Université BORDEAUX 1 I. F. E.

Cours du Diplôme d'Etudes Approfondies de Didactique des Sciences

Fiches de

S T A T I S T I Q U E S
N O N
P A R A M E T R I Q U E S
pour la
D I D A C T I Q U E

par

Guy BROUSSEAU

Laboratoire Aquitain de Didactique des Sciences et des Techniques
Université BORDEAUX 1
et
Institut Universitaire de Formation des Maitres d'Aquitaine

1993

2014
Nouvelle Edition remaniée, augmentée et corrigée

Remerciements

Jean Louis **OYALLON** a effectué la transcription sur ordinateur d'un grand nombre des formules et de tableaux de ce cours et participé à la saisie de certaines leçons. Sa compétence informatique, son aide et ses encouragements ont été décisifs à un moment difficile.

Michèle **ARTIGUE** et Régis **GRAS** m'ont tous les deux encouragé, avec beaucoup de gentillesse, à rassembler et à corriger mes texte. Ils m'ont fait de nombreuses suggestions et des remarques précieuses dont j'espère avoir tenu compte.

Nadine **BROUSSEAU**, comme toujours, m'a aidé à relire le texte et à trouver des conditions favorables pour achever ce travail.

Qu'ils reçoivent tous mes remerciements les plus vifs.

Résumé des intentions de l'ouvrage (1994)

Le présent ouvrage a pour objet de présenter les méthodes statistiques non paramétriques qui peuvent être utilisées pour étudier des questions de didactique sur des populations faibles, de l'ordre d'une classe ou d'une école.

Ces méthodes trouveront leur utilisation, aussi bien auprès des enseignants, qu'auprès des chercheurs. Elles sont assez simples pour être utilisées par les deux sortes d'utilisateurs, et elles sont présentées ici de façon à pouvoir être exploitées et comprises sans bagage scientifique sophistiqué.

Développer une culture statistique minimale, commune aux enseignants et aux chercheurs, serait très utile au développement d'une didactique et d'un enseignement plus modernes, c'est-à-dire, non seulement plus efficaces, mais aussi plus transparents.

Préface à la seconde « édition » (2014)

Cet ouvrage rassemble les fiches polycopiées[i] d'une partie des travaux pratiques de Statistique destinée aux étudiants du DEA de Didactique des Mathématiques, ainsi qu'aux observateurs et aux expérimentateurs du Centre d'Observation et de Recherches sur l'Enseignement des Mathématiques de l'IREM de Bordeaux, entre les années 75 et 80. Il faisait partie d'un ensemble de documents destiné à présenter l'ensemble des méthodes d'analyse quantitatives : inférentielles, descriptives etc. utilisables dans des recherches en Didactique. Un exemple d'utilisation en Didactique accompagnait leur présentation mathématique succincte et leurs programmes de calcul et de traitement informatique des données.

Ces fiches doivent beaucoup, à l'origine, à l'ouvrage de Sidney Siegel, ***non parametric statistics for the behavioral sciences*** que j'ai pillé innocemment, comme il est d'usage pour des travaux pratiques : l'idée du plan de l'ensemble, les tables, quelques exemples, certaines démonstrations… Par contre, l'attirail didactique naïf, les préalables pédagogiques (pour des lecteurs de bonne volonté mais de connaissances très diverses), le texte des fiches, les exemples, les problèmes et leur solution, le formalisme mathématique… sont de ma main, ainsi que les explications que mes difficultés à traduire l'Anglais - à l'époque - m'obligèrent à produire. Les exemples enfin étaient puisés évidemment dans nos travaux au COREM. Ainsi, si les erreurs que le lecteur pourra sans doute trouver, me sont imputables. C'est pour cette raison que cet ouvrage n'a jamais été publié ni distribué en dehors de mes collaborateurs et de mes amis proches. L'inconvénient de cette situation, c'est que je n'ai jamais trouvé le temps et la motivation de le polir et de le terminer. Les rafales de moyens informatiques toujours provisoires et pénibles ont fait que ce n'est qu'en 1993 que les fiches furent rassemblées en un fascicule par l'IREM

Alors que Régis Gras nous initiait aux grandes analyses factorielles de données (AF, ACP, AFC,…), je souhaitais proposer rapidement des méthodes numériques adaptées aux tests d'hypothèses sur des faibles effectifs pour accompagner - je ne dis pas pour conclure – nos réflexions et nous habituer à questionner et à orienter nos observations par des démarches plus scientifiques que les traditionnelles discussions pédagogiques. Notre rencontre provoqua Régis Gras à créer un nouvel indice et un nouveau test de l'implication statistique, qui se déploya par la suite en une nouvelle branche de l'analyse statistique : l'AS Implicative (ASI).

G.B. Octobre 2014.

[i] « Méthodes d'analyse quantitative en didactique des mathématiques » Fascicule 5 Tests d'hypothèses février 1976

SOMMAIRE

-oOo-

A. NOTIONS ET VOCABULAIRE

I. DONNEES ET METHODES STATISTIQUES RELATIVES A L'ENSEIGNEMENT

1. Les données.

Les données qui nous intéressent sont recueillies soit au cours d'une activité d'enseignement, soit en vue d'une recherche sur l'enseignement. Elles ont en commun au moins une même structure mathématique a priori:

Elles sont formées d'une collection d'informations élémentaires. Chaque information élémentaire rapporte en général un comportement d'un élève dans une situation. Une statistique sera donc constituée d'une collection de triplets: (élève, situation, comportement).

a. L'élève appartient à un échantillon E observé, supposé extrait d'une population plus vaste, soit au hasard, soit suivant un système de conditions de contrôle (âge, niveau scolaire, sexe, connaissances personnelles antérieures...)

b. La situation est choisie dans un ensemble S (de questions, d'exercices...) engendré et structuré par des conditions et des paramètres de natures très variées (savoir en jeu, conditions matérielles, conditions didactiques...)

c. Les comportements (typiques des connaissances ou des savoirs visés) sont pris dans un ensemble C de réponses possibles de l'élève aux conditions dans lesquelles il est placé.

2. Les structures.

Ces données, ou collections d'informations sont appréhendées, aussi bien par les enseignants que par les chercheurs grâce aux structures définies a priori sur les composantes E, S, C, par les concepts qu'ils utilisent.

Par exemple, une classe peut être définie comme un ensemble d'élèves E, un cours de mathématiques comme une collection d'exercices S, les résultats des élèves comme une certaine application de E dans l'ensemble S x C où C est l'ensemble des comportements de réussite et d'échec, une note comme une autre application de S x C dans R.

La connaissance d'un certain concept pourra être représentée par une certaine application d'un ensemble de questions dans un ensemble de comportements etc...

3. L'utilisation des statistiques par les enseignants et par les chercheurs.

Les enseignants et les chercheurs s'intéresseront généralement à certaines relations entre des concepts du type ci-dessus.

Par exemple, les élèves de telle classe connaissent-ils telle notion? les élèves possédant tels caractères ont-ils des comportements différents de tels autres sur tel ensemble de questions? Pour réussir telle épreuve vaut-il mieux avoir tel passé ou tel autre?

Mais ces questions, comme toutes les questions essentielles et naïves, sont difficiles. On ne peut pas, le plus souvent, y répondre directement ni de façon générale et il faut conditionner la réponse par de nouveaux concepts et poser de nouvelles questions.

De sorte qu'il y a deux voies, deux dialectiques, très distinctes: celle de l'enseignant et celle du chercheur:

- 3.1. L'enseignant, comme d'ailleurs tous les responsables de l'enseignement, doit prendre rapidement de nombreuses décisions. Mais il peut les corriger assez vite si elles se révèlent à l'usage un peu inadaptées. Par contre, il ne peut pas attendre le retour du traitement statistique de toutes ses questions. D'ailleurs, de toutes les réponses qui lui parviendraient à temps, il ne peut espérer que l'exclusion de quelques possibilités et une diminution de son incertitude sur quelques autres: en aucun cas, les connaissances de didactique ne peuvent relever l'enseignant ou le décideur de sa responsabilité.

Il serait très utile de montrer comment certains traitements statistiques rapides, ne faisant appel qu'aux concepts de didactique les plus connus des élèves et des parents, peuvent néanmoins aider les enseignants non seulement dans le processus de décision à court et à moyen terme mais aussi dans celui de la communication didactique. Il serait indispensable qu'il montre aussi une utilisation des données, telles qu'elles résultent de la pratique de l'enseignement, ainsi que les traitements de statistiques spécifiques de la docimologie et appropriés à l'enseignement. Il devrait présenter enfin quelques concepts fondamentaux de statistique et de mathématiques qui, malgré le caractère très modeste et très pratique de ces fiches, seront supposés connus ici.

- 3.2. Le chercheur doit suivre un processus opposé à celui du professeur: celui du questionnement:

Quelles hypothèses correspondent aux questions qui nous intéressent? Quelles données recueillir? quels traitements utiliser? quelle conclusion tirer?

Plus que la rapidité et l'utilité immédiate, c'est la consistance, la stabilité, la pertinence et la sécurité de ses réponses qui lui importent.

Ces questions requièrent des méthodes et des concepts beaucoup moins spécifiques: il suffit souvent pour y répondre de s'interroger sur la structure mathématique des données, sur les conditions exigées pour l'emploi des épreuves, sur les propriétés mathématiques des indices utilisés etc.

En particulier, il est possible d'utiliser les mêmes méthodes indépendamment du type d'objet didactique étudié. Les ensembles E, S, C sont de ce point de vue assez symétriques. Définies sur des éléments abstraits, les méthodes exposées prendront sens par les applications qu'on en fera à des questions utiles à l'analyse de l'enseignement.

Par exemple, une statistique qui indique comment trois cents élèves répondent à une question permet de dire peu de choses sur la question et rien sur un ensemble de questions. Par contre, elle donne une certaine information sur l'ensemble des élèves. Elle intéresse donc bien un psychologue. L'information sera plus sûre pour lui si l'échantillon est de 1000 élèves.

Considérer les réponses d'un seul élève à trois cents questions intéressera aussi certainement un psychologue car elle lui permettra de savoir beaucoup de choses sur cet élève, mais elle ne constituera pas pour lui une statistique.

Par contre, le didacticien qui étudie principalement les situations didactiques pourra peut-être considérer qu'il possède un échantillon de 300 questions, exploitable s'il peut le structurer.

- 3.3. Mais qui osera prétendre que l'enseignant n'a pas besoin de connaître les réponses aux questions que se sont posées les chercheurs? Au contraire, la connaissance de certaines de ces réponses permettra à ses décisions de gagner en précision, en validité, et en rapidité.

Il est clair que la connaissance des méthodes par lesquelles ces connaissances sont établies lui sera nécessaire, aussi bien pour en comprendre la validité et les limites que pour les employer lui même à l'occasion.

Qui osera prétendre qu'un enseignant ne doit pas participer aux recherches de son époque, dès lors qu'il ne confond pas les deux dialectiques dont nous avons parlé et qu'il emploie des méthodes connues?

Au niveau actuel de recrutement et de formation des professeurs, il serait ridicule de ne pas concevoir qu'ils doivent avoir un rapport aux savoirs spécifiques de leur profession, comparable à celui qui est exigé des étudiants dans les autres domaines des sciences sociales : psychologues, sociologues, linguistes, économistes, anthropologues... ont une formation minimale dans le traitement scientifique des données afin de pouvoir comprendre et discuter au besoin la légitimité des connaissances qui leur sont enseignées.

4. buts et finalités de la formation en statistique pour la didactique

Cette formation devrait avoir pour but de permettre aux enseignants :

- de communiquer entre eux les informations dont ils ont besoin et qu'ils recueillent sur les résultats des élèves, la valeur des méthodes employées...
- d'utiliser aussi avec discernement les résultats des recherches en didactique.
- de connaître les possibilités et les limites des méthodes statistiques et par là, la légitimité des connaissances qu'ils utilisent dans leur profession (et qui ne relèvent pas uniquement du contenu).
- de discuter cette légitimité à l'occasion.
- de formuler eux-mêmes des conjectures susceptibles d'être soumises à l'épreuve de la contingence.
- d'imaginer la plausibilité de ces conjectures.
- de mieux voir ce qui distingue les faits établis des conjectures.
- de savoir comment convertir leur expérience en connaissances.
- de participer à l'occasion à des recherches.

5. Choix des enseignements et des méthodes.

Pour communiquer les résultats, quelle que soit la manière dont ils ont été établis, il semble indispensable d'utiliser seulement les méthodes connues et contrôlées par les destinataires. Enseignants et chercheurs en didactique ont donc besoin d'un vocabulaire et d'un répertoire minimal de méthodes communes, assez faciles à enseigner mais permettant une ouverture sur tous les sujets.

a) D'abord sur le traitement, par les enseignants, des informations scolaires nécessaires à la gestion de l'enseignement : un ouvrage est en préparation à ce sujet.

b) Ensuite sur les méthodes utilisées par les chercheurs. Ceux-ci utilisent actuellement plus volontiers des moyens puissants mais complexes d'analyse de données : analyse en composantes principales, analyse factorielle des correspondances, analyse implicative... Ces méthodes sont pour l'instant peu enseignées et peu acceptées dans les revues d'éducation. Elles seront présentées néanmoins dans un ouvrage spécifique[ii].

c) L'initiation aux statistiques pourrait suivre de nouvelles voies et permettre peut-être une meilleure connaissance et un accès direct à l'ensemble des méthodes, pour un public plus large englobant enfin les éducateurs. Un effort est entrepris dans ce sens dans notre collection. Cependant une transposition de cette ampleur demandera du temps.

d) Les statistiques inférentielles paraissent présenter aujourd'hui le meilleur compromis.

Par rapport à la description, le test d'hypothèse est plus universellement accepté, compris, tenu pour rigoureux, même s'il n'est pas aussi bien connu qu'on pourrait le souhaiter.

e) Plus particulièrement, les méthodes non paramétriques sont assez bien adaptées aux conditions rencontrées par les enseignants : distributions parentes inconnues ou non gaussiennes, petits effectifs...

De plus, elles peuvent être enseignées presque indépendamment des connaissances fondamentales en mathématiques, au niveau actuel de recrutement des professeurs de tous les niveaux scolaires.

Nous avons donc choisi de commencer la collection d'ouvrages de statistiques pour la didactique par un cours de statistiques non paramétriques.

6. Contenus et organisation de l'ouvrage

En fait, il s'agit d'abord d'un inventaire de méthodes, présentées suivant la nature des variables et des échantillons traités. Chacune fait l'objet d'une fiche qui peut être lue, comprise et appliquée

[ii]Brousseau Guy, "L'analyse de données en Didactique", Cours pour le DEA de Didactique des Sciences, LADIST, (1991)

indépendamment des autres. Un ou plusieurs exemples permettent d'exposer l'algorithme à suivre et l'interprétation des résultats.

Des commentaires simples permettent de comprendre le sens de chaque méthode et de mémoriser la formule utilisée.

Les premiers chapitres proposent, sur certaines parties de l'algorithme présenté, des compléments plus généraux qui constituent une initiation à la pratique des statistiques inférentielles et un résumé des connaissances indispensables. Nous avons cru utile toutefois de rappeler au préalable quelques notations mathématiques élémentaires nécessaires pour la lecture et l'utilisation des fiches.

Nous avons repris ici les conceptions de Sidney Siegel (Non paramétric statistics for the behavioral sciences) en les adaptant.

7. Présentation

Cet ouvrage présente cinq sortes de textes.

* Des textes succincts de présentation:

- de l'observation statistique en didactique, de son usage et de son intérêt.

- du vocabulaire de base: objets de l'observation et leur structure mathématique, codification, recueil des données, énumération et dénombrement, notations mathématiques élémentaires correspondantes.

* Un ensemble de fiches présentant les principaux tests non paramétriques, ceux qu'il convient d'utiliser lorsque les observations sont des variables nominales ou des variables ordinales et que les échantillons sont petits ou ne suivent pas une ditribution connue. Chaque fiche consacrée à une méthode précise, comprend: un problème introductif, la structure caractéristique et la disposition des données, la méthodes et ses variantes, les calculs complets sur un exemple concrêt, des exercices. Chaque fiche peut être utilisée directement, indépendamment des autres, et sans apprentissage préalable.

* En situation, des rappels de définitions de quelques notions de statistiques, de brefs commentaires, des explications élémentaires de formules et de leur démonstration qui, pris dans l'odre des fiches, constituent un petit cours d'initiation à la statistique, accessible à des étudiants ne possédant que les connaissances mathématiques communes à l'enseignement secondaire.

* Trois problèmes avec leur correction complète.

* Des compléments sur les stratégies d'analyses, (quelles questions poser...?), sur les instruments du questionnement, sur les méthodes de test d'hypothèse et de preuve expérimentale et sur l'élaboration et le choix des hypothèses nulles.

En résumé le présent ouvrage a pour objet de présenter les méthodes statistiques non paramétriques qui peuvent être utilisées pour étudier des questions de didactique sur des populations faibles, de l'ordre d'une classe ou d'une école.

Ces méthodes trouveront leur utilisation, aussi bien auprès des enseignants, qu'auprès des chercheurs. Elles sont assez simples pour être utilisées par les deux sortes d'utilisateurs, et elles sont présentées ici de façon à pouvoir être exploitées et comprises sans bagage scientifique sophistiqué.

Développer une culture statistique minimale, commune aux enseignants et aux chercheurs, serait très utile au développement d'une didactique et d'un enseignement plus modernes, c'est-à-dire, non seulement plus efficaces, mais aussi plus transparents.

II. TYPES DE VARIABLES ET D'OBSERVATIONS, STRUCTURES, NOTATIONS

1. Les observations.

Les données dont la didactique s'occupe, sont des collections d'informations élémentaires qui, pour la plupart, prennent la forme de triplets. L'exemple "élève, situation, comportement" donné plus haut est le plus courant. Les triplets "enseignant, phase didactique, décision", par exemple, sont moins fréquemment utilisés. Un tel triplet est une observation.

Une **observation** consiste en une attribution d'une valeur à une variable à propos d'un individu: l'objet observé.

La statistique permet principalement de traiter les cas où :

- plusieurs observations sont recueillies,
- et où ces observations dans leur ensemble:
  - i) soit concernent des individus différents pour une même propriété,
  - ii) soit relatent des propriétés différentes pour un même individu,
  - iii) soit les deux.

Elle ne peut rien faire d'observations qui concerneraient à la fois des individus différents et des propriétés différentes.

Donc, de façon générale, pour qu'une observation soit utilisable, il faut pouvoir l'associer à deux éléments de repérage au moins :

- à une propriété ou à une variable, dont l'observation est une valeur
- à un objet d'observation, ou à un sujet.

Par exemple : "18" (valeur observée) est attribuée à "l'élève Michel Dupont" (objet de l'observation), comme "note de mathématiques" (variable observée).

L'objet de l'observation peut être lui-même élément d'une population (dans notre exemple un ensemble d'élèves), et la variable peut être elle-même un élément d'un ensemble de variables ou de propriétés observables sur un même sujet (par exemple un ensemble de renseignements : âge, sexe, notes, ...).

Notion de variable.

Une observation est déterminée par la contingence (par l'expérience) ou choisie (par la décision d'un acteur observé du système éducatif) parmi un **ensemble de valeurs** ou de cas possibles (envisageables théoriquement ou effectivement). Elle est un élément d'un ensemble de possibilités. Cet ensemble de possibilités est appelé la "**variable observée**".

Dans l'exemple en cours, la variable est l'ensemble des valeurs entières comprises entre 0 et 20.

Une variable est caractérisée:

- par l'ensemble de ses valeurs (possibiltés d'observation)
- par la nature mathématique des opérations qu'il est possible d'effectuer sur les valeurs fournies par les observations
- et par des circonstances ou conditions diverses relatives au recueil de ces observations.

Ici deux notes peuvent être rangées, et, par convention, pour certains usages, additionnées. Cette variable est donc numérique, la valeur observée est un nombre. Le fait qu'elle soit une note de mathématiques, obtenue à telle date (il pourrait y en avoir plusieurs), donne des indications nécessaires sur les circonstances du recueil et donc sur les interprétations possibles.

Notion de population et d'échantillons.

Toute observation suppose l'identification du sujet ou de l'objet de l'observation, même si elle est cachée ou réduite au minimum. Il faut par exemple s'assurer que des observations d'une même valeur correspondent soit à des objets distincts soit à des circonstances distinctes (on n'a pas répété le même résultat d'observation).

L'ensemble des objets d'observation est la population observée. Elle s'appellera un échantillon lorsqu'on supposera que les objets observés sont extraits d'une population plus vaste.

Pour résumer, une statistique est une application d'un ensemble d'objets d'observation (en nombre n) dans un ensemble produit de variables (en nombre p).

Chaque statistique est donc un ensemble de n p-uplets et peut être présentée dans un tableau dont chacune des p colonnes est attribuée à une variable différente, chacune des n lignes à un objet différent, chacune des nxp cases reçoit la valeur observée (pour l'objet et la variable ainsi déterminées).

2. Les types de variables.

Les opérations que l'on peut faire avec les valeurs observées jouent un rôle important dans l'analyse et dans l'interprétation des données. Ces opérations dépendent de la structure mathématique de la variable observée, on en distingue quatre principales:

a) une variable est dite NUMERIQUE lorsque ses valeurs sont, non seulement exprimées par des nombres (qui peuvent être des naturels, des décimaux, des fractions, ...), mais encore lorsque les opérations numériques que l'on peut faire sur ces nombres ont un sens pour la variable. En particulier, l'addition de deux valeurs doit être une valeur possible de la variable.

Toutefois, le temps (principalement les dates) constitue une variable particulière qui doit être traitée en général à part.

b) une variable est dite D'INTERVALLE lorsque seules les différences entre les valeurs qui l'expriment ont un sens, alors que la somme n'en a pas. En général, cela vient de ce que le choix de l'origine des valeurs est indifférent.

Par exemple, le score obtenu dans une discipline sportive peut constituer une variable d'intervalle. C'est la différence des scores qui est intéressante car elle correspond à une différence entre les performances, et c'est entre ces différences que sont définies des équivalences pour des épreuves différentes (un gain de tant de dixièmes de secondes au cent mètres équivaut à un gain de tant de centimètres au saut en hauteur). Les valeurs pourraient toutes être augmentées ou diminuées sans changer le résultat : la somme de différences de scores a un sens, la somme de scores quelconques n'en a pas.

c) une variable est dite ORDINALE si ses valeurs expriment seulement un ordre entre les observations, la différence n'a plus de signification précise. Une variable ordinale peut toujours s'exprimer par des rangs. Ces rangs sont habituellement exprimés par des nombres, mais seul l'ordre compte. Ils pourraient être aussi bien transcrits par les lettres de l'alphabet (a pour la première, b pour la suivante...). Les nombres ne représentent alors rien de plus que des lettres.

Dans une variable ordinale, la somme de deux valeurs n'est pas une valeur : par exemple, aucune opération sur l'observation de rang 3 et sur celle de rang 5 ne permet de trouver l'observation qui occupe le rang 8.

d) Une variable est dite NOMINALE si ses valeurs sont des caractères ou des attributs. Cette variable peut être à deux valeurs : un seul attribut et sa négation; ou à plusieurs valeurs, lorsqu'elle est composée de plusieurs attributs exclusifs les uns des autres (par exemple l'appartenance à l'une des classes de 6ème d'un établissement).

Même si elle est exprimée par des nombres comme 0 ou 1, une variable nominale n'est pas numérique : l'addition de deux caractères n'est pas définie, ni leur ordre en général. Les seules opérations que supportent les valeurs des variables nominales sont des opérations logiques (ensemblistes).

Une variable numérique est toujours transformable en variable d'intervalle, ordinale ou nominale (au prix d'une certaine perte d'information); de même, une variable d'intervalle peut être transformée en

variable ordinale ou nominale, une variable ordinale peut être transformée en variable nominale. L'inverse n'est pas vrai.

Certaines variables peuvent être implicites. En particulier, l'ordre d'enregistrement peut être signifiant ou non. (voir exercice ci-dessous). Pour éviter les confusions, il convient de se ramener au cas où on pourrait prendre les objets d'observation dans un ordre aléatoire. Cette opération s'appelle "randomization" ou "brouillage".

3. Les observations de plusieurs variables : p-uplets

Si plusieurs observations de natures différentes sont relatives à un même objet, elles peuvent être regroupées pour constituer un p-uplet, élément d'un produit de variables.

Par exemple: Michel Dupont, échec à l'exercice 4, réussite à l'examen, note de mathématiques: 8,

Mais si rien ne différencie ces observations, sinon le fait qu'il s'agit d'observations distinctes, ni par la nature de la variable qu'elles expriment, ni dans les conditions de recueil, (en particulier l'ordre dans lequel elles sont notées est indifférent) elles sont seulement plusieurs observations d'une même variable.

exemple:
notes de mathématiques au 1er devoir dans tel groupe d'élèves : [7, 12, 11, 5, 14, 9, 10, 10]

exercice:
On veut observer, pendant l'exécution d'un exercice, les allées et venues d'un professeur entre les tables des élèves et la durée des stations auprès d'un élève précis. Préparer un tableau destiné à recevoir les observations recueillies.

Réponse : les renseignements élémentaires sont de la forme : le nième arrêt s'est produit devant tel élève et a duré tant de temps. L'objet de l'observation est un arrêt du professeur et il est donc noté à son propos 3 variables : le rang de l'arrêt (variable ordonnée), l'élève concerné (variable nominale), la durée de l'arrêt (variable numérique temporelle). Il faut donc préparer 3 colonnes : rang, élève, durée ; l'ordre d'écriture des observations devient indifférent.

Mais pratiquement, 2 colonnes seulement suffiront (élève, durée de l'arrêt) car la notation du rang peut être implicitement (et avantageusement) remplacée par l'ordre effectif des enregistrements : première observation sur la première ligne, deuxième sur la deuxième, etc...

Le nombre d'observations, donc de lignes, ne peut être déterminé à l'avance

4. L'analyse statistique

L'analyse statistique consiste, en gros, à envisager que des objets se ressemblent a priori suffisamment pour qu'on puisse essayer de les résumer ou de les modéliser par le moyen d'un petit nombre de renseignements ou même par un objet type, puis de donner une idée de la ressemblance obtenue.

Si cette ressemblance n'est pas satisfaisante, on renonce à considérer ces objets comme des représentants d'un même type, et on cherche une partition de l'ensemble initial en classes plus réduites dont les objets seraient plus ressemblants entre eux et donc plus aisément représentables et explicables.

Cette recherche repose donc d'abord sur l'identification, la désignation et le codage des objets d'observation, des variables et de leurs valeurs, le recueil des valeurs observées, puis sur le classement, l'énumération, et le comptage des classes constituées (comptage des objets qui relèvent d'un même caractère ou d'une même valeur d'une variable).

Le travail commence par la définition et le codage des observations à recueillir, suivi du recueil (ou cueillette) des données, suivi à son tour de leur saisie si les traitements s'effectuent avec un ordinateur.

Il se poursuit par la construction d'une présentation réduite des données, appropriée aux études envisagées.

5. Utilisation de cet ouvrage

Dans chaque chapitre de cet ouvrage, nous supposons que les données à traiter ont déjà été déterminées et que l'utilisateur connaît le nombre et la nature des variables recueillies. Nous commençons par évoquer une disposition de ces données pour leur recueil.

Nous donnons un ou deux exemples conduisant à recueillir des données de ce type. Puis nous suivons pas à pas le déroulement de l'analyse en apportant, à l'occasion, des explications particulières ou des informations plus générales qui constituent une sorte de cours. ces parties sont signalées par une bordure verticale.

De cette manière, l'ouvrage constitue à la fois une initiation très simple et progressive aux statistiques, et un recueil de référence pour l'emploi des méthodes. Chaque chapitre en présente une différente et peut être utilisé et compris directement, sans recours fastidieux aux autres chapitres.

III. IDENTIFICATION ET DESIGNATION DES OBJETS D'OBSERVATION. CODAGES

1. Identification des objets de l'observation

Pour identifier des objets comme distincts, il faut utiliser certaines de leurs propriétés : par exemple, la position qu'ils occupent à un instant donné. Mais il peut être important de les reconnaître, même lorsque certaines de leurs caractéristiques changent (comme par exemple leur position). Il faut alors utiliser d'autres systèmes de propriétés distinctives.

Si un ensemble de propriétés est tel que chaque classe engendrée par leur produit ne contienne qu'un seul objet d'observation, alors ce système est, pour ces observations, un système d'identification : il suffira d'indiquer pour un objet quelconque la suite des valeurs des propriétés qui le déterminent, il ne pourra pas être confondu avec un autre.

Les systèmes d'identification sont donc des systèmes de classification a priori des objets. L'analyse statistique a pour objet de déterminer a posteriori (suivant la composition effective de la contingence) d'autres systèmes de représentation et de classification des mêmes objets.

Les procédés d'identification sont plus ou moins conscients, commodes ou efficaces selon les ensembles d'objets à observer et les utilisateurs.

Pour manipuler les objets de la statistique, on est conduit à les désigner, et pour cela, il faut des noms.

2. Codages et dénominations

Désignation

Concrètement, l'analyse statistique manipulera des noms d'objets et non les objets eux-mêmes. Il faut donc construire des systèmes générateurs de noms (dénomination) qui soient associables à un système d'identification.

Une désignation dans un ensemble est constituée d'une identification et d'une dénomination de ses éléments.

La dénomination des éléments d'un ensemble est construite à l'aide d'un répertoire de signes (alphabet ou vocabulaire) et d'une grammaire plus ou moins complexe.

En général, il faut trouver un compromis entre des conditions contradictoires :

- un bon système de désignation doit être peu coûteux :

a) le nombre total des symboles du répertoire doit être réduit mais la grammaire doit être simple (pour que la dénomination soit vite apprise et distinguée).

b) le nombre de signes utilisés pour désigner un objet doit être réduit (par exemple pour pouvoir figurer sur des graphes).

c) les symboles les plus fréquents doivent être les plus courts

- mais il doit produire des sigles évocateurs.

Codification

Le choix d'un bon codage est essentiel.

L'analyse statistique conduit à la manipulation de nombreux fichiers qui se modifient au fur et à mesure de la progression de l'analyse. Le codage des objets d'étude doit pouvoir se prêter à ces déclinaisons. L'usage d'ordinateurs exige des systèmes de codages rigoureux.

Les systèmes de codage sont nombreux et variés. Le nombre et la variété des types d'objets et de leurs codages manipulés dans une analyse statistique conduit à demander aux mathématiques le soin de soulager leur conception et leur manipulation.

Le codage des propriétés d'identification fournit souvent un système de désignation commode. Inversement, les moyens de caractériser et de classer ou d'ordonner les objets peuvent être considérés comme des variables.

i) Il convient, avant tout, de distinguer les objets et leurs noms (si l'identification est incomplète ou si la dénomination n'est pas bijective, la désignation est incorrecte et conduit parfois à des erreurs de dénombrement).

ii) Il convient de distinguer aussi :

- les dénominations de <u>constantes</u> (ou terminales) qui servent à désigner toujours le même objet d'observation précis et unique, par exemple le nombre 2, Louis XIV, Marc Dumont en tant que personne bien reconnue, etc.

- des dénominations de <u>variables</u> qui désignent un objet générique, pris dans un ensemble déterminé. Par exemple, le nombre x, l'ensemble des classes C, Marcel Dupont, pris comme "un quidam quelconque", etc.

- des dénominations ou des redénominations provisoires mais considérées comme constantes.

iii) Il convient enfin de distinguer, par des notations systématiques, les noms d'éléments, les noms d'ensembles, les noms de familles d'ensembles, les noms des nombres d'éléments de ces ensembles.

Exemple : Codages simples d'ensembles d'objets.

Pour désigner chaque élément d'un ensemble (que l'on compte), on utilisera le plus souvent un nom naturel (par exemple les noms des élèves) abrégé ou fabriqué (comme ci-dessous) ou simplement des numéros (qui ne constituent alors nullement une variable numérique ni ordinale).

Pour désigner les ensembles d'objets que l'on compte, s'ils ne sont pas très nombreux, on utilisera de simples lettres. Il ne s'agit que de distinguer un ensemble d'un "autre" (c'est-à-dire d'un ensemble défini autrement mais qui pourrait avoir des éléments communs, voire être le même ensemble).

Exemple : L'ensemble E des élèves de la classe du CM2 de l'Ecole Jules Michelet de Talence. l'ensemble F des filles du CM2 de ce même CM2.

Choix d'une "grammaire ": exemple.

Si le nombre des ensembles d'objets est plus vaste, il est préférable de construire un codage.

Par exemple, si on considère un ensemble d'écoles, "E" pourra être remplacé par "CM2JM" pour dire : la classe de cours moyen deuxième année de l'école Jules Michelet de Talence..

Cette opération consiste en fait à placer E comme objet dans un ensemble plus vaste engendré par le produit de deux variables: une collection de niveaux (CP, CE1, ...,CM2,...) et une collection d'écoles (MR, PP, PL, GL, JM, etc...).

Si on considère 8 niveaux scolaires et 12 écoles, CM2JM est un élément d'un ensemble comprenant 96 objets.

Le codage 5e construit en accolant le numéro du niveau (1 à 8) et une lettre d'établissement prise parmi 12 (a, b, c, ..., l) est plus court mais moins évocateur.

Ce codage peut ne pas être adapté à la description effective des classes, en particulier s'il existe plusieurs classes de même niveau dans certains établissements. Dans ce cas, il faut ajouter une variable si l'on veut identifier chaque classe: un numéro ou une lettre. La classe CM2AJM et la classe CM2BJM

constituent le niveau CM2 de l'école J.Michelet de Talence. S'il y a deux écoles J.Michelet dans l'échantillon il faudra introduire une nouvelle spécification.

Il peut être nécessaire de désigner l'ensemble de tous les élèves de l'échantillon (par E, par exemple). Il ne faut pas confondre l'ensemble des élèves E, l'ensemble des écoles (par ex F) et l'ensemble C des classes.

Opérations ensemblistes.

Dès lors que des codes ont été attribués à certains ensembles, le langage et les opérations ensemblistes donnent des moyens très précis et relativement efficaces de désigner de nouveaux ensembles.

Ces opérations sont principalement

-la réunion, l'intersection, la différence ensembliste,
-la partition d'ensembles et le quotient
-le produit d'ensembles et la projection

Dans cet ouvrage, nous ne considèrerons pas de manipulations complexes des ensembles à compter, ou alors nous considèrerons qu'elles ont déjà été effectuées au préalable. Aussi, nous n'utiliserons ni le vocabulaire pourtant fondamental de l'analyse des protocoles[iii], ni les symboles mathématiques qui permettent de décrire les opérations ensemblistes sans ambiguïté.

IV. LE DENOMBREMENT DES COLLECTIONS.

1. Symboles de dénombrements

Cette opération bien connue n'est pas toujours simple. On sera donc souvent amené à décrire et à comparer des méthodes de dénombrement. Pour cela, un minimum de métalangage est nécessaire.

Pour indiquer que l'on compte les éléments d'un ensemble spécifié E, on utilise l'un des symboles suivants : #(E) ;|E| ; nbr(E) ; card(E) ;

qui se lisent indifféremment "nombre d'éléments de E" ou "cardinal de E".

Pour spécifier E, on peut donner sa définition, mais le plus souvent il faut disposer de la liste de ses éléments pour pouvoir en faire le dénombrement.

Par exemple, le nombre - l'effectif - des élèves de la classe du CM2AJM pourra s'écrire: #(CM2AJM) ou encore nbr(CM2AJM). En 1990 #(CM2AJM) = 26

Cependant, on utilisera dans le cours des redésignations provisoires plus courtes : l'ensemble à compter sera souvent (pas toujours) désigné par un lettre majuscule (par ex. "A") et le nombre de ses éléments sera indiqué par la lettre minuscule correspondante "a".

S'il n'y a qu'un ou deux échantillons, leur effectif pourra être indiqué par n ou même N.

2. Stratégies de dénombrements : l'énumération

La base du dénombrement d'un ensemble est son énumération. C'est l'opération qui permet de passer d'un ensemble à une liste de ses éléments. Elle consiste à appeler l'un après l'autre tous ses éléments, sans appeler deux fois le même. Il faut savoir choisir un premier élément, puis pour tout élément, soit lui définir un successeur satisfaisant (qui n'a pas déjà été appelé à ce stade de l'énumération), soit déterminer qu'il n'y a plus d'élément non énuméré. L'énumération pratique peut poser des problèmes délicats.

Concrètement, il est d'usage d'effectuer le comptage en deux temps : le dénombrement et le comptage.

- Le dénombrement consiste à tracer de petits traits, groupés par 5, au fur et à mesure de l'énumération de l'ensemble compté. On obtient ainsi un ensemble équipotent au premier mais beaucoup plus facile à compter.

[iii] se reporter aux ouvrages de H. ROUANET et à la présentation de son logiciel d'analyse VAR.

- Le comptage consiste à incrémenter de 1 la suite des nombres naturels à chaque élément énuméré. L'énumération s'opére alors sur le dénombrement ainsi dessiné.

3. Dénombrement suivant une partition

Il est souvent commode de ramener le dénombrement d'un ensemble à des dénombrements plus faciles, plus petits ou connus à l'aide d'une partition en classes.

Une partition est un ensemble E décomposé en sous ensembles disjoints (appelés classes au sens de catégories), non vides, et dont la réunion est égale à E.

Par exemple, pour compter les élèves de l'échantillon E évoqué ci-dessus, il est clair qu'on pourra compter les élèves (les éléments) de chaque école (de chaque classe au sens de catégorie) et en faire ensuite la somme. Il faut, pour cela, disposer d'une liste des élèves pour chaque école et d'une liste des écoles. Appelons F l'ensemble des écoles.

Mais il peut être plus commode de considérer la partition de E en classes (scolaires). Il faudra alors une liste des élèves de chaque classe (scolaire et catégorie) et la liste des classes. Appelons C la liste des classes de l'échantillon.

4. La description du comptage

Si on veut seulement indiquer cette opération et non pas la faire, il faut utiliser le codage. Chaque classe de la partition doit être désignée par son nom (sa "constante", encore faut-il la connaître) :

Dans notre exemple :

#(E) = #(MR) + #(PP) + #(PL) + #(GL) + #(JM) + ...

Si F comprend 12 écoles, il faut 12 codes et la somme comprend 12 termes

5. Le programme de comptage : le symbole Σ

Si l'on ne veut pas ou si on ne peut pas nommer explicitement ces écoles, de nouvelles conventions sont nécessaires :

Il faut pouvoir dire:

-d'abord que l'on effectue une énumération de l'ensemble des écoles. Pour cela, on utilise un symbole générique, par exemple f, qui désigne une école quelconque de la liste F des écoles : fεF se lit alors "f parcourt F" et signifie que l'on fait une énumération de l'ensemble F.

En langage ensembliste, on écrirait :

$$E = U_{f \in F}\, f \qquad (1)$$

E est la réunion des ensembles f de la liste F.

-ensuite que pour chaque école évoquée par cette énumération, on compte le nombre d'élèves de cette école quelconque. Il s'agit, en fait, d'une variable qui ne prend une valeur numérique précise que lorsqu'on indique quelle école désigne f.

On peut alors indiquer à l'aide du symbole Σ, que pour trouver le nombre d'élèves de E, on fait la **somme** des nombres d'élèves de chaque école:

$$nbr(\,E\,) = \sum_{f \in F}\, nbre(f) \qquad (2)$$

La somme des nbr(f) est bien le nbr(E) parce que les f sont disjoints.

Cette formule représente le calcul du cardinal de E suivant la partition F de E.

Elle se lit "somme des nombres de f, f parcourant F".

ou plus prosaïquement dans l'exemple ci-dessus : somme des nombres d'élèves de chaque école f, suivant une énumération de F (ou étendue à tous les éléments de F).

Elle signifie : "somme des nombres d'élèves de toutes les écoles de la liste F".

Pratiquement il est plus commode de dédoubler la désignation, mettre deux lettres pour indiquer qu'on veut désigner de façon générale une école particulière: une lettre "constante", par exemple E, indiquera

qu'il s'agit d'une école, et un indice (ici f) qui est supposé changer dans chaque terme de façon à parcourir une certaine liste. On écrira donc plutôt pour (1) et (2)

$$E = U_{f\in F}\, E_f \qquad \text{et} \qquad nbr(\,E\,) = \sum_{f\in F} nbre(E_r)$$

Remarques sur la notation Σ :

a) Le signe Σ permet de symboliser l'écriture de la somme indiquée plus haut:

$$\#(E) = \#(MR) + \#(PP) + \#(PL) + \#(GL) + \#(JM) + \ldots$$

La formule en est beaucoup plus générale puisqu'elle n'utilise ni les codages des classes (des écoles dans l'exemple), ni leur liste effective, ni bien sûr, le nombre d'élèves qu'il s'agit d'ajouter.

b) Si une énumération, et en particulier le nombre n de classes de la partition F est connu, alors, chacune d'elles peut être désignée par un indice, son rang i dans une énumération de F :

"$f_1, f_2, \ldots, f_i, \ldots\ 0 < i < n$ " ou même " $(f_i)\ _{0<i<n}$ " au lieu de " $E_f \quad f \in F$ "

Par exemple, si on sait qu'il y a douze écoles on peut écrire : nbr(F) = 12

$$nbr(E) = nbr(f_1) + nbr(f_2) + \ldots + nbr(f_i) + \ldots + nbr(f_{12})$$

Mais il est plus commode de résumer ceci ainsi:

$$nbr(E) = \sum_{i=1}^{12} nbr(f_i)$$

qui se lit "somme de i = 1 à 12 des nombres de f_i", ou même en convenant que $nbr(f_i) = n_i$:

$$nbr(E) = \sum_{i=1}^{12} n_i$$

c) un même ensemble E d'élèves peut être dénombré selon des partitions différentes. Evidemment, le résultat reste le même.

Dans notre exemple, si on dispose d'une liste C de toutes les classes d'un secteur scolaire E, c désignant une classe quelconque de la liste C, le nombre d'élèves du secteur est :

$$nbr(E) = \sum_{c\in C} nbr\ (c)$$

On peut de même indiquer, par une somme double, que l'on fait d'abord la somme des effectifs de toutes les classes cf d'une même école f, puis la somme des nombres trouvés pour chacune des écoles f de la liste F :

$$nbr\ (E) = \sum_{f\in F} \sum_{c\in f} nbr(c)$$

exercice. quel nombre représente :

$$\sum_D \sum_S \sum_E \#\ (c_{esd})$$

Réponse : le nombre de classes du département D

6. Dénombrement suivant un produit d'ensembles

6.1. Produit de deux ensembles.

Le dénombrement d'un ensemble sera encore plus facile si la partition est effectuée en catégories de même effectif, il suffira de dénombrer une de ces catégories et de dénombrer la liste des catégories:

effectif total = nbr({catégories}) x nbr (une catégorie)

Par exemple, il s'agit de compter le nombre de classes de l'échantillon. Il y a nbr(F) écoles et dans chaque école le même nombre de classes.

Un moyen de s'assurer que chaque f posséde le même nombre d'éléments consiste à déterminer ces éléments à l'aide d'une énumération standard.

Si toutes les écoles ont la même composition standard de niveaux: S = {CP, CE1, CE2, CM1, CM2} et donc le même nombre de classes nbr(S) = 5, le nombre total de classes nbr(C) est

nbr(S) x nbr(F) = nbr(C) soit concrètement : nbr(C) = 5 x 12

Ce nombre est le cardinal d'un ensemble C obtenu par dénombrement, suivant une double partition: en écoles et en niveaux. L'ensemble C est l'ensemble produit F **X** S. Chaque élément de F X S est un couple (f,s) déterminé par le choix d'un élément f de F et d'un élément s de S.

C = F X S = **{**(f,s)**}** , On peut écrire aussi avec un léger abus de langage,

$$C = \bigcup_{s\in S; f\in F}(s,f) = \bigcup_{i\in SXF} C_i$$

Ainsi nbre(C) = nbr(F X S) = nbr(F) x nbr(S)

Si l'effectif de la classe de niveau s dans l'établissement f est désigné par nbr(C(s,f)) (et n'est plus constant), ou si i parcourt l'ensemble produit SxF, alors on peut écrire :

$$nbr(E) = \sum_{s\in S; f\in F} nbr(s,f) = \sum_{i\in SXF} nbr\ (C_i)$$

ou encore

$$nbr\ (E) = \sum_{i=1}^{nbr(S)Xnbr(F)}(C_i)$$

6.2. Produit de plusieurs ensembles.

Si une statistique présente l'observation d'un ensemble de p variables $(V_i)_{1\leq i\leq p}$, chaque observation est constituée d'un n-uplet: (x_1, x_2, x_3, ..., x_i,...x_p).

L'ensemble A de ces p-uplets constitue le produit ensembliste des p variables. On le note:

A = V_1 x V_2 x V_3 x ...x V_i x...x V_p) noté aussi:

$$A = \prod_{i=1}^{p} V_i$$

Si chacune de ces variables présente ki valeurs, le nombre des identifications possibles avec ces variables, c'est-à-dire le nombre de p-uplets distincts par au moins une de leurs valeurs est :

nbr(I) = k_1 x k_2 x k_3 x ... x k_i x ... x k_p

Par une convention semblable à celle exposée plus haut pour la somme, ce produit s'écrit :

$$nbr(A) = \prod_{i=1}^{p} k_i$$

Le même signe est utilisé pour les produits d'ensembles et pour les multiplications de nombres. Il n'est pas possible d'utiliser le même signe pour la réunion d'ensembles U et pour la somme Σ car il est nécessaire de pouvoir représenter par U des réunions non disjointes d'ensembles (dont le nombre total n'est pas la somme des nombres de chaque ensemble)

* * *

B. VARIABLES NOMINALES, LES EPREUVES DU CHI CARRE

I. VARIABLES NOMINALES: HOMOGENEITE

FICHE 1

UN échantillon, UNE variable nominale à DEUX valeurs.

1) Recueil des données.

Pour chacun des éléments d'un échantillon on relève la valeur d'une variable V. à deux valeurs, notées par exemple 0 et 1. Il faut lire 0 et 1 comme de simples lettres, ces signes ont ici une valeur nominale et non pas numérique.

éléments de l'échantillon	a	b	c	d	e	...
valeurs de la variable	V_a	V_b	V_c	V_d	V_e	...

Exemple.
Un instituteur, Mr A, constate que 17 de ses élèves ont réussi un exercice alors que 10 y ont échoué. Un de ses collègues Mr B, déclare qu'en général la moitié des élèves échouent à cet exercice. Mr A. peut-il penser que ses élèves sont meilleurs que ne le pense Mr B.?

Ici, nous observons, sur 1 échantillon - les 27 élèves de Mr A - une variable: le résultat de l'exercice, cette variable peut prendre deux valeurs: réussite (notée R ou 1) et échec (noté E ou 0).
(Nous aurions pu noter la réussite aussi bien 0, que R, et l'échec 1).

2) Présentation réduite: Distribution observée ou distribution contingente.

La distribution des observations sur les valeurs de la variable permet de présenter les données de façon plus synthétique sans perdre d'information.

valeurs de V	0	1
observations	a	c
(éléments de	b	d
l'échantillon)	e	

La distribution des effectifs: Le nombre d'observations correspondant à chaque valeur de la variable ne retient plus l'identité des éléments observés :

valeurs de V	0	1	Σ
effectifs	n1	n2	n

Dans l'exemple ci-dessus:

valeurs de V	E	R	
effectifs	10	17	27

3) Comparaison avec une population parente: hypothèse nulle.

Cette épreuve permet de comparer deux distributions, ou de comparer la distribution observée avec une distribution parente ou théorique déduite d'une hypothèse: par exemple la distribution uniforme des effectifs sur toutes les valeurs. L'hypothèse nulle correspondante serait:

"chaque valeur de la variable présente le même effectif d'observations. (chaque observation a les mêmes chances de prendre toutes les valeurs)".

Dans l'exemple ci-dessus, plusieurs "hypothèses" pourraient être tirées de l'opinion de Mr B. Traduisons les:

" Le nombre des élèves qui réussissent est égal à celui des élèves qui échouent, cette loi est la même pour toutes les classes (fait ou modèle) et la vôtre ne devrait pas échapper à cette règle (hypothèse)". Nous admettons (sous réserve de vérification) le fait énoncé par Mr B, qui sera ainsi le modèle.

En fait il s'agit de savoir dans cette hypothèse - au sens de condition - si cette classe, où la réussite a été de 63% environ, suit ou non (hypothèse contraire) la loi de Mr B. **L'hypothèse dite nulle associée à un modèle est celle qui affirme que la contingence ne s'écarte pas trop du modèle.** Elle est symbolisée d'habitude par H_0. Ici l'hypothèse nulle dit par conséquent: "la classe de A suit la loi de B". L'hypothèse contraire dit "La contingence s'écarte significativement du modèle.

Si l'hypothèse nulle est contredite, c'est à dire si elle a des conséquences contraires à l'observation (donc si la contingence ne suit pas le modèle), nous pourrons penser (sans pouvoir trancher d'ailleurs), soit que la classe de Mr A. n'est pas "une classe ordinaire", soit qu'elle est une classe ordinaire mais que la loi de Mr B. est fausse.

Si elle la suit pourra-t-on dire le contraire? ne pourra-t-on rien dire?

4) Modèle: La distribution théorique

La loi énoncée par Mr B permet de fabriquer le tableau d'une distribution dite "théorique". Elle présente le nombre d'élèves en échec ou en réussite dans une classe selon Mr B. C'est la distribution uniforme (toutes les valeurs sont égales) dite aussi homogène.

valeurs de V	Va	Vb	
effectifs	n/2	n/2	n

Dans l'exemple ci-dessus il devrait y avoir:
27/2 = 13,5 réussites et autant d'échecs.

valeurs de V	E	R	
effectifs	13,5	13,5	27

5) Distance entre la distribution observée (contingence) et le modèle,

C'est la valeur obtenue avec une formule donnant une distance entre la distribution observée et la distribution théorique.

Nous utiliserons la distance du χ^2, cette distance est donnée dans le cas présent par la formule simplifiée:

$$\chi^2_{\text{observé}} = \sum_i \frac{(E_T - E_O)^2}{E_T}$$

dans laquelle E_T est l'effectif théorique des observations d'une valeur de la variable, E_O est l'effectif observé correspondant, la somme Σ_i porte sur toute les valeurs de la variable.

(χ^2 se lit "khi carré" et s'écrit parfois "chi carré").

Dans notre exemple

$$\chi^2_o = \frac{(13,5 - 10)^2}{13,5} + \frac{(13,5 - 17)^2}{13,5} = \frac{2 \times 3,5^2}{13,5}$$

$$\chi^2_{observé} = 1,9$$

6) Signification de cette distance: mesure de sa rareté.

Si la distance entre le modèle et la contingence est trop grande nous renoncerons à l'hypothèse nulle comme moyen d'expliquer la contingence, on dira qu'on la rejette.

Dans le cas contraire, nous ne pourrons pas la rejeter.

Certains disent alors: "nous acceptons l'hypothèse nulle" mais cette formulation conduit à des erreurs dont nous parlerons plus loin.

Comment savoir si une distance est grande ou petite

Un éléphant de 2m de haut parait grand, mais si on sait que 70% des éléphants sont plus grands que lui, alors il apparait comme un éléphant plutôt petit.

Mais si on vous apprend que 99% des éléphants du même âge que lui sont plus petits que lui alors on comprend qu'il s'agit d'un jeune éléphant exceptionnellement grand.

C'est donc la rareté d'une mesure dans les conditions où on la considère qui dit si elle est grande ou petite. On conviendra de dire qu'une mesure qui n'est dépassée que par k% des mesures de sa catégorie, est **significativement grande** au seuil k.

En psychologie, la valeur k = 0,05 = 5% est la plus fréquemment utilisée. Mais on peut utiliser en médecine des seuils de 0,001 (1 pour mille).

Pour savoir si un khi carré est grand, il faut donc connaître une distribution de nombreux khi carrés, calculés dans les mêmes conditions, et savoir quelle valeur -appelée $\chi^2_{seuil\ 0,05}$ - laisserait d'un côté 95% et de l'autre 5% de ces χ^2.

7) Distribution du Khi carré (χ^2).

Dans notre exemple est-ce-que 1,9 est une valeur grande pour un χ^2?

Pour le savoir par le moyen de la rareté on doit constituer une population de χ^2 que l'on obtiendrait dans les conditions de l'hypothèse nulle:

On suppose qu'on dispose un grand sac contenant des jetons: la moitié, marqués E les autres marqués R. On effectue un très grand nombre de séries (par exemple 100), de 27 tirages (avec remise) qui représentent des classes, comme celle de Mr A, mais qui seraient prises dans une population répondant à l'hypothèse de Mr B.

A chaque série on obtient un n_1 et un n_2 qui ne sont pas exactement égaux à $n_{/2}$. On effectue le calcul du χ^2.

La distribution de ces χ^2 présentée ci-après nous renseigne sur leur rareté.

Distribution du KHI CARRE

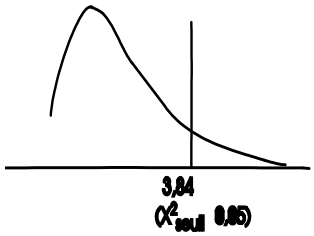

figure 1

Sous l'hypothèse nulle, proportion α des $\chi^2 > \chi^2_{seuil} = \chi\alpha^2$									
0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
0,064	0,15	0,46	1,07	1,64	2,71	3,84	5,41	6,64	10,83
Valeurs des différents χ^2_{seuil}									

Cette table nous apprend que 70% des χ^2 (lire 0,70) seront plus grands que 0,15 , et donc que 30% sont plus petits, ou encore que 50% sont plus grands que 0,46. De même seulement 20% des khi carrés sont plus grands que 1,64.

Nous pouvons représenter cette distribution par un graphe dans lequel le nombre (en fait la proportion) de valeurs de χ^2 supérieures à la valeur seuil choisie est représenté par la surface hachurée. (figure 1)

Exercice:

Sous l'hypothèse nulle, quelle proportion des χ^2 serait comprise entre 0,46 et 1,64?

8) Rareté de la valeur observée.

Comparons le $\chi^2_{observé}$ avec les différents χ^2_{seuils} que présente le tableau: il s'agit de placer le $\chi^2_{observé}$ (ou calculé) par rapport aux χ^2_{seuils}:

Si $\chi^2_{seuil\ \alpha} < \chi^2_{observé} < \chi^2_{seuil\ \beta}$ on dira que le $\chi^2_{observé}$ est significatif au seuil α et non pas au seuil β. Autrement dit $X^2_{observé}$ est plus rare que k% (k = 100α).

Parmi les valeurs des χ^2 échantillonnées dans les conditions de l'hypothèse nulle, on trouve moins de k% de valeurs plus grandes que $\chi^2_{observé}$.

Exercice:

A quel seuil la valeur 5,93 est elle significative?

Dans notre exemple, 1,9 est un peu plus grand que 1,64, donc un peu plus rare que 20%. Par contre il est plus petit que 2,71: donc il y a plus de 10% des échantillons qui ont un χ^2 plus grand que 1,9.

La "rareté" de cette observation se situe entre 20% et 10%;

Puisque 1,64 < 1,9 < 2,71 alors,

$\chi^2_{seuil\ 0,20} < \chi^2_{observé} < \chi^2_{seuil\ 0,10}$

On pourrait dire que notre χ^2 est significatif à 0,20

9) Décision.

Si la distance entre la distribution observée et la distribution théorique est plus grande qu'au moins 95% des χ^2 échantillonnés sous l'hypothèse nulle [autrement dit si χ^2 est significatif à 0,05 (i.e. à 5%), c'est-à-dire si $\chi^2_{seuil\ 0,05} < \chi^2_{observé}$], alors on rejettera l'hypothèse nulle : on n'acceptera pas de croire que la distribution observée aurait pu être obtenue par échantillonnage au hasard suivant cette loi.

On dira simplement que le χ^2 est significatif.

Dans le cas contraire, on ne rejettera pas cette hypothèse: le χ^2 "n'est pas significatif".

Dans notre exemple, il y avait plus de 1 chance sur dix d'obtenir une valeur du χ^2 supérieure ou égale à 1,9 dans la distribution que décrit Mr B. Cette valeur est relativement grande, mais pas assez: il faudrait qu'elle atteigne la valeur $\chi^2_{\text{seuil } 0,05} = 3,84$ pour laisser moins de 5% des valeurs plus grandes qu'elle, alors on pourrait rejeter l'hypothèse de Mr B appliquée à la classe de Mr A.

Autrement dit, puisque la distance entre la distribution observée dans la classe de Mr A. et la distribution modèle calculée d'après l'opinion de Mr B. n'est pas assez grande (1,9 au lieu de 3,84), on doit admettre que la classe de Mr A pourrait être un échantillon tiré au hasard dans une population parente telle que Mr B. la décrit:

Nous dirons qu'**on ne peut pas rejeter l'hypothèse** selon laquelle les élèves de Mr A. appartiennent à une population qui ne réussit qu'une fois sur deux.

Mr A. ne peut pas affirmer que ses élèves sont meilleurs que ce que dit Mr B.

OBSERVATION : 1. Ceci ne veut pas dire que la classe de Mr A. n'avait qu'une probabilité de 50% de réussir cet exercice. Cette probabilité pourrait être de 60%, de 63% (comme elle apparaît) ou peut-être de 70% comme nous le verrons dans les exercices suivants.

C'est pourquoi **il ne faut pas dire qu'on accepte l'hypothèse nulle** (on ne l'a pas prouvée, et il y aurait danger à accepter une hypothèse fausse car beaucoup d'hypothèse contradictoires avec l'hypothèse nulle ne peuvent pas non plus être rejetée), mais seulement qu'**on ne peut pas la rejeter**, ce n'est pas pareil!

OBSERVATION 2: Ceci ne veut pas dire non plus que l'opinion de Mr B. soit vraie: Il faudrait faire une mesure totale ou au moins un important sondage sur la population de tous les élèves pour le savoir. Elle est peut-être vraie et la classe de Mr A suit la norme. Elle est peut-être fausse.

Simplement l'opinion de Mr B. n'est pas contredite par l'expérience de Mr A.

OBSERVATION 3. Si le test avait permis de rejeter H_0, cela n'aurait pas voulu dire que la proposition de Mr B est certainement fausse. On pourrait la mettre en doute et affirmer qu'elle n'est pas universellement vérifiée. Mais elle pourrait être vraie aussi, stochastiquement, et la classe de Mr A serait seulement une bonne classe. Pour montrer que l'opinion de Mr B est fausse il faudrait étudier l'épreuve sur l'ensemble des classes et faire la moyenne. Il suffirait aussi d'avoir un échantillon représentatif de cet ensemble de classes et de montrer que dans cet échantillon la classe de Mr. A, n'occupe pas la bonne place que l'on croit.

10) Exercices

Exercice 1.

Mr A., un peu désabusé par son résultat, réfléchit et se dit que dans ce cas, il se pourrait que ses élèves soient extraits, au hasard, d'une population qui réussit cet exercice à 80%.

Solution.

a), b) comme dans l'exemple ci-dessus.

c) "80% des élèves de la population parente réussissent l'exercice, 20% échouent, et la classe de Mr A aurait pu être tirée au hasard dans cette population parente.

d) Distribution théorique correspondant à cette hypothèse.

Il devrait y avoir 0,8 x 27 = 21,6 réussites et 0,2 x 27 = 5,4 échecs.

Distribution théorique.

valeurs de V	0	1	Σ
effectifs	5,4	21,6	27

e) Distance entre la contingence et la distribution théorique.

$$\chi^2 = \frac{(5,4 - 10)^2}{5,4} + \frac{(21,6 - 17)^2}{21,6} = 4,6^2.(\frac{1}{5,4} + \frac{1}{21,6})$$

$$\chi^2_{\text{observé}} = 4,898$$

f) Signification de cette valeur.

L'examen du tableau du X^2 montre que la valeur observée est comprise entre 3,84 et 5,41 valeurs correspondant respectivement aux seuils de 0,05 et de 0,02. Elle est supérieure au seuil choisi de 5%. Ce serait donc une valeur rare si on admettait l'hypothèse nulle.

g) Conclusion: On préfère rejeter l'hypothèse nulle et accepter l'idée que la population parente dont est extraite cette classe ne peut pas avoir 80% de réussite. Mr A est encore désappointé!

Exercice 2

Pour rassurer Mr A. étudiez les hypothèses suivantes en faisant le moins de calculs possible.

"La réussite des élèves de la classe diffère significativement de P avec P= 85%, P=40%, P=70 %

"La réussite des élèves de Mr A ne diffère pas significativement de Q, avec Q= 75%, Q= 30%, Q= 60%.

FICHE 2

UN échantillon, UNE variable nominale à P valeurs.

Cette épreuve comme la précédente permet de comparer une distribution observée avec une distribution parente ou théorique déduite d'une hypothèse.

1) Recueil des données.

Pour chacun des éléments d'un échantillon on relève la valeur d'une variable V, à p valeurs: notées par exemple v_1, v_2, v_3, ..., v_p. Pour l'élève "a" on relève par exemple la valeur v_3 ,on note donc v_3 dans le tableau de recueil des données, mais ici, nous représentons cette observation par V(a) ou V_a: la valeur observée pour a.

éléments de l'échantillon	a	b	c	d	e	...
valeurs de la variable	V_a	V_b	V_c	V_d	V_e	...

Exemple. Mr A. instituteur a noté le nombre de fréquentations de quatre ateliers au cours des séances d'activités d'éveil. v_1, la bibliothèque, v_2, la peinture, v_3, le remontage de camions, v_4, les expériences avec des miroirs et des lentilles. Ces quatre ateliers sont-ils également intéressants?

Ici, nous observons, sur 1 échantillon: les élèves de Mr A, une variable: le choix d'un atelier. Cette variable peut prendre quatre valeurs: v_1, v_2, v_3, v_4.

2) Présentation réduite. Distribution des observations.

valeurs de V	v_1	v_2	v_3	v_4
	a	c	g	b
observations	e	d	h	f
(éléments de	j		k	i
l'échantillon)	...	...	...	...

DISTRIBUTION DES EFFECTIFS.

valeurs de V	v_1	v_2	v_3	...	v_p	Σ
effectifs	n_A	n_B	n_C	...	n_p	n

Dans l'exemple ci-dessus:

valeurs de V	v_1	v_2	v_3	v_4	Σ
effectifs	8	15	20	9	52

3) Hypothèse nulle.

Une des hypothèses nulles fréquemment examinées est:

H_0: Les observations sont équiréparties entre les différentes valeurs possibles.

Autrement dit, chaque observation a la même probabilité d'appartenir à chacune des catégories déterminées par les valeurs de la variable. Celles-ci devraient se partager également l'effectif.

Dans notre exemple, cette hypothèse exprime que les quatre ateliers sont également attractifs.

4) Modèle et DISTRIBUTION THEORIQUE.

L'hypothèse nulle permet de fabriquer un tableau qui lui correspond: une distribution dite "théorique":

valeurs de V	v_1	v_2	v_3		v_p	Σ
effectifs	n/p	n/p	n/p		n/p	n

Dans l'exemple ci-dessus il devrait y avoir 52/4=13 visites par atelier.

valeurs de V	v_1	v_2	v_3	v_4	Σ
effectifs	13	13	13	13	52

5) Distance entre la contingence et le modèle, entre la distribution observée et la distribution théorique.

Cette distance est donnée par la formule:

$$\chi^2_{\text{observé}} = \sum_i \frac{(E_T - E_O)^2}{E_T}$$

dans laquelle E_T est l'effectif théorique portant une valeur de la variable, E_O l'effectif observé correspondant, la somme Σ porte sur toute les valeurs de la variable.

(X^2 se lit "khi carré" et s'écrit parfois "chi carré").

Dans notre exemple:

$$X^2_o = \frac{(13 - 8)^2}{13} + \frac{(13 - 15)^2}{13} + \frac{(13 - 20)^2}{13} + \frac{(13 - 9)^2}{13} = \frac{(5^2 + 2^2 + 7^2 + 4^2)}{13} = 7{,}23$$

$X^2_{\text{observé}} = 7{,}23$

6) Signification de cette distance: mesure de sa rareté.

Si la distance entre le modèle et la contingence est trop grande nous renoncerons à l'hypothèse nulle comme moyen d'expliquer la contingence, on dira qu'on la rejette. Sinon, nous ne pourrons pas la rejeter.

Pour décider si la distance entre les deux tableaux est trop grande, nous allons évaluer sa "rareté" sous l'hypothèse nulle. Pour cela, comme dans le cas précédent, il faut connaître une distribution de nombreux khi carrés, calculés dans les mêmes conditions et savoir quelle valeur - appelée $X^2_{\text{seuil } 0,05}$ - laisserait d'un côté 95% et de l'autre 5% de ces X^2.

7) Distribution du Khi carré.

Dans notre exemple, est-ce-que 7,23 est une valeur grande pour un X^2?

Pour le savoir, on suppose comme dans le chapitre précédent qu'on dispose d'un grand sac contenant N jetons, chacun représente le choix d'un élève pour un atelier. Donc n/p de ces jetons sont marqués A, n/p sont marqués B, et ainsi de suite (Ainsi chaque valeur est représentée par une proportion de 1/p jetons). On effectue dans ce sac un très grand nombre de séries de n tirages avec remise. (des séries de 52 dans le cas de notre exemple) qui représentent des choix d'ateliers dans des classes comme celle de Mr A, mais dont les élèves seraient pris dans une population où l'on choisit aussi souvent chacun des ateliers. A chaque série de n tirages, on trouve un certain nombre d'élèves pour chacune des catégories et on effectue le calcul du X^2.

La distribution de ces X^2 nous renseigne sur la rareté des différentes valeurs.

Le résultat de ces tirages est présenté dans une table (voir en annexe). Chaque ligne correspond à des conditions différentes des tirages. Ces conditions sont déterminées par le "degré de liberté" du modèle dont on étudie la distribution. C'est une notion qui sera précisée dans le chapitre suivant.

Dans le cas étudié dans ce chapitre: un échantillon, une variable, p valeurs, le degré de liberté est dl = p - 1 .
La distribution des X^2 qui servira de référence est présentée dans la (p-1) ième ligne de la table.

On compare alors le $X^2_{observé}$ (dit aussi "calculé") aux valeurs de la table qui correspondent aux différents seuils. Le meilleur encadrement que l'on peut obtenir de la forme

$$X^2_{seuil\ k} < X^2_{observé} < X^2_{seuil\ j}$$

permet de conclure que le $X^2_{observé}$ est significatif au seuil k et ne l'est pas au seuil j.

Donc dans notre exemple: dl = p - 1 = 3, il faut regarder la 3ème ligne du tableau du X^2.

Table du KHI carré. dl=3

Sous l'hypothèse nulle, proportion des $X^2 > X^2_{seuil}$									
0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,34	16,27
Valeurs des différents X^2_{seuil} pour dl=3									

Cette table nous apprend que 70% des X^2 seront plus grands que 1,42 (et donc que 30% sont plus petits), que 50% encore sont plus grands que 2,37. Seulement 20% des khi carrés sont plus grands que 4,64...

Exercice.
Comparer dans la table la 1ère ligne et la troisième, c'est-à-dire la distribution des X^2 à un et à 3 degrés de liberté.Trouver une valeur X^2_o significative pour l'une et pas pour l'autre.

Comparons notre $X^2_{observé}$ avec les différents X^2_{seuils} que présente le tableau:
Notre mesure, avec 7,23 est un peu plus grande que 6,25 donc un peu plus rare que 10%. Par contre elle est plus petite que 7,82. Donc il y a plus de 5% des échantillons qui ont un X^2 plus grand que cette observation.

8) Rareté de la valeur observée.

Dans notre exemple, la distance 7,23 entre la distribution observée et la distribution théorique est plus grande qu'au moins 90%, mais plus petite qu'au moins 5% de ses semblables.
Elle est grande, mais pas assez selon le critère communément admis. De justesse toutefois: il suffirait qu'elle atteigne la valeur $X^2_{seuil\ 0,05}$ = 7,84 pour laisser moins de 5% des valeurs plus grandes qu'elle et pour être significative.

9) Décision.

Si le meilleur encadrement est:

$$X^2_{seuil\ k} < X^2_{observé} < X^2_{seuil\ j}$$

et si k = 0,05

alors le $X^2_{observé}$ est dit **significatif au seuil k**.

Dans notre exemple, puisque la distance entre la distribution observée dans la classe de Mr A. et la distribution modèle calculée d'après l'hypothèse nulle n'est pas assez grande, on doit admettre que la

classe de Mr A pourrait être un échantillon tiré au hasard dans une population parente où les élèves choisissent également tous les ateliers.

Nous dirons qu'on ne peut pas rejeter l'hypothèse nulle.

La distance euclidienne et la distance du chi carré

A. Pour mesurer la distance entre deux nombres a et b [que nous noterons $d_1(a,b)$], on peut utiliser diverses idées:

1. La valeur absolue de leur différence: $d_1(a,b) = |a-b|$.

Exemple: la distance de 3 à 9 est 6. Celle de 14 à 6 est 8. Si on utilisait la différence seule, il apparaîtrait des "distances négatives", ce qui est contraire à notre idée de la mesure: la distance de a à b est la même que la distance de b à a.

Remarque: Cette distance est aussi égale à $\sqrt{(a - b)^2}$.

On pourrait définir aussi $D_x(a,b) = (a-b)^2$

2. Si la distance entre les deux nombres est susceptible de représenter une erreur de mesure, on peut s'intéresser à la "distance relative": D_r.

Exemple: En voulant réaliser un segment de longueur de 109 cm un robot a réalisé en fait un segment de 103 cm. La distance $d_1(103,109)$ entre ces deux valeurs est 6. C'est la même que la distance entre 3 et 9. On voudrait exprimer que **relativement**, 103 est plus près de 109 que 3 de 9. On posera donc que la distance relative est $D_r(a,b) = |a-b|/a$ si a est la valeur "théorique" ou réelle de la longueur à mesurer et si b représente la valeur "effective" réalisée ou observée. On peut faire la même chose avec le carré de la différence: $D_{x^2}(a,b) = (a-b)^2/a$

B. Il s'agit maintenant de mesurer la distance entre deux suites de nombres a, b, c, ...n et a', b', c', ...n' que nous noterons V et V',

Exemple: Distance entre V = (3, 16, 110) et V'= (5, 14, 91) Il faut fabriquer de nouvelles définitions de la distance. On peut utiliser celles définies pour deux nombres. Si les deux suites comprennent le même nombre de termes (de composantes), on peut dire que les deux suites sont d'autant plus éloignées que la somme des distances entre les nombres correspondants est grande. (Ces distances entre deux nombres correspondants s'appellent contribution à la somme).

Cela donne:

1. Distance de la valeur absolue.

$D = |a-a'| + |b-b'| + |c-c'| + ... + |n-n'|$

sur notre exemple $D = 2 + 2 + 19 = 23$

2. La distance relative.

$Dr(V,V') = |a-a'|/a + |b-b'|/b + |c-c'|/c + ...$

sur notre exemple: $2/3 + 2/16 + 19/110 = 0,964$

3. La distance des carrés.

$D_{c^2} = (a-a')^2 + (b-b')^2 + (c-c')^2 + ... + (n-n')^2$

sur notre exemple :

$D_{c^2} = 4 + 4 + 361 = 369$

4. Le calcul des valeurs absolues et des radicaux n'est pas très facile, on peut aussi définir comme en A: **la distance du χ^2**

Erreur ! Signet non défini. sur notre exemple:
D_{x^2}= 4/3 + 4/16 + 361/110 = 4,865

C. Distances pondérées.

Plus le nombre des éléments des deux suites est grand, plus la somme des erreurs sera grande, pour une même précision relative de chacune des "mesures élémentaires".

Si l'on veut acquérir une certaine expérience d'une distance et pouvoir apprécier immédiatement l'importance d'une valeur obtenue, indépendamment du nombre des données, il est utile de considérer chaque fois non pas la somme des contributions (l'erreur totale) mais leur moyenne (arithmétique) de chaque erreur à l'erreur totale. Pour cela il suffit de diviser la somme des contributions, c'est-à-dire les valeurs des distances obtenues plus haut par le nombre des contributions. On obtient, pour chaque distance définie plus haut, une distance pondérée.

par exemple la distance de la valeur absolue pondérée est:

$$D_p = 1/n\ (|a-a'| + |b-b'| + ... + |n-n'|)$$

Dans notre exemple

$$D_p = 1/3\ (23) = 7,66$$

D.Le choix d'une distance ou d'une autre se justifie par l'usage que l'on veut en faire.

1. Ainsi la distance qui correspond à notre pratique familière de l'espace est définie par le théorème de Pythagore. Nous l'appelons **distance euclidienne**:

$$D_e = \sqrt{[(a - a')^2 + (b- b')^2 + (c - c')^2 + ... + (n - n')^2}$$

Sur notre exemple, V et V' sont les abscisses de deux points, leur distance euclidienne est:
$D_e = \sqrt{D_{C^2}} = \sqrt{369} = 19,20$
La distance euclidienne pondérée serait:
$D_{ep} = 1/n\ D_e$
dans notre exemple: $D_{ep} = 1/3\ (19,20)= 6,4$

2. Le choix de la distance du chi carré pour comparer les suites d'effectifs se justifie par la facilité avec laquelle on peut raisonner et calculer sur sa distribution avec le calcul des probabilités dans les circonstances habituelles.

3. Il existe beaucoup d'autres types d'indices qui ont pour objet de faciliter l'évaluation et la comparaison des résultats des expériences de statistiques. Nous en verrons certains.

---

10) Exercice

Mr A s'attendait à ce que ses ateliers soient inégalement intéressants: dans la revue à laquelle il a emprunté les dispositifs de ses ateliers, les auteurs annonçaient les fréquentations suivantes:

valeurs de V	v_1	v_2	v_3	v_4	Σ
fréquentation (pourcentage)	30	40	20	10	100

1.Mr A peut-il considérer que sa classe se comporte comme la population témoin étudiée par les auteurs?
2.Sinon peut-il étudier d'autres hypothèses?

Solution.
1.a), b) Comme ci-dessus.
c) Hypothèse nulle: les élèves de Mr A pourraient être tirés au hasard dans une population parente distribuée comme la revue l'indique.
d) Distribution théorique correspondant à cette hypothèse.
Il devrait y avoir 10 pour cent des 52 fréquentations pour l'atelier d'optique = 5,2.
Pour l'atelier A: 30x52/100 = 15,6
Pour l'atelier B: 40x52/100 = 20,8
Pour l'atelier C: 20x52/100 = 10,4

Distribution théorique.(modèle)

valeurs de V	v_1	v_2	v_3	v_4	Σ
fréquentation	15,6	20,8	10,4	5,2	52

e) Distance entre la contingence et la distribution théorique:

$$X^2_o = \frac{(5,2-9)^2}{5,2} + \frac{(15,6-8)^2}{15,6} + \frac{(20,8-15)^2}{20,8} + \frac{(10,4-20)^2}{10,4}$$

$X^2_{observé} = 2,77 + 3,70 + 1,62 + 8,86 = 16,95$

f) Signification de cette valeur.
Dans ce cas comme précédemment dl = 3
L'examen de la troisième ligne du tableau du X^2 montre que la valeur observée est supérieure à 16,27, valeur correspondant au seuil de 0,001. Elle est supérieure au seuil de 1 pour mille, donc supérieure à la valeur correspondante au seuil choisi de 5% : 7,82. Si on admettait l'hypothèse nulle, nous aurions moins d'une chance sur mille d'obtenir une valeur de X^2 aussi grande, c'est-à-dire une distribution aussi éloignée du modèle.

g) Conclusion: on préfère rejeter l'hypothèse nulle et accepter l'idée que la classe de Mr A ne peut pas être extraite de la population parente des classes témoins.

2. Peut-être l'environnement culturel des élèves de Mr A est-il différent et présente-t-il un rapport différent aux activités artistiques ou pratiques (B et C) et aux activités plus culturelles ou "intellectuelles" (A et D), ou alors son enseignement est-il orienté différemment...
La première hypothèse conduit à un regroupement des effectifs dans le modèle et dans la table de contingence: A et D d'une part, B et C d'autre part.
On obtient:

	v_1 et v_4	v_2 et v_3	Σ
Modèle	20,8	31,2	52
contingence	17	35	52

Cette hypothèse peut être étudiée comme dans le chapitre précédent.
$\chi^2 = 3,8^2/20,8 + 3,8^2/31,2 = 0,694 + 0,462 = 1,157$.
Cette valeur est insuffisante pour rejeter l'hypothèse nulle: on ne peut pas dire que la classe de Mr A présente un rapport au savoir "intellectuel" ou "pratique" différent de celui de la population témoin. L'explication avancée pour expliquer la différence ne convient pas.

FICHE 3

DEUX échantillons, UNE variable nominale à P valeurs.

Cette épreuve permet de comparer les distributions sur une même variable de deux échantillons afin de savoir si on peut les considérer comme issus d'une même population parente.

1) Recueil des données.

Pour chacun des éléments de chaque échantillon, on relève la valeur d'une variable V, à p valeurs notées par exemple v_1, v_2, v_3, ..., v_p. Pour l'élève "a' " appartenant au 2ème échantillon, on relève par exemple la valeur v_3, on note donc v_3 dans le tableau de recueil des données, mais ici, nous représentons cette observation par V(a') ou $V_{a'}$: la valeur observée pour a'.

éléments 1er échantillon	a	b	c	d	e	...
valeurs de la variable	V_a	V_b	V_c	V_d	V_e	...

éléments 2ème échantillon	a'	b'	c'	d'	e'	...
valeurs de la variable	V_a'	V_b'	V_c'	V_d'	V_e'	...

Exemple: Les élèves, en présence de "l'échec" d'une de leurs anticipations, peuvent présenter trois types de comportements: A: éveil de l'intérêt, B: indifférence, C: découragement. La même épreuve est présentée dans deux écoles différentes: 1 et 2. On note les réactions qu'a provoquées le premier échec chez chaque élève. Les élèves des deux écoles ont-ils des attitudes semblables devant un résultat inattendu ou ceux de l'une d'elles sont-ils plus "amorphes" que ceux de l'autre?

2) Présentation réduite

Distribution observée des EFFECTIFS ou distribution contingente.

	observation					
valeurs de V	v_1	v_2	v_3	...	v_p	Σ
effectifs 1er échantil.	n_1	n_2	n_3	...	n_p	N_1
effectifs 2ème échan.	n_1'	n_2'	n_3'	...	n_p'	N_2
total	Σ_1	Σ_2	Σ_3	...	Σ_p	N

Dans ce tableau,
$\Sigma_1 = n_1 + n_1'$ $\Sigma_2 = n_2 + n_2'$ etc. ;
$n_1 + n_2 + n_3 + ... + n_D = N1$ etc. ...
et $N = N_1 + N_2$
Dans l'exemple ci-dessus:

	observation			
valeurs de V	A	B	C	Σ
effectifs 1er échantil.	12	22	9	43
effectifs 2ème échan.	32	14	6	52

3) Hypothèse nulle.

Une hypothèse très souvent examinée est la suivante:

H: "Les deux échantillons sont tirés du même ensemble parent"

Ce qui signifie que "les proportions des différents types d'observations y sont les mêmes". Cette hypothèse est dite "hypothèse d'homogénéité". Elle s'exprime sous forme d'une hypothèse nulle.

Mais quelles sont les proportions dans l'ensemble parent?

La meilleure représentation de cet ensemble parent est fournie par la réunion des deux échantillons. Elle comprend Σ_1 observations de la valeur v_1, Σ_2 observations de la valeur v_2, Σ_3 observations de la valeur v_3, ..., Σ_P observations de la valeur P dans un échantillon de N observations.

Dans notre exemple, si les deux écoles sont extraites d'un même ensemble parent, **les proportions d'élèves présentant chaque comportement sont les mêmes dans les deux écoles**, à des écarts d'échantillonnage près.

Les proportions des observations A, B, C ... sont les mêmes dans les deux échantillons.

4) Modèle: distribution théorique

Cette hypothèse d'homogénéité permet de fabriquer un tableau qui lui correspond: une distribution dite "théorique":

Dans l'exemple ci-dessus, il y a en tout 95 élèves, et sur ces 95 élèves, 44 ont le comportement A. Il devrait y avoir 44/95 des élèves qui ont le comportement A dans l'école 1, c'est-à-dire 44/95 x 43 et aussi 44/95 des élèves de l'école 2, c'est-à-dire 44/95 x 52.

Plus généralement, t_I, le nombre attendu d'observations d'une valeur I de la variable, dans l'échantillon 1, sous l'hypothèse nulle, sera :

$$t_I = \frac{\Sigma_I \times N_1}{N}$$

Avec Σ_I = le nombre total d'observations de la valeur I sur les deux échantillons,
N_1 = l'effectif du premier échantillon,
N l'effectif des deux échantillons réunis.

On obtient alors le tableau des valeurs théoriques du modèle. Remarquons que les sommes des valeurs par ligne et colonnes restent celles du tableau de contingence.

	modèle					
valeurs de V	v_1	v_2	v_3	...	v_p	Σ
effectifs 1er échantil.	t_1	t_2	t_3	...	t_D	N_1
effectifs 2ème échan.	t_1'	t_2'	t_3'	...	t_D'	N_2
total	Σ_1	Σ_2	Σ_3	...	Σ_D	N

Dans l'exemple ci-dessus, il vient le modèle suivant:

	modèle			
valeurs de V	A	B	C	Σ
effectifs 1er échantil.	19,9	16,3	6,8	43
effectifs 2ème échan.	24,1	19,7	8,2	52
total	44	36	15	95

5) Distance entre la contingence et le modèle:

La distance du χ^2 entre deux tableaux se calcule comme la distance du χ^2 entre deux suites de nombres. On fait la somme des contributions de toutes les cases

Cette distance est donnée par la formule des fiches précédentes:

$$\chi^2_{\text{observé}} = \sum_i \frac{(E_T - E_O)^2}{E_T}$$

dans laquelle E_T est l'effectif théorique d'une case (déterminée par une valeur de la variable et un échantillon), E_O l'effectif observé correspondant. La somme Σ porte sur lignes du tableau correspondant aux deux échantillons.

Et avec les notations ci-dessus, cette formule s'écrit:

$$\chi^2_{\text{observé}} = \Sigma_I \frac{(t_I - n_I)^2}{t_I}$$

Dans notre exemple, il vient:

$$\chi^2_o = \frac{(19,9 - 12)^2}{19,9} + \frac{(24,1 - 32)^2}{24,1} + \frac{(16,3 - 22)^2}{16,3} + \frac{(19,7 - 14)^2}{19,7} + \frac{(6,8 - 9)^2}{6,8} + \frac{(8,2 - 6)^2}{8,2}$$

$\chi^2_o = 3,14 + 2,59 + 1,99 + 0,71 + 1,65 + 0,59$

$\chi^2_o = 10,67.$

6) Signification de cette distance: mesure de sa "rareté"

Pour déterminer la rareté de cette valeur du $\chi^2_{\text{observé}}$, nous la comparons comme dans les cas précédents, à une distribution obtenue dans des conditions comparables, donnée par la table du χ^2. Ces conditions sont déterminées par le "degré de liberté" du modèle dont on étudie la distribution.

Ici, le degré de liberté est **dl = p - 1** (l'explication de cette notion se trouve dans la fiche 4)

Dans l'exemple que nous donnons, dl = 3 - 1 = 2. Il faut donc consulter la deuxième ligne de la table du χ^2. On y trouve que les χ^2_{seuils} correspondant

respectivement à: sont:

Seuil k	5%	1%	0,1%
$\chi^2_{\text{seuil k}}$	5,99	9,21	13,82

Le meilleur encadrement est: $\chi^2_{\text{seuil 1\%}} < \chi^2_{\text{observé}} < \chi^2_{\text{seuil 0,1\%}}$

7) Décision

Si χ^2_o est significatif à 5% ($>\chi^2_{\text{seuil 0,05}}$), ou moins de 5%, on rejette l'hypothèse nulle, sinon on ne peut pas la rejeter.

Dans l'exemple donné, le χ^2_o est significatif à 1%, l'hypothèse nulle doit donc être rejetée: face à l'échec, les élèves ont des réactions différentes dans une école et dans l'autre.

ECHANTILLONS HOMOGENES.

Des échantillons sont homogènes si on peut les considérer comme tirés au hasard dans une même population parente. Si ces observations consistent à noter une variable nominale, la population parente comprend une certaine proportion de sujets pour chacune des valeurs de la variable: soit p_A, p_B ... ces proportions. Si deux (ou plus) échantillons sont homogènes, c'est-à-dire issus de cette population parente, les proportions de A, B, etc. dans chacun d'eux doivent y être les mêmes (aux écarts d'échantillonnage près) que dans la population parente: p_A, p_B ....

Si les échantillons sont tirés du même ensemble parent (inconnu) la meilleure représentation de la distribution de cet ensemble parent est formée par la réunion de tous ces échantillons.

Cette réunion comprendra $\Sigma_A = n_A + n_A'$ observations de la valeur A, $\Sigma_B = n_B + n_B'$ observations de la valeur B, etc et en tout: $N = n_1 + n_2 + ...$ observations.

Par conséquent $p_A = \frac{\Sigma_A}{N}$, $p_B = \frac{\Sigma_B}{N}$, etc...

Donc sur n_1 sujets observés dans le premier échantillon (puisque $n_A + n_B + n_C + n_D = n_1$) il doit y avoir $p_A \times n_1$ sujets dans la catégorie A, $p_B \times n_1$ sujets dans la catégorie B, etc...

Les amateurs de règle de trois peuvent dire aussi que: si sur N sujets, il y en a Σ_A dans la catégorie A, sur 1 sujet il devrait y en avoir N fois moins $\frac{\Sigma_A}{N}$, et sur n_1 sujets, n_1 fois plus $\frac{\Sigma_A \times n_A}{N}$.

Note: une distribution homogène est une distribution uniforme: toutes les catégories ont la même proportion.

TABLE DES VALEURS CRITIQUES DE CHI CARRE

dl	Probabilité sous H_0 que χ^2 >= Chi carré													
	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.00016	.00063	.0039	.016	.064	.15	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.18	.35	.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.43	.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

K échantillons, UNE variable nominale à P valeurs.

Cette épreuve permet de comparer les distributions sur une même variable de K échantillons, afin de savoir si on peut les considérer comme homogènes (issus d'une même population parente).

1) Recueil des données.

Pour chacun des éléments d'un échantillon on relève la valeur d'une variable V. à p valeurs.

éléments 1er échantillon	a_1	b_1	c_1	d_1	e_1	...
valeurs de la variable	V_{a1}	V_{b1}	V_{c1}	V_{d1}	V_{e1}	...
éléments 2ème échantillon	a_2	b_2	c_2	d_2	e_2	...
valeurs de la variable	V_{a2}	V_{b2}	V_{c2}	V_{d2}	V_{e2}	...
...						
éléments kème échantillon	a_k	b_k	c_k	d_k	e_k	...
valeurs de la variable	V_{ak}	V_{bk}	V_{ck}	V_{dk}	V_{ek}	...

Exemple

On veut comparer 3 manuels traitant du même cours, du point de vue de l'importance qu'ils accordent à tel ou tel type ou thème d'exercice. Il y a 5 types d'exercices: A, B, C, D, E, (par exemple les 5 parties d'un programme, ou encore problèmes ouverts, problèmes avec correction, problèmes, exercices d'application, exercices d'entraînement). Les 3 ouvrages ont ils la même composition (à des erreurs d'échantillonnage près?

	observation					
Manuels\Types exercices	A	B	C	D	E	Σ
nbre d'exercices dans le manuel 1	8	15	40	16	2	81
nbre d'exercices dans le manuel 2	4	7	75	107	14	207
nbre d'exercices dans le manuel 3	1	0	31	60	10	102
Total	13	22	146	183	26	390

2) Présentation réduite

Distribution observée des EFFECTIFS ou distribution contingente.

	observation					
valeurs de V	v_1	v_2	v_3	...	v_p	Σ
eff 1er échantillon	$n_{1.1}$	$n_{2.1}$	$n_{3.1}$	...	$n_{P.1}$	N_1
2ème échantillon	$n_{1.2}$	$n_{2.2}$	$n_{3.2}$	...	$n_{P.2}$	N_2
...	...	...	...	...	...	...
Kème échantillon	$n_{1.K}$	$n_{2.K}$	$n_{3.K}$	...	$n_{P.K}$	N_K
Total	Σ_1	Σ_2	Σ_3	...	Σ_P	N

Dans ce tableau,

$\Sigma_1 = n_{1.1} + n_{1.2} + ... + n_{1.K}$ etc. ...;

$N_1 = n_{1.1} + n_{2.1} + n_{3.1} + ... + n_{P.1}$ etc. ... et

$N = N_1 + N_2 + ... + N_K$

3) Hypothèse nulle

Une hypothèse très souvent examinée est la suivante:

H: "Les deux échantillons sont tirés du même ensemble parent"

Ce qui signifie que "les proportions des différents types d'observations y sont les mêmes". Cette hypothèse est dite "hypothèse d'homogénéité". Elle s'exprime sous forme d'une hypothèse nulle.

Mais quelles sont les proportions dans l'ensemble parent?

La meilleure représentation de cet ensemble parent est fournie par la réunion des k échantillons. Le premier comprend N_1 éléments, le second N_2 etc. Elle comprend Σ_1 observations de la valeur v_1, Σ_2 observations de la valeur v_2, Σ_3 observations de la valeur v_3, ..., Σ_P observations de la valeur P dans un échantillon de N observations.

Dans notre exemple, si les trois manuels ont la même composition, ils doivent présenter la même prportion d'exercices de chaque type.

Les proportions des types d'exercices A, B, C, D et E ... sont les mêmes dans les trois ouvrages.

4) Modèle: distribution théorique

L'hypothèse nulle permet de fabriquer un tableau qui lui correspond: une distribution dite "théorique":

Soit $t_{i,j}$ le nombre attendu d'observations d'une valeur i de la variable, dans l'échantillon j. Sous l'hypothèse nulle, ce nombre sera: $t_{i,j} = \frac{\Sigma_i \times N_j}{N}$

Avec Σ_i = le nombre total d'observations de la valeur i sur les k échantillons,

N_j = l'effectif du j-ième échantillon,

N l'effectif des K échantillons réunis.

modèle

valeurs de V Effectifs	v_1	v_2	v_3		v_p	somme
1er échantillon	$\Sigma_1 x\ n_1/N$	$\Sigma_2 x\ n_1/N$	$\Sigma_3 x\ n_1/N$	...	$t_{p,1}$	N_1
2ème échantillon	$\Sigma_1 x n_2/N$	$\Sigma_2 x n_2/N$	$\Sigma_3 x n_2/N$	...	$t_{p,1}$	N_2
...	...	...	...	...	...	...
k-ième échantillon	$t_{1,k}$	$t_{2,k}$	$t_{3,k}$	...	$t_{p,k}$	N_k
Total	Σ_1	Σ_2	Σ_3	...	Σ_p	N

Dans l'exemple ci-dessus, il vient le modèle suivant:

modèle

valeurs de V	A	B	C	D	E	Σ
Types d'exercices du Manuel 1	2,7	4,6	30,3	38	5,4	81
Manuel 2	6,9	11,7	77,5	97	13,8	207
Manuel 3	3,4	5,7	38,2	48	6,8	102
Total	13	22	146	183	26	390

Effectif théorique minimum, regroupements

Lorsque l'effectif théorique d'une case (même celui d'une seule case) devient inférieur à 8, (on tolère même 5 dans le domaine des sciences de l'éducation), il n'est plus possible d'appliquer le test du χ^2. En effet, la contribution de cette case auχ^2 devient, très importante - puisque l'effectif théorique figure au dénominateur - , et très incertaine. Dans ce cas il convient d'effectuer des regroupements afin de dépasser cette valeur DANS TOUTES LES CASES du nouveau tableau, à la condition bien sûr que ces regroupements puissent recevoir une signification. Dans notre exemple, deux cases de la première colonne sont faibles et une de la deuxième. Regroupons ces

deux colonnes en considérant, soit un type commun, soit un thème regroupant les deux thèmes A et B.
Le tableau de la distribution observée devient après regroupement:

valeurs de V	A ou B	C	D	E	Σ
Types d'exercices du Manuel 1	23	40	16	2	81
Manuel 2	11	75	107	14	207
Manuel 3	1	31	60	10	102
Total	35	146	183	26	390

Le tableau de la distribution théorique devient après regroupement:

valeurs de V	A ou B	C	D	E	Σ
Types d'exercices du Manuel 1	7,3	30,3	38	5,4	81
Manuel 2	18,6	77,5	97	13,8	207
Manuel 3	9,1	38,2	48	6,8	102
Total	35	146	183	26	390

Plus aucune valeur n'y est inférieure à 5, le calcul peut continuer.

5) Distance entre la contingence et le modèle

La distance du χ^2 entre deux tableaux se calcule comme la distance du χ^2 entre deux suites de nombres. On fait la somme des contributions de toutes les cases
Cette distance est donnée habituellement par la formule:

$$\chi^2_{\text{observé}} = \sum_i \sum_j \frac{(E_T - E_O)^2}{E_T}$$

dans laquelle E_T est l'effectif théorique d'une case déterminée par une valeur de la variable i et un échantillon j, E_O l'effectif observé correspondant.
Dans notre exemple il vient:…

$$\chi^2_o = \frac{(7,3 - 23)^2}{7,3} + \frac{(18,6 - 11)^2}{18,6} + \ldots = 69,2.$$

6) Signification de cette distance: mesure de sa "rareté"

Pour déterminer la rareté de cette valeur du $\chi^2_{\text{observé}}$, nous la comparons comme dans les cas précédents à une distribution obtenue dans des conditions comparables, donnée par la table du χ^2. Ces conditions sont déterminées par le "degré de liberté" du modèle dont on étudie la distribution.
Ici le degré de liberté est dl = (p - 1)(k-1).

Calcul du degré de liberté.

Le degré de liberté d'un χ^2 est déterminé par le nombre de valeurs que l'on peut fixer de manière arbitraire lors de la construction du modèle. Ainsi dans la première leçon, le total des deux cases est fixé (égal à l'effectif de l'échantillon): on ne peut choisir l'effectif que d'une case, l'autre est alors déterminé; le degré de liberté est 1. Dans les distributions à plusieurs lignes les valeurs des marges sont fixées: effectifs des échantillons, effectifs des comportements dans l'ensemble des échantillons ou des thèmes. On peut choisir p-1 valeurs dans une ligne qui en comprend p. Mais

on ne peut choisir des valeurs que pour k-1 lignes (dans la leçon 4 et 1 ligne dans la leçon 3), car la dernière ligne se déduit des précédentes et du total.

Note. De façon plus générale la distribution du χ^2 est celle de la somme des carrés de dl variables aléatoires centrées réduites.

Dans l'exemple que nous donnons,

dl = (p-1)(k-1) = 3 x 2 = 6

Il faut donc consulter la $6^{ième}$ ligne de la table du χ^2. (p. 33)

On y trouve que les χ^2_{seuils} correspondants

respectivement à sont:

Seuil k	5%	1%	0,1%
$\chi^2_{seuil\ k}$	12,59	16,81	22,46

Le meilleur encadrement donné par la table est: $\chi^2_{seuil\ 0,1\%} < \chi^2_{observé}$

7) Décision

Si χ^2_o est significatif à 5% ou à une valeur plus faible, on rejette l'hypothèse nulle, sinon on ne peut pas la rejeter.

Dans l'exemple donné, le χ^2_o est significatif à 0,1%. Il est énorme, l'hypothèse nulle doit donc être rejetée: les différents manuels ne répartissent pas leur exercices de la même façon.

Le choix de l'hypothèse nulle

On peut essayer de tester n'importe quelle hypothèse, mais

a) Il faut d'abord qu'elle se traduise par un modèle dont on peut calculer les éléments,

b) Il vaut mieux choisir les hypothèses qui donneront le maximum d'information avec le moins de calcul, c'est-à-dire celles qui élimineront à chaque étape, si le test est significatif, le plus de cas envisageables. On tient alors compte du fait suivant: ne pas rejeter une hypothèse n'apporte vraiment pas beaucoup d'information, en rejeter une trop particulière, non plus. D'autre part, plus on regroupe de catégories, plus on perd de l'information et plus les différences ont tendance à s'estomper;

c) Il faut tenir compte de l'intérêt que les conclusions présenteront pour les utilisateurs 1) Toutes les hypothèses ne sont pas également intéressantes, 2) Il faut se demander ce qu'on fera de la réponse lorsque le résultat sera obtenu;

d) Toutes les formulations de l'hypothèse nulle qui aboutissent au même modèle sont équivalentes, (en particulier, qu'elles soient formulées positivement ou négativement). Exemple: Si l'on compare deux échantillons d'élèves du point de vue de la réussite à un exercice, il peut paraître intéressant de savoir s'il n'y aurait pas deux fois plus de sujets d'une certaine catégorie (la réussite) dans l'un que dans l'autre.

Or, Si le test n'est pas significatif, on ne pourra pas pour autant "accepter cette hypothèse" non rejetée, nous l'avons vu. On pourrait recommencer le test avec l'hypothèse qu'il y a trois fois plus de sujets, ou le même nombre sans pouvoir les rejeter non plus.

Si le test est significatif, on peut rejeter l'hypothèse nulle. Et on peut encore en envisager maintenant beaucoup d'autres comme ci-dessus, qui n'ont pas été rejetées. Tester H_0: « *Le nombre des réussites dans le premier échantillon est significativement supérieur à celui des réussites dans le second* » regrouperait en fait un très grand nombre d'hypothèses particulières du genre de celles citées ci-dessus, et serait donc une hypothèse nulle beaucoup plus intéressante:

- Si la réponse est oui on peut essayer de préciser la réponse et on a éliminé la moitié des possibilités: (résultats significativement inférieurs)
- mais si elle est « non », il est inutile d'examiner aucune d'entre elles.

Il faut considérer que l'hypothèse nulle **OPPOSE** un ensemble d'hypothèses à un ensemble d'autres. Plus la taille des deux ensembles est égale et plus le test sera efficace.

Exercice: Pourquoi l'hypothèse nulle "il n'y a pas de différence significative entre les deux échantillons" est-elle encore plus efficace à priori que la précédente ?

FICHE 5

UNE variable nominale à DEUX valeurs et k échantillons appariés, ou UN échantillon et k réplications de la variable, ou UN échantillon et k variables à deux valeurs Test Q de Cochran

Cette épreuve permet de comparer l'effet de k diverses conditions sur une variable à deux valeurs, ou de savoir si un ensemble de k fréquences diffèrent significativement entre elles, ou si k échantillons (appariés), peuvent être considérés comme indépendants et homogènes (issus d'une même population parente).

1) Recueil des données.

Pour chacun des N éléments d'un échantillon, on note les observations de k variables à deux valeurs: a, b, c, ...k. ou k conditions d'une même variable. Chaque variable ne présente que deux valeurs. $V_{i,j}$ représente la valeur,1ou 0, obtenue par le sujet j à la variable ou dans la condition i.

échantillon\conditions	A	b	c	...	k
élément 1 de l'échantillon	V_{a1}	V_{b1}	V_{c1}	...	V_{k1}
élément 2 de l'échantillon	V_{a2}	V_{b2}	V_{c2}	...	V_{k2}
...	...	...	...	...	...
élément N de l'échantillon	V_{aN}	V_{bN}	V_{cN}	...	V_{kN}

On peut aussi bien recueillir les résultats de k échantillons a, b, ...k dont les éléments sont "appariés", c'est à dire mis en correspondance par un moyen quelconque (On espère par ce moyen diminuer les variations parasites). Tous les éléments 1 des différents échantillons ont un caractère ou un lien commun, de même tous les éléments 2 etc.

Remarque sur les échantillons appariés

Qu'il s'agisse d'examiner, grâce à un échantillon de sujets observés,

- l'effet de diverses conditions sur une même variable V/C1, V/C2, V/C3...,
- ou l'évolution d'un comportement avant et après un événement (test et retest: VA, VB) - ou l'évolution d'une variable au cours d'une séquences d'événements (V(T1), V(T2), V(T3)...), le problème est identique. D'ailleurs les données peuvent être présentées de la même façon: les sujets sont disposés dans une première colonne, les données recueillies à leur endroit pour chaque condition, ou chaque variable ou chaque étape sont sur la même ligne.

Il peut toutefois arriver que l'on ne puisse pas présenter toutes les conditions, ou tester plusieurs fois ou noter toutes les variables, pour les mêmes individus, on est alors conduit à utiliser plusieurs échantillons différents. Par exemple on prend au hasard la moitié des sujets (échantillon 1) à qui on présente le test avant l'événement, alors qu'on présentera le test après l'événement à l'autre moitié des sujets (échantillon 2). Il est clair que le risque est grand que les différences observées entre le test Avant et le test Après soient dues non pas à l'effet de l'événement, mais à des différences entre les sujets eux mêmes. Pour diminuer ce risque on peut **MAINTENIR EGALES** dans les deux échantillons les sources présumées de différences en créant des échantillons appariés, c'est-à-dire en associant à chaque sujet a_i du premier échantillon un sujet b_i du second qui présente les mêmes caractères.

On suppose alors que les différences individuelles entre a_i et b_i étant réduites, ils auraient répondu de la même façon aux deux tests. La différence entre la réponse de l'un avant et la réponse de l'autre après serait donc principalement imputable à l'événement survenu entre les deux tests.

La disposition des données devient alors la suivante:

couple	Variable 1	Variable 2	...
(a_1, b_1) ... (a_i, b_i)	valeur obtenue par celui, de a_i ou de b_i, qui a subi V1 ou C1	valeur obtenue par celui, de a_i ou de b_i, qui a subi V2 ou C2	...

2) Exemples

Exemple 1

La "même" question de mathématiques est posée trois fois aux mêmes 20 élèves mais dans des conditions C1, C2, C3 différentes (par exemple à l'issue d'exercices différents, ou à diverses époques de l'année, ou encore posés par des personnes différentes), sans que la correction soit donnée. On note « 1 » les réponses justes et « 0 » les fausses. On veut savoir si ces conditions influent sur les réponses des élèves.

Le tableau des réponses est le suivant, complété par le total Li de la ième ligne, et son carré Li^2, par la somme de chaque colonne Gj, et par la somme des totaux des lignes ΣLi ainsi que la somme des carrés $\Sigma(Li^2)$

élèves\ conditions	C1	C2	C3	Li	Li^2
1	1	1	1	3	9
2	1	1	1	3	9
3	0	0	0	0	0
4	0	1	0	1	1
5	1	1	1	3	9
6	0	1	0	1	1
7	1	0	0	1	1
8	1	1	1	3	9
9	1	0	0	1	1
10	0	0	0	0	0
11	0	0	0	0	0
12	1	0	0	1	1
13	1	0	0	1	1
14	0	1	0	1	1
15	1	1	1	3	9
16	1	1	0	2	4
17	0	1	0	1	1
18	1	1	1	3	9
19	0	1	0	1	1
20	1	1	1	3	9
	G1 = 12	G2 = 13	G3 = 7	$\Sigma Li = 32$	$\Sigma Li^2 = 76$

Exemple 2.

Trois écoles C1, C2, et C3 ont choisi chacune 20 élèves qui s'affrontent amicalement dans un tournoi de mathématiques. Chaque élève reçoit un problème différent, il y a donc 20 énoncés numérotés de 1 à 20. On note dans chaque cas l'information, 1 réussite, 0 échec.

On veut savoir si les résultats des trois groupes de champions diffèrent.

Exemple 3
Trois élèves C1, C2, et C3 comparent leurs réussites sur les vingt exercices qui leur ont été proposés au cours du trimestre. Est-ce qu'ils réussissent différement?

Exemple 4.
20 élèves ont traité trois exercices différents: C1, C2, C3,
Ces exercices différencient-ils les élèves?

Dans ces quatre exemples, les échantillons sont toujours les objets C1, C2, C3 et leurs éléments sont appariés chaque fois par les lignes: c'est le même individu en 1 et 4, ou c'est le même problème, en 2 et 3 qui apparient les éléments de ces échantillons.

3) Hypothèse nulle

H_0: Les succès et les échecs sont distribués au hasard dans les lignes et les colonnes.
Remarquons que cette hypothèse nulle implique les deux suivantes:
H: les trois colonnes devraient contenir le même nombre de succès,
H': toutes les lignes devraient contenir "à peu près le même nombre" de 0, (donc le nombre des lignes composées uniquement de 1 ou de 0 ne devrait pas être très grand).
Nous savons déjà éprouver H, il suffit de comparer, à l'aide du χ^2. :

les effectifs observés	G1	G2	G3
aux effectifs théoriques	k x N / 3	k x N / 3	k x N / 3

Ici on suppose de plus que les réponses dans les conditions différentes sont indépendantes les unes des autres. L'hypothèse nulle pourrait être rejetée même si les pourcentages de réponses justes étaient identiques, par exemple si trop d'élèves répondaient de la même façon (juste ou fausse) aux trois questions.
Cette exigence additionnelle peut surprendre, mais elle permet de conclure en cas de rejet que, soit les conditions des expériences, soit les élèves, soit les deux, différencient les résultats.

4) Distance entre la contingence et le modèle

Cochran a montré que, sous les conditions de l'hypothèse nulle, le coefficient Q suivant est distribué comme χ^2 avec dl = k-1 :

$$Q = \frac{k \times (k-1) \times \Sigma_j (G_j - m(G))^2}{(k \times \Sigma_i Li) - \Sigma_i Li^2}$$

dans lequel
k est le nombre de variables,
G_j est la somme de la j ème colonne,
m(G) est la moyenne des colonnes,
$(G_j - m(G))^2$ est le carré de la différence entre G_j et m(G), pour la jème colonne,
$\Sigma_j (G_j - m(G))^2$ est la somme des carrés de ces différences pour toutes les colonnes,
Li le total de la ième ligne,
Li^2 le carré de Li,
Σ_i Li la somme des Li pour toutes les lignes,
$\Sigma_i Li^2$ la somme des carrés pour toutes les lignes.

La formule ci-dessous, équivalente à la précédente, est plus facile à calculer.

$$Q = \frac{(k-1) \times [k.\Sigma_j G_j^{\,2} - (\Sigma G_j)^2]}{(k \times \Sigma_i Li) - \Sigma_i Li^2}$$

Dans le cas de notre exemple, nous trouvons:

$\Sigma G_j = 32$ $(\Sigma G_j)^2 = 1024$ $k.\Sigma_j G_j^{\,2} = 1086$

$[k.\Sigma_j G_j^{\,2} - (\Sigma G_j)^2] = 62$

$Num = (k-1) \times [k.\Sigma_j G_j^{\,2} - (\Sigma G_j)^2] = 124$

$(k \times \Sigma_i Li) = 96$ $\Sigma_i Li^2 = 76$

$Den = (k \times \Sigma_i Li) - \Sigma_i Li^2 = 20$

$Q = Num / Den = 6,2$

5) Signification de cette distance: mesure de sa rareté

Pour déterminer la rareté de cette valeur Q nous pouvons la considérer comme un $\chi^2_{observé}$. Nous la comparons comme dans les cas précédents à la distribution obtenue dans des conditions comparables et donnée par la table du χ^2. Ces conditions sont déterminées par le "degré de liberté" du modèle dont on étudie la distribution:

Dans notre exemple dl = k-1 = 2 et
$X^2_{observé} > \chi^2_{seuil\ 0,05}$ puisque $6,2 > 5,99$

6) Décision

L'hypothèse nulle doit être rejetée:
Les succès et les échecs ne sont pas distribués au hasard dans les lignes et les colonnes.
Soit les conditions ont influencé les résultats, soit les réponses des élèves ne sont pas indépendantes.

Un examen même superficiel du tableau des résultats laisse supposer que c'est la deuxième hypothèse qui est vérifiée, car la proportion des patrons 1,1,1 ou 0,0,0 est étrangement élevée par rapport à celle des patrons 1,0,0, et surtout 0,0,1.

D'autres part, nous aurions pu mettre à l'épreuve le fait que les réussites dans les trois conditions ne diffèrent pas significativement en utilisant le test d'homogénéité présenté précédemment:

les effectifs observés	12	13	7	
aux effectifs théoriques	10,66	10,66	10,66	
différences	1,33	2,33	3,66	
contributions	1,33/10,66	...	...	
	0,166	0,510	1,26	total: 1,94

$\chi^2 = 1,94 < \chi^2_{seuil\ 0,0\ 5}$ les réussites ne diffèrent pas suivant les conditions.

Ainsi les résultats ne diffèrent pas, mais ils ne sont pas indépendants. Le test de Cochran est plus puissant: il montre des différences que ne détecte pas le test d'homogénéité.

Comparaison avec le test d'homogénéité des proportions. Etude de la formule.

La formule peut se décomposer ainsi :

$\frac{\Sigma_j (G_j - m(G))^2}{m(G)} = \chi^2_{observé}$ correspondant à l'homogénéité des réponses dans les différentes conditions. Il indique lorsqu'il est significatif que le nombre des réponses varie suivant les conditions. Dans l'exemple ci-dessus, il n'est pas significatif: 12,13 et 7 ne sont pas assez différents.

$k \times m(G) = \Sigma_i Li$ donc

$\Sigma_j (G_j - m(G))^2 = m(G).\chi^2_{observé}$

$= \Sigma_i Li / k .\chi^2_{observé}$

d'où

$$Q = \frac{(k-1)\ (\Sigma_i Li).\chi^2_{observé}}{((k \times \Sigma_i Li) - \Sigma_i Li^{2)}}$$

$$Q = \frac{(k-1)\ .(\Sigma_i Li).\chi^2_{observé}}{k\ x\ (\Sigma_i Li - \frac{\Sigma_i Li^2}{k})}$$

$$Q = \frac{(k-1)}{k} \times \frac{\chi^2_{observé}}{[\ 1 - \frac{\Sigma_i Li^2}{k\ (\Sigma_i Li)}]}$$

B. VARIABLES NOMINALES

II. INDEPENDANCE ET DEPENDANCES

FICHE 6

UN échantillon, DEUX variables à DEUX valeurs: Indépendance

Cette épreuve permet de dire si deux variables sont indépendantes ou si elles sont liées, c'est-à-dire si les résultats de l'une dépendent de résultats de l'autre.

1) Recueil des données.

Pour chacun des éléments de l'échantillon on relève les valeurs des 2 variables V_1, V_2, à 2 valeurs. A ou B. Ces valeurs peuvent être notées 1 et 0

éléments de l'échantillon	a	b	c	d	e	...
valeurs de la variable 1	V_{a1}	V_{b1}	V_{c1}	V_{d1}	V_{e1}	...
valeurs de la variable 2	V_{a2}	V_{b2}	V_{c2}	V_{d2}	V_{e2}	...

Exemple 1 : Reprenons l'exemple donné dans la leçon précédente et comparons seulement les réponses des élèves dans les deux premières conditions C1 et C2.

Reprenons les données:

élève	C1	C2
1	1	1
2	1	1
3	0	0
4	0	1
5	1	1
6	0	1
7	1	0
8	1	1
9	1	0
10	0	0
11	0	0
12	1	0
13	1	0
14	0	1
15	1	1
16	1	1
17	0	1
18	1	1
19	0	1
20	1	1
	S1 = 12	S2 = 13

2) Distribution réduite

	C1	C2	Σ
effectif des réussites	12	13	25
effectif des échecs	8	7	15
effectif total	20	20	40

3) Hypothèse nulle.

H: "Les résultats obtenus à une variable ne dépendent pas de ceux obtenus à l'une autre variable".

Dans le modèle correspondant, l'effectif d'une case déterminée est proportionnel à l'effectif de la ligne et aussi à l'effectif de la colonne au croisement desquelles elle se trouve.

H_0: Le tableau des observations ne diffère pas significativement du modèle. On dira dans ce cas que les deux variables observées sont **INDEPENDANTES**.

Dans notre exemple, l'hypothèse nulle veut dire que les élèves qui échouent dans les conditions C1 devraient avoir les mêmes chances que les autres élèves de réussir dans les conditions C2. De même, parmi les élèves qui réussissent dans C1 il devrait y avoir une proportion d'échecs lors de C2 égale à celle que l'on trouve parmi les élèves en échec dans C1.

4) Modèle: Table de contingence.

L'hypothèse nulle porte sur la comparaison des populations suivantes:

* les élèves qui ont réussi lors de C1, et réussi en C2, (1,1). Leur nombre est n(1,1)
* les élèves qui ont réussi lors de C1, et échoué en C2, (1,0). Leur nombre est n(1,0)
* les élèves qui ont échoué lors de C1, et réussi en C2, (0,1). leur nombre est n(0,1)
* les élèves qui ont échoué lors de C1, et échoué en C2, (0,0). Leur nombre est n(0,0)

La distribution réduite ne permet pas de trouver leur effectif: Il faut revenir au tableau initial et le résumer d'une façon différente en faisant apparaître tous les cas possibles pour les valeurs des deux variable. Le résultat est un tableau à double entrée appelé distribution "croisée" ou tri "croisé".

Cette nouvelle table dite "de contingence" présente la distribution des effectifs sur le couple de variables C1 x C2. Ces effectifs sont les valeurs observées.

distribution croisée

effectifs	Echec C1	Réuss C1	Σ
réussites lors de C2	n(0,1)	n(1,1)	n(.1)
échecs lors de C2	n(0,0)	n(1,0)	n(.0)
effectif total	n(0.)	n(1.)	N

Nous retrouvons l'effectif total N, les effectifs des réussites lors de C1, que nous notons n(1.), les réussites lors de C2: n(.1), etc.

Remarque

Le tableau du paragraphe précédent présentait 2 distributions des effectifs l'une sur la variable "réussite dans les deux conditions" l'autre sur la variable "échecs dans les deux conditions", mais il perdait les informations d'appariement entre les observations des deux variables.

Exercice:

Observer les différences entre les deux types de tableaux, s'entraîner à déduire le premier du second et observer les renseignements qui disparaissent.

Dans l'exemple proposé le tableau de contingence est:

effectifs	Echec C1	Réuss C1	Σ
réussites lors de C2	5	8	13
échecs lors de C2	3	4	7
effectif total	8	12	20

5) Modèle: Table des effectifs théoriques

Il faut trouver un modèle, le plus près possible de nos résultats observés: on conservera les mêmes effectifs "marginaux":

effectifs	Echec C1	Réuss C1	
réussites lors de C2			13
échecs lors de C2			7
effectif total	8	12	20

mais ce modèle devra réaliser l'hypothèse H qui engendre l'hypothèse nulle:

Parmi les 12 élèves qui réussissent dans C1 il devrait y avoir une proportion d'échecs lors de C2 [nous allons remplir la case n(1,0)], égale à celle que l'on trouve parmi les élèves en échec dans C1. Dans ces conditions cette proportion serait la même que dans l'ensemble de la population: 7 sur 20.

7/20 de 12 élèves = (7 x 12)/20 = 4,2

est la valeur théorique de n(1,0)

Les valeurs théoriques des autres cases se déterminent de la même manière ou par soustraction aux marges.

De façon générale la valeur théorique $V_T(i,j)$ d'une case dont les marges sont n(i .) et n(.j) est

$$V_T(i,j) = \frac{n(i\ .)\ x\ n(.j)}{N}$$

dans laquelle N est le nombre total de sujets observés.

Finalement le tableau théorique est:

effectifs	Echec C1	Réuss C1	
réussites lors de C2	5,2	7,8	13
échecs lors de C2	2,8	4,2	7
effectif total	8	12	20

6) Distance du modèle théorique au tableau de contingence.

La distance entre deux distributions en tableaux s'évalue comme nous l'avons fait pour les distributions à une variable, à l'aide de la distance du χ^2.

La valeur:

$$\chi^2 = \sum_{i,j} \frac{[V_T(i,j) - V_O(i,j)]^2}{V_T(i,j)} \qquad (1)$$

suit la distribution du χ^2, avec dl = 1 puisqu'il suffit de déterminer dans le modèle la valeur d'une case pour que le modèle soit déterminé.

Formule Réduite.

Dans le cas particulier où le tableau concerne deux variables binaires, ou deux échantillons distribués selon une variable binaire (tableau 2x2), le calcul du peut s'effectuer directement et plus rapidement avec la **formule réduite** ci dessous. Elle est équivalente à (1)

Pour l'appliquer il est commode de désigner les cases du tableau par des majuscules A,B,C et D (avec A et D en diagonale ainsi que B et C):

effectifs	Echec C1	Réuss C1	Σ
réussites lors de C2	A	B	v1= A+B
échecs lors de C2	C	D	v2 = C+D
effectif total	n1 = A+C	n2 = B+D	N

alors:

$$\chi^2 = \frac{N.(A.D - B.C)^2}{(A + B).(C + D).(A + C).(B + D)} \qquad (2)$$

Cas des petits échantillons: Correction de YATES

Si 20 < N < 50 il est préférable d'appliquer à cette formule la correction de YATES:

$$\chi^2 = \frac{N.([A.D - B.C] - N/2)^2}{(A + B).(C + D).(A + C).(B + D)} \qquad (3)$$

Dans l'exemple que nous avons choisi, il n'est pas possible d'utiliser le test du χ^2 car les effectifs sont trop faibles.

REMARQUE: Il existe pour ce cas là une méthode: celle de FISHER

Cas où certaines cases ont un effectif faible

ATTENTION! Il n'est naturellement pas possible de modifier les effectifs observés ! En particulier de multiplier tous les effectifs par un nombre arbitraire, sous le prétexte que les proportions seraient conservées! (dans notre exemple, multiplier par 2 rendrait les V_T supérieures à 5)

Exemple 2.

L'effectif étant trop faible, l'expérience est continuée et donne les résultats suivants sur n=50 élèves.

effectifs	Echec C1	Réuss C1	
réussites lors de C2	16	18	34
échecs lors de C2	5	11	16
effectif total	21	29	50

Calcul des valeurs théoriques:

$V_T(1,1) = 29 \times 34 / 50 = 19,72$

$V_T(1,0) = 29 \times 16 / 50 = 9,28$

$V_T(0,1) = 21 \times 34 / 50 = 14,28$

$V_T(0,0) = 21 \times 16 / 50 = 6,72$

Calcul du χ^2

V_O	V_T	$V_T - V_O$	$(V_T - V_O)^2$	$(V_T - V_O)^2/V_T$
18	19,72	1,72	2,96	0,15
11	9,28	1,72	2,96	0,32
16	14,28	1,72	2,96	0,21
5	6,72	1,72	2,96	0,44

$$\chi^2_{observé} = 0,15 + 0,32 + 0,21 + 0,44 = \mathbf{1,12}$$

7) Signification de cette valeur.

$$\chi^2_{observé} < \chi^2_{seuil\ 0,05} \text{ puisque } 1,12 < 3,84$$

En fait: $\chi^2_{seuil\ 0,30} < \chi^2_{observé} < \chi^2_{seuil\ 0,20}$

8) Décision.

Le χ^2 est trop faible pour que l'on rejette l'hypothèse nulle:

les réponses des élèves dans les conditions C1 et C2 sont indépendantes.

Comme on aurait pu s'attendre à ce que les réponses soient les mêmes, puisque seules les conditions variaient d'une épreuve à l'autre, on peut penser que les changements de conditions ont eu une certaine influence sur les élèves qui "désorientés" n'ont pas reconnu les mêmes questions.

9) Exercices:

exercice 1

Disposer les valeurs suivantes dans un tableau :

n(1,1) = 92 ; n(1,0) = 28 ; n(0,1) = 41 ; n(0,0) = 39.

puis éprouver les hypothèses suivantes:

H1 : Les nombres de 1 dans V1 et dans V2 sont les mêmes,

H2 : Les variables V1 et V2 sont indépendantes

Réponse: $\chi^2(H2) = 12,8$.

exercice 2

Deux expérimentateurs ont obtenus sur des échantillons de tailles différentes les résultats suivants:

Expérimentateur 1. n(1,1) = 18 ; n(1,0) = 11 ; n(0,1) = 16 ; n(0,0) = 5

Expérimentateur 2. n(1,1) = 72 ; n(1,0) = 44 ;n(0,1) = 64 ; n(0,0) = 20

1. Vérifier que les résultats des deux expérimentateurs sont proportionnels
2. Quels sont dans les deux cas les résultats du test du χ^2
3. Conclusion?
4. Démontrer que si dans deux tableaux de contingence de mêmes dimensions sont tels que
A = {n(i,j)} et A' = {n'(i,j)},

on a n'(i,j) = k. n(i,j)

alors, les tests d'homogénéité produiront des valeurs du χ^2 telles que:

$$\chi^2(A') = k.\ \chi^2(A)$$

5. En déduire qu'il faut

- soit n'appliquer le test du χ^2 qu'à des tableaux d'effectifs réels et non à des tableaux de pourcentages ou à des effectifs proportionnels.
- soit appliquer aux valeurs trouvées, une correction que l'on précisera.

exercice 3.

Il ne faut pas croire que la présentation dans un même tableau assure que les formules à utiliser seront les mêmes.

1. Montrez par exemple que les 2 tableaux 2 x 2 ci-dessous se rapportent à des hypothèses et à des formules différentes:

	non V	V
échantillon	a	b
Modèle	n1	n2

	non V	V
1er échantillon	a	b
2nd échantillon	c	d

2. Montrez que, par contre, le tableau de droite peut s'interpréter comme un échantillon à deux variables binaires, et que le test d'égalité des proportions est aussi le test d'indépendance.

Réponses.

H0: l'échantillon est extrait d'un parent conforme au modèle.

$$\chi^2 = \frac{(a - n_1)^2}{n_1} + \frac{(b - n_2)^2}{n_2}$$

Le tableau présente la contingence (1ère ligne) et son modèle.

a + b = n

et n1 + n1 = n

H0: La proportion (ou la fréquence) des V est la même dans les deux échantillons (ou ne dépend pas de l'échantillon)

$a+c = v_1$; $b+d = v_2$; $n_1+n_2 = n$; $a+b = n_1 \neq v_1$

Le tableau ne présente que la contingence, le modèle des valeurs théoriques est à construire et comprendra 4 cases telles que

$T_1 = (n_1 . V_1)/n$ etc

Le χ^2 comprendra 4 termes

$$\chi^2 = = \frac{(a - T_1)^2}{T_1} + \frac{(b - T_2)^2}{T_2} + \frac{(c - T_3)^2}{T_3} + \frac{(d - T_4)^2}{T_4}$$

FICHE 7

UN échantillon, DEUX variables à DEUX valeurs, implication, test de GRAS.

Ce test indique dans quelle mesure une propriété en entraîne une autre. On dit qu'un caractère A implique un caractère B si chaque fois que A est vrai, B l'est aussi. C'est-à-dire si l'ensemble des sujets qui possèdent le caractère A est contenu dans l'ensemble de ceux qui possèdent le caractère B. Ce test est donc utilisable dans le cas où l'on observe deux variables binaires sur un même échantillon.

Exemple: (1)

Un enseignant se demande si, pour réussir à l'exercice A, ses élèves doivent avoir réussi l'exercice B. Ses résultats sont les suivants:

10 élèves ont réussi A et B
1 élève a réussi A et a échoué à B
6 élèves ont échoué à A mais ont réussi à B
8 élèves ont échoué à A et à B.

Exemple (2)

Pour une même affection, un biologiste envisage deux tests de dépistage: A et B. Il se demande si les deux tests sont équivalents ou si le résultat à l'un permet de prévoir le résultat à l'autre.

1) RECUEIL des DONNEES.

Le recueil est fait dans un tableau du genre suivant:

sujets	e_1	e_2	e_3	e_4	...	e_i	...	e_n
propriété A	Oui	Non	Oui	Non	...	Oui	...	Non
propriété B	Oui	Oui	Oui	Non	...	Non	...	Non

2) DISTRIBUTION DES EFFECTIFS

Comme dans la leçon 6 les données recueillies ci-dessus sont regroupées dans un tableau à double entrée: la table de contingence.

propriété A propriété B \	non	oui	TOTAL
oui	n(~A,B)	n(A,B)	b
non	n(~A,~B)	n(A,~B)	n-b
TOTAL	n-a	a	n

Dans ce tableau "~A" symbolise "non A", "la propriété A n'est pas réalisée".

n est le nombre total d'individus (ou d'observations), a le nombre de ceux qui satisfont la propriété A, b le nombre de ceux qui satisfont B, n - b le nombre de ceux qui ne satisfont pas B,

n(~A,B) est donc le nombre de sujets à la fois ~A et B: ils n'ont pas la propriété A, ils ont la propriété B. n(~A,~B) est le nombre de sujets ~A et ~B, etc.

Dans notre exemple 1. le tableau est:

effectifs	Echec A	Réuss A	Σ
réussites à B	6	10	16
échecs à B	8	1	9
effectif total	14	11	25

3) CHOIX DE L'IMPLICATION-MODELE

Il y a lieu d'appliquer le test lorsque l'effectif de l'une des quatre cases du tableau parait particulièrement faible par rapport aux autres, soit ef_{obs} cet effectif.

Si c'est n(A, ~B) qui est faible,.c'est-à-dire si ef_{obs} = n(A et ~B), le modèle auquel il faudra comparer la contingence sera alors A ⇒ B.

A ⇒ B se lit "A implique B", ou "tous les sujets qui sont A, sont aussi B", ou B contient A etc.

Le tableau qui correspond à l'hypothèse A ⇒ B est:

propriété A propriété B \	non	oui	TOTAL
oui	n(~A,B)	n(A,B)	b
non	n(~A,~B)	0	n-b
TOTAL	n-a	a	n

Dans l'exemple 1. la case faible est: ef_{obs} = n(A et ~B) = 1

1 n'est pas très différent de 0, le tableau de contingence ressemble donc beaucoup au modèle logique et l'implication attendue serait: A ⇒ B .

Diverses formulations de l'implication

Remarquons que ce même tableau correspond aussi à l'hypothèse

~B => ~A, c'est-à-dire si non A alors non B (~B => ~A est la contraposée de A=>B).

Il y a ainsi deux façons équivalentes de formuler l'hypothèse correspondant à un même tableau.

Exercice: L'effectif faible peut occuper l'une quelconque des quatre cases. Ecrire dans chaque cas les deux dénominations logiques de l'hypothèse correspondante.

Réponse:

A ==> B ,	(ou non B ⇒ non A)	si c'est n(A , ~B) qui est faible.
A ==> non B,	(ou B => non A)	si c'est n(A , B) qui est faible.
non A => B,	(ou non B => A)	si c'est n(~A , ~B) qui est faible.
non A => non B	(ou B => A)	si c'est n(~A , B) qui est faible.

Remarquons que rien n'empêche de trouver que deux cases (ou même trois) ont des effectifs faibles.

Cela se produit lorsque a est faible, (ou b, ou n-a, ou n-b). Les deux cases faibles se touchent alors par un de leurs côtés. Par exemple si on obtient n-a faible, c'est-à-dire n(A, ~B) et n(~A , ~B)

faibles. les deux cases de ~B sont faibles: on a à la fois A => B et ~A => B. Tous les sujets possèdent la propriété B.

Mais il peut se produire que les deux cases faibles soient disposées en diagonale. Il y a alors équivalence entre certaines valeurs des deux variables

Equivalence entre deux variables

- Si à la fois A ⇒ B et B ⇒ A, on écrit en abrégé A ⇔ B, ce qui se lit "A est équivalent à B".
- Si A ⇒ nonB et nonB ⇒ A, on écrit A ⇔ nonB. (on dit parfois alors que les variables sont contre-équivalentes)

Exercice:
dessiner les tableaux correspondants.

Réponse

0	a
a	0

n-a	0
0	n-a

4) COMPARAISON DU TABLEAU DE CONTINGENCE AVEC LE MODELE

Le tableau de contingence se trouve généralement, comme le montre la figure ci après, entre deux modèles: celui de l'**indépendance** que nous connaissons déjà et celui de l'**implication**. Pour montrer qu'il est près du second, on montre qu'il est loin du premier et pour cela que l'effectif de la case faible a très peu de chances d'être aussi petit sous l'hypothèse d'indépendance.

Le modèle de l'homogénéité, que nous avons utilisé dans les fiches précédentes ne présente pas de case faible. Il exprime que la proportion des A est la même parmi les B que parmi l'ensemble de la population.

Puisqu'il s'agit de comparer deux distributions (même si elles sont croisées croisées), celle observée et celle du modèle, nous pourrions utiliser un test du χ^2.

Ce ne sera possible que si toutes les valeurs théoriques sont supérieures à 5. Celle qui devrait être la plus faible serait ef_T, celle qui correspond à la case présumée faible dans la table de contingence. Sous l'hypothèse nulle, le modèle étant homogène, le calcul des valeurs théoriques s'effectue comme nous l'avons déjà vu. En supposant toujours que la case faible est celle correspondant à A et nonB son effectif sera:

$$ef_T = a.\frac{(n-b)}{n}$$

la proportion des élèves de la classe qui échouent à B est $\frac{n-B}{n}$. Le calcul de ef_T exprime que cette proportion est la même parmi les élèves qui ont réussi A.

Dans notre exemple

$$ef_{obs} = 1 \qquad \text{et} \quad ef_T = 11 \times 9 / 5 = 99 / 25 = 3,96.$$

Cette valeur ne permet donc pas d'utiliser le test du χ^2

Dans le cas ou deux cases en diagonale apparaissent faibles, le modèle de la double implication (équivalence) sera éprouvé par l'application successive du test aux deux modèles:

par exemple d'abord à :	A ⇒ non B	(1er test)
puis à sa réciproque	non B ⇒ A	(2 ème test)

Dans le cas de notre exemple, nous avons donc:

	modèle "indépendance"			contingence			modèle implication		
	~A	A	Σ	~A	A	Σ	~A	A	Σ
B	8,96	7,04	16	6	10	16	5	11	16
~B	5,04	3,96	9	8	1	9	9	0	9
Σ	14	11	25	14	11	25	14	11	25

Remarquons que les valeurs marginales - les sommes - sont les mêmes dans les deux tableaux calculés que dans le tableau de la contingence. Cette conditions est indispensable pour effectuer des comparaisons.

5) HYPOTHESE NULLE

L'hypothèse H étudiée: A ⇒ B, est une implication, mais l'hypothèse nulle est la négation de toute implication:

H_0: Les propriétés A et B sont indépendantes.

Nous venons de voir que la comparaison pourrait se faire à l'aide du χ^2 dans le cas où la taille de l'échantillon permettrait à l'effectif théorique de la case faible d'être supérieure à 5.

L'exemple introductif est un cas de "petit échantillon".

Pourtant, ef_T est presque quatre fois plus grand que l'effectif observé ef_{obs}.

Cette différence est elle assez grande pour que l'on puisse rejeter l'hypothèse nulle? Dans ce cas on préfèrera dire qu'il y a implication.

Pour pouvoir conclure il faudrait placer cette valeur par rapport à une nouvelle distribution théorique. Nous commençons par l'étude du cas d'un petit échantillon dont nous déduirons aussi une nouvelle méthode pour les grands échantillons.

Ainsi, selon que l'échantillon peut être déclaré "grand" ou "petit" les calculs seront différents.

6) CAS D'UN PETIT ECHANTILLON: n < 30 ou n = 30

Nous allons concentrer notre attention sur la case faible. Il s'agit donc de savoir s'il est fréquent ou rare de n'obtenir qu'une seule observation là où on en attendait (près de) quatre dans les conditions de l'expérience.

Sur 25 élèves, on en attendait | 3,96 | élèves A, ~B, On en a trouvé | 1 |

Supposons que chaque élève ait une probabilité p d'appartenir à cette case, que les réponses des élèves soient indépendantes les unes des autres, et que la classe comprenne n élèves, alors, le nombre des élèves S_{obs} d'une telle classe qui appartiennent à cette case, suit une loi Binômiale de **moyenne n.p** et d'**écart type** $\sqrt{\mathbf{n.p.(1 - p)}}$ (voir l'explication fiche 24):

$$S_{obs} = \sum_{k=0}^{ef_{obs}} \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

Une table de la loi binomiale donnant les valeurs cumulées de ef_{obs} pour des valeurs de p échelonnées de 0,05 en 0,05, et pour les valeurs de n de 1 à 30, exigerait 20 pages. Il vaut mieux reprendre le calcul direct.

RARETE DE LA VALEUR ef_{obs}.

La probabilité que moins de ef_{obs} élèves apparaissent dans cette case, au cours de n épreuves répétées, est la probabilité que, en répétant n = 25 fois une expérience où un événement de probabilité 0,16 peut apparaître, cet événement ne se produise qu'une fois au plus au cours de ces 25 expériences. C'est la probabilité qu'il y ait 0 ou 1 élèves:

$p_{(0\ élève)} = (1\text{-}p)^{25} = (0{,}84)^{25} = 0{,}0127$

$p_{(1\ élève)} = 25 \times p \times (1\text{-}p)^{24} = 0{,}0609$

$P1 = p_{(0\ élève)} + p_{(1\ élève)} = 0{,}0736$

Remarquons que

$ef_T = 3{,}96$, $P = \frac{ef_T}{n} = 0{,}1584$ et que $\sqrt{n.p.(1 - p)} = \sqrt{3{,}96.(1 - 0{,}1584)} = \sqrt{3{,}332} = 1{,}825$

La valeur observée s'écarte de la moyenne de près de deux fois l'écart type.

DECISION.

Si cette probabilité était inférieure à 5% (0,05) il conviendrait de préférer l'hypothèse de l'implication A ⇒ B et de rejeter l'hypothèse d'indépendance.

Or dans l'exemple 1, la probabilité d'avoir par hasard, 0 ou 1 élève ayant réalisé A et non B, ces variables étant supposées indépendantes, est supérieure à 7%. On ne peut donc pas rejeter l'hypothèse nulle au seuil de 5% et affirmer qu'il y a implication.

L'hypothèse A ⇒ B n'est pas acceptée, parce que l'hypothèse nulle ne peut pas être rejetée.

7) CAS D'UN GRAND ECHANTILLON: n > 30

EXEMPLE:

Dans deux classes rassemblant 48 élèves le même professeur a obtenu:

à deux exercices les résultats suivants, présentés comme ci-dessus.

	observation		
effectifs	Echec A	Réuss A	Σ
réussites à B	4	21	25
échecs à B	12	11	23
effectif total	16	32	48

Peut on conclure que nonA => non B (ou que B => A)?

CHOIX DE MODELES.

Nous avons signalé que nous pourrions recourir au χ^2 si efT > 5. C(est bien le cas ici car

efT = 16 x 25 / 48 = 8,33

Mais nous pouvons aussi utiliser un modèle plus puissant dérivé du modèle d'urne utilisé ci dessus dans le cas des petits échantillons.

On montre que, si le nombre d' expériences répétées, c'est-à-dire la taille de l'échantillon, croît et tend vers l'infini, la distribution binômiale correspondante de la fréquence observée tend vers une loi de Gauss. (cf fiche 24).

On peut confondre les deux lois dès que n > 30.

Ainsi, sous l'hypothèse nulle, la quantité $f_{obs} = \frac{ef_{obs}}{n}$ suit alors une loi normale

de moyenne $p = \frac{ef_T}{n}$

et d'écart type:

$$\sigma = \sqrt{\frac{1}{n} \cdot \frac{ef_T}{n} \cdot (1 - \frac{ef_T}{n})} = \sqrt{\frac{p.(1-P)}{n}}$$

Dans ces conditions le **coefficient d'implication** défini par:

$$Q_0 = \frac{p - f_{obs}}{\sigma}$$ et plus explicitement par:

$$\boxed{Q_0 = \frac{p - f_{obs}}{\sqrt{\frac{p.(1-P)}{n}}}}$$

suit une loi normale centrée réduite (de moyenne 0 et d'écart type 1, celle donnée par la table).

Remarque,
La formule ci dessous (démontrée dans la fiche 24) permet un calcul plus rapide:

$$Q_0 = \frac{\text{effectif observé} - \mu}{\sigma_\varepsilon} \quad \text{avec } ef_T = \mu \text{ et } \sigma_\varepsilon = \sqrt{\mu \times (1 - \frac{\mu}{n})}$$

APPLICATION:

Dans l'exemple 2. l'effectif de l'échantillon est: n = 48 $nf_{obs} = 4$ $f_T = (25 \times 16) / 48 = 8{,}33$

$$\text{Variance} = \frac{p.(1-P)}{n} = \frac{f_T}{n} \cdot (1 - \frac{f_T}{n}) \cdot \frac{1}{n} = \frac{8{,}33}{48} \cdot (1 - \frac{8{,}33}{48}) \cdot \frac{1}{48} = 0{,}002989$$

$$\sigma = \sqrt{\frac{f_T}{n} \cdot (1 - \frac{f_T}{n}) \cdot \frac{1}{n}} = \sqrt{0{,}002989} = 0{,}05467$$

$$Q_0 = \frac{Eff_T - nf_{obs}}{\sigma} = \frac{8{,}33 - 4}{0{,}05467} = 1{,}6512$$

RARETE DE CETTE VALEUR Q_o.

Si l'hypothèse nulle était vraie, puisque le signe est connu le test est unilatéral, il y aurait:

- moins de 5% de cas où $Q_o > 1{,}65$
- moins de 2.5% de cas où $Q_o > 1{,}96$ et
- moins de 0.5% des cas où $Q_o > 2{,}58$.

Dans l'exemple 2, Q_0 est donc significatif (de justesse) à 5%, il ne l'est pas à 2,5%

DECISION.

Si Q_0 est plus grand que 1,65, c'est une valeur rare, qui n'apparaîtrait par hasard que dans moins de 5% des cas si l'hypothèse nulle était vraie. On préfère alors penser que nous n'avons pas eu la malchance de tomber sur un cas exceptionnel et donc que c'est l'hypothèse nulle qui est à rejeter. Dans ces conditions on préfèrera l'hypothèse de l'implication (sans pour autant l'avoir prouvée) et on conclura avec un certain risque que:

Dans l'exemple 2, l'hypothèse nulle peut être rejetée: **on accepte donc l'hypothèse que B $\Rightarrow$ A**

REMARQUE.

Eprouvons l'hypothèse de l'indépendance des deux variables (réussite à A et réussite à B) à l'aide du χ^2.

Il vient $\chi^2 = 7{,}053$.

Il est bien évident que si A $\Rightarrow$ B alors A et B ne sont pas indépendants. L'inverse n'est pas sûr évidemment. A et B pourraient ne pas être indépendants sans que l'un implique l'autre ou sa négation, ainsi que le montre l'exercice 1 ci après.

8) Exercices

Exercice 1.

Examinez la dépendance et l'indépendance entre deux tests de dépistages A et B sachant que l'on a obtenu les résultats suivants:

13 sujets ont A et B positifs

11 sujets ont A positif et B négatif

5 sujets ont A négatif mais ont B positif

21 sujets ont A négatif et B négatif.

Réponse:

$\chi^2 = 6{,}61$ et $Q_0 = 1{,}580$

Les tests ne sont pas indépendants mais il n'y a aucune implication de l'un vers l'autre.

Exercice 2.

Examinez les dépendances et l'indépendance entre deux tests de dépistage T1 et T2 sachant que l'on a obtenu les résultats suivants:

21 sujets ont T1 et T2 positifs

5 sujets ont T1 positif et T2 négatif

7 sujets ont T1 négatif mais ont T2 positif

20 sujets ont T1 négatif et T2 négatif.

Réponse .

En examinant A $\Rightarrow$ B on trouve $Q_0 = 2{,}36$ et $\chi^2 = 15{,}98$

En examinant B $\Rightarrow$ A on trouve $Q_0 = 2{,}249$

Au seuil de 0,05 (acceptable pour des travaux de psychologie ou d'éducation) on aurait donc A $\Leftrightarrow$ B . Pour un résultat médical ce seuil est très insuffisant.

Exercice 3.

Appliquer au premier exemple de la leçon la méthode décrite pour les grands échantillons.

Réponse

L'application du calcul valable pour n > 30 donnerait la même conclusion: $Q_0 = 1{,}621$

Mais la valeur obtenue est beaucoup plus proche du seuil. La comparaison à la loi binômiale assure une meilleure sécurité.

Table des probabilités associées aux valeurs supérieures aux valeurs extrêmes observées de z dans la **Distribution Normale**.

Le corps de la table donne la probabilité unilatérale sous H_0 de z. La colonne de gauche donne les différentes valeurs de z à la première décimale. La première rangée donne les différentes valeurs de la deuxième décimale.

Ainsi, par exemple, la probabilité p de $z \leq 0.11$ ou $z > -0.11$ est $p = 0.4562$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3321	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2377	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2339	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0963	.0951	.0934	.0918	.0901	.0885	.0369	.0853	.0833	.0823
1.4	.0803	.0793	.0778	.0764	.0749	.0735	.0721	.0703	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007									
3.3	.0005									
3.4	.0003									
3.5	.00023									
3.6	.00016									
3.7	.00011									
3.8	.00007									
3.9	.00005									
4.0	.00003									

UN échantillon, DEUX variables nominales à l et m valeurs, indépendance, dépendances

Il s'agit de savoir si deux variables à plus de deux valeurs sont indépendantes.

1) Recueil des données.

Pour chacun des éléments de l'échantillon, on relève les valeurs des 2 variables:
- V, à l valeur: v_1, v_2,...,v_l,
- et U, à m valeurs: u_1, u_2,...,u_m.

éléments de l'échantillon	a	b	c	d	e	...
valeurs de la variable V	V_a	V_b	V_c	V_d	V_e	...
valeurs de la variable U	U_a	U_b	U_c	U_d	U_e	...

2) Distribution croisée: table de contingence.

L'ensemble de toutes les réponses possibles d'un élément de l'échantillon est formé de tous les couples (v_i,u_i). i varie de 1 à l et j varie de 1 à m.

Le nombre de réponses observées réalisant le couple (i,j) sera noté $ef_{obs}(i,j)$. L'ensemble de ces effectifs observés peut être présenté, comme nous l'avons fait dans les deux fiches précédentes, dans un tableau de contingence dit "croisé". Il comportera maintenant l colonnes et m lignes.

distribution croisée des effectifs observés

U\V	V_1	V_2	V_i	V_l	total
u_1	$ef_{obs}(u_1,v_1)$	$f_{obs}(u_1,v_2)$	$ef_{obs}(u_1,v_i)$	$ef_{obs}(u_1,v_l)$	nu_1
u_2	$ef_{obs}(u_2,v_1)$	$ef_{obs}(u_2,v_2)$	$ef_{obs}(u_2,v_i)$	$ef_{obs}(u_2,v_l)$	nu_2
...					
u_j	$ef_{obs}(u_j,v_1)$	$ef_{obs}(u_j,v_2)$	$ef_{obs}(u_j,v_i)$	$ef_{obs}(u_j,v_l)$	nu_j
...					
u_m	$ef_{obs}(u_m,v_1)$	$ef_{obs}(u_m,v_2)$	$ef_{obs}(u_m,v_i)$	$ef_{obs}(u_m,v_l)$	n_m
total	nv_1	nv_2	nv_i	nv_l	N

Exemple:

On s'intéresse aux interventions qui ont été nécessaires pour le bon fonctionnement d'ateliers de mathématiques. L'échantillon est un ensemble d'interventions du professeur (en plusieurs séances).

Les variables étudiées sont

-d'une part à la variable v = A: l'atelier où l'on observe l'intervention.

$v_1 = A_1$: constructions géométriques, $v_2 = A_2$: statistiques, $v_3 = A_3$: mesures de poids,

$u_4 = A_4$: théorèmes d'arithmétiques)

-d'autre part u = I le type d'interventions du professeur: $u_1 = I_1$ apport d'informations,

$u_2 = I_2$: encouragements sans information, $u_3 = I_3$: aide à l'organisation de la tâche.

La distribution des interventions est donnée ci-dessous.

	A1	A2	A3	A4	Total
I1	7	13	11	22	53
I2	9	6	23	9	47
13	15	11	31	26	83
Total	31	30	65	57	183

Q: Les types d'interventions dépendent-ils des ateliers ou en sont-ils indépendants?

Autrement dit, est ce que tous les ateliers sollicitent de la part du professeur le même type d'interventions, ou certains nécessitent plus d'informations tandis que d'autres requièrent plus d'encouragements

Cette question porte sur l'interaction des deux variables. Elle doit être bien distinguée de questions sur les marges, telles que:

Q1: Le nombre d'interventions (tous types confondus) dépend-il du type d'atelier?

Q2: Y a-t-il des types d'interventions plus fréquents que d'autres (indépendamment des ateliers.

Ces deux dernières questions relèvent du test du χ^2 simple, exposé dans la fiche 2.

3) Hypothèse nulle.

Etudions les hypothèses attachées à la question Q

H: La répartition des interventions entre les différents ateliers dépend du type d'intervention.

Pourquoi une hypothèse quelconque n'est pas toujours une bonne hypothèse nulle?

Il pourrait y avoir bien des façons de dépendre l'une de l'autre pour ces deux variables. Il ne servirait à rien d'en choisir une car:

- ou bien on ne pourrait pas repousser ce modèle, ce qui ne prouverait pas qu'il est le bon et qui laisserait non tranchée la question des autres modèles
- ou bien on pourrait le repousser, mais il faudrait encore examiner tous les autres modèles de dépendance.

Il est beaucoup plus efficace de choisir comme hypothèse le rejet de toute dépendance: il lui correspond un modèle bien déterminé, celui de l'indépendance. Il se présentera alors l'alternative suivante:

- si on peut rejeter cette hypothèse on aura établi (à un risque près) qu'il y a dépendance et il sera temps de se demander laquelle
- si on ne peut pas la rejeter, on n'aura pas pour autant prouvél'indépendance, mais il ne sera pas question alors d'accepter aucun modèle de dépendance.

H_0 sera donc: "les variables type d'intervention" et "genre d'atelier" sont indépendantes.

4) Modèle: l'indépendance

Dans ce cas le nombre d'intervention d'un certain type dans un certain atelier est proportionnel au nombre total d'interventions de ce type et au nombre total d'interventions dans cet atelier. etc. Alors chaque case (i,j) devrait présente une valeur théorique qui est calculée suivant la formule:

$$ef_T(i,j) = \frac{nv_i \times nu_j}{N}$$

Dans l'exemple ci dessus, les valeurs théoriques sont par conséquent:

	A1	A2	A3	A4	Total
I1	8,98	8,69	18,82	16,50	53
I2	7,96	7,7	16,69	14,63	47
I3	14,06	13,6	29,48	25,85	83
total	31	30	65	57	183

5) Distance du modèle à la table de contingence

Il s'agit de dire jusqu'à quel point le tableau de contingence ressemble ou contraire s'éloigne du tableau des effectifs théoriques sous l'hypothèse d'indépendance. Pour cela on mesure leur distance. Cette distance est toujours la distance du χ^2:

$$\chi^2 \text{ observé} = \sum_{i=0}^{l} \sum_{j=0}^{m} \frac{[ef_{observé}(v_i,u_j) - ef_T(v_i,u_j)]^2}{ef_T(v_i, u_j)}$$

Cette formule exprime que l'on fait simplement la somme, pour toutes les cases, des distances du χ^2 entre les effectifs observés et les effectifs théoriques correspondants. La distance $\chi^2_{observé}$ que donne cette formule suit une loi de χ^2 avec **dl = (l -1)(m-1)**

Dans l'exemple de notre exemple les valeurs théoriques sont:

$$\chi^2_{observé} = \frac{(7 - 8,98)^2}{8,98} + \frac{(13 - 8,69)^2}{8,69} + \frac{(11 - 18,82)^2}{18,82} + \frac{(22 - 16,50)^2}{16,50}$$
$$+ \frac{(9 - 7,96)^2}{7,96} + \frac{(6 - 7,70)^2}{7,70} + \frac{(23 - 16,69)^2}{16,69} + \frac{(9 - 14,63)^2}{14,63}$$
$$+ \frac{(15 - 14,06)^2}{14,06} + \frac{(11 - 13,60)^2}{13,60} + \frac{(31 - 29,48)^2}{29,48} + \frac{(26 - 25,85)^2}{25,85}$$

$$\chi^2_{observé} = 0,44 + 2,14 + 3,25 + 1,83 + 0,13 + 0,38 + 2,38 + 2,17 + 0,06 + 0,5 + 0,08 + 0,00$$

$\chi^2_{observé} = 13,36$

dl = 3 x 2 = 6

Dans la sixième ligne de la table du χ^2 on constate que:

$\chi^2_{seuil\ 0,05} < \chi^2_{observé}$ puisque 12,59 < 13,36

6) Conclusions

a) Réponse à la question Q

On doit rejeter l'hypothèse nulle: les interventions du professeur ne sont pas indépendantes du type d'atelier. Pour préciser un peu cette dépendance, il faut examiner les contributions (les termes de la somme) des différentes cases, au $\chi^2_{observé}$ calculé ci dessus et le signe de la différence.

La contribution la plus forte 3,25 (avec l'effectif observé < effectif théorique) est celle de (A3, I1): les mesures de poids requièrent en proportion moins d'informations que les autres d'ateliers. Puis celle de A3 I2, 2,38 indique que ce même atelier reçoit plus d'encouragements

De même avec 2,17 l'atelier d'arithmétique reçoit moins que son compte d'encouragements sans informations. tandis que l'atelier de statistiques (2,14) a reçu plutôt des informations

Aucune case n'est suffisante et il serait vain de vouloir préciser d'avantage. Les ateliers n'ont pas reçu des informations de même nature en mêmes proportions.

Réponse à la question Q1

Les interventions se répartissent entre les ateliers de la façon suivante:

	A1	A2	A3	A4	Σ
Effect.observé	31	30	65	57	183
effect.théori.	45,75	45,75	45,75	45,75	183
différence	14,75	15,75	19,25	11,25	
dif²	217,5	248	370,5	126,6	
contribution	4,75	5,42	8,09	2,76	

d'où $\chi^2 = 21,02$ avec un dl = 3

$\chi^2_{seuil\ 0,05} < \chi^2_{observé}$ puisque $7,82 < 21,02$

Conclusion: les interventions sont significativement plus fréquentes dans certains ateliers que dans d'autres.

Réponse à la question Q2

Les interventions se répartissent de la façon suivante suivant les types:

	I1	I2	I3	Σ
Effect.observé	53	47	83	183
effect.théori.	61	61	61	183
différence	8	14	22	
dif²	64	196	484	
contribution	1,04	3,21	7,93	

d'où $\chi^2 = 12,18$ avec un dl = 2

$\chi^2_{seuil\ 0,05} < \chi^2_{observé}$ puisque $5,99 < 12,18$

Conclusion: les interventions des différents types n'apparaissent pas avec la même fréquence.

7) Conclusion et exercice

L'interaction des deux variables ne résulte pas du fait que les marges n'ont pas une distribution uniforme. Ce n'est pas parce que les ateliers reçoivent des quantités d'interventions différentes, (Q1) ni parce que les différents types d'interventions n'apparaissent pas avec la même fréquence, que l'un influe sur l'autre. Montrez le en construisant sous forme de tableaux, un contre exemple de dimension 2 x 2, et un contre exemple de dimension 3 x 4.

B. VARIABLES NOMINALES

III. SUJETS DE DIDACTIQUE DES MATHEMATIQUES

Enoncés

Devoir n° 1 (Variante A)

Le tableau ci joint (fichier SEDUIR) présente des observations relatives à cinquante élèves résolvant deux problèmes semblables à quelques jours d'intervalles;
problème 1: colonnes 1 à 5, problème 2: colonnes 6 à 10.

Les variables sont les mêmes dans les deux cas:
- méthode utilisée (1 ou 2),
- réussite ou échec (1 ou 0) et dans ce cas,
  - type d'erreur, 1 dans l'une des colonnes
    - "erre1",
    - "erre2",
    - "erre3".

1. Le nombre de réussites a-t-il significativement augmenté?

2. Le choix de la méthode dans le deuxième problème est-il indépendant du choix dans le premier? testez si c'est possible l'hypothèse convenable à ce sujet, à l'aide du test du Chi2.

3. Etudier l'hypothèse suivante par le test de GRAS: "Tout élève qui choisit la méthode 1 pour résoudre le problème 1 la garde pour le problème 2".

4. Formuler des hypothèses raisonnables (le plus grand nombre possible) que ce tableau permettrait d'examiner (si les effectifs étaient suffisants). Etudier celles qui peuvent l'être, compte tenu des effectifs relevés.

Devoir n° 1 (variante B)

Même sujet, mêmes questions sur le fichier SED1FR ci joint.

Maîtrise de Sciences de l'Education U. Bordeaux II. 1990

TITRE : COMPARAISON DE PROBLEMES
NOMBRE D'OBSERVATIONS : 50 NOMBRE DE VARIABLES : 10
FICHIER DE DONNEES : B:SEDU1R

	1 métho	2 réuss	1 errel	4 erre2	5 erre3	6 métho	7 réuss	8 errel	9 erre2	10 erre3
1	2.00	1.00	0.00	0.00	0.00	2.00	0.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
3	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
4	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00
5	2.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
6	2.00	0.00	1.00	0.00	0.00	2.00	0.00	1.00	0.00	0.00
7	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
8	1.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
9	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
10	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
11	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
12	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
13	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
14	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
15	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
16	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
17	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
18	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
19	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
20	2.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
21	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
22	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
23	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
24	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
25	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00
26	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
27	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
29	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00
30	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
31	2.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
31	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
33	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
34	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
35	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
36	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
37	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
38	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
39	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
40	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
41	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
42	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
43	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
44	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
45	2.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
46	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
47	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
48	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
49	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
50	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00

TITRE : COMPARAISON DE PROBLEMES
NOMBRE D'OBSERVATIONS : 50 NOMBRE DE VARIABLES : 10
FICHIER DE DONNEES : B:SED1FR

	1 métho	2 réuss	3 errel	4 erre2	5 errel	6 métho	7 réuss	8 errel	9 erre2	10 erre3
1	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00
2	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00
3	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
4	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
5	2.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
6	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
7	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
8	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
9	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
10	2.00	0.00	0.00	0.00	1.00	2.00	1.00	0.00	0.00	0.00
11	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
12	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
13	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
14	1.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
15	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
16	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00
17	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
18	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
19	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
20	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
21	1.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00
22	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
23	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
24	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
25	2.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
26	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
27	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
28	2.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
29	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
30	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
31	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
32	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
33	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
34	2.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
35	2.00	1.00	0.00	0.00	0.00	2.00	0.00	1.00	0.00	0.00
36	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
37	1.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
38	2.00	0.00	1.00	0.00	0.00	2.00	0.00	1.00	0.00	0.00
39	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
40	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
41	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
42	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
43	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
44	2.00	0.00	0.00	1.00	0.00	2.00	0.00	0.00	1.00	0.00
45	2.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
46	1.00	0.00	1.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
47	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
48	2.00	1.00	0.00	0.00	0.00	2.00	1.00	0.00	0.00	0.00
49	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
50	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00

CORRECTIONS

Devoir n°1 A

1. On considère que l'on a deux échantillons indépendants, la colonne 2 (Pb1) et la colonne 7 (Pb2), et une variable nominale à deux valeurs: Réussite, Echec. Le nombre des réussites au problème 1 est 28, pour 50 élèves et au problème 2 il est 36 pour 50 élèves. Il semble qu'il y ait eu progrès.

Le tableau observé est le suivant:

échantillon\ variable:	Echecs	réussites	sommes
problème 1	22	28	50
problème 2	14	36	50
sommes	36	64	100

Hypothèse nulle: les proportions de réussite sont égales pour les deux problèmes.

Calcul du modèle: L'effectif théorique de chaque case est le produit des marges divisé par l'effectif total (Si les proportions de réussites y sont identiques, la meilleure valeur pour la proportion théorique est celle obtenue sur les deux problèmes: 28+36=64 réussites pour 100 questions, il devrait donc y avoir 32 réussites pour chaque question et 18 échecs).

Le tableau des valeurs théoriques est alors:

échantillon\ variable:	Echecs	réussites	sommes
problème 1	18	32	50
problème 2	18	32	50
sommes	36	64	100

Distance du modèle à la contingence: valeur du Chi^2

$$Chi2 = \frac{(22-18)^2}{18} + \frac{(28-32)^2}{32} + \frac{(14-18)^2}{18} + \frac{(36-32)^2}{32}$$

Chi2= 2,78

Cette distance est elle grande? (i.e rare)

Dl = 1. Dans la table des valeurs de Chi2 apparaissant sous l'hypothèse nulle, on voit que plus de 5 % et moins de 10 % des valeurs de Chi2 seraient plus grandes (la valeur précise est voisine de 9,14 %. Donc 2,78 n'est pas une valeur assez grande (rare) pour que l'on rejette l'hypothèse nulle (il faudrait 3,84 au moins)

Conclusion: Les progrès ne sont pas suffisants pour que l'on puisse rejeter l'idée que la différence de réussite observée est due au hasard. On ne peut pas non plus affirmer qu'il n'y a pas de progrès. Nous dirons que le nombre des réussites n'a pas augmenté significativement.

2. Pour savoir si le choix de la méthode lors du deuxième problème dépend du choix fait lors du premier, il faut considérer que nous avons un échantillon d'élèves et deux variables nominales (pb1 et pb2) à deux valeurs (méthode 1 et méthode 2), ou encore deux échantillons appariés chacun avec une variable.

La table de contingence obtenue est alors

problème 2\ problème 1	Méthode 1	Méthode 2	sommes
Méthode 1	33	7	40
Méthode 2	1	9	10
sommes	34	16	50

hypothèse nulle: Les choix des méthodes sont indépendants dans le Pb1 et dans le Pb2.

Calcul du modèle: d'après H_0, la proportion des élèves qui choisissent la méthode 2 au Pb2 devrait être la même parmi les élèves qui ont choisi la méthode 1 au Pb1 (34) que parmi l'ensemble des 50 élèves (10). Cette proportion est 1/5. Il devrait donc y avoir effT = 1/5 x 34 = 6,8 élèves dans ce cas. Le tableau des valeurs théoriques est alors:

problème 2\ problème 1:	Méthode 1	Méthode 2	sommes
Méthode 1	27,2	12,8	40
Méthode 2	6,8	3,2	10
sommes	34	16	50

Distance du modèle à la contingence: Chi2 = 19,324. Cette valeur est très grande, mais une valeur théorique étant inférieure à 5, il n'est pas possible d'appliquer le test du Chi2, ni de conclure sur l'indépendance cherchée. Cependant les valeurs observées laissent espérer une équivalence.

3. Utilisons l'indice de Gras.

Choix de l'implication modèle. La valeur faible est 1, elle correspond à l'implication modèle M1,Pb1 $\Longrightarrow$ M1,Pb2. (Si un élève choisit la méthode 1 au premier problème, il choisit la méthode 1 au second).

Calcul de la valeur théorique (de la case faible):

effT = 6,8. Sous l'hypothèse d'indépendance, la probabilité de voir apparaître un élève choisissant M1 puis M2 serait alors de p = 6,80/50 = 0,136

Mesurons la rareté de la valeur 1 observée. Puisque

N = 50 > 30 il faut recourir à la distribution normale des fréquences (1ère méthode).

$$p = \frac{effT}{n} = \frac{6,80}{50} = 0,136 \qquad \sigma = \sqrt{p \,.\, (1 - p)\,\frac{1}{n}} = \sqrt{0,00235} = 0,048$$

$$Q = \frac{(\text{Valeur observée} - \text{eff}_T)/n}{\sigma} = \frac{(1 - 6,80)/50}{0,048} = -2,39$$

La table indique qu'il y a moins de 1,7 % de valeurs de Z qui peuvent être inférieures à -2,39. L'hypothèse que le choix des méthodes est indépendant dans les deux problèmes doit être rejetée.

En est-il de même pour l'autre implication: "Tout élève qui choisit M2 persiste dans son choix "?

effT = 12,8

La probabilité de voir un élève choisissant M2 puis M1 serait alors de $p = 12,8/50 = 0,256$

Suivant le calcul rapide (2ième méthode dans le cours):

$$\mu = \text{effT} = 12,8 \qquad \sigma_\varepsilon = \sqrt{\mu \times (1 - \frac{\mu}{n})} = \sqrt{12,8 \times (1 - \frac{12,8}{50})} = 3,08$$

$$Q = \frac{\text{Valeur observée} - \mu}{\sigma_\varepsilon} = \frac{7 - 12,8}{3,08} = -1,88$$

La table indique qu'il n'y a pas plus de 3% de valeurs de Z qui peuvent être inférieures à - 1,88. On ne peut pas accepter l'hypothèse nulle. Les élèves qui ont adopté la méthode 2 la gardent aussi, quoique moins fortement.

Conclusion: Les élèves ont une nette tendance à conserver la même méthode.

4. Propositions d'hypothèses à étudier sur ce tableau.

a) La méthode 1 est plus employée dans le Pb2 que dans le Pb1. (à étudier comme dans la question 1).

b) Les réussites aux Pb1 et aux Pb2 sont indépendantes (à étudier comme dans la question 2).

c) Le choix de la méthode n'influe pas sur la réussite (comme dans 2 mais l'échantillon comprend 100 observations).

d) L'échec favorise le changement de méthode.

e) Le changement de méthode ne favorise pas la réussite.

f) Les types d'erreurs n'ont pas évolué.

g) Les observations relatives à ces deux problèmes sont statistiquement homogènes.

h) Les élèves 1 à 25 ont des comportements homogènes avec ceux des élèves 26 à 50.

Tableau croisé des conjonctions 2 à 2: valeur 1,1

effectif des 1,1	R2	E21	E22	E23
R1	25	1	1	1
E11	6	5	0	1
E12	2	0	4	0
E13	3	0	0	1

CORRECTION du Devoir n°1 B

1. On considère que l'on a deux échantillons indépendants, la colonne 2 (Pb1) et la colonne 7 (Pb2), et une variable nominale à deux valeurs: Réussite, Echec. Le nombre des réussites au problème 1 est 28, pour 50 élèves et au problème 2 il est 36 pour 50 élèves. Il semble qu'il y ait eu progrès.

Le tableau observé est le suivant:

échantillon\ variable:	Echecs	réussites	sommes
problème 1	22	28	50
problème 2	14	36	50
sommes	36	64	100

Hypothèse nulle: les proportions de réussite sont égales pour les deux problèmes.

Calcul du modèle: si les proportions de réussites y sont identiques, la meilleure valeur pour la proportion théorique est celle obtenue sur les deux problèmes: 28+36=64 réussites pour 100 questions, il devrait donc y avoir 32 réussites pour chaque question et 18 échecs.

Le tableau des valeurs théoriques est alors:

échantillon\ variable:	Echecs	réussites	sommes
problème 1	18	32	50
problème 2	18	32	50
sommes	36	64	100

Distance du modèle à la contingence: Valeur du Chi2

$$\text{Chi2} = \frac{(22-18)^2}{18} + \frac{(28-32)^2}{32} + \frac{(14-18)^2}{18} + \frac{(36-32)^2}{32}$$

Chi2 = 2,78

Cette distance est elle grande? (i.e rare)

Dl = 1. Dans la table des valeurs de Chi2 apparaissant sous l'hypothèse nulle, on voit que plus de 5 % et moins de 10 % des valeurs de Chi2 seraient plus grandes (la valeur précise est voisine de 9,14 %. Donc 2,78 n'est pas une valeur assez grande (rare) pour que l'on rejette l'hypothèse nulle (il faudrait 3,84 au moins)

Conclusion: les progrès ne sont pas suffisants pour que l'on puisse rejeter l'idée que la différence de réussite observée est due au hasard. On ne peut pas non plus affirmer qu'il n'y a pas de progrès. Nous dirons que le nombre des réussites n'a pas augmenté significativement.

2. Pour savoir si le choix de la méthode lors du deuxième problème dépend du choix fait lors du premier, il faut considérer que nous avons un échantillon d'élèves et deux variables nominales (pb1 et pb2) à deux valeurs (méthode 1 et méthode 2), ou encore deux échantillons appariés chacun avec une variable.
La table de contingence obtenue est alors :

problème 2\ problème 1:	Méthode 1	Méthode 2	sommes
Méthode 1	29	7	36
Méthode 2	1	13	14
sommes	30	20	50

Hypothèse nulle: Les choix des méthodes sont indépendants dans le Pb1 et dans le Pb2.

Calcul du modèle: d'après H_0, la proportion des élèves qui choisissent la méthode 2 au Pb2 devrait être la même parmi les élèves qui ont choisi la méthode 1 au Pb1 (30) que parmi l'ensemble des 50 élèves (14). Cette proportion est 14/50. Il devrait donc y avoir:

effT = 14/50 x 30 = 8.40 élèves dans ce cas. Le tableau des valeurs théoriques est alors:

problème 2\ problème 1:	Méthode 1	Méthode 2	sommes
Méthode 1	21,60	14,40	36
Méthode 2	8,40	5,60	14
sommes	30	20	50

Distance du modèle à la contingence: Chi2 = 22,636. Cette valeur est très grande. Elle permet de conclure qu'il n'y a pas indépendance entre l'emploi des méthodes au cours du premier et du second problème. Les valeurs observées font penser que les élèves ont tendance à conserver la même méthode.

3. Utilisons l'indice de Gras.

Choix de l'implication modèle. La valeur faible est 1, elle correspond à l'implication modèle:

M1,Pb1 ==> M1,Pb2. (Si un élève choisit la méthode 1 au premier problème, il choisit la méthode 1 au second).

Calcul de la valeur théorique (de la case faible):

effT = 8,40. Sous l'hypothèse d'indépendance, la probabilité de voir apparaître un élève choisissant M1 puis M2 serait alors de p = 8,40/50 = 0,168

Mesurons la rareté de la valeur "1" observée. Puisque N = 50 > 30, on peut utiliser la distribution normale des effectifs (2ièmes méthodes).

μ = effT = 8,40 $\qquad \sigma_\varepsilon = \sqrt{\mu \text{ x } (1 - \frac{\mu}{n})} = \sqrt{8{,}40 \text{ x } (1 - \frac{8{,}40}{50})} = 2{,}64$

$$Q = \frac{\text{Valeur observée} - \mu}{\sigma_\varepsilon} = \frac{1 - 8{,}40}{2{,}64} = -2{,}803 \text{ <<- } 1{,}65$$

La distribution normale est symétrique par rapport à 0, il n'y a aucun inconvénient à considérer les valeurs positives (comme dans le cours, ou négatives comme ici).

La table indique qu'il y a moins de 2,5 pour mille valeurs de Z qui peuvent être inférieures à - 2,803.

Conclusion: l'hypothèse que le choix des méthodes est indépendant dans les deux problèmes doit être rejetée au profit de l'hypothèse: M1,Pb1 $\Longrightarrow$ M1,Pb2.

4. Propositions d'hypothèses à étudier sur ce tableau.

a) La méthode 1 est plus employée dans le Pb2 que dans le Pb1. (à étudier comme dans la question 1).

b) Les réussites aux Pb1 et aux Pb2 sont indépendantes (à étudier comme dans la question 2).

c) Le choix de la méthode n'influe pas sur la réussite (comme dans 2 mais l'échantillon comprend 100 observations).

d) L'échec favorise le changement de méthode.

e) Le changement de méthode ne favorise pas la réussite.

f) Les types d'erreurs n'ont pas évolué.

g) Les observations relatives à ces deux problèmes sont statistiquement homogènes.

h) Les élèves 1 à 25 ont des comportements homogènes avec ceux des élèves 26 à 50.

C. VARIABLES ORDINALES

I. VARIABLES ORDINALES: HOMOGENEITE POUR UNE VARIABLE

UN, DEUX ou k échantillons, UNE variable ordinale
Test de la Médiane

Une variable ordinale est une variable dont les valeurs sont des rangs, c'est-à-dire un ensemble totalement ordonné, avec peut-être des ex aequos. (Voir A. II. 2)

Le test de la médiane est utilisable pour établir dans quelle mesure deux groupes (ou plus de deux groupes) diffèrent par leurs tendances centrales. Il consiste à ramener l'étude d'une variable ordinale à celle d'une variable nominale à deux valeurs: valeurs supérieures à la médiane, valeurs inférieures. Celle ci se fera alors à l'aide d'un simple χ^2.

1) Recueil des données: cas de deux groupes

Les éléments des deux groupes sont réunis en un échantillon puis sont rangés suivant l'ordre de la variable ordinale. Ce tableau est la distribution sur les rangs.

Une **médiane** est un nombre qui ménage dans la population totale autant d'éléments plus petits que d'éléments plus grands qu'elle. Une médiane sera ici appelée "rang médian", même si ce n'est pas une valeur entière

2) Distribution réduite.

- Cas où la variable est un rang sans ex aequo.

On distribue alors chacune des populations ainsi obtenues: I les éléments dont le rang est inférieur à la médiane, S ceux de rang supérieur à la médiane, dans les deux groupes et on note les effectifs obtenus dans une table 2 x 2.

Exemple 1.
Dans une épreuve sportive les élèves de deux équipes: A et B, ont pu être rangés par ordre d'arrivée. Une équipe l'emporte-t-elle sur l'autre?

Rang	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
groupe:	A	B	B	A	A	A	B	A	B	B	A	B	B	A	B	B	B	B

Le rang médian est 9,5. La table est

effectifs	r < 9,5	r > 9,5
Groupe A	5	2
Groupe B	4	7

Ici l'effectif est trop faible pour permettre un test du χ^2 qui déterminerait si 4 sur 9 est significativement différent de 2 sur 9.

- Cas d'une variable ordinale avec de nombreux ex aequo

La variable ordinale utilisée peut ne pas être un rang et/ou il peut y avoir beaucoup d'ex aequo. On considère alors la distribution des effectifs sur les rangs ménagés par la variable ordinale.

Fréquemment cette variable est numérique mais on ne sait pas si la distribution des effectifs correspondante suit la loi "normale" ou non. On ne peut donc pas utiliser les tests "paramétriques" comme le "t" de Student qui présupposent cette condition. Alors seul l'ordre des données sera pris en considération et les différences entre valeurs perdent toute signification.

Exemple.

Un chercheur en didactique veut étudier le rôle du contexte scolaire dans la résolution des problèmes.

Pour cela il constitue deux échantillons de 25 et 27 élèves à qui il fait poser une série de 12 questions difficiles portant sur des sujets étudiés par les élèves la semaine précédente.

Dans le premier échantillon les questions sont posées par l'enseignant qui a fait la leçon, dans le second les questions sont posées par un enseignant qui n'a pas assisté à la leçon mais que les élèves connaissent bien. Le chercheur note pour chaque élève le nombre d'erreurs qu'il a faites, puis il relève les effectifs des élèves sur cette variable considérée comme ordinale. Il obtient alors le tableau suivant.

Résultats												
Nbre d'erreurs	0	1	2	3	4	5	6	7	8	9	10	
échantillon 1	5	10	4	2	4	0	1	0	1	0	0	27
échantillon 2	2	1	4	7	0	4	3	1	2	0	1	25
somme	7	11	8	9	4	4	4	1	3	0	1	52
somme cumulée	7	18	26	35	39	43	47	48	51	51	52	

Le rang médian est $R_m = \frac{N1 + N2}{2} = 52/2 = 26$

La valeur médiane est $M_d = 2$

La table réduite est alors:

effectifs	$n_e < 2$	$n_e > 2$	
Echantillon 1	C = 19	A = 8	27
Echantillon 2	D = 7	B = 18	25
Somme	26	26	52

Nous allons utiliser la formule simplifiée du calcul du χ^2. Pour cela nous désignons les cases du tableau par des majuscules A,B,C et D (avec A et D en diagonale ainsi que B et C). Ici A et B désignent les cases des rangs supérieurs au rang médian.

3) Hypothèse nulle:

H_0: La proportion des observations inférieures à la médiane est la même dans les deux groupes.

Il s'agit de savoir si les deux groupes peuvent être considérés comme extraits "au hasard" d'un même ensemble parent.

Dans l'exemple ci dessus, H0 se traduit par: on peut obtenir souvent 19 objets sur 27, puis 7 sur 25 lors de tirages avec remise dans un même ensemble parent.

4) Modèle

Exercice:

Dresser le tableau des valeurs théoriques, puis calculer le χ^2.selon la méthode (1) exposée dans la leçon 6

Réponse: $\chi 2 = \frac{(C - \frac{N1}{2})^2}{\frac{N1}{2}} + \frac{(A - \frac{N1}{2})^2}{\frac{N1}{2}} + \frac{(D - \frac{N2}{2})^2}{\frac{N2}{2}} + \frac{(B - \frac{N2}{2})^2}{\frac{N2}{2}}$ (1)

dans l'exemple étudié $\chi^2 = 9, 321$

5) Distance des valeurs observées au modèle: Calcul du χ^2

La formule réduite équivalente que nous avons signalée dans la fiche 6 s'applique, :

$$\chi^2 = \frac{N.(A.D - B.C)^2}{(A + B).(C + D).(A + C).(B + D)} \qquad (2)$$

Elle donne le même résultat que (1):
$\chi^2 = 9{,}321$,
exercice:
vérifiez le sur l'exemple proposé.

Si $20 < N < 50$ il est préférable d'appliquer la correction de YATES:

$$\chi^2 = \frac{N.([A.D - B.C] - N/2)^2}{(A + B).(C + D).(A + C).(B + D)} \qquad (3)$$

Le degré de liberté est 1.

Dans l'exemple étudié, la correction de Yates donnerait le résultat:

$$\chi^2 = \frac{52((7x8 - 19x18) - 26)^2}{27 \text{ x } 25 \text{ x } 26 \text{ x } 26}$$

$$\chi^2 = 7{,}70$$

Elle diminue la valeur du qui a tendance à croître excessivement lorsque certains dénominateurs deviennent trop faible, et elle augmente donc la sécurité.

6) Rareté de ce χ^2 et décision

$\chi^2_{\text{seuil } 0,01} = 6{,}64$ et $6{,}64 < \chi^2_{\text{observé}}$

donc l'hypothèse nulle peut être rejetée: La proportion des élèves qui ont un nombre d'erreurs au dessous de la médiane n'est pas le même dans les deux échantillons. Elle est significativement plus élevée dans le premier échantillon que dans le second.

7) Cas de k échantillons, une variable ordinale

On procède comme en ci-dessus:
Les éléments des k groupes sont réunis en un échantillon puis sont rangés suivant l'ordre de la variable ordinale. Ce tableau est la distribution sur les rangs.
Le rang médian est toujours celui qui ménage dans la population totale autant d'éléments plus petits que d'éléments plus grands.
On distribue alors chacune des populations ainsi obtenues: I les éléments dont le rang est inférieur à la médiane, S ceux de rang supérieur à la médiane, dans les k groupes et on note les effectifs obtenus dans une table k x 2.
Le calcul se poursuit comme dans la leçon 2. jusqu'au test du χ^2 avec dl = k-1

8) Exercices

1. On a présenté une épreuve de labyrinthe à des élèves de différents niveaux scolaires E1, E2, E3, E4. Après un entraînement de 10 minutes, on note le nombre d'erreurs dans la recherche de la sortie. La distribution réduite est la suivante.

Formulez l'hypothèse nulle et tirez des conclusions.

	E1	E2	E3	E4	∑
< Md	8	13	12	13	46
> Md	19	12	11	7	49

2. Montrer l'équivalence de la formule (2) avec la formule (1) du χ^2 dans le cas où l'effectif total est réparti également sur les deux valeurs de la variable.

UN échantillon, UNE variable ordinale
Test de KOLMOGOROV-SMIRNOV

Le test de Kolmogorov-Smirnov est un test d'ajustement. Il permet de comparer une distribution de scores à une distribution théorique uniforme, et de déterminer si les scores de l'échantillon peuvent être considérés comme extraits au hasard d'une distribution théorique déterminée, sur une variable ordinale.

Il généralise le t de Student en ce sens qu'il permet de comparer des variables numériques lorsqu'on ignore si la distribution de ces variables est normale ou lorsqu'on sait qu'elle ne l'est pas.

1) Recueil des données

A chaque sujet correspond son score, ou une valeur de la variable ordinale, ou le rang choisi ou obtenu (r(i) rang du sujet i):

Sujets	a	b	c	d	..i	..n
Score	r(a)	r(b)	r(c)	r(d)	..r(i)	..r(n)

2) Distribution réduite.

En regard de chaque score ou rang j se présente le nombre de sujets n(j) ou d'observations qui ont choisi ou obtenu ce rang, (distribution des effectifs):

Rangs	1	2	3	..j	..p
Nombre d'observations	n(1)	n(2)	n(3)	..n(j)	..n(p)

Exemple 1.
Un professeur veut savoir si ses élèves pensent (à priori) que le nombre des figures données pour illustrer une démonstration est un facteur qui favorise la compréhension d'un texte de mathématiques. Pour cela il leur propose 5 fiches présentant chacune une démonstration du même théorème mais qui diffèrent par le nombre de figures qui accompagnent cette démonstration. La fiche 1 ne comporte qu'un dessin, la fiche deux en comporte deux etc. Le nombre des figures se voit immédiatement alors que le texte ne peut pas être lu avant le choix. Les choix des élèves se portent ils de préférence sur les démonstrations les plus illustrées?

La distribution des effectifs est la suivante.

Variable	1	2	3	4	5
Distribution observée	0	3	1	11	10

3) Hypothèse nulle.

L'hypothèse nulle la plus fréquemment examinée suppose que la valeur de la variable ordinale n'influence pas le choix de chaque sujet, ou l'attribution de chaque observation appartenant à l'échantillon, de sorte que chaque sujet choisit l'une des valeurs au hasard.

H_0: La variable n'influence pas les attributions dans l'échantillon. Chaque valeur de la variable présente un effectif de sujets constant et égal à $\frac{n}{p}$, n effectif de l'échantillon, p nombre de valeurs de la variable.

4) Modèle.

Sous l'hypothèse nulle, chaque valeur (rang) de la variable aura un effectif théorique de:

$$n_T(j) = \frac{n}{p} = \frac{\text{effectif total de l'échantillon}}{\text{nombre de valeurs de la variable}}$$

On considèrera, non pas les effectifs, mais les fréquences:

$$f_o(j) = \frac{n(j)}{n} = \frac{\text{effectif observé pour la valeur i de l'échantillon}}{\text{nombre total d'observations}}$$

A chaque valeur, sera affectée une fréquence théorique constante de:

$$f_T(j) = \frac{n_T(j)}{n} = \frac{1}{p}$$

De plus on ne comparera pas les effectifs ni les fréquences, valeur par valeur, comme dans le cas du X^2, car cette méthode ne tiendrait pas compte de l'ordre des valeurs de la variable, on va donc s'intéresser aux effectifs cumulés.

Calculons les fréquences cumulées des observations:

$$S_n(j) = \sum_{k=1}^{j} \frac{n(k)}{n}$$

La fréquence cumulée théorique correspondante, jusqu'à la valeur j sera:

$$F_T(j) = \sum_{k=1}^{j} f_T(k)$$

Dans notre exemple, nous obtenons le tableau suivant:

Variable (valeurs)	1	2	3	4	5
Distribution observée n(j)	0	3	1	11	10
Cumul observé $S_n(j)$	$\frac{0}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{15}{25}$	$\frac{25}{25}$
Distribution théorique $f_T(j)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
Cumul théorique $F_T(j)$	$\frac{5}{25}$	$\frac{10}{25}$	$\frac{15}{25}$	$\frac{20}{25}$	$\frac{25}{25}$

5) Distance du modèle aux fréquences observées

On choisit la distance la plus simple: On considère D(j) les différences entre $S_n(j)$ - $F_T(j)$, puis la plus grande valeur (max) atteinte par la valeur absolue de ces différences sur toutes les valeurs j de la variable

$$D = max_j \in [1, p] \left| S_n(j) - F_T(j) \right|$$

Dans le cas de notre exemple on achève le tableau de la façon suivante:

D(j)	$\frac{5}{25}$	$\frac{7}{25}$	$\frac{11}{25}$	$\frac{5}{25}$	$\frac{0}{25}$

D'où D = 11/25 = 0,44

6) Rareté de cette distance.

La table de KOLMOGOROV-SMIRNOV donne la distribution de la fréquence de ces valeurs sous l'hypothèse nulle en fonction du niveau de signification choisi et de la taille de l'échantillon.

Dans le cas de notre exemple on lit sur la ligne n=25

	Niveau de signification de D				
Taille	0,20	0,15	0,10	0,05	0,01
25	0,21	0,22	0,24	0,27	0,32

On voit que $D > D_{\text{seuil } 0,01}$ puisque 0,44 > 0,32

La valeur observée pour D se rencontrerait moins d'une fois sur 100 dans les conditions de l'hypothèse nulle.

7) Décision.

Si $D > D_{\text{seuil } 0,05}$ l'hypothèse nulle doit être rejetée. Dans le cas contraire, elle ne peut pas être rejetée, mais on n'est pas obligé de l'accepter non plus comme certaine.

Dans le cas de notre exemple, les élèves pensent qu'une démonstration est d'autant plus compréhensible qu'elle est accompagnée de plus de figures.

8) Exercices

1. Quelle différence y a-t-il avec un ajustement de la distribution des effectifs sur les valeurs de la variable traité directement avec un χ^2 ? montrer comment l'ordre des valeur intervient dans la décision.
2. Le test de KOLMOGOROV-SMIRNOV permet de comparer la distribution observée à toute distribution théorique définie. Eprouver les hypothèses nulles suivantes:

a) Le nombre de choix des élèves croit proportionnellement au nombre de figures

b) Le nombre de choix des élèves est constant pour trois figures ou plus, nul pour moins de trois figure.

TABLE DES VALEURS CRITIQUES DE D DANS LE TEST
DE **KOLMOGOROV - SMIRNOV**
à un échantillon

Taille échantillon (N)	Niveau de signification de $D = \max \lvert Sn(j) - F_T(j) \rvert$				
	.20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.828
4	.494	.525	.564	.624	.733
5	.446	.474	.510	.565	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.410	.490
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.392
17	.250	.266	.286	.318	.381
18	.244	.259	.278	.309	.371
19	.237	.252	.272	.301	.363
20	.231	.246	.264	.294	356
25	.21	.22	.24	.27	32
30	.19	.20	.22	.24	.29
35	.18	.19	.21	.23	.27
Au delà de 35	$\frac{1,07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

Deux échantillons, une variable ordinale
Test de KOLMOGOROV-SMIRNOV

Le test de Kolmogorov-Smirnov permet de comparer deux échantillons entre eux, de la même façon qu'on a comparé un échantillon à un ensemble parent déterminé par la distribution théorique. Cependant les distributions de référence seront différentes.

1) Recueil des données

Dans chacun des deux échantillons les données sont recueillies comme dans 1. A chaque sujet correspond son score, ou une valeur de la variable ordinale, ou le rang choisi ou obtenu: r(i) rang du sujet i.

Sujets:	a	b	c	d	...i...	n
Score:	r(a)	r(b)	r(c)	r(d)	...r(i)...	r(n)

2) Distribution réduite.

En regard de chaque score ou rang j se présente, pour chaque échantillon, le nombre de sujets n(j) ou d'observations qui ont choisi ou obtenu ce rang, (distribution des effectifs):

Nombre d'observations:					
Rangs:	1	2	3	j...	p
échantillon1: effectif n_1	$n_1(1)$	$n_1(2)$	$n_1(3)$	$n_1(j)$	$n_1(p)$
échantillon2: effectif n_2	$n_2(1)$	$n_2(2)$	$n_2(3)$	$n_2(j)$	$n_2(p)$

Exemple.

Deux groupes d'élèves G1 et G2, ont appris à disposer leurs opérations différemment. Une première épreuve avait permis de constater que les deux groupes ne différaient pas significativement dans leurs résultats en calcul.

Le groupe G1 utilise la méthode habituelle, le groupe G2 utilise des procédés qui limitent les possibilités d'erreurs de position et la complexité des opérations mentales, mais au prix d'une augmentation du nombre d'opérations élémentaires matérielles. Pour savoir si les deux groupes diffèrent on propose à la fin de l'enseignement, à tous les élèves, une même longue série d'opérations.

Dans chaque groupe, les distributions du nombre des élèves $e_1(i)$ et $e_2(i)$ sur le nombre des erreurs N(E) qu'ils ont commises, sont les suivantes.

	N(E)	0	1	2	3	4	5	6	Σ
G1	$e_1(j)$	6	0	6	11	12	13	6	n1=54
G2	$e_2(j)$	1	11	10	10	6	5	2	n2=45

3) Hypothèse nulle.

Le choix des hypothèses nulles n'est pas limité à l'affirmation d'une égalité ou la négation d'une différence. Examinons diverses possibilités

Test Bilatéral, Test Unilatéral

Il y a deux types d'hypothèses nulles: les hypothèses unilatérales et les hypothèses bilatérales.

a) Hypothèses unilatérales. Elles expriment que l'un des échantillons présents des valeurs fréquemment plus élevées que les autres (ni égales ni plus faibles). On peut en formuler deux, symétriques.

Par exemple

- H_1:le groupe G1 fait plus d'erreurs que le groupe G2.
- H_1': Le groupe G2 fait plus d'erreurs que le groupe G1.

b) hypothèse bilatérale.

L'hypothèse nulle la plus fréquemment examinée suppose que les deux échantillons sont extraits d'une même population parente. Les deux échantillons choisissent leur rang suivant la même loi H_0: les deux échantillons se comportent de la même manière au regard de cette variable ordinale. Chaque valeur de la variable présente des fréquences égales dans les deux échantillons. Cette hypothèse est dite bilatérale car elle rejette à la fois les deux hypothèses unilatérales.

Suivant les types d'hypothèses (uni ou bi) et suivant l'effectif des échantillons:

soit $n1 = n2 < 40$, (petits échantillons)

soit $40 \leq n1 \leq n2$ (grands échantillon)

les modèles et les calculs sont différents.

4) Calculs communs.

On comparera toujours les fréquences observées (puisque dans le cas des grands effectifs n_1 et n_2 peuvent être différents)

$e_1(j)$ = effectif de l'échantillon 1 correspondant à la valeur ou au rang j.

$$f_{o1}(j) = \frac{e_1(j)}{n_1}$$

$e_2(j)$ = effectif de l'échantillon 2 correspondant à la valeur ou au rang j.

$$f_{o2}(j) = \frac{e_2(j)}{n_2}$$

et, comme dans le cas ci-dessus, on s'intéresse aux fréquences cumulées dans les deux groupes:

$$F_{o1}(j) = \sum_{i=1}^{j} f_{o1}(i)$$

$$F_{o2}(j) = \sum_{i=1}^{j} f_{o2}(i)$$

Elles peuvent se calculer en faisant la somme des fréquences comme ci-dessus, mais aussi en considérant d'abord les effectifs cumulés, méthode qui sera préférée dans le cas des petits échantillons:

$$E_1(j) = \sum_{i=1}^{j} e_1(i)$$

alors

$$F_{o1}(j) = \sum_{i=1}^{j} \frac{E_1(i)}{n_1}$$

et de même pour $E_2(j)$ et $F_{o2}(j)$.

Dans notre exemple, nous obtenons le tableau ci après.

Nombre d'erreurs		0	1	2	3	4	5	6	Σ
G1	$e_1(i)$	6	0	6	11	12	13	6	54
G2	$e_2(i)$	1	11	10	10	6	5	2	45
G1	$f_{o1}(j)$	0,11	0,00	0,11	0,20	0,22	0,24	0,11	1
G2	$f_{o2}(j)$	0,02	0,24	0,22	0,22	0,13	0,11	0,04	
G1	$F_{o1}(j)$	0,11	0,11	0,22	0,43	0,65	0,89	1	
G2	$F_{o2}(j)$	0,02	0,27	0,49	0,71	0,84	0,96	1	
DifG1-G2		0,09	-0,16	-0,27	-0,29	-0,20	-0,07	0	
|Dif|		0,09	0,16	0,27	0,29	0,20	0,07	0	

5) Distances du modèle aux fréquences observées

- BILATERAL. Dans le cas de l'hypothèse bilatérale on utilise la distance de la valeur absolue comme dans le cas de la fiche 10

On considère les différences:

$D(j) = F_{o1}(j) - F_{o2}(j)$, puis la plus grande valeur (max), atteinte par la valeur absolue de ces différences sur toutes les valeurs j.

$$\boxed{D_b = \max |F_{o1}(j) - F_{o2}(j)|}$$

Dans notre exemple $D_b = 0,29$

- UNILATERAL. Dans le cas de l'hypothèse unilatérale on utilise la formule:

$$\boxed{D_u = \max (F_{o1}(j) - F_{o2}(j))} \qquad (1)$$

On remarquera d'abord que l'hypothèse H_1':

"le groupe G1 est meilleur que le groupe G2"

doit se traduire par le fait que les petites valeurs de la variable auront tendance à porter des effectifs relatifs de G1 plus élevés que ceux de G2. En effet, la proportion d'élèves qui font moins de fautes est alors plus élevée dans le groupe G1 que dans le groupe G2. Pour j petit, la fréquence f1(j) tend à être supérieure à la fréquence f2(j), et par conséquent en calculant les effectifs cumulés des petites valeurs vers les grandes nous trouverons donc assez vite des différences:

$(F_{o1}(j) - F_{o2}(j))$ POSITIVES, et qui le resteront .

$$G1 > G2 \implies F_{o1}(j) > F_{o2}(j)$$

Au contraire les élèves de G2 faisant plus de fautes, la proportion de ces élèves qui se concentrent sur les grandes valeurs de la variable est plus élevée. La différence tendra à diminuer et s'annulera pour la valeur la plus grande. Elle pourrait s'inverser sans que cela ait d'importance.

Le maximum de $(F_{o1}(j) - F_{o2}(j))$ se produit pour une valeur j. Si ce maximum est positif, il montre un certain avantage du groupe G1 sur le groupe G2 sur les petits nombres d'erreurs. S'il est assez grand, il ne sera pas compensé par un éventuel retard sur les grands nombres d'erreurs.

Contre exemple: G2 est meilleur que G1 par un avantage sur les grands nombres de fautes.

	0	1	2	3	5	6
G1	12	0	0	0	1	11
G2	0	12	10	2	0	0

Remarquons que la formule (1) ne prend en considération que les valeurs des différences et non pas les valeurs absolues. Ainsi deux cas peuvent se présenter:

- soit $D_u = D_b$, lorsque la plus grande différence va dans le sens de l'hypothèse unilatérale envisagée, il reste alors à savoir si cet avantage est significatif;

- soit $D_u < D_b$ dans le cas où la meilleure différence dans le sens envisagé n'est pas la plus grande en valeur absolue. Mais dans ce cas le test sur D_u ne peut pas être positif sinon D_b le serait aussi et on aurait à la fois le groupe 1 strictement meilleur que le groupe 2 et l'inverse.

Mais, le fait de savoir que D_u n'est pas significativement grand, ne renseigne pas sur la significativité de D_b. Il faut donc refaire l'étude unilatérale opposée.

Au contraire si D_b est assez grand pour faire rejeter l'hypothèse bilatérale alors l'une des deux hypothèses unilatérales sera nécessairement acceptée, et il est assez facile en général de voir laquelle à l'oeil nu.

Dans notre exemple les valeurs unilatérales sont :

$D_{1<2} = 0{,}28$

et $D_{2<1} = 0{,}09$

6) Rareté de cette distance.

* CAS DES GRANDS ECHANTILLONS, TEST BILATERAL ($40 \leq n1 \leq n2$)

Calculer la valeur

$$KS_b = D_b \times \frac{\sqrt{n1 \times n2}}{\sqrt{(n1 + n2)}}$$

avec $D_b = \max |F_{o1}(j) - F_{o2}(j)|$

et la comparer à la table suivante:

Si KS_b est supérieur à KS_{seuil} on doit rejeter l'hypothèse nulle.

(Le seuil est le niveau de signification)

Seuils	0,10	0,05	0,025	0,01	0,005	0,001
KS_{seuil}	1,22	1,36	1,48	1,63	1,73	1,95

Notre exemple relève de ce cas là puisque: n1 = 54 et n2 = 45.

Le calcul de KS donne $KS_b = D_b \times 4,95 = 0,29 \times 4,95 = 1,41$.

Dans la table: $1,36 < 1,41 < 1,48$

Donc $KS_{seuil\ 0,05} < KS_b < KS_{seuil\ 0,025}$

L'hypothèse d'homogénéité peut être rejetée au seuil de 5 % mais pas au seuil de 1 %.

** CAS DES GRANDS ECHANTILLONS, TEST UNILATERAL.

Dans ce cas $(2.KS_b)^2$ est distribuée comme un χ^2 avec dl = 2.

Calculer la valeur

$$X_o^2 = 4 \times D_u^2 \times \frac{n_1 \times n_2}{n_1 + n_2}$$

avec bien sûr $D_u = \max (F_{o1}(j) - F_{o2}(j))$

La lecture de la table du χ^2 sur la deuxième ligne nous informe de la place qu'occupe la valeur χ^2_o par rapport aux valeurs seuils.

Décision: si $\chi^2_{seuil\ 0,05} < \chi^2_o$, L'hypothèse nulle doit être rejetée.

Dans notre exemple les valeurs unilatérales sont :

$D_{1<2} = 0,29$ et $D_{2<1} = 0,09$

Les valeurs correspondantes de χ^2_o sont

$\chi^2_{o1<2} = 0,78$

et $\chi^2_{o2<1} = 7,99$

A la lecture de la table du χ^2, ligne 2, puisque
$7,82 < 7,99 < 9,21$ il apparaît que:
$\chi^2_{seuil\ 0,025} < \chi^2_o < \chi^2_{seuil\ 0,020}$

l'hypothèse H_1': G1 > (meilleur que) G2 doit être rejetée au seuil de 2,5 % (elle ne peut l'être au seuil de 1 %).

On peut conclure alors que l'autre hypothèse est vraie:
G2 est significativement meilleur que G1

Il apparaît que la classe qui utilise la méthode fiable fait significativement moins d'erreurs. (Par conséquent les deux distributions sont différentes et les deux classes sont significativement différentes).

*** CAS DES PETITS ECHANTILLONS (N1 = N2 ≤ 40), TEST BILATERAL ET UNILATERAL.

Dans ce cas on compare directement D_b ou D_u à une table.

Comme n1 = n2 il est commode de simplifier le tableau d'analyse et de ne considérer que les effectifs: $e_1(j)$, $e_2(j)$, les effectifs cumulés: $E_1(j)$, $E_2(j)$, leurs différences: $(E_1(j) - E_2(j))$, leur maximum K_u, leurs valeurs absolues: $|E_1(j) - E_2(j)|$, et enfin le maximum de ces valeurs absolues K_b: $K_b = \text{Max } |E_1(j) - E_2(j)|$

La table L ci-après donne les valeurs seuils de K_b et de K_u en fonction de n_1 et du seuil choisi.

Exemple.

Reprenons l'exemple 1 de la fiche 10: après avoir examiné toutes les fiches de démonstration accompagnée de figures, chaque élève choisit celle qu'il trouve la plus commode pour la conserver dans son cahier.

Les choix sont alors les suivants:

nombre de figures:	1	2	3	4	5
nombre de choix 0	2	13	7	3	

En comparant ces choix à ceux qui avaient été faits a priori, dire si la variable "nombre de figures" agit de la même façon a priori et a posteriori, ou s'il y a un cas où le choix d'un plus grand nombre de figures l'emporte sur l'autre.

nombre de figures	1	2	3	4	5	Σ
effectifs G1	0	3	1	11	10	25
Effectifs G2	0	2	13	7	3	25
Eff. cumulésG1	0	3	4	15	25	
Eff. cumulésG2	0	2	15	22	25	
Dif:(G2-G1)	0	-1	11	7	0	
Dif abs	0	1	11	7	0	

D'où KS_b=11 KS_{2-1}=11 KS_{1-2}=1

Il apparaît dans la table à la ligne 25:

	unilatéral		bilatéral	
N	0,05	0,01	0,05	0,01
25	9	11	10	12

KS_b est significatif à 0,05 mais pas à 0,001

$KS_{(1 > 2)}$ est significatif à 0,01

TABLE DES VALEURS CRITIQUES DE D DANS
LE TEST DE **KOLMOGOROV-SMIRNOV**
A DEUX ECHANTILLONS
(petits échantillons)

N	Test unilatéral		Test bilatéral	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
3	3	-	-	-
4	4	-	4	-
5	4	5	5	5
6	5	6	5	6
7	5	6	6	6
8	5	6	6	7
9	6	7	6	7
10	6	7	7	8
11	6	8	7	8
12	6	8	7	8
13	7	8	7	9
14	7	8	8	9
15	7	9	8	9
16	7	9	8	10
17	8	9	8	10
18	8	10	9	10
19	8	10	9	10
20	8	10	9	11
21	8	10	9	11
22	9	11	9	11
23	9	11	10	11
24	9	11	10	12
25	9	11	10	12
26	9	11	10	12
27	9	12	10	12
28	10	12	11	13
29	10	12	11	13
30	10	12	11	13
35	11	13	12	
40	11	14	13	

Deux échantillons, une variable ordinale
TEST U de MANN et WHITNEY

Il indique dans quelle mesure DEUX échantillons d'observations d'UNE variable ordinale diffèrent l'un de l'autre. Il est utilisable dans le cas où l'on veut comparer deux groupes de sujets pour lesquels on a relevé une même variable numérique dont on ignore la distribution parente.

Par exemple, (1) un enseignant se demande si un groupe expérimental d'élèves G_E, obtient de meilleures notes qu'un groupe témoin G_T à un exercice. Il ne sait pas si ses notes suivent une distribution normale, il pense qu'au moins l'ordre des notes attribuées est pertinent.

Le test U de Mann et Whitney est plus puissant que celui de Kolmogorov Smirnov et s'applique dans des cas où ce dernier ne donne pas de résultat.

1) Recueil des données.

Exemple 1:

élèves de G_E	a	b	c		effectif $n_1=3$
notes	9	11	15		
élèves de G_T	a'	b'	c'	d'	$n_2=4$
notes	6	8	10	13	

Les hypothèses et les effectifs ne permettent pas de comparer les moyennes à l'aide du test t de student.

2) Transformation en rangs

Les données recueillies comme ci-dessus sont transformées en rangs de chaque élève dans la réunion des deux échantillons, et regroupées dans un tableau semblable.

Exemple 1:

élèves de G_E	a	b	c		effectif $n_1=3$
notes	5	3	1		
élèves de G_T	a'	b'	c'	d'	$n_2=4$
notes	7	6	4	2	

3) Hypothèse nulle.

H_0: Les deux échantillons ont des résultats "égaux".

Ce qui se traduit par: les observations sont issues d'un même ensemble parent. Aucun des groupes n'est beaucoup plus fort que l'autre. Les rangs, grands ou petits, devraient se trouver "également" répartis entre les deux groupes.

4) Modèle.

On ne juge pas suffisant de calculer des "moyennes de rang" et de les comparer, mais imaginons que l'on organise des compétitions deux à deux entre les élèves des deux groupes et que le rang dans la réussite décide chaque fois qui l'emporte. Il y aura donc n_1 x n_2 compétitions. Sous l'hypothèse nulle on peut s'attendre à trouver à peu près le même nombre de compétitions gagnées dans les deux camps soit 1/2 (n_1 x n_2).

5) Mesure du décalage entre les deux échantillons.

Soit U_E le nombre de compétitions emportées par les éléments de G_E (On aurait pu aussi bien considérer G_T). Ce nombre mesure l'avantage de G_E (on aurait pu prendre aussi bien: $U_E - U_T$, ou:

$$U_E = \frac{(n_1 \text{ x } n_2)}{2}$$

Exemple (1) qui de G_E domine qui de G_T

Paires

G_E \ G_T	a'	b'	c'	d'
a	domine	domine		
b	domine	domine	domine	
c	domine	domine	domine	domine

Ainsi G_E domine G_T 9 fois sur 12. donc $U_E = 9$.
On aurait pu aussi considérer U_T, qui lui, est égal à 3. Remarquons bien que:

$$U_E + U_T = n_1 \text{ x } n_2$$

6) Calcul du décalage dans le cas où n_1 et n_2 sont grands.

La construction du tableau des "compétitions" peut s'avérer fastidieuse. Voici une méthode de calcul plus commode, dont nous donnerons la démonstration plus loin:

$$\boxed{U_E = n_1 \text{ x } n_2 + \frac{[n_1 \text{ x } (n_1 + 1)]}{2} - R_1}$$

dans lequel R1 est la somme des rangs des sujets appartenant à l'échantillon G_E.
Dans le cas de l'exemple, il vient:
R1 = 5 + 3 + 1 = 9
et $U_E = 3 \text{ x } 4 + \frac{1}{2}.[3 \text{ x } 4] - 9 = 9$

7) Rareté de ce décalage.

Cette valeur U_E est-elle étonnamment grande (ou non)? Autrement dit, dans les conditions de l'hypothèse nulle, s'en produit-t-il rarement de plus grandes (ou fréquemment)? Cette question est rigoureusement équivalente à celle-ci sur U_T: U_T est elle étonnamment petite, s'en produit-il rarement de plus petites? Il revient au même de tester la rareté de $U_E > 9$ ou celle de $U_T < 3$

<u>Cas d'un grand échantillon</u> $n_2 > 30$, $n_2 >= n_1$
On démontre que sous l'hypothèse nulle U_E suit une loi normale

de moyenne $m = \frac{n_1 \text{ x } n_2}{2}$ et d'écart type $\sigma = \sqrt{\frac{n_1 \text{ x } n_2 \text{ x } (n_1 + n_2 + 1)}{12}}$

Donc il faut calculer

$$Z_U = \frac{U_E - m}{\sigma} = \frac{U_E - \frac{(n_1 \times n_2)}{2}}{\sqrt{\frac{n_1 \times n_2 \times (n_1 + n_2 + 1)}{12}}}$$

et lire la "rareté" de Z_U dans la table de la loi normale. Puisque le sens de la domination que l'on étudie est connu, le test unilatéral peut être utilisé.

La table donne la probabilité que Z s'écarte de sa moyenne (0) d'une valeur supérieure au seuil choisi.

Cas d'un petit échantillon: $9 < n_2 < 20$, $n_1 < n_2$

La lecture de la rareté se fait dans une table spéciale, la Table K.

Cas d'un très petit échantillon: $n_2 < 8$, $n_1 < n_2$

La lecture de la rareté se fait dans une table spéciale, la Table J. C'est le cas de notre exemple: Il faut lire la table $n_2 = 4$, dans la colonne $n_1 = 3$, et dans la ligne U = 3 (équivalente à U = 9) puisque la ligne U = 9 ne figure pas.

8) Décision.

Au seuil de 5%, la valeur est 1,64.

Donc si $Z_U > 1{,}64$ et si on admet H_0, la probabilité qu'une valeur de U_E plus grande que celle qui est observée apparaisse est inférieure à 5%. Dans ces conditions U_E est assez rare, donc assez grande; les élèves du groupe expérimental dominent trop souvent pour que ce soit par hasard: il convient de rejeter l'hypothèse nulle.

Démonstration de la formule du paragraphe 6

Considérons un sujet quelconque du groupe G_E, Il occupe le rang i dans son groupe et le rang r(i) dans la population entière. Il contribue à la domination de son groupe par le nombre de membres de G_T qu'il domine, c'est-à-dire qui ont un rang plus élevé que le sien:

Combien sont-ils?

Puisque ce sujet est le $i^{ème}$, il a derrière lui $n_1 + n_2 - r(i)$ sujets, dont $n_1 - i$ sont des éléments de son groupe. Il domine donc ΣU sujets de G_T.

$\Sigma U = n_1 + n_2 - r(i) - [n_1 - i] = n_2 - r(i) + i$

Le nombre de dominations obtenues par tous les éléments de G_E est la somme, pour i variant de 1 à n_1

$$U = \sum_{i=1}^{n_1}(n_2 - r(i) + i) = \sum_{i=1}^{n_1} n_2 - \sum_{i=1}^{n_1} r(i) + \sum_{i=1}^{n_1} i$$

Or

$$\sum_{i=1}^{n_1} n_2 = n_2 + n_2 + n_2 + \ldots + n_2 = n_1 \times n_2$$

$$\sum_{i=1}^{n_1} r(i) = R1$$

$$\sum_{i=1}^{n_1} i = 1 + 2 + 3 + \ldots + n_1 = \frac{n_1[n_1 \times (n_1 + 1)]}{2}$$

(en écrivant la somme dans l'ordre inverse, la somme des termes de même rang est toujours n_1+1, il y a n_1 termes; en faisant la somme on trouve $2 \Sigma i$).

D'où la formule.

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS OBSERVEES DE U
DANS LE TEST DE **MANN ET WHITNEY**

$n_2 = 3$

n_1 / U\	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.100
3		.600	.350
4			.500
5			.650

$n_2 = 4$

n_1 / U\	1	2	3	4
0	.100	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

$n_2 = 5$

n_1 / U\	1	2	3	4	5
0	.167	.047	.018	.008	004
1	.333	.095	.036	.016	008
2	.500	.190	.071	.032	016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

$n_2 = 6$

n_1 / U\	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

$n_2 = 7$

n_1 / U\	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS OBSERVEES DE U
DANS LE TEST DE **MANN ET WHITNEY**
(SUITE)

$n_2 = 8$

n_1 U\	1	2	3	4	5	6	7	8	t	Normal
0	.111	.022	.006	.002	.001	.000	.000	.000	3.308	.001
1	.222	.044	.012	.004	.002	.001	.000	.000	3.203	.001
2	.333	.089	.024	.008	.003	.001	.001	.000	3.098	.001
3	.444	.133	.042	.014	.005	.002	.001	.001	2.993	.001
4	.556	.200	.067	.024	.009	.004	.002	.001	2.888	.002
5		.267	.097	.036	.015	.006	.003	.001	2.783	.003
6		.356	.139	.055	.023	.010	.005	.002	2.678	.004
7		.444	.188	.077	.033	.015	.007	.003	2.573	.005
8		.556	.248	.107	.047	.021	.010	.005	2.468	.007
9			.315	.141	.064	.030	.014	.007	2.363	.009
10			.387	.184	.085	.041	.020	.010	2.258	.012
11			.461	.230	.111	.054	.027	.014	2.153	.016
12			.539	.285	.142	.071	.036	.019	2.048	.020
13				.341	.177	.091	.047	.025	1.943	.026
14				.404	.217	.114	.060	.032	1.838	.033
15				.467	.262	.141	.076	.041	1.733	.041
16				.533	.311	.172	.095	.052	1.628	.052
17					.362	.207	.116	.065	1.523	.064
18					.416	.245	.140	.080	1.418	.078
19					.472	.286	.168	.097	1.313	.094
20					.528	.331	.198	.117	1.208	.113
21						.377	.232	.139	1.102	.135
22						.426	.268	.164	.998	.159
23						.475	.306	.191	.893	.185
24						.525	.347	.221	.788	.215
25							.389	.253	.683	.247
26							.433	.287	.578	.282
27							.478	.323	.473	.318
28							.522	.360	.368	.356
29								.399	.263	.396
30								.439	.158	.437
31								.480	.052	.481
32								.520		

TABLE DES VALEURS CRITIQUES DE U
DANS LE TEST DE **MANN ET WHITNEY**

Table 1: test unilatéral à $\alpha = .001$ ou test bilatéral à $\alpha = .002$

n2 n1\	9	10	11	12	13	14	15	16	17	18	19	20
1												
2												
3									0	0	0	0
4		0	0	0	1	1	1	2	2	3	3	3
5	1	1	2	2	3	3	4	5	5	6	7	7
6	2	3	4	4	5	6	7	8	9	10	11	12
7	3	5	6	7	8	9	10	11	13	14	15	16
8	5	6	8	9	11	12	14	15	17	18	20	21
9	7	8	10	12	14	15	17	19	21	23	25	26
10	8	10	12	14	17	19	21	23	25	27	29	32
11	10	12	15	17	20	22	24	27	29	32	34	37
12	12	14	17	20	23	25	28	31	34	37	40	42
13	14	17	20	23	26	29	32	35	38	42	45	48
14	15	19	22	25	29	32	36	39	43	46	50	54
15	17	21	24	28	32	36	40	43	47	51	55	59
16	19	23	27	31	35	39	43	48	52	56	60	65
17	21	25	29	34	38	43	47	52	57	61	66	70
18	23	27	32	37	42	46	51	56	61	66	71	76
19	25	29	34	40	45	50	55	60	66	71	77	82
20	26	32	37	42	48	54	59	65	70	76	82	88

Table 2: test unilatéral à $\alpha = .01$ ou test bilatéral à $\alpha = .02$

n2 n1\	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	10	22
7	9	11	12	14	16	17	19	21	13	24	16	28
8	11	13	15	17	10	22	24	26	18	30	31	34
9	14	16	18	21	13	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

TABLE DES VALEURS CRITIQUES DE U

DANS LE TEST DE **MANN ET WHITNEY** (suite)

Table 3: test unilatéral à $\alpha = .025$ ou test bilatéral à $\alpha = .05$

n2 n1\	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	S7	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	10 6	113	119
20	48	55	62	69	76	83	90	98	10 5	11 2	119	127

Table 4: test unilatéral à $\alpha = .05$ ou test bilatéral à $\alpha = .10$

n2 n1\	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

K- échantillons, une variable ordinale
Test de Kruskal et Wallis
(Analyse de la variance par rangs de dimension 2)

Ce test indique dans quelle mesure **k** échantillons d'observations (K > 2) d'UNE variable ordinale, diffèrent les uns des autres. Il est utilisable dans le cas où l'on veut comparer k groupes de sujets pour lesquels on à relevé une même variable numérique dont on ignore la distribution parente.

Exemple (1) un chercheur se demande si chez les enseignants, l'orientation vers l'administration ne conduit pas à une augmentation de l'autoritarisme. Il constitue 3 groupes d'enseignants: ceux qui envisagent de rester enseignants (EE), ceux qui envisagent d'entrer dans l'administration (EA) de l'éducation, ceux qui sont déjà des administrateurs (A). Il leur fait passer une épreuve d'autoritarisme Il ne sait pas si les scores suivent une distribution normale, il pense qu'au moins leur ordre est pertinent (d'après Siegel).

1) Recueil des données.

Exemple 1: SCORE à l'épreuve

EE	EA	A
96	82	115
128	124	149
83	132	166
61	135	147
101	109	

L'effectif des groupes et l'absence d'informations sur la distribution parente commune ne permettent pas de comparer les moyennes à l'aide de l'analyse de la variance classique.

2) Transformation en rangs.

Les données recueillies comme ci-dessus sont transformées en rangs de chaque élève dans la réunion des deux échantillons, et regroupées dans un tableau semblable.

Exemple 1: RANGS

EE	EA	A
4	2	7
9	8	13
3	10	14
1	11	12
5	6	
R1 = 22	R2 = 37	R3 = 46

3) Hypothèse nulle.

H_0: Les k échantillons ont des résultats "égaux".

Ce qui se traduit par: les observations sont issues d'un même ensemble parent. Aucun des groupes n'est beaucoup plus fort que l'autre. Les rangs, grands ou petits devraient se trouver "également" répartis entre les k groupes.

4) Calcul du décalage entre les groupes.

Calculer

$$H = \left(\frac{12}{N(N+1)} \times \sum_j \frac{R_j^2}{n_j} \right) - 3(N+1)$$

où N est le nombre total d'observations (de sujets)

n_j, le nombre d'observations dans l'échantillon j.

R_j, la somme des rangs dans l'échantillon j

Dans l'exemple ci-dessus

$$H = \frac{12}{14(14+1)} \times \left[\frac{22^2}{5} + \frac{37^2}{5} + \frac{46^2}{4} \right] - 3(N+1)$$

$H = 6{,}4$

5) Cas des ex-æquo.

S'il y a des ex aequo, chacun reçoit un rang égal à la moyenne des rangs que devraient occuper ces ex aequo. L'existence d'ex aequo tend à diminuer la valeur du coefficient H. On peut alors calculer un H corrigé: H_C ,

$$H_C = \frac{H}{1 - \frac{\sum_j T_j}{(N\,(N^2 - 1))}}$$

dans lequel ΣT_j est la somme, pour tous les groupes, du nombre T_j d'ex aequo dans le groupe j.

Si le calcul avec H conduit à rejeter l'hypothèse nulle, il est inutile de calculer H_c, la conclusion ne changera pas. La correction n'a d'intérêt que si le nombre des ex aequo est grand (supérieur à 30%) et si H est assez voisin du seuil de rejet (par ex moins de 10%).

6) Rareté de ce décalage.

Cette valeur H est elle étonnamment grande (ou non)? autrement dit, dans les conditions de l'hypothèse nulle, s'en produit-il rarement de plus grandes (ou fréquemment)?

CAS d'UN GRAND ECHANTILLON: il existe un n_i tel que $n_j > 5$ et $k \leq 3$

On démontre que sous l'hypothèse nulle, H suit une loi de CHI carré
de degré de liberté k - 1
La table du chi^2 donne la probabilité que, sous l'hypothèse nulle, H dépasse la valeur trouvée ci-dessus. En fait, on compare H à la valeur seuil.

CAS D'UN PETIT ECHANTILLON: n_j 5 et k = 3

La lecture de la rareté se fait dans une table spéciale, la Table O. C'est le cas de notre exemple: Il faut lire la table 5, 5, 4, dans la colonne H on lit que sous l'hypothèse nulle H > 5,6429 a une probabilité de se produire inférieure à 0,050. (moins de 5% des cas)

7) Décision.

Au seuil de 5%, la valeur du chi 2 est 5,6429.
Donc si H > 5,6429 et si on admet H_0, la probabilité qu'une valeur de H plus grande que celle qui est observée apparaisse est inférieure à 5%. Ici H = 6,4. Dans ces conditions H est assez rare, donc assez grand: il faut rejeter l'hypothèse nulle: la perspective de l'administration incline les enseignants vers l'autoritarisme.

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS OBSERVEES DE H
DANS LE TEST UNILATERAL DE **KRUSKAL ET WALLIS**
(analyse de la variance par rangs).

taille échant.			H	p	taille échant			H	p	taille échant			H	P
n1	n2	n3			n1	n2	n3			n1	n2	n3		
2	1	1	2.7000	.500	4	4	1	6.6667	.010	5	4	1	6.9545	.008
								6.1667	.022				6.8400	.011
2	2	1	3.6000	.200				4.9667	.048				4.9855	.044
2	2	2	4.5714	.067				4.8667	.054				4.8600	.056
			3.7143	.200				4.1667	.082				3.9873	.098
								4.0667	.102				3.9600	.102
3	1	1	3.2000	.300										
3	2	1	4.2857	.100	4	4	2	7.0364	.006	5	4	2	7.2045	.009
			3.8571	.133				6.8727	.011				7.1182	.010
								5.4545	.046				5.2727	.049
3	2	2	5.3572	.029				5.2364	.052				5.2682	.050
			4.7143	.048				4.5545	.098				4.5409	.098
			4.5000	.067				4.4455	.103				4.5182	.101
			4.4643	.105										
					4	4	3	7.1439	.010	5	4	3	7.4449	.010
3	3	1	5.1429	.043				7.1364	.011				7.3949	.011
			4.5714	.100				5.5985	.049				5.6564	049
			4.0000	.129				5.5758	.051				5.6308	.050
								4.5455	.099				4.5487	.099
3	3	2	6.2500	.011				4.4773	.102				4.5231	.103
			5.3611	.032										
			5.1389	.061	4	4	4	7.6538	.008	5	4	4	7.7604	.009
			4.5556	.100				7.5385	.011				7.7440	.011
			4.2500	.121				5.6923	.049				5.6571	.049
								5.6538	.054				5.6176	.050
3	3	3	7.2000	.004				4.6539	.097				4.6187	.100
			6.4889	.011				4.5001	.104				4.5527	.102
			5.6889	.029										
			5.6000	.050	5	1	1	3.8571	.143	5	5	1	7.3091	.009
			5.0667	.086									6.8364	.011
			4.6222	.100	5	2	1	5.2500	.036				5.1273	.046
								5.0000	.048				4.9091	.053
4	1	1	3.5714	.200				4.4500	.071				4.1091	.086
								4.2000	.095				4.0364	.105
4	2	1	4.8214	.057				4.0500	.119					
			4.5000	.076						5	5	2	7.3385	.010
			4.0179	.114	5	2	2	6.5333	.008				7.2692	.010
								6.1333	.013				5.3385	.047
4	2	2	6.0000	.014				5.1600	.034				5.2462	.051
			5.3333	.033				5.0400	.056				4.6231	.097
			5.1250	.052				4.3733	.090				4.5077	.100
			4.4583	.100				4.2933	.122					
			4.1667	.105						5	5	3	7.5780	.010
					5	3	1	6.4000	.012				7.5429	.010
4	3	1	5.8333	.011				4.9600	048				5.7055	.046
			5.2083	.050				4.8711	.052				5.6264	.051
			5.0000	.057				4.0178	.095				4.5451	.100
			4.0556	.093				3.8400	.123				4.5363	.102
			3.8889	.129										
					5	3	2	6.9091	.009	5	5	4	7.8229	.010
4	3	2	6.4444	.008				6.8218	.010				7.7914	.010
			6.3000	.011				5.2509	.049				5.6657	.049
			5.4444	.046				5.1055	.052				5.6429	.050
			5.4000	.051				4.6509	.091				4.5229	.099
			4.5111	.098				4.4945	.101				4.5200	.101
			4.4444	.102										
4	3	3	6.7455	.010	5	3	3	7.0788	.009	5	5	5	8.0000	.009
			6.7091	.013				6.9818	.011				7.9800	.010
			5.7909	.046				5.6485	.049				5.7800	.049
			5.7273	.050				5.5152	.051				5.6600	.051
			4.7091	.092				4.5333	.097				4.5600	.100
			4.7000	.101				4.4121	.109				4.5000	.102

C. VARIABLES ORDINALES

II. SUJETS DE DIDACTIQUE DES MATHEMATIQUES (2)

Enoncé

Devoir n°2

On a demandé à un élève de ranger une série de 24 problèmes qu'il a résolus depuis le début de l'année scolaire, par ordre de difficulté croissante.
Il range ces problèmes, numérotés de 1 à 24, dans l'ordre suivant:

21, 10, 22, 9, 2, 23, 11, 7, 3, 24, 19, 4, 1, 8, 14, 12, 20, 6, 16, 13, 5, 17,15, 18,

Analysant ces réponses, un didacticien envisage trois sortes de variables (ou de conditions) susceptibles à son avis d'influencer l'opinion de cet élève:
a) La longueur du texte du problème (L)
b) La nature des opérations à effectuer (O)
c) L'ancienneté et la familiarité des problèmes (A) (i.e. le nombre de problèmes semblables rencontrés en classe)
Il classe donc les 24 problèmes en fonction de ces trois variables et obtient:

Pour L:
énoncés très courts: 21, 22, 23, 24, 10, 11, 12, 13
énoncés de longueur moyenne: 9, 2, 3, 14, 8, 1, 4, 15,
énoncés longs: 7, 6, 5, 16, 20, 19, 18, 17,

Pour O:
problèmes d'addition:21, 10, 9, 8, 7, 20,
problèmes de soustraction: 22, 11, 2, 1, 6, 19,
problèmes de multiplication:23, 12, 3, 4, 5, 18,
problèmes de division:24, 13, 14, 15, 16, 17,

Pour A:
problèmes anciens et familiers:21, 22, 23, 24, 9, 2, 3, 14, 7, 6, 5, 16,
problèmes nouveaux ou peu familiers: 10, 11, 12, 13, 8, 1, 4, 15, 20, 19, 18, 17,

A l'aide des tests de KRUSKAL et WALLIS ou de MANN et WHITNEY, établir ce qu'il doit conclure des réponses de cet élève.

(Maîtrise Sciences de l'éducation, Bordeaux 1990)

CORRECTION du Devoir n°2

1. Rangs attribués aux différents problèmes.

Les « nombres » attribués aux problèmes sont des numéros, des noms, ce sont les rangs obtenus qui nous intéressent:

Numéros	rangs	Numéros	rangs
1	13	13	20
2	5	14	15
3	9	15	23
4	12	16	19
5	21	17	22
6	18	18	24
7	8	19	11
8	14	20	17
9	4	21	1
10	2	22	3
11	7	23	6
12	16	24	10

2. La longueur du texte influence-t-elle l'impression de difficulté ressentie par l'élève?

a) H_0: Les rangs des problèmes courts, moyens, ou longs ne diffèrent pas significativement. Ils auraient de bonnes chances d'apparaître ainsi au cours d'un tirage au hasard dans une même population parente.

b) transformation en rangs

énoncés courts	1, 3, 6, 10, 2, 7, 16, 20
énoncés de longueur moyenne	4, 5, 9, 15, 14, 13, 12, 23
énoncés longs	8, 18, 21, 19, 17, 11, 24, 22

c) Calcul des éléments de la formule de KRUSKAL et WALLIS:

$$H = \left(\frac{12}{N(N+1)} \times \Sigma_j \frac{R_j^2}{n_j} \right) - 3(N+1)$$

somme des rangs dans chaque échantillon et contribution de chaque classe, n1= n2= n3= 8

$R_1 = 65$, $R_1^2 = 4225$, $R_1^2/n_1 = 528{,}125$

$R_2 = 95$, $R_2^2 = 9025$, $\frac{R_2^2}{n_2} = 1128{,}125$

$R_3 = 140$, $R_3^2 = 19600$, $\frac{R_3^2}{n_3} = 2450$

$\frac{\Sigma_j R_j^2}{n_j} = 4106{,}25$ $\quad \frac{12}{N(N+1)} = \frac{1}{50} = 0{,}02$

$3(N+1) = 75$

Portons ces valeurs dans la formule

$H = 7{,}125$

d) Rareté de cette valeur.

Puisque l'effectif d'une classe dépasse 5, H suit une loi de Chi 2 avec dl = k-1 = 2

Dans la table du Chi 2, dans la ligne 2, nous trouvons que le H observé est compris entre les deux valeurs 5,99 et 7,82 qui correspondent respectivement à des probabilités de 0,05 et 0,02. Il y a moins de 5% de valeurs (et plus de 2%) de H qui seront plus grande que la valeur 7,125. Cette valeur est rare donc ici grande.

e) Conclusion:

L'hypothèse nulle doit être rejetée: la longueur de l'énoncé influence le rangement des problèmes par cet élève.

3. La nature des opérations à effectuer influe-t-elle sur l'impression de difficulté ressentie par l'élève?

a) H_0. Les rangs des problèmes d'addition, de soustraction, de multiplication ou de division ne diffèrent pas significativement. Ils auraient de bonnes chances d'apparaître dans l'ordre où ils sont au cours d'un tirage au hasard dans une même population parente.

b) transformation en rangs

Additions	1,2,4,14,8,17
soustractions	3,7,5,13,18,11
multiplications	6,16,9,12,21,24
divisions	10,20,15,23,19,22

c) Calcul des éléments de la formule de KRUSKAL et WALLIS:

somme des rangs dans chaque échantillon et contribution de chaque classe,

$n_1 = n_2 = n_3 = n_4 = 6$

$R_1 = 47$, $R_1^2 = 2209$, $\frac{R_1^2}{n_1} = 352{,}66$

$R_2 = 57$, $R_2^2 = 4249$, $\frac{R_2^2}{n_2} = 541{,}5$

$R_3 = 88$, $R_3^2 = 7744$, $\frac{R_3^2}{n_3} = 1290{,}66$

$R_4 = 109$, $R_4^2 = 11881$, $\frac{R_4^2}{n_4} = 1980{,}16$

$\Sigma_j \frac{R_j^2}{n_j} = 4165$ $\qquad \frac{12}{N(N+1)} = \frac{1}{50} = 0{,}02$

$3(N+1) = 75$

Portons ces valeurs dans la formule

$H = 8{,}03$

d) Rareté de cette valeur.

Puisque l'effectif d'une classe dépasse 5, H suit une loi de Chi2 avec dl = k-1 = 3

Dans la table du Chi2, dans la ligne 3, nous trouvons que le H observé est compris entre les deux valeurs 7,82 et 9,84 qui correspondent respectivement à des probabilités de 0,05 et 0,02. Il y a moins de 5% de valeurs (et plus de 2%) de H qui seront plus grande que la valeur 8,03. Cette valeur est rare donc, ici, grande.

e) Conclusion:

L'hypothèse nulle doit être rejetée: la nature des opérations ‘à effectuer’ influence le rangement de cet élève.

4. L'ancienneté des problèmes influe-t-elle sur l'impression de difficulté ressentie par l'élève?

a) H_0 : Les rangs des problèmes anciens et des problèmes nouveaux ne diffèrent pas significativement. Ils auraient de bonnes chances d'apparaître dans cet ordre au cours d'un tirage au hasard dans une même population parente.

b) *Transformation en rangs*

problèmes anciens et familiers: 1, 3, 6, 10, 4, 5, 9, 15, 8, 18, 21, 19

problèmes nouveaux ou peu familiers: 2, 7, 16, 20, 14, 13, 12, 23, 17, 11, 24, 22

Nous pouvons utiliser soit le test U de Mann et Whitney, soit comme ci-dessus celui de Kruskal et Wallis

c) *Calcul des éléments de la formule de KRUSKAL et WALLIS*:

somme des rangs dans chaque échantillon et contribution de chaque classe, $n_1 = n_2 = 12$

$R_1 = 119$, $R_1^2 = 14\,161$, $\frac{R_1^2}{n_1} = 1180$

$R_2 = 181$, $R_2^2 = 32\,761$, $\frac{R_2^2}{n_2} = 2730$

$\Sigma_j \frac{R_j^2}{n_j} = 3910$ $\frac{2}{N(N+1)}1 = \frac{1}{50} = 0{,}02$

$3(N+1) = 75$

En portant ces valeurs dans la formule, il vient : $H = 3{,}20$

d) *Rareté de cette valeur*.

Puisque l'effectif d'une classe dépasse 5, H suit une loi de Chi 2 avec $dl = k-1 = 1$

Dans la table du Chi2, dans la ligne 1, nous trouvons que le H observé est compris entre les deux valeurs 2,71 et 3,84 qui correspondent respectivement à des probabilités de 0,10 et 0,05. Il y a moins de 10% des valeurs et plus de 5% de H qui seront plus grande que la valeur 3,20. Cette valeur n'est pas assez rare donc ici pas assez grande.

e) *Conclusion*:

L'hypothèse nulle ne peut être rejetée: l'ancienneté ou la familiarité du problème n'influencent pas le rangement des problèmes par cet élève.

Calcul du U de MANN et WHITNEY.

$n_1 = 12$ $n_2 = 12$ R2 = 181

$U = n_1 \times n_2 + 1/2(n_1(n_1 + 1)) - R2$

U = 144 + 78 - 181 = 41 (et non pas: U = 144 + 78 - 119 = 103, car R1 est la somme des rangs des sujets de l'échantillon 1).

Le modèle correspondant à l'hypothèse nulle n'exclut aucune des 2 hypothèses: les problèmes anciens sont plus faciles ou au contraire sont moins faciles. La distribution parente est bilatérale.

Rareté de cette valeur: n_1 et n_2 étant inférieurs à 20, U suit une distribution différente de la distribution normale: Il faut recourir aux tables K.

Comme nous avons pris la valeur la plus petite, plus cette valeur est petite (moins il y a de dominations dans le camp le plus faible), plus elle est rare.

Suivant le seuil de "rareté" retenu on trouve à l'intersection des ligne 12, colonne 12 des différents tableaux les valeurs suivantes:

20 correspond au seuil de 0,1% (unilatéral), (1 pour mille, qui s'écrit encore 0,001), ou 0,2% (bilatéral)

31 correspond au seuil de 1% (unilatéral) et 2% (bi)

37 correspond au seuil de 2,5% (uni) et 5% (bi)

42 correspond au seuil de 5% (uni) et 10% (bi).

$$U_{0,1} > U_{\text{observé}} > U_{0,05} \text{ puisque } 42 > 41 > 37$$

f) *Décision*:

Ce résultat confirme les résultats donnés par le test de Kruskal et Wallis:

Les problèmes de la catégorie 1 (textes familiers) ne dominent pas assez souvent ceux de la catégorie 2, l'écart observé peut être dû au hasard, l'hypothèse nulle ne peut être rejetée.

C. VARIABLES ORDINALES

III. HOMOGENEITE POUR PLUSIEURS VARIABLES

FICHE 14

DEUX échantillons appariés, UNE variable ordinale ou UN échantillon et DEUX variables semblables
Épreuve des signes

Cette épreuve est utilisable lorsqu'on possède, pour chaque élément d'un échantillon, un couple d'observations dont on peut dire laquelle est plus grande que l'autre. Elle permet de dire si une des variables l'emporte sur l'autre.

Elle est aussi utilisable avec deux variables numériques, mais elle ne prend en compte qu'une petite partie de l'information disponible (laquelle est supérieure à l'autre). On la choisit donc par défaut de méthodes plus puissantes:

- quand la distribution des variables n'est pas normale, ou quand on l'ignore (on ne peut pas alors appliquer le t de student)
- et si le test de Wilcoxon exposé dans la fiche 15 s'applique mal (par exemple les différences entre les valeurs des deux variables ne prennent qu'un très petit nombre de valeurs différentes).

Exemple (1). On a présenté à des élèves une expression numérique complexe et on demande à chacun de choisir entre deux transformations pour continuer les calculs: l'une utilise seulement les naturels, l'autre conduit à calculer sur des relatifs. On se demande si les élèves sont indifférents à ces transformations ou s'ils en préfèrent une. 19 ont ainsi exprimé leur préférence pour la transformation A, (A>B), 9 ont préféré la transformation B, (B>A).

Exemple (2). Le même test est présenté à des élèves avant et après l'enseignement d'une connaissance capable de renouveler leurs méthodes de résolution. On note le nombre d'erreurs x et y observées dans chacune des épreuves (test et retest). On se demande si l'enseignement a eu pour effet d'améliorer leurs résultats.

1) Recueil et résumé des données.

Soit les résultats s'expriment directement en préférences et donc en signes, comme dans l'exemple 1. soit on effectue la transformation de valeurs en signe en formant la différence d = x - y (qui correspond à X > Y) et en notant le signe de cette différence.

Exemple 1

sujets	choix	A > B
a	B	-
b	B	-
c	A	+
…	…	…
Total	+ : 16 - : 8	

Exemple 2

Observ.	X	Y	d	Signe
a	5	2	3	+
b	2	3	-1	-
c	4	4	0	0
d	9	2	-2	-
…	…	…	…	…
Total	Nombre de + : 34 Nombre de - : 20 Nombre de 0 : 8			

Les résultats se présentent ainsi : sur N observations il y a n(+) avantages pour A, n(-) avantages pour B, n(0) indifférents. Le nombre de préférences déclarées est:

$$n = n(+) + n(-) .$$

et on pose $k = \inf [n(+) ; n(-)]$

Dans l'exemple 1: n = 24, k = 8

Dans l'exemple 2: n = 54, k = 20

2) Hypothèse nulle.

H_0: Il n'y a pas de différence entre les deux variables.

Autrement dit les résultats au test et au retest sont les mêmes,
ou encore: probabilité (A>B) = probabilité (A<B) = 1/2

3) Modèle.

Nous avons déjà étudié une hypothèse de ce type dans la leçon 1 à l'aide du modèle du X^2. Nous allons procéder ici de la même manière mais étudier un autre modèle:

Le choix de chaque élève est représenté par un tirage à pile ou face: pile: A>B, face: B>A. Le choix de n élèves sera représenté par n tirages indépendants.

4) Comparaison de l'observation avec le modèle théorique.

Il s'agit de savoir si le nombre k observé est assez près de n/2 pour qu'on admette le modèle théorique. On va mesurer cette distance directement par sa rareté: Ce nombre k pourrait-il être obtenu, avec une probabilité raisonnable, par la répétition indépendante de n tirages à pile ou face?

Dans l'exemple 1 ci dessus, est-il fréquent de n'obtenir que 8 piles lorsqu'on fait 24 tirs à pile ou face? On devrait en obtenir en moyenne 12.

La distribution de la probabilité d'avoir k ou moins de k piles en n tirages suit une loi binômiale. Pour déterminer la la rareté du résultat observé, on peut utiliser une simulation, un calcul de probabilité ou son résultat sous forme de table de cette distribution comme il est indiqué dans la fiche 7 et en annexe.

5) décision.

Si cette proportion est inférieure à 5 % on rejettera l'hypothèse que
proba (A>B) = proba (A<B)

<u>Cas où n ≤ 25</u>

La table de la loi binomiale (ci-après) indique en fonction de n quel est la valeur k_{seuil} de rejet au seuil choisi:

Si $k \leq k_{seuil}$ le nombre de cas où l'une des valeurs l'emporte sur l'autre est trop petit pour être l'effet du modèle théorique, que l'on rejette alors.

Dans l'exemple 1: n = 24 ; sur la ligne 24 on lit 7 dans la colonne 0,10. Cela veut dire qu'il y a plus 10 % de chances de trouver moins de 7 piles sur 24 tirages à pile ou face. On lit 6 dans la colonne 0,05: il faudrait donc 6 ou moins piles pour que l'on rejette l'hypothèse nulle.

Ce n'est pas le cas, donc on ne rejette pas l'hypothèse nulle (qui n'est pas prouvée pour autant).

<u>Cas où n > 25</u>

Lorsque n croît la loi binômiale B(n,p) tend à se confondre avec la loi normale
de moyenne: m = n.p et donc ici m = 1/2 n

et d'écart type $\sigma = \sqrt{n.p.(1 - p)} = 1/2 \sqrt{n}$

Par conséquent la variable z:

$$\boxed{z = \frac{k - \frac{n}{2}}{\frac{1}{2} \cdot \sqrt{n}}}$$

suit la loi normale centrée réduite:
Il suffit de calculer z_o et de la placer par rapport à la table de la loi normale.

Dans l'exemple 2 on trouve:

$$z = \frac{19 - 27}{\frac{1}{2} \cdot \sqrt{54}}$$

Correction de continuité (Correction de YATES)

Il est préférable d'utiliser la formule corrigée qui diminue la valeur absolue de z_o pour compenser l'effet de la discontinuité de la distribution:

$$z = \frac{(k + 0{,}5) - \frac{n}{2}}{\frac{1}{2} \cdot \sqrt{n}}$$

Dans l'exemple 2, on obtient

$$z = \frac{19{,}5 - 27}{\frac{1}{2} \cdot \sqrt{54}} = -2{,}04$$

or $1{,}96 < 2{,}04$ entraîne $z_{oseuil.05} < z_o$

Donc on rejette l'hypothèse nulle: Les avantages du retest sur le test sont trop nombreux pour que l'on puisse attribuer au hasard la différence observée.

Conclusion: Le retest est significativement meilleur que le test

Exercice. Appliquer aux deux exemples ci dessus le test du χ^2 exposé à la leçon 1 et comparer les résultats des deux méthodes.
Inversement, peut-on appliquer le test des signes aux problèmes proposés dans la leçon 1?
Pourquoi?

Test Unilatéral, test bilatéral.

Le fait de poser k =inf(n(+),n(-)) permet de ne compter que les $j \leq k$ pour conclure. Il faut donc prendre une table pour test unilatéral.

Si nous avions étés conduits à examiner l'écart de k avec N/2 sans en connaître le sens il aurait fallu utiliser le test bilatéral.

TABLE DES VALEURS CRITIQUES DE K
POUR L'**EPREUVE DES SIGNES**.
(test bilatéral)

seuil n	0,10	0,05	0,02	0,01	0,001
5	0	-	-	-	-
6	0	0	-	-	-
7	0	0	0	-	-
8	1	0	0	0	-
9	1	1	0	0	-
10	1	1	0	0	-
11	2	1	1	0	0
12	2	2	1	1	0
13	3	2	1	1	0
14	3	2	2	1	0
15	3	3	2	2	1
16	4	3	2	2	1
17	4	4	3	2	1
18	5	4	3	3	1
19	5	4	4	3	2
20	5	5	4	3	2
21	6	5	4	4	2
22	6	5	5	4	3
23	7	6	5	4	3
24	7	6	5	5	3
25	7	7	6	5	4

FICHE 15

DEUX échantillons appariés, UNE variable ordinale, ou UN échantillon et DEUX variables semblables. Epreuve de WILCOXON

Cette épreuve est utilisable lorsqu'on possède, pour chaque élément d'un échantillon, un couple d'observations dont on peut dire non seulement le sens de préférence mais que l'on peut aussi ranger selon l'importance de la préférence.

Elle est aussi utilisable avec deux variables numériques, et elle prend en compte non seulement le signe mais aussi la grandeur de la différence. Dans ce cas elle utilise donc mieux l'information que l'épreuve des signes. Elle remplace le t de Student, plus puissant, mais inutilisable quand la distribution des variables n'est pas normale ou quand on l'ignore.

Exemple 1.

On enseigne à des élèves qui connaissent déjà une méthode de calcul un procédé supposé plus efficacet. On redoute que le changement de méthode ne fasse diminuer les performances. On propose donc avant et après la leçon, deux séries de 10 opérations de difficultés équivalentes et on note le nombre d'opérations exactes dans chaque exercice. On possède les résultats pour 24 élèves.

Il s'agit de savoir si les nombres de réussites diffèrent significativement entre les deux épreuves.

Exercice 1.

Quelle épreuve permettrait d'assurer que les deux séries d'opérations sont bien de difficultés équivalentes?

Exemple 2.

Comme on redoute aussi que la méthode nouvelle ne demande plus de temps, pour l'exécution, que l'ancienne, on note le temps nécessaire à chaque élève pour effectuer l'exercice. Les résultats pour 44 élèves sont donnés en fin de leçon. Conclusion?

1) Recueil et résumé des données.

Les éléments de l'échantillon sont disposés en colonne et les deux variables X_1 et X_2 sur deux colonnes, trois colonnes supplémentaires sont nécessaires.

On forme la différence $d = x_1 - x_2$,entre les valeurs de la première et les valeurs de la deuxième variable, comme dans le test des signes, mais au lieu de ne conserver que le signe de la différence on conserve ici sa valeur.

Différence nulle: les lignes où d = 0, c'est-à-dire où $x_1 = x_2$, ne contribueront pas au calcul. De sorte que l'effectif n sera celui des <u>différences non nulles</u> et non celui de l'échantillon.

Ainsi dans le tableau ci dessous, qui correspond à l'exemple 1, les différences nulles ignorées, il apparaît que l'effectif actif est de 15 et non de 24.

Puis on range suivant les <u>valeurs absolues</u> croissantes: on attribue un rang à ces différences, sans tenir compte de leur signe, dans l'ordre croissant (le rang 1 est attribué à la plus petite, le rang 2 à la suivante ...).

Ce n'est qu'ensuite que l'on attibue à ce rang le signe de la différence qui l'a déterminé.

Dans le cas d'ex aequo, l'attribution du rang se fait suivant la règle habituelle: somme des rangs ex aequo divisée par nombre d'ex aequo

Exemple 1.

X1 est le nombre d'opérations fausses au test,

X2 le nombre d'opérations fausses au re-test.

sujets	X1	X2	d=X1-X2	val abs	Rang de d	contribution
1	3	2	1	1	4	4
2	5	2	3	3	13	13
3	1	3	-2	2	9,5	-9,5
4	0	1	-1	1	4	-4
5	2	0	2	2	9,5	9,5
6	2	1	1	1	4	4
7	1	0	1	1	4	4
8	3	1	2	2	9,5	9,5
9	2	3	-1	1	4	-4
10	1	2	-1	1	4	-4
11	4	1	3	3	13	13
12	1	3	-2	2	9,5	-9,5
13	6	3	3	3	13	13
14	4	5	-1	1	4	-4
15	3	6	-4	4	15	-15
16à24	a	a	0	0	0	0
Somme T					120	
Somme T +						70
Somme T -						50

Une méthode de calcul pratique du rang des ex aequos est exposée en annexe de cette leçon.

Remarquons que T la somme des rangs est est égale à la somme de T(+), la somme des rangs positifs et de T(-), la somme des rangs négatifs.

$T = T(+) + T(-)$

et que

$T = 1+2+3+ ... + n = ½ (n+1)+(n-1+1)+...+(1+n)$

(en faisant la somme du premier terme et du dernier, ajoutée à la somme du second et de l'avant dernier etc).

donc: $T = \frac{n \times (n+1)}{2}$

Dans notre exemple:

$1+2+3+...+15 = 15 \times (15+1) /2 = 120$

$T(+) + T(-) = 70 + 50 = 120$

2) Hypothèse nulle.

H_0: " il n'y a pas de différence entre les deux variables."

Autrement dit les résultats au test et au retest sont les mêmes,

Si l'hypothèse nulle était vraie, T+ et T- ne devraient pas différer significativement.

Ou encore T(+) ne devrait pas différer trop de T/2. (il en diffère autant que T(-)).

3) Modèle.

Comme dans tous les problèmes de ce type nous allons mesurer la distance entre T(+) et T/2 par sa rareté sous l'hypothèse nulle. Le nombre T(+) - T/2 pourrait-il être obtenu, avec une probabilité raisonnable dans l'expérience modèle:

On répète n matches un contre un de façon indépendante avec des probabilités égales de gain.

Puis on attribue (de façon aléatoire) les rangs de 1 à n aux différents résultats affectés de leurs signes. On calcule alors T(+).

Dans l'exemple 1 ci-dessus, est-il fréquent d'obtenir 50 ou moins de 50 pour T(-) au lieu de 60 (T/2 = 60) lorsqu'on fait 15 matchs à égalité (ou 15 tirages à pile ou face)?

Il s'agit de connaître la probabilité d'avoir une somme des rangs négatifs de T(-) ou moins de T(-), au cours d'un tel match, afin de comparer la valeur de T(-) observée à la fréquence théorique des valeurs ainsi obtenues.

WILCOXON a déterminé cette distribution théorique de T(-) (nous pourrions le refaire dans ce cas aussi par simulation, comme nous l'indiquons en annexes). Il a été conduit à distinguer deux cas:

Pour n > 25 on peut utiliser la table de la loi normale.

Pour n ≤ 25 il faut se reporter à la table de Wilcoxon ci après..

4) Distance de l'observation au modèle

Cas où n ≤ 25

Prenons la valeur la plus faible de T(+) et de T(-). Comparons cette valeur aux différents seuils de la ligne n du tableau. Ce tableau donne les valeurs critiques au-dessous desquelles la valeur de T observée est significativement faible, ce qui conduit à rejeter l'hypothèse d'égalité des scores.

Dans notre exemple T(-) = 50. Or on lit sur la ligne 15 que T(-) devrait être inférieur ou égal à 25 pour que la différence soit significative à .05.

On ne peut donc absolument pas rejeter l'hypothèse que la nouvelle méthode n'a pas permis une amélioration significative des résultats des élèves, malgré l'amélioration de leurs scores, ni non plus celle qu'elle ne les a pas contrariés.

Cas où n > 25.

Si n est supérieur à 25, on a pu montrer que T(-) suit une distribution normale de moyenne

$$m = \frac{T}{2} = \frac{n.(n+1)}{4}$$

et d'écart type $\sigma = \sqrt{\frac{n.(n+1)(2n+1)}{24}}$

Alors on calcule $z = \frac{|T(-)| - m}{\sigma}$

5) Décision.

Cette valeur z_o observée, est comparée à la distribution de la loi normale centrée réduite z ;

si $z_o \leq z_{seuil\ 0.05}$ la valeur de T est suffisamment petite (grande en valeur absolue) ou grande pour que l'on puisse rejeter l'hypothèse nulle.

6) Exemple 2.

Les effectifs sont présentés ci dessous, de façon réduite, dans un tableau différent de l'exemple 1. De plus au lieu de noter les valeurs effectives du temps mis par chaque élève pour chaque exercice, l'observateur note les différences de temps entre le premier et le deuxième exercice. Elles s'échelonnent de -9 minutes à +9 minutes.

Enfin, la méthode de chronométrage assure des différences entières car elle consistait à relever les exercices terminés à intervalles fixes déterminés. Le nombre de minutes s'écoulant entre deux relevés n'a pas d'importance.

Les différences de temps sont donc les valeurs diff = $(t_1 - t_2)$ et l'effectif correspondants est celui des élèves qui ont remis leur devoir dans des "temps présentant cette différence.

Dans le tableau ci dessous l'élève classé dans la colonne -9 a effectué son 2ème devoir en neuf minute de plus que son premier.

diff:	-9	-7	-5	-3	0	3	5	7	9
eff:	1	0	3	8	13	7	6	3	3
efabs					0	15	9	3	4
efcum						15	24	27	31
rang:	-29,5	-26	-20	-8	0	8	20	26	29,5
ctr:	-29,5	0	-60	-64	0	56	120	78	88,5

efabs représente l'effectif obtenu, pour chaque valeur positive de la différence, en comptant tous les élèves qui présentent cette différence de temps, quel qu'en soit le signe. L'élève qui a peiné 9 minutes de plus sur le second problème est compté avec ceux qui ont mis 9 minutes de moins.

« efcum » sont les effectifs cumulés, le rang est obtenu comme dans l'exemple 1 et la contribution ctr d'une valeur de la variable est la somme des rangs obtenus par les élèves de cette catégorie, en fait le produit du rang par l'effectif puisque tous les élèves d'une même colonne sont ex aequos.

T(-) = - 153,5	n = 30	
T(+) = 342,5	m = 248	Z_o = 1,851
T = 496	σ= 51, 02	

Cette valeur de Z est significative à O,04 (4 %) car on lit dans la table des valeurs en Z sur la ligne de 1,8 et dans la colonne ,05 (qui correspond à 1,85) que la probabilité d'avoir une valeur plus grande que z_o = 1,85 est inférieure à 0,0322. donc à 0,04. et à fortiori à 0,05 car

$z_{seuil\ 0,05} \leq z_{seuil\ 0,04} \leq z_o$

Conclusion: Le temps de calcul avec la nouvelle méthode est significativement plus court (puisqu'on a calculé t1 - t2.

Exercice.

Dans l'exemple 2, la distribution des effectifs sur les différences a une allure normale: elle est à peu près unimodale et symétrique. Si on appliquait le test t de Student, quelle conclusion obtiendrions nous?

Rappel: le T de STUDENT

Si l'échantillon de n (n < 30) différences observées, a une moyenne m_d et une variance (débiaisée) $s^2 = \frac{1}{n-1}.\sum_{i=1}^{n}(x_i - m_d)^2$ et qu'il soit extrait d'une population parente distribuée suivant la loi Normale de moyenne m, alors la variable t définie ci-dessous suit une loi de STUDENT à n-1 degrés de liberté.

$$t = \frac{m_d - m}{\frac{s}{\sqrt{n}}} = \frac{(m_d - m).\sqrt{n}}{s}$$

dans l'exemple 2, ef_i est l'effectif de la différence d_i, il vient par conséquent

$$m_d = \frac{\sum_i (ef_i.d_i)}{n} \quad , \quad s = \sqrt{\frac{\sum_i (ef_i d_i - m_d)^2}{(n-1)}} \quad , \quad \text{et}$$

$$t = \sqrt{\frac{n-1}{n}} . \frac{\sum_i (ef_i.d_i)}{\sqrt{\sum_i (ef_i d_i - m_d)^2}}$$

valeur à éprouver dans la table du t de STUDENT à dl = n-1

Dans le cas ou l'échantillon est de taille inférieure à 30, le t de Student permet de savoir si la moyenne m_d des n différences observées diffère significativement de zéro.

Transformation des valeurs en rangs.

Lorsque la différence présente de nombreux ex aequo il peut être commode de calculer les rangs moyens (ou de les faire calculer) par la méthode suivante.

Soit d_i la ième valeur de la différence par ordre de valeur croissante. Dans l'exemple 1, la différence 2 est la deuxième, $d_2 = 2$.

$e(d_i)$ l'effectif de l'échantillon présentant cette différence.

Il y a 4 sujets qui présentent la valeur 2, donc $e(d_2) = 4$

soient n_{i1} les effectifs cumulés pour les différences précédant i.

L'effectif de d_1 est 7 donc $n_{i1} = 7$

Alors $r(d_i)$ le rang moyen de d_i est $n_{i1} + \frac{e(d_i) + 1}{2}$

TABLE DES VALEURS CRITIQUES DE T
DANS LE TEST DE **WILCOXON**

N	niveau de signification (test unilatéral)		
	.025	.01	.005
	niveau de signification (test bilatéral)		
	.05	.02	.01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

FICHE 16

K échantillons appariés, classés selon UNE variable à P valeurs, et UNE variable observée ordinale, Epreuve de Friedman

(Analyse de la variance par rangs de dimension 2)

L'épreuve de Friedman permet de répondre à la question suivante:

k échantillons appariés selon une variable à p valeurs, (et comprenant par conséquent p éléments) pour les éléments desquels on recueille l'observation d'une variable ordinale, sont-ils issus d'une même population parente? Il s'agit donc d'un test d'homogénéité.

Les échantillons sont constitués en fonction d'une seconde variable, il est donc possible alors de dire si cette variable interagit avec la première, ou influence la première.C'est pourquoi on compare ce test à l' "analyse à deux variables de la variance", toutefois ici il s'agit d'analyser des rangs.

Exemple:

On veut étudier le temps mis pour exécuter une tâche par les sujets de 3 groupes, placés dans 5 conditions expérimentales différentes. Chaque groupe comprend 5 sujets affectés au hasard aux conditions I, II, III, IV, V. On obtient ainsi le tableau suivant.

	I	II	III	IV	V
GA	3mn	2mn30	4mn30	2mn	5mn
GB	2mn30	1mn	1mn30	2mn	3mn
GC	4mn	5mn30	7mn	3mn	6mn

1) Présentation des données

Les données se présentent sous forme de tableau de k lignes, chacune pour un échantillon, et de p colonnes, chacune correspondant à une des valeurs de la variable d'appariement (ordinale ou non) Il y a une observation par case du tableau. Ces observations consistent en des rangements dans chaque échantillon (dans chaque ligne), ou en des valeurs numériques dont on ne retient que l'ordre dans chaque échantillon (transformation en rangs).

Exemple 1. Dans chaque échantillons les élèves sont rangés dans l'ordre de rapidité de l'exécution: la grandeur des différences de temps sera donc négligée.

On obtient alors le tableau ci dessous:

	I	II	III	IV	V
GA	3	2	4	1	5
GB	4	1	2	3	5
GC	2	3	5	1	4

2) Hypothèse nulle.

Si les rangs des sujets ne dépendent pas des conditions expérimentales, et si les sujets réagissent de façon différente aux différentes conditions, on peut s'attendre à ce que les rangs se répartissent de façon aléatoire dans les différentes colonnes, avec une tendance à obtenir la même

fréquence des différents rangs dans chaque colonne et des fréquences différentes entre colonnes sur une même ligne.

Si H_1:les trois échantillons sont issus du même ensemble parent.

ET si H_2: il y a des différences entre les effets des conditions expérimentales,

ces différences agiront de la même façon sur tous les échantillons: les rangs auront tendance à se retrouver dans le même ordre au sein de chaque échantillon. Il y aura par conséquent des disparités entre les colonnes.

Par conséquent, s'il n'y a pas de disparités entre les colonnes c'est que soit les conditions expérimentales n'agissent pas sur les sujets, soit qu'elles agissent de façon différenciée sur les différents sujets (et donc que les échantillons ne sont pas homogènes), soit les deux.

La meilleure hypothèse nulle est:

H_0: Les rangs sont distribués de façon homogène entre les colonnes.

En effet, si cette hypothèse est rejetée, alors on peut conclure qu'il y a des disparités, et donc les conditions expérimentales agissent, et qu'elles agissent de la même façon sur tous les échantillons: on prouve à la fois les deux thèses intéressantes.

3) Modèle.

Comment représenter l'équirépartition des rangs? Nous avons vu avec l'épreuve de Mann et Whitney que la somme des rangs est un indicateur intéressant. On considèrera la somme des rangs que l'on obtiendrait dans chaque colonne comme les valeurs théoriques du modèle.

Pour déterminer cette valeur théorique on calculerait la somme des rangs dans une ligne:

$$s_l = 1 + 2 + ... + p = \frac{p \times (p+1)}{2}$$

cette somme est la même pour chaque ligne, le total de tous les rangs du tableau sera donc:

$$S = K \times s_1 = \frac{K \times p \times (p+1)}{2}$$

La somme dans chaque colonne sous l'hypothèse nulle sera donc:

$$S = \frac{K \times (p+1)}{2}$$

Dans notre exemple, $S = \frac{3 \times (5+1)}{2} = 9$

Si on reprenait l'idée de la distance du X^2:

chaque colonne contribuerait à cette distance par le carré de la différence entre la somme observée des rangs de la colonne R_j et sa valeur théorique S, divisée par cette valeur théorique.

4) Distance de la contingence au modèle.

$$\chi^2 = \sum_j \frac{(R_j - S)^2}{S} = \frac{2 \times \sum_j (R_j^2 + S^2 - 2\,S.R_j)}{S}$$

En fait Friedman a proposé un coefficient voisin:

$$\chi^2_r = \frac{12 \times \sum_j (R_j)^2}{K.\,p.(p+1)} - 3\,K.(p+1)$$

dans lequel K = nombre de lignes
p = nombre de colonnes
R_j = somme des rangs de jème colonne
$\Sigma(R_j)^2$ = la somme des carrés de la somme des rangs étendue à toutes les colonnes

Dans l'exemple 1 nous trouvons les résultats suivants:

	I	II	III	IV	V
R_j	9	6	11	5	14
R_j^2	81	36	121	25	196

$\Sigma R_j^2 = 459$ $K.(p+1) = 18$ $K.p.(p+1) = 90$

$$\chi^2_r = \frac{(12 \times 459)}{90} - 3 \times 18 = 61{,}2 - 54 = 7{,}2$$

5) Distribution de cette distance sous l'hypothèse nulle et décision.

χ^2_r est approximativement distribué comme χ^2 avec
dl = p-1 dès que p > 4 ou K > 9
Dans le cas contraire il faut consulter la table de Friedman. Pour p = 1 le test est évidemment inapplicable.

exemple.
Puisque p > 4, consultons la table du χ^2 classique avec dl = p -1 = 4.
On trouve sur la ligne 4 que dans ces conditions le χ^2_r devrait dépasser 9,49 pour être significatif au seuil de 0,05, or il n'est que 7,2.

Remarque. Consultons la table de Friedman pour K = 3.
Nous ne trouvons pas de table pour la valeur p = 5. Examinons la table pour p = 4 qui donne une meilleure sécurité.
La probabilité seuil de 0,05 n'est pas portée dans la colonne K = 3 . En face de 0,053, nous trouvons $\chi^2_r = 7{,}0$ qui signifie que sous l'hypothèse nulle, l'on ne rencontrera des χ^2 supérieurs à 7 que dans moins de 5 % des cas.
La ligne au-dessous indique que les χ^2 supérieurs à 7,4 sont plus rares que 0,033 (3,3%). mais ce sont les résultats pour p=4 et non pour p = 5

6) Décision

La valeur observée de χ^2_r est elle plus rare que 5 % sous les conditions de l'hypothèse nulle? Comparons cette valeur avec la distribution du χ^2 ou avec la table spéciale.
Comme d'habitude si:
$\chi^2_{\text{seuil } 0{,}05} < \chi^2_r$, l'hypothèse nulle sera rejetée.

Or dans l'exemple 1, d'après la table du χ^2 la valeur observée χ^2_r n'atteint pas la valeur seuil. L'hypothèse nulle ne peut pas être rejetée.
Les différentes tâches n'agissent pas de façon suffisamment nette et la distribution des rangs dans les différentes colonnes pourrait être homogène.
Le lecteur trouvera l'exemple 2 dans le devoir n°3 présenté après la fiche 20.

Remarques.

1. A la suite de la transformation en rangs le tableau permet parfois de déterminer un ordre, qui n'était pas défini a priori, entre les valeurs de la variable portée en colonnes.

Ainsi dans l'exemple 1 on peut ordonner les tâches par ordre croissant de difficultés:

IV, II, I, III, et V

2. Remarquons aussi que le passage des valeurs numériques aux rangs, dans les lignes, fait disparaître les différences de scores. Ainsi, les groupes GA, GB, GC dans notre exemple sont visiblement différents du point de vue des temps moyens que mettent leurs membres pour exécuter les différentes tâches. mais cette différence-là n'est pas la cause du rejet de l'hypothèse nulle. Le tableau des rangs aurait pu être le même en partant d'un tableau des temps où tous les groupes auraient eu la même moyenne. Le test de Friedman éprouve donc seulement si la variable des colonnes modifie l'ordre de difficulté des taches (et non la difficulté elle-même), suivant les groupes.

3. Autres interprétations du test de Friedman:

UN échantillon, deux variables à k et p valeurs et une VARIABLE observée ordinale.

n éléments sont rangés selon une variable ordinale, ils sont classés selon deux variables, respectivement à p et k valeurs (alors n = pxk).

Les deux variables sont elles indépendantes? Il s'agirait alors d'une sorte de test de Kruskal et Wallis de dimension 2.

Enfin l'épreuve de Friedman a un usage similaire à l'épreuve du W de Kendall que nous étudierons plus loin.

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS AUSSI GRANDES QUE LES VALEURS OBSERVEES DU χ^2_r DANS L'ANALYSE A DEUX VARIABLES DE LA VARIANCE DE **FRIEDMAN**

Table 1: p = 3

K = 2		K = 3		K = 4		K = 5	
χ^2_r	proba	χ^2_r	proba	χ^2_r	Proba	χ^2_r	proba
0	1.000	.000	1.000	.0	1.000	.0	1.000
1	.833	.667	.944	.5	.931	.4	.954
3	.500	2.000	.528	1.5	.653	1.2	.691
4	.167	2.667	.361	2.0	.431	1.6	.522
		4.667	.194	3.5	.273	2.8	.367
		6.000	.028	4.5	.125	3.6	.182
				6.0	.069	4.8	.124
				6.5	.042	5.2	.093
				8.0	.0046	6.4	.039
						7.6	.024
						8.4	.0085
						10.0	.00077

K = 6		K = 7		K = 8		K = 9	
χ^2_r	Proba	χ^2_r	Proba	χ^2_r	Proba	χ^2_r	Proba
.00	1.000	.000	1.000	.00	1.000	.000	1.000
.33	.956	.286	.964	.25	.967	.222	.971
1.00	.740	.857	.768	.75	.794	.667	.814
1.33	.570	1.143	.620	1.00	.654	.889	.865
2.33	.430	2.000	.486	1.75	.531	1.556	.569
3.00	.252	2.571	.305	2.25	.355	2.000	.398
4.00	.184	3.429	.237	3.00	.285	2.667	.328
4.33	.142	3.714	.192	3.25	.236	2.889	.278
5.33	.072	4.571	.112	4.00	.149	3.556	.187
6.33	.052	5.429	.085	4.75	.120	4.222	.154
7.00	.029	6.000	.052	5.25	.079	4.667	.107
8.33	.012	7.143	.027	6.25	.047	5.556	.069
9.00	.0081	7.714	.021	6.75	.038	6.000	.057
9.33	.0055	8.000	.016	7.00	.030	6.222	.048
10.33	.0017	8.857	.0084	7.75	.018	6.889	.031
12.00	.00013	10.286	.0036	9.00	.0099	8.000	.019
		10.571	.0027	9.25	.0080	8.222	.016
		11.143	.0012	9.75	.0048	8.667	.010
		12.286	.00032	10.75	.0024	9.556	.0060
		14.000	.000021	12.00	.0011	10.667	.0035
				12.25	.00086	10.889	.0029
				13.00	.00026	11.556	.0013
				14.25	.000061	12.667	.00066
				16.00	.0000036	13.556	.00035
						14.000	.00020
						14.222	.000097
						14.889	.000054
						16.222	.000011
						18.000	.0000006

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS AUSSI GRANDES QUE LES VALEURS OBSERVEES DU χ^2_r DANS L'ANALYSE A DEUX VARIABLES DE LA VARIANCE DE **FRIEDMAN**

Table 2: p = 4

K= 2		K= 3		K= 4			
χ^2_r	proba	χ^2_r	proba	χ^2_r	Proba	χ^2_r	proba
.0	1.000	.2	1.000	.0	1.000	5.7	.141
.6	.958	.6	.958	.3	992	6.0	.105
1.2	.834	1.0	.910	.6	.928	6.3	.094
1.8	.792	1.8	.727	.9	.900	6.6	.077
2.4	.625	2.2	.608	1.2	.800	6.9	.068
3.0	.542	2.6	.524	1.5	.754	7.2	.054
3.6	.458	3.4	.446	1.8	.677	7.5	.052
4.2	.375	3.8	.342	2.1	.649	7.8	.036
4.8	.208	4.2	.300	2.4	.524	8.1	.033
5.4	.167	5.0	.207	2.7	.508	8.4	.019
6.0	.042	5.4	.175	3.0	.432	8.7	.014
		5.8	.148	3.3	.389	9.3	.012
		6.6	.075	3.6	.355	9.6	.0069
		7.0	.054	3.9	.324	9.9	.0062
		7.4	.033	4.5	.242	10.2	.0027
		8.2	.017	4.8	.200	10.8	.0016
		9.0	.0017	5.1	.190	11.1	.00094
				5.4	.158	12.0	.000072

C. VARIABLES ORDINALES

IV. DEUX VARIABLES OU PLUS : INDEPENDANCE ET CORRELATIONS

FICHE 17

UN échantillon, K variables ordinales
Test W de Kendall
Epreuve de concordance. Méthode des juges.

Elle indique dans quelle mesure k rangements d'un même ensemble sont concordants.

Elle permet de dire si k observations ordonnées peuvent être issues d'une même variable ordinale, si k observations de ce que les observateurs croient être UNE MEME variable ordinale diffèrent les unes des autres?

Le test est utilisable dans le cas où l'on veut savoir si k "juges" donnent bien le même sens à une variable ordinale en rangeant un ensemble suivant cette variable ou si les rangs attribués aux éléments le sont "au hasard".

Le rangement peut exprimer aussi une variable numérique (ou d'intervalle) dont on ignore la distribution parente comme dans les exemples donnés dans les chapitres précédents: U de Mann et Whitney ou H de Kruskal et Wallis.

Exemple (1)

Un formateur se demande si, chez les enseignants qui ont suivi son cours, la notion de "problème ouvert" correspond bien à une idée commune et exploitable (il ne veut pas pour l'instant vérifier si elle est correcte ni si elle est utilisable): il leur demande de ranger une famille de cinq problèmes semblables "du plus fermé au plus ouvert".

Exemple (2)

Un enseignant demande à ses 25 élèves de classer 24 énoncés "du plus facile au plus difficile" comme dans le Devoir n° 2.

1) Recueil des données.

Un ensemble de k juges range un ensemble de N entités. Exemple 1. N= 5 entités, les problèmes: {a, b, c, d, e}. k=4, les juges: {juge1, juge2, etc}.

Le résultat de leur travail peut se présenter ainsi :

	a	b	c	d	e__
Juge1	2	1	4	3	5__
Juge2	3	2	1	5	4__
Juge3	2	1	4	5	3__
Juge4	2	3	1	5	4__

Le calcul de la somme des rangs Rj obtenus par chaque problème…

R_j	9	7	10	18	16

permet de les ordonner autrement :

	b	a	c	e	d
Juge1	1	2	4	5	3
Juge2	2	3	1	4	5
Juge3	1	2	4	3	5
Juge4	3	2	1	4	5
R_j	7	9	10	16	18

2) Hypothèse nulle.

"Les enseignants n'ont pas une conception homogène de ce qu'est un problème ouvert. Ils ont donc rangé les énoncés de façon non concordante". "Les élèves-juges ont rangé les énoncés au hasard. Chacun des problèmes a les mêmes chances d'avoir n'importe quelle place avec chaque juge"

L'hypothèse nulle: le rangement au hasard, est une des négations de l'hypothèse: "les jugements sont concordants".

3) Modèle.

Sous l'hypothèse nulle, les sommes des rangs devraient être les mêmes dans chaque colonne. Elles devraient donc être égales à la somme de tous les rangs divisé par le nombre de colonnes: $\Sigma R_j / N$.

Pour chaque entité jugée, le carré de la distance (euclidienne) entre la somme des rangs observée et cette valeur théorique sous l'hypothèse d'indépendance, est: $(Rj - \frac{\Sigma R_j}{N})^2$

C'est la contribution de cette entité à la distance

- entre le modèle:

Modèle	$\frac{\Sigma R_j}{N}$	$\frac{\Sigma R_j}{N}$	$\frac{\Sigma R_j}{N}$	$\frac{\Sigma R_j}{N}$	$\frac{\Sigma R_j}{N}$

- et la contingence:

Obser.	ΣR_1	ΣR_2	ΣR_3	ΣR_4	ΣR_5

$\Sigma_j (R_j - \frac{\Sigma R_i}{N})^2$ représente donc la distance totale (son carré) entre le modèle, la discordance totale, et la contingence. W exprime le rapport entre cette distance observée et la valeur la plus grande qu'elle peut prendre. Donc plus W est grand, plus le tableau se rapproche de la concordance parfaite, ainsi que nous le montrons en annexe de cette fiche (en général la concordance n'est pas la négation de "au hasard").

4) W Le coefficient de concordance de KENDALL

La "concordance" entre les juges est "mesurée" par le coefficient W:

$$W = \frac{\Sigma_j (R_j - \frac{\Sigma R_j}{N})^2}{\frac{k^2}{12} \times N \times (N^2-1)}$$

Dans l'exemple 1 la somme de ces rangs est $\Sigma R_j = 60$.
Appelons $\alpha = \Sigma R_j$, alors:

$R_i - \alpha/N$	-5	-3	-2	4	6

et en posant $\beta = R_j - \alpha/N$, alors il vient

β^2	25	9	4	16	36

La somme de ces carrés est S, le numérateur de W
Dans cet exemple S = 90
Le dénominateur est $D = (1/12) \times k^2 \times N \times (N^2-1)$
Ici k = 4 N = 5 D = 160. Alors W = S/D et

$$\underline{W = 0{,}5625}$$

5) Cas des EX AEQUO.

S'il y a de nombreux ex-aequo la valeur calculée du W est artificiellement diminuée. Il faut appliquer la correction - k.Σ T suivante au dénominateur:

$$W = \frac{\Sigma_j (R_j - \frac{\Sigma R_j}{N})^2}{\frac{k^2}{12} \times N \times (N^2-1) - k.\Sigma T}$$

Dans cette formule

$$\Sigma T = \frac{\Sigma_i (t^2 - t)}{12}$$

où t est le nombre d'ex-aequo déterminés par un juge et $\Sigma_i T$ la somme des termes $(t^2 - t)$ relatifs à chaque juge.

6) Rareté des valeurs de W observées.

a) grand échantillon (N > 7)
On démontre que sous l'hypothèse nulle, et à la condition que N > 7 (au moins égal à 8) le coefficient χ^2:

$$\chi^2 = \frac{\Sigma_j (R_j - \frac{\Sigma R_j}{N})^2}{\frac{k^2}{12} \times N \times (N + 1)}$$

est sensiblement distribué comme un CHI CARRE avec un degré de liberté: dl = N-1. Il convient dans ce cas de se rapporter à la table du CHI².
Attention! le dénominateur est différent de celui de W, on a en fait: $\chi^2 = k.(N-1).W$)

b) petit échantillon N < 8

Se rapporter à la table donnée en fin de fiche. La valeur à porter est S.

Dans le cas de l'exemple 1, on trouve dans cette table, pour k = 4 et N = 5, la valeur du CHI2, χ^2_S, , qui, sous l'hypothèse nulle, n'est dépassée que dans 5% des cas: c'est 88,4.

Or la valeur de S observée est 90. Cette valeur, est donc dépassée par moins de 5% des cas. Elle est rare, et donc dans ce cas, grande.

7) Décision

Dans le cas de l'exemple 1, au seuil de 5%, il y a lieu de rejeter l'hypothèse nulle (Par contre au seuil de 1% on ne pourrait pas le faire, puisque 109,3 > 90).

Dans l'épreuve (1), à propos de l'ensemble des énoncés proposés, les enseignants ont un usage relativement concordant de l'expression "problème ouvert"(w= 0,56). Ils n'emploient pas ce terme au hasard.

8) Usage dans une recherche.

Cette conclusion autorise à entreprendre

- l'étude de la pertinence de l'emploi de ce terme si on en possède une norme,
- ou la recherche d'un modèle de cet emploi: un certain nombre de caractères des problèmes en question se traduiront par des variables nominales ou ordonnées et on retiendra celles qui sont liées à l'ordre résultant du rangement des juges. Le modèle du jugement de ces juges sera satisfaisant lorsqu'il permettra de proposer A PRIORI un ordre pour un ensemble de nouveaux problèmes, dont on constatera A POSTERIORI qu'il coïncide avec celui établi par les juges.

Annexe

CALCUL DU DENOMINATEUR DU W DE KENDALL

Calculons S_M la valeur la plus grande que peut prendre la distance observée entre le modèle et les observations. Ce sera le dénominateur du W.

Cette valeur maximum est obtenue lorsque tous les juges donnent le même rangement. La somme des rangs est alors k pour la première entité, k x 2 pour la seconde, 3 x k pour la troisième etc.

La somme pour toutes les entités est

$\Sigma R_j = k + 2k + 3k + ... + Nk = k(1+2+3+...+N)$

$\Sigma R_j = 1/2.k.N.(N+1)$. Donc la moyenne est:

$\frac{\Sigma R_j}{N} = \frac{k.(N+1)}{2}$ et la distance au rang moyen est:

$(R_j - \frac{\Sigma Rj}{N})^2 = R_j^2 + (\frac{\Sigma Rj}{N})^2 - 2R_j.(\frac{\Sigma Rj}{N}) = R_j^2 + 1/4.k^2.(N+1)^2 - k.(N+1).R_j$

La distance totale est: $S_M = \Sigma_j (R_j - \frac{\Sigma Rj}{N})^2$

$S_M = \Sigma_j (R_j)^2 + 1/4.k^2.N.(N+1)^2 - k.(N+1).\Sigma_j (R_j)$

Et puisque $\Sigma_j (R_j^2) = N.(N+1).(2N+1)/6$

et que $\Sigma_j (R_j) = N(N+1)/2$

alors il vient:

$S_M = N.(N+1).(2N+1)/6 + 1/4.k^2.N.(N+1)^2 - k.(N+1).N(N+1)/2$

d'où: $S_M = \frac{k^2 \times N \times (N^2-1)}{12}$

C'est bien la valeur du dénominateur de W. Et W varie donc ainsi de 0 à 1.

TABLE DES VALEURS CRITIQUES DE S DANS LE COEFFICIENT DE CONCORDANCE DE **KENDALL**.

k	N					valeurs additionnelles pour N= 3	
	3 *	4	5	6	7	k	s
Valeurs à .05 de signification							
3			64.4	103.9	157.3	9	54.0
4		49.5	88.4	143.3	217.0	12	71.9
5		62.6	112.3	182.4	276.2	14	83.8
6		75.7	136.1	221.4	335.2	16	95.8
8	48.1	101.7	183.7	299.0	453.1	18	107.7
10	60.0	127.8	231.2	376.7	571.0		
15	89.8	192.9	349.8	570.5	864.9		
20	119.7	258.0	468.5	764.4	1158.7		
Valeurs à .01 de signification							
3			75.6	122.8	185.6	9	75.9
4		61.4	109.3	176.2	265.0	12	103.5
5		80.5	142.8	229.4	343.8	14	121.9
6		99.5	176.1	282.4	422.6	16	140.2
8	66.8	137.4	242.7	388.3	579.9	18	158.6
10	85.1	175.3	309.1	494.0	737.0		
15	131.0	269.8	475.2	758.2	1129.5		
20	177.0	364.2	641.2	1022.2	1521.9		

* Les valeurs critiques supplémentaires de s pour N = 3 sont données dans la colonne de droite ce cette table.

UN échantillon, DEUX variables ordinales COEFFICIENTS DE CORRELATION, Rhô de SPEARMAN, TAU de KENDALL

Les coefficients Rhô de Spearman et Tau de Kendall indiquent dans quelle mesure 2 rangements d'un même ensemble sont concordants. Dans quelle mesure deux variables sont liées l'une à l'autre.

Ils permettent de dire si deux ensembles d'observations ordonnées relatives à un même échantillon peuvent être issus d'une même variable ordinale. Ou encore si 2 observations de ce que les observateurs croient être UNE MEME variable ordinale diffèrent l'une de l'autre?

Ils jouent donc le même rôle que le W de Kendall mais dans le cas particulier où k = 2.

Exemple:
Deux correcteurs rangent une même série de devoirs dans l'ordre décroissant de valeur.
1er Rangement: variable X.
{b, i}(ex aequo), c, {a,c,d}, {h, j}, g, f...
2ème rangement: variable Y.
d, {e, i}, {a, c}, h, b, {g, j}, f

1) Recueil et organisation des données

Les individus sont rangés dans un ordre quelconque et leur rang selon les deux variables sont notés suivant la méthode habituelle:

variables échantillon	X	Y	d	d^2
a	5	4,50	0,5	0,25
b	1,5	7	-5,5	30,25
c	5	4,5	0,5	0,25
d	5	1	4	16
e	3	2,5	0,5	0,25
f	10	10	0	0
g	9	8,5	0,5	0,25
h	7,5	6	1,5	2,25
i	1,5	2,5	-1	1
j	7,5	8,5	-1	1

$N = 10$ $\Sigma d = 0$ $\Sigma d^2 = 51,5$

2) Hypothèse nulle

Deux variables sont corrélées lorsque la connaissance de l'une permet de diminuer l'incertitude sur la valeur de l'autre. Elles ne sont pas indépendantes. L'hypothèse nulle devrait consister à affirmer l'indépendance des variables: les grandes valeurs de Y peuvent être observées aussi bien sur les petites valeurs de X que sur les grandes.

Or contrairement au cas du χ^2 nous ne disposons pas ici d'un modèle de l'indépendance. Toute correspondance entre deux variables pourrait être le résultat d'une relation de dépendance entre elles, éventuellement tourmentée, mais précise. Lorsque deux variables ne sont pas indépendantes, on peut aussi essayer d'exprimer que l'une dépend de l'autre selon une loi mathématique simple mais avec une certain incertitude, une erreur.

Nous sommes donc réduits à envisager les cas de dépendance les plus fréquents ou les plus plausibles et à essayer de les éliminer. Les hypothèses nulles seront donc des hypothèses de dépendance.

Exercice:

Etablir le diagramme de contingence des variables X et Y de l'exemple ci dessus, en représentant les individus par des points, placés dans un plan cartésien, suivant leur rang selon les deux variables. Comment se manifesterait une corrélation étroite? une indépendance?

3) Modèle

Les relations (non constantes) de dépendances les plus simples entre deux variables X et Y sont les relations linéaires et les relations affines: les valeurs (resp. leurs différences) de X sont proportionnelles aux valeurs correspondantes de Y (resp. à leurs différences).

Cette dépendance et la façon de s'en écarter sont représentés par le coefficient de corrélation linéaire de BRAVAIS-PEARSON dont nous rappelons en annexe la justification.

$$r_S = \frac{\sum_{i,j} x_i.y_j}{\sqrt{\sum_i x_i^2 . \sum_j y_j^2}} \quad \text{en abrégé} \quad \frac{\Sigma xy}{\sqrt{(\Sigma x^2 . \Sigma y^2)}}$$

Lorsque les variables sont des rangs, ce coefficient peut être interprété par la formule suivante,

$$r_S = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma_i d_i^2}{2.\sqrt{(\Sigma x^2 . \Sigma y^2)}}$$

et finalement simplifié (s'il n'y a pas trop d'ex-aequos) ainsi que nous le montrons aussi en annexe, en la Formule du Rhô de SPEARMAN:

$$\boxed{r_S = 1 - \frac{6.\sum_i d_i^2}{(N^3-N)}}$$

Ce modèle exprime que rs est d'autant plus proche de 1 que la somme des carrés des différences entre les rangs selon l'une et l'autre variable est petite. En cas de rangement dans le même ordre, rs est égal à 1

Remarquons que le fait d'exclure cette dépendance n'implique pas l'indépendance. Celle ci peut seulement être acceptée (à défaut d'autre contre exemple ou hypothèse.

Alors dans l'exemple ci-dessus il vient: $r_s = 1 - \frac{(6 \times 51,5)}{990} = 0,688$

4. Cas des ex aequo.

Il faut dans ce cas retrancher des termes correcteurs:

- $T_X = \frac{(t^3 - t)}{12}$ dans lequel t est le nombre d'ex-aequo de X pour le rang donné.

– ΣT_X est la somme des T_X pour la variable X

on calcule de même $T_Y = \frac{(t^3 - t)}{12}$ puis ΣT_Y pour la variable Y.

il convient alors de revenir à la formule initiale de BRAVAIS PEARSON et de retrancher ces corrections A CHACUN DES TERMES Σx^2, et Σy^2 dans la formule:

$$r_S = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma_i d_i^2}{2.\sqrt{(\Sigma x^2.\Sigma y^2)}}$$

où $\Sigma x^2 = (N^3 - N) - \Sigma T_X$ et $\Sigma y^2 = (N^3 - N) - \Sigma T_Y$

5. Significativité (rareté) du COEFFICIENT de CORRELATION rs.

Cette valeur du rhô est elle grande? suffit elle pour conclure qu'il existe une liaison entre le jugement du premier professeur et celui du second? Sans cet élément d'appréciation, cet indice ne nous apprendrait pas grand chose.

a) HYPOTHESE NULLE: Les deux rangements sont indépendants, faits "au hasard".

b) DISTRIBUTION: Sous cette hypothèse on peut connaître avec quelle fréquence le coefficient de Spearman s'écarte de sa valeur moyenne (0). (en référence à la distribution du coefficient de Bravais-Pearson)

En fait dès que N > 10 : $$t_R = r_s\sqrt{\frac{(N - 2)}{1 - r_s^2}}$$

est distribué comme le t de student avec dl = N - 2 degrés de liberté.

c) RARETE. Dans la table des valeurs significatives de r_S (annexe de cette fiche), lire pour le seuil choisi (par exemple p = 0,05) et pour l'effectif de l'échantillon d'observations, la valeur au dessus de laquelle le Rhô calculé peut être considéré comme trop grand pour que l'on puisse accepter l'hypothèse nulle.

d) *Exemple.*
Dans notre exemple, $r_{observé} = 0,68$. N =10, p=0,05, $r_{seuil} = 0,564$
On pourrait utiliser la table du r de Bravais Pearson (annexe)
N = 10, p = 0,02, $r_{seuil} = 0,66$. $r_{observé} > r_{seuil2\%}$

L'hypothèse nulle ne peut être retenue, les deux professeurs n'apprécient pas les devoirs de manière indépendante. Cela ne veut-il dire que leurs jugements soient concordants?

Que se passerait-il s'ils les avaient rangés dans un ordre inverse l'un de l'autre par exemple?
La somme des carrés des différences serait très grande, Rs serait exceptionnellement voisin de zéro et nous devrions apprécier sa rareté avec l'autre extrémité de la distribution, qui n'est pas donnée dans la table. Il faudrait alors par exemple tester une variable Y' exactement inverse de Y. Ainsi notre mode de calcul, s'il conduit au rejet de l'indépendance implique la concordance.

TABLE DES VALEURS CRITIQUES DE $\mathbf{r_s}$
LE COEFFICIENT DE CORRELATION DE RANG DE **SPEARMAN**

N	Niveau de signification (test unilatéral)	
	.05	.01
4	1.000	
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

UN échantillon, DEUX variables ordinales
Coefficient de corrélation: Tau de KENDALL

Comme le coefficient Rhô de Spearman présenté dans la fiche 19, le Tau de Kendall indique dans quelle mesure deux rangements d'un même ensemble sont concordants. Dans quelle mesure deux variables sont liées l'une à l'autre.

Deux ensembles d'observations ordonnées relatives à un même échantillon peuvent-ils être issus d'une même variable ordinale. Ou encore deux observations de ce que les observateurs croient être UNE MEME variable ordinale, diffèrent l'une de l'autre?

1) Organisation des données

Disposition pour le calcul du TAU de Kendall:

Trois colonnes:

-les individus, qui peuvent être mis dans un ordre quelconque ou rangés selon l'une ou l'autre des deux variables,

- les rangs selon la première variable,
- les rangs selon l'autre.

Exemple:

Pour évaluer l'impact du contrat didactique sur le jugement d'un élève, on lui demande d'effectuer le rangement d'un ensemble de déclarations douteuses, par ordre décroissant de validité, AVANT (variable X) puis APRES (variable Y) une intervention didactique qui semble évoquer ces déclarations, mais ne contient en fait aucune information directe à leur propos. Il s'agit de savoir dans quelle mesure Y diffère de X.

Les individus (en fait des déclarations) sont rangés dans l'ordre de la variable X.

	X	Y
b	1,5	7
i	1,5	2,5
e	3	2,5
a	5	4,5
c	5	4,5
d	5	1
h	7,5	6
j	7,5	8,5
g	9	8,5
f	10	10

2) Hypothèse nulle

Comme nous l'avons indiqué dans les fiches précédentes, l'hypothèse nulle peut être:

" Les deux rangements sont indépendants"

" Les deux rangements coïncident". Le modèle est celui de la dépendance linéaire, les écarts au modèle doivent alors être représentés par un modèle d'erreur à préciser, une distance par exemple, qui, si elle est trop grande, conduit à rejeter le modèle.

3) Modèle

Il s'agit d'abord de compter pour combien de couples d'individus le rangement d'un élément x selon X coïncide avec le rangement selon Y et pour combien de couples ou entre cet élément x les deux rangements divergent (sont inversés).

Ensuite de calculer la différence S(x) (algébrique) entre ces deux nombres, puis la somme, pour tous les éléments de X de toutes ces différences:

$$\Sigma S = \sum_{x \in X} S(x)$$

Le **Tau de Kendall** est le rapport de cette différence à sa valeur maximum:

$$T_K = \frac{2 \times \Sigma S}{N(N-1)}$$

Le nombre d'inversions correspond au nombre de dominations que nous avons présenté dans la fiche sur le U de MANN et WHITNEY (Le calcul en est rapplé ci-dessous).

Le nombre total des inversions possibles est $\frac{N(N-1)}{2}$;

4) CALCUL DE ΣS

Pour cela,

- On prend successivement les éléments, dans l'ordre du rangement selon X, et pour chacun (par exemple x) on calcule sa contribution S(x).

Pour calculer S(x) on considère successivement les couples qu'il forme avec tous les éléments j qui le suivent et on compte le nombre co(x) de concordances (x et j sont dans le même ordre selon X et selon Y) et le nombre inv(x) d'inversions (x et j sont dans un ordre différent selon X et selon Y).

Ces valeurs sont portées dans le tableau et permettent de calculer S(x) leur différence:

$$S(x) = co(x) - inv(x)$$

- On calcule alors ΣS, la somme de ces contributions pour tous les éléments de l'échantillon.

$$\Sigma S = \sum_{x} S(x) = \sum_{x} [co(x) - inv(x)]$$

(Remarque: on pourrait aussi définir $d_x S(j)$ égale à 1 s'il y a concordance et à -1 s'il y a inversion et

$$S(x) = \sum_{j} d_x S(j)$$

alors

$$\Sigma S = \sum_{x} \sum_{j} d_x S(j) \quad)$$

Exemple:

a) Calcul de S(b), contribution de l'élément b, (le premier selon la variable X) au coefficient S:

On calcule successivement inv(b), le nombre d'inversions de b selon Y, et co(b) le nombre de concordances de rangement.

On compare le rang de b, selon la variable Y (7):

- au rang selon Y , de l'élément i qui le suit selon l'ordre de X; le rang de i selon Y est 2,5. L'ordre entre b et i donné par Y: (7 > 2,5) , il est indifférent par rapport à celui de b et i donné par X (1,5 = 1,5), on le compte comme une inversion.

Alors $d_1S(b) = - 1$

- au rang de l'élément suivant e: (7 > 2,5) que l'on compare à (1,5 < 3): il y a inversion cette fois, alors $d_2S(b) = -1$;

- ainsi de suite on compte, jusqu'à l'élément h, 5 inversions: $d_3S(b) = d_4S(b) = d_5S(b) = d_6S(b) = -1$

donc inv(b) = - 6, car

$inv(b) = d_1S(b)+d_2S(b)+d_3S(b)+d_4S(b)+d_5S(b)+d_6S(b)$

On porte cette valeur dans le tableau en face de b, dans la colonne des inversions,

- puis pour les trois derniers éléments j, g, f, on trouve un ordre concordant: par exemple pour j: (7 > 8,5) dans Y et (1,5 < 7,5) dans X.

Donc $d_7S(b) = d_8S(b) = d_9S(b) = +1$

et $co(b) = d_7S(b) + d_8S(b) + d_9S(b) = 3$

Finalement S(b) = 3 - 6 = -3

Remarquons que

S(b) = Somme des $d_iS(b)$ = co(b) - inv(b)

et plus généralement

S(x) = Somme des $d_iS(X)$ = co(x) - inv(x)

= Nombre total de concordances - nombre total d'inversions.

b) Calcul de S(i) et des suivants.

On opère de même avec l'élément suivant, i tenant lieu de b et on le compare toujours avec tous les éléments qui le suivent selon l'ordre de X, mais pas avec ceux qui le précèdent.

Il y a seulement une inversion (avec d), 1 ex aequo et 6 concordances. S(i) = 5 etc.

c) Calcul de $\sum S$

$\sum S(x) = 23$, N = 10

	X	Y	inv(x)	co(x)	S(x)
b	1,5	7	-6	3	-3
i	1,5	2,5	-1	6	5
e	3	2,5	-1	6	5
a	5	4,5	-1	4	3
c	5	4,5	-1	4	3
d	5	1	0	4	4
h	7,5	6	0	3	3
j	7,5	8,5	0	2	2
g	9	8,5	0	1	1
f	10	10	0	0	0
Σ S = 33 - 10 = 26 - 3 = 23			10	33	23

d) Calcul du Tau de KENDALL

Dans l'exemple proposé, il vient:

$$T_K = \frac{2 \times \Sigma S}{N(N-1)} = \frac{2 \cdot 24}{90} = 0{,}53$$

où Σ S est calculé comme indiqué ci-dessus.

5) Cas des ex aequo

Comme dans la fiche 18, faut retrancher des termes correcteurs,
$\Sigma T_X = ½ \Sigma t(t-1)$ dans lequel t est le nombre d'ex-aequo pour chaque valeur de variable X;
$\Sigma T_Y = ½ \Sigma t(t-1)$ dans lequel t est le nombre d'ex-aequo pour chaque valeur de variable Y.
Ce qui donne:

$$T_K = \frac{\Sigma S}{\sqrt{\left(\frac{N.(N-1)}{2} - T_X\right).\left(\frac{N.(N-1)}{2} - T_Y\right)}}$$

Remarquons que cette valeur est différente du rhô de Spearman.

6) Significativité du coefficient de corrélation TK (rareté)

Sous l'hypothèse nulle, tous les rangements sont également probables.

a) Grands Echantillons

Lorsque $N > 10$, Tau suit une loi normale de moyenne 0 et d'écart type $\sigma_T = \sqrt{\frac{2.(2.N + 5)}{9.N.(N - 1)}}$

donc la variable Z_t

$$Z_t = \frac{T_K}{\sqrt{\frac{2.(2.N + 5)}{9.N.(N - 1)}}}$$

est distribuée comme Z, la variable normale centrée réduite.

Dans l'exemple ci-dessus, $2(2N+5) = 50$; $9N(N-1) = 810$

$\frac{50}{810} = 0{,}0617$; $\sqrt{0{,}0617} = 0{,}248$ $Zt = \frac{0{,}51}{0{,}248} = 2{,}056$

On lit dans la table de la loi normale qu'une valeur égale ou supérieure à cette valeur n'apparait, sous l'hypothèse nulle, qu'avec une probabilité inférieure à 2 %.: trop rarement! L'hypothèse nulle est donc rejetée.

Remarquons que si les valeurs du Rhô et du Tau sont différentes, leur rareté est la même: 2%.

b) Petits Echantillons.

Dans le cas où $N \leq 10$, se rapporter à la table donnée en annexe de cette fiche, en utilisant directement la valeur de ΣS (ce qui évite le calcul effectif de Tau).

Dans notre exemple, Σ S = 23 dans la colonne 10 sur la ligne 23 on lit 0,023. La fréquence avec laquelle on pourra rencontrer, sous l'hypothèse nulle, des valeurs de Σ S plus grandes que 23 est (toujours) 2,3%.

TABLE DES PROBABILITES ASSOCIEES AUX VALEURS
AUSSI GRANDES QUE LES VALEURS OBSERVEES DE ΣS
DANS LE COEFFICIENT DE **CORRELATION DE RANG DE KENDALL**

ΣS	Valeurs de N				ΣS	Valeurs de N		
	4	5	8	9		6	7	10
0	.625	.592	.548	.540	1	.500	.500	.500
2	.375	.408	.452	.460	3	.360	.386	.431
4	.167	.242	.360	.381	5	.235	.281	.364
6	.042	.117	.274	.306	7	.136	191	.300
8		.042	.199	.238	9	.068	.119	.242
10		.0083	.138	.179	11	.028	.068	.190
12			.089	.130	13	.0083	.035	.146
14			.054	.090	15	.0014	.015	.108
16			.031	.060	17		.0054	.078
18			.016	.038	19		.0014	.054
20			.0071	.022	21		.00020	.036
22			0028	.012	23			.023
24			.00087	.0063	25			.014
26			.00019	.0029	27			.0083
28			000025	.0012	29			.0046
30				.00043	31			.0023
32				.00012	33			.0011
34				.000025	35			.00047
36				.0000028	37			.00018
					39			.000058
					41			.000015
					43			.0000028
					45			.00000028

FICHE 20

UN échantillon, DEUX variables ordinales et une variable de contrôle Coefficient de corrélation partielle, Tau de KENDALL

Ce test permet de vérifier une dépendance entre deux variables sous la condition qu'une troisième reste fixe.

Son utilisation est intéressante lorsque l'on rencontre une liaison entre deux variables X et Y qui paraissent étrangères l'une à l'autre. On soupçonne alors que cette liaison s'explique, non pas par une relation causale entre ces deux variables, mais par une liaison qu'elles auraient chacune avec une même troisième Z.

L'exemple célèbre illustrant cette remarque est la corrélation très élevée (supérieur à 0,95) qui existait dans les années trente entre le taux de natalité (Y) et le nombre des nids de cigognes (X) dans les communes d'Alsace. Cette corrélation s'expliquait comme vous le pensez non pas par un effet des cigognes sur les naissances mais par le fait d'une troisième variable Z: le caractère plus ou moins rural ou urbain des communes, qui agissait sur les deux autres variables

Exemple:

On étudie une relation possible entre la mémoire et la résolution de problèmes. On peut envisager a priori le fait qu'il existe (ou non) une liaison entre ces deux variables. Mais si on observe une telle liaison, on prévoit qu'elle pourrait être attribuée à une autre variable: le niveau général en mathématiques (ou l'intelligence ...).

Pour mettre en évidence un tel effet il faudrait limiter le test à des élèves de même niveau mathématique (ou de même Q.I...) et répéter l'expérience pour des niveaux différents.

Le coefficient de corrélation partielle procure une deuxième possibilité: repérer le niveau mathématique des élèves (variable de contrôle) et étudier les corrélations partielles.

Si lorsqu'on annule l'effet de la troisième variable la liaison entre les deux premières disparaît, c'est qu'elle était due aux liaisons de X et de Y avec Z. Si elle subsiste alors, soit il y a relation directe, soit il existe un autre facteur commun agissant et encore inconnu.

1) Recueil et présentation des données.

Les valeurs de trois variables (numériques) X', Y', Z', sont recueillies pour chaque sujet et portées sur la même ligne. Elles constituent ainsi un premier tableau de n lignes (pour n sujets) et de trois colonnes.

Transformation en rangs:

Chaque variable permet d'ordonner tous les sujets dans l'ordre croissant (ou décroissant). Chacun d'eux occupe un rang, ce rang est porté à la place correspondante (en regard du sujet et de la variable considérée) dans un deuxième tableau: celui des variables ordinales X, Y, Z. Dans chaque colonne on trouve, dans un ordre quelconque, la suite des mêmes nombres. Les sommes sont donc égales. On s'arrange pour ranger les sujets dans l'ordre de la variable Z. C'est toujours possible.

Nous retrouvons un tableau comparable à ceux que nous avons traités dans l'épreuve de Friedman et dans la méthode des juges.

Exemple simplifié: La variable X est le rang dans un test de mémoire spécifique, la variable Y est le rang dans une épreuve de résolution de problèmes, Z est le rang dans le niveau général en mathématiques. L'échantillon comprend 4 sujets; résultats:

Tableau I

variables / Sujets \	rangs X	rangs Y	rangs Z
a	3	2	1
b	1	1	2
c	2	3	3
d	4	4	4

2) Méthode

Examinons toutes les paires possibles de sujets (comme dans le test U de Mann et Whitney). Il y en a:

$m = C_2^n = \frac{n \times (n-1)}{2}$ dans notre exemple m = 6

Chaque paire qui, pour une variable, est rangée dans le même ordre que dans Z sera notée +. Elle sera notée - si elle est rangée dans l'ordre inverse.

Formons un tableau de m lignes sur 3 colonnes. Dans chaque colonne on met + pour chaque paire telle que le rang le plus faible précède le rang le plus élevé et moins dans le cas contraire.

Dans l'exemple ci-dessus on obtient:

Tableau II

Variables / Sujets \	rangsX	rangsY	rangsZ
(a,b)	-	-	+
(a,c)	-	+	+
(a,d)	+	+	+
(b,c)	+	+	+
(b,d)	+	+	+
(c,d)	+	+	+

Ce tableau permet de dresser le bilan des désaccords et des accords de X et de Y entre eux, suivant leur accord ou leur désaccord avec Z: on compte les effectifs de tous les patrons de réponses à X et à Y: il y en a quatre possibles:

(-,-) ; (-,+) ; (+,-) ; (+,+) .

Les effectifs de ces quatre patrons sont portés dans les cases du tableau 2x2 (tableau III) ci-dessous où "X : accord avec Z , désaccord avec Z" est croisé avec "Y : accord avec Z , désaccord avec Z"

Tableau III

		X		
		désaccord avec Z	accord avec Z	Σ
Y	accord avec Z	-,+ 1	+,+ 4	5
	désaccord " Z	-,- 1	+,- 0	1
	Σ	2	4	m = 6

On notera b le nombre de (+,+) c le nombre de (-,-)
a le nombre de (+,-) d le nombre de (-,+)

et suivant ce code, les valeurs du tableau III seront portées dans la formule de KENDALL. (lire $\tau_{XY.Z}$: "TAU de X et Y par rapport à Z".

$$\tau_{XY.Z} = \frac{b \times c \; - \; a \times d}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Dans l'exemple ci-dessus on trouverait:

$$\tau_{XY.Z} = \frac{4 \times 1 \; - \; 1 \times 0}{\sqrt{(0+4)(1+1)(0+1)(4+1)}} = \frac{4}{\sqrt{40}} = 0{,}63$$

3) Explication du modèle

Le numérateur de $\tau_{XY.Z}$ est la différence entre:

- le produit du nombre des accords concomitants de X et de Y par celui de leurs désaccords concomitants avec Z
- et le produit du nombre des accords de X avec Z, lorsque Y est en désaccord, par le produit du nombre des accords de Y avec Z, lorsque X est en désaccord.

Ce numérateur est d'autant plus grand que X et Y sont liés entre eux (et avec Z ou non) et que X et Y sont faiblement liés avec Z lorsqu'ils sont en désaccord entre eux.

La valeur maximum du numérateur est atteinte pour a et d nuls et $b = c = \frac{m}{2}$

Le numérateur est alors égal à b x c
Dans notre exemple, dans ce cas b = c = 3 le numérateur est 9.

La valeur minimum du numérateur est atteinte pour b = c = 0 et pour $a = d = \frac{m}{2}$.

Ici le numérateur prend alors la valeur -9

Le dénominateur est choisi pour donner au coefficient de corrélation la valeur 1 au maximum et la valeur -1 au minimum.

4) différents cas

Si $\tau_{XY.Z} = 1$ (ou est voisin de 1) les variables X et Y sont corrélées indépendamment de Z,

Si $\tau_{XY.Z} = -1$ (ou est voisin de -1) les variables X et Y sont chacune corrélées avec Z mais *indépendamment* l'une de l'autre et de ce fait si elles apparaissent comme corrélées, c'est par l'intermédiaire de Z.

Examinons dans l'exemple ci-dessus les différentes corrélations (tau de KENDALL):

- entre X et Y: $\tau_{XY} = 0{,}66$
- entre X et Z: $\tau_{XZ} = 0{,}33$
- entre Y et Z: $\tau_{YZ} = 0{,}66$

Les trois variables paraissent liées deux à deux. Examinons les coefficients de corrélation partielle:

$\tau_{XZ.Y} = - 0{,}22$

$\tau_{YZ.X} = 0{,}63$

Il devient clair que X et Z ne sont pas liées lorsqu'on élimine l'influence de Y, ou si elles l'étaient faiblement, cette corrélation serait négative. Par contre Y est liée à Z, indépendamment de X, et comme X est corrélée avec Y, Y "entraîne" X qui semble corrélée avec Z.

5) Signification (rareté) du $\tau_{XY.Z}$

La valeur 0,68 est elle grande? Avait-on beaucoup de chances de la rencontrer par hasard dans une hypothèse nulle intéressante?

La distribution du $\tau_{XY.Z}$ de Kendall sous l'hypothèse nulle n'était pas mathématiquement connue en 1956. Si elle ne l'était pas aujourd'hui, il serait possible de l'établir par une simulation d'épreuves répétées, après avoir traduit en modèle l'hypothèse nulle:

"H_0: X, Y, Z sont trois variables normales indépendantes dans leur ensemble".

6) Exercice

Etudier le tableau II à l'aide de l'épreuve du W de Kendall, puis du test de Friedman. Formuler les hypothèses qui correspondent à chacun et les comparer avec celle qui peut être avancée dans l'examen de la corrélation partielle.

C. VARIABLES ORDINALES

III . SUJET DE DIDACTIQUE DES MATHEMATIQUES

Devoir n° 3

24 élèves ont été répartis en trois groupes de même taille après un appariement en fonction de leurs résultats en mathématiques (par exemple chaque groupe est composé, pour chaque note, d'un nombre égal d'élèves ayant eu cette même note). Chaque groupe se prépare à une épreuve de mathématiques, à l'aide d'un rappel constitué de quelques exercices suivis d'un corrigé. Chaque rappel portera sur des notions différentes mais connues de l'élève.

Dans le premier groupe, le rappel consiste en exercices sur les propriétés des naturels (recherche d'un successeur, illustration d'un produit, et deux problèmes de divisions)

Le rappel du second groupe porte sur les propriétés des fractions et leurs opérations.

Le rappel du troisième groupe présente des questions de géométrie, sans rapport avec le sujet de l'épreuve finale.

L'épreuve finale comprend cinq questions sur les décimaux, leurs propriétés et leurs opérations.

L'observateur relève les résultats (juste ou faux) de chaque élève sur chaque question et obtient le tableau ci après.

Quelles sont les questions que ce tableau permet de poser? Résumer ce tableau de la manière appropriée à chacune des questions ou des méthodes envisagées.

	Nombre de réussites aprés "rappel" K = 8 p = 3		
Trios appariés	Ex. naturels	Ex. fractions	Ex. géométrie
a	2	1	3
b	2	3	1
c	2	0	2
d	4	2	4
e	5	4	3
f	5	3	4
g	3	3	4
h	3	1	2

Solution

1. Il est possible de répondre à la question suivante:

"le type de rappel a-t-il une influence sur la réussite des élèves? H_0: non.

a) Il n'est pas possible d'appliquer un test classique de χ^2 sur le tableau croisé car les valeurs théoriques sont inférieures à 5, mais il est possible d'effectuer un regroupement: le nombre des réussites s'écarte-t-il de la distribution uniforme (même nombre dans chaque condition)?

Naturels	fractions	géométrie
$\Sigma j = 26$	$\Sigma j = 17$	$\Sigma j = 23$
VT = 23	VT = 23	VT = 23

$\chi^2 = 9/23 + 9/23 < 1$ donc non significatif, l'H_0 ne peut pas être rejetée

b) Plus précisément, on peut alors poser la question

"les rangs des élèves dans leur triade sont ils répartis uniformément dans les trois types de rappels?". Hypothèse nulle: oui. Le test de Friedman permet de répondre:

	Rang dans le trio		
a	2	3	1
b	2	1	3
c	1,5	3	1,5
d	1,5	3	1,5
e	1	2,5	2,5
f	1	3	2
g	2,5	2,5	1
h	1	3	2
Rj	12,5	21	14,5
$(Rj)^2$	156,25	441	210,25
$\Sigma(Rj)^2$			807,5

$K = 8;\ p = 3$ alors $\chi^2_r = 12.\ 807{,}5 \cdot \frac{12}{8.3.(3+1)} - 3.8.(3+1) = 100{,}93 - 96 = 4{,}937$

Dans la partie p = 3 de la table et K = 8 on lit que la probabilité cherchée est comprise entre 0,079 et 0,120 c'est-à-dire entre 7,9% et 12%. Cette valeur est donc supérieure à la valeur seuil de 5% qui permettrait de rejeter l'hypothèse d'homogéïté.

Conclusion: les différences sont insuffisantes pour rejeter l'hypothèse nulle.

Remarquons toutefois que ce test montre que moins de 10% des valeurs obtenues sous l'hyopothèse d'une distribution homogène seraient aussi étranges: ce que ne montrait pas le test beaucoup moins puissant du χ^2 classique.

2. On peut aussi chercher à savoir si une seule des conditions favorise les bonnes réponses car on peut penser a-priori que tel ou tel rappel peut les favoriser:

- que les rappels sur les opérations dans les naturels préparent mieux les élèves aux opérations sur les décimaux, puisque "les calculs sont les mêmes", sauf la position de la virgule.
- Que les décimaux étant des rationnels, les rappels sur les fractions permettront de mieux comprendre - et donc effectuer- les calculs dans cette structure.
- Que les calculs étant de toute manière différents, il vaut mieux stimuler une activité intellectuelle générale et éviter de favoriser la recherche par les élèves d'intentions didactiques cachées, cette attitude étant génératrice d'erreur.

D. VARIABLES D'INTERVALLES ET VARIABLES NUMERIQUES

I. PLUSIEURS VARIABLES D'INTERVALLES: INDEPENDANCE ET CORRELATION

DEUX échantillons, UNE variable d'intervalles
Le Test du brouillage (Randomization test)

Ce test permet de dire si les valeurs observées dans deux petits échantillons peuvent être considérées comme distribuées au hasard dans les deux échantillons ou si au contraire l'un d'entre eux a tendance à rassembler les valeurs les plus petites sans qu'aucune hypothèse ne soit faite sur la distribution de cette variable. (ce test correspond au test t de student pour les variables numériques normalement distribuées).

1) Recueil des données.

Pour chacun des éléments de deux échantillons A et B, on relève la valeur d'une variable exprimée par un nombre.

éléments 1er échantillon A	e_1	e_2	e_3	e_4	...	e_k
valeurs de la variable	a_1	a_2	a_3	a_4	...	a_k

éléments 2ème échantillon B	g_1	g_2	g_3	g_4	...		g_l
valeurs de la variable	b_1	b_2	b_3	b_4	...		b_l

Cette valeur peut n'être pas exactement numérique en ce sens que toutes les opérations numériques et notamment la somme n'ont peut-être pas de signification mais la différence doit en avoir une. Par exemple un score n'est pas une mesure: Il n'y a pas d'épreuve qui corresponde à la somme de deux scores. Par contre la différence entre deux scores correspond à une série d'épreuves réussies dans un cas et non réussies dans l'autre; des différences égales correspondent à des épreuves considérées comme équivalentes, les différences peuvent être comparées; la somme des différences peut recevoir une signification... La différence des scores peut être traitée comme une variable numérique.

Exemple:

Un chercheur a posé à 196 élèves une série de problèmes d'algèbre comprenant des inégalités à résoudre ou à interpréter. Il observe leur tendance à traiter la variable comme un nombre déterminé - à lui attribuer une valeur précise -, ce qui les conduit à des erreurs qualifiées de "valorisation". Cette erreur est observée dans 12 des questions posées. Il s'agit de chercher un facteur qui dans ces douze questions favorise les erreurs par valorisation.

Voici les questions rangées suivant le nombre d'erreurs par valorisation que l'on a pu y détecter:

h	a	g	c	e	j	k	i	f	b	d	l
4	9	15	25	27	27	28	30	40	43	55	63

Ce chercheur formule une première hypothèse: la présence exclusive dans les inégalités de valeurs numériques entières favorise la valorisation.

Parmi les 12 questions, 3 présentent des valeurs non entières: les questions h, g, k.

Ainsi nous constituons deux échantillons de questions. A: les 9 questions qui ne comprennent que des valeurs entières, B: les 3 questions à valeurs non entières.

2) Hypothèse nulle

Les douze effectifs sont des données, mais supposons que la position de chacun, dans A ou dans B soit le fait du hasard (hypothèse nulle).

La classification étant faite au hasard, chaque partition mettant neuf des nombres d'erreurs observés dans une même classe A et 3 dans une autre B, aurait la même chance d'apparaître. Il y aurait très peu de chances dans ces conditions pour que ce soient justement les trois effectifs les plus faibles qui soient dans B et les autres dans A.

Nous pouvons directement évaluer le nombre de ces partitions équiprobables:

Il y a $C_{12}^{3} = \frac{12 \times 11 \times 10}{3 \times 2} = 220$ partitions possibles,

autant que de manières de choisir 3 objets parmi 12. (C'est-à-dire autant que de combinaisons de 12 objets 3 à 3),

Il y a donc 1 chance sur 220 d'obtenir ensemble les effectifs les plus faibles: 4, 9 , 15

3) Modèle.

Appliquons la méthode de l'éléphant (cf fiche 1): nous avons rassemblé les questions h, g, k. Est-ce que le nombre d'erreurs sur ces trois questions est plutôt petit? normal? ou plutôt grand?

La somme des erreurs observées sur les questions correspondant à H_0 est:

$S_0 = 4 + 15 + 28 = 47$

Le nombre total d'erreurs de valorisation observées sur ces 12 questions est 366. La somme "moyenne" pour 3 questions est par conséquent 366 / 4 = 91,5.

La valeur sur les 3 questions h, g, k parait plutôt petite.

Il s'agit en fait de calculer plus exactement combien de partitions donneraient une somme plus petite que 47.

4) Rareté de la valeur observée

Calcul du nombre de partitions qui donnent une somme inférieure à S_o.

Il s'agit en fait d'une énumération: pour trouver les combinaisons les plus faibles on procède par substitution à partir de la plus faible:

1 4,9,15 28
2 4,9,25 38
3 4,9,27 40
4 " "
puisque 27 se présente 2 fois
5 4,9,28 41
6 4,9,30 43
Le suivant de cette suite serait 4,9,40 mais l'augmentation de 10 serait plus grande que celle (6) obtenue en remplaçant 9 par 15
7 4,15,25 44
8 4,15,27 46
9 " "
10 4,15,28 47
11 4,15,30 49
Comme ci-dessus il vaut mieux remplacer 4 par 9 que 30 par 40.
12 9,15,25 49
13 9,15,27 51
14 " "
15 9,15,28 52
16 4,9,40 53
on reprend la suite 4,9,...
17 4,9,43 56
...

5) Test de l'hypothèse nulle. Conclusions

S'il y a moins de 5% des partitions qui donnent des sommes plus petites nous pourrons estimer que notre partition est "étonnamment" petite.

5% x 220 = 11

Or notre triplet de questions figure à la dixième place, il est donc parmi les 11 combinaisons qui donnent les sommes les plus faibles.

Si l'hypothèse nulle était vraie, nous aurions peu de chance de tomber sur une partition donnant une somme aussi (ou plus) petite.

Nous préférons conclure que l'on doit rejeter l'hypothèse nulle. Cette hypothèse nulle était contraire à notre hypothèse.

Nous admettrons donc que "la présence de nombres non entiers rend plus rares les erreurs de "valorisation"".

Remarque 1.

Le regroupement de a,g,c (9,15,25), qui ne laisse au-dessous de lui que la question h, paraît une partition assez marginale et pourtant il ne donne déjà plus une somme significativement petite.

Remarque 2.

Il s'agit du calcul direct de la probabilité d'un événement, obtenu par un calcul "combinatoire": l'événement $S \leq S_o$ avec $S = x + y + z$

où x, y, z parcourent l'ensemble des valeurs de A et de B, $x < y < z$.

Remarque 3.

Dès que le nombre des valeurs à considérer devient supérieur à 12 les calculs deviennent vite très fastidieux. Sous certaines conditions (rapport des effectifs des deux échantillons compris entre 1/5 et 5, convexité de la distribution faible...), il est possible d'utiliser la variable de STUDENT

$$t = \frac{m_A - m_B}{\sqrt{\frac{\Sigma(ai-mA)^2 + \Sigma(bj-mB)^2}{n_a + n_b - 2} \times \left(\frac{1}{n_a} + \frac{1}{n_b}\right)}}$$

où m_A est la moyenne de l'échantillon A, m_B est la moyenne de l'échantillon B

a_i un élément courant de A, b_i un élément courant de B

n_a l'effectif de A. n_b l'effectif de B

Elle suit approximativement une loi de Student avec:

$dl = n_a + n_b - 2$

Exercice 1.

Encouragés par ce résultat, formulons une deuxième hypothèse: les élèves ont tendance à donner une valeur à la variable lorsque l'inégalité s'accompagne d'un dessin ou d'un graphe (ils prennent la valeur effective du dessin par exemple). 4 questions sont accompagnées de représentations graphiques: ce sont i,j,k,l.

La représentation graphique favorise-t-elle les erreurs par "valorisation"?

Exercice 2.

1. Dans l'exemple de la fiche est-ce que les deux échantillons obtenus dans chaque cas peuvent être considérés comme homogènes?

2. Est-ce que le nombre d'élèves à qui on a proposé l'exercice joue un rôle dans le modèle? Si oui, lequel? Comparez-le à celui que joue l'effectif d'un échantillon dans l'estimation d'une mesure.

3. Quelles autres méthodes aurions nous pu appliquer dans le même cas (test du X^2, U de Mann et Whitney)?

FICHE 22

DEUX échantillons appariés, UNE variable d'intervalle
Test du brouillage

Comme dans la fiche précédente, il s'agit de savoir si deux échantillons sont semblables ou différents, mais ces observations sont appariées en couples.

Les différences entre les valeurs obtenues dans chaque couple d'observations sont elles suffisamment concordantes pour que l'on rejette l'hypothèse de différence nulle?

Exemple:

Pour observer le contrat didactique un chercheur a repéré six couples de mots (C_i) à peu près synonymes, le premier mot est supposé appartenir au vocabulaire spontané de l'élève (VS), le second est le terme officiel (TO) dont l'enseignant veut enseigner l'usage (comme par exemple "écart de droites" et "angle de droites"). Il note le nombre d'apparitions de chaque terme dans un genre bien précis de situation d'enseignement.

1) Recueil et disposition des données

Dans notre exemple le chercheur note les données suivantes:

	C1	C2	C3	C4	C5	C6
VS	7	3	2	1	1	7
TO	10	11	5	8	5	2

2) hypothèse nulle

Il se demande si à ce moment de l'apprentissage, les termes sont bien fonctionnellement synonymes, c'est-à-dire s'ils sont employés par les élèves avec la même fréquence.

Cette hypothèse lui paraît plausible au vu de la table de contingence.

3) Modèle

L'idée consiste à calculer les différences d'effectifs d_i, puis à supposer que les valeurs absolues de ces différences sont fixées, mais que leur signe est le fait du hasard.

Cette idée permettra de construire une distribution qui correspond bien à l'hypothèse d'une répartition au hasard des signes, et qui permettra d'apprécier la rareté de la suite (contingente) des différences observées (afin d'appliquer la méthode de l'éléphant et de repérer la position de chaque tirage possible par rapport à l'hypothèse d'indifférence)

Le moyen le plus simple d'intégrer ces différences est d'en faire la somme algébrique:

$S = \Sigma\, d_i$

On place S par rapport à la distribution des différences semblables obtenues en considérant toutes les attributions possibles de signes aux valeurs absolues observées.

Dans l'exemple proposé les différences VS - VO sont

d_i	C1	C2	C3	C4	C5	C6
VS-TO	-3	-8	-3	-7	-4	+5

On suppose qu'on aurait pu aussi bien obtenir la suite:

+3 , -8 , -3 , +7 , +4 , -5 de somme $S = \Sigma\, d_i = -2$

ou encore les suites

-3 , -8 , -3 , -7 , -4 , -5 $S = \Sigma d_i = -29$

+3 , +8 , +3 , +7 , +4 , +5 $S = \Sigma d_i = +29$

-3 , +8 , -3 , +7 , -4 , +5 $S = \Sigma d_i = 0$ etc.

la somme des différences observées est:

S = (-3) + (-8) + (-3) + (-7) + (-4) + 5 = 20

4) Test de l'hypothèse nulle

Sous l'hypothèse nulle, la somme algébrique des différences n'est ni étrangement petite, ni étonnamment grande par rapport à celles obtenues par une autre distribution des signes. Plus de 5% de ces valeurs sont plus grandes, plus de 5% de ces valeurs sont plus petites.

Le nombre total des distributions de signes est 2^n, où n est le nombre de couples considérés.

En effet il y a deux possibilités pour la première différence: $+d_1$ et $-d_1$ puis pour chacune de ces possibilités, deux pour la différence d_2: $+d_2$ et $-d_2$ ce qui fait 2x2 = 4 possibilités... et ainsi de suite. On recommence n fois pour les n couples, il y a donc au total $2x2x...x2 = 2^n$ possibilités .

(On peut aussi considérer qu'il s'agit de déterminer la partie de ces nombres auxquels on attribue le signe +, les signes - étant alors automatiquement attribués aux autres nombres. Le nombre de parties d'un ensemble à n éléments est évidemment encore $2^{n)}$.

Dans notre exemple $2^n = 2^6 = 64$

Le seuil de rejet laissera donc 5% x 64 = 3,2 suites dont la somme algébrique S_i est plus petite que lui.

Il faut donc trouver la suite observée parmi les trois qui possèdent les sommes les plus petites. Donc le nombre 20 devra être dans les trois plus petites sommes.

5) Rareté de la contingence

L'attribution de signes qui donne algébriquement la somme la plus petite est celle qui attribue le signe moins à toutes les différences: $S_{min} = -29$

La suivante n'attribue le signe "+" qu'à la plus petite valeur absolue: 3

donc S = 24

La troisième est identique S = 24

La contingence ne figure dans aucune des trois: Elle serait la suivante.

Elle ne figure pas non plus dans la liste des trois attributions qui donne les sommes les plus grandes.

6) Décision

Si S figure dans la liste des 5% de sommes les plus rares, soit parce qu'elles sont grandes, soit parce qu'elles sont petites par rapport à l'ensemble des attributions possibles, il faudra rejeter l'hypothèse nulle.

Dans notre exemple on ne peut pas rejeter l'hypothèse que les élèves utilisent indifféremment l'un ou l'autre des deux termes à ce moment de leur apprentissage.

7) Remarques générales sur le rôle des épreuves statistiques dans la recherche

Malgré une certaine puissance, le test ne confirme pas l'impression du chercheur, pourtant forte à la lecture des données. Il est près de le faire et pourrait inciter, si la question était d'importance à reprendre la recherche avec des moyens plus sensibles ou sur une population plus importante .

Remarquons que dans de nombreux cas où le nombre des valeurs est faibles, ce test comme tous les test, ne permet pas de découvrir des lois dont on n'aurait pas pris conscience par une observation directe. Il permet au mieux de confirmer les plus "évidentes". C'est donc beaucoup plus un instrument de "preuve" dans un débat serré, qu'un instrument heuristique de recherche.

L'analyse de données est plus appropriée comme moyen suggestif d'investigation. D'autre part, envisager l'épreuve statistique d'une hypothèse exige du chercheur qu'il traduire ses questions et ses convictions en hypothèses falsifiables.

Mais c'est un avantage plus qu'un inconvénient et on aurait grand tort de mépriser ou même de négliger ces modestes "preuves" si utiles pour avancer dans les recherches, et de les rejeter en leur préférerant exclusivement comme arguments acceptables dans la communauté scientifique, des évidences communes ou professionnelles voire de statistique descriptive.

Annexe 1 : Rappels et compléments sur les coefficients de corrélation

Deux variables numériques X et Y sont corrélées si elles se ressemblent assez pour que l'une soit "presque déterminée" lorsque l'autre est connue; l'une des variables est presque une fonction de l'autre.

1. Variables Numériques: Le coefficient de Bravais-Pearson

Il indique dans quelle mesure deux variables numériques sont liées par une loi linéaire. Par exemple lorsque l'une croît l'autre croît aussi,ou décroît, proportionnellement.

$$r = \frac{\Sigma i(Xi - MX) \,.\, (Yi - MY)}{\sqrt{(\Sigma_i(X_i - M_X)^2)}.\sqrt{(\Sigma_i(Y_i - M_Y)^2)}}$$

Explication.

Considérons une observation (x_i, y_i) des deux variables: (X_i,Y_i).

a) Si la valeur x_i est plus grande que la moyenne des X (que nous écrivons M_X),

$x_i - M_x$ est positif.

Si en même temps y_i est plus grande que la moyenne des Y: (M_Y), alors $Y_i - M_y$ est positif lui aussi.

Pour cette observation, les variables X et Y "varient" dans le même sens.

Si les différences $X_i - M_x$ et $Y_i - M_y$ sont grandes, (toutes les deux),

alors le produit $(X_i - M_X).(Y_i - M_Y)$ est grand lui aussi.

Plus ce produit est grand, plus cette observation (x_i, y_i) contribue à accréditer l'idée que X et Y covarient ensemble dans le même sens.

Il en est de même si X et Y sont en même temps inférieurs à leurs moyennes respectives. Le produit $(X_i - M_X). (Y_i - M_Y)$ est encore positif.

b) Par contre si X_i est inférieur à la moyenne des X alors que Y_i est supérieur à celle des Y ou l'inverse, cette observation tend à faire penser que X et Y contre-varient. Alors l'expression $(X_i - M_X). (Y_i - M_Y)$ est négative.

c) En ajoutant les contributions (positives) de toutes les observations qui covarient et les contributions (négatives) de toutes celles qui contrevarient, on obtient une mesure de la tendance à covarier ou à contrevarier des deux variables X et Y.

La covariance $\Sigma_i(X_i - M_X). (Y_i - M_Y)$ indique qui, des unes ou des autres, l'emportent.

Cette somme prend pour valeur maximum

$\sqrt{(\Sigma_i(X_i - M_X)^2)}.\sqrt{(\Sigma_i(Y_i - M_Y)^2)}$

(lorsque pour tous les i, $(X_i - M_X) = a.(Y_i - M_Y)$, c'est-à-dire lorsque Y se calcule exactement en fonction affine de X).

r indique alors quelle proportion de cette valeur maximum est atteinte.

Significativité du coefficient de BRAVAIS-PEARSON

Il reste alors à évaluer la "significativité" c'est-à-dire la rareté de cette valeur.

On suppose que l'échantillon observé est extrait au hasard d'un ensemble parent infini dont chaque élément détermine les valeurs de deux variables X et Y. On suppose que ces variables sont (à peu près) normales et non corrélées, c'est-à-dire indépendantes.

Sous cette hypothèse on peut procéder au calcul (ou à la simulation, par exemple par ordinateur), d'un très grand nombre de tirages, au hasard, d'échantillons de même taille et calculer pour chacun le coefficient de corrélation obtenu.

La distribution de ces valeurs permet de placer le coefficient observé et de lui attribuer une certaine rareté. Pratiquement on peut se rapporter à la table donnée en annexe de cette fiche.

2. Variables ordinales: Coefficient de corrélation rhô de Spearman

Si la distribution de X et de Y n'est pas connue ou si elle s'écarte visiblement trop d'une distribution normale, on peut alors ne plus considérer que l'ordre des observations (et non plus leur valeur) et se ramener au cas suivant.

Nous avons traité les applications dans la fiche 18.

Explication

Que devient la formule (1) lorsque X et Y sont des rangs?

a) Calcul du numérateur.

Il faut calculer $\Sigma_i (x_i - M_X).(y_i - M_Y)$; $\Sigma_i (x_i - M_X)^2$ et $\Sigma_i (y_i - M_Y)^2$ en fonction de N et de d_i différence entre les rangs attribués à une même observation suivant les deux variables:

$d_i = x_i - y_i$.

$$\begin{aligned}\Sigma_i (X_i - M_X)^2 &= \Sigma_i (x_i^2 + M_X^2 - 2x_i.M_X)\\ &= \Sigma_i x_i^2 + N.M_X^2 - 2.M_X.\Sigma_i X_i\\ &= \Sigma_i x_i^2 + N.M_X^2 - 2\, M_X.\, N.M_X\\ &= \Sigma_i x_i^2 - N.M_X^2 \qquad (1)\end{aligned}$$

$$M_X = \frac{\Sigma_i x_i}{N} \quad \text{et} \quad \Sigma_i x_i = \frac{N.(N+1)}{2}$$

car en fait $\Sigma_i x_i$ est la somme des N premiers nombres entiers

$$\text{d'où} \quad M_X = \frac{(N+1)}{2} \text{ et } M_X^2 = \frac{(N+1)^2}{4}$$

$$\text{d'autre part} \quad \Sigma_i x_i^2 = \frac{N.(N+1)(2N+1)}{6}$$

en portant ces valeurs dans (1). il vient

$$\begin{aligned}\Sigma_i (x_i - M_X)^2 &= \frac{N.(N+1)(2N+1)}{6} - \frac{N.(N+1)^2}{4}\\ &= \frac{(2.N.(N+1)(2N+1) - 3.N.(N+1)^2)}{12}\\ &= \frac{N.(N+1)(2(2N+1) - 3(N+1))}{12}\\ &= \frac{N.(N+1)(4.N + 2 - 3.N - 3)}{12} = \frac{(N^3 - N)}{12}\end{aligned}$$

De même: $\Sigma_i (y_i - M_Y)^2 = \frac{(N^3 - N)}{12}$

Calculons : $\Sigma_i (x_i - M_X)(y_i - M_Y)$

$d_i = x_i - y_i$;

Remarquons que puisque $M_X = M_Y$

$d_i = (x_i - M_X) - (y_i - M_Y)$,

$\Sigma_i d_i^2 = ((x_i - M_X) - (y_i - M_Y))^2$

$= \Sigma_i (x_i - M_X)^2 + \Sigma_i (y_i - M_Y)^2 - 2\Sigma_i (x_i - M_X)(y_i - M_Y)$

d'où l'on tire le numérateur du coefficient de corrélation:

$\Sigma_i (x_i - M_X)(y_i - M_Y) = \Sigma_i (x_i - M_X)^2 + \Sigma_i (y_i - M_Y)^2 - \Sigma_i d_i^2$

$= \frac{2.(N^3 - N)}{12} - \Sigma_i d_i^2$

b) Calcul du dénominateur

$\sqrt{\Sigma_i (xi - MX)^2}.\sqrt{\Sigma_i (yi - MY)^2} = \frac{2.(N^3-N)}{12}$

d'où $r_S = \frac{\frac{2.(N^3-N)}{12} - \Sigma_i d_i^2}{\frac{2.(N^3-N)}{12}}$

c) D'où le coefficient de corrélation par rangs: rhô de Spearman

$$r_S = 1 - \frac{6.\Sigma_i d_i^2}{(N^3-N)}$$

Où d_i est $x_i - y_i$ la différence entre les rangs obtenus par la ième observation selon les deux variables et N le nombre d'observations.

3 Le coefficient Tau de Kendall.

Le calcul de ΣS correspond au dénombrement des dominations que nous avons présenté avec le U de Mann et Whitney.

4 Quelques explications sur la corrélation.

a) Deux variables numériques X et Y sont corrélées si elles se ressemblent assez pour que l'une soit "presque déterminée" lorsque l'autre est connue; l'une des variables est presque une fonction de l'autre (par exemple Y # f(X), c'est dire que Y n'est pas très éloignée de f(X), (encore faut-il trouver f).

Montrer que deux variables X et Y sont corrélées consiste donc:

- à choisir une distance générale entre les variables, c'est-à-dire entre les fonctions,
- à choisir parmi une famille de fonctions simples de x (le plus souvent des fonctions linéaires ou affines mais au besoin on peut s'intéresser à des polynômes de degré plus élevé), celle qui, pour cette distance, est la plus près de Y. Appelons la f
- à évaluer cette distance entre Y et f(x)
- à s'assurer que cette distance est assez petite pour qu'on puisse remplacer Y par f(X) dans certaines analyses.

b) Par exemple le coefficient de Bravais Pearson peut être obtenu ainsi:
la distance retenue est la distance euclidienne relative,
la famille de fonctions choisie est celle des fonctions affines,
la meilleure fonction de cette famille est la droite de régression,
la distance entre Y et cette droite est la covariance, r est la distance relative,
mais aussi le cosinus de l'angle de $(X - X_M)$ avec $(Y - Y_M)$,

et enfin la racine carrée du rapport entre la variance expliquée par la régression (par la fonction) et la variance totale.

Cette dernière interprétation permet de déterminer de façon théorique la loi que suit r (c'est une loi de Fisher) et de déterminer la rareté de la valeur observée: la table des valeurs significatives de r

Annexe 2 : Rappels et compléments sur les Tests paramétriques

Dans les fiches précédentes, nous nous sommes référés à plusieurs reprises aux tests paramétriques que nous avons exclus du présent ouvrage. Il nous apparaît néammoins nécessaire de rappeler brièvement les résultats de ce domaine que nous avons utilisés

Rappelons que l'interprétation statistique consiste

- d'abord à envisager un certain nombre de faits à l'aide de structures mathématiques très générales permettant de les décrire et de les recueillir sous formes de données (comme nombres, propriétés, couples, applications, mesures distances, etc.)
- puis de leur chercher des régularités plus spécifiques sous forme de modèles (nombre, hypothèse, distribution, etc.).
- ensuite de représenter la qualité de ce modèle en calculant une certaine distance entre le modèle et les données qu'il doit représenter
- enfin d'évaluer cette distance en la plaçant par rapport à une distribution théorique.

Par exemple dans la fiche 1 nous avons d'abord observé dans une classe le nombre de réussites à un exercice, puis nous lui avons assigné un modèle, un nombre de réussites dans un population parente ou calculé sous une hypothèse quelconque, ensuite nous avons calculé la distance du χ^2 entre les valeurs observées et les valeurs théoriques, et enfin nous avojns lu la rareté de cette valeur dans la distribution du χ^2.

Les tests paramétriques utilisent le même schéma. La différence tient dans le fait que

- les méthodes non paramétriques n'impliquent pas de spécification a priori de lois théoriques dépendant d'un nombre fini de paramètres,
- les méthodes paramétriques sont celles qui obligent à spécifier la loi théorique de l'échantillon (le plus souvent la loi normale) et à estimer les paramètres caractéristiques de ce ces lois. Tels sont les test de comparaisons de fréquences et d'échantillons

Lorsque la taille des échantillons tend vers l'infini les deux types de méthodes donnent les mêmes lois limites ainsi que nous l'avons vu à plusieurs reprises dans ces fiches.

Les tests paramétriques permettent de faire correspondre des modèles probabilistes aux phénomènes étudiés et de les relier entre eux par des hypothèses mathématiques clairement établies (théorème central limite par exemple). Les distributions de référence ont été établies par le calcul et se présentent à l'utilisateur sous forme de tables. Mais il est aujourd'hui possible de les obtenir par simulation informatique sur des appareils très simples. Nous évoquons ce moyen avec l'espoir de rendre plus accessible l'interprétation des modèles.

1. Les comparaisons d'effectifs et de fréquences

1.1. Modèle de l'alternative.

Un individu x possède un caractère C. On sait que d'autres individus possèdent ce caracrtères et certains ne le possèdent pas.

On peut représenter ce fait par le modèle suivant:

x est tiré au hasard dans une population parente infinie comprenant une certaine proportion p d'individus possédant C.

Par exemple dans la fiche 7 nous considérons les élèves qui ont comme caractère d'avoir à la fois réussi l'exercice A et échoué à l'exercice B. Le fait que tel élève, Dupont, ait ce caractère est considéré comme le résultat d'une expérience aléatoire.

La variable "l'élève x possède C" a deux valeurs: {oui, non}: c'est la **variable de l'alternative**, de moyenne p et d'écart-type $\sigma = \sqrt{p(1-p)}$

Test de l'alternative, Seuil de confiance

Un quidam propose une certaine valeur pour p, par exemple p = 0,001 . Si cette affirmation est vraie, est il croyable que l'on obtienne en un seul tirage un individu aussi rare?
On peut accepter que oui (après tout des évènements improbables se produisent "tous les jours") ou refuser (pour ne pas tout admettre et discerner des lois). Pour éviter de laisser la décision fluctuer selon l'humeur ou l'intérêt de l'observateur, on peut fixr un seuil de confiance au delà duquel on refuser d'accepter le modèle. Par exemple si on se donne le seuil de 5% il faut écarter la valeur p comme modèle du fait observé. Si on s'est imposé un seuil de 1 $^{0}/_{00}$ on ne peut pas la rejeter.

Remarquons que la valeur 1 observée se trouve éloignée de la valeur moyenne p = 0,001 de plus de 25 fois l'écart type: 0,0316.
Dans la fiche 7 la valeur proposée par l'hypothèse nulle est p = $\frac{\text{effT}}{\text{n}} = \frac{3,96}{25} = 0,16$ alors
$\sigma = 0,366$ le tirage au premier coup du seul individu de la classe qui réalise A et non B est beaucoup plus crédible quoique la valeur observée se trouve à un peu plus de deux fois l'écart type de la valeur moyenne.

Simulation

Nous allons procéder comme dans la fiche 1 et utiliser un modèle d'urne. Nous disposons dans une urne, 2500 billes dont 396 sont blanches et les autres noires (c'est-à-dire 100 fois les valeurs théoriques). un élève, au lieu de faire les deux exercices A et B vient tirer une bille. Si la bille est blanche on dira que cet élève a réussi A et échoué à B. Si la bille est noire l'élève est dans l'un des autres cas, peu importe lequel (les exigeants pourraient préparer quatre couleurs de billes).

On peut recommencer l'expérience de l'alternative 1000 fois et regarder combien de fois on obtient, "en un coup", un élève qui réalise A et non B. Dans cette nouvelle expérience, ce nombre sera voisin de 160. et elle a pour modèle le modèle binômial exposé ci après. Si on recommence 1000 fois cette nouvelle expérience on pourra alors observer la distribution des fréquences avec laquelle on obtient du premier coup un individu A. son écart type sera voisin de 0,366

1.2. Modèle binômial

Une ensemble de *n* individus dont *a* possèdent un caractère *A* et les autres non, peut être représentée par le modèle suivant:

Une population infinie, dite parente, présente une proportion $p = \frac{a}{n}$ d'individus *A* . On répète sur cette population l'expérience de l'alternative n fois, les tirages étant indépendants. On observe alors dans l'échantillon ainsi obtenu, k individus présentant le caractère A.

On peut imaginer dans l'exemple de la fiche 7 que l'échantillon de 25 tirages représente une classe de 25 élèves dans laquelle le nombre d'élèves ayant à la fois réussi l'exercice A et échoué à l'exercice B est la valeur fournie par l'hypothèse nulle: 3,96 (disons 4 pour une représentation réaliste)

Loi binômiale.

Lorsqu'on répète indépendamment n fois une expérience de l'alternative, au cours de laquelle un évènement A de probabilité p peut se produire, la probabilité de voir cet évènement apparaître k fois est:

$$\text{Proba (A se produit k fois)} = \mathbf{p_k} = \binom{n}{k} \times p^k \times (1-p)^{n-k}$$

dans cette formule: $\binom{n}{k} = \frac{n!}{i! \times (n-i)!} = \frac{n.(n-1).(n-2)...(n-i+1)}{i.(i-1)...1} = C_k^n$

n est le nombre d'expériences, dans l'exemple le nombre d'élèves
p la probabilité que l'évènement E se produixe au cours d'une expérience,
k le nombre de fois que l'évènement E apparaît au cours des n expériences

Dans le cas particulier d'un tirage à pile ou face, p = 1-p = 1/2
Cette formule exprime que la probabilité d'avoir k piles est égale au produit:
- de la probabilité d'une séquence particulière quelconque de n piles ou faces:
p f f p p f p p f f f ...p f p p p
Cette probabilité est $1/(2^n)$ car il y a 2^n séquences différentes possibles et équiprobables.
- par le nombre de séquences contenant k piles.
Ce nombre est égal au nombre de parties à k éléments que l'on peut déterminer dans un ensemble de n places c'est-à-dire: C_k^n.

La variable {on tire k évènements A en n répétitions indépendantes d'une alternative de probabilité p} peut prendre n+1 valeurs: de 0 à n, chacune avec la probabilité P_k.

La distribution $\mathcal{B}(n,p)$ de cette variable est dite distribution (loi) binômiale.

La **moyenne** de la loi binômiale est **n.p** et son **écart type** est $\sqrt{\mathbf{n.p.(1 - p)}}$. ou encore

Comparaison d'un effectif observé à un ensemble parent (test binômial, n $\leq$ 30)

Considérons un échantillon de n observations (n $\leq$ 30) dont ef_{obs} présentent le caractère A. Un devin propose à nouveau une valeur p, comme proportion des A dans la population parente ou une valeur eff_T comme effectif théorique. Est-il raisonnable de le croire?

Remarquons que la **moyenne** de la loi binomiale s'écrit aussi $\mathbf{eff_T}$ et que son **écart type** s'exprime par $\sqrt{\mathbf{eff_T.(1 - \frac{eff_T}{n})}}$.

Dans l'exemple de la fiche 7, on se demandait s'il est raisonnable de croire qu'on n'obtient que 1 individu A, au cours de 25 tirages dans une population parente qui comprend une proportion p = 0,16 d'individus A selon l'hypothèse nulle.

Calculons

$$\text{Proba (A se produit k ou moins de k fois)} = \mathbf{F_k} = \sum_{i=0}^{k} \mathbf{p_i}$$

explicitement $$\mathbf{F_k} = \sum_{i=0}^{k} \binom{n}{i} \times p^i \times (1-p)^{n-i}$$

F_k est la somme des probabilités p_i pour toutes les valeurs de i inférieures à k.

Il est alors possible, de calculer de la même façon que dans les conditions de l'expérience:

proba (A se produit ef_{obs} ou moins de ef_{obs} fois) = $\mathbf{S_{obs}} = \sum_{k=0}^{ef_{obs}} \binom{n}{k} \times p^k \times (1-p)^{n-k}$

et d'appliquer la méthode du seuil de confiance:

Si la probabilité de l'évènement observé, ici S_{obs} est inférieure à un seuil déterminé, par exemple 5%, on préfèrera penser que la valeur f_{obs} est trop éloignée de son modèle « n.p » et que celui ci doit être rejeté.

(Il vaut mieux disposer d'une bonne calculatrice pour effectuer le calcul si k est grand! ou encore lire le résultat dans une table de la loi binomiale.

Dans l'exemple de la fiche 7, l'effectif observé est $f_{obs}= 1$,

la probabilité S_{obs} qu'il y ait 0 ou 1 élèves est :

$p_{(0\ élève)} = (1-p)^{25} = (0,84)^{25} = 0,0127$

$p_{(1\ élève)} = 25 \times p \times (1-p)^{24} = 0,0609$

$S_{obs} = p_{(0\ élève)} + p_{(1\ élève)} = 0,0736$

La probabilité d'avoir par hasard, 0 ou 1 élève ayant réalisé A et non B, est supérieure à 7%. On ne peut donc pas rejeter l'hypothèse nulle au seuil de 5% et affirmer qu'il y a implication.

Simulation

Pour représenter la distribution des élèves de cette classe, nous tirerons 25 fois - avec remise de la boule tirée - dans l'urne ci dessus. On peut obtenir au cours d'une telle expérience 4 boules blanches, mais aussi 6, ou 1, ou aucune.

Nous recommençons l'expérience pour simuler un grand nombre de classes, 10 000 par exemple. Nous examinons la distribution de ces 10 000 tirages: combien ne présentent qu'une blanche? combien 0 blanche? Si ces deux cas réunis sont moins de 500, c'est-à-dire moins de 5 %, on peut estimer que notre résultat $ef_{obs} = 1$ est suffisamment rare pour justifier le rejet de l'hypothèse nulle. Le calcul ci dessus montre que ce n'est pas le cas

Remarque

$ef_T = 3,96$ et que $\sigma = \sqrt{n.p\ (1 - p)} = \sqrt{3,96.(1 - 0,1584)} = \sqrt{3,332} = 1,825$

La valeur observée 1 s'écarte de la moyenne 3,96 de 2,96, c'est à dire moins de deux fois l'écart type.

Comparaison d'un pourcentage observé avec un pourcentage théorique (n ≤ 30)

Il ne faut pas confondre efobs, l'effectif des observations, un nombre naturel, avec la fréquence des observations.

$f_{obs} = \frac{ef_{obs}}{n}$, un nombre inférieur à 1. ni avec le pourcentage $100 .f_{obs}$

Tous les calculs ci dessus concernent la distribution des effectifs. Ils peuvent être traduits pour les fréquences.

Proba ($f_{obs} = k/n$) = $p_k = \binom{n}{k} \times p^k \times (1-p)^{n-k}$ et

Proba ($f_{obs} \leq k/n$) = $\mathbf{F_k} = \sum_{i=0}^{k} \mathbf{p_i}$

La méthode de comparaison est la même.

f_{obs} suit une distribution de moyenne p et d'écart type $\sqrt{\frac{p.(1-p)}{n}}$ qui se déduit de la binômiale: sur les valeurs k/n elle prend la valeur p_k.

1.3 Convergence vers la loi Normale

On montre que, si le nombre d' expériences répétées, c'est-à-dire la taille de l'échantillon, croît et tend vers l'infini, la distribution binômiale correspondante de la fréquence observée tend vers une une loi de Gauss et f_{obs} vers la variable aléatoire normale Z. On peut confondre les deux lois dès que n > 30.

Ainsi, sous l'hypothèse nulle, la quantité $f_{obs} = \frac{ef_{obs}}{n}$ suit alors une loi normale

de moyenne $p = \frac{ef_T}{n}$

et d'écart type:

$$\sigma = \sqrt{\frac{1}{n} \cdot \frac{ef_T}{n} \cdot (1 - \frac{ef_T}{n})} = \sqrt{\frac{p.(1-P)}{n}}$$

Remarque,

$$Z = \frac{\frac{eff_T}{n} - \frac{eff_{obs}}{n}}{\sqrt{\frac{p.(1-P)}{n}}} = \frac{eff_T - eff_{obs}}{n.\sqrt{\frac{p.(1-P)}{n}}} = \frac{eff_T - eff_{obs}}{\sqrt{n.\ p.(1-P)}}$$

Cette dernière formule montre que la loi suivie par eff_{obs} est une loi une loi de Gauss de moyenne eff_T et d'écart type $\sqrt{n.\ p.(1-P)}$ qui prolonge la loi binômiale.

La formule ci dessous permet un calcul plus rapide:

$$Z = \frac{eff_T - eff_{obs}}{\sqrt{n.\frac{eff_T}{n}.(1-\frac{eff_{obs}}{n})}} = \frac{eff_T - eff_{obs}}{\sqrt{eff_T.(1-\frac{eff_{obs}}{n})}} = \frac{eff_T - eff_{obs}}{\sqrt{eff_T.\frac{n - eff_{obs}}{n}}}$$

$$= \frac{\text{Valeur observée - } \mu}{\sigma_\varepsilon}$$

avec $effT = \mu$ et $\sigma_\varepsilon = \sqrt{\mu \times (1 - \frac{\mu}{n})}$

Intervalle de confiance

Nous avons évalué à plusieurs reprises l'éloignement d'un effectif observé f_{obs} par rapport à une distribution binômiale modèle de moyenne ef_T

et d'écart type $\sqrt{n.p\ (1 - p)} = \sqrt{n.\frac{ef_T}{n}\ (1 - \frac{ef_T}{n})} = \sqrt{ef_T\ (1 - \frac{ef_T}{n})}$

en mesurant la distance $f_T - f_{obs}$ ou $ef_T - ef_{obs}$ avec l'écart type comme unité, il vient

$$h = = \frac{x\ ef_T - ef_{obs}\ x}{\sqrt{ef_T.(1 - \frac{ef_T}{n})}} = \frac{x\ f_T - f_{obs}\ x}{\sqrt{\frac{f_T(1- f_T)}{n}}} = \frac{x\ p - f_{obs}\ x}{\sqrt{\frac{p(1-p)}{n}}}$$

Cette formule exprime que ef_{obs} est, soit plus grande, soit plus petite que ef_T, la différence étant exactement h fois l'écart type

Si cette distance h est trop grande on rejettera le modèle.

On peut écrire que

ef_{obs} doit appartenir à un **intervalle de confiance**:

$$ef_T - h\sqrt{n.p(1-p)} \leq ef_{obs} \leq eff_T + h\sqrt{n.p(1-p)}$$

donc f_{obs} doit appartenir à un intervalle de confiance tel que

$$p - h\sqrt{\frac{p(1-p)}{n}} \leq f_{obs} \leq p + h\sqrt{\frac{p(1-p)}{n}} \qquad (1)$$

Il suffit donc de calculer une fois pour toute la distance hs correspondant au seuil choisi (par exemple 5%), puis de comparer à chaque décision h avec h_S. Si h est plus petit que h_S, l'observation n'est pas suffisamment éloignée de la distribution modèle pour qu'on le rejette. S'il est plus grand, le modèle est rejeté.

La table de **la distribution normale** permet de conclure que pour le seuil de 5%, h = 1,96

et que pour le seuil de 1% on a h = 2,576

Cela signifie que si le modèle est vrai on on a moins de 5% de chances de l'écarter (à tort) au vue du résultat d'un échantillon. On dit que le risque est de 5%

L'intervalle de confiance au seuil de 5% est donc

$$p - 1{,}96\sqrt{\frac{p(1-p)}{n}} \leq f_{obs} \leq p + 1{,}96\sqrt{\frac{p(1-p)}{n}}$$

Il est utilisable avec $n \leq 30$

Comparaison d'un pourcentage observé sur un échantillon de taille n avec un pourcentage théorique (n ≤ 30)

Il suffit donc de calculer l'intervalle de confiance au seuil choisi et de placer par rapport à lui la proportion observée.

Exemple.

Dans une école, on a proposé à 52 élèves un exercice d'évaluation qui est réussi dans la population parente par 68 % des élèves. On observe r réussites

Cette école diffère-telle significativement au seuil de 5% de la "moyenne" nationale?

Réponse:

Il vient $1{,}96\sqrt{\frac{p(1-p)}{n}} = 0{,}126$, $\quad 0{,}55 \leq f_{obs} \leq 0{,}80$

Si r est inférieur à 28 élèves (0,55 . 52) ou supérieur à 42 (0,80 . 52) l'échantillon n'est pas extrait de la population parente

Même calcul pour une classe de 26 élèves:

Il vient $1{,}96\sqrt{\frac{p(1-p)}{n}} = 0{,}179$

$$0{,}50 \leq f_{obs} \leq 0{,}859$$

Si r est inférieur à 13 élèves (0,50 . 26) ou supérieur à 23 (0,86 . 26) l'échantillon n'est pas extrait de la population parente. On voit que l'intervalle s'élargit lorsque n diminue.

On sort en principe des limites d'application de la méthode. Il faudrait utiliser le test de la binômiale ou celui du χ^2.

Exercice.

Sur 490 000 enfants 250 000 garçons. Ce résultat est il compatible avec l'hypothèse selon laquelle le nombre de garçons est égal au nombre de filles?

Intervalle de confiance du modèle p

En possession d'un effectif observé ef_{obs}, on peut aussi déterminer l'intervalle de confiance à l'intérieur duquel on peut choisir la valeur théorique: il faut résoudre l'inégalité (1). En pratique on accepte

$$f_{obs} - h\sqrt{\frac{f_{obs}(1-f_{obs})}{n}} \leq p \leq f_{obs} + h\sqrt{\frac{f_{obs}(1-f_{obs})}{n}}$$

Comparaison des pourcentages de deux échantillons.
Test d'homogénéité.

a) Soient deux échantillons d'effectifs n_1 et n_2 tous les deux supérieurs à 30 où les fréquences d'apparition du caractère A sont respectivement f_1 et f_2. Peut on trouver un même modèle pour ces deux échantillons? Autrement dit, peut on les considérer comme extraits d'une même population parente?

La méthode la plus simple théoriquement consiste naturellement à examiner l'intersection des intervalles de confiances.

Il existe toutefois un moyen plus rapide: il consiste à déterminer dans quelle mesure la différence entre les deux proportions est étonnamment grande. Et pour cela d'établir à l'avance la distribution de cette différence $f_1 - f_2$.

b) Si les deux échantillons sont extraits du même ensemble parent où le caractère A est présenté par une proportion p de la population, la distribution des fréquences est respectivement

une loi binomiale $\mathcal{B}(p; n_1)$ et une loi binomiale $\mathcal{B}(p; n_2)$, ou si les effectifs sont assez grands deux lois normales: $\mathcal{N}(p;\ \sigma_1 = \sqrt{\frac{p(1-p)}{n_1}})$ et $\mathcal{N}(p;\ \sigma_1 = \sqrt{\frac{p(1-p)}{n_2}})$

la variable $d = f_1 - f_2$ est la différence de deux variables normales indépendantes.

Elle obéit donc à une loi normale de moyenne $p - p = 0$ et de variance égale à la SOMME des variances

$$(\sigma_d)^2 = (\sigma_1)^2 + (\sigma_2)^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Si les deux échantillons sont extraits d'un même ensemble parent la différence réduite est

$$Z = \frac{d}{\sigma_d} = \frac{f_1 - f_2}{\sqrt{p.(1-p).(\frac{1}{n_1} + \frac{1}{n_2})}}$$

et t suit la loi normale centrée réduite.

c) Il reste à estimer p

Le nombre total d'observations est $n_1 + n_2$ et dans cet échantillon supposée extrait de l'ensemble parent on observe $n_1.f_1 + n_2.f_2$ éléments présentant le caractère A. La meilleure estimation de p est donc

$$p = \frac{n_1.f_1 + n_2.f_2}{n_1 + n_2}$$

d) Le test consiste à calculer t puis à le comparer aux valeurs seuils correspondant au risque choisi ε : Avec la loi normale centrée réduite, nous avons vu plus haut que

si ε = 0,05 alors h = 1,96

Si |t| est supérieur à h, l'hypothèse nulle doit être rejetée.

Remarquons qu'il est possible de comparer les deux distributions avec le test du χ^2, mais le test est dissymétrique: laquelle des deux est prise comme valeur théorique?

2. Les comparaisons de moyennes (grands échantillons)

Considérons un ensemble de valeurs numériques. Peut on considérer ces valeurs comme extraites d'un ensemble parent par un échantillonnage où les tirages sont indépendants?

Le modèle le plus fréquemment utilisé (mais ce n'est pas le seul) est donc une population de valeurs que l'on suppose distribuées selon la loi normale de moyenne m et d'écart-type σ.

Ce modèle s'impose chaque fois que l'on peut penser que les différences de valeurs observées sont l'effet d'une somme de petites erreurs autour d'une valeur connue ou non mais fixe.

a) la population parente est connue.

Un échantilon (x_i) de taille n > 30, est extrait de la population parente de moyenne m et d'écart-type σ, connus. Cet échantillon a pour moyenne m_{obs}. Cette valeur n'est pas nécessairement égale à m.

On démontre que m_{obs} est une variable aléatoire dont la distribution est normale de moyenne m et d'écart-type $\frac{\sigma}{\sqrt{n}}$

Intervalle de confiance de la moyenne d'une variable numérique.

Après avoir déterminé h dans la loi normale centrée réduite, en fonction du seuil de confiance choisi, on peut calculer et utiliser l'intervalle de confiance comme plus haut:

$$m - h\frac{\sigma}{\sqrt{n}} \leq m_{obs} \leq m + h\frac{\sigma}{\sqrt{n}}$$

b) Les paramètres de la population parente ne sont pas connus: estimation.

Un échantilon (x_i) de taille n > 30, est extrait de la population parente normalement distribuée mais de moyenne m et d'écart-type σ, inconnus.

Cet échantillon a pour moyenne m_{obs} et pour variance $\sigma^2_{obs} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_{obs})^2$

On montre que m_{obs} est une bonne estimation de m et que σ_e est une bonne estimation de σ

$$\sigma^2_e = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_{obs})^2$$

c) comparaison de deux "grands" échantillons normalement distribués

α) 1er échantillon: taille n_1, moyenne observée $\underline{x}_1$; la population parente: a pour moyenne m_1, et pour écart-type σ_1 estimé :

$$\sigma_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{i1} - \overline{x_1})}{n_1 - 1}}$$

Alors $\underline{x}_1$ a pour modèle la loi normale $\underline{X}_1$ de moyenne m_1 et de variance $\frac{{\sigma_1}^2}{n_1}$

β) 2ème échantillon: taille n_2, moyenne $\underline{x}_2$;

La population parente a pour moyenne m_2, et pour écart-type σ_2 estimé :

$$\sigma_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(x_{i2} - \overline{x_2})}{n_2 - 1}}$$

Alors $\underline{x}_2$ a pour modèle la loi normale $\underline{X}_2$ de moyenne m_2 et de variance $\frac{{\sigma_2}^2}{n_2}$

γ) hypothèse nulle: $m_1 = m_2 = m$ et $\sigma_1 = \sigma_2 = \sigma$

Dans ces conditions la différence entre les moyennes $d = \underline{x}_1 - \underline{x}_2$ a pour modèle la variable aléatoire $D = \underline{X}_1 - \underline{X}_2$ elle aussi normale de moyenne nulle et de variance la somme des variances ci dessus.

La variable centrée réduite : $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}}$ est donc la variable normale $\mathcal{N}(0;1)$

Il convient alors de comparer |t| avec h au seuil choisi comme ci dessus.

Remarque: Il convient aussi de s'assurer que l'hypothèse que les échantillons comparés sont normalement distribués n'est pas contredite par un test de normalité.

3. Les comparaisons de moyennes (petits échantillons): t de STUDENT

Pour éprouver l'homogénéité de deux échantillons d'effectifs inférieurs à 30 on utilise la propriété suivante:

la valeur t:

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

dans laquelle:

$$S^2 = \frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n_1}(x_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

suit une loi de Student Fisher au degré de liberté

$\nu = n_1 + n_2 - 2$ (voir table ci-après)

On peut alors comparer cette valeur avec la valeur seuil t_S que donne la table et telle que $p(|T| > t_S) = \varepsilon$

Si $|t| > t_S$ l'hypothèse nulle est rejetée.

La distribution t de **Student Fisher** est présentée par la table ci après

Distribution du t de **Student Fisher**

	niveau de signification pour un test unilatéral					
	0,10	0,05	0,025	0,01	0,005	0,0005
	niveau de significationpour un test bilatéral					
	0,20	0,10	0,050	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
infini	1,282	1,645	1,960	2,326	2,576	3,291