Etude de puissance pour la détection d'association directe et indirecte à partir de tables de contingence 2×2.

Mathieu Emily [1] & Chloé Friguet [2]

[1] *Agrocampus Ouest - IRMAR, mathieu.emily@agrocampus-ouest.fr*
[2] *LMBA - Univ. de Bretagne-Sud / IUT de Vannes, chloe.friguet@univ-ubs.fr*

Résumé. L'objectif de cet article est de comparer la puissance des tests d'association du χ^2, de Wald et de déviance, couramment utilisés pour étudier le lien entre deux variables qualitatives à partir de tables de contingence. Nous avons tout d'abord mené une étude de puissance portant sur l'effet conjoint des fréquences observées pour chacune des deux variables. A partir de simulations nous avons pu montrer, d'une part, que le test du χ^2 est systématiquement plus puissant que le test de Wald et d'autre part, que les tests du χ^2 et de déviance ont des comportements différents. L'analyse d'un jeu de données pangénomiques a confirmé les résultats des simulations ouvrant des perspectives intéressantes quant à l'étude de variants rares.

Mots-clés. Test d'association, Table de contingence, Etude d'association pangénomique

Abstract. In this paper, we propose a power study comparing three widely-used association tests: χ^2 test, Wald test and likelihood ratio test (LRT). Using simulations, we demonstrate that χ^2 test always outperforms Wald test, while χ^2 test and LRT display different behaviors with respect to the observed frequencies of the studied variables. The analysis of a Genome-wide association studies dataset has confirmed our simulations results and gives new insights in the detection of rare variants.

Keywords. Association test, Contingency table, Genome-wide association studies

1 Introduction

L'utilisation de tables de contingence permet notamment l'étude de l'association entre deux variables qualitatives. Dans le cas particulier de variables qualitatives à deux modalités, les données peuvent se résumer dans une table de contingence 2×2 croisant ainsi les modalités des deux variables. Largement utilisées dans de nombreux domaines applicatifs, les tables de contingence 2×2 ont fait l'objet de nombreuses recherches permettant le développement de plusieurs mesures d'association (Shieh, 2000). En biologie et médecine, trois tests d'association sont particulièrement utilisés: le test du χ^2, le test de Wald et le test de déviance. En l'absence d'association (hypothèse nulle $\mathcal{H}_0$), ces trois tests sont asymptotiquement équivalents et leurs statistiques de test suivent une loi du χ^2 à 1 degré de liberté. Cependant, en présence d'association, qui constitue notre hypothèse alternative $\mathcal{H}_1$, les statistiques de tests suivent des lois du χ^2 décentrées, pour lesquelles le

coefficient de décentrage, relié directement à la puissance, dépend de plusieurs paramètres inhérents au jeu de données traité.

L'objectif de cet article est de comparer les performances, en termes de puissance, des tests du χ^2, de Wald et de déviance sous l'hypothèse alternative $\mathcal{H}_1$. Nous distinguerons deux types d'association : l'association directe et l'association indirecte. Contrairement à une association directe, le lien détecté par une association indirecte entre deux variables n'existe qu'à travers une troisième variable. Cette troisième variable présente la particularité d'être associée directement aux deux variables testées.

Dans un premier temps, nous étudierons les puissances relatives des tests à partir de simulations, utilisant notamment un modèle logistique, dans le cas d'associations directes et indirectes. Dans un second temps, nous appliquerons notre méthodologie à la détection de marqueurs génétiques associés au développement de maladies complexes. Ces études, appelées études d'association pangénomiques, permettent de tester l'association de plusieurs centaines de milliers de marqueurs de polymorphisme avec un phénotype binaire cas/témoins. Grâce à la structure en bloc du génome, seuls certains marqueurs, appelés *tag*, sont testés. Ces marqueurs *tag*, choisis pour leur forte liaison avec d'autres marqueurs non testés, sont ainsi considérés comme étant représentatif d'un ensemble de marqueurs. En conséquence, les associations détectées par l'intermédiaire des études pangénomiques sont souvent des associations indirectes.

2 Notations et tests d'association

Nous cherchons à tester l'association entre une variable Y et une variable C, telles que Y et C suivent chacune des lois de Bernoulli: $Y \sim \mathcal{B}(\pi_Y)$ et $C \sim \mathcal{B}(\pi_C)$. Considérons un recueil de données pour lequel les populations "$Y = 0$" et "$Y = 1$" ont été échantillonnées de façon indépendante. Notons φ la proportion d'observations telles que $Y = 1$. De façon symétrique, nous pouvons introduire $\widetilde{\pi_C}$ la proportion d'observations telle que $C = 1$. En notant $\pi_{C_{|Y}}$ (resp. $\pi_{\overline{C}_{|\overline{Y}}}$) la probabilité conditionnelle de $C = 1$ (resp. $C = 0$) sachant $Y = 1$ (resp. $Y = 0$), la matrice des comptages attendus peut se représenter par la Table 1.

	$Y = 0$	$Y = 1$	
$C = 0$	$n_{\overline{C}\overline{Y}} = n(1-\varphi)\pi_{\overline{C}_{\|\overline{Y}}}$	$n_{\overline{C}Y} = n\varphi\pi_{\overline{C}_{\|Y}}$	$n_{\overline{C}} = n(1-\widetilde{\pi_C})$
$C = 1$	$n_{C\overline{Y}} = n(1-\varphi)\pi_{C_{\|\overline{Y}}}$	$n_{CY} = n\varphi\pi_{C_{\|Y}}$	$n_C = n\widetilde{\pi_C}$
	$n_{\overline{Y}} = n(1-\varphi)$	$n_Y = n\varphi$	n

Table 1: Table des comptages attendus en croisant les variables Y et C.

L'association entre Y et C, supposée directe, est quantifiée par les paramètres γ_0 et

β_1 tels que :

$$\text{logit}\left(\pi_{Y_{|C}}\right) = \text{logit}\left(\mathbb{P}[Y=1|C=c]\right) = \log\left(\gamma_0\right) + \beta_1 c \tag{1}$$

Afin d'étudier les interactions indirectes, nous introduisons également une variable T, telle que T suit également une loi de Bernoulli ($T \sim \mathcal{B}(\pi_T)$). Deux hypothèses supplémentaires sont introduites:

- C et T sont corrélées par un coefficient r: $Cor(C,T) = r$
- Condionnellement à C, Y et T sont indépendante: $T_{|C} \perp\!\!\!\perp Y$.

Dans notre cas particulier d'une matrice de contingence 2×2, les statistiques, associées aux tests de χ^2, de Wald et de déviance, ont des formules explicites que nous rappelons ci-dessous:

- Statistique du χ^2:

$$\chi^2 = n \times \left(\pi_{C_{|Y}} - \pi_{C_{|\overline{Y}}}\right)^2 \frac{\varphi(1-\varphi)}{\widetilde{\pi_C}\left(1-\widetilde{\pi_C}\right)}$$

- Statistique du test de Wald: $t = \frac{\widehat{\beta_1}}{\sqrt{\mathbb{V}[\widehat{\beta_1}]}}$ avec

$$\widehat{\beta_1} = \log\left(\frac{\pi_{C_{|Y}}\pi_{c_{|\overline{Y}}}}{\pi_{C_{|\overline{Y}}}\pi_{c_{|Y}}}\right) \text{ et } \mathbb{V}[\widehat{\beta_1}] \approx \frac{1}{n(1-\varphi)}\left(\frac{1}{\pi_{|C_{\overline{Y}}}} + \frac{1}{\pi_{c_{|\overline{Y}}}}\right) + \frac{1}{n\varphi}\left(\frac{1}{\pi_{C_{|Y}}} + \frac{1}{\pi_{c_{|Y}}}\right)$$

- Statistique du test de déviance: $LRT = -2\left(\mathcal{L}(\mathcal{M}_{Sat}) - \mathcal{L}(\mathcal{M}_0)\right)$ avec

$$\begin{aligned}\mathcal{L}(\mathcal{M}_{Sat}) &= \sum_C \sum_Y n_{CY}\log(n_{CY}) - \sum_C n_C \log(n_C)\\ \mathcal{L}(\mathcal{M}_0) &= \sum_Y n_Y \log(n_Y) - n\log(n)\end{aligned}$$

3 Etude de simulation

Pour notre étude de simulation, le modèle proposé à l'équation (1) nous a permis d'obtenir les matrices de contingence attendues (voir Table 1). A l'aide des formules développées à la section précédente, nous obtenons ainsi l'espérance des statistiques pour les trois tests comparées : χ^2, Wald et déviance. Ces espérances représentent le coefficient de décentrage des loi du χ^2 suivies par chacune des statistiques sous $\mathcal{H}_1$. Pour chaque test, nous pouvons ainsi calculer la puissance théorique comme la probabilité qu'une variable, suivant la loi du χ^2 décentrée correspondante, soit supérieure au quantile de niveau α de la loi du χ^2 centrée.

Afin de mimer le jeu de données réelles traité dans cet article, nous avons fixé le nombre total d'observations à $n = 2\ 000$. La probabilité d'observer $Y = 1$ est supposée constante et égale à 10^{-5}, ce qui implique que $\gamma = 9\ 999$. De plus, pour simuler une correction de Bonferroni pour un jeu de données avec 500 000 variables, nous avons choisi $\alpha = 10^{-7}$.

3.1 Association directe

Pour étudier les variations de puissance entre les trois tests, nous nous sommes focalisés, d'une part, sur l'effet de la proportion d'observations telles que $Y = 1$ (φ), et d'autre part sur l'effet de la probabilité de l'événement $C = 1$ (π_C). La Table 2 résume les puissances obtenues par les 3 tests pour les valeurs suivantes: $\varphi \in \{0.10, 0.45, 0.90\}$, $\pi_C \in \{0.05, 0.10, 0.50\}$ et $\beta_1 \in \{0.5, 1.0, 1.5\}$.

		$\varphi = 0.10$			$\varphi = 0.45$			$\varphi = 0.90$		
		χ^2	Wald	Dev	χ^2	Wald	Dev	χ^2	Wald	Dev
$\pi_C = 0.05$	$\beta_1 = 0.5$	0	0	0	0	0	0	0	0	0
	$\beta_1 = 1.0$	0.16	0.12	0.07	0.76	0.7	0.76	0.01	0.01	0.03
	$\beta_1 = 1.5$	0.99	0.97	0.89	1	1	1	0.35	0.22	0.62
$\pi_C = 0.10$	$\beta_1 = 0.5$	0	0	0	0.05	0.05	0.05	0	0	0
	$\beta_1 = 1.0$	0.6	0.53	0.4	1	0.99	1	0.15	0.12	0.24
	$\beta_1 = 1.5$	1	1	1	1	1	1	0.92	0.82	0.98
$\pi_C = 0.50$	$\beta_1 = 0.5$	0.02	0.02	0.02	0.56	0.56	0.57	0.03	0.02	0.02
	$\beta_1 = 1.0$	0.81	0.75	0.84	1	1	1	0.93	0.90	0.89
	$\beta_1 = 1.5$	1	1	1	1	1	1	1	1	1

Table 2: Résumé de la puissance des trois tests pour le cas d'une association directe.

Nous pouvons tout d'abord remarquer que le test du χ^2 est toujours plus puissant que le test de Wald. D'autre part, nos résultats montrent que lorsque $\varphi = 0.45$, les tests du χ^2 et de déviance sont équivalents pour toutes les valeurs de π_C et de β_1. Par contre, dans le cas où la proportion observée de $Y = 1$ est faible ($\varphi = 0.10$) et que π_C est faible ($\pi_C \leq 0.10$), le test du χ^2 est plus puissant que le test de déviance. Au contraire, lorsque $\varphi = 0.10$ et π_C est élevé ($\pi_C = 0.50$), le test de déviance devient plus puissant que le test du χ^2. La situation est complètement inversée pour φ élevé ($\varphi = 0.90$). En effet, dans ce cas, le test de déviance devient plus puissant que le test du χ^2 à mesure que π_C devient faible.

3.2 Association indirecte

Pour évaluer la puissance dans le cas d'une association indirecte, nous avons introduit une troisième variable, T, dans les simulations. Nous avons fixé la corrélation entre C et T, telle que $r = 0.80$. Les résultats présentés dans la Table 3 ont été obtenu en fixant $\pi_C = 0.40$ et $\beta_1 = 1.0$. Les autres valeurs testées pour ces deux paramètres n'ont pas montré de différences dans les conclusions obtenues.

D'après la Table 3, nous pouvons tout d'abord remarquer que, pour tous les tests, la puissance augmente lorsque π_T augmente. D'autre part, nous retrouvons les mêmes

conclusions pour la comparaison des méthodes que celles obtenues dans le cas d'une association directe. L'effet indirect n'influence donc pas différemment les trois tests.

	$\varphi = 0.10$			$\varphi = 0.45$			$\varphi = 0.90$		
	χ^2	Wald	Dev	χ^2	Wald	Dev	χ^2	Wald	Dev
$\pi_T = 0.35$	0.15	0.14	0.13	0.95	0.94	0.95	0.11	0.11	0.12
$\pi_T = 0.40$	0.15	0.14	0.14	0.96	0.95	0.96	0.14	0.13	0.14
$\pi_T = 0.45$	0.16	0.14	0.16	0.97	0.96	0.97	0.17	0.16	0.16

Table 3: $\beta_1 = 1.0$, $\pi_C = 0.40$

4 Analyse du jeu de données WTCCC

Nous avons appliqué notre analyse comparative à des données d'association pangénomiques. Ces études ont pour objectif la détection de différences de fréquence d'allèle entre une population d'individus malades et une population d'individus sains. Chaque individu est représenté un phénotype binaire "malade" ou "sain". De plus, nous disposons, pour chaque individu, de mesures de génotypes évaluées par des marqueurs nucléotidiques (SNPs), répartis le long du génome. En pratique, l'association entre le phénotype et chaque marqueur est testée individuellement.

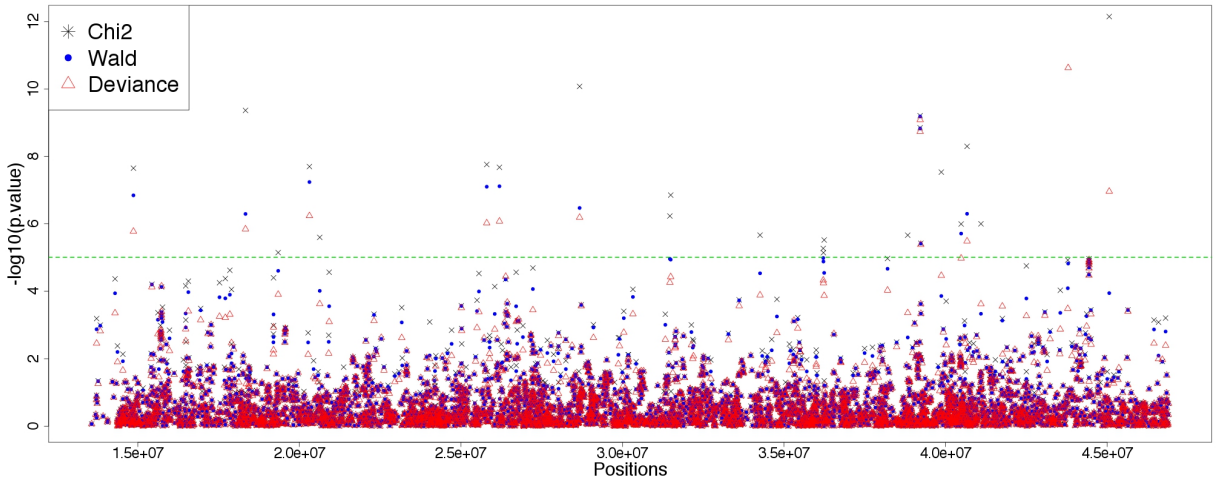

Figure 1: Tests d'associations de 5 000 SNPs du chromosome 21 et la maladie de Crohn à partir des données WTCCC. L'axe des ordonnées représente l'opposé du log en base 10 de la p-valeur. La barre verte correspond à la limite de significatifvé après correction de Bonferroni.

Pour reprendre les notations introduites à la section 2, le phénotype est représenté par la variable Y, où $Y = 1$ code pour le phénotype "malade" et $Y = 0$ pour le phénotype "sain". La variable C représente le marqueur testé. Nous supposerons ici que C modélise la présence ($C = 1$) et l'absence ($C = 0$) de l'allèle mineure. Il est à noter ici que nous ne connaissons pas les marqueurs responsables du développement de la maladie, appelés marqueurs causaux. De plus, du fait de la structuration en bloc du génome, seuls des marqueurs *tag* sont présents dans le jeu de données. Par conséquent, il est fort probable que nous nous trouvions dans une situation de test d'association indirecte. En effet, pour chaque marqueur causal, l'association est vraisemblablement mesurée indirectement par l'intermédiaire du marqueur *tag* qui lui est associé.

Nous avons focalisé notre attention sur le jeu de données du Welcome Trust Case Control Consortium (WTCCC, 2007) et plus particulièrement sur le chromosome 21 pour la maladie Crohn. Les résultats obtenus et représentés à la Figure 1 montre qu'il existe des différences importantes entre les tests notamment pour les associations significatives. En accord avec les simulations, le test du χ^2 est plus puissant que le test de Wald. De plus, le test de χ^2 apparaît être plus puissant que le test de déviance.

5 Discussion

Dans cet article, nous nous sommes intéressé aux variations de puissance qui peuvent exister entre trois tests couramment utilisés (χ^2, Wald et déviance) pour déterminer l'association entre 2 variables qualitatives à 2 modalités. Par une étude par simulation et l'analyse d'un jeu de données réelles, nous avons montré que le test du χ^2 est systématiquement plus puissant que le test de Wald. D'autre part, la comparaison entre le test du χ^2 et le test de déviance dépend fortement de la conception du jeu de données traité.

Nos résultats offrent des perspectives intéressantes quant au choix du test à utiliser pour mettre en avant un effet particulier. Dans le contexte des études pangénomiques, il est bien connu que la puissance est faible pour détecter des variants rares. Nous avons pu montrer que la détection de variants rares (π_C très faible) est très sensible au test utilisé ainsi qu'au design du jeu de données, notamment la valeur de φ. D'autre part, nos résultats ouvrent de nouvelles pistes pour améliorer le processus de *tagging* afin d'améliorer la détection indirecte de variants.

Bibliographie

[1] G. Shieh, On power and sample size calculations for likelihood ratio tests in generalized linear models (2000) Biometrics **56**: 1192–1196.

[2] The Wellcome Trust Case Control Consortium, Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls (2007) *Nature* **447**: 661–678.