Université de Perpignan - IUT de Carcassonne
Nathalie VILLA-VIALANEIX
Année 2013/2014 - Durée : 2 heures
Ce sujet comporte 6 pages

DUT STID, 1ème année
Statistique descriptive II
Devoir du mercredi 18 décembre 2013

Nom : _____ /34

Consignes
- Les réponses sont à donner directement sur le sujet. N'oubliez pas de noter votre nom.
- Toute réponse doit être précisément justifiée. Les réponses insuffisamment justifiées ne donneront droit à aucun point.
- *Matériel autorisé* (à l'exclusion de toute autre chose) : crayons, calculatrices (pas d'ordinateur, pas de téléphone portable), cerveau (pour ceux qui en possèdent un). **Les téléphones portables sont formellement interdits sur les tables, sur vos genoux, dans vos poches : ils doivent être déposés, avec vos sacs, à côté de mon bureau.**
- Les deux exercices sont indépendants ainsi que la plupart des questions à l'intérieur des exercices.
- Il est formellement interdit de parler (même en language des signes et même pour demander une gomme, un crayon, etc à son voisin).

Exercice 1 Titanic ..... /10

Cet exercice a été conçu pour conjurer le mauvais sort, afin que ce devoir ne soit pas un naufrage. Le jeu de données utilisé pour cet exercice fournit des informations sur le sort des passagers du Titanic (sexe, âge, classe de voyage et survie) et provient de : « Dawson, Robert J. MacG. (1995) The 'Unusual Episode' Data Revisited. *Journal of Statistics Education*, **3**. http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html ». Nous nous focalisons ici sur les variables « classe de voyage » et « survie » et étudions la relation existant entre ces deux variables. Les données sont données dans la table de contingence ci-dessous :

Survie :	Non	Oui	Total
1ère classe	122	203	325
2ème classe	167	118	285
3ème classe	528	178	706
Équipage	673	212	885
Total	1 490	711	2 201

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs types ?
   Réponse :
   La population étudiée est l'ensemble des passagers du Titanic, de taille $N = 2\,201$ ($= 122 + 203 + ... + 212$). Les variables étudiées sont la variable « Survie » de type qualitative nominale et la variable « Classe de voyage » de type qualitative ordinale. ..... /1
2. Compléter, dans le tableau ci-dessus, **en bleu** la distribution marginale de la variable « survie » et en **en rouge** la distribution marginale de la variable « classe de voyage ».
   Réponse :
   . ..... /1

3. Doit-on utiliser
   - ☑ La distribution de « survie » conditionnellement à « classe de voyage » ;
   - ☐ La distribution de « classe de voyage » conditionnellement à « survie ».

   si on veut étudier les différences de taux de survie parmi les différentes classes de passagers ?
   Réponse :

. . . . . /1

4. Donner, dans le tableau ci-dessous, la distribution conditionnelle choisie à la question précédente.

Survie :	Non	Oui	
1ère classe	37,5%*	62,5%	100%
2ème classe	58,6%	41,4%	100%
3ème classe	74,8%	25,2%	100%
Équipage	76,0%	24,0%	100%

   Commenter les résultats obtenus.
   Réponse :

* $\frac{122}{325} \simeq 37{,}5\%$ . . . . . /1
Le taux de survie diminue nettement avec la classe du passager : les passagers en 3ème classe ou l'équipage ont un taux de survie environ deux fois moindre que les passagers de 1ère classe. . . . /1

5. Calculer les effectifs théoriques d'indépendance et les contributions au χ^2.
   Quelle paire de modalités contribue le plus au χ^2 ? Est-elle sur/sous-représentée ? Interpréter.
   Réponse :

Les effectifs théoriques d'indépendance sont :

Survie :	Non	Oui
1ère classe	$\frac{325 \times 1\ 490}{2\ 201} \simeq 220{,}01$	104,99
2ème classe	192,94	92,06
3ème classe	477,94	228,06
Équipage	599,11	285,89

. . . . . /1
Les contributions au χ^2 sont

Survie :	Non	Oui
1ère classe	$\frac{(220{,}01-122)^2}{220{,}01} \simeq 43{,}66$	91,50
2ème classe	3,49	7,31
3ème classe	5,24	10,99
Équipage	9,11	19,10

. . . . . /1
La paire de modalités qui contribue le plus au χ^2 correspond aux passagers de 1ère classe qui ont n'ont pas survécu. Cette paire de modalités est très sous-représentée (l'effectif théorique d'indépendance est pratiquement deux fois plus grand que l'effectif observé). Cela signifie que les passagers de première classe qui n'ont pas survécu sont particulièrement rares. . . . . /1,5

6. Calculer le χ^2 puis le C de Cramer. Interpréter.
   Réponse :

$\chi^2 = 43{,}66 + 91{,}50 + ... + 19{,}10 \simeq 190{,}40$ donc $C = \sqrt{\frac{190{,}40}{2\ 201 \times 1}} \simeq 0{,}294$. . . . . /1
La relation constatée entre survie et classe de voyage est donc plutôt faible. . . . . /0,5

Espace supplémentaire (au besoin)

Exercice 2 Qualité de l'air ..... /11

Les données utilisées pour cet exercice sont extraites de « Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983) *Graphical Methods for Data Analysis*. Belmont, CA : Wadsworth ». Elles donnent les statistiques de qualité de l'air à New York entre mai et septembre 1973. Une observation correspond à des mesures effectuées sur un jour de l'année. Nous nous intéresserons ici à deux variables de ce jeu de données :

- le mois de l'année (mai, juin, juillet, août ou septembre) dans lequel se situe la mesure ;
- le taux d'ozone.

Les données sont analysées avec R. Sous R, le fichier de données porte le nom de `airquality` et les variables sont respectivement nommées `Month` et `Ozone`. Les commandes R effectuées ainsi que les résultats numériques sont données ci-dessous :

```
table(airquality$Month)
# mai       juin    juillet       aout  septembre
# 24          9         26         23         29
```

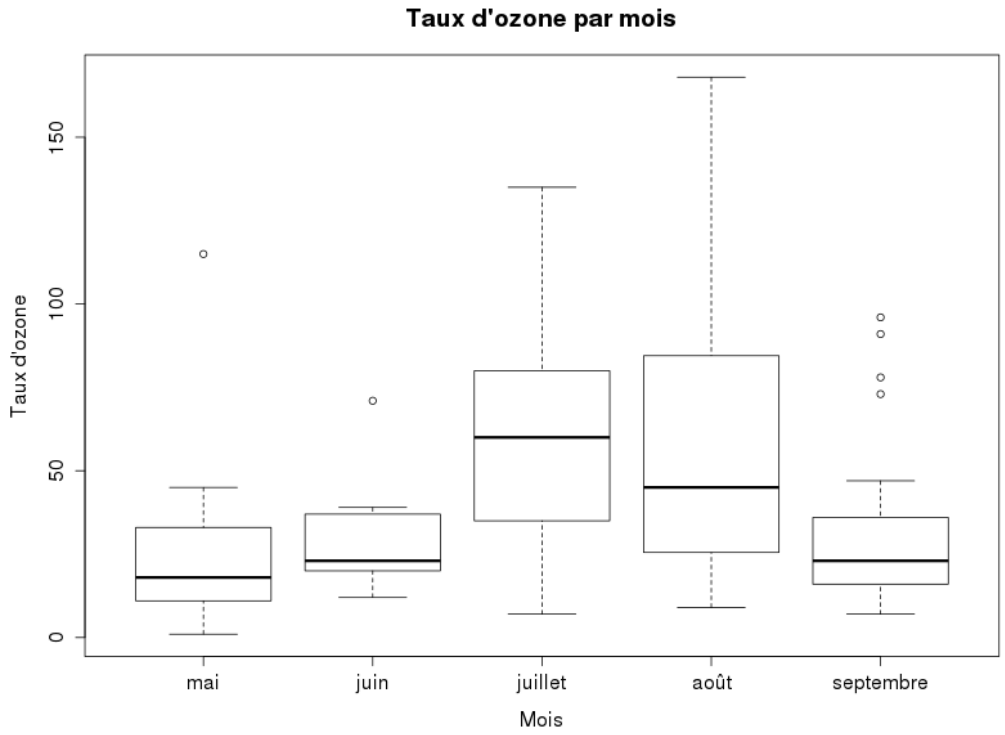
```
by(airquality$Ozone, airquality$Month, mean)
#   airquality$Month: mai
# [1] 24.125
# ------------------------------------------------------------
#   airquality$Month: juin
# [1] 29.44444
# ------------------------------------------------------------
#   airquality$Month: juillet
# [1] 59.11538
# ------------------------------------------------------------
#   airquality$Month: aout
# [1] 60
# ------------------------------------------------------------
#   airquality$Month: septembre
# [1] 31.44828
```

```
by(airquality$Ozone, airquality$Month, var)
#   airquality$Month: mai
```

```
# [1] 523.7663
# ------------------------------------------------------------
#   airquality$Month: juin
# [1] 331.5278
# ------------------------------------------------------------
#   airquality$Month: juillet
# [1] 1000.826
# ------------------------------------------------------------
#   airquality$Month: aout
# [1] 1744.545
# ------------------------------------------------------------
#   airquality$Month: septembre
# [1] 582.8276
```

À partir de ces informations, répondre aux questions suivantes :

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs types ?
   Réponse :
   La population étudiée est l'ensemble des jours sur lesquels des mesures ont été effectuées. Sa taille est $N = 24 + 9 + \ldots + 29 = 111$. Les variables étudiées sont le mois du jour de mesure, de type qualitatif nominal, et le taux d'ozone, de type quantitatif continu. . . . . /1
2. Expliquer ce qui est calculé dans chacune des trois commandes R fournies ci-dessus.
   Réponse :
   La première commande calcule les effectifs marginaux de la variable « mois ». La deuxième commande calcule les moyennes de la variable « ozone » conditionnellement au mois. La troisième commande calcule les variances de la variable « ozone » conditionnellement au mois. . . . . /1,5
3. Entourer toutes les commandes R qui permettent d'obtenir le graphique ci-dessous :



☑

```
plot(Ozone~Month, data=airquality, main="Taux d'ozone par mois",
     xlab="Mois", ylab="Taux d'ozone")
```

☑

```
plot(airquality$Month, airquality$Ozone, main="Taux d'ozone par mois",
```

```
     xlab="Mois", ylab="Taux d'ozone")

plot(Month~Ozone, data=airquality, main="Taux d'ozone par mois",
     xlab="Mois", ylab="Taux d'ozone")

plot(airquality$Ozone, airquality$Month, main="Taux d'ozone par mois",
     xlab="Mois", ylab="Taux d'ozone")
```

. . . . . /1

Commenter ce graphique.

Réponse :

Le graphique montre un taux d'ozone beaucoup plus élevé et surtout beaucoup plus variable durant les mois d'été (juillet et août). . . . . /0,5

4. Quel est le taux d'ozone moyen sur la population ?

Réponse :

Le taux d'ozone moyen sur la population se calcule en faisant la moyenne du taux d'ozone moyen par mois, pondérée par l'effectif du mois correspondant :

$$\overline{X} = \frac{24 \times 24{,}125 + \ldots + 29 \times 31.44828}{111} \simeq 42{,}10$$

. . . . . /1,5

5. Quelle est la variance intra-classes (les classes sont les mois) du taux d'ozone ?

Réponse :

La variance intra-classes se calcule en faisant la moyenne des variances par mois du taux d'ozone, pondére par l'effectif du mois :

$$\mathrm{Var}_{\mathrm{intra}} = \frac{24 \times 523.7663 + \ldots + 29 \times 582.8276}{111} \simeq 888{,}31$$

. . . . . /1,5

6. Quelle est la variance inter-classes du taux d'ozone ?

Réponse :

La variance inter-classes se calcule en faisant la variance des taux d'ozone moyens par mois, pondérée par l'effectif du mois :

$$\mathrm{Var}_{\mathrm{inter}} = \frac{24 \times 24{,}125^2 + \ldots + 29 \times 31.44828^2}{111} - 42{,}10^2 \simeq 246{,}70$$

. . . . . /2

7. Calculer la variance globale du taux d'ozone.

Réponse :

La variance globale du taux d'ozone est :

$$\mathrm{Var}(\texttt{Ozone}) = 246{,}70 + 888{,}31 = 1\ 135{,}01$$

. . . . . /0,5

8. Calculer le rapport de corrélation entre taux d'ozone et mois. Interpréter cette valeur.

Réponse :

Le rapport de corrélation est égal à

$$\eta(\texttt{Ozone}|\texttt{Month}) = \sqrt{\frac{\mathrm{Var}_{\mathrm{inter}}}{\mathrm{Var}_{\mathrm{inter}} + \mathrm{Var}_{\ \mathrm{intra}}}} \simeq 0{,}466$$

. . . . . /1

Le taux d'ozone dépend de manière relativement marquée du mois de mesure. . . . . /0,5

Espace supplémentaire (au besoin)