uOttawa

L'Université canadienne
Canada's university

t

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Marie-Pascal Berthelot
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (mathématiques)
GRADE / DEGREE

Département de mathématiques et statistique
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Tests non-paramétriques de tendance sur les proportions

TITRE DE LA THÈSE / TITLE OF THESIS

M. Alvo
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

D. McDonald

F. Theberge

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

TESTS NON-PARAMÉTRIQUES DE TENDANCE SUR LES PROPORTIONS

Marie-Pascal Berthelot

Thèse soumise à la Faculté des études supérieures et postdoctorales
en vue de l'obtention de la
Maîtrise ès sciences en Mathématiques[1]

Département de mathématiques et de statistique
Faculté des sciences
Université d'Ottawa

[1]le programme de maîtrise est un programme conjoint avec l'Université Carleton, administré par l'Institut d'études supérieures et de recherche en mathématiques et en statistiques d'Ottawa-Carleton

Your file Votre référence
ISBN: 978-0-494-73857-3
Our file Notre référence
ISBN: 978-0-494-73857-3

Résumé

Le problème de tester une tendance monotone dans les proportions a été fréquemment discuté dans la littérature et dans plusieurs applications. Le populaire test de Cochran-Armitage a reçu beaucoup d'attention dans le passé. Il se base sur la mesure de la corrélation entre les proportions de succès observés à n'importe quel temps et un ensemble de constantes monotones qui imitent une tendance dans le temps. Malheureusement, le test de Cochran-Armitage est sensible au choix de ces constantes ainsi qu'aux tailles d'échantillons. Dans cette thèse, on propose deux tests non-paramétriques basés sur les rangs des données. Ces tests se basent sur la notion de compatibilité introduite par Alvo et Cabilio et requièrent la spécification d'une fonction de distance entre les permutations. Nous avons dérivé les tests statistiques qui correspondent aux distances de Spearman, Kendall et Hamming. Il est démontré que les distances de Spearman et de Kendall mènent au même test statistique. On montre que les distances asymptotiques sous l'hypothèse nulle des statistiques de Spearman et de Hamming sont toutes deux normales. On étudie les distributions non nulles par simulation sous des scénarios variés. On conclut que la statistique de Spearman a généralement une puissance plus élevée et est donc le test recommandé. On remarque cependant que la statistique de Hamming modifiée obtient une puissance plus élevée que celle de Hamming et obtient une puissance très près de celle de Spearman dans certain cas.

Abstract

The problem of testing for a monotone trend in proportions has been frequently disccussed in the literature and in various applications. The popular Cochran-Armitage test has received much attention in the past. It is based on measuring the correlation between the proportions of observed success at any time point with a collection of monotone constants which mimic the time trend. Unfortunately the Cochran-Armitage is sensitive to the choice of these constants as well as to the sample sizes. In this thesis, we propose two nonparametric tests based on the ranks of the data. These tests rely on the notion of compatibility introduced by Alvo and Cabilio and required the specification of a distance function between permutations. We derive the statistics tests which correspond to the Spearman, Kendall and Hamming distances. It is shown that the Spearman and Kendall distances lead to the same test statistic. We show that the asymptotic null distributions of the Spearman and Hamming based test statistics are both normal. We study the non-null distributions through simulation under various scenarios. We conclude that the Spearman statistic has generally higher power and is therefore the recommended test. We also note that the modified Hamming based statistic yield a higher power than the Hamming statistic. Furthermore, in some cases the power of the modified Hamming statistic is very close to the power of the Spearman statistic.

Remerciements

Je tiens tout d'abord à exprimer mes profonds remerciements à mon directeur de thèse, le professeur Mayer Alvo, qui a accepté de diriger mes travaux. Ses remarques et questions ont grandement facilité chacune des étapes du processus de recherche. Je le remercie sincèrement pour l'aide compétente qu'il m'a apportée, pour sa patience et son encouragement à finir un travail de longue haleine.

Je tiens également à exprimer ma gratitude envers mes amis les plus chers, qui m'ont été d'un soutien moral précieux et dont mes remerciements leur sont tout naturellement dûs: Luc Barrette, Anick Bruyère, Jamel Chauret, Amay Cheam, Fanny Descary, Jonathan Fillion-Deneault, Élise Lacaille, Olivier Marceau-Robillard, Sébastien Mayer, Annie Philippe et Patrick-André Savard.

Mes plus sincères remerciements sont adressés à mes parents, Jean-Marie Berthelot et Sylvie Morency, dont le soutien moral et matériel fut indispensable à la réussite de cette entreprise. Merci de tous vos encouragements.

Enfin, je voudrais remercier de tout coeur le professeur André R. Dabrowski d'avoir susciter chez moi un intérêt pour les statistiques et de m'avoir donné le goût et le courage de poursuivre des études au niveau de la maîtrise.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

Il y a plusieurs exemples dans la réalité où il serait intéressant de tester s'il y a une tendance dans les proportions. En effet, on peut se demander s'il y a une tendance dans le taux de naissance, le taux de mortalité ou l'incidence d'une certaine maladie dans une région en particulier. Dans cette thèse, comme ensemble de données réelles, nous allons utiliser des données sur les décès en Afrique du Sud pour les années 2000 à 2008.

Année	Nombre de décès	Population
2000	416 155	43 789 115
2001	454 882	43 997 828
2002	502 050	44 187 637
2003	556 779	44 344 136
2004	576 709	42 718 530
2005	598 131	42 768 678
2006	612 778	43 647 658
2007	603 094	43 586 097
2008	592 073	43 421 021

TAB. 1.1 – Données sur les décès en Afrique du Sud entre 2000 et 2008.

Pour d'autres exemples d'applications, le lecteur peut consulter Chen et al [6] et Arase et al [3] pour des essais cliniques contrôles, et Ku et al [13] pour des sondages populationnelles

Dans ce type de problème, on observe au temps t_i un échantillon aleatoire constitué de k variables aléatoires indépendantes $\{y_i\}$ suivant une distribution $Binomiale(N_i\ p_i)$ Notons que les $\{t_i\}$, ou $t_1 < t_2 <\quad < t_k$, représentent des valeurs croissantes de temps Nous sommes intéresses a mesurer et tester la significativité d'une tendance croissante dans les p_i Il existe différentes approches au probleme comme en discute Armitage [4] Une approche basée sur l'analyse de la régression et l'analyse de la variance, qui est décrite dans le chapitre 1, mène au test Cochran-Armitage ([7], [4]) Cette méthode requiert l'utilisation de scores $\{x_i\}$ choisis arbitrairement tel que $x_1 < x_2 <\quad < x_k$ Williams [17] et Chen et al [6] ont noté que lorsque les valeurs espérées de $N_i p_i$, ou $N_i(1-p_i)$ sont petites, l'approximation normale peut devenir peu fiable Par conséquent, le test de Cochran-Armitage devient conservateur et peut mener à un taux d'erreur de type I plus grand que le niveau de signification prescrit Portier et Hoel [16], Horthone et Bretz [12], Corcoran et al [8] ont notes que le test de Cochran-Armitage est très sensible au choix des scores et conclut que l'utilisation général de ce test est suspecte Effectivement, à la page 3045 de l'article [8] on remarque que le seuil du test varie non seulement par rapport au choix des scores, mais également selon les tailles N_i Williams (1998) propose un test conditionnel exact où la statistique est calculée pour chaque table $2 \times k$ possible avec des totaux marginaux egaux Ceci est le modèle d'attribution aléatoire par lequel les groupes ou les traîtements se font assigner aleatoirement les unités expérimentales Les modèles d'attribution aléatoire sont choisis de façon prédominante en recherche biomédicale Neuh*a*user [15] propose une modification de la statistique de Baumgartner-Weiβ-Schindler pour deux echantillons [5] qui utilise des rangs au lieu des scores Une étude par simulation est peu concluante et ne révèle pas clairement lequel est le meilleur entre ce test et le test de Cochran-Armitage

Dans le chapitre 2, nous allons nous attarder aux approches utilisant les méthodes de rangs. Les tests basés sur les distances de Spearman et Hamming sont étudiés au chapitre 3 et 4, respectivement. Finalement, dans le chapitre 5 on fait l'étude des simulations pour étudier la puissance des tests dans plusieurs cas sous l'hypothèse non-nulle, c'est-à-dire lorsqu'il y a une tendance. On traite également d'un exemple avec l'ensemble de données réelles.

1.1 L'approche par l'analyse de la régression et l'analyse de la variance

Considérons les données comme étant sous la forme d'un tableau de contingence $2 \times k$ représenté dans le tableau suivant. Les y_{ij} proviennent des $Bernoulli(p_i)$ et par conséquent les $y_i = \sum_j y_{ij}$ proviennent des $Binomiale(N_i, p_i)$.

	t_1	t_2	t_3	$\cdots$	t_k	Total
Ligne 1	y_1	y_2	y_3	$\cdots$	y_k	$\sum_i y_i = y$
Ligne 2	$N_1 - y_1$	$N_2 - y_2$	$N_3 - y_3$	$\cdots$	$N_k - y_k$	$\sum_i (N_i - y_i) = N - y$
Total	N_1	N_2	N_3	$\cdots$	N_k	$\sum_i N_i = N$

TAB. 1.2 – Table de contingence $2 \times k$

Sans perte de généralité, nous sommes intéressés à mesurer et tester l'importance d'une tendance croissante dans les p_i. Soient $\{x_i\}$ des scores choisis arbitrairement tel que

$$x_1 < x_2 < \cdots < x_k.$$

Alors $\bar{x} = \dfrac{\sum_i N_i x_i}{N}$. $\bar{p}_i = \dfrac{y_i}{N_i}$ est la proportion de succès dans la i^{eme} colonne et $\bar{p} = \dfrac{\sum_i y_i}{N}$ est la proportion totale de succès.
On peut maintenant procéder à l'analyse de la régression selon le modèle suivant

$$E[Y_i|x_i] = p_i = \alpha + \beta(x_i - \bar{x}). \tag{1.1.1}$$

Le modèle en (1.1.1) peut être ajusté à l'aide de la méthode des moindres carrés pondérés avec les poids $\{w_i\}$ satisfaisant à la condition $\sum_i w_i(x_i - \bar{x}) = 0$. Ceci nous permet de tenir compte des différentes tailles d'échantillons. On minimise alors la somme pondéré des erreurs au carré par rapport aux paramètres α et β.

Soit

$$Q = \sum_i w_i(\bar{p}_i - \alpha - \beta(x_i - \bar{x}))^2. \tag{1.1.2}$$

On détermine α et β en dérivant l'équation (1.1.2) une fois par rapport à α et une autre fois par rapport à β et en mettant les équations résultantes égales à zéro.

$$\frac{\partial Q}{\partial \alpha} = -2\sum_i w_i(\bar{p}_i - \alpha - \beta(x_i - \bar{x}))$$

$$\frac{\partial Q}{\partial \beta} = -2\sum_i w_i(x_i - \bar{x})(\bar{p}_i - \alpha - \beta(x_i - \bar{x}))$$

Les estimateurs des paramètres α et β sont les solutions aux deux équations suivantes :

$$\sum_i w_i(\bar{p}_i - \alpha - \beta(x_i - \bar{x})) = 0$$

$$\sum_i w_i(x_i - \bar{x})(\bar{p}_i - \alpha - \beta(x_i - \bar{x})) = 0$$

On a alors que les estimateurs de α et β sont respectivements

$$\hat{\alpha} = \frac{\sum_i w_i\bar{p}_i}{\sum_i w_i}, \qquad \hat{\beta} = \frac{\sum_i w_i\bar{p}_i(x_i - \bar{x})}{\sum_i w_i(x_i - \bar{x})^2}$$

Maintenant, si on prend comme poids $w_i = \dfrac{N_i}{N}$, on obtient

$$\hat{\alpha} = \frac{\sum_i N_i\bar{p}_i}{\sum_i N_i} = \frac{\sum_i y_i}{N} = \bar{p}, \qquad \hat{\beta} = \frac{\sum_i N_i\bar{p}_i(x_i - \bar{x})}{\sum_i N_i(x_i - \bar{x})^2}$$

Donc le modèle de prédiction devient

$$\begin{aligned}\hat{p} &= \hat{\alpha} + \hat{\beta}(x - \bar{x}) \\ &= \bar{p} + \hat{\beta}(x - \bar{x}), \quad \text{pour tout } x\end{aligned}$$

L'hypothèse nulle est qu'il n'y a pas de tendance et que $\beta = 0$. Dans ce cas, $p_i = p$ pour tout i. On sait que

$$E\left[\hat{\beta}\right] = \frac{\sum_i N_i p(x_i - \bar{x})}{\sum_i N_i (x_i - \bar{x})^2} = 0$$

D'autre part $\hat{\beta}$ est un estimateur non biaisé pour β. Donc si $p_i = p$ pour tout i, $\beta = 0$. Aussi,

$$Var\hat{\beta} = \frac{p(1-p)}{\sum_i N_i (x_i - \bar{x})^2}$$

que l'on peut estimer par

$$\widehat{Var\hat{\beta}} = \frac{\bar{p}(1-\bar{p})}{\sum_i N_i (x_i - \bar{x})^2}$$

On peut maintenant procéder à l'analyse de la variance. On peut définir la somme des carrés de la régression, la somme des carrés du manque d'adéquation, la somme des carrés de l'erreur pure et la somme totale des carrés de la manière suivante (Kutner, Nachtsheim, Neter et Li, p.126)[14].

$$\begin{aligned} SSReg &= \sum_i N_i (\hat{p}_i - \bar{p})^2 \\ &= \hat{\beta}^2 \sum_i N_i (x_i - \bar{x})^2 \\ &= \frac{[\sum_i N_i \bar{p}_i (x_i - \bar{x})]^2}{\sum_i N_i (x_i - \bar{x})^2} \end{aligned}$$

$$\begin{aligned} SSLF &= \sum_i N_i (\bar{p}_i - \hat{p}_i)^2 \\ &= \sum_i N_i (\bar{p}_i - \bar{p})^2 - \hat{\beta} \sum_i N_i \bar{p}_i (x_i - \bar{x}) \\ &= \sum_i N_i (\bar{p}_i - \bar{p})^2 - \frac{[\sum_i N_i \bar{p}_i (x_i - \bar{x})]^2}{\sum_i N_i (x_i - \bar{x})^2} \end{aligned}$$

$$SSPE = \sum_i N_i \bar{p}_i (1 - \bar{p}_i)$$

$$SSTO = N\bar{p}(1 - \bar{p})$$

où $SSTO = SSReg + SSLF + SSPE$.

Voici, dans le tableau (1.2), un résumé de l'analyse de la variance

Sources de variation	degrés de liberté	Sommes des carrés (SS)	Carrés moyens (MS)
Régression linéaire	1	SSReg	SSReg/1
Manque d'adéquation	k-2	SSLF	SSLF/(k-2)
Sous Total	k-1	SSReg + SSLF	(SSReg + SSLF)/(k-1)
Erreur pure	N-k	SSPE	SSPE/(N-k)
Total	N-1	SSTO	

TAB. 1.3 – Tableau d'analyse de la variance

Pour expliquer la somme des carrés de l'erreur pure, considérons le tableau (1.3) suivant :

					total
proportions	$\bar{p}_1 = y_1/N_1$	$\bar{p}_2 = y_2/N_2$	$\cdots$	$\bar{p}_k = y_k/N_k$	$\frac{\sum y_i}{\sum x_i} = \bar{p}$
variance	$\bar{p}_1(1-\bar{p}_1)/N_1$	$\bar{p}_2(1-\bar{p}_2)/N_2$	$\cdots$	$\bar{p}_k(1-\bar{p}_k)/N_k$	
SS	$\bar{p}_1(1-\bar{p}_1)$	$\bar{p}_2(1-\bar{p}_2)$	$\cdots$	$\bar{p}_k(1-\bar{p}_k)$	$\bar{p}(1-\bar{p})$

TAB. 1.4 – Tableau de la somme des carrés de l'erreur pure

À chaque temps, t_i, la variance de $\bar{p}_i$ est estimée par $\dfrac{\bar{p}_i(1-\bar{p}_i)}{N_i}$ et donc la somme des

carrés est $\bar{p}_i(1-\bar{p}_i)$. En tenant compte des poids $w_i = N_i$, on a que la somme des carrés de l'erreur pure est

$$SSPE = \sum_i N_i \bar{p}_i (1-\bar{p}_i)$$

De plus, en utilisant l'identité

$$\sum_i N_i (\bar{p}_i - \bar{p})^2 = SSReg + SSLF$$

on peut déduire que

$$SSLF = \sum_i N_i (\bar{p}_i - \hat{p}_i)^2.$$

Nous sommes intéressés à calculer le coefficient de corrélation entre les p_i et les x_i. Pour ce faire, notons que l'on peut écrire $\hat{\beta}$ comme suit

$$\hat{\beta} = \frac{\sum_i N_i (\bar{p}_i - \bar{p})(x_i - \bar{x})}{\sum_i N_i (x_i - \bar{x})^2}.$$

À partir de cette dernière expression de $\hat{\beta}$, on peut calculer le coefficient de corrélation de l'échantillon, $\hat{\rho}$, entre $\{\bar{p}_i - \bar{p}\}$ et $\{x_i - \bar{x}\}$ comme suit (Hogg et Tanis, [11])

$$\begin{aligned}
\hat{\rho} &= \frac{\sum_i N_i (\bar{p}_i - \bar{p})(x_i - \bar{x})}{\sqrt{\sum_i N_i (x_i - \bar{x})^2 \sum_i N_i (\bar{p}_i - \bar{p})^2}} \\
&= \frac{\sum_i N_i (\bar{p}_i - \bar{p})(x_i - \bar{x})}{\sum_i N_i (x_i - \bar{x})^2} \sqrt{\frac{\sum_i N_i (x_i - \bar{x})^2}{\sum_i N_i (\bar{p}_i - \bar{p})^2}} \\
&= \hat{\beta} \sqrt{\frac{\sum_i N_i (x_i - \bar{x})^2}{\sum_i N_i (\bar{p}_i - \bar{p})^2}}
\end{aligned}$$

Notons que le test F habituel est fondé sous l'hypothèse que le terme d'erreur du modèle est distribué normalement, ce qui n'est pas vrai dans le cas de données binaires.

On se retrouve alors face à deux problèmes. Le premier étant de tester que toutes les proportions soient les mêmes. Ceci peut être fait en comparant deux estimations de la variance en calculant le rapport

$$\frac{(SSReg + SSLF)/(k-1)}{(SSPE)/(N-k)} = \frac{[\sum_i N_i (\bar{p}_i - \bar{p})^2]/(k-1)}{[\sum_i N_i \bar{p}_i (1-\bar{p}_i)]/(N-k)}$$

qui a une distribution de $F(k-1, N-k)$ pour de grands échantillons.
Le deuxième problème est de tester la validité de la régression linéaire, ce qui peut être fait en calculant

$$\frac{SSReg}{SSPE/(N-k)} = \frac{\hat{\beta}^2 \sum_i N_i(x_i - \bar{x})^2}{\sum_i N_i \bar{p}_i(1-\bar{p}_i)/(N-k)} \sim F(1, N-k)$$

qui a une distribution de chi-deux avec 1 degré de liberté, i.e. $\chi^2(1)$, pour de grands échantillons. Cependant pour pouvoir utiliser ce test, nous devons spécifier les $\{x_i\}$.
Pour le deuxième problème de tester l'hypothèse que $\beta = 0$, on peut, pour de grande taille d'échantillon, calculer le rapport suivant

$$\begin{aligned} \frac{SSReg}{SSTO/N} &= \frac{[\sum_i N_i \bar{p}_i(x_i - \bar{x})]^2}{\bar{p}(1-\bar{p}) \sum_i N_i(x_i - \bar{x})^2} \\ &= \frac{[\sum_i N_i(\bar{p}_i - \bar{p})(x_i - \bar{x})]^2}{\bar{p}(1-\bar{p}) \sum_i N_i(x_i - \bar{x})^2} \\ &= \frac{\hat{\beta}^2 \sum_i N_i(x_i - \bar{x})^2}{\bar{p}(1-\bar{p})} \\ &= \frac{\hat{\beta}^2}{Var\hat{\beta}} \end{aligned} \tag{1.1.3}$$

qui sous l'hypothèse nulle suit une distribution de chi-carré avec un degré de liberté. Maintenant, si $\bar{p}$ est très petit, d'ordre de $1\% - 2\%$, tel que $\bar{p}^2$ est négligeable, on peut substituer $(1-\bar{p}) \sim 1$ et on trouve que la statistique en (1.1.3) devient

$$\frac{[\sum_i N_i(\bar{p}_i - \bar{p})(x_i - \bar{x})]^2}{\sum_i \bar{p} N_i(x_i - \bar{x})^2} \tag{1.1.4}$$

Au numérateur de l'équation (1.1.4) on a que la différence entre les fréquences observées et les fréquences espérées, $N_i(\bar{p}_i - \bar{p})$. est multiplié par l'effet des scores $(x_i - \bar{x})$, tandis qu'au dénominateur on a une somme de carrés pondérés des écarts à la moyenne des x_i ayant comme poids les fréquences espérées. La statistique donne un test de tendance sur les fréquences au lieu d'un test de tendance sur les proportions.

1.2 L'approche par tableau de contingence

Une alternative pour tester que toutes les proportions soient les mêmes est de traiter les données comme provenant d'un tableau de contingence $2 \times k$. En général, pour des données d'un tableau $2 \times k$ on calcule le test de chi-carré de Pearson de $(k-1)$ degrés de liberté de la manière suivante lorsque nous avons une grande taille d'échantillon. C'est-à-dire, lorsque $N_i \to \infty$ nous avons

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

où les O_{ij} représentent les valeurs observées dans la i^{eme} colonne et la j^{eme} ligne du Tableau (1.1) et les E_{ij} représentent les valeurs espérés correspondantes. Pour le problème qui nous intéresse le test chi-carré de Pearson se calcule comme suit à partir des données du Tableau (1.1). Les valeurs observées de la ligne 1 sont $O_{i1} = y_i$ et les valeurs observées de la ligne 2 sont $O_{i2} = (N_i - y_i)$. Pour obtenir les estimations des valeurs espérées de la ligne 1, on effectue le calcul suivant. Quand $p_1 = p_2 = \cdots = p_k = p$, on peut estimer la proportion p en commun par $\bar{p}$. Par la suite, $E_{i1} = N_i\bar{p}$, $E_{i2} = N_i(1-\bar{p})$.

Par conséquent, le test de chi-carré de Pearson est calculé de la manière suivante

$$\begin{aligned}
\chi^2_{k-1} &= \sum_i \frac{(O_{i1} - E_{i1})^2}{E_{i1}} + \sum_i \frac{(O_{i2} - E_{i2})^2}{E_{i2}} \\
&= \sum_i \frac{(y_i - N_i\bar{p})^2}{(N_i\bar{p})} + \sum_i \frac{((N_i - y_i) - N_i(1-\bar{p}))^2}{N_i(1-\bar{p})} \\
&= \sum_i \frac{(N_i\bar{p}_i - N_i\bar{p})^2}{(N_i\bar{p})} + \sum_i \frac{(N_i(1-\bar{p}_i) - N_i(1-\bar{p}))^2}{N_i(1-\bar{p})} \\
&= \sum_i \frac{N_i(\bar{p}_i - \bar{p})^2}{\bar{p}(1-\bar{p})} \\
&= \frac{SSReg + SSLF}{SSTO/N}
\end{aligned}$$

On rejette l'hypothèse nulle quand $\chi^2_{k-1} > \chi^2_{k-1}(\alpha)$ où $\chi^2_{k-1}(\alpha)$ représente la valeur provenant de la distribution du chi-carré avec $(k-1)$ degrés de liberté pour laquelle l'aire à droite est égale à α.

Il est important de noter que pour utiliser l'approche par la régression il faut spécifier les $\{x_i\}$. Ceux-ci sont habituellement choisis de façon arbitraire. Pour ce qui est de l'approche par tableau de contingence. on utilise cette méthode pour tester l'hypothèse nulle $H_0 : p_i = p$ pour tout i, contre l'alternative $H_1 : p_i \neq p$ pour au moins un i Ce test ne sera pas aussi puissant pour l'hypothèse alternative d'une tendance dans les p_i car cette hypothèse est trop générale.

Chapitre 2

L'approche par les méthodes des rangs

2.1 L'approche par les méthodes des rangs, sans exactos

Dans le chapitre précédent nous nous sommes attardés au problème de tester une tendance dans les proportions en attribuant des scores. Dans cette section, nous allons examiner le problème sous un autre angle, en utilisant des méthodes basées sur les rangs sans avoir besoin de spécifier des scores. Nous commencerons par une brève introduction aux méthodes statistiques basées sur les rangs sans exactos.

Un classement de N objets identifiés par $1, \cdots . N$ est une permutation des entiers $(1, \cdots, N)$. Pour deux classements $\mu = (\mu(1), \cdots, \mu(N))'$ et $\nu = (\nu(1), \cdots, \nu(N))'$, on peut définir, respectivement, les distances de Spearman et Kendall de manière suivante

$$d_s(\mu. \nu) = \frac{1}{2} \sum_{i=1}^{N} (\mu(i) - \nu(i))^2 \tag{2.1.1}$$

$$d_k(\mu,\nu) = \sum_{i<j}[1 - sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j))]$$
$$\text{où } sgn(x) = \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{si } x < 0 \end{cases} \tag{2.1.2}$$

Ces deux distances peuvent être écrites sous la forme

$$d(\mu,\nu) = c - \mathcal{A}(\mu,\nu)$$

où $\mathcal{A}(\mu,\nu)$ représente une fonction de similarité centrée à la moyenne entre μ et ν, tandis que c représente une constante. Dans le cas de Spearman, on a

$$\begin{aligned} d_s(\mu,\nu) &= \frac{1}{2}\sum_{i=1}^{N}(\mu(i) - \nu(i))^2 \\ &= \frac{1}{2}\sum_{i=1}^{N}\left[\left(\mu(i) - \frac{N+1}{2}\right) - \left(\nu(i) - \frac{N+1}{2}\right)\right]^2 \\ &= \sum\left(\mu(i) - \frac{N+1}{2}\right)^2 - \sum_{i=1}^{N}\left(\mu(i) - \frac{N+1}{2}\right)\left(\nu(i) - \frac{N+1}{2}\right) \\ &= c_s - \mathcal{A}_s(\mu.\nu) \end{aligned} \tag{2.1.3}$$

où $c_s = \dfrac{N(N^2-1)}{12}$ et $\mathcal{A}_s = \mathcal{A}_s(\mu,\nu) = \sum_{i=1}^{N}\left(\mu(i) - \frac{N+1}{2}\right)\left(\nu(i) - \frac{N+1}{2}\right)$.
Pour le cas de Kendall, on a

$$\begin{aligned} d_k(\mu,\nu) &= \sum_{i<j}\{1 - sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j))\} \\ &= \sum_{i<j}1 - \sum_{i<j}sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j)) \\ &= \frac{N(N-1)}{2} - \sum_{i<j}sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j)) \end{aligned} \tag{2.1.4}$$

Donc $c_k = \dfrac{N(N-1)}{2}$ et $\mathcal{A}_k = \mathcal{A}_k(\mu,\nu) = \sum_{i<j} sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j))$.
La corrélation entre deux classements μ et ν est définie par

$$\rho(\mu,\nu) = 1 - \frac{2d(\mu,\nu)}{\max_{\mu,\nu} d(\mu,\nu)}$$

où le maximum est pris sur les $N!$ valeurs possibles de μ et ν. Notons que

$$-1 \leq \rho(\mu, \nu) \leq 1.$$

Calculons la corrélation entre μ et ν sous la distance de Spearman et Kendall, respectivement. Pour ce faire nous devons calculer la valeur maximale de $d(\mu. \nu)$ dans chaque cas. Pour la distance de Spearman, on a

$$\begin{aligned}
\max_{\mu,\nu} d_s(\mu. \nu) &= \max_{\mu\ \nu} \frac{1}{2} \sum_{i=1}^{N} (\mu(i) - \nu(i))^2 \\
&= \frac{1}{2} \sum_{i=1}^{N} (i - (N + 1 - i))^2 \\
&= \frac{1}{2} \sum_{i=1}^{N} (2i - (N + 1))^2 \\
&= \frac{N(N^2 - 1)}{6} \\
&= 2\left(\frac{N(N^2 - 1)}{12}\right) \\
&= 2c_s
\end{aligned}$$

Donc la corrélation $\rho_s(\mu, \nu)$ est

$$\begin{aligned}
\rho_s(\mu, \nu) &= 1 - \frac{2d_s(\mu, \nu)}{\max_{\mu,\nu} d_s(\mu. \nu)} \\
&= 1 - \frac{2[c_s - \mathcal{A}_s]}{2c_s} \\
&= \frac{\mathcal{A}_s}{c_s}
\end{aligned}$$

Dans le cas de Kendall, on a

$$\begin{aligned}
\max_{\mu,\nu} d_k(\mu, \nu) &= \max_{\mu,\nu} \sum_{i<j} \{1 - sgn(\mu(i) - \mu(j)) sgn(\nu(i) - \nu(j))\} \\
&= \frac{N(N-1)}{2} - \sum_{i<j} (-1) \\
&= N(N-1) \\
&= 2c_k
\end{aligned}$$

La corrélation $\rho_k(\mu, \nu)$ est donc

$$\begin{aligned}\rho_k(\mu,\nu) &= 1 - \frac{2d_k(\mu,\nu)}{\max_{\mu,\nu} d_k(\mu,\nu)} \\ &= 1 - \frac{2[c_k - \mathcal{A}_k]}{2c_k} \\ &= \frac{\mathcal{A}_k}{c_k}\end{aligned}$$

On remarque que dans le cas de ces deux distances la corrélation peut être définie comme suit

$$\rho(\mu,\nu) = \frac{2\mathcal{A}(\mu,\nu)}{\max_{\mu,\nu} d(\mu,\nu)} = \frac{\mathcal{A}(\mu,\nu)}{c} \tag{2.1.5}$$

Il est possible d'obtenir des corrélations définies en (2.1.5) en utilisant des fonctions de distance autres que celles de Spearman et Kendall définies respectivement aux équations (2.1.1) et (2.1.2). Diaconis et Graham [9], nous proposent trois autres exemples de distances pouvant être utilisées. Alvo et Cabilio [1] nous offrent un apperçu des résultats concernant l'analyse des données de rang basée sur le concept de distance.

Définition 2.1.1 *Une statistique de rang linéaire est une statistique qui peut être écrite de la forme suivante*

$$S = \sum_{i=1}^{N} a(i, \nu(i))$$

où $\{a(i,j)\}$ *est une matrice* $N \times N$ *arbitraire.*

La statistique de Spearman peut être écrite sous cette forme, i.e.

$$\mathcal{A}_s = \sum_{i=1}^{N} \left(i - \frac{N+1}{2}\right)\left(\nu(i) - \frac{N+1}{2}\right)$$

Par contre, la statistique de Kendall ne peut être écrite sous cette forme. donc elle n'est pas une statistique de rang linéaire. Nous trouverons sa projection sur la famille des statistiques de rang linéaire. Ceci nous permettra de démontrer certains résultats comme on le verra.

Théorème 2.1.2 *(Hájek-Šıdák : [10])*

Soıt une statıstıque de rang

$$T = t(\nu(1), \cdots, \nu(N))$$

et posons $\hat{a}(\imath, \jmath) = E\{T|\nu(\imath) = \jmath\}$ *où* $1 \leq \imath, \jmath \leq N$.

Alors la statıstıque

$$\hat{T} = \frac{N-1}{N} \sum_{\imath=1}^{N} \hat{a}(\imath, \nu(\imath)) - (N-2)E(T)$$

est la projectıon de T *sur la famılle des statıstıques de rang lınéaıre.*

On veut obtenir la projection de la statistique de Kendall sur la famille des statıstiques de rang linéaire Soit la statıstıque de Kendall définıe en (2.1 4)

$$\begin{aligned} \mathcal{A}_k &= \sum_{\imath<\jmath} sgn(\mu(\imath) - \mu(\jmath)) sgn(\nu(\imath) - \nu(\jmath)) \\ &= \sum_{\imath<\jmath} sgn(\imath - \jmath) sgn(\nu(\imath) - \nu(\jmath)) \end{aligned}$$

sı on prend. sans perte de généralıté, $\mu = (1, \cdots, N)'$

Pour commencer, calculons

$$E[sgn(\nu(\imath) - \nu(\jmath))|\nu(k) = h] = \begin{cases} 0 & \text{si } k \neq \imath \text{ et } k \neq \jmath \\ \dfrac{2h - N - 1}{N - 1} & \text{si } k = \imath \text{ et } 1 \leq h \leq N \\ \dfrac{N + 1 - 2h}{N - 1} & \text{si } k = \jmath \text{ et } 1 \leq h \leq N \end{cases}$$

Alors

$$\begin{aligned} E[\mathcal{A}_k|\nu(k) = h] &= E\left[\sum_{\imath<\jmath} sgn(\imath - \jmath) sgn(\nu(\imath) - \nu(\jmath))|\nu(k) = h\right] \\ &= \frac{1}{2} E\left[\sum_{\imath\neq\jmath} sgn(\imath - \jmath) sgn(\nu(\imath) - \nu(\jmath))|\nu(k) = h\right] \\ &= \frac{1}{2}\left[\sum_{\jmath} sgn(k - \jmath) E(sgn(h - \nu(\jmath))) + \sum_{\imath} sgn(\imath - k) E(sgn(\nu(\imath) - h))\right] \\ &= \frac{1}{2}\left[\sum_{\jmath} sgn(k - \jmath)\left(\frac{2h - N - 1}{N - 1}\right) + \sum_{\imath} sgn(\imath - k)\left(\frac{N + 1 - 2h}{N - 1}\right)\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\left[\left(\frac{2h-N-1}{N-1}\right)\sum_{j} sgn(k-j) + \left(\frac{N+1-2h}{N-1}\right)\sum_{i} sgn(i-k)\right] \\
&= \frac{1}{2}\left(\frac{2h-N-1}{N-1}\right)\left[\sum_{j=1}^{k-1} 1 + \sum_{j=k+1}^{N}(-1)\right] \\
&+ \frac{1}{2}\left(\frac{N+1-2h}{N-1}\right)\left[\sum_{i=1}^{k-1}(-1) + \sum_{i=k+1}^{N} 1\right] \\
&= \frac{4}{N-1}\left[\left(k-\frac{N+1}{2}\right)\left(h-\frac{N+1}{2}\right)\right] \\
&= \hat{a}(k,h)
\end{aligned}
$$

En appliquant le théorème (2.1.2) on obtient

$$
\begin{aligned}
\hat{\mathcal{A}}_k &= \frac{N-1}{N}\sum_{i=1}^{N}\hat{a}(i,\nu(i)) - (N-2)E[\mathcal{A}_k] \\
&= \frac{4}{N}\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)\left(\nu(i)-\frac{N+1}{2}\right) = \frac{4}{N}\mathcal{A}_s
\end{aligned}
$$

Notons que $E[\mathcal{A}_k] = 0$ sous l'hypothèse que toutes les permutations sont équiprobables.

Définition 2.1.3 *Une statistique de rang linéaire, S, est dite simple si*

$$S = \sum_{i=1}^{N} c_i a(\nu(i)) \tag{2.1.6}$$

où $(c_1, \cdots, c_N)$ et $(a(1), \cdots, a(N))$ sont des vecteurs.

Soient

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a(i),$$

$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i,$$

et

$$\sigma_a^2 = \frac{1}{N-1}\sum_{i=1}^{N}(a(i)-\bar{a})^2$$

Théorème 2.1.4 *[10] Sous H_0 on a pour une statistique de rang linéaire simple que :*

(i)

$$E[S] = \bar{a} \sum_{i=1}^{N} c_i$$

(ii)

$$Var[S] = \sigma_a^2 \sum_{i=1}^{N} (c_i - \bar{c})^2$$

On souhaite calculer

$$A = \frac{Var(\hat{\mathcal{A}}_k)}{Var(\mathcal{A}_k)} \tag{2.1.7}$$

Si cette expression tend vers 1 lorsque $N \to \infty$, alors les deux statistiques $\mathcal{A}_k$ et $\hat{\mathcal{A}}_k$ sont asymptotiquement équivalentes. Puisque $\hat{\mathcal{A}}_k$ est une projection sur la famille des statistiques de rang linéaire, elle peut s'écrire sous la forme (2.1.6) où

$$c_i = \frac{4}{N}\left(i - \frac{N+1}{2}\right) \text{ et } a(\nu(i)) = \left(\nu(i) - \frac{N+1}{2}\right).$$

On peut donc utiliser le théorème (2.1.4) pour calculer sa variance. Nous devons avant tout calculer $\bar{a}$, $\bar{c}$ et σ_a^2 définies à la définition (2.1.3).

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a(i) = \frac{1}{N}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right) = 0$$

$$\bar{c} = \frac{1}{N}\sum_{i=1}^{N} c_i = \frac{1}{N}\sum_{i=1}^{N}\frac{4}{N}\left(i - \frac{N+1}{2}\right) = 0$$

$$\begin{aligned}
\sigma_a^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(a(i) - \bar{a})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(a(i))^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 \\
&= \frac{N(N+1)}{12}
\end{aligned}$$

Donc la variance de $\hat{\mathcal{A}}_k$ est

$$\begin{aligned} Var(\hat{\mathcal{A}}_k) &= \sigma_a^2 \sum_{i=1}^{N} (c_i - \bar{c})^2 \\ &= \sigma_a^2 \sum_{i=1}^{N} c_i^2 \\ &= \frac{N(N+1)}{12} \sum_{i=1}^{N} \left[\frac{4}{N} \left(i - \frac{N+1}{2} \right) \right]^2 \\ &= \frac{N(N+1)}{12} \left(\frac{4}{N} \right)^2 \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 \\ &= \frac{N(N+1)}{12} \left(\frac{4}{N} \right)^2 \frac{N(N+1)(N-1)}{12} \\ &= \frac{1}{9}(N+1)^2(N-1) \end{aligned}$$

On doit maintenant calculer la variance de $\mathcal{A}_k$. Tout d'abord, définissons

$$\Psi(h) = \frac{1}{2}\left[1 + sgn(h)\right] = \begin{cases} 1, & h > 0 \\ 0, & h < 0 \end{cases} \quad \text{où} \quad sgn(h) = \begin{cases} 1, & h > 0 \\ -1, & h < 0 \end{cases}$$

Pour un échantillon $X_1, \cdots, X_N$, on a que le rang de X_i est

$$Rang\ X_i = \sum_{i=1}^{N} \Psi(X_i - X_N) + 1$$

Donc,

$$\begin{aligned} \sum_{i=1}^{N-1} sgn\left[\nu(i) - \nu(N)\right] &= \sum_{i=1}^{N-1} \left[2\Psi(\nu(i) - \nu(N)) - 1\right] \\ &= 2 \sum_{i=1}^{N-1} \Psi(\nu(i) - \nu(N)) - (N-1) \\ &= 2\left[\nu(N) - 1\right] - (N-1) \end{aligned}$$

$$
\begin{aligned}
&= 2\nu(N) - 2 - N + 1 \\
&= 2\nu(N) - (N+1) \\
&= 2\left[\nu(N) - \frac{N+1}{2}\right]
\end{aligned}
$$

Pour le calcul de la variance de $\mathcal{A}_k$, notons que $\mathcal{A}_k$ peut s'écrire comme suit :

$$\mathcal{A}_k = \sum_{i<j} sgn(\nu(i) - \nu(j))$$

Écrivons maintenant $\mathcal{A}_k$ sous forme d'une équation récursive :

$$
\begin{aligned}
\mathcal{A}_k(N) &= \sum_{i<j}^{N} sgn(\nu(i) - \nu(j)) \\
&= \left[\sum_{i<j}^{N-1} sgn(\nu(i) - \nu(j)) + \sum_{i=1}^{N-1} sgn(\nu(i) - \nu(N))\right] \\
&= \sum_{i<j}^{N-1} sgn(\nu(i) - \nu(j)) + \sum_{i=1}^{N-1} sgn(\nu(i) - \nu(N)) \\
&= \mathcal{A}_k(N-1) + 2\left(\nu(N) - \frac{1}{2}(N+1)\right)
\end{aligned}
\tag{2.1.8}
$$

On a alors que les deux termes du côté droit sont indépendants. En utilisant l'équation (2.1.8) prouvons que

$$Var(\mathcal{A}_k(N)) = Var(\mathcal{A}_k(N-1)) + \frac{1}{3}(N^2 - 1) \tag{2.1.9}$$

En prenant la variance des deux côtés de l'équation (2.1.8), on obtient

$$
\begin{aligned}
Var(\mathcal{A}_k(N)) &= Var\left(\mathcal{A}_k(N-1) + 2\left(\nu(N) - \frac{1}{2}(N+1)\right)\right) \\
&= Var(\mathcal{A}_k(N-1)) + 4Var\left(\left(\nu(N) - \frac{1}{2}(N+1)\right)\right) \\
&= Var(\mathcal{A}_k(N-1)) + 4Var(\nu(N)) \\
&= Var(\mathcal{A}_k(N-1)) + \frac{1}{3}(N^2 - 1)
\end{aligned}
\tag{2.1.10}
$$

Soit $u_i = \sigma_i^2 = Var(\mathcal{A}_k(i))$. L'équation (2.1.10) devient

$$u_i = u_{i-1} + \frac{1}{3}(i^2 - 1) \tag{2.1.11}$$

Soit la série télescopique suivante

$$\begin{aligned}\sum_{i=1}^{N}[u_i - u_{i-1}] &= (u_1 - u_0) + (u_2 - u_1) + \cdots + (u_{N-1} - u_{N-2}) + (u_N - u_{N-1}) \\ &= u_N - u_0, \qquad \text{où } u_0 = 0\end{aligned} \tag{2.1.12}$$

En isolant u_N de l'équation (2.1.12) et en utilisant l'équation (2.1.11), on obtient

$$\begin{aligned}u_N &= \sum_{i=1}^{N}[u_i - u_{i-1}] + u_0 \\ &= \frac{1}{3}\sum_{i=1}^{N}(i^2 - 1) \\ &= \frac{1}{3}\left[\sum_{i=1}^{N} i^2 - N\right] \\ &= \frac{1}{3}\left[\frac{N(N+1)(2N+1)}{6} - N\right] \\ &= \frac{1}{18}N(N-1)(2N+5)\end{aligned}$$

Ce qui nous donne

$$\sigma_N^2 = \frac{1}{18}N(N-1)(2N+5) = Var(\mathcal{A}_k(N)) = Var(\mathcal{A}_k)$$

Il est maintenant possible de calculer A définie à l'équation (2.1.7).

$$\begin{aligned}A &= \frac{Var(\hat{\mathcal{A}}_k)}{Var(\mathcal{A}_k)} \\ &= \frac{\frac{1}{9}(N+1)^2(N-1)}{\frac{1}{18}N(N-1)(2N+5)} \\ &= \frac{2(N+1)^2}{N(2N+5)}\end{aligned}$$

Maintenant regardons ce qui se passe lorsque N tend vers l'infini.

$$\lim_{N\to\infty} A = \lim_{N\to\infty} \frac{2(N+1)^2}{N(2N+5)} \longrightarrow 1 \tag{2.1.13}$$

Donc la distribution de $\mathcal{A}_k$ est asymptotiquement celle de $\hat{\mathcal{A}}_k$. Nous savons grâce à Hájek-Šidák ([10]) que premièrement

$$Var(\mathcal{A}_s) = (N-1)\frac{[N(N+1)]^2}{144} \quad \text{(p. 61)}$$

et

$$\frac{12\mathcal{A}_s}{N(N+1)\sqrt{N-1}} \longrightarrow_{dist} N(0,1) \text{ lorsque } N \to \infty \quad \text{(p.160)}$$

Donc par cet énoncé et le résultat obtenu à l'équation (2.1.13) et celui de la page 15. nous avons que les deux mesures sont équivalentes et asymptotiquement normales lorsque N est grand.

Les tests non-paramétriques basés sur les distances de Spearman et de Kendall ci-dessus peuvent être utilisé pour tester l'hypothèse qu'il n'y a pas de tendance croissante dans des données continues. Dans le cas de Spearman, on mesure la corrélation entre les rangs des données et la permutation $(1, 2, \cdots, N)$ selon la fonction $\mathcal{A}_s$ tandis que dans le cas de Kendall on utilise la fonction $\mathcal{A}_k$. Dans le chapitre 3, on utilisera le concept de compatibilité pour modifier les fonctions $\mathcal{A}_s$ et $\mathcal{A}_k$ dans le cas des exactos.

Chapitre 3

L'étude du test basé sur la distance de Spearman avec exactos

3.1 Concept de compatibilité

Définition 3.1.1 *Un classement complet de t objets est compatible avec un classement exacto de ces mêmes objets, si chaque paire d'objets qui reçoit un rang distinct obtient le même rang relatif dans les deux classements.*

Définition 3.1.2 *La classe de compatibilité d'un classement est un ensemble qui contient tout les classements compatibles avec celui-ci.*

Le concept de compatibilité nous permet de généraliser la définition de distance entre les deux permutations dans le cas des exactos ([2]). Par conséquent, la mesure $\mathcal{A}_s$ sera remplacée par l'espérance conditionnelle sur l'espace des classements complets compatibles,

$$\begin{aligned} S &= E[\mathcal{A}_s | C(\mu)C(\nu)] \\ &= \sum_i E\left[\mu(i) - \frac{N+1}{2}\middle| C(\mu)\right] E\left[\nu(i)|C(\nu)\right] \end{aligned}$$

Les espérances conditionnelles à l'intérieur de la sommation représentent des moyennes prises sur les classes de compatibilité.

Voici un exemple pour illustrer ceci. Soient $\mu = (110|10)$, le classement des observations et $\nu = (123|45)$ est le classement du temps. Dans cet exemple on a deux temps, t_1 et t_2, donc $k = 2$. Il y a trois observations au temps t_1 ($N_1 = 3$) et deux observations au temps t_2 ($N_2 = 2$). Pour $\mu = (110|10)$, il y a deux succès ('1') suivi d'un échec ('0') au temps t_1 ($y_1 = 2$, $N_1 - y_1 = 1$), et un succès ('1') suivi d'un échec ('0') au temps t_2, ($y_2 = 1$, $N_2 - y_2 = 1$). On se retrouve donc avec un total de trois succès ($y = 3$) et de deux échecs ($N - y = 2$). On a donc des exactos aux rangs 1, 2 et des exactos aux rangs 3, 4, 5.

Pour le classement du temps, $\nu = (123|45)$, on a que la classe de compatibilité correspondant à ν contient les 12 classements obtenus en permutant les rangs 1, 2, 3 entre eux et les rangs 4, 5 entre eux.

$$C(\nu) = \left\{ \begin{array}{l} (123|45), (132|45), (213|45), (231|45), (312|45), (321|45), \\ (123|54), (132|54), (213|54), (231|54), (312|54), (321|54) \end{array} \right\}.$$

L'espérance conditionnelle $E[\nu(i)|C(\nu)]$ sera donnée par la moyenne des rangs sur la classe $C(\nu)$. Dans cet exemple on obtient

$$E[\nu(i)|C(\nu)] = \left\{ \begin{array}{ll} 2, & \text{pour } i = 1, 2, 3 \\ 4.5, & \text{pour } i = 4, 5 \end{array} \right.$$

Dans notre exemple, il y a un total de 6 configurations de rangs avec (2 succès et 1 échec) dans le bloc 1 et (1 succès et 1 échec) dans le bloc 2 comme suit :

$$(341|52), (351|42), (451|32), (342|51), (352|41), (452|31)$$

En permutant les entrés dans les blocs 1 et 2 respectivement, on obtient un total de 72 rangs compatibles. Donc la classe de compatibilité correspondant à $\mu = (110|10)$ est

$$C(\mu) = \left\{ \begin{array}{l} (341|52), (431|52), (143|52), (314|52), (413|52), (134|52), \\ (341|25), (431|25), (143|25), (314|25)\ \ (413|25), (134|25), \\ (351|42), (531|42), (153|42), (315|42), (513|42), (135|42), \\ (351|24), (531|24), (153|24), (315|24), (513|24), (135|24), \\ (352|41), (325|41), (523|41), (532|41), (235|41), (253|41), \\ (352|14), (325|14), (523|14), (532|14), (235|14), (253|14), \\ (452|31), (425|31), (542|31)\ \ (524|31), (254|31), (245|31), \\ (452|13), (425|13), (542|13), (524|13), (254|13), (245|13), \\ (451|32), (415|32), (514|32), (541|32), (145|32), (154|32), \\ (451|23), (415|23), (514|23), (541|23), (145|23), (154|23), \\ (342|51), (324|51), (423|51), (432|51), (234|51), (243|51), \\ (342|15), (324|15), (423|15), (432|15), (234|15), (243|15) \end{array} \right\}$$

Dans la section 3 2, on calcule ces espérances conditionnelles en général

3.2 Calcul des espérances conditionnelles

Soit un echantillon aléatoire constitue de k variables aléatoires indépendantes suivant une distribution $Binomiale(N_i, p_i)$ On s'intéresse maintenant au cas où il y a des exactos dans les données A cet effet, supposons que les données sont obtenues aux temps $t_1, \quad , t_k$ avec N_i observations au temps t_i En procédant au classement des donnees de temps, on voit qu'il y a N_i rangs qui sont exactos au même rang Par conséquent, on assigne aux N_1 premières observations les rangs $1, 2, \quad , N_1$, en utilisant le concept de compatibilité, on assigne aux N_1 premieres observations un rang égal à la moyenne des rangs $1, 2, \quad , N_1$ Ce qui nous donne le rang

$$g_1 = \frac{N_1 + 1}{2},$$

les N_2 observations suivantes auront le rang moyen

$$g_2 = N_1 + \frac{N_2 + 1}{2},$$

et ainsi de suite jusqu'au N_k dernières observations qui prendront le rang

$$g_k = \sum_{i=1}^{k-1} N_i + \frac{N_k + 1}{2}.$$

Notons que

$$\sum_{i=1}^{k} N_i g_i = \frac{N(N+1)}{2}, \quad \text{où} \quad N = \sum_{i=1}^{k} N_i$$

Soit ν le classement du temps de longueur N

$$\nu = (g_1 \cdots g_1 | \cdots | g_k \cdots g_k)'$$

où chaque g_i est répété N_i fois.

Supposons que y observations ont la valeur 1 et donc que $N-y$ observations prennent la valeur 0. On a alors que $N-y$ observations sont exactos et obtiennent donc le même rang moyen $l_1 = \dfrac{N-y+1}{2}$ tandis que les y autres observations sont exactos au rang $l_2 = N - y + \dfrac{y+1}{2} = N - \dfrac{y}{2} + \dfrac{1}{2}$. Soit μ le classement des observations de longueur N

$$\mu = (\underbrace{l_1 \cdots l_1}_{N_1 - y_1} \underbrace{l_2 \cdots l_2}_{y_1} | \cdots | \underbrace{l_1 \cdots l_1}_{N_k - y_k} \underbrace{l_2 \cdots l_2}_{y_k})'.$$

3.3 Développement de la statistique de Spearman

La distance de Spearman définie en (2.1.1) peut s'écrire de la manière suivante

$$\begin{aligned} d_s(\mu, \nu) &= \frac{1}{2} \sum_{i=1}^{N} (\mu(i) - \nu(i))^2 \\ &= \frac{N(N+1)(2N+1)}{6} - \sum_{i=1}^{N} \mu(i)\nu(i) \end{aligned}$$

En omettant la constante, la somme dans le second terme devient

$$\begin{aligned}\sum_{i=1}^{N}\mu(i)\nu(i) &= (N_1-y_1)l_1g_1 + y_1l_2g_1 + \cdots + (N_k-y_k)l_1g_k + y_kl_2g_k \\ &= \sum_{i=1}^{k}(N_i-y_i)g_il_1 + \sum_{i=1}^{k}y_ig_il_2 \\ &= \frac{N}{2}\left[(N+1)l_1 + \sum_{i=1}^{k}N_ig_i\bar{p}_i\right]\end{aligned}$$

Nous avons vu en (2.1.3) que la distance de Spearman pouvait être écrite sous la forme

$$d_s(\mu,\nu) = c_s - \mathcal{A}_s$$

En omettant la constante c_s on a que la statistique de Spearman $S = \mathcal{A}_s$ devient

$$\begin{aligned}\mathcal{A}_s &= \sum_{i=1}^{N}\left(\mu(i) - \frac{N+1}{2}\right)\left(\nu(i) - \frac{N+1}{2}\right) \\ &= \sum_{i=1}^{N}\left(\mu(i) - \frac{N+1}{2}\right)\nu(i)\end{aligned}$$

$$\begin{aligned}S &= E[\mathcal{A}_s|C(\mu),C(\nu)] \\ &= \sum E\left[\mu(i) - \frac{N+1}{2}\middle|C(\mu)\right]E[\nu(i)|C(\nu)] \\ &= \sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right]\left[l_2\sum_j Y_{ij} + \left(N_i - \sum_j Y_{ij}\right)l_1\right] \\ &= \sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right][(l_2-l_1)y_i + N_il_1] \\ &= \sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right]\left[\frac{N}{2}y_i + N_il_1\right] \\ &= \frac{N}{2}\sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right]y_i + \sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right]N_il_1 \\ &= \frac{N}{2}\sum_{i=1}^{k}\left[g_i - \frac{N+1}{2}\right]y_i\end{aligned}$$

On calcule de la même façon la statistique de Kendall, K, d'après l'équation vu en (2.1.4) et en omettant la constante c_k. Notons qu'au cours d'une période de temps donnée, la différence entre les exactos est zéro et donc il n'y a pas de contribution à la distance. Entre les périodes de temps différentes, nous avons :

$$\mathcal{A}_k = \sum_{i<j}^{N} sgn(\mu(i) - \mu(j))sgn(\nu(i) - \nu(j))$$

$$\begin{aligned} K &= E[\mathcal{A}_k | C(\mu)C(\nu)] \\ &= \sum_{i<j}^{k} [(N_i - y_i)y_j - (N_j - y_j)y_i] \\ &= \sum_{i<j}^{k} [N_i y_j - N_j y_i] \end{aligned}$$

Notons que

$$\begin{aligned} \sum_{i<j}^{k} N_j y_i &= \sum_{i=1}^{k} y_i \sum_{j=i+1}^{k} N_j \\ &= \sum_{i=1}^{k} y_i \left[\sum_{j=1}^{k} N_j - \sum_{j=1}^{i} N_j \right] \\ &= \sum_{i=1}^{k} y_i \left[N - \sum_{j=1}^{i} N_j \right] \\ &= \sum_{i=1}^{k} y_i \left[N - \left(\sum_{j=1}^{i-1} N_j + N_i \right) \right] \\ &= \sum_{i=1}^{k} y_i \left[N - \left(g_i - \frac{N_i + 1}{2} + N_i \right) \right] \\ &= \sum_{i=1}^{k} y_i \left[N - \left(g_i + \frac{N_i - 1}{2} \right) \right] \\ &= \sum_{i=1}^{k} y_i \left[N - g_i - \frac{N_i - 1}{2} \right] \end{aligned}$$

et que

$$\begin{aligned}\sum_{i<j}^{k} N_i y_j &= \sum_{j=2}^{k} y_j \sum_{i=1}^{j-1} N_i \\ &= \sum_{j=2}^{k} y_j \left[g_j - \frac{N_j + 1}{2} \right] \\ &= \sum_{i=2}^{k} y_i \left[g_i - \frac{N_i + 1}{2} \right] \\ &= \sum_{i=1}^{k} y_i \left[g_i - \frac{N_i + 1}{2} \right]\end{aligned}$$

car $g_1 - \dfrac{N_1 + 1}{2} = 0.$

Donc la statistique de Kendall devient

$$\begin{aligned}K &= \sum_{i<j} [N_i y_j - N_j y_i] \\ &= \sum_{i<j} N_i y_j - \sum_{i<j} N_j y_i \\ &= \sum_{i=1}^{k} y_i \left[g_i - \frac{N_i + 1}{2} \right] - \sum_{i=1}^{k} y_i \left[N - g_i - \frac{N_i - 1}{2} \right] \\ &= 2 \sum_{i=1}^{k} \left[g_i - \frac{N + 1}{2} \right] y_i\end{aligned}$$

À présent, si on écrit les statistiques de Spearman et de Kendall côte à côte, on voit clairement la similitude entre les deux

$$S = \frac{N}{2} \sum_{i=1}^{k} \left[g_i - \frac{N + 1}{2} \right] y_i \tag{3.3.1}$$

$$K = 2 \sum_{i=1}^{k} \left[g_i - \frac{N + 1}{2} \right] y_i \tag{3.3.2}$$

3.4 Distribution asymptotique de Spearman

Soient k variables indépendantes, $Y_i \sim Binomiale(N_i, p_i)$, $1 \leq i \leq k$. Supposons que l'on veut tester l'hypothèse nulle où il n'y a pas de tendance, c'est-à-dire, H_0 : $p_i = p$ où $i = 1, \cdots, k$. En premier lieu, considérons la statistique de Spearman (3.3.1) qui est sous la forme d'une combinaison linéaire de binomiales indépendantes, $\sum c_i Y_i$, où $c_i = g_i - \dfrac{N+1}{2}$ et $g_i = \sum_{j=1}^{i-1} N_j + \left(\dfrac{N_j + 1}{2}\right)$. Notons que sous l'hypothèse nulle, $H_0 : p_i = p$, $\sum_{i=1}^{k} c_i N_i p_i$ devient

$$\begin{aligned}
\sum_{i=1}^{k} c_i N_i p &= p \sum_{i=1}^{k} \left(g_i - \frac{N+1}{2} \right) N_i \\
&= p \left[\sum_{i=1}^{k} N_i g_i - \frac{N+1}{2} \sum_{i=1}^{k} N_i \right] \\
&= p \left[\frac{N(N+1)}{2} - \frac{(N+1)N}{2} \right] \\
&= 0
\end{aligned}$$

Dans la plupart des applications, la situation asymptotique d'intérêt se présente lorsque

$$N_i \to \infty, \quad \text{avec } \frac{N_i}{N} \to \lambda_i > 0, \quad i = 1, \cdots, k \tag{3.4.1}$$

Donc, sous (3.4.1), nous avons approximativement

$$Y_i \approx_d N\left(N_i p_i, N_i p_i (1 - p_i)\right)$$

Par l'indépendance des $\{Y_i\}$, on a

$$S = \sum_i c_i Y_i \approx_d N\left(\sum_i c_i N_i p_i, \sum_i c_i^2 N_i p_i (1 - p_i) \right)$$

et donc

$$\frac{\left(\sum c_i Y_i - \sum c_i N_i p_i\right)}{\sqrt{\sum c_i^2 N_i p_i (1 - p_i)}} \approx_d N(0.1)$$

Puisque sous H_0, $\sum_i c_i N_i p_i = 0$, on a donc

$$\frac{\sum_i c_i Y_i}{\sqrt{\sum c_i^2 N_i p_i (1-p_i)}} \approx_d N(0,1)$$

On peut estimer p_i par soit $\bar{p}_i$ ou $\bar{p}$. Dans le premier cas, la statistique correspond exactement à (1.1.3) lorsqu'on choisi $x_i = g_i$. Le test rejette l'hypothèse nulle lorsque

$$\frac{\sum c_i Y_i}{\sqrt{\sum c_i^2 N_i \bar{p}_i (1-\bar{p}_i)}} \geq z_\alpha$$

où z_α est le point au niveau $100(1-\alpha)\%$ d'une distribution normale standard. Une forme alternative du test statistique de Spearman qui sera utilisé plus tard pour une comparaison avec le test de Hamming est

$$\frac{\sum N_i c_i (\bar{p}_i - \bar{p})}{\sqrt{\sum N_i c_i^2 \bar{p}_i (1-\bar{p}_i)}} \tag{3.4.2}$$

De plus, l'expression pour l'estimation de la puissance asymptotique devient

$$\begin{aligned} 1 - \Phi\left(z_\alpha - \frac{\sum c_i N_i \bar{p}_i}{\sqrt{\sum c_i^2 N_i \bar{p}_i (1-\bar{p}_i)}} \right) &= \Phi\left(\frac{\sum c_i N_i \bar{p}_i}{\sqrt{\sum c_i^2 N_i \bar{p}_i (1-\bar{p}_i)}} - z_\alpha \right) \\ &\approx \Phi\left(\frac{\sqrt{N} \sum a_i \lambda_i \bar{p}_i}{\sqrt{\sum a_i^2 \lambda_i \bar{p}_i (1-\bar{p}_i)}} - z_\alpha \right) \end{aligned}$$

où $a_i = \lim \frac{c_i}{N_i}$. La puissance semble converger vers 1 très rapidement lorsque N croit.

3.5 Étude par simulations

3.5.1 L'étude par simulations sous l'hypothèse nulle

Pour procéder à l'étude par simulations, on génère des binomiales (N_i, p_i) On choisi p egale a 0 1 jusqu'à 0 5, des valeurs de $k = 5$, les valeurs 10, 20, 30, 50, 100 pour N_i et un seuil de significance de 5% Sous l'hypothese nulle qu'il n'y a pas de tendance, on a que $p_i = p$

Les simulations ont éte effectuées à l aide du logiciel statistique SAS Après avoir généré les binomiales et calculé la statistique de Spearman pour les différentes valeurs de k, N_i et p, on remarque que l'histogramme des valeurs de la statistique, sous l'hypothese nulle, semble tel que prévu être distribué normalement, comme on peut le voir dans les figures 3 1 a 3 9

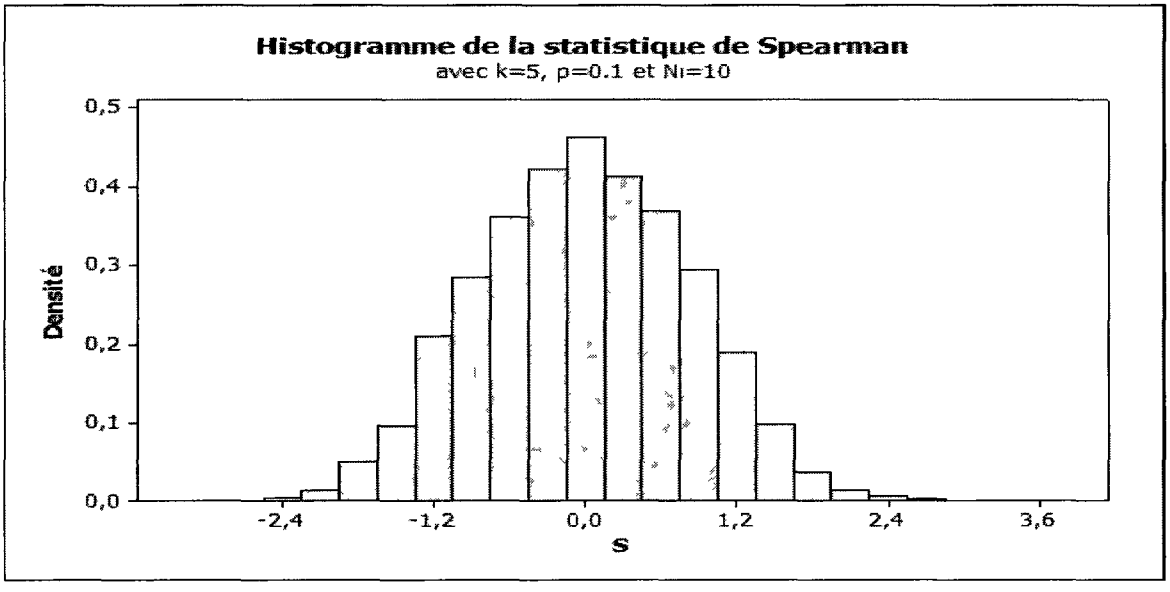

FIG 3 1 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 10$ et $p = 0\,1$

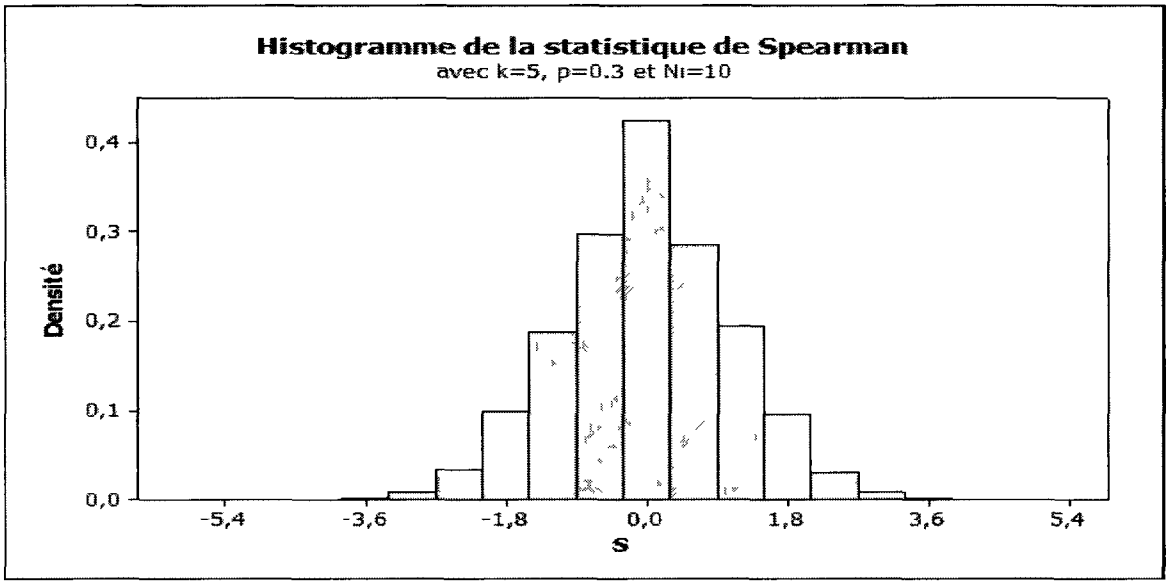

FIG. 3.2 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 10$ et $p = 0.3$.

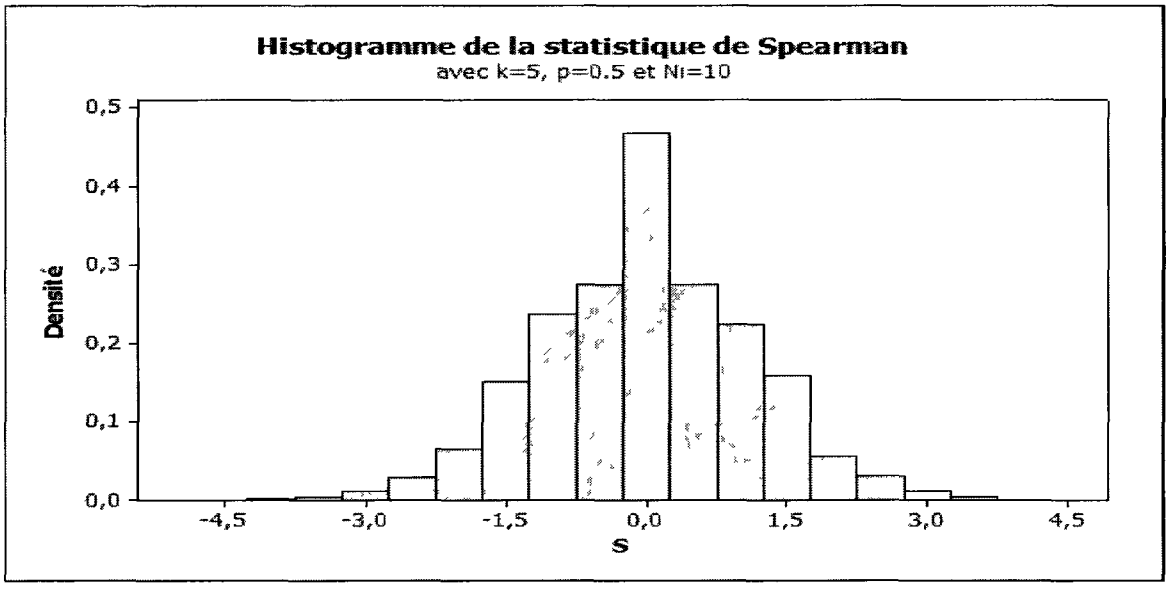

FIG. 3.3 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 10$ et $p = 0.5$.

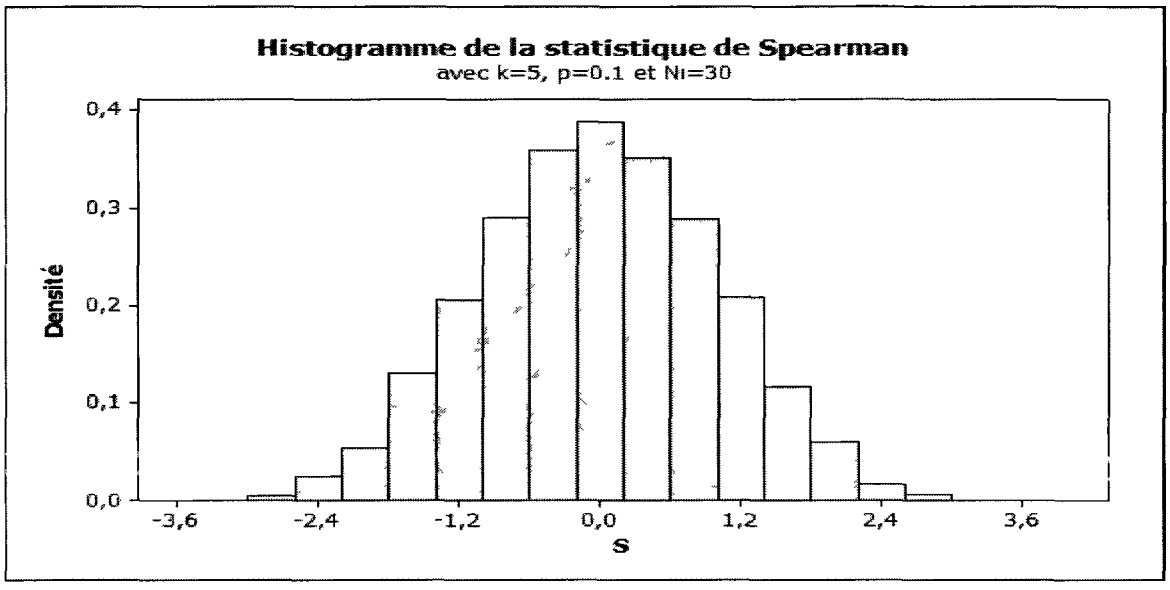

FIG 3 4 Histogramme de la Statistique de Spearman lorsque $k = 5$ $N_i = 30$ et $p = 0\,1$

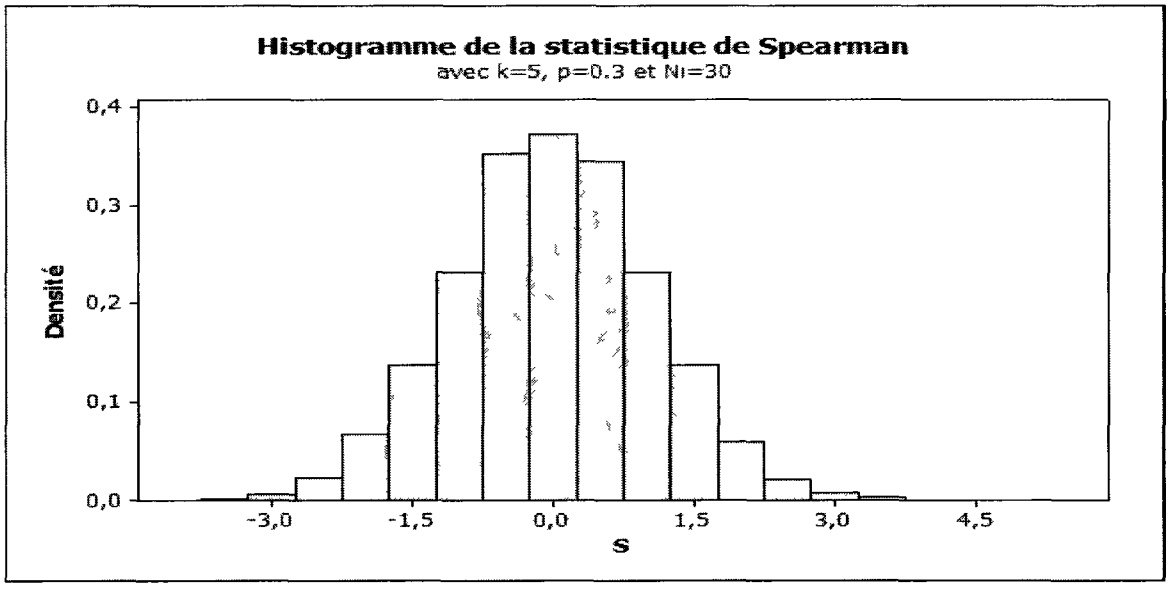

FIG 3 5 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 30$ et $p = 0\,3$

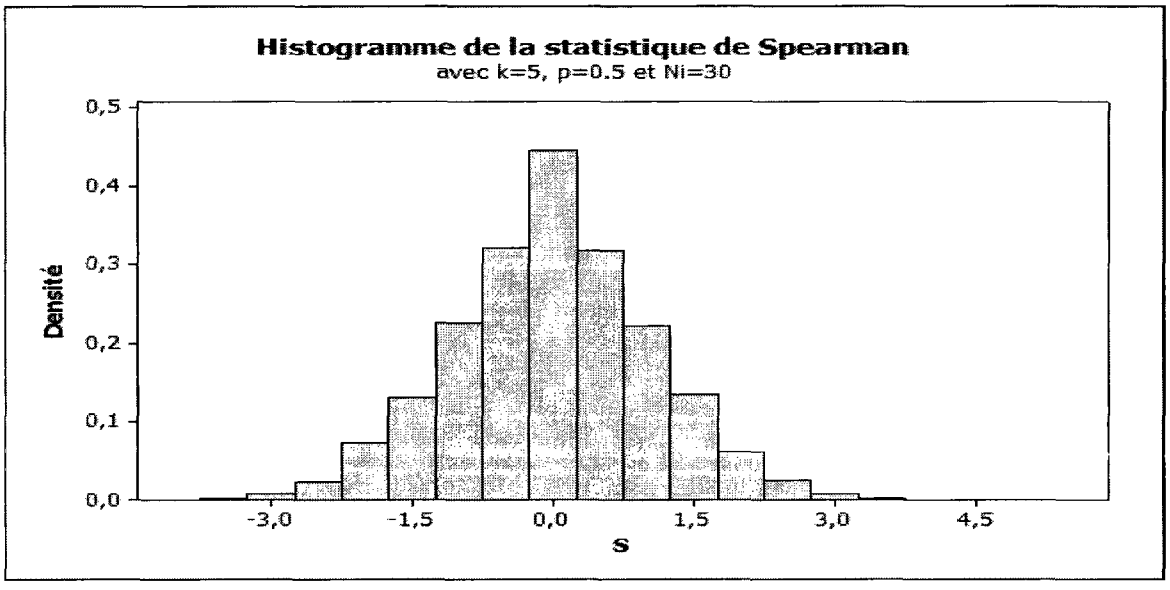

FIG. 3.6 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 30$ et $p = 0.5$.

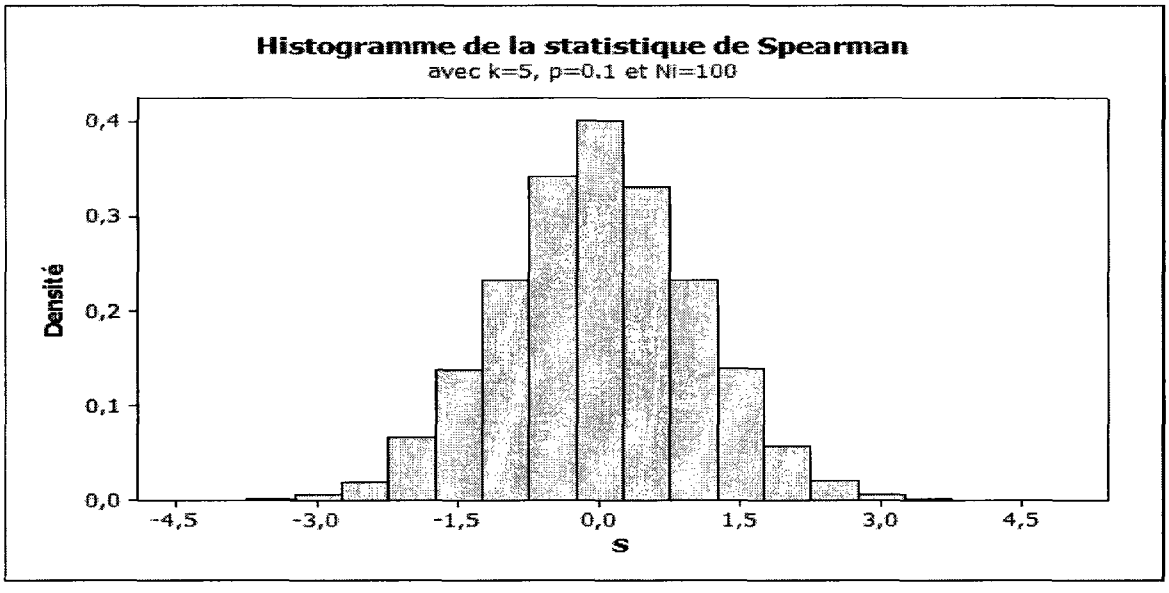

FIG. 3.7 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 100$ et $p = 0.1$.

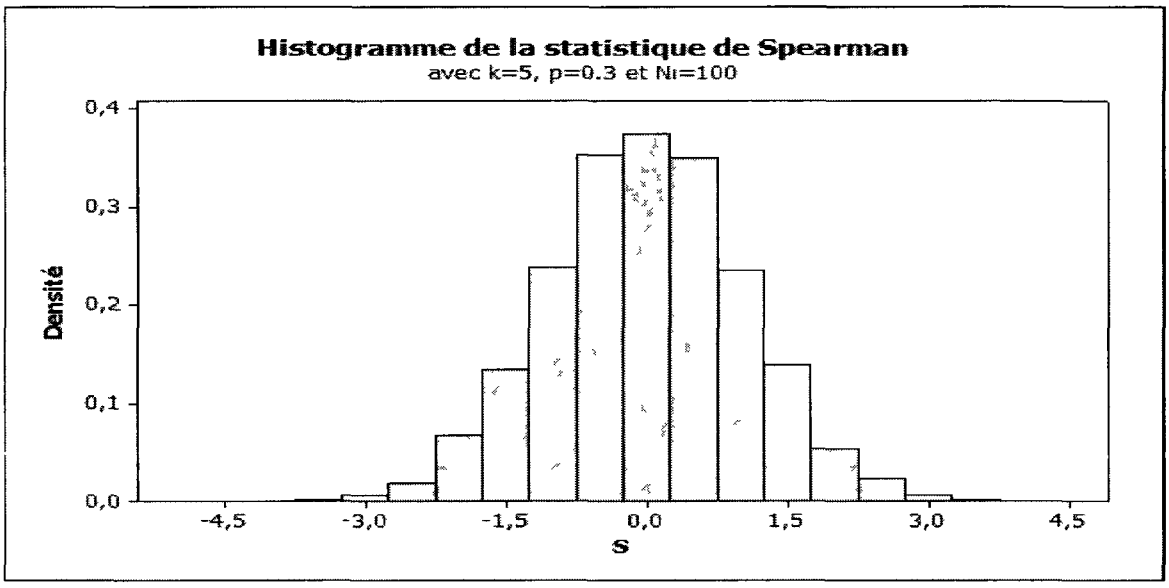

FIG. 3.8 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 100$ et $p = 0.3$.

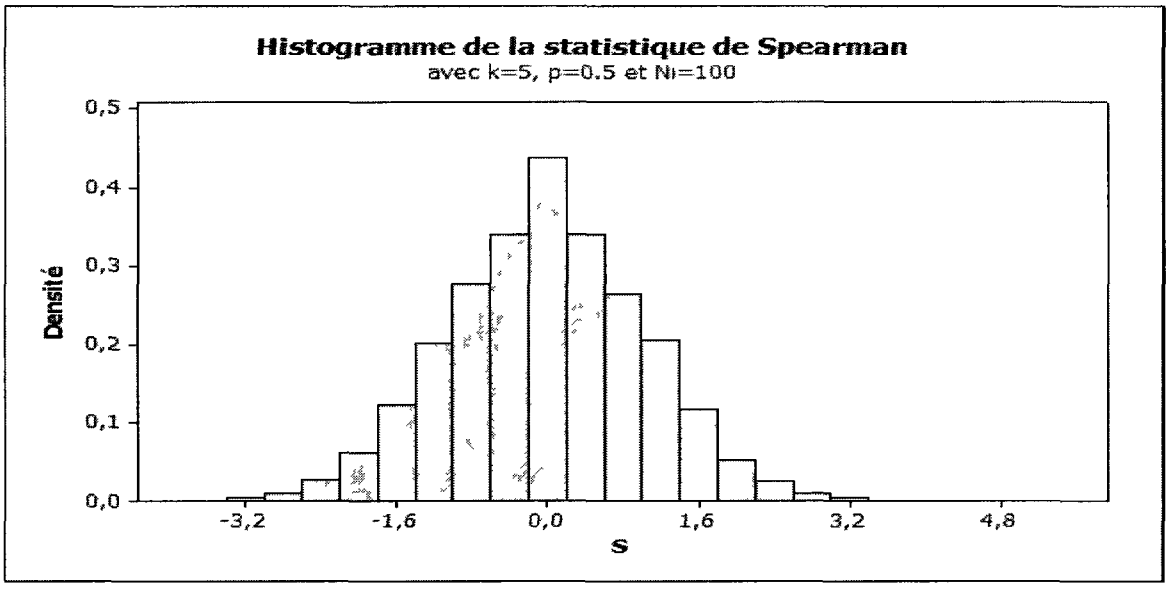

FIG. 3.9 – Histogramme de la Statistique de Spearman lorsque $k = 5$, $N_i = 100$ et $p = 0\,5$.

Ces histogrammes nous donnent confiance que la distribution asymptotique de Spearman est une normale sous l'hypothèse nulle

3.5.2 L'étude de la puissance

La puissance d'un test est la probabilité de rejetter l'hypothèse lorsque la statistique dépasse une certaine valeur, que l'on appelle le point critique. Dans le cas de Spearman, la distribution asymptotique est approximée par une $Normale(0, 1)$, donc à un seuil de significance de 5% on obtient le point critique théorique unilatéral de 1.645.

Nous devons tout d'abord vérifier que sous l'hypothèse nulle le test atteint effectivement le seuil de 5%. Dans le cas de Spearman, on peut approximer p_i par soit $\bar{p}_i$ ou $\bar{p}$ et donc on doit vérifier que la statistique atteigne le seuil. Voici les seuils obtenus lors des simulations lorsque $k = 5$:

$p \diagup N_i$	10	20	30	50	100
0.1	0.0182	0.0347	0.048	0.0544	0.0532
0.2	0.048	0.0564	0.0554	0.0547	0.0536
0.3	0.0616	0.0598	0.0576	0.054	0.0529
0.4	0.067	0.0569	0.0574	0.0553	0.0535
0.5	0.071	0.0549	0.0585	0.0527	0.0559

TAB. 3.1 – Seuils de la statistique de Spearman lorsque p_i est estimé par $\bar{p}_i$

On constate ici que le seuil n'est pas très bon lorsque la taille d'échantillon est petite, mais qu'il se rapproche de 5% lorsque la taille d'échantillon croît. Jetons maintenant un coup d'oeil au seuil lorsqu'on estime p_i par $\bar{p}$:

$p \diagdown N_i$	10	20	30	50	100
0.1	0.0442	0.0477	0.0495	0.0527	0.0512
0.2	0.0513	0.0535	0.0517	0.0503	0.0531
0.3	0.0517	0.0533	0.0523	0.0516	0.0519
0.4	0.0519	0.0492	0.0522	0.0516	0.0526
0.5	0.0477	0.0524	0.0492	0.0513	0.051

TAB. 3.2 – Seuils de la statistique de Spearman lorsque p_i est estimé par $\bar{p}$

Dans ce cas-ci, le seuil est atteint beaucoup plus rapidement et est également beaucoup plus près de de 5% lorsque la taille d'échantillon est grande Puisque la statistique atteint le seuil de 5%, on peut alors procéder aux simulations sous les hypothèses alternatives. Dans ce cas, les simulations sont calculées lorsque p_i est estimé par $\bar{p}$. Dans le chapitre 5, nous comparerons les puissances obtenues en utilisant la statistique de Spearman avec celles obtenues en utilisant la statistique de Hamming.

Chapitre 4

L'étude du test basé sur la distance de Hamming

4.1 Développement de la statistique de Hamming

Dans ce chapitre, nous étudierons un test statistique basé sur la distance de Hamming. La distance de Hamming entre deux classements est le nombre de disparités entre les deux classements. Pour deux classements μ et ν, on peut définir la distance de Hamming comme suit :

$$d_H(\mu, \nu) = N - \sum_{i=1}^{N} \sum_{j=1}^{N} I[\mu(i) = j] I[\nu(i) = j].$$

où $I[\mu(i) = j]$ représente la fonction indicatrice prenant la valeur '1' si le ième objet reçoit le rang j. Nous allons nous concentrer sur la statistique de Hamming définie de la manière suivante lorsqu'il n'y a pas d'exactos.

$$\mathcal{A}_H = \sum_{i=1}^{N} \sum_{j=1}^{N} I[\mu(i) = j] I[\nu(i) = j] \tag{4.1.1}$$

Alvo et Cabilio [2] ont étudiés le cas où il y a des exactos en utilisant le concept de compatibilité. Nous sommes intéressés à trouver une formule pour pouvoir calculer cette statistique plus facilement.

Nous devons tout d'abord définir le concept de compatibilité Pour ce faire regardons un exemple où $\mu = (110|10)$ est le classement des observations et $\nu = (123|45)$ est le classement du temps Dans cet exemple on a deux temps, t_1 et t_2, donc $k = 2$ Il y a trois observations au temps t_1 ($N_1 = 3$) et deux observations au temps t_2 ($N_2 = 2$) Pour $\mu = (110|10)$, il y a deux succès ('1') suivi d'un échec ('0) au temps t_1 ($y_1 = 2$, $N_1 - y_1 = 1$), et un succès ('1') suivi d'un echec ('0') au temps t_2 ($y_2 = 1$, $N_2 - y_2 = 1$) On se retrouve donc avec un total de trois succès ($y = 3$) et de deux échecs ($N - y = 2$) On a donc des exactos aux rangs 1, 2 et des exactos aux rangs 3, 4, 5

Pour le classement du temps, $\nu = (123|45)$, on a que la classe de compatibilité correspondant à ν contient les 12 classements obtenus en permutant les rangs 1, 2, 3 entre eux et les rangs 4, 5 entre eux

$$C(\nu) = \left\{ \begin{array}{l} (123|45), (132|45), (213|45), (231|45), (312|45), (321|45), \\ (123|54), (132|54), (213|54), (231|54), (312|54), (321|54) \end{array} \right\}$$

Les rangs 1, 2, 3 apparaissent chacun deux fois pour les objets 1, 2, 3 et aucune fois pour les objets 4, 5 Les rangs 4, 5 apparaissent six fois pour les objets 4, 5 et aucune fois pour les objets 1, 2, 3 La compatibilité nécessite que l'on prenne la moyenne de la fréquence d'apparition ce qui nous mène à la matrice suivante

i/j	1	2	3	4	5
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
2	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
4	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
5	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$

où les lignes représentent les objets et les colonnes représentent les rangs

En général, pour l'objet l dans le bloc de temps i, on a

$$E[I(\nu(l) = j|C(\nu)] = \left\{ \begin{array}{cc} \frac{1}{N_i}, & \sum_{q=1}^{i-1} N_q + 1 \leq j \leq \sum_{q=1}^{i} N_q \\ 0, & \text{sinon} \end{array} \right\}$$

où $C(\nu)$ est la classe de compatibilité correspondant au rang ν.

On considère maintenant le classement des données selon le modèle binomial. Notons premièrement que le nombre de configurations différentes par lesquelles les rangs correspondant aux succès (et de même aux échecs) sont assignés aux k blocs est donné par

$$\frac{y!}{y_1! \cdots y_k!} \frac{(N-y)!}{(N_1 - y_1)! \cdots (N_k - y_k)!}$$

Le nombre de permutations à l'intérieur de chaque bloc de temps, pour chaque configuration est donné par

$$N_1! \cdots N_k!$$

Donc le nombre total de classements compatibles est donné par le produit

$$\frac{y!(N-y)!}{\Pi y_i!(N_i - y_i)!} \Pi N_i! = N! \frac{\Pi \binom{N_i}{y_i}}{\binom{N}{y}}$$

Supposons maintenant que l'on veut compter le nombre de configurations dans lesquels le i^e succès occupe une certaine position. Ceci est donné par

$$\frac{(y-1)!(N-y)!}{\Pi_{j \neq i} y_j!(y_i - 1)! \Pi (N_i - y_i)} \Pi_{j \neq i} N_j!(N_i - 1)! = N! \frac{\Pi \binom{N_i}{y_i}}{\binom{N}{y}} \frac{y_i}{N_i y}$$

On procède de même pour trouver le i^e échec. Alors, pour l dans le bloc de temps i, $l = 1, \cdots, N_i$

$$\begin{aligned} E[I(\mu(l) = j)|C(\mu)] &= P(\mu(l) = j|C(\mu)) \\ &= \left\{ \begin{array}{cc} \frac{(N_i - y_i)}{N_i(N-y)}, & j \in A_{0Y} \\ \frac{y_i}{N_i y}, & j \in A_{1Y} \end{array} \right. \end{aligned}$$

où $A_{0Y} = \{1, 2, \cdots, N-y\}$ et $A_{1Y} = \{N-y+1, \cdots, N\}$.

Dans notre exemple, il y a un total de 6 configurations de rangs avec (2 succès et 1 échec) dans le bloc 1 et (1 succès et 1 échec) dans le bloc 2 comme suit :

$$(341|52), (351|42), (451|32), (342|51), (352|41), (452|31)$$

En permutant les entrés dans les blocs 1 et 2 respectivement, on obtient un total de 72 rangs compatibles. Donc la classe de compatibilité correspondant à $\mu = (110|10)$ est

$$C(\mu) = \left\{ \begin{array}{l} (341|52), (431|52), (143|52), (314|52), (413|52), (134|52), \\ (341|25), (431|25), (143|25), (314|25), (413|25), (134|25), \\ (351|42), (531|42), (153|42), (315|42), (513|42), (135|42), \\ (351|24), (531|24), (153|24), (315|24), (513|24), (135|24), \\ (352|41), (325|41), (523|41), (532|41), (235|41), (253|41), \\ (352|14), (325|14), (523|14), (532|14), (235|14), (253|14), \\ (452|31), (425|31), (542|31), (524|31), (254|31), (245|31), \\ (452|13), (425|13), (542|13), (524|13), (254|13), (245|13), \\ (451|32), (415|32), (514|32), (541|32), (145|32), (154|32), \\ (451|23), (415|23), (514|23), (541|23), (145|23), (154|23), \\ (342|51), (324|51), (423|51), (432|51), (234|51), (243|51), \\ (342|15), (324|15), (423|15), (432|15), (234|15), (243|15) \end{array} \right\}.$$

En prenant la moyenne des fréquences d'apparitions de chaque rangs, ceci nous mène à la matrice suivante :

i/j	1	2	3	4	5
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$
2	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$
3	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{2}{9}$
4	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
5	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Définissons l'ensemble d'entiers

$$B_i = \left\{ \sum_{q=1}^{i-1} N_q + 1, \cdots, \sum_{q=1}^{i} N_q \right\}, \quad i = 1, \cdots, k$$

et soient

$$w_{0i} = Card(A_{0Y} \cap B_i)$$

$$w_{1i} = Card(A_{1Y} \cap B_i)$$

où $w_{0i} = 0$ si $Y = N$ et $w_{1i} = 0$ si $Y = 0$.

On note que le modèle dans des lignes dans le i^e bloc sont répétés N_i fois. La statistique de Hamming devient alors,

$$\begin{aligned}
H &= E\left[\mathcal{A}_H(\mu,\nu)|C(\mu)C(\nu)\right] \\
&= \sum_{i,j} E\left[I(\mu(i)=j|C(\mu)\right] E\left[I(\nu(i)=j)|C(\nu)\right] \\
&= \sum_{i,j} P[\mu(i)=j|C(\mu)]P[\nu(i)=j|C(\nu)] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N} I_{[\sum_{q=1}^{i-1} N_q+1\le j\le \sum_{q=1}^{i} N_q]}\left[\frac{y_i}{N_i y} I_{[j\in A_{1y}]} + \frac{(N_i-y_i)}{N_i(N-y)} I_{[j\in A_{oy}]}\right] \\
&= \sum_{i=1}^{k}\left[w_{1i}\frac{y_i}{N_i y} + w_{0i}\frac{(N_i-y_i)}{N_i(N-y)}\right] \\
&= \sum_{i=1}^{k}\left[w_{1i}\frac{\bar{p}_i}{y} + w_{0i}\frac{(1-\bar{p}_i)}{(N-y)}\right]
\end{aligned}$$

Puisque $w_{0i} + w_{1i} = N_i$ et $\sum_{i=1}^{k} w_{1i} = y$, il en suit que

$$\begin{aligned}
H &= \frac{1}{N-y}\sum_{i=1}^{k}(1-\bar{p}_i)w_{0i} + \frac{1}{y}\sum_{i=1}^{k}\bar{p}_i w_{1i} \\
&= \frac{1}{N-y}\sum_{i=1}^{k}(1-\bar{p}_i)(N_i - w_{1i}) + \frac{1}{y}\sum_{i=1}^{k}\bar{p}_i w_{1i} \\
&= \frac{1}{N-y}\sum_{i=1}^{k}[N_i(1-\bar{p}_i) - (1-\bar{p}_i)w_{1i}] + \frac{1}{y}\sum_{i=1}^{k}\bar{p}_i w_{1i} \\
&= \sum_{i=1}^{k} w_{1i}\left(\frac{\bar{p}_i}{y} - \frac{1-\bar{p}_i}{N-y}\right) + \sum_{i=1}^{k} N_i\frac{(1-\bar{p}_i)}{(N-y)} \\
&= \sum_{i=1}^{k}\left[w_{1i}\left(\frac{\bar{p}_i}{y} - \frac{1-\bar{p}_i}{N-y}\right) + N_i\frac{(1-\bar{p}_i)}{(N-y)}\right] \\
&= \sum_{i=1}^{k}\left[w_{1i}\left(\frac{N\bar{p}_i - y}{y(N-y)}\right) + N_i\frac{(1-\bar{p}_i)}{(N-y)}\right] \\
&= \sum_{i=1}^{k}\left[w_{1i}\left(\frac{N\bar{p}_i - y}{y(N-y)}\right)\right] + 1 \\
&= \frac{1}{N\bar{p}(1-\bar{p})}\sum_{i=1}^{k}[w_{1i}(\bar{p}_i - \bar{p})] + 1 \\
&= \frac{1}{N\bar{p}(1-\bar{p})}\sum_{i=1}^{k} w_{1i}\bar{p}_i + \frac{1-2\bar{p}}{1-\bar{p}} \\
&= \frac{1}{N\bar{p}(1-\bar{p})}\sum_{i=1}^{k} w_i y_i + \frac{1-2\bar{p}}{1-\bar{p}},
\end{aligned} \tag{4.1.2}$$

où $w_i = \dfrac{w_{1i}}{N_i}$. On voit que la statistique de Hamming en (4.1.2) ressemble au test de chi-carré du manque d'adéquation sauf que les poids sont aléatoires. On peut alors écrire les poids de la façon suivante :

$$w_{1i} = \begin{cases} 0, & N - y + 1 > \sum_{q=1}^{i} N_q \\ \sum_{q=1}^{i} N_q - (N-y), & \sum_{q=1}^{i-1} N_q + 1 \leq N - y + 1 \leq \sum_{q=1}^{i} N_q \\ N_i, & N - y + 1 < \sum_{q=1}^{i-1} N_q + 1. \end{cases}$$

4.2 Distribution asymptotique de Hamming

Théorème 4.2.1 *La moyenne et la variance conditionnelle de* $\sum_{i=1}^{k} w_i Y_i$ *sous l'hypothèse nulle et sous la condition que* $\sum_{i=1}^{k} Y_i = y$ *sont données par* $y\bar{p}$ *et* $N\bar{p}(1-\bar{p})\hat{\sigma}_y^2$, *respectivement, où* $\hat{\sigma}_y^2 = \sum_{i=1}^{k}(w_i - \bar{p})^2\lambda_i$.

Preuve 4.2.2 On remarque que sous H_0,

$$f(y_1, \cdots, y_k) = \prod_{i=1}^{k} \binom{N_i}{y_i} p^{y_i}(1-p)^{N_i - y_i}.$$

En conditionnant sur $\sum Y_i = y$, la distribution conjointe des $\{Y_i\}$ devient

$$\begin{aligned} f_{\sum_i y_i = y}(y_1, \cdots, y_k) &= \frac{\prod_{i=1}^{k} \binom{N_i}{y_i} p^{y_i}(1-p)^{N_i - y_i}}{\binom{N}{y} p^y (1-p)^{N-y}} \\ &= \frac{\prod_{i=1}^{k} \binom{N_i}{y_i}}{\binom{N}{y}} \end{aligned} \tag{4.2.1}$$

qui est une hypergéométrique multivariée.

Il en suit que

$$E(Y_i|H_0) = \frac{N_i}{N} y \tag{4.2.2}$$

$$Var(Y_i|H_0) = \frac{N_i(N-N_i)}{N^2(N-1)} y(N-y) \tag{4.2.3}$$

$$Cov(Y_i, Y_j|H_0) = -\frac{N_i N_j}{N^2(N-1)} y(N-y), \quad i \neq j \tag{4.2.4}$$

Donc,

$$\begin{aligned} E\left(\sum_{i=1}^{k} w_i Y_i \middle| y, H_0\right) &= \left(\sum_{i=1}^{k} w_i \frac{N_i}{N} y\right) \\ &= y\bar{p} \end{aligned} \tag{4.2.5}$$

et

$$
\begin{aligned}
Var\left(\sum_{i=1}^{k} w_i Y_i \middle| y, H_0\right) &= \sum_i w_i^2 Var(Y_i|H_0) + \sum_{i\neq j} Cov(Y_i, Y_j|H_0) \\
&= \frac{1}{N^2(N-1)} y(N-y)\left[\sum_i N_i(N-N_i)w_i^2 - \sum_{i\neq j} N_i N_j w_i w_j\right] \\
&= \frac{1}{N^2(N-1)} y(N-y)\left[N\sum_i N_i w_i^2 - (\sum N_i w_i)^2\right] \\
&= \frac{1}{N^2(N-1)} y(N-y)\left[N^2\sum_i \lambda_i w_i^2 - N^2(\sum \lambda_i w_i)^2\right] \\
&\approx N\bar{p}(1-\bar{p})\sum_{i=1}^{k}(w_i-\bar{p})^2\lambda_i \qquad \blacksquare
\end{aligned}
\tag{4.2.6}
$$

Dans le prochain théorème, on démontre qu'asymptotiquement la statistique de Hamming converge vers une distribution normale.

Théorème 4.2.3 *Sous (3.4.1) et conditionnel sur l'hypothèse nulle,* $H_0 : p_i = p$, $1 \leq i \leq k$, *la statistique* $\sum_{i=1}^{k} w_i Y_i$ *converge vers une distribution normale pour de grande valeur de* N.

Preuve 4.2.4 Soit (3.4.1), il en suit que

$$
\frac{\prod_{i=1}^{k}\binom{N_i}{y_i}}{\binom{N}{y}} \to \binom{y}{y_1 y_2 \cdots y_k}\lambda_1^{y_1}\lambda_2^{y_2}\cdots\lambda_k^{y_k}
$$

Donc on voit que la densité conjointe des $\{Y_i\}$ conditionnelle sur $\sum_i Y_i = y$ tend vers une distribution multinomiale indépendante de p.

On considère maintenant la distribution de la statistique de Hamming. Supposons que les poids $\{w_i\}$ sont les valeurs obtenues à partir d'un échantillon i.i.d. provenant d'une population ayant une moyenne μ. Considérons un échantillonage bootstrap où on pige de l'ensemble des poids $\{w_i\}$ un échantillon aléatoire $W_1^*, \cdots, W_k^*$ avec remise qui suit la distribution

$$P(W^* = w_i) = \lambda_i, \quad i = 1, \cdots, k.$$

La moyenne et la variance sont respectivement

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^{k} w_i \lambda_i \\ &= \sum_{i=1}^{k} \frac{w_{i1}}{N} = \bar{p}\end{aligned}$$

et

$$\hat{\sigma}_y^2 = \sum_{i=1}^{k} (w_i - \bar{p})^2 \lambda_i$$

Donc la somme de l'échantillon bootstrap est

$$S^* = \sum_{i=1}^{y} W_i^*$$

Le théorème central limite affirme que pour y grand mais fixe. S^* est asymptotiquement normale. Maintenant y deviendra grand presque certainement puisque $N \to \infty$, autrement $y/N \not\to p$ presque certainement, ce qui contredit la loi forte des grands nombres.

Nous voulons faire le lien entre la distribution de Hamming et celle du bootstrap. Soit X_i le nombre de temps que w_i est sélectionné dans le rééchantillonage. Il en suit que le vecteur $(X_1, \cdots, X_k)$ a une distribution multinomiale$(y, \lambda_1, \cdots, \lambda_k)$ et S^* peut être écrit comme suit :

$$S^* = \sum_{i=1}^{k} w_i X_i =_d \sum_{i=1}^{k} w_i Y_i.$$

Par conséquent. sous l'hypothèse nulle pour de grands N,

$$\frac{\sum_{i=1}^{k} w_i Y_i - y\bar{p}}{\sqrt{N\bar{p}(1-\bar{p})\hat{\sigma}_y^2}} \to_d N(0,1)$$

Il en suit que sous l'hypothèse nulle et en conditionnant sur y,

$$E[H|y] = E\left[\frac{S^*}{N\bar{p}(1-\bar{p})} + \frac{1-2\bar{p}}{1-\bar{p}}\right] = 1 \tag{4.2.7}$$

et

$$\begin{aligned} Var[H|y] &= \frac{N\bar{p}(1-\bar{p})\hat{\sigma}_y^2}{N^2[\bar{p}(1-\bar{p})]^2} \\ &= \frac{\hat{\sigma}_y^2}{N[\bar{p}(1-\bar{p})]} \end{aligned} \tag{4.2.8}$$

Notons qu'en (4.2.7), $E[H] = 1$. Il en suit que

$$\begin{aligned} \frac{H - E[H|y]}{\sqrt{Var[H|y]}} &= \frac{H-1}{\sqrt{\dfrac{\hat{\sigma}_y^2}{N[\bar{p}(1-\bar{p})]}}} \\ &= \frac{\dfrac{1}{N\bar{p}(1-\bar{p})}\sum_{i=1}^{k}[w_{1i}(\bar{p}_i - \bar{p})]}{\sqrt{\dfrac{\hat{\sigma}_y^2}{N[\bar{p}(1-\bar{p})]}}} \\ &= \frac{\sum_{i=1}^{k}[N_i w_i(\bar{p}_i - \bar{p})]}{\hat{\sigma}_y\sqrt{N[\bar{p}(1-\bar{p})]}} \rightarrow_d N(0,1) \end{aligned} \tag{4.2.9}$$

et alors la statistique en (4.2.9) tend vers une distribution normale standard non conditionnelle sur y. Un test semblable à (3.4.2) qui rejette l'hypothèse nulle sera

$$\frac{\sum_{i=1}^{k} N_i(w_i - \bar{p})(\bar{p}_i - \bar{p})}{\sqrt{\sum_{i=1}^{k} N_i(w_i - \bar{p})^2}\sqrt{\bar{p}(1-\bar{p})}} \geq z_\alpha, \quad y \neq 0, N. \tag{4.2.10}$$

où les poids

$$\sum_{i=1}^{k} N_i(w_i - \bar{p}) = \left(\sum_{i=1}^{k} w_{1i}\right) - n\bar{p} = 0$$

4.3 Étude par simulations

4.3.1 L'étude par simulations sous l'hypothèse nulle

Comme pour la statistique de Spearman, on génère des binomiales (N_i, p_i). On choisi p égale à 0.1 jusqu'à 0.5, des valeurs de $k = 5$, les valeurs 10, 20, 30, 50, 100 pour N_i et un seuil de significance de 5%. Sous l'hypothèse nulle qu'il n'y a pas de tendance, on a que $p_i = p$.

Les simulations ont été effectuées à l'aide du logiciel statistique SAS Après avoir généré les binomiales et calculé la statistique de Hamming pour les différentes valeurs de k, N_i et p, on remarque que l'histogramme des valeurs de la statistique, sous l'hypothèse nulle, semble tel que prévu être distribué normalement, comme on peut le voir dans les figures 4.1 à 4 9.

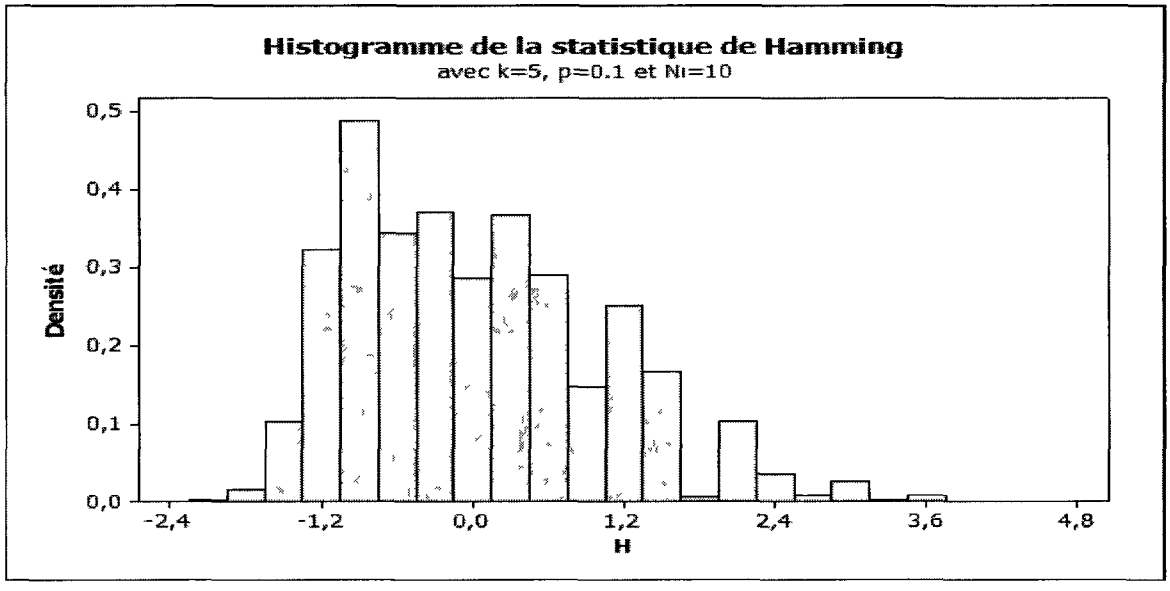

FIG. 4.1 – Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 10$ et $p = 0.1$.

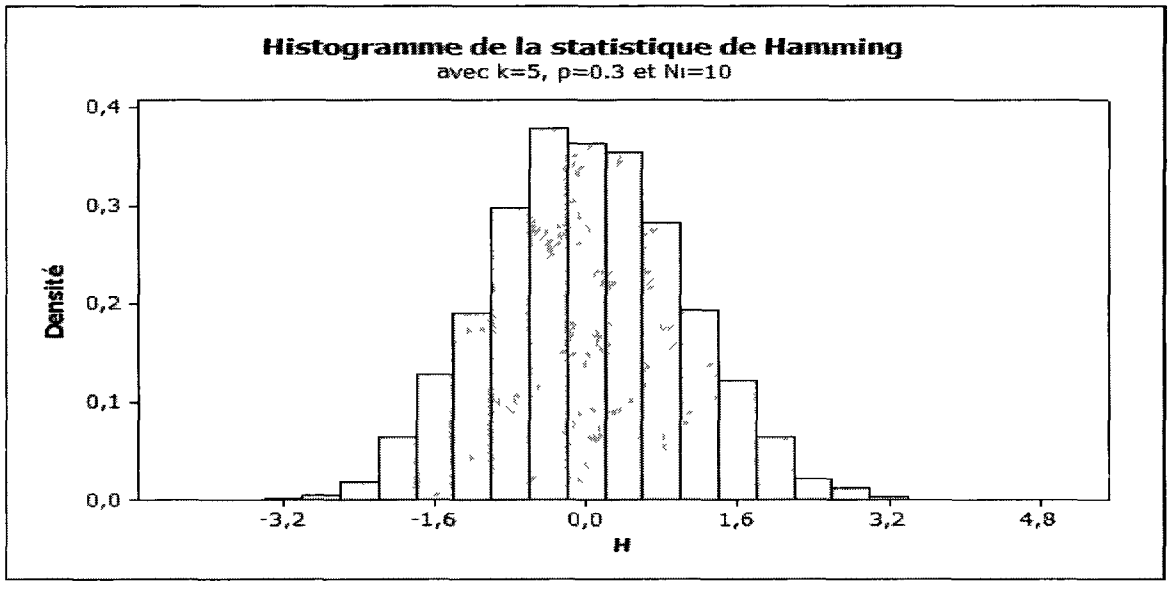

FIG. 4.2 – Histogramme de la Statistique de Hammıng lorsque $k = 5$, $N_i = 10$ et $p = 0.3$.

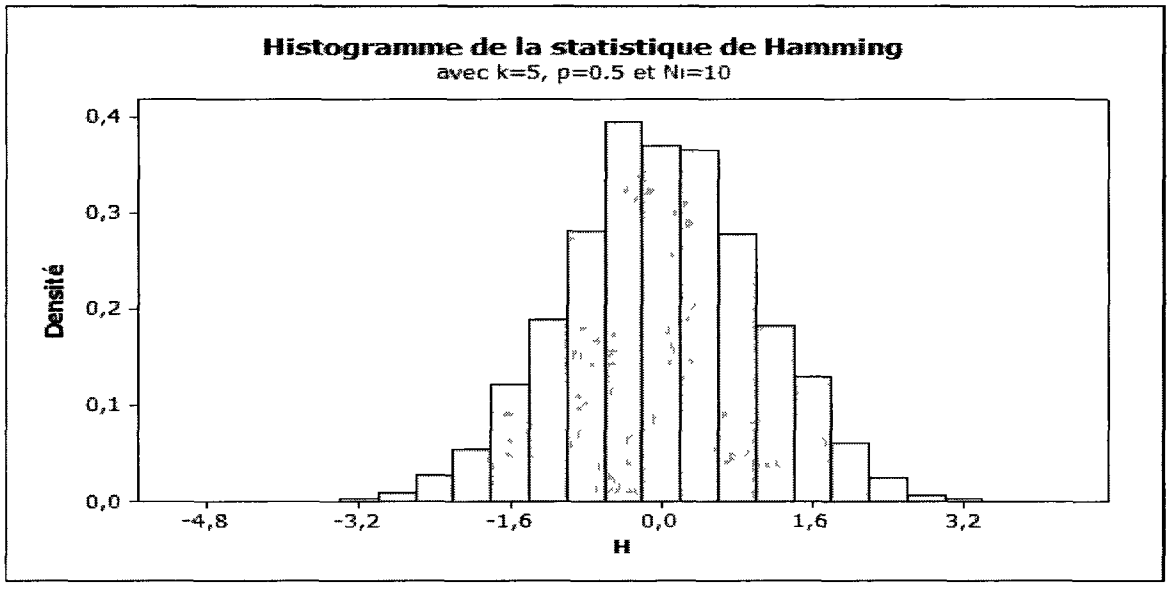

FIG. 4.3 – Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 10$ et $p = 0.5$.

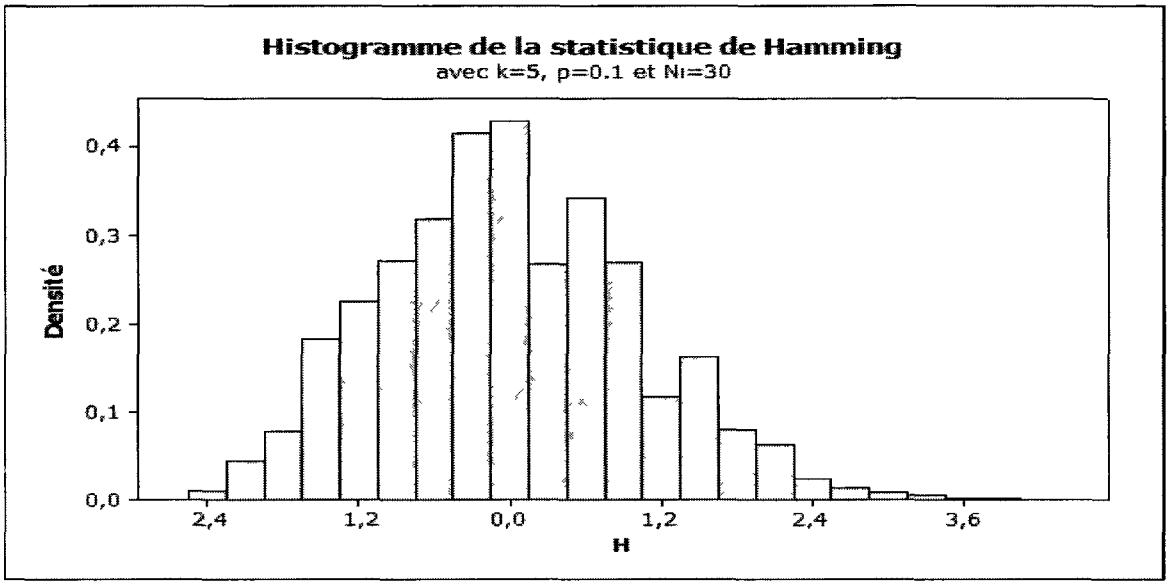

FIG 4 4 Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 30$ et $p = 0\,1$

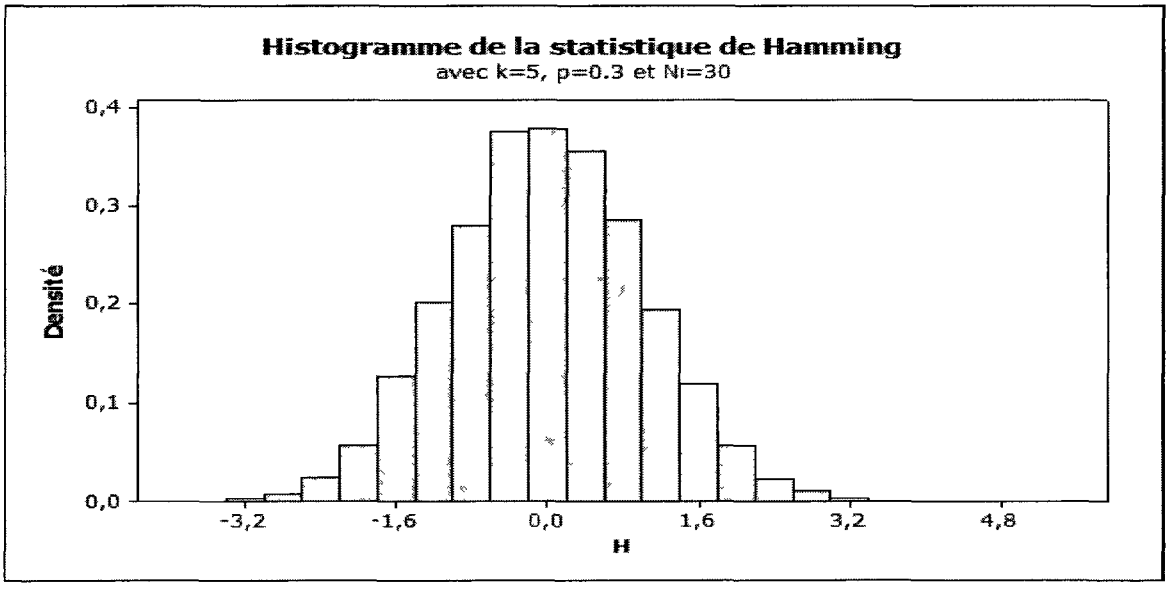

FIG 4 5 Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 30$ et $p = 0\,3$

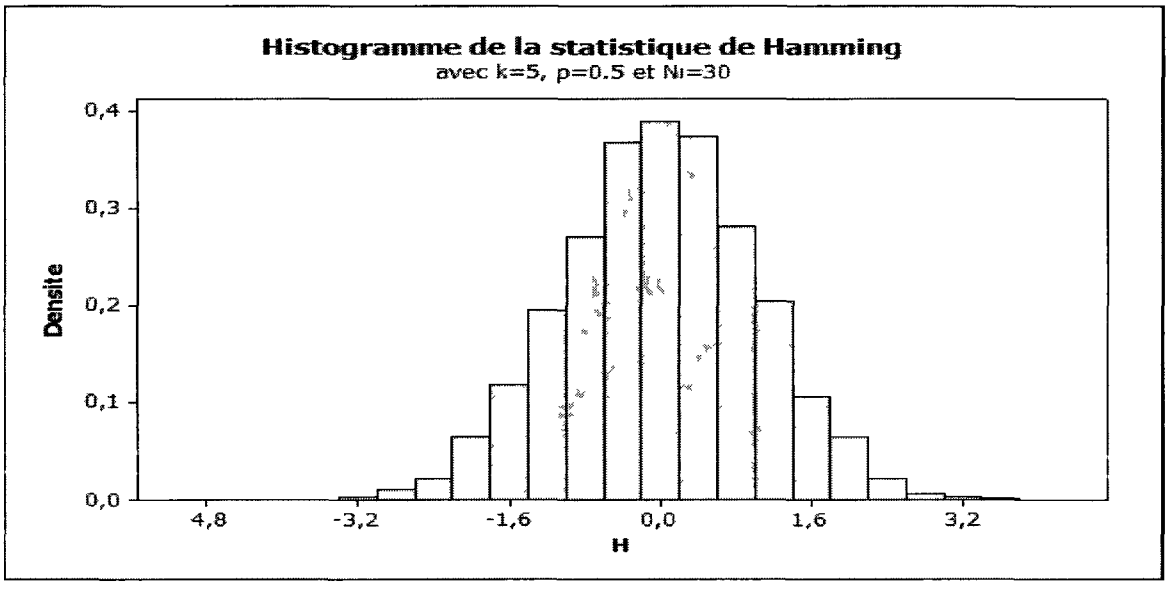

FIG 4 6 Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 30$ et $p = 0\,5$

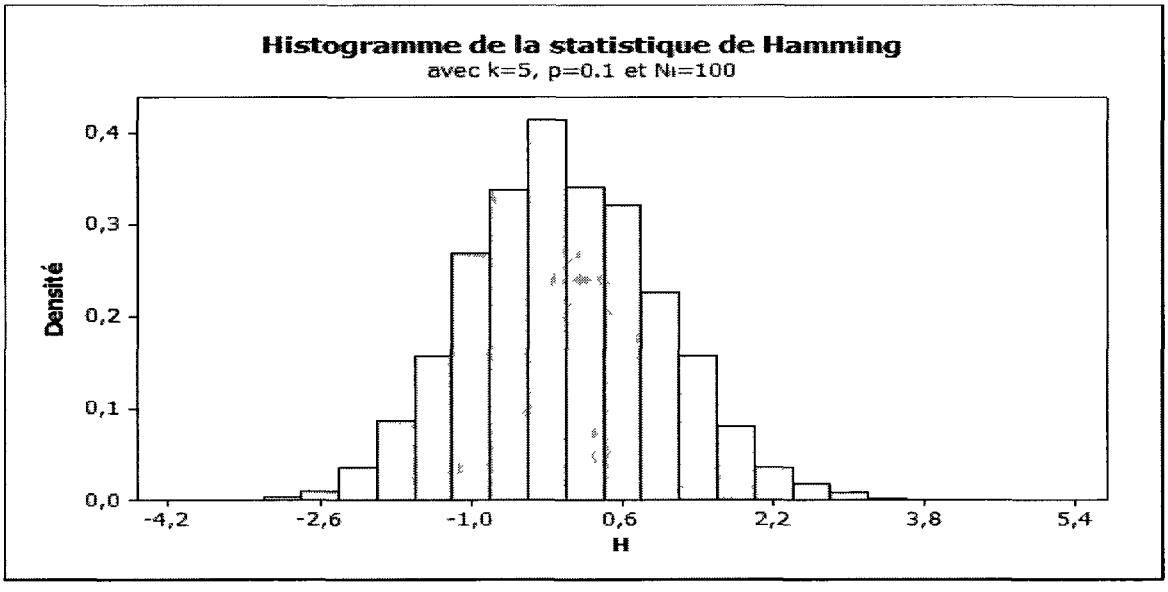

FIG 4 7 – Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 100$ et $p = 0\,1$

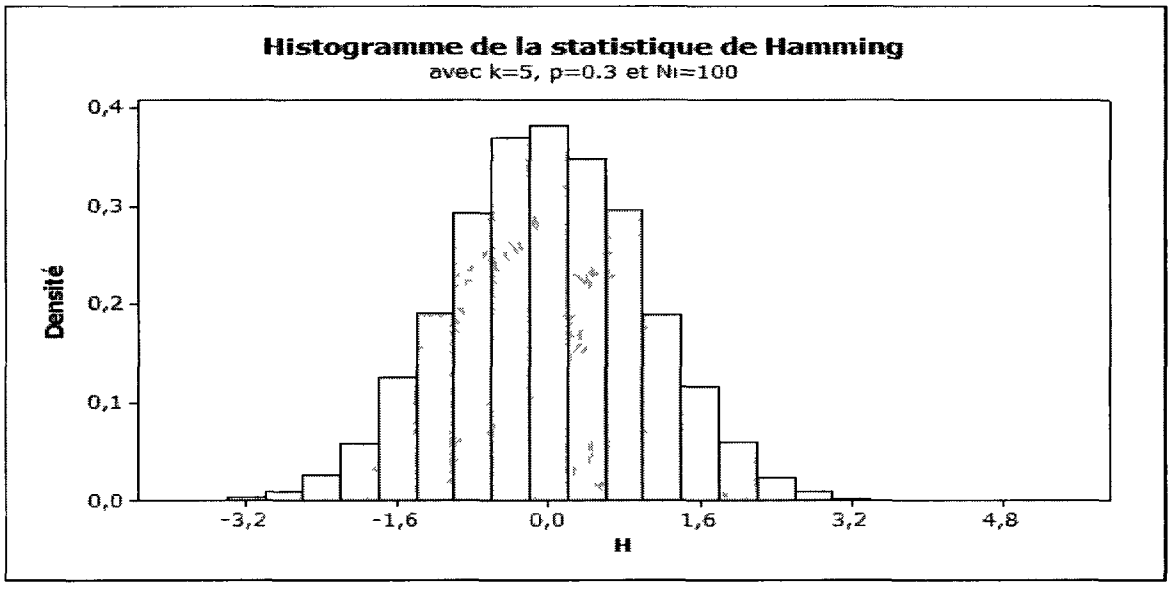

FIG 4 8 – Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 100$ et $p = 0\,3$

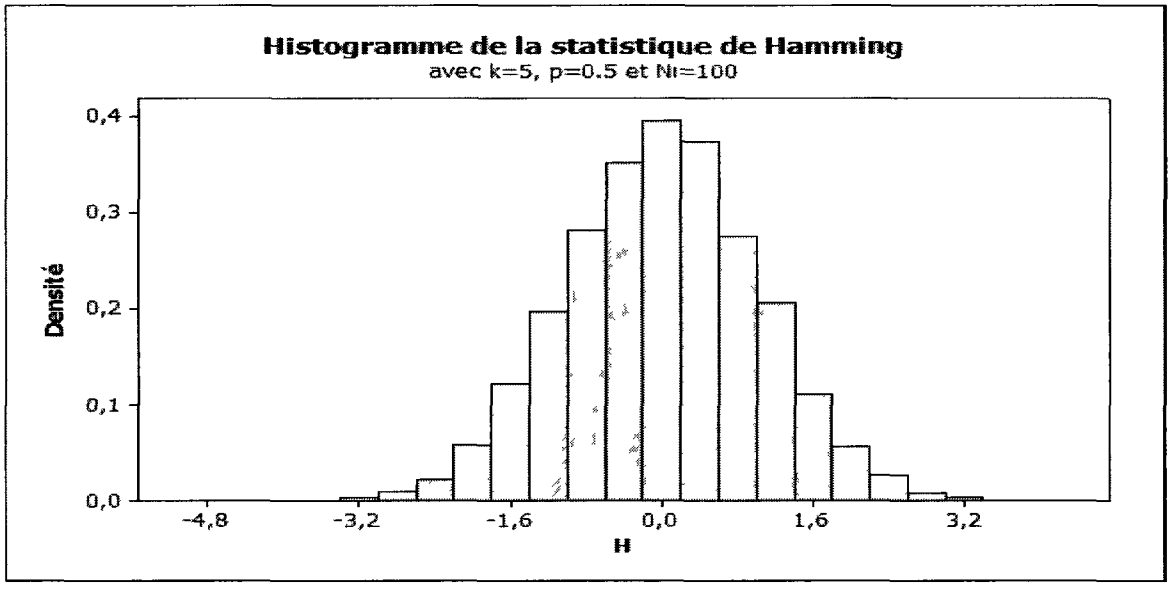

FIG 4 9 Histogramme de la Statistique de Hamming lorsque $k = 5$, $N_i = 100$ et $p = 0\,5$

Ces histogrammes nous donnent confiance que la distribution asymptotique de Hamming est une normale sous l'hypothèse nulle.

4.4 L'étude de la puissance

La puissance d'un test est la probabilité de rejetter l'hypothèse lorsque la statistique dépasse une certaine valeur, que l'on appelle le point critique. Dans le cas de Hamming, la distribution asymptotique est approximée par une $Normale(0, 1)$, donc à un seuil de significance de 5% on obtient le point critique théorique unilatéral de 1.645. Nous devons tout d'abord vérifier que sous l'hypothèse nulle le test atteint effectivement le seuil de 5%. Pour la statistique de Hamming voici les seuls lorsque le test en (4.2.10) est utilisé :

$p \diagup N_i$	10	20	30	50	100
0.1	0.0558	0.0638	0.0608	0.0564	0.0558
0.2	0.0544	0.0561	0.0561	0.0547	0.0534
0.3	0.0540	0.0524	0.0520	0.0519	0.0530
0.4	0.0543	0.0554	0.0516	0.0526	0.0516
0.5	0.0525	0.0526	0.0503	0.0500	0.0499

TAB. 4.1 – Seuils de la statistique de Hamming selon l'équation (4.2.10).

Ici les seuils sont près de 5%. Donc, on peut alors procéder aux simulations sous les hypothèses alternatives. Dans le chapitre 5, nous comparerons les puissances obtenues en utilisant la statistique de Spearman avec celles obtenues en utilisant la statistique de Hamming.

Chapitre 5

Analyse et discussion

5.1 Étude par simulations

Un des objectifs de l'étude par simulations est de comparer les méthodes de Spearman et Hamming selon leur seuil de signification sous l'hypothèse nulle. On veut également comparer leur puissance sous différentes alternatives.

5.1.1 L'étude par simulations sous l'hypothèse nulle

Nous avons constaté dans les chapitres 3 et 4 que les statistiques de Spearman et de Hamming atteignent toutes deux le seuil de 5%. Dans le cas de Spearman, les simulations sont calculées lorsque p_i est estimé par $\bar{p}$, puisque le seuil est plus près de 5% dans ce cas. Pour Hamming, les simulations sont calculées en utilisant le test défini en (4.2.10).

5.1.2 L'étude de la puissance

Lorsqu'on fait des simulations sous l'hypothèse alternative, on compare la puissance des tests de Spearman et de Hamming. Comme nous avons vu au chapitre 3 et au chapitre 4, le point critique théorique unilatéral est de 1.645. On procède maintenant aux simulations sous les hypothèses alternatives. Dans le cas de Spearman, les

simulations sont calculées lorsque p_i est estimé par $\bar{p}$. Pour l'étude des puissances, on considère plusieurs cas :

1 Les p_i sont strictement croissants monotones

2 Les p_i sont non-décroissants avec répétitions

3 Les p_i ne sont pas croissants ou non-décroissants

Pour chacun de ces cas, on a choisi $k = 5$, $N_i = 10, 30$ ou 100. Les résultats sont présentés dans les tableaux (5.1), (5.2) et (5.3). On remarque dans les tableax (5.1) et (5.2) que Spearman a une plus grande puissance que Hamming, cependant lorsqu'on a de grandes tailles d'échantillons la puissance des deux tests est semblable et très élevé. Regardons maintenant les puissance lorsque les p_i ne sont pas en ordre croissant ou non décroissant. Pour les cas présenté dans le tableau (5.3), lorsque les p_i ne sont pas croissants ou non-décroissants, les résultats sont mixtes. Dans certains cas Hamming a une puissance plus élevée que Spearman et vice-versa. Il se peut que la statistique de Hamming puisse détecter d'autre alternatives que la croissance monotone des p_i. Il ne semble pas y avoir de modèle pour prédire lequel de Hamming ou Spearman aura la meilleure puissance.

					$N_i = 10$		$N_i = 30$		$N_i = 100$	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S
0.05	0.2	0.3	0.4	0.5	0.6387	0.818	0.9606	0.9981	1	1
0.05	0.1	0.15	0.2	0.25	0.2693	0.3875	0.5275	0.7989	0.9003	0.9984
0.05	0.1	0.2	0.3	0.4	0.552	0.7346	0.8815	0.9916	0.9994	1
0.1	0.2	0.4	0.5	0.6	0.7668	0.8926	0.9917	0.9996	1	1
0.1	0.2	0.3	0.4	0.5	0.5943	0.7285	0.9422	0.9897	1	1
0.15	0.2	0.3	0.4	0.5	0.5521	0.6352	0.9134	0.9606	1	1
0.2	0.5	0.6	0.7	0.8	0.8305	0.9024	0.9978	0.9998	1	1
0.2	0.3	0.4	0.5	0.6	0.5879	0.6622	0.9344	0.975	0.9999	1
0.2	0.25	0.3	0.35	0.4	0.2512	0.2886	0.4965	0.601	0.9028	0.9623
0.3	0.6	0.7	0.8	0.9	0.8775	0.9398	0.9996	1	1	1
0.3	0.4	0.5	0.6	0.7	0.5768	0.6328	0.9458	0.9669	0.9999	1
0.3	0.4	0.6	0.7	0.8	0.8141	0.8493	0.997	0.9988	1	1
0.35	0.5	0.7	0.8	0.9	0.8741	0.9296	0.9992	0.9999	1	1
0.35	0.7	0.75	0.8	0.9	0.8295	0.8665	0.9988	0.9992	1	1
0.4	0.5	0.6	0.7	0.8	0.5789	0.6631	0.932	0.9742	0.9998	1
0.4	0.55	0.6	0.65	0.7	0.3472	0.3985	0.7118	0.7874	0.992	0.9984
0.45	0.6	0.7	0.75	0.8	0.5295	0.5704	0.9108	0.9331	1	1
0.5	0.6	0.7	0.8	0.9	0.583	0.7247	0.9395	0.9907	1	1
0.55	0.6	0.7	0.8	0.9	0.4949	0.6562	0.8716	0.9741	0.9994	1
0.6	0.65	0.8	0.85	0.9	0.482	0.6015	0.8276	0.9538	0.9972	1

TAB. 5.1 – Puissances de Hamming et Spearman lorsque les p_i sont strictement croissants monotones

					$N_i = 10$		$N_i = 30$		$N_i = 100$	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S
0.05	0.1	0.1	0.2	0.3	0.4529	0.5084	0.7889	0.9009	0.9949	1
0.1	0.3	0.3	0.4	0.5	0.5054	0.6373	0.8796	0.9653	0.9999	1
0.1	0.2	0.4	0.4	0.6	0.6852	0.8506	0.9769	0.9982	1	1
0.15	0.2	0.2	0.3	0.4	0.3949	0.4002	0.7439	0.7781	0.9928	0.9966
0.2	0.3	0.3	0.4	0.5	0.4084	0.4444	0.7773	0.8259	0.9969	0.9989
0.2	0.2	0.3	0.3	0.5	0.4427	0.4625	0.823	0.8429	0.9985	0.9992
0.2	0.2	0.3	0.4	0.5	0.502	0.5417	0.8783	0.9087	0.9999	1
0.2	0.3	0.3	0.5	0.6	0.658	0.6738	0.9723	0.9761	1	1
0.3	0.5	0.5	0.5	0.9	0.5778	0.8096	0.9431	0.9975	1	1
0.3	0.4	0.5	0.5	0.7	0.45	0.5502	0.8551	0.9341	0.9996	1
0.35	0.5	0.6	0.6	0.8	0.5329	0.6517	0.9161	0.9731	0.9998	1
0.35	0.6	0.6	0.6	0.7	0.3026	0.3966	0.628	0.7854	0.977	0.9985
0.4	0.5	0.6	0.6	0.8	0.4625	0.5692	0.8549	0.9415	0.9994	0.9999
0.4	0.5	0.6	0.8	0.8	0.6231	0.7498	0.9457	0.9918	0.9999	1
0.45	0.5	0.5	0.7	0.75	0.3965	0.4825	0.7556	0.8822	0.9929	0.9998
0.5	0.6	0.7	0.8	0.8	0.4894	0.5359	0.8777	0.9101	0.9999	1
0.55	0.6	0.6	0.7	0.9	0.3315	0.5404	0.6524	0.9253	0.9807	0.9999
0.6	0.7	0.8	0.8	0.9	0.4371	0.5026	0.7986	0.8957	0.9961	0.9999
0.6	0.7	0.7	0.8	0.9	0.3579	0.486	0.6918	0.8884	0.9958	0.9999
0.6	0.6	0.7	0.8	0.9	0.3962	0.5679	0.7387	0.9425	0.992	1

TAB. 5.2 – Puissances de Hamming et Spearman lorsque les p_i sont non-décroissants

					$N_i = 10$		$N_i = 30$		$N_i = 100$	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S
0.05	0.4	0.2	0.5	0.9	0.979	0.9967	1	1	1	1
0.05	0.3	0.2	0.1	0.45	0.5441	0.4179	0.917	0.8248	1	0.9994
0.1	0.7	0.9	0.2	0.5	0.0284	0.095	0.1032	0.2180	0.0021	0.6245
0.1	0.9	0.2	0.6	0.5	0.0884	0.1986	0.2328	0.5234	0.667	0.9696
0.2	0.5	0.4	0.7	0.3	0.2022	0.1700	0.4688	0.3935	0.9180	0.8408
0.2	0.7	0.05	0.5	0.3	0.0935	0.0427	0.1727	0.0431	0.3784	0.0425
0.2	0.8	0.1	0.3	0.5	0.0596	0.0603	0.0949	0.0855	0.2062	0.1387
0.2	0.5	0.3	0.8	0.4	0.5139	0.3904	0.9283	0.7986	1	0.9989
0.3	0.9	0.7	0.5	0.6	0.0863	0.0944	0.0796	0.1609	0.0777	0.3489
0.3	0.5	0.2	0.6	0.4	0.2473	0.1467	0.5615	0.2764	0.9669	0.6098
0.35	0.9	0.5	0.8	0.7	0.2310	0.3500	0.3903	0.7123	0.7443	0.9937
0.4	0.6	0.3	0.7	0.5	0.1578	0.1352	0.294	0.2622	0.6153	0.5914
0.5	0.4	0.1	0.8	0.3	0.2856	0.0443	0.0921	0.0436	0.9929	0.0438
0.55	0.9	0.6	0.8	0.7	0.1689	0.1226	0.2861	0.1984	0.5601	0.4121
0.6	0.2	0.4	0.8	0.5	0.4596	0.1837	0.8771	0.3925	0.9997	0.8266
0.6	0.5	0.1	0.9	0.4	0.1925	0.0389	0.3233	0.0408	0.6097	0.0428
0.6	0.1	0.8	0.4	0.5	0.1524	0.0648	0.2248	0.0872	0.4121	0.1551
0.6	0.9	0.55	0.7	0.6	0.0412	0.0228	0.025	0.0106	0.0074	0.0018
0.7	0.5	0.4	0.8	0.6	0.0716	0.0683	0.0664	0.0899	0.0668	0.1496
0.9	0.05	0.3	0.5	0.7	0.1337	0.0171	0.4034	0.0257	0.9301	0.0435

TAB. 5.3 – Puissances de Hamming et Spearman lorsque les p_i ne sont pas croissants ou non-décroissants

Sachant que pour le test de Cochran-Armitage, le seuil varie selon les valeurs de N_i (Corcoran, [8]), nous avons également fait l'étude par simulations avec des valeurs de N_i différentes. Nous avons considérés quatre cas pour $k = 5$:

1. $N_1 = 10$, $N_2 = 10$, $N_3 = 5$, $N_4 = 15$, $N_5 = 10$
2. $N_1 = 5$, $N_2 = 8$, $N_3 = 5$, $N_4 = 16$, $N_5 = 16$
3. $N_1 = 16$, $N_2 = 16$, $N_3 = 5$, $N_4 = 8$, $N_5 = 5$
4. $N_1 = 3$, $N_2 = 8$, $N_3 = 13$, $N_4 = 18$, $N_5 = 8$

p=0.1	N_1	N_2	N_3	N_4	N_5	H	S
	10	10	5	15	10	0.0552	0.0633
	5	8	5	16	16	0.055	0.053
	16	16	5	8	5	0.0771	0.069
	3	8	13	18	8	0.0634	0.0741

TAB. 5.4 – Seuils des statistiques de Hamming et de Spearman avec p=0.1.

p=0.2	N_1	N_2	N_3	N_4	N_5	H	S
	10	10	5	15	10	0.0551	0.068
	5	8	5	16	16	0.0582	0.071
	16	16	5	8	5	0.0648	0.0703
	3	8	13	18	8	0.0555	0.0696

TAB. 5.5 – Seuils des statistiques de Hamming et de Spearman avec p=0.2.

p=0.3	N_1	N_2	N_3	N_4	N_5	H	S
	10	10	5	15	10	0.053	0.0662
	5	8	5	16	16	0.0549	0.0702
	16	16	5	8	5	0.0597	0.0678
	3	8	13	18	8	0.0519	0.0705

TAB. 5.6 – Seuils des statistiques de Hamming et de Spearman avec p=0.3.

p=0.4	N_1	N_2	N_3	N_4	N_5	H	S
	10	10	5	15	10	0.053	0.0661
	5	8	5	16	16	0.052	0.068
	16	16	5	8	5	0.0589	0.0701
	3	8	13	18	8	0.0565	0.0697

TAB. 5.7 – Seuils des statistiques de Hamming et de Spearman avec p=0.4.

p=0.4	N_1	N_2	N_3	N_4	N_5	H	S
	10	10	5	15	10	0.0555	0.066
	5	8	5	16	16	0.0553	0.0672
	16	16	5	8	5	0.0586	0.0677
	3	8	13	18	8	0.0574	0.071

TAB. 5.8 – Seuils des statistiques de Hamming et de Spearman avec p=0.5.

Il est intéressant de constater dans ces cas que la statistique de Hamming est beaucoup plus près du seuil de 5% que la statistique de Spearman.

Pour ce qui est des hypothèses alternatives, nous avons testés les mêmes valeurs de p_i que celles utilisées dans les tableaux (5.1), (5.2) et (5.3). Voici quelques-uns des résultats. On constate que même avec des tailles différentes, la conclusion initiale ne change pas. C'est-à-dire que lorsque les p_i sont strictement croissant ou non

décroissant, la statistique de Spearman a une meilleure puissance que celle de Hamming. Par contre dans le cas où les p_i ne sont pas croissants ou non-décroissants, il est difficile de prédire lequel des tests aura la meilleure puissance.

					Cas 1		Cas 2		Cas 3		Cas 4	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S	H	S
0.05	0.2	0.3	0.4	0.5	0.6450	0.8799	0.5165	0.7656	0.6976	0.8893	0.4569	0.6205
0.05	0.1	0.15	0.2	0.25	0.2563	0.4798	0.2472	0.4042	0.3413	0.4607	0.1966	0.3379
0.05	0.1	0.2	0.3	0.4	0.5389	0.8055	0.4682	0.7415	0.6222	0.7402	0.4162	0.6281
0.05	0.3	0.5	0.7	0.9	0.9974	0.9999	0.9909	0.9988	0.9931	0.9997	0.9317	0.9830
0.05	0.15	0.3	0.6	0.9	0.9989	0.9999	0.9986	0.9999	0.9962	0.9989	0.9846	0.9975
0.35	0.5	0.7	0.8	0.9	0.8982	0.9472	0.8605	0.9059	0.8248	0.9227	0.7169	0.8090
0.45	0.6	0.7	0.75	0.8	0.5534	0.6137	0.4694	0.4983	0.5228	0.6264	0.3448	0.4079
0.4	0.55	0.6	0.65	0.7	0.3748	0.4636	0.2965	0.3716	0.4181	0.4874	0.2293	0.2941
0.35	0.7	0.75	0.8	0.9	0.8475	0.8926	0.6984	0.7772	0.8949	0.9236	0.4329	0.5795
0.55	0.6	0.7	0.8	0.9	0.5342	0.7285	0.5864	0.7177	0.3582	0.6004	0.4412	0.6087

TAB. 5.9 – Puissances de Hamming et Spearman lorsque les p_i sont strictement croissants monotones

					Cas 1		Cas 2		Cas 3		Cas 4	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S	H	S
0.1	0.3	0.3	0.4	0.5	0.5083	0.7091	0.4143	0.5812	0.5155	0.7385	0.3543	0.4431
0.05	0.1	0.4	0.4	0.6	0.8590	0.9706	0.7934	0.9430	0.8549	0.9472	0.5521	0.8068
0.05	0.1	0.1	0.2	0.3	0.3933	0.5801	0.4035	0.5618	0.4483	0.4836	0.3436	0.4709
0.05	0.2	0.2	0.5	0.7	0.9605	0.9904	0.9347	0.9810	0.9376	0.9715	0.9129	0.9416
0.05	0.05	0.2	0.6	0.9	0.9998	1	0.9999	0.9999	0.9994	0.9988	0.9988	0.9997
0.15	0.2	0.2	0.3	0.4	0.3723	0.4540	0.3477	0.4526	0.3722	0.3719	0.3197	0.3820
0.2	0.3	0.3	0.4	0.5	0.4056	0.4983	0.3659	0.4629	0.3646	0.4521	0.3242	0.3793
0.2	0.2	0.3	0.3	0.5	0.4837	0.4978	0.5297	0.5738	0.3779	0.3654	0.3484	0.3968
0.2	0.3	0.5	0.5	0.7	0.7086	0.8464	0.7030	0.8086	0.7220	0.7965	0.3801	0.6124
0.5	0.6	0.7	0.8	0.8	0.5344	0.5825	0.4966	0.4580	0.4267	0.5748	0.3992	0.4397

TAB. 5.10 – Puissances de Hamming et Spearman lorsque les p_i sont non-décroissants

					Cas 1		Cas 2		Cas 3		Cas 4	
p_1	p_2	p_3	p_4	p_5	H	S	H	S	H	S	H	S
0.4	0.9	0.5	0.6	0.7	0.0865	0.2039	0.0271	0.1354	0.6180	0.3610	0.0140	0.0728
0.35	0.9	0.5	0.8	0.7	0.4245	0.4415	0.1182	0.1843	0.8741	0.6830	0.0592	0.2312
0.5	0.8	0.4	0.6	0.9	0.1228	0.4663	0.1623	0.6623	0.3415	0.3183	0.1189	0.3781
0.35	0.5	0.4	0.9	0.7	0.8192	0.8069	0.7746	0.5750	0.6827	0.8037	0.8138	0.8014
0.7	0.5	0.4	0.8	0.6	0.1207	0.1112	0.2526	0.1130	0.0120	0.0480	0.4873	0.3594
0.1	0.45	0.2	0.3	0.15	0.0067	0.0468	0.0023	0.0062	0.0275	0.2671	0.0080	0.0192
0.55	0.9	0.6	0.8	0.7	0.2567	0.1601	0.0561	0.0724	0.5841	0.3047	0.0252	0.1090
0.6	0.9	0.55	0.7	0.6	0.0508	0.0336	0.0073	0.0162	0.2914	0.0904	0.0071	0.0265
0.7	0.55	0.6	0.9	0.8	0.2461	0.4359	0.4468	0.3992	0.0530	0.2521	0.4833	0.6119
0.05	0.3	0.2	0.1	0.45	0.6157	0.4804	0.7320	0.5406	0.2422	0.5543	0.4617	0.1789

TAB. 5.11 – Puissances de Hamming et Spearman lorsque les p_i ne sont pas croissants ou non-décroissants

5.2 La statistique de Hamming modifiée

Nous avons vu que les tests de tendance prennent tous la forme

$$\frac{\sum N_i(x_i - \bar{x})(\bar{p}_i - \bar{p})}{\sqrt{\sum N_i(x_i - \bar{x})^2}\sqrt{\bar{p}(1-\bar{p})}}, \quad \text{où } \bar{x} = \frac{\sum N_i x_i}{\sum N_i}$$

Dans les cas de la régression et de Spearman, $x_1 < x_2 < \cdots < x_k$. Ce n'est pas le cas pour Hamming. On considère donc une modification. Soit x_i = le nombre d'entier en commun entre $\{1, \cdots, \sum_{q=1}^{i} N_q\}$ et $\{N - y + 1, \cdots, N\}$. De cette manière $\{x_i\}$ représente une suite non-décroissante monotone.

Dans un exemple pratique, Neuhäuser [15] utilise des données sur le temps de réaction des souris à un stimuli sur leur queue. Il y a quatre groupes qui contiennent 10 animaux chacun. Un temps de réaction d'au moins 3 secondes est considéré comme un succès. Dans son article, il effectue des simulations avec les valeurs des p_i suivantes :

1. $p_1 = p_2 = p_3 = 0.1, p_4 = 0.1 + x$
2. $p_1 = p_2 = 0.1, p_3 = 0.1 + \frac{x}{2}, p_4 = 0.1 + x$
3. $p_1 = p_2 = 0.1, p_3 = p_4 = 0.1 + x$
4. $p_1 = 0.1, p_2 = 0.1 + \frac{x}{3}, p_3 = 0.1 + \frac{2x}{3}$
5. $p_1 = 0.1, p_2 = p_3 = 0.1 + \frac{x}{2}, p_4 = 0.1 + x$
6. $p_1 = 0.1, p_2 = p_3 = p_4 = 0.1 + x$

On procède donc à des simulations avec les mêmes valeurs de p_i pour les statistiques de Spearman, Hamming et la statistique de Hamming modifiée et on construit un graphique de la puissance des différents tests pour chacun des cas. Les graphiques (5.2) et (5.3) sont construits en suivant le même modèle mais en utilisant 0.3 et 0.5, respectivement, comme valeur de départ. Aussi, $N_i = 20$ pour tout i.

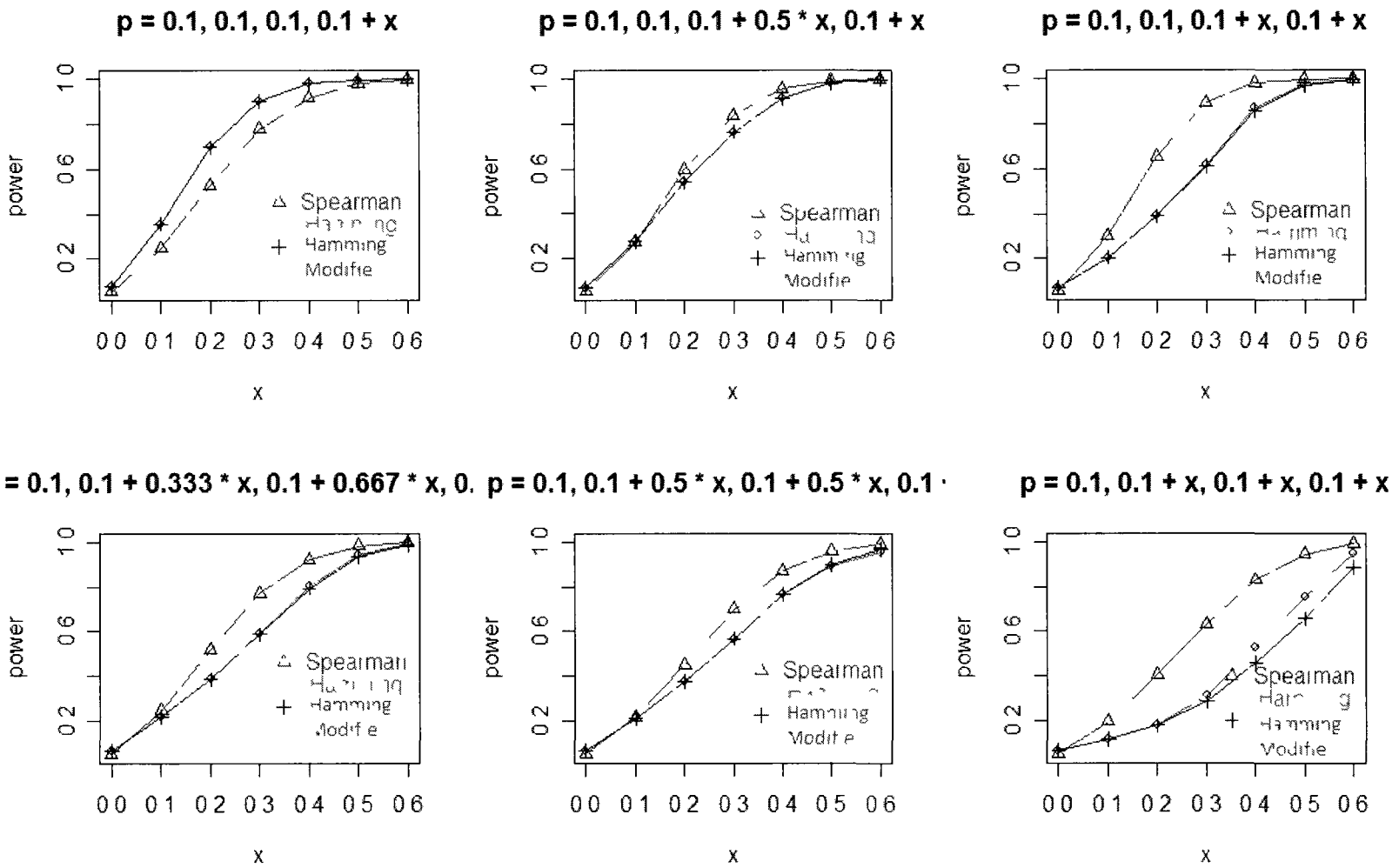

FIG 5 1 – Graphe de la puissance avec $\hat{p}_i = 0\,1$

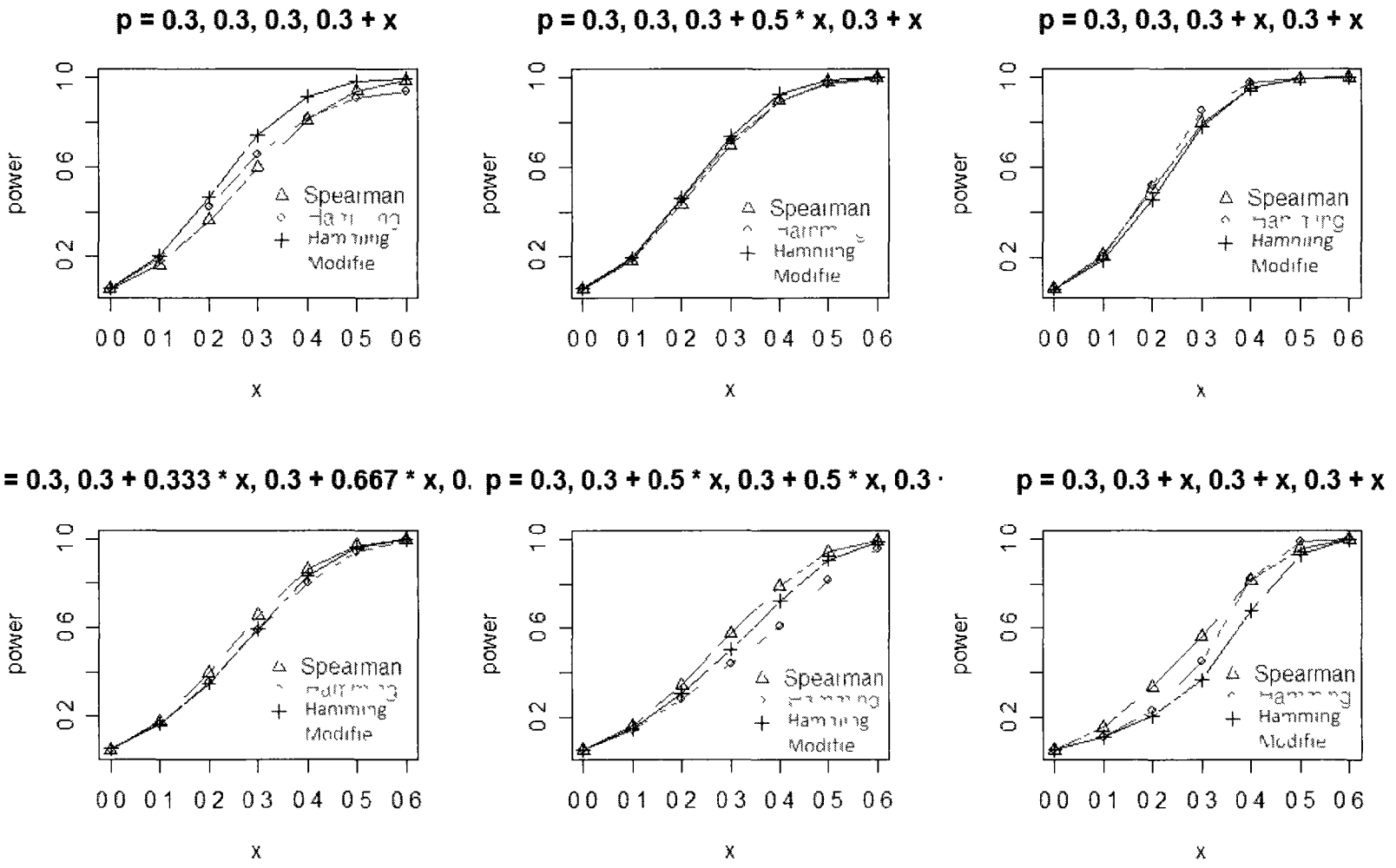

FIG. 5.2 – Graphe de la puissance avec $\hat{p}_i = 0\,3$

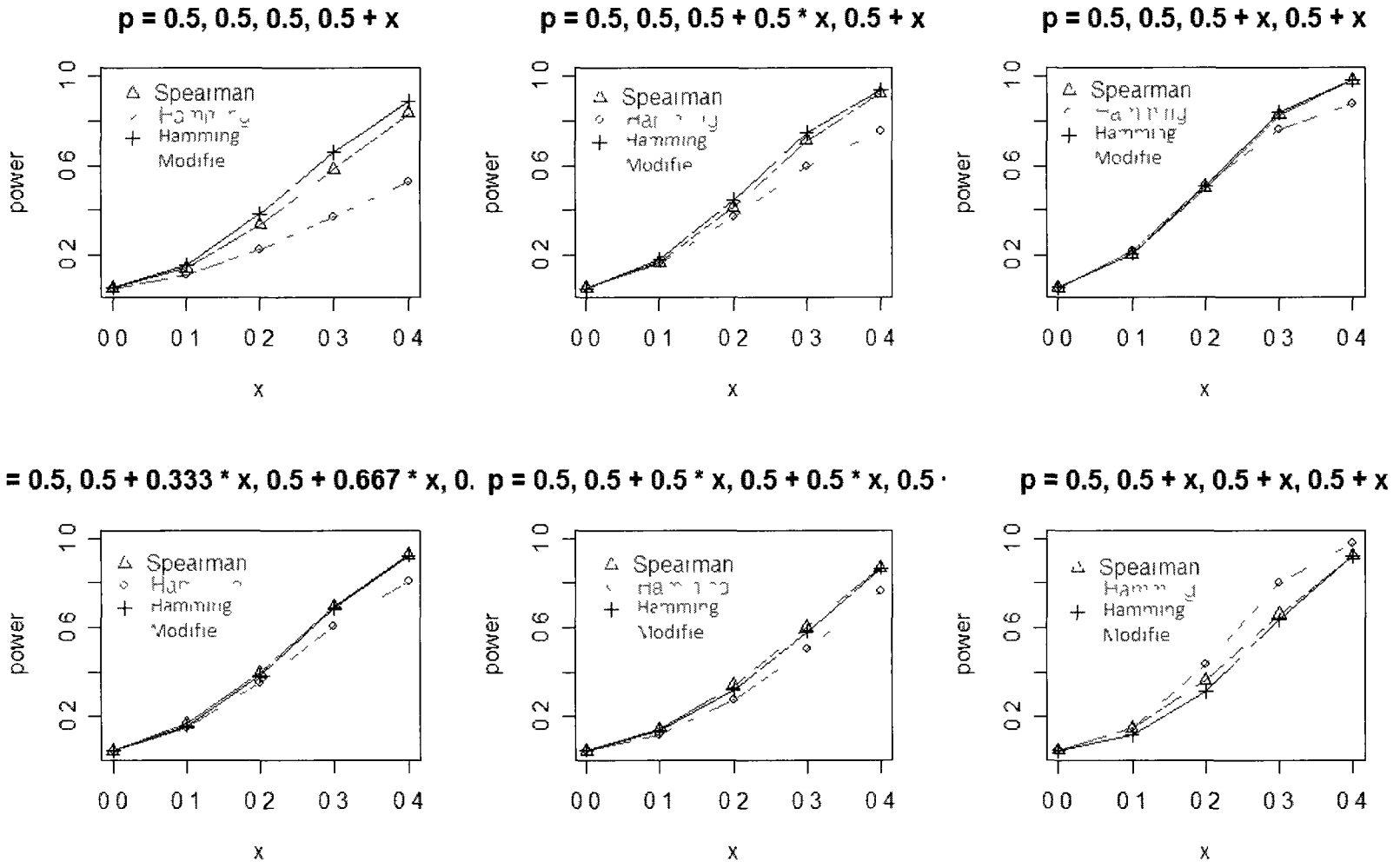

FIG. 5.3 – Graphe de la puissance avec $\hat{p}_i = 0.5$

On constate que la puissance de la statistique de Hamming modifiée est presque toujours plus élevée que celle de Hamming. De plus, lorsque la valeur de départ est plus grande, la puissance de la statistique de Hamming modifiée est presque toujours égale à celle pour la statistique de Spearman.

5.3 Exemple

Nous voulons comparer les différents tests de tendance présentés dans cette thèse. Nous avons utilisé des données sur les décès en Afrique du Sud pour les années 2000 à 2008.

Année	Nombre de décès	Population
2000	416 155	43 789 115
2001	454 882	43 997 828
2002	502 050	44 187 637
2003	556 779	44 344 136
2004	576 709	42 718 530
2005	598 131	42 768 678
2006	612 778	43 647 658
2007	603 094	43 586 097
2008	592 073	43 421 021

TAB. 5.12 – Données sur les décès en Afrique du Sud entre 2000 et 2008.

On veut tester l'hypothèse que tous les p_i sont les mêmes, i.e. $H_0 : p_i = p$ contre l'hypothèse qu'il y a une tendance croissante. Notons que dans cet exemple les y_i représentent le nombre de décès à chaque année et N_i représentent la population à chaque année. Puisqu'il y a 9 années, alors $k = 9$.

On procède donc au test de tendance en utilisant les méthodes suivantes : la méthode par tableau de contingence, le test de Spearman, le test de Hamming et le

test de Hamming modifiée.

5.3.1 La méthode par tableau de contingence

Cette méthode consiste à calculer

$$\chi^2_{k-1} = \sum_{i=1}^{k}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

et on rejette l'hypothèse nulle si $\chi^2_{k-1} > \chi^2_{k-1}(\alpha)$ avec $\alpha = 0.05$. Donc nous devons calculer

$$\chi^2_8 = \sum_{i=1}^{9}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 83353.07 > \chi^2_8(0.05) = 15.5$$

avec un p-value < 0.0001. On rejette donc très fortement l'hypothèse nulle.

5.3.2 Le test de Spearman

Nous devons calculer la statistique de Spearman :

$$S = \frac{\sum_i N_i \left(g_i - \dfrac{N+1}{2}\right)(\bar{p}_i - \bar{p})}{\sqrt{\sum_i N_i \left(g_i - \dfrac{N+1}{2}\right)^2}\sqrt{\bar{p}(1-\bar{p})}}$$

et on rejette l'hypothèse nulle si $S \geq z_\alpha$, pour $\alpha = 0.05$. On a que $S = 260.0833 > z_{0.05} = 1.645$ avec un p-value < 0.0001. donc on rejette très fortement l'hypothèse nulle.

5.3.3 Le test de Hamming

Nous devons calculer la statistique de Hamming :

$$H = \frac{\sum_i [N_i(w_i - \bar{p})(\bar{p}_i - \bar{p})]}{\sqrt{\sum_i N_i(w_i - \bar{p})^2}\sqrt{\bar{p}(1-\bar{p})}}$$

et on rejette l'hypothèse nulle si $H \geq z_\alpha$, pour $\alpha = 0.05$. On a que $H = 70.26753 > z_{0\,05} = 1.645$ avec un p-value < 0.0001. donc on rejette très fortement l'hypothèse nulle.

5.3.4 Le test de Hamming modifiée

On calcule la statistique de Hamming modifiée

$$H_M = \frac{\sum N_i(x_i - \bar{x})(\bar{p}_i - \bar{p})}{\sqrt{\sum N_i(x_i - \bar{x})^2}\sqrt{\bar{p}(1-\bar{p})}}, \quad \text{où } \bar{x} = \frac{\sum N_i x_i}{\sum N_i}$$

et x_i = le nombre d'entier en commun entre $\{1, \cdots, \sum_{q=1}^{i} N_q\}$ et $\{N - y + 1, \cdots, N\}$. et on rejette l'hypothèse nulle si $H_M \geq z_\alpha$, pour $\alpha = 0.05$. On a que $H_M = 70.26753 > z_{0\,05} = 1.645$ avec un p-value < 0.0001, donc on rejette très fortement l'hypothèse nulle.

Dans cet exemple, la statistique de Hamming et celle de Hamming modifiée donnent exactement le même résultat. On constate ici que tous les tests rejettent l'hypothèse nulle qu'il n'y a pas de tendance Puisque la statistique de Spearman et celle de Hamming sont toutes deux asymptotiquements des $N(0,1)$ et que la valeur de la statistique de Spearman est plus grande que celle de Hamming, on en conclut que la statistique de Spearman est meilleure que celle de Hamming ou celle de Hamming modifiée.

5.4 Discussion

En résumé, les tests statistiques étudiés sont :

Régression : $$\frac{\sum_i N_i(x_i - \bar{x}_k)(\bar{p}_i - \bar{p})}{\sqrt{\sum_i N_i(x_i - \bar{x}_k)^2}\sqrt{\bar{p}(1-\bar{p})}} = \frac{\hat{\beta}}{\sqrt{Var\hat{\beta}}}$$

Spearman : $$\frac{\sum_i N_i \left(g_i - \frac{N+1}{2}\right)(\bar{p}_i - \bar{p})}{\sqrt{\sum_i N_i \left(g_i - \frac{N+1}{2}\right)^2}\sqrt{\bar{p}(1-\bar{p})}}$$

Hamming : $$\frac{\sum_i [N_i(w_i - \bar{p})(\bar{p}_i - \bar{p})]}{\sqrt{\sum_i N_i(w_i - \bar{p})^2}\sqrt{\bar{p}(1-\bar{p})}}$$

Rappelons-nous que la régression mène au test de Cochran-Armitage et que pour pouvoir l'utiliser nous devons tout d'abord spécifier les scores. Portier et Hoel ([16]), Horthone et Bretz ([12]) et Corcoran et al ([8]) ont notés que le test de Cochran-Armitage est très sensible au choix des scores et conclut que l'utilisation générale de ce test n'est pas recommandé. Les tests statistiques de Spearman et de Hamming, qui sont basés sur les rangs, ont étés dérivés en utilisant le concept de compatibilité. C'est ce concept qui permet de justifier, dans le cas de Spearman, l'utilisation usuelle d'un rang moyen pour les exactos. À l'aide des simulations, on conclut que la statistique de Spearman a généralement une puissance plus élevée que celle de Hamming, et est donc le test recommandé. Par contre, la statistique de Hamming modifiée obtient une puissance plus élevée que celle de Hamming et dans certain cas, elle obtient une puissance très près de celle de Spearman. Il serait intéressant d'approfondir l'étude de la statistique de Hamming modifiée.

Chapitre 6

Conclusion

Cette thèse présente différents tests et méthodes pour étudier les tendances dans les proportions. Le test de Cochran-Armitage par exemple nécessite que l'on spécifie des scores. Corcoran et al. [8] mentionnent que le test de Cochran-Armitage est très sensible au choix des scores et que son utilisation générale n'est pas recommandée. Ce test est également sensible au choix des N_i. C'est pourquoi on explore deux autres tests basés sur les distances de Spearman et de Hamming. L'avantage de ces tests est que nous n'avons pas besoin de spécifier des scores, ces tests sont basés sur les rangs et sont donc plus adéquats pour une utilisation générale.

Les contributions principales de cette thèse sont le développement d'un test de tendance non paramétrique sur les proportions basé sur la distance de Hamming et la comparaison avec un test similaire basé sur la distance de Spearman. On a démontré que les tests de régression, de Spearman et de Hamming ont tous une forme semblable à l'exception du choix de pondération. Dans le cas de Spearman, la pondération est basée sur les tailles d'échantillons cumulatives. Sous l'hypothèse nulle, les statistiques de Spearman et de Hamming sont asymptotiquement normales. Pour comparer les deux tests, nous avons décidé de procéder à des simulations et de comparer les puissances obtenues pour la statistique de Hamming et celle de Spearman, pour des p_i croissants monotones lorsque les N_i sont tous égaux et également lorsqu'ils sont différents. Les résultats nous montrent que lorsque les p_i sont croissants monotones,

la statistique de Spearman obtient une puissance plus élevée Par contre, lorsque les p_i ne sont pas strictement croissants monotones, il est impossible de savoir lequel de Hamming ou Spearman aura la plus grande puissance Il serait donc intéressant de faire une étude plus approfondie de la puissance selon différentes alternatives Il serait également intéressant d'étudier plus profondémment la statistique de Hamming modifiée car on remarque que sa puissance est presque toujours plus grande que celle de la statistique de Hamming et que dans certaines circonstances celle-ci est très près de la puissance pour la statistique de Spearman

Bibliographie

[1] M. Alvo et P. Cabilio, Rank Correlations and the Analysis of Rank-Based Experimental Designs, *Probability Models and Statistical Analyses for Ranking Data*, Lecture Notes in Statistics **80**, Springer-Verlag, pp. 140-154, 1993.

[2] M. Alvo et P. Cabilio, Applications of Hamming Distance to the Analysis of Block Designs, *Asymptotic Methods in Probability and Statistics* B. Szyszkowicz ed., Elsevier Science, pp. 787-800, 1998.

[3] Y. Arase et al., Randomized controlled clinical trial of lymphoblastoid interferon-alpha for chronic hepatitis C, *Hepatology*, **21(1)**, pp.55-66, 2001.

[4] P. Armitage, Tests for Linear Trends in Proportions and Frequencies, *Biometrics*, **11**, No. 3, pp. 375-386, 1955.

[5] W. Baumgartner et al., A Nonparametric Test for the General Two-Sample problem, *Biometrics*, **54(3)**, pp.1129-1135, 1998.

[6] J.J. Chen et al., Significance Levels of Randomization Trend Tests in the Event of Rare Occurrences, *Biometrical Journal*, **39(3)**, pp.327-337, 1997.

[7] W.G. Cochran, Some Methods for Strengthening the Common χ^2 Tests, *Biometrics*, **10**, pp.417-451, 1954.

[8] C. Corcoran et al., Power Comparisons for Tests of Trend in Dose-Response Studies, *Statistics in Medicine*, **19(22)**, pp.3037-3050, 2000.

[9] P. Diaconis et R.L. Graham, Spearman's Footrule as a Measure of Disarray, *Journal of the Royal Statistical Society*, **B39**, pp.262-268, 1977.

[10] J. Hájek, Z. Šidák, **Theory of Rank Tests**, Academic Press, 1967.

[11] R.V. Hogg et E.A.Tanis, **Probability and Statistical Inference**, Macmillan Publishing Company, 1993.

[12] L.A. Hothorn et F. Bretz, Evaluation of Animal Carcinogenicity Studies : Chochran-Armitage Trend Test vs. Multiple Contrast Tests, *Biometrical Journal*, **42(5)**, pp. 553-567, 2000.

[13] J.H. Ku et al., The Prevalence of Chronic Prostatitis-Like Symptoms in Young Men : A Community-Baser Survey, *Urological Research.* **29(2)**, pp.108-112, 2001.

[14] M.H. Kutner, W. Li, C.J. Nachtsheim et J. Neter, **Applied Linear Statistical Models, 5th Edition**, McGraw-Hill/Irwin, 2005.

[15] M. Neuhäuser, An Exact Test for Trend Among Binomial Proportions Based on a Modified Baumgartner-Weiβ-Schindler Statistic, *Journal of Applied Statistics*, **Vol. 33, No. 1**, pp. 79-88, 2006.

[16] C. Portier et D. Hoel, Type I Error of Trend Tests in Proportions and the Design of Cancer Screens, *Communications in Statistics - Theory and Methods*, **13(1)**, pp. 1-14, 1984.

[17] D.A. Williams, Tests for Differences Between Several Small Proportions, *Applied Statistics*, **37(3)**, pp.421-434, 1988.