IDS - Assignment 1

Bjarke Kingo Iversen - SWD457

February 2018

1 Introduction

For all my code, please see the Jupyter notebook file ***smoking.ipynb*** in the ***src.zip*** folder.

2 Exercise 1: Reading and processing data

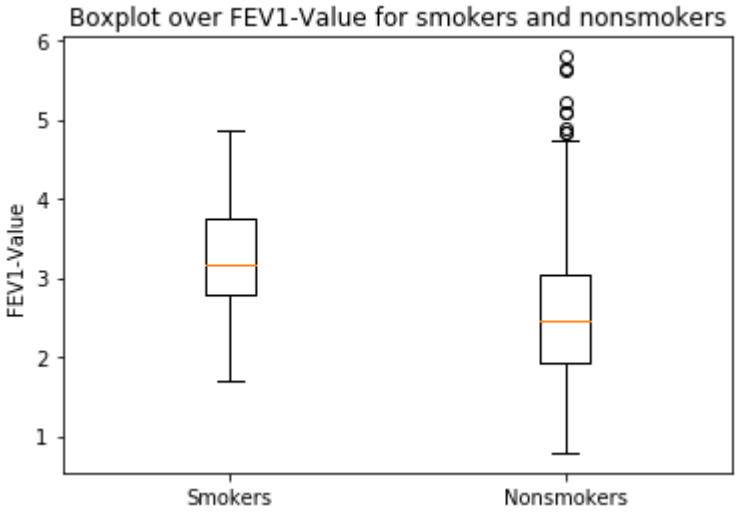
2.1 a

See ***smoking.ipynb*** for implementation.

2.2 b

My computed averages are **(3.2768615384615383, 2.5661426146010187)** for respectively smokers and nonsmokers.
We see that the FEV-1 lung capacity value actually has a higher mean for smokers. Intuitively you may think that smoking decreases the FEV-1 capacity value. However this dataset sugests that this is not the case. This may be because long-term smoking enlarges the lungs when inhaling huge amounts of smoke. However this does not mean that smokers have better working lungs then nonsmokers.

3 Exercise 2: Boxplots



By this boxplot we confirms what we found in exercise 1, that we average FEV-1 value is higher for smokers, for this particular dataset. We also see that non-smokers have more outliers, and a lot more variance in the FEV-1 value.

4 Exercise 3: Hypothesis testing

4.1 a

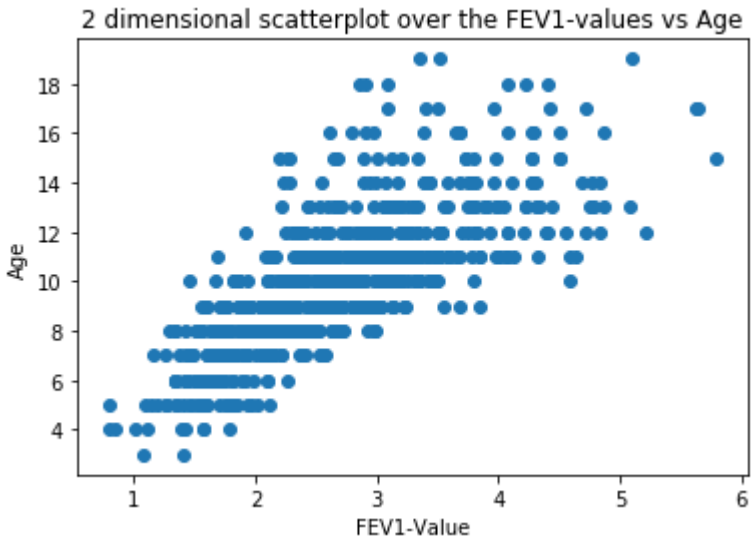
See ***smoking.ipynb*** for implementation.

4.2 b

- Let X and Y be the random variable describing the FEV1-value og smokers and nonsmokers.
- Now, my Null-hypothesis is that mean(X) = mean(Y).
- That is, there is no significant difference between the FEV1-value of smokers vs. nonsmokers.
- Standard deviation of smokers: 0.00852039899499, nonsmokers: 0.00122607587324.
- Value of t-statistic: 7.199031861.
- Number of freedom degrees: 83.0.
- P value: 2.49456448153e-10.
- I reject my Null-hypothesis.

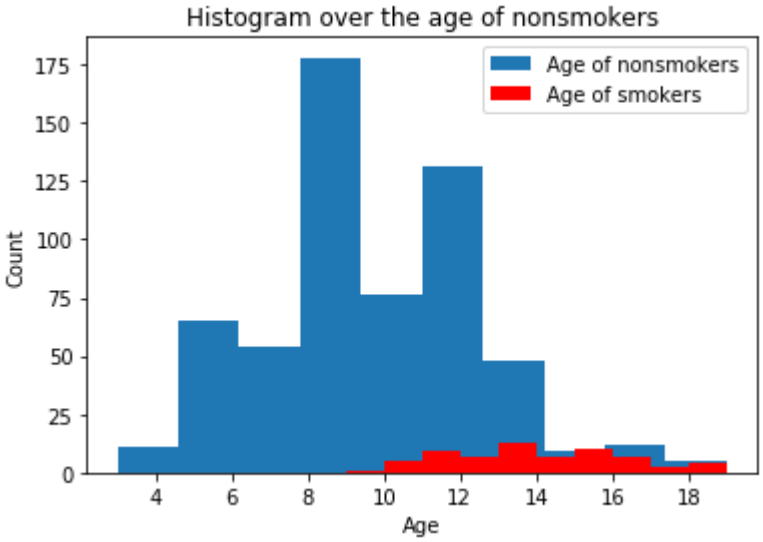
Print-snippet run from my implementation. We reject the Null-hypothesis since P were way smaller than $\alpha = 0.05$. This also means that the observed difference between the mean-values cannot be attributed to chance alone.

5 Exercise 4: Correlation



By this scatterplot we see that the older a person gets, the more the FEV-1 value can differ from person to person. Also it seems that the FEV-1 value is much likely to increase with age.
The correlation is computed to be **0.75645898998959993**.

6 Exercise 5: Histograms



This histogram has a much relevant observation. We observed earlier that smokers seemed to have, on avereage, higher FEV-1 value than those of nonsmokers. Also we discovered that the average FEV-1 value seemed to increase with age. Now we see that this dataset is over-sampled on nonsmokers and early to middle aged persons. However we see that its primarily the older people that smoke, and thus we have a huge bias. In order to fix this bias, we would have to have to consider a more equal distribution of smokers and nonsmokers in correlation with their age. Then we'd properly be able to determine whether the FEV-1 value really is higher for smokers.