Université Paris Ouest - Nanterre La Défense
Master M1 SES
Enquêtes et méthodes d'analyses quantitatives
A.K. FERMIN et C. Hardouin

Test d'indépendance de deux variables qualitatives (Chi2) Rapport de chances et Odds-ratio

Exercice 1 *Considérons la table de mobilité sociale issue de l'enquête sur l'emploi de juin 1953, obtenue pour les hommes français et étrangers actifs âgés de 40 à 59 ans, dans la nomenclature suivante : 1- Paysan (agriculteur exploitant ou salarié agricole); 2- Autre.*

On nomme I la variable placée en ligne (position sociale du père) et J la variable placée en colonne (position sociale du fils).

$I \setminus J$	*1-Paysan*	*2-Autre*	*Total*
1-Paysan	*657*	*447*	*1104*
2-Autre	*73*	*1370*	*1443*
Total	*730*	*1817*	*2547*

1. *Quelle est leur nature des variables ?*
2. *On effectue un test d'indépendance de Chi2 entre les deux variables.*
   (a) *Préciser les hypothèse nulle et alternative du test.*
   (b) *Donner le tableau des effectifs théoriques.*
   (c) *Donner les conditions d'application du test. Sont-elles vérifiées ?*
   (d) *Donner la statistique du Chi2 test et sa loi sous l'hypothèse nulle. Donner la valeur observée de la statistique Chi2.*
   (e) *Énoncer la règle de décision du test.*
   (f) *La p-valeur associée au test donnée par le logiciel* `R` *est* `p-value < 2.2e-16`*. Que pouvez-vous conclure au risque 5% ?*

Exercice 2 *On intérroge* 1873 *étudiants de M2 sur la catégorie socio-professionnelle de leur parents. Les étudiants suivent différents cursus: écoles d'ingénieurs, écoles de commerce, universités scientifiques, médecine. Les résultats sont les suivants :*

	Ouvriers	*Employés*	*Cadres*	*Professions libérales*
Ecoles d'ingénieurs	*50*	*280*	*120*	*20*
Ecoles de commerce	*8*	*29*	*210*	*350*
Universités Scientifiques	*150*	*230*	*100*	*40*
Médecine	*26*	*80*	*80*	*100*

On veut étudier l'influence du milieu socio-professionnel des parents sur le type d'étude des enfants.

1. *Quelles sont les variables étudiées ? Quelle est leur nature ?*

2. *On effectue un test d'indépendance du Chi2 entre les deux variables*

   (a) *Préciser les hypothèse nulle et alternative du test.*

   (b) *Donner le tableau des effectifs théoriques.*

   (c) *Donner les conditions d'application du test. Sont-elles vérifiées ?*

   (d) *Donner la statistique du Chi2 test et sa loi sous l'hypothèse nulle.*

   (e) *Vérifier que la valeur observée de la statistique Chi2 vaut 853,26.*

   (f) *Énoncer la règle de décision du test.*

   (g) *La p-valeur associée au test donnée par le logiciel* `R` *est* `p-value < 2.2e-16`*. Que pouvez-vous conclure au risque 5% ?*

Exercice 3 *Les auteurs d'une étude sur la gestion des ressources humaines dans les entreprises réunionnaises ont constitué un échantillon de 136 entreprises de l'île. Pour chaque entreprise, ils ont relevé la présence ou l'absence d'un DRH et la taille de l'entreprise. La répartition des 136 entreprises selon les modalités de ces 2 variables est la suivante :*

	< 50	*50 à 99*	*100 à 249*	*> 249*
avec DRH	*16*	*16*	*18*	*16*
sans DRH	*20*	*21*	*28*	*1*

On veut étudier l'existence d'un lien entre les deux variables.

1. *Définir les variables et leur type.*

2. *On effectue un test d'indépendance du chi 2 entre les deux variables*

   (a) *Préciser les hypothèses nulle et alternative du test.*

   (b) *Donner le tableau des effectifs théoriques.*

   (c) *Donner les conditions d'application du test. Sont-elles vérifiées ?*

   (d) *Donner la statistique du Chi2 test et sa loi sous l'hypothèse nulle.*

   (e) *La p-valeur associée au test donnée par le logiciel* `R` *est* `p-value = 0.0009273`*. Que pouvez-vous conclure au risque 5% ?*

Exercice 4 *Un étude fait par la baraserie Alber de Tucson, en Arizone. Cette brasserie produit trois types de bières : légère, normale et brune. Les données obtenues sont regroupées dans le tableau de contingence ci-dessous.*

	Légère	*Normale*	*Brune*
Homme	*20*	*40*	*20*
Femme	*30*	*30*	*10*

Lors de l'analyse de la segmentation du marché de la bière entre ces trois catégories , le groupe de recherche marketing de la firme s'est demandé si il existe un lien entre la consommation et le sexe.

On effectue un test d'indépendance du chi 2 entre les variables consommation et sexe.

1. *Préciser les hypothèses nulle et alternative du test.*

2. *Utiliser les sorties de* `R` *pour répondre aux questions suivantes :*

   (a) *Donner le tableau des effectifs théoriques. Les conditions d'application sont-elles vérifiées ?*

   (b) *Donner la statistique du Chi2 test et sa loi sous l'hypothèse nulle. Donner la valeur observée de la statistique Chi2. Que pouvez-vous conclure au risque 5% ?*

```
> nij=rbind(c(20,40,20),c(30,30,10))
> test=chisq.test(nij)
> test$expected
         [,1]     [,2] [,3]
[1,] 26.66667 37.33333   16
[2,] 23.33333 32.66667   14
> test
Pearson's Chi-squared test
data:  nij
X-squared = 6.1224, df = 2, p-value = 0.04683
```

Exercice 5 *L'étude sur les abonnés à* The Wall Street Journal *de 1996 a fourni des données sur le status professionel des abonés. Les résultats sont presentés dans le tableau de contingence.*

	Région	
Statut professionel	*Edition de l'Est*	*Edition de l'Oest*
Employé à plein temps	*1105*	*574*
Employé à temps-partiel	*31*	*15*
Profession libérale	*229*	*186*
Sans emploie	*485*	*344*

1. *Définir les variables et leur type.*

2. *On effectue un test d'indépendance du chi 2 entre les deux variables considérées.*

   (a) *Préciser les hypothèses nulle et alternative du test.*

   (b) *Utiliser les sorties de* `R` *pour répondre aux questions suivantes :*

      i. *Donner le tableau des effectifs théoriques. Les conditions d'application sont-elles vérifiées ?*

      ii. *Donner la statistique du Chi2 test et sa loi sous l'hypothèse nulle. Donner la valeur observée de la statistique Chi2. Que pouvez-vous conclure au risque 5% ?*

```
> nij=rbind(c(1105,574),c(31,15),c(229,186),c(485,344));
> test=chisq.test(nij)
> test$expected
           [,1]      [,2]
[1,] 1046.19400 632.80600
[2,]   28.66285  17.33715
[3,]  258.58875 156.41125
[4,]  516.55440 312.44560
```

```
> test
Pearson's Chi-squared test
data:  nij
X-squared = 23.373, df = 3, p-value = 3.376e-05
```

Exercice 6 *Une étude a été réalisée sur 100 patients d'un service hospitalier afin de vérifier la relation entre le tabac et les problèmes pulmonaires. Pour cela, nous avons demandé à chaque personne son âge, son sexe, sa situation familiale (célibataire, mariée,...), sa consommation de tabac (nombre de cigarettes par jour), la présence de tabagisme passif, et la présence de problème pulmonaire (par exemple cancer du poumon ou broncho-pneumopathie chronique obstructive) chez cette personne. Les données sont enregistrées dans le fichier*

Exercice 7 *Les données correspondant à* 30 *observations de deux variables qualitatives x et Y son présentées ci-dessous. Les catégories pour X sont a, b et c. Les catégories pour Y sont* 1 *et* 2*.*

observation	X	Y	*observation*	X	Y
1	*a*	*1*	*16*	*b*	*2*
2	*b*	*1*	*17*	*c*	*1*
3	*b*	*1*	*18*	*b*	*1*
4	*c*	*2*	*19*	*c*	*1*
5	*b*	*1*	*20*	*b*	*1*
6	*c*	*2*	*21*	*c*	*2*
7	*b*	*1*	*22*	*b*	*1*
8	*c*	*2*	*23*	*c*	*2*
9	*a*	*1*	*24*	*a*	*1*
10	*b*	*1*	*25*	*b*	*1*
11	*a*	*1*	*26*	*c*	*2*
12	*b*	*1*	*27*	*c*	*2*
13	*c*	*2*	*28*	*a*	*1*
14	*c*	*2*	*29*	*b*	*1*
15	*c*	*2*	*30*	*b*	*2*

1. *Effectuez une tabulation croisée pour les données en utilisant X en ligne et Y en colonne.*
2. *Y a-t-il une relation entre les variables ? Utilisez un seuil de signification de 5%.*

Exercice 8 *Pour étudier la mobilité sociale et son évolution, on va s'intéresser à des tableaux d'enquête croisant la position sociale d'individus adultes (fils) avec celle de leur père au cours de leur jeunesses. Le tableau (à trois dimensions) contient $2 \times 2 \times 2$ cellules.*

	Date T1			*Date T2*	
Origine\ Position	*Cadre*	*Ouvrier*	*Origine \ Position*	*Cadre*	*Ouvrier*
Cadre	*125*	*75*	*Cadre*	*150*	*50*
Ouvrier	*125*	*675*	*Ouvrier*	*200*	*600*

1. *Calculer la proportion de fils d'ouvrier qui sont cadres et la proportion de fils d'ouvrier qui sont ouvrier (en t2 et en t1).*
2. *Calculer la proportion de fils de cadre qui sont cadres et la proportion de fils de cadre qui sont ouvrier (en t2 et en t1).*

3. *Calculer les odds-ratio. Que peut-on en conclure ?*

Exercice 9 *On a relevé pour 535 étudiants les valeurs prises par deux variables qualitatives X et Y. La variable X représente l'origine scolaire des étudiants de 1ère année et la variable Y le passage en L2. Pour étudier l'inégalité due à l'origine scolaire, on va s'intéresser à des tableaux d'enquête croisant la variable X à 2 modalités (G = Bac Générale et TP = Bac Technologique ou professionnel) avec la variable Y à 2 modalités (Oui = Passe en 2ème année, Non = Ne passe pas). Les résultats sont présentés dans les tableaux de contingence ci-dessous*

	Sociologie			*Psychologie*	
	Oui	*Non*		*Oui*	*Non*
G	*60*	*30*	*G*	*150*	*100*
TP	*15*	*25*	*TP*	*45*	*90*

1. *Calculer P(Y=Oui|X=G), P(Y=Oui|X=TP), P(Y=Non|X=G) et P(Y=Non|X=TP) (en Sociologie et en Psychologie).*

2. *Calculer le rapport de chances de Oui/Non pour les bacheliers G (en Sociologie et en Psychologie). Calculer le rapport de chances de Oui/Non pour les bacheliers TP (en Sociologie et en Psychologie). Que peut-on en conclure ?*

3. *Calculer l'odds-ratio (en Sociologie et en Psychologie). Quelle interprétation peut-on donner de l'inégalité due à l'origine scolaire ?*

Exercice 10 *Considérons le tableau de contingence suivant, relatant le nombre d'accidents mortels de la circulation et le port de la ceinture de sécurité en 1988 dans un état des Etats-Unis*

	Région	
Accidents	*fatal*	*non fatal*
port de ceinture	*510*	*412368*
pas de port de ceinture	*1601*	*162527*

Notons X la variable qui vaut 1 (Oui) si l'individu portait une ceinture de sécurité et 0 (Non) sinon. La variable réponse Y est la variable dichotomique indiquant un accident mortel (1) ou non mortel (0).

1. *Quel est le rapport de chances des personnes portant une ceinture*

2. *Quel est le rapport de chances des personnes qui ne portant pas de ceinture.*

3. *Est-ce que le sujets qui ne portant pas de ceinture auront plus de chance d'avoir un accident fatal que ceux qui en portant ?*

Exercice 11 *Considérons les données recueillies en 1969 par le sociologue suisse Roger Girod auprès d'un échantillon d'hommes du canton de Genève : I désigne l'origine sociale (donnée par la profession du père); J désigne la position sociale à l'entrée dans la vie active; K désigne la position sociale à l'enquête (en 1969). Chaque variable utilise la nomenclature suivante: 1- Manuel; 0- Non manuel.*

Le tableau (à trois dimensions) contient $2 \times 2 \times 2$ *cellules.*

	I=1			*I=2*	
$J \setminus K$	*1*	*0*	$J \setminus K$	*1*	*0*
1	*92*	*26*	*1*	*57*	51
0	*5*	*50*	*2*	*3*	*154*

Nous voulons étudier les relations entre les variables : origine sociale, I, position sociale à l'entrée dans la vie active, J, et position sociale à l'enquête K.

Calculer les odds-ratio pour chaque tableau de contingence. Que peut on en conclure ?

Exercice 12 *Dellatolas et collaborateurs (1997), dans une étude portant sur 1155 enfants âgés de 2 ans et demi à 6 ans (560 garçons et 595 filles), ont constaté que 78 garçons et 52 filles étaient gauchers. Les ensembles de modalités des deux variables* ***Sexe*** *et* ***Latéralité manuelle*** *sont notées* $\{M, F\} = \{0, 1\}$ *et* $\{G, D\} = \{0, 1\}$ *respectivement.*

1. *Donner le tableau de contingence croisant la variable* ***Sexe*** *et* ***Latéralité manuelle****.*
2. *On cherche à voir dans quelle mesure les gauchers seraient plus fréquents chez les garçons que chez les filles.*
   (a) *Déterminer le rapport de chances de l'évènement "être gaucher" chez les garçons et le rapport de chances de l'évènement "être gaucher" chez les filles. Commentez.*
   (b) *Calculer le rapport de rapport de chances (****odds ratio****). Peut-on dire que les chances d'être gaucher plutôt que droitier sont plus élevées chez les garçons ? Justifier.*
3. *Calculer l'odds-ratio.*