الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

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

Université 8 Mai 1945 de Guelma
Faculté des sciences de la nature et de la vie
et des sciences de la terre et de l'univers

Département des sciences de la nature et de la vie

Cours de Biostatistiques

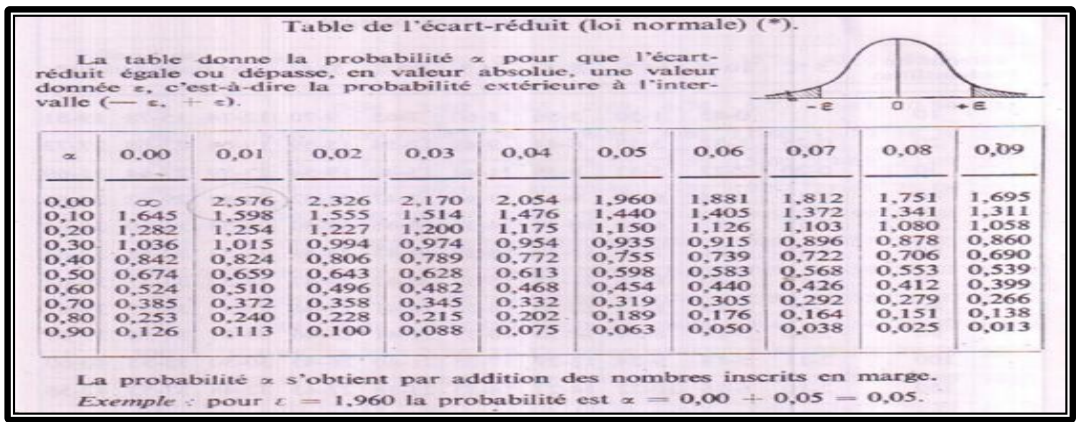


Table de l'écart-réduit (loi normale) (*).

La table donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée ε, c'est-à-dire la probabilité extérieure à l'intervalle $(-\varepsilon, +\varepsilon)$.

α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

La probabilité α s'obtient par addition des nombres inscrits en marge.
Exemple : pour $\varepsilon = 1{,}960$ la probabilité est $\alpha = 0{,}00 + 0{,}05 = 0{,}05$.

Destiné aux étudiants en deuxième année sciences biologiques

Elaboré par : Dr DERBAL Nora

Objectifs pédagogiques

L'objectif de cette brochure est d'apporter certains outils méthodologiques classiquement utilisés pour décrire et tester des phénomènes biologiques.

Sommaire

Introduction

Toute étude statistique peut être décomposée en deux phases au moins : le rassemblement ou la collecte des données, d'une part, et leur analyse ou leur interprétation, d'autre part.

La collecte des données peut être décomposée en deux étapes, l'une déductive ou descriptive et l'autre inductive.

La statistique descriptive a pour but de mesurer et de présenter les données observées d'une manière telle qu'on puisse en prendre connaissance aisément, par exemple sous la forme de tableaux ou de graphiques.

L'inférence statistique permet d'étudier ou de généraliser dans certaines conditions les conclusions ainsi obtenues à l'aide de tests statistiques en prenant certains risques d'erreur qui sont mesurées en utilisant la théorie des probabilités (DAGNELIE.P., 1986).

1) **La collecte des données :** il existe deux façons de collecter des données qui sont :
    a) ***Une simple observation*** : permet d'acquérir une 1ère connaissance des phénomènes de la nature ex : médiologie, économie……
    b) ***Une expérimentation*** : l'expérience a pour but d'apporter de nouvelles informations aux connaissances que l'on possède déjà. Les phénomènes étudiés peuvent être provoqués facilement en biologie, en physique et/ou en biochimie et, pour réaliser une expérience il faut suivre les étapes suivantes (DAGNELIE.P., 2003) :
        - Planification
        - Réalisation
        - Collecte des données
        - Analyse des données
        - Interprétation des résultats
        - Conclusion

Ce plan constitue la pierre angulaire de la méthode expérimentale.

2) **Analyse statistique :** c'est l'application des tests statistiques pour analyser les données collectées. Ces tests sont expliqués dans les chapitres suivants.

Chapitre 01 : Statistique descriptive à une dimension ou statistique univariée

C'est une partie de la statistique qui a pour but :

- de rassembler des données numériques ;
- d'en donner des représentations graphiques ;
- d'en résumer l'information sous une forme condensée plus accessible, plus commode.

Pour cela, on définit des valeurs caractéristiques qui sont les paramètres de position et de dispersion (LEGRAS. B.,1998).

Généralité

La population est un ensemble de sujets (objets = éléments) qui ont au-moins une propriété en commun.

L'échantillon de la population est un sous-ensemble de la population. Cet échantillon doit être représentatif de la population.

L'unité statistique est l'élément de la population sur lequel on travaille. Par exemple, si on s'intéresse aux étudiants d'une école, l'unité sera l'étudiant.

Variable statistique : c'est une caractéristique des individus constituant la série étudiée.

On peut répartir les variables en deux catégories principales.

- **Variables qualitatives** : ce sont des variables non mesurables.

  On distingue deux catégories secondaires :

  - Les variables ordinales (ou semi-quantitatives), elles peuvent bénéficier d'un classement ordonné.

    Exemple : intensité de la douleur (nulle, légère, forte, ...) ;
  - Les variables qualitatives pures.

    Exemple : couleur des yeux.
- **Variables quantitatives** : ce sont des variables mesurables. On distingue deux catégories secondaires (LEGRAS. B.,1998):

  - Les variables discontinues, elles ne peuvent prendre qu'un nombre fini de valeurs.

    Exemple : nombre d'enfants… ;
  - Les variables continues, elles peuvent prendre un nombre infini de valeurs.

    Exemple : taille, pression artérielle, ….

1.1 Tableaux statistiques

Un tableau statistique est une représentation chiffrée d'un fait social ou économique construit à partir d'une ou plusieurs variables (ligne, colonne) chacune caractérisée par une ou plusieurs modalités (âge, sexe, année,…), par exemple un tableau du poids de 5000 animaux :

I	y_i
$i = 1$	$y_1 = 75{,}2$
$i = 2$	$y_2 = 82{,}5$
:	:
:	:
:	:
:	:
$n = 5000$	$y_{5000} = 65{,}0$

Tableau simple dans un cahier de 96 pages

Simplifier

Classe de poids	n_i
50 – 55	60
55 – 60	2500
60 – 65	470
:	:
:	:
:	:

Tableau de distribution de fréquence (une seule page)

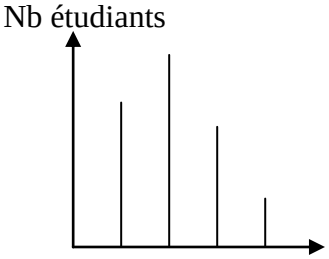
1.2 Représentation graphiques : la représentation graphique donne la 1ère idée de l'aspect général des distributions étudiées. Il y a 3 types de graphiques

a) Diagramme en bâtons : pour les variables discontinues avec les valeurs discontinues.

Ex : nombre d'étudiants par filière :

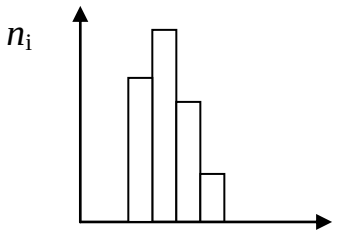
B.V ⟶ 53 Biochimie ⟶ 25

B.M ⟶ 30 Microbiologie ⟶ 80



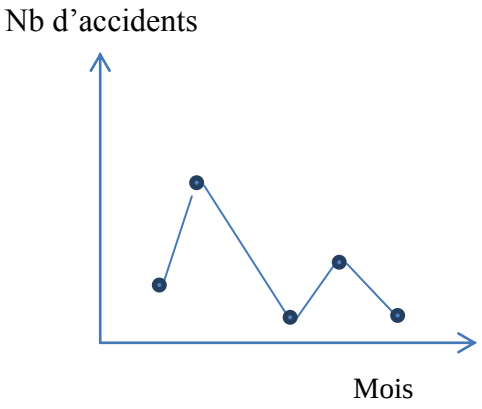
Filière (discontinue)

b) Histogrammes : pour les variables continues avec des valeurs continues de -∞ à +∞.



(Continue) **classe**

c) Polygones de fréquences : pour variables discontinues avec des fréquences.
Ex : nombre d'accidents par mois :



1.3 Réduction des données

La réduction des données permet de condenser les données sous forme des paramètres typiques qui sont les suivants :

- Paramètres de position
- Paramètres de dispersion
- Paramètres de dissymétrie et d'aplatissement

1.3.1 Paramètres de position

Ce sont des valeurs moyennes qui servent à caractériser l'ordre de grandeur des observations. Ce sont principalement :

- La moyenne arithmétique
- La moyenne géométrique
- La moyenne harmonique
- La moyenne quadratique
- La médiane
- Le mode

a) La moyenne arithmétique : $\bar{y}$

y_i = les valeurs observées

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Propriétés :

*) la somme des $y_i - \bar{y}$ est nulle

*) c'est par rapport à cette moyenne que la somme de carrés des écarts est la plus petite.

b) La moyenne géométrique : $\bar{y}_g$

La moyenne géométrique $\bar{y}_g$ d'une série statistique composée de n valeurs positives $y_1 + y_2 + \cdots + y_n$ est par définition, la racine n^{ième} du produit de ces n valeurs :

$$\bar{y}_g = \sqrt[n]{x_1 x_2 \ldots x_n} \quad (x_i \geq 0)$$

c) La moyenne harmonique : $\bar{y}_h$

la moyenne harmonique d'une série de n valeurs positives est égale à l'inverse de la moyenne arithmétique des inverses

$$\bar{y}_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \quad (x_i \neq 0)$$

d) La moyenne quadratique : la moyenne quadratique d'une série des n valeurs positives, nulles ou négative, est la racine carrée de la moyenne arithmétique des carrés

$$\bar{y}_q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

e) La médiane :

La médiane est la valeur qui laisse 50% des observations en-dessous et 50% des observations au-dessus. On l'appelle également parfois "percentile 50" : c'est la valeur centrale par excellence.

Pour la calculer, il faut d'abord trier l'échantillon. Ensuite,

– si l'effectif n de l'échantillon est impair,

$$M = x_{(\frac{n+1}{2})}$$

– si l'effectif n de l'échantillon est pair,

$$M = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

La médiane a, comme propriété, d'être peu sensible aux valeurs extrêmes.

f) Le mode est la valeur la plus fréquente dans l'échantillon. C'est la valeur dominante = la valeur observée qui a la fréquence maximum. On peut avoir des séries unimodales, bimodales et plurimodales

Remarque : la moyenne arithmétique est la plus couramment utilisée.

1.3.2 Paramètre de dispersion

Ce sont des paramètres qui permettent de chiffrer la variabilité des valeurs observées, autour d'un paramètre de position, ce sont principalement :

La variance

L'écart-type

Le coefficient de variation

L'écart-moyen absolu

L'amplitude

a) La variance : c'est la moyenne arithmétique des carrés des écarts par rapport à la moyenne

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{SCE}{n}$$

Et la formule pratique de calcul est :

$$S^2 = \frac{1}{n}\left[\underbrace{\sum_{i=1}^{n} y_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}_{\text{SCE}}\right]$$

b) L'écart-type : ou l'écart quadratique moyen c'est la racine carrée de la variance :

$$S = \sqrt{S^2} = \sqrt{\frac{SCE}{n}}$$

c) Le coefficient de variation : il est utilisé pour comparer la variabilité relative de plusieurs séries statistiques

$$CV = \frac{S}{\bar{x}} \text{ ou } 100 \times \frac{S}{\bar{x}} = \%$$

Remarques

- ✓ s^2,s et CV sont nuls si et seulement si tous les écarts $y_i - \bar{y}$ sont nuls, c'est-à-dire si toutes les valeurs observées sont égales entre elles, et à leur moyenne ou plus simplement, s'il n'y a pas de variabilité dans les observations

- ✓ l'unité suit l'unité des donnée observée s^2(unité au carré), s (unité), CV est sans unité ou en pourcentage.

d) L'écart-moyen absolu : e_m

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

e) L'amplitude : w

C'est l'écart entre les valeurs extrêmes d'une série d'observation classées par ordre croissants.

$$w = x_{(n)} - x_{(1)}$$

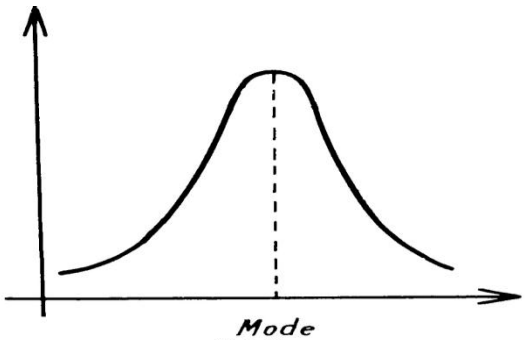
1.3.3 Paramètre de forme ou de dissymétrie et d'aplatissement : coefficient de PEARSON et FISHER

a) Moments centrés d'ordre k

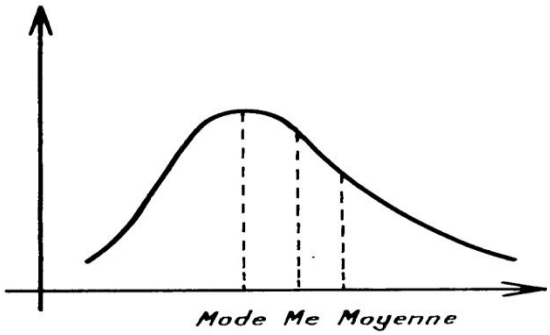
- moyenne arithmétique des écarts à la moyenne élevée à la puissance k.
- si k pair => paramètre de dispersion.
- si k impair => paramètre de symétrie.

b) Coefficient de Pearson et de Fisher

- b1 pour caractériser la *symétrie* de la courbe; b2 pour caractériser l'*aplatissement.*
- b1 = M_3^2 / M_2^3 : est voisin de 0 si la distribution est symétrique.
- b2 = M_4 / M_2^2 : est voisin de 3 si la distribution suit une loi normale (plus aplatie qu'elle si b2 < 3).



Distribution symétrique



Distribution asymétrique

Chapitre 2 : La régression : statistique descriptive à deux dimensions ou à deux variables ou bi variées

Ce chapitre a pour but de mettre en évidence les relations qui existent entre deux séries d'observations considérées simultanément. Ici aussi on distingue 3 méthodes (DAGNELIE. P., 2006) :

2.1 L'élaboration de tableaux statistiques: permettant de condenser les données sous formes de distribution de fréquences

2.2 La représentation graphique:

a) Diagrammes de dispersion ou nuage de points qui sont obtenus en représentant chaque couple d'observation (x_i,y_i) par un points dans un plan (x,y).

b) Diagrammes en bâtons } Pour distribution de fréquences à deux dimensions

c) Les stéréogrammes

2.3-Réduction des données: il existe deux types de paramètres:

- les paramètres relatifs à une seule variable: sont des paramètres qui ne concernent qu'une variable à la fois: la moyenne, la variance, et l'écart-type.
- les paramètres relatifs aux deux variables: qui servent à décrire les relations entre les deux variables prises simultanément sont:

- La covariance

- Le coefficient de corrélation

- Le coefficient de détermination

- Les variances résiduelles

- Les droites de régression des moindres carrés

2.3.1 La covariance: caractérise simultanément les deux séries d'observation. Elle est positive ou négative selon que la relation entre les deux séries de données est croissante ou décroissante, c'est-à-dire selon que les valeurs élevées d'une série correspondant, dans l'ensemble, aux valeurs élevées ou aux valeurs peu élevées de l'autre.

Ex : Le tableau suivant montre les données observées de poids et de taille de 5000 étudiants :

	Poids y_i	Taille x_i
$i = 1$	$y_1 = 75,2$	$x_1 = 1,55$
$i = 2$	$y_2 = 82,5$	$x_2 = 1,82$
:	:	:
:	:	:
$n = 5000$	$y_{5000} = 65,0$	$x_{5000} = 1,92$

$$Cov(x.y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

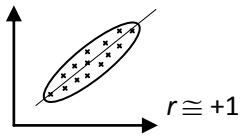
$$Cov(x.y) = \frac{SPE}{n}$$

$$Cov(x.y) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i \times y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \times \left(\sum_{i=1}^{n} y_i\right)\right]$$

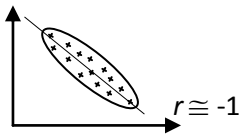
$$\boxed{SPE = \sum xi \times yi - \frac{1}{n}\left(\sum xi\right) \times \left(\sum yi\right)}$$

	Poids y_i	Taille x_i	y_i^2	x_i^2	$x_i \times y_i$
$i = 1$	$y_1 = 75,2$	$x_1 = 1,55$	$(75,2)^2$	$(1,55)^2$	$75,2 \times 1,55$
$i = 2$	$y_2 = 82,5$	$x_2 = 1,82$	$(82,5)^2$	$(1,82)^2$	$82,5 \times 1,82$
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
$n = 5000$	$y_{5000} = 65,0$	$x_{5000} = 1,92$	$(65,0)^2$	$(1,92)^2$	$65,0 \times 1,92$
	$\sum y_i$	$\sum x_i$	$\sum y_i^2$	$\sum x_i^2$	$\sum x_i \times y_i$

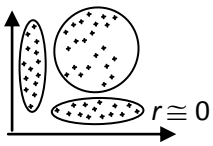
2.3.2 Le coefficient de corrélation r : sert à mesurer l'intensité de la relation qui existe entre les deux séries de données pour autant que cette relation soit linéaire ou approximativement linéaire.



Forte corrélation positive



Forte corrélation négative



Pas de corrélation

$$\boxed{-1 \le r \le +1}$$

-1 0 +1

$$r = \frac{Cov(x.y)}{Sx\ Sy}$$

Test de signification de la corrélation

Le premier test qui vient à l'esprit est la significativité de la corrélation c'est-à-dire le coefficient de corrélation est-il significativement différent de 0 ?

Le test s'écrit :

$$H_0 : r=0$$

$$H_1 : r \neq 0$$

Remarque : (autres hypothèses alternatives). On peut vouloir définir une hypothèse alternative différente ($H_1 : r<0$ ou $H_1 : r > 0$). Les caractéristiques des distributions restent les mêmes.Pourun risque α donné, seul est modifié le seuil de rejet de H_0 puisque le test est unilatéral dans ce cas.

Le test étudié est paramétrique. On suppose a priori que le couple (X.Y) suit une loi normale bi-variée.Dans ce cas : la distribution sous H_0 de la statistique du test que nous présenterons plus bas est exact : le test de significativité équivaut à un test d'indépendance.

Cette restriction est moins contraignante lorsque n est suffisamment grand (RAKOTOMALALA R., 2015). A partir de 25 observations, l'approximation est bonne ,même si nous écartons (un peu) de la distribution normale conjointe. La distribution est valable sous l'hypothèse r=0. Mais le test de significativité revient simplement à tester l'absence ou la présence de corrélation.

Statistique du test : Sous H_0, la statistique :

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-2}}}$$

Suit une loi de Student à (n-2) degrés de liberté. L'hypothèse nulle est rejetée si :

$$|t| > t_{1-\frac{\alpha}{2}}$$

2.3.3 Le coefficient de détermination r^2: est égale à la part de la variation de y qui est expliquée par la régression de y en x.

$$\boxed{r^2 = (r)^2 = \text{Coefficient de détermination (\%)}}$$

$$\boxed{0 \leq r^2 \leq 1}$$

2.3.4 Les droites de régression au sens des moindres carrés et les variances résiduelles

La droite de régression de y en x a pour but de résumer le nuage de points c'est-à-dire de représenter sur le plan l'allure de la distribution à deux caractères. Cette droite donne une idée de la façon dont varie en moyenne la variable y dite dépendante ou variable à expliquer, en fonction de la variable x, dite indépendante ou explicative. La droite est appelée également diagramme de régression.
Lorsque le diagramme de régression est linéaire ou approximativement linéaire, on peut s'efforcer de rechercher l'équation de la droite qui s'y ajuste le mieux (DAGNELIE. P., 2006).

2.3.5 La variance résiduelle de y apparait ainsi comme un indice de dispersion des points observés autour de la droite de régression de y en x

La quantité $cov^2(x,y)/s^2_x$ est considéré comme la part de la variance de y qui est expliquée ou justifiée par régression de y en x, tandis que la variance résiduelle est la part de cette variance qui ne peut être expliquée de la sorte

$$S^2_{y.x} = \frac{\sum (yi_{obs} - \hat{y}i_{est})^2}{n}$$

L'écart-type résiduel: c'est la racine carrée de la variance résiduelle, ce paramètre mesure la dispersion des points observés autour de la droite de régression. C'est l'erreur que l'on connaitrait si l'on estime à l'aide del'équation

$$\boxed{y = a + bx \Leftrightarrow \text{Equation de régression}}$$

$$S_{y.x} = \sqrt{S^2_{y.x}}$$

Remarque: L'équation de la régression est utilisée le plus souvent dans un but de prévision ou d'estimation.

e_i: l'erreur d'ajustement

L'équation de la droite est alors de la forme y=a+bx cette droite passe par le point moyen($\bar{x}$, $\bar{y}$).

$$\bar{y} = a + b\bar{x} \Rightarrow a = \bar{y} - b\bar{x} \quad \text{et} \quad b = \frac{\text{cov}(x,y)}{s_x^2}$$

On appelle résidus de y par rapport à x les écarts: y_i-$\hat{y}(x_i)$ entre les points correspondants de la droite de régression de y en x. Ces écarts sont de somme et de moyenne nulle ce sont ces valeurs qu'on ne peut pas expliquer.

La variance résiduelle de y est la variance de ces résidus.

Chapitre 3 : les méthodes statistiques relatives aux moyennes

Ces méthodes sont subordonnées à deux conditions :

- la normalité des populations
- le caractère aléatoire et simple des échantillons

Dans le cas de certains tests relatifs aux moyennes ces méthodes nécessitent une troisième condition :

- l'égalité des variances des populations.

3.1 Intervalle de confiance et le test de conformité d'une moyenne

3.1.1 Intervalle de confiance

Estimation de la moyenne : il est dit que la meilleure estimation de la moyenne $\hat{m}$ de la population est donnée par la moyenne $\bar{x}$ de l'échantillon $\boxed{\hat{m} = \bar{x}}$

Intervalle de confiance de la moyenne estimée

- Dans le cas où la variance de la population parent est connue les limites de l'intervalle de confiance sont alors :

$$\bar{x}_{\inf} = \bar{x} - \mu_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \qquad \bar{x}_{\sup} = \bar{x} + \mu_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

- Dans le cas où la variance de la population parent n'est pas connue alors il faut l'estimer à partir de la variance de l'échantillon $\hat{\sigma}^2 = \frac{SCE}{n-1} = \frac{n \cdot S^2}{n-1} \Rightarrow \hat{\sigma} = \sqrt{\frac{n \cdot S^2}{n-1}}$

donc : $\bar{x}_{\inf} = \bar{x} - \mu_{1-\frac{\alpha}{2}} \times \sqrt{\frac{SCE}{n(n-1)}} \qquad \bar{x}_{\sup} = \bar{x} + \mu_{1-\frac{\alpha}{2}} \times \sqrt{\frac{SCE}{n(n-1)}}$

Dans la pratique quand *n* est supérieur à 30 ($n > 30$) on remplace $\mu_{1-\frac{\alpha}{2}}$ par la valeur $2 \Rightarrow \mu_{1-\frac{\alpha}{2}} = 2$.

Cependant pour des effectifs inférieurs à 30 ($n < 30$) on utilise la loi t de STUDENT et on remplace $\mu_{1-\frac{\alpha}{2}}$ par $t_{1-\frac{\alpha}{2}} \Rightarrow \mu_{1-\frac{\alpha}{2}} = t_{1-\frac{\alpha}{2}}$ pour $\begin{cases} \alpha = 0{,}05 \\ (n-1)\text{ddl} \end{cases}$

Application : dans une forêtdistincte on a pris au hasard 12 arbres et on a mesuré leurs hauteurs :

20,4 ; 25,4 ; 25,6 ; 25,6 ; 26,6 ; 28,6 ; 28,7 ; 29,0 ; 29,8 ; 30,5 ; 30,9 ; 31,1.

Estimer la moyenne et calculer l'intervalle de confiance ?

$$\bar{x} = \frac{1}{12}\sum xi = 27{,}68\,\text{m} \qquad \boxed{\hat{m} = \bar{x} = 27{,}68\,\text{m}}$$

$$\hat{\sigma} = \sqrt{\frac{SCE}{n-1}} = \sqrt{\frac{106{,}16}{11}} = \sqrt{9{,}63} = 3{,}107\,\text{m}$$

$$\bar{x}_{\inf} = 27{,}68 - 2{,}201 \times \frac{3{,}107}{\sqrt{12}} = 25{,}70$$

$$\bar{x}_{\sup} = 27{,}68 + 2{,}201 \times \frac{3{,}107}{\sqrt{12}} = 29{,}65$$

3.1.2 Test de conformité d'une moyenne

Test de conformité d'une moyenne: le test de conformité d'une moyenne à pour but de vérifier si la moyenne *m* d'une population est ou n'est pas égale à une valeur donnée m_0. $H_0 : m = m_0$ et on rejette cette hypothèse lorsque la moyenne observé $\bar{x}$ est trop différente de la moyenne théorique m_0

Le test se réalise en calculant les quantités suivantes :

$$\text{si } t_{obs} \geq t_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow m \neq m_0 \begin{cases} \alpha = 0{,}05 \\ (n-1)\,\text{ddl} \end{cases}$$

$$t_{obs} = \frac{|\bar{x} - m_0|}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{|\bar{x} - m_0|}{\sqrt{\frac{SCE}{n(n-1)}}}$$

Application : supposant que l'on souhaite vérifier si la forêt dans laquelle les 12 mesures ont été réalisées appartient, quand à la hauteur, à un type de forêt donné, dont la moyenne est parfaitement connue et égale à 29 m.

$$H_0 : 27{,}68 = 29 \qquad t_{obs} = \frac{|27{,}68 - 29|}{\frac{3{,}107}{\sqrt{12}}} = 1{,}47 \text{ et } t_{1-\frac{\alpha}{2}} = 2{,}201 \qquad \text{donc } t_{obs} < t_{1-\frac{\alpha}{2}} \Rightarrow AH_0 \Rightarrow m = m_0$$

3.2 Le test de signification et l'intervalle de confiance d'une différence de deux moyennes : échantillons indépendants

Le cas de populations de même variance : On supposant satisfaite les conditions précédentes et en admettant que les échantillons sont indépendants et que les populations sont de même variance on peut donc tester l'hypothèse d'égalité suivante : $H_0 : m_1 = m_2$

Si $n_1 \neq n_2$

$$t_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad t_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{Si } t_{obs} \geq t_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow m_1 \neq m_2 \quad \text{pour} \quad \begin{cases} \alpha = 0{,}05 \\ (n_1 + n_2 - 2)\,\text{ddl} \end{cases}$$

Ce test est appelé test t de STUDENT ou test de STUDENT - FISHER.

Si $n_1 = n_2 = n$ la formule se simplifie considérablement et elle sera comme suit :

$$t_{obs} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}}}$$

$$\text{Si } t_{obs} \geq t_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow m_1 \neq m_2 \quad \text{pour} \quad \begin{cases} \alpha = 0{,}05 \\ 2(n-1)\,\text{ddl} \end{cases}$$

Application : Dans 2 types de forêt distincte on a mesuré les hauteurs respectivement de 13 et 14 arbres choisis au hasard et indépendamment, dans le but de vérifier si les hauteurs moyennes des 2 types de forêt sont ou ne sont pas égale, les valeurs observées sont les suivantes :

Type 1	Type 2
23,4	22,5
24,4	22,9
24,6	23,7
24,9	24,0
25,0	24,4
26,2	24,5
26,3	25,3
26,8	26,0
26,8	26,2
26,9	26,4

27,0	26,7
27,6	26,9
27,7	27,4
	28,5

$\bar{x}_1 = 25{,}97$ $\bar{x}_2 = 25{,}39$

$SCE_1 = 22{,}15$ $SCE_2 = 40{,}88$

$$t_{obs} = \frac{\left|25{,}97 - 25{,}39\right|}{\sqrt{\frac{22{,}15 + 40{,}88}{25}\left(\frac{1}{13} + \frac{1}{14}\right)}} = 0{,}95$$

$$t_{1-\frac{\alpha}{2}} = t_{1-\frac{0{,}05}{2}} = t_{0{,}975} = 2{,}060$$

$$n_1 + n_2 - 2 = 25 \text{ ddl}$$

On constate que $t_{obs} = 0{,}95$ est inférieure à $t_{1-\alpha/2} = 2{,}060$ par conséquent on accepte l'hypothèse d'égalité de la moyenne des 2 types de forêt ; c'est-à-dire qu'il n'existe pas de différences significatives entre m_1 et m_2

Remarque **:** quand $n_1 \neq n_2$ avant d'appliquer le test *t* de STUDENT d'égalité de 2 moyennes, il faut toujours vérifier l'hypothèse d'égalité des 2 variances.

Chaque fois qu'on rejette l'hypothèse d'égalité de 2 moyennes il faut alors estimer la différence des 2 moyennes et calculer sont intervalle de confiance.

Quand $n_1 \neq n_2$: $\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{1-\frac{\alpha}{2}} \sqrt{\dfrac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$ $\begin{cases} \alpha = 0{,}05 \\ \left(n_1 + n_2 - 2\right) ddl \end{cases}$

Quand $n_1 = n_2 = n$: $\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{1-\frac{\alpha}{2}} \sqrt{\dfrac{SCE_1 + SCE_2}{n(n-1)}}$ $\begin{cases} \alpha = 0{,}05 \\ 2\left(n-1\right) ddl \end{cases}$

Le cas de populations de variance inégale : Plusieurs auteurs ont montré que l'hypothèse de normalité est secondaire dans le test d'égalité de 2 moyennes. De même l'hypothèse d'égalité des variances n'est pas fondamentale lorsque les effectifs des échantillons sont égaux $\left(n_1 = n_2\right)$. Quand le test est non sensible à la non normalité et à l'inégalité de variance, le test est alors robuste, par contre, lorsque $n_1 \neq n_2$ il est absolument indispensable de s'assurer de l'égalité de variance. Si cette hypothèse n'est pas vérifiée, il est indispensable d'utiliser une méthode adaptée à ces circonstances, on peut procéder à une transformation de variable destinée à stabiliser la variance et utiliser ensuite le test *t* de STUDENT (DAGNELIE.P .,1999).

.

3.3 Le test de signification et l'intervalle de confiance d'une différence de deux moyennes : échantillons associés par paires.

3.3.1 Introduction

Un autre cas important de comparaison de moyenne est relatif aux échantillons dont les individus sont associés par paires ou par couple ex : les mêmes individus soumis à 2 méthodes différentes (comparaison de 2 méthodes). Pour tester l'égalité des moyennes, on doit alors considérer la population de différence et vérifier la nullité de la moyenne de ces différences, les conditions d'application du test sont alors :

- le caractère aléatoire simple des échantillons.
- La normalité de la population de différence.

Le test d'égalité de moyennes s'écrit alors : $H_0 : m_1 = m_2$ ou $\delta = 0$

L'hypothèse se réalise en calculant les différences suivantes :

n_1 y_i	n_2 x_i		di
y_1 y_2 : y_n	x_1 x_2 : x_n	$y_1 - x_1$ $y_2 - x_2$: $y_n - x_n$	d_1 d_2 : d_n
$\bar{y}$	$\bar{x}$		$\bar{d}$

$$t_{obs} = \frac{|\bar{d}|}{\sqrt{\frac{SCEd}{n(n-1)}}} = \frac{|\bar{y} - \bar{x}|}{\sqrt{\frac{SCEd}{n(n-1)}}}$$

$$SCEd = \sum di^2 - \frac{1}{n}\left(\sum di\right)^2$$

Si $t_{obs} \geq t_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow m_1 \neq m_2$ pour $\begin{cases} \alpha = 0{,}05 \\ (n-1)ddl \end{cases}$

Ce test est appelé test *t* de STUDENT pour échantillons associés par paires ou par couple.

* Dans le cas du rejet de l'hypothèse d'égalité des 2 moyennes il faut alors estimer la différence des 2 moyennes est calculer l'intervalle de confiance de cette différence.

$\hat{\delta} = m_1 - m_2 \quad \bar{d} \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{SCEd}{n(n-1)}}$ pour $\begin{cases} \alpha = 0{,}05 \\ (n-1)ddl \end{cases}$

Application : On a mesuré les hauteurs de 12 arbres avec 2 méthodes différentes :

Arbres debout	**Arbres abattus**	*Di*
20,4	21,7	-1,3
25,4	26,3	-0,9
25,6	26,8	-1,2
25,6	28,1	-2,5
26,6	26,2	+0,4
28,6	27,3	+1,3
28,7	29,5	-0,8
29,0	32,0	-3,0
29,8	30,9	-1,1
30,5	32,3	-1,8
30,9	32,3	-1,4
31,1	31,7	-0,6
$\bar{x}_1 = 27{,}68$	$\bar{x}_2 = 28{,}76$	$\bar{d} = -1{,}08$

$$H_0 : m_1 = m_2$$

$$SCEdi = 28{,}45 - \frac{(12{,}9)^2}{12} = 14{,}58$$

$$t_{obs} = \frac{1{,}08}{\sqrt{\frac{14{,}58}{132}}} = 3{,}25 \qquad t_{1-\frac{0{,}05}{2}} = t_{0{,}975} = 2{,}201 \text{ pour } \begin{cases} \alpha = 0{,}05 \\ (n-1)ddl \end{cases}$$

$$t_{obs} = 3{,}25 > t_{1-\frac{\alpha}{2}} = 2{,}201 \Rightarrow RH_0 : m_1 \neq m_2$$

donc : $d_{\text{inf}} = 1{,}81$ m et $d_{\text{sup}} = 0{,}35$ m

Chapitre 4 : Les méthodes statistiques relatives aux variances

4.1. Introduction : toutes les méthodes proposées sont applicables dans le cas des conditions suivantes :

- les échantillons sont aléatoires, simples et indépendants.
- la normalité des populations parents.

4.2 Estimation de la variance de la population et intervalle de confiance

$$S^2 = \frac{SCE}{n} \qquad \hat{\sigma} = \sqrt{\frac{SCE}{n-1}} = \sqrt{\frac{n \cdot S^2}{n-1}}$$

En pratique, et dans les conditions définies précédemment, les limites de confiance ($S_{inf} - S_{sup}$) sont donc pour un niveau de signification α donné :

- **Quand** $n \leq 30$: $S_{inf} = \sqrt{\frac{SCE}{\chi^2_{1-\frac{\alpha}{2}}}}$ et $S_{sup} = \sqrt{\frac{SCE}{\chi^2_{\frac{\alpha}{2}}}}$ pour $\begin{cases} \alpha = 0{,}05 \\ (n-1)ddl \end{cases}$

- **Quand** $n > 30$: $\chi^2 = \frac{\left(\mu + \sqrt{2n-1}\right)^2}{2}$ $S_{inf} = \frac{\sqrt{2 \cdot SCE}}{\sqrt{2n-3} + \mu_{1-\frac{\alpha}{2}}}$

et $S_{sup} = \frac{\sqrt{2 \cdot SCE}}{\sqrt{2n-3} - \mu_{1-\frac{\alpha}{2}}}$

4.3 Test de conformité d'une variance

$H_0 : \sigma^2 = \sigma_0^2$ ⟶ Valeur théorique si $k \leq 30$

En pratique on calcule le rapport suivant : $\chi^2_{obs} = \frac{SCE}{\sigma_0^2}$

$$\left.\begin{array}{l} \text{si} \quad \chi^2_{obs} < \chi^2_{\frac{\alpha}{2}} \\ \quad ou \\ \text{si} \quad \chi^2_{obs} \geq \chi^2_{1-\frac{\alpha}{2}} \end{array}\right\} \Rightarrow RH_0 \Rightarrow \sigma^2 \neq \sigma_0^2$$

Quand $k > 30$: le test peut être réalisée d'une manière approchée en calculant la quantité suivante :

$$\mu_{obs} = \left| \sqrt{\frac{2SCE}{\sigma_0^2}} - \sqrt{2n-3} \right| \qquad \text{si} \quad \mu_{obs} \geq \mu_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow \sigma^2 \neq \sigma_0^2$$

Application 1: reprenant une fois encore les 12 mesures de hauteur des arbres considérés précédemment et estimons la variance des hauteurs de toute la forêt et calculons ses limites de confiance.

$$\hat{\sigma}^2 = \frac{SCE}{n-1} = \frac{106{,}16}{11} = 9{,}651\,\text{m}^2 \Rightarrow \hat{\sigma} = 3{,}106\,\text{m}$$

$$S_{\text{inf}}^2 = \frac{106{,}16}{21{,}9} = 4{,}85 \Rightarrow S_{\text{inf}} = \sqrt{4{,}85} = 2{,}202$$

$$S_{\text{sup}}^2 = \frac{106{,}16}{3{,}82} = 27{,}8 \Rightarrow S_{\text{sup}} = \sqrt{27{,}8} = 5{,}272$$

Application 2: supposer que pour une race bovine donnée dans une région donnée on ait mesuré la production laitière de 50 bêtes choisis au hasard et indépendamment et que l'on ait obtenu comme SCE par rapport à la masse : *SCE* = 37.173.400 Kg^2

$$\hat{\sigma} = \sqrt{\frac{SCE}{n-1}} = \sqrt{\frac{37.173.400}{49}} = 871\,\text{Kg}$$

Calculons l'intervalle de confiance :

$$S_{\text{inf}} = \frac{\sqrt{2(37.173.400)}}{\sqrt{97} + 1{,}96} = 730\,\text{Kg}$$

$$S_{\text{sup}} = \frac{\sqrt{2(37.173.400)}}{\sqrt{97} - 1{,}96} = 1093\,\text{Kg}$$

$S_{\text{inf}} = 728Kg \,//\, 730Kg$

$S_{\text{sup}} = 1085Kg \,//\, 1093Kg$

Supposons maintenant que l'écart type estimé appartient à une production laitière ayant un écart type égal à 1000 Kg.

$$H_0 : \underset{\sigma}{871} = \underset{\sigma_0}{1000} \qquad \mu_{obs} = \left| \sqrt{\frac{2(37.173.400)}{1.000.000}} - \sqrt{97} \right| = 1{,}227 \qquad \mu_{1-\frac{\alpha}{2}} = 1{,}96$$

$$\mu_{obs} < \mu_{1-\frac{\alpha}{2}} \Rightarrow AH_0 \Rightarrow \sigma = \sigma_0 \Rightarrow 871 = 1000$$

4.4 Les tests de comparaisons et l'intervalle de confiance de rapport de 2 variances

4.4.1 Le test d'égalité de 2 variances $H_0 : \sigma_1^2 = \sigma_2^2$

Dans la pratique, on calcule le rapport des 2 variances en mettant la variance maximale en numérateur et la variance minimale en dénominateur et on calcule le rapport suivant :

Quand : $n_1 \neq n_2 \Rightarrow$

$$F_{obs} = \frac{\hat{\sigma}^2_{\max}}{\hat{\sigma}^2_{\min}} \quad \text{si} \quad F_{obs} \geq F_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2 \quad \text{pour} \quad \begin{cases} \alpha = 0{,}05 \\ k_1(n_1 - 1)ddl \\ k_2(n_2 - 1)ddl \end{cases}$$

Quand : $n_1 = n_2 \Rightarrow \quad F_{obs} = \frac{\hat{\sigma}^2_{\max}}{\hat{\sigma}^2_{\min}} \Rightarrow F_{obs} = \frac{SCE_{\max}}{SCE_{\min}}$ si $F_{obs} \geq F_{1-\frac{\alpha}{2}} \Rightarrow RH_0 \Rightarrow \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2$

pour $\begin{cases} \alpha = 0{,}05 \\ k_1(n-1)ddl \\ k_2(n-1)ddl \end{cases}$

Ce test est appelé test *F* de FISHER

Application:

Les données de l'application : $n_1 = 13$ $\quad \bar{x}_1 = 25{,}97$ $\quad SCE_1 = 22{,}15$ $\quad n_2 = 14$

$\bar{x}_2 = 25{,}39$

$SCE_2 = 40{,}88$

$\hat{\sigma}_1^2 = \frac{SCE}{n-1} = \frac{22{,}15}{12} = 1{,}846$; $\sigma_2^2 = \frac{SCE}{n-1} = \frac{40{,}88}{13} = 3{,}145 \Rightarrow \hat{\sigma}_1^2 = 1{,}846 = \hat{\sigma}^2_{\min}$;

$\sigma_2^2 = 3{,}145 = \sigma^2_{\max}$

$F_{obs} = \frac{\hat{\sigma}^2_{\max}}{\hat{\sigma}^2_{\min}} = \frac{3{,}145}{1{,}846} = 1{,}703$; $F_{1-\frac{\alpha}{2}} \approx 3{,}26$ pour $\begin{cases} \alpha = 0{,}05 \\ k_1 = 13 \\ k_2 = 12 \end{cases}$

$F_{obs} < F_{1-\frac{\alpha}{2}}$

On constate que $F_{obs} < F_{1-\frac{\alpha}{2}}$ donc AH_0 d'égalité des 2 variances ($\hat{\sigma}_1^2 = \hat{\sigma}_2^2$).

4.4.2 Deux tests d'égalité de plusieurs variances : deux méthodes sont utilisées pour tester l'égalité de variance de plusieurs populations :

- le test de BARTLETT : pour les échantillons *d'effectifs différents* et aussi pour des *effectifs constants*. Ce test est long à réaliser.

- le test de HARTLEY : il est d'un usage beaucoup plus rapide mais il ne s'applique qu'à des échantillons de *même effectifs*. Dans les deux cas les conditions d'application doivent être très strictes.

a) Le test de BARTLETT : considérant *p* échantillons aléatoires, simples des effectifs $n_1, n_2, ..., n_P$

$H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = ... = \hat{\sigma}_p^2$ Calculons pour chaque échantillon la *SCE* et la $\hat{\sigma}^2$, calculons aussi la *SCE* de la variance estimée relative à l'ensemble des observations

$n\bullet = n_1 + n_2 + ... + n_P \quad n\bullet = \sum_{i=1}^{p} n_i$

$$SCE = \sum_{i=1}^{k} SCE_i = SCE_1 + SCE_2 + ... + SCE_p \qquad \hat{\sigma}^2 = \frac{SCE}{n\bullet - p}$$

En pratique le test se réalise en calculant la quantité suivante :

Quand : $n_1 \neq n_2 \neq ... \neq n_P$

$$\chi^2_{obs} = \frac{2{,}3026\left\{(n\bullet - p)\log_{10}\hat{\sigma}^2 - \sum_{i=1}^{p}\left[(n_i - 1)\log_{10}\hat{\sigma}_i^2\right]\right\}}{1 + \frac{1}{3(p-1)}\left[\sum_{i=1}^{p}\frac{1}{n_i - 1} - \frac{1}{n\bullet - p}\right]}$$

Quand : $n_1 = n_2 = ... = n_P = n$ $\qquad H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = ... = \hat{\sigma}_p^2$

$$\chi^2_{obs} = \frac{2{,}3026(n-1)\left[p\log_{10}\frac{SCE}{p} - \sum_{i=1}^{p}\log_{10} SCE_i\right]}{1 + \frac{p+1}{3p(n-1)}}$$

si $\chi^2_{obs} \geq \chi^2_{1-\alpha} \Rightarrow RH_0 \Rightarrow \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2 \neq ... \neq \hat{\sigma}_p^2$ pour $\begin{cases} \alpha = 0{,}05 \\ (p-1)ddl \end{cases}$

Application : considérons 3 types de forêts dans lesquelles nous avons prélevé des échantillons d'effectifs inégaux et pour lesquels nous avons mesuré les hauteurs.

n = 13	n = 14	n = 10
Type 1	Type 2	Type 3
23,4	22,5	18,9
24,4	22,9	21,1
24,6	23,7	21,2
24,9	24,0	22,1
25,0	24,4	22,5
26,2	24,5	23,6
26,3	25,3	24,5
26,8	26,0	24,6
26,8	26,2	26,2
26,9	26,4	26,7
27,0	26,7	
27,6	26,9	
27,7	27,4	
	28,5	

Paramètres	Type 1	Type 2	Type 3	Totaux
n_i	13	14	10	$n\bullet$ = 37
SCE_i	22,15	40,88	53,62	SCE = 116,65
$\hat{\sigma}_i^2$	1,846	3,145	5,958	_
$\log_{10} \hat{\sigma}_i^2$	0,26623	0,49762	0,77510	_
$\frac{1}{(n_i - 1)}$	0,0833	0,0769	0,1111	0,2713

- Testez l'égalité des variances des hauteurs des 3 types de forêts

$$H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2$$

$\hat{\sigma}^2$ = 3,4309 $\log_{10} \hat{\sigma}^2$ = 0,53541

$$\chi^2_{obs} = \frac{2{,}3026 \{ 34(0{,}53541) - [12(0{,}26623) + 13(0{,}49762) + 9(0{,}77510)] \}}{1 + \frac{0{,}2713 - 0{,}0294}{6}}$$

$\chi^2_{obs} = 3{,}46$ $\chi^2_{1-0{,}05} = 5{,}99$ (Pour 2ddl) $\Rightarrow \chi^2_{obs} < \chi^2_{1-\alpha} \Rightarrow AH_0 \Rightarrow \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \hat{\sigma}_3^2$

La variance des hauteurs est égale pour les 3 types de forêts.

Remarques

- le test de BARTLETT est très sensible à la non normalité des populations parents quels que soient les effectifs des échantillons, de plus, il s'agit d'une méthode approximative, qui n'est satisfaisante que si les effectifs $n_i \geq 4$ et si le nombre d'échantillons p n'est pas trop élevé par rapport aux effectifs n_i ce test ne permet donc pas de comparer les variances d'un grand nombre de petits échantillons

- enfin, signalons que pour 2 populations, le test de BARTLETT n'est strictement équivalent au test F que si les 2 échantillons sont de même effectifs.

b) Le test de HARTLEY : lorsque les effectifs des échantillons sont constants et égaux à n le test de HARTLEY permet de vérifier plus rapidement l'hypothèse d'égalité des variances $H_0 : \hat{\sigma}_1^2 = \hat{\sigma}_2^2 = ... = \hat{\sigma}_p^2$ quand $n_1 = n_2 = ... = n_P = n$ en effet, ce test nécessite seulement le calcul des différentes SCE_i et le calcul des cautions ou du rapport des valeurs extrêmes.

$$H_{obs} = \frac{SCE_{\max}}{SCE_{\min}} \text{ si } \quad H_{obs} \geq H_{1-\alpha} \Rightarrow RH_0 \Rightarrow \hat{\sigma}_1^2 \neq \hat{\sigma}_2^2 \neq ... \neq \hat{\sigma}_P^2 \text{ pour} \begin{cases} \alpha = 0{,}05 \\ p \\ k(n-1)ddl \end{cases}$$

Chapitre 5 : L'analyse de la variance à un et à deux critères de classification

5.1 Analyse de la variance à un critère de classification

Le test d'analyse de la variance à un critère ou à un facteur de classification consiste à comparer plus de deux moyennes de plusieurs populations à partir des données d'échantillons aléatoires simples et indépendants (DAGNELIE.P., 1999).

La réalisation du test se fait soit en comparant la valeur de F_{obs} avec une valeur théorique $F_{1-\alpha}$ extraite à partir de la table F de FISHER pour un niveau de signification α=0,05, 0,01 ou 0,001 et pour K_1 et K_2 degrés de liberté, soit en comparant la valeur de la probabilité P avec toujours les différentes valeurs de α=5%, 0,1 % ou 1‰. Selon que cette hypothèse d'égalité des moyennes est rejetée au niveau α=0,05; 0,01 ou 0,001, on dit conventionnellement que l'écart observé est significatif, hautement significatif ou très hautement significatif (DAGNELIE.P.,1999).

$H_0 : m_1 = m_2 \Leftrightarrow$ test de STUDENT - FISHER

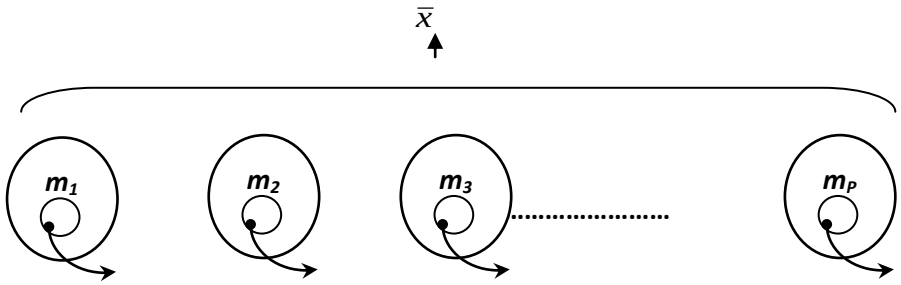
$H_0 : m_1 = m_2 = ... = m_P$

L'analyse de la variance à un critère de classification, ou à un facteur, a pour but de comparer les moyennes de plusieurs populations supposées normales et de même variance, à partir d'échantillons aléatoires simples et indépendants les uns des autres.

C'est une généralisation du test *t* de STUDENT pour deux échantillons indépendants.

5.1.1 Principe de l'analyse de la variance

Pour tester l'hypothèse d'égalité des moyennes de *p* populations $H_0 : m_1 = m_2 = ... = m_P$



$\hat{m}_1 = \bar{x}_1$

n_1	n_2	n_3	n_P
x_{n_1} $\bar{x}_1$	x_{n_2} $\bar{x}_2$	x_{n_3} $\bar{x}_3$	x_{n_p} $\bar{x}_P$

On prélève un échantillon aléatoire et simple dans chaque population, les moyennes de ces *p* échantillons et la moyenne générale de l'ensemble des observations $\bar{x}$ permettent de définir deux types de variations : les écarts existants entre les différents échantillons (variation entre échantillon ou variation factorielle) et les écarts existants à l'intérieur des échantillons (variation dans les échantillons ou variation résiduelle). L'importance de ces deux sources de variations est mesurée par deux quantités : carré moyen ou variance :

- le carré moyen factoriel est défini à partir des écarts entre les moyennes des différents échantillons et la moyenne générale (CM_f).
- le carré moyen résiduel (CM_r) est défini à partir des écarts existants chaque fois entre les valeurs observées et la moyenne de l'échantillon correspondant.

Lorsqu'il existe des différences importantes entre les moyennes des populations, on doit s'attendre à ce qu'il en soit de même pour les échantillons, on doit donc s'attendre aussi à observer un carré moyen factoriel élevé, par comparaison avec le carré moyen résiduel : le rapport du carré moyen factoriel au carré moyen résiduel est une mesure observée du degré de fausseté de l'hypothèse nulle.

Dans le cas où le test d'égalité de plusieurs moyennes concerne un nombre infini de population alors on fait un échantillonnage à deux degrés. Au 1er degré on choisit P populations d'individus ou de mesures (unité de 1er de degré). Au 2ème degré on choisit un échantillon de plusieurs individus ou de plusieurs observations (unité de 2ème degré) dans chacune de ces P populations.

L'H_0 concerne alors l'égalité des moyennes de l'infinité de populations.

$H_0 : m_1=m_2=\ldots=m_p$

- Quand le nombre de populations est fini le modèle est fixe
- Quand le nombre de populations est infini le modèle est aléatoire

5.1.2 Le modèle fixe

Considérons P échantillons d'effectifs $n_1, n_2, \ldots, n_P$: $n.=\sum ni$

Désignons par x_{ik} les valeurs observées, le symbole x_{ik} représentant d'une manière générale la $k^{ème}$ observation ($k= 1,2,..,n_i$) de l'échantillon extrait de la $i^{ème}$ population ($i=1,2,\ldots,p$) désignons en autre par $\bar{x}_i$ les moyennes des différents échantillons et $\bar{X}$ la moyenne générale :

$$\bar{X} = \frac{\sum \bar{x}_{ik}}{n_i} \text{ et } \bar{x}. = \left(\frac{1}{n_i}\right)\sum\sum x_{ik} = \left(\frac{1}{n_i}\right)\sum n_i x_i$$

On peut alors écrire le modèle observé « la variation totale se divise en deux composantes additives » : l'écart par rapport à la moyenne générale (variation totale)= l'écart des moyennes des échantillons par rapport à la moyenne générale (variation factorielle) + les écarts existants à l'intérieur des échantillons (variation résiduelle)

En élevant au carré les deux membres de cette identité et en sommant pour toutes les valeurs observées, on obtient l'équation de l'analyse de la variance à un critère de classification :

La somme des carrés des écarts totale peut elle aussi se diviser en deux composantes additives : la SCE factorielle (entre échantillons) et la SCE résiduelle (dans les échantillons)

$$SCE_t = SCE_f + SCE_r$$

Pour utiliser cette formule dans le test de l'analyse de la variance, nous supposerons que :

- Les P populations sont normales et de même écart-type
- Les échantillons sont aléatoires, simples et indépendants les uns des autres

Les carrés moyens sont : $CM_t = SCE_t/n.-1$ $\quad CM_f = SCE_f/P-1$

$$CM_r = SCE_r/n.-P$$

Lorsque l'hypothèse nulle est vraie, le quotient (résultat de division) $F_{obs}=CM_f/CM_r$ est donc une valeur observée d'une variable F de SNEDECOR à $k_1=P-1$ et $k_2= n.-P$ddl

La moyenne de cette variable F est voisine de l'unité, les valeurs attendues du numérateur et du dénominateur étant égales. Par contre lorsque l'H° est fausse, la valeur attendue du numérateur est supérieure à celle du dénominateur et d'autant plus grande que l'H° est plus fausse : la moyenne de la variable F est donc supérieure à l'unité et, elle aussi, d'autant plus grande que cette H° est plus fausse.

On rejette donc l'H° lorsque au niveau α : $F_{obs} \geq F_{1-\alpha}$ avec $\begin{cases} k_1= P-1\text{ddl} \\ k_2= n.-P\text{ddl} \end{cases}$

5.1.3 Le modèle aléatoire

Dans le cas du modèle aléatoire tout ce qui concerne les observations se présente sous la même forme que pour le modèle fixe :

- Le modèle observé.
- L'équation de l'AVI.
- Les Sommes des Carrés des Ecarts.
- Les nombres de ddl.
- Les Carrés Moyens.

5.1.4 Réalisation de l'Analyse de la variance

- **Quand les échantillons sont d'effectifs inégaux**

Les calculs peuvent être réalisés conformément au tableau en utilisant les notations et les formules suivantes :

- Pour l'effectif total : $n.= \sum n_i$
- Pour les totaux par échantillon : $X_{i.} = \sum x_{ik}$ (pour tout i)
- Pour le total général : $X_{..} = \sum X_{i.}$
- Pour la somme des carrés général : $T = \sum\sum x_{ik}^2$
- Pour la somme des carrés des écarts par échantillon : $SCE_i = \sum x_{ik}^2 - \frac{X_{i.}^2}{n_i}$
- Pour la somme des carrés des écarts résiduelle : $SCE_r= \sum SCE_i$

- $\bar{X} = \frac{X..}{n.}$ et $\bar{x}_i = \frac{X_i.}{n_i}$

Pour pouvoir dresser le tableau d'AVI, il reste à calculer :

- Le terme correctif : $C = X^2../n.$
- $SCE_t = T-C$
- $SCE_f = SCE_t - SCE_r$
- Les carrés moyens : $CM_f = SCE_f/P-1$ $CM_r = SCE_r/n.-P$
- Ainsi que le rapport $F_{obs} = CM_f/CM_r$

On peut alors réaliser le test de l'H° nulle, par comparaison de F_{obs} avec la valeur $F_{1-\alpha}$ dont les nombres de degrés de liberté sont : $k_1 = P-1$ et $k_2 = n.-P$

- **Quand les échantillons sont d'effectifs égaux**

Les calculs peuvent être simplifiés comme le montre le tableau de l'analyse de la variance. Les notions et les formules principales sont alors les suivantes :

- Pour l'effectif total : $n. = Pn$
- Pour les totaux par échantillon : $X_{i.} = \sum_{k=1}^{n} x_{ik}$
- Pour le total général : $X.. = \sum X_i.$
- Pour la somme des carrés général : $T = \sum\sum_{k=1}^{n} x_{ik}^2$
- Pour la somme des carrés des écarts par échantillon : $SCE_i = \sum x_{ik}^2 - X_i.^2/n$
- Pour la somme des carrés des écarts résiduelle : $SCE_r = \sum SCE_i$

Le tableau d'AVI est dressé comme dans le cas général grâce aux relations suivantes :

- Pour le terme correctif : $C = X..^2/Pn$
- Pour la somme des carrés des écarts totale : $SCE_t = T-C$
- Pour la somme des carrés des écarts factorielle : $SCE_f = SCE_t - SCE_r$
- Pour les carrés moyens : $CM_f = SCE_f/P-1$ $CM_r = SCE_r/P(n-1)$
- Et le rapport final : $F_{obs} = CM_f/CM_r$

5.2 Analyse de variance à deux critères de classification : Modèle croisés et Echantillons de mêmes effectifs

Objectifs

1. De réaliser une analyse de variance à deux critères de classification
2. De définir concrètement une interaction entre deux facteurs
3. D'interpréter les résultats d'une analyse de variance à deux critères de classification

Position du problème

Dans l'analyse de variance à un critère de classification, le principe consistait à diviser la variation totale en deux composantes :

- Factorielle
- Résiduelle

Cette façon de procéder être étendue à deux critère de classification, la variation totale étant alors divisée en plus de deux composantes : l'une résiduelle et les autres liées aux deux critères de classification.

Les deux facteurs considérés peuvent être placé sur le même pied (modèles croisés) ou subordonnés l'un à l'autre (modèles hiérarchisés).

Dans chaque cas, on doit distinguer un modèle fixe, un modèle aléatoire et un modèle mixte selon que les deux critères de classification sont fixes. Aléatoire, ou l'un fixe ; l'autre aléatoire.

Dans ce qui suivra, nous ne considérons que la réalisation et l'interprétation de l'analyse de variance à deux critères de classification pour des modèles croisés et des échantillons de mêmes effectifs.

5.2.1 Réalisation et interprétation de l'analyse de variance à deux critères de classification échantillons de plusieurs observations

Présentation des données et des calculs

La présentation des tableaux des données et des calculs se fera en deux parties.

A. Première partie

Tableau d'analyse de la variance à deux critères de classification.

I	1			.....	p			Totaux
J K	1	.....	Q	.....	1	.....	q	
1 2 . . . N	x_{111} x_{112} . . . x_{11n}	....	x_{1q1} x_{1q2} . . . x_{1qn}	....	x_{p11} x_{p12} . . . x_{p1n}	....	x_{pq1} x_{pq2} . . . x_{pqn}	
$X_{ij.}$ $\sum_{k=1}^{n} x^2_{ijk}$ $X^2_{ij}./n$ SCE_{ij}	$X_{11.}$ $\sum_{k=1}^{n} x^2_{11}k$ $X^2_{11}./n$ SCE_{11}	....	$X_{1q.}$ $\sum_{k=1}^{n} x^2_{1qk}$ $X^2_{1q.}/n$ SCE_{1q}	....	$X_{p1.}$ $\sum_{k=1}^{n} x^2_{p1k}$ $X^2_{p1}./n$ SCE_{p1}	....	$X_{pq.}$ $\sum_{k=1}^{n} x^2_{pqk}$ $X^2_{pq.}/n$ SCE_{pq}	$X_{...}$ T SCE_r

Réalisation des calculs avec les principales notions et formules suivantes:

- Pour les totaux par échantillon : $X_{ij.} = \sum_{k=1}^{n} x_{ijk}$ pour tout i et tout j
- Pour les totaux généraux: $X_{...} = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij.}$
- Pour la somme des carrés générale : $T = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} x^2_{ijk}$

- Pour le terme correctif : $C = \frac{X_{...}^2}{pqn}$
- Pour la somme des carrés des écarts totale: $SCEt = T - C$
- Pour la somme des carrés des écarts par échantillon : $SCE_{ij} = \sum_{k=1}^{n} x_{ijk}^2 - \frac{X_{ij.}^2}{n}$

  (pour tout i et tout j)
- Pour la somme des carrés des écarts résiduelle: $SCEr = \sum_{i=1}^{p}\sum_{j=1}^{q} SCE_{ij}$

  La différence (SCEt –SCEr) est une somme des carrés des écarts relative à l'ensemble des facteurs contrôlés.

  Cette différence sera en fait divisée en trois composantes factorielles:
  - Les deux premières liées chacune à l'un des 2 facteurs contrôlés
  - La troisième à l'interaction des 2 facteurs.

  Pour cela, il faut dresser un deuxième tableau de la façon suivante:

  - en reportant le contenu de la ligne $X_{ij.}$ du tableau précédant ;
  - et en calculant les sommes suivantes :

    $X_{i..} = \sum_{j=1}^{q} X_{ij.}$ (pour tout i)

    $X_{.j.} = \sum_{i=1}^{p} X_{ij.}$ (pour tout j)

B. Deuxième partie

Tableau d'analyse de la variance à deux critères de classification:

Suite de la réalisation des calculs

J i	1	.....	q	Xi..
1 . . . P	X_{11} . . . $X_{p1.}$	....	$X_{1q.}$ . . . $X_{pq.}$	$X_{1..}$ . . . $X_{p..}$
$X_{.J.}$	$X_{.1.}$	.....	$X_{.q.}$	X...

Dans ces conditions, on aura évidemment

$$\sum_{i=1}^{p} X_{i..} = \sum_{j=1}^{q} X_{.j.} = X_{...}$$

Les sommes des carrés des écarts liées aux deux facteurs seront:

$$SCEf = \frac{1}{qn} \sum_{i=1}^{n} X_{i..}^2 - C$$

$$SCEb = \frac{1}{pn} \sum_{j=1}^{q} X_{.j.}^2 - C$$

On obtient alors par différance:

$$SCEfb = SCEt - SCEr - SCEf - SCEb$$

Ou encore, l'équation de l'analyse de variance pour deux facteurs contrôlés simultanément:

$$SCEt = SCEf + SCEb + SCEb + SCEr$$

Cette équation indique donc que la variation totale (SCEt) peut être décomposée en 4 composantes principales:

Variation due au facteur a : *SCEf*

Variation due au facteur b : *SCEb*

Variation due à interaction entre les facteurs a et b : *SCEfb* .

Une somme résiduelle.

La notion d'interaction sera précisée plus loin

A ces différentes sommes des carrées sont affectés des nombres de degrés de liberté par la formule suivants :

$$pqn - 1 = (p-1) + (q-1) + (p-1)(q-1) + pq(n-1)$$

A cette étape, nous pouvons dresser le tableau d'analyse de la variance en calculant les nombres de degrés de liberté, les carrés moyens et les valeurs de F_f ,F_b et F_{ab}

$$Ff = \frac{\text{CMf}}{\text{CMr}} \qquad Fb = \frac{\text{CMb}}{\text{CMr}} \qquad Ffb = \frac{\text{CMfb}}{\text{CMr}}$$

Tableau d'analyse de variance

Source de variation	**ddl**	**SCE**	**CM**	**F**
Facteur a Facteur b Interaction Variation résiduelle	p-1 q-1 (p-1) (q-1) pq (n-1)	SCEf SCEb SCEfb SCEr	CMf CMb SCEfb CMr	Ff Fb Ffb
Totaux	pqn-1	SCEt		

5.2.2 Application pratique

Supposons que l'on veuille comparer, chez deux races bovines différentes (critère 1), les effets de 3 régimes alimentaire caractérisés par des teneurs énergétiques différentes (critère 2) : haut (H), bas (B) et moyen (M).

Le tableau suivant donne les résultats de la production laitière (en kg de lait/jour)

Obtenus avec chacun de ces 3 régimes. Pour chaque combinaison entre ces 2 critère, 4 valeurs sont données.

Tableau de comparaison des productions laitières (en Kg de lait/j), chez 2 races bovines différentes recevant 3 régimes énergétiques différents (H, B ou M).

	H (j = 1)	**B** (j= 2)	**M** (j=3)	Moyennes
Race 1 (i= 1)	33 35 36 43 **36,75**	31 32 33 34 **32,50**	32 34 36 38 **35,00**	34,75
Race 2 (i= 2)	30 30 30 33 **30,75**	25 27 30 30 **28,00**	27 29 30 30 **29,00**	29,25
Meyennes	33,75	30,25	32,00	32,00

Ainsi calculées, ces moyennes montrent une influence **considérable** du facteur (race). En effet, tous régimes confondus, la race 2 présente une moyenne de 29,25 Kg de lait contre 34,75 pour la race 1 soit une différence de 5,5 Kg.

Calculées par rapport à la moyenne générale, les différences dues à ce premier critère de classification sont:

34,75-32,00=02,75

29,25 -32,00 =-2,75

Vous remarquerez que la somme de ces deux termes est forcément nulle.De la même façon, si on considère le deuxième critère, on aura:

33,75-32,00 =1,75

30,25 -32,00 =-1,75

32,00 -32,00= 0

La somme de ces 3 termes étant également nulle.

Considérons à présent l'interaction entre le facteur (race) et le facteur (régime) .

Le tableau suivant illustre ce phénomène.

Tableau de calcul des termes de l'interaction entre les 2 facteurs

	H	**B**	**M**	**Somme**
Race 1	**0,25**	-0,50	0,25	0
Race 2	-0,25	0,50	**-0,25**	0
Somme	0	0	0	0

La première case ombrée (0,25) est obtenue ainsi:

36,75 - 34,17 - 33,75 + 32=0,25

De la même façon, la dernière case ombrée (-0,25) est obtenue ainsi:

29-29,25-32+32= -0.25

Et ainsi de suite pour les autres cases.....

Ces valeurs ainsi obtenues représentent les termes de l'inter action entre les deux facteurs étudiés.

Dans le cas présent l'interaction entre le facteur (race) et le facteur (régime) peut être considérée comme étant faible. On le confirmera plus loin par des calculs.

Imaginons à présent, des valeurs différentes pour la race 2 avec le régime B (valeurs en gras dans le tableau suivant).

	H (j=1)	**B** (j = 2)	**M** (j=3)	**M**
Race 1 (i =1)	33 35 36 43 36,50	31 32 33 34 32,50	32 34 36 38 35,00	 28,75
Race 2 (i = 2)	30 30 30 33 **30,75**	**25,5** **28,5** **24,5** **27,5** **26,5**	27 29 30 30 **29,00**	 28,75
Moyennes	33,75	**29,50**	**32,00**	31,75

Avec de telles valeurs, **tous les termes de l'interaction seraient exactement nuls.**

Exemple:

Race 1, régime H :36,75-34,75-33,75+31,75=0

Race2, régime B : 26,5-28,75-29,50+31,75=0

Etc...

Ce cas particulier traduit l'absence totale d'interaction entre les 2 facteurs.

Concrètement cela signifie que les 3 types de régimes donnent exactement la même différence entre les 2 races. Cet écart s'obtient tout simplement par différence entre les valeurs moyennes obtenues pour chaque race. Dans notre cas, ce sera:

36,75-30,75(colonne1)=**32,50 - 26,50**(colonne 2) = **35,00 - 29,00**(colonne3) = **6**

Bien évidemment, cette valeur peut être également obtenue directement par la différence entre X (soit 34,75 pour la race 1) et X (soit 28,75 pour la race 2). En effet :

34 ,75 -28,75 =6

Vous remarquez aussi que l'absence d'interaction signifie aussi que les différences entre les races sont indépendantes des régimes.

Exemples: 36,75 – 32,50=30,75-26,50=4,25 pour les régimes H et B

36 ,75 -35 ,00 = 30,75 -29,00 = 1,75 pour le régime H et M, ….. Etc.

En revanche, la présence de termes d'interaction non nuls signifie qu'il existe une (dépendance) entre les 2 facteurs étudiés.

Reprenons à présent les données du tableau et effectuons l'analyse de variance.

	Race 1 (i=1)			Race 2 (i=2)			Totaux
	H (j=1)	**B (j=2)**	**M (j=3)**	**H (j=1)**	**B (j=2)**	**M (j=3)**	
1	33	31	32	30	25	27	
2	35	32	34	30	27	29	
3	36	33	36	30	30	30	
4	43	34	38	33	30	30	
Moyenne	36,8	32,5	35,0	30,8	28,0	29,0	-
Xij	147	130	140	123	112	116	X..+768
	5459	4230	4920	3789	3154	3370	T=24922
	5402,25	4225	4900	3782,25	3136	3364	-
SCEij	56,8	5,0	20,0	6,8	18,0	6,0	SCEr=112,5

	H (j=1)	**B (j=2)**	**M (j=3)**	**Xi..**
Race 1(i=1)	147	130	140	417
Race 2(i=2)	123	112	116	351
	270	242	256	768

Terme correctif: $C = \frac{X_{...}^2}{pqn} = 24576$

Somme des carrées des écarts totale: SCEt = T – C + 346,00

Somme des carrées des écartes résiduelle: $SCEr = \sum_{i=1}^{p}\sum_{j=1}^{q} SCE_{ij} = 112,5$

$$SCEf = \frac{\left(417^2 + 351^2\right)}{12} - 24576 = 181,50$$

$$SCEb = \frac{\left(270^2 + 242^2 + 256^2\right)}{8} - 24576 = 49,00$$

$$SCEfb = 346,00 - 112,50 - 181,50 - 49,00 = 3,00$$

Tableau de l'analyse de variance

Sources de variation	ddl	SCE	CM	F_{obs}	F_{table}
Race	1	181,5	181,5	29,04***	$F_{1\ ;18\ ;0,05}$=4,41
Régime	2	49,0	24,5	3,92*	$F_{2\ ;18\ ;0,05}$=3,55
Interaction	2	3,00	1,5	0,24	$F_{2\ ;18\ ;0,05}$=3,55
Variance résiduelle	18	112,5	6,25		
Totaux	23	346,0			

Conclusion

- Il existe un effet race très important (F_{obs}>>>f_{table})
- Il existe un effet régime mais faible (F_{obs}>Ftable):
- Il n'y a pas d'interaction entre la race et le régime (F_{obs}<F_{table}).

Références Bibliographiques

LEGRAS.B., 1998. Eléments de statistique à l'usage des étudiants en médecine et en biologie. Édition marketing S.A, Paris, pp 222.

DAGNELIE.P., 1986. Analyse statistique à plusieurs variables. Gembloux, Pesses agronomiques, pp. 362.

DAGNELIE.P., 2003. Principes d'expérimentation. Planification des expériences et analyse de leurs résultats. Les Presses Agronomiques de Gembloux, Belgique, pp.397.

DAGNELIE. P., 2006. Statistique théorique et appliquée. Tomme 2: inférences à une et à deux dimensions. Bruxelles-université DE BOECK et LARCIER: pp.659.

RAKOTOMALALA R., 2015. Analyse de corrélation. Étude des dépendances - Variables quantitatives. Version 1.1. Université Lumière Lyon 2. Notes de cours. pp99.