IMS Inleveropdracht 3

by Gijs Turkenburg (s3573532) and Sam Noordam (s3680657)

2023-11-08

We want to check whether an increase of the minimum wage has an effect on the number of: a) business closures, b) employees. We therefore look at the dataset FastFood.Rdata of fast-food stores in New Jersey and Pennsylvania. It contains the number of FTE (full-time equivalent) employees at 391 stores in March 1992 and December 1992. If a restaurant reports 0 FTE it means that it is closed. The first column indicates where the store is located (0=Pennsylvania, 1=New Jersey). New Jersey raised the minimum wage on April 1, 1992, from 4.25 to 5.05 per hour. In Pennsylvania the minimum wage remained at 4.25 over that time period. Choose suitable (parametric or non-parametric) statistical models for the data and apply hypothesis tests to answer the initial questions. Do not forget to formulate null- and alternative hypotheses and to justify your test choices.

Introduction

. . .

Organizing the data

The available data is first organised so it can be used to calculate with. The current dataset has a specific order of elements, however when extra data is added this can cause problems. That's why for the case of formality, the data will be separated without assumptions on order or size of the data, so the dataset can be expanded, or datasets from different stores/years can be used.

```r
#organising the data
setwd("/Users/gijs/Desktop/Universiteit Leiden/Programmas/SamCollab/R-collab/Inleveropdracht 3")
load("FastFood.Rdata") #loading the data

pennCounter <- 0 #amount of rows that contains information about Pennsylvania
njCounter <- 0 #amount of rows that contains information about New Jersey
dataFTE <- list(pennBefore = numeric (0), pennAfter = numeric (0), pennDiff = numeric(0),
                njBefore = numeric (0), njAfter = numeric (0), njDiff = numeric(0))
for(i in 1:length(FastFood$NewJersey)){#iterating over the list
  if (FastFood$NewJersey[i] == 0){#information about a Pennsylvania store
    pennCounter = pennCounter + 1
    dataFTE$pennBefore <- c(dataFTE$pennBefore, FastFood$FTEbefore[i])
    dataFTE$pennAfter <- c(dataFTE$pennAfter, FastFood$FTEafter[i])
    dataFTE$pennDiff <- c(dataFTE$pennDiff, FastFood$FTEafter[i] - FastFood$FTEbefore[i])
  }
  else{#information about a New Jersey store
    njCounter = njCounter + 1
    dataFTE$njBefore <- c(dataFTE$njBefore, FastFood$FTEbefore[i])
    dataFTE$njAfter <- c(dataFTE$njAfter, FastFood$FTEafter[i])
    dataFTE$njDiff <- c(dataFTE$njDiff, FastFood$FTEafter[i] - FastFood$FTEbefore[i])
  }
}
print(pennCounter)
```

```
## [1] 76
```

```
print(njCounter)
```

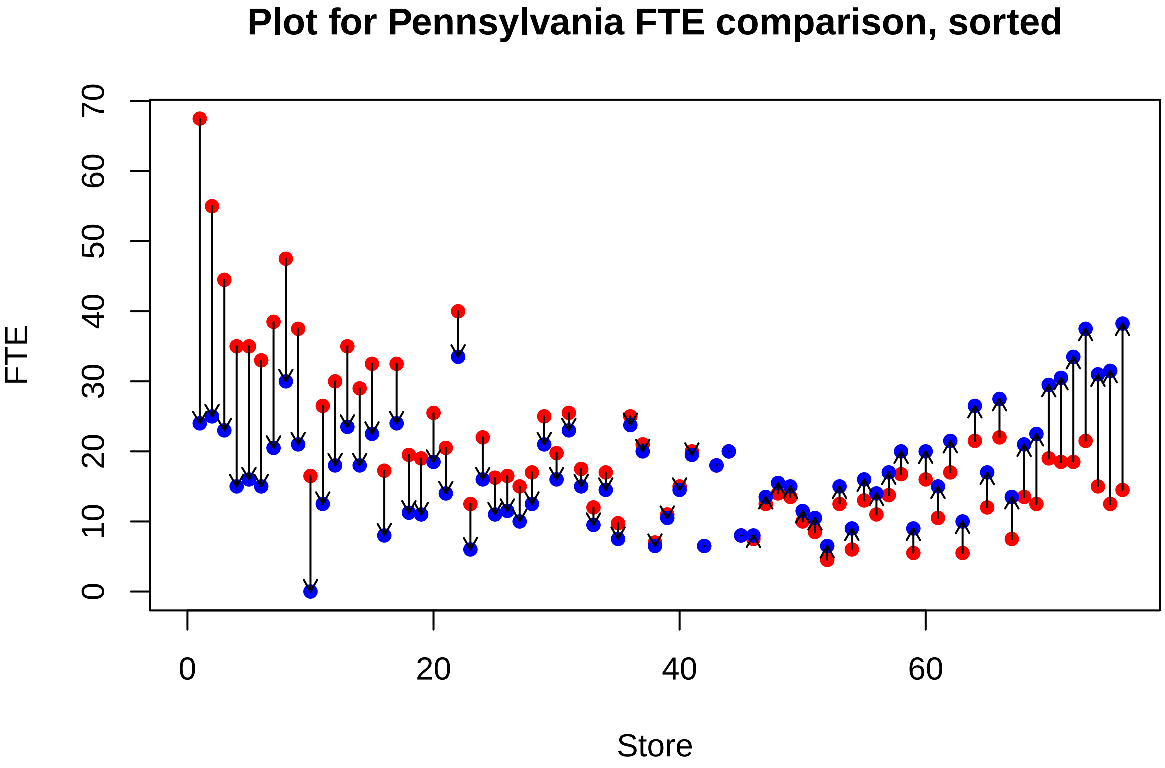
```
## [1] 315
```

```
#an element in our dataset can be accessed by dataFTE$datatype[rownumber]
```

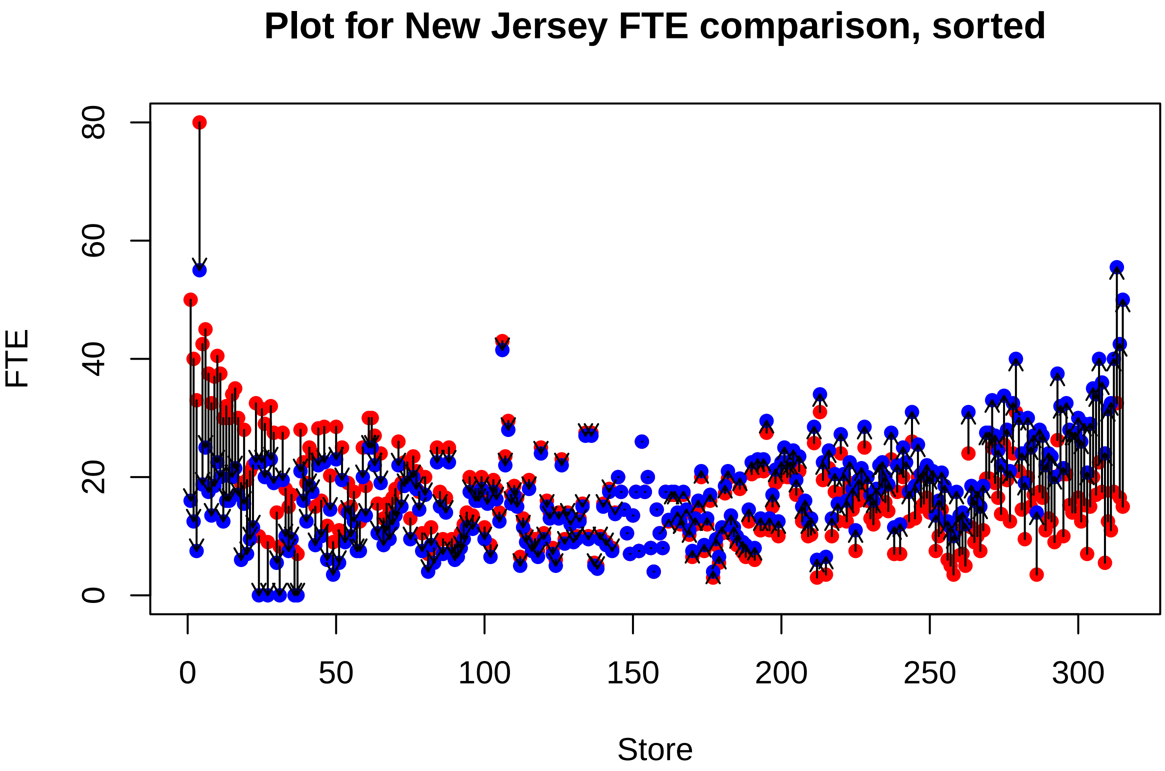
Visualising the data

Before the coice of a paramatric or a non-parametric model can be made, the data is visualised to obtain clues about it's properties. We start with a plot that compares the values of pennBefore and pennAfter.

```
#plot for the Pennsylvania data
max_penn <- max(dataFTE$pennBefore, dataFTE$pennAfter)
sort_index_penn <- order(dataFTE$pennDiff)
pennBeforeSorted <- dataFTE$pennBefore[sort_index_penn]
pennAfterSorted <- dataFTE$pennAfter[sort_index_penn]
pennDiffSorted <- dataFTE$pennDiff[sort_index_penn] #we also sort the differences to say consistent
plot(x = 1,
     type = "n",
     xlim = c(0, pennCounter),
     ylim = c(0, max_penn),
     pch = 16,
     xlab = "Store",
     ylab = "FTE",
     main = "Plot for Pennsylvania FTE comparison, sorted")
points(pennBeforeSorted, pch = 16, lwd = 1, col = "red")
points(pennAfterSorted, pch = 16, lwd = 1, col = "blue")
suppressWarnings({arrows(x0 = (1:pennCounter), y0 = pennBeforeSorted, x1 = (1:pennCounter), y1 = pennAf
```



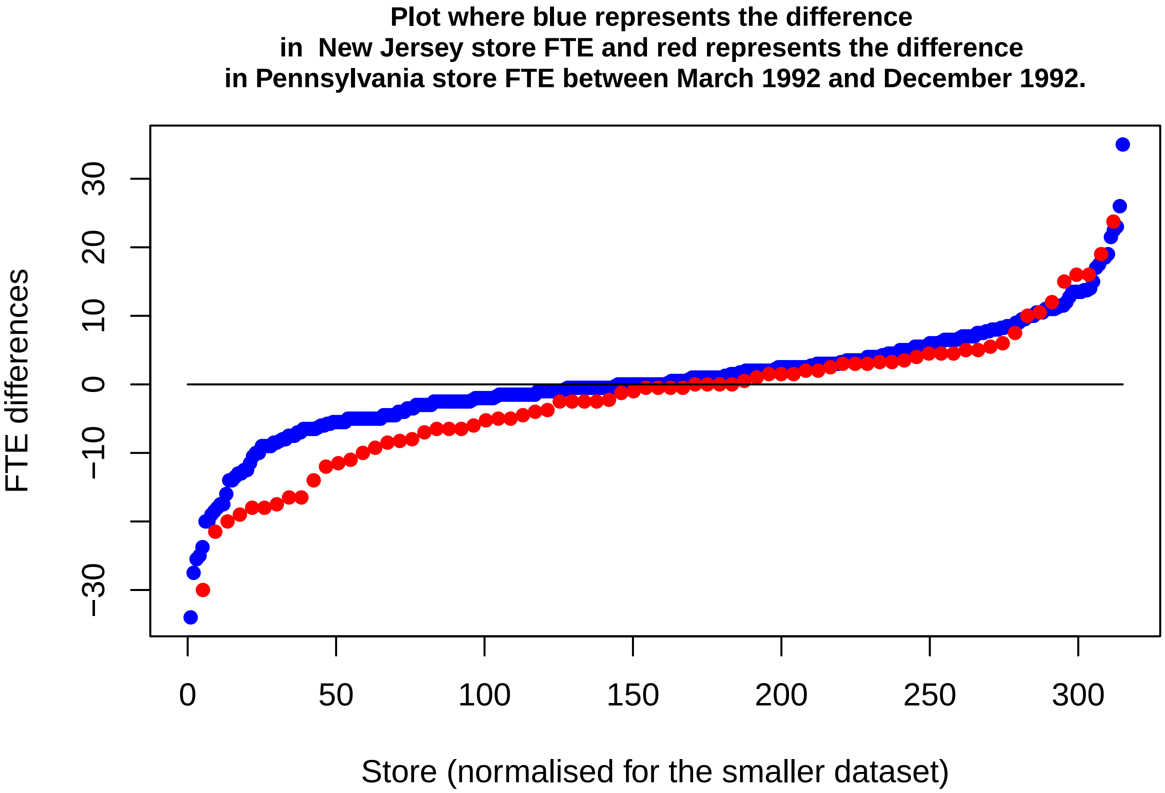
```r
#plot for the New Jersey data
max_nj <- max(dataFTE$njBefore, dataFTE$njAfter)
sort_index_nj <- order(dataFTE$njDiff)
njBeforeSorted <- dataFTE$njBefore[sort_index_nj]
njAfterSorted <- dataFTE$njAfter[sort_index_nj]
njDiffSorted <- dataFTE$njDiff[sort_index_nj] #we also sort the differences to say consistent
plot(x = 1,
     type = "n",
     xlim = c(0, njCounter),
     ylim = c(0, max_nj),
     pch = 16,
     xlab = "Store",
     ylab = "FTE",
     main = "Plot for New Jersey FTE comparison, sorted")
points(njBeforeSorted, pch = 16, lwd = 1, col = "red")
points(njAfterSorted, pch = 16, lwd = 1, col = "blue")
suppressWarnings({arrows(x0 = (1:njCounter), y0 = njBeforeSorted, x1 = (1:njCounter), y1 = njAfterSorte
```



Since no conclusions can be drawn from these points graphs, we zoom in on the pennDiff and njDiff data. Let's plot them in one plot.

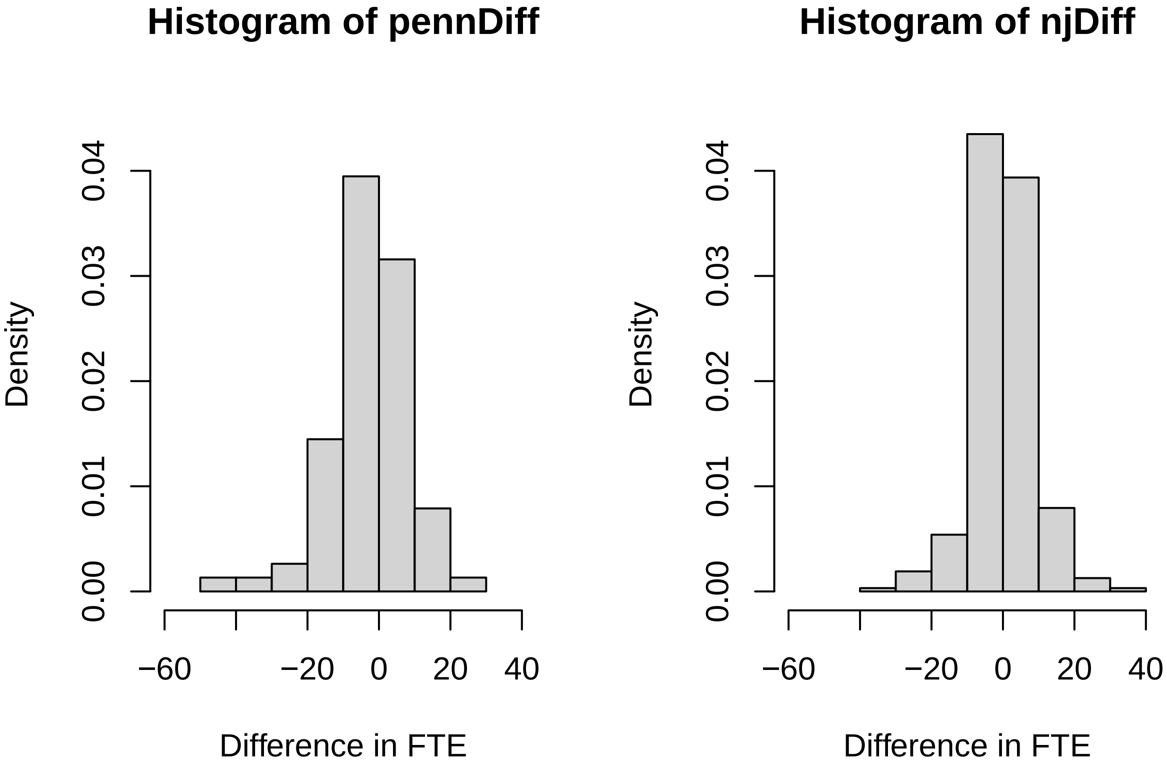
```r
if(njCounter > pennCounter){
  blue_name <- "New Jersey"
  red_name <- "Pennsylvania"
} else{
  blue_name <- "Pennsylvania"
  red_name <- "New Jersey"
}
plot(x = 1,
     type = "n",
     xlim = c(0, njCounter),
     ylim = c(min(njDiffSorted), max(njDiffSorted)),
     pch = 16,
     xlab = "Store (normalised for the smaller dataset)",
     ylab = "FTE differences")
title(
  main = paste("Plot where blue represents the difference \nin ", blue_name , "store FTE and red represe
  cex.main = 0.8  # You can adjust the value as needed
)

if(njCounter > pennCounter){
  points(njDiffSorted, pch = 16, lwd = 1, col = "blue")
  points(seq(from = 1, to = njCounter, by = njCounter/pennCounter), pennDiffSorted, pch = 16, lwd = 2,
} else{
  points(pennDiffSorted, pch = 16, lwd = 1, col = "blue")
  points(seq(from = 1, to = pennCounter, by = pennCounter/njCounter), njDiffSorted, pch = 16, lwd = 2,
}
segments(x0 = 0, y0 = 0, x1 = max(njCounter, pennCounter), y1 = 0)
```



From our data we can clearly see that the red datapoints are (almost) always below the blue datapoints, which will be our motivation to test if raising the minimum wage does have an impact on the amount of FTE, since it looks like not raising the minimum wage (which is represented by the red points) results in a bigger decline of FTE (even when the amount of FTE increased in the year 1992, the red dots show that the amount increased less). Finally, we plot this data in two histograms/

```
par(mfrow=c(1, 2))
pennHist <- hist(dataFTE$pennDiff, freq = FALSE, breaks = 6, main = "Histogram of pennDiff", xlim = c(-
     ,xlab = "Difference in FTE")
njHist <- hist(dataFTE$njDiff, freq = FALSE, breaks = 6, main = "Histogram of njDiff", xlim = c(-60,40)
      ,xlab = "Difference in FTE")
```



Researching the amount of employees: Choosing the test

We need to test on the null hypothesis H_0 that says that there is no difference between the two datasets. First we have to make a decision to either use a parametric test, or use a non-parametric test. Because we have limited datapoints (one of our datasets only has 76 samples) and parametric test rely on convergence which will be inaccurate with our smaller dataset, we choose a non-parametric test. Note that we have a different amount of samples for the New Jersey data and the Pennsylvania data, so we can't use tests that rely on paired samples. When we look at the histograms generated, we can see that they don't really reassemble a normal distribution. Sure, they have a clear mean and some sort of variance, however the tails on both sides differ (in both histograms). Because we aren't really close to a normal distribution (where if we were we could choose a test like the two-sample t-test), we want to choose a test that takes inconsistent tails into account. Also the graphs still have the same shape, this would mean that doing a median-based test would work. This leads us to choose a Wilcoxin rank-sum test. This test is median based, and can be used on distributions that have inconsistent tails, but still the same shape. One last motivation is that, looking back at the point plot, we clearly want to show that the difference between differences almost always results in the blue difference being higher. This can also be linked tot the behavior of the median of both differences.

Researching the amount of employees: The Wilcoxin rank-sum test

We will use the Wilcoxin rank-sum test as a permutation test. Our H_0 (null hypothesis) is that $F_X = F_Y$, where the distribution on $(X_1, ..., X_76)$ (representing the Pennsylvania stores) has a distribution function F_X and $(Y_1, ..., Y_)$ (representing the New Jersey stores) has a distribution function F_X. Our H_1 (the alternative hypothesis) is that $F_X \neq F_Y$. We know that there are

```
choose(pennCounter + njCounter, pennCounter)#we choose pennCounter as k,
```

```
## [1] 2.179679e+82
```

```
                                            #as this is the smallest value
```

, which will be our $\mathcal{X}_1, ..., \mathcal{X}_m$. Then, following the definition of the two-sample permutation test, we denote $\mathcal{Y}_1, ..., \mathcal{Y}_m$ as the respective unselected random variables. We can reject H_0 if $t(\mathcal{X}, \mathcal{Y} < t_{1+\lfloor \alpha m \rfloor})$ or $t(\mathcal{X}, \mathcal{Y}) > t_{m-\lfloor \alpha m \rfloor}$. Because the amount of combinations is too high, we select randomly 10^7 out of all the combinations.

```
tstat <- numeric(10^6)
for (i in 1:10^6) {
    Index <- sample(1 : (pennCounter + njCounter), pennCounter)
    X <- c(dataFTE$pennDiff, dataFTE$njDiff)[Index]
    Y <- c(dataFTE$pennDiff, dataFTE$njDiff)[-Index]
    tstat[i] <- sum(rank(c(X,Y))[1:pennCounter])
  }
tst <- sum(rank(c(dataFTE$pennDiff, dataFTE$njDiff))[1:pennCounter])
sort(tstat)[1+floor(10^6*0.05)]
```

```
## [1] 13443
```

```
sort(tstat)[10^6-floor(10^6*0.05)]
```

```
## [1] 16348.5
```

```
min(mean(tst <= tstat), mean(tst >= tstat))*2
```

```
## [1] 0.032748
```

Researching the amount of employees: Conclusions

Researching the business closures: Choosing the test

Reaserching the business closures: …

Researching the business closures: Conclusions

Final conclusions

Sources

For this assignment the handbook of distributions has been used to determine the family of distributions that was used, and for the density function of the Poisson distribution. The following page has also been used in the decision to choose the Wilcoxin rank-sum test: Wikipedia contributors. (2023, September 7). Mann–Whitney U test. In Wikipedia, The Free Encyclopedia. Retrieved 16:05, November 8, 2023, from https://en.wikipedia.org/w/index.php?title=Mann%E2%80%93Whitney_U_test&oldid=1174323392