TP R Introduction à la statistique

Mohamed FOFANA | *L1 MIASHS - Université Grenoble Alpes*

Contents

1 Séance de TP 2: Statistique du Khi - carré χ^2 **1**
1.1 Distributions bivariées 1
1.2 test du χ^2 5
1.3 Normalité et Anova 6

1 Séance de TP 2: Statistique du Khi - carré χ^2

1.1 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

1.1.1 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("Gar¸cons","filles")
```

(iii) On ajoute la ligne modalité de Y :

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

(iv) Afficher poidstaille.

```
head(poidstaille)
```

```
##          Faible Moyen Elevé Tr.élv
## Gar¸cons    830  8615 30784   4839
## filles      862 11183 27566   2348
```

1.1.2 Tableau des fréquences

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

```
sum(poidstaille)
```

```
## [1] 87027
```

(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)
```

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

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

```
##               Faible      Moyen     Elevé     Tr.élv
## Gar¸cons 0.009537270 0.09899227 0.3537293 0.05560343
## filles   0.009904972 0.12850035 0.3167523 0.02698013
```

1.1.3 1.3 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 d2rouler les colonnes).
```

```
## Faible  Moyen  Elevé Tr.élv
##    830   8615  30784   4839
```

En calculant les fréquences par rapport à l'effectif de garçons

```
sum(poidstaille[1,])
```

```
## [1] 45068
```

On obtient la distribution conditionnelle de Y étant donné X =garçon :

```
poidstaille[1,]/sum(poidstaille[1,])
```

```
##     Faible      Moyen      Elevé     Tr.élv
## 0.01841661 0.19115559 0.68305671 0.10737108
```

(ii) De même, on obtiendra la distribution conditionnelle de Y étant donné X =fille par

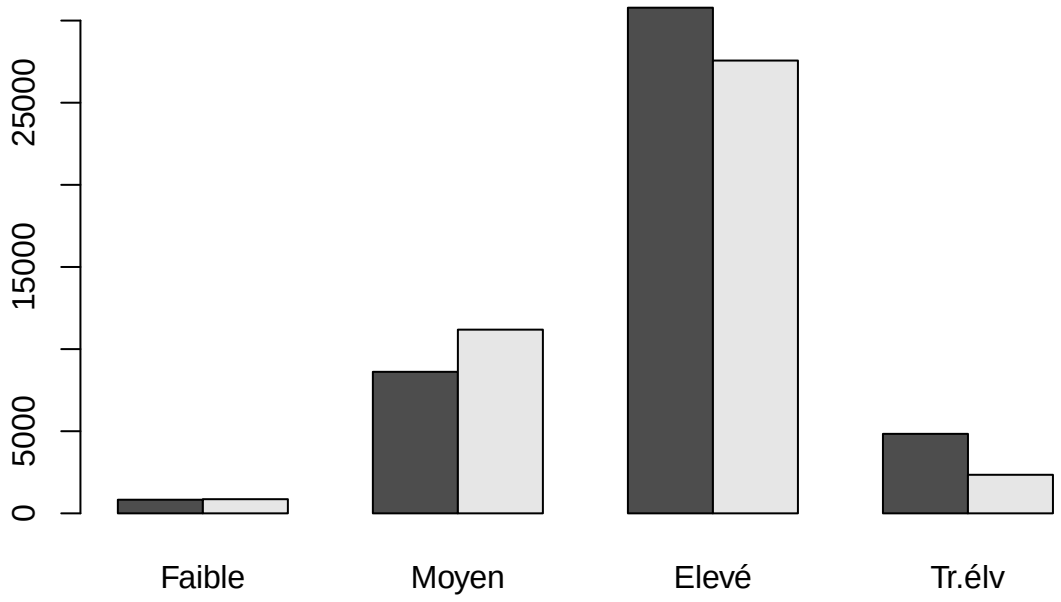
```
poidstaille[2,]/sum(poidstaille[2,])
```

```
##     Faible      Moyen      Elevé     Tr.élv
## 0.02054386 0.26652208 0.65697467 0.05595939
```

(iii) Tracer sur le même diagramme en bâtons ces deux distributions conditionelles. Que peut-on conclure intuitivement sur la d´ependance 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.
```

1.1.4 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 :

```
frequences[,5] #  Il s'agit de la colonne sum que l'on extrait par la commande frequences[,5] (on note
```

```
##  Gar¸cons    filles       Sum
## 0.5178623 0.4821377 1.0000000
```

(iii) De même, on obtient la distribution marginale de Y à partir de la ligne sum en extrayant de frequences la derni'ere ligne (on fixe la troisième ligne et on déroule les colonnes) grâce à:

```
frequences[3,]
```

```
##     Faible      Moyen      Elevé     Tr.élv        Sum
## 0.01944224 0.22749262 0.67048157 0.08258357 1.00000000
```

1.1.5 ν de Cramér

On se propose d'utiliser un indicateur d'association appelé ν de Cramér pour vérifier l'observation de la dépendance entre les variables observée dans la section 4.3. Plus ν est proche de zéro, plus il y a ind´ependance 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 χ^2. La statistique du χ^2 est disponible via le test du même nom :

```
chisq.test(poidstaille)
```

```
##
##  Pearson's Chi-squared test
##
## data:  poidstaille
## X-squared = 1265.1, df = 3, p-value < 2.2e-16
```

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 :

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

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 !)

La fonction s'appelle Cramer, la variable à qui elle s'applique sera nommée table durant la programmation de la fonction, Le test du χ^2 est stocké dans la variable test, Nous ne prenons que la variable statistic dans test, on l'affecte à χ^2, L'effectif total stocké dans n, Le nombre de colonnes est la longueur d'une ligne de table, Le nombre de lignes est la longueur d'une colonne de table, Ne pas oublier de faire afficher ν, Fin de la déclaration de la fonction

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

```
cramer=function(table){
  test=chisq.test(table)
  chi2=as.numeric(test$statistic)
   n=sum(table)
    c=length(table[1,])
r=length(table[,1])
 m=min(c,r)
V=sqrt(chi2/(n*(m-1)))
V
}
```

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

```
cramer(poidstaille)
```

```
## [1] 0.1205687
```

On donne le tableau suivant pour l'interpr´etation de la valeur du V de Cramér :

1.2 test du χ^2

1.2.1 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 donn´ees séparées par des virgules entre diff´erents logiciels comme excel, python, R...

Le fichier "diplome sexe.csv" recense le sexe et le niveau de diplˆome obtenu d'un échantillon aléeatoire de 1367 diplômés d'université. L'objectif de ce paragraphe est d'étudier la 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 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("C:/Users/DVE ICAMPUS/Desktop/MIASHS-UGA/STAT L1/Donnees/diplome_sexe.csv",header=TRUE,sep
```

Avec la fonction head(), vérifions que les données 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 diplˆome et sexee et affichons le.

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

```
##           Diplome
## Sexe       Doctorat Licence Maîtrise
##   Féminin        11     515      141
##   Masculin       22     534      144
```

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.

1.3 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 ´eviter 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'´etiquettes 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("C:/Users/DVE ICAMPUS/Desktop/MIASHS-UGA/STAT L1/Donnees/notesCC1.csv",header=TRUE,sep=";
```

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
```

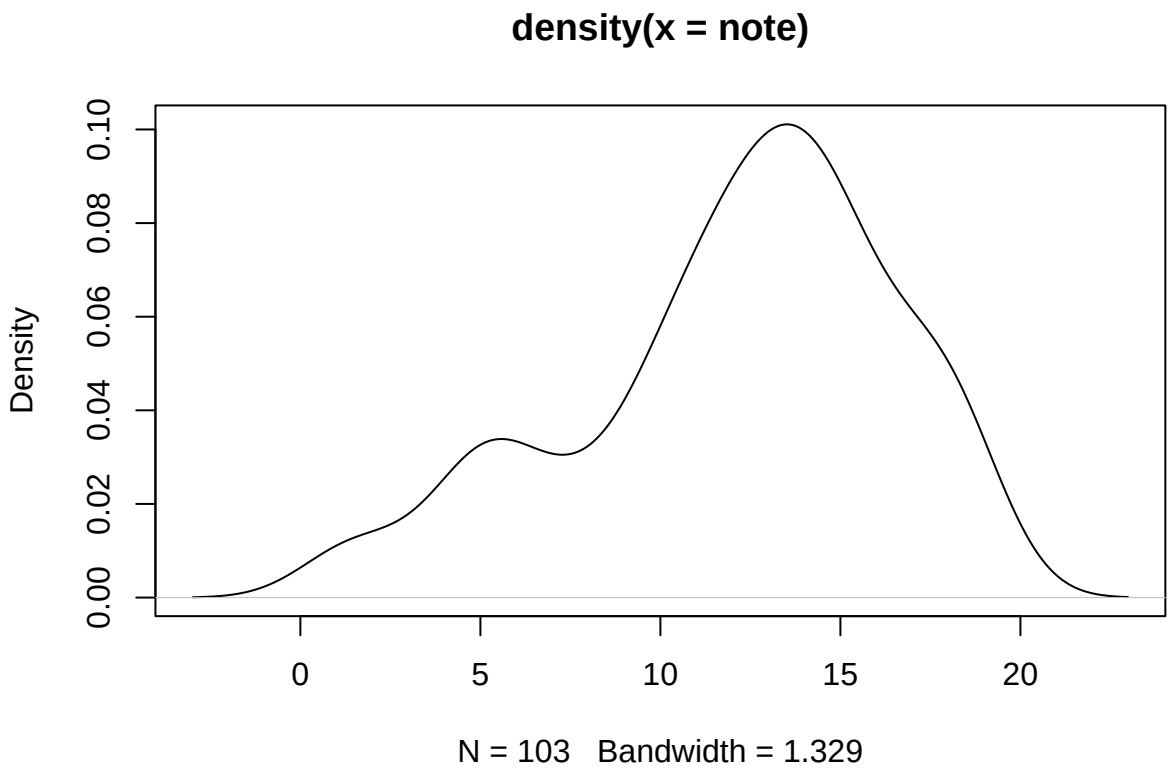
```
sujet=factor(sujet)
head(sujet)
```

```
## [1] 1 1 2 1 1 1
## Levels: 1 2
```

1.3.1 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)
```

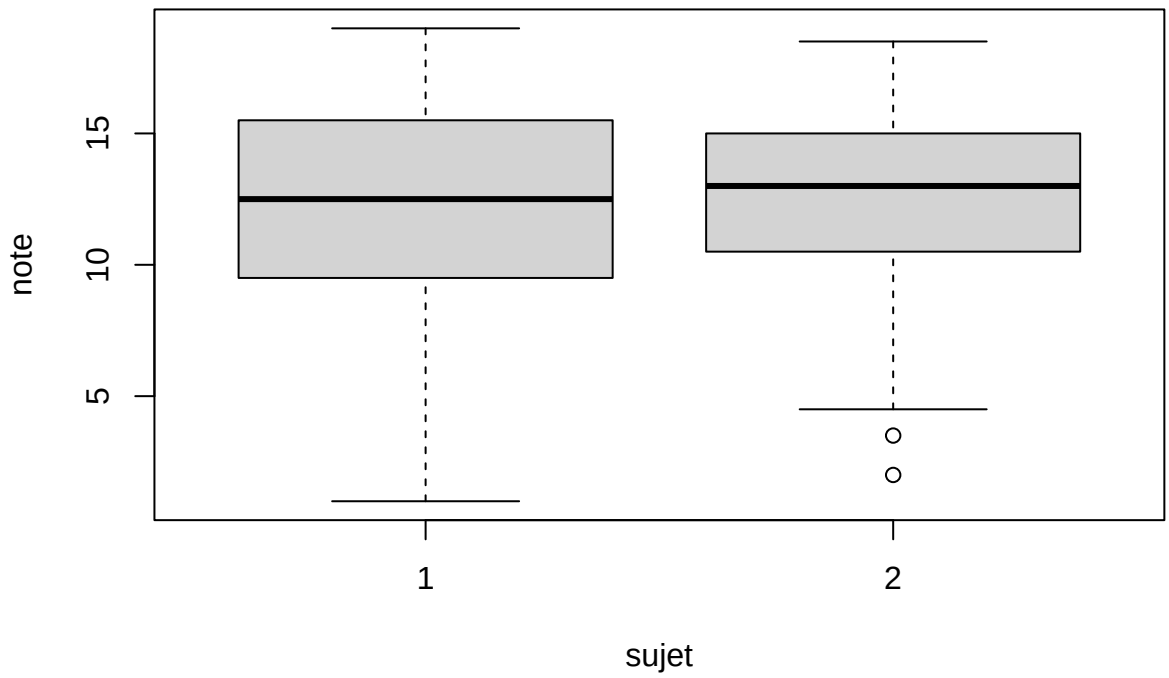
```
##
##  Shapiro-Wilk normality test
##
## data:  note
## W = 0.95003, p-value = 0.0006783
```

Réponse : La p−value est proche de zéro. On en d´eduit que l'on rejette significativement la normalité.

1.3.2 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.