Chapitre III. L'estimation statistique et les test d'hypothèse

Cette partie de la statistique, qui, contrairement à la statistique descriptive, ne se contente pas de décrire des observations mais extrapole des observations faites sur un ensemble limité à un ensemble plus large, permet de tester des hypothèses sur cet ensemble et de prendre des décisions à leur sujet.

1. Méthodes statistiques par relatives aux valeurs moyennes

Toutes les méthodes que nous allons présenter en relation avec les moyens supposent que les conditions suivantes soient remplies :

- Normalité des populations initiales,
- Aléa et simplicité des échantillons tirés,
- Pour certains tests, en plus, égalité des variances des populations.

La première condition n'est pas indispensable pour les grands nombres (N > 30).

1.1. Intervalle de confiance et test de conformité d'une moyenne

1.1.1. Intervalle de confiance

- Dans le cas où la variance de la population parente est connue, les limites de confiance de la moyenne $\bar{X}$ estimée sont :

$$\bar{\bar{X}} = \bar{X} \pm \mu_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

α peut prendre habituellement la valeur (0.05, 0.01 ou 0.001). $\mu_{1-\alpha/2}$ est tirée à partir de la table de la distribution normale réduite.

- Dans le cas où la variance de la population parente est inconnue, alors il faut l'estimer et ceci provoque l'élargissement de l'intervalle de confiance. Cet élargissement est obtenu pour un degré de confiance (1- α) en remplaçant la valeur $\mu_{1-\alpha/2}$ de la distribution normale par la valeur $t_{1-\alpha/2}$ de la distribution de Student à (n – 1) degrés de liberté. Les limites de confiance de la moyenne $\bar{X}$ estimée sont ainsi :

$$\bar{\bar{X}} = \bar{X} \pm t_{1-\alpha/2} \frac{S}{\sqrt{n}}$$

S représente l'écart type estimé et $t_{1-\alpha/2}$ est tirée à partir de la table de la distribution de Student pour (n – 1) ddl.

En pratique cette formule est remplacée par :

$$\bar{\bar{X}} = \bar{X} \pm t_{1-\alpha/2} \sqrt{\frac{SCE}{n(n-1)}}$$

Lorsque n est supérieur à 30 et α = 0.05

$$\bar{\bar{X}} = \bar{X} \pm 2 \sqrt{\frac{SCE}{n(n-1)}}$$

1.1.2. Test de conformité

Le test de conformité d'une moyenne a pour but de vérifier si la moyenne $\bar{X}$ d'une population est ou n'est pas égale à une valeur donnée $\bar{X}_0$. On rejette l'hypothèse d'égalité ($\bar{X} = \bar{X}_0$) lorsque la moyenne observée est trop différente de la valeur théorique.

Le test est réalisé en calculant la valeur suivante :

$$t_{obs.} = \frac{|\bar{X} - \bar{X}_0|}{\sqrt{\frac{SCE}{n(n-1)}}}$$

On rejette l'hypothèse H0 : $\bar{X} = \bar{X}_0$ si $t_{obs.} \geq t_{1-\alpha/2}$ pour (n-1) ddl. et l'accepter si $t_{obs.} < t_{1-\alpha/2}$ pour (n-1) ddl.

1.2. Test de signification d'une différence de deux moyennes : échantillons indépendants

Le fait que les variances soient inconnues provoque, comme pour l'intervalle de confiance, dans la comparaison de deux moyennes, l'introduction d'une valeur estimée des variances et le remplacement de la distribution normale par la distribution t de Student.

1.2.1. Populations de même variance

En supposant satisfaites les conditions générales (Populations normales et Echantillons aléatoires et simples), alors nous pouvons tester l'hypothèse nulle suivante :

H0 : $\bar{X}_1 = \bar{X}_2$ en calculant la quantité suivante :

$$t_{obs.} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{SCE1+SCE2}{n_1+n_2-2}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}$$

On rejette l'hypothèse H0 : $\bar{X}_1 = \bar{X}_2$ si $t_{obs.} \geq t_{1-\alpha/2}$ pour $(n_1 + n_2 - 2)$ ddl. et l'accepter si $t_{obs.} < t_{1-\alpha/2}$ pour $(n_1 + n_2 - 2)$ ddl. n_1

Ce test est appelé test t de Student ou de Student-Fisher.

Lorsque les deux échantillons sont de même effectif (n), la condition relative à l'égalité des variances n'est pas nécessaire et l'expression précédente devient :

$$t_{obs.} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{SCE1+SCE2}{n(n-1)}}}$$ et le nombre de degrés de liberté est de 2(n – 1).

Dans ces conditions, le test de Student est qualifié de « Robuste »

D'autre part, lorsque l'effectif total ($n_1 + n_2$) est suffisamment élevé (de l'ordre de 30 au moins), on peut remplacer $t_{1-\alpha/2}$ par la valeur correspondante $\mu_{1-\alpha/2}$ de la distribution normale réduite, l'hypothèse de normalité des populations parentes est alors d'importance secondaire.

Quand on rejette l'hypothèse nulle d'égalité des deux moyennes, nous pouvons calculer l'intervalle de confiance de la différence $\overline{X}_1 - \overline{X}_2$.

- Cas des effectifs différents :

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\alpha/2} \sqrt{\frac{SCE1+SCE2}{n_1+n_2-2}\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

- Cas des effectifs égaux :

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\alpha/2} \sqrt{\frac{SCE1+SCE2}{n\,(n-1)}}$$

1.2.2. Populations de variances différentes

Plusieurs auteurs ont montré que l'hypothèse de normalité est secondaire dans le test d'égalité de deux moyennes. De même l'hypothèse d'égalité des variances n'est pas fondamentale lorsque les effectifs sont égaux. Pour ces deux raisons (insensibilité à la non normalité et inégalité des variances), on dira que le test de Student est robuste lorsque les effectifs sont égaux.

Par contre, lorsque les effectifs sont inégaux, il est indispensable de s'assurer de l'égalité des variances. Si cette hypothèse n'est pas vérifiée, il faut procéder à une transformation de la variable mesurée afin de stabiliser les variances et utiliser ensuite le test t de Student tel que décrit précédemment sur les données transformées. On peut aussi utiliser les méthodes de la statistique non paramétrique dans le cas où une condition quelconque du test fait défaut. Dans ce sens, plusieurs tests sont disponibles :test de Mood, test de Wilcoxon pour échantillons indépendants, test de White, test de Mann et Whitney etc…

1.3. Test de signification d'une différence de deux moyennes : échantillons associés

Un autre cas important de comparaison de moyennes est relatif aux échantillons dont les individus sont associés par paires ou par couples. Ce cas se présente, par exemple, quand on compare deux méthodes de mesure en soumettant à ces deux méthodes les mêmes individus, tirés d'une population donnée. A chacune des méthodes correspond alors une population de mesures, mais les populations et les échantillons extraits ne sont pas indépendants.

Pour tester l'égalité des moyennes, on doit alors considérer la population des différences et vérifier la nullité de sa moyenne. On remplace alors le test d'égalité de deux moyennes par un test de conformité d'une moyenne.

Les conditions d'application du test sont :

- caractère aléatoire et simple de l'échantillon

- normalité de la population des différences.

Le test se réalise comme suit :

On pose l'hypothèse nulle : (H0) : $\bar{d} = 0$

On calcule les différences :

E_1	E_2	Différences (d_i)
x_{11}	x_{21}	$x_{11} - x_{21} = d_1$
x_{12}	x_{22}	$x_{12} - x_{22} = d_2$
…	…	…
x_{1n}	x_{2n}	$x_{1n} - x_{2n} = d_n$

On calcule la quantité suivante :

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

On rejette l'hypothèse H0 : ($\bar{d} = 0$) si $t_{obs.} \geq t_{1-\alpha/2}$ pour (n – 1) ddl.

et l'accepter si $t_{obs.} < t_{1-\alpha/2}$ pour (n – 1) ddl.

Ce test est appelé test de Student par couples.

Quand on rejette l'hypothèse de nullité de la moyenne des différences, nous pouvons calculer l'intervalle de confiance de la différence.

$$\bar{d} \pm t_{1-\alpha/2} \sqrt{\frac{SCEd}{n\,(n-1)}}$$

Dans le cas où les conditions d'application ne sont pas satisfaites, la normalisation de la distribution par une transformation de la variable mesurée est indispensable. Si la transformation ne donne pas un résultat, on s'oriente alors vers les méthodes de la statistique non paramétrique : test de Wilcoxon pour échantillons associés, test de Page, test des signes etc…

1.4. Analyse de la variance (ANOVA)

Dans le cadre des tests d'hypothèses, nous avons émis des hypothèses concernant la moyenne d'une population (test de conformité) puis comparé les moyennes de deux populations (test d'homogénéité de Student). Ce chapitre a trait à la comparaison des moyennes de plusieurs populations (> 2), test appelé communément analyse de la variance.

L'analyse de variance à un facteur ou Anova a pour objectif de tester l'effet d'un facteur A sur une variable quantitative mesurée sur plus de 2 populations. Ceci revient à comparer les moyennes de plusieurs populations normales et de même variance à partir d'échantillons aléatoires simples et indépendants les uns des autres. Chaque échantillon extrait d'une

population indépendante correspond à une modalité du facteur A. Le terme Anova indique que la comparaison des moyennes de plusieurs populations correspond en fait à la comparaison de deux variances (la variance factorielle et la variance résiduelle).

Les données relatives à une analyse de variance à un facteur sont structurées dans un tableau du type suivant :

i / *k*	*1*	*2*	…….	*p*	*Totaux*
1	x_{11}	x_{12}		x_{1p}	
2	x_{21}	x_{22}		x_{2p}	
. . .	. . .	. . .	. …….. .	. . .	
n	x_{n1}	x_{n2}	……	x_{np}	

Le facteur A présente p modalités ($1 \leq i \leq p$). On parle aussi de niveaux ou traitements. Le nombre de répétitions k pour une modalité i est noté ni. Le nombre de répétitions pour chaque modalité du facteur n'est pas forcément le même.

La valeur prise par la variable x pour la modalité i du facteur et la répétition k est notée x_{ik}.

Conditions d'application de l'ANOVA

Indépendance et caractère aléatoire des échantillons : L'indépendance entre les différentes valeurs de la variable mesurée x_{ik} est un préalable essentiel à la réalisation d'une analyse de variance. Les échantillons p comparés sont aléatoires et indépendants.

Normalité : la variable quantitative étudiée a une distribution normale dans les p populations comparées.

La normalité de la variable peut être testée, par exemple, avec le test de Shapiro et Wilk.

Remarque : Si la normalité de la variable n'est pas vérifiée, soit elle est transformée pour la normaliser, soit l'équivalent non paramétrique de l'Anova est appliqué, le test de Kruskall-Wallis pour les échantillons indépendants, ou le test de Friedman dans le cas de données liées.

Homoscédasticité ou égalité des variances des populations : Les p populations comparées ont la même variance. Différents tests permettent de vérifier l'égalité des variances des p populations comparées.

• Le test de Lévène est le test le plus satisfaisant pour effectuer la comparaison multiple de variances mais sa réalisation est assez longue.

• Le test de Bartlett est dédié à la comparaison multiple de variances avec un nombre de répétitions *n*i différent.

• Le test de Hartley est dédié à la comparaison multiple de variances avec un nombre de répétitions *n*i identiques

Remarque : Si l'hétérogénéité entre les variances est trop importante, on peut faire une transformation de la variable mesurée pour égaliser les variances, sinon on s'oriente vers l'analogue non paramétrique (test de Kruskal-Wallis pour les échantillons indépendants ou celui de Friedman dans le cas de données associées.).

Modèle de l'analyse de variance

L'analyse de variance à un facteur teste l'effet d'un facteur *A* ayant *p* modalités sur les moyennes d'une variable quantitative *x*.

L'hypothèse nulle testée est la suivante : il n'y a pas d'effet du facteur A et les *p* moyennes sont égales à une même moyenne $\bar{x}$.

H0 : $\bar{x}_1 = \bar{x}_2 = \ldots\ldots = \bar{x}_i = \ldots\ldots = \bar{x}_p = \bar{x}$

L'hypothèse alternative est la suivante : il y a un effet du facteur A et il existe au moins deux moyennes significativement différentes.

H1 : il existe au moins deux moyennes différentes ($\bar{x}i \neq \bar{x}j$)

Equation fondamentale de l'analyse de variance

Variation totale = Variation factorielle + Variation résiduelle

SCE totale = SCE factorielle + SCE résiduelle

Réalisation de l'ANOVA : La réalisation des calculs suit les étapes suivantes

i / k	1	2	…….	p	Totaux
1	x_{11}	x_{21}		x_{1p}	
2	x_{21}	x_{22}		x_{2p}	
. . .	. . .	. . .	. ……. .	. . .	
n	x_{n1}	x_{n2}	……	x_{np}	
n_i	n_1	n_2		n_p	**n.**
X_i**.**	X_1**.**	X_2**.**		X_p**.**	**X..**
$\sum x_{ik}^2$	$\sum x_{1k}^2$	$\sum x_{2k}^2$		$\sum x_{pk}^2$	T
$X_i.^2/n_i$	$X_1.^2/n_i$	$X_2.^2/n_i$		$X_p.^2/n_i$	
SCE_i	SCE_1	SCE_2		SCE_p	SCE_r

Pour pouvoir dresser le tableau d'analyse de la variance, il reste à calculer :

- Le terme correctif : $C = X_{..}^2/n_.$
- La SCE totale : $SCEt = T - C$
- La SCE factorielle : $SCE_f = SCE_t - SCE_r$
- Le carré moyen factoriel : $CM_f = SCE_f / (p-1)$
- Le carré moyen résiduel : $CM_r = SCE_r / (n_.-p)$
- Enfin, le rapport $F_{obs} = CM_f / CM_r$.

Nous pouvons enfin, réaliser le test en comparant *F*obs à $F_{1-\alpha}$ lue dans la table de la loi de Fisher-Snedecor pour un risque d'erreur α fixé et (p-1, n.-p) degrés de liberté et dresser le tableau de variation suivant :

Source de variation	ddl	SCE	CM	F_{obs}
Variation factorielle	(P – 1)	SCEf	CMf = SCEf/(P – 1)	CMf / CMr
Variation résiduelle	(n. – P)	SCEr	CMr = SCEr/(n. – P)	
Variation totale	(n. – 1)	SCEt		

• si *F*obs ≥ $F_{1-\alpha}$ l'hypothèse H0 est rejetée au risque d'erreur α : le facteur A a un effet significatif en moyenne sur les valeurs de la variable étudiée.

• si *F*obs < $F_{1-\alpha}$ l'hypothèse H0 est acceptée: le facteur A n'a pas d'effet significatif en moyenne sur les valeurs de la variable étudiée.

En analyse de la variance, il est intéressant de mesurer l'intensité de la différence dans le cas où l'hypothèse d'homogénéité est refusée au seuil de 5%. Pour ce faire, on refait la comparaison au seuil de 1% et dans le cas d'un second rejet, on passe au seuil $\alpha = 0.001$.

Voici les cas de figure que nous pouvons avoir :

- Accepter H0 au seuil $\alpha = 0.05$ ⟹ **le groupe des moyennes est homogène**.
- Refuser H0 au seuil $\alpha = 0.05$ ⟹ le groupe des moyennes est hétérogène, il faut passer aux seuils suivants pour mesurer l'intensité de la différence.
- Accepter H0 au seuil $\alpha = 0.01$ ⟹ les moyennes présentent entre elles des **différences significatives**.
- Refuser H0 au seuil $\alpha = 0.01$ ⟹ le groupe des moyennes est hétérogène, il faut passer au seuil $\alpha = 0.001$ pour mesurer toujours l'intensité de la différence.
- Accepter H0 au seuil $\alpha = 0.001$ ⟹ les moyennes présentent entre elles des **différences hautement significatives**.
- Refuser H0 au seuil $\alpha = 0.001$ ⟹ les moyennes présentent entre elles des **différences très hautement significatives**.

Remarque 1 : Les moyennes ne sont équivalentes que lorsque l'hypothèse d'égalité est acceptée au seuil de 5%. Dans le cas contraire, les moyennes sont différentes et la conclusion doit être tirée toujours par rapport au dernier rejet de l'hypothèse H0.

Remarque 2 : Le rejet de l'égalité des moyennes ne permet pas de savoir quelles sont les moyennes significativement différentes. Pour cela, la méthode de la comparaison multiple des moyennes permet de répondre à cette question. Un bon nombre de tests de comparaison multiples peuvent être utilisés (LSD, Student Newman et Keuls, Tukey etc.). Nous présenterons dans le chapitre VII, le test de la Plus Petite Différence Significative (PPDS) ou Low Significant Difference (LSD).

5. La comparaison des fréquences

Toutes les données recueillies par le biologiste peuvent être exprimées en pourcentages, en proportions ou en fréquences et faire ainsi l'objet des comparaisons que nous allons décrire.

Ce chapitre nous permet de traiter d'une façon optimale les données de nature qualitative. Le principe de la comparaison consiste à savoir si les écarts observés entre les estimations de deux ou plusieurs pourcentages, proportions ou fréquences peuvent être uniquement le résultat de fluctuations d'échantillonnage.

Les données générales du problème partent toujours d'un tableau de distribution de fréquences comportant **k** colonnes correspondantes aux **k** échantillons à comparer et **r** lignes relatives aux **r** classes ou catégories inventoriées.

Les données de base indispensables au calcul du test sont les fréquences absolues.

Nous présenterons le cas de la conformité d'une distribution observée à une distribution théorique et puis le test khi-deux d'homogénéité entre deux ou plusieurs distributions.

5.1 Le test χ^2 d'homogénéité

Ce test permet de comparer deux ou plusieurs populations normales à partir d'échantillons aléatoires et simples en mesurant l'écart qui existe entre des fréquences observées et des fréquences théoriques ou attendues et de tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.

Considérons le cas général de k échantillons aléatoires simples d'effectifs n_1, n_2, ….., n_k dont les éléments peuvent être classés en r catégories.

Echantillons / Catégories	1	2	….	j	….	k	∑
1	f_{11}	f_{12}		f_{1j}		f_{1k}	$f_{1.}$
2	f_{21}	f_{22}		f_{2j}		f_{2k}	$f_{2.}$

…							
i	f_{i1}	f_{i2}		f_{ij}		f_{ik}	$f_{i.}$
…							
r	f_{r1}	f_{r2}		f_{rj}		f_{rk}	$f_{r.}$
$\sum$	n_1	n_2		n_j		n_k	$n_.$

+ Hypothèse nulle (H0) : Les k échantillons constituent un groupe homogène

+ Calculer la quantité suivante :

$$\chi^2_{obs} = \sum \frac{(fobs - fth)^2}{fth}$$

Pour calculer la quantité χ^2_{obs} on doit d'abord chercher les fréquences théoriques ou attendues dans le cas de l'acceptation de l'hypothèse d'homogénéité. Les différentes fréquences théoriques correspondantes à chaque fréquence observée peuvent être obtenues à l'aide de la formule suivante :

$$fth_{ij} = \frac{\text{Total de la ligne i} * \text{Total de la colonne j}}{\text{Total général}} = \frac{fi. * nj}{n.}$$

+ Après calcul des fréquences théoriques, on calcul la quantité χ^2_{obs} et on la compare à la valeur théorique $\chi^2_{1-\alpha}$ tirée de la table du test khi-deux pour un risque d'erreur α choisi et un ddl = (k-1)(r-1).

Si $\chi^2_{obs} < \chi^2_{1-\alpha} \Rightarrow$ H0 est acceptée $\Rightarrow$ les distributions de fréquences des k échantillons sont identiques et le groupe et homogène.

Si $\chi^2_{obs} \geq \chi^2_{1-\alpha} \Rightarrow$ H0 est refusée $\Rightarrow$ les distributions de fréquences des k échantillons sont différentes et le groupe et hétérogène.

+*conditions d'application*

+ Si ddl = 1 :

- si l'effectif total se situe entre 20 et 40 et si toutes es fréquences théoriques sont ≥ 5, le test est valide à condition d'appliquer la correction de continuité de Yates

$$\chi^2_{cor} = \sum \frac{(|fobs - fth| - 0.5)^2}{fth}$$

- si l'effectif total est > 40, le test est valide à condition que toutes les fréquences théoriques soient ≥ 5.

+ Si ddl > 1 : Le test est valide à condition que toutes les fréquences théoriques soient ≥ 5.

Ces conditions sont à considérer en plus de la normalité des distributions et du caractère aléatoire et simple des échantillons extraits.

6. La corrélation

La notion de corrélation se rapporte au degré de liaison qui unit deux ou plusieurs variables. Selon la nature et le nombre de variables impliquées, on utilise une terminologie propre correspondant à des définitions différentes du même concept.

+ Liaison entre deux variables quantitatives distribuées normalement ⇒ corrélation linéaire simple.

+ Intensité de la relation liant une variable dépendante à un ensemble de variables indépendantes quantitatives ⇒ corrélation multiple.

+ Lien entre deux ensembles de variables quantitatives ⇒ corrélation canonique.

+ Relation entre deux variables semi quantitatives ⇒ corrélation de rang.

+ Relation entre deux variables qualitatives ⇒ association.

+ Relation entre deux variables qualitatives binaires ⇒ corrélation de point.

1. Corrélation entre deux variables quantitatives : corrélation de Pearson

La corrélation de Pearson ou de Bravais – Pearson est une mesure de la liaison linéaire existant entre deux variables quantitatives normales.

$$r_{xy} = \frac{Cov.xy}{Sx * Sy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{SCEx * SCEy}}$$

Toutes les valeurs de $r_{x,y}$ sont comprises entre -1 et +1. r = -1 ou +1 si tous les points du diagramme de dispersion sont situés sur une ligne droite et r = 0 lorsque le nuage de dispersion ne montre aucune tendance de relation entre les deux variables. Aussi, si r est positif ⇒ les deux variables augmentent au même temps et si r est négatif ⇒ l'une des variables augmente quand l'autre diminue.

6.1 Test de signification du coefficient de corrélation de Pearson

A partir d'un échantillon de n sujets sur lesquels on relève les couples de valeurs (X, Y), on estime r et on vérifie si l'estimation obtenue est suffisamment distante de 0 pour rejeter l'hypothèse d'indépendance (r = 0). H0 : r = 0 Les variables ne sont pas corrélées

Le r peut être comparé directement à la valeur critique donnée par la table de signification du coefficient de corrélation pour un ddl = au nombre de couples d'observation (x, y) diminué de 2 soit : (n-2) ddl, soit lorsque celle-ci est insuffisante (cas ou ddl > 100 et $\alpha < 1\%$), en calculant :

$$tr = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ et comparer cette valeur à $t_{1-\alpha/2}$ pour un ddl = n-2.

Si $t_r > t_{1-\alpha/2} \Rightarrow$ H0 est rejetée $\Rightarrow$ la liaison est significative.

Si $t_r \leq t_{1-\alpha/2} \Rightarrow$ H0 est acceptée $\Rightarrow$ la liaison n'est pas significative.

7. La régression

La régression est une méthode statistique qui permet de résumer la relation existant entre une variable aléatoire dépendante et une ou plusieurs variables aléatoires ou contrôlées appelées variables explicatives. La liaison entre les variables se résume en une équation et en l'estimation de quelques paramètres dont le coefficient de corrélation est le plus important. La régression permet de décrire la forme de la relation entre les variables : linéaire, polynômiale, hyperbolique, exponentielle, logistique etc. Elle permet également de prévoir les variations de la variable aléatoire dépendante « y » à partir de celles de « x » ou des « x_i » si nous avons plusieurs variables explicatives.

7.1 Régression linéaire simple

On parle de régression linéaire simple lorsqu'on désire calculer une fonction du premier degré liant les variables « x » et « y ». Cette fonction linéaire est de la forme :

$$\boldsymbol{y = ax + b}$$

Elle correspond à l'équation d'une ligne droite qui traverse au mieux le nuage de points et permet de calculer une valeur estimée « ŷ » pour chaque point de l'axe des x correspondant à la variable prédictive. Cette droite porte le nom de droite d'estimation ou droite de régression de « y » en « x ». Cette droite est obtenue par la méthode des moindres carrés dont le principe consiste à choisir la pente « a » et l'ordonnée à l'origine « b » de la droite qui minimisent la somme des carrés des erreurs. L'erreur est représentée par l'écart entre la valeur observée « y_i »

et la valeur prédite par la droite « $\hat{y}_i$ ». Cet écart « $e_i = y_i - \hat{y}_i$ » est appelé également « résidu ». La méthode des moindres carrés consiste à réduire au maximum la somme des carrés des écarts (résidus) : $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (yi - \hat{y}i)^2$.

$b = \bar{y} - a\bar{x}$ et $a = \frac{S_{xy}}{S_x^2}$ avec $S_{xy} = \frac{n \sum xy - (\sum x)(\sum y)}{n(n-1)}$

8. La comparaison multiple

8.1 La comparaison multiple des moyennes

L'analyse de la variance (ANOVA) constitue la première étape d'une comparaison des moyennes de plusieurs échantillons indépendants. Dans le cas du rejet de l'hypothèse d'homogénéité (Hypothèse nulle), une question supplémentaire se pose. En effet, il est intéressant de savoir quelles sont les moyennes qui différents significativement entre elles. Autrement, il faut poursuivre l'analyse par un test de comparaison multiple des moyennes pour rechercher les groupes homogènes éventuellement. Plusieurs tests nous permettent de répondre à cette question : test LSD (Least Significant Difference), test de Duncan (Duncan's multiple range test), test SNK (méthode de Student-Newman-Keuls), test HSD de Tukey (Honestly Significant Difference) etc. Nous présenterons dans le cadre de ce cours, le test LSD ou test de la plus petite différence significative (PPDS) équivalente au test HSD dans le cas d'effectifs égaux.

La réalisation du test se fait selon les étapes suivantes :

- Ordonner les moyennes des échantillons par ordre croissant de valeurs
- Calculer : $LSD = t_{1-\alpha/2}\sqrt{2\,CMr/n}$

où le ddl à considérer pour la valeur de $t_{1-\alpha/2}$ est celui du CMr = SCEr/n.-P), n l'effectif de chaque échantillon et n. l'effectif total.

Rejeter l'Hypothèse d'égalité des moyennes à chaque fois que la différence ($\overline{Xi} = \overline{Xj}$) est supérieure ou égale à la quantité LSD.

9. Test non paramétrique de comparaisons multiples

Si le test de Kruskal-Wallis indique une hétérogénéité au sein du groupe d'échantillons analysés, on se demande quels sont les échantillons qui différents les uns des autres ou quels groupes d'échantillons se révèlent significativement différents des autres. Le test non paramétrique de comparaisons multiples nous permet de répondre à cette question.

La réalisation du test se fait de la même manière que le test SNK selon les étapes suivantes :

- Ordonner les sommes des rangs des différents échantillons par ordre croissant,
- Faire une série de comparaisons en commençant avec la plus grande différence entre les sommes des rangs prises deux à deux,
- Calculer : $q_{KW} = \frac{Y.max - Y.min}{SR}$

où $(Y_{\bullet max} - Y_{\bullet min})$ correspond à la différence entre les sommes des rangs et S_R à l'erreur type donnée par la formule suivante : $S_R = \sqrt{\frac{n(n\ p)\ (n\ p + 1)}{12}}$

où n représente l'effectif de l'échantillon, qui doit être constant d'un échantillon à l'autre et p = 2 + (le nombre d'échantillons dont la valeur de $Y_{\bullet}$ est comprise entre $Y_{\bullet max} - Y_{\bullet min}$ considérés). Au premier pas de la démarche (P = k), au deuxième pas à (k-1) au troisième pas à (k-2) et ainsi de suite.

- Comparer la valeur calculée q_{KW} à la valeur critique q_{α} fournie par la table des valeurs critiques de l'étendue de Student pour α choisi en fonction de la valeur de k et du ddl = (∞).

Si $q_{kw} > q_{\alpha}$ H0 est rejetée et les deux sommes des rangs comparées sont significativement différentes au seuil considéré.