \frac{3266.28}{4} = 816.57$$

[4]Ce sont ces SCE par traitements que nous utiliserons pour vérifier les hypothèses d'égalité des variances

5.7 Tests d'hypothèses (paramétriques)

Le test d'hypothèses réalisé dans l'analyse de variance teste les hypothèses suivantes :

$$H_0 : \forall i = 1..5, \tau_i = 0$$

$$H_1 : \exists i \in \{1, 5\} / \tau_i \neq 0$$

Avant de calculer la statistique de test, notons que ce test ne nous informe que sur le fait que tous les traitements ne sont pas équivalents. En effet, si le test rejette l'hypothèse nulle, nous ne savons pas quel(s) traitement(s) diffère(nt) des autres. Aussi, le test réalisé au cours de l'analyse de variance n'est utilisable que si :
- il est non significatif
- il a une puissance suffisante pour détecter une différence.

Pour tester les hypothèses ci-dessus, on compare la variance expliquée par le facteur traitement à la variance non expliquée par le modèle soit :

$$F = \frac{\hat{\sigma}_T^2}{\hat{\sigma}^2 erreur}$$

Si l'hypothèse nulle est vraie, cette quantité suit une loi de Fisher à 4 et 45 degrés de libertés. Donc si F est supérieur à $f_{4,45}^{1-\alpha}$ (valeur qui se trouve dans la table de la loi de Fisher à 4 et 45 ddl), on rejette l'hypothèse nulle. En regardant la valeur de P, on constate que l'hypothèse nulle est rejetée avec un risque $\alpha < 0.001$. Nous venons d'apporter la réponse à la première question posée : tous les traitements ne sont pas équivalents. ß5.4 Puissance du test F Nous venons de fixer une règle de décision pour rejeter l'hypothèse H_0 et le risque de rejeter H_0 lorsque cette hypothèse est vraie est contrôlé. Supposons que la règle de décision ne nous ait pas permis de rejeter H_0, une question de pose alors : était il possible, compte tenu des effectifs de rejeter cette hypothèse ?

Pour répondre correctement à cette question, il faut se fixer une hypothèse H_1 particulière. Nous allons calculer la puissance du test de Fisher pour l'hypothèse H_1 suivante :

$$H_1 : \tau_1 = \tau_{01},\ \tau_2 = \tau_{02}, \ldots,\ \tau_5 = \tau_{05}$$

Les quantités τ_{0i} sont des quantités fixées *a priori*. Supposons maintenant que l'hypothèse H_1 que nous venons de nous fixer est vraie, alors la statistique de test

$$F = \frac{\hat{\sigma}_T^2}{\hat{\sigma}^2 erreur}$$

suit une loi de Fisher **décentrée** à 4 et 45 degrés de libertés et le paramètre de décentrage ϕ est donné par

$$\phi = \sqrt{\frac{n\sum \tau_{0i}^2}{k\sigma_{erreur}^2}} = \sqrt{\frac{10\sum \tau_{0i}^2}{5\sigma_{erreur}^2}}$$

n est le nombre d'observations par traitement, et k est le nombre de traitements. La puissance est donnée par

$$P(F_{4,45}(\phi) \geq f_{4,45}^{1-\alpha})$$

Comme la variance résiduelle (de l'erreur) est inconnue, nous nous servirons de son estimation $\hat{\sigma}_{erreur}^2$ pour calculer la puissance. Le calcul de la puissance ne peut pas se faire facilement, aussi utilise t-on des tables qui fournissent cette quantité en fonction des degrés de liberté, de α et de ϕ.

5.7.1 Méthode des contrastes

Une fonction linéaire des effets des traitements est une expression de la forme :

(1) $$\Psi = a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k$$

où les a_i sont des constantes arbitraires.Si on ajoute aux a_i la contrainte supplémentaire

$$\sum_{i=1}^{k} a_i = 0$$

alors l'expression (1) s'appelle un contraste. On voit ici que dans le cas de deux traitements, tester l'hypothèse

$$H_0 : \tau_1 = \tau_2 \text{ contre } H_1 : \tau_1 \neq \tau_2$$

est équivalent à tester

$$H_0 : \tau_1 - \tau_2 = 0 \text{ contre } H_1 : \tau_1 - \tau_2 \neq 0.$$

L'hypothèse H_0 s'écrit donc sous la forme d'un contraste (il suffit de prendre $a_1 = 1$ et $a_2 = -1$).On peut noter que tester $\tau_1 - \tau_2 = 0$ est strictement équivalent à tester $2\tau_1 - 2\tau_2 = 0$ où plus généralement $a\tau_1 - a\tau_2 = 0$ $a \neq 0$. On dit que deux contrastes sont équivalents s'ils diffèrent d'une constante multiplicative. Comme un contraste est une combinaison linéaire de paramètres inconnus, un estimateur sans biais de Ψ est donné par la combinaison linéaire des estimateurs des τ_i soit

$$\hat{\Psi} = a_1\hat{\tau}_1 + a_2\hat{\tau}_2 + \ldots + a_k\hat{\tau}_k$$

Avec cette remarque, il est maintenant très facile de construire un intervalle de confiance d'un contraste de sécurité $1 - \alpha$. Voyons dans le détail la technique de construction. Notons se_i l'écart type de $\hat{\tau}_i$, alors

$$Var\hat{\Psi} = \sum a_i^2 se_i^2$$

ceci n'est vrai que si les estimateurs $\hat{\tau}_i$ sont indépendants. Dans le cas contraire, il faut tenir compte des corrélations entre les τ_i. En notant $N - p$ le degré de liberté avec lequel est estimée la variance $\hat{\sigma}^2_{erreur}$, on en déduit que

$$\hat{\Psi} - t_{N-p}^{1-\alpha/2}\sqrt{Var(\hat{\Psi})} \leq \Psi \leq \hat{\Psi} + t_{N-p}^{1-\alpha/2}\sqrt{Var(\hat{\Psi})}$$

est un intervalle de confiance de sécurité $1 - \alpha$ de Ψ.

5.7.2 Orthogonalité et indépendance

Deux contrastes

$$\Psi^1 = a_1^1\tau_1 + a_2^1\tau_2 + \ldots + a_k^1\tau_k, \ \sum a_i^1 = 0$$

$$\Psi^2 = a_1^2\tau_1 + a_2^2\tau_2 + \ldots + a_k^2\tau_k, \ \sum a_i^2 = 0$$

sont orthogonaux (dans le cas équilibré) si

$$\sum a_i^1 a_i^2 = 0.$$

Par exemple les contrastes

$$[2, -1, -1] \text{ et } [0, 1, -1]$$

sont orthogonaux. L'orthogonalité est une façon élégante de dire que les SCE associées à ces contrastes (ou encore les variances de ces contrastes) sont indépendantes, en d'autres termes que les informations apportées par un contraste sont indépendantes des informations apportées par l'autre. En choisissant des contrastes indépendants, on peut décomposer la SCE des traitements en $SCE_{contrastes}$ et les tester de façons complètement indépendantes. En étant astucieux, on peut notamment chercher dans la réponse des traitements des effets linéaires, quadratiques, cubiques ...
Très souvent, on veut être capable de construire des "groupes homogènes" c'est à dire des groupes pour lesquels les effets du facteur sont du même ordre de grandeur. Certaines techniques sont tout spécialement réservées à certaines comparaisons. Rappelons que l'hypothèse fondamentale sur laquelle repose ces tests est l'hypothèse d'égalité des variances des populations dont sont issues les moyennes à comparer. Nous noterons :$\hat{\sigma}^2$ un estimateur sans biais de cette variance, et nous supposerons que cette variance est estimée avec k degrés de liberté.

5.7.3 Plus petite différence significative (PPDS)

Dans cette méthode, une succession de tests de Student est réalisée pour constituer des groupes homogènes. Supposons que p moyennes $(m_1, m_2, ..., m_p)$ soient à comparer, que ces p moyennes soient respectivement estimées par:$\bar{X}_1, \bar{X}_2, ...\bar{X}_p$, et que ces moyennes soient estimées sur des échantillons de tailles respectives $n_1, n_2, .., n_p$. En comparant les moyennes deux à deux, il faut faire $\frac{p(p-1)}{2}$ comparaisons.
Chaque comparaison de 2 moyennes est effectuée en utilisant la règle de

décision suivante: si

$$\frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\hat{\sigma}^2(1/n_i + 1/n_j)}} > t_k^{1-\alpha/2} \tag{4.1}$$

alors, on rejette l'hypothèse $H_0 : m_i = m_j$.
Remarquons que si les effectifs des échantillons sont égaux, (en d'autres termes si $n_1 = n_2 = .. = n_p = n$ la règle de décision (4.1) peut se réécrire:

$$\frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{2\hat{\sigma}^2}{n}}} > t_{1-\alpha/2}^k$$

ou encore, on rejette l'hypothèse H_0 si

$$|\bar{X}_i - \bar{X}_j| > t_{1-\alpha/2}^k \sqrt{\frac{2\hat{\sigma}^2}{n}}$$

Si une analyse de variance a au préalable été effectuée, on dispose d'une estimation sans biais de la variance: elle est donnée par la variance résiduelle. Prenons un exemple pour illustrer cette méthode. On veut comparer 5 moyennes m_1, m_2, m_3, m_4, m_5. Les estimations respectives de ces moyennes (obtenues sur des échantillons de taille $n = 7$ sont: $\bar{X}_1 = 8.2$, $\bar{X}_2 = 10.34$, $\bar{X}_3 = 7.53$, $\bar{X}_4 = 9.64$, $\bar{X}_5 = 7.49$
La variance de population est estimée à l'aide d'une analyse de variance avec 30 degrés de liberté, l'estimation est:$\hat{\sigma}^2$=0.4683
Chaque différence devra donc être comparée à

$$t_{1-\alpha/2}^k \sqrt{\frac{2\hat{\sigma}^2}{n}} = 2.042\sqrt{\frac{2(0.4683)}{7}} = 0.75$$

Pour être sûr de ne pas oublier de comparaison, il est d'usage de construire le tableau des différences entre moyennes (classées) qui, sur notre exemple donne:

On en conclut que: On en conclut que les moyennes m_1, m_3 et m_5 ne peuvent pas être considérées comme différentes, la même conclusion peut être tirée pour les moyennes m_2, m_4.
IMPORTANT

	$\bar{X}_3$	$\bar{X}_1$	$\bar{X}_4$	$\bar{X}_2$
	7.53	8.2	9.64	10.34
$\bar{X}_5 = 7.49$	0.04	0.71	2.15	2.85
$\bar{X}_3 = 7.53$		0.67	2.11	2.81
$\bar{X}_1 = 8.2$			1.44	2.14
$\bar{X}_4 = 9.64$				0.7

$\bar{X}_5$ $\bar{X}_3$ $\bar{X}_1$ $\bar{X}_4$ $\bar{X}_2$

Cette méthode est de moins en moins utilisée car le risque global de première espèce pris en affirmant une telle décomposition en groupes n'est pas égal à 5% (il est de l'ordre de 40%). Ceci provient du fait qu'une succession de tests de risque α ne permet pas de prendre une décision **globale** avec ce même risque α.[5].

5.7.4 Méthode de Bonferroni

Comme nous venons de le voir dans le paragraphe précédent, il est possible de contrôler le risque de première espèce pour le test de n'importe quel contraste. Mais qu'arrive t-il lorsque l'on multiplie les tests ? Si deux comparaisons sont réalisées avec un risque de première espèce de α, il est faux de penser que la décision globale peut être prise avec un risque α. Le risque que vous prenez dans la décision globale est difficile à calculer, en revanche, Bonferroni a proposé une majoration de ce risque. La méthode de Bonferroni est une méthode *a maxima*: elle ne permet pas un strict contrôle de α, mais en revanche elle en donne une majoration (qui peut être énorme). L'idée de Bonferroni est de se placer dans "le pire des cas" (pour α) .
Supposons que p moyennes doivent être comparées avec un risque global α. En utilisant des comparaisons deux à deux, $r = \frac{p(p-1)}{2}$ comparaisons

[5]On dit dans ce cas la que le test n'est pas **conservatif**

sont nécessaires. Par exemple, si $p = 5$, il faut effectuer $\frac{5\times4}{2} = 10 = r$ comparaisons. Pour avoir un risque **global** α, il faut que chacune des r comparaisons soit effectuée avec un risque α'.
Le calcul de α' peut-être fait selon 2 méthodes selon que les comparaisons sont indépendantes (orthogonales) ou pas (qui conduisent à des résultats sensiblement identiques quand α est petit).
1)Si les comparaisons sont indépendantes, alors $\alpha' = 1 - (1 - \alpha)^{\frac{1}{r}}$
2)Si les comparaisons sont dépendantes (ou indépendantes) $\alpha' = \frac{\alpha}{r}$
On applique alors la méthode de la PPDS en utilisant cette fois, $t_k^{1-\alpha'/2}$ (k est le degré de liberté avec lequel la variance est estimée).

5.7.5 Méthode de Newman-Keuls

La méthode de Newman-Keuls (NK) est basée sur la comparaison des amplitudes observées pour des groupes de 2,3,...,p moyennes avec l'amplitude maximum attendue à un niveau de signification donnée. Pour effectuer ces comparaisons, on doit d'abord calculer la **plus petite amplitude significative** relative à des groupes de 2,3,...,p moyennes.
Ce calcul nécessite l'utilisation de tables particulières (Tables de NK données en annexe) à 3 entrées comportant:
1) risque globale de première espèce α
2) le nombre de degrés de liberté (k) avec lesquels est estimée la variance de population
3) le nombre de moyennes à comparer (i)
La table fournit alors la valeur $q_{1-\alpha}^{i,k}$
Chaque amplitude est alors comparée à $q_{1-\alpha}^{i,k}\sqrt{\frac{\hat{\sigma}^2}{n}}$
Un exemple illustrera le principe de cette méthode.
Reprenons l'exemple précédent avec exactement les mêmes données. Les plus petites amplitudes significatives sont au niveau $\alpha = 5\%$ pour $k = 30$ degrés de liberté:
Rangeons dans un premier temps les moyennes:

$$\bar{X}_5 \leq \bar{X}_3 \leq \quad \bar{X}_1 \leq \quad \bar{X}_4 \leq \bar{X}_2$$

$$\begin{array}{ll}
\text{pour 2 moyennes} & q_{0,95}^{2,30}\sqrt{\frac{\hat{\sigma}^2}{n}} = 2.89\sqrt{\frac{0.4683}{7}} = 0.75 \\
\text{pour 3 moyennes} & q_{0,95}^{3,30}\sqrt{\frac{\hat{\sigma}^2}{n}} = 3.49\sqrt{\frac{0.4683}{7}} = 0.90 \\
\text{pour 4 moyennes} & q_{0,95}^{4,30}\sqrt{\frac{\hat{\sigma}^2}{n}} = 3.85\sqrt{\frac{0.4683}{7}} = 1.00 \\
\text{pour 5 moyennes} & q_{0,95}^{5,30}\sqrt{\frac{\hat{\sigma}^2}{n}} = 4.10\sqrt{\frac{0.4683}{7}} = 1.06
\end{array}$$

L'amplitude calculée sur les 5 moyennes vaut:
$\bar{X}_2 - \bar{X}_5 = 10.34 - 7.49 = 2.85 > 1.06$
L'hypothèse $H_0 : m_1 = m_2 = m_3 = m_4 = m_5$ n'est donc pas être acceptée.
Passons alors, aux calculs des amplitudes sur 4 moyennes: $\bar{X}_4 - \bar{X}_5 = 9.64 - 7.49 = 2.15 > 1.00$
$\bar{X}_2 - \bar{X}_3 = 10.34 - 7.53 = 2.81 > 1.00$
Les hypothèses $H_0 : m_1 = m_3 = m_4 = m_5$ et $H_0 : m_1 = m_2 = m_3 = m_4$ sont donc rejetées, il faut passer aux calcul des amplitudes sur 3 moyennes:
$\bar{X}_1 - \bar{X}_5 = 8.20 - 7.49 = 0.71 < 0.90$
$\bar{X}_4 - \bar{X}_3 = 9.64 - 7.53 = 2.11 > 0.90$
$\bar{X}_2 - \bar{X}_1 = 10.34 - 8.27 = 2.14 > 0.90$
L'hypothèse $H_0 : m_1 = m_3 = m_5$ ne peut pas être rejetée, en revanche les hypothèses $H_0 : m_1 = m_3 = m_4$ et $H_0 : m_1 = m_2 = m_4$ sont rejetées. Il est inutile de tester de calculer les amplitudes sur 2 moyennes dans le groupe qui n'a pas été déclaré hétérogène (qui peut le plus peut le moins).
Il ne reste donc plus que deux amplitudes sur 2 moyennes à calculer : $\bar{X}_4 - \bar{X}_1 = 9.64 - 8.2 = 2.11 > 0.85$
$\bar{X}_2 - \bar{X}_4 = 10.34 - 9.64 = 0.7 < 0.85$
L'hypothèse $H_0 : m_1 = m_4$ est donc refusée et l'hypothèse $H_0 : m_2 = m_4$ ne peut pas être refusée.
On obtient *in fine*: On en conclut que: ce qui dans ce cas particulier donne exactement le même résultat que la méthode de la $PPDS$ avec, ici, moins de doute quant à la valeur effective du risque de première espèce α. [6]

[6]Dans certains cas, on observe des chevauchements entre les groupes ce qui complique un peu l'interprétation.

$\bar{X}_5 \quad \bar{X}_3 \quad \bar{X}_1 \quad \bar{X}_4 \quad \bar{X}_2$

5.7.6 Méthode de Duncan

Le principe de la méthode de Duncan est en tout point similaire à celle de NK, seule la valeur $q_{1-\alpha}^{i,k}$ est différente (inférieure à celle de NK). Ainsi, cette méthode est caractérisée par des risques de première et de seconde espèce respectivement supérieur et inférieur à la méthode de NK.
Il en résulte que les résultats déduits de Duncan sont dans l'ensemble plus proches (que ceux de NK) des résultats de la PPDS.

5.7.7 Méthode de Tuckey

Tuckey dans le but de bien contrôler le risque de première espèce, a suggéré de prendre comme valeur de $q_{1-\alpha}^{i,k}$, une valeur indépendante de i (nombre de moyennes sur lesquelles on calcule l'amplitude). Pour être sûr de bien contrôler α, Tuckey a proposé de prendre la valeur maximale utilisée par NK soit $q_{1-\alpha}^{p,k}$ (ou p est le nombre total de moyennes à comparer.)
Cette technique permet en effet de bien contrôler α, mais elle a des conséquences fâcheuses sur le risque le seconde espèce.
Dans certains cas, on ne s'intéresse qu'à la comparaison de p moyennes à un témoin. C'est l'objet de la méthode de Dunnett.

5.7.8 Méthode de Dunnett

La méthode ressemble à celle de la PPDS et à NK, mais comme il n'y a que p comparaisons à effectuer, des tables spéciales (celles de Dunnett) ont été conçues spécialement à cet effet.
Voyons sur notre exemple l'utilisation de la méthode.
Supposons que le traitement de référence soit le traitement numéro 1 de moyenne $\bar{X}_1 = 8.2$

Quatre comparaisons avec le témoin sont à considérer en voici la liste:
$\bar{X}_1 - \bar{X}_5 = 0.71$
$\bar{X}_1 - \bar{X}_3 = 0.67$
$\bar{X}_4 - \bar{X}_1 = 1.44$
$\bar{X}_2 - \bar{X}_1 = 2.14$
Il reste maintenant à définir la valeur à laquelle il faut comparer ces différences. La forme de cette valeur est de la même forme que celle que nous avons utilisé pour la PPDS soit :

$$d^k_{1-\alpha/2}\sqrt{\frac{2\hat{\sigma}^2}{n}} = 2.58\sqrt{\frac{2(0.4683)}{n}} = 0.9437$$

La quantité $d^k_{1-\alpha/2}$ est trouvée dans une table de Dunnett.
On conclue donc (avec un risque $\alpha = 5\%$) que les traitements 5 et 3 ne sont pas significativement différents du traitement 1, et que les traitements 4 et 2 sont significativement différents du traitement de référence.

5.8 Quelques tests non parametriques

On qualifie de non paramétriques, les méthodes applicables, quelque soit la distribution de la population. L'expression anglaise "distribution free" dit bien mieux que "non paramétrique", ce dont il s'agit.
Aucune hypothèse n'est donc faite sur la distribution, il ne faut pas en conclure pour autant que les méthodes non paramétriques peuvent s'utiliser sans aucune hypothèses.
Pour tous les tests que nous allons voir, il faut que les variables étudiées soient **continues** et, dans certains cas, indépendantes (nous le préciserons le temps venu);
Une autre caractéristique essentielle des tests non paramétriques, est leur **faible puissance** pour les petits effectifs, par rapport à leurs analogues paramétriques. Aussi, nous ne conseillons d'utiliser ces méthodes, que lorsque les hypothèses des tests paramétriques sont violées.

5.8.1 Tests sur échantillons appariés

Le test du signe

Il est relatif au cas de deux échantillons appariés.
Il est uniquement basé sur le signe des différences observées entre les paires.
L'hypothèse nulle est :

$$H_0 : P(+) = P(-) = \frac{1}{2}$$

où $P(+)$ est la probabilité d'observer une différence positive et $P(-)$ est la probabilité d'observer une différence négative.
Lorsque l'hypothèse nulle est vraie, le nombre de différences positives[7] est une variable binomiale de paramètres n (nombre de paires) et 1/2.
Si x est le nombre de différences positives observées, il est assez facile de calculer la proba pour que le nombre de différences positives soit inférieur ou égal à celui que nous avons observé en calculant:

$$P(X \leq x) = (1/2)^n \sum_{i=0}^{x} C_n^i$$

Pour un test bilatéral, on rejette l'hypothèse nulle avec un risque α si:

$$P(X \leq x) \leq \frac{\alpha}{2}$$

Pour des échantillons de taille élevée, on peut utiliser l'approximation:

$$u_{obs} = \frac{|x - n/2| - 1/2}{\sqrt{n/4}}$$

et on rejette l'hypothèse nulle avec un risque de première espèce α si

$$u_{obs} \geq u_{1-\alpha/2}$$

où $u_{1-\alpha/2}$ est la valeur limite au seuil $1 - \alpha/2$ d'une loi $\mathbf{N}(0, 1)$.
Quand certaines différences sont nulles, les paires d'observations correspondantes sont éliminées du test, la valeur de n étant par conséquent réduite.

[7] le nombre de différences négatives pourrait aussi être utilisé.

Le test des rangs appliqué au cas des échantillons appariés.
Il est aussi appelé test de Wilcoxon, il tient compte non seulement du signe des différences, mais aussi de leur rang. La réalisation du test nécessite le calcul des différences observées entre paires d'individus,la détermination du rang de ces différences **en faisant abstraction du signe**, et le calcul de la somme des rangs des différences positives (Y_+) et celui de la somme des rangs des différences négatives Y_-.
L'hypothèse testée est ici comme pour le test des signes:

$$H_0 : P(+) = P(-) = 1/2$$

On rejette cette hypothèse si la plus petite des quantités (Y_+) et (Y_-) est supérieure à la valeur trouvée dans la table de Wilcoxon.
Quand n (le nombre de paires) est assez grand (supérieur à 30) on peut calculer:

$$u_{obs} = \frac{|Y_+ - n(n+1)/4|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

et on rejette l'hypothèse nulle avec un risque de première espèce α si

$$u_{obs} \geq u_{1-\alpha/2}$$

où $u_{1-\alpha/2}$ est la valeur limite au seuil $1 - \alpha/2$ d'une loi $\mathbf{N}(0, 1)$.
Quand certaines différences sont nulles, les paires d'observations correspondantes sont éliminées du test, la valeur de n étant par conséquent réduite.

5.8.2 Tests sur échantillons indépendants

Test de Mann-Withney
La réalisation du test est basée sur le classement de l'ensemble des observations par ordre croissant, la détermination du rang de chacune d'elles, et le calcul de la somme des rangs U relative à l'échantillon qui comporte le plus petit nombre d'observations.
Supposons que cet échantillon soit d'effectif m, et soit n l'effectif de l'autre

échantillon, alors on rejette l'hypothèse nulle H_0 :*les distributions sont égales* avec un risque de première espèce α si

$$U \geq MW_{1-\alpha/2} \quad ou \quad si \quad U \leq MW_{\alpha/2}$$

où $MW_{1-\alpha/2}$ et $MW_{\alpha/2}$ sont les valeurs lues dans la table de Mann-Withney pour m et n fixés.
Quand $n+m$ est assez grand (supérieur à 30) on calcule

$$u_{obs} = \frac{|U - m(m+n+1)/2|}{\sqrt{\frac{nm(n+m+1)}{12}}}$$

et on rejette l'hypothèse nulle avec un risque de première espèce α si

$$u_{obs} \geq u_{1-\alpha/2}$$

où $u_{1-\alpha/2}$ est la valeur limite au seuil $1-\alpha/2$ d'une loi $\mathbf{N}(0,1)$.
Test de Kruskal-Wallis
L'application du test des rangs a été étendue au cas de plusieurs échantillons indépendants par Kruskal et Wallis. Comme pour deux échantillons, la réalisation du test est basée sur le classement de l'ensemble des observations par ordre croissant, la détermination du rang de chacune d'elle et le calcul des sommes des rangs Y_i relatives aux différents échantillons.
A partir de ces sommes, on obtient la valeur:

$$\chi^2_{obs} = \frac{12}{n(n+1)} \sum_{i=1}^{p} \frac{Y_i^2}{n_i} - 3(n+1)$$

où n_i est la taille de l'échantillon i, p est le nombre d'échantillons à comparer et $n = \sum_{i=1}^{p} n_i$.
On rejette l'hypothèse nulle d'égalité des distributions avec un risque de première espèce α si:

$$\chi^2_{obs} \geq \chi^2_{1-\alpha},$$

où $\chi^2_{1-\alpha}$ est la valeur limite au seuil $1-\alpha$ d'une loi du χ^2 à $p-1$ degrés de liberté.

Ce test est asymptotique, et l'approximation est "satisfaisante" quand n est assez grand.
Pour les petites valeurs de n ($p < 4, n_i \leq 5$), on utilise les tables de Kruskal-Wallis.