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

Josiane Confais, Yvette Grelet, Le Guen Monique

LA PROCÉDURE FREQ DE SAS ®
TESTS D'INDEPENDANCE ET MESURES D'ASSOCIATION DANS UN TABLEAU DE CONTINGENCE

Josiane CONFAIS *(UPMC-ISUP)* [1]
Yvette GRELET *(CEREQ-IDL-LASMAS)* [2]
Monique LE GUEN *(CNRS-MATISSE)* [3]

Résumé
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 : nominale, ordinale 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-clés : Tableau de contingence, tests d'indépendance, mesures d'association

Summary
This paper presents, in a pedagogical way, the different tests and association measurements available in PROC FREQ of SAS software, distinguishing them according to the nature of the variables in presence: categorical, ordinal-scaled. They are then described with comments and applied to various examples. The probabilistic approach based on odd-ratios and logit model is tackled. In order to point out the possible doubts when using a unique test, a "curiosity" is reported revealing the differences in the results obtained from various standpoints. A history of the exact Fisher test allows the reader to confort his opinion.

Note : *Une première version de ce document fut d'abord publié en 1992 à l'Université d'Orléans ; il est maintenant épuisé. En 1996 il a été complété et réédité à l'INSEE sous forme d'un document de travail de la Direction des Statistiques Démographiques et Sociales, n° 9603, 78 pages, toujours disponible. En 1997, il fut de nouveau réédité dans la revue MODULAD de Juin 1997 dans une version un peu raccourcie (50 pages). C'est cette version remaniée qui est publiée par la Revue MODULAD dans son numéro 33.*

® SAS, le système SAS sont les marques déposées de SAS Institute Inc., Cary, NC, USA

[1] JOSIANE CONFAIS, Université Pierre et Marie Curie (Paris 6) - ISUP, Boîte 157, 4 Place Jussieu, 75252 Paris Cedex 05 confais@ccr.jussieu.fr

[2] YVETTE GRELET, CEREQ-LASMAS-IdL, MRSH, Université de Caen, Esplanade de la Paix, 14032 Caen Cedex grelet@mrsh.unicaen.fr

[3] MONIQUE LE GUEN, CNRS-MATISSE, Maison des Sciences Economiques, 106-112 Bd de l'Hôpital, 75647 Paris Cedex 13 leguen@univ-paris1.fr

SOMMAIRE

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 vivacités du capitaine TIC "16 Mars 1861
de Eugène Labiche (1815-1888)

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

Mais au XXI^{ème} siècle, 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.

Introduction

La procédure FREQ de SAS 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.**

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-ratios et le modèle logit.
Afin de montrer les doutes que l'on doit avoir lors d'un test unique nous rapporterons en annexe une « curiosité », révélant les discordances des résultats selon les points de vue. En annexe également, un historique sur le test de FISHER permettra au lecteur de conforter son opinion.

Remarque : *nous faisons l'inventaire des tests et mesures de Proc FREQ pour SAS Version 6. D'autres mesures ont été ajoutées [4] dans les versions 8 et 9, dont nous ne parlerons pas ici. Par contre, les exemples et les sorties listing sont exécutées avec la version 8 de SAS, version encore la plus couramment utilisée.*

[4] on trouvera dans les références un article de YELLANKI et SULIGAVI présentant les améliorations de la procédure FREQ dans la version 9.

I – Terminologie

Dans de ce chapitre, nous allons préciser la terminologie élémentaire utilisée par SAS, en montrant les liens entre le codage informatique des données et les traitements statistiques souhaités par l'utilisateur.

I - 1 Variables

Les objets de base traités dans la Procédure FREQ sont des variables.

Exemples:
COULEUR = 'bleu' ; SEXE='1' ; ou SEXE=1 ;
COULEUR ou SEXE représentent le nom de la variable.
'bleu', '1' ou 1 sont des valeurs de la variable appelées **modalités** de la variable.

Dans SAS, les variables peuvent être segmentées selon 2 statuts. Le premier dépend du codage informatique utilisé (variable numérique/ variable caractère[5]), le deuxième dépend du type de traitements statistiques envisagés pour la variable (variable nominale, ordinale, intervalle, ratio, catégorisée).

I - 1 . 1 Le codage informatique

Dans SAS, une variable est soit une variable caractère, soit une variable numérique[6]. Son codage informatique est définit dès la création de la variable de manière implicite ou de manière explicite, et ce statut ne peut être modifié au cours des traitements.

- Par une instruction implicite c'est à dire par contexte
  Exemples d'instructions SAS
  SEXE= 1 ; ➔ définit une variable numérique
  SEXE='1' ; ➔ définit une variable caractère

- Par une instruction explicite
  Exemples d'instructions SAS

`Data table1;`
`input a b c $;`
`cards;`
`1 2 3`
`;`
	Par l'instruction input :
	
	a et b sont des variables numériques
	c est une variable caractère

`data table2;`
`length f $4;`
`input d e f ;`
`cards;`
`4 5 6`
`;`
	Par l'instruction length :
	
	d et e sont des variables numériques
	f est une variable caractère

[5] On dit aussi variable alphanumérique, pour signifier que la variable peut avoir des lettres ou des chiffres comme valeurs.
[6] Cette restriction surprend les utilisateurs de EXCEL qui peuvent mélanger dans une colonne d'un tableur des modalités caractères et des modalités numériques.

I - 1 . 2 Approche liée aux techniques de traitement

L'approche liée aux techniques de traitement statistique fait référence à l'échelle de mesures utilisée pour évaluer la variable.

Terminologie SAS :

- variable **nominale**
- variable **ordinale**
- variable d'**intervalle**
- variable de **rapport**
- variable **catégorisée**

Dans certains modules comme SAS/INSIGHT, cette terminologie doit être connue car elle conditionne les types de traitements statistiques adaptés selon les variables.

• **Variable nominale** *(nominal data)*

Exemples:

SEXE = 'Masculin' ; ➔ variable caractère
SEXE = '1' ; ➔ variable caractère
ou
SEXE = 1 ; ➔ variable numérique

Cette variable est nominale.
Les codes utilisés 'Masculin' '1' ou 1 sont totalement arbitraires
Il n'existe aucune notion de mesure ni de comparabilité entre les **modalités** de la variable sexe.
Une variable nominale est une variable de **classification**.

• **Variable ordinale** *(ordinal data)*

Exemples :

OPINION = 'un peu' ; avec comme autres modalités possibles: 'beaucoup', 'passionnément', etc., 'pas du tout'.

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

Axe des opinions

-	pas du tout	➔	un peu	➔	beaucoup	➔	passionnément	+	
	1		5		10		15		← code
	1		2		3		4		← rang

Une variable ordinale est une variable dont les modalités sont graduelles. On peut leur affecter une valeur numérique en utilisant une échelle. L'échelle peut être un rang.

Pour les variables nominales, les analyses statistiques doivent prendre en compte l'**ordre** des valeurs, et **non les distances** entre les valeurs numériques. Les écarts entre graduations n'ont aucun sens.

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

Exemple:
TEMPERATURE = 10 ;
Une température est une variable d'intervalle. La valeur 10 est une valeur exprimée dans une certaine unité : Celsius, ou Fahrenheit ou Kelvin.

Pour une variable d'intervalle, les valeurs sont ordonnées mais la valeur 0 est une valeur arbitraire. Le 0°C est une référence ici exprimée en Celsius, transposée en Kelvin elle donnerait 273° Kelvin.
La différence entre deux valeurs distinctes de la variable a un sens. La différence entre 5°C et 10°C est comparable à la différence entre 15°C et 20°C.
Par contre faire le rapport de 2 valeurs n'a aucun sens. 30°C n'est pas 2 fois plus élevé que 15°C, c'est seulement beaucoup plus chaud.

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

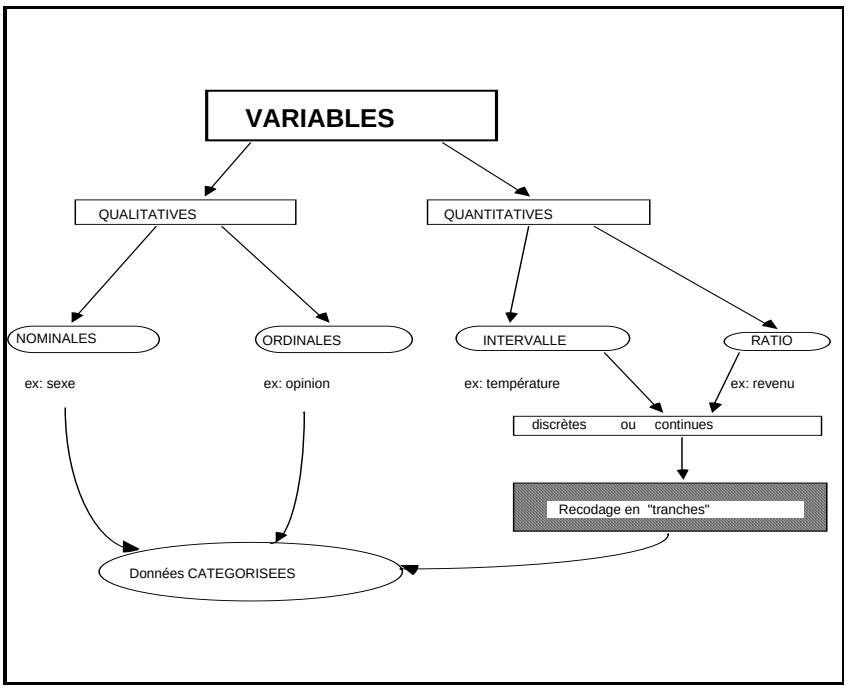
Exemple:
Revenu = 10232.32 ;

On parle de variable de rapport (*ratio data*) lorsque les valeurs sont ordonnées et lorsque la mesure du rapport entre deux valeurs de la variable a un sens.
Un revenu de 10000 francs par exemple est 2 fois plus élevé, qu'un revenu de 5000 francs.
De même 0 Franc même traduit en Deutschemark donne toujours 0 DM !

Remarque :
Depuis quelques années, la plupart des logiciels ne font plus la différence entre variables d'intervalle et variable de rapport. Ainsi SAS/INSIGHT utilise la terminologie *Interval Variable* pour désigner à la fois les variables d'intervalle et les variables de rapport.

• **Variables catégorisées** (*Categorical Data*)

Le schéma de la page suivante résume ce que SAS appelle *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".



I - 1 . 3 Liens entre statut informatique et échelles de mesures

Toute variable SAS définit en caractère est forcément une variable nominale. Par défaut une variable numérique n'est pas nominale. C'est à l'utilisateur de choisir le type d'échelles de mesures. Les propriétés des variables nominales, ordinales, intervalle et ratio étant elles aussi graduelles, l'utilisateur peut selon les besoins abaisser le niveau. Ainsi une variable ratio peut être traitée comme une variable ordinale ou une variable nominale (si le nombre de modalités n'est pas très élevé). L'inverse n'est pas possible. On trouvera une schématisation de ces propriétés en annexe 4.

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

Age	Effectifs
1	22
3	25
6	12
7	13

Tableau d'effectifs ou de fréquences [7]

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

[7] On notera la différence de terminologie : pour les anglo-saxons, un tableau d'effectifs est appelé "frequency table" (c'est un tableau de fréquences absolues), tandis que pour les francophones, un tableau de fréquences est un tableau de fréquences relatives.

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.

Remarque : 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 différentes méthodes disponibles dans plusieurs procédures de SAS.

Types de variables et méthodes

Variables	tableau de fréquences	Statistiques descriptives
nominales	*	
ordinales	*	*
intervalle	*	*
rapport	*	*

⬇ ⬇

Proc FREQ **Proc UNIVARIATE**
Proc MEANS

La procédure FREQ concerne plutôt les variables nominales et ordinales.

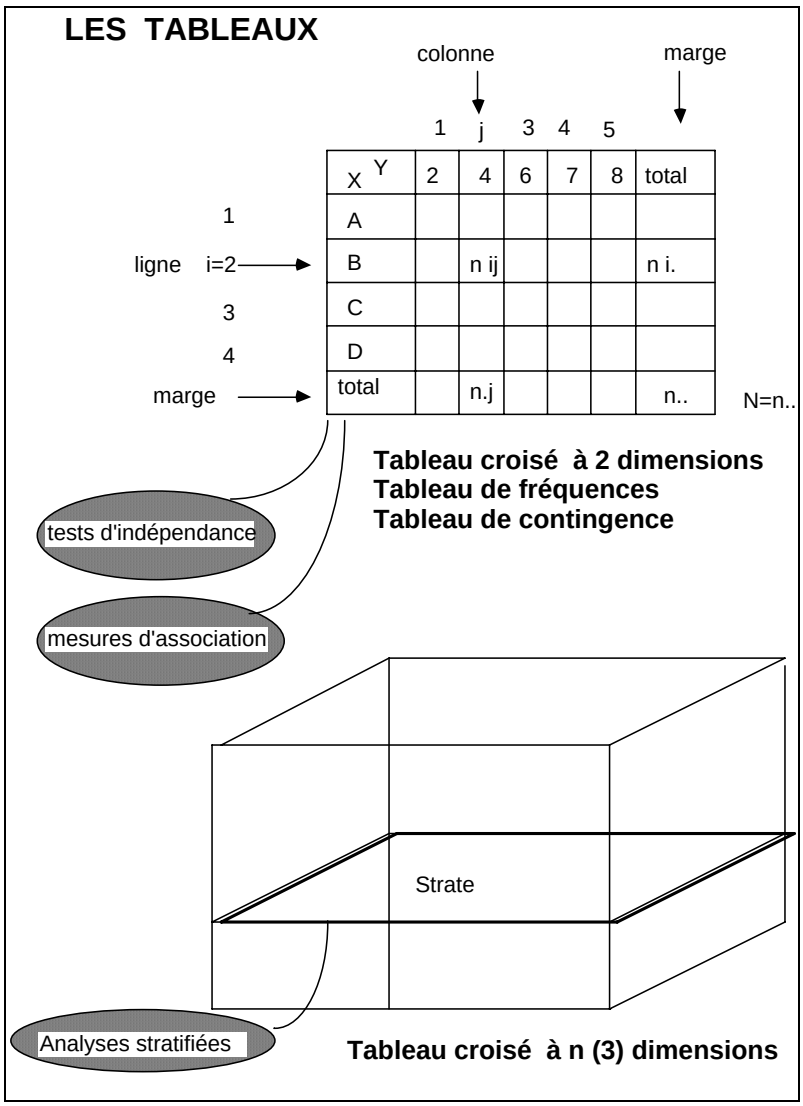
I - 2 . 2 Tableaux de fréquences pour 2 ou n variables

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 ci-dessous, la variable X en ligne prend 4 modalités (A,B,C,D) et la variable Y en colonne 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	$n_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} n_{ij}$

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



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 [8].

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

[8] exemple cité par D. SCHWARTZ p79.

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 ?*

• 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%***
employ**é**	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 des 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 pourcentages.
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 : une liaison particulière, l'association parfaite

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

Epreuve de course à pied

Ici existe une association ou une liaison évidente : plus on s'entraîne plus on court vite.
C'est une structure très forte qui conduit à une **structure de linéarité,** ou de corrélation linéaire 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[9] 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, qui permet 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 annexes A1 A2 A3 les sorties listing de la Proc FREQ effectuées sur les 3 exemples du § I.3.

[9] 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 mesure les liaisons structurelles éventuelles.
C'est pourquoi H. ROUANET distingue les statistiques inférentielles, qui dépendent de la taille de l'échantillon, des statistiques descriptives, indépendantes de la taille de l'échantillon.

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 (SAS version 6)

• 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 coefficient d'incertitude

champ d'application **:** 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 tables (2*2)
• relative risk
• odds ratio

II - Analyse d'un tableau de contingence

II – 1. Description élémentaire du tableau

Les intructions SAS **ci-après** produisent un tableau de **contingence** à partir des données issues de l'enquête d'insertion du CEREQ/ DEP effectuée en 1990.

```
Data CEREQ;
 input DIPLOME $ 1-8 SITU $ 11-17 poids;
 put _infile_;
 cards;
 NON DIPL  CHOMAGE 54
 NON DIPL  MESURE 52
 NON DIPL  EMPLOI 40
 DIPLOMES  CHOMAGE 122
 DIPLOMES  MESURE 97
 DIPLOMES  EMPLOI 133
 ;

Proc FREQ data=cereq order=data;
 tables DIPLOME*SITU;
 weight poids;
 TITLE 'tables DIPLOME*SITU';
 Title3 'Atelier SAS PROC FREQ';
 Title4 "SOURCE: Enquête d'insertion CEREQ-DEP";
 title5 'de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)';
```

Le tableau obtenu 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: Enquête d'insertion CEREQ-DEP
de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)

The FREQ Procedure

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-lignes 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'est 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'est à dire de la colonne marginale.
Le test approprié pour comparer une distribution observée à une distribution théorique est le **test du χ^2** . 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) [10]

[10] SCHWARTZ précise : $N.p \geq 10$ et $N.(1-p) \geq 10$

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, et donc la largeur de l'intervalle est moindre.

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 0.353 *(proportion théorique),* peut-on dire que le pourcentage observé chez les non diplômés (0.37) 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 0.353 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 $p_1 = 0.37$ et $p_2 = 0.347$. Cet écart est-il «significatif» ?
On teste l'hypothèse qu'il n'y a pas de différence entre p_1 et p_2 , c'est à dire que les deux échantillons sont extraits de la même population dans laquelle on suppose que la proportion est p = 0.353.

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 $\sigma = \sqrt{\left[p(1-p)\left((1/N_1)+(1/N_2)\right)\right]}$.
Au risque de 5% on rejettera l'hypothèse si $|p_1 - p_2| > 2\sigma$.

Ici $N_1 = 146$, $N_2 = 352$, $p_1 - p_2 = 0.0233$, $\sigma = 0.047$,
$\Rightarrow$ 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 (KHI-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.

$$\chi^2 = \sum_i \sum_j \frac{\left(n_{ij} - \left(n_{i.}\ n_{.j}/N\right)\right)^2}{\left(n_{i.}\ n_{.j}/N\right)}$$

L'**option CHISQ** de l'instruction TABLES édite, après le tableau croisé, la **valeur de cette statistique** (et d'autres informations qu'on détaillera plus loin) ainsi que le nombre de degrés de liberté [11], et la probabilité associée (voir plus loin le test du χ^2).

```
tables DIPLOME*SITU/CHISQ CELLCHI2 DEVIATION EXPECTED

                    Atelier SAS PROC FREQ
              SOURCE: Enquête d'insertion CEREQ-DEP
de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)

                      The FREQ Procedure

                    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
```

[11] le nombre de degrés de liberté est le nombre de cases n_{ij} qu'il suffit de connaître pour en déduire toutes les autres connaissant $n_{i.}$ et $n_{.j}$

Statistics for Table of DIPLOME by SITU

Statistic	DF	Value	Prob
Chi-Square	2	5.6035	0.0607
Likelihood Ratio Chi-Square	2	5.6809	0.0584
Mantel-Haenszel Chi-Square	1	2.3757	0.1232
Phi Coefficient		0.1061	
Contingency Coefficient		0.1055	
Cramer's V		0.1061	

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) = $\left(n_{ij} - (n_{i.} n_{.j}/N)\right)$

CELLCHI2 (contribution au χ^2) = $\dfrac{\left(n_{ij} - (n_{i.} n_{.j}/N)\right)^2}{(n_{i.} n_{.j}/N)}$

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 factorielle des correspondances sur tableau de contingence.

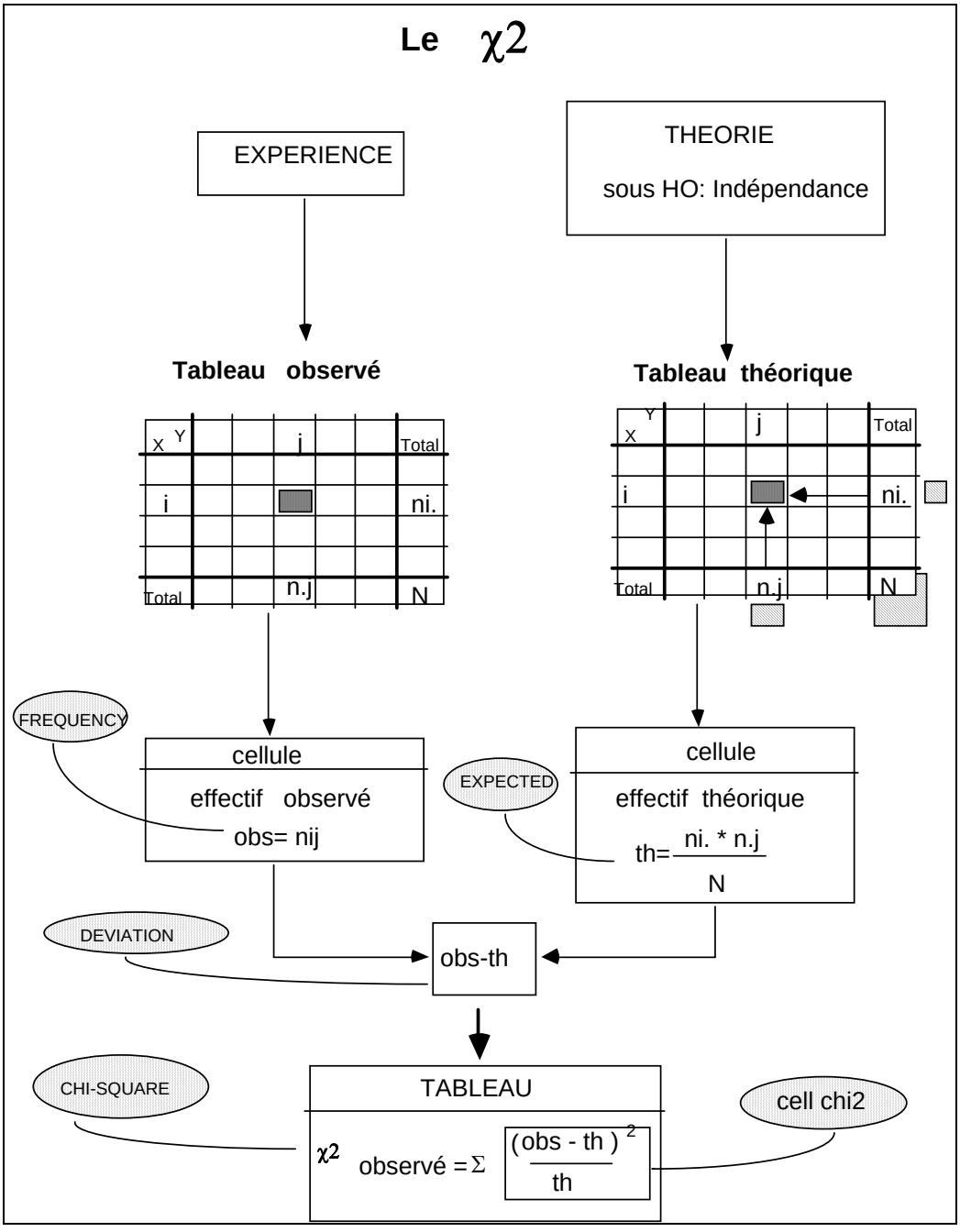
III- Indépendance-Association entre variables nominales

III – 1. Le Test du χ^2

Dans le cas d'une recherche d'indépendance entre la variable ligne et la variable colonne d'un tableau de contingence, on compare 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» du schéma font référence au vocabulaire SAS en sortie de la Proc FREQ.



Le nombre de **d**egrés **d**e **l**iberté d'une table de contingence à r lignes et c colonnes est donné par :

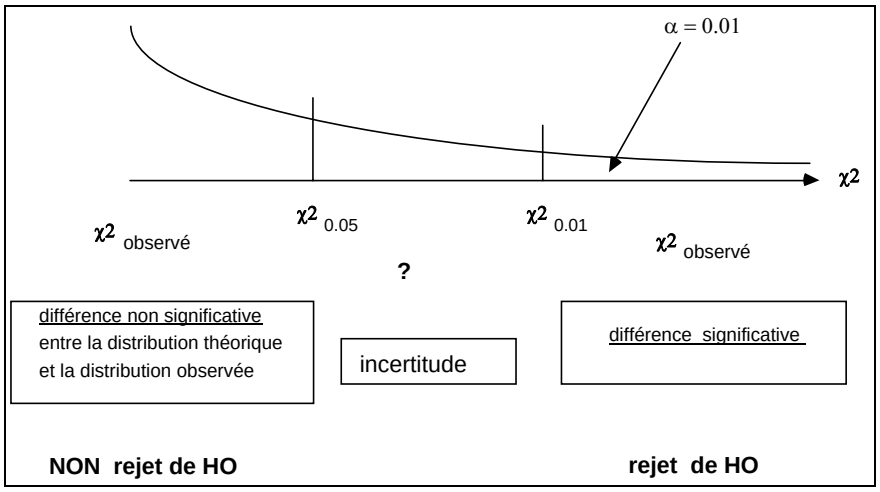
ddl= (r-1) * (c-1)
DF : Degree of Freedom

Comme pour tous les tests de SAS, à chaque valeur d'une statistique calculée (observée) SAS associe la probabilité appelée **p-value**, qui est ici la probabilité d'obtenir une valeur au moins égale à la valeur observée du χ^2 si les deux variables étaient indépendantes. 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**



Conditions de validité : quelques rappels

• Le test du χ^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 toutes les cases doivent être **au moins égaux à 5** pour que le test du χ^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 χ^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 χ^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 χ^2 n'a pas de sens.

• Le test du χ^2 **est sensible à la taille de l'échantillon.**
Pour l'analyse des tableaux obtenus en seconde main les tests du χ^2 doivent être effectués sur les tableaux **avant redressement**.

Jean-Marie GROSBRAS fait la remarque malicieuse **:**
« Il y a toujours moyen d'obtenir un χ^2 significatif (c'est à dire dépassant les valeurs critiques de la table à 5% ou 1%), c'est d'avoir un gros échantillon ».

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

• **association ne signifie pas causalité**

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

χ^2 obs = 21 p-value = 0.000 $\Rightarrow$ test significatif $\Rightarrow$ 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.

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** (ou TABLES).

Les 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 (n_{ij}) avec $N = \sum_{ij} n_{ij}$ effectifs observés

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \left(n_{i.}\ n_{.j}/N\right)\right)^2}{\left(n_{i.}\ n_{.j}/N\right)}$$

L'hypothèse H_0 est l'indépendance, c'est à dire $n_{ij} = \left(n_{i.}\, n_{.j}\right)/N$

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

Le χ^2 de PEARSON prend des valeurs positives.
Il est nul sous H_0 ; en cas d'association «parfaite» entre les 2 variables, il prend une valeur qui dépend de N et du nombre 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 KHI-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».
La loi de ϕ n'étant pas connue de façon théorique, on ne peut l'utiliser pour tester l'indépendance.

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 : $C = \sqrt{\chi^2 / \left(N + \chi^2\right)} = \sqrt{\phi^2 / \left(1 + \phi^2\right)}$

C 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 $\sqrt{1 - 1/r}$), et elle 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 : $V = \phi / \sqrt{\min(r-1, c-1)}$

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 :

- ces mesures sont symétriques en lignes et colonnes ;
- elles sont invariantes par permutations de 2 lignes et/ou 2 colonnes ; il faut donc choisir d'autres mesures si les lignes et/ou colonnes sont ordonnées (variables ordinales) comme on le verra au chapitre IV.
- elles dépendent des valeurs r et c, c'est-à-dire de la taille de la table (sauf V de Cramer) : 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.
- 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 ($n_{i.}$ /N et $n_{.j}$ /N).

SAS donne, dans l'option CHISQ, deux autres mesures qui ont la propriété de suivre des lois du KHI-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 : $G^2 = -2 \text{ Log(RVM)}$.

$$G^2 = 2\sqrt{n_{ij}\, Log\left(n_{ij} \Big/ \left(n_{i.}\, n_{.j}\right) \Big/ N\right)}$$

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

Remarque : La valeur du G^2 est proche de celle du χ^2 d'indépendance si on est «proche» de l'indépendance H_0, ou si N est grand.

- **Qmh appelé Mantel-Haenszel Chi-Square :**

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 KHI-2 à 1 degré de liberté quelle que soit la taille de la table.
On retrouvera cette mesure au chapitre V.

Exemple : enquête d'insertion CEREQ-DEP (cf. § II - 1)
X = diplômes à deux modalités et Y = situations à trois modalités.

Statistics for Table of DIPLOME by SITU

Statistic	DF	Value	Prob
Chi-Square	2	5.6035	0.0607
Likelihood Ratio Chi-Square	2	5.6809	0.0584
Mantel-Haenszel Chi-Square	1	2.3757	0.1232
Phi Coefficient		0.1061	
Contingency Coefficient		0.1055	
Cramer's V		0.1061	

Dans ce tableau le χ^2 et le G^2 sont proches et tous deux légèrement non significatifs : p-value légèrement supérieure à 0.05. Qmh n'a pas de sens car les variables ne sont pas ordinales. Comme r = 2 et c = 3, alors V = ϕ. De plus ϕ est petit donc C est proche de ϕ.

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

Les variables X et Y sont dichotomiques : la table devient $\begin{bmatrix} n_{11} & n_{12} & n_{1.} \\ n_{21} & n_{22} & n_{2.} \\ n_{.1} & n_{.2} & N \end{bmatrix}$

La formule du χ^2 d'indépendance se simplifie alors : $\chi^2 = N\dfrac{\left(n_{11}n_{22} - n_{12}n_{21}\right)^2}{n_{1.}n_{2.}n_{.1}n_{.2}} = N\phi^2$

- Le χ^2 prend des valeurs comprises entre 0 (indépendance) et N (association parfaite). Le test d'indépendance se fait avec une loi du KHI-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 ($\phi = V$).
- Le coefficient C de contingence a des valeurs comprises entre :

0 (sous indépendance) et $\sqrt{1/2} \approx 0.707$ (sous association parfaite).

Remarque : les variables étant dichotomiques, $\phi^2 = r^2$ (r = 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 KHI-2) à une quantité qui est discontinue, YATES a proposé une correction au calcul du χ^2 d'indépendance suivant la formule suivante :

$$Q_c = N\frac{\left(\left|n_{11}n_{22} - n_{12}n_{21}\right| - N/2\right)^2}{n_1 n_2 n_1 n_2} \quad \text{si} \quad \left|n_{11}n_{22} - n_{12}n_{21}\right| > N/2$$

$$Qc = 0 \qquad \text{sinon.}$$

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

- **Qmh Mantel-Haenszel Chi-Square :**

Ici il peut être calculé indifféremment par la formule : $(N-1)\,r^2$ 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 KHI-2 à 1 degré de liberté.

Exemple : Extrait de RADELET (1981) et 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

```
/* Exemple des verdicts de 326 procès en Floride (1976 … 1977)
 issu de Radelet (1981) et cit, dans Agresti (1990)*/
data proces ;
input race $  verdict  $ effectif ;
cards ;
a_blanc  oui 19
a_blanc  non 132
a_blanc  oui 0
a_blanc  non 9
a_noir  oui 11
a_noir  non 52
a_noir  oui 6
a_noir  non 97
;
title 'Etude d''une table 2x2 : RACE / VERDICT ' ;
proc freq data = proces order = data ;
weight effectif ;
table race * verdict / CHISQ ;
run ;
```

Etude d'une table 2x2 : RACE / VERDICT

The FREQ Procedure

Table of race by verdict

race	verdict		
Frequency Percent Row Pct Col Pct	oui	non	Total
a_blanc	19 5.83 11.88 52.78	141 43.25 88.13 48.62	160 49.08
a_noir	17 5.21 10.24 47.22	149 45.71 89.76 51.38	166 50.92
Total	36 11.04	290 88.96	326 100.00

Statistics for Table of race by verdict

Statistic	DF	Value	Prob
Chi-Square	1	0.2214	0.6379
Likelihood Ratio Chi-Square	1	0.2215	0.6379
Continuity Adj. Chi-Square	1	0.0863	0.7689
Mantel-Haenszel Chi-Square	1	0.2208	0.6385
Phi Coefficient		0.0261	
Contingency Coefficient		0.0261	
Cramer's V		0.0261	

Dans ce tableau le χ^2 et le G^2 sont très proches et tous deux non significatifs : p-value très élevée. Qc est lui aussi non significatif, comme Qmh. On retrouve la propriété d'égalité de ϕ, C et V.

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 théorique inférieur à 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.

($n_{1.}$, $n_{2.}$) et ($n_{.1}$, $n_{.2}$) étant fixés, on peut calculer sous l'hypothèse d'indépendance H_0 la probabilité d'obtenir le tableau de contingence :

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \begin{matrix} n_{1.} \\ n_{2.} \end{matrix}$$
$$n_{.1} \quad n_{.2} \quad N$$

(a b c 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 n_{11} suit une loi hyper-géométrique **H** (N, $n_{1.}$, $n_{.1}$ /N) (tirage sans remise de $n_{.1}$ individus parmi N, dans une population où des individus ayant un caractère particulier sont en proportion $n_{1.}$ / N).

$$Prob(n_{11} = a) = \frac{n_{1.}!\ n_{2.}!\ n_{.1}!\ n_{.2}!}{N!\ a!\ b!\ c!\ d!}$$

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 H_0 est l'indépendance, l'alternative H_1 étant «non indépendance» qui peut se dire «les cases du tableau ne sont pas chargées comme sous indépendance».
Dans le cas unilatéral, on fixe une alternative H_1 «non indépendance» du genre «les cases sont plus (ou moins) chargées que sous indépendance».

Choix du «sens» de l'alternative H_1 :

soit $\delta(n_{ii}) = n_{ii} - \frac{n_{i.} n_{.i}}{N}$ $\left[= -\left(n_{ij} - \frac{n_{i.} n_{.j}}{N} \right) \quad si \quad i \neq j \right]$

= «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)
	droit	= l'alternative à H_0 est $\delta > 0$ (plus chargé)
	bilatéral	= l'alternative à H_0 est $\delta \neq 0$ (différent)

On regardera donc δ (a) pour choisir l'alternative, c'est-à-dire la différence (effectif observé-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 :

- test gauche *left* Prob1 = somme des probabilités des tables telles que l'effectif $a \geq n_{11}$;
- test droit *right* Prob2 = idem pour les tables $n_{11} \geq a$;
- 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 ($N \geq 20$) car alors Prob ($n_{11} = a$) = 0.
On n'a donc pas en général : Prob3 = 2 Prob1 (ou 2 Prob2).
Par contre on a toujours : Prob1 + Prob2 = 1 + Prob($n_{11} = a$).

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.
Nota-bene : Dans SAS, les calculs sont aussi possibles si r ou c sont supérieurs à 2, mais ils sont très longs. D'autre part, on a du mal à concrétiser l'alternative dans le cas de table de dimension supérieure à 2x2, car il faut alors plus d'une case pour pouvoir déterminer totalement la table.

Exemple détaillé issu de SIEGEL « Non parametric statistics for the behavioral sciences», repris par SAUTORY ; soit la tables ci-dessous, avec des modalités G1 et G2 en lignes, - + en colonnes :

	-	+	$n_{.j}$
	1	6	7
	4	1	5
$n_{i.}$	5	7	12

L'effectif sous indépendance serait 7x5/12 = 2.916 ; $\delta(n_{11})$ vaut -1.916 donc $\delta(n_{11})$ 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, 7/12) c'est-à-dire la loi de tirage sans remise de 5 individus parmi 12, ayant une caractéristique en proportion 58.33 %.

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 :

$\text{Prob}(n_{11} = 1) = 0.044$
$\text{Prob}(n_{11} = 0) = 0.001 \Rightarrow$ 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 *

* = valeur à choisir pour la table de l'alternative

$\Rightarrow$ Prob (2-tail) = 0.001 + 0.044 + 0.027 = 0.072

Cette probabilité est à comparer à la probabilité du test du χ^2 (non valide) qui est 0.023, comme on le voit dans la sortie de PROC FREQ ci-dessous.

exemple SIEGEL

The FREQ Procedure

Table of group by val

group val

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

Statistics for Table of group by val

Statistic	DF	Value	Prob
Chi-Square	1	5.1820	0.0228
Likelihood Ratio Chi-Square	1	5.5550	0.0184
Continuity Adj. Chi-Square	1	2.8310	0.0925
Mantel-Haenszel Chi-Square	1	4.7502	0.0293
Phi Coefficient		-0.6571	
Contingency Coefficient		0.5492	
Cramer's V		-0.6571	

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

Fisher's Exact Test

Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.0455
Right-sided Pr >= F	0.9987
Table Probability (P)	0.0442
Two-sided Pr <= P	0.0720

Sample Size = 12

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 $Y \mid X$ *qui se lit Y sachant X*
2. λ asymétrique $X \mid Y$
3. λ symétrique

1. λ asymétrique $Y \mid X$

Idée

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 :

- sans aucune information complémentaire ;
- en connaissant la modalité i de la variable X.

• 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	$\Rightarrow$ en effectif	$\underset{j}{Max}\left(n_{.j}\right)$
	$\Rightarrow$ en fréquence	$\underset{j}{Max}\left(n_{.j}/N\right)$

avec une **probabilité d'erreur**

$$p_1 : \quad \text{Proba (erreur(Y))} = 1 - \underset{j}{Max}\left(n_{.j}/N\right)$$

• 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.

$$Y \; Max \Big| X_i \Rightarrow en\, frequence \; \underset{J}{Max}\left(n_{ij}/n_{i.}\right)$$

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

$$p_2 : \quad \text{Proba (erreur Y } | \text{ X)} = 1 - \sum_i \underset{j}{Max}\left(n_{ij}/N\right)$$

A partir de ces deux probabilités (p_1 : proba d'erreur sans information sur X) et (p_2 : proba d'erreur si on connaît X), on définit le ratio appelé Lambda asymétrique Y | X .

$$\lambda \; Y | X = (p_1 - p_2) / p_1$$

$$\lambda Y|X = \frac{\left(\sum_i \underset{j}{Max}\, n_{ij}\right) - \underset{j}{Max}\, n_{.j}}{\left(N - \underset{j}{Max}\, n_{.j}\right)}$$

Lecture de la formule de λ Y | X

$\sum_i \underset{j}{Max}\, n_{ij}$: représente la somme sur toutes les lignes des valeurs maximum des effectifs des cellules sur les colonnes.

$\underset{j}{Max}\, n_{.j}$: représente la valeur maximum des totaux sur les lignes (marge)

Interprétation de λ Y | X

Le ratio **λ 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.

λ 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. Il est toujours compris entre 0 et 1.

• Si **λ Y | X =0** $\Rightarrow$ (p_1=p_2)
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 maxima sont tous repérés sur une même colonne .

• Si **λ Y | X =1** $\Rightarrow$ (p_2=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-ci, 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 :

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

λ 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

Soient 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 :

Q2-Ciné **Q1-Télé**	OUI	NON	**Total**
OUI	20	680	700

NON	80	220	300
Total	100	900	1000

λ **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.

λ **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 fournit 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 deux λ asymétriques.

$$\lambda = \frac{\left(\sum_i Max_i\, n_{ij}\right) + \left(\sum_j Max_i\, n_{ij}\right) - \left(Max_j\, n_{.j} + Max_i\, n_{i.}\right)}{2 * N - \left(Max_j\, n_{.j} + Max_i\, n_{i.}\right)}$$

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

Calcul pour l'exemple couleur des yeux et des cheveux : λ = **(**10 + 12)/ (79 + 77) =0.141.

on a bien 0.141 compris entre λ **Y | X**= 0.127 et λ **X | Y**= 0.156.

Remarque importante pour l'interprétation

- s'il y a **indépendance** alors λ **= 0**

Mais attention le raisonnement réciproque est faux : avoir λ = 0 ne signifie pas toujours avoir indépendance.

- λ **= 1** $\Leftrightarrow$ **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*).

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 :
- Coefficient d'Incertitude **Y** | **X** ;
- Coefficient d'Incertitude **X** | **Y** ;
- Coefficient d'Incertitude symétrique.

Son invention prend origine dans l'approche de la théorie de l'information de SHANNON (1940), dans le 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).
Soit un ensemble d'événements possibles (E_1, ... , E_n), dont les probabilités de réalisation sont (p_1, ..., p_n), supposées connues. Le problème est de 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 :

$$H = -K\sum_{i=1}^{n} p_i Log(p_i)$$

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.

Remarque : 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 économie

- **Entropie**

de X $H(X) = -\sum_i (n_{i.}/n) Log(n_{i.}/n)$

de Y $H(Y) = -\sum_j (n_{.j}/n) Log(n_{.j}/n)$

de XY $H(XY) = -\sum_i \sum_j (n_{ij}/n) Log(n_{ij}/n)$

- **Incertitude Asymétrique**

de **Y** | **X** $U(Y/X) = (H(X) + H(Y) - H(XY)) / H(Y)$

de $\mathbf{X} \mid \mathbf{Y}$ $\quad U(X/Y) = (H(X) + H(Y) - H(XY)) / H(X)$

- **Incertitude symétrique**

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

Comparaisons entre U sur λ

- 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.
- 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[12]), utilisent cette propriété que les modalités[13] 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

[12] on ne traitera pas le coefficient polychorique (cf. l'ouvrage de O. SAUTORY).

[13] dans FREQ on peut définir les valeurs des codages des modalités par l'option SCORES.

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 :

$$\frac{\sum_h \sum_{h'} a_{hh'} b_{hh'}}{\sqrt{\left(\sum_h \sum_{h'} a_{hh'}^2\right)\left(\sum_h \sum_{h'} b_{hh'}^2\right)}}$$

(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 $b_{hh'} = y_h - y_{h'}$.

- **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 $r_X(h)$ qui vaut $\sum_{i=1}^{k-1} n_{i.} + (n_{k.} + 1)/2$.

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 $b_{hh'}$.

Remarque : 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 :

$$\sum_h \sum_{h'} a_{hh'} \quad b_{hh'} = 2(C - D)$$

Remarque : 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 (cf. 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, ainsi :

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

Calcul de C et D

	1, ... , j-1	j	**j+1, ..., n**
1 ... i-1	**Concordances**		**Discordances**
i		n_{ij}	
i+1 ... n	**Discordances**		**Concordances**

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.

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

Le nombre C de paires concordantes est donc : $C = (1/2)\sum_i \sum_j n_{ij} C_{ij}$

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 :

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

Le nombre D de paires discordantes est alors : $D = (1/2)\sum_i \sum_j n_{ij} D_{ij}$

En fin de ce paragraphe, on illustrera les calculs sur un exemple.

- **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 $n_{i.}$ ($n_{i.}$ - 1) paires nulles ; le nombre total de termes $a_{hh'}$ nuls est donc $\sum_i n_{i.}(n_{i.}-1)$, et :

$$\sum_h \sum_{h'} a_{hh'}{}^2 = n(n-1) - \sum_i n_{i.}(n_{i.}-1) = n^2 - \sum_i n_{i.}^2$$

De même le nombre total de termes $b_{hh'}$ nuls vaut $\sum_j n_{.j}(n_{.j}-1)$, et $\sum_h \sum_{h'} b_{hh'}{}^2 = n^2 - \sum_j n_{.j}^2$.

$$\tau_b = \frac{2(C-D)}{\sqrt{\left(n^2 - \sum_i n_{i.}^2\right)\left(n^2 - \sum_j n_{.j}^2\right)}}$$

Remarques

- τ_b est plus approprié au cas des tableaux carrés. τ_b=1 en cas de concordance parfaite (tableau chargé sur la diagonale majeure) et -1 en cas de discordance (diagonale mineure).
- 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)** $\gamma = \dfrac{C-D}{C+D}$

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** $\tau_C = \dfrac{2(C-D)}{n^2(m-1)/m}$

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 (cf J.M. 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) :

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

On définira de même :

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

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: Enquête d'insertion CEREQ-DEP
de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)

                    The FREQ Procedure

                Table of DIPLOME by SITU

DIPLOME      SITU

Frequency|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                                 0.1229   0.0773
Kendall's Tau-b                       0.0650   0.0411
Stuart's Tau-c                        0.0683   0.0432

Somers' D C|R                         0.0823   0.0520
Somers' D R|C                         0.0513   0.0325

Pearson Correlation                   0.0691   0.0435
Spearman Correlation                  0.0689   0.0436

Lambda Asymmetric C|R                 0.0342   0.0487
Lambda Asymmetric R|C                 0.0000   0.0000
Lambda Symmetric                      0.0235   0.0337

Uncertainty Coefficient C|R           0.0052   0.0043
Uncertainty Coefficient R|C           0.0094   0.0078
Uncertainty Coefficient Symmetric     0.0067   0.0056

                Sample Size = 498
```

Pour cette table, les étapes du calcul de C (concordances) et D (discordances) sont les suivantes :

- $C_{11} = 97 + 133$, $C_{12} = 133$, $C_{22} = 54$, $C_{23} = 54 + 52$

Donc : $C = (1/2)(54\ C_{11} + 52\ C_{12} + 97\ C_{22} + 133\ C_{23})$

Soit : $C = 54\ (97+133) + 52\ (133) = 12\ 420 + 6\ 916 = 19\ 336$

- $D_{12} = 122$, $D_{13} = 122 + 97$, $D_{21} = 52 + 40$, $D_{22} = 40$

Donc : $D = 40\ (97+122) + 52\ (122) = 8\ 760 + 6\ 344 = 15\ 104$

- $C - D = 4232$, $C + D = 34\ 440$, $n(n-1) = 247\ 506$

On en déduit les valeurs des différentes statistiques. Au vu de leur ASE, on en déduit qu'aucune n'est significative, ce qui est en accord avec les résultats obtenus par l'option CHISQ (cf. § III.2.1).

Exemple : table 2 x 2 - données insertion des jeunes

```
/* table 2*2 */
proc format fmtlib;
value $regroup
   'CHOMAGE ','MESURE '='CHOMAGE & MESURE';

Proc FREQ data=cereq ORDER=data;
tables DIPLOME*SITU /norow nocol nopercent measures;
weight poids;
format SITU $regroup.;
TITLE 'tables DIPLOME*SITU/norow nocol nopercent measures';
Title3 'Atelier SAS PROC FREQ';
Title4 "SOURCE: Enquête d'insertion CEREQ-DEP";
title5 'de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)';
run;
```

tables DIPLOME*SITU/norow nocol nopercent measures

Atelier SAS PROC FREQ
SOURCE: Enquête d'insertion CEREQ-DEP
de Terminale CAP ou BEP Commerce en L.P. (SN, Apprentis exclus)

The FREQ Procedure

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	0.2335	0.1020
Kendall's Tau-b	0.0993	0.0430
Stuart's Tau-c	0.0861	0.0375
Somers' D C|R	0.1039	0.0451
Somers' D R|C	0.0949	0.0413
Pearson Correlation	0.0993	0.0430
Spearman Correlation	0.0993	0.0430
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0078	0.0069

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 χ^2 (cf. article de 1978 de LANDIS, HEYMAN et KOCH, International Statistical Review, 1978, 46 , pages 237-254, cité en référence dans SAS).

On les obtient par l'option CMH de FREQ, qui donne trois statistiques que l'on notera CMH1 CMH2 CMH3.

	Y ordinale	Y nominale
X ordinale	CMH1	
X nominale	CMH2	CMH3

- **CMH1** ***Nonzero correlation*** **: 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 χ^2 à 1 degré de liberté (ddl). C'est la mesure Qmh de l'option CHISQ (cf. III - 2.1).

- **CMH2** ***Row Mean Score Differ*** **: cas où X est nominale (r modalités) et Y ordinale (à c modalités)**

Ce test est basé sur la comparaison des r moyennes des scores de Y, calculées 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 une loidu χ^2 à (r-1) degrés de liberté.

- **CMH3** ***General Association*** **: Cas où X et Y sont nominales**

C'est un test «d'association» entre X et Y.

La statistique suit un χ^2 à (r-1)(c-1) degrés de liberté ;

elle est égale à $\frac{N}{N-1}\chi^2$, où χ^2 est la valeur du KHI-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 alors 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 : Il est identique à celui du § III - 2 . 1

The FREQ Procedure

Table of DIPLOME by SITU

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

Summary Statistics for DIPLOME by SITU

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

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	2.3757	0.1232
2	Row Mean Scores Differ	1	2.3757	0.1232
3	General Association	2	5.5923	0.0610

Total Sample Size = 498

C'est une 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 : la p-value étant légèrement supérieure à 0.5, l'association n'est donc pas significative, comme on l'a déjà constaté avec d'autres statistiques.

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

Plutôt que la table de contingence, on étudie ici la table de probabilité $p_{ij} = n_{ij} / N$

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{matrix} p_{1.} \\ p_{2.} \end{matrix}$$
$$\begin{matrix} p_{.1} & p_{.2} & 1 \end{matrix}$$

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_{ij}).
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 aléatoire
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 (le rapport de chances) peut lui se traduire par «cote» comme dans les paris.

Lois conditionnelles sachant les lignes $(p_{j/i} = p_{ij} / p_{i.})$

- Sachant la ligne 1

la probabilité d'être en colonne 1 : $p_{1/1} = \frac{p_{11}}{p_{1.}}$

la probabilité d'être en colonne 2 : $p_{2/1} = \frac{p_{12}}{p_{1.}}$

➔ rapport = odds $\Omega 1 = \frac{p_{1/1}}{p_{2/1}} = \frac{p_{11}}{p_{12}}$

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

- idem sachant ligne 2 $\Omega 2 = \frac{p_{1/2}}{p_{2/2}} = \frac{p_{21}}{p_{22}}$

$$\text{Odds} - \text{Ratio } \theta = \frac{\Omega_1}{\Omega_2} = \frac{p_{1/2} \cdot p_{2/2}}{p_{2/1} \cdot p_{1/2}} = \frac{p_{11} \cdot p_{22}}{p_{12} \cdot p_{21}}$$

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 :

$$\text{Indépendance} \Leftrightarrow \theta = 1 \Leftrightarrow \text{Log}(\theta) = 0 \qquad (\text{si } p_{ij} \neq 0 \text{ pour tout i et j})$$

$\theta > 1 \Rightarrow \Omega 1 > \Omega 2$: 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'avoir une modalité de Y, selon la modalité de la variable X. C'est ce que SAS appelle *Relative Risk (Row1/Row2)*.

Risque relatif : $\text{col1 Risk} = \dfrac{p_{1/1}}{p_{1/2}} = \dfrac{\dfrac{p_{11}}{p_{1.}}}{\dfrac{p_{21}}{p_{2.}}}$: c'est le 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.

Par symétrie : $\text{col2 Risk} = \dfrac{p_{2/1}}{p_{2/2}}$

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 : $\dfrac{\text{col1 Risk}}{\text{col2 Risk}} = \text{Odds} - \text{ratio } \theta$

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 (n_{hij}) où h varie de 1 à k ; i vaut 1 ou 2 ; j vaut 1 ou 2 (l'effectif total est N_h)

L'odds-ratio de la table h est $$OR_h = \frac{n_{h11} \cdot n_{h22}}{n_{h12} \cdot n_{h21}}$$

On peut estimer l'odds-ratio «global» par 2 estimateurs :

Mantel-Haenszel : $$\frac{\sum_h \frac{n_{h11} n_{h22}}{N_h}}{\sum_h \frac{n_{h12} n_{h21}}{N_h}}$$

Logit : $$exp\left[\frac{\sum_h w_h log(OR_h)}{\sum_h w_h}\right] \quad \text{où} \quad w_h = \frac{1}{var(log(OR_h))}$$

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 KHI-2 à (k-1) degrés de liberté. Ce test n'est applicable que si N_h 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 (avec l'option MEASURES si on veut l'odds-ratio de chaque table)

Exemple : croisement RACE et VERDICT pour l'exemple du § III - 2 . 2 .
On est alors dans un cas X fixé (RACE) Y aléatoire (VERDICT), c'est-à-dire *cohort.*

```
proc freq data = proces order = data ;
weight effectif ;
table race * verdict / noprint MEASURES ;
run ;
```

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	1.1811	0.5902	2.3634
Cohort (Col1 Risk)	1.1596	0.6255	2.1496
Cohort (Col2 Risk)	0.9818	0.9094	1.0600

Sample Size = 326

- θ est plus grand que 1, donc les Blancs ont plus de « chance » d'encourir 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'encourir la peine de mort : il est plus fréquent pour les Blancs, mais non significatif.

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	p_1	$1-p_1$
X = 0	p_0	$1-p_0$

On étudie ici Y = **présence de la maladie**, qui prend 2 modalités :
M+ = (Y = oui) et M- = (Y = non) .

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

Risque relatif = Prob (M+ | X=1) / Prob(M+ | X=0)=p_1/p_0

Odds-ratio θ = Prob (M+|X=1) Prob(M-|X=0) /Prob(M -|X=1) Prob(M+|X=0)
= $p_1 (1- p_0) / p_0 (1- p_1)$
= $(p_1 / (1- p_1)) /(p_0 / (1- p_0))$

Par définition on nomme ***logit de p***, la fonction de p qui s'exprime Log[p/(1-p)].

⇒ $\text{Log}(\theta) = \text{logit}(p_1) - \text{logit}(p_0)$

Dans le ***Modèle Logistique***, on pose : probabilité (M+ | X) = $f(X) = p = \frac{1}{1+\exp(a+bX)}$

c'est à dire : logit (p) = a+bX (d'où le nom "modèle logistique")

Dans ce modèle particulier, Log (θ) est un estimateur de b, car logit (p_1)=a+b , logit (p_0)=a.

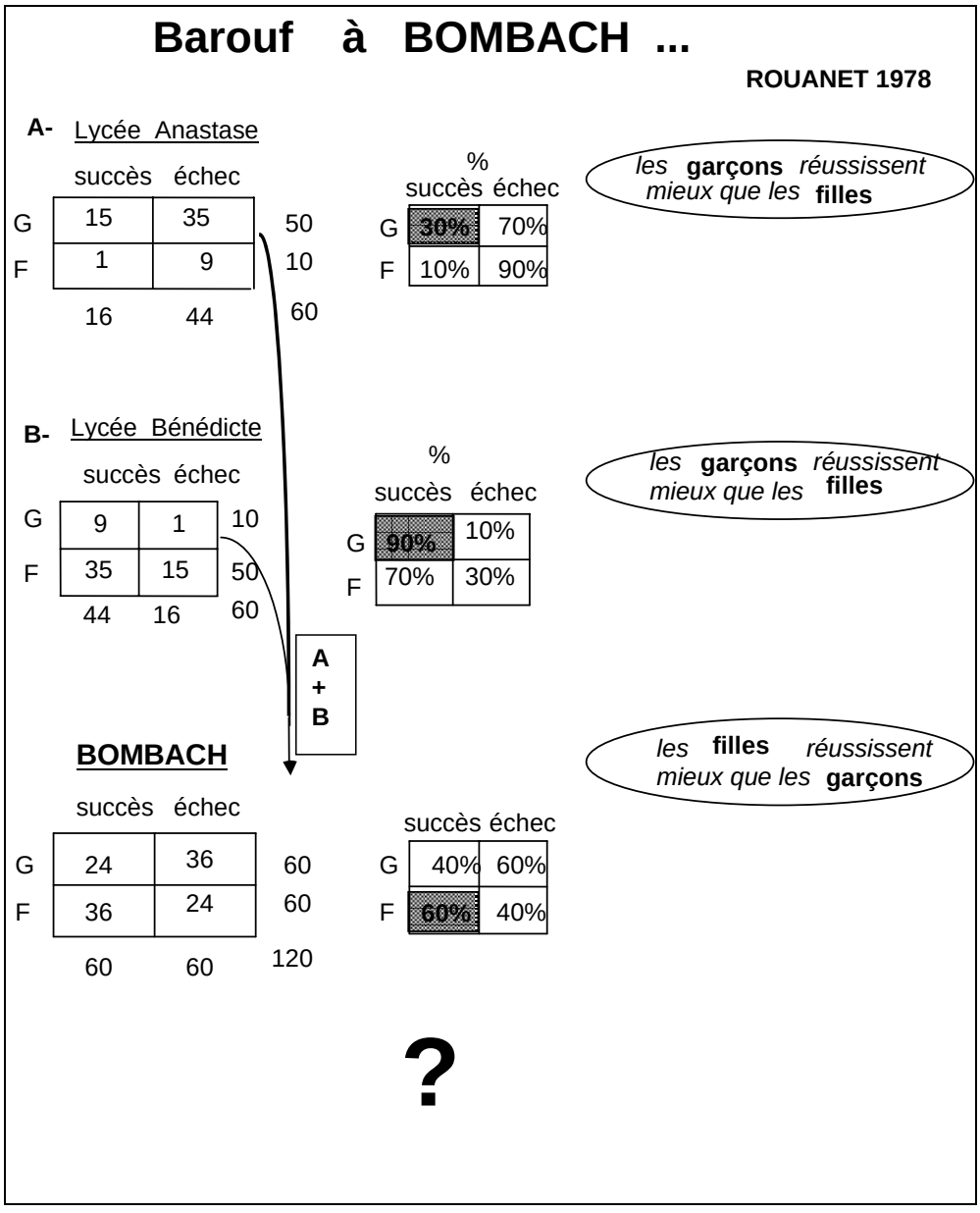
VII. Curiosité

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 de méthodologie sociologique (n°6, avril 1985, pages 3-27).

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.

C'est un exemple de situation statistique conflictuelle, appelé Paradoxe de Simpson.



Annexes

Annexe 1 : Exemple d'indépendance

```
DATA ELEVES;
   INPUT PCS $  NIVEAU $6.
EFFECTIF;
 CARDS;
CADRE  - -   2
CADRE   -    2
CADRE   +   12
CADRE  + +  24
EMPLO  - -   2
EMPLO   -    2
EMPLO   +   12
EMPLO  + +  24
;
 PROC FREQ DATA=ELEVES
ORDER=DATA;
      TABLES PCS*NIVEAU/  CHISQ
MEASURES;
      WEIGHT EFFECTIF;
TITLE1"EXEMPLE 1: PCS PERE *
NIVEAU DES ELEVES  ";
TITLE2"--------------------------
--------------";
run;
```

EXEMPLE 1: PCS PERE * NIVEAU DES ELEVES

The FREQ Procedure

Table of PCS by NIVEAU

PCS / 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	0.0000	1.0000
Likelihood Ratio Chi-Square	3	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0000	1.0000
Phi Coefficient		0.0000	
Contingency Coefficient		0.0000	
Cramer's V		0.0000	

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

Statistics for Table of PCS by NIVEAU

Statistic	Value	ASE
Gamma	0.0000	0.2061
Kendall's Tau-b	0.0000	0.1076
Stuart's Tau-c	0.0000	0.1123
Somers' D C|R	0.0000	0.1123
Somers' D R|C	0.0000	0.1030
Pearson Correlation	-0.0000	0.1118
Spearman Correlation	-0.0000	0.1118
Lambda Asymmetric C|R	0.0000	0.0000
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0000	0.0000
Uncertainty Coefficient C|R	0.0000	0.0000
Uncertainty Coefficient R|C	0.0000	0.0000
Uncertainty Coefficient Symmetric	0.0000	0.0000

Sample Size = 80

On vérifie que toutes les statistiques sont nulles : cas d'indépendance « parfaite ».

Annexe 2 : Exemple de dépendance

```
DATA COULEUR;
    INPUT YEUX $ CHEVEUX $
EFFECTIF;
CARDS;
BLEUS BLONDS 25
BLEUS BRUNS   9
BLEUS NOIRS   3
BLEUS ROUX    7
VERTS BLONDS 13
VERTS BRUNS  17
VERTS NOIRS  10
VERTS ROUX    7
MARRONS BLONDS  7
MARRONS BRUNS  13
MARRONS NOIRS   8
MARRONS ROUX    5
;
 PROC FREQ DATA=COULEUR
ORDER=DATA;
      TABLES YEUX*CHEVEUX/ CHISQ
MEASURES;
      WEIGHT EFFECTIF;
TITLE1"EXEMPLE 2: COULEURS YEUX *
COULEURS DES CHEVEUX";
TITLE2"-----------------------------------
-----------";
run;
```

EXEMPLE 2: COULEURS YEUX * COULEURS DES CHEVEUX

The FREQ Procedure

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	15.0666	0.0197
Likelihood Ratio Chi-Square	6	15.5592	0.0163
Mantel-Haenszel Chi-Square	1	4.7210	0.0298
Phi Coefficient		0.3486	
Contingency Coefficient		0.3292	
Cramer's V		0.2465	

Les statistiques dérivées du KHI-2 sont significatives (p-value faibles).

Statistics for Table of YEUX by CHEVEUX

Statistic	Value	ASE
Gamma	0.2928	0.1088
Kendall's Tau-b	0.2069	0.0784
Stuart's Tau-c	0.2134	0.0804
Somers' D C|R	0.2157	0.0813
Somers' D R|C	0.1984	0.0757
Pearson Correlation	0.1959	0.0889
Spearman Correlation	0.2377	0.0889
Lambda Asymmetric C|R	0.1266	0.0837
Lambda Asymmetric R|C	0.1558	0.0736
Lambda Symmetric	0.1410	0.0646
Uncertainty Coefficient C|R	0.0475	0.0231
Uncertainty Coefficient R|C	0.0577	0.0280
Uncertainty Coefficient Symmetric	0.0521	0.0253

Sample Size = 124

Avec les ASE, on vérifie que les statistiques ci-dessus reflètent une association entre les 2 variables.

Annexe 3 : Exemple d'association parfaite

```
DATA COURSE ;
    INPUT ENTR $  PERFO $6.
EFFECTIF;
CARDS;
2FOIS  < 4    0
2FOIS  4-5    0
2FOIS  > 5   10
4FOIS  < 4    0
4FOIS  4-5   12
4FOIS  > 5    0
8FOIS  < 4   15
8FOIS  4-5    0
8FOIS  > 5    0
;
 PROC FREQ DATA=COURSE
ORDER=DATA;
      TABLES ENTR*PERFO  /  CHISQ
MEASURES;
      WEIGHT EFFECTIF;
TITLE1"EXEMPLE 3: ENTRAINEMENT *
PERFORMANCE";
TITLE2"-------------------------------------
";
run;
```

EXEMPLE 3: ENTRAINEMENT * PERFORMANCE

The FREQ Procedure

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

Statistic	DF	Value	Prob
Chi-Square	4	74.0000	<.0001
Likelihood Ratio Chi-Square	4	80.2770	<.0001
Mantel-Haenszel Chi-Square	1	36.0000	<.0001
Phi Coefficient		1.4142	
Contingency Coefficient		0.8165	
Cramer's V		1.0000	

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

Les statistiques dérivées du KHI-2 sont hautement significatives (p-value proches de 0).

The FREQ Procedure

Statistics for Table of ENTR by PERFO

Statistic	Value	ASE
Gamma	1.0000	0.0000
Kendall's Tau-b	1.0000	0.0000
Stuart's Tau-c	0.9861	0.0276
Somers' D C|R	1.0000	0.0000
Somers' D R|C	1.0000	0.0000
Pearson Correlation	1.0000	0.0000
Spearman Correlation	1.0000	0.0000
Lambda Asymmetric C|R	1.0000	0.0000
Lambda Asymmetric R|C	1.0000	0.0000
Lambda Symmetric	1.0000	0.0000
Uncertainty Coefficient C|R	1.0000	0.0000
Uncertainty Coefficient R|C	1.0000	0.0000
Uncertainty Coefficient Symmetric	1.0000	0.0000

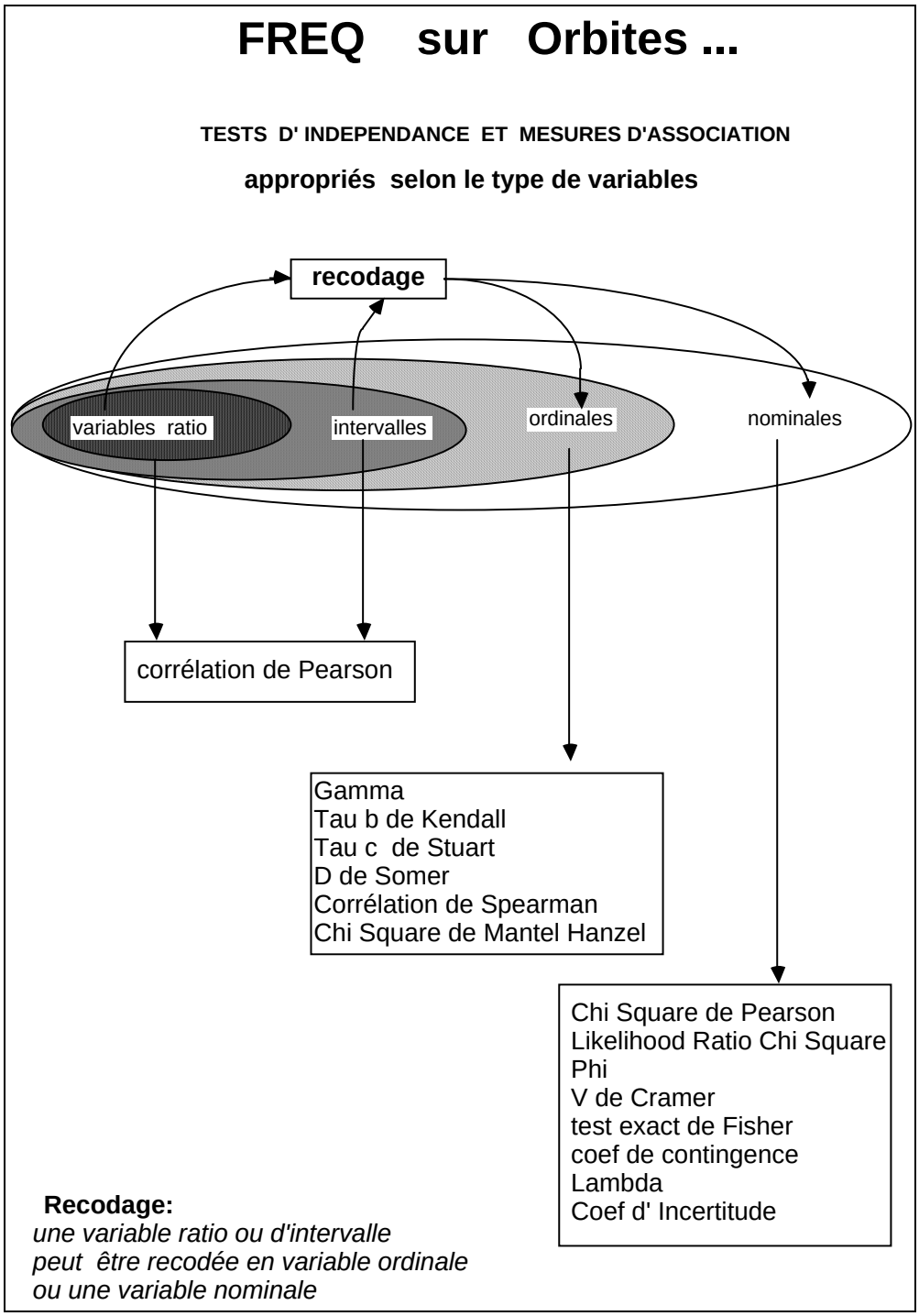
Sample Size = 37

Avec les ASE très petits, on vérifie que les statistiques ci-dessus reflètent une association parfaite entre les 2 variables.

Annexe 4 : Tests et mesures appropriés selon les types de variables

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

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



Annexe 5 : Historique de la polémique autour du test exact de FISHER

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

1890	- Naissance de R. FISHER
1900	- K. PEARSON publie le test du χ^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 χ^2 appliqué à la table 2*2 possède 3 degrés de liberté.
1922	- 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. - PEARSON n'avouera jamais son erreur.
1925	- FISHER propose une règle pratique pour l'application valide du test du χ^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 χ^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 rétracte et reconnaît publiquement que son test n'est pas adéquat.
1962	- Mort de 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 après la parution du test exact de FISHER, BERKSON publie un article dans lequel il expose les raisons pour lesquelles le χ^2 de PEARSON a de meilleures propriétés que le test exact de FISHER et le χ^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 à une table 2*2 et aboutit aux mêmes conclusions que celles de BERKSON et KEMPTHORNE.
1984	- 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.

Annexe 6 : Vocabulaire de la Proc FREQ

attendu (espéré)	expected
degré de liberté (ddl)	degree of freedom (DF)
effectif=fréquence absolue[14]	frequency
effectif théorique (fréquence absolue attendue)	expected frequency
erreur-type asymptotique	asymptotic Standard error
fréquence (relative)	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

[14] 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
- CONFAIS J., GRELET Y., LE GUEN M. (1996), La procédure FREQ de S.A.S., document de travail "Méthodologie statistique" de l'INSEE n° F9610
- FRIENDLY M. (2000), Visualizing Categorical Data Analysis, SAS Institute.
- GROSBRAS J.M. (1990), Notes de cours ENSAE
- LANCRY P.J. (1982), Théorie de l'Information et Economie, ECONOMICA
- MORICE & CHARTIER (1954), Méthodes statistiques, INSEE
- NOVI M. (1998), Pourcentages et tableaux Statistiques, Série "Que Sais-Je ?", n° 3337, PUF.
- PARTRAT C. (1991), support de cours 2ème 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
- SAS ® - STAT User's Guide -The FREQ Procedure - version 6, SAS Institute
- SAUTORY O. (1995), La Statistique Descriptive avec le système SAS®, INSEE-GUIDE n° 1-2
- SCHLOTZHAUER S.D. et LITTELL R.C. (1987), SAS ® System for Elementary Statistical Analysis, SAS Institute
- 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.
- YELLANKI J. N. & SULIGAVI R., (2005), "*What's New in Proc Freq Procedure, Version 9*", SAS Conference Proceedings: PharmaSUG 2005, May 22-25, 2005, Phoenix, Arizona. http://www.lexjansen.com/pharmasug/2005/coderscorner/cc23.pdf

Articles

- BOUYER J. (1991), La régression logistique en épidémiologie, Revue Epidémiologie et Santé Publique, MASSON Partie I, 1991,39,79-87 et 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, S.A.S.-Club 1990
- LE GUEN M . (2003), Tableaux croisés et Diagrammes en Mosaïque, pour Visualiser les probabilités marginales et conditionnelles, BMS, n°77, pp 62-79, January 2003
http://matisse.univ-paris1.fr/doc2/leguen1491.pdf

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

Sites Internet

Home page de Friendly M., spécialiste de la visualisation des données catégorielles
http://www.math.yorku.ca/SCS/friendly.html
CatTrees : Dynamic Visualization of Categorical Data Analysis using Treemaps
http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Kolatch_Weinstein/
Visualizing Categorical Data
http://www.math.yorku.ca/SCS/vcd/

Liens à partir de la SFDS –Société Française de Statistique
http://www.sfds.asso.fr/liens/c_lien01.htm

Cours ST@TNET Multimedia du CNAM
http://www.agro-montpellier.fr/cnam-lr/statnet/cours1.htm

Logiciel SEL Statistiques en ligne de l'INRIA
http://www.inrialpes.fr/sel/

Logiciel TRI2 de Philippe CIBOIS pour le dépouillement d'enquêtes
http://perso.wanadoo.fr/cibois/SitePhCibois.htm

Lexique anglais-français d'écologie numérique et de statistique
http://www.bio.umontreal.ca/Casgrain/lex/lexique.html

Dictionnaire électronique de Statsoft
Une mine d'informations pour apprendre les statistiques
http://www.statsoftinc.com/textbook/stathome.html

Glossaire de statistiques de l'Université de Lancaster
http://www.cas.lancs.ac.uk/glossary_v1.1/
Avec entrée par alphabet :
http://www.cas.lancs.ac.uk:80/glossary_v1.1/Alphabet.html
Un exemple sur le thème des Données Catégorielles :
http://www.cas.lancs.ac.uk/glossary_v1.1/presdata.html#catdat