Statistiques inférentielles

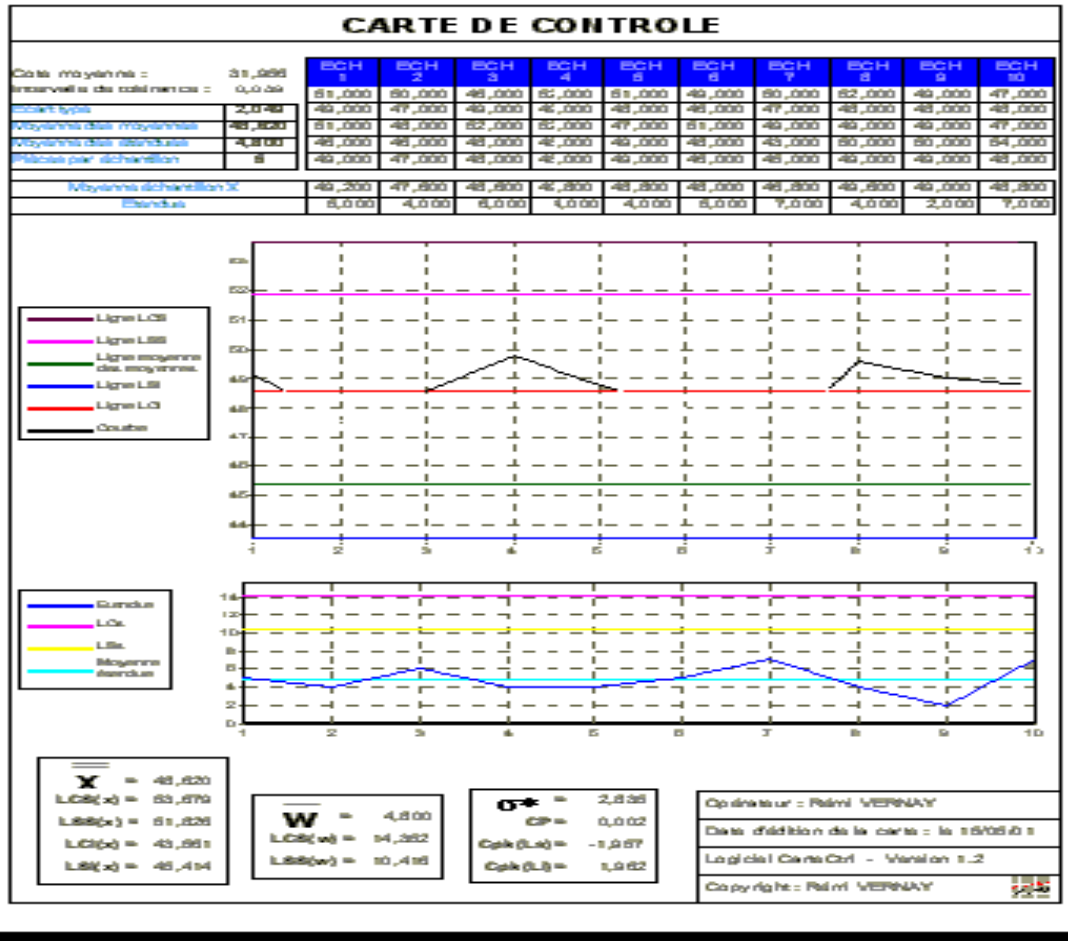
Prof: Najoua Fezzaa Ghriss

Université de Tunis

Institut Supérieur de l'Education et de la Formation Continue

Département de Sciences de l'Education

ED 209/1



Avril - 2009

Table des matières

Chapitre **Page**

STATISTIQUE INFERENTIELLE

Introduction

La statistique a envahit aujourd'hui tous les champs scientifiques. Son enseignement dans le cadre d'une formation en sciences de l'éducation s'y inscrit donc tout naturellement.

Tôt ou tard, vous serez amenés, dans vos projets d'apprentissage ou vos activités professionnelles, à recueillir des données, à les traiter, à les décrire et dans la mesure du possible les généraliser. L'utilisation des statistiques inférentielles s'avérera alors indispensable.

C'est justement dans ce cadre que s'insère ce module destiné aux étudiants du deuxième cycle de la maîtrise en sciences de l'éducation; il constitue le prolongement du module ED117 "Statistiques descriptives" et se propose comme objectif de permettre à l'étudiant d'acquérir les savoirs et les savoir-faire techniques et de les mobiliser à bon escient dans une situation de recherche éducative.

Objectifs spécifiques:

- Identifier le statut des variables dans une hypothèse de recherche;
- Choisir le test statistique approprié;
- Appliquer la procédure algorithmique de résolution
- Rédiger les conclusions dans un langage clair et précis.

Le module se compose de six chapitres. Pour construire les situations statistiques nous nous sommes appuyée sur des données statistiques réelles provenant de divers domaines, sur un ensemble de documents ou de questions inspirées d'ouvrages de statistiques ou de travaux de recherche. Nous avons jalonné le parcours par une série d'exercices afin de vous permettre de fixer vos acquis.

Premier chapitre: les bases de statistique inférentielle

1/ Définitions

D'après le dictionnaire Petit Robert (édition 1996): *inférer* : tirer (d'un fait, d'une proposition) une conséquence, c'est arguer, conclure, déduire, induire.

Inférence: opération logique par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies; c'est une déduction, induction.

Appliqué au terme statistique, on pourrait dire que les *statistiques inférentielles* servent à déduire à partir d'informations connues sur un ou plusieurs cas particuliers des conclusions sur ce qui se passe en général, qui est inconnu.

Autrement dit, les *Statistiques inférentielles* **:** Ensemble des méthodes et des théories permettant de généraliser à une population de référence des conclusions obtenues à partir de l'étude d'un échantillon extrait de cette population. Elles visent à:

- évaluer un paramètre ou une relation
- prédire une valeur
- déterminer si les différences sont dues au hasard
- déterminer si deux échantillons sont issus d'une même population

2/ Notion de base

2-1 Notion de variable: la variable est une propriété, un caractère qui permet de décrire et de classer les sujets ou les individus (objets d'étude). Le choix d'un caractère détermine le critère qui servira à classer les individus de la population en deux ou plusieurs sous-ensembles. Le nombre de ces sous-ensembles correspond aux diverses situations possibles ou *modalités* de ce caractère ou cette variable.

Ces variables peuvent se présenter sous 3 types d'échelles

1. Variables nominales: sont des variables de nature qualitative dont les modalités ne sont pas hiérarchisées.

- elles expriment l'appartenance d'un individu à un ensemble ou une catégorie non hiérarchique
- elles échappent à la mesure: elles peuvent seulement être constatées (par exemple, sexe, nationalité, profession)
- la relation qui définit une variable nominale est une relation d'appartenance à un ensemble.

2. Variables ordinales: sont des variables de nature qualitative dont les modalités sont hiérarchisées. Pour de telles variables, les modalités peuvent être classées par ordre de grandeur (par exemple

classe sociale, niveau d'études…). On distingue trois types de variables ordinales:

2.1. les variables rangées, qui se composent d'un nombre limité de modalités ordonnées les unes par rapport aux autres; par exemple, degré de concentration estimé sur une échelle à 4 degrés: 1 = non concentré; 2 = un peu concentré; 3 = moyennement concentré: 4 = non concentré;

2-2. les rangs, obtenus après un classement des unités d'observation de la première à la dernière, par exemple, d'après les résultats à un examen ou à une course; s'il n'y a pas d'ex æquo, il y aura autant de modalité que d'unité d'observation;

2.3. les scores rangés: mesures quantitatives classiques pour lesquelles on ne tient compte que des propriétés d'équivalence et d'ordre et pour lesquelles on ne prend pas en compte les autres propriétés arithmétiques du nombre (additivité, zéro vrai, intervalles numériques égaux).

3. Variables métriques ou d'intervalles: sont des variables de nature quantitative.

On peut attribuer à chaque élément évalué un nombre qui mesure ses propriétés. Ce nombre doit être tel que des intervalles numériques égaux représentent des distances égales dans la propriété mesurée.

Deux situations:

a. le zéro de l'échelle ne correspond pas à l'absence de la propriété chez l'élément caractérisé par la mesure zéro. Le zéro est arbitraire: on dit alors qu'il s'agit d'une donnée métrique dans une échelle d'intervalles. Dans ce cas, on peut soustraire des données, mais non les additionner, les multiplier ou les diviser. Par contre, on peut additionner, multiplier ou diviser des intervalles de données;

b. le zéro de l'échelle correspond à l'absence chez l'unité d'observation de la propriété observée : on dit alors qu'il y a un zéro vrai et on parlera de variable métrique (ou mesure) dans une échelle rationnelle; toutes les opérations arithmétiques ont un sens sur ce genre de données;

Exemple: la taille, l'âge, le nombre d'élèves par classe.

Remarque: on peut distinguer des variables métriques discrètes, qui ne peuvent prendre que des valeurs discrètes, c'est-à-dire séparées les unes des autres et correspondant à des nombres entiers indivisibles (par exemple, le nombre d'enfants dans une classe), des variables métriques continues, qui peuvent prendre toutes les valeurs possibles dans un intervalle (par exemple, la taille).

<u>2-2 Notion d'hypothèse:</u> l'hypothèse est une relation hypothétique (provisoire, postulée par le chercheur) entre une variable indépendante et une variable dépendante.

On distingue deux formes d'hypothèse:

- Hypothèse nulle (H0), postulant l'absence de différences entre les caractéristiques de

l'échantillon et celles de la population de référence.

- Hypothèse significative (H1), postulant l'existence de différences entre les caractéristiques de l'échantillon et celles de la population de référence.

L'hypothèse alternative peut être de deux types: soit dirigés (postule l'existence de la différence et précise le sens qu'elle prendrait) soit non dirigée (postule la différence sans précision du sens). Le premier type est di*t unilatéral* et le second est *bilatéral*.

3/ Concepts fondamentaux de l'inférence statistique

3-1 Population et échantillon

- Population: l'ensemble des sujets ou des événements visés par l'étude (les enseignants du primaire, scores des étudiants, revenus des personnes**,** etc.) = population de référence. Elle peut aller d'un ensemble de nombre relativement réduit, et donc facile à rassembler, à un ensemble de nombre important, fini ou infini, qui serait, en pratique difficile à rassembler dans son entièreté. Il en résulte que les chercheurs ont recours généralement à prélever de la population (de référence ou mère) un nombre déterminé de sujets ou d'observations = l'échantillon.
- Echantillon: un ensemble de sujets ayant les mêmes caractéristiques de la population-mère, utilisé en vue d'inférer quelque chose à propos de cette population. Il y a plusieurs types d'échantillon dont l'échantillon aléatoire, l'échantillon stratifié, l'échantillon par quotas, etc.

3-2 Tests d'hypothèse

C'est une fonction des variables aléatoires représentant l'échantillon dont la valeur numérique obtenue pour l'échantillon considéré permet de distinguer entre H_0 vraie et H_0 fausse.

Autrement dit, c'est une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données (échantillon). Il s'agit d'émettre, à partir de calculs réalisés sur des données observées, des conclusions sur la population, en leur rattachant des risques de se tromper.

Définir les hypothèses de travail, constitue un élément essentiel des tests d'hypothèses de même que vérifier les conditions d'application de ces dernières (normalité de la variable, égalité des variances)

- ✓ 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 te médiane), la procédure de test subséquente

ne porte alors que sur ces paramètres. L'hypothèse de normalité sous jacente des données est le plus souvent utilisée, la moyenne et la variance suffisent pour caractériser complètement la distribution. Concernant les tests d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances.

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

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

- ✓ Puissance d'un test:

C'est une évaluation de sa sensibilité, de sa capacité à détecter les effets significatifs dans les données quand, en fait, ils sont présents; lors de chaque test, nous acceptons une hypothèse et nous refusons l'autre

- On commet une erreur de type 1 quant on rejette H0, alors qu'elle n'est pas valable
- On commet une erreur de type 2 quand on accepte H0, alors qu'elle n'est pas valable

On dit qu'un test qui conduit à peu d'erreurs de type 2 est un test qui possède une haute puissance ou très sensible; à l'inverse, on dit qu'un test qui conduit à peu d'erreurs de type 1 et beaucoup d'erreurs de type 2 à une faible puissance. Il est souhaitable d'utiliser un test de haute puissance chaque fois que c'est possible; la puissance d'un test augmente avec la taille de l'échantillon.

3-3 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 ou de catégorique, 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. Au plus le seuil est petit, au moins il est probable que nous nous trompions quand nous prononçons pour le rejet ou l'acceptation d'une hypothèse; généralement , on travaille avec un seuil de 5%.

Deuxième chapitre: Rappel de statistique descriptive

1/Les distributions de fréquence

Une distribution de fréquence est l'ensemble du nombre de données qu'il y a dans chaque classe de résultats. Chaque distribution de fréquence doit comporter les classes d'observation (x) et la fréquence de chaque classe.

On y trouve donc, dans les colonnes successives:

a. les fréquences brutes (effectifs - nombres absolus)

b. les fréquences cumulées

2/ Les tableaux de contingence

Les tableaux de contingence (ou d'association) sont des tableaux croisés présentant les fréquences de distribution des observations sur deux ou plusieurs caractères qualitatifs.

3/ Les indices de tendance centrale

Les indices de tendance centrale permettent de résumer en quelque sorte la distribution de la variable en question.

Le mode (Mo) : C'est la valeur de la variable qui présente l'effectif le plus élevé.

La médiane (Md): C'est la valeur de la variable qui partage en deux groupes d'effectifs égaux l'ensemble des individus rangés par ordre de valeurs croissantes ou décroissantes de la variable; c'est la première modalité ou valeur dont la fréquence relative cumulée (Fi) dépasse 0,500.

La moyenne arithmétique: est obtenue en rapportant la somme des valeurs prises par la variable pour chaque observation au nombre total de celles-ci. Pour l'échantillon:

$$\overline{X} = (\Sigma x)/n$$

4/ Les indices de dispersion

La variance: elle exprime la dispersion des notes autour de la moyenne. La variance de l'échantillon est notée s2 et est donnée par la formule suivante:

$$S^2 = (\Sigma x\text{-}M)^2 /n$$

L'écart-type: c'est la racine carrée de la variance; il est utilisé de préférence à la variance quand il s'agit uniquement de décrire une distribution car il est exprimé dans la même unité que celles des valeurs.

5/ le coefficient de dissymétrie

Lorsque la distribution est symétrique, la moyenne et la médiane sont égales (ou presque). Cependant, lorsqu'elle est dissymétrique, la moyenne se déplace plus rapidement que la médiane et ce, dans le sens de l'étalement. Par conséquent, on prend, comme mesure de dissymétrie, la distance entre ces deux mesures de tendance centrale, pondérée par l'écart type.

La formule générale du coefficient de dissymétrie (*CD*) est

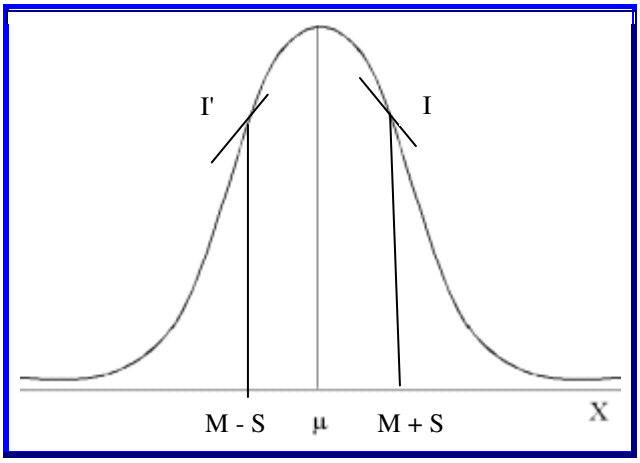
$$CD = \frac{3(\mu - Md)}{\sigma}$$

Le signe de ce coefficient indique le type de dissymétrie (positive ou négative) Ce coefficient est nul lorsque la distribution est symétrique

6/ Distribution et courbe "normale"

Une distribution normale correspond à la distribution de probabilités d'une variable aléatoire continue dont la courbe est parfaitement symétrique, unimodale et en forme de cloche. Elle possède deux points d'inflexion (I et I') situés à égale distance de l'axe de symétrie de la courbe.

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.



Troisième chapitre: comparaisons de moyennes

I- Test T de Student

Ce test permet de comparer deux distributions extraites d'une population normale ou approximativement normale au niveau de leurs moyennes.

Il s'agit de décider si la différence observée entre les moyennes des deux échantillons de comparaison est attribuable à la variable indépendante testée ou si elle peut être considérée comme l'effet du hasard.

1/ Cas 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.

Soit la situation suivante:

Variables:

VI = variable indépendante: nominale dichotomique

VD = Variable dépendante: variable d'intervalle

Hypothèse:

H0: m1 = m2 (c'est-à-dire les deux groupes de comparaison appartiennent à des populations qui possèdent des moyennes identiques)

H1: m1 ≠ m2 (hypothèse bilatérale) ou m1 < m2 ou m1 > m2 (Hypothèses unilatérales)

✓ Conditions d'application

- La distribution des données de chaque échantillon ne peut pas différer fortement de la normale, et, en particulier, ne pas être trop dissymétrique, surtout si les échantillons sont petits
- Les variances des populations de provenance ne peuvent pas être extrêmement différentes
- Les tailles des échantillons ne peuvent pas extrêmement différentes

✓ Algorithme de résolution

$$t = \frac{m_1 - m_2}{\sqrt{Vc(\frac{1}{n_1} + \frac{1}{n_2})}}$$ où

m_1: moyenne du premier échantillon

m_2: moyenne du deuxième échantillon

n_1: nombre de mesures (sujets) du premier échantillon

n_2: nombre de mesures (sujets) du deuxième échantillon

Vc: la variance commune, c'est une sorte de moyenne des deux variances (V_1 et V_2) pondérée par le nombre de mesures n1 et n2; sa formule est:

Si $n_1 \neq n_2$ alors $Vc = \frac{\sum (x_1 - m_2)2 + \sum (x_2 - m_2)2}{n_1 + n_2 - 2}$ cette formule peut être transformée en:

$$Vc = \frac{V_1(n_1 - 1) + V_2(n2 - 1)}{n_1 + n_2 - 2}$$

Si $n_1 = n_2$ alors $Vc = \frac{V_1 + V_2}{2}$

Une fois on a la valeur de t calculé, se rapporter à la table de t de Student pour comparer le "t calculé" au "t critique" , et ce, au ddl = $n_1 + n_2 - 2$ et au seuil 0,05. La différence est significative si "t calculé" est supérieur ou égal au "t critique".

Exemple:

Soit deux groupes de sujets ayant subi une expérience sur la mémoire (retenir une série de mots).

n_1=27 ont obtenu une moyenne de 63,5, une médiane = 63 et un écart-type de 15, 6

n_2= 18 ont obtenu une moyenne de 48,7, une médiane= 49 avec un écart-type de 16,4

Question: y a-t-il une différence entre la performance de deux groupes?

Solution:

Pour pouvoir appliquer t de Student on doit vérifier la normalité de deux distribution, l'homogénéité des variances et calculer la variance commune.

1) Vérification de la normalité des distributions

$$CD_1 = \frac{3(63,5 - 63)}{15,6} = 0,096 \qquad CD_2 = \frac{3(48,7 - 49)}{16,4} = -0,054$$

2) Vérification de l'homogénéité des variances

$F = \frac{16,4^2}{15,6^2} = 1,105$ La valeur critique de F à ddl horizental = 17 (18-1) et ddl vertical = 26 (27-1) égal 1,89 (17 et 26 ne figurent pas sur la table, on prendra les valeurs immédiatement supérieures.

F calculé étant inférieur à F critique, on conclue donc que la différence entre les variance n'est pas significative

3) $Vc = \frac{(15,6)2\times 26 + (16,4)2\times 17}{27+18-2} = 253,48$

Alors $t = \frac{63,5-48,7}{\sqrt{253,48(\frac{1}{27}+\frac{1}{18})}} = \frac{14,8}{4,84} = 3,06$

Nous devons maintenant chercher la valeur critique de t.

ddl = 27+18-2 = 43; au seuil 0,05 t = 2,02 (43 n'existe pas sur la table, on choisira le degré de liberté juste inférieur c'est-à-dire 40).

3,06 étant > 2,02, nous rejetons H0 et nous admettons l'existence d'une différence significative entre m1 et m2.

Exercice d'application: (inspiré de G. Langouet et J-C. Porlier, 1998, p.92)

A l'issue des épreuves de mathématiques du baccalauréat, on a relevé les notes (de 0 à 20) de deux échantillons de sujets:

- le premier G1, ayant réussi cet examen, a eu les mesures suivantes: n_1= 30; m_1=11,2; σ_1=3,5
- le second G2, ayant réussi cet examen, a eu les mesures suivantes: n_2= 25; m_2=12,49; σ_2=4,06

Question: Supposant l'homogénéité vérifiée, existe-t-il entre les deux groupes une différence significative?

Réponse:

$Vc = \frac{(3,5)2\times 29 + (4,06)2\times 24}{30+25-2} = 14,16$

Alors $t = \frac{|11,2-12,49|}{\sqrt{14,16(\frac{1}{30}+\frac{1}{25})}} = \frac{1,29}{1,016} = 1,269$

Nous devons maintenant chercher la valeur critique de t.

ddl = 30+25-2 = 53; au seuil 0,05 t = 2,00 (53 n'existe pas sur la table, on choisira le degré de liberté le plus proche c'est-à-dire 60).

1,269 étant < 2,00, nous retenons H0 et nous admettons l'absence d'une différence significative entre m1 et m2.

2/ Cas de deux échantillons dépendants ou appareillés

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. C'est le cas par exemple de deux séries de notes

relevées auprès d'un échantillon d'élèves, la première avant les vacances et la deuxième à la rentrée; il y a donc une correspondance parfaite puisque c'est le même groupe qui effectue les deux épreuves.

Là encore on va calculer un t qui indique si les deux moyennes sont significativement différentes. La formule sera légèrement modifiée par rapport à la précédente:

$$t = \frac{|m_1 - m_2|}{\sigma_d / \sqrt{N}} \quad \text{ou encore} \quad t = \frac{md}{\sigma_d / \sqrt{N}} \quad \text{où}$$

- md: la moyenne des différences
- σd: l'écart-type de la distribution des différences
- N: le nombre de sujets

Exemple (tiré de S. Ehrlich et C. Flament, 1970, p.122): dans une expérience sur la perception du langage, on fait subir deux épreuves à un même groupe de 40 sujets. On a obtenu les mesures suivantes qui désignent le nombre de mots correctement reproduits:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Epreuve1	3	5	5	7	7	7	4	6	6	7	4	8	5	8	6	8	6	7	6	7
Epreuve2	5	2	4	2	6	3	4	1	3	4	1	3	3	2	5	3	2	7	3	3

Sujets	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Epreuve1	9	7	6	5	7	5	7	9	6	7	6	6	8	4	6	4	8	5	5	8
Epreuve2	3	3	5	2	4	2	6	3	2	4	3	5	2	4	2	4	5	3	4	4

1- On commence par calculer les différences entre les mesures de l'épreuve1 et celles de l'épreuve2 (d), puis les relever au carré (d^2). On obtient la distribution suivante:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
d	-2	+3	+1	+5	+1	-4	0	+5	+3	+3	+3	+5	+2	+6	+1	+5	+4	0	+3	+4
d^2	4	9	1	25	1	46	0	25	9	9	9	25	4	30	1	25	16	0	9	16

Sujets	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
d	+6	+4	+1	+3	+3	+3	+1	+6	+4	+3	+3	+1	+6	0	+4	0	+3	+2	+1	+4
d^2	36	16	1	9	9	9	1	36	16	9	9	1	36	0	16	0	9	4	1	16

2- on calcule ensuite la somme des différences (∑d) et la moyenne de ces différences (m_d)

$$m_d = \frac{\sum d}{N} = \frac{114}{40} = 2{,}85$$

3- on calcule la variance puis l'écart-type de cette distribution des différences (V_d et σ_d)

$$V_d = \frac{\sum d2 - \frac{(\sum d)2}{N}}{N-1} = \frac{474 - \frac{(114)2}{40}}{39} = 3,82$$

L'écart-type des différences = $\sigma_d = \sqrt{V_d} = \sqrt{3,82} = 1,95$

4- On applique la formule de t pour échantillons appareillés

$$t = \frac{md}{\sigma d / \sqrt{N}} = \frac{2,85}{1,95 / \sqrt{40}} = 9,23$$

5- dans la table de t, on cherche la valeur critique au ddl N-1 = 40-1 =39 et au seuil 0,05; on trouve t= 2,02 ce qui est largement inférieur à t calculé, la différence entre les deux moyennes est donc très significative

Exercice d'application: (tiré de D. C. Howell, 1998, p.238)

Pour mesurer l'effet d'une compagne de sensibilisation contre le tabagisme, on a demandé à 15 sujets de noter le nombre moyen de cigarettes fumées par jour durant la semaine qui a précédé et la semaine qui a suivi la campagne; les données sont les suivantes:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Avant	45	16	20	33	30	19	33	25	26	40	28	36	15	26	32
Après	43	20	17	30	25	19	34	28	23	41	26	40	16	23	34

Question: la campagne de sensibilisation a-t-elle réduit le tabagisme?

<u>Réponse</u>

∑d= 5 md= 5/15= 0,333

$$V_d = \frac{\sum d2 - \frac{(\sum d)2}{N}}{N-1} = \frac{117 - \frac{(5)2}{15}}{14} = 8,238 \qquad \sigma_d = \sqrt{8,238} = 2,870$$

$$t = \frac{md}{\sigma d / \sqrt{N}} = \frac{0,333}{2,870 / \sqrt{15}} = 0,029$$

II- L'analyse de la variance (ANOVA)

Supposant que l'on ait fait passer un test de connaissance à tous les élèves des classes de 6ème d'une école. S'il y a quatre classes, les résultats peuvent être résumés comme suit:

classes	1	2	3	4	
Notes (nbre de mesures peut être différent selon les groupes	- - -	- - -	- - -	- - -	
Effectifs	n1	n2	n3	n4	N = n1 + n2 + n3 + n4
Total	T1	T2	T3	T4	Tg = T1 + T2 + T3 + T4
Moyenne	m1	m2	m3	m4	mg = Tg / N

Les deux conditions suivantes sont supposées vérifiées:

a) Condition de normalité c'est-à-dire la variable (note au test) se distribue normalement dans les ensembles parents des 4 classes.

b) Homogénéité des variances. Les ensembles parents des 4 classes ont la même variance

✓ **Principes de solution**

La variance totale de la variable (la note au test) mesure la dispersion de toutes ses valeurs (N valeurs) autour de la moyenne générale (mg) de la distribution de l'ensemble des notes des 4 classes. Cette variance totale dépend de deux sources de variation:

- La variation à l'intérieur des classes dite variance intra-groupe: c'est la dispersion des valeurs de la variable à l'intérieur d'une classe, autour de la moyenne m_i de cette classe. On l'appelle Ve

- la variation entre les classes dite variance inter-groupe: c'est la mesure de la dispersion des moyennes des 4 classes. On l'appelle Vg

✓ **Algorithme de résolution**

1- Calculer la variance totale

$$\frac{\sum (x - mg)_2}{N-1} \quad \text{soit} \quad \frac{\sum x2 - \frac{(Tg)2}{N}}{N-1}$$

2- Calculer la variance inter-groupe

$$\text{Vg} = \frac{\sum nimi - \frac{Tg2}{N}}{K-1} \quad \text{soit} \quad \text{Vg} = \frac{\sum \frac{Ti2}{ni} - \frac{Tg2}{N}}{K-1}$$

2- Calculer la variance intra-groupe

$$\text{Ve} = \frac{\sum x2 - \sum (\frac{Ti2}{ni})}{N-K}$$

- ✓ **Méthode d'analyse de la variance**

Snedecor a étudié le rapport entre la variance inter-groupe et la variance intra-groupe appelé F de Snedecor. Ce rapport est d'autant plus grand que la variance inter-groupe est élevée et que la variance intra-groupe est faible. Et réciproquement, il est d'autant plus faibles que la variance inter-groupe est faible et la variance intra-groupe est élevée

La table F indique la valeur lue en fonction du nombre de degrés de liberté de la variance inter-groupe et du nombre de degrés de liberté de la variance intra-groupe. Si F calculé s'avère inférieur à F critique, on rejette H1; on retient H0 et l'analyse touche sa fin. Par contre si F calculé s'avère supérieu à F critique, on rejette H0; on retient H0 et on poursuit l'analyse.

- ✓ **Où faut-il poursuivre l'analyse?**

Si F prouve des différences entre les moyennes des groupes de comparaison, il ne saurait localiser les lieux de ces différence, il est alors nécessaire de poursuivre l'analyse pour comparer les moyennes des groupes deux à deux en appliquant le test t de Student où la variance intra-groupe remplacera la variance commune.

$$t = \frac{|m1 - m2|}{\sqrt{Ve(\frac{1}{n1} + \frac{1}{n2})}}$$

Exemple: on souhaite comparer trois méthodes d'enseignement pour adultes: une méthode d'enseignement livresque, une méthode d'enseignement audio-visuel et une méthode d'enseignement assisté par ordinateur. Trois groupes de même niveau initial sont constitués. A l'issue de l'enseignement, on relève les performances des sujets des 3 groupes. Elles se résument ainsi:

- Groupe 1.Enseignement livresque: $n_1=18$ $T_1= 198$ $m_1= 11$
- Groupe 1.Enseignement audio-visuel: $n_2=20$ $T_2= 240$ $m_2= 12$
- Groupe 1.Enseignement assisté par ordinateur: $n_3=22$ $T_3= 308$ $m_3= 14$

En outre nous savons que $\sum x^2 = 9967$

Question de recherche: Y a-t-il une différence d'efficacité entre ces trois méthodes?

Solution:

On peut récapituler les calculs nécessaires dans le tableau suivant:

Sources de variation	**Somme des carrés des écarts**	**Nombre de degrés de liberté**	**Carrés moyens (variances)**	**F**
Entre les groupes (inter-groupe)	$\sum(\frac{Ti2}{ni}) - \frac{Tg2}{N}$	K-1	$\frac{\sum\frac{Ti2}{ni} - \frac{Tg2}{N}}{K-1}$	F= $\frac{Vg}{Ve}$
A l'intérieur des groupes (intra-groupe)	$\sum x2 - \sum(\frac{Ti2}{ni})$	N-K	$\frac{\sum x2 - \sum(\frac{Ti2}{ni})}{N-K}$	
Total	$\sum x2 - \frac{Tg2}{N}$	N-1		

Sources de variation	**Somme des carrés des écarts**	**Nombre de degrés de liberté**	**Carrés moyens (variances)**	**F**
inter-groupe	94,73	3-1=2	47,36	4,52
intra-groupe	597,00	60-3=57	10,47	
Total	691,00			

On vérifie que 94,73 + 597 = 691,73

Pour la lecture de F, nous avons K-1 = 2 et N-K = 57. La table du F de Snedecor pour K = 2 et N-K = 40 (valeur inférieure de N-K) à P=0,05 nous donne F=3,23 (à P=0,01, F=5,18)

Nous concluons que, à P=0,05, F calculé est supérieur à F critique (lu dans le tableau), nous pouvons alors affirmer que ces trois méthodes d'enseignement ne sont pas d'égale efficacité.

On poursuit alors l'analyse et on compare les trois groupes deux à deux:

G1 et G2 $t= \frac{|11-12|}{\sqrt{10,47(\frac{1}{18}+\frac{1}{20})}} \cong 0,954$

G2 et G3 $t= \frac{|12-14|}{\sqrt{10,47(\frac{1}{20}+\frac{1}{22})}} \cong 2,00$

G1 et G3 $t= \frac{|11-14|}{\sqrt{10,47(\frac{1}{18}+\frac{1}{22})}} \cong 2,932$

La lecture de t montre que, à P=0,05 et à ddl = 57 (N-K), t critique = 2, 02, nous pouvons donc conclure que:

- La différence entre le groupe 1 et le groupe 2 n'est pas significative;
- La différence entre le groupe 2 et le groupe 3 n'est pas significative;
- La différence entre le groupe 1 et le groupe 3 est significative.

Exercices d'application

Exercice1: Soit les deux distributions suivantes des notes obtenues par les élèves d'une classe:

Filles	8	10	10	10	13	13	13	15	15	15	15	15	15	16	16	16	16	17	17
Garçons	7	7	11	11	12	12	12	14	14	14	14	14	14	14	15	15	15	15	18

Question-problème: y a-t-il une différence entre les notes obtenues par les filles et celles des garçons?

Exercice2:

Question problème suivante: le QI moyen des élèves inscrits dans la classe1 est-il différent de celui des élèves de la classe2?

Soit la distribution suivante: n1 = 12 et n2 = 10

Classe1	90	95	80	100	130	115	135	70	65	105	100	130
Classe2	60	55	80	70	75	110	120	65	55	110		

Exercice3:

Pour mesurer l'efficacité de son enseignement, un enseignant a évalué les performances de ses élèves au début et en fin d'année scolaire. Sur 36 élèves, il a obtenu les notes suivantes:

Sujets	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Début	5	7	15	8	19	12	4	11	10	17	12	13
Fin	9	7	12	9	17	15	6	15	11	16	13	14

Sujets	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Début	16	6	9	16	3	4	9	12	11	17	12	14
Fin	15	9	6	18	5	10	11	10	13	17	13	12

Sujets	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Début	9	15	4	2	6	7	7	9	8	10	14	12
Fin	10	14	8	4	4	9	7	11	9	11	12	15

Question: peut-on affirmer qu'il a eu progrès?

Quatrième chapitre: comparaisons de fréquences

Test de Khi-deux (X^2)

Il permet de comparer deux ou plusieurs groupes caractérisés par une répartition de leurs effectifs respectifs.

1) Cas des échantillons indépendants

✓ **Principes d'application:**

1. ce test n'est applicable que si les catégories sont les mêmes dans les différents échantillons
2. les données doivent être indépendantes d'une colonne à l'autre ou d'une rangée à l'autre (pas d'échantillons appareillés).
3. les groupes doivent avoir une taille suffisante, ce test ne pas être appliqué si 20% ou plus des fréquences attendues sont inférieures à 5, sinon il faut apporter la correction de Yates.

✓ **Algorithme de résolution:**

1. Calculer l'effectif théorique pour chaque case
2. Calculer la statistique khi-deux pour chaque case
3. Faire la somme des khi-deux obtenus
4. Comparer ce résultat avec la valeur tabulaire correspondant au seuil de signification choisi et au nombre de degré de liberté que comporte la situation. Si le résultat est supérieur ou égal à cette valeur, alors on rejette H0

Soit une variable nominale trichotomique VA formée de 2 modalités: a1 et a2

Soit une variable ordinale de catégories rangées VB à 3 modalités: b1; b2 et b3

1/Dresser le tableau des effectifs observés

	b1	**b2**	**b3**	**Total**
a1	n_1	n_2	n_3	L1
a2	n_4	n_5	n_6	L2
Total	C1	C2	C3	**N**

2/ Calculer les effectifs théoriques (appelés également attendus)

	b1	**b2**	**b3**	**Total**
a1	n'_1	n'_2	n'_3	L1
a2	n'_4	n'_5	n'_6	L2
Total	C1	C2	C3	**N**

L: Total ligne

C: Total colonne

N: Effectif total

	b1	**b2**	**b3**
a1	$n'_1 = \frac{C1 \times L1}{N}$	$\frac{C2 \times L1}{N} n'_{2=}$	$\frac{C3 \times L1}{N} n'_{3=}$
a2	$\frac{C1 \times L2}{N} n'_{4=}$	$\frac{C2 \times L2}{N} n'_{5=}$	$\frac{C3 \times L2}{N} n'_{6=}$

3/ Calculer le Khi-deux des cases

Pour chaque case, appliquer la formule suivante: *(son effectif observé – son effectif théorique)^2 / sur son effectif théorique*

	b1	**b2**	**b3**
a1	$\frac{(n_1 - n'_1)2}{n'_1}$	$\frac{(n_2 - n'_2)2}{n'_2}$	$\frac{(n_3 - n'_3)2}{n'_3}$
a2	$\frac{(n_4 - n'_4)2}{n'_4}$	$\frac{(n_5 - n'_5)2}{n'_5}$	$\frac{(n_6 - n'_6)2}{n'_6}$

Si 20% au plus des effectifs théoriques sont inférieurs à 5, on apporte la correction de Yates et la formule devient: *(|son effectif observé – son effectif théorique| - 0.5)^2 / sur son effectif théorique*

4/ Calculer le Khi-deux (la somme de chacune des cases de l'étape précédente).

$$X^2 = \frac{(n_1 - n'_1)2}{n'_1} + \frac{(n_2 - n'_2)2}{n'_2} + \frac{(n_3 - n'_3)2}{n'_3} + \frac{(n_4 - n'_4)2}{n'_4} + \frac{(n_5 - n'_5)2}{n'_5} + \frac{(n_6 - n'_6)2}{n'_6}$$

5/ Déterminer les degrés de liberté de Khi-deux en appliquant la formule suivante:

(Nombre de colonnes – 1) (Nombre de lignes – 1)

6/ La valeur de Khi-deux calculée doit être comparée à la valeur critique de Khi-deux (sur la table) au seuil 0,05: si Khi-deux calculé est supérieur au Khi-deux théorique, on considère la différence significative, c'est-à-dire qu'il y a influence de la variable indépendante sur la variable dépendante.

Remarque: cette procédure est générale, qu'il s'agisse d'un tableau à quatre cases ou plus.

Exemple:

Question problème: le choix de la filière dépend t-il de la catégorie socioprofessionnelle?

Réponse:

H0: Il n'y a pas effet de la catégorie socioprofessionnelle sur le choix de la filière

H1: La catégorie socioprofessionnelle influence le choix de la filière.

1/ Effectifs observés

	Maths	Lettres	Sciences	Technique	Total
Très défavorisée	7	2	1	3	13
Défavorisée	6	3	3	0	12
Moyenne	4	3	3	2	12
Favorisée	5	1	6	1	13
Total	22	9	13	6	**50**

2/ Effectifs théoriques

	Maths	Lettres	Sciences	Technique	Total
Très défavorisée	5.72	2.34	3.38	1.56	13
Défavorisée	5.28	2.16	3.12	1.44	12
Moyenne	5.28	2.16	3.12	1.44	12
Favorisée	5.72	2.34	3.38	1.56	13
Total	22	9	13	6	**50**

3/ Khi-2 des cases

Puisque 12 fréquences théoriques sur 16 sont inférieures à 5, on applique la correction de Yates on obtient le tableau suivant:

	Maths	Lettres	Sciences	Technique
Très défavorisée	0.10	0.01	1.04	0.56
Défavorisée	0.009	0.05	0.04	0.61
Moyenne	0.11	0.05	0.04	0.0025
Favorisée	0.008	0.3	1.32	0.002

4/ Khi-deux = 4.25

5/ Degrés de liberté = (4 -1)(4 -1) = 9

6/ 4.25 <16.9, donc H0 est retenue. La catégorie socioprofessionnelle n'a pas d'effet sur le choix de la filière.

2) Cas des échantillons dépendants

Il s'agit de comparer un tableau de fréquences construit sur des dichotomies (fréquences recueillies auprès d'un seul échantillon à des moments différents ou dans deux situations différentes).

Supposant, par exemple que l'on veuille étudier la différence entre le nombre d'élèves accédant à deux types de formation

Formation A

Formation B

	Admis	Refusés	Total
Admis	n_1	n_2	N1
Refusés	n_3	n_4	N2
Total	N3	N4	N

Ce sont les mêmes candidats (ayant participé à l'examen de la formation A et l'examen de la formation B), on veut comparer la proportion des admis à la première formation avec la proportion des admis à la deuxième formation), c'est-à-dire les fréquences:

$p_1 = \frac{N3}{N}$ et $p_2 = \frac{N1}{N}$

Pour ce faire, on calcule un X^2 assez différent du précédent:

$$X^2 = \frac{(n_2 - n_3)2}{n_2 + n_3}$$

Remarquons que cette formule ne s'intéresse qu'aux effectifs des cases hétérogènes (admis à une formation et refusés à une autre).

Exemple (tiré de S. Ehrlich et C. Flament, 1970, p.158):

On a posé à 300 personnes deux questions: "allez-vous souvent au cinéma?" et "allez-vous souvent au théâtre?". Les réponses sont "oui" ou "non". On observe les résultats suivants:

		Cinéma		
		oui	**non**	**Total**
Théâtre	**oui**	$n_1 = 42$	$n_2 = 48$	N1= 90
	non	$n_3 = 78$	$n_4 = 132$	N2= 210
	Total	N3= 120	N4= 180	N= 300

- 42 personnes vont souvent au cinéma **et** au théâtre;
- 78 personnes vont souvent au cinéma **et rarement** au théâtre;
- 120 personnes vont souvent au cinéma;
- 90 personnes vont souvent au théâtre

La question: la différence entre ces deux nombres est-elle significative?

Réponse: $X^2 = \frac{(n2 - n3)2}{n2 + n3} = \frac{(48 - 78)2}{48 + 78} = \frac{900}{126} = 7{,}14$

La table du X^2 pour 1 ddl et 0,05 probabilité d'erreur donne 3,84; la différence est dons significative.

Exercices d'application

Exercice1:

Question: le degré de motivation à la lecture est –il fonction du sexe?

	+ motivé	**± motivé**	**- motivé**	**Total**
Fille	12	7	8	27
Garçon	10	8	5	23
Total	22	15	13	50

Exercice2:

165 élèves passent deux épreuves scolaires. 63 d'entre eux réussissent les deux épreuves; 32 réussissent à la première et échouent à la seconde; 12 échouent à la première et réussissent à la seconde; 58 échouent aux deux.

L'une de deux épreuves est-elle plus difficile que l'autre?

Exercice3:

A l'issue d'un examen, on relève, en fonction du genre, les résultats de 164 élèves. ils se répartissent ainsi:

Résultats / Sexe	Admis définitivement	Admis à l'oral	Eliminés	Total
Garçons	32	28	10	70
Filles	48	34	12	94
Total	80	62	22	164

Question: la réussite varie-telle en fonction du genre?

Exercice4

Au début de l'année scolaire, on a classé 302 enfants suivant qu'ils manifestent un déficit de l'attention ou pas. A la fin de la même année, on a examiné leurs bulletins scolaires et on a trouvé les résultats suivants:

	Mauvaise performance	Bonne performance	Total
Normaux	22	187	209
Déficit de l'attention	19	74	93
Total	41	261	303

Question: l'état des élèves au début de l'année influence-t-il leur performance?

Cinquième chapitre: étude de la relation entre deux variables

Se sont les techniques qui permettent d'étudier la relation qui pourrait exister entre deux variables quantitatives X et Y:

- Corrélation positive, c'est-à-dire à toute augmentation au niveau de X correspond une augmentation au niveau de Y. Les deux variables varient dans le même sens et avec une intensité similaire. Exemple: la taille et le poids
- Corrélation négative, c'est-à-dire à toute augmentation au niveau de X correspond une diminution au niveau de Y. Les deux variables varient dans deux sens opposés et avec une intensité similaire. Exemple: l'absentéisme et la réussite scolaire
- Corrélation faible ou nulle, c'est-à-dire X et Y sont indépendantes l'une de l'autre. Exemple: la pluviométrie et la réussite scolaire

Le coefficient de corrélation qui exprime cette relation varie entre **-1** (négative) et **+1** (positive) passant par **0** (nulle)

1- R de Bravais-Pearson

✓ **Principes d'application:**

- L'échantillon doit être tiré aléatoirement
- Les deux distributions doivent être normalement distribuées.

✓ **Algorithme de résolution:**

Deux distributions de scores X et Y

1/ calculer somme X (ΣX) et somme Y (ΣY), puis relever au carré $(\Sigma X)^2$ et $(\Sigma Y)^2$

2/ calculer X x Y et leur somme (ΣXY)

3/calculer X^2et Y^2 et leurs sommes ΣX^2 et ΣY^2

4/ calculer la covariance de XY

$$\text{cov}_{XY} = \sum XY - \frac{\sum X \sum Y}{N}$$

5/ calculer $r = \dfrac{\text{cov}\, XY}{\sqrt{(\sum X2 - \frac{(\sum X)2}{N})(\sum Y2 - \frac{(\sum Y)2}{N})}}$

6/ consulter la table des valeurs critiques de r de Spearman pour voir si la corrélation peut être considérée significative

Exemple

			Etape 2	**Etape 3**	
Sujet	VariableX	VariableY	X x Y	X^2	Y^2
1	30	99	2970	900	9801
2	27	94	2538	729	8836
3	9	80	720	81	6400
4	20	70	1400	400	4900
5	3	100	300	9	10000
6	15	109	1635	225	11881
7	5	62	310	25	3844
8	10	81	810	100	6561
9	23	74	1702	529	5476
10	34	121	4114	1156	14641
Total	176	890	**16499**	**4154**	**82340**
	Etape1				

Etape4: calcul de la covariance

$$\sum XY - \frac{\sum X \sum Y}{N} = 16499 - \frac{176 \times 890}{10} = 835$$

Etape5: calcul de r

$$r = \frac{\operatorname{cov} XY}{\sqrt{(\sum X2 - \frac{(\sum X)2}{N})(\sum Y2 - \frac{(\sum Y)2}{N})}} = \frac{835}{\sqrt{(4154 - \frac{30976}{10})(82340 - \frac{792100}{10})}} = \mathbf{0.46}$$

2- Rô de Spearman (r_s)

✓ **Conditions d'application:**

Il est utilisé quand les données sont présentées sous la forme de rangs, de classements.

✓ **Algorithme de résolution**

1/ ranger les scores de X par ordre croissant et leur faire correspondre les scores obtenus au niveau de Y; c'est-à-dire pour chaque sujet il faut avoir un paire de score (x, y).

2/ attribuer un rang, de 1 à N, à chaque score de X

3/ attribuer un rang, de 1 à N, à chaque score de Y

4/ calculer pour chaque paire la différence (d) entre les rangs de X et de Y; puis relever au carré (d^2)

5/ calculer la somme des différences (Σd^2)

6/ calculer le coefficient de corrélation de Spearman

$$r_s = 1- \frac{6\times\sum d2}{N(N2-1)}$$

7/ consulter la table des valeurs critiques de r de Spearman pour voir si la corrélation peut être considérée significative

Exemple:

Q-P: existe-t-il une relation entre les résultats au test de mathématiques et ceux du test de physique?

Tableau de données (déjà rangées pour V1)

Sujet	1	2	3	4	5	6	7	8
V1	6	8	9	10	13	14	18	19
V2	14	12	13	9	11	12	19	20

Réponse

	Etape1		**Etape2**	**Etape3**	**Etape4**	
Sujet	Sc_Maths	Sc_Physique	R_maths	R_physique	d	d^2
1	6	14	1	6	-5	25
2	8	12	2	3.5	-1.5	2.25
3	9	13	3	5	-2	4
4	10	9	4	1	+3	9
5	13	11	5	2	+3	9
6	14	12	6	3.5	+2.5	6.25
7	18	19	7	7	0	0
8	19	20	8	8	0	0

Etape5: $\Sigma d^2 = 55.5$

Etape6: $r_s = 1- \frac{6\times 55.5}{8(8-1)} = 0.34$

Etape7: ddl = N-1 = 7 au seuil 0,05 la valeur de r_s est .67.

Donc r_s calculé < au r_s de la table, H0 est à retenir c'est-à-dire il n'y a pas de corrélation significative entre les scores en mathématiques et ceux en physique

Exercices d'application:

Exercice1:

Voici le tableau de données obtenues d'une recherche sur le rendement des élèves de terminales:

Sujets	V1	V2	V3	V4
1	12	13	100	105
2	13	15	115	95
3	4	3	80	100
4	16	18	120	110
5	17	20	135	130
6	7	3	85	80
7	10	11	110	100
8	19	15	130	120
9	15	13	120	125
10	12	10	110	115
11	8	4	125	120
12	13	13	105	95
13	11	9	95	105
14	8	7	90	100
15	2	5	80	85

V1: scores obtenus à un test de langue **V2**: scores obtenus à un test d'expression écrite

V3: QI obtenus au test de la WISC **V4**: QI obtenus des cubes de Kohs

Questions:

1/ Y a-t-il une relation entre le rendement des élèves en langue et leur rendement en expression écrite?

2/ Existe-t-il une relation entre les résultats de QI obtenus au test de la WISC et ceux obtenus des cubes de Kohs?

Exercice2

Un chercheur s'intéresse à la relation entre l'absentéisme et les troubles relationnels à l'école. Il a rassemblé les donnés suivantes:

Sujets	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Jours d'absence	19	2	7	17	12	10	11	8	15	13	12	4	13	16	8
Troubles	17	7	3	20	13	11	13	9	14	15	10	3	6	18	4

Question: Quelle est la relation entre les deux variables étudiées? La relation est-elle significative?

Sixième chapitre: les tests non paramétriques

Il existe de tests moins "exigeants" en conditions d'applications, notamment en ce qui concerne la taille de l'échantillon, la normalité de la distribution et l'égalité des variances**,** ces tests sont dits non paramétriques. Le principe de base de ces tests est de transformer les données en rangs et à mesurer l'accord entre les rangs observés et ce que devrait être ces rangs sous une hypothèse nulle. Parmi ces tests nous allons voir:

- le test de Mann-Wihtney, l'alternative non paramétrique de t de Student pour deux échantillons indépendants;
- le test de Wilcoxon, l'alternative non paramétrique de t de Student pour deux échantillons dépendants;
- le test de Kruskal-Wallis, l'alternative non paramétrique de l'analyse de variance.

I- U de Mann-Withney

✓ Principes d'application

Ce test est destiné à étudier si une variable indépendante nominale dichotomique influence une variable dépendante ordinale de scores rangés ou d'intervalle.

Ce test doit être préféré au test t de student lorsque la distribution n'obéit pas à la loi normale (donc remarquablement dissymétrique)

✓ Algorithme de résolution

1er cas: n_A et n_B sont supérieurs à 8

Supposant les donnés suivantes:

A ($n_A = 12$)	11	9	7	12	12	40	5	4	15	10	10	14
B ($n_B = 11$)	13	15	15	14	35	18	13	25	20	6	5	

1/ Transformer les scores en rangs

- Mélanger les donnés de deux groupes
- Ordonner la série obtenue en ordre croissant
- Accorder des rangs; pour les ex-æquo attribuer à chacun le rang moyen
- Reconstruire les deux groupes avec données et les rangs correspondants

Ce qui donnera la répartition suivante:

A	Scores	11	9	7	12	12	40	5	4	15	10	10	14
	Rangs	*9*	*6*	*5*	*10.5*	*10.5*	*23*	*2.5*	*1*	*17*	*7.5*	*7.5*	*14.5*
B	Scores	13	15	15	14	35	18	13	25	20	6	5	
	Rangs	*12.5*	*17*	*17*	*14.5*	*22*	*19*	*12.5*	*21*	*20*	*4*	*2.5*	

2/ calculer la somme des rangs de A et la somme des rangs de B
$T_A = 114$ et $T_B = 162$

4/ Calculer $U = n_A n_B + \frac{n_A(n_A+1)}{2} - T_A$ $\quad U' = n_A n_B + \frac{n_B(n_B+1)}{2} - T_B$

5/ Mann et Whitney ont montré que la variable U se distribue selon une loi approximativement normale. Calculer donc:

la moyenne de la distribution de U est: $m_U = \frac{n_A \times n_B}{2}$ et

l'écart-type est: $\sigma_U = \sqrt{\frac{(n_A \times n_B)(n_A + n_B + 1)}{12}}$

6/ Il en est de même pour U', valeur symétrique de U. il suffit donc de tester l'écart entre U et m_U (ou entre m_U et U'), soit:

$$|Z| = \frac{|U - m_U|}{\sigma_U}$$

Si nous revenons à notre exemple, nous aurons donc:

$U = (12 \times 11) + \frac{12(12+1)}{2} - 114 = 96$ $\quad U' = (12 \times 11) + \frac{11(11+1)}{2} - 162 = 36$

$mU = \frac{12 \times 11}{2} = 66$; $\quad \sigma_U = \sqrt{\frac{(12 \times 11)(12+11+1)}{12}} = 12{,}25$

On peut vérifier que $\frac{U + U'}{2} = \frac{96 + 66}{2} = \frac{132}{2} = 66 \ (= m_U)$

Par conséquent $|Z| = \frac{|96 - 66|}{16{,}25} \approx 1{,}85$

7/ Vérifier la signification de la valeur Z:

- Si Z calculé est supérieur ou égal à 1,96; la différence est significative au P=0,05
- Si Z calculé est supérieur ou égal à 2,56; la différence est significative au P=0,01

1er cas: n_A et n_B sont inférieurs à 8

Dans ce cas, la distribution n'est pas gaussienne, le modèle précédent ne peut pas être appliqué.

Mann et Withney ont construit des tables avec des valeurs critiques qu'il est possible de consulter directement en fonction de:

- de **U** si U est inférieur à U'
- de **U'** si U' est inférieur à U

Supposons les mesures de deux groupes et leurs rangs:

A	Scores	6	3	10	5	14	n_A=5
	Rangs	*7*	*9*	*5*	*8*	*3*	T_A= 32
B	Scores	12	8	16	18		n_B= 4
	Rangs	*4*	*6*	*2*	*1*		T_B= 13

$$U = 5\times 4 + \frac{5(5+1)}{2} - 32 = 3$$

$$U' = 5\times 4 + \frac{4(4+1)}{2} - 13 = 17$$

La table est consultée en fonction de l'effectif n2 du plus grand de deux échantillons (ici, n2 =5). Pour n1=4 et U=3, nous lisons P=.056

L'hypothèse nulle n'est pas rejetée, il n'existe pas une différence entre les moyennes des rangs.

Remarque: pour simplifier les calculs on prendra toujours la somme des rangs dans la situation comportant le moins de sujets. Lorsqu'il y aura le même nombre de sujets dans les deux conditions, il sera possible de prendre l'une ou l'autre des deux conditions pour calculer la somme des rangs.

Exemple

X	**Y**	**Rx**	**Ry**
3	15	1	12.5
5	18	2	15
12	15	9.5	12.5
10	12	7	9.5
13	7	11	3
8	10	4.5	7
17	10	14	7
	8		4.5
nx=7	ny=8	Rx = 49	Ry = 71

$$U = 7 \text{ x } 8 + \frac{7(7+1)}{2} - 49 = 35 \qquad U' = 7 \text{ x } 8 + \frac{8(8+1)}{2} - 71 = 21$$

Nous avons n2=8. Pour n1=7 et U=21, nous lisons P=.198; La différence entre les moyennes des rangs est donc non significative

II- Le test de Kruskal-Wallis

C'est la généralisation du test de Mann-Withney à trois échantillons ou plus . Les scores sont remplacés par les rangs obtenus à l'intérieur d'un seul groupe constitués à partir des échantillons à comparer.

Supposant 4groupes de sujets reçoivent un enseignement selon quatre méthodes différentes. On souhaite comparer leurs résultats sur la base des données suivantes:

Groupes	1	2	3	4
Scores	8	15	18	4
	20	14	16	7
	13	7	15	12
	14	9	19	10
	17	12		8
		10		6
				11
Effectifs	n_1=5	n_2=6	n_3=4	n_4=7

- On mélange les 4 groupes (k=4) et on ordonne les scores:

4 - 6 - 7 - 7 - 8 - 8 - 9 - 10 - 10 - 11 - 12 - 12 - 13 - 14 - 14 - 15 - 15 - 16 - 17 - 18 - 19 - 20

Groupe1		Groupe2		Groupe3		Groupe4	
Scores	Rangs	Scores	Rangs	Scores	Rangs	Scores	Rangs
8	5.5	15	16.5	18	20	4	1
20	22	14	14.5	16	18	7	3.5
13	13	7	3.5	15	16.5	12	11.5
14	14.5	9	7	19	21	10	8.5
17	19	12	11.5			8	5.5
		10	8.5			6	2
						11	10
T_1= 74 m_1= 14.8		T_2=61,5 m_2=10,25		T_3=75.5 m_2=18.875		T_4=42 m_4=6	

Notons que ∑Ti (somme des totaux des rangs) = $\frac{N(N+1)}{2}$

$74+61.5+75.5+42 = \frac{22(22+1)}{2} = 253$

- On applique la formule de Kruskal et Wallis:

$H=\left[\frac{12}{N(N+1)}\times\frac{Ti2}{ni}\right]-3(N+1)$ donc H=

$$\frac{12}{22(22+1)}\times(\frac{(74)2}{5}+\frac{(61.5)2}{6}+\frac{(75.5)2}{4}+\frac{(42)2}{7})-3(22+1)=11,64$$

Cette variable H suit une loi de X^2. Il suffit donc de revenir à la table de X^2 et de comparer H calculé à la valeur critique de X^2 au ddl = k – 1 (c'est-à-dire nombre de groupes – 1).

K = 4; k – 1= 3. La valeur critique de X^2 au ddl = 3 et P=0,05 est égale à 7,82. La valeur calculée est supérieure à la valeur théorique, on rejette donc H0. Autrement dit il existe des différences entre les moyennes des rangs des 4 groupes.

III- Le test de Wilcoxon

Il permet la comparaison les moyennes des rangs de deux échantillons appareillés. Son principe consiste à classer les sujets dans l'ordre croissant des valeurs absolues des différences non nulles.

Supposons les données suivantes portant sur les notes (variant de 0 à 10) obtenues par un groupe d'élèves à deux moments différents de l'année scolaire:

Eléves	a	b	c	d	e	f	g	h	i	j
Première note (A)	1	1	2	2	7	2	3	6	4	5
Deuxième note (B)	9	6	9	2	5	7	8	8	7	4

1/ calculer pour chaque élève la différence entre la première et la deuxième note

d=B-A

ce qui donnera la distribution suivante

Eléves	a	b	c	d	e	f	g	h	i	j
d	+8	+5	+7	0	-2	+5	+5	+2	+3	-1

Nous constatons que:

- Un élève n'a ni régressé ni progressé (d=0)
- Deux élèves ont régressé (d négative)
- Sept élèves ont progressé (d positive)

2/ Classer les sujets dans l'ordre croissant des valeurs absolues des différences non nulles c'est-à-dire dans cet exemple 9 valeurs (on élimine l'élève d car sa différence =0)

Ce qui donnera le classement suivant:

d	1	2	3	5	7	8
sujets	j	e - h	i	b – f - g	c	a
rangs	1	2,5	4	6	8	9

3/ calculer la sommes des rangs des différences positives (T^+)

calculer la sommes des rangs des différences négatives (T^-)

Dans notre exemple T^+ est la somme des différences des sujets a, b, c, f, g, h et i. T^- est la somme des différences des sujets e et j. nous aurons donc:

$T^+ = 9 + 6 + 8 + 6 + 6 + 2{,}5 + 4 = 41{,}5$

$T^- = 2{,}5 + 1 = 3{,}5$

Notons que $T^+ + T^- = \frac{n(n+1)}{2}$ $\qquad 41{,}5 + 3{,}5 = \frac{9(9+1)}{2} = 45$

4/ tester la plus petite valeur des T^+ ou T^-

1) Cas où le nombre de couples dont les différences non nulles est inférieur ou égal à 20 (n≤20)

La distribution de T n'est pas normale; la valeur théorique de T est tabulée et on rejette H0 si T Calculé (T+ ou T-) est inférieur à T lu sur la table.

Dans notre exemple, n = 9 et T calculé = 3,5 (nous avons pris T- parce qu'elle est plus petite que T+); à P = 0,05 T théorique = 6. L'hypothèse nulle est alors rejetée.

2) Cas où le nombre de couples dont les différences non nulles est supérieur à 20 (n>20)

La distribution de T tend vers une distribution normale.

Sa moyenne est: $m_T = \frac{n(n+1)}{4}$

Son écart-type est: $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

On calcule alors $|Z| = \frac{|T - m_T|}{\sigma_T}$

La valeur calculée sera comparée à la valeur critique de Z (table de la loi normale réduite). L'hypothèse nulle est rejetée si la valeur calculée de Z est supérieur à la valeur lue à un seuil donné.

Exemple:

Les données suivantes représentent les notes obtenues par un groupe d'élèves avant et après les vacances. On voulait vérifier l'hypothèse d'une déperdition des acquis:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Avant	10	12	12	19	5	13	20	8	12	10	8	19	5	11	8	7	4	7	16	2	5
Après	8	10	8	18	8	7	12	10	7	10	3	12	8	11	5	3	5	7	9	8	14

Réponse:

- On commence par calculer d = B-A, ce qui donnera:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
d	2	2	4	1	-3	6	8	-2	5	0	5	7	-3	0	3	4	-1	0	7	-6	-9

- On classe les sujets dans l'ordre croissant des valeurs absolues des différences non nulles; on obtient la distribution suivante:

Sujets	1	2	3	4	5	6	7	8	9	11	12	13	15	16	17	19	20	21
I d I	1	1	2	2	2	3	3	3	4	4	5	5	6	6	7	7	8	9
Rd	1.5	1.5	4	4	4	7	7	7	9.5	9.5	11.5	11.5	13.5	13.5	15.5	15.5	17	18

- On calcule T- et T+

  $T^- = 51$ et $T^+ = 120$

On peut vérifier que $T^- + T^+ = \frac{n(n+1)}{2} = 51 + 120 = \frac{18(18+1)}{2} = 171$

- On calcule $m_T = \frac{n(n+1)}{4} = \frac{18\times 19}{4} = 85{,}5$
- On calcule $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{(18\times 19)(36+1)}{24}} = \sqrt{527{,}25} = 22{,}961$
- On calcule $Z = |Z| = \frac{|T - m_T|}{\sigma_T} = \frac{|51 - 85.5|}{22{,}961} = 1{,}502$
- On compare la valeur calculée de Z à la valeur lue sur la table de la loi normale réduite; Z lue au P=0,05 égale 0,121, on rejette alors l'hypothèse nulle; il y a une différence entre les scores des élèves avant les vacances et ceux d'après.

Exercices d'application

Exercice1:

On a relevé les notes attribuées par trois enseignants à un groupe d'élèves:

Enseignant1	82	71	56	58	63	64	62	53
Enseignant2	55	88	85	83	71	70	68	72
Enseignant3	65	54	66	68	72	78	65	73

Question: existe-t-il une différence significative entre ces trois séries de notes?

Exercice2:

Dans ce qui suit les scores obtenus à une échelle d'indépendance par deux groupes d'adolescents provenant de deux milieux différents: un milieu favorisé culturellement et l'autre défavorisé:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
G1	12	18	13	17	8	15	16	5	8	12	13	5	14	20	19	17	2	5	15	18	7
G2	10	12	15	13	9	12	13	8	10	8	8	9	8	10	14	11	7	7	13	12	5

Question: le milieu a-t-il une influence sur le niveau d'autonomie des adolescents?

Exercice3: (tiré de G. Langouet et J-C Porlier, 1991, p.147)

Pour vérifier l'influence de l'origine socio-professionnelle sur la réussite à un test, on a constitué au hasard trois groupes d'enfants de même age.

- Groupe A: 7 enfants de parents ouvriers
- Groupe B: 10 enfants de parents cadres moyens
- Groupe C: 8 enfants de parents cadres supérieurs

Les notes ont été les suivantes:

A: 20 - 18 - 22 - 17 - 15 - 23 - 16

B: 25 - 20 - 32 - 19 - 15 - 35 - 30 - 18 - 25 - 24

C: 39 - 28 - 36 - 30 - 38 – 33- 40 – 35

Consigne: comparer ces résultats sachant que l'hypothèse de normalité et l'hypothèse d'homogénéité des variables ont été rejetées.

Bibliographie

- De Ketele, J.M. et Bonhivers, B. (1986), *Pratique de la statistique. Bruxelles*: De Boeck

- De Ketele, J-M. et Roegiers, X (1993). *Méthodologie du recueil d'informations*. Bruxellles : De Boeck

- Elrich, S. et Flament, C. (1966), Précis de statistique, Paris: PUF

- Grégoire, J et Laveault, D. (1997), *Introduction aux théories des tests en sciences humaines*, Bruxelles: De Boeck

- Howell, D C. (1998), Méthodes *statistiques en sciences humaines*, Bruxelles: De Boeck

- Jones, R A. (2000). *Méthodes de recherche en sciences humaines*. Bruxelles : De Boeck

- Lacoure, L. et al. (1996). *Méthodologie de la recherche en sciences humaines, une initiation par la pratique*. Fascicule : La problématique. Québec : Bibliothèque Nationale.

- Mace, G. (1988).*Guide d'élaboration d'un projet de recherche*. Bruxelles : De Boeck

- Mengal, P. (1984), *Statistique descriptive*, Berne: Peter Lang

- Mialaret, G. (1991), *Statistiques appliquées aux sciences humaines*, Paris: PUF

- Murray, R. et Spiegel (1981) *Probabilités et statistique, Cours et problèmes*. Série Schaum. Ed groupe McGraw Hill. Paris.

- Reuchlin, M. (1976), *Précis de statistique*. Paris: PUF

- عبد الرحمان عدس (1980). مبادئ الإحصاء. الأردن. منشورات النهضة الإسلامية

- فؤاد أبو حطب و آمال صادق (1996). مناهج البحث وطرق التحليل الإحصائي في العلوم النفسية والتربوية والاجتماعية. القاهرة: مكتبة الأنقلومصرية.