République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abbes Laghrour Khenchela

Faculté des Sciences de la Nature et de la Vie

Département des sciences agronomiques

Polycopie du Cours

Destiné aux étudiants en M1 :

* Biotechnologie Végétale

* Production Végétale

BIOSTATISTIQUE

Préparé par :

Dr. ADDAD Dalila

Maitre de conférences classe B

Université de Khenchela, 2020/2021

Avant-Propos

L’étape de l'analyse des résultats est souvent vécue comme une contrainte, un passage obligé mais désagréable, voire même parfois un calvaire. Pourtant, le premier objectif des statistiques est bien de révéler ce que les données ont à nous dire. Passer à côté d'une bonne analyse par manque de temps, de motivation ou de compétence, c'est surtout prendre le risque de rater un phénomène intéressant qui était pourtant là, sous nos yeux.

L’objectif assigné à ce cours est la formation des étudiants de master en biotechnologie et production végétale aux traitements des données liées à leurs thématiques de travail via les biostatistiques. Ce document permet à l’étudiant de voir différents exemples d’application de la biostatistique dans les sciences expérimentales, et lui permettre de passer du stade d’observation vers le stade de description et de calculs statistiques.

Ce polycopié de cours est destiné également à toute personne souhaitant connaître et surtout utiliser les principales méthodes de la statistique.

Vue l’importance accordée à la biostatistique dans l’université de khenchela, ce module est enseigné presque à tous les niveaux (Mathématique et statistique en première année et Bistatistique en deuxième année et troisième année licence et en Master 1)

Sommaire

Liste des tableaux

Liste des figures

Introduction générale

La biostatistique est l'application des statistiques en biologie ; sachant que, la statistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation de phénomènes aléatoires, c'est-à-dire dans lesquels le hasard intervient. Dans l'histoire de la statistique, la discipline biologique a souvent été à l'origine d'idées nouvelles importantes. La biostatistique nous permet de d'écrire une population donnée, selon ses attributs et ses qualités, de mesurer la précision d'une estimation ou de définir le degré d'association entre une série de caractères et d'évènements.

Les méthodes statistiques permettent d'éprouver la validité des résultats, en fonction même de leur variabilité avec la plus grande rigueur scientifique. Elles permettent une analyse base de toute interprétation.

La bistatistique englobe :

- La conception d'expériences biologiques ;
- La collecte d'informations ;
- L'analyse des données chiffrées ;
- L'interprétation des résultats et conclusion.

Les biostatistiques permettent de confirmer ou d'infirmer une hypothèse avec une marge d'erreur la plus petite possible, et/ou prédire un évènement à l'aide d'outils. Il existe deux types de statistiques:

• Les statistiques descriptives, permettant de décrire une série de données

• Les statistiques inférentielles, consistant en des tests permettant de confirmer ou infirmer une hypothèse.

Le polycopie est structuré en onze chapitres, dont le premier aborde **un rappel sur les principales lois de probabilité :** loi Normale, loi Normale centrée réduite, loi de Student et loi de Khi deux, le deuxième chapitre est consacré aux différents types de **représentations graphiques,** histogramme, diagramme, boite à moustaches, le troisième chapitre donne idée sur **les tests classiques paramétriques et non paramétriques** surtout de point de vu définition et différences, puis le quatrième chapitre dans lequel on va voir quelques **tests de conformité**, alors que le chapitre cinq est consacré à **la comparaison de moyennes** que ce soit des échantillons indépendants ou des échantillons appariés. Dans les chapitres six et sept, on va toucher respectivement **les tests d'indépendance**, où on va prendre comme exemple le test de Khi-deux d'indépendance, et **la comparaison de distribution** puis on va prendre les deux types de **régression linéaire simple et multiple** qu'est une méthode statistique de modélisation des relations entre différentes variables dépendantes et indépendantes, par la suite on va passer à **l'analyse de variance à un et deux facteurs** tout en passant par la comparaison multiples des moyennes et finalement on va voir **un rudiments d'analyses factorielles : AFC, ACP, CAH et cladistique.**

1. Rappel des principales lois de probabilité : loi normale, t, F, Khi deux

1.1.Loi normale ou de Laplace-Gauss $N(\mu, \sigma)$

On parle de **loi normale** lorsque l'on a affaire à une variable aléatoire continue dépendant d'un grand nombre de causes indépendantes dont les effets s'additionnent et dont aucune n'est prépondérante. Cette loi acquiert sa forme définitive avec **Gauss** (en 1809) et **Laplace** (en 1812). C'est pourquoi elle porte également les noms de : **loi de Laplace, loi de Gauss et loi de Laplace-Gauss.**

Exemple : la taille d'un animal dépend des facteurs environnementaux (disponibilité pour la nourriture, climat, prédation, etc.) et génétiques. Dans la mesure où ces facteurs sont indépendants et qu'aucun n'est prépondérant, on peut supposer que la taille suit une loi normale.

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. La loi normale est caractérisée par sa densité de probabilité. Pour une loi normale de moyenne *m* et de variance$\boldsymbol{\sigma^2}$, elle est donnée par la formule

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

La courbe représentative de la densité a la forme d'une courbe en cloche symétrique

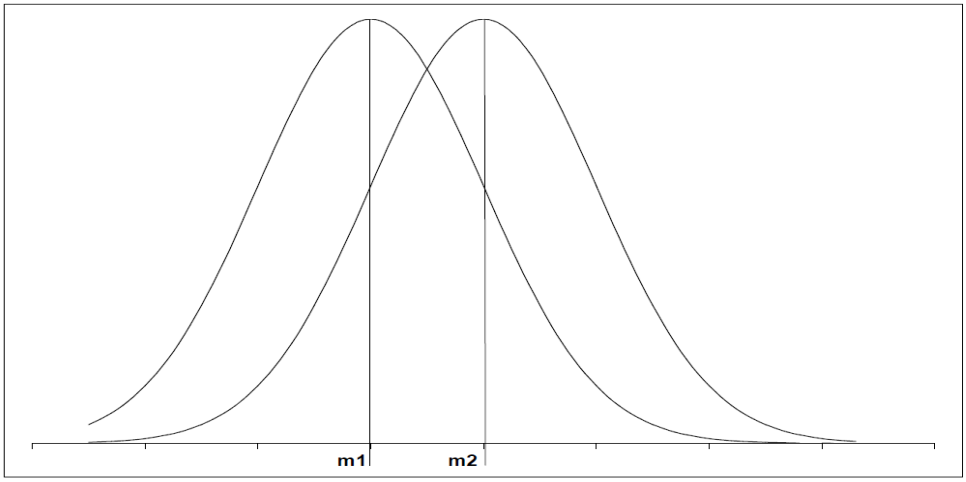


Figure 1. Exemple de deux lois normales

Les deux lois ont la même variance. La moyenne *m*1 de la première loi est inferieure à celle *m*2 de la seconde.

Les propriétés de la densité

- ✓ Elle varie de $-\infty$ *à* $+\infty$

- ✓ Elle est symétrique par rapport à la valeur Moyenne μ.
- ✓ Elle représente deux points d'inflexion en $\mu - \delta \; et \; \mu + \delta$

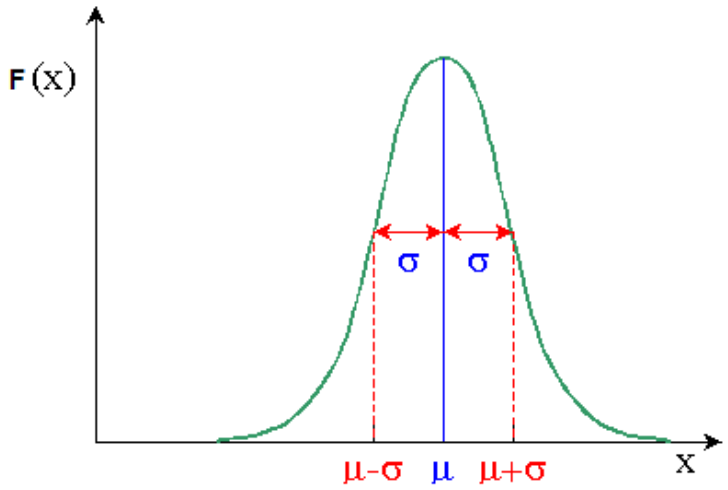


Figure 2. Propriété symétrique de la loi normale

La fonction f est **paire** autour d'un axe de symétrie $x = m$ car $f(x + m) = f(m - x)$

Remarque : Le paramètre m ou $\bar{x}$ représente **l'axe de symétrie** et s **le degré d'aplatissement** de la courbe de la loi normale dont la forme est celle d'une courbe en cloche.

Le paramètre d'asymétrie $G_1 = \frac{\mu_3}{S^3} \; et \; \mu_3 = \frac{1}{n} \sum_{i=1}^{n} x_i^3 - \bar{x}^3$

$G_1 = 0 \Rightarrow la \; distribution \; est \; symétrique \; au \; tour \; de \; la \; moyenne$

Le paramètre d'aplatissement $G_2 = \frac{\mu_4}{S^4} - 3 \; et \; \mu_4 = \frac{1}{n} \sum_{i=1}^{n} x_i^4 - \bar{x}^4$

$G_2 = 0 \Rightarrow la \; distribution \; est \; normale \; ou \; gaussienne$

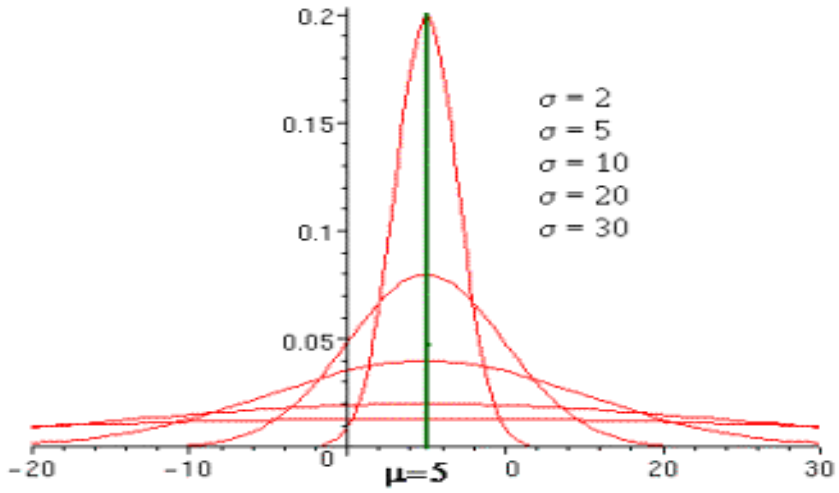


Figure 3. Exemples des lois normales avec la même moyenne (μ=5) et des écarts types σ croissants

Géométriquement une probabilité peut s'interpréter comme la surface sous la courbe densité comme l'indique le graphique.

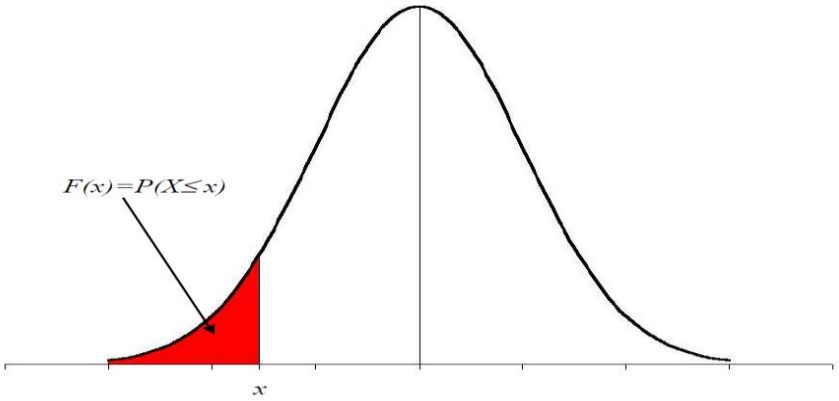
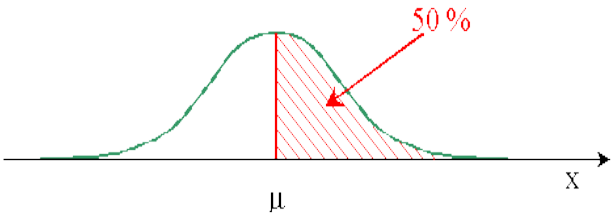


Figure 4. Interprétation géométrique de la probabilité

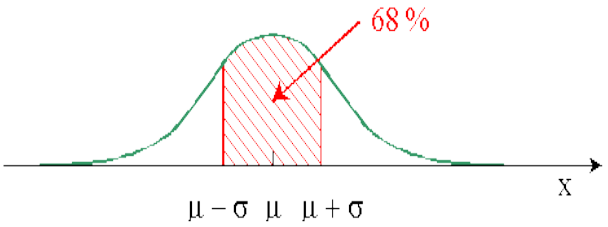
Une probabilité s'interprète comme la surface sous la courbe représentant la densité.

Lorsque la distribution des individus dans une population obéit à la loi normale, on trouve:

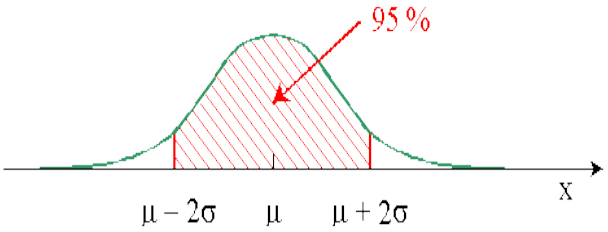
A. 50 % des individus en-dessous de la moyenne μ et 50 % au-dessus (la loi normale est symétrique)



B. 68 % des individus entre $\mu-\sigma$ et $\mu+\sigma$



C. 95 % des individus entre $\mu-1{,}96\sigma$ et $\mu+1{,}96\sigma$, que nous arrondirons à l'intervalle $[\mu-2\sigma, \mu+2\sigma]$



D. 99,7 % des individus entre $\mu-3\sigma$ et $\mu+3\sigma$ (il y a donc très peu de chances qu'un individu s'écarte de la moyenne de plus de 3σ)

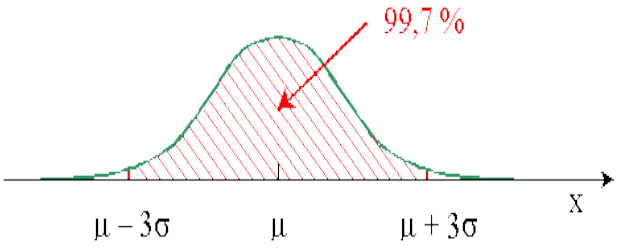


Figure 5. pourcentages relatifs de la loi normale

1.2. Cas particulier : La loi normale réduite N(0,1)

Une variable aléatoire continue *X* suit une **loi normale réduite** si sa densité de probabilité est donnée par :

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

L'axe de symétrie correspond à l'axe des ordonnées (x=0) et le degré d'aplatissement de la courbe de la loi normale réduite est 1.

Données centrées réduites:

• Centrage: $x_i - \bar{x}$ • Réduction: diviser par S

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

Remarques :

- Lorsque l'on suppose qu'une variable X suit le modèle de la loi normale N ($\bar{X}$, σ), on écrit $X \sim (\mu, \sigma)$
- Dans la loi normale centrée réduite on note N (0, 1).
- Dans N (0, 1) si on cherche population ou pourcentage P (X ≤ α) (rappel : on écrit aussi F(α)), on cherche la valeur de α dans le tableau
- P (X ≥ α) = 1 − P (X ≤ α) = 1 − F (α)
- P (X ≤ − α) = P (X ≥ α)

Exercice

1. Soit une variable aléatoire de loi N (2 , 1,4). Calculez : P(X≤2.3)

2. X ~ N (27, 1,9): Calculez P (X ≤ 30,5)

3. Quelle est la valeur de a telle que P(X>a)=0,6517 pour une loi centrée réduite, puis déduire le quantile, supposant que la variable étudiée suit une loi normale

Solution

1. Nous avons X ~ N (2 , 1,4), On centre et réduit X : $\frac{X-2}{1,4}$~ N (0,1)

P (**X** ≤ 2,3) = P ($\frac{X-2}{1,4} \leq \frac{2,3-2}{1,4}$)

P (**Z** ≤ 0,214)= 0,5823 (58,23%) (Lecture sur table)

2. Nous avons X ~ N (27 , 1,9), On centre et réduit X : $\frac{X-27}{1,9}$~ N (0,1)

$$P(\mathbf{X} \leq 30{,}5) = P\left(\frac{X-27}{1{,}9} \leq \frac{30{,}5-27}{1{,}9}\right)$$

$$P(\mathbf{Z} \leq 1{,}842) = 0{,}9671\ (96{,}71\%)\ \text{(Lecture sur table)}$$

3. On cherche le quantile à 65 % pour la N (0,1)

Cela revient à trouver à tel que P ($\mathbf{z} \leq a$) = 0,6517.On lit la table **à l'envers :**

Donc P($X \leq 0{,}39$) = 0,6517 ➔Le quantile recherché est donc 0,39.$\mathbf{Z_{0,6985} = 0,39.}$

La valeur réelle du quantile dans une loi N (12,5. 2,4) est :

$$Q_{0,6517} = \bar{X} + \sigma \times Z_{0,6517} \rightarrow Q_{0,6517} = 12{,}5 + (2.4 * 0{,}52)$$
$$\mathbf{Q_{0,6985} = 13,75}$$

Exemple : Dans un lac, on a trouvé que la longueur des truites d'un an est distribué à peu près normalement autour d'une moyenne μ=15 cm avec un écart-type de 2.1 cm.

Quelle est la proportion de ces truites qui :

a. excèdent 20 cm?

b. n'excèdent pas 17.5 cm?

1.3. Loi F de Fisher et sa distribution

La **loi de Fisher** est utilisée pour comparer deux variances observées et sert surtout dans les très nombreux tests d'analyse de variance et de covariance.

Si X et Y sont deux variables aléatoires indépendantes de lois respectives χ^2_{n1} et χ^2_{n2}, alors la variable aléatoire Z suivra une loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté.

$$Z = \frac{X/n_1}{Y/n_2}$$

Si X est une variable aléatoire de loi de Fisher à n_1 et n_2 degrés de liberté, on a :

$$E(X) = \frac{n_2}{n_2 - 2} \quad \text{avec} \quad n_2 > 2$$

$$Var(X) = \frac{2n_1^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2} \quad \text{avec} \quad n_2 > 4$$

Si X est une variable aléatoire de Fisher à n_1 et n_2 degrés de liberté, alors la variable aléatoire Y=(1/X) suit une loi de Fisher à n_2 et n_1 degrés de liberté.

Dans le cas où des échantillons de taille n_1 et n_2 sont tirés au hasard d'une population dans la distribution est normale, le rapport de variance de ces échantillons suit la distribution du **F** = $\frac{S_1^2}{S_2^2}$ avec n-1 = ddl.

Cette distribution du F a les propriétés suivantes :

* Elle est non symétrique.
* Elle est continue.
* Sa limite inferieur à 0.
* La forme de la distribution est déterminée par les ddl des numérateurs et des dénominateurs liés au rapport des variances utilisées.

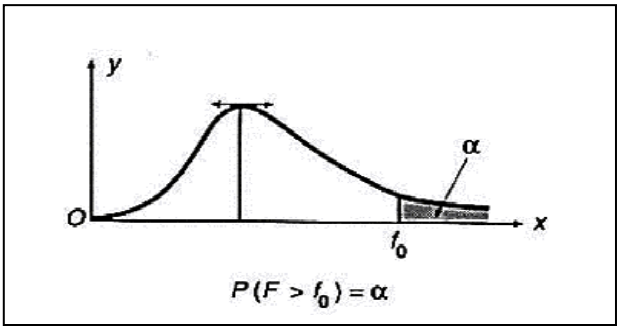


Figure 6. Représentation graphique de la loi F

Le test F est utilisé pour tester si deux variances de deux échantillons indépendants viennent d'une même population de variance ou variance commune σ^2. ce test s'exprime comme suit :

$$F = \frac{S_G^2 \ plus \ grande}{S_P^2 plus \ petit}$$

Exemple

$S_1^2 = 40.2 \ ; \ S_2^2 = 6.97$

H_0 $\rightarrow S_1^2 = S_2^2$

H_1 $\rightarrow S_1^2 \neq S_2^2 \rightarrow S_1^2 > S_2^2$

$F = \frac{S_G^2 \ plus \ grande}{S_P^2 plus \ petit} = \frac{40.2}{6.97} = 5.76$ $F_{(5\%,6;6)} = 4.28$

$F_{obs} > F_{tab}$ ⇒on rejette

$$H_0 \ et \ on \ accepte \ H_1 \ donc \ les \ deux \ variance \ sont \ inégales$$

1.4. Loi du χ^2 de Pearson Khi deux

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

Soit X_1, X_2, …, X_i,…, X_n, n variables **normales centrées réduites**, on appelle

c^2 la variable aléatoire définie par :

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

On dit que c^2 suit une **loi de Pearson** à *n* **degrés de liberté** (d.d.l.).

On peut remarquer qu'une variable 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. la fonction densité de probabilité est de la forme :

$$f(\chi^2) = C(n)\,\chi^{2\frac{n}{2}-1}\,e^{-\frac{\chi^2}{2}} \quad \text{AVEC} \quad C(1) = \frac{1}{\sqrt{2\pi}}$$

Pour *n* > 1, on utilise **la table du Khi 2**

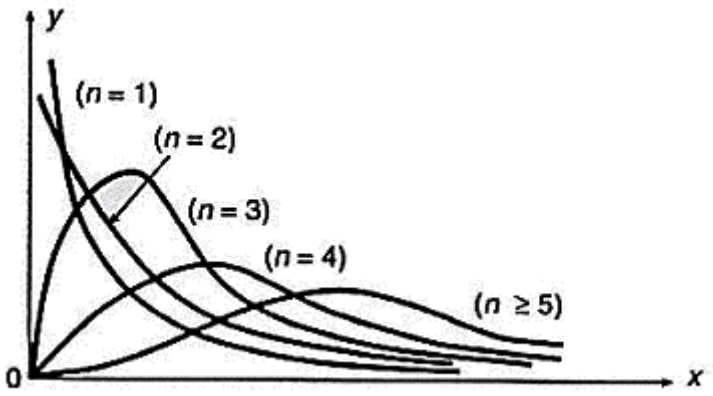


Figure 7. Densité de fonction de la loi khi-deux

Remarque : La distribution du χ^2 est asymétrique et tend à devenir symétrique lorsque *n* augmente en se rapprochant de la **distribution normale** N(μ=n, σ =$\sqrt{2 * n}$) à laquelle elle peut être assimilée lorsque *n* > 30.

Le test χ^2 est un test qui répond à la question existe-il une association entre les variables portées sur les colonnes et celles portées sur les rangs ? ou il ya indépendance entre ces variables ou non ?

Le test χ^2 est exprimé par la formule :

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Avec :

O : valeurs observées, E : valeurs attendues

Le test χ^2 a plusieurs applications

• Comparer plusieurs groupes indépendants décrits par une variable qualitative: équivalent qualitatif de l'ANOVA.

• Mesurer la liaison entre deux variables qualitatives: équivalent qualitatif de la corrélation.

• Estimer la conformité entre une distribution observée et une distribution théorique.

Le test χ^2 teste deux hypothèses :

• H_0: désigne que les groupes constituent un groupe homogène. Ils peuvent provenir de la même population statistique.

• H_1: désigne que les groupes ne constituent pas un ensemble homogène.

Exemple

Des arbres de pin blanc ont été classés selon leurs âges et la réaction à un champignon causant la rouille, on vous demande de tester l'association entre l'âge et la réaction au champignon ?

Age / Réaction	4	10	20	40	Total
Résistant	7	6	11	15	39
sensible	14	11	5	8	38
total	21	17	16	23	77

• $H_0 \Rightarrow$ l'indépendance entre l'âge et la réaction à un champignon causant la rouille

• $H_1 \Rightarrow$ la réaction à un champignon causant la rouille est dépendante de l'âge des arbres.

Solution

1- Le tableau donné est de type tableau de contingence, contient les fréquences absolues observées a_{ij} :

Tableau 1. Table de contingence

États de classement	*Groupes*				$\sum$
	1	2	(j)	g	
1	a_{11}	a_{12}	…	a_{1g}	m_1
2	a_{21}	a_{22}	…	a_{2g}	m_2
3	a_{31}	a_{32}	…	a_{3g}	m_3
(i)	:	:	a_{ij}	:	(m_i)
k	a_{k1}	a_{k2}	…	a_{kg}	m_k
Σ	n_1	n_2	(n_j)	n_g	n

Évidemment, $n = \Sigma\, a_{ij} = \Sigma\, n_j = \Sigma\, m_i$.

La valeur attendue dans chaque case, sous l'hypothèse nulle, se calcule par

$$n_j \times m_i / n$$

Exemple : E_i = 21x39/77=10.6

réaction	âge	O_i	E_i	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Résistant	1	7	21x39/77=10.6	12.96	1.22
	2	6	8.6	6.76	0.79
	3	11	8.1	8.41	1.04
	4	15	11.6	11.56	0.99
sensible	1	14	10.4	12.96	1.25
	2	11	8.4	6.76	0.80
	3	5	7.9	8.41	1.06
	4	8	11.4	11.56	1.01

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 8.16$$

$x_{obs}^2 = 8.16$

$x_{tab}^2 = 7.81\ pour\ 5\%\ seuil\ d'erreur\ et\ ddl = (k-1)(g-1) = (2-1)(4-1) = 3$

$x_{obs}^2 > x_{tab}^2 \Rightarrow$ n accepte H_1 donc la réaction à un champignon causant la rouille est dépendante de l'âge des arbres.

1.5. Loi du t de student (William Gosset)

Le test *t* de Student est un rapport entre deux variables aléatoires indépendantes (la technique vérifiant cette indépendance est le test de Fisher-Snedecor)...

Au numérateur se trouve une variable aléatoire qui suit une loi normale centrée réduite. Au dénominateur figure la racine carrée d'une autre variable aléatoire qui suit quant à elle une loi du khi² à *n* degrés de liberté (ddl) divisée par la racine carrée de l'effectif *n* :

Si X et Y sont deux variables aléatoires indépendantes de lois respectives N(0;1) et $\chi^2 n$, alors la variable aléatoire Tn suivra une loi de Student à n degrés de liberté.

$$Tn = \frac{X}{\sqrt{\frac{Y^2}{n}}}$$

Si X est une variable aléatoire suivant une loi de Student à n degrés de liberté, on a :

$$E(X) = 0$$

$$Var(X) = \frac{n}{n-2} \quad \text{avec} \quad n > 2$$

La loi de Student c'est une loi de probabilité dont la fonction de densité a une forme algébrique un peu compliquée à écrire. Mais sa courbe représentative est simple à visualiser.

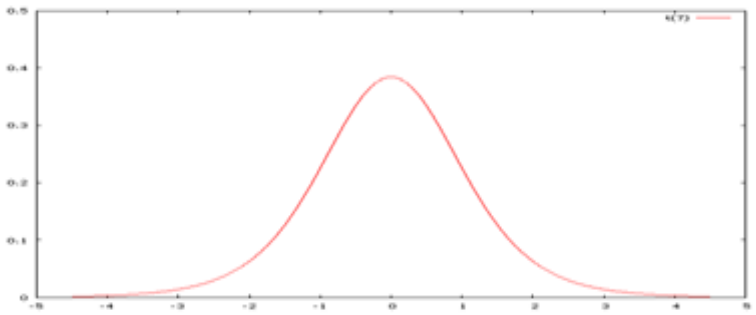


Figure 8. Courbe représentative de loi de Student

La formule de la loi de Student peut s'écrire comme suit :

$$t_{obs} = \frac{\bar{x} - \mu}{\sqrt{\frac{S^2}{n}}}$$

Applications **du test *t***

• Estimation des paramètres d'une population à partir de renseignements portant sur un échantillon.

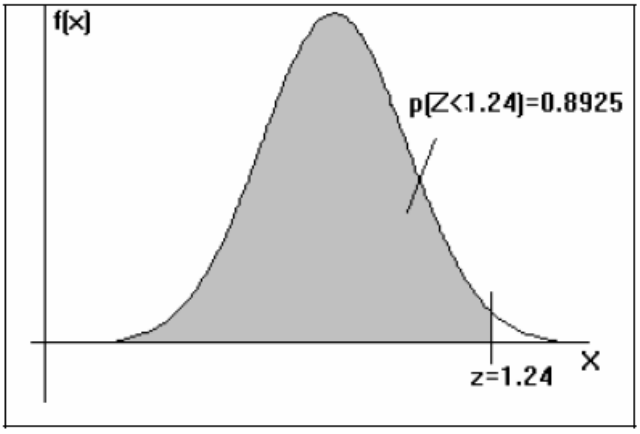
• Test de comparaison des moyennes.

On admettra que pour n ≥ 30, on peut approcher une loi de Student à n degrés de liberté par une N(0;1).

Tableau 2. Tables représentatives des principales lois de probabilité

TABLE DE LA LOI NORMALE CENTREE REDUITE

Lecture de la table: Pour z=1.24 (intersection de la ligne 1.2 et de la colonne 0.04), on a la proportion P(Z < 1,24) = 0.8925



P(Z > 1,96) = 0,025
P(Z > 2,58) = 0,005
P(Z > 3,29) = 0,0005

Rappels:

1/ P(Z > z) = 1 - P(Z < z) et 2/ P(Z < -z) = P(Z > z)

Exemple: Sachant P(Z < 1,24) = 0,8925, on en déduit:

1/ (P(Z > 1,24) = 1 - P(Z < 1,24) = 1- 0,8925 = 0,1075

2/ P(Z < -1,24) = P(Z > 1,24) = 0,1075

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997
4,0	0,99997	0,99997	0,99997	0,99997	0,99997	0,99997	0,99998	0,99998	0,99998	0,99998

DISTRIBUTION DU KHI2

La table donne les valeurs critiques de χ^2 pour un nombre de degrés de liberté (ddl) et pour un seuil repère donnés (α).

Par exemple:

Pour ddl = 3 et α= 0,05 la table indique $\chi^2 = 7{,}81$

Ceci signifie que: P ($\chi^2_{[3]} > 7{,}81$) = 0,05

ddl \ α	0,05	0,01	0,001
1	3,84	6,63	10,83
2	5,99	9,21	13,82
3	7,81	11,34	16,27
4	9,49	13,28	18,47
5	11,07	15,09	20,52
6	12,59	16,81	22,46
7	14,07	18,48	24,32
8	15,51	20,09	26,12
9	16,92	21,67	27,88
10	18,31	23,21	29,59
11	19,68	24,72	31,26
12	21,03	26,22	32,91
13	22,36	27,69	34,53
14	23,68	29,14	36,12
15	25,00	30,58	37,70
16	26,30	32,00	39,25
17	27,59	33,41	40,79
18	28,87	34,81	42,31
19	30,14	36,19	43,82
20	31,41	37,57	45,31
21	32,67	38,93	46,80
22	33,92	40,29	48,27
23	35,17	41,64	49,73
24	36,42	42,98	51,18
25	37,65	44,31	52,62
26	38,89	45,64	54,05
27	40,11	46,96	55,48
28	41,34	48,28	56,89
29	42,56	49,59	58,30
30	43,77	50,89	59,70

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
400	3.865	3.018	2.627	2.394	2.237	2.121	2.032	1.962	1.903	1.854

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	11	12	13	14	15	16	17	18	19	20
1	243.0	243.9	244.7	245.4	245.9	246.5	246.9	247.3	247.7	248.0
2	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660
4	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803
5	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558
6	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874
7	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445
8	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150
9	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936
10	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774
11	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646
12	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544
13	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459
14	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388
15	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328
16	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276
17	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230
18	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191
19	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155
20	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124
21	2.283	2.250	2.222	2.197	2.176	2.156	2.139	2.123	2.109	2.096
22	2.259	2.226	2.198	2.173	2.151	2.131	2.114	2.098	2.084	2.071
23	2.236	2.204	2.175	2.150	2.128	2.109	2.091	2.075	2.061	2.048
24	2.216	2.183	2.155	2.130	2.108	2.088	2.070	2.054	2.040	2.027
25	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007
26	2.181	2.148	2.119	2.094	2.072	2.052	2.034	2.018	2.003	1.990
27	2.166	2.132	2.103	2.078	2.056	2.036	2.018	2.002	1.987	1.974
28	2.151	2.118	2.089	2.064	2.041	2.021	2.003	1.987	1.972	1.959
29	2.138	2.104	2.075	2.050	2.027	2.007	1.989	1.973	1.958	1.945
30	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932
40	2.038	2.003	1.974	1.948	1.924	1.904	1.885	1.868	1.853	1.839
50	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784
60	1.952	1.917	1.887	1.860	1.836	1.815	1.796	1.778	1.763	1.748
70	1.928	1.893	1.863	1.836	1.812	1.790	1.771	1.753	1.737	1.722
80	1.910	1.875	1.845	1.817	1.793	1.772	1.752	1.734	1.718	1.703
90	1.897	1.861	1.830	1.803	1.779	1.757	1.737	1.720	1.703	1.688
100	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676
150	1.853	1.817	1.786	1.758	1.734	1.711	1.691	1.673	1.656	1.641
200	1.837	1.801	1.769	1.742	1.717	1.694	1.674	1.656	1.639	1.623
400	1.813	1.776	1.745	1.717	1.691	1.669	1.648	1.630	1.613	1.597

Quantiles d'ordre 0.95 de la loi de Fisher

Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	21	22	23	24	25	26	27	28	29	30
1	248.3	248.6	248.8	249.1	249.3	249.5	249.6	249.8	250.0	250.1
2	19.45	19.45	19.45	19.45	19.46	19.46	19.46	19.46	19.46	19.46
3	8.654	8.648	8.643	8.639	8.634	8.630	8.626	8.623	8.620	8.617
4	5.795	5.787	5.781	5.774	5.769	5.763	5.759	5.754	5.750	5.746
5	4.549	4.541	4.534	4.527	4.521	4.515	4.510	4.505	4.500	4.496
6	3.865	3.856	3.849	3.841	3.835	3.829	3.823	3.818	3.813	3.808
7	3.435	3.426	3.418	3.410	3.404	3.397	3.391	3.386	3.381	3.376
8	3.140	3.131	3.123	3.115	3.108	3.102	3.095	3.090	3.084	3.079
9	2.926	2.917	2.908	2.900	2.893	2.886	2.880	2.874	2.869	2.864
10	2.764	2.754	2.745	2.737	2.730	2.723	2.716	2.710	2.705	2.700
11	2.636	2.626	2.617	2.609	2.601	2.594	2.588	2.582	2.576	2.570
12	2.533	2.523	2.514	2.505	2.498	2.491	2.484	2.478	2.472	2.466
13	2.448	2.438	2.429	2.420	2.412	2.405	2.398	2.392	2.386	2.380
14	2.377	2.367	2.357	2.349	2.341	2.333	2.326	2.320	2.314	2.308
15	2.316	2.306	2.297	2.288	2.280	2.272	2.265	2.259	2.253	2.247
16	2.264	2.254	2.244	2.235	2.227	2.220	2.212	2.206	2.200	2.194
17	2.219	2.208	2.199	2.190	2.181	2.174	2.167	2.160	2.154	2.148
18	2.179	2.168	2.159	2.150	2.141	2.134	2.126	2.119	2.113	2.107
19	2.144	2.133	2.123	2.114	2.106	2.098	2.090	2.084	2.077	2.071
20	2.112	2.102	2.092	2.082	2.074	2.066	2.059	2.052	2.045	2.039
21	2.084	2.073	2.063	2.054	2.045	2.037	2.030	2.023	2.016	2.010
22	2.059	2.048	2.038	2.028	2.020	2.012	2.004	1.997	1.990	1.984
23	2.036	2.025	2.014	2.005	1.996	1.988	1.981	1.973	1.967	1.961
24	2.015	2.003	1.993	1.984	1.975	1.967	1.959	1.952	1.945	1.939
25	1.995	1.984	1.974	1.964	1.955	1.947	1.939	1.932	1.926	1.919
26	1.978	1.966	1.956	1.946	1.938	1.929	1.921	1.914	1.907	1.901
27	1.961	1.950	1.940	1.930	1.921	1.913	1.905	1.898	1.891	1.884
28	1.946	1.935	1.924	1.915	1.906	1.897	1.889	1.882	1.875	1.869
29	1.932	1.921	1.910	1.901	1.891	1.883	1.875	1.868	1.861	1.854
30	1.919	1.908	1.897	1.887	1.878	1.870	1.862	1.854	1.847	1.841
40	1.826	1.814	1.803	1.793	1.783	1.775	1.766	1.759	1.751	1.744
50	1.771	1.759	1.748	1.737	1.727	1.718	1.710	1.702	1.694	1.687
60	1.735	1.722	1.711	1.700	1.690	1.681	1.672	1.664	1.656	1.649
70	1.709	1.696	1.685	1.674	1.664	1.654	1.646	1.637	1.629	1.622
80	1.689	1.677	1.665	1.654	1.644	1.634	1.626	1.617	1.609	1.602
90	1.675	1.662	1.650	1.639	1.629	1.619	1.610	1.601	1.593	1.586
100	1.663	1.650	1.638	1.627	1.616	1.607	1.598	1.589	1.581	1.573
150	1.627	1.614	1.602	1.590	1.580	1.570	1.560	1.552	1.543	1.535
200	1.609	1.596	1.583	1.572	1.561	1.551	1.542	1.533	1.524	1.516
400	1.582	1.569	1.556	1.545	1.534	1.523	1.514	1.505	1.496	1.488

Quantiles d'ordre 0.95 de la loi de Fisher

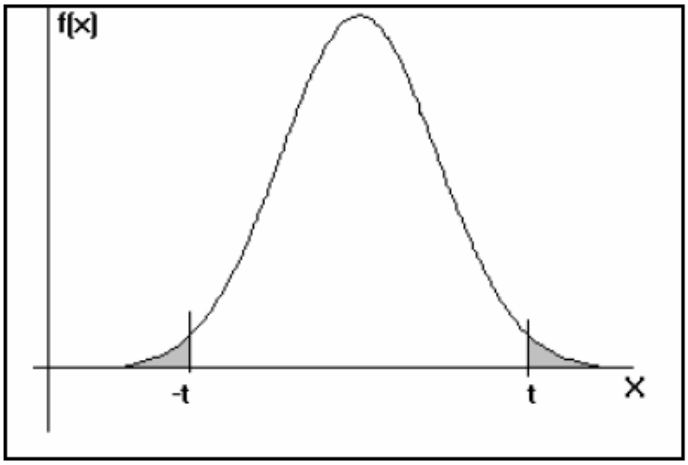
Degrés de liberté du numérateur sur la première ligne
Degrés de liberté du dénominateur sur la colonne de gauche

	40	50	60	70	80	90	100	150	200	400
1	251.1	251.8	252.2	252.5	252.7	252.9	253.0	253.5	253.7	253.8
2	19.47	19.48	19.48	19.48	19.48	19.48	19.49	19.49	19.49	19.49
3	8.594	8.581	8.572	8.566	8.561	8.557	8.554	8.545	8.540	8.537
4	5.717	5.699	5.688	5.679	5.673	5.668	5.664	5.652	5.646	5.643
5	4.464	4.444	4.431	4.422	4.415	4.409	4.405	4.392	4.385	4.381
6	3.774	3.754	3.740	3.730	3.722	3.716	3.712	3.698	3.690	3.686
7	3.340	3.319	3.304	3.294	3.286	3.280	3.275	3.260	3.252	3.248
8	3.043	3.020	3.005	2.994	2.986	2.980	2.975	2.959	2.951	2.947
9	2.826	2.803	2.787	2.776	2.768	2.761	2.756	2.739	2.731	2.726
10	2.661	2.637	2.621	2.610	2.601	2.594	2.588	2.572	2.563	2.558
11	2.531	2.507	2.490	2.478	2.469	2.462	2.457	2.439	2.431	2.426
12	2.426	2.401	2.384	2.372	2.363	2.356	2.350	2.332	2.323	2.318
13	2.339	2.314	2.297	2.284	2.275	2.267	2.261	2.243	2.234	2.229
14	2.266	2.241	2.223	2.210	2.201	2.193	2.187	2.169	2.159	2.154
15	2.204	2.178	2.160	2.147	2.137	2.130	2.123	2.105	2.095	2.089
16	2.151	2.124	2.106	2.093	2.083	2.075	2.068	2.049	2.039	2.034
17	2.104	2.077	2.058	2.045	2.035	2.027	2.020	2.001	1.991	1.985
18	2.063	2.035	2.017	2.003	1.993	1.985	1.978	1.958	1.948	1.942
19	2.026	1.999	1.980	1.966	1.955	1.947	1.940	1.920	1.910	1.903
20	1.994	1.966	1.946	1.932	1.922	1.913	1.907	1.886	1.875	1.869
21	1.965	1.936	1.916	1.902	1.891	1.883	1.876	1.855	1.845	1.838
22	1.938	1.909	1.889	1.875	1.864	1.856	1.849	1.827	1.817	1.810
23	1.914	1.885	1.865	1.850	1.839	1.830	1.823	1.802	1.791	1.784
24	1.892	1.863	1.842	1.828	1.816	1.808	1.800	1.779	1.768	1.761
25	1.872	1.842	1.822	1.807	1.796	1.787	1.779	1.757	1.746	1.739
26	1.853	1.823	1.803	1.788	1.776	1.767	1.760	1.738	1.726	1.719
27	1.836	1.806	1.785	1.770	1.758	1.749	1.742	1.719	1.708	1.701
28	1.820	1.790	1.769	1.754	1.742	1.733	1.725	1.702	1.691	1.683
29	1.806	1.775	1.754	1.738	1.726	1.717	1.710	1.686	1.675	1.667
30	1.792	1.761	1.740	1.724	1.712	1.703	1.695	1.672	1.660	1.652
40	1.693	1.660	1.637	1.621	1.608	1.597	1.589	1.564	1.551	1.542
50	1.634	1.599	1.576	1.558	1.544	1.534	1.525	1.498	1.484	1.475
60	1.594	1.559	1.534	1.516	1.502	1.491	1.481	1.453	1.438	1.428
70	1.566	1.530	1.505	1.486	1.471	1.459	1.450	1.420	1.404	1.394
80	1.545	1.508	1.482	1.463	1.448	1.436	1.426	1.395	1.379	1.368
90	1.528	1.491	1.465	1.445	1.429	1.417	1.407	1.375	1.358	1.348
100	1.515	1.477	1.450	1.430	1.415	1.402	1.392	1.359	1.342	1.331
150	1.475	1.436	1.407	1.386	1.369	1.356	1.345	1.309	1.290	1.278
200	1.455	1.415	1.386	1.364	1.346	1.332	1.321	1.283	1.263	1.249
400	1.425	1.383	1.352	1.329	1.311	1.296	1.283	1.242	1.219	1.204

DISTRIBUTIONS DU t DE STUDENT

Table des valeurs critiques bilatérales usuelles

Pour une distribution de Student à ddl degrés de liberté et pour une proportion α (.05, .01 ou .001), la table indique *t* tel que P($|T| > t$) = α



Exemple: Pour ddl = 5, on a P($|T| > 2.571$) = .05 (on note $t_{[5].05}$ cette valeur.).

α / α/2 / ddl	0,05 / 0,025	0,01 / 0,005	0,001 / 0,0005
1	12.706	63.657	636.619
2	4.303	9.925	31.599
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
21	2.080	2.831	3.819
22	2.074	2.819	3.792
23	2.069	2.807	3.768
24	2.064	2.797	3.745
25	2.060	2.787	3.725
26	2.056	2.779	3.707
27	2.052	2.771	3.690
28	2.048	2.763	3.674
29	2.045	2.756	3.659
30	2.042	2.750	3.646
40	2.021	2.704	3.551
60	2.000	2.660	3.460
120	1.980	2.617	3.373
30000	1.960	2.576	3.291

2. Représentations graphiques, histogramme, diagramme

La présentation des données a pour but de faciliter la **compréhension** et l'**interprétation** des données par le lecteur ou l'auditeur. Une **représentation graphique de données statistiques** est un résumé visuel des données chiffrées. Elle permet en un seul coup d'œil d'en saisir la tendance générale. Il existe 2 grands **types de présentation**:

a) Tableaux

- tableau de fréquences
- tableau de corrélation
- tableau de contingence

b) Figures

- diagramme en bâtons
- polygone de fréquences
- histogramme
- diagramme de dispersion

2.1. Règles générales de présentation

2.1.1. Toute représentation doit avoir un TITRE situé:

- en **haut** pour un tableau
- en **bas** pour une figure

Le titre doit contenir l'information suivante:

- le numéro du tableau ou de la figure

ex: Figure 1: ...

ex: Tableau 4: ...

- le type de représentation

ex: histogramme

ex: diagramme en bâton

- les variables représentées et leurs unités (Quoi?)
- le lieu d'échantillonnage (Où?)
- le moment de l'échantillonnage (Quand?)
- la provenance des données s'il y a lieu (Par qui?)

2.1.2. Les axes des figures sont:

- **identifiés** (titres et unités)
- gradués avec des valeurs **simples**

ex: ne pas faire des graduations du genre:

32,234245 ; 35,384236 ; 37, 938402 ; etc.

2.1.3. Les axes représentant des variables quantitatives **commencent à 0**.

Tableau 1. Tableau de fréquences de la longueur totale du crâne pour 60 souris sylvestres récoltées par Landry(2000) Lac Duparquet (Québec)

Longueur totale du crâne (mm)	Fréquence absolue	Fréquence cumulée	Fréquence relative	Pourcentage
22,5	6	6	0,100	10,0
23,0	6	12	0,100	10,0
23,5	21	33	0,350	35,0
24,0	9	42	0,150	15,0
24,5	10	52	0,167	16,7
25,0	6	58	0,100	10,0
25,5	2	60	0,033	3,3
Total	60	60	1,000	100,0

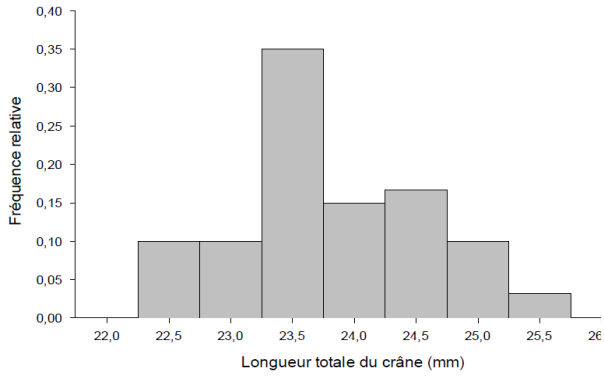


Figure 1. Histogramme représentant la distribution de fréquences relatives de la longueur totale du crâne de 60 souris récoltées par Landry (2000) Lac Duparquet (Québec)

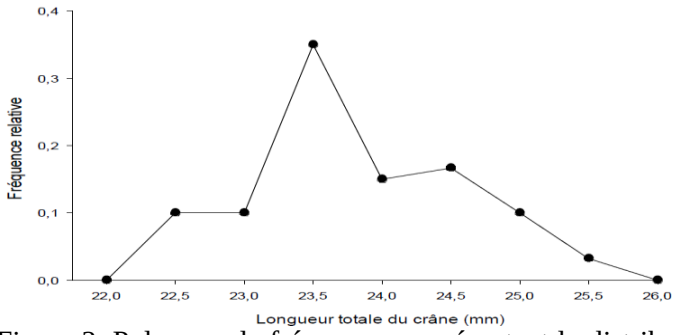


Figure 2. Polygone de fréquences représentant la distribution de fréquences relatives de la longueur totale du crâne de 60 souris récoltées par Landry (2000) Lac Duparquet (Québec)

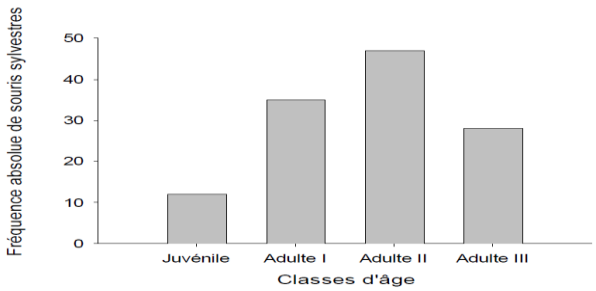


Figure 3. Diagramme en bâtons représentant la distribution de fréquences du nombre de souris sylvestres en fonction de leur âge récoltées par Landry (2000) Lac Duparquet (Québec)

Tableau 2 : Matrice de corrélation ou tableau de corrélation des critères notés sur les cidres

	odeur	sucre	acide	Amer	parfum
odeur	1,00				
sucre	0,08	1,00			
acide	-0,16	-0,29	1,00		
amer	0,49	-0,60	-0,08	1,00	
parfum	-0,29	0,87	-0,40	-0,63	1,00

Tableau 3. Tableau de contingence de la couleur des yeux et des heveux de 100 étudiants en 2004 ramassé par Corinne Tastayre et Marie-Hélène Ouellette.

	blond	*Brun*	*noir*	*somme*
Bleu	10	13	5	28
Vert	6	3	2	11
noisette	24	32	5	61
somme	40	48	12	100

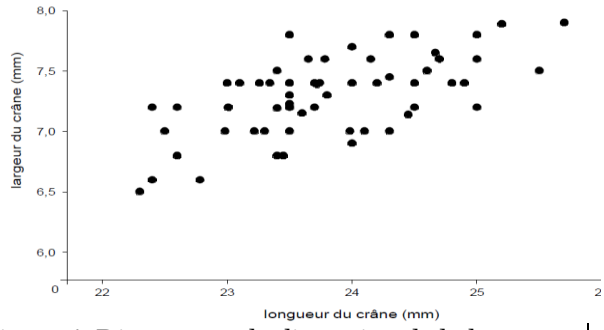


Figure 4. Diagramme de dispersion de la largeur (mm) et de la longueur totale (mm) du crâne chez 60 souris sylvestres (Québec; Landry, 2000).

Figure 9. Exemples des représentations graphiques et tabulaires

2.2. Représentation des effectifs et des fréquences

- ✓ Pour les variables qualitatives, on utilise fréquemment les **diagrammes circulaires** dits « en camembert ».

Figure 10. Diagramme circulaire

- ✓ Dans le cas d'une série discrète, la représentation graphique associée est un diagramme en bâtons. Il se présente dans un repère orthogonal où figurent sur l'axe des abscisses les valeurs du caractère étudié, et sur l'axe des ordonnées les effectifs. Au niveau de chaque valeur en abscisse, on trace un segment vertical de longueur la valeur de l'effectif correspondant à cette valeur. On peut compléter ce diagramme par le polygone des effectifs, formé des segments joignant les sommets des bâtons comme on peut le voir sur le graphique ci-dessous :

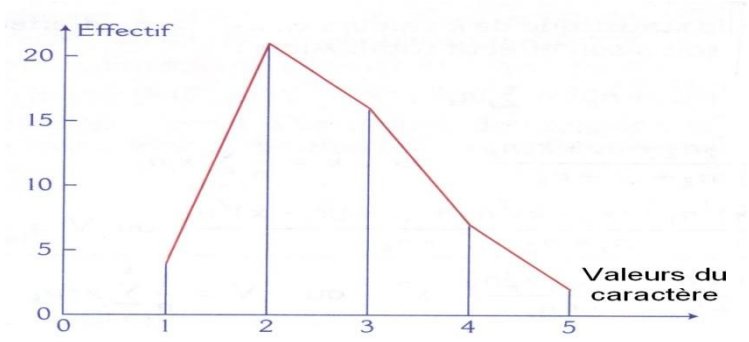


Figure 11. Diagramme en bâtons

Ce diagramme correspond au tableau suivant :

Tableau 3. Tableau correspondant au Diagramme précédent

Valeur du caractère	1	2	3	4	5
Effectifs	4	22	15	7	2

- ✓ Pour représenter une série statistique continue, on utilise un histogramme qui est un graphique composé de barres rectangulaires dont la largeur est égale à l'amplitude de la classe et dont l'aire est proportionnelle à l'effectif de la classe.

  On fait toujours l'hypothèse que les valeurs observées sont réparties uniformément à l'intérieur de chaque classe.

  Si les classes sont d'amplitudes égales, il suffit de prendre pour hauteur de chaque rectangle l'effectif de la classe qu'il représente. Par contre, si les classes n'ont pas toutes la même amplitude, on se ramène au cas précédent en partageant chaque classe en sous-classes ayant

toutes la même amplitude. On prend ensuite pour hauteur de chaque rectangle l'effectif de la sous-classe qu'il représente. L'amplitude de chaque sous-classe est appelée unité d'amplitude.

Considérons le tableau suivant, regroupant les poids nets de légumes observés dans un échantillon de 100 boîtes de conserves.

Tableau 4. Table d'effectifs relatifs à un caractère continu

Poids en g	[240 ; 244[	[244 ; 246[	[246 ; 248[	[248 ; 252[	[252 ; 260[
Effectifs	12	20	24	36	8

On constate que les classes ont parfois une amplitude de 2 , parfois de 4 , et la dernière classe a quant à elle une amplitude de 8. On va alors choisir pour unité d'amplitude la valeur 2. Etant donné qu'on suppose que les valeurs du caractère sont réparties uniformément dans chaque classe, on construit alors le nouveau tableau suivant :

Tableau 5. Table d'effectifs relatifs à un caractère continu (classes d'étendues égales)

Poids en g	[240;242[	[242;244[	[244;246[	[246;248[	[248;250[	[250;252[	[252;254[	[254;256[	[256;258[	[258;260[
Effectifs	6	6	20	24	18	18	2	2	2	2

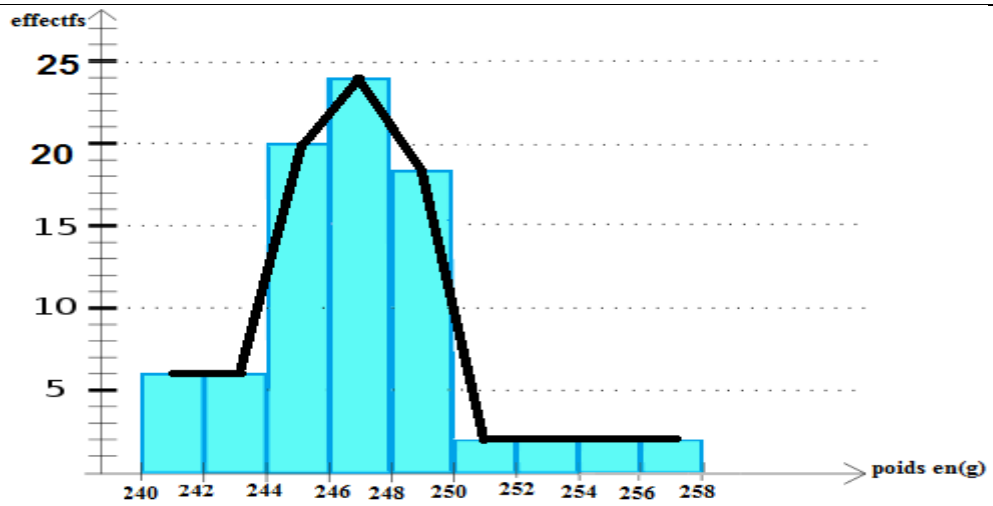


Figure 12. Histogramme et polygone de fréquences

2.3. Représentation des effectifs cumulés.

Ce graphique est obtenu à partir du tableau des **effectifs cumulés croissants.** On place dans un repère orthogonal les points dont l'abscisse est la classe, et dont l'ordonnée est l'effectif cumulé correspondant à cette classe. On relie ensuite ces différents points par des segments de droites pour obtenir une ligne brisée.

Prenons l'exemple des tailles d'un groupe de 100 personnes. Le tableau des effectifs cumulés croissants est :

Tableau 6. tableau des effectifs cumulés croissants

Valeur du caractère	<165	<170	<175	<180	<185	<190
Effectifs	12	37	44	72	92	100

Les points à placer sur le graphique ont pour coordonnées : (165 ; 12) ; (170 ; 37) ; (175 ;44) ; (180 ; 72) ; 185 ; 92) ; (190 ;100). On obtient alors le polygone suivant :

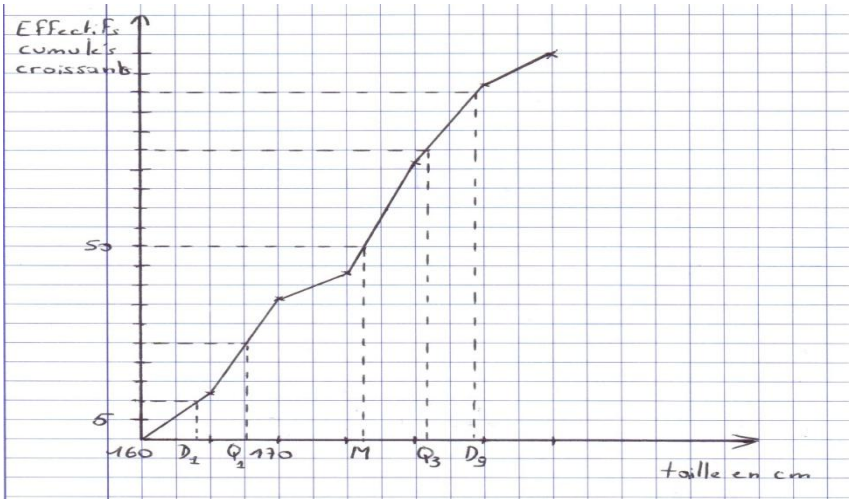


Figure 13. Polygone des effectifs cumulés

Remarquons que le point de départ de ce polygone est l'origine du repère.

Obtention de médiane, quartiles et déciles

On remarquera les pointillés sur le graphique précédent, permettant de donner les abscisses des points d'ordonnées respectives : 50 ; 25 ; 75 ; 10 ; 90. Ces valeurs sont des valeurs approchées obtenues graphiquement de la médiane, des 1er et 3ème quartiles, des 1er et dernier déciles. En lisant ces valeurs sur l'axe des abscisses, on a donc :

M = 176 ; Q1 = 167,5 ; Q3 = 181 ; D1 = 164 ; D9 = 184

2.4. Nuage de points

On rencontre principalement cette représentation dans les **séries statistiques à deux variables**. Elle apparaît aussi de manière moins identifiable dans les cartes géographiques ou météorologique (densité de population, présence d'industries,...). L'effectif est alors associé à une taille de point ou une couleur de

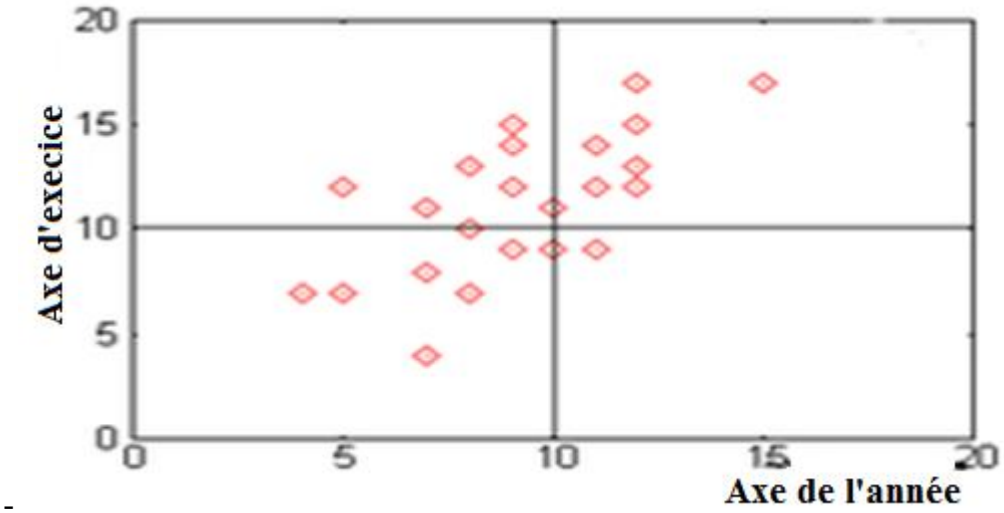


Figure 14. Représentation graphique sous forme de nuage de points

2.5. Diagramme en boîte à moustaches

Le diagramme en **boîte à moustaches** résume seulement quelques caractéristiques de **position** du caractère étudié (médiane, quartiles, min/max). Il est utilisé principalement pour comparer un même caractère dans deux populations de tailles différentes. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. On ajoute parfois des segments aux extrémités menant jusqu'aux valeurs min/max ou jusqu'au premier et neuvième décile. On parle alors de diagramme en boîte à moustaches ou à pattes.

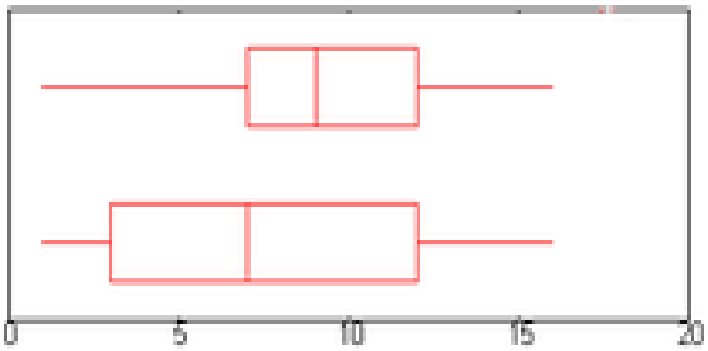


Figure 15. Diagramme en boîte à moustaches

Exercices sur la statique descriptive (Rappel des cours de 3eme année)

Exercice 01

Le centre de recherche d'une société spécialisée dans les biotechnologies a récemment mis au point une nouvelle variété céréalière. Le rendement (en quintaux/hectare) de cette nouvelle variété a été relevé dans la région Centre sur un échantillon de parcelles témoin en condition de culture "BIO".

On vous demande d'exploiter rapidement et au mieux ces résultats (en définissant la taille de l'échantillon, type caractère, type de distribution, Mode, Médiane, Moyenne , Ecart-type, Quartiles, Fréquence, Fréquence cumulée ainsi que les différentes représentations graphiques utiles etc.)..

Rendements mesurés (en quintaux/hectare) :

102 ; 104,5 ; 121,63 ; 122,5 ; 103,23 ; 106 ; 107,25 ; 106,44 ; 110 ; 111,3 ; 112,23 ; 96,6 ; 120 ; 102 ; 103,2 ; 104 ; 127,56 ; 113,5 ; 95 ; 101,125 ; 125 ; 96 ; 114 ; 109 ; 92,3 ; 106,65 ; 107,7 ; 113,8 ; 115 ; 109,3 ; 123,325 ; 124,1 ; 113,5 ; 112,21 ; 111,56 ; 110 ; 112,22 ; 129 ; 133,25 ; 102 ; 114,5 ; 94,25 ; 116 ; 115,2 ; 125 ; 109,4 ; 108,55 ; 96,2 ; 116,6 ; 117,3 ; 97,5 ; 117,66 ; 117,8 ; 117,5 ; 118,22 ; 118 ; 117,5 ; 119 ; 114,3 ; 90 ; 96,2 ; 116,6 ; 123,56 ; 115 ; 120,23

Exercice 02

Voici le gain moyen quotidien d'un animal exprimés en g

$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	$x9$	$x10$
737	630	573	615	718	620	820	763	786	529

- calculez les paramètres de position ainsi que ceux de dispersion.
- représentez ces paramètres graphiquement (boites à moustaches).
- donnez les paramètres de forme le degré de symétrie et le coefficient d'aplatissement.
- représentez graphiquement la distribution des données.

3. Tests classiques paramétriques et non paramétriques

3.1. Définition des tests paramétriques et non paramétriques

On parle de **tests paramétriques** lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres. L'hypothèse de normalité sous jacente des données est le plus souvent utilisée, la moyenne et la variance suffisent pour caractériser complètement la distribution. Concernant les tests d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances.

Les **tests non paramétriques** ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests *distribution free*. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire.

Lorsque les données sont quantitatives, les tests non paramétriques transforment les valeurs en rangs. L'appellation **tests de rangs** est souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables.

La distinction paramétrique – non paramétrique est essentielle. Elle est systématiquement mise en avant dans la littérature. Les tests non paramétriques, en ne faisant aucune hypothèse sur les distributions des données, élargissent le champ d'application des procédures statistiques. En contrepartie, ils sont moins puissants lorsque ces hypothèses sont compatibles avec les données.

En conclusion les différences entre les deux types de tests sont données dans le tableau ci-dessous

Tableau 7. Différences entre les tests paramétriques et non paramétriques

○ Les tests paramétriques nécessitent des conditions de validité	○ Les tests non paramétriques :
✓ Hypothèses sur la distribution des observations (ex : $X \sim N(\mu;\sigma)$) ✓ Distributions caractérisées par des paramètres (moyenne, variance, . . .) ✓ Ces paramètres sont estimés	✓ Distribution free : pas d'hypothèse sur la distribution des observations ✓ Tests adaptés aux variables quantitatives et qualitatives (nominales et ordinales) ✓ La plupart du temps : tests bases sur la notion de rangs

3.2. Liste des tests usuels

Les tests paramétriques et non paramétriques usuels snt donnés dans le tableau suivant :

Tableau 8. Liste des tests usuels

Type de test	Tests paramétriques	Tests non paramétriques
Problème à 1 échantillon		
Tests de conformité à un standard	• Test de conformité d'une moyenne (test de Student), d'un écart type et d'une proportion	
Tests d'adéquation à une loi		• Test de Kolmogorov-Smirnov • Test d'adéquation du χ^2 • Test de Shapiro-Wilks, test de Lilliefors, test d'Anderson-Darling, test de D'Agostino, Test de Jarque Bera
Tests de symétrie des répartitions		• Test de Wilcoxon • Test de Van der Waerden
Comparaison de (K ≥ 2) populations		
Tests omnibus de comparaison de populations, les fonctions de répartition sont les mêmes dans les groupes		• Test de Kolmogorov - Smirnov • Test de Kuiper • Test de Cramer - von Mises
Tests de comparaison de K échantillons indépendants (différenciation selon les caractéristiques de tendance centrale, modèle de localisation)	• Test de comparaison de moyennes (K = 2) • ANOVA (analyse de variance) à 1 facteur	• Test de la somme des rangs de Wilcoxon (K=2) • Test de Mann - Whitney (K=2) • Test de Kruskal - Wallis • Test des médianes • Test de Van der Waerden • Test de Jonckheere - Terpstra (alternatives ordonnées)
Tests de comparaison de K échantillons indépendants (différenciation selon les caractéristiques de dispersion, modèle d'échelle)	• Test de Fisher (K=2) • Test de Bartlett • Test de Cochran • Test F-max de Hartley • Test de Levene • Test de Brown-Forsythe	• Test de Ansari - Bradley • Test de Klotz • Test de Mood • Test de Siegel-Tukey • Test des différences extrêmes de Moses
Tests pour K échantillons appariés (mesures répétées ou blocs aléatoires complets)	• Test de Student de comparaison de moyennes pour échantillons appariés (K=2)	• Test des signes (K=2) • Test des rangs signés de Wilcoxon (K=2) • Test de Friedman • Test de Page (alternatives ordonnées) • Test de McNemar (K=2, variables

	• Test de comparaison de variances pour échantillons appariés (K=2) • ANOVA pour blocs aléatoires complets	binaires) • Test Q de Cochran (variables binaires)
Tests multivariés pour K échantillons indépendants	• T^2 de Hotelling, comparaison de K=2 barycentres (vecteur des moyennes) • MANOVA (analyse de variance mutlivariée), comparaison de K barycentres : Lambda de Wilks, Trace de Pillai, Trace de Hotelling-Lawley, La plus grande valeur propre de Roy • Test M de Box de comparaison de matrices de variance covariance	
Association entre variables		
Association entre p=2 variables quantitatives	• Coefficient de corrélation de Pearson	• Rho de Spearman • Tau-a de Kendall
Association entre p = 2 variables ordinales		• Gamma de Goodman - Kruskal • Tau-b et Tau-c de Kendall • d de Sommers • Test de Mantel - Haenszel (variables binaires)
Association entre p=2 variables nominales		• Test d'indépendance du χ^2 • t de Tschuprow et v de Cramer • Coefficient phi (variables binaires) • Coefficient Q de Yule (variables binaires) • Lambda de Goodman - Kruskal • Tau de Goodman - Kruskal • U de Theil
Association entre ($p \geq 2$) variables		• Coefficient de concordance de Kendall (variables quantitatives ou ordinales) • Coefficient Kappa de Fleiss, concordance de p jugements (variables ordinales ; Kappa de Cohen pour p = 2)

3.3. Avantages et inconvénients des tests non paramétriques

* Avantages

- Pas d'hypothèse sur la distribution⇒ champ d'application a priori plus large
- Tests adaptés aux variables ordinales (ex : degré de satisfaction)
- Robustesse par rapport aux données atypiques

Exemple:

4	5	8	7	3	38	$\bar{x}_1 = 10{,}8$
4	5	8	7	3	6	$\bar{x}_2 = 5{,}5$

- Différence due à une seule observation!! La transformation en rangs permet de gommer cette difference
- Tests adaptés aux petits échantillons (n<30)

* Inconvénients

- Lorsque les conditions d'applications sont vérifiées : Tests non paramétriques moins puissants que les tests paramétriques
- Difficulté d'interprétation: on ne compare plus des paramètres (moyenne, proportion, variance, …)

3.4. Exemple de test non paramétrique -Test de Mann-Whitney-Wilcoxon –

Ce test regroupe 2 tests équivalents : Test U de Mann-Whitney et le test W de Wilcoxon. Son objectif est la comparaison de 2 échantillons indépendants par rapport à une variable X de nature : quantitative et qualitative ordinale.

Soient : $F_1(x)$ la fonction de répartition de x dans la population 1

$F_2(x)$ la fonction de répartition de x dans la population 2

Les hypothèses de test sont :

$H_0 \Rightarrow F_1(x) = F_2(x+\Theta)$; $\Theta = 0$ Distributions identiques.

$H_1 \Rightarrow F_1(x) = F_2(x+\Theta)$; $\Theta \neq 0$ Distributions différentes.

Θ paramètre de translation : décalage entre les fonctions de répartition

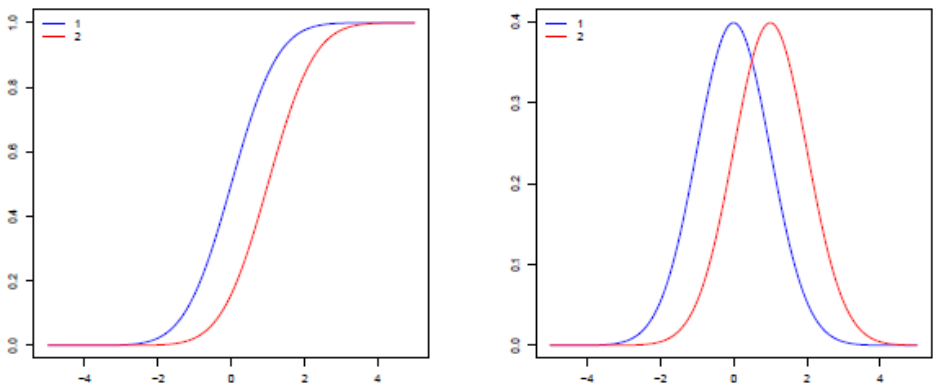


Figure 16. Translation entre les fonctions de répartition exemple décalage Θ≠0

Soient les deux groupes G_1et G_2

$$\left.\begin{array}{l} G_1 : \{x_{11}, x_{12}, \ldots, x_{1n_1}\} \\ G_2 : \{x_{21}, x_{22}, \ldots, x_{2n_2}\} \end{array}\right\} = \text{Transformation en rangs } (G_1 \cup G_2)$$

$R(X_1) = x$ les rangs des valeurs du groupe 1

$R(X_2) = o$ les rangs des valeurs du groupe 2

2 configurations extrêmes

$\xrightarrow[rangs]{x\,o\,x\,o\,x\,o\,x\,o\,x\,o\,x\,o}$ (1) $\rightarrow$ $\mathcal{H}_0$ vraie (mélange total)

$\xrightarrow[rangs]{x\,x\,x\,x\,x\,x\,x\,o\,o\,o\,o\,o\,o\,o}$ (2) $\rightarrow$ $\mathcal{H}_0$ "totalement" fausse

Notons : S_1 est la somme des rangs des observations du groupe 1

Et U_1 est le nombre de couples $\{(x_{1i}, x_{2j})/\ x_{1i}>x_{2j}\}$

$U_1 = S_1 - \frac{n_1(n_1+1)}{2}$

Notons : S_2 est la somme des rangs des observations du groupe 2

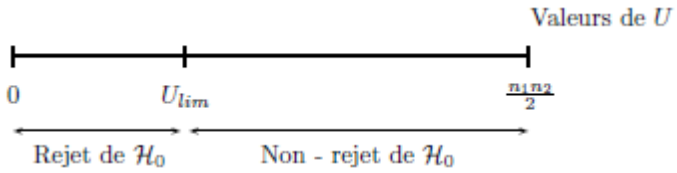
Et U_2 est le nombre de couples $\{(x_{1i}, x_{2j})/\ x_{1i}>x_{2j}\}$

$U_2 = S_2 - \frac{n_2(n_2+1)}{2}$

Statistique de test : U= min (U_1,U_2)

Cette valeur est comparée à la valeur de $U_{tab\ ou\ lim}$ comme suit :

* $n_1$ou n_2<10

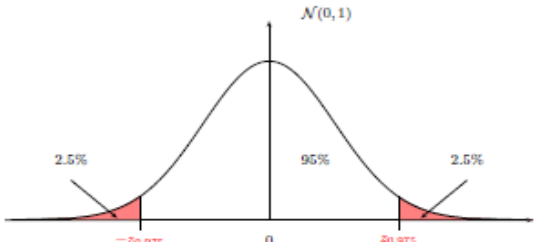


Valeur de U_{lim} lue dans la table U

* Cas particulier si $n_1$et n_2>10

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \sim \mathcal{N}(0,1)$$

$$R.C. : |Z| \geq z_{1-\alpha/2}$$



Exemple

G_1 : n_1=7 : {11, 21, 21, 25, 52, 71, 79}

G_2 : n_2=5 : {22, 43, 72, 91, 100}

Passage aux rangs :

	11	21	21	22	25	43	52	71	72	79	91	100
Rangs	1	2,5	2,5	4	5	6	7	8	9	10	11	12
groupes	1	1	1	2	1	2	1	1	2	1	2	2

Calcul de U_1 et U_2

$U_1 = S_1 - \frac{n_1(n_1+1)}{2} =$ 1+2,5+2,2+…..+10 - $\frac{7(7+1)}{2}$ = 8

$U_2 = S_2 - \frac{n_2(n_2+1)}{2}$ =4+6+9+11+12$\frac{5(5+1)}{2}$ = 27

Umin (U_1,U_2) =U_1= 8

U_{lim}= 5 (Table) et U> U_{lim} = donc non rejet de H_0 donc on peut dire que les distributions sont identiques.

Tableau 9. Table de Mann Whitney

TABLE DE MANN-WHITNEY

Valeurs critiques (U_{crit}) à comparer avec la valeur observée (U_{obs}) à partir de vos 2 échantillons pour un test **bilatéral** au seuil α = **0.05** ou **0.01**.

NB : n_1 et n_2 représentent le nombre d'observations dans chaque échantillon.

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

4. Tests de conformité

Les tests de conformité sont dits "tests à 1 échantillon". Ils ont pour but de vérifier si un échantillon peut être considéré comme représentatif de la population dont il est extrait. On étudie une variable quantitative X et on cherche à établir si les observations sont en accord avec la loi théorique de cette variable. En général, il s'agit de tester si un paramètre (tel que la moyenne, la fréquence ou la variance) calculé dans l'échantillon est conforme à sa valeur au niveau de la population. Ceci suppose que la loi théorique du paramètre est connue au niveau de la population.

4.1. Test de conformité d'une moyenne

Les tests sur la moyenne sont très proches des notions d'intervalles de confiance. En effet, leur règle de décision pour établir si l'hypothèse H_0 doit être acceptée ou rejetée repose sur le calcul d'une statistique qui n'est autre que la moyenne empirique et consiste à vérifier si la valeur calculée de cette statistique appartient ou pas à un intervalle.

Cet intervalle est formulé comme suit :

Cas 1. Petit échantillon $\mathbf{IC=[\bar{x} - t_{(n-1,\ \alpha/2)} \frac{s}{\sqrt{n}} ;\ \bar{x} + t_{(n-1,\ \alpha/2)} \frac{s}{\sqrt{n}}]}$.

Cas 2. Grand échantillon $\mathbf{IC = [\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} ;\ \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}]}$.

Ce test sont destiné à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne observée.

On tire de la formulation de l'intervalle de confiance, le test de signification suivant:

Ce test permet de vérifier deux hypothèses qui sont:

H_0 : $\bar{x} = \mu$ c'est –à-dire l'échantillon appartient à la population cible

H_1 : $\bar{x} \neq \mu$ c'est-à-dire l'échantillon n'appartient pas à la population cible (appartient à une autre population que la population cible).

- Comparaison de la moyenne à une valeur donnée : μ dans le cas où σ est connu

$$t_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

A comparer la valeur observée à celle théorique lue de la table de t avec α seuil d'erreur pour arriver à des conclusions.

Les valeurs seuils sont lues dans la table de la loi normale centrée réduite (dernière ligne de la table de Student...)

- Comparaison de la moyenne à une valeur donnée : μ dans le cas où σ est inconnu. La variance est inconnue, il nous faut donc l'estimer. Le bon estimateur est S^2. On sait que l'introduction de cet estimateur a pour effet de changer la loi normale en la loi de Student.

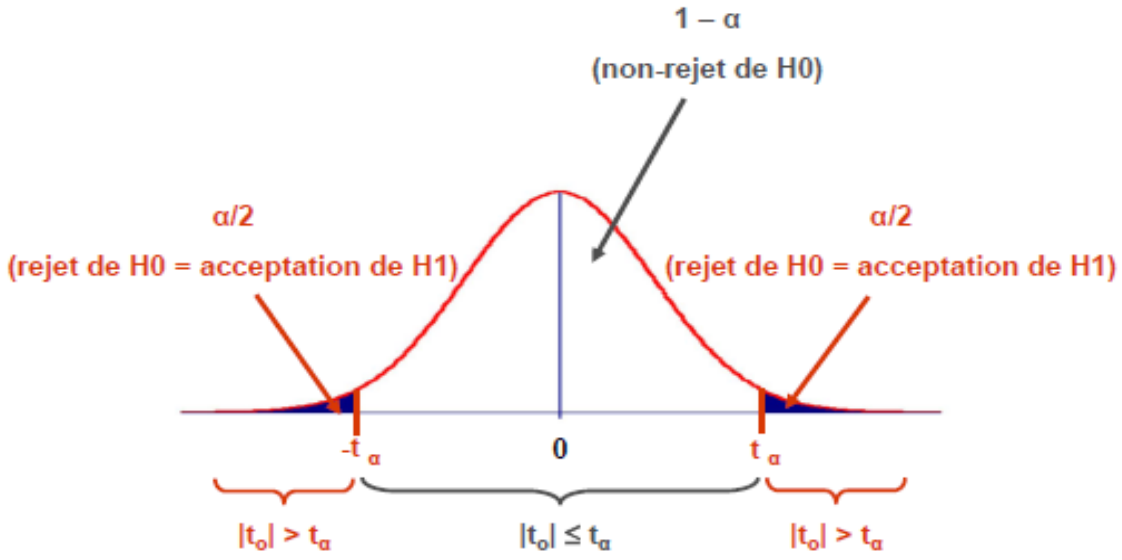
$$t_{obs} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

- En cas de Petit échantillon, sous l'hypothèse H_0, la statistique T suit une loi de Student à (n−1) degrés de liberté: T~t(n−1). Si l'échantillon est de taille n inférieure à 30, on cherche donc les bornes de l'intervalle d'acceptation dans une table de la loi de Student.
- En cas de grand échantillon, sous l'hypothèse H_0, la statistique T suit une loi normale centrée réduite Z (0,1).

On peut récapituler les différents cas du test de conformité comme suit :

Tableau 10. Différents cas du test de conformité d'une moyenne

n<30 : cas des petits échantillons		**n>30 : cas des grands échantillons**	
σ connu	**σ inconnu**	**σ connu**	**σ inconnu**
$t_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$t_{obs} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$	$z_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$z_{obs} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$
$t_{5\%}$: valeur lue à partir de la table t avec α= 5% et ddl =n-1		$Z_{5\%}$: valeur lue à partir de la dernière ligne de la table t avec α= 5 %	



Si $t_{obs} < t_{tab}$ ⟶ on accepte H_0

Si $t_{obs} > t_{tab}$ ⟶ on accepte H_1

Figure 17. Zones d'acceptation et de rejet de l'hypothèse H_0

Exemple : Le taux de cholestérol dans la population est connu et vaut 4,3 mg/l avec une Précision de 1,8 mg/l. Après 15 pesées, on trouve une moyenne de 4,6 mg/l.
Ce résultat est-il conforme à la valeur de la population ?

Solution

σ est connu σ = 1,8 mg/l, $\mu = 4.3\ mg/l$; n= 15 ; $\bar{x} = 4.6$

$$t_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{4.6 - 4.3}{\frac{1.8}{\sqrt{15}}} = 0.64$$

$$t_{tab\ (1\%et\ ddl=14)} = 3.977$$

$t_{obs} < t_{tab}$ ⇒ On accepte H_0 et on rejette H_1 donc Ce résultat est conforme à la valeur de la population

Exemple : Après 15 pesées, on trouve une moyenne de 1,42 g et un écart-type estimé de 0,5 g.
Ce résultat est-il conforme à la spécification de l'emballage à savoir poids net 1.50 g ?

Solution

σ est inconnu S = 0.5 g, $\mu = 1.5\ g$; n = 15 ; $\bar{x} = 1.42\ g$

$$t_{cal} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{1.42 - 1.5}{\frac{1.5}{\sqrt{15}}} = 0.20$$

$$t_{tab\ (5\%et\ ddl=14)} = 2.145$$

$t_{cal} < t_{tab}$ ⇒ On accepte H_0 et on rejette H_1 donc Ce résultat est conforme à la spécification de l'emballage à savoir poids net 1.50 g.

4.2. Test de conformité d'une distribution

Il s'agit de comparer une distribution d'un caractère observé sur un échantillon donné et une distribution théorique basée sur un modèle susceptible de décrire la probabilité d'observer une valeur du caractère. On dit parfois que l'on cherche à ajuster une distribution expérimentale à une distribution théorique.
L'hypothèse nulle consiste à supposer que l'n a concordance des deux distributions. Le critère du test est $X^2_{obs} = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$
Où O : est l'effectif observé

E : l'effectif théorique

Sous l'hypothèse nulle, le X^2_{obs} ainsi calculé devrait être nul (négligeable). Il sera d'autant plus grand que les deux distributions divergent.

Les hypothèses :

H_0⇒ la distribution observée conforme la distribution théorique.

H_1⇒ la distribution observée ne conforme pas la distribution théorique.

La statistique X^2_{obs} est par suite comparée à la valeur table X^2_{tab} (cette valeur lue de la table à un ddl= k-1)

Si $X^2_{obs}<X^2_{tab}$ alors l'hypothèse H_0 est jugée acceptable

Si $X^2_{obs}>X^2_{tab}$ alors on rejette l'hypothèse H_0 au risque α de se tromper.

Exemple 1

Les résultats des épreuves d'un examen à l'échelle nationale sont : 60% de reçus, 20%admissibles (admis à passer les épreuves orales)et 15% éliminés

Un établissement présente 160 élèves et obtient 75 reçus, 53 admissibles et 32 éliminés

Y a-t-il conformité entre ces résultats et ceux valables à l'échelle nationale ?

Solution 1

* Il s'agit d'un test X^2de conformité

* Les hypothèses

H_0⇒ il y a une conformité entre les résultats de cet établissement et ceux valables à l'échelle nationale

H_1⇒ il n'y a pas une conformité entre les résultats de cet établissement et ceux valables à l'échelle nationale

* Pour le calcul de X^2, on peut utiliser le tableau suivant :

Résultat	O	E	$(O-E)^2$	$(O-E)^2/E$
Reçus	75	160x 0,6=96	441	4,593
Admissibles	53	160x 0,25=40	169	4,225
Eliminés	32	160x 0,15=24	64	2,666

$$X^2_{obs} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} = 11{,}484$$

$X^2_{tab(0,05;ddl=3-1=2)}$ =5,99 ; $X^2_{tab(0,01;ddl=3-1=2)}$ = 9,21 ; $X^2_{tab(0,001;ddl=3-1=2)}$ = 13,82

* X^2_{obs} > X^2_{tab} ⇒ on accepte H_1 donc il n'y a pas une conformité entre les résultats de cet établissement et ceux valables à l'échelle nationale au seuil 0,05 et 0,01 et il y a une conformité au seuil 0,001 ($X^2_{obs} < X^2_{tab}$)

Exemple 2

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

Phénotype	Rouge	Rose	Blanc
Effectif	141	325	134

Donner les proportions théoriques de la répartition mendélienne pour les trois couleurs. Calculer la statistique de test pour le test du khi-deux.

Solution 2

Notons R l'allèle induisant la couleur rouge et B l'allèle induisant la couleur blanche. On suppose que les phénotypes "fleurs rouges", "fleurs roses" et "fleurs blanches" correspondent respectivement aux génotypes RR, RB et BB. Si on croise deux individus de génotypes respectifs RR et BB, on obtient forcément des individus de génotype RB à la première génération. À la seconde génération, on obtiendra théoriquement un quart de génotypes RR, la moitié de génotypes RB et un quart de génotypes BB ; on devrait donc observer théoriquement un quart de plantes à fleurs rouges, la moitié à fleurs roses, et un quart à fleurs blanches.

Les effectifs théoriques correspondants sont 150, 300, 150.

La statistique de test du khi-deux prend la valeur :

$$x^2_{obs} = \sum \frac{(O-E)^2}{E} = \frac{(141-150)^2}{150} + \frac{(325-300)^2}{300} + \frac{(134-150)^2}{150} = 4.33$$

Cette valeur doit être comparée aux quantiles de la loi du khi-deux de paramètre $3 - 1 = 2$. D'après le tableau de x^2, la valeur de x^2_{tab} égale à 5.99, et dépasse 4.33. On accepte l'hypothèse d'adéquation de la loi observée avec la loi théorique.

5. Comparaison des moyennes

5.1. Comparaison des moyennes de deux échantillons indépendants

L'objectif du test T de Student est de tester la différence entre les moyennes de deux échantillons indépendants. Soient deux échantillons avec n_1 et n_2 éléments respectivement : nous voulons savoir si la différence entre $\bar{x}_1$ et $\bar{x}_2$ reflète une différence significative des moyennes des populations statistiques dont sont extraits les échantillons, ou si l'écart observé n'est dû qu'aux fluctuations naturelles de l'échantillonnage. Nous calculerons alors une statistique t de Student à partir des données et nous déterminerons la probabilité de cette valeur à l'aide de la distribution de Student à ddl=n_1+n_2-2 degrés de liberté.

Conditions d'application :

- indépendance des observations
- normalité des distributions de données des échantillons

- homoscédasticité des échantillons (voir mini-test de F).

En fonction de la taille des échantillons, il existe deux méthodes de calcul :

• Petits échantillons (n_1 ou n_2 <30), le test statistique utilisé est un test t de Student.

• Grands échantillons (n_1 et n_2 >30), le test statistique utilisé est un test Z.

Student t test

Pour chaque série statistique, on calcule et dresse au sein d'un tableau :

* Moyenne $\bar{x}$

* Variance S^2

* Effectif n

Les hypothèses

H_0 : $\bar{x}_1 = \bar{x}_2$ c'est-à-dire les deux échantillons appartiennent à la même population
H_1 : $\bar{x}_1 \neq \bar{x}_2$ c'est-à-dire les deux échantillons n'appartiennent pas à la même population (ils appartiennent à deux populations différentes).

La valeur de test à calculer est donnée par les formules suivantes :

$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{S_{\bar{D}}}$ avec $\boldsymbol{S_{\bar{D}}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

.

Tableau 11. Différents cas du test t de Student (comparaison des moyennes)

<table>
<tr><td colspan="2">1. σ^2_1 et σ^2_2 connues ou estimées avec $n_1$et n_2 > 30</td></tr>
<tr><td>Conditions</td><td>lois</td></tr>
<tr><td>* Echantillons indépendants.
* Normalité.
* σ^2_1 et σ^2_2 connues ou estimées avec n_1 et n_2 grands> 30.</td><td>$Z_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$</td></tr>
<tr><td colspan="2">2. σ^2_1 et σ^2_2 inconnues</td></tr>
<tr><td>$n_1$et n_2 < 30 cas des petits échantillons</td><td>$n_1$et n_2 > 30 des grands échantillons</td></tr>
<tr><td>Cas 1. $S_1^2 = S_2^2$ et $n_1 = n_2$
$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{2s^2}{n}}}$ avec : ddl=2(n-1)</td><td rowspan="3">$Z_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$</td></tr>
<tr><td>Cas 2. $S_1^2 \neq S_2^2$ et $n_1 = n_2$
$t_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$ avec : ddl=2(n-1)</td></tr>
<tr><td>$n_1 \neq n_2$ ⇒on procède à un test F de fisher (test d'égalité des variances)
On teste les hypothèses : $H_0 \Rightarrow S_1^2 = S_2^2$ vis-à-vis $H_0 \Rightarrow S_1^2 \neq$ tout en calculant la statistique F_{obs}</td></tr>
</table>

$F_{obs}=\frac{S_g^2}{S_p^2}$ Puis on compare cette valeur à la valeur F_{tab} (valeur lue de la table F avec α seuil d'erreur ; ddl du numérateur ; ddl dénominateur) **Cas 3.** $S_1^2 = S_2^2 \; et \; n_1 \neq n_2$ $t_{obs} = \frac{\lvert\bar{x}_1-\bar{x}_2\rvert}{\sqrt{S_{\bar{p}}^2(\frac{1}{n_1}+\frac{1}{n_2})}}$ avec : ddl=n_1+ n_2-2 Et $S^2{}_{\bar{p}} = \left[(n_1-1)S_1^2+(n_2-1)S_2^2\right] / [n_1+n_2-2]$ variance pondérée	
***Cas* 4**. $S_1^2 \neq S_2^2$ et $n_1 \neq n_2$ $t_{obs} = \frac{\lvert(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)\rvert}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ Et **ddl**=$\frac{\left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1-1}+\frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2-1}}$ → appelé ddl corrigé	
t_{obs} : lue de la table t avec α seuil d'erreur et ddl selon les cas	Z_{tab} : avec α seuil d'erreur et dernière ligne de la table t

5.2. Comparaison des moyennes de deux échantillons appariés

Dans le cas d'appariement des observations (par exemple lorsque le même paramètre est mesuré sur les mêmes sujets avant et après un traitement ou si on visite plusieurs sites d'échantillonnage en mer et que l'on a fait des prélèvements aux mêmes profondeurs, ou encore si on recherche à comparer le rendement de deux variétés sur différents lieux), Le test t des échantillons appariés se déroule comme suit :

- ✓ calcul des d_i les différences entre les couples d'observations des deux échantillons
- ✓ calcul de la moyenne des différences $\bar{d} = \frac{\sum d_i}{n}$
- ✓ calcul de la variance des différence $S_d^2 = \frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}$
- ✓ calcul de l'écart type de la moyenne de différences $S_{\bar{d}} = \sqrt{\frac{S_d^2}{n}}$
- ✓ calcul de $t_{obs} = \frac{\bar{d}}{S_{\bar{d}}}$ puis on compare le t_{obs} au t_{tab} lue de la table t à n- 1 ddl

Puis on test les hypothèses suivantes :

$H_0 \Rightarrow \bar{d} = 0$ c'est-à-dire que les deux échantillons appartiennent à la même population

$H_1 \Rightarrow \bar{d} \neq 0$ c'est-à-dire les deux échantillons n'appartiennent pas à la même population (ils appartiennent à deux populations différentes).

Exercice 1

On a chargé un médecin de répondre à la question suivante : l'aspirine (acide acétylsalycilique = AAS) diminue-t-elle l'espérance de vie des patients asthmatiques ?

Ce médecin a récolté des données selon les critères suivants : individus asthmatiques et décédés de façon naturelle au cours des 5 dernières années. Les informations retenues sont l'âge au décès et si de l'aspirine a été recommandée au patient (Oui : O ; Non : N). Le tableau suivant présente un échantillon aléatoire des milliers de réponses obtenues. La distribution des données est normale.

Age au décès	AAS	Age au décès	AAS
45,6	O	69,7	N
45,85	O	51,48	O
48,45	O	51,56	O
48,63	O	55,19	O
48,74	N	55,32	N
49,6	N	57,8	O
51,4	O	58,59	O
60,86	N	58,63	N
52,06	O	58,89	O
53,16	N	59,18	O
54	O	59,24	O
65,16	N	60,53	O
56,93	N	64,86	N
57,38	O	65,81	N
57,94	N	67,72	O
67,96	N	68,8	N
58,24	O	69,58	N
68,61	N	72,66	N

- Réalisez un test statistique adapté afin de répondre à la question posée.

Solution 1

Il faut faire un test de comparaison de moyennes entre les groupes avec ou sans prise d'AAS : test t si les conditions d'applications sont respectées. Ces conditions sont :

- Variable quantitative. C'est le cas.
- Echantillon de taille suffisante. Il y a 36 observations, 19 dans un groupe et 17 dans l'autre.
- Normalité de la distribution: déjà supposée
- Indépendance des observations : elle est supposée. Elle dépend de l'échantillonnage qui a été bien réalisé dans ce sens puisque c'est un échantillon aléatoire parmi des milliers de

réponses que l'on étudie. Exemples de non indépendance : auto-corrélation spatiale des mesures, parenté des patients (proximité génétique confondante), etc.

- Homogénéité des variances Celle-ci doit être préalablement testée à l'aide d'un test F.
  Les hypothèses.
- $n_1 \neq n_1 < 30$

$H_0 \Rightarrow$ la consommation d'aspirine n'a pas un effet sur l'âge du décès des patients asthmatiques.

$H_1 \Rightarrow$ la consommation d'aspirine a un effet sur l'âge du décès des patients asthmatiques.

	O(avec AAS)	N(sans AAS)
1	45,6	69,7
2	45,85	48,74
3	51,48	55,32
4	48,45	49,6
5	51,56	60,86
6	48,63	58,63
7	55,19	53,16
8	57,8	65,16
9	51,4	56,93
10	58,59	64,86
11	52,06	65,81
12	58,89	57,94
13	59,18	67,96
14	54	68,8
15	59,24	69,58
16	60,53	68,61
17	57,38	72,66
18	67,72	
19	58,24	
$\sum_{i=1}^{n} x_i$	1041,79	1054,32
$\bar{x}$	54,83	62,02

Test F :

- Calcul des variances

$$S_1^2 = \frac{\sum_{i=1}^{n} x_i{}^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}}{n-1} = \frac{57704{,}1771 - \frac{1041{,}79^2}{19}}{19-1} = 32{,}319$$

$$S_2^2 = \frac{\sum_{i=1}^{n} x_i{}^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}}{n-1} = \frac{66283{,}3976 - \frac{1054{,}32^2}{17}}{17-1} = 55{,}982$$

- Calcul du F_{obs}

$F_{obs} = S_g^2 / S_p^2 = 55{,}982/32{,}319 = 1{,}732$

F_{tab} (16, 18) = 2,64 pour un test bilatéral à un seuil de 5 %

$F < F_{tab}$: on ne peut pas rejeter l'hypothèse nulle, on considère que les deux variances sont homogènes.

Test t :

On a moins de 30 observations par groupes, on utilise donc la formule du test t pour les petits échantillons. Le test est unilatéral car on veut savoir si l'espérance de vie est diminuée.

✓ $S_1^2 = S_2^2$ et $n_1 \neq n_2 \Rightarrow S^2{}_{\bar{D}} = S^2{}_{\bar{p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$.

$\Rightarrow S^2{}_{\bar{D}}\ S^2{}_{\bar{p}}\left(\frac{n_1+n_2}{n_1 n_2}\right)$.

Avec $S^2{}_{\bar{p}} = \left[(n_1-1)S_1^2 + (n_2-1)S_2^2\right] / \left[n_1+n_2-2\right]$

$S^2{}_{\bar{p}} = \left[(19-1)32{,}319 + (17-1)55{,}982\right] / \left[19+17-2\right] = 43.455$

$S^2{}_{\bar{D}} = S^2{}_{\bar{p}}\left(\frac{n_1+n_2}{n_1 n_2}\right) = 43.455\left(\frac{19+17}{19x17}\right) = 4.843$

$S_{\bar{D}} = \sqrt{S^2{}_{\bar{D}}} = 2.2007$

$t_{obs} = \frac{(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)}{S_{\bar{D}}} = \frac{(62{,}019-54{,}831)-0}{2.2007} = 3.266$

t_{tab} (34) = 1,691 pour un test unilatéral et α = 5 %

t_{tab} (34) = 2,032 pour un test bilatéral et α = 5 %

$t_{obs} > t_{tab}$: on rejette H_0, les deux moyennes sont significativement différentes, la consommation d'aspirine a un effet sur l'âge du décès des patient asthmatiques.

Exercice 2

Dans le cadre d'une étude sur l'efficacité d'un nouveau type de fertilisant, un chercheur a mesuré le rendement agricole (en Kg/ha) de 9 parcelles cultivables sélectionnées aléatoirement, après traitement avec l'ancien et le nouveau fertilisant. On admet que les données suivent une distribution normale.

Parcelle	1	2	3	4	5	6	7	8	9
Ancien fertilisant	1920	2020	2060	1960	1960	2140	1980	1940	1790
Nouveau fertilisant	2250	2410	2260	2200	2360	2320	2240	2300	2090

- Réalisez un test statistique approprié pour savoir si le nouveau fertilisant permet d'obtenir un meilleur rendement que l'ancien. Prenez un niveau α de 5 %.

Solution 2

Dans ce cas, nous avons affaire à des données appariées, et il faut utiliser le test t pour de telles données. Il demande les mêmes conditions d'applications que le test t pour les groupes indépendants.

On travaille sur les différences di calculées deux à deux entre groupes. Comme on veut savoir si le nouveau traitement est plus efficace que l'ancien (test unilatéral), on considère les différences dans le sens Nouveau – Ancien, que l'on attend positives en moyenne.

Parcelle	1	2	3	4	5	6	7	8	9
Ancien fertilisant	1920	2020	2060	1960	1960	2140	1980	1940	1790
Nouveau fertilisant	2250	2410	2260	2200	2360	2320	2240	2300	2090
Di	330	390	200	240	400	180	260	360	300

$\sum_{i=1}^{n} d_i = 2\,660$

Il faut calculer $t_{obs} = \bar{d}/S_{\bar{d}}$, avec $S_{\bar{d}}$ = erreur standard (de la moyenne des d_i)

$S_{\bar{d}} = S_d/\sqrt{n}$

$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ = 2 660 / 9 = 295,556; $S_d = \frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n}$ = 80,640; $S_{\bar{d}}$ = 80,640/$\sqrt{9}$ = 26,880

Donc t_{obs} = 295,556/26,880 = 10,995

t_{tab} = 1,860 pour un test unilatéral (car on teste un meilleur rendement du nouveau fertilisant) à un seuil de 5%

$t_{obs} > t_{tab}$: on rejette H_0, le nouveau fertilisant est plus efficace que l'ancien.

Exercice 3.

Deux groupes de 100 souris ont reçu les traitements A et B respectivement. Après traitement, la durée de vie moyenne dans le groupe A était de $\bar{x}$= 114j avec S_1^2= 410 tandis que dans le groupe B, la durée de vie moyenne était de $\bar{y}$= 119j avec S_2^2= 490. On veut tester

H_0: "la durée de vie moyenne est identique dans les groupes A et B" contre

H_1: "la durée de vie moyenne est différente dans les groupes A et B".

Solution 3.

Si l'on note X la variable représentant la durée de vie d'une souris dans le groupe A et si l'on note Y la variable représentant la durée de vie dans le groupe B, on veut tester

H_0:E[X] = E[Y] contre H_1:E[X] ≠ E[Y]. On fixe le risque à α= 0.05. Les 2 échantillons sont indépendants, n_1=n_2=n=100 donc on utilise un z-test avec variances estimées.

On calcule z_{obs}=$\frac{|\bar{x}-\bar{y}|}{\sqrt{\frac{S_1^2+S_2^2}{1}}} = \frac{|114-119|}{\sqrt{\frac{410+490}{100}}}$=1.67.

On détermine $Z_{\alpha/2}$ par P[N(0,1)≤$z_{\alpha/2}$] = 1−0.05/2 = 0.975, on lit dans la table t (dernière ligne) $z_{0,05/2}$= 1.96. On voit que |z|≤1.96 donc on ne rejette pas H_0.

6. Test d'indépendance

Il s'agit de comparer entre elles des distributions relatives à plusieurs échantillons afin de déterminer si les différences observées sont significatives, ou si elles sont dues à des fluctuations d'échantillonnage.

C'est un cas particulier du test du khi-deux d'ajustement, qui permet de tester l'indépendance de deux caractères discrets.

Dans le cas, les données figurent en général sur un tableau à double entrée, appelé tableau de contingence. Ce tableau présente les effectifs conjoints. À la ligne *i*, colonne *j*, on trouve n_{ij} qui est le nombre d'individus dans la classe *i* pour le premier caractère et dans la classe *j* pour le second. Si le nombre de modalités des deux caractères sont *r* et *s*, la table a *r* lignes et *s* colonnes. Les *effectifs marginaux* sont les sommes par ligne ou par colonne de la table de contingence ; $n_{i.} = \sum_j n_{ij}$ est le nombre total d'individus dans la classe *i* pour le premier caractère ; $n_{.j} = \sum_i n_{ij}$ est le nombre total d'individus dans la classe *j* pour le second caractère. Le nombre total d'individus est n= $\sum_i n_{i.}$= $\sum_j n_{.j}$

La statistique du test est :

$$T = n\left(-1 + \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i.}n_{.j}}\right) \text{ ou } x_{obs}^2 = \sum_{i=0}^{k}\frac{(O-E)^2}{E}$$

Sous l'hypothèse nulle où les deux caractères sont indépendants, *T* suit la loi du khi-deux de paramètre *ddl* = (*r*−1)(*s*−1).

Exercice.

Le centre de transfusion sanguine de Pau a observé la répartition suivante sur 5000 donneurs.

$O^+ \rightarrow$ 2291 $A^+ \rightarrow$ 1631 $B^+ \rightarrow$ 282 $AB^+ \rightarrow$ 79

$O^- \rightarrow$ 325 $A^- \rightarrow$ 332 $B^- \rightarrow$ 48 $AB^- \rightarrow$ 12

1. Écrire la table de contingence correspondant à ces observations.

2. Calculer la valeur prise par la statistique du test du khi-deux de contingence.

3. Au seuil de 1% que concluez-vous ?

Solution

1. la table de contingence de ces observations est la suivante :

Groupe Rhésus	O	A	B	AB	Total
Rhésus +	2291	1631	282	79	4283
Rhésus -	325	332	48	12	717
Total	2616	1963	330	91	5000

2. Calcul de la valeur prise par la statistique du test du khi-deux de contingence.

Groupes	O	E	O- E	$(O-E)^2$	$(O- E)^2/E$
O^+	2291	4283x2616/5000=2240,8656	50.1344	2513,45806	1,121646056
A^+	1631	4283x1963/5000=1681,5058	-50.5058	2550,83583	1,516994966
B^+	282	4283x330/5000=282,678	-0.678	0,459684	0,001626175
AB^+	79	4283x91/5000=77,9506	1.0494	1,10124036	0,014127414
O^-	325	717x2616/5000=375,1344	-50.1344	2513,45806	6,700153501
A^-	332	717x1963/5000=281,4942	50.5058	2550,83583	9,061770486
B^-	48	717x330/5000=47,322	0.678	0,459684	0,00971396
AB^-	12	717x91/5000=13,0494	-1.0494	1,10124036	0,084390114

$x^2_{obs} = \sum_{i=0}^{k} \frac{(O-E)^2}{E} = 18{,}5104$ ou

$$\boldsymbol{x^2_{obs}} = \mathbf{n}\left(-1 + \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i.}n_{.j}}\right) = 5000\left(-1 + \frac{2291^2}{2616\ x\ 4283} + \frac{1631^2}{1963\ x\ 4283} + \ldots + \frac{12^2}{717\ x\ 91}\right) = 18{,}104$$

Sous l'hypothèse d'indépendance, la statistique de test suit la loi de khi-deux de paramètre :
ddl= $(4-1)(2-1) = 3$. Le quantile d'ordre 0,99 de cette loi est 11,34.
Comme 18,5104 est supérieur, on conclut qu'il y a dépendance entre le groupe sanguin et le rhésus, au vu de ces données.

7. Comparaisons de deux distributions

Il s'agit de comparer les distributions d'un même caractère dans deux populations, observées sur deux échantillons. Les techniques statistiques utilisées dépendent du type de caractère étudié, qualitatif ou quantitatif, des tailles des échantillons et de s'ils sont indépendants ou non (appariés).
Pour un caractère qualitatif (à deux modalités ou plus) et des tailles d'échantillons suffisamment grandes (> 30) on utilise des tests du khi-deux (ou khi-carré X^2) qui consistent à comparer les proportions des différentes modalités.
Pour un caractère quantitatif, lorsque les distributions sont supposées normales, il suffit pour les comparer, de comparer leurs moyennes (indice de position ou de valeur centrale) et donc de procéder à un test de comparaison de deux moyennes basé sur la loi de Student, ou lorsque les tailles des échantillons sont suffisamment grandes (> 30) d'utiliser des tests basés sur les approximations normales des moyennes empiriques.
En revanche lorsque les distributions ne peuvent pas être considérées comme normales, et en général pour de petites tailles d'échantillons (< 30), il est préférable d'utiliser des tests dits non-paramétriques (distribution free) qui ne font pas d'hypothèse sur la forme des distributions et consistent à comparer l'ensemble des distributions (les fonctions de répartition) ou les médianes (indice de position ou de valeur centrale) de ces distributions.
La plupart de ces techniques se généralisent à la comparaison de plus de deux distributions.

7.1. Comparaison d'une distribution à une loi de référence : test d'ajustement de Kolmogorov-Smirnov

* Soit X une variable quantitative de loi F.

* On veut tester : $H_0 \Rightarrow F=F_0$ contre $H_1 \Rightarrow F \neq F_0$ i.e. pour au moins un x dans R, on a $F(x) \neq F_0(x)$.

* Soit $(X_1,\ldots,X_n)$ un échantillon de variable parente X. Soit F_n la fonctionde répartition empirique de X, égale au point x à $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i \leq x) = \frac{\text{nb d'observations inférieures ou égales à x}}{\text{n}}$.

* On rejette H_0 au risque α lorsque la statistique de test

$D=\sup_{x\in R}|F_n(x)-F_0(x)|$ dépasse une valeur dα définie par $P[Y_n>D_\alpha] =\alpha$ où la loi de la variable Y_n est tabulée selonαetn. Intuitivement, D mesure en chaque point la distance entre la loi empirique et la loi théorique proposée.

NB : dans le cas d'un ajustement à une loi normale, on préfèrera le testde Shapiro-Wil

Exercice

Un appareil de radiographie admet 5 réglages possibles en ce qui concerne le tirage, allant du plus clair au plus foncé. On veut tester l'hypothèse

H_0 selon laquelle la lisibilité de la radiographie est la même pour les 5 tirages possibles, au risque $\alpha=0,05$. Autrement dit, sous H_0, les préfèrences des médecins en ce qui concerne la lisibilité des radios est uniformément répartie sur les 5 tirages. Si l'on note F la loi théorique des préférences des médecins, on veut tester $H_0\Rightarrow F=U\{1,2,3,4,5\}$ contre $H_1\Rightarrow F\neq U\{1,2,3,4,5\}$. On demande à 10 médecins d'observer les 5 tirages différents d'une même radiographie. On obtient :

tirage sélectionné	1	2	3	4	5
nb de médecins	0	1	0	5	4

solution

On calcule

tirage sélectionné	1	2	3	4	5
nb de médecins	0	1	0	5	4
F_n	0/10	1/10	1/10	6/10	10/10
F_0	1/5	2/5	3/5	4/5	5/5
$\lvert F_n-F_0\rvert$	2/10	3/10	5/10	2/10	0

* On obtient la réalisation

$D_{obs}=\sup_{x\in R}|F_n(x)-F_0(x)|\Rightarrow D_{obs}= 5/10 = 0,500$

On lit dans la table que $P[Y_n>D_\alpha] = 0,05$ pour $D_\alpha= 0,409$ avec n= 10.

* On voit que $D_{obs}>D_\alpha$ donc on rejette H_0.

7.2. Comparaison deux distributions des échantillons indépendants : test de Kolmogorov-Smirnov

* Soit une variable X de fonction répartition F et soit Y une variable de f.r. G.

* On veut tester au risque α $H_0\Rightarrow F= G$ contre $H_1\Rightarrow F\neq G$ i.e. pour au moins un x dans R, on a $F(x) \neq G(x)$.

* Soit (X_1,...,X_{n1}) échantillon de variable parente X et soit (Y_1,...,Y_{n2}) échantillon de variable parente Y. Soit F_{n1} la f.r. empirique de X, définie au point x par

$$\mathbf{F_{n1}(x)} = \frac{1}{n1}\sum_{i=1}^{n1}(X_i \leq x) = \frac{\textbf{nb d'observations inférieures ou égales à x}}{\textbf{n}}.$$

et soit G_{n2} la f.r. empirique de Y, définie au point x par

$$\mathbf{G_{n2}(x)} = \frac{1}{n1}\sum_{i=1}^{n2}(X_i \leq x) = \frac{\textbf{nb d'observations inférieures ou égales à x}}{\textbf{n}}.$$

On rejette H_0 au risque α lorsque la réalisation de la statistique de test $\mathbf{D_{obs}} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\,\mathbf{sup_{x\in R}|F_{n1}(x) - G_{n2}(x)|}$ dépasse une valeur D_α. Pour n_1>12 et n_2>12, on considère que pour α= 0,05, D_α= 1.36.

* Intuitivement, d mesure en chaque point la distance entre la loi empirique de X et la loi empirique de Y.

Exercice

On s'intéresse à l'effet d'un médicament sur les infections des souris par une larve. 16 souris sont infectées par le même nombre de larves, puis réparties au hasard entre 2 groupes égaux. Le premier groupe reçoit le traitement, pas le second. Au bout d'une semaine, toutes les souris sont sacrifiées et les nombres suivants de versadultes ont été retrouvés dans les intestins :

souris traitées	44	47	49	53	57	60	62	67
souris non traitées	51	55	62	63	68	71	75	79

On veut conclure sur l'éventuelle efficacité du traitement.

Solution

On note X la variable représentant le nombre de vers chez une souris traitée et Y la variable représentant le nombre de vers chez une souris non traitée.

Soit F la loi de X et soit G la loi de Y.

On veut tester $H_0 \Rightarrow F=G$ contre $H_1 \Rightarrow F \neq G$ au risque α= 0,05.

nb de vers	44	47	49	51	53	55	57	60	62	63	67	68	71	75	79
G_n (SNT)	0	0	0	1/8	1/8	2/8	2/8	2/8	3/8	4/8	4/8	5/8	6/8	7/8	8/8
F_n (ST)	1/8	2/8	3/8	3/8	4/8	4/8	5/8	6/8	7/8	7/8	8/8	8/8	8/8	8/8	8/8
$\lvert Gn - Fn \rvert$	1/8	2/8	3/8	2/8	3/8	2/8	3/8	4/8	4/8	3/8	4/8	3/8	2/8	1/8	0

On obtient la réalisation :

$$D_{obs} = \sqrt{\frac{n_1 n_2}{n_1+n_2}}\,sup_{x\in R}|F_{n1}(x) - G_{n2}(x)| = \sqrt{\frac{8x8}{8+8}}x4/8 = 1$$

Pour n_1>12 et n_2>12, on considère que pour α= 0,05, d_α= 1.36

$D_{obs} < D_\alpha$ donc on ne rejette pas H_0

Tableau 12. Table statistique de Kolmogorov Smirnov

Taille de l'échantillon (n)	Seuils critiques $D_a(n)$				
	a = .20	a = .15	a = .10	a = .05	a = .01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	.356
25	.210	.220	.240	.270	.320
30	.190	.200	.220	.240	.290
35	.180	.190	.210	.230	.270
> 35	$1.07 / \sqrt{n}$	$1.14 / \sqrt{n}$	$1.22 / \sqrt{n}$	$1.36 / \sqrt{n}$	$1.63 / \sqrt{n}$

8. Régresion linéaire simple et multiple, changement de variable

8.1. Régression simple linéaire

Soit une distribution deux variables quantitatives. La régression linéaire simple permet de chercher l'éventuelle relation fonctionnelle linéaire qui existerait entre une valeur Explicative (ou indépendante) x et une variable aléatoire Expliquer (ou dépendante) y. x est remplacé par t s'il s'agit d'une mesure du temps. Graphiquement, on représente cette éventuelle relation dans un plan orthogonal. L'axe des abscisses indique la variable qui explicative et l'axe des ordonnées celle que l'on cherche à expliquer. L'ensemble des données figure sous forme de nuage de points (Il y a autant de points que d'observations différentes). Si les données sont disponibles en fourchettes de valeurs, on remplace ces dernières par les valeurs centrales des classes. Une relation linéaire déterministe entre les deux variables se traduit par des points parfaitement alignés. En mathématiques, on dit que la droite qui les relie représente une fonction affine (en statistiques, on emploie le terme linéaire plutôt qu'affine).

La régression linéaire simple cherche à modéliser cette relation par une équation et l'analyse de corrélation vise à en évaluer la qualité. Ce type d'analyse peut d'ailleurs être utilisé pour des relations non linéaires mais transformables en fonctions affines à condition d'utiliser des variables

auxiliaires (voir régression simple sur tendance exponentielle). En pratique, il est toute fois rare de passer par là puisque n'importe quel logiciel réalise des régressions non linéaires.
La régression permet de prédire la valeur de la variable dépendante pour une valeur donnée de la variable indépendante, implique une relation causale. Si on trace une ligne droite on obtient une droite dite droite de régression.

8.1.1. Droite de régression

Nous observons un nuage de forme plus ou moins rectiligne. Comment trouver l'équation de la droite qui le résume au mieux? En minimisant les distances qui la séparent des points. Quelles distances? Généralement les carrés des distances euclidiennes.
Plusieurs droites peuvent s'ajuster à un nuage de points mais parmi toutes ces droites on peut retenir celle qui jouit d'une propriété remarquable : celle qui minimise la somme carré des écarts des ordonnées observées.

Exemple : Si on projette des points M_1 à M_4 parallèlement à l'axe des y sur la droite on obtient les points P_1 P_4. Le critère retenu pour déterminer la droite D passant au peut près de tous les points sera tel que la somme des carrés des écarts des points observées M_i à la droite solution soit minimum.

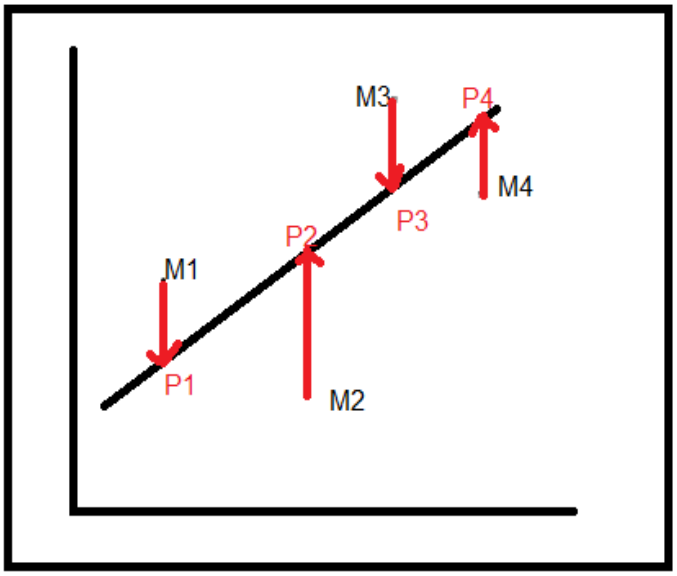


Figure 18. droite de régréssion et les écarts des points observées

La droite solution sera appelée la droite de régression de y sur x
La formule de la droite de régression peut s'écrire comme suit :

$$\hat{y} = b_0 + b_1 x \qquad \text{et} \qquad \hat{y} = \bar{y} + b_1 (x - \bar{x})$$

Avec : $\boldsymbol{b_1} = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n} = \frac{COV_{xy}}{S_x^2}$ est le coefficient de régression

$$\boldsymbol{b_0} = \bar{\boldsymbol{y}} - \boldsymbol{b_1}\bar{\boldsymbol{x}}$$

Le graphique suivant montre l'explication géométrique de la décomposition de la formule de la droite de régression

- La relation entre le coefficient de régression (b_1) et de corrélation (r)

$$b_1 = \frac{COV_{xy}}{S_x^2} \; ou \; b_1 = \frac{COV_{xy}}{\sigma_x^2}$$

$$r = \frac{COV_{xy}}{\sigma_x \sigma_y} = b_1 \frac{\sigma_x}{\sigma_y} = b_1 \frac{S_x}{S_y}$$

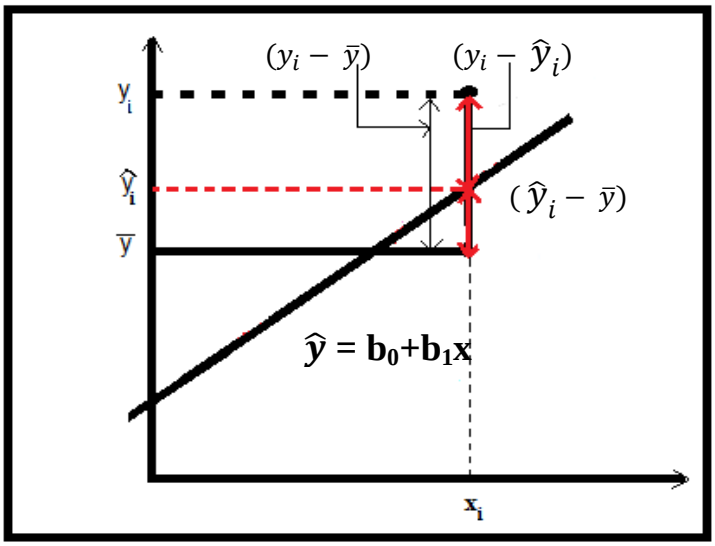


Figure 19. l'explication géométrique de la décomposition de la formule de la droite de régression

8.1.2. Interprétation

- La droite de régression exprime la meilleure prévision de la variable dépendante (*Y*), compte tenu des variables indépendantes (*X*). La nature étant rarement parfaitement prévisible (si toutefois elle l'est), il existe souvent des écarts substantiels entre les points observés autour de la droite de régression ajustée. L'écart d'un point particulier à la droite de régression (sa valeur prévue) est appelé résidu.
- Plus faible sera la dispersion des résidus autour de la droite de régression par rapport à la dispersion relative globale, meilleure sera notre prévision. Par exemple, s'il n'y a pas de relation entre les variables *X* et *Y*, le ratio entre la dispersion des résidus de la variable *Y* et la variance initiale sera égal à 1,0. Si *X* et *Y* sont parfaitement dépendantes, il n'y aura aucune variance des résidus et le ratio des variances sera égal à 0,0. Dans la plupart des cas, le ratio se situera entre ces deux extrêmes, c'est-à-dire entre 0 et 1. Ce ratio est appelé **R^2 ou coefficient de détermination** ($R^2 = \frac{SCEr}{SCEt} = \frac{b_1 COV_{xy}}{S_y}$). Cette valeur est immédiatement interprétable de la manière suivante. Si nous avons un R^2 de 0,4, nous savons que la dispersion des valeurs de *Y* autour de la droite de régression est 1-0,4 fois la variance initiale ; en d'autres termes, nous avons expliqué 40% de la dispersion initiale, et il reste 60% de

dispersion résiduelle. Dans l'idéal, nous souhaitons expliquer le plus possible, voire toute la dispersion initiale. La valeur du R^2 est un indicateur de la qualité d'ajustement du modèle aux données (par exemple, un R^2 proche de 1,0 indique que nous avons réussi à expliquer quasiment toute la dispersion grâce aux variables spécifiées dans le modèle).

- Habituellement, l'intensité de la relation entre deux prédicteurs ou plus (variables indépendantes ou *X*) et la variable dépendante (*Y*) s'exprime par **le coefficient de corrélation r (r = $\sqrt{R^2}$ ⇒ r = $\frac{cov_{xy}}{S_x.S_y}$),** qui est la racine carrée du R^2. En régression multiple, *r* peut prendre des valeurs comprises entre 0 et 1. Pour interpréter le sens de la relation entre des variables, il faut examiner le signe (plus ou moins) de la régression ou des coefficients b_1. Si un coefficient b_1 est positif, la relation entre cette variable et la variable dépendante est positive et si le coefficient b_1 est négatif, la relation sera négative. Naturellement, si le coefficient b_1 est égal à 0, il n'y aura aucune relation entre les variables.

8.1.3. Test de la pente de régression

Si la droite de régression est horizontale ($b_1 = 0$), alors cela signifie qu'il n'y a pas de lien entre x et y.

Les hypothèses testées : $H_0 \Rightarrow b_1 = 0$

$H_1 \Rightarrow b_1 \neq 0$

Pour tester ces hypothèses précédentes on doit calculer la variance :

$$S_b^2 = \frac{\frac{s_y^2}{s_x^2} - b_1^2}{n-2}$$

Puis on calcule $t_{obs} = \frac{b_1^2}{\sqrt{S_b^2}}$ (suit une loi de Student à (n-2) ddl)

Et on le compare avec $t_{\alpha/2}$ à (n-2) ddl

- Si $t_{obs} \geq t_{tab} \Rightarrow$ H_0 est rejetée, la pente est différente de l'horizontale (il y a un lien entre x et y)
- Si $t_{obs} < t_{tab} \Rightarrow$ H_0 est acceptée, la droite de régression ne s'écarte pas significativement de l'horizontale (il n'y a pas de lien entre x et y).

8.1.4. Test de linéarité

***** Le test de la pente ou test du coefficient de corrélation, suppose que le nuage de points observé a une allure linéaire, ou au moins une variation monotone (croissance ou décroissance)

***** Supposant que Y soit lié à X mais non de manière linéaire.

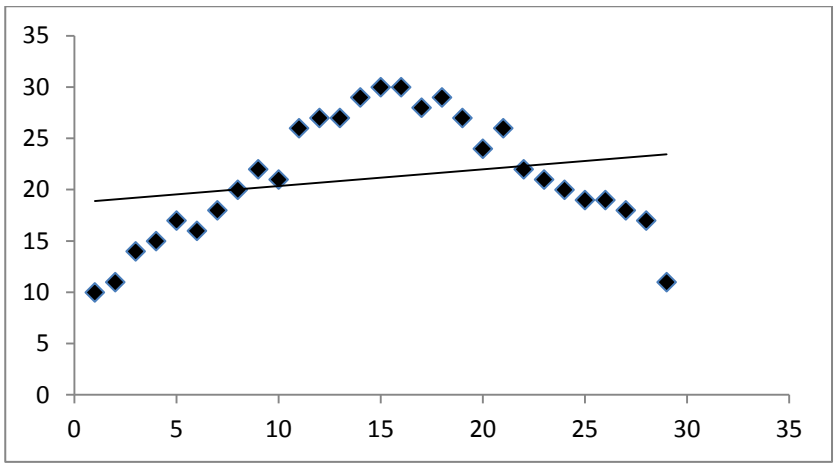


Figure 20. Régression non linéaire

***** La pente de la droite de régression observée risque d'être proche de 0, faisant conclure à tort à une liaison non significative. le test de linéarité permet de savoir si la courbe de régression vraie, dans l'intervalle étudié peut être considéré comme une droite. le test n'est valable que si l'on dispose de plusieurs valeurs de Y pour un X donné. pour ce test on teste deux hypothèses :

H_0 : La courbe de régression vraie est une droite

H_1 : La courbe de régression vraie n'est pas une droite

* Le principe de ce test consiste alors à comparer la variance « déviation par rapport à la droite de régression » et la variance résiduelle.

Apartir de la figure (19) on peut conclure que:

$\sum(y_i - \bar{y})^2 = \sum(\hat{y_i} - \bar{y})^2 + \sum(y_i - \hat{y_i})^2$

c'est-à-dire que $SCE_t = SCE_r + SCE_e \rightarrow$(r : régression modèle).

$SCE_t = \sum(y_i - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = \sum y^2 - n\bar{y}^2$

$SCE_{r\ ou\ m} = \sum(\hat{y_i} - \bar{y})^2 = \frac{[\sum x_i y_i - \sum x_i \sum y_i / n]^2}{\sum x_i^2 - (\sum x_i)^2/n} = \frac{COV_{xy}^2}{S_x^2} = b_1 COV_{xy}$

$SCE_e = SCE_t - SCE_{r\ ou\ m}$

ANOVA pour une régression simple (2 variables régression Y sur X) :

Tableau 13. Table d'ANOVA de la régression simple

Sources de variation	ddl	SCE	CME
Totale	n-1	$\sum x_i^2 - (\sum x_i)^2/n$	$\frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}$
Régression y sur x	1	$\frac{[\sum x_i y_i - \sum x_i \sum y_i /n]^2}{\sum x_i^2 - (\sum x_i)^2/n}$	$\frac{[\sum x_i y_i - \sum x_i \sum y_i /n]^2}{\sum x_i^2 - (\sum x_i)^2/n}/1$
Erreur	n-2	$SCE_t - SCE_s$	$\frac{SCE_e}{n-2}$

$ddl_r = 1$ car nous avons deux variables.

x : variable indépendante et y variable dépendante.

n : nombre d'observations

* pour tester la linéarité, on calcule le rapport $F_{obs} = \frac{S_r^2}{S_e^2}$

* On compareF_{obs} à F_{n-2}^1 lu sur la table de Fisher au seuil 5%

* Si $F_{obs} \geq F_{n-2}^1$ $\Rightarrow$ H_0 est regetée ; il y a un écart significatif par rapport à la linéarité(on précise le degré de liberté).

* Si $F_{obs} < F_{n-2}^1$ $\Rightarrow$ H_0 n'est pas regetée

Exercice

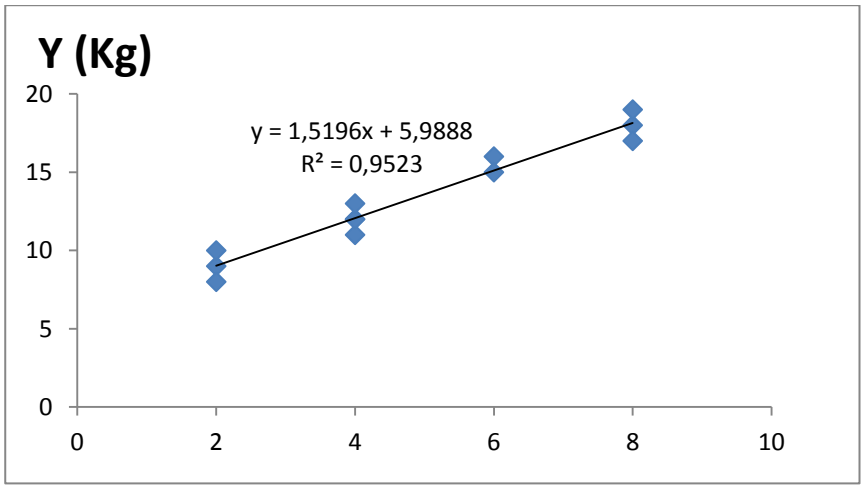
On s'intéresse à 12 moutons, traités par anabolisant et on veut savoir si l'augmentation de poids observée y est liée à la dose d'anabolisant ingéré x. les données sont résumées dans le tableau suivant :

mouton	1	2	3	4	5	6	7	8	9	10	11	12
x (mg/j)	2	2	2	4	4	4	4	6	6	8	8	8
Y (Kg)	9	10	8	12	11	12	13	16	15	18	17	19

- Représenter graphiquement ces résultats
- Donner la formule de la droite de régression puis tracer la
- Tester la ponte de régression

Solution

* Représentation graphique des données



- la formule de la droite de régression

Calcule de :

$\sum x_i = 58$; $\sum x_i^2 = 340$; $\sum y_i = 160$; $\sum x_i^2 = 2278$; $\sum x_i y_i = 864$

$$b_1 = \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sum x_i^2 - (\sum x_i)^2 / n} = \frac{864 - \frac{85x160}{12}}{340 - \frac{58^2}{12}} = 1,52$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{160}{12} - 1,52\frac{58}{12} = 5,99$$

La formule de la droite de régression est la suivante :

$\hat{y} = b_0 + b_1 x \Rightarrow \hat{y} = 5,99 + 1,52\,x$

- Test de pente de régression

On veut tester s'il y a un lien entre x et y

Les hypothèses

$H_0 \Rightarrow b_1 = 0$

$H_1 \Rightarrow b_1 \neq 0$

Pour tester ces hypothèses précédentes on doit calculer les variances :

- $(\sum x_i^2 - \frac{(\sum x_i)^2}{n})/(n-1) = 5,42$
- $(\sum y_i^2 - \frac{(\sum y_i)^2}{n})/(n-1) = 13,15$
- $S_b^2 = \frac{\frac{S_y^2}{S_x^2} - b_1^2}{n-2} = \frac{\frac{13,15}{5,42} - 1,52^2}{12-2} = 0,012$

Puis on calcule $t_{obs} = \frac{b_1}{\sqrt{S_b^2}} = \frac{1,52}{\sqrt{0,012}} = 13,8$

Et on le compare avec $t_{\alpha/2}$ à (12-2=10) ddl

$t_{obs} \geq t_{tab} \Rightarrow H_0$ est rejetée, la pente est différente de l'horizontale (il y a un lien entre l'augmentation de poids et la dose d'anabolisant).

8.2. Régression multiple

L'exemple développé à partir de deux variables permet de comprendre la logique de la théorie de la régression mais il ne peut être généralisé de la sorte aux régressions multiples. Le système à deux équations à deux inconnus présenté se résolvait facilement comme on l'a vu. Les équations se compliquent avec plusieurs régresseurs, deux méthodes distinctes permettent de résoudre les équations. La première repose sur la connaissance des coefficients de corrélation linéaire simple de toutes les paires de variables entre elles, de la moyenne arithmétique et des écarts-types de toutes les variables. La seconde repose sur des calculs matriciels.

8.2.1. Qunad utiliser la régression multiple ?

***** pour estimer la relation entre une variable dépendante (Y) et plusieurs variables indépendantes (X_1, X_2, ….X_P)

***** exemples :

- Expliquer le rendement d'une culture par les doses d'engrais, des apports d'eau et des densités de semis

- Expliquer la note moyenne en fin d'année par le QI quotient intellectuel et l'effectif de la classe.

8.2.2. Modèle général de régression multiple

* L'équation de la régression multiple précise la façon dont la variable dépendante est reliée aux variables explicatives :

$$\mathbf{Y=b_0+ b_1X_1+ b_2X_2+\ldots+ b_pX_p+\varepsilon}$$

Où : b_0, b_1, b_2, …b_p sont les paramètres ou pentes et ε est la constante d'ajustement ou un bruit aléatoire représentant le terme d'erreur.

* les termes de l'équation

$$\mathbf{Y_i=b_0+ b_1X_{1i}+ b_2X_{2i}+\ldots+ b_pX_{pi}+\varepsilon}$$

Y_i :la i^{ème} observation de Y la vriable à expliquée

b_0 : Terme constant

b_1X_{1i} : Influence de la variable expliquative X_1

b_2X_{2i} : Influence de la variable expliquative X_2

b_pX_{pi} : Influence de la variable expliquative X_p

ε : Résidu de la i^{ème} observation et $\varepsilon = \bar{y} - \mathbf{b_1}\,\bar{x}_1 - \mathbf{b_1}\,\bar{x}_2 - \ldots - \mathbf{b_1}\,\bar{x}_p$

Avec deux variables explicatives X1 et X2 et une variable à expliquer Y on a par exemple :

$$b_1 = \frac{(Var_{X_2} * Cov_{YX_1}) - (Cov_{YX_2} * Cov_{X_1X_2})}{(Var_{X_1} * Var_{X_2}) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * (r_{YX_1} - (r_{YX_2} * r_{X_1X_2}))}{\sigma_{X_1} * (1 - r^2_{X_1X_2})}$$

$$b_2 = \frac{(Var_{X_1} * Cov_{Y_{21}}) - (Cov_{YX_1} * Cov_{X_1X_2})}{(Var_{X_1} * Var_{X_2}) - Cov^2_{X_1X_2}} = \frac{\sigma_Y * (r_{YX_2} - (r_{YX_1} * r_{X_1X_2}))}{\sigma_{X_{21}} * (1 - r^2_{X_1X_2})}$$

* Le coefficient de corrélation multiple est alors donnée par :

$$R_{Y,X_1X_2} = \sqrt{\frac{(r^2_{YX_1} + r^2_{YX2} - 2(r_{YX_1} * r_{YX_2} * r_{X_1X_2}))}{1 - r^2_{X_1X_2}}} = r_{YY'}$$

8.2.3. La notation matricielle

Nous recherchons les estimations : $\hat{b}_0$, $\hat{b}_1$,……$\hat{b}_p$ des paramètres : b_0, b_1, …., b_p permettant de reconstituer au mieux les données y_i à partir des observations des p variables pour l'individu i.

Determiner les estimations : $\hat{b}_0$, $\hat{b}_1$,……$\hat{b}_p$ minimisant $\sum_{i=1}^{n}(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \ldots - \hat{b}_p x_{ip})^2$

En utilisant les notations matricielles ci-dessous (cas de deux variables explicatives):

$$\mathbf{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & \vdots & \vdots \\ 1 & x_{1,n-1} & x_{2,n-1} \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix}, \quad \mathbf{b}=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}, \quad \boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix}$$

Il s'agit dès lors de calculer le vecteur des estimateurs c: défini par l'égalité suivante :

$\hat{b}=(X'\mathrm{X})^{-1}\,X'y$

En notation matricielle X' signifie la matrice X transposée et X^{-1} la matrice inverse.

Donc $(X'X)^{-1}X'y$ donne les termes de l'équation multiple

Exercice

Vous avez les données suivantes :

y_i	X_{i1}	X_{i2}
4	1	6
4	3	6
8	4	7
12	6	7
12	6	9

Estimer les coeffitions de régression multiple b_0, b_1, b_2

Solution

L'équation de la régression multiple est donnée comme suit :

$$\mathbf{Y_i=b_0+ b_1X_{1i}+ b_2X_{2i}+\ldots+ b_pX_{pi}+\varepsilon}$$

On veut estimer les paramètres de cette équation donc :

$\hat{y}_i = \hat{b}_0+\hat{b}_1x_{i1}+\hat{b}_2x_{i2}$ sachant que l'estimation de l'erreur égale à zéro $\hat{\varepsilon}$ =0

Nous avons : $\hat{b}=(\mathrm{X}X')^{-1}\,X'y$

$$\begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = (X'\mathrm{X})^{-1}\,X'y$$

X=$\begin{pmatrix}1&1&6\\1&3&6\\1&4&7\\1&6&7\\1&6&9\end{pmatrix}$, $X'=\begin{pmatrix}1&1&1&1&1\\1&3&4&6&6\\6&6&7&7&9\end{pmatrix}$ ‘⇒transposer

$(X'\mathrm{X})=\begin{pmatrix}1&1&1&1&1\\1&3&4&6&6\\6&6&7&7&9\end{pmatrix}.\begin{pmatrix}1&1&6\\1&3&6\\1&4&7\\1&6&7\\1&6&9\end{pmatrix}=\begin{pmatrix}5&20&35\\20&98&148\\35&148&251\end{pmatrix}$

(3 ;5) (5 ;3)

$(X'\mathrm{X})^{-1}=\frac{1}{det(X'\mathrm{X})}(\mathrm{Com}(X'\mathrm{X}))'$ avec $(X'\mathrm{X})^{-1}$ $estl'inverse\ de\ la\ matrice$ $(X'\mathrm{X})$ et det : déterminant de la matrice (X’X) et Com : la comatrice

$\det(X'X)=\det\begin{pmatrix}5&20&35\\20&98&148\\35&148&251\end{pmatrix}=\begin{vmatrix}5&20&35\\20&98&148\\35&148&251\end{vmatrix}$

$\det(X'X)=+5\begin{vmatrix}98&148\\148&251\end{vmatrix}-20\begin{vmatrix}20&35\\148&251\end{vmatrix}+35\begin{vmatrix}20&35\\98&148\end{vmatrix}$

$\det(X'X)=+5(98*251-148*148)-20(20*251-35*148)+35(20*148-35*98)=1347+3200-1645=220\neq0$

on peut calcule l’inverse de (X’X) puisque det(X’X)≠0 c-à-d (X’X) est réversible

$\mathrm{Com(X'X)}=\begin{pmatrix}+\begin{vmatrix}98&148\\148&251\end{vmatrix}&-\begin{vmatrix}20&148\\35&251\end{vmatrix}&+\begin{vmatrix}20&98\\35&148\end{vmatrix}\\-\begin{vmatrix}20&35\\148&251\end{vmatrix}&+\begin{vmatrix}5&35\\35&251\end{vmatrix}&-\begin{vmatrix}5&20\\35&148\end{vmatrix}\\+\begin{vmatrix}20&35\\98&148\end{vmatrix}&-\begin{vmatrix}5&35\\20&148\end{vmatrix}&+\begin{vmatrix}5&20\\20&98\end{vmatrix}\end{pmatrix}=\begin{pmatrix}2694&160&-470\\160&30&-40\\-470&-40&90\end{pmatrix}$

$\mathrm{Com(X'X)}=\begin{pmatrix}2694&160&-470\\160&30&-40\\-470&-40&90\end{pmatrix}\Rightarrow(\mathrm{Com(X'X)})'=\begin{pmatrix}2694&160&-470\\160&30&-40\\-470&-40&90\end{pmatrix}$

La prime = inverser (Com(X’X)=⇒ $(\mathrm{Com(X'X)})'$ puisque la matrice est symétrique)

$(\mathrm{X'X})=\frac{1}{det(X'\mathrm{X})}(\mathrm{Com}(X'X))'=\frac{1}{220}\begin{pmatrix}2694&160&-470\\160&30&-40\\-470&-40&90\end{pmatrix}$

$(\mathrm{X'y})=\begin{pmatrix}1&1&1&1&1\\1&3&4&6&6\\6&6&7&7&9\end{pmatrix}.\begin{pmatrix}4\\4\\8\\12\\12\end{pmatrix}=\begin{pmatrix}40\\192\\296\end{pmatrix}$

(3 ;5) (5 ;1)

$\begin{pmatrix}\hat{b}_0\\\hat{b}_1\\\hat{b}_2\end{pmatrix}=(X'\mathrm{X})^{-1}X'y=\frac{1}{220}\begin{pmatrix}2694&160&-470\\160&30&-40\\-470&-40&90\end{pmatrix}.\begin{pmatrix}40\\192\\296\end{pmatrix}=\frac{1}{220}\begin{pmatrix}-640\\320\\160\end{pmatrix}$

$$\begin{pmatrix}\hat{b}_0\\ \hat{b}_1\\ \hat{b}_2\end{pmatrix}=\begin{pmatrix}-2{,}91\\ 1{,}455\\ 0{,}727\end{pmatrix}$$

$$\hat{y}_i = -2{,}91 + 1{,}455x_{i1} + 0{,}727x_{i2}$$

8.2.4. ANOVA pour une régression multiple (+ 2 variables)

Analyse de la variance d'une régression multipe teste deux hypothèses qui sont :

H_0 : $b_1=b_2=\ldots= b_p$

H_0 : au moins un des $b_i\neq 0$

Tableau 14. Table d'ANOVA des régression multiple

Sources de variation	Ddl	SCE	CME
Totale	n-1	$\sum(\hat{y}_i - \bar{y})^2$	–
Régression y sur x	p-1	$\sum(y_i - \hat{y}_i)^2$	$SCE/(p-1)$
Erreur	n-p-1	$\sum(y_i - \bar{y})^2$	$\frac{SCE_e}{n-p-1}$

Comment apprécier globalement la régression ?

Les deux quantité SCE_t et $SCE_{m\ ou\ r}$ sont des sommes de carrés donc toujours positives ou nulles et tel que $SCE_{m\ ou\ r} \leq SCE_t$.

Décomposition de la somme des carrés des écart se fait comme suit :

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

S.C. totale = S.C. résiduelle + S.C. expliquée

pour tester la linéarité, on calcule le rapport $F_{obs} = \frac{SCE/(p-1)}{\frac{SCE_e}{n-p-1}}$

* On compareF_{obs} à F_{n-p-1}^{p-1} lu sur la table de Fisher au seuil 5%

* Si $F_{obs} \geq F_{n-2}^{1}$ ⇒ H_0 est regetée et Si $F_{obs} < F_{n-2}^{1}$ ⇒ H_0 n'est pas regetée

*** Coefficient de détermination**

La formule précédente nous suggère de mesurer la qualité de la régression multiple par le rapport :

$R^2 = \frac{SC\ expliquée}{SC\ totale} \Rightarrow R^2 = 1 - \frac{SC\ résiduelle}{SC\ totale}$

La racine carrée R de R^2 est le coefficient de corrélation multiple entre la variable à expliquer Y et les variables explicatives X_1, ... X_p.

* $0 \leq R^2 \leq 1$: R^2 proche de 1 signifie : Y est "bien expliquée" par les variables X_1 ... X_p

*** Coefficient de détermination justé**

Prenons garde au fait que le coefficient de détermination dont les coefficients de régression constituent en quelque sorte la contribution, croit avec le nombre de variable.Par conséquent, ce comportement déterministe lié aux propriétés des variables aléatoires doit être compenser, on calcule alors le coefficient ajusté

$$R^2 ajusté = 1 - \frac{(n-1)}{n-(p-1)-1}(1-R^2)$$

p :nombre de variables explicatives

R_a^2: permet de comparer des modèles où le nombre de variables explicatives. Il tient compte du nombre de degrés de liberté du modèle. Plus R_a^2 ajusté est élevé, plus la variance des résidus est faible

8.3. Changement de variable

En mathématiques, et plus précisément en analyse, l'**intégration par changement de variable** est un procédé d'intégration qui consiste à considérer une nouvelle variable d'intégration, pour remplacer une fonction de la variable d'intégration initiale. Ce procédé est un des outils principaux pour le calcul explicite d'intégrales. Il est parfois appelé **intégration par substitution** en lien avec le nom anglais du procédé.

9. Analyse de variance à un et deux facteurs

lorsque le nombre d'échantillon augmente (+3), la comparaison des moyennes utilisant le test t n'est plus commode. Dans ces conditions l'utilisation du dispositif expérimental qui se base sur le test F pour la comparaison des moyennes multiples est indispensable.

L'ANOVA est un test statistique de comparaison de moyenne qui généralise le test de comparaison de deux moyennes.

L'ANOVA test deux hypothèses qui sont :

$H_0 \Rightarrow \mu_1 = \mu_2 = \ldots\ldots = \mu_p$

$H_1 \Rightarrow \mu_1 \neq \mu_2 \neq \ldots\ldots \neq \mu_p$ donc il y a au moins une moyenne qui diffère des autres

Pour savoir la ou lesquelles , il faut avoir recours par la suite aux tests de comparaisons multiples.

9.1. Notion de base en expérimentation

* Un facteur : une série d'éléments de même nature susceptibles d'influencer les résultats d'une expérience. C'est l'effet qu'on veut étudié ses ou niveaux sont volontairement choisi

* Un traitement :procédure dont l'effet mesuré, la combinaison de deux ou plusieurs modalités variantes ou niveaux des facteurs étudiés constitue un traitement.

* Répétition: elle a pour fonction de permettre une estimation de l'erreur

* Notion d'erreur : l'erreur expérimentale est une imprécision qui entraine une hétérogénéité inévitable dans l'expérience. Elle regroupe la résultante de toutes les causes non contrôlées de l'essai, elle a pour origine : le manque de l'uniformité du terrain, le manque l'uniformité, le manque de précision des appareils et de l'expérimentateur lui-même.

* Notion unité expérimentale : C'est l'élément de base d'une expérience qui est considéré individuellement durant tout le processus expérimental. Une unité est soumise à un même traitement et conduit à la même observation.

* Notion d'interaction : On parle d'interaction entre deux facteurs A et B quand l'effet du facteur A sur la réponse va dépendre de la valeur du facteur B. L'interaction peut être synergique ou antagoniste. Elle peut, aussi, être significative ou non significative.

* Randomisation : l'allocation des traitements aux unités doit être faite par un tirage aléatoire.

9. 2. Conditions d'application d'ANOVA

- Variable dépendante quantitative
- indépendance des observations
- normalité de la distribution de la population d'où tiré chaque groupe et cela se fait de plusieurs manières – histogramme de fréquence – les deux coefficients d'asymétrie et d'aplatissement – la méthode des pourcentages (50%, 68%, 99% et 99,7%) (voir chapitre 1)

NB: dans le cas d'asymétrie du distribution, on peut faire une transformation ou changement de variable (logarithme, racine ….)

Ces transformations peuvent se faire via :

- Logarithme $\ln(x)$ pour des données strictement positives, ou encore $\ln(x+1)$ si la variable x prend des valeurs positives ou nulles
- Racine $\sqrt{x}$ pour des données positives ou nulles, ou encore $\sqrt{x+c}$ si x prend une ou des valeurs négatives
- Inverse $\frac{1}{x}$

Il ne nous reste qu'à effectuer à nouveau un test de normalité afin de vérifier si la transformation est adéquate.

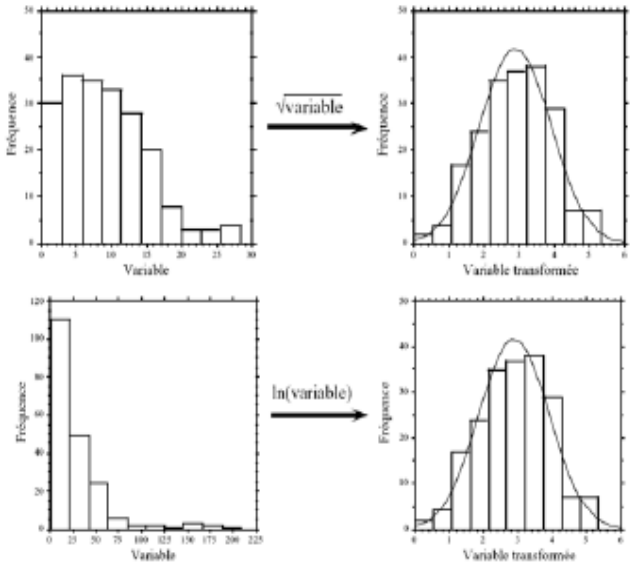


Figure 21.Différentes transformations des dnnées

- Homoscédasticité: ou test d'égalité des variances

* si les groupes ont la même taille on utilise le test Hartley := $H_{obs} = \frac{S_G^2}{S_P^2}$ à comparer avec H_{tab} (α=5%, na , n ou ddl)

* Si les groupes des tailles différentes on utilise le test Fisher ($F_{obs}=\frac{S_G^2}{S_P^2}$ à comparer avec F_{tab} (α=5%, ddl_{Sg}, ddl_{Sp})

9.3. Notion du dispositif expérimental

Un dispositif expérimental est un ensemble des parcelles (ou unités expérimentales) dont la répartition particulière et caractéristique permet d'étudier un ou plusieurs facteurs. L'objectif du dispositif expérimental est d'obtenir un essai d'une puissance maximale et d'une interprétation simple

Le choix d'un dispositif expérimental est fonction de trois critères ; le nombre de facteurs étudiés, le nombre de gradients d'hétérogénéité et les contraintes liées à l'expérimentation (mise en place, conduite, observations…). Sur la base de ces trois critères nous avons :

- 1 facteur étudié + aucun gradient d'hétérogénéité = dispositif en randomisation totale
- 1 facteur étudié + 1 gradient d'hétérogénéité = dispositif en blocs aléatoires complets.
- 1 facteur étudié + 2 gradients d'hétérogénéités = dispositif en carré latin.
- 2 facteurs étudiés + 1 gradient d'hétérogénéité = dispositif en factoriel bloc.
- 2 facteurs étudiés + 1 gradient d'hétérogénéité + 1 contrainte expérimentale = dispositif en split-plot

9.4. Rappel sur les dispositifs expérimentaux à un facteur étudié

- L'analyse de variance à un facteur (*one-way analysis of variance*) va consister à chercher le rapport entre la variance entre les groupes (V. inter-groupe) et la variance à l'intérieur des groupes (V. intra-groupe).
- La valeur de ce rapport appelé F [*attention* : ce F n'a rien a voir avec le F du test de vérification de l'homogénéité des variances] est comparée à celle d'une table de F de Snedecor, table à double entrée.

Tableau 15. Principaux dispositifs expérimentaux à un facteur étudié

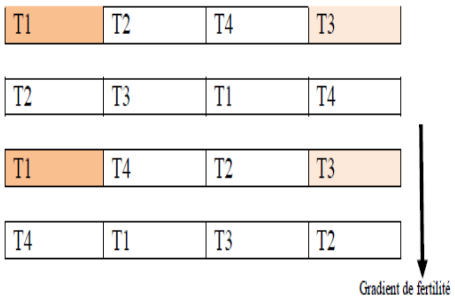
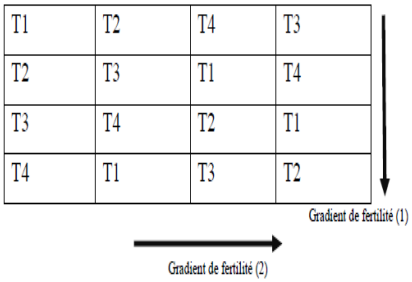
dispositif complètement randomisé (DCR)	**dispositif en bloc complètement randomisé (DBCR)**	**dispositif en carré latin ou le double bloc**
Condition d'utilisation * appliqué dans les terrains assez homogènes. * L'affectation des traitements se fait par tirage au sort complètement aléatoire. *Le même traitement peut apparaître plus d'une fois dans la même ligne et dans la même colonne.	**Condition d'utilisation** *appliqué dans le cas des terrains hétérogènes avec la présence d'un gradient d'hétérogénéité (pente, cours d'eau, drains, brises vents etc…). *Les blocs doivent être allongés perpendiculairement au sens du gradient d'hétérogénéité Chaque bloc doit comporter l'ensemble des traitements.	**Condition d'utilisation** *appliqué dans le cas des terrains est très hétérogènes avec la présence d'un gradient d'hétérogénéité (pente, cours d'eau, *Nombre de colonnes = N^{bre} de rangs = N^{bre} de traitements= N^{bre} de répétitions.
Plan d'expérience 	**Plan d'expérience** 	**Plan d'expérience** 
Modèle additif $x_{ij}= \mu + \tau_i + e_{ij}$ e_{ij}=erreur relative au traitement i et à la répétition j μ=moyenne générale de l'essai τ_i = effet du niveau i du facteur F	**Modèle additif** $x_{ij} = \mu + \tau_i + B_j + e_{ij}$ e_{ij}=erreur relative au traitement i et à la répétition j μ=moyenne générale de l'essai τ_i = effet du niveau i du facteur F B_j= effet du niveau j du facteur bloc	**Modèle additif** $x_{ij} = \mu + \tau_i + L_j + C_k + e_{ijk}$ e_{ij}=erreur relative au traitement i et à la répétition j μ=moyenne générale de l'essai τ_i = effet du niveau i du facteur F L_j= effet du niveau j du facteur ligne. C_k= effet du niveau k du facteur ligne.

Tableau 16. Table d'ANOVA du dispositif ou randomisation totale (DCR)

Sources de variation	ddl	SCE	CME	F_{obs}	F_{tab}
Totale	tr-1	SCE_t	-	-	-
Traitements	t-1	SCE_{tr}	$\mathrm{SCE}_{tr}/\mathrm{t}-1$	$\mathrm{CME}_{tr}/\mathrm{CME}_e$	$F_{\mathrm{t}(\mathrm{r}-1)}^{\mathrm{t}-1}$
Erreur	t(r-1)	SCE_e	$\mathrm{SCE}_e/\mathrm{t}(\mathrm{r}-1)$	-	-

Avec :

- $C = \frac{(\sum X_{ij})^2}{t.r}$
- $SCE_t = \sum x_{ij}^2 - C$
- $SCE_{tr} = \frac{\sum x_{i.}^2}{r} - C$
- $SCE_e = SCE_t - SCE_{tr}$

Tableau 17. Table d'ANOVA pour le dispositif en bloc complètement randomisé (DBCR)

Sources de variation	ddl	SCE	CME	F_{obs}	F_{tab}
V. totale	tr-1	SCE_t	-	-	-
V. traitements	t-1	SCE_{tr}	$\mathrm{SCE}_{tr}/\mathrm{t}-1$	$\mathrm{CME}_{tr}/\mathrm{CME}_e$	$F_{tab}{}_{(\mathrm{t}-1)(\mathrm{r}-1)}^{\mathrm{t}-1}$
Bloc	r-1	SCE_b	$\mathrm{SCE}_b/\mathrm{r}-1$	$\mathrm{SCE}_b/\mathrm{CME}_e$	$F_{tab}{}_{(\mathrm{t}-1)(\mathrm{r}-1)}^{\mathrm{r}-1}$
Erreur	(t-1)(r-1)	SCE_e	$\mathrm{SCE}_e/(\mathrm{t}-1)(\mathrm{r}-1)$	-	-

Avec :

- $C = \frac{(\sum X_{ij})^2}{t.r}$
- $SCE_t = \sum x_{ij}^2 - C$
- $SCE_{tr} = \frac{\sum x_{i.}^2}{r} - C$
- $SCE_b = \frac{\sum x_{.j}^2}{t} - C$
- $SCE_e = SCE_t - SCE_{tr} - SCE_b$

Tableau 18. Table d'ANOVA pour le dispositif en carré latin ou le double bloc (DCL)

Sources de v	ddl	SCE	CME	F_{obs}	F_{tab}
Totale	r^2-1	SCE_t	-	-	-
Rangs	r-1	SCE_r	$\frac{SCE_r}{ddl_r} = \frac{SCE_r}{r-1}$	$\frac{CME_r}{CME_e}$	$F_{tab}{}^{r-1}_{(r-1)(r-2)}$
Colonnes	r-1	SCE_c	$\frac{SCE_c}{r-1}$	$\frac{CME_c}{CME_e}$	$F_{tab}{}^{r-1}_{(r-1)(r-2)}$
Traitements	r-1	SCE_{tr}	$\frac{SCE_{tr}}{r-1}$	$\frac{CME_{tr}}{CME_e}$	$F_{tab}{}^{r-1}_{(r-1)(r-2)}$
Erreur	(r-1)(r-2)	SCE_e	$\frac{SCE_e}{(r-1)(r-2)}$	-	-

Avec :

- $C = \frac{(\sum X_{ij})^2}{t.r}$
- $SCE_t = \sum x_{ij}^2 - C$
- $SCE_{tr} = \frac{\sum x_i^2}{r} - C$
- $SCE_c = \frac{\sum x_c^2}{r} - C$
- $SCE_r = \frac{\sum x_r^2}{r} - C$
- $SCE_e = SCE_t - SCE_{tr} - SCE_c - SCE_r$

Comparaison des moyennes multiples

Si on accepte H_1 n doit procéder à la comparaison des moyennes en utilisant la plus petite différence significative ppds

$ppds_{5\%} = t_{5\%}\sqrt{\frac{2.CME_e}{r}}$

Exercice 1

Pour définir l'impact de la nature du sol sur la croissance d'une plante X, un botaniste a mesuré la hauteur des plantes dans 4 types de sol. Pour chaque type de sol, il disposait de 3 réplicas.

Type I	Type II	Type III	Type IV
15	25	17	10
9	21	23	13
4	19	20	19

Que peut-on conclure sur cette expérience?

Solution 1

Il s'agit d'un DCR (le manque d'un facteur contrôlé)

Analyse de la variance

RAPPORT DÉTAILLÉ

Groupes	Nombre d'échantillons	Somme	Moyenne	Variance
Type I	3	28	9,33	30,3333333
Type II	3	65	21,67	9,33333333
Type III	3	60	20	9
Type IV	3	42	14	21

- $C = \frac{(\sum X_{Ij})^2}{T.r} = \frac{195^2}{4x3} = 3108,75$
- $SCE_T = \sum X_{Ij}^2 - C$ = 3597-3108,75= 428,25
- $SCE_{Tr} = \frac{\sum X_L^2}{R} - C = \frac{10373^2}{3} - 3108,75 = 288,92$
- $SCE_E = SCE_T - SCE_{Tr} = 428,25 - 288,92 = 139,33$

ANALYSE DE VARIANCE

Source des variations	SCE	Ddl	CME	Fobs	Probabilité	Ftab
Entre Groupes	288,92	3	96,31	5,53	0,02370	4,066
A l'intérieur des groupes	139,33	8	17,42			
Total	428,25	11				

$F_{obs} > F_{obs} \Rightarrow On\ accepte\ H_1\ donc\ effet\ type\ de\ sol\ est\ significatif\ sur\ la\ croissance\ des\ plante$

Comparaison des moyennes

$ppds_{5\%} = t_{5\%}\sqrt{\frac{2.CME_e}{r}}$ = 2,306 $\sqrt{\frac{2x17,42}{3}}$ = 7,86

Type II	65	] Groupe A
Type III	60	
Type IV	42	Groupe B
Type I	28	Groupe C

Exercice 02.

Pendant la cuisson, les croissants absorbent la graisse en quantité variable. Nous avons relevé la quantité de graisse absorbée lors de la cuisson de six fournées de croissants pour quatre types de graisse. Les mesures sont présentées dans le tableau ci-dessous.

graisse / fournée	1	2	3	4
1	64	78	75	55
2	72	91	93	66
3	68	97	78	64
4	77	82	71	64
5	56	85	63	70
6	65	77	76	68

1. Quels sont les modèles que vous pouvez utiliser pour analyser ces données ?

2. Nous nous intéressons uniquement à ces quatre types de graisse. Par contre, nous cherchons à savoir si il y a un effet des fournées en général sur la quantité de graisse utilisée. Pour cela, quel est donc le modèle à choisir parmi les deux étudiés ? Détailler des hypothèses relatives à ce modèle.

Nous allons maintenant travailler avec le modèle que nous venons de choisir. Construire le tableau d'analyse de la variance pour ce modèle.

3. La fournée a-t-elle un effet significatif sur la quantité de graisse absorbée ?

4. La graisse a-t-elle un effet significatif sur la quantité de graisse absorbée ? Que concluez-vous ?

Solution 2

1. les modèles que nous pouvons utiliser pour analyser ces données sont

$x_{ij} = \mu + \tau_i + e_{ij}$ pour un dispositif totalement randomisé et $x_{ij} = \mu + \tau_i + B_j + e_{ij}$ pour un dispositif en blocs complets

2. le modèle choisi est le modèle en blocs complets

Les hypothèses :

$H_0 \rightarrow$ les types de graisse n'ont pas un effet significatif sur la quantité de graisse absorbée par les croissants hors de la cuisant.

$H_1 \rightarrow$ les types de graisse ont un effet significatif sur la quantité de graisse absorbée par les croissants hors de la cuisant.

Pour répondre aux questions 3 et 4 on dit faire l'analyse de la variance

Analyse de variance: deux facteurs sans répétition d'expérience

RAPPORT DÉTAILLÉ	*Nombre d'échantillons*	*Somme*	*Moyenne*	*Variance*
fournée 1	4	272	68	111,333333
fournée 2	4	322	80,5	183
fournée 3	4	307	76,75	216,916667
fournée 4	4	294	73,5	60,3333333
fournée 5	4	274	68,5	153,666667

fournée 6	4	286	71,5	35
graisse 1	6	402	67	52
graisse 2	6	510	85	60,4
graisse 3	6	456	76	97,6
graisse 4	6	387	64,5	27,1

- $C = \frac{(\sum x_{ij})^2}{t.r} = \frac{(1755)^2}{24}$ = 128334,38
- $SCE_t = \sum x_{ij}^2 - C$= 131087-128334,38 =2752,63
- $SCE_b = \frac{\sum x_j^2}{t} - C = \frac{515225^2}{4} - 128334{,}38$= 471,88
- $SCE_{tr} = \frac{\sum x_i^2}{r} - C = \frac{779409^2}{6} - 128334{,}38$= 1567,13
- $SCE_e = SCE_t - SCE_{tr} - SCE_b$=2752,63-471,88-1567,13=713,63

ANALYSE DE VARIANCE

Source des variations	*Somme des carrés*	*Degré de liberté*	*Moyenne des carrés*	*F*	*Valeur critique pour F*
Effet fournée	471,875	5	94,375	1,98	2,90
Effet type de graisse	1567,125	3	522,375	10,98	3,29
Erreur	713,625	15	47,575		
Total	2752,625	23			

3. La fournée n'a pas un effet significatif sur la quantité de graisse absorbée

4. La graisse a un effet très hautement significatif sur la quantité de graisse absorbée, on doit donc procéder à une comparaison des moyennes

$ppds_{5\%} = t_{5\%}\sqrt{\frac{2.\mathrm{CME}_e}{r}}$

$ppds_{5\%} = 2{,}1448\sqrt{\frac{2x47{,}575}{6}} = 8{,}56$

Graisse 2 : 85] groupe A

Graisse 3 : 76] groupe B

$\left.\begin{matrix}\text{Graisse1} : 67 \\ \text{Graisse4} : 64{,}5\end{matrix}\right]$ groupe C

On peut conclure que le type de graisse 2 donne la plus grande quantité de graisse absorbées par les croissants suivi par le type 3, alors que les deux autres type de graisse 1 et 4 représentent des quantités de graisse relativement plus faibles.

Exercice 03

Dans une expérience de plantation forestière conduite en plan carré latin, nous mesurons la taille des arbres de 04 variétés

Les résultats sont montrés dans le tableau suivant :

C : 11	B : 23	A : 31	D : 12
B : 24	A : 33	D : 12	C : 14
A : 32	D : 11	C : 13	B : 21
D : 13	C :12	B : 22	A : 33

Quelle variété des 04 variétés qui possède la hauteur la plus grande ?

Solution

	I	II	III	IV	$\sum colonne$
I	11,00	23,00	31,00	12,00	77,00
II	24,00	33,00	12,00	14,00	83,00
III	32,00	11,00	13,00	21,00	77,00
IV	13,00	12,00	22,00	33,00	80,00
$\sum ligne$	80,00	79,00	78,00	80,00	317,00

	A	B	C	D
$\sum Y_{i.}$	129,00	90,00	52,00	46,00
$\bar{Y}_{i.}$	32,25	22,50	13,00	11,50

- C= $\frac{(\sum Y_{ij})^2}{rxr} = \frac{317^2}{4X4}$ = 6280,56
- SCE_t= $\sum(Y_{ij})^2$ - C = 1120.44
- $SCE_{tr} = \frac{\sum(Y_{i.})^2}{r}$ – C = $\frac{32.25^2+22.5^2+13^2+11.54^2}{4}$ – 162.56= 1109.69
- SCE_l= $\frac{\sum l^2}{r} - c = \frac{80^2+79^2+78^2+80^2}{4} - 6280{,}56$= 0.69
- SCE_c= $\frac{\sum c^2}{r} - c = \frac{77^2+83^2+77^2+80^2}{4} - 6280{,}56$= 6.19
- SCE_e=SCE_t-$SCE_l - SCE_c - SCE_{tr}$= 3.88

ANOVA :

Sources de v	ddl	SCE	CME	F_{obs}	F_{tab}	
Totale	4^2-1=15	1120,44	-	-	-	
Traitements	4-1=3	1109,69	369,90	572,7	4.76	$\frac{3}{6}$
Rangs	4-1=3	0,69	0,23	0,35	4.76	$\frac{3}{6}$
Colonnes	4-1=3	6,19	2,06	3,19	4.76	$\frac{3}{6}$
Erreur	(4-1)(4-2) =6	3,88	0,65	-	-	

$F_{tab} > F_{obs} \Rightarrow$ On accepte H_1 et on rejette H_0 donc effet traitement est significatif c-à-d qu'ils y a des différences significatives entre les moyennes des hauteurs des variétés étudiées

Comparaison des moyennes

$$ppds_{5\%} = t_{5\%}\sqrt{\frac{2CME_e}{r}} = 2.447\sqrt{\frac{2X0.69}{4}} = 1.39$$

A: 32.25 → *A*

***B*: 22.50 →B**

C: 19.56 → C

***D*: 31.74 → D**

9.5. Dispositifs expérimentaux à deux facteurs à étudiés

9.5.1. Dispositif factoriel

Un facteur est un traitement qui contribue à l'expérience avec plusieurs niveaux exp : si la ration est un facteur à étudier, on trouve plusieurs types de rations dans l'expérimentation.

Ce dispositif peut être illustré ainsi : supposant qu'on cherche à évaluer la capacité de production de plusieurs variétés de soja, dans l'expérimentation avec un seul facteur étudier, tous les autres facteurs seront gardés constants (sauf les variétés), mais si on cherche à évaluer en plus un second facteur, la distance entre les rangs de semis, dans ce cas précis toutes les variétés étudiées seront prises à tous les niveaux de second facteur c-à-d que si le second facteur comporte les niveaux 40cm, 60cm et 80cm, alors les niveaux variétés seront évaluées à 40, 60 et 80.

Donc une expérimentation factorielle étudié les combinaisons de plusieurs niveaux de différents facteurs.

9.5.1.1. Factoriel randomisé

On considère ici que les traitements sont issus de la combinaison de deux facteurs : le facteur A à I modalités et le facteur B à J modalités. Cela fait IJ traitements en tout, répétés chacun r fois.

Donc N = IJr. Une notion très importante pour étudier la combinaison de ces facteurs est la notion d'interaction. Un exemple va permettre de la mettre en lumière. On considère deux facteurs croisés,

la variété avec deux modalités (V1 et V2) et la dose d'azote avec deux modalités aussi (N1 et N2), et on observe le rendement d'une culture.

Il y'a une conséquence importante : la première chose à tester, dans un essai étudiant deux facteurs A et B, est l'interaction entre A et B, car en présence d'interaction les effets principaux de A et de B peuvent ne pas avoir de sens.

* L'analyse statistique est faite au moyen du modèle suivant :

$$\mathbf{Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}}$$

où Y_{ijk} est la donnée observée sur la k^{eme} répétition du traitement (i,j), c'est à dire le traitement constitué de la modalité i du facteur A et de la modalité j du facteur B,

μ est la moyenne générale,

α_i est l'effet principal de la modalité i deA .

β_j est l'effet principal de la modalité j de B

$(\alpha\beta)_{ij}$ est l'interaction du traitement (i,j),

e_{ijk} est l'erreur résiduelle

* plan d'essai

N1V1	N2V1	N1V1	N1V2
N2V2	N1V2	N2V1	N2V2
N1V2	N2V2	N1V1	N2V1

* Analyse de la variance ANOVA :

Tableau 19. Table d'ANOVA du dispositif factoriel randomisé

Sources de V.	ddl	SCE	CME	F_{obs}	F_{tab}
V. totale	A.B.r -1	SCE_t	-	-	-
Traitements	A.B-1	SCE_{trait}	-	-	-
Facteur A	A-1	SCE_A	$\frac{SCE_A}{A-1}$	$\frac{CME_A}{CME_e}$	$F_{tab}{}^{A-1}_{A.B(r-1)}$
Facteur B	B -1	SCE_B	$\frac{SCE_B}{B-1}$	$\frac{CME_B}{CME_e}$	$F_{tab}{}^{B-1}_{A.B(r-1)}$
Interaction AxB	(A-1)(B-1)	$SCE_{A.B}$	$\frac{SCE_{A.B}}{(A-1)(B-1)}$	$\frac{CME_{A.B}}{CME_e}$	$F_{tab}{}^{(A-1)(B-1)}_{A.B(r-1)}$
Erreur	A.B(r-1)	SCE_e	$\frac{SCE_e}{A.B(r-1)}$	-	-

- $C=\frac{(\sum Y_{ijk})^2}{rxAxB}$
- $SCE_t=\sum Y_{ijk}{}^2 - C$
- $SCE_{tr}=\frac{\sum Y_{ij.}{}^2}{r} - C$
- $SCE_A=\frac{\sum Y_{i..}{}^2}{rxB} - c$

- $SCE_B = \frac{\sum Y_{j.}^2}{r x A} - c$
- $SCE_{AxB} = SCE_{tr} - SCE_A - SCE_B$
- $SCE_e = SCE_t - SCE_{tr}$

Exercice

Les résultats suivants proviennent d'une expérimentation dont l'objectif est d'étudier l'effet de la période (matin /soir) de l'injection d'une matière chimique (sans injection/ injection) sur la concentration en phospholipides du plasma des agneaux (5 agneaux/ traitement).

	Matin		Soir	
	Sans injection	Injection	Sans injection	injection
1	8.53	17.53	39.14	32.00
2	20.53	21.07	26.20	23.80
3	12.53	20.80	31.33	28.87
4	14.00	17.33	45.80	25.06
5	10.80	20.07	40.20	29.33

1- Tester les hypothèses $H_0 \ \ vis - à - vis \ H_1$

2- Comparer les moyennes en utilisant le test $ppds_{5\%}$

Solution

	Matin		Soir	
	Sans injection	Injection	Sans injection	injection
1	8.53	17.53	39.14	32.00
2	20.53	21.07	26.20	23.80
3	12.53	20.80	31.33	28.87
4	14.00	17.33	45.80	25.06
5	10.80	20.07	40.20	29.33
$\sum Y_{i.}$	66.39	96.80	182.67	139.06
$\bar{Y}_{i.}$	13.28	19.36	36.53	27.81

$\sum Y_{ij}$= 484.92 , $\bar{Y}_{..} = 24.25$

Facteur B effet	Facteur A effet période		
	Matin	Soir	$\sum Y_{.j.}$
Sans injection	66.39	182.67	249.06
injection	96.8	139.06	235.86
$\sum Y_{i..}$	163.19	321.73	484.92

- C= $\frac{(\sum Y_{ijk})^2}{rx\text{AxB}} = \frac{484.92^2}{5x2x2}$ = 11757.37
- SCE_t= $\sum {Y_{ijk}}^2$ - C = 13676.7– 11757.37=1919.33
- SCE_{tr}= $\frac{\sum {Y_{ij.}}^2}{r}$ – C= $\frac{66.39^2+182.67^2+96.8^2+139.06^2}{5}$ – 11757.37=1539.41
- SCE_A= $\frac{\sum {Y_{i..}}^2}{rx\text{B}} - \text{c} = \frac{163.19^2+321.73^2}{5x2} - 11757.37$=1256.75
- SCE_B= $\frac{\sum {Y_{.j.}}^2}{rx\text{A}} - \text{c} = \frac{249.06^2+235.86^2}{5x2} - 11757.37$=8.71
- SCE_{AxB}=SCE_{tr}-SCE_A-SCE_B=1539.41-1256.75-8.71=273.95
- SCE_e=SCE_t-SCE_{tr}= 1919.33-1539.41=379.92

Sources de v	Ddl	SCE	CME	F_{obs}	F_{tab}
Totale	5x2x2-19	1919.33	-	-	-
Traitements	2x2-1=3	1539.41	-	-	-
Facteur A	2-1=1	1256.75	$\frac{1256.75}{1} = 1256.75^{**}$	$\frac{1256.75}{23.75} = 53$	4.49
Facteur B	2-1=1	8.71	$\frac{8.71}{1} = 8.71^{ns}$	$\frac{8.71}{23.75} < 1$	4.49
Interaction AxB	(2-1)(2-1)=1	273.95	$\frac{273.95}{1} = 273.95^{*}$	$\frac{273.95}{23.75} = 11.53$	4.49
Erreur	(2x2)(5-1) =16	379.22	$\frac{379.22}{16} = 23.75$	-	-

On accepte H_1 et on rejette H_0 donc on observe que l'effet période significatif, l'effet injection non significatif et l'effet interaction significatif.

Comparaison des moyennes :

- Comparaison des moyennes des périodes:

ppds = $t_{5\%}\sqrt{\frac{2CME_e}{rxB}}$ = 2.120$\sqrt{\frac{2X23.75}{3x2}}$ = 4.62.

Soir: 32.17]A

Matin : 16.32]B

- Comparaison des moyennes de l'interaction:

ppds = $t_{5\%}\sqrt{\frac{2CME_e}{r}}$ = 2.120$\sqrt{\frac{2X23.75}{5}}$ = 6.53.

	36.53	27.81	19.36	13.28
36.53	0	8.72*	17.17*	23.25*
27.81		0	8.45*	14.53*
19.36			0	6.08
13.28				0

Soir-sans injection: 36.53]A

Soir- injection: 27.81]B

$$\left.\begin{matrix}\text{Matin} - \text{ injection: } 19.36 \\ \text{Matin} - \text{ sans injection: } 13.26\end{matrix}\right] C$$

9.5.1.2. factoriel en blocs

Da ce cas plus courant, le terrain expérimental est d'abord divisé en blocs qui vont jouer le rôle de répétition (1 bloc= 1 répétition) chaque bloc reçoit une fois et une seul fois une combinaison des différents niveaux des facteurs étudiés cette combinaison est affectée à une parcelle élémentaire donnée.

* L'analyse statistique est faite au moyen du modèle additif suivant :

$$\mathbf{Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + b_k + e_{ijk}}$$

où Y_{ijk} est la donnée observée sur la k^{eme} répétition du traitement (i,j), c'est à dire le traitement constitué de la modalité i du facteur A et de la modalité j du facteur B,

μ est la moyenne générale,

α_i est l'effet principal de la modalité i deA .

β_j est l'effet principal de la modalité j de B

$(\alpha\beta)_{ij}$ est l'interaction du traitement (i,j),

b_k est l'effet du bloc k

e_{ijk} est l'erreur résiduelle

* plan d'essai

N1V1	N2V1	N2V2	N1V2

N2V2	N1V2	N2V1	N1V1

N1V2	N2V2	N1V1	N2V1

* Analyse de la variance ANOVA :

Tableau 20. Table d'ANOVA du dispositif factoriel en blocs complètement randomisés

Sources de V.	ddl	SCE	CME	F_{obs}	F_{tab}
V. totale	A.B.r -1	SCE_t	-	-	-
Traitements	A.B-1	SCE_{trait}	-	-	-
Facteur A	A-1	SCE_A	$\frac{SCE_A}{A-1}$	$\frac{CME_A}{CME_e}$	$F_{tab}{}^{A-1}_{(A.B-1)(r-1)}$
Facteur B	B -1	SCE_B	$\frac{SCE_B}{B-1}$	$\frac{CME_B}{CME_e}$	$F_{tab}{}^{B-1}_{(A.B-1)(r-1)}$
Interaction AxB	(A-1)(B-1)	$SCE_{A.B}$	$\frac{SCE_{A.B}}{(A-1)(B-1)}$	$\frac{CME_{A.B}}{CME_e}$	$F_{tab}{}^{(A-1)(B-1)}_{(A.B-1)(r-1)}$
Bloc b	b-1	SCE_b	$\frac{SCE_b}{b-1}$	$\frac{CME_b}{CME_e}$	$F_{tab}{}^{b-1}_{(A.B-1)(r-1)}$
Erreur	(A.B-1)(r-1)	SCE_e	$\frac{SCE_e}{(A.B-1)(r-1)}$	-	-

- C= $\frac{(\sum Y_{ijk})^2}{rx\text{AxB}}$

- $SCE_t = \sum Y_{ijk}^2$ - C
- $SCE_{tr} = \frac{\sum Y_{ij.}^2}{r} - C$
- $SCE_A = \frac{\sum Y_{i..}^2}{rxB} - c$
- $SCE_B = \frac{\sum Y_{.j.}^2}{rxA} - c$
- $SCE_{AxB} = SCE_{tr} - SCE_A - SCE_B$
- $SCE_b = \frac{\sum Y_{..k}^2}{AxB} - C$
- $SCE_e = SCE_t - SCE_{tr} - SCE_b$

Exercice

Il s'agit d'expérimenter le comportement de 03 variétés de blé (facteur A à 03 niveaux V_1, V_2, V_3) avec 02 densit2 de semis (facteur B à 02 niveaux D_1, D_2) à travers un dispositif expérimental de type factoriel à 4 répétitions. La variable est le rendement exprimé en qx/ha

I	V_1D_2 :50.15	V_2D_1 :42.75	V_3D_1 :30.15	V_1D_1 :48.50	V_3D_2 :40.25	V_2D_2 :40.3
II	V_2D_2 :41.65	V_3D_2 :41.25	V_1D_1 :53.25	V_3D_1 :32.20	V_2D_1 :40.65	V_1D_2 :51.8
III	V_2D_1 :38.60	V_1D_2 :53.90	V_3D_1 :31.40	V_1D_1 :52.25	V_3D_2 :40.6	V_2D_2 :43.4
IV	V_2D_2 :42.20	V_3D_1 :29.70	V_1D_1 :50.70	V_3D_2 :41.60	V_1D_2 :50.75	V_2D_1 :39.70

On vous demande de faire

1- L'analyse de la variance

2- La comparaison des moyennes.

Solution

	I	**II**	**III**	**IV**	$\sum y_{i.}$
V_1D_1	48,5	53,25	52,25	50,7	**204,7**
V_1D_2	50,15	51,8	53,9	50,75	**206,6**
V_2D_1	42,75	40,65	38,6	39,7	**161,7**
V_2D_2	40,3	41,65	43,4	42,2	**167,55**
V_3D_1	30,15	32,2	31,4	29,7	**123,45**
V_3D_2	40,25	41,25	40,6	41,6	**163,7**
$\sum y_{.j}$	**252,1**	**260,8**	**260,15**	**254,65**	**1027,7**

	D1	D2	$\sum y_{i..}$	$\bar{y}_{i..}$
V1	204,7	206,6	411,3	51,41
V2	161,7	167,55	329,25	41,16
V3	123,45	163,7	287,15	35,89
$\sum y_{.j.}$	489,85	537,85	1027,7	
$\bar{y}_{.j.}$	40,82	44,82		

- C= $\frac{(\sum Y_{ij})^2}{\text{A}x\text{B}xr} = \frac{1027.7^2}{3x2x4}$ = 44006.97
- SCE_T= $\sum(Y_{ij})^2$ - C = 1244.13
- SCE_b= $\frac{\sum(tot_b)^2}{t} - C = \frac{264095.69}{6}$ – 44006.97 = 8,98
- SCE_{tr}= $\frac{\sum(tot_{tr})^2}{r}$ – C= $\frac{180843.13}{4}$ – 44006.97=1203.81
- SCE_A= $\frac{\sum(tot_A)^2}{rx\text{B}} - c = \frac{411.3^2+329.25^2+287.15^2}{4x2} - 44006.97$=996,58
- SCE_B= $\frac{\sum(tot_B)^2}{rx\text{A}} - c = \frac{489.85^2+537.85^2}{4x3} - 44006.97$=96,00
- SCE_{Sxt}= $SCE_{tr} - SCE_A - SCE_B$=111,24
- SCE_e=SCE_t-$SCE_{tr} - SCE_b$= 31,34

ANOVA :

Sources de v	ddl	SCE	CME	F_{obs}	F_{tab}	
V. totale	23	1244,13	-			
v. blocs	3	8,98	2,99	1,43	3,29	$^{3}_{15}$
traitements	5	1203,81	240,76	-		
Effet variété ou effet A	2	996,58	498,29	238,47	3,68	$^{2}_{15}$
Effet densité de semis ou effet B	1	96,00	96,00	45,94	4,54	$^{1}_{15}$
Interaction AxB	2	111,24	55,62	26,62	3,68	$^{2}_{15}$
Erreur	15	31,34	2,09	-		

On accepte H_1et on rejetteH_0, donc effet variété est significatif, effet densité est hautement significatif et effet interaction SxT est significatif.

Comparaison des moyennes

- **Variété** : ppds = $t_{5\%}\sqrt{\frac{2CME_e}{rxb}}$ = 2.131$\sqrt{\frac{2X2.09}{4x2}}$ = 1.54

$V_1 = 51.41]A$

$V_2 = 41.16]B$

$V_3 = 35.89]C$

- **Densité de semis** : ppds = $t_{5\%}\sqrt{\frac{2CME_e}{rxA}}$ = $2.131\sqrt{\frac{2X2.09}{4x3}}$ = 1.26

$D_1 = 44.82\ ]A$

$D_2 = 40.82]B$

*** Interaction SxT :**

ppds = $t_{5\%}\sqrt{\frac{2CME_e}{r}} = 2.131\sqrt{\frac{2X2.09}{4}}$ = 2.18

		V_1D_1	V_1D_2	V_2D_1	V_2D_2	V_3D_1	V_3D_2
		51,18	51,65	40,43	41,89	30,86	40,93
V_1D_1	51,18	0	-0,47	10,75*	9,29*	20,32*	10,25*
V_1D_2	51,65		0	11,22*	9,76*	20,79*	10,72*
V_2D_1	40,43			0	-1,46	9,57*	-0,5
V_2D_2	41,89				0	11,03*	0,96
V_3D_1	30,86					0	-10,07*
V_3D_2	40,93						0

V_1D_2	51,65	
V_1D_1	51,18	A
V_2D_2	41,89	
V_3D_2	40,93	
V_2D_1	40,43	B
V_3D_1	30,86	C

9.5.2. Le dispositif expérimental de type Split-Plot

Ce dispositif est appliqué dans les expérimentations qui nécessitent des surfaces importantes à ce titre on peut citer comme exemple techniques d'irrigation, dates et doses de semis, techniques culturales, travail de sol...etc.

Exemple d'un Split-Plot à 02 facteurs étudiés

Chaque facteur a 03 niveaux avec 03 répétitions

* Facteur principal = facteur A ayant 03 niveaux (A1, A2 et A3) =03 bloc=03 grandes parcelles

* Facteur secondaire = facteur B ayant 03 niveaux=03 (B1, B2 et B3) sous bloc=03 parcelles élémentaire

Plan d'expérience

A1			A2			A3		
B2	B1	B3	B3	B1	B2	B2	B3	B1

A3			A1			A2		
B1	B2	B3	B2	B3	B1	B3	B2	B1

A3			A2			A1		
B2	B3	B1	B2	B1	B3	B2	B1	B3

Table d'ANOVA

Tableau 21. Table d'ANOVA du dispositif Split-Plot

Sources de v	ddl	SCE	CME	F_{obs}	F_{tab}
V. totale	A.B.r-1	SCE_t	-	-	-
V. parcelles principales	A.r-1	SCE_{pp}	-	-	-
V. bloc	r-1	SCE_b	$\frac{\mathrm{SCE}_b}{\mathrm{r-1}}$	$\frac{\mathrm{SCE}_b}{\mathrm{SCE}_{ea}}$	$F_{tab}{}^{\mathrm{r-1}}_{\mathrm{(A-1)(r-1)}}$
Effet Facteur A	A-1	SCE_A	$\frac{\mathrm{SCE}_A}{\mathrm{A-1}}$	$\frac{\mathrm{SCE}_A}{\mathrm{SCE}_{ea}}$	$F_{tab}{}^{\mathrm{A-1}}_{\mathrm{(A-1)(r-1)}}$
Erreur a	(A-1)(r-1)	SCE_{ea}	$\frac{\mathrm{SCE}_{ea}}{\mathrm{(A-1)(r-1)}}$	-	-
Effet Facteur B	B-1	SCE_B	$\frac{\mathrm{SCE}_B}{\mathrm{B-1}}$	$\frac{\mathrm{SCE}_B}{\mathrm{SCE}_e}$	$F_{tab}{}^{\mathrm{B-1}}_{\mathrm{A(B-1)(r-1)}}$
Interaction AxB	(A-1)(B-1)	SCE_{AxB}	$\frac{\mathrm{SCE}_{AxB}}{\mathrm{(A-1)(B-1)}}$	$\frac{\mathrm{SCE}_{AxB}}{\mathrm{SCE}_e}$	$F_{tab}{}^{\mathrm{(A-1)(B-1)}}_{\mathrm{A(B-1)(r-1)}}$
Erreur	A(B-1)(r-1)	SCE_e	$\frac{\mathrm{SCE}_e}{\mathrm{A(B-1)(r-1)}}$	-	-

Exemple Voici un dispositif de type split-plot, dans lequel on veut étudier l'effet de l'engrais azotée (N_0 et N_{120}) sur la production de la betterave, précédée par des différentes cultures (vesce V, fétuque F, brome B et mélange brome vesce B/V)

	N_0				N_{120}			
I	V	F	B	B/F	B/F	V	B	F

	N_{120}				N_0			
II	F	B	V	B/F	V	F	B/V	F

	N_0				N_{120}			
III	F	B/F	B	V	V	B/V	F	B

N	**FO**	**I**	**II**	**III**	**Tot trait**	**moyennes**
N_0	**F**	13.8	13.5	13.2	40.5	13.5
	B	15.5	15.0	15.2	45.7	15.2
	V	21.0	22.7	22.3	66.0	22.0
	B/V	18.9	18.3	19.6	56.8	18.9
Tot parcelle principale		**69.2**	**69.5**	**70.3**	**209.0**	**14.4= $\overline{y}_{i.}$$N_0$**
N_{120}	**F**	19.3	18.0	20.5	57.8	19.3
	B	22.2	24.2	25.5	71.8	23.9
	V	25.3	24.8	28.4	78.5	26.2
	B/V	25.9	26.7	27.6	80.2	26.7
Tot parcelle principale		**92.7**	**93.7**	**101.9**	**288.3**	**24.0 = $\overline{y}_{i.}$$N_{120}$**
Tot bloc		**161.9**	**163.2**	**172.2**	**497.3**	**20.7= $\overline{y}_{..}$**

	F	**B**	**V**	**B/V**
Tot	**98.3**	117.5	144.5	137.0
Moy	**16.4**	19.6	24.1	22.8

- C= $\frac{(\sum Y_{ij})^2}{\mathrm{N}x\mathrm{FO}xr} = \frac{497.3^2}{2x4x3}$ = 10304.47

- $SCE_T = \sum(Y_{ij})^2$ - C = 516.12
- $SCE_{pp} = \frac{\sum(tot_{par\ prin})^2}{FO} - C = \frac{69.2^2+69.5^2+70.3^2+92.7^2+93.7^2+101.6^2}{4} - 10304.47 = 274.92$
- $SCE_b = \frac{\sum(tot_b)^2}{NxFO} - C = \frac{161^2+163.2^2+172.2^2}{2x4} - 10304.47 = 7.87$

- $SCE_N = \frac{\sum(tot_N)^2}{rxFO} - C = \frac{209^2+228.3^2}{3x4} - 10304.47 = 262.02$

❖ $SCE_{ea} = SCE_{pp} - SCE_N - SCE_b = 274.92 - 262.02 - 7.87 = 5.03$

- $SCE_{FO} = \frac{\sum(tot_{FO})^2}{rx\mathrm{N}} - c = \frac{98.3^2+117.5^2+144.5^2+137.0^2}{3x2} - 10304.47 = 215.26$
- $SCE_B = \frac{\sum(tot_B)^2}{rx\mathrm{A}} - c = \frac{489.85^2+537.85^2}{4x3} - 44006.97 = 96{,}00$
- $SCE_{NxFO} = \frac{\sum(tot_{tr})^2}{r} - c - SCE_N - SCE_{FO} =$

$\frac{40.5^2+45.7^2+66^2+56.8^2+57.8^2+71.8^2+78.5^2+80.2^2}{4} - 10304.47 - 262.02 - 215.26 = 18.70$

- $SCE_e = SCE_t - SCE_b - SCE_N - SCE_{ea} - SCE_{FO} - SCE_{NxFO} = 7.24$
- $ou\ SCE_e = SCE_t - SCE_{pp} - SCE_{FO} - SCE_{NxFO}$

Table d'ANOVA

Sources de v	ddl	SCE	CME	F_{obs}	F_{tab}	
V. totale	23	516.12	-	-		
V. parcelles principales	5	274.92	-	-		
V. bloc	2	7.87	3.935	1.56	19.0	$\frac{2}{2}$
Effet engrais ou effet N	1	262.02	262.02	104.18	18.5	$\frac{1}{2}$
Erreur a	2	5.03	2.515	-	-	
Effet précédent cultural Effet FO	3	215.26	71.753	118.99	3.49	$\frac{3}{12}$
Interaction NxFO	3	18.70	6.233	10.34	3.49	$\frac{3}{12}$
Erreur	12	7.24	0.603	-		

On accepte H_1et on rejetteH_0, donc effet engrais azotée est significatif, effet précédent cultural est hautement significatif et effet interaction SxT est significatif.

Comparaison des moyennes

<u>Effet N</u>

Ppds= $t_{5\%}\sqrt{2\,CME_{ea}/_{r.FO}} = 4.303\sqrt{2(2.515)/_{3x4}} = 2.78$

$$\left.\begin{matrix} \text{N120} : 24.0 \ \rightarrow \ \text{A} \\ \text{N0} \ : 17.4 \ \rightarrow \ \text{B} \end{matrix}\right] 6.6^{**}$$

Donc l'apport de l'azote améliore globalement la production de la betterave quelque soit la fumure organique

Effet FO Ppds= $t_{5\%}\sqrt{2\,CME_e/_{r.N}}$= 2.179 $\sqrt{2(0.603)/_{3x2}}$=0.98

V : 24.1 → A

B/V : 22.8 →B

B : 19.6 →C

F : 16.4 →D

La fumure organique améliore la production de la betterave avec un meilleur effet de la vesce suivi par le mélange vesce / brome puis le brome et le fétuque

Cette amélioration égale à 47% pour la vesce et 39% chez le mélange

X= $\frac{24.1\ x\ 16.4}{16.4} = 47\ \%\ \ et$ X= $\frac{22.8\ x\ 16.4}{16.4} = 47\ \%$

Effet N x FO $\text{Ppds}_{5\%} = t_{5\%}\sqrt{2\,CME_e/_{r}}$ = 2.179 $\sqrt{2(0.603)/_{3}}$ = 1.38

	F	B	V	B/V
N_0	13.5	15.2	22	18. 9
N_{120}	19.3	23.9	26.2	26.7

Les différences sont significatives entre les différentes combinaisons entres les deux facteurs

10. Rudiments d'analyses factorielles :AFC, ACP, CAH

10.1. Analyse en composantes principales « ACP »

10.1.1. Notion de Correlation :

- ✓ La corrélation est le degré d'association entre deux variables X et Y, pas de relation causale impliquée.

- ✓ En statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. Dans le cas de deux variables numériques, il s'agit de régression linéaire.
- ✓ Une mesure de cette corrélation est obtenue par le calcul du coefficient de corrélation linéaire, ce coefficient égal au rapport de leur covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation est compris entre -1 et +1

*** Coefficient de correlation**

La formule est : r = $\frac{\sigma_{xy}}{\sigma_x \sigma_y} \Rightarrow$ r = $\frac{COV_{(x,y)}}{\sigma_x \sigma_y}$.

Par exemple, nous allons calculer le coefficient de corrélation entre deux séries de même longueur (cas typique : une régression), on suppose qu'on a les tableaux de valeurs suivants : X ($X_1, \dots, X_n$) et Y ($Y_1, \dots, Y_n$) pour chacune des deux séries, alors pour connaitre le coefficient de corrélation liant ces deux séries. On applique la formule suivante :

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\cdot\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Avec: $\sigma_{xy}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$ et la covariance entre X et Y ou

$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$ est l'écart type de X. $\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Y_i}-\bar{\mathrm{Y}})^2}$ est l'écart type de Y.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est la moyenne de X et $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ est la moyenne de Y.

*** Interprétation**

r =1 dans le cas ou l'une des variables est fonction affine croissante de l'autre variable.

r = - 1 dans le cas ou l'une des variables est fonction affine décroissante de l'autre.

r = 0 signifie que les variables sont indépendants linéairement

Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1 plus la corrélation entre les variables est forte "on emploi simplement l'expression fortement corrélées pour qualifier les deux variables".

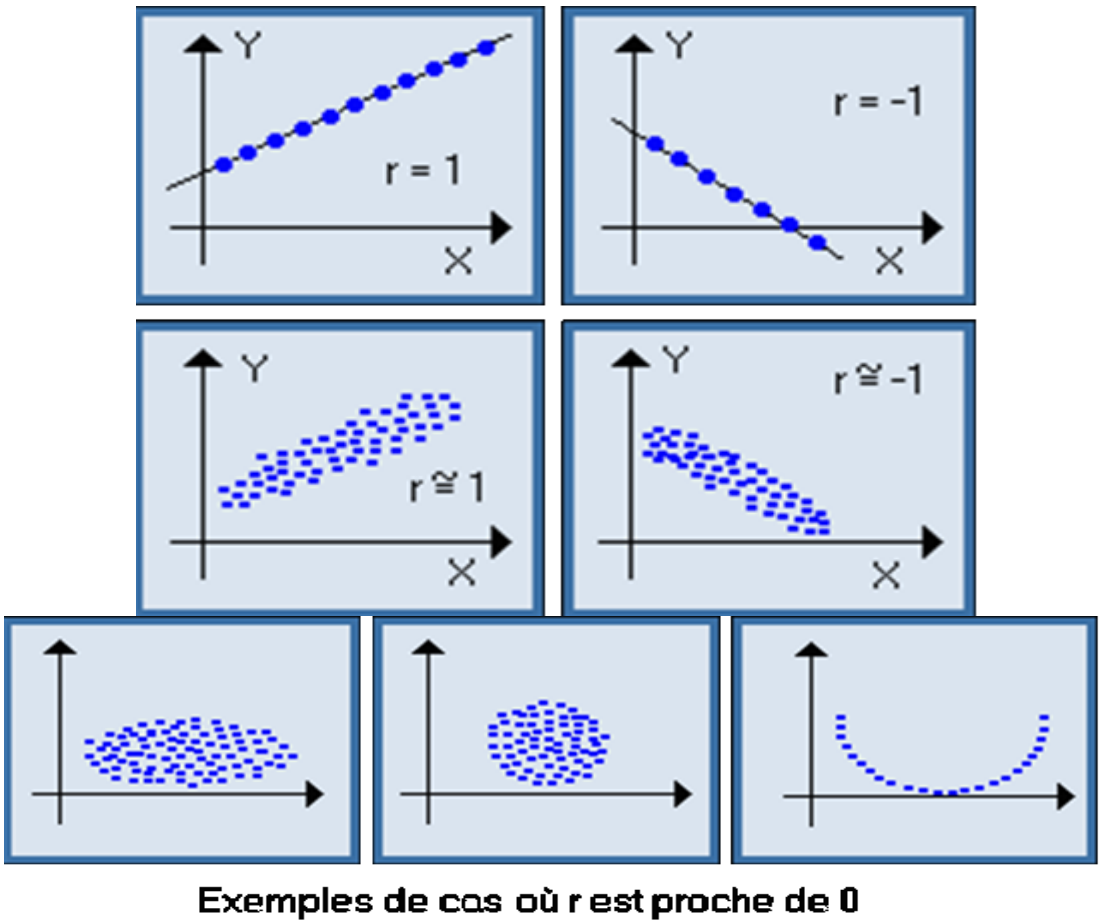


Figure 22. Exemples de nuage de points avec des différents coefficients de corrélation

10.1.2. Nuage de points

Ensemble de points isolés représentés dans un graphique cartésien : points M_1 , M_2 , ... , M_n de coordonnées (x_1 , y_1) ; (x_2 , y_2) ; ... ; (x_n , y_n).

Exemples :taille et poids de 60 enfant

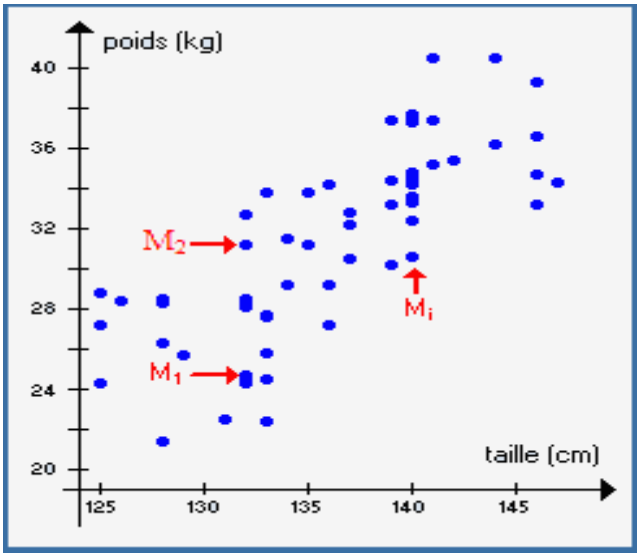


Figure 23. exemple d'un nuage de points (taille/poids)

10.1.3. Cercle des correlations en acp

- L'analyse en composantes principales (A.C.P) est une méthode statistique essentiellement descriptive : son objectif est de présenter sous une forme graphique le maximum de l'information contenue dans un tableau des données.

✓ Ce tableau doit être constitué en lignes par des individus (exp variétés , animaux …etc) sur lesquels sont mesurés des variables quantitatives ou pouvant être considérées comme telles : rendements gain , le poids , note …etc disposées en colonnes .

✓ L'ACP apporte l'avantage de traiter un groupe important de variables.

✓ L'ACP permet d'identifier les variables qui vont ensemble "ressemblance" et celles qui s'opposent "dissemblance"

✓ L'ACP est un traitement Multivariés des données.

✓ Le cercle de corrélation est fondamental en A.C.P

Remarque : on peut distinguer :

- **traitement univarié** : on peut calculer la moyenne et l'écarte type

ainsi que les *quantiles* (*médiane, quartiles, déciles, centiles...*). .

- **traitement bivarié** : lorsqu'on s'intéresse à la liaison entre deux variables, on peut représenter le nuage des points Mi(Xi , Yi) et examiner sa frome. la covariance et le coefficient de corrélation sont des indicateurs de l'intensité de la liaison linéaire éventuelle de ces deux variables.
- **traitement multivariés** : lorsqu'on s'intéresse aux liaisons entre plus de deux ou trois variables .on ne peut plus représenter graphiquement le nuage des points Mi , L'A.C.P nous permet de l'observer sous ses angles les plus intéressants , en examinant les projections du nuage sur des plants , elle permet également de repérer les groupes de variables ou d'individus fortement corrélées entre elles .

✓ Le cercle de corrélation et composé de :

- **les axes** : les axes du cercle représentent les facteurs étudies en générale on choisi 02 axe appelés axes factoriels en doit retenir autant d'axes qu'il le faut pour atteindre le seuil de variance expliquée désiré (80% par exemple)

Axe 01 : est la direction de plus grand allongement du nuage ou de plus grande dispersion, lorsque on projette les points du nuage sur cet axe, leurs projections sont plus dispersées qu'elles ne le seraient sur n'importe quel autre axe.

Axe 02 : est la 2 ^{eme} direction d'allongement du nuage c'est-à-dire celle qui explique après le 1 ^{er} axe le maximum de dispersion résiduelle .cet axe est choisi orthogonal sur le premier axe.

- les variables associées aux axes factoriels sont appelées facteurs ou composantes principales.

- ✓ **les points variables :** à chaque point – variable on associe un point dont la coordonnée sur un axe factoriel est une mesure de la corrélation entre cette variable et le facteur par projection sur un plan .les points-variables s'inscrivent dans un cercle de rayon 1, et sont d'autant plus proche du bord du cercle que le point-variable est bien représenté par le plan factoriel , c'est-à-dire que la variable est bien corrélée avec les deux facteurs constituant ce plan .

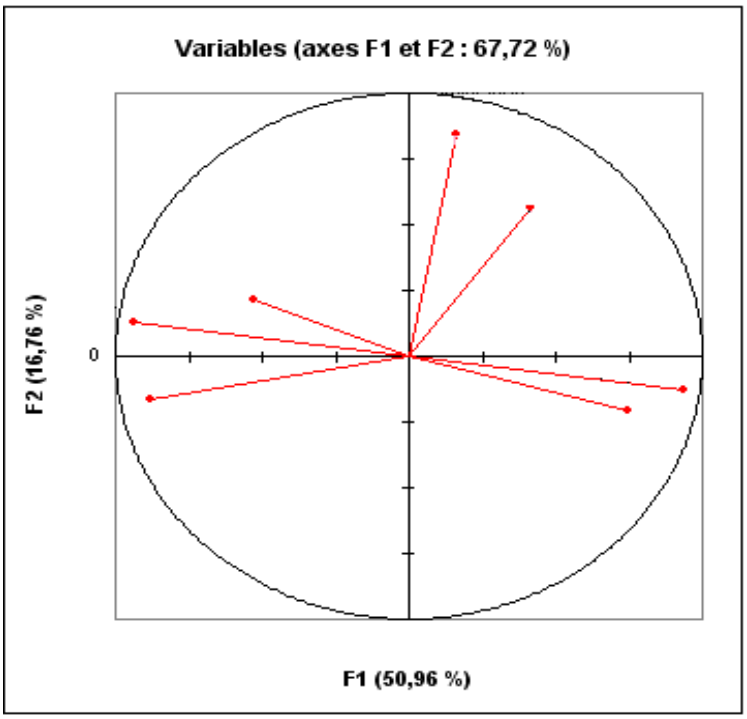


Figure 24 Exemple d'un cercle de correlation en ACP

Attention : les variables qui ne sont pas situées au bord du cercle dans un plan factoriel ne sont pas corrélées avec les deux facteurs représentés, elles ne servent pas à l'interprétation (Voir d'autres plans factoriels ou la corrélation sera plus forte).

L'angles entre deux point-variables mesuré par son cosinus est égale au coefficient de corrélation linéaire entre les deux variables : cos α = r (x_1,x_2)

- Si les points sont très proches ($\alpha\ peu\ différentde 0$): cos α =1 donc x_1 et x_2 sont très fortement corrélés positivement.

- Si α = 90° ; cos α = 0 alors pas de corrélation linéaire entre x_1 et x_2

Si α =180° ; cos α = -1 $\rightarrow$ *les points sont opposés* donc x_1 et x_2 sont très fortement corrélés négativement.

Le cercle des corrélations permet de voir, parmi les variables, les groupes de variables très corrélées entre elles.

- ✓ **Les points- individus**

La position d'un point-individu par rapport à un axe factoriel ainsi que les proximités entre les individus, peuvent être interprétées dès lors que ces points sont bien représentés par le plan factoriel observé. Certains individus seront bien représentés par le plan 1-2 (les très fort ou les très faible) d'autre seront représentés par d'autres plans exemple 1-3..etc.

10.2. Analyse factorielle des correspondances

10.2.1. Définition

Analyse factorielle des correspondances notée (AFC) est une méthode statistique qui permet de transformer un tableau de données ou de nombres en un graphique, cette méthode est très utilisée dans les études de type enquêtes. Elle a été développée par Jean-paul Benzécri (les années 70).

L'Analyse Factorielle des Correspondances (AFC) est une méthode qui permet d'étudier l'association entre deux variables qualitatives. Cette méthode est basée sur l'inertie.

Le but de l'Analyse Factorielle des Correspondances consiste à représenter un maximum de l'inertie totale sur le premier axe factoriel, un maximum de l'inertie résiduelle sur le second axe, et ainsi de suite jusqu'à la dernière dimension.

Les méthodes d'analyse factorielle des correspondances (AFC) tout comme celles d'analyse en composantes principales (ACP) s'utilisent pour décrire et hiérarchiser les relations statistiques qui peuvent exister entre des individus placés en ligne et des variables placées en colonnes dans un tableau rectangulaire de données. Les logiciels d'AFC fournissent donc en sortie une ou plusieurs figures de plans factoriels sur lesquels sont positionnés à la fois les individus et les variables. Par exemple: si 5 espèces d'insectes se répartissent entre 20 zones d'étude, on obtient par AFC une carte comprenant 25 points, dont 5 représentent chacun des espèces et les 10 autres représentent chacune des 20 sites ou zones d'étude.

10.2.2. Principe de l'AFC

L'analyse factorielle traite des tableaux de nombres. Elle remplace un tableau de nombres difficile à analyser par une série de tableaux plus simples qui sont une bonne approximation de celui-ci » Ces

tableaux sont « simples », car ils sont exprimables sous forme de graphiques - Pourquoi « des correspondances » ?

Pour des variables numériques on parle de la corrélation

Pour des variables nominales on parle de la correspondance

- Pourquoi « factorielle » ?

Il s'agit de décomposer le tableau original en une somme de tableaux/matrices qui sont chacun le produit de facteurs simples. Autrement dit, on les « met en facteurs ».

L'analyse factorielle des correspondances (AFC) a été conçue pour étudier des tableaux appelés couramment tableaux de contingence. Il s'agit de tableaux d'effectifs obtenus en croisant les modalités de deux variables qualitatives définies sur une même population de n individus et il permet d'exprimer la liaison entre ces deux variables.

10.2.3. Tableau de contingence et nuages associés

Soient deux variables qualitatives X et Y, comportant respectivement p et q modalités. On observe les valeurs de ces variables sur une population et on dispose d'un tableau de contingence à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison de la modalité i de X et de la modalité j de Y. Soit N la matrice des effectifs :

Tableau 22. Constitution d'une table de contingence

Y / X	y_1	...	y_j	...	y_q	
x_1	n_{11}	...	n_{1j}	...	n_{1q}	$n_{1.}$
...	...	...	...	...	...	...
x_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i.}$
x_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p.}$
	$n_{.1}$	...	$n_{.j}$	...	$n_{.q}$	N

Soit N la matrice des effectifs :

$$N = \begin{pmatrix} n_{11} & n_{12} & . & . & n_{1q} \\ n_{21} & . & . & . & n_{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ n_{p1} & n_{p2} & . & . & n_{pq} \end{pmatrix}$$

Les $n_{i.}$ et les $n_{.j}$ s'appellent respectivement marges en lignes et marges en colonnes et ils sont calculées comme suit :$n_{i.} = \sum_j n_{ij}$ et $n_{.j} = \sum_i n_{ij}$

10.2.4. Représentation des profiles associés à un tableau de contingence

* **Tableau des profils-lignes** : On appelle tableau des profils-lignes le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$. Ainsi ce tableau est définie par : D_1^{-1} N avec :

$$N = \begin{pmatrix} n_{11} & n_{12} & . & . & n_{1q} \\ n_{21} & . & . & . & n_{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ n_{p1} & n_{p2} & . & . & n_{pq} \end{pmatrix} \quad et \quad D_1 = \begin{pmatrix} n_{1.} & . & . & . & 0 \\ 0 & n_{2.} & . & . & 0 \\ . & . & n_{3.} & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & n_{p.} \end{pmatrix}$$

Tableau des profils-colonnes :

On appelle tableau des profils-colonnes le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{.j}}$ Ainsi ce tableau est définie par : ND_2^{-1} avec :

$$D_2 = \begin{pmatrix} n_{.1} & . & . & . & 0 \\ 0 & n_{.2} & . & . & 0 \\ . & . & n_{.3} & . & 0 \\ . & . & . & . & . \\ 0 & . & . & . & n_{.p} \end{pmatrix}$$

On appelle profil marginale ligne (resp profil marginale colonne) les quantités $\frac{n_{i.}}{n}$ (resp $\frac{n_{.j}}{n}$). l'écriture matricielle sera alors : $\frac{D_1}{n}$ (resp $\frac{D_2}{n}$).

Centre de gravité des deux profils : les profils lignes forment un nuage de p points dans R^q. Le centre de gravité de ce nuage est donné par :

$$g_l = \frac{1}{n}(D_1^{-1}N)'D_1 1 = \begin{pmatrix} \frac{n_{.1}}{n} \\ \frac{n_{.2}}{n} \\ . \\ . \\ \frac{n_{.q}}{n} \end{pmatrix} = \begin{pmatrix} p_{.1} \\ p_{.2} \\ . \\ . \\ p_{.q} \end{pmatrix}$$

Ainsi le centre de gravité de nuage des profils colonnes est définie par :

$$g_c = \frac{1}{n}(D_2^{-1}N')'D_2 1 = \begin{pmatrix} \frac{n_{1.}}{n} \\ \frac{n_{2.}}{n} \\ . \\ . \\ \frac{n_{p.}}{n} \end{pmatrix} = \begin{pmatrix} p_{1.} \\ p_{2.} \\ . \\ . \\ p_{p.} \end{pmatrix}$$

10.2.5. La métrique de χ^2

Comment mesurer la dispersion de ces nuages de profils ? Autrement dit, quelle métrique choisir dans chacun des espaces pour obtenir une bonne analyse ?

La distance entre deux profils lignes : Pour calculer la distance entre deux profils lignes i et i' on utilise la formule suivante :

$$\mathrm{d}^2_{\chi^2}(i,i') = \sum_{j=1}^{q} \frac{n}{n_{.j}}\left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2$$

Il s'agit donc de la métrique diagonale nD_2^{-1}.

La distance entre deux profils colonnes : Par analogie, on définit la distance entre deux profils colonnes j et j' par :

$$\mathrm{d}^2_{\chi^2}(j,j') = \sum_{i=1}^{p} \frac{n}{n_{i.}}\left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2$$

Ici on a utilisé la matrice nD_1^{-1}

L'inertie totale : L'inertie totale du nuage de point est donnée par la formule suivante :

$$\phi^2 = \frac{1}{n}\sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}-n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

On a aussi :

$$\phi^2 = \sum_i \frac{n_{i.}}{n} d^2_{\chi^2}(i, g_l) = \sum_i \frac{n_{i.}}{n} \sum_j \frac{n}{n_{.j}} (\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n})^2$$

10.2.6. La liaison entre deux variables qualitatives

Lorsque tous les profils-lignes sont identiques c'est à dire ∀j on a :

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = ... = \frac{n_{pj}}{n_{p.}}$$

on peut parler d'indépendance entre les deux variables X et Y. On remarque aussi que :

$$\sum_{i=1}^{p} \frac{n_{ij}}{n_{i.}} (\frac{n_{i.}}{n}) = \frac{n_{j.}}{n}$$

Ce qui entraine que dans le cas d'indépendance on a : $\boldsymbol{n_{ij} = \frac{n_{i.}n_j}{n}}$

Remarques :

- **Caractère significatif de l'écart à l'indépendance** : Lorsque on étudie un tableau de contingence, c'est-à-dire une population de n individus à travers de deux variables qualitatives, il est classique de mesurer le caractère significatif de la liaison entre ces deux variables à l'aide de la statistique χ^2. Appliquée à un tableau d'effectifs, cette statistique mesure l'écart entre les effectifs observés et les effectifs théoriques. Le test de χ 2 permet de s'assurer du caractère significatif de cette liaison. La démarche du test et déjà vue dans le chapitre (6).
- **Rapports entre ACP et AFC**

• Si on a des données permettant de faire une AFC, peut-on y appliquer une ACP ? – Non

• Si on a des données permettant de faire une ACP, peut-on y appliquer un AFC ? – Oui !

• .. Mais alors ?

–Alors on traite les données numériques, les nombres comme des catégories

– Si par exemple on travaille sur des notes, 18/20 n'est plus « supérieur à » 10/20, il n'est pas non plus « plus proche » de 16/20 que de 10/20.

- Quelques applications de l'AFC

On ne vous demande pas de savoir faire les calculs à la main mais seulement de savoir interpréter les informations portées par les logiciels comme Minitab,Statistica…. C'est d'ailleurs ce dernier point qui pourra être demandé à l'examen et en pratique.

10.2.7. Règles générales d'interprétation de l'AFC

* Pour interpréter l'AFC, la première étape consiste à évaluer s'il existe une dépendance significative entre les lignes et les colonnes. Une méthode rigoureuse consiste à utiliser la statistique de khi2 pour examiner l'association entre les modalités des lignes et celles des colonnes.

***** L'examination des valeurs propres permet de déterminer le nombre d'axes principaux à considérer. Les valeurs propres correspondent à la quantité d'informations retenue par chaque axe. Elles sont grandes pour le premier axe et petites pour l'axe suivant.

* Notez qu'une analyse est bonne lorsque les premières dimensions (axes) représentent une grande partie de la variabilité.

* Dans le graphique nuage de point représentant l'AFC, les lignes et colonnes sont représentées par des marqueurs (points, triangles) et des couleurs (bleus, rouges) différentes.

* La distance entre les points lignes ou entre les points colonnes donne une mesure de leur similitude (ou dissemblance). Les points lignes avec un profil similaire sont proches sur le graphique. Il en va de même pour les points colonnes.

10.3. Classification ascendante hiérarchique (CAH)

10.3.1. C'est quoi qu'une classification ascendante hiérarchique (CAH)

C'est une méthode de classification automatique utilisée en analyse des données ; à partir d'un ensemble de *n* individus, son but est de répartir ces individus dans un certain nombre de classes.
La méthode suppose qu'on dispose d'une mesure de dissimilarité entre les individus; dans le cas de points situés dans un espace euclidien, on peut utiliser la *distance* comme mesure de dissimilarité. La classification ascendante hiérarchique est dite ascendante car elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes.

.10.3.2. Principe de CAH

* Cette analyse consiste à regrouper progressivement les individus dans un groupe
* Il faut d'abord mettre les individus les plus proches ensemble ensuite rejeté les individus les plus éloignés. Cette méthode convient plus au cas des variables explicatives de type quantitatives.
* L'état de rapprochement ou l'éloignement entre les individus est mesuré souvent par le biais de la distance euclidienne.

* La classification ascendante hiérarchique (CAH) conduit à la construction d'un **arbre de classification** (ou **dendrogramme**) montrant le passage des *n* individus au groupe «total» par une succession de regroupements

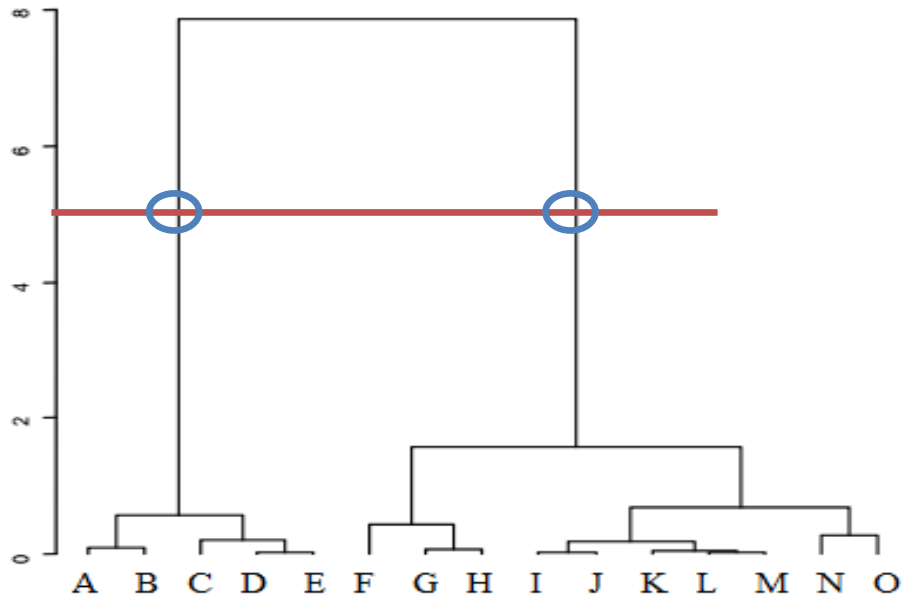


Figure 25. Exemple d'un arbre de classification (ou **dendrogramme**)

* On peut choisir un niveau de coupure et les croisements avec les bras du dendrogramme représentent les groupes dans notre exemple on peut identifier deux groupes.*

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches.
- Si les individus de 2 classes différentes sont éloignés

Et mathématiquement ça se traduit par ?

- Variabilité intra-classe petite

_ Variabilité inter-classes grande.

⇒ Deux critères, lequel choisir ?

* Qualité d'une partition

$\bar{x}_k$ moyenne de x_k, $\bar{x}_{qk}$ moyenne de x_k dans la classe q

$$\underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_k)^2}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_{qk})^2}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_{qk}-\bar{x}_k)^2}_{\text{Inertie inter}}$$

La qualité d'une partition est mesurée par :

$$0 \leq \frac{\text{Inertie inter}}{\text{Inertie totale}} \leq 1$$

$\frac{\text{Inertie inter}}{\text{Inertie totale}} = 0 \Rightarrow \forall k, \forall q,\ \bar{x}_{qk} = \bar{x}_k$ C'est-à-dire par variable, les classes ont mêmes moyennes (ne permet pas de classer)

$\frac{\text{Inertie inter}}{\text{Inertie totale}} = 1 \Rightarrow \forall k, \forall q, \forall_i \bar{x}_{iqk} = \bar{x}_k$ C'est-à-dire les individus d'une même classe sont identiques (idéal pour classifier).

Attention: ce critère ne peut être jugé en absolu car il dépend du nombre d'individus et du nb de classe.

10.3.3. Etapes de construction d'un dendrogramme

* la méthode de CAH utilise la matrice de distance comme critère de segmentation.

* cette méthode ne nécessite pas de spécifier a priori le nombre de groupes k (tel que k means) mais elle nécessite une critère d'arrêt.

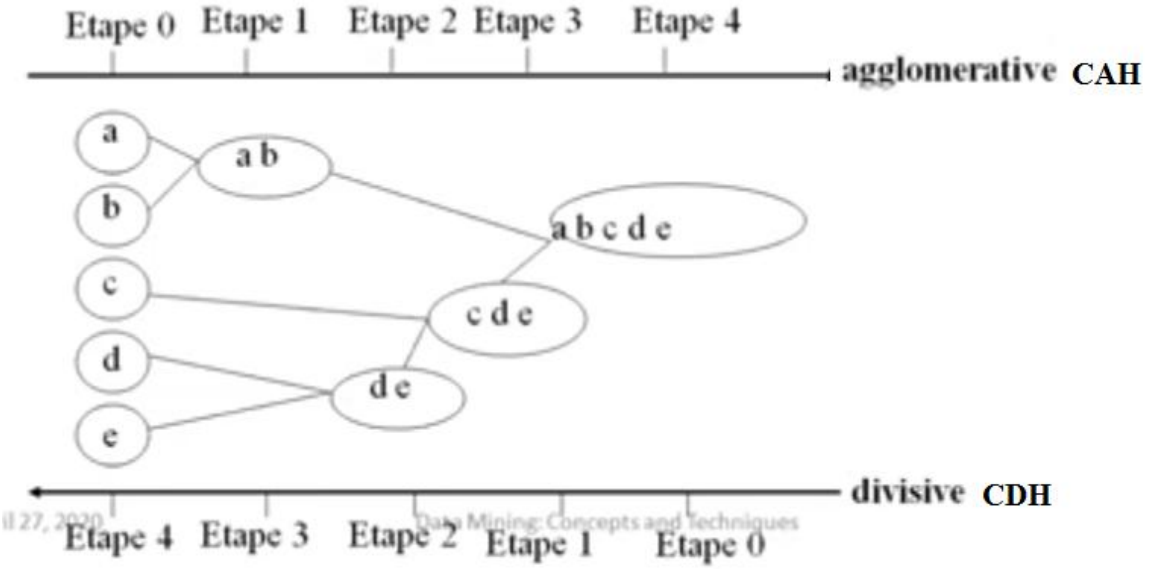


Figure 26. Deux méthodes de classification (CAH et CDH)

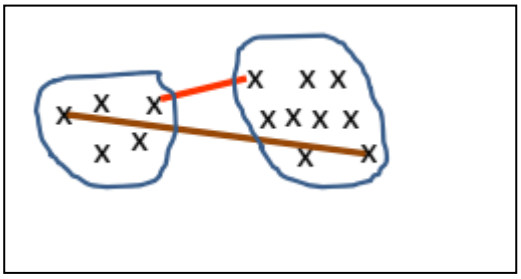
* on peut distinguer 4 étapes de la méthode CAH :

- ✓ *Initialisation : les classes initiales sont les n singletons individus. Calculer la matrice de leurs distances deux à deux*
- ✓ Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe :
  - ➢ Regrouper les deux éléments (classes) les plus proches au sens de la distance entre groupes choisie.
  - ➢ Mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes.
- ✓ Nécessité de définir une distance entre groupes d'individus (appelé stratégie d'agrégation). Nécessite de choisir le nombre de classes à retenir

10.3.4. Méthodes de classification ascendante hiérarchique ou Stratégies d'agrégation

* **Stratégie du saut minimum** ou single linkage (la distance entre parties est la plus petite distance entre éléments des deux parties):

* **Stratégie du saut maximum ou du diamètre** ou complete linkage (la distance entre parties est la plus grande distance entre éléments des deux parties



* **Méthode du saut Ward** (en espace euclidien),

$$\text{Inertie}(a) + \text{Inertie}(b) = \text{Inertie}(a \cup b) - \underbrace{\frac{m_a m_b}{m_a + m_b} d^2(a,b)}_{\text{à minimiser}}$$

A chaque itération, on agrège de manière à avoir un gain minimum d'inertie intra-classe : perte d'inertie interclasse due à cette agrégation

10.3.5. Interprétation d'une classification

Une partition est considérablement enrichie par une description des classes à l'aide des individus et des variables

- **Interprétation par les individus**

Pour chaque classe, on examine:

- son effectif,
- son diamètre (distance entre les 2 points les plus éloignés),
- la séparation (distance minimale entre la classe considérée et la classe la plus proche) et le numéro de la classe la plus proche,
- les identités des individus les plus proches du centre de gravité de la classe ou «parangons»,
- les identités des individus les plus éloignés du centre de gravité de la classe ou «extrêmes».

- **Interprétation par les variables : une par une**

On calcule un critère mesurant la pertinence de chaque variable de façon isolée pour interpréter la classe.

Est-ce que tous les éléments de la classe ont certaine (s) valeur(s) de cette variable (condition nécessaire d'appartenance à la classe) ?

Est-ce que certaine(s) valeur(s) de cette variable ne se rencontrent que dans cette classe (condition suffisante d'appartenance à la classe) ? ...

Interprétation par les variables continues

Comparaison de la moyenne et de l'écart-type *Sk* d'une variable *X* dans la classe *k* à la moyenne générale et à l'écart-type général.

Exercice (cours data mining et analyse des données Mili (2020)

On compte représenter 8 objets {A,B,C,D,E,F,G,H} sous la forme de 3 groupes notés {G1, G2, G3}, en se basant sur 2 critères X et Y tels que représentés dans le tableau suivant:

Objets	X	Y
A	2	10
B	2	5
C	8	4
D	5	8
E	7	5
F	6	4
G	1	2
H	4	9

1/Représenter les objets dans un graphique d'abscisse X et d'ordonnée Y. Que peut-on conclure sur le nombre des groupes à choisir.

2/ Dressez la matrice des distances

3/ Utiliser la méthode hiérarchique pour tracer le dendrogramme en se basant sur la méthode d'agrégation de saut minimum.

4/Interpréter les résultats

Solution

Les étapes à suivre

- Dresser la matrice des distances D
- Sélection la distance minimale entre les individus i et j et joint ces individus dans le même groupe.

Calcul des distances euclidiennes $d_{(i,j)}$ de chaque objet Ii autour de chaque objet j.

	A (2,10)	B (2,5)	C (8,4)	D (5,8)	E (7,5)	F (6,4)	G (1,2)	H (4,9)
A(2,10)	0	**5**	**8,48**	**3,6**	**7,07**	**7,21**	**8,06**	**2,23**
B(2,5)	5	0	**6,08**	**4,24**	**5**	**4,12**	**3,16**	**4,47**
C(8,4)	8,48	6,08	0	**5**	**1,4**	**2**	**7,28**	**6,4**
D(5,8)	3,6	4,24	5	0	**3,6**	**4,12**	**7,21**	**1,4**
E(7,5)	7,07	5	1,4	3,6	0	**1,4**	**6,7**	**5**
F(6,4)	7,21	4,12	2	4,12	1,4	0	**5,38**	**5,38**
G(1,2)	8,06	3,16	7,28	7,21	6,7	5,38	0	5,38
H(4,9	2,23	4,47	6,4	1,4	5	5,38	5,38	0

Objets	X	Y
A	2	10
B	2	5
C	8	4
D	5	8
E	7	5
F	6	4
G	1	2
H	4	9

$$d_{(A,B)} = \sqrt{(2-2)^2 + (10-5)^2} = \sqrt{25} = 5$$

$$d_{(A,C)} = \sqrt{(2-8)^2 + (10-4)^2} = \sqrt{72} = 8{,}48$$

$$d_{(A,C)} = \sqrt{(2-5)^2 + (10-8)^2} = \sqrt{13} = 3{,}8$$

La distance minimale est 1,4 donc on peut regroupé F,E et C dans un groupe et D et H dans un groupe

Puis on calcule les distances entre les nouvelles classes :

	A (2,10)	B (2,5)	C (8,4)	D (5,8)	E (7,5)	F (6,4)	G (1,2)	H (4,9)
A(2,10)	0	5	8,48	3,6	7,07	7,21	8,06	2,23
B(2,5)		0	6,08	4,24	5	4,12	3,16	4,47
C(8,4)			0	5	1,4	2	7,28	6,4
D(5,8)				0	3,6	4,12	7,21	1,4
E(7,5)					0	1,4	6,7	5
F(6,4)						0	5,38	5,38
G(1,2)							0	5,38
H(4,9								0

Objets	X	Y
A	2	10
B	2	5
C	8	4
D	5	8
E	7	5
F	6	4
G	1	2
H	4	9

	A (2,10)	B (2,5)	{C,E,F}	{D, H}	G (1,2)
A(2,10)	0	5	7,07	2,23	8,06
B(2,5)		0	4,12	4,24	3,16
{C,E,F}			0	3,6	5,38
{D, H}				0	5,38
G(1,2)					0

Pour les distances entre les groupes formés et les classes on prend toujours les distances minimales entre les différentes combainaisons.

La nouvelle distance minimale est 2,3 donc on peut regroupé A avec (H,D)

	{A, D, H}	B (2,5)	{C,E,F}	G (1,2)
{A, D, H}	0	4,24	3,6	7,21
B(2,5)		0	4,12	3,16
{C,E,F}			0	5,38
G(1,2)				0

	{A, D, H}	{B,G}	{C,E,F}
{A, D, H}	0	4,24	3,6
{B,G}		0	4,12
{C,E,F}			0

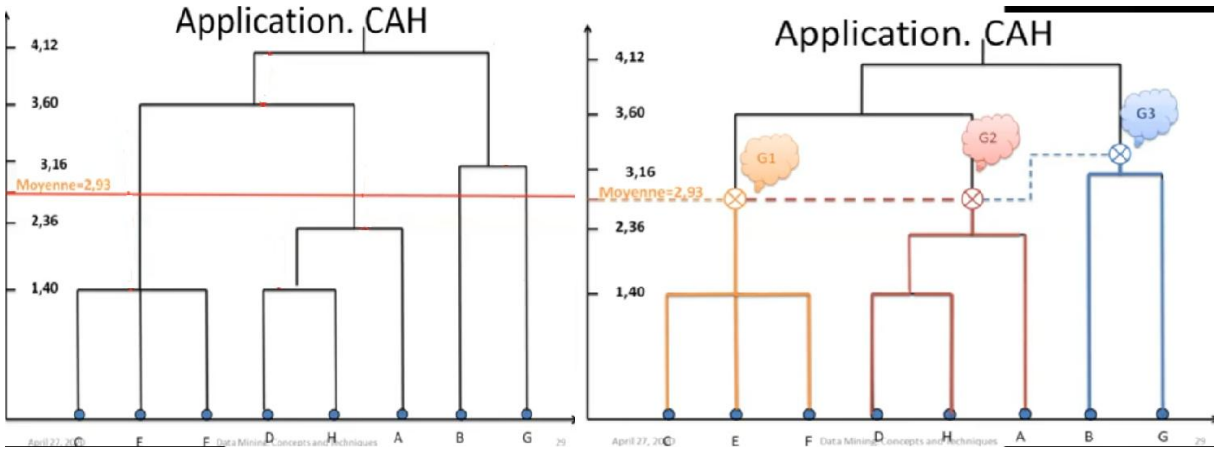
De la même façon on regroupe les autre classes

	{A, D, H}	{B,G}	{C,E,F}
{A, D, H}	0	4,24	3,6
{B,G}		0	4,12
{C,E,F}			0

	{A, D, H, C,E,F}	{B, G}
{A, D, H, C,E,F}	0	4,12
{B, G}		0

	{A, D, H, C,E,F, B, G}
{A, D, H, C,E,F, B, G}	0

Et finalement on aboutit au dendrogramme suivant



11. Cladistique

La **cladistique** est une méthode de classification par clade, la classification cladiste ou cladisme, basée sur la plus ou moins grande proximité des liens de parenté, et dans laquelle tous les taxons (voir ce terme) doivent être monophylétiques (certains auteurs produisant des classifications évolutives mais pas strictement cladistiques, se démarquent partiellement de cette méthode en admettant des taxons paraphylétiques).

Se dit aussi pour qualifier une classification utilisant la méthode cladistique. Elle utilise des cladogrammes qui montrent les relations de parenté entre les taxons relativement à certains caractères préalablement sélectionnés. Le cladogramme est remplacé par un phénogramme lorsque seuls les phénotypes sont pris en compte, par un phylogramme en phylogénétique.

Références

Ayache A. et Hamonier J. 2017. Statistique Descriptive. HCE Maroc. 39.

Benzecri J. P. 1980. Analyse des données (Tome 2). L'analyse des correspondances. Dunod.

Bouroche J.M. 1978. L'analyse des données. Pour la Science : 5, 23-35

Carpentier F.G. 2008. Introduction aux analyses multidimensionnelles. Exercices de synthèse corrigés. PSR83B.115-126.

Corre E. 2013. Introduction aux méthodes de phylogénie. Formation Biogenouest. Rennes. 373p.

Celeux G., Diday E., Govaert G., Lechevallier Y. et Ralambondrainy H. 1989. Classification automatique des données. Ed. Dunod.

Depiereux E., De Hertogh B. et Vincke G. 2008. Biostatistiques. Sciences module 105. FUNDP. 13p.

Dervin C. 1988. Comment interpréter les résultats d'une analyse factorielle de correspondances ? ITCF.

Ducay S. 2012. Estimation, intervalle de confiance, test statistique (suite). Cas d'une ou de deux moyennes, d'une ou de deux variances. Unv. Picardie Jules Verne.17p.

ENFA 2000. Comparaison de deux moyennes. Bulletin du GRES n°9 :23p.

Escofier B. 1969. L'analyse factorielle des correspondances. Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, tome 13 : 25-59.

Escofier-Cordier B. et Pagées J. 1990. Analyse factorielle simple et multiples objectifs, Méthodes et interprétations. Ed. Dunod.

Grais, B. 1979. Méthodes statistiques. *Dunod, Paris,* 381p.

Labarere J. 2012. UE4 : Biostatistique. Tests paramétriques de comparaison de 2 moyennes. Exercices commentés. 66p.

Saporta G. 2006. Probabilités et analyse des données statistiques, 3*eme* édition, Ed. Technip.

Spiegel M.R. 1985. Théorie et applications de la statistique. *Série Schaum, McGaw Hill, Paris,* 358p.

Tallur B. 1983 : Méthode d'interprétation d'une classification hiérarchique d'attributs-modalités pour l'explication d'une variable ; application à la recherche d'un seuil critique de la tension artérielle systolique et des indicateurs de risque cardiovasculaire Revue de statistique appliquée, 31(1) : 25-43.

Wonnacott, T.H. & Wonnacott, J. 1990. Statistique. Economie-Gestion-Sciences Médecine. *Economica,* Paris 4*eme* édition, 919p.

Annexe : Rappel sur la représentation numérique des données

1. Paramètres de position

- **La moyenne**

Lorsque x désigne la variable statistique, la valeur moyenne, ou moyenne de la série se note m ou . Elle est l'analogue d'un centre de gravité.

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

- **La médiane**

Notée *Me*, consiste en la valeur de la variable qui se trouve au centre de la série statistique, classée en ordre croissant. Elle sépare la série en deux groupes égaux. S'il y a un nombre impair d'observations, *Me* est une observation de la série. Sinon, la médiane est située entre les deux observations centrales de la série. Par convention, on utilise la moyenne de ces deux valeurs.

- Si la variable est discrète :

n est impair $\longrightarrow$ $Me = x_{\frac{n+1}{2}}$

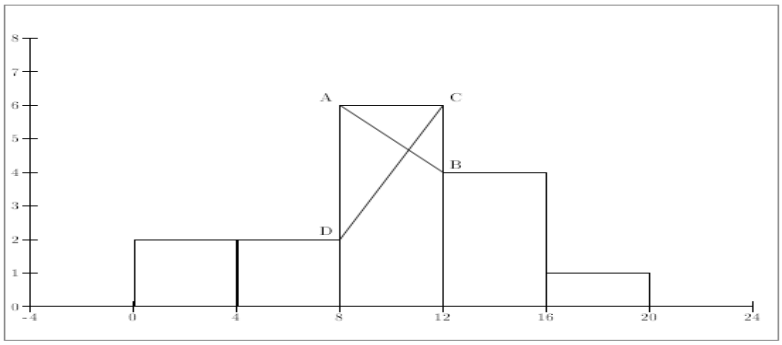
n est pair $\longrightarrow$ $Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

- Si la variable est continue

Vérifie F(Me) = 0.5, où F est la fonction de répartition de la variable. On détermine alors un intervalle médian (intervalle contenant la médiane), puis on procède à l'intérieur de cette classe à une interpolation linéaire.

- **Le mode** désigné par Mo est la valeur de la variable statistique la plus fréquente. Dans le cas d'une variable statistique continue, on parle plutôt de **classe modale**.

NB : Le mode ou la classe modale n'est pas obligatoirement unique.



$$Mo = L + i\left(\frac{\Delta i}{\Delta i + \Delta S}\right)$$

L = borne inférieure de la classe modale
i = intervalle de classe
Δi = excédent de fréquence entre la classe modale et la classe inférieure
ΔS = excédent de fréquence entre la classe modale et la classe supérieure

3.1. Paramètres de dispersion

- **L'étendue**

L'étendue E de variable x est la différence entre la plus grande et la plus petite des valeurs observées : E = max- min

- **Variance et Écart-type**

La variance permet d'estimer la variabilité des valeurs se trouvant autour de la moyenne, donc. Cette dernière peut alors être d'ordre biologique, ou peut être causée par la mauvaise qualité ou le faible nombre des mesures expérimentales. La variance d'une population sera notée σ^2, et la variance d'un échantillon s^2.

- Pour une population $$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}{N}$$
- Pour un échantillon $$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

L'écart-type de la variable X, noté par X, est, par définition, la racine carrée de la variance de cette variable. Signalons au passage que l'écart-type est la mesure de la dispersion la plus couramment utilisée.

- **Le coefficient de variation**, noté CV permet de comparer la variation de variables exprimées originellement dans des unités physiques différentes. Il est donné par :

$$cv(\%) = \frac{S}{\overline{x}} 100$$

Lorsque les échantillons sont de petite taille (n<20), on applique la correction suivante :

$$cv'(\%) = \left(1 + \frac{1}{4n}\right) cv$$

- **Covariance est** une mesure de la variabilité conjointe de deux variables aléatoires qui s'obtient par la somme des produits rectangulaires des écarts des valeurs de deux variables par rapport à leurs moyennes.

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{n-1}$$

La covariance indique si, et indirectement dans quelle mesure, les valeurs d'une variable augmentent ou diminuent avec les valeurs croissantes de l'autre.

- **Coefficient de corrélation** permet de détecter la présence ou l'absence d'une relation linéaire entre deux caractères quantitatifs, il est donné par la formule suivante :

$$r_{xy} = \frac{S_{xy}}{S_x . S_y}$$

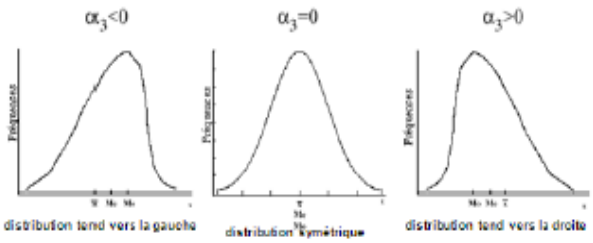
Le **coefficient d'asymétrie** ou skewness mesure l'asymétrie d'une distribution, c'est-à-dire la façon dont la représentation de la distribution penche d'un côté ou de l'autre. Ce coefficient est noté α_3 et est égal à :

$$\alpha_3 = \frac{k_3}{s_x^3}$$

où s^3_x est le cube de l'écart-type de la distribution, et où

$$k_3 = \frac{n.\sum_{i=1}^{n}(x_i - x)^3}{(n-1)(n-2)}$$

En fonction de la valeur de α_3, nous obtenons les représentations ci-dessous :



Le **coefficient d'aplatissement**, aussi appelé kurtose ou kurtosis, est noté α_4. Il est donné par la formule suivante :

$$\alpha_4 = \frac{k_4}{s_x^4}$$

Où s^4_x est la quatrième puissance de l'écart-type de la distribution, et où :

$$k_4 = \frac{n(n+1)\sum_{i=1}^{n}(x_i - x)^4 - 3(n-1)\left(\sum_{i=1}^{n}(x_i - x)^2\right)^2}{(n-1)(n-2)(n-3)}$$

En fonction de la valeur de α_4, nous obtenons les représentations ci-dessous :