Université Paris Nanterre
Statistiques L3 Gestion (App)

Exercice 1

Cet exercice porte sur les données de la mortalité routière en Europe. Nous disposons d'un échantillon de taille 27. Les variables étudiées ici sont les suivantes :

— `MortsPM` : Mortalité sur les routes par million selon les données de l'UE.
— `Transp` : Transparence selon *Heritage Foundation.*
— `Alcool` : Taux d'alcoolémie permis par la loi.
— `Nvdemo` : Nouvelle démocratie, Ancienne démocratie.

Il est important remarquer que le terme Transparence utilisé dans la variable ***`Transp`*** *correspond à un indice de perception de la corruption.*

```
MortsPM          Transp          Alcool           NvDemo
Min.   : 29.00   Min.   :3.50   Min.   :0.0000   Ancienne:17
1st Qu.: 49.50   1st Qu.:4.65   1st Qu.:0.2000   Nouvelle:10
Median : 68.00   Median :6.30   Median :0.5000
Mean   : 67.67   Mean   :6.30   Mean   :0.4222
3rd Qu.: 82.50   3rd Qu.:7.90   3rd Qu.:0.5000
Max.   :112.00   Max.   :9.30   Max.   :0.9000
```

1. Relever la valeur de la mortalité sur les routes par million d'habitants en dessous duquel se situent 50% des pays de l'échantillon et la valeur de la mortalité sur les routes par million d'habitats au-dessus duquel se situent 25% des pays de l'échantillon.
2. Commenter **brièvement** le graphique ci-dessous.
3. Nous avons déterminé la matrice de corrélation.

```
          MortsPM  Transp  Alcool
MortsPM     1.000  -0.759 -0.363
Transp     -0.759   1.000  0.420
Alcool     -0.363   0.420  1.000
```

   (a) Pourquoi n'y a-t-il pas la variable `NvDemo` dans le matrice de corrélation ?
   (b) Que peut-on dire de la corrélation linéaire entre `MortsPM` et `Transp` ?

Exercice 2

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

1. On effectue un test d'indépendance de Chi2 entre les deux variables.

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

(b) Comment a été calculé la valeur 1029.4193 du tableau des effectifs théoriques ci-dessous ?

I \ J	1-Paysan	2-Autre
1-Paysan	316.4193	787.5807
2-Autre	413.5807	1029.4193

(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.

```
Pearson's Chi-squared test with Yates' continuity correction
data:  nij
X-squared = 904.35, df = 1, p-value < 2.2e-16
```

(e) Énoncer la règle de décision du test. Que pouvez-vous conclure au risque 5% ?

Exercice 3

Nous utilisons ici le jeu de données hdv2003 extrait de l'enquête **histoire de vie** réalisée par l'INSEE en 2003. Il contient 2000 individus et 20 variables parmi lesquelles : d'une part des variables décrivant les caractéristiques socio-démographiques des individus (age, sexe, nivetud, etc.), et d'autre part des variables décrivant leurs pratiques de loisirs (hard.rock, lecture.bd, peche.chasse, etc.) Cette enquête part du postulat de départ que pour comprendre comment un individu s'intègre dans la société, il faut disposer à la fois d'informations objectives (situation professionnelle, situation familiale, état de santé, centres d'intérêts,etc.) mais aussi tenir compte d'éléments plus subjectifs.

Description des certaines variables :

— age : Age
— hard.rock : Ecoute du Hard rock ou assimilés
— lecture.bd : Lecture de bandes dessinées
— cinema : Cinéma au cours des 12 derniers mois

1. On veut voir si il existe un lien entre les variables cinéma et lecture BD.

Table 1 – Table des valeurs observées

	lecture.bd	
cinema	Non	Oui
Non	1156	18
Oui	797	29

(a) Préciser les hypothèses nulle et alternative du test d'indépendance.

(b) Nous avons utilisé le logiciel R pour calculer avec les données le tableau des effectifs théoriques (voir tableau ci-dessous)

Table 2 – Table des valeurs attendues (effectifs théoriques)

	lecture.bd	
cinema	Non	Oui
Non	1146.411	27.589
Oui	806.589	19.411

Expliquer pourquoi on peut appliquer le test du chi 2 entre ces deux variables ?

(c) Nous avons effectuer le test d'indépendance du chi 2 entre les deux variables (voir les sorties de R ci-dessous). Combien vaut la réalisation de la statistique sur les données ? Que pouvez-vous conclure au seuil $\alpha = 5\%$?

```
Pearson's Chi-squared test with Yates' continuity correction
data:  hdv2003.bis$cinema and hdv2003.bis$lecture.bd
X-squared = 7.4246, df = 1, p-value = 0.006434
```

Exercice 4

La association *HEC Sondages* avait organisé, début 2007, un sondage pour les élections présidentielles. 286 étudiants de première, deuxième et troisième année on été interrogés et le tableau obtenu est le suivant :

TABLE 3 – Table des valeurs observées

	Année		
Intention de vote	1A	2A	3A
Royal	9	8	6
Sarkozy	38	36	76
Autre	14	22	0
Indécis	29	26	14
NSPP	3	5	0

1. Définir la population étudiée. Quelles sont les variables étudiées et quelle est leur nature ?
2. Préciser les hypothèses nulle et alternative du test.
3. Donner les conditions d'application du test.
4. Calculer à la main les effectifs théoriques e_{51}, e_{52} et e_{53} du tableau ci-dessus. Est-ce qu'on peut appliquer le test d'indépendance du chi 2 entre les deux variables ? Justifier votre réponse.

TABLE 4 – Table des valeurs attendues (effectifs théoriques)

	Année		
Intention de vote	1A	2A	3A
Royal	7.479	7.801	7.720
Sarkozy	48.776	50.874	50.350
Autre	11.706	12.210	12.084
Indécis	22.437	23.402	23.161
NSPP	e_{51}	e_{52}	e_{53}

5. Nous avons procédé à un regroupement. Nous avons regroupé ceux qui ne se prononcent pas avec les indécis et le tableau obtenu est le suivante :

   (a) Les conditions d'application du test d'indépendance du chi 2 sont-elles vérifiées ?

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

TABLE 5 – Table des valeurs observées

	Année		
Intention de vote	1A	2A	3A
Royal	9	8	6
Sarkozy	38	36	76
Autre	14	22	0
Indécis ou NSPP	32	31	14

(c) Nous avons utilisé le logiciel R pour calculer avec les données le tableau des effectifs théoriques et pour effectuer le test d'indépendance du chi 2 entre les deux variables (voir les sorties de R ci-dessous). Combien vaut la réalisation de la statistique sur les données ? Que pouvez-vous conclure au seuil $\alpha = 5\%$?

```
1A     2A     3A
Royal             7.479  7.801  7.720
Sarkozy          48.776 50.874 50.350
Autre            11.706 12.210 12.084
Indécis ou NSPP 25.038 26.115 25.846

Pearson's Chi-squared test

data:  nij
X-squared = 49.157, df = 6, p-value = 6.936e-09
```

Exercice 5

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 6

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% ?