Assignment 1 - Introduction to Data Science

Ninell Oldenburg

February 20, 2022

1 Exercise 1

```
average lung function for smokers:  3.2768615384615387
average lung function for non smokers:  2.5661426146010164
```

In the calculation of the mean lung capacity (FEV1 score) for the two groups, smokers and non-smokers, we can see a higher FEV1 score for smokers that implies a larger lung volumne for this group. These results are definitely surprising because it is expected that smokers have a smaller lung volume.

2 Exercise 2

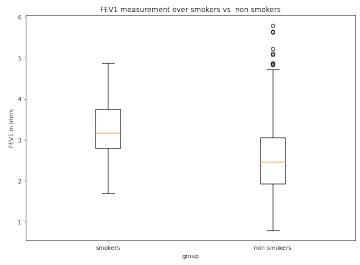

Figure 1: Boxplot of the FEV1 in the two groups

As expected form Exercise 1, the box plot supports the findings of smokers having a higher lung volume in average. The variance, however, seems to be higher for non smokers, as well as the number of outliers, which implies that the samples are far more spread out over the group of non-smokers.

3 Exercise 3

```
t:  7.149608129503807,
df:  83,
p-val:  3.11735748326214e-10,
accept?:  False
```

Performing the Welsh T-test for two a two-sample hypothesis, we're coming to the conclusion that we can reject the null hypothesis, which is: the smokers and non-smokers do not significantly differ from each other in their lung volume capacity. I think I don't find this extremely surprising as the two means are quite a bit different, also taking the different variance into account. However, I find the degree to which they differ, with a unexpectedly low p-value from ca. 3.11e-10, very interesting.

4 Exercise 4 - Confounders

```
Pearson Correlation coefficient:  0.75645899
Spearman Correlation coefficient:  0.53284167
```

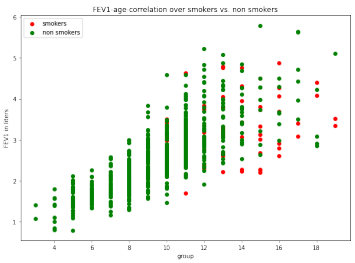

Figure 2: 2D scatter plot of age versus FEV1

On the scatter plot we can see a significant difference between the distribution of the two groups. First of all, the non-smokers seem to have an increased lung volume with growing age up until the age group of 13. Here after the trend is not so clear anymore, also because of there being less records which leads to a more spread out distribution. Same goes for the smokers, even though the record points seem to be spread out from for every age group, which leads to the conclusion that there are in general less data points in this group. Furthermore, in contrast to the non-smokers, there is not such a clear trend of growing lung volume with growing age observable for the smokers. Lastly, we can see that there appear to be no smokers until an age of 10, whereas in the older ages from 14 on, seem to be less non-smokers.

For the confounders I chose to look at the Spearman Correlation coefficient because the scatter plot is visually indicating that we could have a monotonic relation in our data, rather than a pure linear one. A value of ca. 0.53 implies, that this positive relationship is indeed observable. Comparing it to Pearson Correlation coefficient of ca. 0.75 indicates, that the latter one is overgeneralizing over our data.

5 Exercise 5

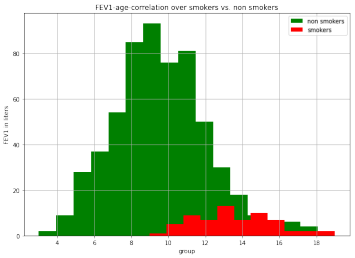

Figure 3: Histogram over the age of subjects in each of the two groups

Overall, three things become even more clear than in the previous analysis.

1. The number of non smokers exceeds the number of smokers drastically.
2. The distribution over the age groups is very different from smokers to non smoker is two ways:
   - The plot supports the findings from the earlier exercise that the mean age is higher in the group of smokers. That is, the red plotted bars for the smokers are more towards the right side of the plot
   - The distribution of the both groups over the parameter of age is different: where there are rather high spikes for the ages 9, 8 und 11 in the non smokers plot, we see a more flat curve throughout the smokers age groups. However, this could also be due to the general less datapoints in that group with some classes (age 9) only having 1 record.
3. Interesting side note: even though we have way more examples of non smokers (589 vs. 65), there still seem to be more smokers at the ages of 15 and 16 than there are non-smokers.