

Intitulé du cours : Production et Traitement des données quantitatives
Code cours : 4PAKC011 Nom de l'enseignant : Jean-Claude Régnier
Une calculatrice scientifique est indispensable

Document n°2

Situations Problèmes & Documents autocorrectifs

Compléments théoriques

Situation Problème n°2.1

Document n° 1 : situation de l'étude

La présente situation s'inspire des travaux d'une thèse de doctorat en sciences de l'éducation, intitulée CONSTRUÇÃO DE CONHECIMENTOS EM SITUAÇÕES DE EXCLUSÃO SOCIAL E QUESTÕES DE GÊNERO, préparée en cotutelle par Valdir Pretto (*voir* site cyberthèses Lyon2) et soutenue en août 2009 au Brésil. La thématique centrale porte les connaissances alternatives construites par des sujets confrontés à des situations d'exclusion sociale, pour survivre. Un échantillon de 103 unités a été extrait d'une population habitant dans une zone périphérique d'une ville du Sud du Brésil. Cet échantillon a été constitué en deux étapes : tirage aléatoire des rues, tirage aléatoire des foyers. Une unité est alors le ou les sujets qui ont répondu au questionnaire qui a été administré par le chercheur. Dans un premier temps une enquête par questionnaire a été réalisée afin de recueillir des informations relatives aux individus mais aussi pour identifier les problèmes qu'ils estimaient être ceux qu'ils rencontraient et dont les résolutions nécessitaient des connaissances. D'autres méthodes de construction ont été mises en oeuvre fondées sur des entretiens audio-vidéographiés individuels puis collectifs. En ce qui concerne les problèmes, 14 catégories furent identifiées. Le questionnaire comportait une vingtaine de grandes questions portant sur les caractéristiques individuelles et sociales des sujets, la situation familiale, la composition de la famille, les ressources, le logement, la formation de chacun des membres de la famille, etc.. Dans le cadre de cette étude, nous ne retenons ici que quelques questions que nous présentons ci-dessous. Les tableaux statistiques fournis ensuite contiennent des valeurs réelles mais aussi des valeurs fictives adaptées à la situation d'étude présente.

<i>code de la variable</i>	<i>descriptif succinct</i>
[V1]	Sexe : 1_homme 2_femme
[V2]	Age à la date de l'enquête du sujet qui répond aux questions
[V3]	Lieu de naissance
[V4]	Ancienneté de résidence dans la ville
[V5]	Ancienneté de résidence dans le quartier
[V6]	Etat civil : 1_Célibataire – 2_Marié(e) – 3_Séparé(e) – 4_Divorcé(e) – 5_Vit avec les parents – 6_Veuf(ve)
[V7]	Nombre de personnes qui vivent dans le foyer
[V8]	Ressources du foyer
[V9]	<i>Savez-vous lire et écrire ?</i> 1_Pas du tout ; 2_Un peu ; 3_Moyennement ; 4_Asez facilement ; 5_Très facilement
[V10]	Niveau de scolarité : 1_Analphabète – 2_Semi-analphabète – 3_Enseignement fondamental incomplet – 4_ Enseignement fondamental complet – 5_ Enseignement moyen (lycée) incomplet – 6_ Enseignement moyen (lycée) complet – 7_ Enseignement supérieur incomplet – 8_ Enseignement supérieur complet
[V11]	<i>Quels sont les problèmes que vous rencontrez actuellement là où vous habitez ?</i>
[V12]	<i>Comment vous affrontez, résolvez les problèmes que vous rencontrez là où vous habitez ?</i>
[V13]	Nombre de catégories de problèmes évoquées

Nous rapportons ci-dessous les tableaux statistiques des variables

Tableau statistique de la variable V1

(Tableau n° T1) V1=Sexe	Homme	Femme
Échantillon	34	69

Tableau statistique de la variable V4

(Tableau n° T2) V4= Temps de résidence dans la ville (années)	[1 ; 10[[10 ; 19[[19 ; 30[[30; 35[[35; 50[
Échantillon	19	44	22	9	9

Tableau statistique de la variable V6

(Tableau n° T3) V6=Etat civil	1_Célibataire	2_Marié(e)	3_Séparé(e)	4_Divorcé(e)	5_Vit avec les parents	6_Veuf(ve)
Échantillon	6	82	1	8	1	5

Tableau statistique de la variable V9

(Tableau n° T4) V9=Lire et Écrire Échantillon (valeurs estimées)	1_Pas du tout	2_Un peu	3_Moyennement	4_Asez facilement	5_Très facilement.
	8	5	54	24	12

Tableau des catégories de problèmes identifiés

(Tableau n° T5) codages des problèmes	Nombre d'évocations	Catégories	
Pb01	86	Santé	Saúde
Pb02	3	Alimentation	Alimentação
Pb03	3	Économie familiale	Economia familiar
Pb04	2	Situation familiale	Situação familiar
Pb05	8	Travail	Trabalho
Pb06	68	Éducation, formation scolaire	Educação
Pb07	93	Sécurité	Segurança
Pb08	91	Loisir	Lazer
Pb09	8	Questions religieuses	Religiosidade
Pb10	28	Transport	Transporte
Pb11	15	Recyclage	Reciclagem
Pb12	5	Habitation	Habitação
Pb13	79	Assainissement	Saneamento
Pb14	13	Relations de voisinage	Relacionamento com os vizinhos

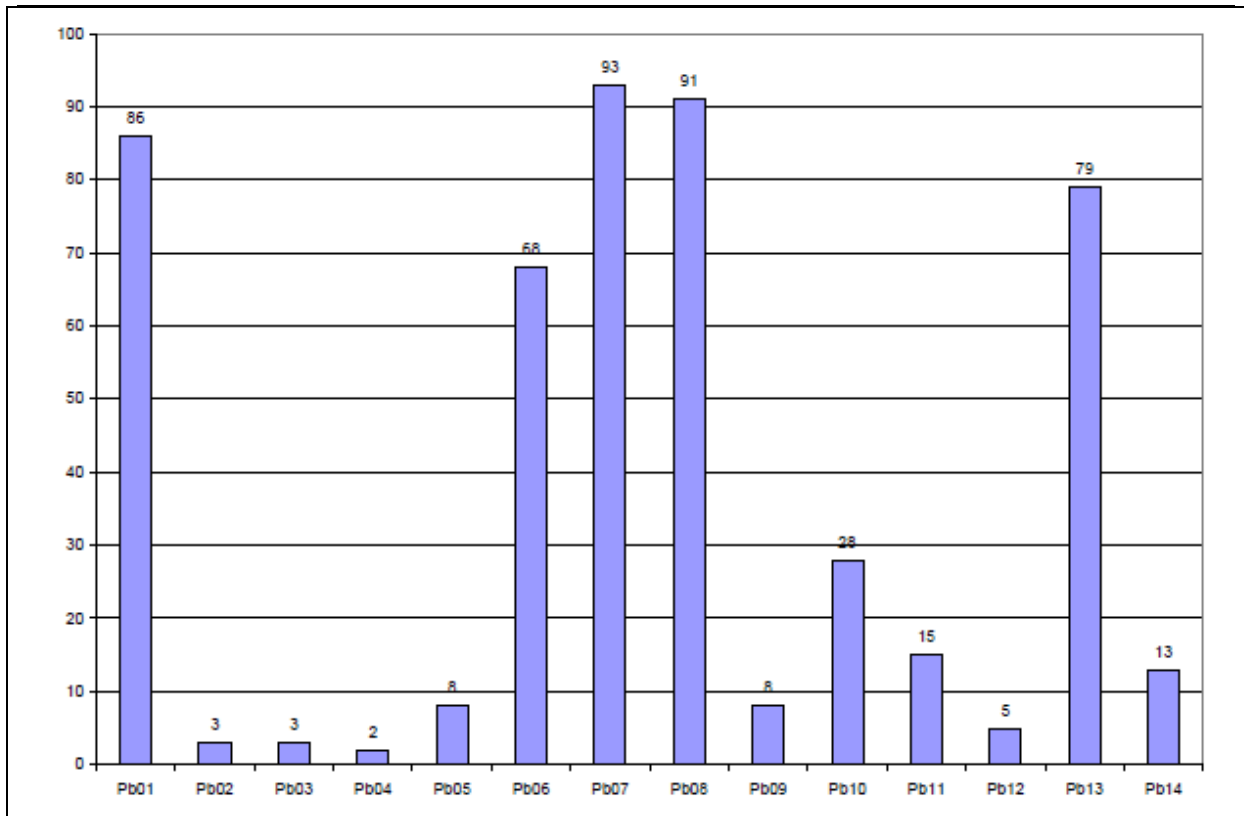


Tableau statistique de la variable V13

(Tableau n° T6) V13=nombre de catégories évoquées	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Échantillon	0	1	5	11	22	31	24	5	1	2	1	0	0	0	0

(Tableau n° T7) Pb08 cité / Pb08 n'est pas cité

Homme	27	7	34
Femme	64	5	69
	91	12	103

Document n° 2 : traitement demandé

TQ1- De quoi s'agit-il ?

Q101- Après avoir pris connaissance des informations fournies dans le document n°1, identifier et caractériser “succinctement” mais très clairement, les éléments de cette étude statistique : individus, population(s), échantillon(s), variables.

Population :

Ensemble des habitant dans une zone périphérique d'une ville du Sud du Brésil (**individus**)

Échantillons:

Sous-ensemble des 103 individus extrait de la population

Variables.

Variables quantitatives		Variables qualitatives		
discrètes	continues	ordinales	nominales	Textuelles
[V7]	[V2]	[V9]	[V1]	[V11]
[V8]	[V4]		[V3]	[V12]
[V13]	[V5]		[V6]	
			[V10]	

TQ2-Etude de la variable V4 « Temps de résidence dans la ville (années) »

Q201 Sur l'échantillon, quelle est la valeur m de la durée moyenne de résidence dans la ville et quelle est la valeur de l'écart-type ?

Intervalles	valeurs centrales (xi)	effectifs (ni)	$n_i x_i$	$n_i(x_i-m)^2$
[1 ; 10[5,5	19	104,5	3460,260156
[10 ; 19[14,5	44	638	889,0787068
[19 ; 30[24,5	22	539	666,6752757
[30 ; 35[32,5	9	292,5	1641,429824
[35 ; 50[42,5	9	382,5	4972,30361
		<i>effectif total</i> 103	<i>somme</i> 1956,5	<i>somme</i> 11629,74757
			<i>moyenne</i> 18,9951	<i>variance</i> 112,9101706
				<i>écart-type</i> 10,62591975

Q202- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

Ainsi la durée moyenne de résidence sur l'échantillon est d'environ 19 ans avec un fluctuation moyenne donnée par l'écart-type de 10 ans 8 mois environ

TQ3- Étude du lien entre l'évocation de la catégorie Pb08 – Loisirs et la variable V01 Sexe

Q301- A partir de l'échantillon des 103 réponses, peut-on considérer qu'il y a ou non un lien entre ces deux variables?

En exploitant les données du tableau T7, tester les hypothèses correspondantes (Ho et H1 qui devront être explicitées) au seuil de risque $\alpha=0.05$ puis $\alpha=0.01$ à l'aide du test d'indépendance du Khi-deux.

Nous mettons en œuvre un test d'indépendance (Test du Khi-deux) pour décider laquelle des deux hypothèses est la plus plausible :

Ho Les hommes et les femmes ont des points de vue homogènes relatifs à la question de l'évocation de la catégorie Pb08 – Loisir

H1 Les hommes et les femmes ont des points de vue hétérogènes relatifs à la question de l'évocation de la catégorie Pb08 – Loisir

Tableau des effectifs observés :

(Tableau n° T7a) Pb08 cité / Pb08 n'est pas cité			
Homme	27	7	34
Femme	64	5	69
	91	12	103

Tableau des effectifs théoriques sous l'hypothèse Ho :

(Tableau n° T7b) Pb08 cité / Pb08 n'est pas cité			
Homme	30,039	3,961	34
Femme	60,961	8,039	69
	91	12	103

Tableau des contributions absolues :

(Tableau n° T7c) Pb08 cité / Pb08 n'est pas cité			
Homme	0,3074	2,3313	2,6387
Femme	0,1515	1,1487	1,3002
	0,4589	3,4800	3,9389

Q302- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

Au seuil de $\alpha = 5\%$ la valeur critique lue dans la table du Khi-deux ddl = 1 est de $k = 3,84$. la valeur empirique est de 3,93 et donc supérieure à la valeur théorique sous Ho. On rejette l'hypothèse d'homogénéité/indépendance à ce niveau de risque de 1^{ère} espèce.

Au seuil de $\alpha = 1\%$ la valeur critique lue dans la table du Khi-deux ddl = 2 est de $k = 6,63$. la valeur empirique est de 3,93 et donc inférieure à la valeur théorique sous Ho. On ne rejette pas l'hypothèse d'homogénéité à ce niveau de risque de 1^{ère} espèce. Cette conservation de Ho contre H1 est alors prise avec un risque de seconde espèce de niveau β inconnu

TQ4 - Étude de l'homogénéité des points de vue relatifs à l'ordre d'importance des problèmes rencontrés

Nous rapportons les rangements des 14 catégories de problèmes produits par les 8 sujets qui ont constitué le groupe des sujets auprès desquels ont été réalisés les entretiens. Les résultats qui utilisent les rangs *ex aequo*, sont reportés dans le tableau n°T8 ci-dessous :

Tableau T8		Pb_01	Pb_02	Pb_03	Pb_04	Pb_05	Pb_06	Pb_07	Pb_08	Pb_09	Pb_10	Pb_11	Pb_12	Pb_13	Pb_14
individus	Sujet_01	4	10	10	10	10	5	2	3	10	10	10	10	1	10
	Sujet_02	4	10,5	10,5	10,5	10,5	6	1	10,5	10,5	2	10,5	5	3	10,5
	Sujet_03	4	10,5	10,5	10,5	10,5	6	1	10,5	10,5	2	10,5	5	3	10,5
	Sujet_04	5	11,5	11,5	11,5	11,5	7	2	6	11,5	8	4	11,5	3	1
	Sujet_05	1	3	9	9	9	9	9	9	2	9	9	9	9	9
	Sujet_06	1	3	9	9	9	9	9	9	2	9	9	9	9	9
	Sujet_07	2	1	9,5	9,5	9,5	9,5	3	4	9,5	9,5	9,5	9,5	9,5	9,5
	Sujet_08	2	1	9,5	9,5	9,5	9,5	3	4	9,5	9,5	9,5	9,5	9,5	9,5
	Totaux	23	50,5	79,5	79,5	79,5	61	30	56	65,5	59	72	68,5	47	69
Ho	60	60	60	60	60	60	60	60	60	60	60	60	60	60	
Ecarts	1369	90,25	380,3	380,3	380,3	1	900	16	30,25	1	144	72,25	169	81	

Q401 - Quelle est la valeur empirique du coefficient W de concordance de Kendall sans faire la correction liée à l'existence des ex aequo ?

Q402- Peut-on considérer qu'il existe un classement ordonné significatif correspondant à une certaine homogénéité des points de vue de ces étudiants? Réaliser le test W de Kendall de confrontation des deux hypothèses Ho et H1 qui devront être explicitées, à niveau de risque de 1^{ère} espèce $\alpha=0.01$.

Ho Les rangements déterminés par l'ordre d'évocation des problèmes sont hétérogènes au sein de l'échantillon,

H₁ Il existe une tendance à l'homogénéité des rangements déterminés par l'ordre d'évocation des problèmes au sein de l'échantillon,

Nous utilisons le test W de Kendall.

$S_{Kmax} =$	$S_K =$	$W = S_K / S_{Kmax}$	$W_{critique}$ seuil 1%
14560	4014,5	0,27572115	0,26623254

Ho = hypothèse d'hétérogénéité des rangements et d'absence de concordance est rejetée au seuil de 1%. Nous acceptons à ce niveau de risque le rangement suivant comme représentant celui avec lequel l'ensemble des individus serait plutôt en accord.

Q403- Donner le rangement qui serait celui le plus concordant.

objectif	n°1	n°2	n°3	n°4	n°5	n°6	n°7	n°8	n°9	n°10	n°11	n°12	n°13	n°14
RANGEMENT significatif	1	4	12	12	12	7	2	5	8	6	11	9	3	10

Q404- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ5- Comparaison des rangements réalisés par les hommes et les femmes relativement aux catégories de problèmes évoqués.

Nous considérons les deux rangements obtenus à partir du test de concordance W de Kendall explicitant les ordres concordants respectivement du groupe des femmes et du groupe des hommes. Les résultats sont consignés dans le tableau T9 ci-dessous (avec présence d'ex aequo comme cela se présente dans la thèse).

Tableau T9	Pb_01	Pb_02	Pb_03	Pb_04	Pb_05	Pb_06	Pb_07	Pb_08	Pb_09	Pb_10	Pb_11	Pb_12	Pb_13	Pb_14
Rangement concordant du groupe Hommes	1	12	13,5	13,5	9	3	2	5	8	6	11	10	4	7
Rangement concordant du groupe Femmes	2	13	12	14	9	5	1	3	10	6	7	11	4	8
écarts au carré	1	1	2,25	0,25	0	4	1	4	4	0	16	1	0	1

$$\Delta^2 = 35,5$$

Q501-Peut-on considérer qu'il existe une tendance à l'accord significative entre ces deux groupes ?

Tester les hypothèses correspondantes (Ho et H1) au seuil de 5% à l'aide du test de Rhô de Spearman.

Ho Les rangements déterminés par l'ordre d'évocation des problèmes sont indépendants au sein de l'échantillon entre les hommes et les femmes

H1 Il existe un lien entre les rangements déterminés par l'ordre d'évocation des problèmes au sein de l'échantillon par les hommes et par les femmes.

Calcul de la valeur empirique du coefficient ρ de Spearman

$$R_s = 1 - \frac{\Delta^2}{n(n^2 - 1)} = 1 - \frac{35,5}{14(14^2 - 1)} = 1 - \frac{35,5}{455} = \frac{419,5}{455} \approx 0,922$$

La valeur critique de Rs au seuil $\alpha=0.05$ (test bilatéral) est de 0,538.

Q502- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

Nous rejetons l'hypothèse Ho au profit de H1. C'est-à-dire que nous retenons l'idée d'un lien entre l'ordre déterminé par les évocations des problèmes chez les hommes et celui chez les femmes.

Situation Problème n°2.2

Document n° 1 : situation de l'étude

Considérons une enquête par questionnaire visant la construction de données pour une étude sur *l'impact d'une formation réalisée dans le cadre d'un Master professionnel en enseignement des sciences physiques et des mathématiques*. Il s'agit d'une recherche en cours de réalisation dans une université brésilienne.

Un questionnaire comportant près de 100 questions a été soumis à un échantillon. Pour des raisons didactiques, les données ici rapportées sont des données fictives tant pour la taille de l'échantillon aléatoire : 125 que pour les réponses aux questions. Nous ne prenons en compte que les 9 questions suivantes mais le questionnaire complet utilisé est accessible à l'URL suivant :

<http://www.pagesperso-orange.fr/jean-claude.regnier/>

Tableau des variables

code de la variable	descriptif succinct
V01	Sexe : 1_femme ; 2_homme
V02	Âge à la date de l'enquête
V03	Dominante dans le Master : 1_enseignement des sciences physiques 2_enseignement des mathématiques
V04	Aujourd'hui qu'évoque pour vous l'enseignement des sciences physiques ou des mathématiques ?
V05	Quel est le degré d'utilité attribué à la discipline : Hypermédia en Éducation ? Min = 0 1 2 3 4 5 6 7 8 9 Max =10
V06	Quel est le degré d'intérêt attribué à la discipline : Théories de l'apprentissage et de l'enseignement ? Min = 0 1 2 3 4 5 6 7 8 9 Max =10
V07	Auto-évaluation du niveau de compétences atteint dans le cadre de la formation en Master (Mestrado) en ce qui concerne l'objectif : Être capable de produire des connaissances dans le domaine de l'enseignement des sciences physiques ou des mathématiques 1_Tout à fait faible 2_Plutôt faible 3_Plutôt fort 4_Tout à fait fort
V08 (fictive)	Nombre de réussites obtenues à un QCM de 20 items portant sur des connaissances scientifiques et mathématiques
V09	Rangement par ordre décroissant d'importance de 10 disciplines du curriculum de diplôme de Mestrado (voir Tableau T9)

Nous rapportons ci-dessous les tableaux statistiques des variables :

(Tableau n° T1) Tableau statistique de la variable V01 = sexe

Modalités de V01	SEXE	
	1_femme	2_homme
Effectifs	70	55

(Tableau n° T2) Tableau statistique de la variable V02 = âge à la date de l'enquête

Valeurs de V02	[23; 30[[30; 36[[36; 42[[42; 48[[48; 53[
Effectifs	23	46	19	23	14

(Tableau n° T5) Tableau statistique de la variable V05 =

Degré d'utilité attribué à la discipline : **Hypermédia en Éducation**

Modalités de V05	0	1	2	3	4	5	6	7	8	9	10
	le plus faible le plus fort										
Effectifs	0	4	0	0	0	7	0	11	26	22	55

(Tableau T7) Tableau statistique de la variable V07 :

Auto-évaluation du niveau de compétences...

Modalités de V07	1_ Tout à fait faible	2_ Plutôt faible	3_ Plutôt fort	4_ Tout à fait fort
Effectifs	25	30	42	28

(Tableau T8 fictif) Tableau statistique de la variable V08= Connaissances
Nombre de réussites obtenues à un QCM de 20 items portant sur des connaissances scientifiques et mathématiques

scores/20	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
effectifs	0	0	2	5	1	0	0	3	7	7	10	12	8	12	3	15	3	12	10	8	7

Document n° 2 : résultats du traitement demandé

TQ1- De quoi s'agit-il ?

Q101- Après avoir pris connaissance des informations fournies dans le **document n°1**, identifier et caractériser “succinctement” et clairement, les éléments de cette étude statistique : population, échantillon, individus, variables.

Q102 – Identifier clairement la nature des 9 variables retenues

Population :

Ensemble de ceux ou celles (**individus**) qui suivent ou ont suivi la formation mais aussi ceux qui la suivront

Échantillons:

Échantillon aléatoire de 125 individus qui ont été soumis à l'enquête

Variables.

Variables quantitatives			Variables qualitatives		
discrètes	continues	rangs	ordinales	nominales	textuelles
[V08]	[V02]	[V09]	[V05] [V06] [V07]	[V01] [V03]	[V04]

La question V09 est un vecteur-variable à 10 composantes dont chaque réalisation est une permutation des nombres de 1 à 10.

TQ2-Étude de la variable V02 sur l'échantillon

Q201 Quel est l'âge **moyen** des individus avec quel **écart-type** ?

Cours de **Jean-Claude Régnier**
Situations Problèmes & Documents autocorrectifs
 10/43

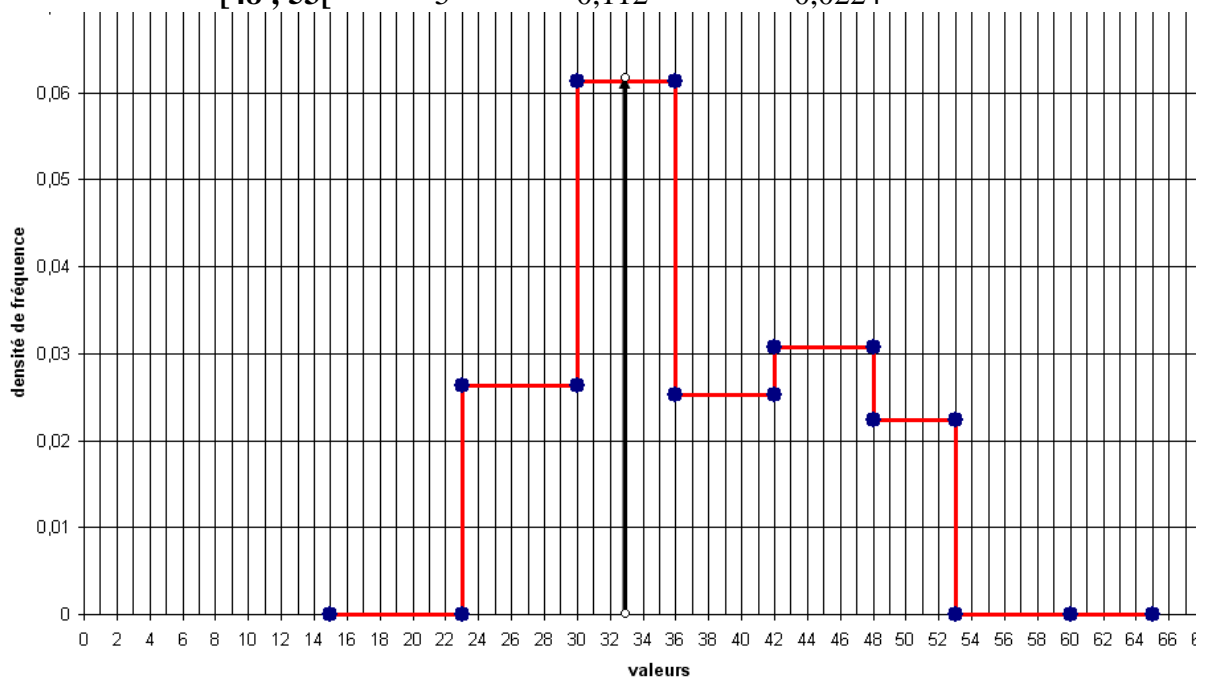
Intervalles	valeurs centrales (xi)	effectifs (ni)	$n_i x_i$	$n_i(x_i-m)^2$
[23; 30[26,5	23	609,5	2480,031488
[30 ; 36[33	46	1518	693,930976
[36 ; 42[39	19	741	85,071664
[42 ; 48[45	23	1035	1514,997488
[48 ; 53[50,5	14	707	2595,536384
		<i>effectif total</i>	<i>somme</i>	<i>somme</i>
		125	4610,5	7369,568
			<i>moyenne</i>	<i>variance</i>
			36,884	58,956544
				<i>écart-type</i>
				7,678316

L'âge moyen est de 36,88 années soit environ 36 ans 11 mois avec un écart-type de l'ordre de 7 ans et 8 mois.

Q202 Construire un **histogramme** de la distribution des fréquences

Tableau T11

Intervalles	amplitude	fréquences	densité de fréquences
[23; 30[7	0,184	0,026285714
[30 ; 36[6	0,368	0,061333333
[36 ; 42[6	0,152	0,025333333
[42 ; 48[6	0,184	0,030666667
[48 ; 53[5	0,112	0,0224



Q203 Quel est l'**âge modal** ?

Nous pouvons situer la valeur modale dans l'intervalle [30 ; 36[sur lequel la densité de fréquence est maximale : 0,0613. Nous pouvons fournir comme estimation ponctuelle de l'âge modal, le centre de l'intervalle, c'est-à-dire 33 ans.

Q204 Quel est l'**âge médian** ?

En prenant appui sur l'histogramme et recourant à un raisonnement par interpolation linéaire, déterminer une estimation de l'âge médian des bacheliers.

On peut déjà estimer que la durée médiane est située dans l'intervalle [30 ; 36[.

On pose l'équation suivante pour obtenir la médiane Me :

$$(Me - 30) \frac{46}{125} \frac{1}{6} + \frac{23}{125} = 0,5$$

$$(Me - 30) \frac{46}{125} \frac{1}{6} = 0,5 - \frac{23}{125} \text{ d'où } (Me - 30) \frac{46}{6} = (0,5)(125) - 23$$

$$\text{ou encore } (Me - 30) = \frac{6}{46} [62,5 - 23] = \frac{6}{46} (39,5) \approx 5,152$$

$$Me = 30 + \frac{237}{46} \text{ soit } Me \approx 35,152$$

On peut estimer que l'âge médian est de l'ordre de 35 ans et 2 mois

Q205- A partir de cette étude, énoncer sous une forme synthétique votre **conclusion** et votre **interprétation**.

TQ3 - Étude de l'homogénéité des points de vue relatifs à l'ordre décroissant d'importance de 10 disciplines constituant le curriculum de la formation

Nous rapportons ici les réponses fournies (fictivement) par un sous-échantillon de 14 individus extrait de l'échantillon global des diplômés du Mestrado de la spécialité mathématiques. Chaque répondant ignore le rangement proposé par les autres. Les résultats concernant les rangs sont reportés dans le tableau T9 ci-dessous :

Tableau T9

Disciplines										
individus	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10
Mestr_001	6	1	2	3	8	7	4	10	9	5
Mestr_122	6	9	1	4	5	7	10	8	3	2
Mestr_073	1	2	8	7	5	4	3	6	9	10
Mestr_014	3	2	10	8	1	5	6	7	4	9
Mestr_105	3	10	7	5	6	8	4	2	1	9
Mestr_026	7	1	6	3	2	5	4	8	9	10
Mestr_017	6	1	3	4	8	5	2	7	10	9
Mestr_038	9	1	2	3	6	8	10	4	5	7
Mestr_019	7	5	2	4	6	8	1	9	10	3
Mestr_015	7	1	10	3	6	8	2	4	5	9
Mestr_111	5	2	3	8	1	6	4	7	10	9
Mestr_012	6	9	4	2	3	7	1	8	5	10
Mestr_003	1	9	3	4	7	6	5	8	10	2
Mestr_064	3	4	6	10	5	9	2	8	1	7

Énoncés des disciplines
D01 Hypermédia en éducation
D02 Théories de l'apprentissage et de l'enseignement
D03 Théorie et systématisation de l'enseignement
D04 Méthodologie de la recherche
D05 Séminaires communs
D06 Accompagnement de l'élaboration du mémoire
D07 Thèmes de méthodologie d'enseignement des mathématiques
D08 Fondements de Calcul
D09 Fondements de Mathématiques
D10 Fondements de Géométrie

Q301 Quel nom donne-t-on à ce tableau T9 ? Quelles informations contient-il ?

C'est un tableau de rangs qui donne les rangements réalisés par les $p = 14$ individus relativement à $n = 10$ disciplines. Chaque ligne est donc une permutation de $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

Disciplines												Somme totale de tous rangs attribués
	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10		
Sommes des rangs attribués à chaque item	70	57	67	68	69	93	58	96	91	101	770	

Q302 Peut-on considérer qu'il existe un classement ordonné significatif correspondant à une certaine homogénéité des points de vue des étudiants encore en formation ou qui en sont issus? Avec le test W de Kendall tester respectivement au seuil de 1% et au seuil de 5% afin de choisir laquelle des deux hypothèses H_0 et H_1 qu'il faut bien expliciter auparavant, est la plus vraisemblable.

Deux hypothèses sont en concurrence :

H_0 Les rangements attribués par les bacheliers interrogés sont indépendants et constituent un tout hétérogène.

H_1 Il existe une tendance à l'accord entre les rangements attribués.

Pour mettre en place la procédure de test fondée sur celle proposée par Kendall à partir de la statistique W , il nous faut obtenir la somme totale des $p=14$ rangs obtenus par chacun des $n=10$ items.

Sous l'hypothèse H_0 ces sommes seraient toutes identiques et seraient égales à

$$\frac{p \frac{n(n+1)}{2}}{n} = \frac{14(55)}{10} = 77$$

On mesure ensuite un écart global S_K entre ces valeurs obtenues et les valeurs espérées sous H_0 .

$$S_K = (70-77)^2 + (57-77)^2 + (67-77)^2 + (68-77)^2 + (69-77)^2 + (93-77)^2 + (58-77)^2 + (96-77)^2 + (91-77)^2 + (101-77)^2$$

$$S_K = 49+400+100+81+64+256+361+361+196+576$$

$$S_K = 2444$$

On mesure l'écart global maximum qui correspond à l'accord parfait c'est-à-dire le cas où déca-uplet des sommes serait une permutation de (14, 28, 42, 56, 70, 84, 98, 112, 126, 140) :

$$S_{K \max} = (14-77)^2 + (28-77)^2 + (42-77)^2 + (56-77)^2 + (70-77)^2 + (84-77)^2 + (98-77)^2 + (112-77)^2 + (126-77)^2 + (140-77)^2$$

$$S_{K \max} = 16170$$

L'écart maximum peut être obtenu directement par la formule :

$$S_{K \max} = \frac{1}{12} p^2 (n^3 - n)$$

$$S_{K \max} = \frac{1}{12} 14^2 (10^3 - 10) = 16170$$

On calcule ensuite la valeur empirique de la statistique W de Kendall :

$$W = \frac{S_K}{\frac{1}{12} p^2 (n^3 - n)}$$

Pour information, retournons aux propriétés de la variable W . On connaît un certain nombre de caractéristiques :

D'une part : $0 \leq W \leq 1$

D'autre part :

Espérance	Variance	Moment centré d'ordre 3	Moment centré d'ordre 4
$E(W) = \frac{1}{p}$	$\sigma^2(W) = \frac{2(p-1)}{p^3(n-1)}$	$\mu_3(W) = \frac{8(p-1)(p-2)}{p^5(n-1)^2}$	$\mu_4(W) = \frac{12(p-1)^2}{p^6(n-1)^2} + \frac{48(p-1)(p-2)(p-3)}{p^7(n-1)^3} - \frac{48(p-1)}{p^7(n+1)(n-1)^2}$

Notons que $w = 1$ correspond à la situation d'accord parfait sur le rangement des n objets pour les p critères.

Si $w=0$ correspond au cas où chaque colonne a même total, c'est-à-dire que les objets seraient tous *ex aequo*. De faibles valeurs de W suggèrent encore cette indiscernabilité entre les objets, c'est-à-dire l'idée d'une indépendance des rangements.

Pour décider du rejet ou non de l'hypothèse nulle H_0 d'indépendance des p rangements, on compare la valeur empirique obtenue avec la valeur critique correspondant à un niveau α du risque de 1^{ère} espèce :

a) Pour $n \leq 7$ on utilise la table ci-dessous qui fournit les valeurs critiques w_k telles que $P(W \geq w_k) = \alpha = 0,05$:

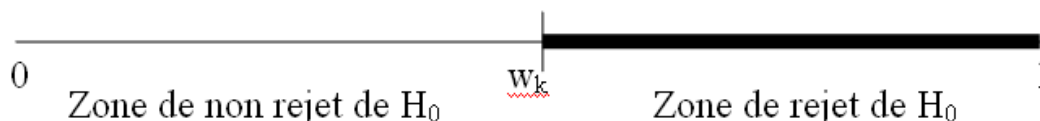
p	n	3	4	5	6
3	3	1	0,750	0,600	0,500
4	4	0,822	0,619	0,500	0,421
5	5	0,716	0,553	0,449	0,377
6	6	0,660	0,512	0,418	0,351
7	7	0,626	0,484	0,395	0,332
8	8	0,595	0,461	0,378	0,319
9	9	0,576	0,447	0,365	0,307
10	10	0,560	0,434	0,354	0,299
11	11	0,548	0,425	0,346	0,287
12	12	0,535	0,415	0,336	0,287
13	13	0,527	0,409	0,332	0,280
14	14	0,520	0,402	0,327	0,275
15	15	0,514	0,395	0,322	0,272
20	20	0,49	0,37	0,30	0,25
40	40	0,43	0,33	0,26	0,22
60	60	0,41	0,31	0,25	0,21
100	100	0,38	0,29	0,24	0,20
∞	∞	0,33	0,25	0,20	0,17

b) Pour $n \leq 7$ et $2 < p \leq 20$ la variable $\frac{(p-1)W}{1-W}$ est distribuée comme la variable de Fisher-Snédecor $F(n_1 = n-1-\frac{2}{p}; n_2 = (p-1)(n-1-\frac{2}{p}))$

c) Pour $n > 7$ on utilise la variable $p(n-1)W$ qui est une variable de Pearson χ^2_{n-1}

La table fournie en annexe qui donne les valeurs critiques w_k telles que $P(W > w_k) = \alpha$ est déduite de cette propriété.

Globalement la règle de décision s'appuie sur la configuration suivante de l'espace de décision :

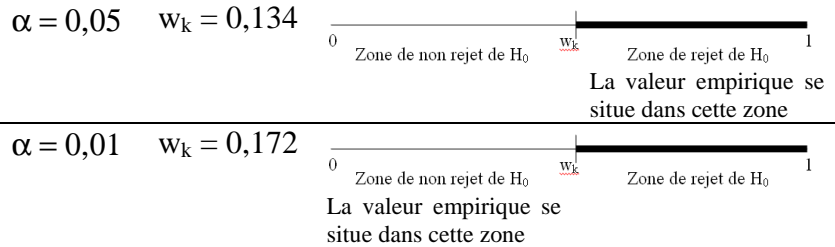


Dans le cas où l'on est conduit à rejeter l'hypothèse d'indépendance des rangements, on utilise souvent la règle de classement suivante:

Les objets sont classés dans l'ordre défini par les totaux des colonnes.

Retour à la situation étudiée :
$$W = \frac{S_K}{\frac{1}{12} p^2 (n^3 - n)} = \frac{2444}{16170} \approx 0,15114$$

Dans la table de valeurs critiques, nous pouvons lire :



Il ressort qu'au niveau de risque $\alpha = 0,05$ nous rejetons l'hypothèse H_0 c'est-à-dire que nous pouvons retenir l'idée d'une tendance à l'accord qui pourrait être réalisé autour du rangement suivant :

	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10
Rang	6	1	3	4	5	8	2	9	7	10

En revanche, au niveau $\alpha = 0,01$ nous ne rejetons pas l'hypothèse H_0 c'est-à-dire que nous ne pouvons retenir l'idée d'une tendance à l'accord qui pourrait être réalisé autour du rangement obtenu. Nous prenons alors un risque de 2nde espèce, de niveau β , dont nous pouvons préciser la valeur.

Q303 S'il existe un rangement significatif pour un niveau de risque donné, quel est ce rangement ?

Rang	1	2	3	4	5	6	7	8	9	10
Disciplines	D02	D07	D03	D04	D05	D01	D09	D06	D08	D10

Ou encore :

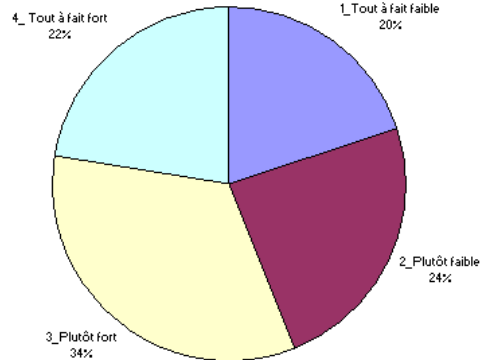
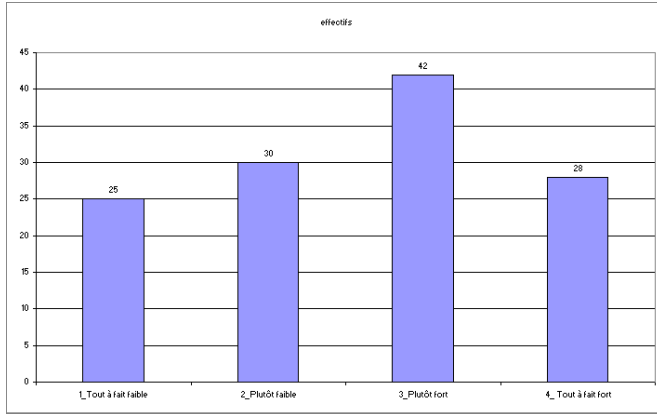
- | | | |
|----|-----|---|
| 1 | D02 | Théories de l'apprentissage et de l'enseignement |
| 2 | D07 | Thèmes de méthodologie d'enseignement des mathématiques |
| 3 | D03 | Théorie et systématisation de l'enseignement |
| 4 | D04 | Méthodologie de la recherche |
| 5 | D05 | Séminaires communs |
| 6 | D01 | Hypermédia en éducation |
| 7 | D09 | Fondements de Mathématiques |
| 8 | D06 | Accompagnement de l'élaboration du mémoire |
| 9 | D08 | Fondements de Calcul |
| 10 | D10 | Fondements de Géométrie |

Q304- A partir de cette étude, énoncer sous une forme synthétique votre **conclusion** et votre **interprétation**.

TQ4 - Étude de la variable V07 = Auto-évaluation du niveau de compétences

Q401 De quoi s'agit-il dans cette étude guidée par les items Q402 et Q403 ? Quel est l'objectif de cette procédure ? Quelles précisions sont requises dans les énoncés des résultats ?

Modalités de V07	1_ Tout à fait faible	2_ Plutôt faible	3_ Plutôt fort	4_ Tout à fait fort
effectifs	25	30	42	28
fréquences	0,2	0,24	0,34	0,22
%	20%	24%	33,6%	22,4%



Q402 Donner la proportion des individus qui se considèrent comme «3_plutôt fort» dans l'échantillon.

Nombre d'individus ayant répondu «3_Plutôt fort » est 42 parmi les 125 qui ont répondu.
Proportion dans l'échantillon :

$$f = \frac{42}{125} = 0.336 = 33.6\%$$

Q403 Donner une estimation ponctuelle de la proportion des individus qui se considèrent comme «3_plutôt fort» dans la population.

Estimation ponctuelle de la proportion dans la population :
33.6%

Q404 Donner une estimation par intervalle de confiance à 95% ou à 99% (selon votre choix) de la proportion des individus qui se considèrent comme «3_plutôt fort» dans la population.

$$\left[f - k \sqrt{\frac{f(1-f)}{n-1}}; f + k \sqrt{\frac{f(1-f)}{n-1}} \right]$$

Au seuil de confiance $1-\alpha=0,95$ la valeur critique $k=1,96$

$$\left[0.336 - 1.96 \sqrt{\frac{0.336(1-0.336)}{125-1}}; 0.336 + 1.96 \sqrt{\frac{0.336(1-0.336)}{125-1}} \right]$$

or $1,96 \left(\sqrt{\frac{0.336(0.664)}{124}} \right) \approx 1,96(0,042417281) \approx 0,083137872$

d'où [0,2528 ; 0,4192] c'est à dire entre 25,28% et 41,92%

Avec un niveau de confiance de 95%, nous estimons que la proportion se situe environ entre 25,28% et 41,92%

Au seuil de confiance $1-\alpha=0,99$ la valeur critique $k=2,57$

$$\left[0.336 - 2.57 \sqrt{\frac{0.336(1-0.336)}{125-1}}; 0.336 + 2.57 \sqrt{\frac{0.336(1-0.336)}{125-1}} \right]$$

or $2,57 \left(\sqrt{\frac{0,336(0,664)}{124}} \right) \approx 2,57(0,0424417281) \approx 0,10901214$

d'où [0,2269 ; 0,4451] C'est à dire entre 22,69% et 44,51%

Avec un niveau de confiance de 99%, nous estimons que la proportion se situe environ entre 22,69% et 44,51%

Q405 Quelle incidence a le niveau de confiance ?

Accroître le niveau de confiance conduit à accroître l'amplitude de l'intervalle de confiance, c'est-à-dire diminuer le niveau de précision.

Q406- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ5- Étude du lien entre sexe et les spécialités mathématiques et sciences physiques (croisement des variables V01 et V03)

Tableau T10		V03 = Spécialité		
		Sciences Physiques	Mathématiques	
V01 Sexe	Femme	18	52	70
	Homme	33	22	55
		51	74	125

Q501 Quel nom donne-t-on à ce tableau ? Quelles informations contient-il ?

C'est un tableau de contingence. Tableau de distribution des effectifs de la conjonction de deux variables. Tableau des effectifs conjoints.

Q502 A partir de ces données issues de l'échantillon des répondants, peut-on considérer qu'il y a une différence selon le sexe ?

Il s'agit de tester les hypothèses correspondantes (Ho et H1 qu'il faut expliciter) au seuil de 5% ou au seuil de 1% selon votre choix, à l'aide du test d'indépendance du Khi-deux.

Nous mettons en œuvre un test d'indépendance (Test du Khi-deux) pour décider laquelle des deux hypothèses est la plus plausible :

Ho Les hommes et les femmes ont des points de vue homogènes relatifs au choix en mathématiques et sciences physique

H1 Les hommes et les femmes ont des points de vue hétérogènes relatifs à relatifs au choix en mathématiques et sciences physique

Tableau des effectifs observés :

Tableau T10		V03 = Spécialité		
		Sciences Physiques	Mathématiques	
V01 Sexe	Femme	18	52	70
	Homme	33	22	55
		51	74	125

Cours de **Jean-Claude Régnier**
Situations Problèmes & Documents autocorrectifs
 17/43

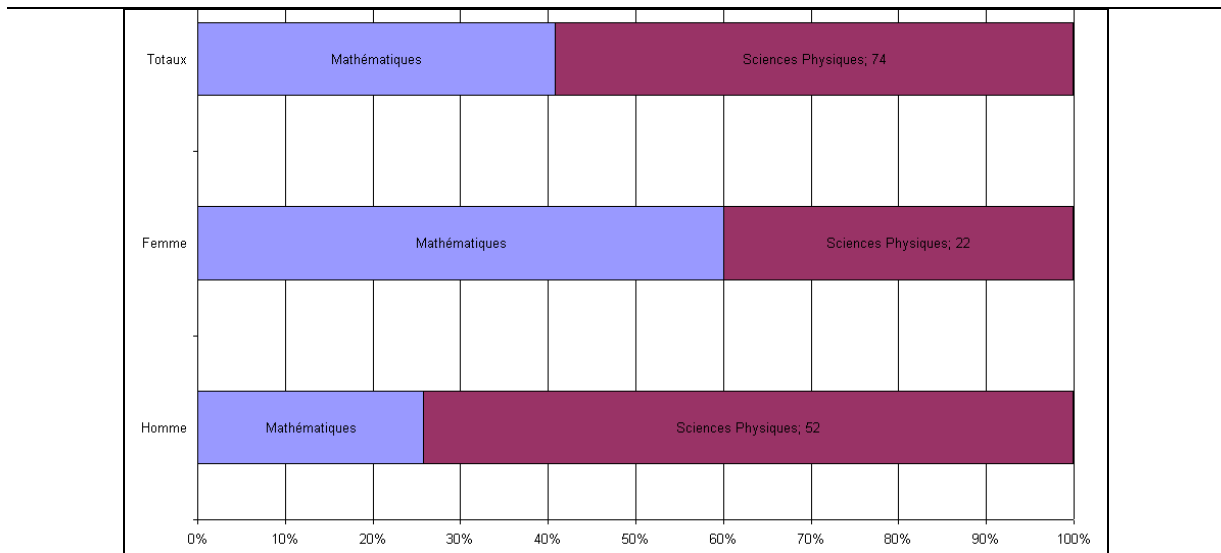


Tableau des effectifs théoriques sous l'hypothèse H_0 :

		V03 = Spécialité		
		Sciences Physiques	Mathématiques	
V01 Sexe	Femme	28,56	41,44	70
	Homme	22,44	32,56	55
		51	74	125

Tableau des contributions absolues :

		V03 = Spécialité		
		Sciences Physiques	Mathématiques	
V01 Sexe	Femme	3,9045	2,6909	6,5954
	Homme	4,9694	3,4249	8,3943
		8,8739	6,1158	14,9897

Au seuil de $\alpha = 5\%$ la valeur critique lue dans la table du Khi-deux ddl = 1 est de $k = 3,84$.

Au seuil de $\alpha = 1\%$ la valeur critique lue dans la table du Khi-deux ddl = 1 est de $k = 6,64$.

Ici la valeur empirique est de **14,98** :

Elle est donc supérieure à la valeur théorique sous H_0 au niveau de risque 0.05. On rejette l'hypothèse d'homogénéité/indépendance à ce niveau de risque de 1^{ère} espèce.

Mais elle est aussi, au niveau de risque 0.01, supérieure à la valeur théorique sous H_0 . On rejette l'hypothèse d'homogénéité/indépendance à ce niveau de risque de 1^{ère} espèce.

Q503- A partir de cette étude, énoncer sous une forme synthétique votre **conclusion** et votre **interprétation**

TQ6-Étude de la variable V08 sur l'échantillon

Q601 Quel est le **score moyen** des individus avec quel **écart-type** ?

Q602 Construire un **diagramme** de la distribution des fréquences

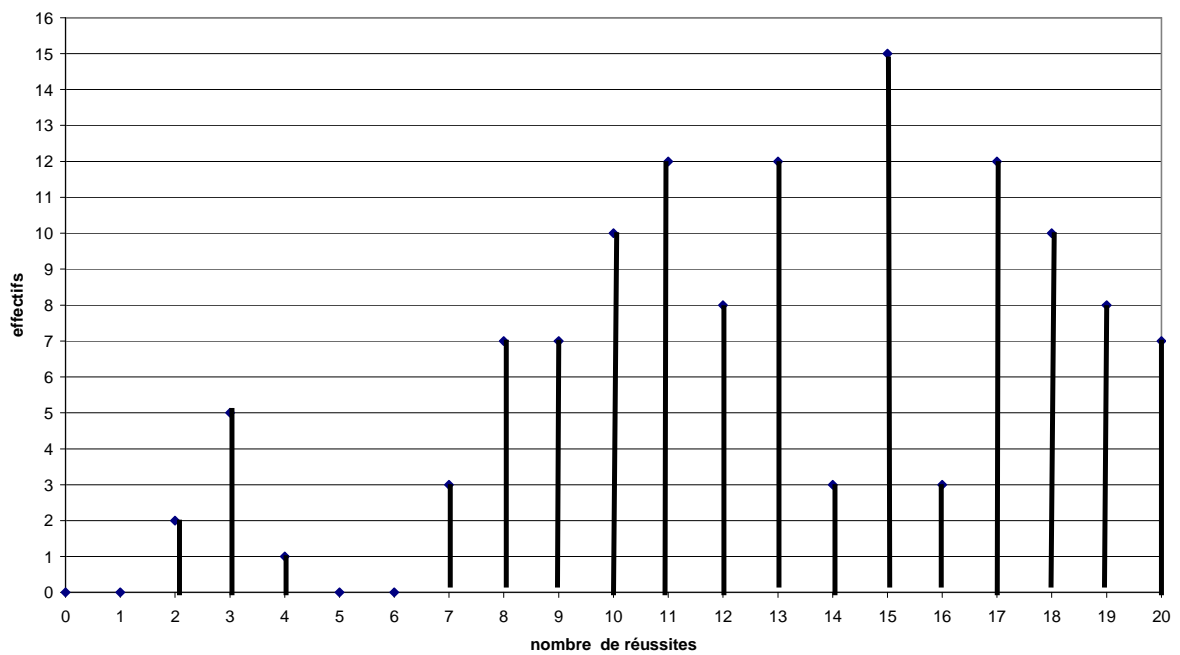
Q603 Quel est le **score modal** ?

Q604 Quel est le **score médian** ?

Q605- A partir de cette étude, énoncer sous une forme synthétique votre **conclusion** et votre **interprétation**.

Moyenne	13,104
Variance	20,205184
Ecart-Type	4,49501769
Minimum	2
1 ^{er} quartile Q1	10
2 ^{ème} quartile Q2=médiane	13
3 ^{ème} quartile Q3	17
Maximum	20
Mode	15

scores/20



Situation Problème n°2.3

Document n° 1 : situation de l'étude

La présente situation s'inspire d'une recherche fondée pour partie sur une enquête par entretien auprès de 20 individus et pour le reste sur une enquête par questionnaire impliquant 430 individus visant la construction de données sur les représentations sociales des « maîtres généralistes » et des « maîtres E » à l'égard de l'enfant « aidé ». Cette recherche a fait l'objet d'une thèse réalisée par Jean-Bruno Bernard et soutenue le 24 avril 2007 à l'Université Lyon2. Elle est consultable sur le site des thèses en ligne de Lyon2. La population d'étude est celle des « enseignants du premier de degré », plus particulièrement les deux sous-populations constituées par les « maîtres généralistes » et par les « maîtres E ». Les données ont été construites par le moyen d'un questionnaire comportant 28 questions pour les « maîtres généralistes » et 23 pour les « maîtres E ». Le questionnaire a été soumis à deux échantillons qui furent considérés comme représentatifs des sous-populations respectives : 300 « maîtres généralistes » et 130 « maîtres E ».

Parmi toutes les variables étudiées, nous ne considérons ici que quelques-unes réellement étudiées auxquelles nous ajoutons d'autres dont nous suggérons la mise en œuvre dans une poursuite d'exploration.

Les deux questionnaires que nous nommons QMG et QME, utilisés pour cette recherche portaient, entre autre, sur les variables suivantes :

<i>code de la variable</i>	<i>descriptif succinct</i>
S01	Sexe : 1 = homme 2 = femme
S02	Ancienneté dans la fonction (années)
S03	Statut : Instituteur (P1) Professeur d'école (P2).
S04	Zone géographique d'intervention : rurale (Z1) ; urbain (Z2) ; ZEP (Z3)
V02	Différence entre difficulté d'apprentissage et échec scolaire : 1=ooui ; 2=non ; 3=je ne sais pas
V 04	Certaines matières sont-elles plus représentatives que d'autres quand on parle de difficultés scolaires ? 1=ooui ; 2=non ; 3=je ne sais pas
V07	Considérez-vous qu'il existe un profil type d'enfants prédisposés à être en difficulté d'apprentissage ? 1=ooui ; 2=non ; 3=je ne sais pas
V08 QME	Vous est-il arrivé des sentir des tensions entre vous et le « maître généraliste » ? 1=ooui ; 2=non ; 3=je ne sais pas
V12 QMG	Vous est-il arrivé de sentir des tensions entre vous et le « maître E » ? 1=ooui ; 2=non ; 3=je ne sais pas
V12 QME	Vous considérez-vous en tant que « maître E » comme un enseignant ? 1=ooui ; 2=non ; 3=je ne sais pas
V16 QMG	Considérez-vous le « maître E » comme un enseignant ? 1=ooui ; 2=non ; 3=je ne sais pas
V21QMG V18QME	De quoi dépend l'efficacité du travail en équipe ?

Nous rapportons ci-dessous les tableaux statistiques des variables

Tableaux statistiques de la variable S01

Tableau n° T1	S01=Sexe	homme	femme
« Maîtres généralistes »	Effectifs	100	200
« Maîtres E »	Effectifs	76	44

Tableaux statistiques de la variable S02

Tableau n° T2	S02= Ancienneté	[1 ; 10[[10 ; 20[[20 ; 30[[30; 40[
« Maîtres généralistes »	effectifs	95	68	66	71
« Maîtres E »	effectifs	27	40	43	20

Tableaux statistiques de la variable S03

Tableau n° T3	S03=Statut	Instituteur	Professeur des écoles
« Maîtres généralistes »	Effectifs	88	212
« Maîtres E »	Effectifs	31	99

Tableaux statistiques de la variable S04

Tableau n° T4	S04=Zone	rurale	urbaine	ZEP
« Maîtres généralistes »	Effectifs	166	85	49
« Maîtres E »	Effectifs	44	45	41

Tableaux statistiques de la variable V02

Tableau n° T5	V02=différence	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	262	29	5
« Maîtres E »	Effectifs	129	0	1

Tableaux statistiques de la variable V04

Tableau n° T6	V04=matières	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	244	46	9
« Maîtres E »	Effectifs	125	4	0

Tableaux statistiques de la variable V07

Tableau n° T7	V07=profil-type	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	108	160	30
« Maîtres E »	Effectifs	52	57	14

Tableaux statistiques de la variable « Tensions ressenties »

Tableau n° T8	V08QME/V12QMG	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	61	221	11
« Maîtres E »	Effectifs	101	26	0

Tableaux statistiques de la variable « Maître E est un enseignant »

Tableau n° T9	V12QME/V16QMG	oui	non	Je ne sais pas
« Maîtres généralistes »	Effectifs	286	4	2
« Maîtres E »	Effectifs	123	4	0

Document n° 2 : traitement demandé

TQ1- De quoi s'agit-il ?

Q101- Après avoir pris connaissance des informations fournies dans le document n°1, identifier et caractériser "succinctement" et clairement, les éléments de cette étude statistique : population, échantillon, individus, variables.

Q102- Quel traitement proposeriez-vous pour le tableau T7 ?

TQ2-Étude de la variable S02 « ancienneté »

Q201 Quelle est la valeur de l'ancienneté moyenne et celle de l'écart-type ?

Q202- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ3- Étude de la question de l'existence de profil-type

Q301- A partir de l'échantillon des 321 réponses, peut-on considérer qu'il y a ou non une différence de point de vue entre les deux sous-populations ?

En exploitant les données du tableau T7, tester les hypothèses correspondantes (Ho et H1) au seuil de risque 5% à l'aide du test d'homogénéité du Khi-deux.

Q302- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ4 - Étude de la variable V07 « Existence de profil type»

Q401 Donner la proportion des individus déclarant « oui, il existe un profil type... » dans l'échantillon des « maîtres généralistes ».

Q402 Donner une estimation ponctuelle de la proportion des individus déclarant « oui, il existe un profil type... » dans la population. des « maîtres généralistes »

Q403 Donner une estimation par intervalle de confiance à 95% de la proportion des individus « oui, il existe un profil type... » dans la population des « maîtres généralistes »

Q404- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ5 - Étude de l'homogénéité des points de vue relatifs l'efficacité du travail en équipe

À partir d'un sous-échantillon aléatoire sans remise de 12 enseignants extrait de l'échantillon des 300 « maîtres généralistes », nous supposons avoir pu recueillir les rangements par ordre d'importance décroissante de 8 items caractérisant ce dont dépend un travail en équipe efficace

	L'efficacité du travail en équipe dépend avant tout...
Item 1	De l'âge des enseignants
Item 2	De l'ancienneté des enseignants
Item 3	Du sexe des enseignants
Item 4	De la formation des enseignants
Item 5	Du statut des enseignants
Item 6	Du lieu géographique d'enseignement
Item 7	D'autres facteurs
Item 8	De facteurs que j'ignore

Les résultats (fictifs) concernant les rangs sont reportés dans le tableau n°T10 ci-dessous :

Tableau T10 individus	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
MG18	2	1	3	4	7	6	5	8
MG32	3	7	4	5	2	1	8	6
MG15	8	2	6	1	4	5	3	7
MG06	2	1	3	8	5	4	7	6
MG65	4	7	1	6	5	2	8	3
MG152	3	2	1	4	7	6	5	8
MG251	1	5	7	6	8	2	3	4
MG99	6	2	3	4	5	1	8	7
MG63	6	1	3	2	7	4	5	8
MG25	3	2	1	4	7	6	5	8
MG17	3	1	5	6	2	4	8	7
MG211	2	4	8	1	5	7	3	6

Q501 - Quelle est la valeur empirique du coefficient W de concordance de Kendall ?

Q502- Peut-on considérer qu'il existe un classement ordonné significatif correspondant à une certaine homogénéité des points de vue de ces enseignants? Tester cette hypothèse à niveau de risque de 1^{ère} espèce $\alpha=0.01$

Q503- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

TQ6- Représentations des difficultés d'apprentissage et de l'échec scolaire.

Nous considérons l'échantillon des 20 enseignants ayant participé à l'enquête par entretien.

Dans les discours produits, nous avons dénombré pour chaque individu, les mots qui caractérisent les difficultés scolaires en les différenciant nettement de l'échec scolaire. Ensuite un rangement de ces 20 enseignants a été réalisé à partir de la variable "nombre de mots utilisés" Les résultats sont consignés dans le tableau T11 (étude et résultats fictifs)

Tableau T11 individus	Catégorie	Rang	Tableau T11 individus	Catégorie	Rang
Ind01	MG	1	Ind11	MG	11
Ind02	MG	2	Ind12	MG	12
Ind03	MG	3	Ind13	ME	13
Ind04	ME	4	Ind14	MG	14
Ind05	MG	5	Ind15	ME	15
Ind06	MG	6	Ind16	ME	16
Ind07	ME	7	Ind17	ME	17
Ind08	MG	8	Ind18	MG	18
Ind09	MG	9	Ind19	MG	19
Ind10	MG	10	Ind20	MG	20

Q601-Peut-on considérer qu'il existe une différence significative entre ces deux groupes « maîtres généralistes »[MG] et « maîtres E » [ME] ?

Tester les hypothèses correspondantes (Ho et H1) au seuil de 5% à l'aide du test de Mann et Whitney.

Q602- A partir de cette étude, énoncer sous une forme synthétique votre conclusion et votre interprétation.

Retour sur les tableaux de contingence

1. Test du Khi-2 d'indépendance ou Test du Khi-2 d'homogénéité

Quand il s'agit de deux variables X et Y qui possèdent plus de deux modalités, nous avons alors à traiter un tableau à q colonnes et p lignes donc à pq cases.

	y1	...	yj	...	yq	effectifs	totaux relatifs à X
x1	n11	...	n1j	...	n1q		n1.
...
xi	n_{i1}	...	n_{ij}	...	n_{iq}		n_{i.}
...
xp	np1	...	n _{iq}	...	npq		np.
effectifs totaux relatifs à Y	n.,1	...	n.,j	...	n.,q		n= n _{..}

On calcule la valeur empirique d^2 de D^2 par la formule

$$d^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \quad \text{avec } n_{ij} = \text{effectif observé relatif à la réalisation}$$

simultanée de la modalité i de la première variable X et de la modalité j de la seconde variable Y.

1.1. Contribution d'une "case" à la valeur prise par $D^2 = \chi^2$:

On peut calculer les **contributions absolues** de chaque "case" :

$$CTA_{ij} = \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Puis les **contributions relatives** de chaque "case" :

$$CTR_{ij} = \frac{1}{d^2} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Cette information peut être présentée sous forme d'un tableau analogue au tableau de contingence. Cela fait apparaître les croisements qui contribuent le plus à écarter le tableau observé du tableau sous l'hypothèse d'indépendance.

1.2. Mesures d'association dérivées de la valeur prise par $D^2 = \chi^2$:

On peut tout d'abord énoncer une propriété remarquable :

$$\frac{d^2}{n} \leq \inf \{p-1, q-1\}$$

où p est le nombre de lignes et q est le nombre de colonnes du tableau de contingence.

Il existe toute une série de coefficients proposés pour obtenir une mesure comprise entre 0 (dans le cas de l'indépendance) et 1 (dans le cas d'une liaison fonctionnelle).

1.2.1. Coefficient de contingence de Karl Pearson :

$$C = \sqrt{\frac{d^2}{n+d^2}}$$

1.2.2. Coefficient de Tschuprow :

$$T = \sqrt{\frac{d^2}{n\sqrt{(p-1)(q-1)}}$$

1.2.3. Coefficient de Cramer :

$$V = \sqrt{\frac{d^2}{n \inf\{p-1, q-1\}}}$$

1.2.4. Apport des coefficients d'association

Ces coefficients permettent alors de comparer des niveaux d'association entre deux variables étudiées selon des modélisations différentes à partir d'échantillons de tailles distinctes. En d'autres termes cela revient à pouvoir comparer des tableaux de contingence qui n'ont ni le même nombre de lignes ni le même nombre de colonnes ni le même effectif total

1.3. Autres utilisations de la mesure d'association χ^2

Il y a d'autres formes d'utilisation de la mesure du χ^2 en fonction des questions que l'on se pose. Tout d'abord, s'il s'agit de comparer des proportions relatives aux modalités d'une variable qualitative au sein de plusieurs populations, il suffit de se ramener au cas précédent en écrivant un tableau de contingence où les lignes correspondent aux divers échantillons et les colonnes représentent les modalités de la variable qualitative. **C'est le test d'homogénéité du χ^2 .** La procédure mathématique et la démarche sont rigoureusement les mêmes que le test du χ^2 d'indépendance

S'il s'agit de comparer la distribution de fréquence d'une variable qualitative ou d'une variable quantitative, observée sur un échantillon, à une distribution théorique connue sur la population. **C'est le test du χ^2 d'adéquation.**

1.4. Aide à l'analyse d'un tableau de contingence et à l'interprétation

		Variable B			effectifs
		B1	B2	B3	
Variable A	A1				q ₁
	A2		$o_{12} = \text{Card}(A1 \& B2)$		q ₂
	A3				q ₃
	A4				q ₄
effectifs		p ₁	p ₂	p ₃	n

Effectif théorique espéré sous l'hypothèse Ho de l'indépendance des deux modalités A1 et B2 :

$$t_{12} = \frac{p_1 q_2}{n}$$

1^{er} cas : si $o_{12} = t_{12}$ alors les deux modalités sont indépendantes

2^{ème} cas : si $o_{12} < t_{12}$ alors les deux modalités sont dépendantes, associées, liées dans le sens d'une répulsion

3^{ème} cas : si $o_{12} > t_{12}$ alors les deux modalités sont dépendantes, associées, liées dans le sens d'une attirance

Mise en évidence de la structure d'un tableau de contingence par le recours à des couleurs :

1^{er} cas : blanc ; 2^{ème} cas : rouge ; 3^{ème} cas : vert

2. Test du χ^2 d'adéquation

S'il s'agit de comparer la distribution de fréquence d'une variable qualitative ou d'une variable quantitative, observée sur un échantillon, à une distribution théorique connue sur la population. C'est le **test du χ^2 d'adéquation** qui sera mis en pratique. Nous exposerons ce test à l'aide de deux exemples.

2.1. Présentation du premier exemple

Ayant relevé la répartition des diplômes dans un échantillon de 250 personnes représentatif (au sens du sondage aléatoire) de la population française, on cherche à la comparer à une répartition connue de ces mêmes diplômes dans la population entière. On se demande donc, si l'échantillon est représentatif (au sens du modèle réduit) relativement aux diplômes.

<i>Diplômes</i>	<i>Répartition dans la Population</i>	<i>Effectifs de l'échantillon</i>
Aucun ou Certificat d'études	34%	72
BEPC	8%	24
CAP-BEP	25%	65
Bac ou équivalent	11%	25
Diplôme supérieur au Bac	22%	64
	Total : 100%	Total : 250

2.2. La démarche du test du χ^2 d'adéquation

Nous suivons les quatre grandes étapes :

- Étape 1 : formulation d'hypothèses

Les deux hypothèses à prendre en compte sont :

H_0 : Il n'y a pas de différence entre la répartition des effectifs observés dans l'échantillon et la répartition des effectifs théoriques de la population.

H_1 : Il y a une différence entre la répartition des effectifs observés et la répartition des effectifs théoriques.

- Étape 2 : calcul du χ^2 et du nombre de degrés de liberté (ddl)

Le calcul du χ^2 est fait à partir des effectifs observés et des effectifs théoriques. On considère que le total des effectifs théoriques et le total des effectifs observés sont égaux. Ce χ^2 est la statistique du test. On calcule également le nombre de degrés de liberté associé. Ici, il est égal au nombre de modalités diminué de 1. On vérifie également que les effectifs théoriques sont bien supérieurs ou égaux à 5. Dans cet exemple, les effectifs théoriques sont calculés à partir des fréquences connues sur la population appliquées à un échantillon de 250 personnes. Ainsi l'effectif théorique de 85 personnes correspond au 34% de 250 etc..:

Diplômes	Répartition dans la Population	Effectifs théoriques attendus sur un échantillon de taille 250	Effectifs observés sur l'échantillon
Aucun ou Certificat d'études	34%	85	72
BEPC	8%	20	24
CAP-BEP	25%	62,5	65
Bac ou équivalent	11%	27,5	25
Diplôme supérieur au Bac	22%	55	64
	Total : 100%	Total 250	Total : 250

$$\chi^2 = \frac{(72 - 85)^2}{85} + \frac{(24 - 20)^2}{20} + \frac{(65 - 62,5)^2}{62,5} + \frac{(25 - 27,5)^2}{27,5} + \frac{(64 - 55)^2}{55} = 4,58$$

avec ddl=5-1= 4

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

Dans la table statistique du χ^2 , sur la ligne ddl=4, on peut lire que pour un risque α donné de 0,05 (5%), le χ^2 théorique ou lu, noté χ^2_{α} , est égal à 9,49. C'est à dire qu'il n'y a que 5 chances sur 100 pour que la statistique du χ^2 dépasse cette valeur de 9,49.

- Étape 4 : décision statistique

La valeur du χ^2 calculé est inférieure au χ^2 théorique, alors la différence entre les effectifs observés et les effectifs théoriques n'est pas significative, elle résulte plutôt des seules fluctuations d'échantillonnage. Il est possible, avec un risque β inconnu de retenir l'hypothèse H_0 d'absence de différence entre la répartition des diplômes observés dans l'échantillon et des diplômes théoriques de la population.

2.3. Présentation du second exemple d'adéquation à la distribution de fréquences uniforme

Situation-problème paradigmatique : étude de l'équilibrage d'un dé à 6 faces

Il s'agit de se donner une méthode pour décider entre deux hypothèses alternatives concernant l'état d'un dé cubique

H_0 le dé est bien équilibré	H_0 la distribution de fréquences des résultats est uniforme, les résultats sont équiprobables
H_1 le dé est truqué	H_1 au moins un des six résultats a une probabilité différente d'apparition des cinq autres

Pour pouvoir décider, on recourt à une expérience aléatoire.

On jette au hasard et de manière identique 336 fois le dé	Extraction d'un échantillon aléatoire avec remise de taille n=336, de la population constituée par les six faces $P = \{1 ; 2 ; 3 ; 4 ; 5 ; 6\}$
---	---

Description des résultats obtenus sur l'échantillon :

Faces	Face1	Face2	Face3	Face4	Face5	Face6
Effectifs observés	47	49	78	56	60	46

Description des résultats théoriques que l'on peut ESPERER si le dé est parfaitement équilibré, sur cet échantillon

Faces	Face1	Face2	Face3	Face4	Face5	Face6
Effectifs théoriques	56	56	56	56	56	56

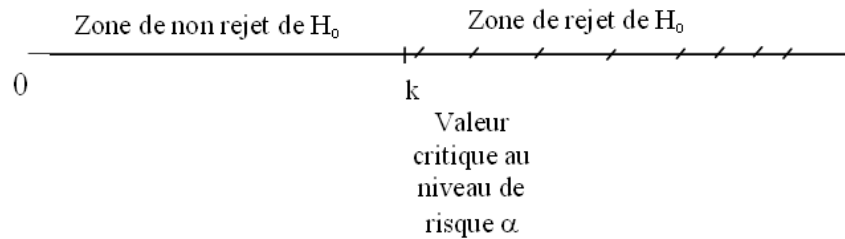
Mesure synthétique D^2 de l'écart entre ces deux tableaux

Faces	Face1	Face2	Face3	Face4	Face5	Face6
Écart exacts	$\frac{(47-56)^2}{56}$	$\frac{(49-56)^2}{56}$	$\frac{(78-56)^2}{56}$	$\frac{(56-56)^2}{56}$	$\frac{(60-56)^2}{56}$	$\frac{(46-56)^2}{56}$
Écart approchés	1,446	0,875	8,643	0	0,286	1,786

Valeur de D^2 dans cette expérience $d2 \approx 13,0357143$

Règle de décision : on utilise la table du Khi-deux (χ^2) de ddl = 6-1= 5

Il faut fixer le niveau de risque α puis lire dans la table. En choisissant un de deux niveaux habituels 5% ou 1%, nous obtenons	0,05	11,07
	0,01	15,08



2.4. Présentation du troisième exemple : test d'adéquation à une distribution de la Laplace-Gauss

Ce test est parfois appelé test de normalité en raison du fait que la variable de Laplace-Gauss est souvent nommée variable normale. Nous avons laissé de côté cette dénomination ambiguë.

Comparer la variable V3 « hauteur du saut. » à une variable de Laplace-Gauss en utilisant la forme du polygone des fréquences et les effectifs théoriques sous l'hypothèse V3 est une variable de Laplace-Gauss. Tester les hypothèses correspondantes (H_0 et H_1) au seuil de 5% avec le test du Khi-Deux d'adéquation.

Nous traitons la variable V3 comme une variable continue.

intervalles	[95;105[[105;110[[110;115[[115;120[[120;125[[125;135]	Totaux
effectifs	2	4	8	6	3	2	25
fréquences	0,08	0,16	0,32	0,24	0,12	0,08	1

Le détail des calculs permettant d'aboutir à ces résultats est fourni après

	<i>paramètres</i>	<i>valeurs</i>	<i>valeurs approchées à utiliser</i>
étendue		40	40 cm
moyenne		114,5	114,5 cm
variance		55	55 cm ²
écart-type		7,41	7,4 cm
Q2 = médiane = second quartile		114,0625	114 cm

Il s'agit de comparer la variable V3 "hauteur du saut.." à une variable de Laplace-Gauss en utilisant la forme du polygone des fréquences et les effectifs théoriques sous l'hypothèse V3 est une variable de Laplace-Gauss. Le polygone des fréquences (lissage de la densité de fréquences) suggère une forme proche de la courbe de Laplace-Gauss. La hauteur

médiane:114 cm, la hauteur modale : 112,5 cm et la hauteur moyenne 114,5 cm confirme par leur proximité cette remarque.

Nous allons procéder à un test d'hypothèse à l'aide du test du Khi-deux d'adéquation.

- Étape 1 : formulation d'hypothèses

L'hypothèse H_0 est alors formulée ainsi :

La variable V_3 « hauteur du saut » est une variable de Laplace-Gauss $LG(\mu, \sigma)$ de paramètres estimés $m = 114,5$ et $s = 7,5$

L'hypothèse alternative H_1 peut être énoncée comme suit :

La variable V_3 « hauteur du saut » n'est pas une variable de Laplace-Gauss

- Étape 2 : calcul du χ^2 et du nombre de degrés de liberté (ddl)

La variable de décision est la variable $D = \chi^2 = \sum_{i=1}^{i=s} \frac{(O_i - T_i)^2}{T_i}$ où O_i = effectif observé et T_i

= effectif théorique sous la contrainte de la loi de référence choisie pour chacune des s classes.

Si aucun des effectifs théoriques n'est inférieur à 5, la variable D suit approximativement une loi du Khi-Deux à $s-1-e$ degrés de liberté (où e désigne le nombre de paramètres estimés parmi les r paramètres ici $e=r=2$ ce sont μ et σ).

Si un des effectifs est inférieur à 5, on procède à des regroupements de classes pour se mettre dans les conditions d'application.

On calcule les effectifs théoriques sous l'hypothèse H_0 puis la valeur de la variable D critère de décision

Les intervalles de la variable centrée réduite $Z = \frac{V_3 - 114,5}{7,5}$	effectifs théoriques	effectifs théoriques après regroupement	effectifs observés après regroupement	effectifs observés	$\frac{(O_i - T_i)^2}{T_i}$
$]-\infty; -2,6[$	0,12			0	
$[-2,6; -1,27[$	2,43	6,86	6	2	0,1078
$[-1,27; -0,6[$	4,31			4	
$[-0,6; 0,07[$	6,34	6,34	8	8	0,4346
$[0,07; 0,73[$	5,98	5,98	6	6	0,00006
$[0,73; 1,4[$	3,8			3	
$[1,4; 2,73[$	1,94	5,82	5	2	0,1155
$[2,73; +\infty[$	0,08			0	
Totaux	25	25	25	25	0,65796

Ainsi la valeur expérimentale de la variable de décision est $d = \chi^2 = \mathbf{0,65796}$

Le calcul du degré de liberté donne $ddl = 4 - 2 - 1 = 1$

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

Choisir la valeur α du risque de première espèce. Rechercher dans une table du Khi-Deux à $s-1-e$ degrés de liberté, la valeur critique k telle que $\text{Prob}\{\chi^2_{(s-1-e)} > k\} = \tilde{\alpha}$

Au seuil de risque de 0,05 nous pouvons lire la valeur critique $k = 3,84$

- Étape 4 : décision statistique

Ayant calculé la valeur empirique de D avec les s valeurs O_i observées sur l'échantillon extrait.

Si $d < k$, c'est à dire si $d \in A = [0, k]$ région d'acceptation, alors on ne rejette pas l'hypothèse nulle H_0 , on accepte alors H_0 en prenant un risque β d'erreur de seconde espèce

Si $d > k$, c'est à dire si $d \in K = [k, +\infty[$, région critique, alors on rejette l'hypothèse nulle H_0 et on accepte H_1 .

La valeur expérimentale $d^2 \approx 0,657$ obtenue sur l'échantillon est inférieure à $k= 3,84$. Par conséquent, nous ne sommes pas dans les conditions de rejet de l'hypothèse H_0 . Il nous est possible d'affirmer qu'il est raisonnable de tenir la variable V_3 pour une variable de Laplace-Gauss $LG(114,5 ; 7,5)$. Toutefois cette affirmation nous place dans la situation d'une prise de risque de seconde espèce de valeur β inconnue.

3. Test du χ^2 de Mac Nemar

Si l'on cherche à comparer deux séries successives de données de type présence-absence ou échec-réussite relevées sur le même échantillon, on pourra utiliser le test du χ^2 de Mac Nemar

Si l'on veut comparer la difficulté de deux épreuves scolaires passées par le même groupe d'individus (séries dites appariées), on peut résumer le problème par l'étude d'un tableau croisé simple à 2 lignes et 2 colonnes, à condition de ne s'intéresser qu'à l'échec ou à la réussite.

Mac Nemar a montré que le test du χ^2 d'indépendance n'est pas approprié car il ne faut prendre en compte que les cas de discordances entre les épreuves, c'est à dire le cas de réussite à l'une et d'échec à l'autre et son complémentaire. Cela revient à savoir si ces deux nombres sont ou non égaux.

Prenons un exemple :

50 élèves passent deux épreuves A et B. Le tableau suivant résume l'ensemble des résultats :

		Épreuve B : Réussite	Échec	Total
Épreuve A:				
Réussite		16	12	28
Échec		10	12	22
Total		26	24	50

Les fréquences de réussite aux deux épreuves sont-elles différentes significativement ?

Nous présentons dans le tableau ci-dessous les notations de ce problème :

		Épreuve 2		Total
		Réussite	Échec	
Épreuve 1	Réussite	N1	N2	N1 + N2
	Échec	N3	N4	N3 + N4
Total		N1 + N3	N2 + N4	N1 + N2 + N3 + N4

Il s'agit de comparer N_2 et N_3 , ou bien de comparer $N_2 / (N_2 + N_3)$ à $1/2$ dans l'hypothèse d'une équivalence entre les épreuves. On compare une fréquence observée à une fréquence théorique de $1/2$.

Mac Nemar a montré qu'il suffisait de calculer la valeur du χ^2 suivante :

$$\chi^2_{(\text{Mac Nemar})} = \frac{(N_2 - N_3)^2}{N_2 + N_3}$$

Cette valeur ne peut se calculer que pour un tableau de 4 cases. De plus le test requiert également que $N_2 + N_3$ soit au moins égale à 10

Les 4 étapes de la démarche du test sont identiques aux précédentes :

- Étape 1 : formulation d'hypothèses
 H_0 : égalité des changements d'état entre les deux épreuves
 H_1 : Non-égalité des changements d'état entre les deux épreuves
- Étape 2 : calcul du $\chi^2_{(\text{Mac Nemar})}$ et du nombre de degrés de liberté (ddl)

Le calcul du $\chi^2_{\text{(Mac Nemar)}}$ donne : $\chi^2_{\text{(Mac Nemar)}} = (12-10)^2 / (12 + 10) = 4 / 22 = 0,18$

Quant au nombre de degrés de liberté (ddl) il est donc toujours de 1 dans le cas de cette statistique..

- Étape 3 : lecture du χ^2 théorique pour un risque α donné

La lecture du χ^2_t dans la table du χ^2 pour un risque α donné, sur la ligne ddl=1 permet de conclure. Ainsi, si l'on prend un risque α de 5%, la valeur théorique au-delà de laquelle l'hypothèse H_0 sera rejetée est 3,84. Pour un risque α de 1%, elle vaut 6,63.

- Étape 4 : décision statistique

Comme $0,18 < 3,84$ Nous ne rejetons pas l'hypothèse nulle. Les deux épreuves sont équivalentes pour un risque de 2^{ème} espèce β inconnu.

Variables statistiques et biais

Exemple des recherches dans le domaine de la santé

Les **indicateurs de santé** sont des variables qui peuvent être mesurées directement et qui permettent de décrire l'état de santé des individus d'une communauté. Ils sont utilisés pour la mise au point d'indices plus complexes établis selon des formules spécifiques.

1. A propos des biais

D'après *Biostatistique Clinique, Epidémiologie et Essais cliniques*. ©Faculté de Médecine Necker -Enfants Malades 2002

Un **bias dans une enquête** épidémiologique désigne tout effet qui altère la représentativité des résultats. Il est caractérisé par une erreur systématique sur la représentation d'un effet. Il entraîne que la mesure de la relation maladie-facteur d'exposition, sur la population étudiée n'est pas égale à la mesure de cette relation dans la population-cible. Il existe de nombreuses sources de biais que l'on peut schématiquement regrouper en 3 groupes :

1.1. Les biais de sélection

Ce type de biais réfère à une distorsion dans l'estimation d'un effet résultant de la façon par laquelle les sujets de la population étudiée ont été sélectionnés. Il peut concerner :

- les erreurs sur le choix des groupes à comparer dans tous les types d'enquêtes,
- les erreurs dans le choix du mode d'échantillonnage dans les enquêtes cas-témoins ou dans les études transversales,
- le biais de Berkson dans les études cas-témoins lorsque les cas et les témoins sont issus d'une population hospitalière non représentative de la population générale,
- les sujets perdus de vue dans les études de cohortes et les sujets qui ne répondent pas aux demandes des enquêteurs, répétées au fil du temps, et indispensables au suivi d'une cohorte,
- la survie sélective dans les enquêtes cas-témoins ou dans les études transversales. La survie sélective concerne les profils différentiels de mortalité des cas et des témoins qui pourraient être à tort ignorés dans une étude rétrospective ou dans une étude transversale,
- les biais de détection dans les études cas-témoins, quand la procédure d'identification de la maladie varie avec l'exposition. Un exemple est illustré par la relation entre le cancer de l'endomètre et l'utilisation des oestrogènes. Ceux-ci peuvent entraîner des ménorrhagies qui poussent les femmes à consulter et augmenter ainsi la possibilité de diagnostiquer un cancer de l'endomètre chez les sujets exposées par rapport aux non-exposées.

Ainsi il existe de très nombreuses sources de biais de sélection dont il faut tenir compte à l'élaboration du protocole des enquêtes, et particulièrement dans les enquêtes cas-témoins.

1.2. Les biais de classification

Ils concernent une distorsion ou erreur systématique dans l'estimation d'un effet quand la mesure de la condition d'exposition ou de la maladie est systématiquement impropre. Ce biais représente des erreurs d'information qui peuvent conduire à une classification impropre des

sujets aussi bien sur la maladie que sur le facteur d'exposition. Les origines les plus importantes sont liées à :

- l'utilisation d'appareils de mesure défectueux ou improprement réglés qui conduisent à des erreurs systématiques de classification,
- des critères diagnostiques impropres pour définir la maladie,
- des omissions ou des imprécisions sur des données enregistrées dans le passé,
- une surveillance inégale des sujets exposés et des sujets non exposés dans les études de cohortes.

Il faut prévenir ces biais de classification pour obtenir une meilleure estimation de la relation maladie-facteur d'exposition.

1.3. Les facteurs de confusion

La relation entre 2 variables peut être affectée par une troisième variable. Un facteur de confusion représente une variable qui est associée aussi bien au facteur étiologique vrai qu'à la maladie. Par exemple l'âge est un facteur de confusion dans l'étude de la relation entre tabagisme et cancer bronchique.

Nous avons vu précédemment qu'il fallait tenir compte des biais potentiels de sélection ou de classification au moment de l'élaboration du protocole d'une enquête. Il est illusoire, voire impossible, d'en tenir compte au moment de l'analyse.

Pour les facteurs de confusion plusieurs méthodes permettent de les prendre en compte soit lors de l'élaboration du protocole, soit au moment de l'analyse. A l'élaboration du protocole il s'agit de la *stratification*.

La stratification consiste dans un échantillon de malades et de témoins à former des classes de sujets par rapport aux facteurs de confusion. L'âge et le sexe sont deux facteurs de confusion pour l'étude de la relation entre tabac et cancer broncho-pulmonaire. On classera les sujets par tranche d'âge et par sexe.

L'*appariement* consiste à neutraliser les facteurs de confusion en groupant les sujets de telle sorte que ceux d'un même groupe partagent le(s) même(s) facteur(s) de confusion. Dans l'exemple précédent, chaque cas sera apparié avec un (ou plusieurs) témoin(s) de même âge et de même sexe. L'appariement n'est qu'une modalité particulière de stratification. Lors de l'analyse on peut utiliser des techniques statistiques dites d'*ajustement*. L'ajustement est un procédé qui vise à éliminer d'une comparaison de série d'observations le lien entre un effet et une ou plusieurs causes autres que celles qui sont le sujet propre de l'étude. La prévention des biais ou leur recherche lors de l'analyse constituent des étapes importantes dans une enquête afin de juger de la causalité entre un ou plusieurs facteurs de risque et une maladie.

1.4. L'identification des biais

Afin d'identifier des biais potentiels, il est bon de se poser quelques questions :

- la population de l'étude a-t-elle été bien définie ?
- est-ce que la population étudiée représente de manière adaptée la population cible ? (la population cible est la population pour laquelle on souhaitera généraliser les résultats de l'étude).
- les définitions de la maladie et de l'exposition sont-elles claires ?
- la définition des cas est-elle précise ?
- quels sont les critères d'inclusion et d'exclusion ?
- les contrôles représentent-ils de manière adéquate la population dont sont issus les cas ?
- l'identification ou la sélection des cas ou des contrôles a-t-elle pu être influencée le statut

- d'exposition ?
- les cohortes sont-elles similaires à l'exclusion du statut de l'exposition ?
- les mesures sont-elles aussi objectives que possibles ?
- l'étude est-elle réalisée le plus en aveugle possible ?
- le suivi est-il adapté ?
- le suivi est-il identique pour toutes les cohortes ?
- l'analyse est-elle appropriée ?
- l'interprétation qui en est faite est-elle étayée par les résultats ?

2. L'enquête par entretien et quelques biais possibles

En considérant comme biais, tout fait susceptible de diminuer la qualité de représentativité de la réalité, d'un élément choisi par le chercheur, il est clair que dans le cadre de l'enquête par entretien, le chercheur lui-même peut être un inducteur de tel facteur.

En tant que procédure donnant au chercheur un accès à des expériences particulières de l'informateur, l'entretien est sensible aux biais. (Sudman & Bradburn 1974) Les caractéristiques personnelles du chercheur-enquêteur telles que sexe, âge, classe sociale, appartenance religieuse, traits de personnalité, etc., viennent se confronter à celles du sujet-informateur avec lesquelles elles interagissent. Neuman propose six catégories de biais susceptibles d'apparaître à l'occasion d'un entretien :

- Les erreurs du répondant issues des oublis, de la mauvaise compréhension des questions, de la gêne éprouvée, des mensonges liés à la présence d'autrui.
- Les erreurs non intentionnelles dues à la négligence du chercheur-enquêteur comme la rencontre d'un informateur ne correspondant pas à la recherche, la mauvaise lecture d'une question, la modification non pertinente de l'ordre des questions, la mauvaise compréhension de la réponse de l'informateur, la prise en compte d'une réponse non pertinente ou incorrecte fournie à une question par l'informateur.
- Les altérations intentionnelles du chercheur-enquêteur qui modifie les réponses fournies, en oublie certaines, ou reformule des questions.
- Les influences directement dues au chercheur-enquêteur comme son apparence, son ton de voix, son attitude, ses réactions aux réponses, ses commentaires effectués hors du contexte de l'entretien, etc..
- Les influences dues aux attentes du chercheur-enquêteur en fonction de l'apparence, de la situation de vie de l'informateur ou de ses réponses préalables.
- Les erreurs dues à une exploitation insuffisante ou défectueuse des résultats de l'exploration.

Références :

Neuman, W. L. (1991). *Social research methods : Qualitative and quantitative approaches*. Boston, MA : Allyn and Bacon

Sudman, S., & Bradburn, N.M. (1974). *Response effects in surveys*. Chicago : Aldine.

Liaison entre deux facteurs. Paradoxe de Simpson

Lucette Carter (Université Paris 10 – Nanterre) et Jean-Louis Piednoir (Inspecteur Général)

1. L'analyse de tables de contingence permet de mettre en évidence la liaison pouvant exister entre deux facteurs.

Les exemples qui suivent sont issus d'une étude analysant les performances scolaires d'enfants entrés au CP en 1978 et âgés de 11 ans au moment de l'étude (Jacqueline Goapper, Service de la Prévision, des Statistiques et de l'Évaluation. Ministère de l'Éducation Nationale). La performance est un facteur à plusieurs modalités que nous regrouperons ici en 2 modalités : Bon parcours (correspondant à un redoublement au plus), parcours difficile (plus d'un redoublement ou enseignement spécialisé).

Le tableau ci-dessous répartit la population étudiée selon la performance et la nationalité de l'enfant (Française ou Étrangère) :

Tableau 1 :

	Bon parcours	Parcours difficile
Français	15348	1685
Étranger	1372	345

La fréquence d'un bon parcours est égale à 0,90 pour les élèves français alors qu'elle n'est égale qu'à 0,80 pour les élèves étrangers. Mais on sait que la catégorie socio-professionnelle de la famille est un facteur explicatif de la réussite scolaire (voir, par exemple, les études de Bourdieu et Passeron dans « Les héritiers »). On va donc introduire cette variable et comparer les réussites dans les différentes catégories socioprofessionnelles. On sait, par ailleurs, que les étrangers sont plus proportionnellement plus nombreux que les français dans les CSP les plus modestes.

A titre d'exemple, le tableau ci-dessous qui concerne uniquement la catégorie « ouvriers non qualifiés » :

Tableau 2 : Ouvriers non qualifiés

	Bon parcours	Parcours difficile
Français	2594	551
Étranger	766	213

Les fréquences de « Bon parcours » valent respectivement :

$Fr (BP | F) = 0,82$ pour les élèves français

$Fr (BP | E) = 0,78$ pour les élèves étrangers.

On constate, que la prise en compte de la CSP ne modifie pas le sens de la liaison entre performance et nationalité mais que l'écart entre le niveau de performance entre élèves français et étrangers diminue (l'auteur de l'étude remarque que ce phénomène est dû à un effet de structure de population).

A la suite de la publication de ces résultats, on a émis l'hypothèse que la différence de réussite était essentiellement due à un déficit linguistique que l'on pouvait diminuer par de la remédiation en français. Cette interprétation fut partiellement infirmée par l'introduction d'un troisième facteur correspondant à la taille de la fratrie.

On va voir que l'introduction du facteur correspondant au nombre d'enfants de la famille modifie le sens de la liaison entre performance et nationalité : dans les familles nombreuses, les élèves étrangers ont de meilleures performances scolaires que les élèves français.

A titre d'exemple, prenons la catégorie « Ouvriers non qualifiés ». Les deux tableaux ci-dessous correspondent à la partition du tableau 2 en deux sous-tableaux correspondant au nombre d'enfants (« 3 enfants au plus » et « plus de 3 enfants » :

Tableau 3 « familles ayant au plus 3 enfants »

	Bon parcours	Parcours difficile
Français	1879	250
Etranger	329	77

Tableau 4 « familles ayant plus de 3 enfants »

	Bon parcours	Parcours difficile
Français	715	301
Etranger	437	136

Les fréquences de « bon parcours » sont égales à :

$Fr(BP | F, \leq 3 ENF) = 0,88$ pour les français de familles peu nombreuses ;

$Fr(BP | E, \leq 3 ENF) = 0,81$ pour les étrangers de familles peu nombreuses ;

$Fr(BP | F, >3 ENF) = 0,70$ pour les français de familles nombreuses.

$Fr(BP | E, >3 ENF) = 0,76$ pour les étrangers de familles nombreuses.

Comment interpréter ces différences ? Une hypothèse provisoire a été formulée. Dans les familles nombreuses, l'enfant a plus de difficultés à apprendre efficacement ses leçons et à faire ses devoirs. Cette difficulté serait plus accentuée dans les familles françaises. Dans ce contexte, en termes de remédiation, des études surveillées seraient probablement efficaces.

Les données de cette étude mettent bien en évidence la nécessité d'introduire ce que l'on appelle « facteurs de confusion », facteurs pouvant modifier le sens ou le degré de liaison existant entre deux facteurs donnés. Le paradoxe de Simpson correspond à un cas extrême où le sens de la liaison entre deux facteurs donnés est inversé pour tous les niveaux d'un troisième facteur. Un exemple de ce paradoxe est présenté dans une étude d'homicides aux Etats-Unis (Radelet, M. 1981), répartis suivant l'ethnie de l'accusé (Noir, Blanc), le type de condamnation (Peine de mort, Autre peine) et l'ethnie de la victime (Noir, Blanc). Voici les données :

Table 5 : « Victime noire »

	Peine de mort	Autre peine
Accusé Noir	6	97
Accusé Blanc	0	9

Table 6 : « Victime blanche »

	Peine de mort	Autre peine
Accusé Noir	11	52
Accusé Blanc	19	132

Dans chaque cas (victime noire et victime blanche), la fréquence de peine de mort est plus élevée lorsque l'accusé est noir. Par contre, si l'on ne tient plus compte de l'ethnie de la victime, le calcul des fréquences (à partir du tableau agrégé) montre que la liaison s'inverse et que la fréquence de peine de mort est plus élevée lorsque l'accusé est blanc. Le paradoxe s'explique ainsi : on applique davantage la peine de mort lorsque la victime est blanche ; or il y a une liaison positive importante entre l'ethnie de l'accusé et celle de la victime (lorsque la victime est blanche, l'accusé est blanc dans 58% des cas ; lorsque la victime est noire, l'accusé est noir dans 92% des cas). On remarque qu'en anglais, le facteur modifiant la liaison entre deux facteurs donnés (dans l'exemple du paradoxe de Simpson, l'ethnie de la victime) s'appelle « effect modifier factor » ce qui est plus approprié que le terme « facteur de confusion » généralement utilisé en épidémiologie.

Bases théorique pour l'étude statistique des rangs

Jean-Claude Régnier

1. Le coefficient de corrélation des rangs R_s de Spearman

A chaque observation, nous associons son rang k dans l'ordre du recueil des observations et son rang R_k dans l'ordre de l'échantillon.

Rangs des observations	1	2	...	k	...	n
Rangs dans l'échantillon ordonné	R1	R2	...	Rk	...	Rn

La statistique est

$$S_n = R_s = 1 - \frac{\sum_{k=1}^{k=n} [R_k - k]^2}{n(n^2 - 1)}$$

l'espérance de R_s est $E(R_s) = 0$

la variance de R_s est $V(R_s) = \frac{1}{n-1}$

Pour tester la significativité de la valeur obtenue, on prend en référence la situation théorique d'indépendance des deux rangements dans la population, c'est à dire que les $n!$ classements sont équiprobables. Dans ce cas de l'indépendance, la valeur est $\rho_s = 0$. Cependant il convient de rappeler que la réciproque est fautive.

Dans le cas d'une tendance monotone croissante parfaite, les classements sont identiques

$$R_k = k$$

$\rho_s = 1$: les deux classements sont identiques

Dans le cas d'une tendance monotone décroissante parfaite, les classements sont inversés

$$R_k = n + 1 - k$$

$\rho_s = -1$: les deux classements sont inverses

Il s'agit alors de prendre une décision sur la base d'une valeur r_s , réalisation de la variable R_s sur un échantillon. A l'aide de la table du coefficient de corrélation de Spearman ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle H_0 postulant l'indépendance des deux rangements. La région critique W , région de rejet de H_0 au profit de H_1 , est définie selon H_1 , l'une des trois hypothèses alternatives usuelles.

H1 tendance monotone croissante dépendance positive $W = \{ r_s, r_s > c \}$	H1 tendance monotone dépendance quelconque $W = \{ r_s, r_s > c \}$	H1 tendance monotone décroissante dépendance négative $W = \{ r_s, r_s < -c \}$
---	---	--

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W/H_0) = \alpha$

Asymptotiquement la variable $Z_s = R_s \sqrt{n-1}$ suit la loi de la variable de Laplace-Gauss $LG(0,1)$. Cette approximation est jugée acceptable pour $n > 30$.

Asymptotiquement la variable $T_s = R_s \sqrt{\frac{n-2}{1-R_s^2}}$ suit la loi de la variable de Student à $n-2$ ddl. Cette approximation est jugée acceptable pour $n > 10$.

2. Test de concordance de p rangements de n objets de M.G. Kendall

Objets	1	2	...	i	...	n	
Critères							totaux
1	R11	R21		Ri1		Rn1	R.1
2	R12	R22		Ri2		Rn2	R.2
...							
j	R1j	R2j		Rij		Rnj	R.j
...							
p	R1p	R2p		Rip		Rnp	R.p
Totaux	R1.	R2.		Ri.		Rn.	r..

2.1. Méthode de calcul:

Chaque ligne est une permutation des nombres entiers de 1 à n dont la somme est constante et vaut $\frac{n(n+1)}{2}$. Ainsi $r.. = p \frac{n(n+1)}{2}$. Dans l'hypothèse d'une concordance parfaite, les totaux des colonnes seraient égaux respectivement à p, 2p, 3p, ..., np à une permutation près. On utilise alors la statistique $S_K = \sum_{i=1}^{i=n} (R_i - \frac{r..}{n})^2$ dont la valeur maximale est $S_{\max} = \frac{p^2(n^3 - n)}{12}$

Le coefficient de concordance de Kendall est:

$$W = \frac{S_K}{S_{\max}} = \frac{S_K}{\frac{1}{12} p^2 (n^3 - n)}$$

On peut aussi $W = \frac{1}{p} + \frac{2}{p^2} \sum_{i=1}^{i=p-1} \sum_{j=i+1}^j=p R_{ij}$ D'où

- l'espérance de W : $E(W) = \frac{1}{p}$

- la variance de W : $V(W) = \frac{2(p-1)}{p^3(n-1)}$

$w = 0$ correspond au cas où chaque colonne a même total. De faibles valeurs de W suggèrent l'indépendance des classements.

Pour tester l'hypothèse nulle H_0 d'indépendance des p rangements, on procède selon les procédures suivantes:

- Pour $n \leq 7$ on utilise une table,

- Pour $n \leq 7$ et $2 < p \leq 20$ la variable $\frac{(p-1)W}{1-W}$ est distribuée comme la variable de Fisher-

Snédécour $F(n_1 = n-1-\frac{2}{p}; n_2 = (p-1)(n-1-\frac{2}{p}))$

- Pour $n > 7$ on utilise la variable $p(n-1)W$ qui est une variable de Pearson χ^2_{2n-1}

Dans le cas où l'on est conduit à rejeter l'hypothèse d'indépendance des classements, on utilise souvent la règle de classement suivante:

les objets sont classés dans l'ordre défini par les totaux des colonnes.

Lorsqu'il y a des *ex aequo* on remplace le rang de ceux-ci par la moyenne arithmétique des rangs qu'ils auraient eu sans *ex aequo*.

$$W = \frac{S_k}{S_{\max}} = \frac{S_k}{\frac{1}{12} p^2 (n^3 - n) - p \sum_{j=1}^{j=p} (t_j^3 - t_j)}$$

avec t_j = nombre d'*ex æquo* au $j^{\text{ème}}$ classement

Cette table fournit les valeurs critiques k telles que $P(W \geq k) = \alpha = 0.05$

3. Le test de Mann et Whitney

Rangs des observations X dans l'échantillon global ordonné de taille $N = m + n$	R1	R2	...	Rk	...	Rm	
Rangs des observations Y dans l'échantillon global ordonné de taille $N = m + n$	Q1	Q2	...	Qk	Qn

Ce test est fondé sur la statistique suivante dérivée de celle de Wilcoxon :

$$SN = U_N = W_N - \frac{m(m+1)}{2} = \sum_{k=1}^{k=m} R_k - \frac{m(m+1)}{2}$$

on peut en déduire que :

- U_N varie entre 0 et mn

- l'espérance est $E(U_N) = \frac{mn}{2}$

- la variance est $V(U_N) = V(W_N) = \frac{mn(m+n+1)}{12}$

3.1. Conditions d'utilisation:

Etant donnés les deux échantillons indépendants $(X_1, X_2, X_3, \dots, X_m)$ et $(Y_1, Y_2, Y_3, \dots, Y_n)$ issus de deux populations P_1 et P_2 .

On mélange ces deux échantillons et on réordonne les valeurs.

On dénombre les couples (X_p, Y_q) tels que X_p a un rang plus grand que Y_q

U_N est la variable qui à chaque situation associe ce nombre. Elle varie entre 0 et nm selon les deux cas extrêmes:

$x_1, x_2, x_3, \dots, x_m, y_1, y_2, y_3, \dots, y_n$ et $y_1, y_2, y_3, \dots, y_n, x_1, x_2, x_3, \dots, x_m$,

Sous l'hypothèse de l'identité des distributions des deux variables X et Y , la loi exacte de U_N peut être calculée pour de faibles valeurs de n et de m . Toutefois dès que $n > 8$ et $m > 8$ on peut l'approcher par une loi de Laplace-Gauss.

$$\frac{U_N - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \approx LG(0;1)$$

3.2. Statistique et variable de décision

Pour plus de faciliter, on utilise de façon intermédiaire la variable WX , somme des rangs de la variable X , puis on calcule $U_N = WX - \frac{m(m+1)}{2}$

3.3. Test bilatéral :

H_0 (identité des deux distributions $F = G$) contre H_1 (Les deux distributions sont différentes $F \neq G$)

3.4. Test unilatéral :

Ho (identité des deux distributions $F = G$) contre H1 (Les deux distributions sont différentes soit $F < G$, soit $F > G$)

Il s'agit alors de prendre une décision sur la base d'une valeur u_N , réalisation de la variable U_N sur un échantillon. A l'aide d'une table ou d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss, on peut fonder le rejet ou le non-rejet au seuil α de l'hypothèse nulle Ho postulant l'identité des deux distributions. La région critique W, région de rejet de Ho au profit de H1, est définie selon H1, l'une des trois hypothèses alternatives usuelles.

H1 $F > G$	H1 $F \neq G$	H1 $F < G$
$W = \{ u_N, u_N \leq c_1 \}$	$W = \{ u_N, u_N \leq c_1 \text{ ou } u_N \geq c_2 \}$	$W = \{ u_N, u_N \geq c_2 \}$

La valeur critique c ($c \geq 0$) s'obtient en résolvant l'équation $\text{Prob}(W/H_0) = \alpha$

A l'aide d'une formule d'approximation permettant l'usage de la loi de la variable de Laplace-Gauss $\frac{U_N - E(U_N)}{\sigma(U_N)}$ tend vers LG(0,1)

Dans ce cas, il convient de tenir du passage d'une variable discrète à une variable continue en réalisant une correction (*correction de continuité*) selon la démarche suivante : $\text{Prob} \{ U_N = s \} = \text{Prob} \{ s - \frac{1}{2} < U_N < s + \frac{1}{2} \}$ La valeur critique au seuil α s'obtient alors par la résolution d'équation

4. Test d'homogénéité

4.1. Comparaison de k échantillons indépendants Test H de Kruskal-Wallis

Rangs des observations X_1 dans l'échantillon global ordonné de taille N	R_{11}	R_{12}	...	R_{1k}	...	R_{1n1}	
Rangs des observations X_2 dans l'échantillon global ordonné de taille N	R_{21}	R_{22}	...	R_{2k}	...	R_{2n2}	
Rangs des observations X_p dans l'échantillon global ordonné de taille N	R_{p1}	R_{p2}	...	R_{pk}	...	R_{pnp}	
Rangs des observations X_k dans l'échantillon global ordonné de taille N	R_{k1}	R_{k2}	...	R_{kk}	...	R_{kkn}	

4.2. Conditions d'utilisation:

étant donnés les k échantillons indépendants respectivement de taille n_1, n_2, \dots, n_k issus de k populations $P_1, P_2 \dots P_k$. La variable étudiée est une variable ordinaire

4.3. Statistique et variable de décision

On mélange ces k échantillons et on réordonne les $N = \sum_{p=1}^{p=k} n_p$ valeurs.

On prend en compte le rang de chaque observation dans le classement global

Sous l'hypothèse H_0 de l'identité des k distributions de la variable ordinale, les rangs sont distribués au hasard dans chaque échantillon. Considérons:

- la variable S_p = somme des rangs des n_p observations de l'échantillon n° p

- la variable "rang moyen" est alors: $\frac{S_p}{n_p}$

- l'espérance de cette variable "rang moyen" sous H_0 :

$$E\left(\frac{S_p}{n_p}\right) = \frac{1}{n_p} E(S_p) = \frac{1}{n_p} \left(n_p \frac{N+1}{2}\right) = \frac{N+1}{2}$$

On mesure l'écart entre les résultats attendus sous l'hypothèse H_0 et les observations par la variable H :

$$H = \frac{1}{N(N+1)} \sum_{p=1}^{p=k} n_p \left(\frac{S_p}{n_p} - \frac{N+1}{2}\right)^2$$

La distribution exacte de H s'obtient par le dénombrement des m configurations équiprobables

$$m = \frac{N!}{n_1! n_2! \dots n_k!}$$

L'espérance de la variable H : $E(H) = k-1$

La variance de la variable H : $V(H) = 2(k-1) - \frac{2(3k^2 - 6k + N(2k^2 - 6k + 1))}{5N(N+1)} - \frac{6}{5} \sum_{p=1}^{p=k} \frac{1}{n_p}$

Cependant la distribution de H peut être approchée par la distribution de la variable de Pearson à ddl = $k-1$.

4.4. Test unilatéral : H_0 (identité des K distributions) contre H_1 (deux distributions au moins sont différentes)

On choisit un niveau de risque de 1ère espèce α . A ce seuil, on détermine la valeur critique c :

- soit à l'aide d'une table du H de Kruskal et Wallis
- soit à l'aide de la table de la variable de Pearson à ddl = $k-1$

Dans ce cas, on détermine la valeur c telle que $\Pr \text{ob}[\chi^2_{(k-1)} > c] = \alpha$.

On détermine la valeur expérimentale h de H par l'une des deux possibilités décrites ci-dessus: **si** $h > c$ alors **on rejette H_0** en prenant un risque de première espèce de niveau α **sinon on ne rejette pas H_0** , ce qui revient à accepter l'hypothèse H_1 en prenant un risque de seconde espèce β .

4.5. En cas d'existence d'*ex aequo*

En cas d'*ex aequo* chaque observation reçoit un rang égal à la moyenne des rangs qu'elles occupent. Pour chaque groupe de q observations *ex aequo*, on pose $Q = (q-1)q(q+1)$. On

calcule la somme $\sum_{i=1}^{i=r} Q_i$ des valeurs Q obtenues pour les r groupes d'*ex aequo*. Et on utilise la variable $H^* = \frac{\sum_{i=1}^{i=r} Q_i}{1 - \frac{\sum_{i=1}^{i=1} Q_i}{n(n-1)}}$

Lire un article d'une revue scientifique

La lecture de cet extrait d'article mobilise à l'évidence des compétences du domaine de la statistique pour s'assurer de la compréhension et contrôler la validité des énoncés de conclusion

Yanakou Koffiwai GBATI ¹

Climat affectif familial et réussite scolaire : étude auprès d'élèves de cours moyens première année à Lomé

Résumé :

Cet article est une première tentative de recherche de liens entre l'affectivité familiale et la réussite scolaire des élèves dans un milieu africain en particulier à Lomé (TOGO).

L'étude a porté sur des élèves des classes de CMI à partir d'un questionnaire sur les relations parents-enfants.

Il ressort des investigations que le résultat scolaire est lié au climat affectif dans lequel vit l'enfant: le climat affectif positif va de pair avec le bon rendement scolaire même dans les conditions matérielles difficiles d'existence. Néanmoins, dans un climat affectif défavorable, les résultats des garçons sont supérieurs à ceux des filles.

Abstract :

This article is a first attempt of the research of relationships between the familial affection and the school success of the pupils in an African environment, particularly in Lomé (TOGO).

The study is carried on the pupils in the primary school, class five, based on questions about the relationship parents-children.

According to the investigations, school result is linked with the affective environment where the child lives: the positive affective environment goes with the good school yield despite difficult material conditions. However, in a defavorable affective environment, the boys results are superior to those of the girls.

Mots clefs : climat affectif familial, climat affectif positif, climat affectif négatif, réussite et échec scolaires.

Key words : *Familial affective environment, positive affective environment, negative affective environment, school success and failure.*

- Age des enfants et climat affectif

Tableau 5 : Répartition des réponses en fonction de l'âge et du climat affectif

Age	Affectivité	Affectivité positive		Affectivité négative		Total	
		Effectif	%	Effectif	%	Effectif	%
9 – 11 ans		33	63,46	19	36,54	52	100
12 – 14 ans		54	52,43	49	47,57	103	100
15 – 18 ans		14	31,11	31	68,89	45	100
χ^2 cal = 10,38		ddl = 2	P = . 01 DS	C = . 22		200	

Le tableau 5 nous indique que l'âge de l'enfant est liée au type de relation affective avec les parents. Les plus jeunes bénéficient d'une affectivité plus positive que les plus âgés ($P < . 01$).

- Taille de la famille et climat affectif

Tableau 4 : Répartition des réponses en fonction de la taille de la famille

Famille	Affectivité	Affectivité positive		Affectivité négative		Total	
		Effectif	%	Effectif	%	Effectif	%
Famille réduite		73	63,00	43	37,00	116	100
Famille nombreuse		28	33,33	56	66,70	84	100
χ^2 cal = 17,06		ddl = 1	P < . 001 DS	C = . 28		200	

Il ressort de ce tableau que la taille de la famille est liée au climat affectif dans lequel vit l'enfant : les relations affectives parents-enfants sont davantage négatives dans les familles nombreuses que dans les familles restreintes ($P < . 001$). D'autre part, le climat affectif familial apparaît plus corrélé à la taille de la famille ($. 28$) qu'à la classe sociale

5.3. Facteurs sociologiques, climat affectif et réussite scolaire

- Catégorie sociale, climat affectif et résultats scolaires

Tableau 8 : Répartition des réponses selon la catégorie sociale, le climat affectif et les résultats scolaires

Résultat	Affectivité positive				Affectivité négative				TOTAL		
	Réussite		Echec		Réussite		Echec		Eff	%	
	Eff	%	Eff	%	Eff	%	Eff	%			
Classe sociale											
MF	18	45	10	25	11	27,5	1	2,5	40	100	
MD	28	17,5	45	28,2	6	3,7	81	50,6	160	100	
χ^2 cal = 5,47					ddl = 1					P < . 05 DS	
					χ^2 cal = 47,7					ddl = 1	
					Co					P < . 001 DS	
										200	

Il ressort du tableau 8 que, quel que soit le climat affectif, les élèves des milieux favorisés réussissent plus que ceux des milieux défavorisés. Néanmoins, lorsque le climat affectif est négatif, les disparités entre les milieux sont encore plus importantes : les élèves des milieux défavorisés échouent davantage que ceux des milieux favorisés (50,6 % contre 2,5 %).

On peut donc déduire que lorsque le climat affectif est positif, la classe sociale a un effet moins important, mais quand le climat affectif est défavorable, la classe sociale a un effet additif plus prépondérant.