JG

# Première année de Licence MIASHS

# TP R – Introduction à la statistique[1]

**Julien GREPAT**[2]

## Contents





[2]Contact : julien.grepat@univ-grenoble-alpes.fr

# Part I

# Séance de TP 1–Statistiques descriptives – Boîtes de distribution

## 1 Prise en main de R

L'objectif de ce TP est de vous familiariser avec les objets et les commandes du logiciels R. R est un logiciel libre et collaboratif de statistique qui s'enrichit au fur et à mesure de l'apport gracieux de chercheurs du monde entier. Il est devenu un outil incontournable de statistique et de visualisation de données, tant dans le monde universitaire que dans le monde de l'entreprise.

### I. R est une calculatrice scientifique

R permet de faire les opérations de calcul élémentaire. Essayez les commandes suivantes :

```
4*5-2^4
1+14/4
```

R permet de faire des calculs plus élaborés. Il utilise pour cela des fonctions. Plusieurs fonctions prédéfinies sont disponibles. Que font les fonctions suivantes ?

```
sqrt(5)
abs(-4)
log(1)
cos(pi)
exp(1)
round(pi,2)
```

### II. R est un langage de programmation

#### I. Création de variables

```
x<-2.5
x
y<-2*x
y
y<-1+log(y)^2
y
Y<-floor(y)       # partie entière de x
Y
```

Dans la première ligne, On a créé une variable de nom $x$ contenant la valeur 2.5. La flèche, obtenu en tapant $<$ et $-$ est appelée opérateur d'affectation. Cet opérateur dit que la variable $x$ doit contenir la valeur 2.5. Si cette variable n'existe pas, elle est créée. Si elle existe, son contenu est remplacé par la nouvelle valeur. Pour rappeler le contenu d'une variable, il suffit de taper son nom (ligne 2).

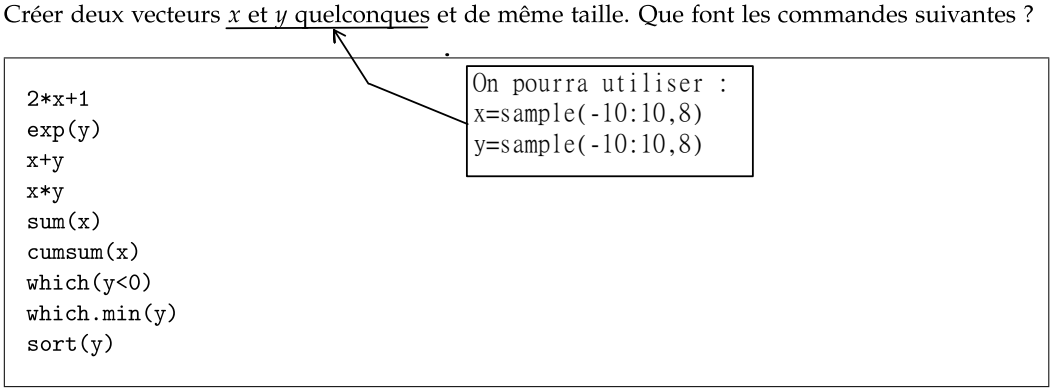
## II. Création d'un vecteur

Il y a différentes manières de créer des vecteurs dans R. Étudiez les lignes de commande suivantes

```
vec<-c(-1,5,2,0.5,-5)
vec
vec1<-c(3,10,vec)
vec1
length(vec1)
rep(c(3,5),2)
1:20
5:1
seq(from=1, to=16, by=3)
seq(from=1, to=2, length=10)
```

## III. Opérations sur les vecteurs

Créer deux vecteurs *x* et *y* quelconques et de même taille. Que font les commandes suivantes ?




```
2*x+1
exp(y)
x+y
x*y
sum(x)
cumsum(x)
which(y<0)
which.min(y)
sort(y)
```

## IV. Accès aux éléments d'un vecteur

On peut accéder aux éléments d'un vecteur grâce à l'opérateur "[]".

```
x<-c(3,-1,5,7,0,3,2,-9)
y<-c("rouge","noir","vert")            # vecteur de chaines de charactères
x[1:3]
x[-1]
x[c(2,4)]
x[x>0 & x<=5]
x[1]<-1.25
x
z=y[y!="rouge"]
z
```

# 2 R pour les statistiques

## 2.1 Les dataframes

Tout comme les vecteurs, les dataframes sont des objets, mais plus élaborés et spécialement désignés pour le stockage de données. Il s'agit plus ou moins d'un tableau à 2 dimensions (une matrice), mais avec en plus des noms de colonnes, des noms de lignes, etc. Très généralement, et ce sera toujours le cas en ce qui nous concerne, les lignes d'un dataframe seront des individus (dans l'exemple ci-dessous ce sont des sujets) et les colonnes des variables (dans l'exemple ci-dessous ce sont le poids, la taille et l'Indice de Masse Corporelle des sujets).

```
poids= c(65,82,45,63,70)
taille= c(1.75,1.78,1.52,1.57,1.80)
IMC= poids/(taille^2)                  # IMC est l'indice de masse corporelle
data=cbind(poids,taille,IMC)
# cbind permet de coller des colonnes les unes à la suite des autres
# et rbind permet de coller des lignes les unes à la suite des autres
```

De la même manière que pour les vecteurs, les crochets permettent d'aller chercher des éléments d'une matrice ou d'un tableau. Il y a alors 2 paramètres à définir, le premier pour les lignes et le second pour les colonnes à sélectionner (les 2 séparés par une virgule). Si rien n'est préciser pour le premier paramètre, cela signifie que l'on prend toutes les lignes. Si rien n'est préciser pour le deuxième paramètre, cela signifie que l'on prend toutes les colonnes.

```
data[3,2]       # taille du troisième sujet
data[ ,1]       # poids de tous les sujets
dim(data)
nrow(data)
ncol(data)
 rownames(data)
  colnames(data)
```

## 2.2 Les fonctions statistiques de R

On considèrera la série statistique

```
x=c(1,2,4,8,9,15,6,8,18,7,5,2,4,6,8,9)
```

R possède plusieurs fonctions statistiques. Que font les fonctions suivantes ?

```
min(x)
max(x)
mean(x)
sd(x)
quantile(x,0.25)
median(x)
var(x)
summary(x)
```

## II. Représentations graphiques

**La fonction plot**
La fonction plot est la fonction générique de graphe.

```
x=seq(-2,2,by=0.1)
y=exp(x)
plot(x, y)
```

Il est possible de paramétrer l'affichage obtenu (couleurs, etc). Voici un bref aperçu des options de la fonction plot, vous pouvez regarder l'aide (en tapant ?plot) pour plus de détails. Nous aurons de toute façon l'occasion de revenir sur l'utilisation de cette fonction.

- **type** (chaîne de caractères) "p" pour points, "l" pour lignes, "b" pour les deux (both) ;
- **main** (chaîne de caractères) un titre pour le graphique ;
- **sub** (chaîne de caractères) un sous-titre pour le graphique ;
- **xlab** (chaîne de caractères) un titre pour l'axe des abscisses ;
- **ylab** (chaîne de caractères) un titre pour l'axe des ordonnées ;
- **col** (vecteur de chaînes de caractères) la ou les couleur(s) à utiliser ;
- etc . . .

```
plot(x, y, pch = 1, type="l", col = "red", main = "Courbe de la fct exp",
xlab = "abscisse", ylab = "ordonnée")
```

# 3 Criminalité – Peine de mort

Le but de l'étude statistique de cette partie est de s'éclairer sur le lien potentiel entre abolition de la peine de mort et taux de criminalité. Nous resterons néanmoins très prudent sur les conclusions que l'on peut rendre avec des outils de statistiques descriptives.

Nous vous conseillons dès à présent d'allumer votre ordinateur et de lancer le logiciel R.

En annexe 1, on trouvera des données issues du site du FBI. Elle donnent le nombre de meurtres dans les différents états des États Unis en 2009 ainsi que la population.

Afin de s'affranchir des différences de tailles de populations dans chaque état, nous contruisons le taux de criminalité pour 10 000 habitants.

J'ai malheureusement renversé mon café sur la colonne *taux de criminalité*.

## 3.1 Construction de la variable *taux de criminalité* pour les états ayant aboli la peine de mort

(i) À partir des colonnes 2 et 3 de l'annexe, comment calculer le taux de criminalité pour 10 000 habitants ?

(ii) Nous allons créer la variable série statistique `abolie09` contenant le taux de criminalité pour 10 000 habitants des états ayant aboli la peine de mort. Pour cela, nous allons saisir les variables `murder` et `population`. Nous calculerons alors la variable `abolie09`.

```
murder=c(7,34,72,9,21,26,144,31,22,319,76,625,144)
population=c(622,3008,5266,647,1395,1318,5655,1053,698,8708,1820,9970,2010)
```

*On rappelle que dans R, une série statistique est un vecteur horizontal. On peut affecter une valeur à une variable en utilisant au choix* `<-` *ou* `=`*.*

Calculons la variable `abolie09` en divisant terme à terme `murder` par `population` et en multipliant par 10 (pour un taux par dizaine de milliers d'habitants).

```
abolie09=10*murder/population
```

On pourra afficher en tapant `abolie09` puis entrée pour vérifier l'adéquation avec ce qui reste visible dans le tableau en annexe.

## 3.2 Comparaison des variables *taux de criminalité* pour les états ayant aboli ou non la peine de mort

Nous avons en mémoire la variable `abolie09`. Je vous propose de mettre en mémoire la variable `nonabolie09` qui porte le taux de criminalité dans les pays n'ayant pas aboli la peine de mort en 2009.

```
nonabolie09=c(0.075,0.133,0.135,0.142,0.202,0.217,0.223,0.254,0.256,0.287,0.304,
0.332,0.371,0.394,0.399,0.419,0.435,0.440,0.456,0.463,0.497,0.512,0.512,0.522,
0.534,0.535,0.552,0.590,0.592,0.610,0.627,0.636,0.675,0.732,0.768,1.082,2.401)
```

Vous noterez l'indispensable d'utiliser d'autres méthodes d'acquisition des données pour les grandes séries. Nous pouvons vérifier qu'il ne manque pas de valeur en demandant le nombres de valeurs dans la série tapée :

```
length(nonabolie09)
    37
```

(i) À l'aide de la fonction `mean()` calculer les moyennes de `abolie09` et `nonabolie09`. Comparer. Que peut-on en conclure ?

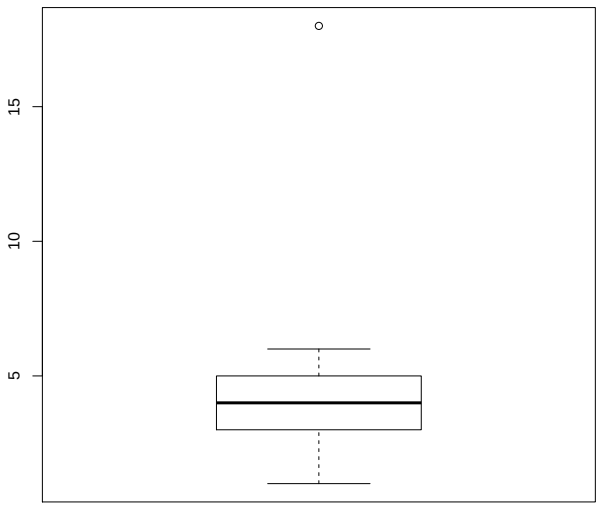
(ii) On rappelle ce qu'est le diagramme *boîte de distribution*.

---

La boîte de distribution (ou boxplot) est une représentation graphique synthétique de la distribution des données. Elle résume quelques caractéristiques de position et de dispersion du caractère étudié (médiane, quartiles, minimum et maximum). Ce diagramme est utilisé essentiellement pour comparer un même caractère dans des populations différentes, ou une évolution au cours du temps.

(a) Tracer un rectangle qui s'étend du premier quartile au troisième.

(b) Séparer ce rectangle en deux à la hauteur de la médiane. On obtient alors une boîte.




(c) On complète ce rectangles par deux segments. Pour cela, on calcule

$$a = q_{0.25} - 1.5IQ \quad \text{et} \quad b = q_{0.75} + 1.5IQ,$$

avec la distance inter-quartile

$$IQ = q_{0.75} - q_{0.25}.$$

On repère les valeurs :

$$x_a = \min\{x_i : x_i \geq a\} \quad \text{et} \quad x_b = \max\{x_i : x_i \leq b\}.$$

Ces valeurs sont appelées *valeurs adjacentes*. On relie ces valeurs aux cotés de la boîte.

(d) Les valeurs qui ne sont pas comprises entre les valeurs adjacentes sont représentées par des points et sont appelées *valeurs extrêmes*.

---

On va tracer les boîtes de distribution des deux séries `abolie09` et `nonabolie09` côte à côte pour avoir une meilleur vision de leur distribution.

```
boxplot(abolie09,nonabolie09)
```

On pourra modifier les paramètres du graphe, par exemple ajouter un titre principal avec le paramètre `main` :

```
boxplot(abolie09,nonabolie09,main="Taux de criminalité").
```

*Les accents sont pris en charge. Les chaînes de caractère doivent être mises entre guillemets.* On peut également changer le nom de chaque variable et sa couleur avec les paramètres `names` et `col` dont les arguments sont des vecteurs portant la caractéristique souhaitée pour chaque série dans l'ordre.

```
boxplot(abolie09,nonabolie09,main="Taux de criminalité",col=c("orange","red"),
names=c("abolie09","nonabolie09")).
```

Que peut-on déduire de ce graphe ?

(iii) Peut-on déduire un lien de cause à effet avec ce type d'étude ?

### 3.3 Autre démarche pour obtenir le boxplot

Supposons que, comme c'est dans la présentation de l'annexe, les deux séries `abolie09` et `nonabolie09` soient concaténées dans un seul vecteur de données.

```
txmeurtres=c(abolie09,nonabolie09)
```

Faites afficher `txmeurtres` pour vous en persuader. On peut créer la variable qui discrimine l'abolition ou non de la peine de mort.

```
pdm=c(rep("oui",13),rep("non",37))
```

Faites afficher `pdm` pour vérifier que nous avons le statut de l'abolition ou non de la peine de mort pour chaque état dans le même ordre que le taux de criminalité. On pourra obtenir le même boxplot que précdédemment en discriminant les donnés sur le facteur `pdm`.

```
boxplot(txmeurtres∼pdm)
```

# Part II

# Séance de TP 2 – Statistique du $\chi^2$

## 4 Distributions bivariées

On souhaite se familiariser avec le tableau de contingence d'une série double. On prend pour exemple la série des poids et tailles de bébés :

| | $Y$ : **Poids à la naissance** | | | | |
|---|---|---|---|---|---|
| $X$ : **Sexe** | **Faible (0,5-2 kg)** | **Moyen (2-3 kg)** | **Élevé (3-4 kg)** | **Très élevé (4 kg et +)** | **Total** |
| Garçons | 830 | 8 615 | 30 784 | 4 839 | 45 068 |
| Filles | 862 | 11 183 | 27 566 | 2 348 | 41 959 |
| Total | 1 692 | 19 798 | 58 350 | 7 187 | 87 027 |

SOURCE : Bureau de la Statistique du Québec

### 4.1 Saisie du tableau de contingence

Le tableau de contingence est une matrice sous R à laquelle nous allons rajouter quelques éléments cosmétiques.

(i) On saisit la matrice ayant pour coefficients les effectifs dans l'ordre du tableau ci dessus. On peut choisir entre deux façon de déclarer :

- concaténer des vecteurs verticaux ("c" pour colonne) :

```
poidstaille=cbind(c(830,862),c(8615,11183),c(30784,27566),c(4839,2348))
```

- concaténer des vecteurs verticaux ("r" pour raw (ligne)) :

```
poidstaille=rbind(c(830,8615,30784,4839),c(862,11183,27566,2348))
```

(ii) On ajoute la colonne modalité de $X$ :

```
rownames(poidstaille)=c("Garçons","filles")
```

(iii) On ajoute la ligne modalité de $Y$ :

```
colnames(poidstaille)=c("Faible","Moyen","Elevé","Tr.élv")
```

(iv) Afficher `poidstaille`.

### 4.2 Tableau des fréquences

(i) On peut obtenir l'effectif total de la population par la fonction `sum()` :

```
sum(poidstaille)
```

.

(ii) On peut obtenir le tableau des fréquences. Il ne faut pas oublier que notre tableau est une matrice et que l'on peut effectuer toutes les opérations usuelles comme diviser la matrice `poidstaille` terme à terme par l'effectif total `sum(poidstaille)`

```
frequences=poidstaille/sum(poidstaille)
```

.

On obtient :

| | Faible | Moyen | Eleve | Tr. elv |
|---|---|---|---|---|
| Garçons | 0.009537270 | 0.09899227 | 0.3537293 | 0.05560343 |
| filles | 0.009904972 | 0.12850035 | 0.3167523 | 0.02698013 |

**Remarque 4.1** *On aurait pu obtenir ce tableau par la commande* `prop.table()` *:*

```
prop.table(poidstaille)
```

.

### 4.3 Distributions conditionnelles

(i) La distribution de $Y$ pour les garçons est donnée par la première ligne de la matrice `poidstaille` soit

```
poidstaille[1,]
```

(on note qu'on fixe la première ligne par le 1, mais on fait dérouler les colonnes). On obtient

| Faible | Moyen | Eleve | Tr. elv |
|---|---|---|---|
| 830 | 8615 | 30784 | 4839 |

En calculant les fréquences par rapport à l'effectif de garçons `sum(poidstaille[1,])`, on obtient la distribution conditionnelle de $Y$ étant donné $X$ =garçon :

| Faible | Moyen | Eleve | Tr. elv |
|---|---|---|---|
| 0.01841661 | 0.19115559 | 0.68305671 | 0.10737108 |

obtenu par `poidstaille[1,]/sum(poidstaille[1,])`.

(ii) De même, on obtiendra la distribution conditionnelle de $Y$ étant donné $X$ =fille par `poidstaille[2,]/sum(poidstaille[2,])`.

(iii) Tracer sur le même diagramme en bâtons ces deux distributions conditionelles. Que peut-on conclure intuitivement sur la dépendance entre $X$ et $Y$ ?

**Réponse :** On note qu'utiliser des fréquences ou des effectifs ne change que l'échelle. On choisit donc de tracer les effectifs. On utilisera la commande

`barplot(poidstaille,beside=TRUE)`.

On observe un léger décallage du pois des bébés filles vers les faibles poids. Ce n'est pas flagrant.

### 4.4 Distributions marginales

(i) On peut obtenir les colonnes et lignes "Total" (les marges) par la commande `addmargins` :

`frequences=addmargins(frequences)`.

(ii) La série de fréquences de Garçon et de filles dans la population est appelée distribution marginale de $X$ :

| Garçons | filles | Sum |
|---|---|---|
| 0.5178623 | 0.4821377 | 1.0000000 |

.

Il s'agit de la colonne `sum` que l'on extrait par la commande `frequences[,5]` (on note qu'on fixe la dernière colonne par le 5, mais on fait dérouler les lignes).

(iii) De même, on obtient la distribution marginale de $Y$ à partir de la ligne `sum` en extrayant de `frequences` la dernière ligne (on fixe la troisième ligne et on déroule les colonnes) grâce à `frequences[3,]` :

| Faible | Moyen | Eleve | Tr. elv | Sum |
|---|---|---|---|---|
| 0.01944224 | 0.22749262 | 0.67048157 | 0.08258357 | 1.00000000 |

### 4.5 $V$ de Cramér

On se propose d'utiliser un indicateur d'association appelé $V$ de Cramér pour vérifier l'observation de la dépendance entre les variables observée dans la section 4.3. Plus $V$ est proche de zéro, plus il y a indépendance entre les deux variables $X$ et $Y$ étudiées. Il vaut 1 en cas de complète dépendance.

Le coefficient $V$ de Cramér nécessite l'utilisation de la statistique du $\chi^2$. La statistique du $\chi^2$ est disponible via le test du même nom :

chisq.test(poidstaille).

On s'aperçoit que R donne plusieurs valeurs et non seul la statistique. Nous verrons la signification de ces valeurs dans la suite. Nous rappelons la formule du $V$ de Cramér :

$$V = \sqrt{\frac{D^2_\chi}{n \times min\{l-1; c-1\}}},$$

où $n$ est l'effectif total de la population, $c$ est le nombre de colonnes (nombre de modalités de $Y$) et $l$ le nombre de lignes (modalités de $X$).

On se propose de définir une fonction `Cramer` : (*taper sans fautes !*)

| | |
|---|---|
| `cramer=function(table){` | La fonction s'appelle Cramer, la variable à qui elle s'applique sera nommée `table` durant la programmation de la fonction |
| `test=chisq.test(table)` | Le test du $\chi^2$ est stocké dans la variable `test` |
| `chi2=as.numeric(test$statistic)` | Nous ne prenons que la variable `statistic`[1] dans `test`, on l'affecte à `chi2` |
| `n=sum(table)` | L'effectif total stocké dans `n` |
| `c=length(table[1,])` | Le nombre de colonnes est la longueur d'une ligne de `table` |
| `r=length(table[,1])` | Le nombre de lignes est la longueur d'une colonne de `table` |
| `m=min(c,r)` | |
| `V=sqrt(chi2/(n*(m-1)))` | |
| `V` | Ne pas oublier de faire afficher `V` |
| `}` | Fin de la déclaration de la fonction |

1 Taper `help(chisq.test)` pour obtenir le mode d'emploi en ligne de `chisq.test`.

Il reste à appliquer notre fonction `Cramer` à notre tableau de contingence `poidstaille` pour lire le $V$ de Cramér :

```
cramer(poidstaille)
```.

On donne le tableau suivant pour l'interprétation de la valeur du $V$ de Cramér :

| Valeur du $V$ de Cramér | Intensité de la relation entre les variables |
|---|---|
| inférieur à 0,10 | relation nulle ou très faible |
| entre 0,10 et 0,20 | relation faible |
| entre 0,20 et 0,30 | relation moyenne |
| au dessus de 0,30 | relation forte |

Interpréter le résultat.

# 5 test du $\chi^2$

## 5.1 Acquisition de fichier `.csv`

Il faut bien reconnaître que pour de grandes séries statistiques, on préfèrerait éviter d'avoir à retaper les valeurs. Pour cela, on a créé le fichier `.csv` (*coma separated values*) qui permet

de communiquer des listes données séparées par des virgules entre différents logiciels comme excel, python, R...

Le fichier "diplome_sexe.csv" recense le sexe et le niveau de diplôme obtenu d'un échantillon aléatoire de 1367 diplômés d'université. L'objectif de ce paragraphe est d'étudier le relation entre ces deux variables qualitatives. Nous allons pour cela effectuer un test d'indépendance de Chi-deux. Puis pour quantifier cette relation, nous utiliserons le coefficient de Cramer.

Enregistrer le fichier `diplome_sexe.csv` et sélectionner le dossier dans lequel il est enregistré comme répertoire de travail dans R (Fichier => changer le répertoire courant).

En ouvrant le fichier `diplome_sexe.csv` dans un éditeur de texte on s'aperçoit que les données sont organisées verticalement sous *Sexe* et *Diplome* qui servent d'étiquettes de liste (`header=TRUE`) et que les colonnes sont délimitées par des ";", (`sep=";"`).

Ouvrons le fichier dans R et affectons le à la variable `data`.

```
data=read.csv("diplome_sexe.csv",header=TRUE,sep=";").
```

Avec la fonction `head()`, vérifions que les donnés ont été correctement importées :

```
> head(data)
      Sexe  Diplome
1 Masculin  Licence
2 Masculin  Licence
3 Masculin  Licence
4  Féminin Maîtrise
5 Masculin  Licence
6 Masculin Doctorat
```

Utilisons la fonction `table()` pour éditer le tableau de contingence des variables *diplôme* et *sexee* et affichons le.

```
contingences=table(data)
contingences
```

Utilisons la fonction `chisq.test()` pour effectuer le test du Chi2 et conclure quant à la dépendance entre le sexe et le niveau d'étude obtenu.

```
> chisq.test(contingences)

        Pearson's Chi-squared test

data:  contingences
X-squared = 3.2476, df = 2, p-value = 0.1971
```

On note que la $p-$value est supérieure à 0.05. On ne rejette donc pas l'hypothèse nulle qui est l'indépendance. On en conclut qu'il y a significativement indépendance des variables.

Utiliser la fonction `cramer()` pour calculer le $V$ de Cramér et confirmer l'indépendance des deux variables.

```
> cramer(contingences)
[1] 0.04874159
```

**Réponse :** Puisque le `V` est inférieur à 0,1, cela confirme l'indépendance des variables.

# 6 Normalité et Anova

Enregistrer le fichier `notesCC1.csv` et sélectionner le dossier dans lequel il est enregistré comme répertoire de travail dans R (Fichier => changer le répertoire courant). Ce fichier regroupe vos notes de $CC1$ par ordre décroissant, pour éviter toute perte de confidentialité, et le numéros du sujet sur lequel portait la copie. On se propose d'étudier la normalité des notes et de faire une ANOVA pour savoir s'il y a une différence significative des notes en fonction du sujet donné (et ainsi se questionner sur l'égalité de vos chances).

En ouvrant le fichier `notesCC1.csv` dans un éditeur de texte on s'aperçoit que les données sont organisées verticalement :

- sous *note* et *sujet* qui servent d'étiquettes de liste (`header=TRUE`),
- séparées par un ";" (`sep=";"`),
- avec pour séparateur décimal une virgule (`dec=","`).

Ouvrons le fichier dans R et affectons le à la variable `data`.

```
data=read.csv("notesCC1.csv",header=TRUE,sep=";",dec=",").
```

Vérifions l'acquisition

```
> head(data)
  note sujet
1 19.0     1
2 19.0     1
3 18.5     2
4 18.5     1
5 18.5     1
6 18.0     1
```

Affectons les variables `sujet` et `note` dans notre environnement R.

```
sujet=data$sujet
note=data$note
```

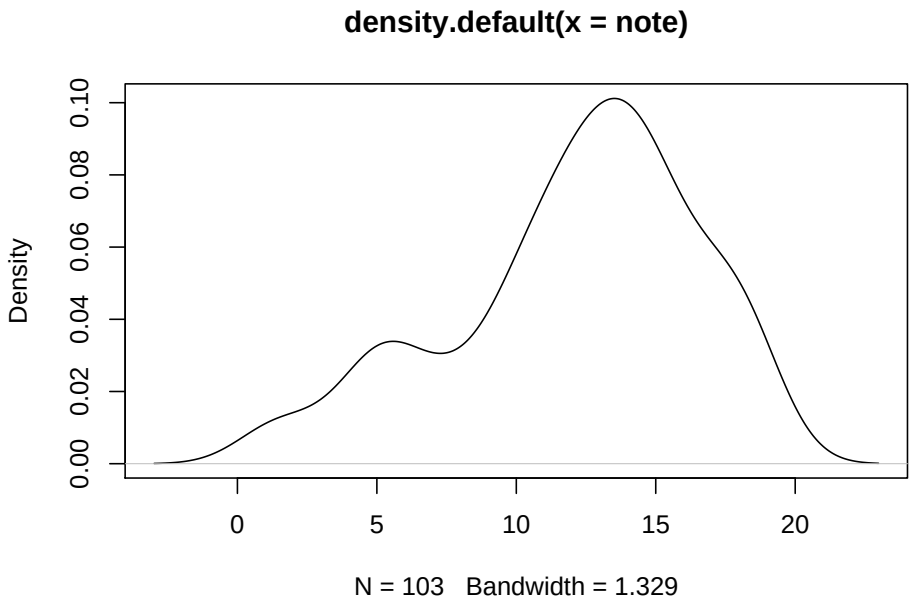
Il y a fort à parier que R prend la variable sujet pour une variable quantitative. Or, c'est une variable qualitative (facteur) à deux valeurs. Forçons R a la considérer comme tel.

```
> sujet=data$sujet
> head(sujet)
[1] 1 1 2 1 1 1
> sujet=factor(sujet)
> head(sujet)
[1] 1 1 2 1 1 1
Levels: 1 2
```

## 6.1 Normalité des notes

Testons la normalité des notes avec le test de Shapiro–Wilk. Nous pouvons d'abord regarder la distribution des notes

```
plot(density(note))
```




La cloche semble bosselée, ils n'y a sans doute pas normalité... Rappelons que l'hypothèse nulle du test de Shapiro–Wilk est "la série est normalement distribuée". Effectuer le test

```
> shapiro.test(note)

        Shapiro-Wilk normality test

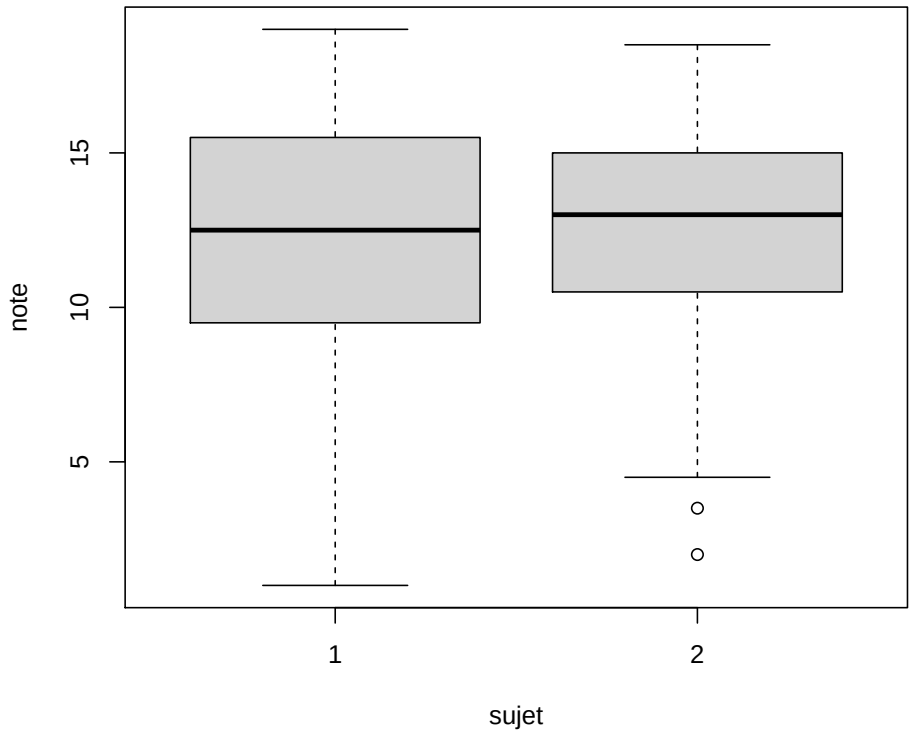
data:  note
W = 0.95003, p-value = 0.0006783
```

**Réponse :** La $p-$value est proche de zéro. On en déduit que l'on rejette significativement la normalité.

## 6.2 Influence du sujet sur la note

Affichons les boîtes de distribution côte à côte avec ces nouvelles séries.

```
boxplot(note~sujet)
```




Honnêtement... peu de différences.

Tester l'hypothèse nulle "Le facteur(sujet) n'a pas d'effet sur les notes" par l'ANOVA.

```
anova(lm(note~sujet))
```

```
Analysis of Variance Table

Response: note
           Df  Sum Sq Mean Sq F value Pr(>F)
sujet       1    0.95  0.9503  0.0485 0.8262
Residuals 101 1980.37 19.6077
```

**Réponse :** La $p-$value 0.8262 est grande, on ne rejette pas significativement l'hypothèse nulle. On retient que le sujet n'influence pas les notes.

# Part III
# Séance de TP 3–Régressions linéaires

## 2 Dépendance non linéaire

On considère les données suivantes :

| $x_i$ | 1 | 4 | 4 | 4 | 8 | 4 | 1 | 1 | 1 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | -16 | 16 | 16 | 16 | 9 | 16 | -16 | -16 | -16 | 9 |

### 2.1 Saisie des séries `x` et `y` dans R

```
x=c(1,4,4,4,8,4,1,1,1,8)
y=c(-16,16,16,16,9,16,-16,-16,-16,9)
```

.

Tracer `y` en fonction de `x`,

```
plot(x,y)
```

### 2.2 Calcul du coefficient de corrélation $r_{xy}$

```
cor(x,y)
```

.

### 2.3 Tableau de contingences et test du $\chi^2$

```
contingence=table(x,y)
```

.

Calculer la statistique du Chi-2. Commenter.

*On peut calculer la statistique du $\chi^2$ avec R en utilisant* `chisq.test(contingence)`*. Néanmoins, R nous informe qu'il ne peut pas calculer une p-valeur fiable à cause d'effectifs trop petits. On va chercher la valeur du quantile 0.95 du $\chi^2$ :*

```
qchisq(0.95,df=4)
```

.

> **Réponse :** On en déduit que les variables ne sont pas indépendantes (au seuil 5%) car $D_{\chi^2} = 20$ est supérieure au quantile 0.95 du $\chi^2$ qui vaut 9,49, mais que leur dépendance n'est pas linéaire.

**Remarque 2.1** *On note que la p-value donnée par le test du $\chi^2$ est très petite :* `p-value = 0.0004994`*. Inférieure à 5%, elle confirme la dépendance des variables.*

# 3 Régression linéaire simple

On va utiliser le fichier `GPA.csv` qui contient des notes d'études secondaires et universitaires pour les diplômés en informatiques dans une école publique locale. Notre objectif est de déterminer la droite des moindres carrés permettant d'expliquer linéairement la note universitaire d'un étudiant par sa note du secondaire.

## 3.1 Acquisition des données

Quelle est la variable explicative (indépendante) et la variable à expliquer (dépendante) ?

> **Réponse :** Variable explicative : note secondaire, variable expliquée : note universitaire.

Stockons les variable. Nous pouvons voir que le format du fichier `.csv` suit les standards francophones (header, ";" pour séparer les colonnes, "," comme séparateur décimal). Pour éviter le détail de compléter tous les paramètres dans la fonction `read.csv()`, on peut utiliser la fonction `read.csv2()`.

```
data=read.csv2("GPA.csv")
x=data$high_GPA
y=data$univ_GPA
```

.

## 3.2 Nuage de points

Tracer le nuage de points et commenter.

```
plot(x,y)
```

.

> **Réponse :** Le nuage est relativement rectiligne, la régression linéaire a du sens, elle donnera une tendance moyennement précise de `y` en fonction de `x`.

## 3.3 Corrélation linéaire

Compléter la fonction suivante permettant de calculer le coefficient de corrélation linéaire entre `x` et `y`.

```
correlation=function(u,v){
cov=cov(u,v)
denom=sqrt(var(u)*...)
corr=.../...
corr
}
```

Appliquer la fonction à `x` et `y`:

```
correlation(x,y)
```

Vérifier que l'on obtient la même valeur avec la fonction `cor` :

```
cor(x,y)
```

### 3.4 Régressions linéaire

Calculer les coefficients de la droite de régression $y = ax + b$ par la méthode des moindres carrés.

```
a=cov(x,y)/var(x)
b=mean(y)-a*mean(x)
```

.

On utilise la fonction `abline` pour tracer la droite de régression :

```
abline(b,a,col="red")
```

.

On peut calculer les valeurs ajustées et les erreurs :

```
yajust=a*x+b
erreurs=y-yajust
```

.

La fonction `lm` permet d'effectuer directement ce que nous avons fait :

```
lm(y∼x)
model=lm(y∼x)
b=model$coefficients[1]
a=model$coefficients[2]
yajustes=model$fitted.values
erreurs=model$residuals
```

.

### 3.5 Résidus

Représenter les résidus (erreurs) en fonction des valeurs ajustées.

```
plot(yajust,erreurs)
```

.

Voit-on une structure particulière des résidus ?

> **Réponse :** Il ne semble pas y avoir de structure particulière des erreurs car dans une régression linéaire on a une hypothèse : les erreurs sont distribuées selon une loi normale $\mathcal{N}(0, \sigma)$ et sont indpendantes.

### 3.6 Prévision

Un diplomé a eu 2,5. Donner une prédiction de sa note universitaire.

> **Réponse :** On prévoit 7,55.

### 3.7 Anova

Nous avons vu que la proportion de variance de `y` expliquée par la droite est assez faible. À l'aide l'ANOVA décider si la variable `x` explique significativement la variable `y`.

```
> anova(lm(y~x))
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 227.199 227.199  297.71 < 2.2e-16 ***
Residuals 84  64.105   0.763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> **Réponse :** Puisque l'hypothèse nulle du test de l'ANOVA est "`x` n'explique pas `y`" et que la $p-$value est proche de 0, on peut conclure que `x` a un effet significatif sur `y`.

## 4 Régression linéaire multiple

Pour mesurer les performances auditives d'un individu, on le soumet à un signal sonore d'une fréquence donnée dont l'intensité va croissante. On note alors le seuil à partir duquel le signal est perçu. Ce seuil, exprimé en décibel, mesure la performance d'audition à partir du seuil moyen. Ainsi, un seuil haut correspond à un trouble de l'audition pour la fréquence.

Les données représentent les mesures obtenues pour quatre fréquences (500Hz, 1000Hz, 2000Hz et 4000Hz). On prendra le seuil maximal pour l'oreille gauche et l'oreille droite pour 1000 personnes de plus de 39 ans.

On ajoutera la donnée `globale` qui est une auto-évaluation par le patient.

Le tableau `audition2.csv` donne les valeurs des 1000 relevés auditifs.

L'enjeu est de prédire `globale` par les variables `A5, A10, A20, A40` et de savoir si le patient peut prédire lui-mêm l'état de son audition.

### 4.1 Import des données

Importons les données et stockons les dans les variables du même nom.

```
audition=read.csv2("audition2.csv")
A5=audition$A5
A10=audition$A10
A20=audition$A20
A40=audition$A40
globale=audition$globale
```

## 4.2 Les coefficients de la régression linéaire

Donner l'équation de la régression linéaire multiple et discuter de l'effet des variables explicatives sur la variable expliquée.

```
> lm(globale~A5+A10+A20+A40)

Call:
lm(formula = globale ~ A5 + A10 + A20 + A40)

Coefficients:
(Intercept)           A5          A10          A20          A40
     4.1423       0.7082       0.8602       0.6355       0.5106
```

**Réponse :** Ainsi, on peut lier les variables par l'équation

$$globale = 4,14 + 0,7A5 + 0,86A10 + 0,64A20 + 0,51A40.$$

On voit que le niveau d'audition globale évalué par le sujet augmente quand chaque seuil augmente avec sensiblement la même importance.

## 4.3 Tests statistiques

Faire les tests statistiques classiques de la régression linéaire et conclure.

```
> summary(lm(globale~A5+A10+A20+A40))

Call:
lm(formula = globale ~ A5 + A10 + A20 + A40)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63918 -0.19985  0.04163  0.27016  0.72287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.14226    0.01274  325.10   <2e-16 ***
A5           0.70820    0.01808   39.17   <2e-16 ***
A10          0.86024    0.01948   44.17   <2e-16 ***
A20          0.63553    0.01439   44.17   <2e-16 ***
A40          0.51059    0.01333   38.30   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3613 on 995 degrees of freedom
Multiple R-squared:  0.9653,    Adjusted R-squared:  0.9651
F-statistic:  6911 on 4 and 995 DF,  p-value: < 2.2e-16
```

**Réponse :**

- Le test de Fisher (ANOVA), donné dans la dernière ligne du résumé, montre que, dans leur globalité, les variables explicatives ont un effet signifcatif sur `globale`.
- Tous les tests de Student amènent à dire que les coefficients de la régression linéaire multiple sont significativement non nuls. Chaque variable explique significativement la variable `globale`.
- De plus, le coefficient $R^2$ vaut 0,97. Ainsi, 97% de la variance de `globale` est expliquée par le modèle linéaire. Ainsi, il semble que la prévision soit proche de l'évaluation par le sujet. On peut penser qu'il a évaluation lucide de son état d'audition.

## Part IV

# Annexe 1

| State | Population (en milliers) | Total murders (2009) | Nombre de meutres pour 10000 habitants | Peine de Mort abolie en 2009 |
|---|---|---|---|---|
| Vermont | 622 | 7 | 0,113 | oui |
| Iowa | 3 008 | 34 | [redacted] | oui |
| Minnesota | 5 266 | 72 | | oui |
| North Dakota | 647 | 9 | | oui |
| Hawaii | 1 295 | 21 | | oui |
| Maine | 1 318 | 26 | | oui |
| Wisconsin | 5 655 | 144 | | oui |
| Rhode Island | 1 053 | 31 | | oui |
| Alaska | 698 | 22 | | oui |
| New Jersey | 8 708 | 319 | | oui |
| West Virginia | 1 820 | 76 | | oui |
| Michigan | 9 970 | 625 | | oui |
| New Mexico | 2 010 | 144 | 0,717 | oui |
| Florida | 18 538 | | | |
| New Hampshire | 1 325 | 10 | 0,075 | |
| Utah | 2 785 | 37 | 0,133 | |
| South Dakota | 812 | 11 | 0,135 | |
| Idaho | 1 546 | 22 | 0,142 | |
| Wyoming | 544 | 11 | 0,202 | |
| Oregon | 3 826 | 83 | 0,217 | |
| Nebraska | 1 797 | 40 | 0,223 | |
| Washington | 6 664 | 169 | 0,254 | |
| Massachusetts | 6 594 | 169 | 0,256 | |
| Montana | 975 | 28 | 0,287 | |
| Connecticut | 3 518 | 107 | 0,304 | |
| Colorado | 5 025 | 167 | 0,332 | |
| Illinois | 12 910 | 479 | 0,371 | |
| Kentucky | 4 314 | 170 | 0,394 | |
| New York | 19 541 | 779 | 0,399 | |
| Kansas | 2 819 | 118 | 0,419 | |
| Ohio | 11 543 | 502 | 0,435 | |
| Virginia | 7 883 | 347 | 0,440 | |
| Indiana | 6 423 | 293 | 0,456 | |
| Delaware | 885 | 41 | 0,463 | |
| Arizona | 6 596 | 328 | 0,497 | |
| Mississippi | 2 952 | 151 | 0,512 | |
| North Carolina | 9 381 | 480 | 0,512 | |
| Pennsylvania | 12 605 | 658 | 0,522 | |
| California | 36 962 | 1 972 | 0,534 | |
| Texas | 24 782 | 1 325 | 0,535 | |
| Georgia | 9 829 | 543 | 0,552 | |
| Nevada | 2 643 | 156 | 0,590 | |
| Arkansas | 2 889 | 171 | 0,592 | |
| Oklahoma | 3 687 | 225 | 0,610 | |
| South Carolina | 4 561 | 286 | 0,627 | |
| Missouri | 5 988 | 381 | 0,636 | |
| Alabama | 4 709 | 318 | 0,675 | |
| Tennessee | 6 296 | 461 | 0,732 | |
| Maryland | 5 699 | 438 | 0,768 | |
| Louisiana | 4 492 | 486 | 1,082 | |
| District of Columbia | 600 | 144 | 2,401 | |