Université de Perpignan - IUT de Carcassonne
Nathalie VILLA-VIALANEIX
Année 2013/2014 - Durée : 2 heures
Ce sujet comporte 8 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	
2ème classe	167	118	
3ème classe	528	178	
Équipage	673	212	
Total			

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs

types ?

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 ».

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 ?

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

Survie :	Non	Oui	
1ère classe 2ème classe 3ème classe Équipage			

Commenter les résultats obtenus.

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.

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

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
```

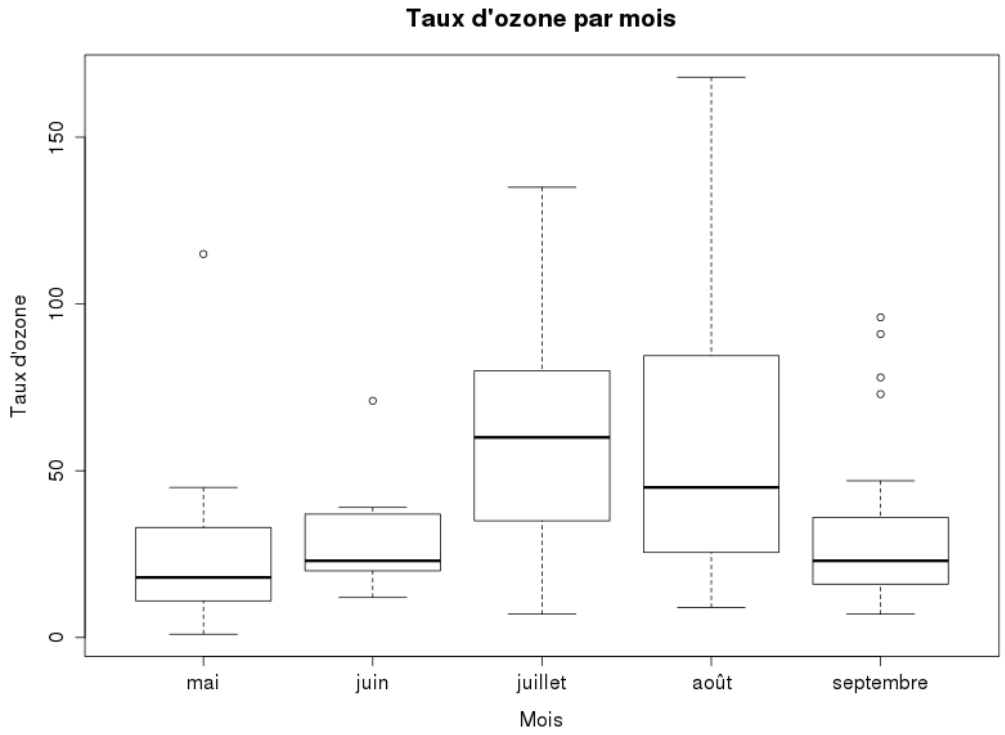
```
by(airquality$Ozone, airquality$Month, var)
#   airquality$Month: mai
# [1] 523.7663
# ------------------------------------------------------------
#   airquality$Month: juin
# [1] 331.5278
# ------------------------------------------------------------
#   airquality$Month: juillet
# [1] 1000.826
# ------------------------------------------------------------
#   airquality$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 ?

2. Expliquer ce qui est calculé dans chacune des trois commandes R fournies ci-dessus.

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")
```

Commenter ce graphique.

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

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

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

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

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

Espace supplémentaire (au besoin)