Rapport PST - TP1

Alexis Martins

2023-01-04

Introduction

Ce rapport concerne le premier et le seul travail pratique que nous avons réalisé durant le cours de PST (Probabilités et statistiques) de l'HEIG-VD.

Ce travail pratique avait plusieurs buts. Dans un premier temps que nous puissions découvrir un outil très répandu dans le milieu de la donnée, mais aussi dans le milieu scientifique en général. Cet outil s'appelle R et permet de réaliser diverses opérations sur les données, les afficher dans différents contextes (graphiques) et aussi de réaliser des documents comme des rapports.

Dans un second temps, ce TP nous a permis de faire un parallèle avec le cours pour mettre en pratique ce que nous avions vu. Notamment pour les analyses des graphiques en comprenant ce qu'indiquent les différents indicateurs ou lors de la visualisation de données directement.

Exercice 1

Ce premier exercice était surtout présent pour la prise en main de R et l'explication de certaines commandes. Il n'y avait aucune partie d'analyse contrairement aux parties qui suivent. Cette partie ne sera donc pas traitée.

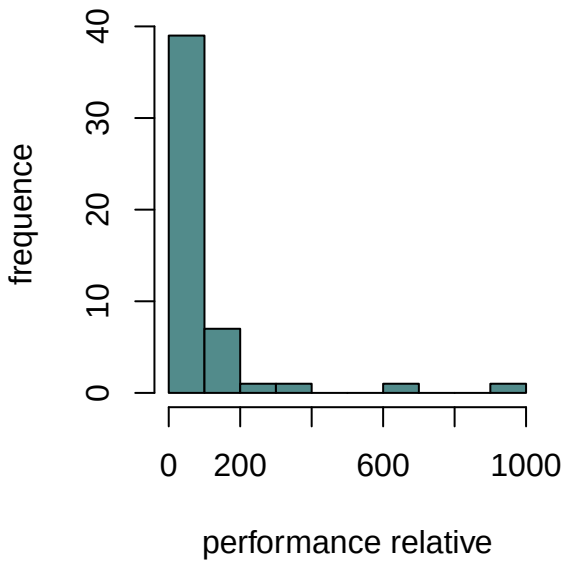
Exercice 2

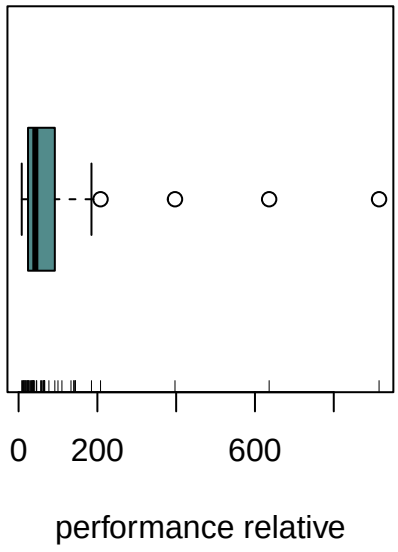
Dans ce second exercice, on va créer nos premiers graphiques et surtout réaliser nos premières analyses.

a)

Consigne : Construire un diagramme branche-et-feuilles, un histogramme et une boîte à moustaches des données observées à l'aide des commandes ci-dessous.

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 1111122222222233333444444566666777789
##   1 | 01344449
##   2 | 1
##   3 |
##   4 | 0
##   5 |
##   6 | 4
##   7 |
##   8 |
##   9 | 2
```





Ces trois graphiques représentent les mêmes données :

- Le premier est un diagramme branche-et-feuilles, il permet de voir la distribution des données de façon textuelle. Pour retrouver les valeurs correspondantes, le chiffre à gauche de la barre représente les centaines et chaque chiffre à droite de la barre est à prendre de façon individuelle et à multiplier par 10 pour avoir les dizaines et les unités. Les valeurs résultantes sont alors uniquement des approximations multiples de 10.
- Le second graphique est un histogramme, il permet de voir la dispersion et la fréquence des données pour certaines valeurs.
- Le dernier est une boîte à moustaches, elle est très utile pour voir en un coup d'oeil la dispersion des données et la position des différents indicateurs (médiane, quartiles, min, etc...).

b)

Consigne : Commenter la distribution des valeurs observées : valeur(s) atypique(s), asymétrie

Grâce au diagramme branche-et-feuilles ou à l'histogramme, on observe directement que le graphe est asymétrique positif et qu'il est unimodal. La grande majorité des valeurs sont concentrées dans cette partie gauche du graphe. On retrouve quelques valeurs atypiques lorsque l'on passe la valeur 200 sur l'axe des abscisses.

PS : Si on creuse un peu plus avec les valeurs, on remarquera qu'il est en réalité plûtot trimodal.

c)

Consigne : Calculer la performance relative médiane, la performance relative moyenne et le(s) mode(s) des valeurs observées. Est-il plus approprié d'utiliser la médiane ou la moyenne ?

```
## mediane : 42.5
```

```
## mean : 93.78
```

```
## modes : 24 36 66
```

On remarque directement une différence assez grande entre la médiane et la moyenne. Si on ne connaissait pas les données, on pourrait déterminer qu'il y a sûrement des valeurs atypiques présentes dans le set. On ayant vu les graphiques précédemment, on sait alors que c'est le cas. Cela explique la raison pour laquelle la moyenne est aussi haute comparée à la médiane.

On peut donc déterminer que la médiane est un indicateur beaucoup plus approprié dans ce cas. Elle l'est d'ailleurs dans la majorité des cas étant donné qu'elle est moins affectée par des valeurs atypiques. Elle est donc plus fiable étant donnée sa robustesse accrue.

d)

Consigne : Que fait la commande suivante ?

```
summary(cpus)
```

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

Summary signifie résumé lorsqu'on le traduit en français. Cette commande va alors faire un résumé du dataset. C'est-à-dire qu'elle va nous afficher les principaux indicateurs que l'on aimerait savoir sur un dataset. Dans ce cas, elle affiche notamment la moyenne, la médiane, les quartiles, min, etc...

e)

Consigne : En effectuant aucun calcul, décrire l'effet sur la moyenne et sur la médiane des trois interventions suivantes :

1. Ajouter un processeur de performance relative 45

La moyenne va légèrement diminuer, car celle-ci est plus haute que 45. La médiane ne va que très peu changer étant donné qu'elle est déjà très proche de la valeur rajoutée. Si on regarde d'ailleurs le set de plus près, on remarque que 45 sera la nouvelle médiane.

2. Soustraire 9 à chaque valeur observée

Les valeurs de la moyenne et de la médiane vont diminuer de neuf unités.

3. Diviser chaque observation par 4

Les valeurs de la moyenne et de la médiane vont se retrouver divisées par quatre.

f)

Consigne : Déterminer l'écart-type des performances relatives une fois avec les valeurs atypiques et une fois sans.

```
## Ecart-type avec toutes les valeurs : 158.3789
```

```
## Ecart-type sans les valeurs atypiques : 43.91173
```

L'écart-type n'est pas du tout un indicateur robuste. On remarque qu'en enlevant simplement les quelques valeurs atypiques celui-ci est presque divisé par quatre. On peut expliquer cela en regardant la manière dont il est calculé.

$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

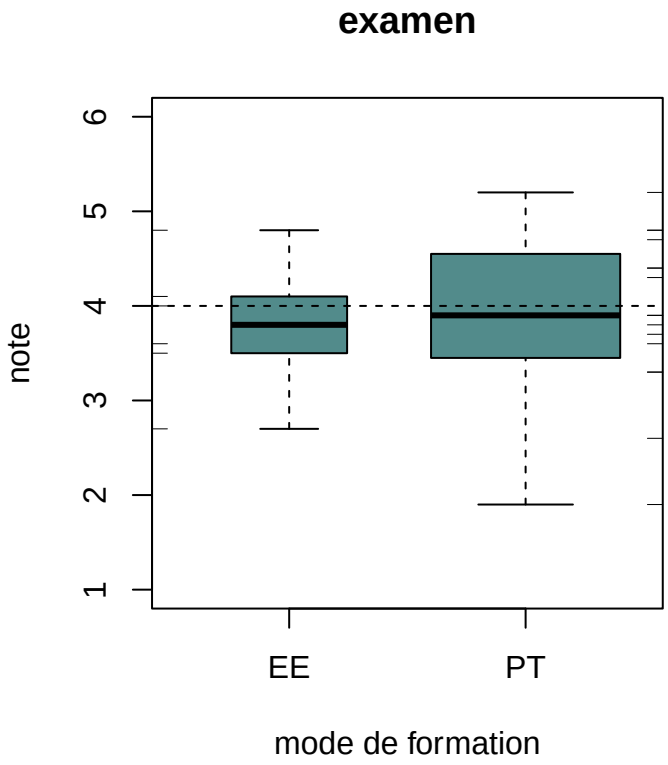
C'est donc normal qu'il varie, car c'est simplement la différence de la moyenne avec les observations le tout élevé au carré. La moyenne n'étant déjà pas un indicateur très robuste de base, cela se répercute sur l'écart-type.

Exercice 3

Dans ce troisième exercice, on va se concentrer sur la comparaison de deux boîtes à moustaches. Il s'agira surtout de comparer leur dispersion l'une par rapport à l'autre. On terminera par la comparaison avec un autre type de graphique pour définir lequel est le plus adapté à la situation.

a)

Consigne : En se basant sur ce graphique, existe-t-il une différence significative entre les deux groupes à l'examen de fin d'unité ?



Dans un premier temps, on peut comparer leur hauteur. Les deux boîtes ont à peu près la même hauteur, le changement est surtout sur leur largeur. La médiane quant à elle est aussi extrêmement proche. Comme dit précédemment, ce que l'on distingue immédiatement c'est la dispersion des données dans les deux boîtes.Les EE sont beaucoup plus compactés, tandis que les PT sont répandus que ça soit dans la largeur de la boîte ou dans les moustaches. Il ne faut tout de même pas oublier que le nombre d'observations pour chaque groupe n'était pas similaire, cela peut jouer sur la représentation.

b)

Consigne : Observe-t-on sur les boîtes à moustaches une différence entre les dispersions des deux groupes ?

Oui, une dispersion plus grande est observée du côté des PT. Le premier quartile est la médiane de chaque boîte son assez similaire, ce qui n'est pas le cas pour le troisième quartile ou la taille des moustaches.

c)

Consigne : En se basant sur les écarts-types, existe-t-il une différence en dispersion entre les deux groupes à l'examen de fin d'unité ?

```
## Ecart-type pour les EE : 0.7026142
```

```
## Ecart-type pour les PT : 0.8624577
```

L'écart-type des PT est légèrement plus grand que celui des EE. Cela veut dire que les PT sont plus dispersés que les EE, mais ce n'est pas une grande différence. Encore une fois, cela pourrait être dû au fait qu'il y ait plus de monde en PT, donc plus de chance d'avoir de la dispersion.

d)

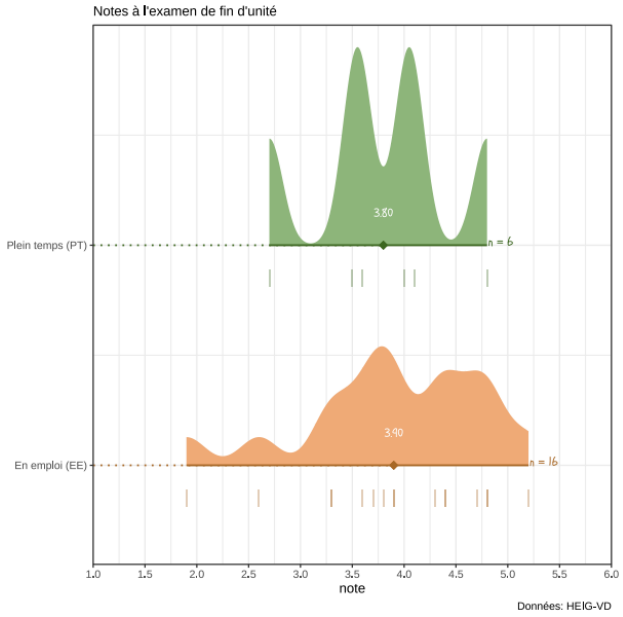
Consigne : Que peut-on déduire en comparant les conclusions établies en b) et en c) ?

Que les deux classes sont assez similaires au niveau des résultats. On observe légèrement une plus grande dispersion pour les PT, mais rien d'alarmant.

e)

Consigne : Un autre graphique pour étudier les éventuelles différences entre les deux groupes à l'examen de fin d'unité se trouve ci-dessous.

À votre avis, entre les boîtes à moustaches en parallèle et le graphique tracé ci-dessus, lequel est le plus approprié ?



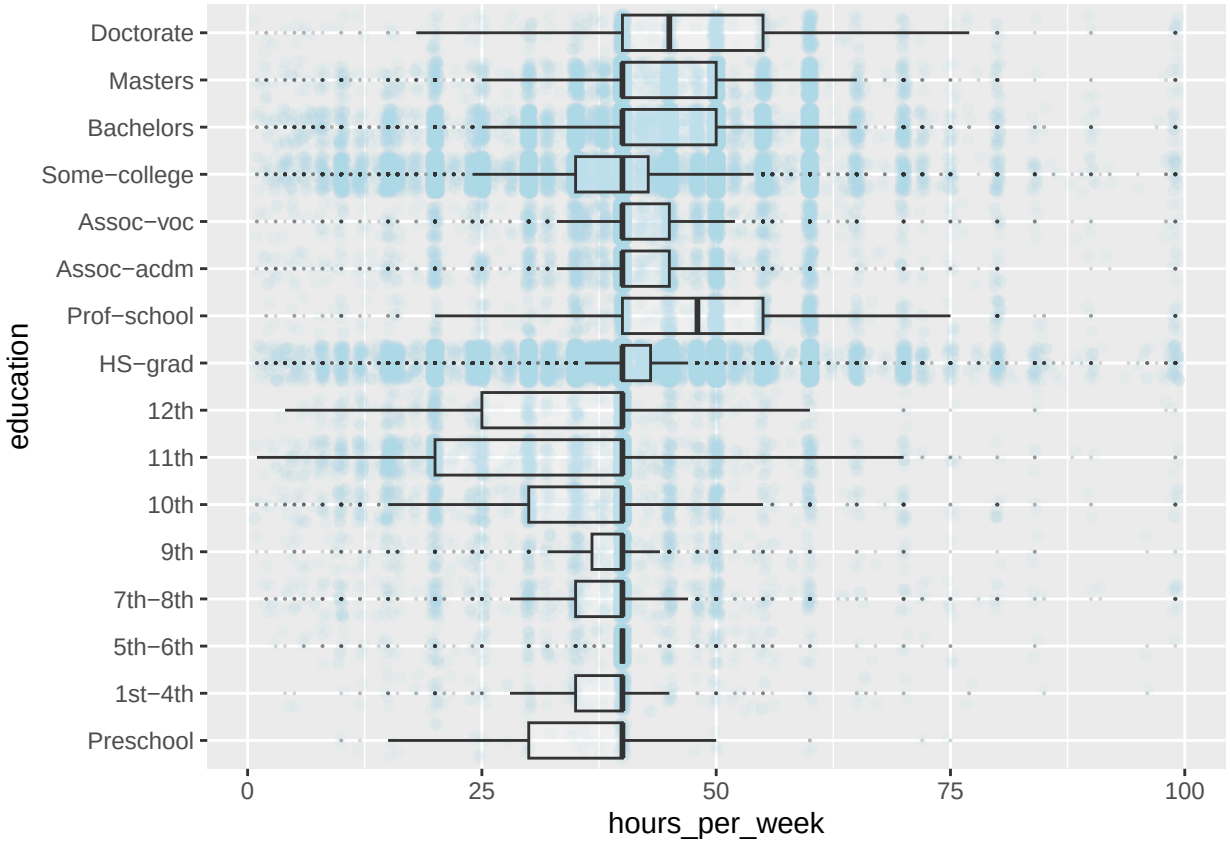
Personnellement je trouve les boîtes à moustaches plus adaptées. Je n'aime pas trop la représentation en courbe, je trouve que l'on perd de l'information notamment au niveau des quartiles. Alors que dans une boîte à moustaches, on voit directement les indicateurs tels que les quartiles, médianes, etc...

Le second graphique est peut-être adapté si on souhaite avoir une idée rapide sur la modalité de notre graphique et la répartition de façon vague. Mais si on souhaite avoir des données plus justes et praticables pour de l'analyse, je pense que les boîtes à moustaches restent plus appropriées.

Exercice 4

a)

Consigne : Commenter le graphique obtenu



C'est un graphique composé de multiples autres graphiques. Dans ce cas, ce sont des boîtes à moustaches classiques, mais on aperçoit des tracés bleutés qui indiquent la concentration d'observations à cet instant.

On voit alors qu'il y a différents types de scolarités (16 catégories) et la boîte à moustaches représente le nombre d'heures que chaque type travaille par semaine. Le graphique permet donc de comparer si un certain type de scolarité impacte ou non le nombre d'heures réalisées par un individu sur une période d'une semaine lorsque celui-ci est dans la vie active.

A première vue, on remarque surtout que pour presque tous les groupes la médiane est placée au même endroit. On remarque aussi que l'on pourrait former deux groupes distincts selon le placement des boîtes. La séparation se situant entre HS-grad et 12th.

Le premier groupe est composé des individus entre Preschool et 12th et on remarque que ~75% des observations se situent sous la médiane. Pour le second groupe composé des individus entre HS-grad et Doctorate, on remarque plutôt que ~75% des observations se situent au-dessus de la médiane. Exception faite pour les groupes Prof-school, Some-college et Doctorate.

Pour les deux groupes où la médiane n'est pas au même endroit, c'est-à-dire Doctorate et Prof-school, on pourrait expliquer cela grâce au type de travail que fournissent ces deux groupes. Ils doivent souvent travailler durant les heures "classiques", mais par exemple pour les enseignants, une partie du travail est aussi réalisée en dehors de ces heures. Par exemple, pour de la correction, suivi de travaux, etc...

Finalement du point de vue de la dispersion et des valeurs atypiques, qui sont représentées par les zones

bleues sur le graphique. On remarque que tous les groupes sont plutôt constants, mais on retrouve tout de même deux exceptions qui sont les Some-college et HS-grad.

PS question dans le code : Renommer les heures par semaines "-" -> "_". Pas de confusion au niveau de l'opération arithmétique

b)

Consigne : 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

La plus grande dispersion se joue entre les HS-Grad et les Some-college. La médiane est similaire pour toutes les catégories, sauf pour les doctorants et les profs-school où elle est plus élevée. Cela indique qu'il y a plus de valeurs élevées, donc plus de gens qui travaillent plus d'heures. Les explications se retrouvent dans la réponse à la question précédente.

c)

Consigne : Pour chaque type de formation, on peut déterminer puis afficher à l'écran le temps maximal de travail hebdomadaire, est-ce surprenant ?

```
## Temps maximum par type de formation :

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

##
## Formation(s) avec le temps maximum ( 99 ) :

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

Non, car les données sont très diversifiées et on en a énormément. Il se peut donc peut-être que certaines de ces valeurs maximales soient vraies, mais il est aussi très probable que ça soit des valeurs atypiques. Je pense que même pour des personnes avec un agenda très chargé, il est rare de travailler autant.

d)

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

```
## Temps maximum par type de formation :
```

```
## Preschool    :  11.434
## 1st-4th      :  12.22667
## 5th-6th      :  11.37219
## 7th-8th      :  14.56277
## 9th          :  11.46478
## 10th         :  13.91801
## 11th         :  13.9968
## 12th         :  12.62027
## HS-grad      :  11.42384
## Prof-school  :  14.98344
## Assoc-acdm   :  12.1991
## Assoc-voc    :  10.94331
## Some-college :  12.79618
## Bachelors    :  11.42306
## Masters      :  12.14094
## Doctorate    :  14.9196

##
## Formation(s) avec le temps minimum ( 10.94331 ) :

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

On aurait en effet pu s'attendre à ce résultat lorsque l'on met en parallèle les deux graphiques qui nous sont présentés.

e)

Consigne : Observe-t-on un résultat similaire en utilisant l'étendue interquartiles à l'aide de la fonction IQR() ?

```
## Tableau contenant les étendues interquartiles :

##       education hours_per_week
## 1     Preschool          10.00
## 2       1st-4th           5.00
## 3       5th-6th           0.00
## 4       7th-8th           5.00
## 5           9th           3.25
## 6          10th          10.00
## 7          11th          20.00
## 8          12th          15.00
## 9       HS-grad           3.00
## 10  Prof-school          15.00
## 11   Assoc-acdm           5.00
## 12    Assoc-voc           5.00
## 13 Some-college           7.75
## 14    Bachelors          10.00
## 15      Masters          10.00
## 16    Doctorate          15.00

## Minimum des IQR :

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

## [1] 0
```

Ce ne sont pas les mêmes groupes qui ressortent, car le calcul réalisé pour ces deux indicateurs n'est pas le même. L'écart-type résulte de la somme de la différence à la moyenne comme on l'a vu précédemment. Il est

d'ailleurs très exposé aux valeurs atypiques.

Alors que l'étendue interquartile représente la différence entre le premier quartile et le troisième quartile. Cet indicateur est déjà plus robuste que l'écart-type. Le résultat n'est donc pas choquant, car on remarque sur le graphique que les quartiles du groupe 5th-6th sont tous à la même position.

Exercice 5

Dans ce dernier exercice, il n'y a pas non plus d'analyse comme les questions précédentes. Il permet juste de présenter quelques librairies utilisées couramment en R. Elles permettent notamment de réaliser des graphes interactifs.

Conclusion

Cette conclusion sera divisée en deux parties. La première traitera du travail pratique avec un ressenti par rapport à celui-ci, le sentiment par rapport aux objectifs fixés initialement. Dans une seconde partie, on abordera un point de vue un peu plus personnel en faisant un parallèle avec ce que l'on habitude de faire dans notre formation dans le cadre de l'HEIG-VD.

Donc pour cette première partie, je pense que tout c'est bien déroulé dans le cadre de la réalisation du travail pratique. Je n'ai pas vraiment trouvé de difficultés au cours de celui-ci, les documents qui étaient mis à notre disposition étaient largement suffisant pour répondre à toutes mes questions. De plus les sessions que nous avons réalisées en classe permettaient de renforcer ce sentiment en plus de faire de bons parallèles avec le cours. Je pense que les objectifs fixés qui étaient la découverte de R et le renforcement des notions théoriques sont plus qu'atteints.

Pour la partie personnelle, j'ai bien aimé réaliser ce travail pratique pour plusieurs raisons. La première étant que cela m'a rappelé le cours d'ISD qui m'avait beaucoup plus, mais en enlevant les parties dérangeantes. C'est-à-dire la partie où nous étions livrés à nous-mêmes sans aucune notion sur le langage Python. Pour ceux qui n'avaient jamais fait de Python c'était assez difficile de se concentrer sur la partie analyse de l'exercice. Alors que pendant ce TP, nous avions toutes les ressources à disposition pour pouvoir nous concentrer sur le coeur de cet exercice. Le second point que j'ai particulièrement apprécié, c'est que le TP était bien orienté sur ce que nous avions vu en classe et sur ce que nous étions en train de voir lors de sa réalisation. Il servait vraiment de complément au cours pour vérifier que l'on ait bien compris.