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

DUT STID, 1ème année
Statistique descriptive I
Devoir du mercredi 6 novembre 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 ..... /17

Cet exercice utilise les données `precip` extraites du livre : McNeil, D.R. (1977) *Interactive Data Analysis*. New York : Wiley. Elles contiennent les précipitations annuelles, en pouces, de 70 villes des États-Unis :

Mobile : 67,0 ; Juneau : 54,7 ; Phoenix : 7,0 ; Little Rock : 48,5...

(pour simplifier, seules les premières observations sont reproduites dans l'énoncé). On donne :
- le total des précipitations sur l'ensemble des 70 villes est 2 442 pouces ;
- la somme des carrés des précipitations par ville est 98 154,1 pouces^2.

La plupart des questions suivantes sont indépendantes :

1. Quelle est la population étudiée, quelle est sa taille ?
   Réponse :
   La population étudiée est constitutée des 70 villes américaines du fichier de données ; sa taille est donc 70. ..... /1
2. Quelle est la variable étudiée, quel est son type ?
   Réponse :
   La variable étudiée est le total des précipitations annuelles de la ville (en pouces), de type quantitative continue. ..... /1
3. Quelle est la moyenne des précipitations sur ces 70 villes ?
   Réponse :
   La précipitation moyenne est égale à :
   $$\overline{X} = \frac{2\ 442}{70} \simeq 34{,}89 \text{ pouces.}$$
   ..... /1

4. Quel est l'écart type des précipitations sur ces 70 villes ?
Réponse :

La variance des précipitations est égale à :

$$\mathrm{Var}(X) = \frac{98\ 154{,}1}{70} - 34{,}89^2 \simeq 184{,}8465$$

et l'écart type des précipitations est donc égal à :

$$\sigma(X) = \sqrt{184{,}8465} \simeq 13{,}60 \text{ pouces.}$$

. . . . . /2

Dans la suite de l'exercice, on utilisera le regroupement en classes suivant des données :

précipitation	[0,20[	[20,30[	[30,35[	[35,40[	[40,45[	[45,70[
effectifs	13	5	11	14	13	14

5. À partir du regroupement en classes, recalculer la moyenne des précipitations sur les 70 villes. Comparer à la valeur trouvée dans la question **??** : que peut-on en conclure ?
Réponse :

La moyenne des précipitations selon le regroupement en classes est égale à :

$$\overline{X} = \frac{10 \times 13 + 25 \times 5 + 32{,}5 \times 11 + 37{,}5 \times 14 + 42{,}5 \times 13 + 57{,}5 \times 14}{70} = \frac{2\ 495}{70} \simeq 35{,}64 \text{ pouces.}$$

. . . . . /1

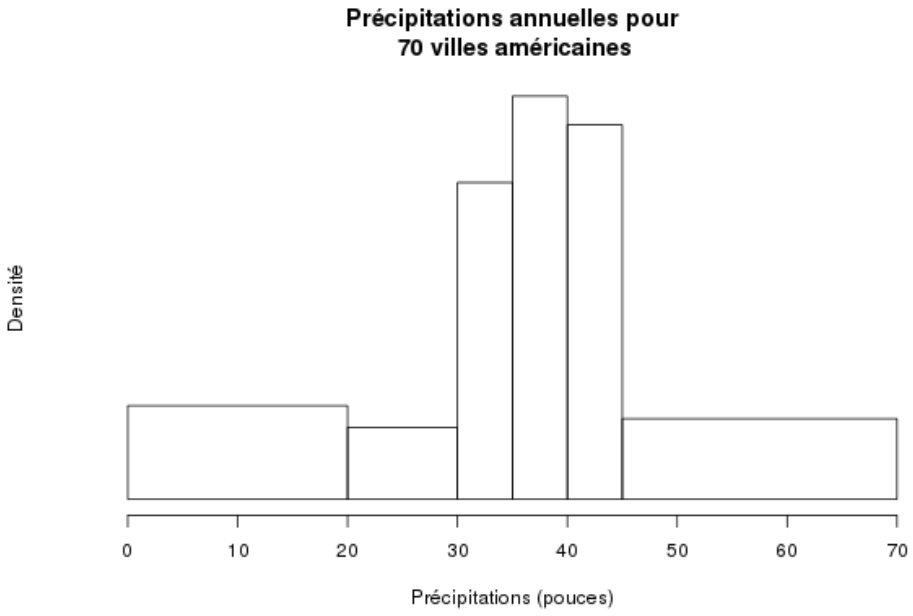
La moyenne trouvée à partir du regroupement en classes est légèrement plus élevée que la vraie moyenne : le regroupement en classes conduit à une perte d'information sur les données. . . . . /0,5

6. Sur le quadrillage suivant, construire l'histogramme des données selon le regroupement en classes données précédemment. *On n'oubliera pas de donner les détails de la construction.*

Réponse : Les densités des différentes classes sont données ici :

précipitation	[0,20[	[20,30[	[30,35[	[35,40[	[40,45[	[45,70[
densité	$1320 = 0{,}65$	0,50	2,20	2,80	2,60	0,56

d'où l'on déduit l'histogramme ci-dessous :



. . . . . /2,5

7. À partir du regroupement en classes, calculer la médiane. Comparer la valeur de celle-ci à la moyenne et commenter.

Réponse :

Le tableau des effectifs cumulés est donné ci-dessous :

précipitation	[0,20[	[20,30[	[30,35[	[35,40[	[40,45[	[45,70[
eff. cumulés	13	18	29	43	56	70

$\frac{N}{2} = 35$ donc la médiane est située dans la classe [35,40[. Sa valeur est :

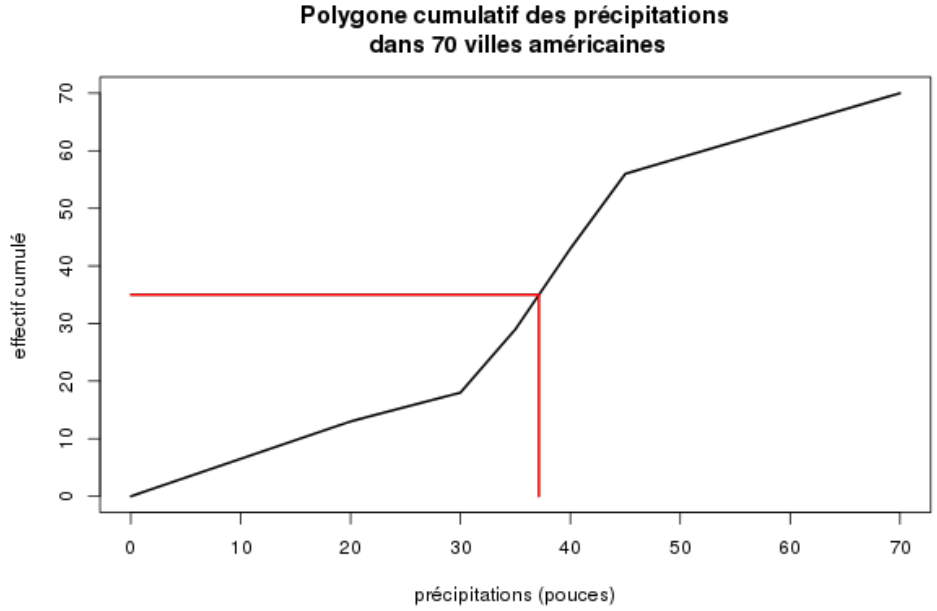
$$m = 35 + (40 - 35) \times \frac{35 - 29}{43 - 29} \simeq 37{,}14 \text{ pouces.}$$

. . . . . /2

La médiane est très similaire à la moyenne : la distribution de la distribution ne semble pas présenter d'étalement particulier. Elle est relativement symétrique (comme l'histogramme l'a déjà montré).

. . . . . /0,5

8. Sur le quadrillage suivant, construire le polygone cumulatif et retrouver graphiquement la valeur de la médiane. *On fera clairement apparaître les traits de construction et où la valeur de la médiane est lue.* **Réponse** : D'après le tableau d'effectifs cumulé de la question précédente, le polygone cumulatif est donné dans le graphique ci-dessous :



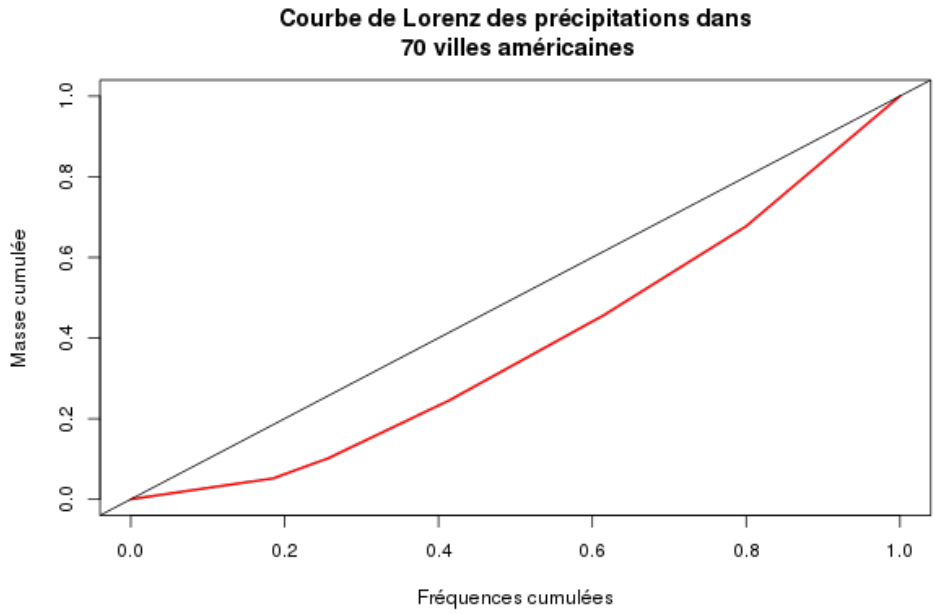
. . . . . /2

La médiane est l'abscisse du point du polygone cumulatif dont l'ordonnée est $N/2 = 35$. On retrouve bien la valeur de la médiane trouvée précédemment, soit environ 37 pouces. . . . . /0,5

9. Déterminer la courbe de Lorenz de cette distribution statistique sur le quadrillage ci-dessous. Commenter son allure. *On n'oubliera pas de donner les détails de la construction.* **Réponse** : Les points de la courbe de Lorenz ont pour coordonnées (f_i^*, v_i) donnés dans le tableau ci-dessous :

précipitation	[0,20[	[20,30[	[30,35[	[35,40[	[40,45[	[45,70[
f_i^*	$\frac{13}{70} \simeq 0{,}186$	$\frac{18}{70} \simeq 0{,}257$	0,414	0,614	0,800	1
v_i	$\frac{10\times13}{2\ 495} \simeq 0{,}052$	$\frac{10\times13+25\times5}{2\ 495} \simeq 0{,}102$	0,245	0,456	0,677	1

La courbe de Lorenz est donc donnée dans la figure ci-dessous :



. . . . . /2,5

La courbe de Lorenz est proche de la diagonale : la répartition des précipitations sur les différentes villes américaines est plutôt égalitaire ; il n'existe pas de concentration forte des précipitations sur un faible nombre de villes. . . . . /0,5

10. Compléter la phrase suivante : 1/5 des villes américaines les plus pluvieuses totalisent **environ 33%** des précipitations des 70 villes étudiées.

Réponse :

. . . . . /1

Espace supplémentaire (au besoin)

Exercice 2 ..... /17

La plupart des questions de cet exercices sont indépendantes.

1. Entourer la commande permettant d'importer dans R le fichier de données au format texte partiellement reproduit ci-dessous :

```
"Nom  ","Gout  "," Amertume  ","Soif  ","TxAlcool  "," Fermen
"""109""",3.0,0.0,1.5,6.0,"Haute ","Ambrée","France","Non","S
"1356 Jean le Bon ",2.5,0.5,1.0,5.6,"Haute ","Blonde","France
"1664 Gold ",1.0,1.0,2.0,6.1,"Basse ","Blonde","France","Non",
"""1845""",3.0,1.5,0.5,6.3,"Haute ","Autre","Iles britanniques
"732 Charles Martel ",2.5,0.5,1.0,5.6,"Haute ","Blonde","Franc
"A.K. Damm ",2.0,0.5,2.0,4.8,"Basse ","Blonde","Espagne","Non
"Abbaye d'Aulnes Blonde 6 ",1.5,1.0,1.5,6.0,"Haute ","Blonde",
"Abbaye d'Aulnes Blonde 8 ",2.0,1.0,0.5,8.0,"Haute ","Blonde",
"Abbaye d'Aulnes Brune 6 ",3.0,1.0,0.5,6.0,"Haute ","Brune","E
"Abbaye d'Aulnes Brune 8 ",3.5,0.5,0.5,8.0,"Haute ","Brune","E
"Abbaye d'Aulnes Val de Sambre ",2.5,1.0,0.5,7.0,"Haute ","Amb
"Abbaye de Gembloux ",1.5,0.5,0.5,8.0,"Haute ","Blonde","Belgi
"Abbaye de St Landelin Spéciale ",2.0,0.0,0.5,6.8,"Haute ","Ar
"Abbaye des Rocs Brune ",4.0,0.5,0.0,9.0,"Haute ","Brune","Be
"Abbaye des Rocs Grand Cru ",4.0,1.0,0.5,10.0,"Haute ","Brune
"Abbaye des Rocs Spéciale ",4.0,1.0,0.0,9.0,"Haute ","Brune","
"Abbaye du Val Dieu Grand Cru (l') ",4.0,1.0,0.0,10.5,"Haute
"Abbaye du Val Dieu (l') ",3.5,0.0,2.0,7.0,"Haute ","Blonde","
"Abdis Brune ",2.0,0.5,0.0,6.5,"Haute ","Brune","Belgique","D
```

Réponse :

```
donnees <- read.table("dataset.txt", header=TRUE, sep=",", dec=".")
```

..... /1

La suite des questions porte sur le fichier de données `CO2` qui est disponible dans R en tapant `data(CO2)`. Il n'est pas nécessaire de savoir ce que contient ce fichier de données pour répondre aux questions. La commande `summary(CO2)` donne

```
    Plant             Type         Treatment       conc          uptake
 Qn1    : 7   Quebec     :42   nonchilled:42   Min.   :  95   Min.   : 7.70
 Qn2    : 7   Mississippi:42   chilled   :42   1st Qu.: 175   1st Qu.:17.90
 Qn3    : 7                                    Median : 350   Median :28.30
 Qc1    : 7                                    Mean   : 435   Mean   :27.21
 Qc3    : 7                                    3rd Qu.: 675   3rd Qu.:37.12
 Qc2    : 7                                    Max.   :1000   Max.   :45.50
 (Other):42
```

2. Quel est le type de chacune des variables du fichier de données (selon la typologie des variables vue en cours). *Remarque : lorsqu'une ambiguïté sur le sous-type de la variable existe, on ne donnera pas ce sous-type.*

   Réponse :

   La variable `Plant` est qualitative, la variable `Type` est qualitative nominale, la variable `Treatment` est qualitative nominale, la variable `conc` est quantitative et la variable `uptake` est quantitative continue. ..... /2

3. Quelle est la taille de la population de ce fichier de données ?

   Réponse :

   $N = 42 + 42 = 84$. ..... /1

4. Quelle valeur renvoie la commande `quantile(CO2$conc,0.75)` ?

   Réponse :

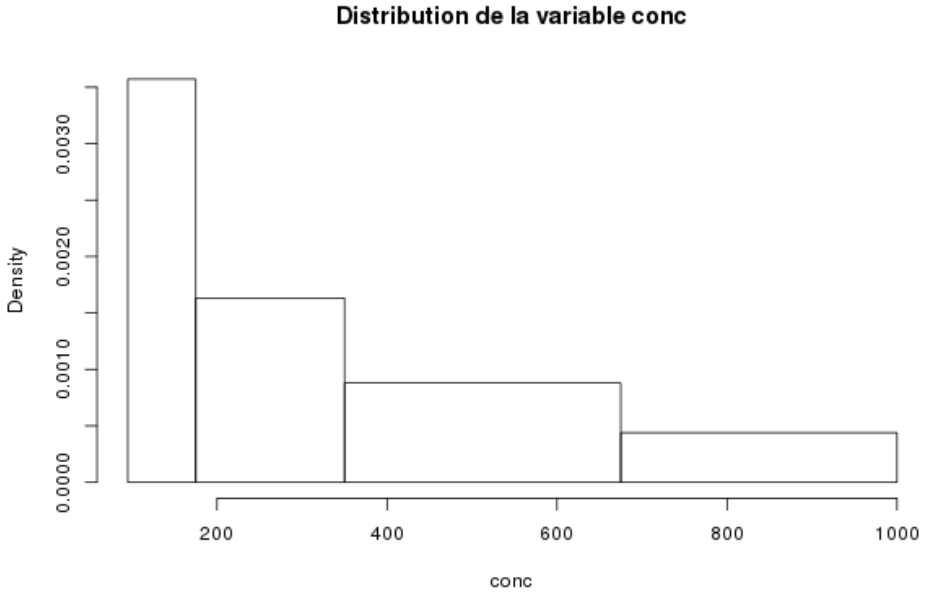
   675 ..... /1

5. Quelle est la moyenne de la variable `conc` ? Quelle est sa médiane ? Commenter la différence.

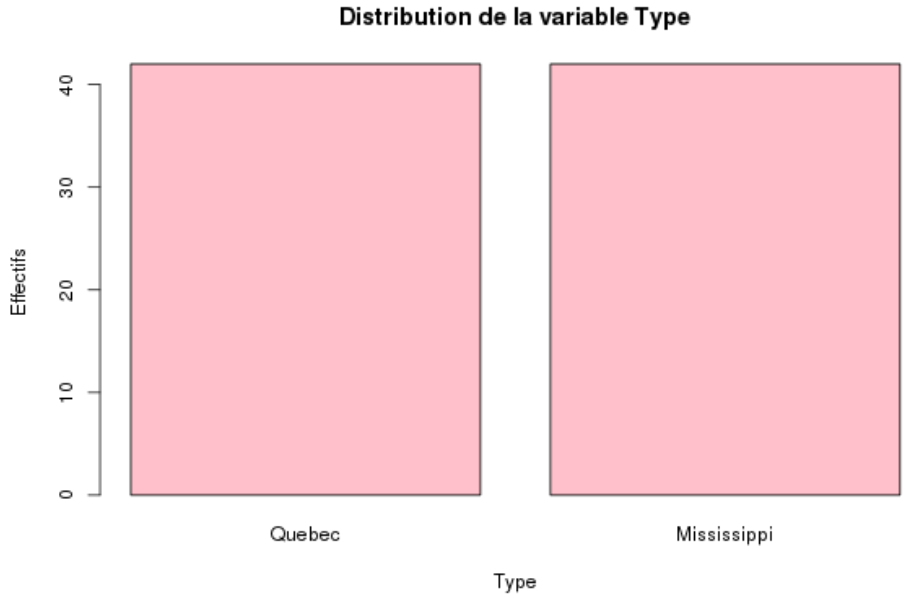
   Réponse :

   La moyenne vaut 435, la médiane vaut 350. La moyenne est sensiblement supérieure à la médiane : la distribution doit présenter des valeurs atypiques vers les fortes valeurs ou un étalement vers la droite. ..... /2

6. Quelle commande permet d'obtenir l'écart type de la variable `uptake` ?
   Réponse :
   ```
   sd(CO2$uptake)
   ```
   ..... /1
7. Quelle commande permet d'otenir le tableau d'effectifs complet de la variable `Plant` ?
   Réponse :
   ```
   table(CO2$Plant)
   ```
   ..... /1
8. Quelles sont les commandes qui permettent de stocker dans une nouvelle variable du jeu de données, que l'on nommera `uptakeC`, le découpage en classes de la variable `uptake` délimitées par les bornes 7, 20, 30, 40 et 50 puis d'obtenir le tableau d'effectifs de ces classes ?
   Réponse :
   ```
   CO2$uptakeC <- cut(CO2$uptake,breaks=c(7,20,30,40,50))
   table(CO2$uptakeC)
   ```
   ..... /2
9. Quelle commande permet de stocker dans une nouvelle variable du jeu de donnée, que l'on nommera `uptakeC2`, le découpage en 5 classes de même amplitudes de la variable `uptake` ?
   Réponse :
   ```
   CO2$uptakeC2 <- cut(CO2$uptake,breaks=5)
   ```
   ..... /1
10. L'histogramme ci-dessous est celui de la variable `conc`.



    Quelle commande permet de l'obtenir ? (Les bornes des classes sont 95, 175, 350, 675 et 1 000.) Commenter la forme de la distribution.
    Réponse :
    ```
    hist(CO2$conc, main="Distribution de la variable conc", xlab="conc",
    ylab="Density", breaks=c(95,175,350,675,1000), freq=FALSE)
    ```
    ..... /2
    La distribution de la variable `conc` est très disymétrique avec un fort étalement vers la droite et une concentration des individus vers les faibles valeur de la variable. ..... /1
11. Le diagramme en tuyau d'orgues ci-dessous est celui de la variable `Type`.



Quelle commande permet de l'obtenir ? (La couleur des tuyaux d'orgue est le rose.)
Réponse :

```
barplot(table(CO2$Type), main="Distribution de la variable Type",
xlab="Type", ylab="Effectifs", col="pink")
```

. . . . . /2