HEIG-VD — PST

Laboratoire 1 – Compte-rendu

Loïc Herman

5 janvier 2023

1 Introduction

Le but de ce travail pratique est de se familiariser avec les commandes de base du logiciel **R** et d'utiliser des techniques d'analyse statistique pour une variable. Nous avons choisi de travailler avec **R** car c'est un logiciel de statistique populaire et puissant qui est largement utilisé dans de nombreux domaines. En réalisant plusieurs exercices avec **R** , nous espérons acquérir une meilleure compréhension de ses fonctionnalités et de la manière dont elles peuvent être utilisées pour l'analyse statistique.

En fin de compte, nous espérons que ce travail pratique nous aidera à mieux comprendre l'utilisation de **R** pour l'analyse statistique et à développer nos compétences dans ce domaine. Nous sommes impatients de mettre en pratique ce que nous avons appris et de voir comment ces techniques peuvent nous aider à mieux comprendre les données et à prendre des décisions éclairées.

2 Exercices

2.1 Exercice 1

L'objectif de cet exercice est de se familiariser avec les commandes de base du logiciel **R** afin de pouvoir manipuler des objets. Pour atteindre cet objectif, nous avons utilisé les fonctions `scan()` et `read.table()` pour charger les données nécessaires à notre travail pratique.

2.1.1 Import des données

Le code suivant utilise `scan()` pour lire un vecteur à partir du fichier `cpus.txt` et `read.table()` avec l'option `header=T` pour lire une table à partir du fichier `examen.txt`, avec la première ligne du fichier utilisée comme en-tête de colonne.

2.1.b Utiliser les commandes suivantes dans **R** pour charger les données :

```
cpus<-scan(paste(datadir, "cpus.txt", sep = "/"))
examen<-read.table(paste(datadir, "examen.txt", sep = "/"), header=T)
```

2.1.2 Lecture du contenu des données

Le code suivant utilise les méthodes élémentaires du logiciel **R** pour l'exploration des données chargées précédemment.

2.1.c Pour voir le contenu de l'objet `cpus`, taper l'instruction

```
cpus
```

```
##  [1]  77  18  22 144  12 185  38  24  45  38  65 141 208  58  12 636 397  66 915
## [20]  27  30  66  36  14  26  92   8  36 133  66  24  10  36 100  60  33  40  19
## [39]  16 140  63  21  32  64  24 110  16  56  46 144
```

2.1.d Pour accéder à la 10ème composante du vecteur `cpus`, utiliser la commande

```
cpus[10]
```

```
## [1] 38
```

2.1.e Pour obtenir une partie du vecteur `cpus` comme par exemple les éléments du vecteur compris entre la 3ème et la 17ème composante, taper l'instruction

```
cpus[3:17]

##  [1]  22 144  12 185  38  24  45  38  65 141 208  58  12 636 397
```

2.1.f Pour extraire du vecteur `cpus` ses éléments supérieurs à 185, utiliser la commande

```
cpus[cpus>185]

## [1] 208 636 397 915
```

2.1.g Pour afficher la composante `note` de l'objet `examen`, on peut utiliser la commande

```
examen$note

##  [1] 3.3 3.3 3.6 4.8  NA 2.6 3.9 4.4 4.4 4.8 5.2 3.8  NA 1.9 3.9 4.3 3.7 4.0 4.1
## [20] 4.7 3.6 2.7 3.5 4.8
```

2.1.h On peut aussi accéder en profondeur aux composantes comme par exemple par la commande

```
examen$note[7]

## [1] 3.9
```

2.1.i Création d'objets La méthode la plus simple pour créer un vecteur consiste à énumérer ses éléments à l'aide de la fonction `c()`

2.1.i.1 Un vecteur de données numériques

```
mesdonnees<-c(2.9, 3.4, 3.4, 3.7, 3.7, 2.8, 2.8, 2.5, 2.4)
mesdonnees

## [1] 2.9 3.4 3.4 3.7 3.7 2.8 2.8 2.5 2.4
```

2.1.i.2 Un vecteur de données textuelles

```
couleurs<-c("bleu", "vert", "blanc", "noir")
couleurs

## [1] "bleu"  "vert"  "blanc" "noir"
```

2.1.j On peut ôter des composantes d'un vecteur en indiquant entre crochets les indices précédés du signe négatif comme par exemple

```
mesdonnees[-c(3:5)]

## [1] 2.9 3.4 2.8 2.8 2.5 2.4
```

2.1.k Finalement, le contenu de votre environnement de travail est affiché à l'aide de la commande

```
ls()

## [1] "couleurs"   "cpus"       "datadir"    "examen"     "mesdonnees"
```

2.2 Exercice 2

L'objectif de cet exercice est de faire une analyse élémentaire des valeurs de la performance relative au processeur `IBM 370/158-3` afin de consolider notre connaissance des valeurs observées et des variables statistiques.

2.2.a Construire un diagramme branche-et-feuilles, un histogramme et une boîte à moustaches des données observées à l'aide des commandes

2.2.a.1 Création d'un diagramme branche-et-feuilles

```
stem(cpus)

##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 11111222222222333334444445566666777789
##   1 | 01344449
##   2 | 1
##   3 |
##   4 | 0
##   5 |
##   6 | 4
##   7 |
##   8 |
##   9 | 2
```

2.2.a.2 Création d'un histogramme avec une boîte à moustaches

```
par(mfrow=c(1,2), pty="s")
hist(cpus, xlab="performance relative", ylab="fréquence", main="", col="darkslategray4")
rug(cpus)
boxplot(cpus, xlab="performance relative", col="darkslategray4", horizontal=T)
par(mfrow=c(1,1))
```

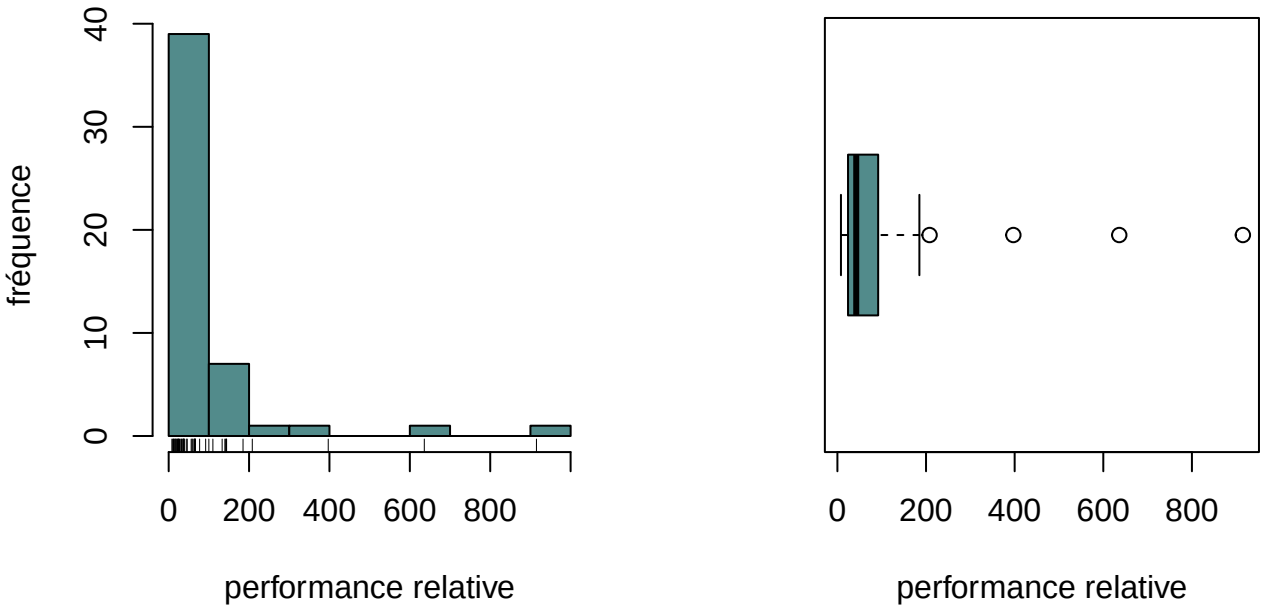


FIGURE 1 – Un histogramme et une boîte à moustaches des données observées

2.2.b Commentaires

L'analyse de la distribution des données montre une asymétrie positive, c'est-à-dire que la queue de la distribution à droite est plus longue que celle à gauche. De plus, il y a quelques valeurs qui semblent atypiques, c'est-à-dire qui sont éloignées de la plupart des autres valeurs. Ces valeurs se trouvent généralement après environ 200.

2.2.c Calculer la performance relative médiane, la performance relative moyenne et le(s) mode(s) des valeurs observées

2.2.c.1 Calcul de la performance relative médiane

```
median(cpus)

## [1] 42.5
```

2.2.c.2 Calcul de la performance relative moyenne

```
mean(cpus)

## [1] 93.78
```

2.2.c.3 Calcul de la mode trimodale, les valeurs les plus observées

```
n.cpus <- table(cpus)
as.numeric(names(n.cpus)[n.cpus == max(n.cpus)])

## [1] 24 36 66
```

2.2.c.4 Questions

L'analyse des statistiques descriptives montre que la moyenne et la médiane sont très différentes l'une de l'autre. Cela indique la présence de valeurs atypiques dans les données, c'est-à-dire de valeurs qui sont éloignées de la plupart des autres valeurs.

Il est donc plus approprié d'utiliser la médiane, qui accordera moins d'importance à ces valeurs atypiques.

2.2.d Que fait la commande suivante ?

La commande `summary()` est une fonction de base de **R** qui permet de calculer et d'afficher un résumé des principales statistiques descriptives d'un objet. Dans le cas présent, la commande `summary(cpus)` calcule et affiche un résumé des statistiques descriptives de la variable `cpus`.

Plus précisément, cette commande calcule et affiche la moyenne, la médiane, le minimum, le maximum, ainsi que le premier et le troisième quartile.

```
summary(cpus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   24.00   42.50   93.78   88.25  915.00
```

2.2.e En effectuant aucun calcul, décrire l'effet sur la moyenne et sur la médiane des trois interventions suivantes

2.2.e.1 ajouter un processeur de performance relative 45

L'ajout d'une valeur de 45 à la distribution des données ne devrait pas avoir un grand impact sur la médiane et la moyenne. En effet, comme 45 se trouve environ au milieu de la distribution, cela signifie que cette valeur n'est pas très éloignée de la plupart des autres valeurs. Ainsi, l'ajout de cette valeur ne va pas provoquer un décalage important de la médiane et de la moyenne. Toutefois, cela pourrait avoir un impact sur l'étendue de la distribution et sur la variabilité des données.

2.2.e.2 soustraire 9 à chaque valeur observée

Le déplacement de la médiane et de la moyenne de 9 unités indique que ces mesures de tendance centrale sont sensibles à la position des valeurs dans la distribution des données. En effet, lorsque la médiane et la moyenne sont décalées de manière importante, cela signifie que les valeurs de la distribution ont été modifiées de manière significative par rapport à leur position initiale.

Par conséquent, si la médiane et la moyenne étaient initialement de 43 et que l'on soustrait 9 à chaque valeur de la distribution, cela va déplacer la médiane et la moyenne à 34.

2.2.e.3 diviser chaque observation par 4

La division de la moyenne et de la médiane par 4 indique que les valeurs de la distribution ont été modifiées en changeant l'échelle de mesure. Cette opération s'appelle une homotétie, c'est-à-dire une transformation qui conserve les proportions entre les éléments d'un objet tout en modifiant leur taille. Lorsqu'une homotétie est appliquée à une distribution de données, la moyenne et la médiane sont divisées par le même coefficient, ce qui a pour effet de réduire l'étendue de la distribution.

Par exemple, si la moyenne et la médiane étaient initialement de 43 et que l'on divise chaque valeur de la distribution par 4, cela va diviser la moyenne et la médiane par 4 et réduire l'étendue de la distribution de 4 fois.

2.2.f Déterminer l'écart-type des performances relatives une fois avec les valeurs atypiques et une fois sans en utilisant les commandes

```
sd(cpus)

## [1] 158.3789

sd(cpus[cpus<=185])

## [1] 43.91173
```

2.2.f.1 Questions

L'écart-type est une mesure de la variabilité ou de la dispersion des données autour de la moyenne. Comme il est basé sur la moyenne, il est fortement influencé par les valeurs extrêmes de la distribution. En effet, lorsque les données comportent des valeurs atypiques ou éloignées de la plupart des autres valeurs, cela peut avoir un impact important sur la moyenne et, par conséquent, sur l'écart-type.

Ainsi, l'écart-type peut être considéré comme moins robuste que d'autres mesures de dispersion, comme par exemple l'écart interquartile, qui est moins sensible aux valeurs atypiques. Cela signifie que l'écart-type peut être moins stable et moins fiable que d'autres mesures de dispersion lorsque les données comportent des valeurs atypiques.

2.3 Exercice 3

L'exercice consiste à analyser les résultats de l'examen de fin d'unité d'un cours de Probabilités et Statistique passé par 24 étudiants répartis en deux groupes, en emploi et à plein temps, et à déterminer si les résultats des deux groupes sont similaires ou différents.

2.3.a Tracer les boîtes à moustaches en parallèle en utilisant les commandes suivantes :

```
note.EE<-split(examen$note, examen$mode_de_formation)$EE
note.PT<-split(examen$note, examen$mode_de_formation)$PT
lblue<-"#528B8B"
par(pty="s")
boxplot(note~mode_de_formation, data=examen, ylim=c(1,6),
        xlab="mode de formation", varwidth=T, col=lblue, main="examen")
abline(h=4, lty=2)
rug(note.EE, side=2)
rug(note.PT, side=4)
```

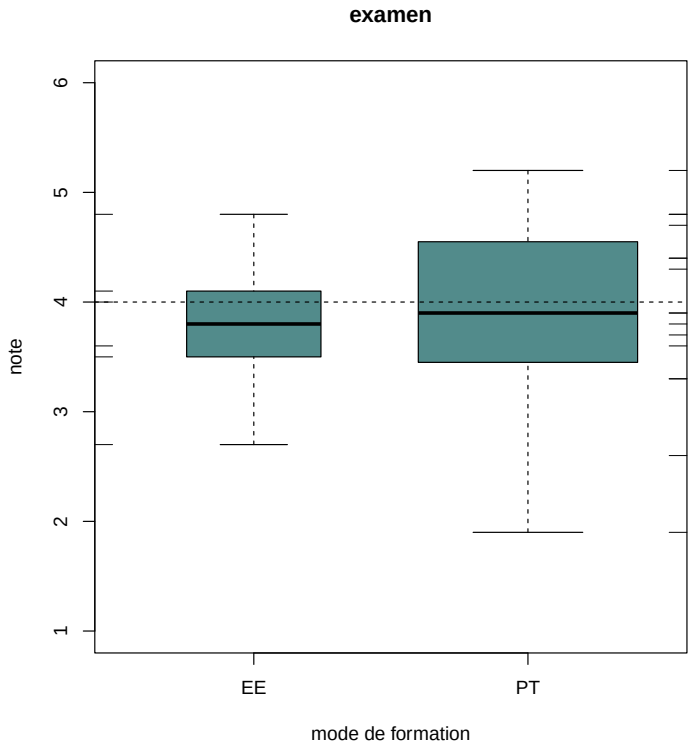


FIGURE 2 – Boîte à moustache des notes par mode de formation

2.3.a.1 Questions

En analysant les résultats de l'examen de fin d'unité pour les deux groupes d'étudiants, on constate qu'il n'y a pas de différence significative en termes de position. Cela signifie que les notes moyens des deux groupes sont similaires et que la distribution des notes est centrée autour de valeurs proches.

Toutefois, on observe une différence significative de dispersion entre les deux groupes. Cela signifie que la variabilité des notes est plus importante dans un des groupes que dans l'autre. Pour illustrer cette différence, on peut utiliser une mesure de dispersion comme l'écart-type ou l'écart interquartile. Si l'écart-type est plus grand pour un des groupes, cela signifie que la dispersion des notes est plus importante dans ce groupe. De même, si l'écart interquartile est plus grand pour un des groupes, cela indique que les notes sont plus dispersés dans ce groupe.

Cette différence de dispersion peut être due au fait qu'il y a plus d'étudiants à plein temps dans l'échantillon. En effet, plus il y a de données, plus l'étendue de la distribution est grande, toutes choses égales par ailleurs. Mais, cette justification n'est pas robuste.

2.3.b Observe-t-on sur les boîtes à moustaches une différence entre les dispersions des deux groupes ?

L'étendue de la distribution des résultats des étudiants à plein temps est beaucoup plus grande que celle des résultats des étudiants en emploi. Cela peut être interprété comme une indication que les résultats des étudiants à plein temps sont plus variables ou plus dispersés autour de la médiane que ceux des étudiants en emploi.

2.3.c Calculer les écarts-types des deux groupes à l'aide de la commande `sd()`

```
sd(note.EE, na.rm=TRUE)

## [1] 0.7026142
```

```r
sd(note.PT, na.rm=TRUE)

## [1] 0.8624577
```

2.3.c.1 Questions

En comparant les résultats de l'examen de fin d'unité des étudiants en emploi et à plein temps, on observe que la dispersion des données est significativement plus grande dans le groupe à plein temps. Dispersion ici désigne la variabilité ou l'étendue des données autour de la médiane. Cette différence de dispersion peut être expliquée par le fait qu'il y a plus d'étudiants à plein temps dans l'échantillon. Plus il y a de données, plus l'étendue de la distribution est grande, toutes choses égales par ailleurs. Ainsi, la différence de dispersion observée entre les deux groupes peut être attribuée à la différence de taille des échantillons. Il n'y a pas d'autre différence notable entre les deux groupes.

2.3.d Que peut-on déduire en comparant les conclusions établies en 2.3.b et en 2.3.c

En analysant les résultats de l'examen de fin d'unité des étudiants en emploi et à plein temps, on observe que la dispersion des données est significativement plus grande dans le groupe à plein temps.

2.3.e Un autre graphique pour étudier les éventuelles différences entre les deux groupes à l'examen de fin d'unité est proposé. À votre avis, entre les boîtes à moustaches en parallèle et le graphique tracé ci-dessus, lequel est le plus approprié ?

Le deuxième graphique présente les données sous forme de histogramme, c'est-à-dire en regroupant les valeurs par intervalle de classe et en représentant le nombre de valeurs dans chaque intervalle par une colonne. Un histogramme permet de visualiser la distribution des données et de se rendre compte de leur répartition. Il permet également de mettre en évidence les valeurs extrêmes, c'est-à-dire les valeurs les plus élevées et les plus basses de l'ensemble des données.

Le premier graphique, en revanche, présente les données sous forme de diagramme en boîte (aussi appelé "boîte à moustaches"). Un diagramme en boîte permet de représenter la distribution des données et de mettre en évidence certains aspects de cette distribution, comme la médiane, les quartiles et les valeurs extrêmes.

Dans ce cas précis, le deuxième graphique est considéré comme plus pertinent car il permet de mieux visualiser la distribution des notes et de mettre en évidence les valeurs extrêmes. Le diagramme en boîte, en revanche, ne permet pas de se rendre compte de la répartition des données et ne met pas en évidence les valeurs extrêmes de la même manière.

2.4 Exercice 4

Dans cet exercice, nous allons utiliser `ggplot2` pour générer des graphiques.

2.4.a Installation

```r
library(ggplot2)
library(arules, warn.conflicts=F, quietly=T)
```

2.4.a.1 Import des données

La commande `colnames(dframe) \<-c("education", "hours_per_week")` permet de renommer les colonnes de `dframe`.

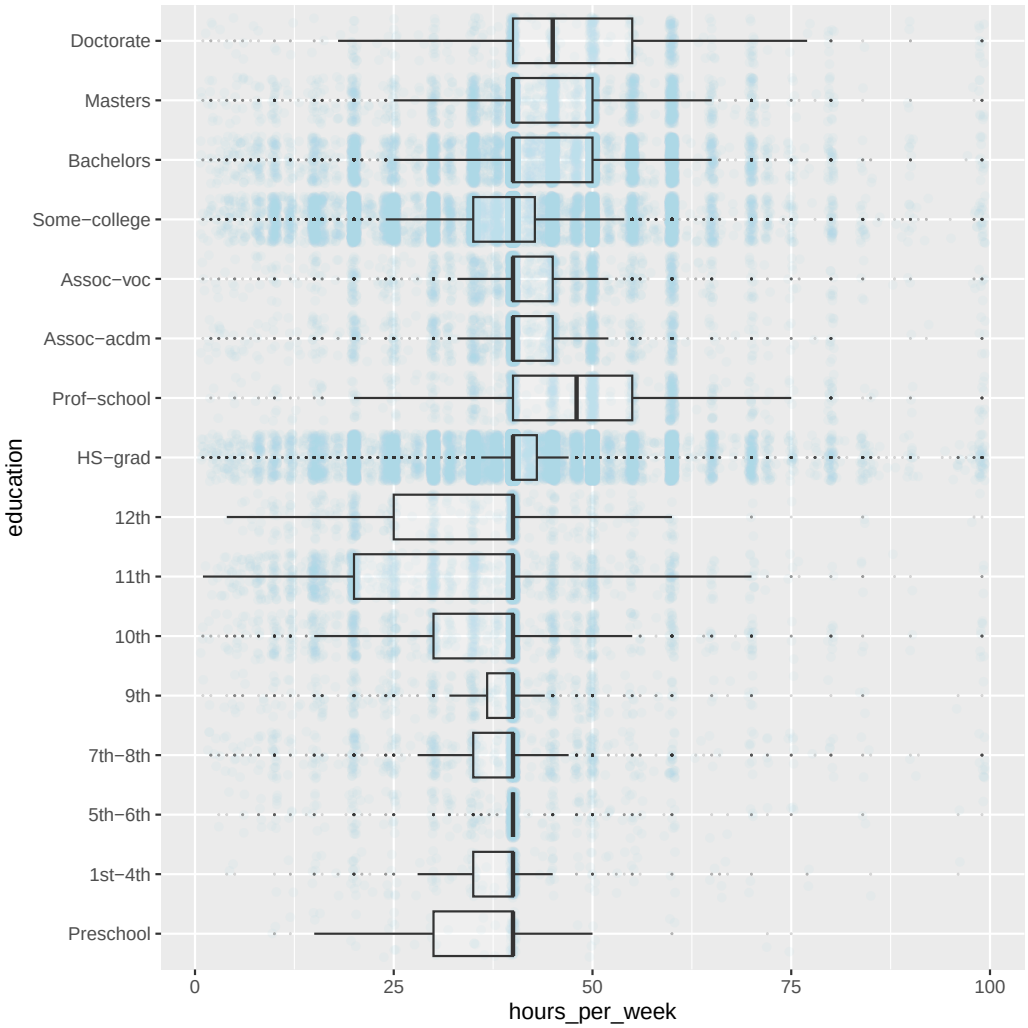
```
data("AdultUCI")
dframe<-AdultUCI[, c("education", "hours-per-week")]
colnames(dframe)<-c("education", "hours_per_week")
str(dframe)

## 'data.frame': 48842 obs. of  2 variables:
##  $ education     : Ord.factor w/ 16 levels "Preschool"<"1st-4th"<..: 14 14 9 7 14 15 5 9 15 14
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
```

2.4.a.2 Graphique

Le graphique en question est une combinaison de deux types de représentation graphique : une boîte à moustaches et un nuage de points. La boîte à moustaches permet de visualiser la distribution des données et de mettre en évidence certains aspects de cette distribution, comme la médiane, les quartiles et les valeurs extrêmes. Elle est particulièrement utile pour avoir une vue d'ensemble de la répartition des données. Le nuage de points, quant à lui, permet de représenter l'ensemble des données individuellement sous forme de points sur un graphique en deux dimensions. Il permet de mettre en évidence la densité des données et de visualiser les valeurs extrêmes de manière plus détaillée.

En combinant ces deux types de représentation graphique sur la même ligne, le graphique obtenu est très complet et permet d'avoir une bonne idée de la répartition générale des données, autant sur le plan global (grâce à la boîte à moustaches) que sur le plan détaillé (grâce au nuage de points). Il est donc très utile pour comprendre la distribution des données et pour identifier les tendances et les particularités de cette distribution.

```
ggplot(dframe, aes(x=education, y=hours_per_week)) +
geom_point(colour="lightblue", alpha=0.1, position="jitter") +
geom_boxplot(outlier.size=0, alpha=0.2) + coord_flip()
```



2.4.b Pour quel type de formation observe-t-on la plus grande dispersion du temps de travail ? Existe-t-il une différence entre les médianes des types de formation ? En donner brièvement la raison.

"11th" a la plus grande dispersion du temps de travail.

Sauf pour les valeurs de "Doctorate" et "Prof-school", les médianes sont pratiquement égales.

Dans cette étude, les données portent sur le nombre d'heures travaillées par semaine par des individus de différents niveaux de formation. Si les médianes du nombre d'heures travaillées sont similaires pour la plupart des niveaux de formation, cela peut s'expliquer par le fait qu'il y a peut-être un certain nombre d'individus qui travaillent un nombre standard d'heures par semaine, comme par exemple 40 heures. Cette hypothèse peut être vérifiée en regardant la distribution des données et en analysant si une proportion importante des individus travaille environ 40 heures par semaine. Si c'est le cas, cela pourrait expliquer pourquoi les médianes sont similaires. Il serait également intéressant de vérifier si cette hypothèse tient compte de facteurs tels que le genre, l'âge et la profession des individus, qui peuvent également influencer le nombre d'heures travaillées par semaine.

2.4.c Pour chaque type de formation, on peut déterminer puis afficher à l'écran le temps maximal de travail hebdomadaire à l'aide des commandes

On remarque que chaque variable a une valeur maximale de 99, ce qui pourrait indiquer que c'est une valeur qui remplace les observations pour lesquelles la valeur est manquante.

```
nx<-tapply(dframe$hours_per_week, dframe$education, max, na.rm=T)
nx

##    Preschool      1st-4th      5th-6th      7th-8th          9th         10th
##           75           96           99           99           99           99
##         11th         12th      HS-grad  Prof-school   Assoc-acdm    Assoc-voc
##           99           99           99           99           99           99
## Some-college    Bachelors      Masters    Doctorate
##           99           99           99           99

max(nx)

## [1] 99

names(nx)[nx==max(nx)]

##  [1] "5th-6th"      "7th-8th"      "9th"          "10th"         "11th"
##  [6] "12th"         "HS-grad"      "Prof-school"  "Assoc-acdm"   "Assoc-voc"
## [11] "Some-college" "Bachelors"    "Masters"      "Doctorate"
```

2.4.d En s'inspirant des commandes utilisées en 2.4.c, déterminer la formation pour laquelle la distribution des temps de travail se caractérise par le plus petit écart-type

```
nx.sd<-tapply(dframe$hours_per_week, dframe$education, sd, na.rm=T)
names(nx.sd)[nx.sd==min(nx.sd)]

## [1] "Assoc-voc"
```

2.4.e Observe-t-on un résultat similaire en utilisant l'étendue interquartiles à l'aide de la fonction `IQR()` ?

```
nx<-aggregate(hours_per_week~education, dframe, IQR, na.rm=T)
min(nx[,2])

## [1] 0

nx[,1][nx[,2]==min(nx[,2])]

## [1] 5th-6th
## 16 Levels: Preschool < 1st-4th < 5th-6th < 7th-8th < 9th < 10th < ... < Doctorate
```

L'écart interquartile (aussi appelé "intervalle interquartile" ou "IQR") est une mesure de dispersion qui s'applique à une distribution de données quantitatives. Elle correspond à la différence entre le premier quartile (Q1) et le troisième quartile (Q3) de la distribution. L'IQR est une mesure robuste qui prend en compte la majorité des données, c'est-à-dire les données comprises entre le premier quartile (Q1) et le troisième quartile (Q3). Elle ne prend pas en compte les valeurs extrêmes, c'est-à-dire les valeurs situées en dehors de l'intervalle Q1-Q3.

L'écart-type, quant à lui, est une autre mesure de dispersion qui s'applique à une distribution de données quantitatives. Il correspond à la racine carrée de la variance de la distribution. L'écart-type prend en compte

toutes les valeurs de la distribution, y compris les valeurs extrêmes. Il est donc moins robuste que l'IQR car il est plus sensible aux valeurs extrêmes qui peuvent influencer fortement la valeur de l'écart-type.

Donc, non, les résultats ne seront pas les mêmes.

2.5 Exercice 5

Dans cet exercice, nous allons nous intéresser à deux librairies de R : `plotly` et `magrittr`.

La librairie `plotly` permet de construire des diagrammes en boîte (aussi appelés "boîtes à moustaches") accompagnés par les statistiques élémentaires usuelles.

La librairie `magrittr`, quant à elle, fournit un mécanisme grâce auquel des commandes peuvent être efficacement enchaînées.

Nous verrons comment utiliser ces deux librairies pour manipuler des objets de **R** de manière efficace et produire des graphiques de qualité.

```
library(plotly)
library(magrittr)
```

```
plot_ly(examen, y=~note, x=~mode_de_formation, type="box") %>%
  layout(yaxis=list(range=c(1, 6)))
```

3 Conclusion

Pour conclure, ce travail pratique nous a permis de nous familiariser avec les commandes de base du logiciel **R** et d'utiliser des techniques d'analyse statistique pour une variable. Nous avons réalisé plusieurs exercices qui nous ont aidé à mieux comprendre comment utiliser **R** pour l'analyse de données et à développer nos compétences dans ce domaine.

En général, ce travail pratique nous a permis de mieux comprendre l'utilisation de **R** pour l'analyse statistique et de développer nos compétences dans ce domaine. Nous avons maintenant une base solide pour continuer à utiliser **R** et à explorer d'autres fonctionnalités de ce logiciel.

En outre, ce travail pratique nous a montré l'importance de bien comprendre les données et d'utiliser des techniques statistiques appropriées pour analyser ces données. En effet, une mauvaise interprétation des données peut conduire à des conclusions erronées et à des décisions inappropriées.

Enfin, ce travail pratique nous a donné l'occasion de mettre en pratique ce que nous avons appris au long du cours de Probabilités et Statistiques de la HEIG-VD et de voir comment ces techniques peuvent nous aider à mieux comprendre les données en prenant des décisions éclairées.