L'odds ratio et la mesure de l'association

Michel Novi – 6 juin 2006

Introduction : les odds ratios, une méthode d'analyse locale

Figure 1. **Origine scolaire et réussite en L1 (psychologie, Nice)**

	réussit	AJAC	ajourné	abandon S2	abandon S1	non ins. ex.	TOTAL	Θ(S,x)
S	a = 41	1	b = 17	3	0	2	64	= (a/b)/(c/d)
ES	52	1	18	8	5	4	88	0,8
L	55	0	27	9	1	6	98	1,2
STI SMS	c = 23	2	d = 16	5	0	1	47	1,7
STT	14	3	30	6	3	3	59	5,2
Bac Pro	3	0	14	7	4	1	29	11,3
TOTAL	188	7	122	38	13	17	385	

Les scientifiques ont plus de chances de réussir que d'être ajournés : 2,4 fois plus.
Les STI-SMS aussi : 1,4 fois plus.
Les scientifiques sont toutefois plus avantagés (1,7 fois).

1. Un paradigme classique : la différence des pourcentages

Une méthodologie additive fondée sur $\Delta = a/(a+b) - c/(c+d)$.
Exemple : passage en L2 des étudiants de sociologie (Nice) selon l'origine scolaire.
G = bac général ; TP = bac technologique ou professionnel ; P = passe en 2ème année ; NPP = ne passe pas.

Figure 2. **En socio… Δ et la lignée du X²**

	P	NPP	Total
G	a = 60	b = 30	90
TP	c = 15	d = 45	60
Total	75	75	150

X²cor = 23,361
X² = 25,000
Φ² = X²/N = 0,167
Φ = 0,408

% en lignes			
	0,667	0,333	1
	0,250	0,750	1
Δij	**0,417**		

% en colonnes			
	0,800	0,400	**0,400**
	0,200	0,600	
	1	1	

Δji

Δij x Δji = 0,167

- Δ est donc asymétrique : il est différent, calculé en lignes ou en colonnes.
  - On ne peut le prendre pour un indice d'association qui attribuerait le même rôle aux deux variables.
  - Avantage dans une visée causale (VI et VD) ou praxéologique (prédictive) et d'ailleurs Δ est un coefficient de régression.
- X^2 est symétrique mais n'est pas un indice d'association ; Φ^2 (carré moyen de contingence) si.
  - Ainsi que les indices construits à partir de Φ^2 : T de Tschuprow, V de Cramér (différents de Φ^2 si ddl > 1), C de Pearson...
  - On peut aussi tenir compte des marges (cf. Φ/Φmax, coef. de Benini, PEM de Cibois…) pour relativiser les mesures…

2. Des quotients, rien que des quotients

2.1. Définition

Le calcul du rapport des chances (*odds*, coefficient concurrentiel du 1er ordre) :

RC(G) = a/(a+b) / b/(a+b) = (60/90) / (30/90) = 2 = a/b
Les bacheliers G ont 2 fois plus de chances de passer que de [chances de] ne pas passer

RC(TP) = c/(c+d) / d/(c+d) = (15/60) / (45/60) = 0,333 = c/d
Les bacheliers TP ont 3 fois moins de chances de passer que de ne pas passer

On en déduit l'*odds ratio* : rapport des rapports des chances, coefficient concurrentiel du 2ème ordre :

Θ = (a/b) / (c/d) = 2 / 0,333 = 6

« Les G ont 6 fois plus de chances que les TP de passer plutôt que de ne pas passer. »
Ou : « Par rapport aux TP, les G ont 6 fois plus de chances de passer que de ne pas passer. »

Figure 3. **En psycho…**

	P	NPP	Total
G	150	100	250
TP	45	90	135
Total	195	190	385

X^2cor = 23,884
X^2 = 24,939
$\Phi^2 = X^2/N$ = 0,065
Φ = 0,255

% en lignes	0,600	0,400	1
	0,333	0,667	1
Δij	**0,267**		

% en colonnes	0,769	0,526	**0,243**
	0,231	0,474	
	1	1	

Δji

Δij x Δji = 0,065

Θ = (150/100) / (45/90) = 1,5 / 0,5 = 3.
L'inégalité (due à l'origine scolaire) est deux fois plus forte en sociologie.

Quatre manières de procéder :

(a) Calculer des pourcentages. Donc rapporter aux marges. D'ici, évaluer l'inégalité par une différence (Δ) ou par un rapport (R). R est le rapport des pourcentages (opération très fréquente : rapport des risques – *risk ratio*). Dans l'exemple *socio* :
R = [a/(a+b)] / [c/(c+d)] = 2,67 ou 2,25… le ratio n'est plus le même sur les proportions complémentaires.
(b) Rapporter l'effectif à un autre effectif : tenir compte à la fois des pourcentages et de leurs complémentaires.
On aurait pu faire ensuite une différence (δ). Non ! On fait l'odds ratio : que des rapports !

Figure 4. **Les quatre manières**

	marges oui	marges non
-	Δ	δ = a/b - c/d
÷	R	Θ

2.2. Propriétés immédiates

(1) Indépendance ⇔ [Θ = 1] (tableau 2x2)

Figure 5. **Les effectifs d'indépendance**

	P	NPP	Total	
G	45	45	90	Δ = 0
TP	30	30	60	Θ = 1
Total	75	75	150	Log Θ = 0

(2) Interprétation probabiliste : Θ = ad/bc

Hypothèse de travail : formation plus adaptée des bacheliers généraux
Configuration représentative de l'hypothèse : le bachelier G passe ; le bachelier TP ne passe pas.
Probabilité (praxéologique) : P_1 = a/(a+b) x d/(c+d) = 0,667 x 0,75 = 0,5
Configuration inverse : le bachelier G ne passe pas ; le bachelier TP passe.
Probabilité (praxéologique) : P_0 = b/(a+b) x c/(c+d) = 0,333 x 0,25 = 0,0833
Rapport des probabilités :
P_1/P_0 = **ad / bc** = 6 = Θ = rapport du produit croisé (*cross product ratio*)
Si on tire au hasard un étudiant parmi les G et si on tire au hasard un étudiant parmi les TP, on a 6 fois plus de chances (Θ) de les trouver dans la configuration attendue que dans la configuration inverse.

(3) Invariance par symétrie et permutation (Figure 6)

	calcul	valeur	sens	commentaire	Log
Θ_1 =	a/b / c/d	6	en ligne	"6 fois plus"	1,792
Θ_2 =	a/c / b/d	6	en colonne		1,792
Θ_3 =	c/d / a/b	0,167	en ligne	"6 fois moins"	-1,792
Θ_4 =	b/d / a/c	0,167	en colonne		-1,792

(4) Domaine de variations : $0 \leq \Theta < +\infty$ $\quad -\infty < \text{Log}\,\Theta < +\infty$

Avantage : très pratique pour les *modèles*, la « réponse » sera toujours dans l'intervalle.
Cf. le « logit », Log du RC : logit(p) = Log[p/(1-p)] ∈ [- ∞ , + ∞] ; le Log-odds : « taux logistique », idem.
Inconvénient : sensibilité aux variations sur les petits effectifs ; caractère « absorbant » du cas *n* = 0.

Figure 7. **Cas extrêmes** (calcul de Θ =ad / bc)

	P	NPP	
G	75	15	$\Theta = +\infty$
TP	0	60	$\text{Log}\,\Theta = +\infty$

	P	NPP	
G	15	75	$\Theta = 0$
TP	60	0	$\text{Log}\,\Theta = -\infty$

Nota : un cas encore plus extrême… deux zéros sur la même ligne ou la même colonne !

(5) Propriété fondamentale d' « homothétie ». Invariance de Θ pour **H : $n_{ij} \rightarrow r_i\ s_j\ n_{ij}$**

Figure 8. **Exemple** ($r_1 = 0{,}5$; $s_1 = 3$)

	P	NPP	Total	
G	30	15	45	
TP	15	45	60	
Total	45	60	105	$\Theta = 6$

	P	NPP	Total	
G	90	15	105	
TP	45	45	90	
Total	135	60	195	$\Theta = 6$

- avec Θ on peut toujours calculer dans les deux sens même lorsqu'un des deux facteurs est « contrôlé » (et même mieux : possibilité d'estimer un rapport de risque par un odds ratio)
- Θ est laissé invariant par tout *algorithme* fondé sur des transformations de type H (voir RAS, Annexe).

(6) Associativité de l'odds ratio (Figure 9)

	P	NPP				
G	60	30	$\Theta_{G/T}$	5,0		
Technol.	14	35			$\Theta_{G/P}$	20
Prof.	1	10	$\Theta_{T/P}$	4,0		

Il y a 3 odds ratios *différents* possibles : l(l-1)/2 x c(c-1)/2,
mais seulement 2 odds ratios *indépendants* : (l-1)(c-1) =ddl.

(7) Compatibilité avec l'analyse log-linéaire

Logit, odds ratio, log-odds : les outils de la régression logistique et des modèles log-linéaires.

2.3. Tests et intervalles de confiance

- **intervalle de confiance pour un Θ**

En socio : $\Theta = 6$ $\quad \boxed{E_\Theta = \Theta \times \sqrt{(1/a + 1/b + 1/c + 1/d)}}$

$$= 6 \sqrt{(1/60 + 1/30 + 1/15 + 1/45)} = 2{,}236$$

$$L_1 = 6 - 1{,}96 \times 2{,}236 = 1{,}617 \qquad L_2 = 6 + 1{,}96 \times 2{,}236 = 10{,}383$$

En psycho : $\Theta = 3$ $\quad E_\Theta = 0{,}671$ $\quad L_1 = 1{,}685$ $\quad L_2 = 4{,}315$

- **test de Θ = 1 (indépendance) dans un tableau 2x2**

$$\boxed{X^2 = (\text{Log}\,\Theta)^2 / (1/a + 1/b + 1/c + 1/d)}$$

$$= (1{,}792)^2/0{,}139 = 23{,}115 \quad \text{(socio)}$$
$$= (1{,}099)^2/0{,}050 = 24{,}139 \quad \text{(psycho)}$$

- **comparaison de deux odds ratio (test de $\Theta_1/\Theta_2 = 1$)** (socio contre psycho)

$$\boxed{z = (\text{Log}\Theta_1 - \text{Log}\Theta_2) / \sqrt{[(1/a + 1/b + 1/c + 1/d) + (1/e + 1/f + 1/g + 1/h)]}}$$

$$z = (1{,}792 - 1{,}099)/ \sqrt{(0{,}139 + 0{,}050)} = \mathbf{1{,}595} < 1{,}96$$

Pour $\alpha = .05$, z ne détecte pas d'effet d'interaction (même en test unilatéral).

3. Que choisir ?

Les mesures ne sont ni exemptes de contradictions (entre elles) ni de défauts intrinsèques.

Sur les données, l'inégalité est moins forte en psycho, quelle que soit la mesure : pas de contradiction entre Δ, R et Θ ; ni d'incohérence entre R et R*.

Figure 10.

socio	P	NPP	Total
G	60	30	90
TP	15	45	60
Total	75	75	150

psycho	P	NPP	Total
G	150	100	250
TP	45	90	135
Total	195	190	385

			l'inégalité…			
Δ	0,417		**diminue**	(0,64)	Δ	0,267
R et R*	2,667	et 2,25	**diminue**	(0,68)	R et R*	1,800 et 1,667
Θ	6		**diminue**	(0,50)	Θ	3

Si on fait varier n – le taux de passage des TP de psycho – on observe des contradictions et des incohérences pour n de 25 à 30 (taux entre 18,5% et 22,2%) :

Figure 11.

socio	P	NPP	Total
G	60	30	90
TP	15	45	60
Total	75	75	150

psycho	P	NPP	Total
G	150	100	250
TP	n = 27	108	135
Total	177	208	385

			l'inégalité…			
Δ	0,417		**diminue**	(0,96)	Δ	0,400
R et R*	2,667	et 2,25	**augmente**	(1,13)	R et R*	3,000 et 2,000
Θ	6		**est stable**	1,00)	Θ	6,000

Mais est-ce un domaine de variations plausible ?

Annexe. Calage sur marges d'un échantillon par l'algorithme de Deming Stephan

	Cadre	Autre	Total	Cadrage	
Homme	**80**	320	400	350	
Femme	30	570	600	650	
Total	110	890	1000		Θ
Cadrage	150	850		1000	**4,75**
(1a)	70,000	280,000	350	350	
70 =	32,500	617,500	650	650	
350x80/400	102,500	897,500	1000		Θ
	150	850		1000	**4,75**
(1b)	**102,439**	265,181	367,620	350	
102,439 =	47,561	584,819	632,380	650	
150x70/102,5	150	850	1000		Θ
	150	850		1000	**4,75**
(2a)	97,529	252,471	350	350	
	48,886	601,114	650	650	
	146,415	853,585	1000		Θ
	150	850		1000	**4,75**
(2b)	99,917	251,411	351,328	350	
	50,083	598,589	648,672	650	
	150	850	1000		Θ
etc	150	850		1000	**4,75**