REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECEHERCHE SCIENTIFIQUE

UNIVERSITE LARBI TEBESSI- TEBESSA

Faculté
des Sciences Exactes et des Sciences de la Nature et de la Vie

Département de Biologie Appliquée

Intitulé de la matière :

Biostatistique

Pour les étudiants de troisième année Licence (Biologie moléculaire et Biochimie)

Présentée par :

Dr. DRIS Djemaa

Sommaire

Introduction

La statistique a envahi aujourd'hui tous les champs scientifiques. Les statistiques, dans le sens populaire du terme, traitent des populations (dans des études démographiques), ce qui est très difficile dans le cas de la Bio-statistique. La statistique constitue, en science, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence.**[1]**

- Le traitement A est-il plus efficace que le traitement B ?
- Quelle sont les valeurs normales de grandeurs biologiques (taille,poids, glycémie, ...) ?
- Les modifications de poids d'un individu sont-elles liées auxmodifications de cholestérolémie ?
- Un test de dépistage est-il fiable ?

Leur objectif consiste à caractériser une population à partir d'une image plus ou moins floue constituée à l'aide d'un échantillon issu de cette population. *On peut alors chercher à extrapoler une information obtenue à partir de l'échantillon*.

On trouve des applications de la statistique dans tous les domaines : industrie, environnement, médecine, finance, marketing, sport, ... etc.

Le programme se compose de trois parties.

1) *Généralités et notions de base ;*
2) *Statistiques descriptives (à une et à deux dimensions) ;*
3) *Statistiques inférentielles (tests de comparaison).*

Objectifs de cours

- Connaître le vocabulaire particulier de la biostatistique **;**
- Comprendre les principes du traitement des données ;
- Le choix de la méthode statistique opportune à chaque situation particulière ;
- La réalisation des calculs et des tests de base pour une et deux variables.

Chapitre I : Généralités et notions de base

La Bio-statistique

Définition

Ensemble de méthodes à partir desquelles on recueille, organise, résume, présente et analyse des données afin d'en tirer des conclusions et de prendre des décisions avec prudence.**[2]**

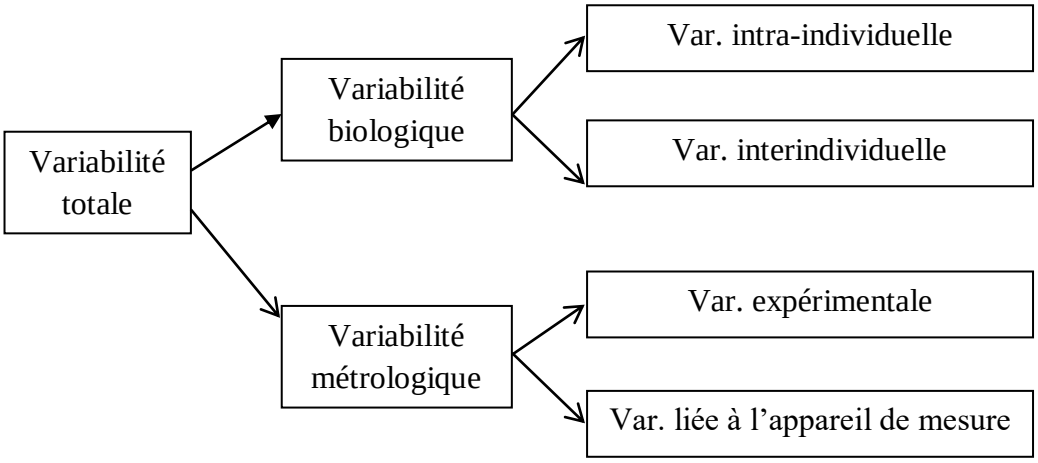
Notions importantes

Parmi les notions importantes nous avons :

La variabilité

Disposition à varié, qualité de ce qui est variable.

La variabilité en biologie est la somme d'une variabilité métrologique et d'une variabilité proprement biologique.



Population :Ensemble des individus objets de l'étude, ou Ensembles des ***éléments*** ou ***d'individus*** de même nature, visés par une problématique scientifique.

Elément : Les éléments sont les unités qui composent une population.

Synonymes : Objet, individu, unité statistique, unité d'échantillonnage, sujet, événement, comportement,

Echantillon : C'est un sous ensemble de la population considérée, prélevé pour juger de cet ensemble.

Echantillon représentatif : Échantillon qui reflète fidèlement la complexité et la composition de la population. *Le tirage au sort* ainsi que *l'inventaire exhaustif* (recensement), sont deux façons d'obtenir un échantillon représentatif d'une population.**[2]**

Caractère statistique (ou variable statistique)

C'est ce qui est observé ou mesuré sur les individus d'une population statistique.Cette variable peut être quantitative (numérique) ou qualitative (non numérique).

Variable quantitative :C'est un paramètre expérimental qui s'exprime par un nombre.

Pouvant être classées en ***variables continue*** (taille, poids) ou ***discontinue*** **(discrète)** (nombre d'enfantsdans une famille, nombre d'œufs pondus par un oiseau).

Variable qualitative : Une variable qualitative se caractérise par un ensemble discontinu d'états.

Pouvant être classées en ***variables catégorielles***(nominales) (couleurs des plumes des oiseaux) ou ***ordinales*** (résistance d'une plante vis-à-vis un ravageur classée en faible, moyenne, importante).**[1]**

Notion d'hypothèse

L'hypothèse est une relation hypothétique (provisoire, postulée par le chercheur).

On distingue deux formes d'hypothèses :

Hypothèse nulle (H0) et *Hypothèse significative (H1)* ou *alternative*.

- **Hypothèse nulle (H0)**: $m_1 = m_2$ou l'absence d'une différence significative entre les moyennes ;
- **Hypothèse alternative (H1)**: $m_1 \neq m_2$ ou l'existence d'une différence significative entre les moyennes.

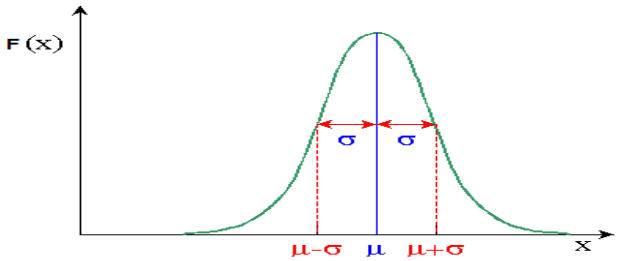
Seuil de signification

En statistique, il n'existe pas de règle rigide permettant de tirer une conclusion concernant les hypothèses ; aucun test ne nous fournit une réponse en terme de oui ou non, mais indique dans quelle mesure nous pouvons être certain de tirer des conclusions ; cette mesure se nomme niveau ou seuil de signification, ou encore probabilité d'erreur. **[3]**

La loi normale

Une distribution normale correspond à la distribution de probabilités d'une variable aléatoire continue dont la courbe est parfaitement symétrique et en forme de cloche.

Lorsqu'une variable (x) se distribue de telle sorte que les fréquences de ses différentes éventualités suivent la loi normale, alors elle est dite variable normale.



Types de test

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 (moyenne, mode et médiane). La distribution des données suit la loi normale.

Les ***tests non paramétriques***ne font aucune hypothèse sur la distribution sous-jacente des données (la distribution des données ne suit pas la loi normale). On les qualifie souvent de tests *distribution free*.**[3]**

Chapitre II : Statistiques descriptives

Les méthodes statistiques peuvent être classées en deux groupes:

1) **Les Statistiques descriptives :** Elle regroupeles méthodesdontl'objectif principal est ladescriptiondesdonnées étudiées.Cettedescriptiondesdonnéessefaitàtraversleurreprésentation graphique,etlecalculderésumésnumériques.Danscetteoptique,onnefaitpas appelàdesoutilsde typeprobabiliste.

   Oncitetrois typesdesstatistiquesdescriptives:**Statistique descriptive uni-variée**:étudedela populationselonuneseule variable.**Statistique descriptive bi-variée**:étudedescorrélationsetrelationséventuellesentredeuxvariablesdelamême population.**Statistique descriptive multi-variée**:étudedesrelationséventuellesentreplusieursvariablesdelamême population.**[2]**

2) **La statistique inférentielle** : Ceterme regroupeles méthodesdontl'objectifprincipalestdepréciserunphénomènesurune populationglobale,àpartirdesonobservationsurunéchantillondecette population.Cepassagenesefaitquemoyennantdes hypothèsesde typeprobabiliste.**[1]**

II.1. Statistiques descriptives à une dimension

Paramètres de position	Paramètres de dispersion
La moyenne m ou$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} nixi$	**L'étendue**E= x_{max} - x_{min}
Le mode (Mo) : C'est la valeur ou classe correspondant à l'effectif (ou fréquence) le plus élevé.	**La variance** S^2_X= S^2_X= 1/n $\sum f\,(x_i - x)^2$
La médiane (Me) : valeur centrale de la série statistique.	**L'écart-type** S_x = Racine carrée de la variance
Les quartiles Les quartiles partagent la série en quatre groupes. Le premier quartile : C'est la plus petite donnée de la liste telle qu'au moins un quart des données de la liste sont inférieures ou égales à Q_1 = N/4 Le troisième quartile : C'est la plus petite donnée de la liste telle qu'au moins les trois quarts des données de la liste sont inférieures ou égales à Q_3 = N× ¾	**Le coefficient de variation C.V. =** S/m ×100. 1) CV < 5% : Les valeurs sont très homogènes. 2) 5%<CV< 10% : Les valeurs sont homogènes. 3) 10%<CV< 15% : Les valeurs sont moyennement homogènes. 4) 15%<CV< 30% : Les valeurs sont hétérogènes. 5) CV> 30% : Les valeurs sont très hétérogènes.

Exercices de révisions[4]

Exercice1: On donne les couleurs de n=15 plantes.**VVRNRRVRRRJNNNN**

1. De quel type est la variable couleur des plantes?
2. Construire le tableau statistique et en déduire le mode.
3. Déterminer le mode.
4. Construire le diagramme en secteurs.

Solution :

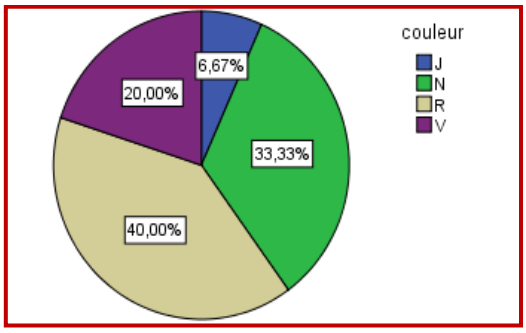
1- La variable couleur des plantes est une variable qualitative nominale.

2- Le tableau statistique :

Xi	**V**	**R**	**N**	**J**
ni	3	6	5	1
fi (ni/N)	0,2	0,4	0,33	0,07

3- Le mode est : **R**

4-Le diagramme en secteurs:



Exercice 2 :

Trenteéprouvettesd'acier spécialontsoumisesàdesessaisderésistance.Pourchacune,onnotelenombrede chocsnécessairespourobtenirlarupture.Lesrésultatsobtenussontlessuivants :

2	2	3	1	2	1	4	2	3	2
3	2	3	3	4	1	1	4	2	3
2	3	2	2	3	4	3	2	3	2

1. Dequel typeestcettevariable?
2. Construireletableaustatistiqueetendéduirele mode.
3. Construirelediagrammeenbâtonnetsdes effectifs.
4. Déterminerlamédiane,lamoyenne,lavarianceetl'écarttype decettevariable.

5. Déterminerlafonctionderépartitionettracersa courbe.

Solution :

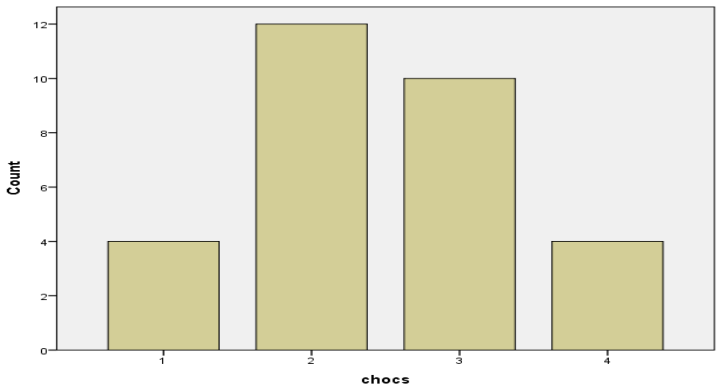
1- Le nombre de chocs nécessaires pour obtenir la rupture est une variable quantitative discrète.

2- Le tableau statistique :

Xi	**1**	**2**	**3**	**4**
ni	4	12	10	4

Le mode est : 2

3- Le diagramme en bâtonnets des effectifs:



4- La médiane (Me): N/2=30/2=15, la quinzième valeur est 2 donc Me=2

La moyenne=1/30(1*4+2*12+3*10+4*4)=2,47.

La variance=0,809 et l'écart type=0,9.

Exercice 3 : On

pèselesn=50élèvesd'uneclasseetnousobtenonslesrésultatsrésumésdansletableausuivant:

43	43	43	47	48	48	48	48	49	49
49	50	50	51	51	52	53	53	53	54
54	56	56	56	57	59	59	59	62	62
63	63	65	65	67	67	68	70	70	70
72	72	73	77	77	81	83	86	92	93

1. Dequel typeestlavariable poids?
2. Construireletableaustatistiqueenadoptantquatreclasses.
3. Déterminerlafonctionderépartitionettracersa courbe.
4. Déterminerlamoyenne,lavarianceetl'écart typedelavariable poids

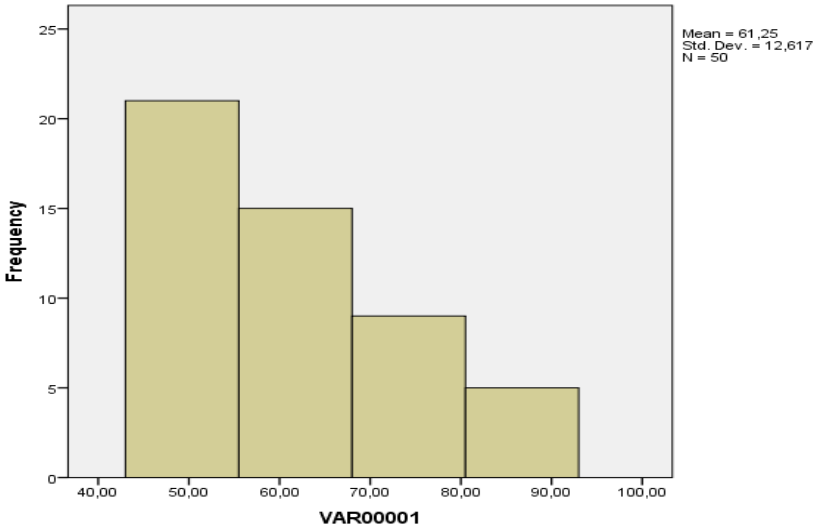
Solution :

1- Le poids est une variable quantitative continue.

2- Le tableau statistique :

Xi Classe	[43-55,5[	[55,5-68[	[68-80,5[	[80,5-93]
ni	21	15	9	5
Ci	49,25	61,75	74,25	86,75

3- La fonction de répartition est l'histogramme.



4- La moyenne:61,25 ; l'écart-type=12,61 et la variance=159,18

II.2. Statistiques descriptives à deux dimensions

La statistique descriptive à deux dimensions a essentiellement pour but de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors ***liaison*.** Dans certains cas, cette liaison peut être considérée *a priori* comme *causale*, une variable *X* expliquant l'autre *Y* ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité. Sont ainsi introduites les notions de covariance, coefficient de corrélation linéaire, régression linéaire, rapport de corrélation, indice de concentration, khi-deux et autres indicateurs qui lui sont liés. De même, nous présentons les graphiques illustrant les liaisons entre variables : nuage de points (*scatterplot*), diagrammes-boîtes parallèles, diagramme de profils.

Lasériestatistiqueestalorsunesuitedencouplesdesvaleursprisesparlesdeuxvariablessurchaqueindividu:(x1,y1),.....,(xn,yn). L'effectif associeàl'observation(xi,yj)estnoténij.Etsafréquencenotée:fij=nijn.Lesrésultatssont regroupésdansuntableau appelétableaudecontingence.**[3]**

II.2.1. Deux variables qualitatives

Tableau de contingence des effectifs

Ons'intéresseàuneéventuellerelationentreX:lesexeden=200 personnesetY:lacouleurdesyeux.

X/Y	Bleu	Vert	Marron	Total
Homme	n11=10	n12=50	n13=20	n1•=80
Femme	n21=20	n22=60	n23=40	n2•=120
Total	n•1=30	n•2=110	n•3=60	n=200

n1•,n2•etn•1,n•2,n•3sont appelés effectifsmarginaux.

n11+n12+n13=n1•,

n21+n22+n23=n2•,

n11+n21=n•1,

n12+n22=n•2,

n13+n23=n•3,

n11+n12+n13+n21+n22+n23=n.

Tableau de contingence desfréquences

X/Y	Bleu	Vert	Marron	Total
Homme	f11=0,05	f12=0,25	f13=0,10	f1•=0,40
Femme	f21=0,10	f22=0,30	f23=0,20	f2•=0,60
Total	f•1=0,15	f•2=0,55	f•3=0,30	1

f1•,f2•etf•1,f•2,f•3sont appeléesfréquencesmarginales.

fij=nij/n,fi•=ni•/n,f•j=n•j/n

f11+f12+f13=f1•,

f21+f22+f23=f2•,

f11+f21=f•1,

f12+f22=f•2,

f13+f23=f•3,

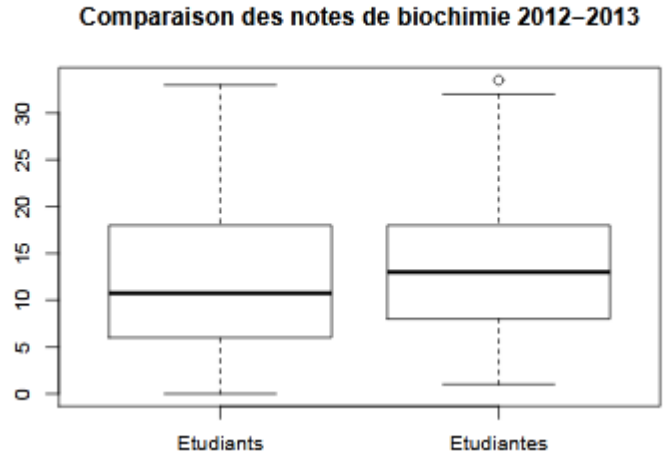
f11+f12+f13+f21+f22+f23=1.**[5]**

II.2.2. Une variable quantitative et une variable qualitative

Diagramme à boîtes à moustaches :

La boîte,(verticaleoubienhorizontale),estlapartiedugraphiquecompriseentrelespremierettroisièmequartiles,(cesquartilesséparentla populationenquatrepartieségauxen effectifs).Lamédianeestsituéeàl'intérieurdela boîteetreprésentéeparuntraithorizontal.Danslespartiesbasseethautedugraphique figurentlesmoustaches,joignantleminimumaupremierquartileetletroisièmequartileaumaximum. **[5]**

Exemple : X= notes des étudiants (quantitative) et Y= sexe (qualitative)



II.2.3. Deux variables quantitatives

La corrélation

La corrélation est la netteté ou l'intensité de la relation existante entre deux séries de données.

II.2.3.1. Notion de covariance

Nous notons parCov(X,Y)la covariance entre les variablesXetY. La covariance estun paramètre qui donne la variabilité deXpar rapport àY. La covariance se calcule par l'expression suivante :$Cov(X,Y) = \overline{xy} - \bar{x}\bar{y} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{J} n_{ij}x_i y_j - \bar{x}\bar{y}$**[6]**

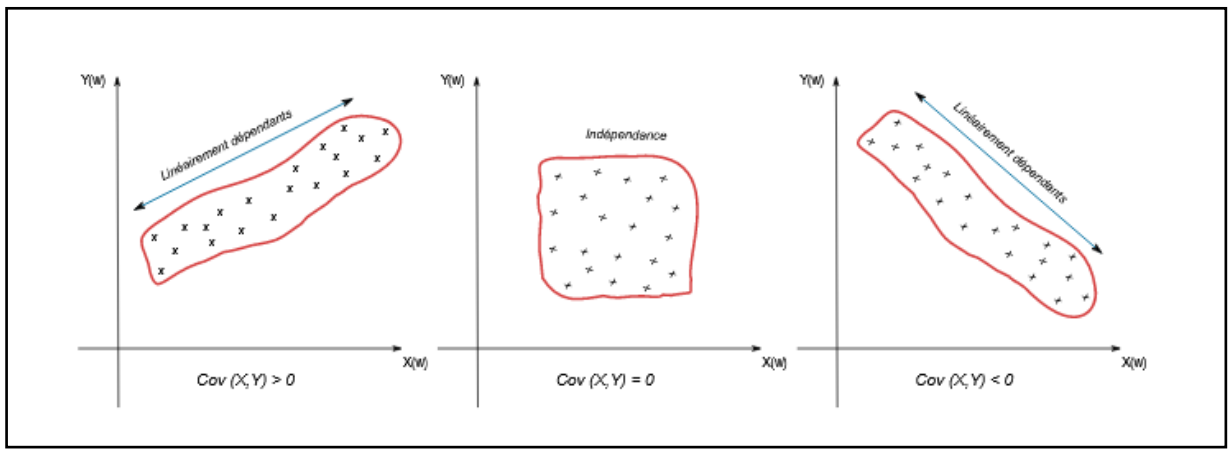


Figure : La covariance et la variabilité

II.2.3.2. Le coefficient de corrélation

Pour des variables quantitatives, choisissez le coefficient de corrélation de Pearson ou Spearman.

Coefficient de corrélation de Pearson (r) : pour les données avec une distribution normale.

Rho de Spearman (ρ) : pour les données qui ne suit pas la loi normale.**[7]**

Propriétés

- Si le coefficient de corrélation est positif, les points du nuage sont alignés le long d'une droite croissante. Dans ce cas X et Y évoluent dans le même sens (**figure 1**).
- Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante. Dans ce cas X et Y évoluent dans des sens opposés (**figure 1**).
- Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire (**figure 1**).**[7]**

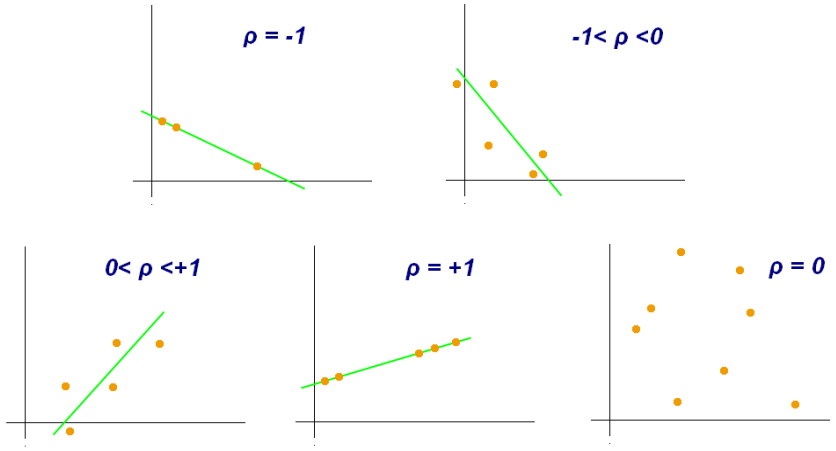


Figure 1: Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation.

II.2.3.3. Droite de régression

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite. Cette méthode vise à expliquer un nuage de points par une droite qui lie **y** à **x**(**figure 2**).

y = a.**x** + b avec a=cov(X, Y)/Var(X) ; $b = \bar{y} - a\bar{x}$**[7]**

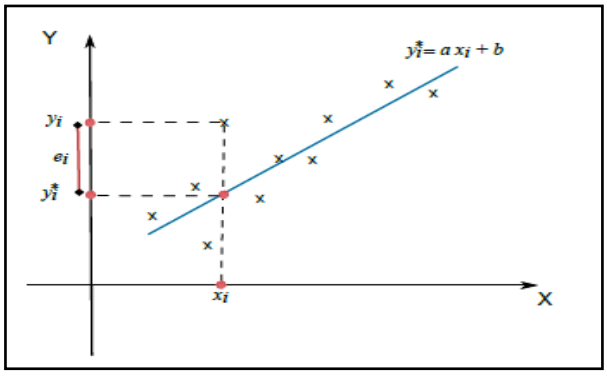


Figure 2: La droite la plus proche possible de chacun des points.

Ajustement linéaire

L'ajustement linéaire consiste à remplacer le nuage de points par une droite à l'aide d'une équation de la régression.

Remarque

Le coefficient de corrélation permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir **figure 3**).

Si | r |< 0,7 ; alors l'ajustement linéaire est refusé (droite refusée).

Si | r | ≥ 0,7 ; alors l'ajustement linéaire est accepté (droite acceptée).**[5]**

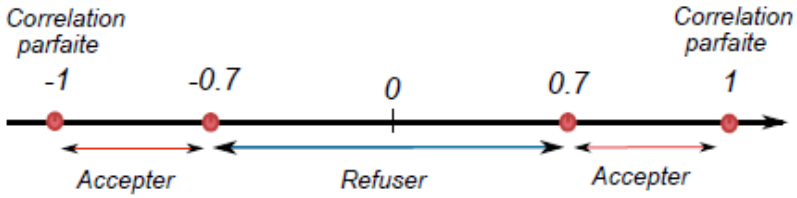


Figure 3: La zone d'acceptation ou de refus de l'ajustement linéaire.

Exemple d'application : La fécondité du poisson *Scorpaenichtysmarmoratus*s'avère être un paramètre fastidieux à définir. Afin de simplifier une étude sur la dynamique de population de cette espèce, le nombre y d'œufs (en milliers) présent chez 11 femelles matures a été compté en relation avec leur poids (kg).

Poids X	Nbre d'œufs Y
14	61
17	37
24	65
25	69
27	54
33	93
34	87
37	89
40	100
41	90
42	97

1. Calculer le coefficient de corrélation r, quelle conclusion en tirez-vous ?
2. Déterminer l'équation de la droite de régression y en x.
3. Combien d'œufs pondue prévoyez-vous pour une femelle pèse 50 kg ?

Solution:

1- La moyenne de X : $\bar{X} = 30{,}36$; Sx= 9,657 et la moyenne de Y :$\bar{Y} = 76{,}55$; Sy=20,466.

$$\overline{XY} = \frac{1}{n}\sum\sum nijxiyj = \frac{1}{11}(14 * 61 + 17 * 37 + 24 * 65 \dots \dots \dots \dots + 42 * 97)$$
$$= \frac{1}{11}(854 + 624 + 1560 + 1725 + 1458 + 3069 + 2958 + 3293 + 4000 + 3690 + 4074) = 2482{,}27$$

La covariance : $Cov(X,Y) = \overline{xy} - \bar{x}\bar{y} = 2482{,}27 - (30{,}36 * 76{,}55) = 158{,}212$

$$r = \frac{Cov(X,Y)}{\delta_x \delta_y} = \frac{158{,}212}{9{,}657 \times 20{,}466} = 0{,}80$$

Il existe une relation positive et forte entre le poids et le nombre d'œufs

2- L'équation de la droite de régression :

r=0,8 : alors l'ajustement linéaire est accepté (droite acceptée).

y = a.**x** + b avec a=cov(X, Y)/Var(X)=158,212/(9,657)^2=1,696 ;

$$b = \bar{y} - a\bar{x} = 76{,}55 - 1{,}696(30{,}36) = 25{,}059$$

L'équation est : Y=1,696*X+25,059

3- Pour X=50, Y=1,696*50+25,059=110 œufs

Chapitre III :Échantillonnage et estimation

DEFINITIONS

On considère une population sur laquelle on dispose d'informations concernant un paramètre relatif à un certain caractère. L'échantillonnage consiste à passer de la population totale à un échantillon provenant de cette population. C'est à dire à déduire, à partir des informations sur la population, des informations concernant le paramètre sur l'échantillon.

On considère, cette fois, un échantillon sur lequel on dispose d'informations concernant un paramètre relatif à un certain caractère. L'estimation consiste à passer de l'échantillon à la population, c'est à dire à induire, à partir des résultats observes sur l'échantillon, des résultats concernant la population.**[7]**

III. 1. L'échantillonnage

En statistiques, les méthodes d'échantillonnage correspondent aux différentes manières de constituer un échantillon de la population étudiée.

Si l'échantillon n'est pas constitué de manière aléatoire, il ne peut pas être représentatif de la population c'est-à-dire ne pas posséder les mêmes caractéristiques que la population que l'on souhaite étudier. Les résultats obtenus sur l'échantillon ne peuvent alors être extrapolés à la population. L'étude est dite biaisée et non valide (le biais désigne une erreur systématique dans l'estimation d'un paramètre).

Il existe différentes méthodes d'échantillonnage, aléatoires ou non : Lorsqu'on souhaite effectuer un sondage ou une enquête, il n'est pas toujours possible d'interroger chaque membre de la population de par des contraintes géographiques, monétaires ou temporelles. Par contre, il est tout

demême possible d'en apprendre plus à propos de la population visée notamment en analysant un échantillon. Pour ce faire, il est primordial de choisir la bonne méthode de construction d'un tel échantillon.**[6]**

1. **L'échantillonnage aléatoire simple**

Chaque personne ou objet de la population a la même probabilité de faire partie de l'échantillon puisqu'ils sont tous pigés au hasard.De façon plus générale, cette méthode présente un avantage et un inconvénient majeurs.

Avantage

- De par les différentes lois en probabilité, cet échantillon sera représentatif de la population.

Inconvénient

- Il faut avoir la liste complète de la population pour ensuite faire le tirage au sort.**[3]**

2. **L'échantillonnage systématique**

Chaque élément qui compose l'échantillon est choisi de façon régulière, selon un intervalle régulier, à l'intérieur de la population ciblée.Tout comme la méthode précédente, on peut dégager les principaux avantages et inconvénients d'une telle méthode de sélection.

Avantages

- On peut facilement prédéterminer la taille et les éléments faisant partie de l'échantillon.
- L'échantillon est distribué dans des proportions égales dans la population.

Inconvénient

De par sa caractéristique d'intervalles réguliers pour choisir les éléments, cela ne garantit pas un échantillon représentatif.**[3]**

3. **L'échantillonnage par grappes**

En se basant sur la position géographique de la population ciblée, on la divise d'abord en grappes (sous-groupes de la population) pour ensuite en sélectionner un certain nombre de façon aléatoire afin de former l'échantillon.Malgré son application à l'air plutôt simpliste, il n'en demeure pas moins que cette méthode possède des bons et des mauvais côtés.

Avantages

- Il n'est pas nécessaire d'avoir une liste officielle de tous les membres de la population ciblée.
- Idéal pour sonder une population qui est géographiquement étendue.

Inconvénients

- Généralement, les éléments d'une même grappe possèdent des caractéristiques semblables sans nécessairement être celles de la population ciblée.
- Il est très difficile de prédire la taille de l'échantillon étant donné que les grappes n'ont pas toutes la même quantité d'individus.**[3]**

4. L'échantillonnage stratifié

En se basant sur une caractéristique de la population ciblée, on la divise d'abord en strates (sous-groupes de la population) pour ensuite sélectionner de façon aléatoire des membres de chacune des strates en respectant leur proportionnalité dans la population.

Avantage

- Cette méthode assure une assez bonne représentativité de la population due à son critère de proportionnalité.

Inconvénient

- Il faut avoir une bonne connaissance de la population afin d'établir les strates avec lesquelles il faudra travailler.**[3]**

III. 2. Estimation de la moyenne

1. Principe de l'estimation

La théorie de l'échantillonnage consistait à déterminer des propriétés sur des échantillons tirés au hasard parmi une population dont on connait les propriétés.Le principe de l'estimation est de faire exactement l'inverse, c'est-à-dire que l'on accès à des informations

sur des échantillons (sondages, tests de confomité,...) et l'on souhaite déterminer certaines propriétés sur la population entière. **[2]**

Remarque 1. Il est clair que l'on ne pourra jamais obtenir, à partir d'un échantillonréduit, des données exactes sur la population entière, c'est pourquoi il sera importantdans la suite de donner des estimations de certaines données mais en précisanttoujours la marge d'erreur ou le risque que l'on prend.**[1]**

2. Type d'estimation

L'estimation peut être :

- Une estimation ponctuelle
- Une estimation par intervalle de confiance

2.1.<u>Estimation ponctuelle</u> : Une estimation ponctuelle de la moyenne **μ** est la réalisation d'une moyenne d'un échantillon de taille **n** tiré aléatoirement d'une population **P.**

$$m = \overline{x_e}\ est\ un\ estimateur\ de\ \mu$$

2.2.<u>Estimation par intervalle de confiance</u>

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible dans l'estimation due aux fluctuations d'échantillonnage.

La théorie des intervalles de confiance (IC) consiste à construire, autour de l'estimation ponctuelle, un intervalle qui aura une grande probabilité $(1-\alpha)$ de contenir la vraie valeur du paramètre.

3. Distribution d'échantillonnage et intervalle de confiance d'une moyenne

3.1. Cas des grands échantillons ($n \geq 30$)

A. Distribution d'échantillonnage d'une moyenne

On considère une population de moyenne μ et d'écart-type δ_p relatif à un caractère quantitatif. Si on prélève au hasard k échantillons de même taille n par exemple, on constate que les moyennes m_1, m_2, …….m_k de ces k échantillons font apparaitre des différences, parfois importantes, dues aux fluctuations d'échantillonnage.

On désigne par $\bar{X}$, la variable aléatoire qui peut prendre pour valeur la moyenne d'un échantillon prélevé au hasard de la population $\bar{X}$ est appelée moyenne d'échantillonnage.

On détermine la loi de probabilité de $\bar{X}$appelée distribution d'échantillonnage de la moyenne.

On démontre que $\bar{X}$ suit une loi normale de moyenne μ et de variance $\frac{\delta_p^2}{n}$lorsque la taille des échantillons n ≥ 30

$$\text{Donc } \mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}$$

$$\text{Avec} \quad P\left(\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}\right) = 1 - \alpha$$

La probabilité pour que $\bar{X}$soit dans l'intervalle$\left[\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} , \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}\right]$, cette intervalle est appelée intervalle fluctuation de la moyenne.

* $1 - \alpha$ est appelé seuil de confiance.

* α est appelé risque d'erreur.

$t_\alpha$est une valeur donnée par la table de la loi normale centrée réduite

On générale, on choisit $\alpha = 5\%$ et dans certains cas assez particuliers $\alpha = 1\%$

D'après les propriétés de la loi normale on a :

Pour $\alpha = 5\%$, $t_\alpha = 1.96$ $\quad P\left(\mu - 1.96\frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + 1.96\frac{\delta_p}{\sqrt{n}}\right) = 0.95$

Pour $\alpha = 1\%$, $t_\alpha = 2.60$ $\quad P\left(\mu - 2.60\frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + 2.60\frac{\delta_p}{\sqrt{n}}\right) = 0.99$**[6]**

Exemple

Une machine est destinée à fabriquer des comprimés de poids moyen de 200mg avec un écart-type de 10mg. On extrait au hasard un échantillon de 50 comprimés.

Entre quelles limites varie le poids moyen des comprimés de cette échantillon au risque de 5%

Solution

Population : $\mu = 200\text{mg}$, $\delta_p = 10mg$

Echantillon : $n = 50 > 30$

Le poids moyen d'un échantillon varie d'un échantillon à un autre, c'est donc une variable que l'on désigne par $\bar{X}$ (la moyenne d'échantillonnage)

Comme n> 30 alors :

$$\text{I de F}\left[\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} , \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}\right]$$

Au risque $\alpha = 5\%$, $t_\alpha = 1.96$

$$\text{IP} = [197.22\,, 202.77] \quad \text{au risque } \alpha = 5\%$$

Le poids moyen d'un échantillon de 50 comprimés est compris entre 197.22 et 202.77 avec un risque de 5% de se tromper (d'erreur)

B. Intervalle de confiance d'une moyenne

Soit à étudier dans une population un certain caractère quantitatif. Désignons par μ et δ écart-type du caractère étudié (μ et δ sont inconnus). On prélève au hasard un échantillon de taille n et on détermine la moyenne $\bar{X}$ et l'écart-type S. le problème qui se pose est d'estimer la moyenne μ de la population a partir de n, $\bar{X}$ et S, c'est à dire de trouver un intervalle dans lequel se trouve la moyenne de la populationμ.

D'après ce qui précède $\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}$

Puisque la moyenne $\bar{X}$ de l'échantillon est connue alors $\bar{X}$ est la valeur prise par la moyenne d'échantillonnage $\bar{X}\bar{X} - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \mu \leq \bar{X} + t_\alpha \frac{\delta_p}{\sqrt{n}}$

$\delta_p$est inconnu et on démontre que, lorsque n ≥ 30, la variance de la population δ_p^2 est estimée par $\frac{n}{n-1}\delta_e^2$ en d'autre terme, lorsque n ≥ 30, on a : $\delta_p^2 \approx \frac{n}{n-1}\delta_e^2$

$$\text{Donc} \qquad \bar{X} - t_\alpha \frac{\delta_e}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_\alpha \frac{\delta_e}{\sqrt{n-1}}$$

$$P\left(\bar{X} - t_\alpha \frac{\delta_e}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_\alpha \frac{\delta_e}{\sqrt{n-1}}\right) = 1 - \alpha$$

C'est la probabilité pour que l'intervalle : $\left[\bar{X} - t_\alpha \frac{\delta_e}{\sqrt{n-1}}\,, \bar{X} + t_\alpha \frac{\delta_e}{\sqrt{n-1}}\right]$

Cette intervalle est appelé l'intervalle de confiance de la moyenne μ

On choisit généralement $\alpha = 5\%$ ou $\alpha = 1\%$

Pour $\alpha = 5\%$, $t_\alpha = 1.96$, Pour $\alpha = 1\%$, $t_\alpha = 2.60$

D'une manière générale, on écrit : $\text{IC}(\mu) = \left[\bar{X} - t_\alpha \frac{\delta_e}{\sqrt{n-1}}\,, \bar{X} + t_\alpha \frac{\delta_e}{\sqrt{n-1}}\right]$

Au risque α (ou au seuil de confiance $1 - \alpha$)**[6]**

Exemple :

Dans une population de personnes, on extrait au hasard un échantillon de taille 40 dont le poids moyen est de 70 kg et l'écart-type de 15.4 kg. Quel est au risque de 5%, l'intervalle de confiance du poids moyen de la population.

Solution :

Echantillon : $n = 40 > 30$, $\bar{X} = 70$, $\delta_e = 15.4$

Désignons par μ le poids moyen de la population à estimer. L'intervalle de confiance de μ est donc : $\text{IC}(\mu) = \left[\bar{X} - t_\alpha \frac{\delta_e}{\sqrt{n-1}}\,, \bar{X} + t_\alpha \frac{\delta_e}{\sqrt{n-1}}\right]$

Au risque α donné , pour $\alpha = 5\%$, $t_\alpha = 1.96$

IC(μ) $= [65.16 \ , \ 74.83]$ au risque $\alpha = 5\%$

Ceci veut dire qu'il y a 95% de chance pour que l'intervalle de confiance $[65.16 \ , \ 74.83]$ contienne le poids moyen μ de la population.

C. Précision de l'estimation

Il convient de remarquer que la précision de l'estimation est d'autant meilleure que la taille de l'échantillon est assez grande car la longueur de l'intervalle de confiance diminue quand n croit.

On a : $\mu = \bar{X} \pm t_\alpha \frac{\delta_e}{\sqrt{n-1}}$

La précision de l'estimation est donc : $h = t_\alpha \frac{\delta_e}{\sqrt{n-1}}$/ h pour un risque α donné

Dans l'exemple précédent, la précision de l'estimation du poids moyen de la population est : $h = 1.96 \frac{15.4}{\sqrt{39}} = 4.83$

D'autre part, si on diminue le risque α (donc t_α augmente), la longueur de l'intervalle de confiance augmente, par conséquent on perd la précision de l'estimation. **[7]**

3.2. Cas des petits échantillons (n < 30)

A. Distribution d'échantillonnage d'une moyenne

Comme dans le cas des grands échantillons on a : $\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}$

Avec : $P\ (\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}) = 1 - \alpha$

C'est la probabilité pour que la variable $\bar{X}$ soit dans l'intervalle $\left[\mu - t_\alpha \frac{\delta_p}{\sqrt{n}} \ , \mu + t_\alpha \frac{\delta_p}{\sqrt{n}}\right]$ **[6]**

B. Intervalle de confiance d'une moyenne

La différence avec le cas des grands échantillons commence au moment où on remplace la variance δ_p^2 par son estimation $\frac{n}{n-1}\delta_e^2$ obtenue d'après l'échantillon observé. Cette façon de faire n'est acceptable que dans le cas des grands échantillons, mais elle ni l'est plus dans le cas des petits échantillons.

Suit une loi de probabilité peu différente de la loi normale appelée loi de Student-Fisher qui dépend de la taille de l'échantillon n. la loi de Student-Fisher est une loi dont la courbe de densité de probabilité est plus aplatie que celle de la loi normale.

En fonction du risque α et le nombre de ddl $v = n - 1$ et comme $\bar{X}$ est connue alors :

$$\bar{X} - t_\alpha^* \frac{\delta_e}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_\alpha^* \frac{\delta_e}{\sqrt{n-1}}$$

Avec : $P\ (\bar{X} - t_\alpha^* \frac{\delta_e}{\sqrt{n-1}} \leq \mu \leq \bar{X} + t_\alpha^* \frac{\delta_e}{\sqrt{n-1}}) = 1 - \alpha$

C'est la probabilité pour que l'intervalle :

$\left[\bar{X} - t_\alpha^* \frac{\delta_e}{\sqrt{n-1}} \ , \ \bar{X} + t_\alpha^* \frac{\delta_e}{\sqrt{n-1}}\right]$contienne la moyenne μ de la population.

Cette intervalle appelé intervalle de confiance de la moyenne de la population.

On choisit généralement $\alpha = 5\%$ ou, dans certains cas assez particuliers $\alpha = 1\%$ et d'une manière générale on écrit $IC(\mu) = \left[\bar{X} - t_\alpha^* \frac{\delta_e}{\sqrt{n-1}} \ , \ \bar{X} + t_\alpha^* \frac{\delta_e}{\sqrt{n-1}}\right]$ au risque α (ou au seuil de confiance $1 - \alpha$).**[6]**

Exemple :

Un dosage de sucre dans une solution effectué sur 8 prélèvements provenant d'une même population a donné les résultats suivants exprimés en g/l

19.5, 19.7, 19.8, 20.2, 20.3, 20.4, 20.4, 20.8

01- Calculer la moyenne et l'écart-type de cette distribution.

02- Quel est l'intervalle de confiance de la moyenne au risque de 5%

Solution:

01- Calcule de la moyenne et de l'écart-type

$$\bar{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 20.11 \delta_e = \sqrt{\frac{1}{8}\sum_{i=}^{8}(xi - \bar{x})^2} = 0.395$$

02- Désignons par μ le dosage moyen du sucre de la population à estimer. En supposant que le dosage du sucre est distribué dans la population selon une loi normale, l'intervalle de confiance de la moyenne μ est donc : $IC(\mu) = \left[\bar{X} - t_\alpha^* \frac{\delta_e}{\sqrt{n-1}} \ , \ \bar{X} + t_\alpha^* \frac{\delta_e}{\sqrt{n-1}}\right]$

Au risque α et le nombre de ddl $v = 8 - 1 = 7$, la table de Student-Fisher nous donne $t_\alpha^* = 2.365$

$IC(\mu) = [19.75 \ , \ 20.46]$, au risque $\alpha = 5\%$

<u>Applications numériques</u>

A. Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir 1625mg de bicarbonate de sodium. Afin de contrôler la fabrication de ces médicaments, on a prélevé un échantillon de 150 comprimés et on a mesuré la quantité de bicarbonate de sodium pour chacun d'eux:

Classes	[1610 ; 1615[	[1615 ; 1620[	[1620 ; 1625[	[1625 ; 1630[	[1630 ; 1635[
Effectifs	7	8	42	75	18

1)- En convenant que les valeurs mesurées sont regroupées au centre de chaque classe, donner une estimation ponctuelle de la moyenne et de la variance de la quantité de bicarbonate de sodium dans la population formée de l'ensemble de tous les comprimés fabriqués et supposée très grande.

2)-Déterminer un intervalle de confiance pour la moyenne à 95% de la quantité moyenne de bicarbonate de sodium dans la population.

B. On admet que le taux de cholestérol chez une femme suit une loi normale. Sur un échantillon de10 femmes, on a obtenu les taux de cholestérol (eng/l) suivants:

3	1.8	2.1	2.7	1.4	1.9	2.2	2.5	1.7	2

1)-Déterminer une estimation ponctuelle de la moyenne et de la variance du taux.

2)-Déterminer un intervalle de confiance pour la moyenne du taux au seuil 5%.

Chapitre IV : Tests de comparaison

Introduction :

La majorité des tests repose sur le principe suivant : On définit une hypothèse nulle notée H0 contre l'hypothèse alternative H1. Letesta pourobjectifd'accepterouderejeterH0avecunrisqueconnuàpartirdesdonnéesdonton dispose.

- Ondéterminealorsunestatistiquequiestunevariablealéatoireconstruiteàpartirdesdonnées.
- Sous H0, cette variable suit une loi de probabilité connue (normale, student, khi-deux,...)
- On détermine alors l'intervalle de confiance dont le quel doit tomber la statistique avec une probabilité donnée (1−α), (le plus souvent 95% pour α=5%).
- On définitalorslarèglededécisionsuivante:
  - Silastatistique tombedansl'intervalle,onacceptеH0.Attention,celaneveutpasdirequeH0estvraie maisqueletestetlesdonnéesne permettentpasdevoirunécart significatifàH0.
  - Silastatistiquene tombepasdansl'intervalle,onrejetteH0aveclerisqueαdese tromper,(parexempleα=5%).**[1]**

IV.1. Comparaison de deux moyennes

1) **Test t ou test Student**

Il existe plusieurs formes du test-t de Student:

• Le test-t de Student pour échantillon unique.

• Le test-t de Student comparant deux groupes d'échantillons indépendants (on parle de test de Student non apparié).

• Le test-t de Student comparant deux groupes d'échantillons dépendants (on parle de test de Student apparié).

Ces différents tests peuvent être utilisés seulement sous certaines conditions :

➢Dans le cas du test de Student pour un échantillon unique:

✓Si les données suivent la loi normale.

➢Dans le cas du test de Student indépendant:

✓Si les deux groupes d'échantillons (x et y), à comparer, suivent une loi normale.

✓Si les variances des deux groupes sont égales ou pas.

➢Pour le test de student apparié:

✓Si la différence d (= x-y) suit une loi normale.

2) **Comment tester la normalité des données ?**

La normalité peut être vérifiée par une inspection visuelle [Histogramme] ou par destests de significativité.

•L'histogrammepermet un jugement visuel à savoir si la distribution est une courbe en cloche (courbe de Gauss).

•Letest de significativitécompare la distribution d'un échantillon donné à celle de la loi normale et renvoie une p-value.Plusieurs méthodes existent pour letest de normalité, notamment le test deKolmogorov-Smirnov(K-S)et le test deShapiro-Wilk.**[7]**

a. Comment tester l'égalité des variances ?

Le test de Student indépendant classique suppose l'homogénéité des variancesdes deux groupes à comparer. Si les deux échantillons suivent une loi normale, le test F peut être utilisé pour comparer les variances.**[7]**

b. Que faire lorsque les conditions d'application du test de Student ne sont pas remplies ?

La procédure suivante, à deux étapes, est largement acceptée:

➢Si la normalité est acceptée,le test de Student est utilisé.

➢Si les échantillons à comparer ne sont pas distribués selon une loi normale, un test non-paramétrique tel que le test de Wilcoxon est recommandé comme une alternative au test de Student.

➢Si les deux groupes d'échantillons suivent une loi normale, mais de variances inégales, letest t de Welchpeut être utilisé.[**7**]

1.1) Test t pour échantillon unique (test de conformité)

But: Les tests de conformité sont destinés 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.

Il s'agit de comparer une moyenne observée à une moyenne théorique (μ).

Soit X une série de valeurs de taille n, de moyenne m et d'écart-type S. La comparaison de la moyenne observée (m) à une valeur théorique μ est donnée par la formule : $t = \frac{m-\mu}{\delta/\sqrt{n}}$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au risque alpha = 5% pour un degré de liberté:d.d.l=n−1**[6]**

Exemple : Le taux de cholestérol dans la population est connu et vaut 4,3 mg/l. Les résultats de 15 pesées sont présentés dans le tableau suivant.
La moyenne de l'échantillon conforme la valeur de la population ?

4,5	5	3,8	4,5	4,9
4,8	5,2	5,1	4,6	5,2
4,3	3,9	5,2	4,8	5,1

Solution :

On utilise le test de Student pour un seul échantillon.

On suppose que la condition de normalité est vérifiée « c'est-à-dire que le taux de cholestérol est distribué suivant une loi normale »

1- On pose l'hypothèse nulle

H_0 :m=µ=4,3mg/L.

2- On calcule la statistique t :

$m = 4{,}726\ ;\ \delta = 0{,}455\ ;\ n = 15$

$$t = \frac{m-\mu}{\delta/\sqrt{n}} = \frac{4{,}726-4{,}3}{0{,}455/\sqrt{15}} = 3{,}626$$

3- Conclusion

Au risque $\alpha = 5\%$ avec le nombre de ddl $V = n_1 - 1 = 14$. La table de student-Fisher, nous donne $t_\alpha^* = 2{,}144$

$t > t_\alpha^*$alors on rejette H_0 et on conclut que la différence est significative.La moyenne de l'échantillon ne conformepas la valeur de la population.

1.2)Comparaison de deux moyennes pour deux échantillons

1.2.1) Test de deux échantillons indépendants

Il s'agit de deux séries de mesure pour lesquelles il n'y a aucune correspondance entre les éléments de la première série et ceux de la deuxième ; les deux séries de mesures sont obtenues avec des sujets différents. Dans ce cas le but de l'application du test t est de voir si les deux moyennes calculées sur les deux échantillons diffèrent significativement.**[6]**

***Exemple 01**

On veut comparer l'efficacité de deux traitements T_1 et T_2 ayant pour but de diminuer le taux d'urée dans le plasma. Le but de l'essai est de répondre à la question suivante:

*les deux traitements T_1 et T_2 sont-ils d'efficacité différente ?

Dans ce but, on prend un échantillon de malades que l'on sépare par tirage au sort en deux groupes de même effectif ou d'effectifs différents (avant toute expérience, les deux groupes doivent avoir sensiblement le même taux d'urée dans le plasma).

On administre le traitement T_1aux malades du premier groupe et le traitement T_2 à ceux du deuxième groupe.

A l'issue de l'expérience, on obtient les taux moyens respectifs d'urée $\overline{x_1}$ et $\overline{x_2}$ ainsi que les écarts-types δ_1 et δ_2

***Exemple 02**

On veut évaluer l'efficacité d'un traitement T ayant pour but d'augmenter le taux de glycémie. Le but de l'essai est de répondre à la question suivante:

Le traitement T est-il est efficace ?

Dans ce but, on sépare par tirage au sort un échantillon de malades en deux groupes ayant au départ sensiblement le même taux de glycémie. Les malades de premier groupe reçoivent le traitement T (appelé groupe traité) et ceux du deuxième groupe reçoivent un placebo (appelé groupe témoin).

A l'issue de l'expérience, on obtient les taux moyens respectifs de glycémie $\overline{x_1}$ et $\overline{x_2}$ ainsi que les écarts-types $\delta_1$et δ_2.

Pour pouvoir répondre à ces questions. On procède au test de comparaison de deux moyennes pour des observations indépendantes.

A- $n_1=n_2$

On calcule la valeur t observée (Tobs) qui suit une variable aléatoire de student aux degrés de liberté ddl=2n-2 et α%.

$$T_{obs} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{SCE_1+SCE_2}{n(n-1)}}}$$

Exemple d'application

Dans le cadre d'une étudeécotoxicologique, la concentration en DDT et en ses dérivés a été mesurée chez les brochets âgés respectivement de 2 et 3 ans. Les résultats obtenus sont les suivants :

[Pesticides] 2 ans	[Pesticides] 3 ans
0,144	0,285
0,171	0,295
0,178	0,321
0,184	0,354
0,193	0,359
0,197	0,361
0,198	0,362
0,199	0,364

Les moyennes de ces deux échantillons prélevés indépendamment différent elles de façon significative ?

Solution :

Echantillon 1 :([Pesticides] 2 ans) $n_1 = 8$

Echantillon 2 :([Pesticides] 3 ans) $n_2 = 8$

$n_1 = n_2$: on utilise le test t de Student.

On doitvérifier la condition de normalité : On suppose que la première condition est vérifiée « c'est-à-dire que la concentration en pesticide est distribuée suivant une loi normale »

1- On pose l'hypothèse nulle.

H_0 : Les deux échantillons ont la même moyenne $\overline{x}_1 = \overline{x}_2$

$$\overline{x}_1 = 0{,}183 \quad ; \quad \delta_1 = 0{,}018 \quad ; \quad \delta_1^2 = 0{,}000324$$
$$\overline{x}_2 = 0{,}337 \quad ; \quad \delta_2 = 0{,}032 \quad ; \quad \delta_2^2 = 0{,}001024$$

La somme des carrés des écarts : $SCE = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$

$T_{obs} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}}} = 11{,}634$

2. Conclusion

Au risque $\alpha = 5\%$ avec le nombre de ddl $V = n_1 + n_2 - 2 = 14$ La table de student-Fisher, nous donne $t_\alpha^* = 2.144$

$t > t_\alpha^*$alors on rejette H_0 et on conclut que la différence est significative. La concentration en pesticide est différente selon l'âge.

B- n1≠n2 où n1≥30 etn2≥30test de l'écart-réduit

1- on pose l'hypothèse nulle.

H_0 : Les deux échantillons proviennent de deux populations de même moyenne $\mu_1 = \mu_2$

Dans ex 01: H_0 : Les deux traitements T_1 et T_2 ont la même efficacité.

Dans ex 02: H_0 : Le traitement T n'est pas efficace.

02- on calcule la statistique ε:

$$\varepsilon = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\delta_1^2}{n_1 - 1} + \frac{\delta_2^2}{n_2 - 1}}}$$

03- conclusion

Au risque $\alpha = 5\%$

- Si $\varepsilon < 1.96$ alors on ne rejette pas H_0 et on conclut qu'il n'y a pas de différence significative entre les moyennes de deux échantillons.

- Si $\varepsilon \geq 1.96$ alors on rejette H_0 et on conclut qu'il y a une différence significative entre les moyennes de deux échantillons.

Remarque : pour un risque $\alpha = 1\%$, on compare la valeur de ε à 2.6**[6]**

Exemple d'application

On a prélevé deux échantillons de pommes pour le peser. Le premier échantillon, constitué de 100 pommes cueillies au début de la récolte, a pour moyenne 120g et pour écart type estimé 20g, le second, constitué de 150 pommes cueillies à la fin de la récolte, a pour moyenne 150g et pour écart type estimé 10g.

* La différence entre les poids moyens ces deux époques de la récolte est-elle significative ?

C- n1≠n2 où n1<30 etn2<30 : test t de student

01- on pose l'hypothèse nulle H_0, definie comme dans le 1er cas

02- on calcule la statistique t:

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

s^2:est la variance estimée de la population sous la condition d'égalité des variances des deux populations (condition d'utilisation du test t de student).

03- conclusion

On compare la valeur de t ainsi calculée avec celle t_α^* de la table de student-Fisher en fonction du nombre de ddl $v = n_1 + n_2 - 2$ et le risque $\alpha = 5\%$ ou $\alpha = 1\%$

- Si $t < t_\alpha^*$ alors on ne rejette pas H_0 et on conclut que la différence n'est pas significative.

- Si $t \geq t_\alpha^*$alors on rejette H_0 et on conclut que la différence est significative.**[6]**

***Condition d'utilisation du test t de student**

Dans le cas des petits échantillons ($n_1 < 30$ ou $n_2 < 30$), le test t de student n'est utilisable que si le caractère étudié est distribué dans les deux populations d'où proviennent les échantillons selon des lois normales et de même variance $\delta_{p_1}^2 = \delta_{p_2}^2$

La première condition est souvent vérifiée : les caractères rencontrés en biologie, du fait qu'ils résultent de l'addition d'un grand nombre d'effets indépendants, obéissent à des lois de probabilités proches de la loi normale.

Pour savoir si la 2eme condition est vérifiée, on utilise le test de comparaison de deux variances appelée test F de Fisher-snedecor.**[5]**

Test F de Fisher-snedecor

1- On pose l'hypothèse nulle

H_0 :les deux populations ont la même variance ($\delta_{p_1}^2 = \delta_{p_2}^2$)

02- On calcule

La variance estimée de $\delta_{p_1}^2 : s_1^2 = \frac{n_1}{n_1-1}\delta_1^2$

La variance estimée de $\delta_{p_2}^2 : s_2^2 = \frac{n_2}{n_2-1}\delta_2^2$

$Si:\ s_1^2 > S_2^2, on\ calcule\ la\ statistique\ F: F = \frac{S_1^2}{S_2^2}$

$Si:\ s_1^2 < S_2^2, on\ calcule\ la\ statistique\ F: F = \frac{S_2^2}{S_1^2}$

03- Conclusion

- $Si: F = \frac{S_1^2}{S_2^2}$, on la compare avec la valeur de Fs de la table de Fisher-snedecor en fonction du nombre de ddl $V_1 = n_1 - 1$ (colonne) et $V_2 = n_2 - 1$ (ligne) et le risque $\alpha = 5\%$
- $Si: F = \frac{S_2^2}{S_1^2}$, on la compare avec la valeur de Fs de la table de Fisher-snedecor en fonction du nombre de ddl $V_2 = n_2 - 1$ (colonne) et $V_1 = n_1 - 1$(ligne) et le risque $\alpha = 5\%$

- Si $F < F_s$ alors on ne rejette pas H_0 et on conclut que la différence n'est pas significative.
- Si $F \geq F_s$ alors on rejette H_0 et on conclut que la différence est significative.**[5]**

Remarque : Si les variances des deux populations d'où proviennent les deux échantillons sont différentes, on utilise le test de l'écart-réduit amélioré par WELCH. Ce test consiste à calculer, comme dans le test de l'écart-réduit, la quantité:$\mathcal{E} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\delta_1^2}{n_1-1} + \frac{\delta_2^2}{n_2-1}}}$

On compare la valeur de $\mathcal{E}$ avec la valeur donnée par la table de student-Fisher en fonction du risque de α = 5 ou α = 1% et le nombre de ddl V est l'entier le plus proche de k donné par la relation.

$$k = \frac{\left(\frac{\delta_1^2}{n_1-1} + \frac{\delta_2^2}{n_2-1}\right)^2}{\frac{1}{n_1-1}\left(\frac{\delta_1^2}{n_1-1}\right)^2 + \frac{1}{n_2-1}\left(\frac{\delta_2^2}{n_2-1}\right)^2}$$

Exemple 1:

Dans des études d'anesthésie, on peut comparer l'effet de deux somnifères. On a noté les durées de sommeil qui ont suivi les injections d'une dose bien définie. On a obtenu les résultats suivants :

Echantillon	Durée de sommeil exprimé en minutes
Somnifère 1	170 ; 175 ; 187 ; 180 ; 190 ; 165 ; 175 ; 174 ; 173 ; 181
Somnifère 2	155 ; 160 ; 164 ; 150 ; 160 ; 159 ; 154 ; 156 ; 160 ; 167 ; 153 ; 158

Comparer les deux somnifères.

Solution :

Echantillon 1 :(Somnifère 1) $n_1 = 10$

Echantillon 2 :(Somnifère 2) $n_2 = 12$

n_1 et $n_2<30$: on utilise le test de Student.

On doitvérifier les deux conditions : On suppose que la première condition est vérifiée « c'est-à-dire que la durée de sommeil est distribuée suivant une loi normale » et on teste l'égalité des deux variances au moyen du test F de Fisher-Snedecor.

$$\overline{x}_1 = 177 \quad ; \quad \delta_1 = 7.21 \quad ; \quad \delta_1^2 = 52$$

$$\overline{x}_2 = 158 \quad ; \quad \delta_2 = 4.54 \quad ; \quad \delta_2^2 = 20.66$$

Test F de Fisher-snedecor

1. On pose l'hypothèse nulle

H_0 :les deux populations ont la même variance ($\delta_{p_1}^2 = \delta_{p_2}^2$)

2. On calcule

La variance estimée de$\delta_{p_1}^2$: $s_1^2 = \frac{n_1}{n_1-1}\delta_1^2 = 57.77$

La variance estimée de$\delta_{p_2}^2$: $s_2^2 = \frac{n_2}{n_2-1}\delta_2^2 = 22.54$

$s_1^2 > S_2^2, on\ calcule\ la\ statistique\ F: F = \frac{S_1^2}{S_2^2} = 2.56$

3. Conclusion

Au risque $\alpha = 5\%$ avec le ddl $V_1 = n_1 - 1 = 9$ (colonne) et $V_2 = n_2 - 1 = 11$ (ligne), la table de Fisher-snedecor, nous donne Fs=2.9.

$F < F_s$alors on ne rejette pas H_0 et on conclut que la différence n'est pas significative.

Test t de student

4. On pose l'hypothèse nulle

H_0 :les deux somnifères ont le même effet.

5. On calcule la statistique t : t=7.17 avec δ^2=38.36

6. Conclusion

Au risque $\alpha = 5\%$ avec le nombre de ddl $V = n_1 + n_2 - 2 = 20$. La table de student-Fisher, nous donne$t_\alpha^* = 2.086$

$t > t_\alpha^*$alors on rejette H_0 et on conclut que la différence est significative. Les deux somnifères ont des effets différents : le premier provoque des sommeils de plus longue durée que le second ($\bar{x}_1 > \bar{x}_1$).

Exemple 2 d'application

Pour déterminer le poids moyen d'épis de blé appartenant à deux variétés, on a procédé 10 pesées pour chacune. Les moyennes obtenues sont X1= 107.7 et X2= 168.5. On admet que le poids de graines est distribué de chaque variété suivant la loi de Gauss et que les variances des deux distributions peuvent être considérées comme égales. Les estimations obtenus pour celles-ci sont S^2_1= 432.9 et S^2_2= 182.7.

* Les deux moyennes sont-elles significativement différentes au risque α=5% ?

1.2.2) Test de deux échantillons dépendants ou appariés

Il s'agit de deux séries de mesures pour lesquelles il y a une correspondance stricte, terme à terme, entre les éléments de l'une et les éléments de l'autre.

Exemple : on veut comparer deux somnifères S1 et S2 sur un échantillon de n sujets. Les n sujets ont reçu le somnifère S_1un soir et on a obtenu les durées de sommeil X1, X2 ..., Xn exprimé en heures.

Une semaine plus tard, les mêmes sujets ont reçu le somnifère S_2 et on a obtenu les durées de sommeil Y1, Y2,…, Yn est exprimée en heures.

Dans ce cas, les observations sont appariées puisque chaque sujet et son propre témoin (chaque sujet fait l'objet d'une durée de sommeil avec les deux somnifères).

Quel est des deux somnifères le plus efficace. **[7]**

Test de comparaison :

1. On pose l'hypothèse nulle

H_0 : il n'y a pas de différence entre les moyennes $\bar{x}\ et\ \bar{y}$

- Les deux somnifères en la même efficacité

2. On calcule :

- Les différences :$d_1 = x_1 - y_1\ ;\ d_2 = x_2 - y_2\ ;\ \ldots\ ;\ d_n = x_n - y_n$
- La moyenne de ces différences :$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} di = \bar{x} - \bar{y}$
- L'écart type de ces différences : $\delta_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2 - \bar{d}^2}$
- Ensuite on calcule la statistique$\varepsilon = \frac{\bar{d}}{\frac{\delta_d}{\sqrt{n-1}}}$

03. Conclusion

1er Cas : n≥30 : test de l'écart réduit

Au risque $\alpha = 5\%$

- - Si $\varepsilon < 1.96$ alors on ne rejette pas H_0 et on conclut qu'il n'y a pas de différence significative entre les moyennes.
- - Si $\varepsilon \geq 1.96$ alors on rejette H_0 et on conclut qu'il y a une différence significative entre les moyennes.
- Remarque : pour un risque $\alpha = 1\%$, on compare la valeur de ε à 2.6

2ème Cas : n<30 : test t de student

On compare la valeur de t ainsi calculée avec celle t_α^* de la table de student-Fisher en fonction du nombre de ddl v $= n - 1$ et le risque $\alpha = 5\%$ ou $\alpha = 1\%$

- Si $t < t_\alpha^*$ alors on ne rejette pas H_0 et on conclut que la différence n'est pas significative.

- Si $t \geq t_\alpha^*$alors on rejette H_0 et on conclut que la différence est significative.**[5]**

Conditions d'utilisation du test t de student :

Le test t de student n'est utilisable que si la différence des valeurs du caractère étudié est distribuée dans la population d'où provient l'échantillon selon la loi normale.

Exemple :

On veut comparer les effets de deux médicaments provoquant un retardement de battements du cœur sur un échantillon de 8 chats. Une expérience à donner les résultats ci-après :

N° de chat		**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**
Changement de	**Médicament A**	-22	-14	-36	-28	-8	-22	-8	+2

battements du cœur	Médicament B	-14	-12	-22	-30	+10	0	-8	+24

Solution

n = 8 inférieur à 30, on utilise donc le test t de student en supposant que la différence du changement des battements cardiaques par les deux médicaments est distribué dans la population selon une loi normale.

Test t de student

1. On pose l'hypothèse nulle

H_0 : Les deux médicaments ont la même efficacité

2. On calcule

	Changement de battements du cœur par			
N° de chat	**Médicament A (xi)**	**Médicament B (yi)**	$d_i = x_i - y_i$	d_i^2
01	-22	-14	-8	64
02	-14	-12	-2	04
03	-36	-22	-14	196
04	-28	-30	+2	04
05	-8	+10	-18	324
06	-22	0	-22	484
07	-8	-8	0	0
08	+2	+24	-22	484
Total	$\sum_{i=1}^{8} xi = -136$	$\sum_{i=1}^{8} yi = -52$	$\sum_{i=1}^{8} di = -84$	$\sum_{i=1}^{8} d_i^2 = 1560$

$\bar{x} = -17$; $\bar{y} = -6.5$

- La moyenne de ces différences : $\bar{d} = \frac{1}{8}\sum_{i=1}^{8} di = -10.5$
- L'écart type de ces différences : $\delta_d = \sqrt{\frac{1}{8}\sum_{i=1}^{8} d_i^2 - \bar{d}^2} = 9.2$
- D'où : $\varepsilon = \frac{\bar{d}}{\frac{\delta_d}{\sqrt{n-1}}} = -3.01$

03. Conclusion

Au risque $\alpha = 5\%$, avec v=n-1=7, la table de student-Fisher, nous donne la valeur $t_{\alpha}^{*} = 2.365$

$t > t_{\alpha}^{*}$alors on rejette H_0 et on conclut que la différence est significative. Les effets des deux médicaments sont différents. Le médicament A est plus efficace que le médicament B dans le retardement du battement cardiaque.

Exemple d'application : La quantité de bactéries par cm^3 de lait provenant de 8 vaches différentes est estimée juste après la traite et 24 h plus tard.

Vache	Juste après la traite	24h après la traite
1	12000	14000
2	13000	20000
3	21500	31000
4	17000	28000
5	15000	26000
6	22000	30000
7	11000	16000
8	21000	29000

Existe-t-il un accroissement significatif du nombre de bactéries par cm^3 de lait au cours du temps ?

IV.2. Comparaison des variances

1. **Comparaison de deux variances**

La comparaison de deux variance peut être réalisée sans difficulté et d'une manière exacte par les distributions F de snedecor, lorsque les échantillons sont aléatoires, simples, indépendants et sont tirés de populations normales « voir chapitre précédent ».

2. **Comparaison de plusieurs variances**

- Test de Hartley

- **Effectif des échantillons égaux :**

Lorsque les effectifs des échantillons sont égaux à n, ce test permet de vérifier rapidement l'hypothèse nulle d'égalité des variances par le calcul suivant : $H_{cal} = \frac{SCE_{max}}{SCE_{min}}$

Conclusion :

- Si $H_{cal} < H_{théo}$, on accepte l'hypothèse nulle et on rejette l'hypothèsealternative.
- Si $H_{cal} \geq H_{théo}$, on rejette l'hypothèse nulleet on conclut qu'il y a des différences significatives entre les variances.
- Pour k échantillons, le ddl$V = (n_1 - 1) + (n_2 - 1) \ldots\ldots + (n_k - 1)$.

- **Effectif des échantillons inégaux**

Lorsque les effectifs des échantillons sont inégaux le test est toujours valable selon la formule suivante : :$H_{cal} = \frac{\delta^2{}_{max}}{\delta^2{}_{min}}$

Conclusion :

- Si $H_{cal} < H_{théo}$, on accepte l'hypothèse nulle et on rejette l'hypothèsealternative.
- Si $H_{cal} \geq H_{théo}$, on rejette l'hypothèse nulle et on conclut qu'il y a des différences significatives entre les variances.
- Pour k échantillons, le ddl$V = (n_1 - 1) + (n_2 - 1) \ldots\ldots + (n_k - 1)$.

La somme des carrés des écarts : $SCE = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$**[3]**

V. ANALYSE DE LA VARIANCE « ANOVA »

INTRODUCTION

Nous avons traité dans le chapitre précédent, des problèmes de comparaison de deux séries d'observation. Cependant il existe de très nombreux problèmes posés en biologie et en médecine où on est amené à comparer plus de deux séries d'observations. Ce type de problème pourrait nous faire croire qu'il est légitime d'effectuer la comparaison deux à deux en utilisant des tests statistiques vu dans le chapitre précédent. Cependant la comparaison multiple entraîne comme nous l'avons indiqué dans le chapitre de la théorie des tests, une inflation du risque d'erreur. Il faut donc effectuer un test unique et global permettant de traiter ce problème.

Dans ce chapitre, on examine la comparaison de plusieurs moyennes afin de savoir si, dans l'ensemble, elle différents significativement ou non. C'estl'analyse de la variance qui va nous permettre de traiter ce problème au moyen d'un test unique et global appelé test F de Fisher-Snedecor. **[7]**

01- Analyse de la variance à un facteur « ANOVA I »

L'analyse de la variance à un facteur ou à un seul critère de classification, a pour but d'étudier les effets d'un facteur ou traitement par le biais de comparaison des moyennes de plusieurs populations supposées distribuées normalement et de même variance, à partir des échantillons aléatoires, simples et indépendants, les unes des autres. On considère R échantillon E1, E2,

..... avec R≥3 de tailles respectives n1, n2,.....,nk desquels on étudie un certain caractère quantitatif. L'ensemble des observations est désigné dans le tableau suivant :

Echantillon	**E1**	**E2**	**……Ej…..**	**E_R**
Valeur du caractère	X_{11} X_{21} X_{31} . . X_{i1} X_{ni1}	X_{12} X_{22} X_{32} . . X_{i2} X_{ni2}	X_{1j} X_{2j} X_{3j} . . X_{ij} X_{nij}	X_{1R} X_{2R} X_{3R} . . X_{iR} X_{niR}
Tailles	n_1	n_2	n_j	n_p
Moyenne	$\bar{x}_{i1}$	$\bar{x}_{i2}$	$\bar{x}_j$	$\bar{x}_{np}$
Variance	δ_1^2	δ_2^2	δ_j^2	δ_p^2

- Les moyennes des échantillons et la moyenne générale de l'ensemble des observations permettent de définir deux type de variation : Les écarts existent entre les différents échantillons « variation entre échantillons ou variation factorielle » et les écarts existant à l'intérieur des échantillons « variation résiduelle ». **[5]**

- L'importance de ces deux sources de variation est mesurée par deux quantités appelé communément carrés moyens ou variances : ce sont **le carré moyen factoriel**, défini à partir des écarts entre les moyennes des échantillonset la moyenne générale, et **le carré moyen résiduel** qui est fonction des écarts existant entre les observations et la moyenne de l'échantillon correspondant.

- S'il y a des différences importantes entre les moyennes des populations, cela doit se refléter à travers les moyennes des échantillons. On doit donc s'attendre à observer un carré moyen factoriel plus élevé que le carré moyen résiduel. Le rapport de ces deux termes est une mesure du degré de fausseté et de l'hypothèse nulle.

- Aussi la somme des carrés des écarts par rapport à la moyenne générale, appelé aussi somme des carrés des écarts totale, peut-être divisée en deux composantes additives : une somme des carrés des écarts factorielle et une somme des carrés des écarts résiduelle.

$$SCE_t = SCE_f + SCE_r$$

- Test F de Fisher-snedecor :

1- On pose l'hypothèse nulle

H_0 : Les P échantillons proviennent des P populations de même moyennes

$$\bar{x}_1 = \bar{x}_2 = \cdots \ldots \ldots \ldots = \bar{x}_p$$

2- On calcule la statistique F :

$F = CM_f/CM_r$

3- Conclusion

Au risque alpha = 5 % on compare la valeur de F calculée avec la valeur de F théorique de la table de Fisher-snedecor en fonction du nombre de ddl, V_A = P- 1 (colonne), V_B=P(n-1).

- Si $F_{cal} \geq F_{théo}$: Alors on rejette H_0 et on conclut que les moyennes différentes significativement dans l'ensemble.
- Si $F_{cal} < F_{théo}$: Alors on accepte H_0 et on conclut que les moyennes ne diffèrent pas significativement dans l'ensemble.

En général, les résultats d'analyse de la variance sont représentés dans le tableau suivant :

Tableau d'analyse de la variance à un facteur

Source de variation	**ddl**	**SCE**	**CM**	**F test**
Variation factorielle	P-1	SCE_f	SCE_f/P-1	CM_f/ CM_r
Variation résiduelle	P(n-1)	SCE_r	SCE_r/P(n-1)	
Variation totale	Pn-1	SCE_t	/	

$SCE_t = SCE_f + SCE_r \rightarrow SCE_f = SCE_t - SCE_r$

$$SCE_t = T - \frac{X_.^2}{np} \; ; \qquad SCE_r = \sum_{i=1}^{p} SCE_i \; ;$$

$$SCE_i = \sum_{i=1}^{ni} x_i^2 - \frac{\left(\sum_{i=1}^{ni} x_i\right)^2}{n} \; ; \quad T = \sum_{j=1}^{p} \sum_{i=1}^{ni} x_i^2 \textbf{[6]}$$

Conditions d'utilisation du test F

Le test de Fisher-snedecor n'est utilisable que si le caractère étudié est reparti dans les populations selon des lois de probabilités proche de la loi normale de même variance c'est-à-dire : $\delta_{p1}^2 = \delta_{p2}^2 = \cdots\cdots\cdots\cdots = \delta_{pp}^2$**[7]**

Exemple

Une équipe de chercheurs à mener une étude doit l'objectif est de tester des techniques d'inactivation in vivo de plusieurs solutions d'œstrogène. Le poids utérins (mg) de 4 souris par traitement « 6 plus le contrôle » est pris comme mesure de l'activité hormonale d'œstrogène. Les données sont représentées dans le tableau suivant :

Rép	Traitements						
	Contrôle	1	2	3	4	5	6
01	89,8	84,8	64,4	75,2	88,4	56,4	65,6
02	93,8	116	79,8	62,4	90,2	83,2	79,4
03	88,4	84	88	62,4	73,2	90,4	65,6
04	68,6	68,6	69,4	73,8	87,8	85,6	70,2

- Tester les effets des traitements au niveau alpha = 0,05.

Solution :

1- On pose l'hypothèse nulle

$H_0 : \bar{x}_c = \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \cdots\cdots = \bar{x}_6$

H_1 : Il existe au moins une différence significative entre les moyennes

2- On calcule la statistique F :

Fcal = CMf/CMr

On calcule : $\sum x_i$; $\sum x_i^2$; $\frac{(\sum x_i)^2}{n}$; $SCEi$; $\sum\sum x_i = X.$; $\sum\sum x_i^2 = T$

$X. = 2249$; $SCEr = 3062{,}57$ $T = 186121{,}4$

$$SCE_t = T - \frac{X_.^2}{np} = 186121{,}4 - \frac{5058001}{28} = 5478{,}51$$

$$SCE_f = SCE_t - SCE_r = 5478{,}51 - 3062{,}57 = 2415{,}94$$

Tableau de l'ANOVA

Source de variation	ddl	SCE	CM	F test
Traitement	6	2415,94	402,65	2,76
Erreur	21	3062,57	145,83	
total	27	5478,51	/	

α = 5 % on compare la valeur de F calculée avec la valeur de F théorique de la table de Fisher-snedecor en fonction du nombre de ddl, V_A = 7- 1=6 (colonne), V_B=P(n-1), V_B=7(4-1)=21 (ligne).

$F_{théo}$=(2,60+2,55)/2=2,57<F_{cal}

Alors on rejette H_0 et on accepte H1 : Il y a des différences significatives entre les traitements.

Exemple d'application : La quantité d'oxygène consommée par la patelle *Acmaeascabra*(µlO_2/mg/min) a été mesurée pour différentes conditions expérimentales. Les résultats sont lessuivants :

100% eau de mer	75% eau de mer	50% eau de mer
12,1	10,4	14,6
8,9	7,2	11,7
13,6	6,4	16,9
9,7	13,3	18,8
9,7		
7,2		
14,1		
8,3		

La salinité affecte-t-elle la respiration des patelles (α= 0,05) ?

02. Analyse de la variance à deux facteurs « ANOVA II »

Nous avons vu que l'analyse de la variance à un critère de classification a notamment pour principe de repartir la variance totale en deux composantes : L'une factorielle et l'autre résiduelle. Pour l'analyse de la variance à deux critères de classification, la variance totale étant divisée en plus de deux composantes, l'une résiduelle et les autres liées aux deux facteurs de classification.

Considérant p, q population à partir desquelles sont prélevés des échantillons d'effectif n :désignons par X_{ijk} les valeurs observées.

- L'indice i servant à distinguer pvariantes du premier critère de classification (i=1,……,p).

- L'indice j indique q variantes du deuxième critère de classification (j=1,……,q).

- Et k désignant pour chaque échantillon, les numéros d'ordre des différentes observation (k=1,……,n).**[6]**

L'ensemble des observations est consigné dans le tableau suivant :

I	1	…….	p
j / k	1……..q	…….	1…….q
1	X_{111}……X_{1q1}	…….	X_{p11}……X_{pq1}
2	X_{112}……X_{1q2}	…….	X_{p12}……X_{pq1}
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
n	X_{11n}……X_{1qn}	…….	X_{p1n}……X_{pqn}

* Dans ce cas la somme des carrés des écarts par rapport à la moyenne générale ou la somme des carrés des écarts totale, peut être divisée en plus de deux composantes : une somme des carrés des écarts factorielle « facteur 1 », une somme des carrés des écarts factorielle « facteur 2 », une somme des carrés des écarts de l'interaction entre les deux facteurs et une somme des carrés des écarts résiduelle.

$SCE_t = SCE_{f1} + SCE_{f2} + SCE_{int} + SCE_r$**[6]**

1/ Test F de Fisher-snedecor : « Cas de modèle fixe »

- On pose l'hypothèse nulle

$H_0 : a_1 = a_2 = \cdots\cdots\cdots\cdots = a_p$ ou $\bar{x}_{1.} = \bar{x}_{2.} = \cdots \ldots \ldots \ldots = \bar{x}_{p.}$l'absence de l'effet du premier facteur.

$H_0 : b_1 = b_2 = \cdots\cdots\cdots\cdots = b_q$ ou $\bar{x}_{.1} = \bar{x}_{.2} = \cdots \ldots \ldots \ldots = \bar{x}_{.q}$ l'absence de l'effet du deuxième facteur

$H_0 : C_{11} = C_{12} = \cdots\cdots\cdots\cdots = C_{pq}$l'absence de l'effet de l'interaction.

2. On calcule la statistique F :

F_{cal}pour le 1er facteur : $F_{cal} = \frac{CM_{f1}}{CM_r}$

F_{cal}pour le 2ème facteur : $F_{cal} = \frac{CM_{f2}}{CM_r}$

F_{cal}pour l'interaction entre les deux facteurs : $F_{cal} = \frac{CM_{int}}{CM_r}$

3. Conclusion

Au risque α= 5 % ou α= 1 %, on compare la valeur de F calculée avec la valeur de F théorique de la table de Fisher-snedecor en fonction du nombre de ddl,

- Pour le premier facteur : V_A = p- 1 (colonne), V_B=pq(n-1) ligne.
- Pour le deuxième facteur : V_A = q- 1 (colonne), V_B=pq(n-1) ligne.
- Pour l'interaction : V_A = (p- 1)(q-1) (colonne), V_B=pq(n-1) ligne

- Si $F_{cal} \geq F_{théo}$: Alors on rejette H_0
- Si $F_{cal} < F_{théo}$: Alors on accepte H_0

Tableau d'analyse de la variance à deux facteurs

Source de variation	ddl	SCE	CM	F test
Facteur 01	p-1	SCE_{f1}	$CM_{f1}=SCE_{f1}/p-1$	$F_{cal\ f1}$
Facteur 02	q-1	SCE_{f2}	$CM_{f2}=SCE_{f2}/q-1$	$F_{cal\ f2}$

Interaction	(p-1)(q-1)	SCE_{int}	$CM_{int}=SCE_{int}/(p-1)(q-1)$	F_{calint}
Résiduelle	pq(n-1)	SCE_r	$CM_r=SCE_r/pq(n-1)$	
Totale	pqn-1	SCE_t	/	

$SCE_t = SCE_{f1} + SCE_{f2} + SCE_{int} + SCE_r \rightarrow SCE_{int} = SCE_t - [SCE_{f1} + SCE_{f2} + SCE_r]$

- $SCE_t = T - \frac{X_{..}^2}{pqn}$; $SCE_r = \sum_{i=1}^{p} SCE_{i.}$;
- $SCE_i = \sum_{i=1}^{ni} x_i^2 - \frac{\left(\sum_{i=1}^{ni} x_i\right)^2}{n}$; $T = \sum_{i=1}^{p}\sum_{j=1}^{q} x_{i.}^2$; $X_{..} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_{i.}$
- $SCE_{f1} = \frac{1}{qn}\sum_{i=1}^{p} x_{i.}^2 - \frac{X_{..}^2}{pqn}$; $SCE_{f2} = \frac{1}{pn}\sum_{j=1}^{q} x_{.j}^2 - \frac{X_{..}^2}{pqn}$

Conditions d'utilisation du test F

Ce test F de Fisher-snedecor n'est utilisable que si le caractère étudié est reparti dans les populations selon des lois de probabilités proche de la loi normale de même variance.**[3]**

Remarque : Cas d'échantillons à une seule observation : Une observation par échantillon (n = 1). Les données théoriques de l'analyse de la variance nesont modifiées en aucune façon, sauf que le terme résiduel disparaît puisque pour chaque échantillon sa moyenne se confond avec son unique observation.

Tableau d'analyse de la variance

Source de variation	ddl	SCE	CM	F test
Facteur 01	p-1	SCE_{f1}	$CM_{f1}=SCE_{f1}/p-1$	$F_{cal\ f1=}\ CM_{f1}/CM_{int}$
Facteur 02	q-1	SCE_{f2}	$CM_{f2}=SCE_{f2}/q-1$	$F_{cal\ f2=}\ CM_{f2}/CM_{int}$
Interaction	(p-1)(q-1)	SCE_{int}	$CM_{int}=SCE_{int}/(p-1)(q-1)$	
Totale	pq-1	SCE_t	/	

$SCE_t = SCE_{f1} + SCE_{f2} + SCE_{int} \rightarrow SCE_{int} = SCE_t - [SCE_{f1} + SCE_{f2}]$

- $SCE_t = T - \frac{X_{..}^2}{pq}$;
- $SCE_{f1} = \frac{1}{q}\sum_{i=1}^{p} x_{i.}^2 - \frac{X_{..}^2}{pq}$; $SCE_{f2} = \frac{1}{p}\sum_{j=1}^{q} x_{.j}^2 - \frac{X_{..}^2}{pq}$ **[2]**

Exemple : Les données du tableau ci-après sont celles d'une expérimentation ayant pour objectif de tester les effets du temps, et de l'atropine susceptible de retarder l'atrophie musculaires suite à la dénervation d'un des principaux muscles d'une patte arrière de rat. 48 animaux sont répartis aléatoirement en 4 groupes recevant chacun une forte dose d'atropine (a), une dose modérée de quinidine(b), une dose modérée d'atropine (c) et une solution saline témoin (d). Pour chaque traitement 4 rats sont choisis aléatoirement et le poids de membre dénervé est mesuré le 4ème, 8ème et 12ème jour.

Durée	4 jours				8 jours				12 jours			
	Atropine											
Rép	a	b	c	d	a	b	c	d	a	b	c	d
01	0,94	1,19	1,22	0,99	0,91	0,87	0,67	0,97	0,34	0,41	0,57	0,81
02	1,16	1,15	0,90	1,51	0,73	1,04	0,72	1,07	0,43	0,87	0,80	1,01
03	1,26	0,85	1,00	1,55	0,52	0,88	1,08	1,16	0,41	0,91	0,69	0,97
04	0,85	1,21	1,00	0,98	0,65	0,96	0,75	1,04	0,48	0,87	0,84	0,87

En considérant les facteurs fixes, tester les effets des traitements au niveau α=0,05. **[6]**

Solution :

1. On pose les hypothèses nulles

- Facteur 1 : Traitement

$H_0 : a_1 = a_2 = a_3 = a_4$ ou $\bar{x}_{1.} = \bar{x}_{2.} = \bar{x}_{3.} = \bar{x}_{4.}$* l'absence de l'effet de traitement.

- Facteur 2 : Durée

$H_0 : b_1 = b_2 = b_3$ ou $\bar{x}_{.1} = \bar{x}_{.2} = \bar{x}_{.3}$ * l'absence de l'effet Temps

- Interaction : traitement*durée

$H_0 : C_{11} = C_{12} = \cdots\cdots\cdots = C_{pq}$* l'absence de l'effet de l'interaction.

2. On calcule la statistique F :

- Facteur 1 : Dose $F_{cal} = \frac{CM_D}{CM_r}$; Comparaison avec le $F_{théo}$de la table de Fisher-snedecor.
- Facteur 2 : Temps $F_{cal} = \frac{CM_t}{CM_r}$; Comparaison avec le $F_{théo}$de la table de Fisher-snedecor.
- Interaction : $F_{cal} = \frac{CM_{int}}{CM_r}$; Comparaison avec le $F_{théo}$de la table de Fisher-snedecor.

On passe à l'analyse de la variance à deux facteurs :

$\sum x_{i.}$	4,21	4,4	4,12	5,03	2,81	3,75	3,22	4,24	1,66	3,06	2,9	3,66
$\sum x_{i.}^2$	4,539	4,925	4,298	6,623	2,053	3,534	2,696	4,513	0,699	2,51	2,146	3,374
$\frac{(\sum x_{i.})^2}{n}$	4,431	4,84	4,243	6,325	1,974	3,516	2,592	4,494	0,688	2,34	2,10	3,348
$SCE_{i.}$	0,108	0,085	0,055	0,298	0,079	0,018	0,104	0,019	0,011	0,17	0,046	0,026

- $\sum_{i=1}^{p} SCE_{i.} = SCE_r = 1{,}019$;
- $\sum_{i=1}^{p}\sum_{j=1}^{q} x_{i.} = X_{..} = 43{,}06$; $\sum_{i=1}^{p}\sum_{j=1}^{q} x_{i.}^2 = T = 41{,}91$;
- $SCE_t = T - \frac{X_{..}^2}{pqn} = 41{,}91 - \frac{(43{,}06)^2}{48} = 3{,}282$;

Calcule de SCE_{f1} etSCE_{f2}

	a	**b**	**c**	**d**	**Total**

4 J	4,21	4,4	4,12	5,03	$\sum x_{.1} = 17{,}76$
8 J	2,81	3,75	3,22	4,24	$\sum x_{.2} = 14{,}02$
12 J	1,66	3,06	2,9	3,66	$\sum x_{.3} = 11{,}28$
Total	$\sum x_{1.} = 8{,}68$	$\sum x_{2.} = 11{,}21$	$\sum x_{3.} = 10{,}24$	$\sum x_{4.} = 12{,}93$	$X_{..} = 43{,}06$

$$\sum x_{i.}^2 = (8{,}68)^2 + (11{,}21)^2 + (10{,}24)^2 + (12{,}93)^2 = 473{,}049$$

$$\sum x_{.j}^2 = (17{,}76)^2 + (14{,}02)^2 + (11{,}28)^2 = 639{,}216$$

$$SCE_{f1} = \frac{1}{qn}\sum_{i=1}^{p} x_{i.}^2 - \frac{X_{..}^2}{pqn} = \frac{1}{3\times4} \times 473{,}049 - \frac{(43{,}06)^2}{48} = 39{,}420 - 38{,}628 = 0{,}792\ ;$$

$$SCE_{f2} = \frac{1}{pn}\sum_{j=1}^{q} x_{.j}^2 - \frac{X_{..}^2}{pqn} = \frac{639{,}216}{4\times4} - \frac{(43{,}06)^2}{48} = 39{,}951 - 38{,}628 = 1{,}323$$

$$SCE_{int} = SCE_t - [SCE_{f1} + SCE_{f2} + SCE_r] = 3{,}282 - [0{,}792 + 1{,}323 + 1{,}019]$$
$$= 3{,}282 - 3{,}134 = 0{,}148.$$

Source de variation	**ddl**	**SCE**	**CM**	**F_{cal}**
Dose	4-1=3	0,792	0,792/3=0,264	9,428
Temps	3-1=2	1,323	1,323/2=0,661	23,607
Interaction	(3-1)(4-1)=6	0,148	0,148/6=0,024	0,857
Résiduelle	3*4(4-1)=36	1,019	1,019/36=0,028	
Totale	3*4*4-1=47	3,282		

- $F_{cal\ f1} > F_{théo} = 2{,}88$: Alors on rejette H_0, il y a un effet significatif du dose
- $F_{cal\ f2} > F_{théo} = 3{,}275$: Alors on rejette H_0, il y a un effet significatif du temps
- Si $F_{cal\ I} < F_{théo} = 2{,}38$: Alors on accepte H_0pas d'effet significatif de l'interaction.

2/ **Test F de Fisher-snedecor : « Cas de modèle Aléatoire »**

- On pose l'hypothèse nulle

H0 : $a_1 = a_2 = \cdots\cdots\cdots\cdots = a_p$ ou $\bar{x}_{1.} = \bar{x}_{2.} = \cdots \ldots \ldots \ldots = \bar{x}_{p.}$l'absence de l'effet du premier facteur.

H0 : $b_1 = b_2 = \cdots\cdots\cdots\cdots = b_q$ ou $\bar{x}_{.1} = \bar{x}_{.2} = \cdots \ldots \ldots \ldots = \bar{x}_{.q}$ l'absence de l'effet du deuxième facteur

H0 : $C_{11} = C_{12} = \cdots\cdots\cdots\cdots = C_{pq}$l'absence de l'effet de l'interaction.

2. On calcule la statistique F :

F_{cal}pour le 1er facteur : $F_{cal} = \frac{CM_{f1}}{CM_{int}}$

F_{cal}pour le 2ème facteur : $F_{cal} = \frac{CM_{f2}}{CM_{int}}$

F_{cal}pour l'interaction entre les deux facteurs : $F_{cal} = \frac{CM_{int}}{CM_r}$

3. Conclusion

Au risque α= 5 % ou α= 1 %, on compare la valeur de F calculée avec la valeur de F théorique de la table de Fisher-snedecor en fonction du nombre de ddl,

- Pour le premier facteur : V_A = p- 1 (colonne), V_B=(p- 1)(q-1)ligne.
- Pour le deuxième facteur : V_A = q- 1 (colonne), V_B=(p- 1)(q-1)ligne.
- Pour l'interaction : V_A = (p- 1)(q-1) (colonne), V_B=pq(n-1) ligne**[6]**

Tableau de l'ANOVA

Source de variation	**ddl**	**SCE**	**CM**	**F_{cal}**
Facteur 01	p-1	SCE_{f1}	$CM_{f1}=SCE_{f1}/p-1$	$F_{cal} = \frac{CM_{f1}}{CM_{int}}$
Facteur 02	q-1	SCE_{f2}	$CM_{f2}=SCE_{f2}/q-1$	$F_{cal} = \frac{CM_{f2}}{CM_{int}}$

Interaction	(p-1)(q-1)	SCE_{int}	$CM_{int}=SCE_{int}/(p-1)(q-1)$	$F_{cal}=\frac{CM_{int}}{CM_r}$
Résiduelle	pq(n-1)	SCE_r	$CM_r=SCE_r/pq(n-1)$	
Totale	pqn-1	SCE_t	/	

3/ Test F de Fisher-snedecor : « Cas de modèle mixte »

Tableau de l'ANOVA : « Facteur 01 fixe et facteur 2 aléatoire »

Source de variation	**ddl**	**SCE**	**CM**	**F_{cal}**
Facteur 01	p-1	SCE_{f1}	$CM_{f1}=SCE_{f1}/p-1$	$F_{cal}=\frac{CM_{f1}}{CM_{int}}$
Facteur 02	q-1	SCE_{f2}	$CM_{f2}=SCE_{f2}/q-1$	$F_{cal}=\frac{CM_{f2}}{CM_r}$
Interaction	(p-1)(q-1)	SCE_{int}	$CM_{int}=SCE_{int}/(p-1)(q-1)$	$F_{cal}=\frac{CM_{int}}{CM_r}$
Résiduelle	pq(n-1)	SCE_r	$CM_r=SCE_r/pq(n-1)$	
Totale	pqn-1	SCE_t	/	

On compare la valeur de F calculée avec la valeur de F théorique de la table de Fisher-snedecor en fonction du nombre de ddl :

- Pour le premier facteur : V_A = p- 1 (colonne), V_B=(p- 1)(q-1)ligne.
- Pour le deuxième facteur : V_A = q- 1 (colonne), V_B=pq(n-1) ligne.
- Pour l'interaction : V_A = (p- 1)(q-1) (colonne), V_B=pq(n-1) ligne**[7]**

VI. ANALYSE DE LA COVARIANCE (ANCOVA)

01. INTRODUCTION

L'analyse de covariance procède à la fois de l'analyse de variance et de la régression linéaire simple. Plus généralement, elle permet d'étudier les influences d'une variable qualitative à plusieurs classes et d'une variable explicative quantitative sur une variable dépendante quantitative. On suppose qu'on a k droites de régression linéaire indépendantes entre deux variables quantitatives x et y issues de k expériences différentes.

Tout comme l'Anova, la procédure Ancova vise à déterminer l'effet d'une variable catégorielle (indépendante) sur une variable quantitative (dépendante). La particularité de l'Ancona est de calculer cet effet en contrôlant l'effet d'une autre variable (une ou plusieurs ; quantitatives ou qualitatives) qui a un impact présumé sur la relation initiale. **[3]**

Remarque :

L'intérêt du modèle d'ANCOVA, c'est qu'il permet de

1- séparer l'effet spécifique du facteur étudié de l'effet de la covariable

2- réduire la variance résiduelle, ce qui augmente la puissance du test du facteur étudié

Exemple :

On veut comparer trois méthodes de traitement de dyslexie « difficultés d'apprentissage de lecture » chez des enfants d'école primaire. On mesure, chez chaque enfant, son âge et le progrès. On désire estimer un modèle de régression pour expliquer le progrès en fonction de l'âge, du traitement et de l'interaction âge-traitement car on croit que l'effet du traitement peut dépendre de l'âge.**[6]**

Traitement 01		**Traitement 02**		**Traitement 03**	
Age	**Progrès**	**Age**	**Progrès**	**Age**	**Progrès**
6	42	6	44	6	80
7	70	7	50	7	88
8	105	8	65	8	118
9	123	9	72	9	136

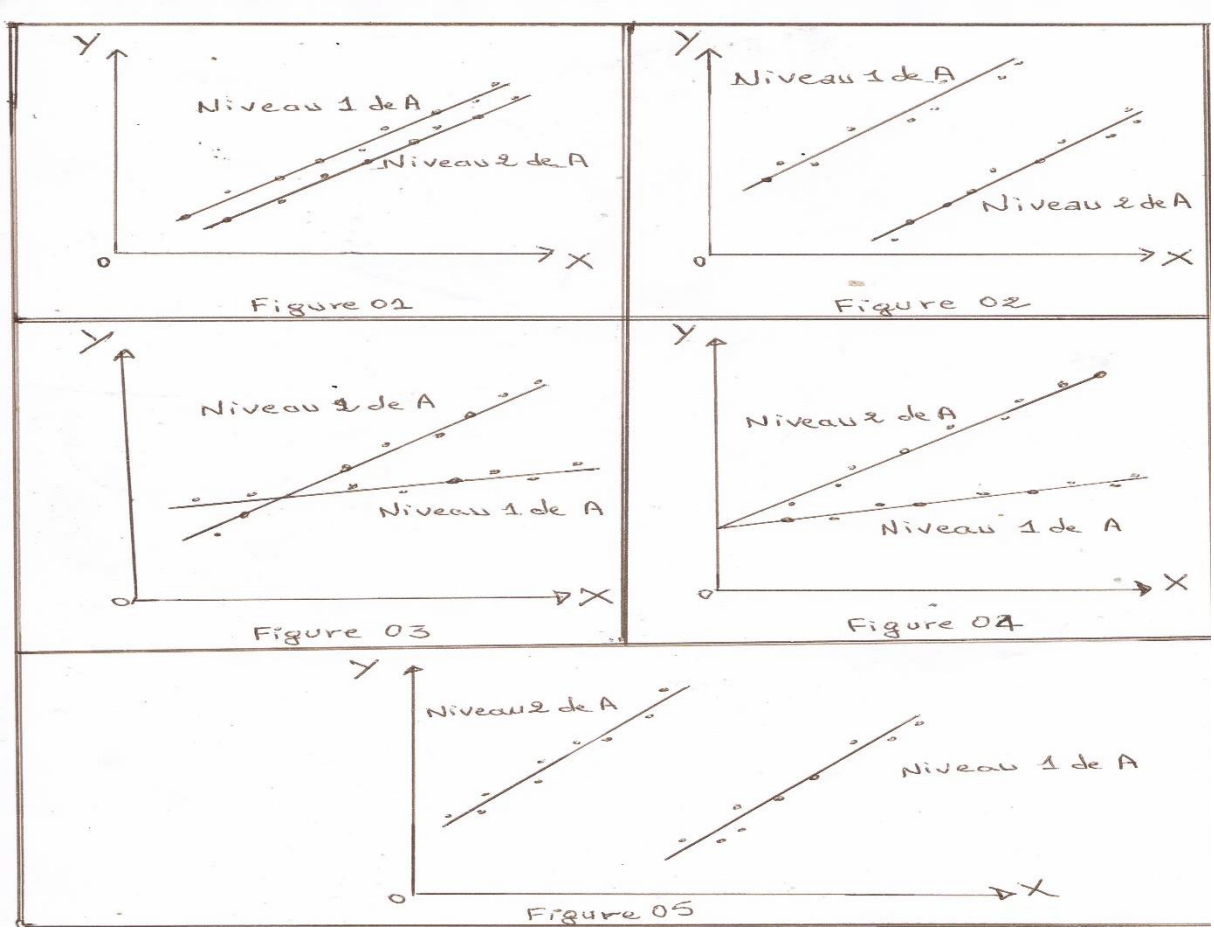
On se propose de répondre à la question suivante : Les k droites obtenues sont-elles ? Autrement dit, peut-on dire que la quantité de variation en regroupant les données des k groupes est la même que celle en calculant k régression ? Si la réponse est affirmative, on conclura que les trois méthodes de traitement donnent le même progrès et il n'y a pas d'interaction entre laméthode de traitement est le rapport entre le progrès et l'âge de l'enfant. S'il existe une différence significative entre les k droites de régression, les questions se posent alors à trois niveau :

- Le Progrès change-t-il en fonction de l'âge indépendamment de la méthode de traitement ?
- Sinon, y a-t-il une interaction entre la méthode de traitement et le rapport entre le progrès et l'âge de l'enfant.
- Sinon, les méthodes de traitement donnent-elles le même progrès indépendamment de l'âge de l'enfant ?

La méthode d'analyse permettant de répondre à ces trois questions s'appelle de covariance cette méthode d'analyse est une méthode d'ajustement puisqu'il s'agit d'étudier la régression.**[6]**

02. Interprétation graphique de l'analyse de la covariance

L'exemple graphique ci-dessous présente 5 situations possibles avec diverses combinaisons de pentes et d'ordonnées à l'origine. On étudier la variable quantitative à expliquer y en fonction de la covariablex pour chaque niveau d'une variable d'ajustement A.



On interprète, d'une manière descriptive, ces situations sans que les testsaient été faits en réalité :

- Figure 01 : Les pentes et les ordonnées à l'origine sont identiques. Il n'y a donc ni interaction ni effet de la variabled'ajustement A sur la variable y. Il y a seulement la liaison entre la variable y et la covariable x.

- Figure 02 : Les pentes sont parallèles (il n'y a pas donc une interaction). Il y a l'effet de la variable A sur la variable Y (les ordonnées à l'origine sont différentes), dans ce cas, les pentes parallèles indiquent que la variable A n'influe pas sur la relation entre la variable y et la covariable x, par conséquent, la différence ordonnée à l'origine peut être interprétée comme un effet de la variable A sur la variable y, indépendamment à la covariable X.

- Figure 03 : Les pentes et les ordonnées à l'origine sont différentes. Il y a donc une interaction entre ni la variable A et le rapport entre la variable y et la covariable x. (la manière dans la variable y réagit à la covariable dépend du niveau de la variable A). Dans ce cas, on ne pourra pas rechercher s'il y a l'effet de la variable A sur la variable Y puisque que les ordonnéesà l'origine dépendant à la fois de la covariable X et la variable A

- Figure 4 : Les pentes sont différentes et les ordonnées à l'origine sont identiques. Comme dans la figure 3, il y a une interaction et la recherche éventuelle de l'effet de la variable A sur la variable Y ne peut pas se faire.

- Figure 5 : Les valeurs de la covariable X ne sont pas les mêmes pour les deux niveaux de la variable A. Il y a donc un effet de la variable A sur la covariable X. **[6]**

Chapitre VII : Statistiques descriptives multidimensionnelle

1. Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) fait partie des analyses descriptives multi-variées. Le but de cette analyse est de résumer le maximum d'informations possibles et en perdant le moins possible pour :

• Faciliter l'interprétation d'un grand nombre de données initiales ;

• Donner plus de sens aux données réduites.

L'ACP permet donc de réduire des tableaux de grandes tailles en un petit nombre de variables (2 ou 3 généralement) tout en conservant un maximum d'information. **[6]**

2. L'Analyse Factorielle des Correspondances (AFC)

Définition

Il s'agit d'une méthode de description statistique multidimensionnelle d'un tableau de données qualitatives. Représentation des similitudes entre les individus et entre les modalités des variables qualitatives.

Sonobjectifestd'analyserla liaisonexistantentre deuxvariablesqualitatives(siondisposede plusde deuxvariablesqualitatives,on aura recours à l'AnalysedesCorrespondances Multiples, AFCM).Ainsi,avant de mettre en œuvre uneA.F.C.,il fauts'assurerquecetteliaisonexistebien.Onnotera qu'ondisposeaussid'unteststatistique,le testdu khi-deux d'indépendance, basé surl'indice khi-deux,permettantde testers'ilexisteounonuneliaison significativeentre deuxvariablesqualitatives.Cetestest très simple à mettre en œuvre maisne relèvepas de la statistiquedescriptive.L'A.F.C.est,enfait,uneAnalyseenComposantesPrincipales particulière, réalisée surlesprofils associés à la tablede contingence croisantlesdeuxvariables considérées.Plus précisément,l'A.F.C. consiste à réaliseruneA.C.P.surles profils-ligneset uneautre sur les profils-colonnes. Les résultatsgraphiquesde ces deuxanalysessontensuite superposéspourproduire ungraphique (éventuellementplusieurs)de type nuagede points,danslequelsont réuniesles modalités desdeuxvariables considérées,ce quipermet d'étudierlescorrespondances entre ces modalités,autrementditla liaisonentre lesdeuxvariables.**[6]**

Référence :

[1] :Motulsky H. J. BIOSTATISTIQUE UNE APPROCHE INTUITIVE. De boeck. P 447.

[2] :Balan, R et Lamothe, G. PREVOIR L'IMPREVISIBLE-UNE INTRODUCTION A LA BIOSTATISTIQUE. Presses de l'Université de Québec. P 276

[3] :Vessereau, A. 1992. METHODES STATISTIQUES EN BIOLOGIE ET EN AGRONOMIE. Imprimé en France, Paris.

[4] : Mallet, A. et Morice, V. 2012. QCM corrigées et commentées de biostatistique. Editions ellipses.

[5] :Laberche, J-C. STATISTIQUES ET EXPERIMENTATION EN BIOLOGIE : Outils- statistiques inférentielles.Editions ellipses.

[6] :Borsali, F., 2010. STATISTIQUES ET MEDICALES ET BIOLOGIQUES.Editions ellipses. P 409

[7] :Triola, Marc M et Triola, Mario F. 2012. BIOSTATISTIQUE pour les Sciences de la vie et de la santé. Pearson France, Paris. P 367.