La procédure FREQ de SAS. Tests d'indépendance et mesures d'association dans un tableau de contingence

Josiane Confais, Yvette Grelet, Monique Le Guen

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail
'Méthodologie Statistique'

N° 9603

La procédure FREQ de SAS®

Tests d'indépendance et mesures d'association dans un tableau de contingence

J. Confais, Y. Grelet, M. Le Guen

Ce document est aussi enregistré sous le n° F9610 de la Série de Documents de Travail de la Direction des Statistiques Démographiques et Sociales

La procédure FREQ de SAS®

Tests d'indépendance et mesures d'association dans un tableau de contingence

Josiane CONFAIS
Institut de Statistique de l'Université Pierre et Marie Curie - Paris VI -

Yvette GRELET
Centre d'Etudes et de Recherches sur les Qualifications

Monique LE GUEN
Centre National de la Recherche Scientifique
INSEE - Unité Méthodes Statistiques

RESUME

Ce document présente de manière pédagogique les divers tests et mesures d'association disponibles dans la procédure FREQ de SAS. Ces tests et mesures sont classés selon le type nominal, ordinal des variables étudiées ; puis ils sont décrits, commentés et appliqués sur des exemples variés.

L'approche probabiliste basée sur les odds-ratio et le modèle logit est abordée.

Afin de montrer les doutes que l'on doit avoir lors d'un test unique, une «curiosité » est rapportée ; celle-ci révèle les discordances des résultats selon les points de vue.

Un historique sur le test exact de Fisher permet au lecteur de conforter son opinion.

MOTS CLES : Tableau de contingence, tests d'indépendance, mesures d'association

La procédure

FREQ de SAS®

Tests d'Indépendance et Mesures d'Association

dans un tableau de contingence

Josiane Confais (UPMC-ISUP)
Yvette Grelet (CEREQ-LES)
Monique Le Guen (CNRS-INSEE-UMS)

Ce document a déjà été publié en 1992 à l'Université d'Orléans.

Révision Mai 1996

Coordonnées des auteurs :

Josiane Confais Université UPMC- ISUP Tour 45 Boîte 157 4, Place Jussieu 75 252 Paris Cedex 5	Yvette Grelet CEREQ-LES Université Paris Tolbiac 90 rue de Tolbiac 75634 Paris Cedex 13	Monique Le Guen INSEE-UMS/MEAD Timbre F410 18, Boulevard A. Pinard 75675 Paris Cedex 14

PRÉFACE

Le travail de Josiane Confais, Yvette Grelet et Monique Le Guen arrive à point nommé. L'analyse des variables catégorielles, classique en biostatistique, trouve maintenant des applications nouvelles et de plus en plus nombreuses dans des domaines aussi variés que l'économie, la finance, l'assurance, les sciences humaines, ... Des méthodes inférentielles puissantes sont appliquées : analyses logit, log-linéaire, modèles Glim.

Comme dans toute analyse multivariée, il est au préalable nécessaire d'analyser les données étudiées et d'évaluer l'intensité (et le sens) des liaisons entre les variables catégorielles prises 2 à 2. De nombreuses mesures d'association, se substituant aux mesures de corrélation du cas continu, sont disponibles et permettent de ne plus se limiter au seul calcul du chi-2 de contingence.

Ces mesures sont fournies dans la plupart des logiciels statistiques standards, et en particulier dans S.A.S., qui tend à devenir la référence dans ce domaine.

Ce polycopié présente de manière très pédagogique, très vivante mais en même temps très exacte, les divers tests et mesures d'association disponibles dans la procédure FREQ de S.A.S. Ceux-ci sont décrits, commentés et appliqués sur des exemples variés.

Il faut remercier les auteurs pour le travail considérable qu'a nécessité leur présentation claire et synthétique des nombreux résultats statistiques présents dans cette procédure. Il est à souhaiter que d'autres procédures subissent le même "traitement" (par exemple REG, NLIN, ou la gigantesque procédure CATMOD) afin d'aider de manière aussi efficace les utilisateurs de S.A.S.

Christian PARTRAT
Directeur de l'ISUP

SOMMAIRE

Avant Propos 2

Introduction 3

I - Terminologie 4
I - 1 Variables 4
I - 1 . 1 Approche par les “qualités” ou "propriétés" d’une variable 4
I - 1 . 2 Approche liée aux techniques de traitement 5
I - 1 . 3 Approche plus française 7
I - 2 Tableaux de fréquences - Tables de contingence 9
I - 2 . 1 Tableaux de fréquences pour 1 variable 9
I - 2 . 2 Tableaux de fréquences pour 2 ou n variables 10
I - 3 Exemples de structure dans des tableaux 12
I - 4 . Mesures d’association - Tests d’indépendance 15
I - 4 . 1 Qu’est-ce qu’une association? 15
I - 4 . 2 Qu'est-ce qu'un test d'indépendance? 15
I - 5 Inventaire des Tests et Mesures 17

II - Analyse d'un tableau de contingence 18
II - 1 Description élémentaire du tableau 18
II - 2 Inférences sur les proportions 20
II - 2 . 1 Estimation d'une proportion 20
II - 2 . 2 Comparaison à une proportion théorique 20
II - 2 . 3 Comparaison de deux proportions 21
II - 3 Association entre variables ligne et colonne 22
II - 3 . 1 Indicateur global d'association le $\chi 2$ 22
II - 3 . 2 Analyse locale des associations 24

III Indépendance-Association entre variables nominales 24
III - 1 Le Test du $\chi 2$ 24
III - 1 . 1 Logique des Tests d’Hypothèse 24
III - 1 . 2 Démarche du test du $\chi 2$ (test de K. Pearson) 25
III - 1 . 3 Interprétation des résultats 27
Positionnement d’un $\chi 2$ observé 29
III - 1 . 4 Les conditions d’application du $\chi 2$ 29
III - 2 Mesures dérivées du $\chi 2$ d'indépendance 32
III - 2 . 1 Cas général d'une table rxc 32
III - 2 . 2 Cas d'une table 2x2 36
III - 3 Test exact de Fisher dans le cas 2x2 38
III - 4 Mesures orientées vers la prédiction 42

III - 4 . 1 Coefficient Lambda....42
III - 4 . 2 Coefficient d' Incertitude U....48

IV - Indépendance et association entre variables ordinales....51
IV - 1 Coefficients dérivés de la Formule de Daniels....51
IV - 1 . 1 Approche formelle....51
IV - 1 . 2 Coefficients de corrélation....51
IV - 1 . 3 Les coefficients de Kendall t et tb....52
IV - 2 Autres coefficients basés sur les concordances et discordances.54

V - Tests d'association de Cochran-Mantel-Haenszel....58

VI - Approche probabiliste dans le cas d'une table 2x2....60
VI - 1 Odds-ratio....60
VI - 2 Risque relatif....61
VI - 3 Analyse stratifiée....62
VI - 4 Lien avec les modèles LOGIT....65

VII. Curiosités....66
Barouf à Bombach....66
Le "chameau" de J.P FENELON....68

Conclusion....69

Annexes....70
A0. Exemples de sorties de PROC FREQ....70
A1. Tests et mesures appropriés selon les types de variables....74
A2. Historique de la polémique autour du test exact de FISHER....76
A3. Vocabulaire de la PROC FREQ....77

Bibliographie....78

PROC FREQ

Avant Propos

"La statistique est une science moderne et positive.
Elle met en lumière les faits les plus obscurs.

Ainsi, dernièrement, grâce à des recherches laborieuses, nous sommes arrivés à connaître le nombre exact de veuves qui ont passé le Pont-Neuf pendant le cours de l'année 1860.

Il y en avait treize mille quatre cent cinquante trois..., dont une douteuse."

extrait de la pièce "Les vicissitudes du capitaine TIC"- 16 Mars 1861-
de Eugène Labiche (1815-1888)

La procédure FREQ de S.A.S permet ainsi de dénombrer.

Mais en 1991, dénombrer ne suffit plus, et FREQ permet de faire beaucoup plus, au prix comme pour toute la Statistique, d'une sophistication logique et technique nécessitant une bonne culture statistique si on veut en comprendre les possibilités et les finesses.

Notre but est de vous mettre sur la voie en vous montrant les premiers pas. A vous de poursuivre.

Note: Ce document rassemble les présentations faites par les auteurs au cours des ateliers SAS de l'année universitaire 91-92, pour la version 6.04 de SAS.

Josiane Confais (ISUP) Yvette Grelet (CEREQ) Monique Le Guen (CNRS-IOF) Ateliers SAS-IRESCO 91-92

PROC FREQ

Introduction

Etant donné 2,3 ou n variables non quantitatives l'objectif est d'étudier une co-variation entre les variables.

Tout comme il n'existe pas de mesure unique permettant de porter un diagnostic sur l'intelligence d'un individu, **il n'existe pas** de mesure d'association à portée universelle.

Une mesure d'association répond à un point de vue. Elle a ses qualités et ses défauts, qu'une nouvelle mesure permet bien souvent de compenser, mais en perdant les qualités de la précédente.

Dans ce document nous avons essayé de montrer la logique qui a présidé à l'invention de la plupart des mesures d'association rapportées ici.

Ce fil logique permet de rendre le "catalogue" des mesures plus cohérent et plus appréhendable dans son ensemble.

Après avoir précisé la terminologie employée au chapitre I , et présenté le type de tableaux sur lequel nous voulons porter un diagnostic au chapitre II, nous passerons en revue le catalogue des tests et mesures d'association disponibles dans la procédure FREQ de SAS, selon les grands types de variables **nominales** au chapitre III , ou **ordinales** au chapitre IV.

Au chapitre V nous présenterons les tests d'association de Cochran-Mantel-Haenszel qui s'appliquent aux 2 types de variables.

Au chapitre VI nous aborderons l'approche probabiliste basée sur les Odds- Ratio et le modèle logit.

Afin de montrer les doutes que l'on doit avoir lors d'un test unique nous rapporterons une "curiosité", révélant les discordances des résultats selon les points de vue.

En annexe un historique sur le test de Fisher permettra au lecteur de conforter son opinion.

La Procédure FREQ de S.A.S permet:

- de produire des tableaux de fréquences à une dimension, et des tableaux croisés.
- d'analyser des **associations** entre **variables** dans des **tables de contingence.**

Au cours du premier chapitre nous allons préciser la terminologie élémentaire.

I - Terminologie.

I - 1 Variables

Dans FREQ les objets de base sur lesquels on travaille sont des variables.
Exemple: couleur='bleu'; ou couleur=1;

couleur est le nom de la variable, 'bleu' ou 1 est une valeur de la variable ou une modalité de la variable.

Les variables peuvent être segmentées selon plusieurs critères:

- approche par les qualités ou propriétés des variables
- approche liée aux techniques de traitement
- approche plus française

I - 1 . 1 Approche par les "qualités" ou "propriétés" d'une variable

- variable *caractère / numérique*
- variable *qualitative / quantitative*
- variable *discrète / continue*

Plutôt que de commenter nous allons donner des exemples:

- qualité ou propriété informatique

variable caractère	couleur='bleu'
variable numérique	couleur=1

• qualité ou propriété de mesurabilité

variable qualitative	sexe='féminin' ou '2'
variable quantitative	revenu=10 KF

• qualité de l'échelle de mesure utilisée

variable discrète	poids=10 kg
variable continue	poids=10,236...kg

I - 1 . 2 Approche liée aux techniques de traitement

Terminologie S.A.S

• variables **nominales**
• variables **ordinales**
• variables d' **intervalle**
• variables de **rapport**
• variables **catégorisées**

• **Variables nominales**

Exemple: sexe='Masculin' / 'Féminin' (variable caractère)
ou sexe= '1' / '2' (variable caractère)
ou sexe= 1 / 2 (variable numérique)

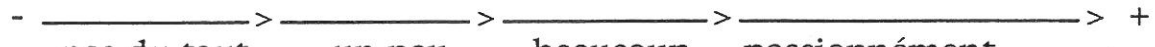
Ici il n'existe aucune notion de mesure ni de comparabilité entre les **modalités** de la variable sexe. Cette variable est dite nominale.

• **Variables ordinales**

Exemple: opinion = 'un peu'
= 'beaucoup'
= 'passionnément' etc.... 'pas du tout'

On peut positionner ces modalités les unes par rapport aux autres, en les représentant sur un axe:

Axe des opinions



1	5	10	15	<---- code
1	2	3	4	<---- rang

Les variables ordinales sont des variables pour lesquelles il existe une graduation. On peut donc affecter une valeur numérique en utilisant une échelle.

Cas particulier : l'échelle peut être un rang.

Pour ces variables les analyses statistiques doivent prendre en compte l'**ordre** des valeurs, et **non les distances** entre les valeurs numériques.

• **Variables d' intervalle** (*interval data*)

Exemples: température=10
10 est une valeur exprimée dans une certaine unité

Une température est une variable d'intervalle.

On parle de variable d'intervalle (*interval data*) lorsque la différence entre deux valeurs distinctes de la variable a un sens.
Dans le cas de la variable température, la différence de température entre 5° et 10° est comparable à la différence entre 15° et 20°.

• **Variables de rapport** (*ratio data*)

Exemple: revenu=10232 valeur exprimée dans une certaine unité

On parle de variable de rapport (*ratio data*) lorsque la mesure du rapport entre deux valeurs dictinctes de la variable a un sens.

Dans le cas de la variable revenu, un revenu de 10000 francs par exemple est 2 fois plus élevé, qu'un revenu de 5000 francs.

Dans le cas de la variable température, cela n'aurait aucun sens de dire que 30° est 2 fois plus élevé que 15°, c'est seulement beaucoup plus chaud.

De plus la valeur 0° n'a pas le même statut que 0 francs. Le 0° est une référence exprimée en Celsius qui transposée en Kelvin donnerait 273° Kelvin.
Tandis que 0 francs même traduit en Deutsche-Mark donnerait toujours 0 DM !

• **Variables catégorisées** (*categorical data*)

Le schéma de la page suivante résume ce que S.A.S appelle les Categorical Data.
Les variables catégorisées peuvent être soit des variables nominales, soit des variables ordinales, ou encore des variables, à l'origine, d'intervalle ou de ratio, qui ont été recodées en "tranches".

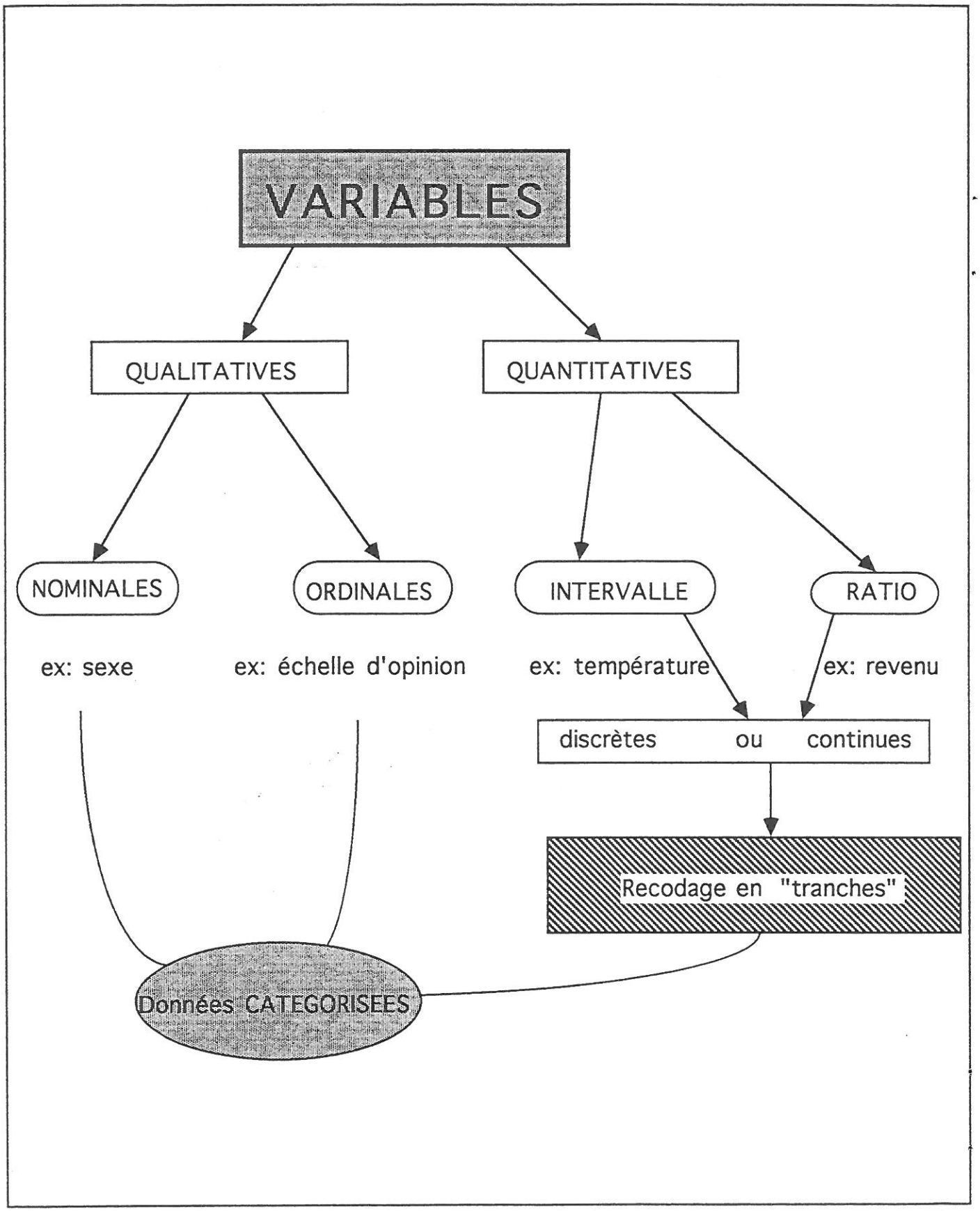
Remarque : Dans la procédure Freq de SAS la distinction entre variables d'intervalle et variables de rapport n'est jamais faite.

I - 1 . 3 Approche plus française

• **Variables nominales**
même définition que S.A.S

• **Variables ordinales**
variables numériques, discrètes avec un faible nombre de modalités et pouvant être ordonnées.

• **Variables "mesures"**
Ce sont des variables numériques à valeurs continues, pour lesquelles il existe une échelle de mesures.
Dans la littérature française la distinction entre variables d'intervalle et variables de rapport n'est pas toujours faite.


VARIABLES
QUALITATIVES
QUANTITATIVES
NOMINALES
ORDINALES
INTERVALLE
RATIO
ex: sexe
ex: échelle d'opinion
ex: température
ex: revenu
discrètes ou continues
Recodage en "tranches"
Données CATEGORISEES

I - 2 Tableaux de fréquences - Tables de contingence

A partir des objets de base: les variables, on peut constituer des tableaux. Le tableau le plus élémentaire que l'on puisse construire est un tableau d'effectifs dit aussi tableau de fréquences.

I - 2 . 1 Tableaux de fréquences pour 1 variable

Tableau d'effectifs ou de fréquences

Age	effectifs fréquences
1	22
3	25
6	12
7	13

Un tableau de fréquences associe à chaque valeur de la variable, ici l'âge, l'effectif ou fréquence absolue, totalisé dans l'échantillon observé.

Un tableau de fréquences apparaît comme une structure qui résume ou condense une partie de l'information contenue dans les données. Il permet d'avoir une vue synthétique de l'information apportée par la variable, mais en perdant les détails individuels.

Note: Pour des variables d'intervalle ou des variables ratio il est aussi possible d'avoir un tableau de fréquences à **condition** que la variable soit mesurée sur une **échelle discrète** et que le nombre d'occurrences de la variable ne soit pas trop élevé. Cependant pour ces deux types de variables il existe des méthodes d'analyse mieux adaptées.

Aussi selon les types de variables on utilisera certaines méthodes "résumé" que l'on trouve dans plusieurs procédures de SAS.

Types de variables et méthodes "résumé".

Variables	tableau de fréquences	Statistiques descriptives	
nominales	*		
ordinales	*	* <--	certaines statistiques
intervalle	*	*	
rapport	*	*	

V V

Proc FREQ Proc UNIVARIATE
Proc MEANS

La procédure FREQ concerne plutôt les variables nominales et ordinales.

I - 2 . 2 Tableaux de fréquences pour 2 ou n variables

Un tableau de fréquences croisant 2 variables encore appelé tableau de contingence, est un tableau qui croise les modalités x_i d'une variable ligne X, avec les modalités y_j d'une variable colonne Y.
Dans le schéma de la page suivante, la variable X peut prendre 4 modalités: A,B,C,D et la variable Y respectivement 5 modalités : 2,4,6,7,8.

Par convention on note:

n_{ij}	l'effectif de la cellule de rang i en ligne et de rang j en colonne.	
$n_{i.}$	l'effectif total sur la ligne i	$n_{i.} = \sum_{j=1}^{p} n_{ij}$
$n_{.j}$	l'effectif total sur la colonne j	$n_{.j} = \sum_{i=1}^{n} n_{ij}$
$n_{..}$	l'effectif total global	$n_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} nij$

Le tableau de base analysé par la procédure FREQ est un tableau qui croise 2 variables.

LES TABLEAUX

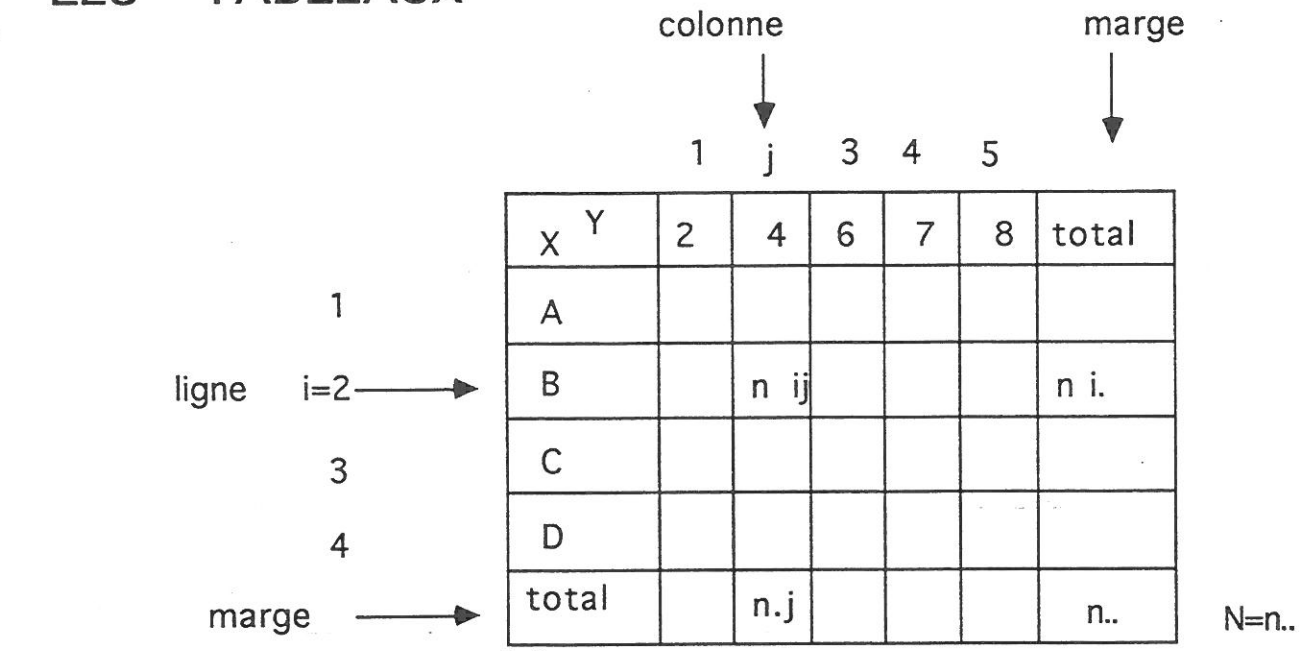
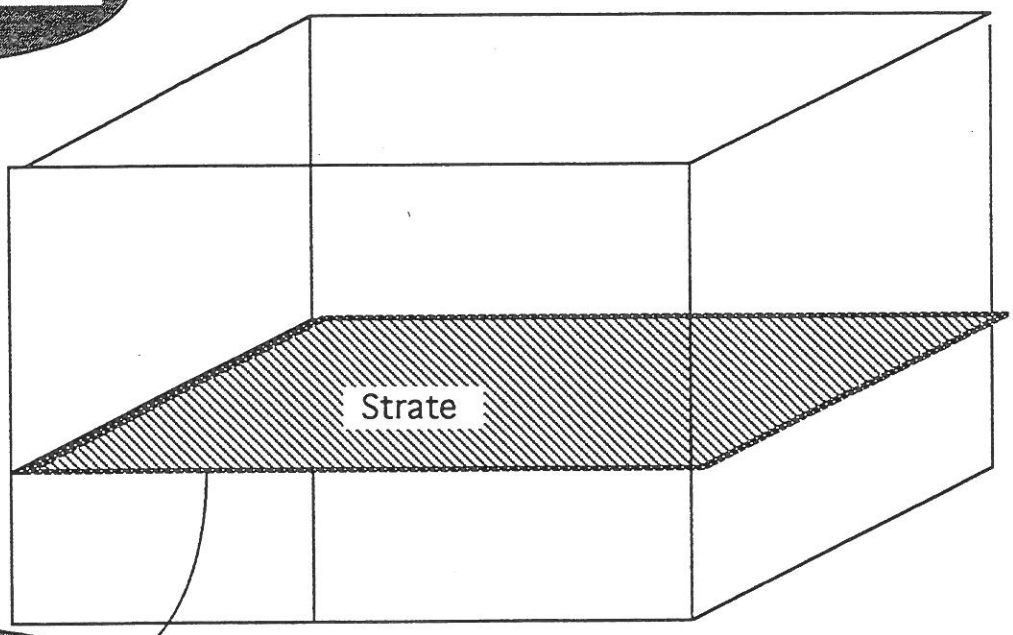


Tableau croisé à 2 dimensions
Tableau de fréquences
Tableau de contingence

tests d'indépendance

mesures d'association



Analyses stratifiées

Tableau croisé à n (3) dimensions

Si on croise plus de 2 variables, on obtient un hyper-tableau. Il faut alors effectuer des analyses stratifiées. Chaque section de dimension 2 définit une **strate.**

I - 3 Exemples de structure dans des tableaux

Un tableau de contingence permet de révéler une éventuelle structure.
Nous allons donner 3 exemples.

• Exemple 1: couleur des yeux et couleur des cheveux[1].

Soit un échantillon de 124 individus pour lesquels on a relevé la couleur des yeux et la couleur des cheveux.

cheveux yeux	blond	brun	noir	roux	Total
bleu	25	9	3	7	44
vert	13	17	10	7	47
marron	7	13	8	5	33
Total	45	39	21	19	124

Si on regarde en colonnes le tableau croisé ci-dessus, on remarque que la distribution des blonds est différente de la distribution des roux. Il y a des points d'accumulation (attractions) ou des vides (répulsions) à des endroits différents.
Les cheveux blonds et les yeux bleus sont souvent associés (25), comme le sont les cheveux bruns (17) avec les yeux marrons (13).
On parle alors d'une association entre modalités des lignes et modalités des colonnes.

Lecture: Il y a **une dépendance** entre la variable ligne et la variable colonne du tableau.

La question que l'on se pose, est *Comment mesurer cette dépendance* ?

[1] exemple cité par D. Schwartz p79.

• Exemple 2: Niveau des élèves selon la CSP du père

Niveau élèves CSP père	- -	-	+	++	Total
cadre	2	2	12	24	40
% en ligne	*5%*	*5%*	*30%*	*60%*	*100%*
employé	1	1	6	12	20
% en ligne	*5%*	*5%*	*30%*	*60%*	*100%*
Total	3	3	18	36	60
% en ligne	*5%*	*5%*	*30%*	*60%*	*100%*

La comparaison à partir de effectifs n'est pas facile lorsque les marges sont trés différentes (ici 40,20,60).
Pour rendre les profils de ligne homogènes, on compare les pourcentages.

Règle: Pour comparer 2 distributions on compare les %.

Lecture: On remarque alors que les profils sont dans le tableau précédent strictement identiques.

Il y a **indépendance** entre la variable ligne et la variable colonne du tableau

• Exemple 3 liaison particulière: l'association parfaite

Epreuve de course à pied

performances entraînement	<4'	4-5'	>5'
2 fois/semaine	0	0	13
4 fois / semaine	0	12	0
8 fois / semaine	15	0	0

Ici existe une association ou une liaison évidente : plus on s'entraîne plus on court vite.

C'est une structure qui conduit à une **liaison très forte,** lorsque le nombre de modalités est plus important.

Conclusion

Les 3 exemples précédents ont montré qu'il existe des "organisations" différentes dans les tableaux croisés. A partir d'un tableau croisé on peut se poser différentes questions.

Questions que l'on peut se poser sur un tableau de contingence

- Existe-t-il une structure dans le tableau?

- Quels liens existent entre la variable ligne et la variable colonne du tableau?

- Existe-il des points d'accumulation et/ou des vides?

- Le fait d'être dans la modalité i de la variable ligne permet-il de prévoir avec une certaine probabilité, l'appartenance à une modalité j de la variable colonne?

- La structure d'un tableau peut-elle être comparée à celle d'un autre tableau?

- Comment comparer deux structures?

Toutes ces questions peuvent trouver partiellement une réponse en ayant recours à des indicateurs globaux que sont les **mesures d'association** et les **tests d'indépendance.**

Nous avons vu qu'il y a différentes formes d'"organisation" dans un tableau croisé aussi, il y a différentes manières d'évaluer. D'où la multiplicité des mesures et des tests.

I - 4 . Mesures d'association - Tests d'indépendance

I - 4 . 1 Qu'est-ce qu'une association?

On dit qu'il y a association si la répartition des modalités d'une variable c'est à dire la distribution diffère selon les modalités de la deuxième variable.

Une **mesure d'association** indique avec quelle force deux variables sont reliées entre elles sur la base de l'échantillon étudié. Mais une mesure d'association ne permet pas d' inférer[2] sur la population dont est issu l'échantillon.

I - 4 . 2 Qu'est-ce qu'un test d'indépendance?

Le rôle d'un **test** est de fournir une significativité statistique, c'est à dire d'étendre à la population les résultats obtenus sur l'échantillon.

Un **test d'indépendance** sert à tester la vraisemblance d'une absence de liaison, dans une population, à partir d'un échantillon.

Il renseigne sur la **force de l'évidence** et non sur la force de l'association.

La difficulté est qu'un nombre unique ne peut représenter les différentes facettes des liaisons entre 2 variables. Chaque test, chaque mesure, a une capacité plus ou moins orientée à révéler un phénomène.

Aussi **l'utilisateur est-il totalement désorienté** devant la multiplicité des tests et des mesures proposés dans la Proc FREQ.

Pour un premier coup d'oeil, on trouvera en Annexe A0, la sortie listing de la Proc FREQ effectuée sur les 3 exemples du § I.3.

---

[2] Dans la démarche inférentielle, on considère un échantillon de N individus comme tiré d'une population plus large, sur laquelle on peut faire des déductions d'autant meilleures que l'échantillon est grand.
Dans le cadre descriptif, ces individus constituent l'univers observé; on y constate et on mesure les liaisons structurelles éventuelles.
C'est pourquoi Rouanet distingue les statistiques inférentielles, qui dépendent de la taille de l'échantillon, des statistiques descriptives.

RETENIR

Test d'Indépendance	=	maîtrise la proba d'erreur lors de l'inférence
Mesure d'association	=	mesure la force d'une liaison

Nous verrons bien cette différence d'objectif entre un test d'indépendance comme le χ^2 et une mesure d'association dans les chapitres suivants.

Pour le lecteur sceptique, citons une remarque de D. Schwartz :

" On notera qu'un χ^2 très élevé permet de rejeter avec une grande sécurité l'hypothèse d'indépendance, mais ne prouve pas que la liaison soit très forte, car lorsqu'il existe une liaison, la valeur de χ^2 augmente avec l'effectif de l'échantillon. Le χ^2 ne mesure pas l'intensité de la liaison, intensité qu'il est d'ailleurs difficile de définir."

Avec cette dernière phrase de D. Schwartz nous voilà prévenus pour la suite, l'intensité d'une liaison est difficile à définir.
C'est pour cette raison qu'il existe un grand nombre de tests et de mesures, et les plus courants sont disponibles dans la Proc FREQ.

Le chapitre suivant en dresse l'inventaire selon les champs d'application, c'est à dire le type des variables.

I - 5 Inventaire des Tests et Mesures

•Le χ^2 et ses dérivés

Champ d'application : Tous Types de variables traitées **nominales**	TEST
• Chi-Square	oui
• Likelihood ratio Chi-Square	oui
• Continuity Adj Square (TABLE 2*2)	oui
• Fisher's Exact test 1-tail /2-tail	oui
• Phi	
• Contingency Coef	
• Cramer's V	

• Mesures d'association: Lambda et coef d'incertitude

champ d'applications: tous types de variables traitées **nominales**
• Lambda Asymétrique C/R
• Lambda Asymétrique R/C
• Lambda Symétrique
• Coefficient d'Incertitude C/R
• Coefficient d'Incertitude R/C
• Coefficient Symétrique

• Autres Mesures

champ d'application : Variables au minimum **ordinales**	
• Gamma	
• Tau b de Kendall	
• Tau c de Stuart	
• DC/R de Somer	
• DR/C de Somer	
• Corrélation de Pearson	
• Corrélation de Spearman	
• Mantel-Haenszel Chi-Square	oui

champ d'application : tous types de variables	
• Cochran-Mantel-Haenszel : 3 statistiques	oui

champ d'application : variables dichotomiques pour les TABLE(2*2)
• relative risk
• odds ratio

II - Analyse d'un tableau de contingence

II - 1 Description élémentaire du tableau

Considérons le tableau ci-dessous (source : enquête d'insertion 1990 CEREQ/ DEP).
Le tableau croise en ligne la variable DIPLOME qui définit deux groupes:

- les jeunes sortis de l'école *sans diplôme* ,
- les *diplômés* d'un CAP ou d'un BEP,

et en colonne la variable SITUATION, qui définit quant à elle trois classes de jeunes selon qu'ils sont, au moment de l'enquête:

- au *chômage*,
- sur une *mesure* d'aide à l'insertion des jeunes,
- en *emploi* ordinaire.

TABLES DIPLOME*SITU;

ATELIER SAS PROC FREQ
SOURCE: ENQUETE D'INSERTION CEREQ-DEP
DE TERMINALE CAP OU BEP COMMERCE EN L.P. (SN, APPRENTIS EXCLUS)

TABLE OF DIPLOME BY SITU

DIPLOME SITU

Frequency Percent Row Pct Col Pct	CHOMAGE	MESURE	EMPLOI	Total
NON DIPL	54 10.84 36.99 30.68	52 10.44 35.62 34.90	40 8.03 27.40 23.12	146 29.32
DIPLOMES	122 24.50 34.66 69.32	97 19.48 27.56 65.10	133 26.71 37.78 76.88	352 70.68
Total	176 35.34	149 29.92	173 34.74	498 100.00

Le quadrant supérieur gauche du tableau indique (en anglais) le contenu de chaque case (i,j), à savoir:

- l'effectif n_{ij} ("Frequency")
- le pourcentage ("Percent") correspondant à $f_{ij} = n_{ij}/N$
- le pourcentage-ligne ("Row Pct") correspondant à $n_{ij}/n_{i.}$
- le pourcentage-colonne ("Col Pct") correspondant à $n_{ij}/n_{.j}$

° Ligne et colonne marginales

Sur la ligne "Total" on peut lire:
- les effectifs $n_{.j}$ des modalités de la variable colonne,
- les pourcentages ligne correspondant aux proportions $f_{.j}=n_{.j}/N$

C'est la **ligne marginale** donnant la distribution (le tri-à-plat) de la variable SITUATION sans distinction du diplôme.

Sur la colonne "Total", **colonne marginale**, on lit de même la distribution de la variable DIPLOME dans l'ensemble de la population (effectifs $n_{i.}$ et pourcentages colonne correspondant à $f_{i.}=n_{i.}/N$.

° Distribution conditionnelle

Pour une modalité i de la variable DIPLOME, **l'ensemble des pourcentages-ligne correspondant aux fréquences** $n_{ij}/n_{i.}$ aussi notées $f_{j|i}$ (qui se lit "f de j sachant i") donne la **distribution conditionnelle** de la variable SITUATION, c-à-d la distribution de cette variable, conditionnée par le fait qu'on se trouve dans la sous-population définie par cette modalité i. On parlera aussi du **profil** de la sous-population i.

De même pour une modalité j de la SITUATION: **l'ensemble des pourcentages-colonnes** correspondant aux f_{ij}, donne la distribution du DIPLOME **conditionnellement** à la modalité j de la SITUATION.

On pourra par exemple se demander si la distribution d'une colonne j diffère de la distribution observée dans l'ensemble de la population, c-à-d de la colonne marginale.

Le test approprié pour comparer une distribution observée à une distribution théorique est le **test du χ^2** dont on verra plus loin une définition. La répartition diplômés/ non diplômés est-elle la même chez les jeunes chômeurs que dans l'ensemble de la population des jeunes sortants ?

II - 2 Inférences sur les proportions

II - 2 . 1 Estimation d'une proportion

Le pourcentage de chômeurs dans l'échantillon est de 35,3%, soit une **proportion** p_0=0.353.

Ce chiffre donne une **estimation** de la vraie proportion p de chômeurs dans la population des jeunes sortants, avec une certaine marge d'erreur qu'on peut calculer aisément aux conditions que :

- l'échantillon soit issu d'un tirage aléatoire,
- p (théorique) ne soit pas trop proche de 0 ni de 1,
- N soit assez grand (>=30). *(Schwartz précise Np>=10 et Nq>=10)*

Sous ces conditions en effet, la proportion de chômeurs, observée dans un échantillon de taille N suivant une loi binomiale B(N,p), peut être approximée par **une loi normale de moyenne p et d'écart-type** $\sigma = \sqrt{\frac{p(1-p)}{N}}$.

Avec 5% de risque de se tromper, on peut dire que p est dans l'intervalle:

$[p_0 - 2s, p_0 + 2s]$, avec $s = \sqrt{\frac{p_0(1-p_0)}{N}}$.

C'est l'**intervalle de confiance** de la proportion p, calculé à partir de l'échantillon.

On remarque que plus N est grand, plus s est petit, donc aussi la largeur de l'intervalle.
Ici N=498, p_0=0.353, et 2s=0.043; soit $0.31 < p < 0.396$.

II - 2 . 2 Comparaison à une proportion théorique

Si on *suppose* que la proportion de chômeurs dans l'ensemble de la population est de .353 *(proportion théorique),* peut-on dire que le pourcentage observé chez les non diplômés (.37) s'en écarte "significativement"?

Ici N=146, p_0=0.37, et 2s=0.0799; soit $0.29 < p < 0.45$.
Au risque de 5% la réponse est non, puisque .353 tombe dans l'intervalle calculé ci-dessus.
Peut-être un échantillon plus grand aurait-il amené à conclure à une différence significative.

II - 2 . 3 Comparaison de deux proportions

Les proportions de chômeurs observées dans les échantillons correspondant aux non-diplômés et aux diplômés sont respectivement p_1=.37 et p_2=.347. Cet écart est-il "significatif"?

On teste l'hypothèse qu'il n'y a pas de différence entre p_1 et p_2, c-à-d. que les deux échantillons sont extraits de la même population dans laquelle on suppose que la proportion est p=.353.

Sous les conditions édictées plus haut, d'approximation normale de la loi binômiale, la différence p_1-p_2 suit une loi normale de moyenne 0 et d'écart-type σ=RAC[p(1-p)((1/N_1)+(1/N_2))].

Au risque de 5% on rejettera l'hypothèse si $|p_1-p_2| > 2\sigma$.

Ici N_1=146 , N_2=352 , p_1-p_2=.0233 , σ=.047 ,

====> l'écart n'est pas significatif.

II - 3 Association entre variables ligne et colonne

II - 3 . 1 Indicateur global d'association : le χ^2

Si on n'a pas conclu à une différence entre diplômés et non diplômés sur le taux de chômage, l'examen de l'ensemble du tableau laisse à penser qu'il y a pourtant un lien entre le diplôme et l'insertion de ces jeunes sur le marché du travail.
Pour tester ce lien, on va calculer le χ^2 ("CHI-2") associé au tableau: c'est la **somme**, sur toutes les cases (i,j) du tableau, **des carrés des écarts entre l'effectif observé n_{ij} et l'effectif théorique $n_{i.}n_{.j}/N$** qu'on aurait dans la case si les deux variables étaient indépendantes ; de plus, pour ne pas donner trop d'importance aux cases lourdes, on **divise l'écart-carré par l'effectif théorique** - comme pour un calcul de variance, on élève au carré pour que les écarts ne s'annulent pas.

$$\chi^2 = \Sigma_{ij} \frac{[\, n_{ij} - (n_{i.}n_{.j}/N)]^2}{(n_{i.}n_{.j}/N)}$$

L'**option CHISQ** de l'instruction TABLES édite, après le tableau croisé, la **valeur de cette statistique** (et d'autres informations qu'on verra plus loin) ainsi que le nombre de degrés de liberté, ou nombre de cases n_{ij} qu'il suffit de connaître pour en déduire toutes les autres connaissant $n_{i.}$ et $n_{.j}$; et aussi la probabilité associée (voir test de χ^2 plus loin).

```
TABLES DIPLOME*SITU/ CHISQ CELLCHI2 DEVIATION EXPECTED;

                    ATELIER SAS PROC FREQ
              SOURCE: ENQUETE D'INSERTION CEREQ-DEP
DE TERMINALE CAP OU BEP COMMERCE EN L.P. (SN, APPRENTIS EXCLUS)

                    TABLE OF DIPLOME BY SITU

DIPLOME          SITU

Frequency
Expected
Deviation
Cell Chi-Square
Percent
Row Pct
Col Pct          |CHOMAGE  |MESURE   |EMPLOI   |  Total
-----------------+---------+---------+---------+
NON DIPL         |      54 |      52 |      40 |    146
                 |  51.598 |  43.683 |  50.719 |
                 |  2.4016 |  8.3173 |  -10.72 |
                 |  0.1118 |  1.5836 |  2.2653 |
                 |   10.84 |   10.44 |    8.03 |  29.32
                 |   36.99 |   35.62 |   27.40 |
                 |   30.68 |   34.90 |   23.12 |
-----------------+---------+---------+---------+
DIPLOMES         |     122 |      97 |     133 |    352
                 |   124.4 |  105.32 |  122.28 |
                 |  -2.402 |  -8.317 |  10.719 |
                 |  0.0464 |  0.6568 |  0.9396 |
                 |   24.50 |   19.48 |   26.71 |  70.68
                 |   34.66 |   27.56 |   37.78 |
                 |   69.32 |   65.10 |   76.88 |
-----------------+---------+---------+---------+
Total            |     176 |     149 |     173 |    498
                 |   35.34 |   29.92 |   34.74 | 100.00

                STATISTICS FOR TABLE OF DIPLOME BY SITU

Statistic                          DF       Value        Prob
-------------------------------------------------------------
Chi-Square                          2       5.604       0.061
Likelihood Ratio Chi-Square         2       5.681       0.058
Mantel-Haenszel Chi-Square          1       2.376       0.123
Phi Coefficient                             0.106
Contingency Coefficient                     0.105
Cramer's V                                  0.106

Sample Size = 498
```

II - 3 . 2 Analyse locale des associations

Chaque case contribue au χ^2, d'autant plus fortement qu'il y a attraction (écart positif) ou répulsion (écart négatif) entre les modalités i et j.

Les **options** EXPECTED, DEVIATION et CELLCHI2 de l'instruction TABLES donnent dans chaque case respectivement:

- l'**effectif attendu** (théorique) dans la case sous l'hypothèse d'indépendance,
- la valeur (signée) de l'**écart entre effectifs observé et attendu,**
- la **contribution** de la case ("cell") au χ^2.

EXPECTED (attendu ou théorique) = $N f_{i.} f_{.j} = N (n_{i.}/N)(n_{.j}/N) = n_{i.}n_{.j}/N$
DEVIATION (observé - théorique) = $n_{ij} - n_{i.}n_{.j}/N$

$$\text{CELLCHI2 (contribution au } \chi^2) = \frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N}$$

Pour sélectionner les cases les plus contributives on se basera sur le CELLCHI2 moyen (χ^2 divisé par le nombre de cases).

Ces informations sont très précieuses pour analyser finement la structure du tableau. Si le tableau est de grande dimension, la lecture peut cependant en être difficile et on gagnera à tenter une analyse des correspondances ...

III Indépendance-Association entre variables nominales

III - 1 Le Test du χ^2

Avant de présenter la démarche du test du χ^2, rappelons ce qu'est la logique d'un test d'hypothèse.

III - 1 . 1 Logique des Tests d'Hypothèse

Pour tester une hypothèse il faut :

- formuler l'hypothèse nulle H_0 et l'hypothèse alternative.
- élaborer une statistique de test.

• connaître la distribution d'échantillonnage de cette statistique.
Cette information est fournie par les statisticiens : loi Normale, loi de Student-Fisher, loi du $\chi 2$ etc.

• connaître les conditions de validité de ce test pour chercher à les vérifier

• interpréter le test en lisant la **p-value** associée à la statistique de test. La p-value mesure le **risque de rejeter à tort** H_0.

III - 1 . 2 Démarche du test du $\chi 2$ (test de K. Pearson)

Dans le cas d'une recherche d'indépendance entre la variable ligne et la variable colonne d'un tableau de contingence, on va comparer la distribution statistique observée dans l'échantillon, à une distribution théorique.

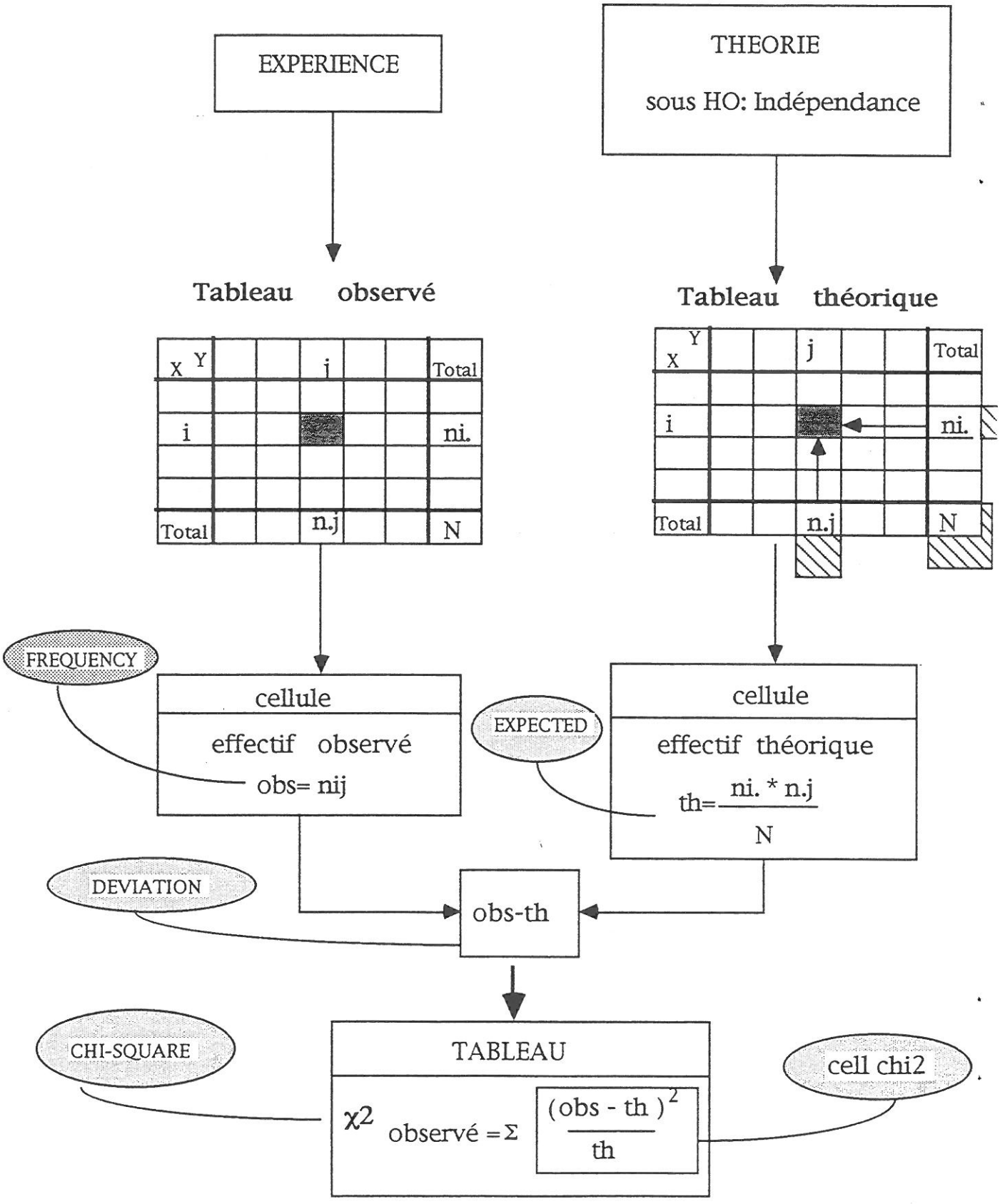
Cette distribution théorique est celle que l'on doit avoir **si les 2 variables sont indépendantes**, c'est à dire sous l'hypothèse H_0.

On veut savoir si les écarts entre ces deux distributions sont imputables aux fluctuations d'échantillonnage, ou si au contraire, les écarts sont trop importants pour que l'on puisse accepter l'hypothèse H_0.

Le schéma de la page suivante montre le parallèle qui est fait entre l' **Expérience**, partie gauche du graphique et la **Théorie**, partie droite du graphique.

Note : les "bulles" font référence au vocabulaire SAS en sortie de la Proc FREQ.

Le $\chi 2$


EXPERIENCE
THEORIE
sous HO: Indépendance
Tableau observé
Tableau théorique
X Y
i
Total
ni.
n.j
N
j
FREQUENCY
cellule
effectif observé
obs= nij
EXPECTED
effectif théorique
th= ni. * n.j / N
DEVIATION
obs-th
CHI-SQUARE
TABLEAU
χ2 observé = Σ (obs - th)2 / th
cell chi2

III - 1 . 3 Interprétation des résultats

La statistique appelée $\chi 2$ **observé**, porte ce nom par ce que les statisticiens nous ont dit qu'elle suit une loi du $\chi 2$, mais on aurait très bien pu l'appeler: **Distance.**
Sous Ho cette distance est nulle: **$\chi 2$ observé =0**

Comme la valeur du $\chi 2$ observé correspond à une distance :

• Plus cette **distance** est **grande** plus on s'écarte de l'hypothèse nulle.
Dans une telle situation on rejette l'hypothèse Ho : **indépendance.**
On vient de mettre en évidence une **association** entre la variable ligne et la variable colonne.

Conclusion : **L'expérience a contredit l'hypothèse.**

• Si la **distance** est **petite** on conclut que l'on n'a aucune preuve d'association ce qui ne veut pas dire qu'il y a non-association.

Comme dans tous les tests on dit qu'*on ne rejette pas l'hypothèse nulle*.

Note: *Ce qui ne veut pas dire qu'on accepte l'hypothèse nulle.*

Inconvénient de cette distance:

L'inconvénient de cette distance, est qu'elle est fonction de la taille du tableau c'est à dire du nombre de lignes et du nombre de colonnes.
On peut cependant s'en affranchir en tenant compte des degrés de liberté de la table.

Les statisticiens nous disent que le nombre de degrés de liberté d'une table de contingence à "r" lignes et "c" colonnes est donné par:

ddl= (r-1) * (c-1)
DF

DF : Degree of Freedom

Les statisticiens ont montré que la statistique du test suit une loi du CHI-2 à (r-1)(c-1) degrés de liberté, sous certaines conditions dites *asymptotiques* (N grand, et au moins 5 comme effectif théorique (ni. * n.j)/N dans chaque case).

Comme pour tous les tests de SAS, à chaque valeur d'une statistique calculée (observée) SAS associe la probabilité appelée **p-value.** Cette p-value résulte du calcul automatique fait dans SAS en utilisant la fonction PROBCHI.

p-value = 1 - PROBCHI (χ2 obs , ddl)

Raisonnement sur la p-value:

Si p-value petit ====> rejet de Ho ====> association Si p-value grand ====> non rejet de Ho

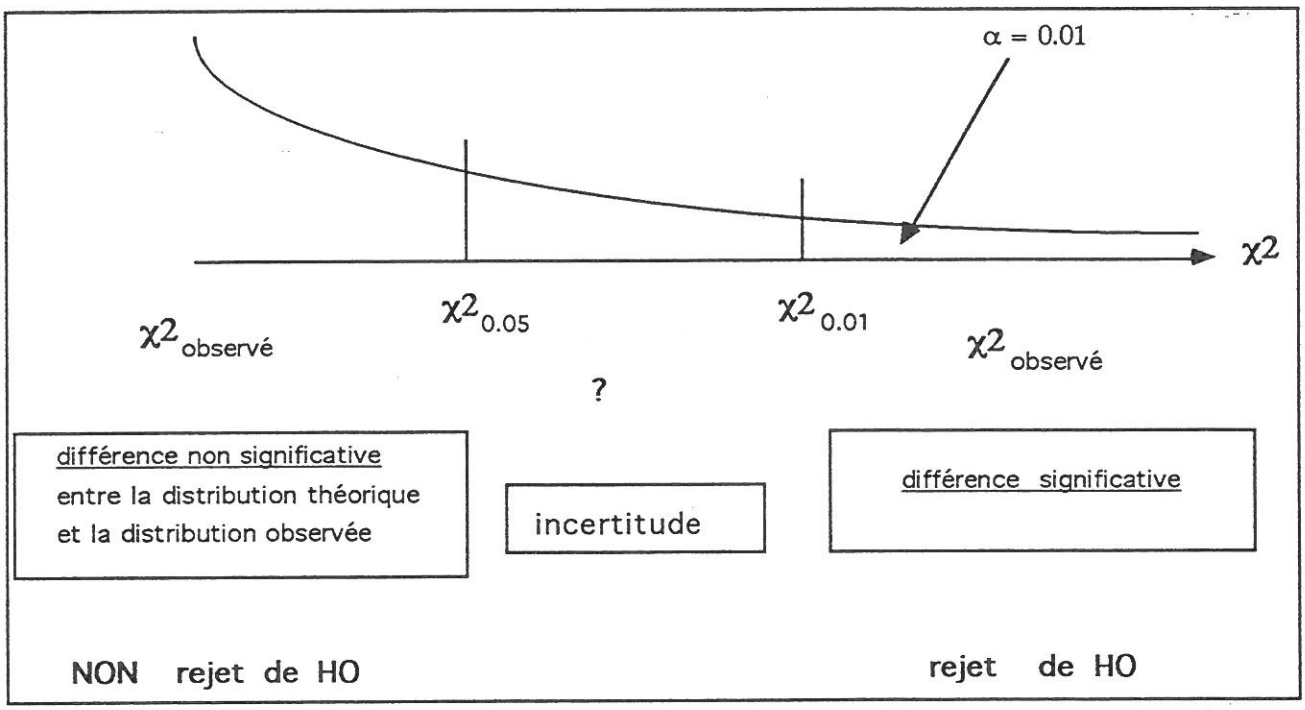
Par habitude on prend comme bornes de p-value : α=0.05 ou α=0.01.

Que signifie la p-value?

•La p-value représente la probabilité d'observer une statistique au moins égale au χ2 obs avec une hypothèse Ho vraie.

• En clair **p-value=0.05** signifie que l'on a 5 chances sur 100 de se tromper en rejetant l'hypothèse d'indépendance. On dit aussi que c'est le risque de se tromper, que l'on accepte de prendre.

Positionnement d'un $\chi 2$ observé



III - 1 . 4 Les conditions d'application du $\chi 2$

Conditions de validité

•Le test du $\chi 2$ peut s'appliquer sur **tous les types de variables**, variables nominales, variables ordinales, variables d'intervalle ou de ratio. Cependant pour les 3 derniers types il existe d'autres indicateurs ou mesures d'association mieux adaptés.

•Les effectifs théoriques dans toute les cases doivent être **au moins égaux à 5** pour que le test du $\chi 2$ soit valide. Cette règle fait a peu près l'unanimité des théoriciens de la Statistique.

Si cette règle n'est pas vérifiée SAS le signale. On peut alors procéder à des regroupements de modalités, si cela est possible et a un sens, ou utiliser le test exact de Fisher.

• Pour appliquer le test du $\chi 2$ on fait la supposition que les proportions marginales dans la population totale sont les mêmes que celles observées sur l'échantillon.

• Le test du $\chi 2$ **ne s'applique que dans un cadre d'inférence.** C'est à dire lorsque l'on dispose d'un échantillon, et que l'on souhaite étendre les résultats observés à la population totale.

Si l'échantillon recouvre toute la population, faire un test du $\chi 2$ n'a pas de sens.

• Pour l'analyse des tableaux obtenus en seconde main les tests du $\chi 2$ doivent être effectués sur les tableaux **avant redressement**.

•Le test du $\chi 2$ **est sensible à la taille de l'échantillon**.

Grosbras JM met en garde contre cette assertion de certains praticiens malicieux :

"Il y a toujours moyen d'obtenir un $\chi 2$ significatif, (c'est à dire dépassant les valeurs critiques de la table à 5% ou 1%), c'est d'avoir un gros échantillon".

ATTENTION

Les 3 encadrés ci-après, sont à mémoriser.

La statistique du $\chi 2$ ne mesure pas la force de la liaison

Preuve

Si on multiple tous les effectifs des cellules d'un tableau par 100 par exemple, alors la statistique du $\chi 2$ est multipliée aussi par 100 et pourtant la force de la liaison n'a pas changé.

ASSOCIATION ne signifie pas CAUSALITE

Exemple:

Complications lors d'un accouchement, en présence ou absence de médecin.

Complications avec médecin	oui	non	Total
oui	60	440	500
non	20	480	500
Total	80	920	1000

$\chi 2$ obs = 21 p-value= 0.000 ==> significatif ==> rejet de l'indépendance

Interprétation sans bon sens:

Les complications sont plus fréquentes en présence d'un médecin.

Le médecin est-il la cause?

En fait, les 2 groupes "avec médecin" et "sans médecin" sont constitués de cas inégalement graves . Les 2 groupes ne sont pas comparables

Autre exemple: "70% des individus meurent au lit."
===> pour vivre vieux, vivons debout!

L'interprétation causale nécessite des groupes comparables

Remède

Vers les plans expérimentaux ..... A suivre.....

III - 2 Mesures dérivées du χ^2 d'indépendance

III - 2 . 1 Cas général d'une table rxc

Ces mesures sont obtenues par l'option CHISQ de l'instruction TABLE.

Rappel des notations : 2 variables nominales X à r modalités et Y à c modalités (r=row nombre de lignes et c=column nombre de colonnes).

table (nij) avec $N = \sum_{ij} nij$ effectifs observés

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(nij - (ni.\ n.j / N))^2}{(ni.\ n.j / N)}$$

L'hypothèse H0 est l'indépendance $\leftrightarrow$ nij = (ni. n.j) / N

Différentes mesures

* Rappel des propriétés du χ^2 d'indépendance :

Le χ^2 de Pearson prend des valeurs positives.
Il est nul sous H0 ; en cas d'association "parfaite" entre les 2 variables, il prend une valeur qui dépend de N et des nombres de modalités : N x minimum (r-1, c-1).
Il permet de tester l'hypothèse d'indépendance à l'aide d'une statistique de test qui suit aymptotiquement (c'est-à-dire si N est grand) une loi du CHI-2 à (r-1) (c-1) degrés de liberté.

Pour remédier à l'influence de N dans le calcul, Pearson a proposé le coefficient phi^2.

*** phi^2 :**

$$\phi^2 = \chi^2 / N$$

Le ϕ^2 ne dépend pas de la taille de l'échantillon. C'est une statistique descriptive. Il prend lui aussi la valeur 0 sous indépendance ; sous association parfaite, il vaut minimum (r-1, c-1).

SAS donne $\phi = \sqrt{\phi^2}$ avec un signe : positif si dans la table l'association se retrouve suivant la diagonale, négatif si celle-ci se retrouve sur "l'anti-diagonale".

Pour obtenir une autre mesure qui ne dépende pas de l'effectif total N, et soit plus petite que 1, Pearson a proposé le coefficient de contingence C.

* C Contingency coefficient :

Il est calculé suivant la formule suivante :
$C = RAC (\chi^2 / (N + \chi^2)) = RAC (\phi^2 / (1+\phi^2))$.

Il est compris entre 0 et 1, mais la valeur 1 n'est pas atteinte : s'il vaut encore 0 sous indépendance, sa valeur sous association parfaite dépend de r et c (si r=c, c'est RAC (1-1/r)), et peut être très éloignée de 1.
La loi de C n'étant pas connue, on ne peut l'utiliser pour tester l'indépendance.

Pour obtenir un coefficient qui puisse atteindre la valeur 1, Cramer a proposé le coefficient V.

* V de Cramer :

Il est obtenu par la formule suivante :

$V = \phi / RAC (minimum(r-1 , c-1))$.

Ses valeurs possibles sont donc comprises entre -1 et +1 ; il vaudra 0 sous indépendance et +1 ou -1 sous association parfaite.

C'est donc une mesure d'association ressemblant au coefficient de corrélation linéaire entre variables quantitatives. On ne connait pas la loi suivie par V, donc on ne peut pas l'utiliser pour tester l'indépendance.

Remarques sur les mesures dérivées du χ^2 d'indépendance :

1/ ces mesures sont symétriques en lignes et colonnes ;

2/ elles sont invariantes par permutations de 2 lignes et/ou 2 colonnes ; il faut donc choisir d'autre mesures si les lignes et/ou colonnes sont ordonnées (variables ordinales) comme on le verra au chapitre IV.

3/ elles dépendent des valeurs r et c, c'est-à-dire de la taille de la table (sauf V) : on ne peut donc comparer 2 mesures que pour des tables de dimensions voisines ;
ϕ^2 , V et C ne dépendent pas de l'effectif total N. Ce sont des statistiques descriptives.

4/ elles ne sont pas marginalement invariantes (c'est-à-dire changent si les lignes et/ou les colonnes sont multipliées par des constantes).
Le test d'indépendance associé est fait à marges fixées qui sont déterminées par celles observées (ni./N et n.j/N).

SAS donne dans l'option CHISQ deux autres mesures qui ont la propriété de suivre des lois du CHI-2 , mais qui ne sont pas dérivées du χ^2 d'indépendance.

*** G^2 likelihood ratio :**

Il s'agit de la statistique du test d'indépendance construite à partir du Rapport de Vraisemblance Maximum de l'échantillon (RVM) .

$$G^2 = -2 \text{ Log (RVM)}$$
$$G^2 = 2 \sum_{ij} nij \text{ Log } (nij / (ni. \ n.j)/N)$$

Ses valeurs possibles sont positives. Il vaut 0 sous indépendance. Asymptotiquement (si N grand), il suit une loi du CHI-2 à (r-1) x (c-1) degrés de libertés, et donc peut être utilisé pour tester l'indépendance.

Remarque : G^2 est proche du χ^2 d'indépendance si on est "près" de l'indépendance H_0, ou si N est grand.

*** Qmh dit Mantel-Haenszel CHI-2 :**

Qmh mesure l'association entre les variables X et Y. Il est calculé à partir du coefficient de corrélation linéaire ρ entre les variables dont les modalités sont codées numériquement (ce codage est défini par l'option SCORES) : il n'est donc à utiliser que si les variables sont ordinales.

$$Qmh = (N-1) \ \rho^2.$$

Il vaut 0 sous indépendance et ((N-1)/N) x minimum (r-1, c-1) sous association parfaite.
Il a la propriété de suivre une loi du CHI-2 à 1 degré de liberté quelle que soit la taille de la table.

On le retrouvera au chapitre V.

Exemple : enquête d'insertion CEREQ-DEP (cf. § II - 1)

X = diplômes à deux modalités
Y = situations à trois modalités.

Ici r = 2 et c = 3 donc V = ϕ.
De plus ϕ est petit donc C est proche de ϕ.

STATISTICS FOR TABLE OF DIPLOME BY SITU

Statistic	DF	Value	Prob
Chi-Square	2	5.604	0.061
Likelihood Ratio Chi-Square	2	5.681	0.058
Mantel-Haenszel Chi-Square	1	2.376	0.123
Phi Coefficient		0.106	
Contingency Coefficient		0.105	
Cramer's V		0.106	

Sample Size = 498

III - 2 . 2 Cas d'une table 2x2

Les variables X et Y sont dichotomiques. La table devient :

n11	n12	| n1.
n21	n22	| n2.
n.1	n.2	N

La formule du χ^2 d'indépendance se simplifie alors :

$$\chi^2 = N \frac{(n11n22 - n12n21)^2}{n1.\ n2.\ n.\ 1n.2} = N\,\phi^2$$

Le χ^2 prend des valeurs comprises entre 0 (indépendance) et N (association parfaite). Le test d'indépendance se fait avec une loi du CHI-2 à 1 degré de liberté.

Le ϕ^2 prend ses valeurs dans [0, 1]. Il est strictement égal au carré du V de Cramer (soit $\phi = V$).

Le coefficient C de contingence a des valeurs comprises entre 0 (indépendance) et $\sqrt{1/2}$ (environ 0.707) sous association parfaite.

Remarque : les variables étant dichotomiques, $\phi^2 = \rho^2$ coefficient de corrélation linéaire, quel que soit le codage numérique associé aux modalités.

* Qc continuity adjusted χ^2 :

Pour corriger le fait qu'on applique une loi continue (le CHI-2) à une quantité qui est discontinue, Yates a proposé une correction au calcul du χ^2 d'indépendance suivant la formule suivante :

$$Qc = N \frac{(|\ n11\ n22\ -\ n12\ n21\ |\ -\ N/2)^2}{n1.\ n2.\ n.1\ n.2}$$

si $|n11\ n22 - n12\ n211\ | > N/2$

$Qc = 0$ sinon.

Qc a les mêmes propriétés que le χ^2 d'indépendance.

* Qmh Mantel-Haenszel CHI-2 :

Ici il peut être calculé indifféremment par la formule :

$(N-1)\ \rho^2$ ou $((N-1)/N)\ \chi^2$.

il vaut 0 sous indépendance et (N-1) sous association parfaite. Le test d'indépendance utilise une loi du CHI-2 à 1 degré de liberté.

Exemple : Radelet (1981) cité dans Agresti (1990)

Verdict de 326 procès en Floride de 1976 à 1977.
X = race de l'accusé à deux modalités blanc / noir
Y = verdict de mort à deux modalités oui / non

etude d'une table 2x2 : RACE / VERDICT

TABLE OF RACE BY MORT

RACE / MORT	OUI	NON	Total
BLANC	19	141	160
Percent	5.83	43.25	49.08
Row Pct	11.88	88.12	
Col Pct	52.78	48.62	
NOIR	17	149	166
Percent	5.21	45.71	50.92
Row Pct	10.24	89.76	
Col Pct	47.22	51.38	
Total	36	290	326
	11.04	88.96	100.00

STATISTICS FOR TABLE OF RACE BY MORT

Statistic	DF	Value	Prob
Chi-Square	1	0.221	0.638
Likelihood Ratio Chi-Square	1	0.221	0.638
Continuity Adj. Chi-Square	1	0.086	0.769
Mantel-Haenszel Chi-Square	1	0.221	0.638
Fisher's Exact Test (Left)			0.741
(Right)			0.384
(2-Tail)			0.725
Phi Coefficient		0.026	
Contingency Coefficient		0.026	
Cramer's V		0.026	

III - 3 Test exact de Fisher dans le cas 2x2

Le test exact de Fisher s'obtient dans FREQ avec l'option CHISQ si la table est 2x2 (sinon ajouter l'option EXACT).

Il s'applique quand les conditions de validité du test du χ^2 d'indépendance sont violées : si N est petit (N < 20) ou s'il existe des cases d'effectif < 5.
Il s'applique également au cas où le test donne une probabilité critique voisine du seuil 5 % (donc la conclusion du test est "délicate").

Théorie : il s'agit d'un test à marges fixées.
(n1., n2.) et (n.1, n.2) étant fixés, on peut calculer sous l'hypothèse d'indépendance H0 la probabilité d'obtenir le tableau de contingence :

a	b	n1.
c	d	n2.
n.1	n.2	N

(a b c et d effectifs observés)

Remarque : si l'effectif en case (1,1) est donné, les 3 autres sont déterminés puisque les marges sont fixées.

On montre que n11 suit une loi hyper-géométrique H (N, n.1, n1./N) (tirage sans remise de n.1 individus parmi N, dans une population où des individus ayant un caractère particulier sont en proportion n1./N).

$$\text{Prob (n11 = a)} = \frac{n1.!\ \ n2.!\ \ n.1!\ \ n.2!}{N\ \ a!\ \ b!\ \ c!\ \ d!}$$

On peut donc calculer de façon exacte la probabilité du test.
SAS permet de faire le test bilatéral et le test unilatéral.

Dans le test bilatéral, l'hypothèse nulle H0 est l'indépendance, l'alternative H1 étant "non indépendance" qui peut se dire "les cases du tableau ne sont pas chargées comme sous l'indépendance".
Dans le cas unilatéral, on fixe une alternative H1 "non indépendance" du genre "les cases sont plus (ou moins) chargées que sous l'indépendance".

Choix du "sens" de l'alternative H1 :

$$\text{soit delta (nii)} = nii - \frac{ni.\ n.i}{N} \quad [= -(nij - \frac{ni.\ n.j}{N}) \text{ si } i \neq j]$$

= "écart à l'indépendance"

delta = 0 caractérise l'indépendance H_0
(case aussi chargée que sous indépendance)

test gauche = l'alternative à H_0 est delta < 0 (moins chargé)
droite = l'alternative à H_0 est delta > 0 (plus chargé)
bilatéral = l'alternative à H_0 est delta ≠ 0 (différent)

On regardera donc delta (a) pour choisir l'alternative, c'est-à-dire la différence (effectif observé) moins (effectif attendu) de la case (1,1) [On obtient cette différence par l'option DEVIATION] .
On peut aussi regarder celle des probabilités des 2 tests unilatéraux qui est plus petite que 0.5 : c'est l'alternative à choisir.

Calcul des probabilités des 3 tests :
Nous allons préciser la "région critique" de chacun des 3 tests:

- test gauche "left" Prob1 = somme des probabilités des tables telles que l'effectif n11<= a ;
- test droite "right" Prob2 = idem pour les tables n11 >= a ;
- test bilatéral "2-tail" Prob3 = idem pour les tables dont la probabilité est inférieure ou égale à celle de la table observée.

Les calculs peuvent être longs puisqu'il faut dénombrer les tables répondant à l'hypothèse alternative, puis en calculer les probabilités par la loi hypergéométrique.

D'autre part la loi hypergéométrique n'est pas symétrique, sauf si les marges des 2 lignes et des 2 colonnes sont égales, ou si N est grand (N>=20) car alors Prob (n11=a) = 0.
On n'a donc pas en général : Prob3 = 2 Prob1 (ou 2 Prob2).
Par contre on a toujours : Prob1 + Prob2 = 1 + Prob(n11=a).

Dans le cas du test du χ^2 d'indépendance, on l'applique si N est grand, et donc alors la probabilité du test unilatéral est la moitié de celle du test bilatéral qui est donnée par SAS.

Remarque à propos de SAS version 5 (d'après O. Sautory) : il ne donnait que 2 probabilités, calculées ainsi :

Prob (2-tail) = 2 Prob(1-tail) si Prob (1-tail) < 0.5
= 2 (1-Prob(1-tail)) si Prob (1-tail) > 0.5

Le calcul était donc faux sauf si N est grand, ou si les marges lignes et colonnes sont égales.

Nota-bene : Dans SAS, les calculs sont aussi possibles si r ou c > 2, mais ils sont très longs : à éviter si rxc > 5. D'autre part, on a du mal à concrétiser l'alternative dans ce cas, car il faut ici plus d'une case pour pouvoir déterminer totalement la table.

Exemple détaillé issu de SIEGEL " Non paramétric statistics for the behavioral sciences", repris par Sautory :

table

	-	+	n.j
G1	1	6	7
G2	4	1	5
ni.	5	7	12

L'effectif sous indépendance serait 7x5/12 = 2.916 ; delta (n11) vaut -1.916 donc delta (n11) est négatif : l'alternative est "case moins chargée que sous indépendance" c'est-à-dire qu'il faut faire un test unilatéral "gauche".

La loi hypergéométrique est la loi H (12, 5, 58.33%) c'est-à-dire de tirage sans remise de 5 individus parmi 12, ayant une caractéristique en proportion 7/12.

Pour effectuer le test "gauche", on va calculer les probabilités des tables pour lequelles la case (1,1) est au plus aussi chargée que celle observée :

Prob(n11 = 1) = 0.044
Prob(n11 = 0) = 0.001 ⇒ somme = 0.045 = Prob-left.

Pour effectuer le test bilatéral, on va dénombrer parmi toutes les tables possibles celles dont la probabilité est inférieure ou égale à celle de la table observée :

Prob (n11=0) = 0.001 *	Prob (n11=3) = 0.442
Prob (n11=1) = 0.044 * (table observée)	Prob (n11=4) = 0.221
Prob (n11=2) = 0.265	Prob (n11=5) = 0.027 *

* = table à choisir pour l'alternative

⇒ Prob (2-tail) = 0.001 + 0.044 + 0.027 = 0.072

(à comparer à la probabilité du test du χ^2 (non valide) qui est 0.023)

Exemple : extrait de Siegel, repris par O. Sautory.

TABLE OF GROUP BY VAL

GROUP VAL

Frequency / Expected / Percent / Row Pct / Col Pct	-	+	Total
G1	1	6	7
	2.9167	4.0833	
	8.33	50.00	58.33
	14.29	85.71	
	20.00	85.71	
G2	4	1	5
	2.0833	2.9167	
	33.33	8.33	41.67
	80.00	20.00	
	80.00	14.29	
Total	5	7	12
	41.67	58.33	100.00

STATISTICS FOR TABLE OF GROUP BY VAL

Statistic	DF	Value	Prob
Chi-Square	1	5.182	0.023
Likelihood Ratio Chi-Square	1	5.555	0.018
Continuity Adj. Chi-Square	1	2.831	0.092
Mantel-Haenszel Chi-Square	1	4.750	0.029
Fisher's Exact Test (Left)			4.55E-02
(Right)			0.999
(2-Tail)			7.20E-02
Phi Coefficient		-0.657	
Contingency Coefficient		0.549	
Cramer's V		-0.657	

Sample Size = 12
WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

III - 4 Mesures orientées vers la prédiction

III - 4 . 1 Coefficient Lambda (λ)

Approche de GUTTMANN (1941) et de GOODMAN & KRUSKAL (1954)

Lambda est une mesure d'association pour des tableaux croisés, les variables étant traitées comme **nominales.**

Il existe 3 formes de coefficicient λ

1. λ asymétrique Y|X *qui se lit Y sachant X*
2. λ asymétrique X|Y
3. λ symétrique

1. λ asymétrique Y|X

IDEE

On veut essayer de pronostiquer la modalité de Y prise par un individu tiré au hasard parmi les N individus, et ceci dans 2 situations:

1. sans aucune information complémentaire
2. en connaissant la modalité i de la variable X

•1 Aucune information complémentaire

En l'absence de toute information on choisira la **modalité de Y la plus fréquente** sur la marge.
C'est la meilleure stratégie puisqu'en faisant ce choix on minimise la probabilité de se tromper.

Y max	-------->	en effectif	$\max_j (n.j)$
	-------->	en fréquence	$\max_j (n.j / N)$

avec une **probabilité d'erreur**

p1:	Proba (erreur(Y)) = $1 - \max_j (n.j / N)$

•2 On connaît la modalité prise sur la variable X et on veut pronostiquer celle prise par la variable Y.

X: joue le rôle de variable indépendante
Y: joue le rôle de variable dépendante.

Si on connaît la modalité i de X pour l'individu tiré au hasard, on choisira la **modalité de Y dont la fréquence est maximum** (fréquence max sur la ligne i) , toujours dans le but de minimiser la probabilité d'erreur.

Y max |Xi ------> en fréquence $\max_j$ (nij / ni.)

On démontre que la probabilité d'erreur de Y sachant X pour toutes les modalités de X est:

p2: Proba (erreur Y |X) $= 1 - \sum_i \max_j$ (nij / N)

A partir de ces deux proba
p1: proba d'erreur sans information sur X
p2: proba d'erreur si on connaît X

on définit le ratio appelé Lambda asymétrique Y|X

$$\lambda Y|X = (p1 - p2) / p1$$

$$\lambda Y|X = \frac{(\sum_i \max_j nij) - \max_j n.j}{(N - \max_j n.j)}$$

Lecture de la formule de $\lambda Y|X$

$(\sum_i \max_j nij)$: représente la somme sur toutes les lignes des valeurs maximum des effectifs des cellules sur les colonnes.

$\max_j n.j$: représente la valeur maximum des totaux sur les lignes (marge)

Interprétation de $\lambda\, Y|X$

Le ratio $\lambda\, Y|X$ représente le pourcentage de réduction de l'erreur de pronostic entre:
- la prévision de Y sans connaissance sur X
- et la prévision de Y connaissant X.

$\lambda\, Y|X$ est une mesure du % d'amélioration du pronostic de Y apporté par la connaissance de X.

De par sa construction ce ratio est indépendant de la taille de l'échantillon. Ce ratio est toujours compris entre 0 et 1.

• Si $\lambda\, Y|X = 0$ ===> (p1=p2)
Connaître X n'est d'aucune utilité pour prédire Y
On prédit toujours la même modalité de Y

forme du tableau :

• M • • •
• M • • •
• M • • •
• M • • •

Les maximum sont tous repérés sur une même colonne

• Si $\lambda\, Y|X = 1$ ===> (p2=0)
La prédiction dans ce cas est effectuée sans erreur
A chaque modalité i de la variable indépendante X est associée une seule modalité j de la variable dépendante Y.

forme du tableau

0 X 0 0 0
0 0 X 0 0
0 0 X 0 0
X 0 0 0 0
0 0 0 X 0

chaque ligne du tableau n'a qu'une seule cellule non nulle.

2. λ asymétrique X|Y

On peut faire le même raisonnement en inversant les rôles de X et de Y, ce qui donne Lambda asymétrique de X sachant Y noté λ X|Y.

X devient la variable dépendante
Y devient la variable indépendante

Cette fois-çi c'est Y qui est susceptible d'apporter de l'information au pronostic de X.

Exemples Pratiques

• Soit le tableau croisant la couleur des yeux et celle des cheveux.

cheveux yeux	blond	brun	noir	roux	Total
bleu	25	9	3	7	44
vert	13	17	10	7	47
marron	7	13	8	5	33
Total	45	39	21	19	124

Nous avons vu précédemment que la distribution des blonds est différente de la distribution des roux. Il y a des points d'accumulation (attractions) ou des vides (répulsions) à des endroits différents.
Les cheveux blonds et les yeux bleus sont souvent associés, comme le sont les cheveux bruns avec les yeux marrons. Le calcul de λ donne:

Calcul:

$\lambda Y|X = (25 + 17 + 13 - 45) / (124 - 45) = 10 / 79 = 0.127$

$\lambda X|Y = (25 + 17 + 10 + 7 - 47) / (124 - 47) = 12 / 77 = 0.156$

Notation de SAS

λ Y|X noté λ C|R

λ X|Y noté λ R|C

- **Exemple et analyse empruntés à J.M GROSBRAS**

Soit les 2 questions Q1 et Q2 posées lors d'une enquête:
Q1: possédez-vous un téléviseur ?
Q2: fréquentez-vous le cinéma ?

Le tableau croisant les réponses à Q1 et Q2 est le suivant:

Ciné Télé	OUI	NON	Total
OUI	20	680	700
NON	80	220	300
Total	100	900	1000

$\lambda Y|X = (680 + 220 - 900) / (1000 - 900) = 0$

"Quelle que soit la réponse à la question sur la possession d'un téléviseur, la fréquentation du cinéma est minoritaire, et on peut toujours pronostiquer la réponse 'NON' pour Q2."

$\lambda X|Y = (80 + 680 - 700) / (1000 - 700) = 0.2$

"Savoir qu'un individu a, ou non été au cinéma influence le pronostic sur le fait qu'il a, ou non, un téléviseur".

Variance de l'estimateur λ : ASE Asymptotic Standard Error

SAS fourni l'erreur-type pour chaque λ, ce qui permet d'accorder une certaine confiance à la valeur de cette mesure.

3. λ symétrique

Afin d'établir une symétrie entre X et Y un coefficient "artificiel" est calculé par SAS. C'est une sorte de moyenne sur les 2 λ asymétriques.

$$\lambda = \frac{(\sum_i \max_j n_{ij}) + (\sum_j \max_i n_{ij}) - (\max_j n_{.j} + \max_i n_{i.})}{2N - (\max_j n_{.j} + \max_i n_{i.})}$$

De par sa construction la valeur de ce λ est comprise entre les 2 λ asymétriques.

Calcul pour l'exemple couleur des yeux et des cheveux:

$\lambda = (10 + 12)/(79 + 77) = 0.141$

on a bien 0.141 compris entre $\lambda Y|X = 0.127$ et $\lambda X|Y = 0.156$

Remarque importante pour l'interprétation

- s'il y a **indépendance** alors $\lambda = 0$

Attention: le raisonnement réciproque est faux :

avoir $\lambda = 0$ **ne signifie pas** toujours avoir indépendance

- $\lambda = 1$ <===> **Association parfaite**

Deux cases non nulles du tableau de contingence ne sont jamais sur la même ligne ni sur la même colonne. (*cf la forme du tableau ci-dessous*)

forme du tableau

0 X 0 0 0
0 0 X 0 0
X 0 0 0 0
0 0 0 X 0
0 0 0 0 X

chaque ligne et chaque colonne du tableau n'a qu'une seule cellule non nulle

III - 4 . 2 Coefficient d' Incertitude U

Tout comme le Lambda, le coefficient d'incertitude est utilisé pour des tableaux croisés, les variables étant traitées comme **nominales.**

Il existe 3 formes de coefficient d'Incertitude

1. Coefficient d'Incertitude Y|X
2. Coefficient d'Incertitude X|Y
3. Coefficient d'Incertitude symétrique

Son invention prend origine dans l'approche de la Théorie de l'information de Shannon (1940). Domaine des communications.

Historique

Lorsque Shannon a proposé en 1940 une mesure de l'incertitude, il se plaçait dans une situation de transfert d'information en télécommunications depuis une source (émetteur) jusqu'à sa réception (récepteur).

Shannon considère un ensemble d'événements possibles $E_1, \ldots, E_n$, dont les probabilités de réalisation sont $p_1, \ldots, p_n$,supposées connues.

Comment trouver " une mesure du nombre de *choix* impliqués dans la sélection de l'événement ou celle de *l'incertitude* du résultat"?

Shannon a démontré que la seule fonction H vérifiant certaines propriétés (continuité, monotonie etc...) est de la forme:

$$H = - K \sum_{i=1}^{n} pi \, Log \, pi$$

K étant une constante positive dépendant des unités de mesure.

La quantité H introduite par Shannon comme mesure du choix et de l'incertitude joue un rôle central comme mesure de l'information. Cette mesure a été étendue depuis à d'autres domaines de la connaissance comme en Statistique, en Economie, en Biologie etc...

Shannon a donné le nom d' **Entropie** à cette mesure de l'information, du choix et de l'incertitude.

Note

Selon les auteurs et les domaines de connaissance il y a une certaine confusion entre les mots, *Entropie*, *Incertitude*, et même *Information.* Pour plus d'information, voir le livre de P.J LANCRY "Théorie de l'information et Economie".

L'INCERTITUDE EN STATISTIQUE ET ECONOMIE

Entropie

de X $\quad H(X) = - \Sigma_i \, (ni. / n) \; Log \, (ni. / n)$

de Y $\quad H(Y) = - \Sigma_j \, (n.j / n) \; Log \, (n.j / n)$

de XY $\quad H(XY) = - \Sigma_i \Sigma_j \, (nij / n) \; Log \, (nij / n)$

Incertitude Asymétrique

de Y|X $\quad U(Y|X) = (\; H(X) + H(Y) \; - H(XY) \;) \; / \; H(Y)$

de X|Y $\quad U(X|Y) = (\; H(X) + H(Y) \; - H(XY) \;) \; / \; H(X)$

Incertitude symétrique

$U \; = 2 \; (\; H(X) + H(Y) \; - H(XY) \;) \; / \; (\; H(X) + H(Y) \;)$

L'approche des deux mesures U et λ est un peu similaire. On cherche la réduction de l'erreur de pronostic lorsque l'une des variable peut apporter une information sur l'autre.

Avantage de U sur λ

L'avantage de la mesure U sur λ, est qu'elle prend en compte toute la distribution de la variable et non seulement le mode.

Interprétation de U

U est compris entre 0 et 1

- Si U=0 il n'y a aucune possibilité d'améliorer la connaissance de la variable dépendante à partir de la variable indépendante.

- Si U=1 on élimine complètement l'incertitude
  Ceci n'est réalisé que lorsque chaque modalité de la variable indépendante est associée à une modalité unique de la variable dépendante.

Nous terminons ici l'inventaire des tests et mesures afférant aux variables nominales. Dans le chapitre suivant nous traiterons le cas des variables ordinales.

IV - Indépendance et association entre variables ordinales

Les mesures d'association entre variables ordinales calculées par la PROC FREQ (Gamma, Tau-b de Kendall, Tau-c de Stuart, D de Somer, coefficients de corrélation de Pearson et de Spearman), utilisent cette propriété que les modalités[3] des variables sont ordonnées, et cherchent à mesurer une relation monotone entre elles : croissent-elles dans le même sens, ou en sens contraire?
Avant de définir ces mesures faisons un détour par une approche formelle qui nous permettra de mieux comprendre, croyons-nous, à la fois ce qu'elles doivent au coefficient de corrélation de Pearson, et le développement des calculs.

IV - 1 Coefficients dérivés de la Formule de Daniels

IV - 1 . 1 Approche formelle

Soit un échantillon de n individus sur lesquels on mesure deux variables X et Y. Si à toute paire (h,h') d'individus on associe un nombre noté $a_{hh'}$ (resp. $b_{hh'}$) correspondant à la variable X (resp. Y), la formule générale de Daniels s'écrira alors:

$$\frac{\Sigma_h \Sigma_{h'} a_{hh'} b_{hh'}}{\text{Rac}[(\Sigma_h \Sigma_{h'} a_{hh'}^2)(\Sigma_h \Sigma_{h'} b_{hh'}^2)]} \quad \text{(où h et h' varient de 1 à n)}$$

Ce coefficient varie entre -1 et +1 (inégalité de Schwarz).
On obtient des coefficients différents selon le choix de $a_{hh'}$ et $b_{hh'}$

IV - 1 . 2 Coefficients de corrélation

. Coefficient de Pearson (1896)

pour X et Y quantitatives, il s'obtient en prenant
$a_{hh'} = x_h - x_{h'}$ et $b_{hh'} = y_h - y_{h'}$.

[3] Dans FREQ on peut définir les codages des modalités par l'option SCORES

. **Coefficient de corrélation des rangs de Spearman (1904)**

Il s'obtient avec $a_{hh'} = r_X(h) - r_X(h')$ et $b_{hh'} = r_Y(h) - r_Y(h')$
où $r_X(h)$ désigne le rang de l'individu h sur la variable X.

Il peut y avoir des individus "ex-aequo" sur l'une ou l'autre des variables, c'est-à-dire prenant la même valeur. Soit $n_{k.}$ le nombre d'individus prenant la valeur k sur la variable X (pour reprendre les notations usuelles d'un tableau croisant X en ligne et Y en colonne) : leur rang sera alors le rang moyen $r_X(h) = \Sigma_{(i=1,...k-1)}\, n_{i.} + (n_{k.}+1)/2$.

IV - 1 . 3 Les coefficients de Kendall τ et τ_b

. **Le Tau de Kendall (1938)** s'obtient en prenant
$a_{hh'}$ = signe de $(x_h - x_{h'})$ et $b_{hh'}$ = signe de $(y_h - y_{h'})$.
$a_{hh'}$ vaut alors:
1 si $x_h > x_{h'}$, -1 si $x_h < x_{h'}$,
0 si h et h' sont ex-aequo sur la variable X.

Et de même pour $b_{hh'}$.
(La PROC FREQ ne calcule pas le Tau de Kendall).

. **Concordances et discordances**
Le produit $a_{hh'}\, b_{hh'}$ vaut 1 si les rangs de h et de h' sont en **concordance** sur les deux variables:
$(x_h < x_{h'}$ et $y_h < y_{h'})$ ou $(x_h > x_{h'}$ et $y_h > y_{h'})$,

Le produit vaut -1 si les rangs sont en **discordance** :
$(x_h < x_{h'}$ et $y_h > y_{h'})$ ou $(x_h > x_{h'}$ et $y_h < y_{h'})$.

Si on note C le nombre de paires hh' concordantes et D le nombre de paires discordantes on a donc:

$$\Sigma_h \Sigma_{h'}\, a_{hh'}\, b_{hh'} = 2\,(C - D)$$

(on compte dans la somme double deux fois la même paire, comme hh' et h'h).

C-D est nul "si les concordances équilibrent les discordances, ce qui est en particulier le cas s'il y a indépendance au sens des profils" (J-M. Grosbras). La différence est positive si X et Y varient plutôt dans le même sens, négative si elles varient en sens contraire.

Tous les coefficients qui suivent sont calculés à partir de cette différence C-D au numérateur. Ils diffèrent par le dénominateur choisi.
Ils s'interpréteront comme la différence entre la proportion (probabilité) de concordances et la proportion (probabilité) de discordances $\Pi_C - \Pi_D$.

Dans le cas du τ **de Kendall** on considère qu'il n'y a pas d'ex-aequo et tous les $a_{hh'}{}^2$ et les $b_{hh'}{}^2$ valent 1, si bien qu'on a au dénominateur le nombre n(n-1) de paires hh' ou h'h d'individus distincts :

$$\tau = \frac{2(C - D)}{n(n-1)}$$

Calcul de C et D

Cij		Dij
	n_{ij}	
Dij		Cij

Reprenons les notations usuelles pour le tableau croisant les variables X et Y, Le nombre d'individus en concordance avec ceux de la case ij est obtenu en sommant toutes les cases du coin supérieur gauche du tableau, et du coin inférieur droit.

$C_{ij} = \Sigma_{k>i} \Sigma_{l>j} n_{kl} + \Sigma_{k<i} \Sigma_{l<j} n_{kl}$

Le nombre C de paires concordantes est donc:

$C = (1/2) \Sigma_i \Sigma_j n_{ij} C_{ij}$
Le coefficient 1/2 vient de ce que l'on a comptabilisé 2 fois chaque paire d'individus.

De même le nombre d'individus en discordance avec ceux de la case ij est obtenu en sommant toutes les cases du coin inférieur gauche et toutes celles du coin supérieur droit du tableau :

$D_{ij} = \Sigma_{k>i} \Sigma_{l<j} n_{kl} + \Sigma_{k<i} \Sigma_{l>j} n_{kl}$

Le nombre D de paires discordantes est alors:

$D = (1/2) \Sigma_i \Sigma_j n_{ij} D_{ij}$

Pour le calcul sur un exemple voir en fin de ce paragraphe.

. Le Tau-b de Kendall

S'il y a des ex-aequo sur l'une ou l'autre des variables, certains termes $a_{hh'}$ ou $b_{hh'}$ sont nuls .

Pour chaque valeur i de la variable X il y a $n_{i.}$ individus ex-aequo, donc n_i $(n - 1)$ paires nulles ; le nombre total de termes $a_{hh'}$ nuls est donc $\Sigma_i\ n_{i.}\ (n_{i.}-1)$, et:

$\Sigma_h \Sigma_{h'}\ a_{hh'}^2 = n(n-1) - \Sigma_i\ n_{i.}\ (n_{i.}-1) = n^2 - \Sigma_i\ n_{i.}^2$.

De même le nombre total de termes $b_{hh'}$ nuls est-il $\Sigma_j\ n_{.j}\ (n_{.j}-1)$, et $\Sigma_h\Sigma_{h'}\ b_{hh'}^2 = n^2 - \Sigma_j n_{.j}^2$.

$$\tau_b = \frac{2(C - D)}{\text{Rac}[(n^2 - \Sigma_i\ n_{i.}^2)(n^2 - \Sigma_j\ n_{.j}^2)]}$$

Remarques

1°) τ_b est plus approprié au cas des tableaux carrés. $\tau_b=1$ en cas de concordance parfaite (tableau chargé sur la diagonale majeure) et -1 en cas de discordance (diagonale mineure).

2°) Dans le cas des tables 2 x 2 on a $|\tau_b|=\Phi$. L'avantage de τ_b est qu'il indique par son signe la tendance de l'association.

IV - 2 Autres coefficients basés sur les concordances et discordances

Comme τ et τ_b ces mesures reposent sur le nombre de concordances C et de discordances D comptées sur toutes les paires d'observations.

° Gamma (Goodman & Kruskal - 1954)

$$\gamma = \frac{C - D}{C + D}$$

Ce coefficient ne tient pas compte des ex-aequo ; il varie entre -1 et +1 mais on peut avoir $|\gamma| = 1$ sans que le tableau soit diagonal.

° Tau-c de Stuart

$$\tau_C = \frac{2(C - D)}{n^2 \ (m-1)/m}$$ où m est la plus petite des dimensions (r,c).

τ_C est approprié aux tableaux rectangulaires puisqu'il tient compte de ses dimensions. "$|\tau_C|$ est voisin de 1 quand les seules cases non nulles sont celles des diagonales les plus longues" (Grosbras).

Comme C+D <= n(n+1)/2 on a en général $\gamma > \tau_b$, $\gamma > \tau_C$.

° D asymétrique de Somer

Ce coefficient tient compte aussi des ex-aequo dans le calcul du dénominateur, mais de façon dissymétrique: si la variable ligne X est considérée comme dépendante, on compte au dénominateur le nombre de paires non ex-aequo sur la variable Y (i.e. on déduit les ex-aequo sur la variable indépendante) :

$$D(X/Y) = \frac{2\ (C - D)}{(n^2 - \Sigma_j n_{.j}^2)}$$

On définira de même :

$$D(Y/X) = \frac{2\ (C - D)}{(n^2 - \Sigma_i n_{i.}^2)}$$

Toutes ces statistiques sont asymptotiquement normales : SAS en calcule l'ASE (Asymptotic Standard Error) dont on déduit un intervalle de confiance.

Exemple: table 2 X 3 - données insertion des jeunes

```
TABLES DIPLOME*SITU/ NOROW NOCOL NOPERCENT MEASURES;
```

ATELIER SAS PROC FREQ
SOURCE: ENQUETE D'INSERTION CEREQ-DEP
DE TERMINALE CAP OU BEP COMMERCE EN L.P. (SN, APPRENTIS EXCLUS)

TABLE OF DIPLOME BY SITU

DIPLOME / SITU	CHOMAGE	MESURE	EMPLOI	Total
NON DIPL	54	52	40	146
DIPLOMES	122	97	133	352
Total	176	149	173	498

STATISTICS FOR TABLE OF DIPLOME BY SITU

Statistic	Value	ASE
Gamma	0.123	0.077
Kendall's Tau-b	0.065	0.041
Stuart's Tau-c	0.068	0.043
Somers' D C|R	0.082	0.052
Somers' D R|C	0.051	0.033
Pearson Correlation (Rank Scores)	0.069	0.043
Spearman Correlation	0.069	0.044
Lambda Asymmetric C|R	0.034	0.049
Lambda Asymmetric R|C	0.000	0.000
Lambda Symmetric	0.024	0.034
Uncertainty Coefficient C|R	0.005	0.004
Uncertainty Coefficient R|C	0.009	0.008
Uncertainty Coefficient Symmetric	0.007	0.006

Sample Size = 498

$C_{11} = 97 + 133$, $C_{12} = 133$, $C_{22} = 54$, $C_{23} = 54+52$
$C = (1/2)(54\, C_{11} + 52\, C_{12} + 97\, C_{22} + 133\, C_{23})$
ou $C = 54\,(97+133) + 52\,(133) = 12\,420 + 6\,916 = 19\,336$

$D_{12} = 122$, $D_{13} = 122+97$, $D_{21} = 52+40$, $D_{22} = 40$
$D = 40\,(97+122) + 52\,(122) = 8\,760 + 6\,344 = 15\,104$
$C - D = 4232$, $C + D = 34\,440$, $n(n-1) = 247\,506$

Exemple: table 2 X 2 - données insertion des jeunes

```
TABLES DIPLOME*SITU/ NOROW NOCOL NOPERCENT MEASURES;
```

ATELIER SAS PROC FREQ
SOURCE: ENQUETE D'INSERTION CEREQ-DEP
DE TERMINALE CAP OU BEP COMMERCE EN L.P. (SN, APPRENTIS EXCLUS)

TABLE OF DIPLOME BY SITU

DIPLOME SITU

Frequency	CHOMAGE MESURE	EMPLOI	Total
NON DIPL	106	40	146
DIPLOMES	219	133	352
Total	325	173	498

STATISTICS FOR TABLE OF DIPLOME BY SITU

Statistic	Value	ASE
Gamma	0.234	0.102
Kendall's Tau-b	0.099	0.043
Stuart's Tau-c	0.086	0.038
Somers' D C|R	0.104	0.045
Somers' D R|C	0.095	0.041
Pearson Correlation (Rank Scores)	0.099	0.043
Spearman Correlation	0.099	0.043
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.000	0.000
Lambda Symmetric	0.000	0.000
Uncertainty Coefficient C|R	0.008	0.007
Uncertainty Coefficient R|C	0.008	0.007
Uncertainty Coefficient Symmetric	0.008	0.007

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.609	1.055	2.456
Cohort (Col1 Risk)	1.167	1.026	1.327
Cohort (Col2 Risk)	0.725	0.539	0.975

Sample Size = 498

V - Tests d'association de Cochran-Mantel-Haenszel

Ce sont des tests qui utilisent la loi hypergéométrique multiple, pour calculer la moyenne et la matrice de variance-covariance d'un vecteur mesurant la différence entre les fréquences observées et les fréquences attendues (en général sous indépendance). Si les cases sont d'effectifs suffisamment grands, on peut appliquer le théorème central limite et le vecteur est distribué selon une loi normale ; les statistiques de tests suivent alors des lois du CHI-2.
(cf. article de 1978 de Landis, Heyman et Koch Landis Heyman Koch International Statistical Review,1978, 46 , pp 237-254 cité en référence dans SAS)

On les obtient par l'option CMH de FREQ, qui peut se séparer en CMH1 CMH2 CMH3.

	Y ordinale	Y nominale
X ordinale	CMH1	
X nominale	CMH2	CMH3

CMH1 : Cas où les 2 variables X et Y sont ordinales

Ce test est basé sur le coefficient de corrélation entre X et Y, codées numériquement selon des valeurs définies par l'option SCORES (cf. Chapitre IV).

SCORES = TABLE : cas où les modalités sont numériques et donc leurs valeurs sont utilisées dans le calcul ;
SCORES = RANK : cas de modalités ordinales dont le rang est utilisé dans le calcul.

Lorsqu'il n'y a qu'une seule strate, la statistique vaut (N-1) r^2, qui suit un CHI-2 à 1 dd1.
C'est la mesure Qmh de l'option CHISQ (cf. III - 2.1).

CMH2 : cas où X est nominale (r modalités) et Y ordinale (c modalités)

Ce test est basé sur la comparaison des r moyennes des scores de Y, calculés pour les r modalités de X.
Il s'agit donc d'une analyse de variance (ou d'un test non paramétrique dit de Kruskall-Wallis si SCORES = RANK).
La statistique suit un CHI-2 à (r-1) degrés de liberté.

CMH3 : Cas où X et Y sont nominales

C'est un test "d'association" entre X et Y.
La statistique suit un CHI-2 à (r-1)(c-1) degrés de liberté ;
elle est égale à $\frac{N-1}{N} \chi^2$ d'indépendance.

Intérêt : les test CMH sont des tests non paramétriques dans le cas où SCORES = RANK.

Condition d'application : Il faut que les effectifs par case soient "assez grands" pour que le théorème central limite soit applicable.

Cas où il y a plusieurs tables : si on ajoute une troisième variable Z à k modalités, on parle d'analyse stratifiée.
SAS calcule une statistique CMH "ajustée" sur les k tables, permettant de vérifier si l'association se retrouve dans les k tables (les degrés de liberté ne changent pas).
SAS précise que cette statistique est peu efficace dans le cas de distorsions entre le sens des associations des k tables.

Exemple : Identique au § III - 2 . 1

Table 2 x 3 , avec option CMH.

Ici CMH1 est égal au Mantel-Haenszel chi-square de l'option CHISQ du § III - 2 . 1.
Par contre aucune des variables n'étant ordinale, c'est CMH3 qu'il faut utiliser.

enquête CEREQ_DEP 1990 - option CMH

SUMMARY STATISTICS FOR DIPLOME BY SITU

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	2.376	0.123
2	Row Mean Scores Differ	1	2.376	0.123
3	General Association	2	5.592	0.061

Total Sample Size = 498

VI - Approche probabiliste dans le cas d'une table 2x2

Plutôt que la table de contingence, on étudie ici la table de probabilité p i j = ni j/N.

p11	p12	p1.
p21	p22	p2.
p.1	p.2	1

Modèle : Ici, le modèle probabiliste est multinomial :
on tire un échantillon de taille N, avec remise dans une population possédant 4 types d'individus répartis selon les proportions (p i j).

On distingue les modèles d'échantillonnage du type "case control" (X aléatoire, Y fixé) des modèles du type "cohort" (X fixé, Y aléatoire).

	Y fixé	Y aléaloire
X fixé		cohort
X aléatoire	case control	

Remarque : On est souvent amené à privilégier le rôle de la variable Y, notamment dans les études médicales, lorsqu'il s'agit de la variable *présence/absence* d'une maladie (cf VI-4). Aussi, lorsqu'une des deux variables est de type *oui/non,* on l'appelle variable "Réponse" (à la maladie dans le cas médical) et on la place en variable colonne Y. Quand les deux sont du type *oui/non*, on place en général la modalité *oui* en premier; c'est-à-dire que la case (1,1) correspond à (X=oui et Y=oui).

VI - 1 Odds-ratio

Odds se traduit par "chance" ; odds-ratio peut lui se traduire par "cote" comme dans les paris.

Lois conditionnelles sachant les lignes $(p\,j|i = \frac{p\,i\,j}{p\,i.}$ pour $j = 1,2)$

* Sachant la ligne 1

la probabilité d'être en colonne $1 = \frac{p11}{p1.}$

la probabilité d'être en colonne $2 = \frac{p12}{p1.}$

$\Rightarrow$ rapport odds $\quad \Omega 1 = \frac{p1\,|\,1}{p2\,|\,1} = \frac{p11}{p12}$

Pour les individus de la ligne 1, Ω1 est le rapport de "chances" entre les 2 réponses en colonne.

* idem sachant ligne 2 $\quad \Omega 2 = \frac{p1 \mid 2}{p2 \mid 2} = \frac{p21}{p22}$

ODDS-RATIO $\theta = \Omega 1 / \Omega 2 = \frac{p1 \mid 1 \; p2 \mid 2}{p2 \mid 1 \; p1 \mid 2} = \frac{p11 \; p22}{p12 \; p21}$

Remarques :

* La table est entièrement déterminée par :
- les deux lois de probabilité marginales en ligne et en colonne,
- l'odds-ratio.

* θ ne change pas si on inverse le rôle des lignes et des colonnes.

SAS indique "case control" pour l'odds-ratio (X aléatoire, Y fixé).
Le logarithme de θ s'appelle logit (cf. le lien avec les modèles logit en VI-4).

Interprétation de l'odds-ratio :

Indépendance $\leftrightarrow \theta = 1 \leftrightarrow \text{Log}(\theta) = 0$
(si pij $\neq$ 0 pour tout i et j)

$\theta > 1 \Rightarrow \Omega 1 > \Omega 2$

Les individus ayant la modalité 1 en ligne ont alors plus de "chance" d'avoir la réponse 1 en colonne que ceux ayant la modalité 2 en ligne : θ est donc la "cote" de la modalité 1 en ligne.

Intervalle de confiance : on peut calculer un intervalle de confiance (à 95% dans FREQ) qui permettra de conclure si l'odds-ratio diffère de 1.

VI - 2 Risque relatif

La variable Y est ici privilégiée: on va calculer la probabilité d'en avoir une modalité de Y, selon la modalité de la variable X. C'est ce que SAS appelle "Relative Risk" (Row1/Row2).

Risque relatif :

$$\text{col1 - risk} = \frac{p1 \mid 1}{p1 \mid 2} = \frac{\frac{p11}{p1.}}{\frac{p21}{p2.}} = \text{rapport du 1er élément des deux profils-lignes}$$

Si col1-risk = 1, les individus ont autant de risque d'avoir la réponse 1 en colonne, quelle que soit leur caractéristique en ligne : ceci est équivalent à l'indépendance.

$$\text{col2-risk} = \frac{p2 \mid 1}{p2 \mid 2}$$

Remarque : SAS indique **cohort (col1 risk ou col2 risk)** pour préciser qu'on est dans le cadre "X fixé et Y aléatoire".

on retrouve la relation évidente : $\frac{\text{col1 risk}}{\text{col2 risk}} = \text{odds-ratio}$

Par définition, si la variable Y est du type réponse et si la modalité 1 en colonne est (Y=oui), col1-risk est appelé **Risque Relatif.** Si elle est du type (Y=non), le **Risque Relatif** est col2-risk.

Dans le cas d'un échantillonnage "case control" (X aléatoire, Y fixé), c'est l'odds-ratio qui estime le risque relatif. Dans le cas "cohort" (X fixé, Y aléatoire), c'est col1-risk (ou col2-risk).

Intervalle de confiance : on peut calculer des intervalles de confiance (à 95% dans FREQ) qui permettront de conclure si les risques diffèrent de 1.

VI - 3 Analyse stratifiée

S'il y a trois variables Z X Y , on est dans le cadre d'une analyse dite stratifiée (il y a autant de strates, et donc de tables, que de modalités de la 3ème variable Z). On peut alors estimer un risque relatif commun aux différentes tables.

Cet estimateur diffère selon le modèle : dans le cas "case control" (X aléatoire, Y fixé), c'est l'odds-ratio qui est un estimateur du risque relatif commun. Dans le cas "cohort" (X fixé, Y aléatoire) ou dans le cas où X et Y sont aléatoires, il y a un estimateur direct du risque relatif commun.

SAS donne donc deux estimateurs qu'il nomme :
"case control" (Mantel-Haenszel et logit) pour l'odds-ratio,
"cohort" (Mantel-Haenszel et logit) pour les risques relatifs.

Exemple : formule pour le modèle "Case control" (odds-ratio)

Si la variable Z a k modalités il y a donc k tables 2 x 2 :

La table h est (nhij) où h=1 à k ; i=1,2 ; j=1,2 (effectif total Nh)

L'odds-ratio de la table h est $ORh = \frac{nh11\ nh22}{nh12\ nh21}$

On peut estimer le odds-ratio "global" par 2 estimateurs :

Mantel-Haenszel :

$[\ \sum_h\ nh11\ nh22\ /\ Nh\ ]\ /\ [\ \sum_h\ nh12\ nh21\ /\ Nh\ ]$

Logit

$\exp\ [\ \sum_h\ wh \log\ (ORh))\ /\ \sum_h\ wh\ ]$ où $wh = 1\ /\ var\ (\log\ (ORh))$

SAS calcule des intervalles de confiance à 95% autour de ces deux estimateurs.

On peut aussi tester l'égalité des odds-ratios dans les k tables par un **test de Breslow-Day**, basé sur une statistique suivant une loi du CHI-2 à (k-1) degrés de liberté. Ce test n'est applicable que si Nh est grand pour tout h.

Utilisation : dans l'instruction TABLES de la procédure FREQ, pour avoir risques et odds-ratios, il faut ajouter :

- l'option MEASURES (ou CMH) dans le cas d'une seule table 2X2
- l'option CMH dans le cas de plusieurs tables 2x2 à comparer (+ option MEASURES si on veut l'odds-ratio de chaque table)

Exemple : Identique au § III - 2 . 2 avec option MEASURES
On est alors dans un cas X fixé (RACE) Y aléatoire(VERDICT), c'est-à-dire "cohort":

- θ est plus grand que 1, donc les "blancs" ont plus de "chance" d'avoir la peine de mort que les "noirs", mais cette différence est non significative d'après l'intervalle de confiance.
- Col1 Risk = risque de (Y=oui) = risque d'obtenir la peine de mort: il est plus fréquent pour les "blancs", mais non significatif.

etude d'une table 2x2 : RACE/VERDICT , option MEASURES

STATISTICS FOR TABLE OF RACE BY MORT

Statistic	Value	ASE
Gamma	0.083	0.176
Kendall's Tau-b	0.026	0.055
Stuart's Tau-c	0.016	0.035
Somers' D C|R	0.016	0.035
Somers' D R|C	0.042	0.088
Pearson Correlation	0.026	0.055
Spearman Correlation	0.026	0.055
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.013	0.037
Lambda Symmetric	0.010	0.030
Uncertainty Coefficient C|R	0.001	0.004
Uncertainty Coefficient R|C	0.000	0.002
Uncertainty Coefficient Symmetric	0.001	0.003

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.181	0.590	2.363
Cohort (Col1 Risk)	1.160	0.625	2.150
Cohort (Col2 Risk)	0.982	0.909	1.060

Sample Size = 326

VI - 4 Lien avec les modèles LOGIT

Dans un exemple médical (cf. Jean Bouyer) où Y est la variable présence ou absence d'une maladie (M+ / M-) et X est la variable dichotomique exposition ou non exposition à un facteur déclenchant de la maladie (exposition X=1 , non exposition X=0), on peut définir la table des lois conditionnelles selon l'exposition :

	M+	M-
X = 1	p1	1-p1
X = 0	p0	1-p0

On étudie ici Y=*présence de la maladie*:

(M+)=(Y=oui) donc le **Risque relatif** est Col1 Risk.

Risque relatif : Prob (M+ | X=1)/Prob(M+ | X=0)=p1/p0

Odds-ratio = Prob (M+ | X=1) Prob(M- | X=0) / Prob(M - | X=1) Prob(M+ | X=0)
= p1(1-p0)/p0(1-p1)
= (p1/(1-p1))/(p0/(1-p0))

Par définition, on nomme logit : Log(p/(1-p))

$\Rightarrow$ Log(θ) = logit(p1) - logit(p0)

Dans le MODELE LOGISTIQUE, on pose :

probabilité (M+| X) = f (X) = p
= 1/(1+ exp(a+bX))

$\leftrightarrow$ logit (p) = a+bX

Dans ce modèle, Log (θ) est un estimateur de b
(car logit (p1)=a+b , logit (p0)=a).

VII. Curiosités

A titre de curiosité nous reprenons un exemple de tableaux et sous tableaux de ROUANET paru dans l'Echo des Messaches Nov 78 n° 8 et repris dans le Bulletin Méthodologie Sociologie n°6 avril 1985 pp3-27.

Barouf à Bombach

Dans la ville de Bombach existent deux lycées: (A)nastase et (B)énédicte. Les résultats au Bac, succès et échecs selon le sexe, sont donnés dans les tableaux de la page suivante.

La lecture des pourcentages de réussite permet de conclure:
Les garçons réussissent mieux que les filles quel que soit le lycée.

Mais le tableau résumé A+B pour la ville de Bombach aboutit à la conclusion surprenante:
Les filles réussissent mieux que les garçons.

Analyse de Rouanet: C'est un exemple de situation statistique conflictuelle.

Barouf a BOMBACH ...

ROUANET 1978

Lycée Anastase

	succès	échec	
G	15	35	50
F	1	9	10
	16	44	60

	% succès	% échec
G	30%	70%
F	10%	90%

les **garçons** *réussissent mieux que les* **filles**

B - Lycée Bénédicte

	succès	échec	
G	9	1	10
F	35	15	50
	44	16	60

	% succès	% échec
G	90%	10%
F	70%	30%

les **garçons** *réussissent mieux que les* **filles**

regroupement A+B

BOMBACH

	succès	échec	
G	24	36	60
F	36	24	60
	60	60	120

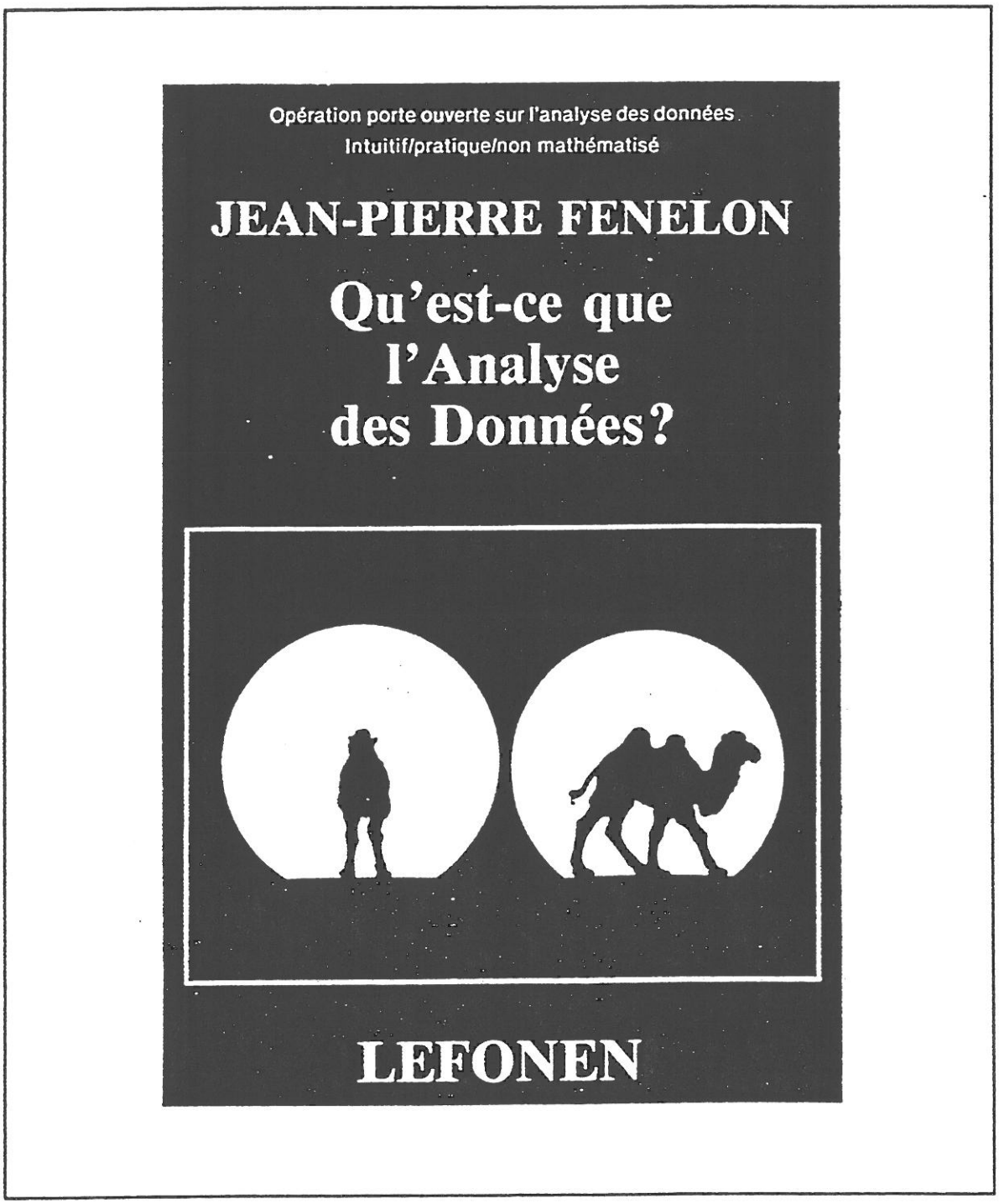
	succès	échec
G	40%	60%
F	60%	40%

les **filles** *réussissent mieux que les* **garçons**

?

Cet exemple de discordances des résultats selon les points de vue est bien symbolisé par le "Chameau" de J.P. Fénelon et ce dernier nous invite à l'Analyse de Données lorsque le nombre de variables en inter-relation dépasse 2.

Le "chameau" de J.P Fénelon



Conclusion

Pour finir, citons Agresti, qui lui-même cite dans Categorical Data Analysis :

28 DESCRIBING TWO-WAY CONTINGENCY TABLES

Goodman and Kruskal (1959) summarized the historical development of measures of association for contingency tables. Their book (1979) reprints four classic papers (1954, 1959, 1963, 1972) they published on this topic. The 1959 paper contains the following quote from a paper by M. H. Doolittle in 1887, which undoubtedly clarified the meaning of *association* for his contemporaries:

> Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.

Que nous nous risquons à traduire ainsi:

"Etant donné le nombre de cas où respectivement les choses sont à la fois comme-ci et comme-çà, sont comme-ci mais non pas comme-çà, sont comme-çà et non comme-ci, et ne sont ni comme-ci ni comme-çà, il convient d'éliminer le lien quantitatif général inhérent à la simple chosité des choses, et de déterminer le lien quantitatif spécifique résiduel entre les caractères comme-ci et comme-çà des choses".

Annexes

A0. Exemples de sorties de PROC FREQ

1. Exemple n° 1 : indépendance
2. Exemple n° 2 : dépendance
3. Exemple n° 3 : association parfaite

1. Exemple n° 1 : indépendance

EXEMPLE 1: PCS PERE * NIVEAU DES ELEVES

EXEMPLE DE STRUCTURE DANS UN TABLEAU DE CONTINGENCE

I N D E P E N D A N C E

TABLE OF PCS BY NIVEAU

PCS NIVEAU

Frequency / Percent / Row Pct / Col Pct	- -	-	+	+ +	Total
CADRE	2 2.50 5.00 50.00	2 2.50 5.00 50.00	12 15.00 30.00 50.00	24 30.00 60.00 50.00	40 50.00
EMPLO	2 2.50 5.00 50.00	2 2.50 5.00 50.00	12 15.00 30.00 50.00	24 30.00 60.00 50.00	40 50.00
Total	4 5.00	4 5.00	24 30.00	48 60.00	80 100.00

STATISTICS FOR TABLE OF PCS BY NIVEAU

Statistic	DF	Value	Prob
Chi-Square	3	0.000	1.000
Likelihood Ratio Chi-Square	3	0.000	1.000
Mantel-Haenszel Chi-Square	1	0.000	1.000
Phi Coefficient		0.000	
Contingency Coefficient		0.000	
Cramer's V		0.000	

Statistic	Value	ASE
Gamma	0.000	0.206
Kendall's Tau-b	0.000	0.108
Stuart's Tau-c	0.000	0.112
Somers' D C|R	0.000	0.112
Somers' D R|C	0.000	0.103
Pearson Correlation	0.000	0.112
Spearman Correlation	0.000	0.112
Lambda Asymmetric C|R	0.000	0.000
Lambda Asymmetric R|C	0.000	0.000
Lambda Symmetric	0.000	0.000
Uncertainty Coefficient C|R	0.000	0.000
Uncertainty Coefficient R|C	0.000	0.000
Uncertainty Coefficient Symmetric	0.000	0.000

Sample Size = 80
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

2. Exemple n° 2 : dépendance

EXEMPLE 2: COULEURS YEUX * COULEURS DES CHEVEUX

EXEMPLE DE STRUCTURE DANS UN TABLEAU DE CONTINGENCE

D E P E N D A N C E

TABLE OF YEUX BY CHEVEUX

YEUX CHEVEUX

Frequency / Percent / Row Pct / Col Pct	BLONDS	BRUNS	NOIRS	ROUX	Total
BLEUS	25 20.16 56.82 55.56	9 7.26 20.45 23.08	3 2.42 6.82 14.29	7 5.65 15.91 36.84	44 35.48
VERTS	13 10.48 27.66 28.89	17 13.71 36.17 43.59	10 8.06 21.28 47.62	7 5.65 14.89 36.84	47 37.90
MARRONS	7 5.65 21.21 15.56	13 10.48 39.39 33.33	8 6.45 24.24 38.10	5 4.03 15.15 26.32	33 26.61
Total	45 36.29	39 31.45	21 16.94	19 15.32	124 100.00

STATISTICS FOR TABLE OF YEUX BY CHEVEUX

Statistic	DF	Value	Prob
Chi-Square	6	15.067	0.020
Likelihood Ratio Chi-Square	6	15.559	0.016
Mantel-Haenszel Chi-Square	1	4.721	0.030
Phi Coefficient		0.349	
Contingency Coefficient		0.329	
Cramer's V		0.246	

Statistic	Value	ASE
Gamma	0.293	0.109
Kendall's Tau-b	0.207	0.078
Stuart's Tau-c	0.213	0.080
Somers' D C|R	0.216	0.081
Somers' D R|C	0.198	0.076
Pearson Correlation	0.196	0.089
Spearman Correlation	0.238	0.089
Lambda Asymmetric C|R	0.127	0.084
Lambda Asymmetric R|C	0.156	0.074
Lambda Symmetric	0.141	0.065
Uncertainty Coefficient C|R	0.048	0.023
Uncertainty Coefficient R|C	0.058	0.028
Uncertainty Coefficient Symmetric	0.052	0.025

Sample Size = 124

3. Exemple n° 3 : association parfaite

EXEMPLE 3: ENTRAINEMENT * PERFORMANCE

EXEMPLE DE STRUCTURE DANS UN TABLEAU DE CONTINGENCE

ASSOCIATION PARFAITE ====> LINEARITE

TABLE OF ENTR BY PERFO

ENTR PERFO

Frequency / Percent / Row Pct / Col Pct	> 5	4-5	< 4	Total
2FOIS	10 27.03 100.00 100.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	10 27.03
4FOIS	0 0.00 0.00 0.00	12 32.43 100.00 100.00	0 0.00 0.00 0.00	12 32.43
8FOIS	0 0.00 0.00 0.00	0 0.00 0.00 0.00	15 40.54 100.00 100.00	15 40.54
Total	10 27.03	12 32.43	15 40.54	37 100.00

STATISTICS FOR TABLE OF ENTR BY PERFO

Statistic	DF	Value	Prob
Chi-Square	4	74.000	0.000
Likelihood Ratio Chi-Square	4	80.277	0.000
Mantel-Haenszel Chi-Square	1	36.000	0.000
Phi Coefficient		1.414	
Contingency Coefficient		0.816	
Cramer's V		1.000	

Statistic	Value	ASE
Gamma	1.000	0.000
Kendall's Tau-b	1.000	0.000
Stuart's Tau-c	0.986	0.028
Somers' D C|R	1.000	0.000
Somers' D R|C	1.000	0.000
Pearson Correlation	1.000	0.000
Spearman Correlation	1.000	0.000
Lambda Asymmetric C|R	1.000	0.000
Lambda Asymmetric R|C	1.000	0.000
Lambda Symmetric	1.000	0.000
Uncertainty Coefficient C|R	1.000	0.000
Uncertainty Coefficient R|C	1.000	0.000
Uncertainty Coefficient Symmetric	1.000	0.000

Sample Size = 37
WARNING: 89% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

A1. Tests et mesures appropriés selon les types de variables

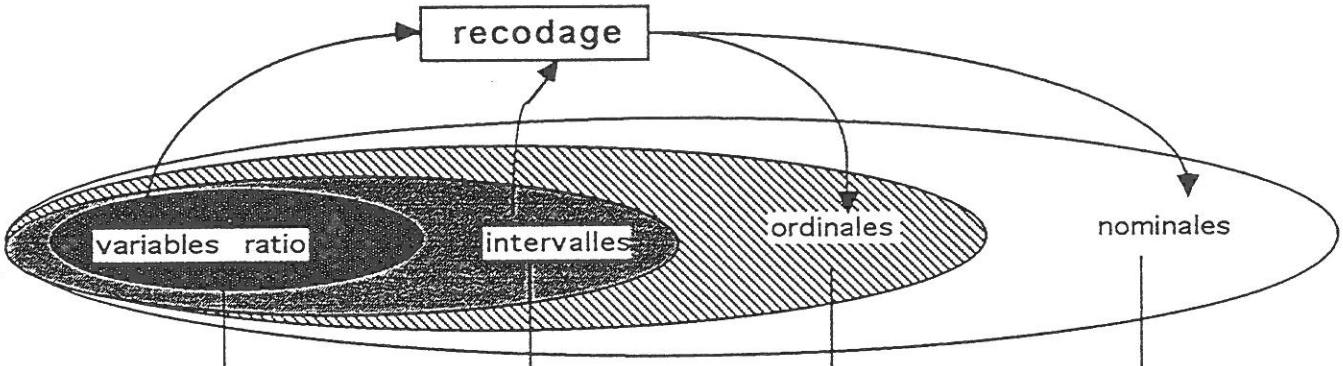
Le schéma de la page suivante synthétise l'ensemble des tests et mesures disponibles dans Proc FREQ selon le type de variables:

- variables nominales
- variables ordinales
- variables intervalles
- variables ratio

FREQ sur Orbites ...

TESTS D' INDEPENDANCE ET MESURES D'ASSOCIATION

appropriés selon le type de variables



corrélation de Pearson

Gamma
Tau b de Kendall
Tau c de Stuart
D de Somer
Corrélation de Spearman
Chi Square de Mantel Haenszel

Chi Square de Pearson
Likelihood Ratio Chi Square
Phi
coef de contingence
V de Cramer
test exact de Fisher
Lambda
Coef d' Incertitude

Recodage:
une variable ratio ou d'intervalle peut être recodée en variable ordinale ou une variable nominale

A2. Historique de la polémique autour du test exact de FISHER

1890 Naissance de R. Fisher

1900 K Pearson publie le test du $\chi 2$ pour une table r*c. Mais il propose pour son test un nombre incorrect de degrés de liberté: r*c -1. En particulier, le test du $\chi 2$ appliqué à la table 2*2 possède 3 degrés de liberté.

1922 R. Fisher, une vingtaine d'années plus tard, propose le nombre correct de degrés de liberté: (r-1) * (c-1) pour une table r*c et donc seulement un degré de liberté pour la table 2*2. K. Pearson n'avouera jamais son erreur.

1925 Fisher propose une règle pratique pour l'application valide du test du $\chi 2$: tous les effectifs théoriques doivent être supérieurs à 5.

1934-1938 Fisher publie son test pour une table 2*2. Fisher précise les raisons pour lesquelles les marges d'une table 2*2 sont des statistiques ancillaires et doivent être donc considérées comme fixées. Yates propose une correction du test du $\chi 2$.

1945-1947 Barnard propose, à la sortie de la guerre, un test fondé sur la distribution binomiale et plus puissant que le test de Fisher.

1949 Convaincu par les arguments que Fisher lui propose, Barnard se retracte et reconnaît publiquement que son test n'est pas adéquat.

1962 Mort de R. Fisher. Celui-ci est considéré comme le plus grand statisticien depuis le début du siècle;
Il est en effet le père incontesté de l'analyse inductive des données.

1978 Près d'un demi-siècle plus tard après la parution du test exact de Fisher, Berkson publie un article dans lequel il expose les raisons pour lesquelles le $\chi 2$ de Pearson a de meilleures propriétés que le test exact de Fisher et le $\chi 2$corrigé de Yates.

1979 Kempthorne précise que l'emploi systématique du test de Fisher n'est pas toujours adéquat selon la situation expérimentale.

1982 Upton fait une revue exhaustive de l'ensemble des tests usuels appliqués a une table 2*2 et aboutit aux mêmes conclusions que celles de Berkson et Kempthorne.

1984 F. Yates publie un article pour faire une mise au point sur le débat.

Actuellement, d'autres articles sur le sujet continuent d'être publiés dans les revue de statistiques: la polémique se poursuit.

Source: GROUIN JM. Test Usuels de Signification dans une table de contingence 2*2 à l'aide de la procédure FREQ. SAS CLUB 1990

A3. Vocabulaire de la Proc FREQ

attendu (espéré)	expected
degré de liberté (ddl)	degree of freedom (DF)
effectif=fréquence absolue[4]	frequency
effectif théorique (fréquence absolue)	expected frequency
erreur-type asymptotique	Asymptotic Standard error
fréquence (relative)[4]	percent - proportion
fréquence absolue = effectif	frequency
rapport des chances (cote)	odds ratio
tableau de contingence	contingency table
tableau croisé	cross tabulation
tableau de fréquences à une dimension	one-way frequency
tableau de fréquences à deux dimensions	two-way frequency
tableau de fréquences à n dimensions	n-way frequency

[4] La littérature anglo-saxonne dénomme les effectifs: FREQUENCY, alors que le mot fréquence en français correspond aux fréquences relatives.

Bibliographie

Ouvrages

- AGRESTI A,(1984), Analysis of Ordinal Categorical Data , WILEY
- AGRESTI A, (1990), Categorical Data Analysis, WILEY
- GROSBRAS JM, livre en préparation
- LANCRY PJ, (1982), Théorie de l'Information et Economie, ECONOMICA
- MORICE et CHARTIER, (1954), Méthodes statistiques, INSEE
- PARTRAT C,(1991), support de cours 2ième année ISUP
- ROUANET H. et alii, (1990), Statistique en sciences humaines: Analyse Inductive des Données, DUNOD
- SAPORTA G, (1990), Probabilités Analyse de Données Statistique, TECHNIP
- SAUTORY O, (1983), La Statistique Descriptive et S.A.S, Manuel ENSAE-INSEE
- SAUTORY O, (1995), Statistique Descriptive avec le Système S.A.S, INSEE -GUIDES n°1-2
- SCHWARTZ D, (1963), Méthodes statistiques à l'usage des médecins et des biologistes, FLAMMARION
- SIEGEL S, (1956), Non parametric Statistics for the Behavioral Sciences, WILEY.
- SAS System for Elementary Statistical Analysis, SAS Institute 1987
- SAS User's Guide STATISTICS -The FREQ Procedure- version 5 et version 6 SAS Institute

Articles

- BOUYER J, (1991), La régression logistique en épidémiologie, Revue Epidémiologie et Santé Publique, MASSON
  Partie I, 1991,39,79-87
  Partie II,1991,39,183-196

- GROUIN J.M, (1990), Tests usuels de signification dans une table de contingence 2*2 à l'aide de la procédure FREQ, SAS-Club 1990

- ROUX M, (1988), Pondération des contributions en analyse des correspondances quand les nombres de modalités diffèrent grandement: Application en écologie, Les cahiers de l'Analyse de Données Vol XIII 1988 n°4 pp. 459-468.

Série des Documents de Travail 'Méthodologie Statistique'

9601 : "Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population"
G. DECAUDIN, J.-C. LABAT

9602 : "Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises"
N. CARON, P. RAVALET, O. SAUTORY

9603 : "La procédure FREQ de SAS® - Tests d'indépendance et mesures d'association dans un tableau de contingence"
J. CONFAIS, Y. GRELET, M. LE GUEN