Assignment 1 - Introduction to Data Science

Ninell Oldenburg

February 10, 2022

1 Exercise 1

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

These results, that the average lung volume for smokers is higher than for non smokers, are definitely surprising because it is expected/hypothetically known that smokers have a smaller lung volume in litres.

2 Exercise 2

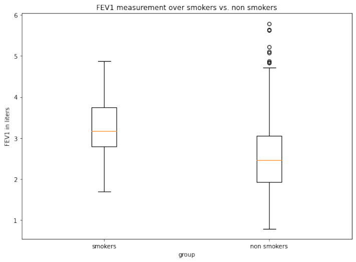

Figure 1: Boxplot of the FEV1 in the two groups

As expected form the average value in Exercise 1, the box plot supports the findings of smokers having a higher lung volume. The variance, however, seems to be higher for non smokers, as well as the number of outliers.

3 Exercise 3

```
t:  6.473411327744889,
df:  652,
p-val:  0.05,
accept?:  False
```

From the results in exercise 2 can be derived that the mean is significantly different throughout the two groups, so rejecting the null hypothesis here and concluding that the groups differ doesn't come surprisingly. However, I find the degree to which they differ, with a t-value from ca. 6.47, unexpectedly high for such a relatively small difference in the means.

4 Exercise 4

We can see two curves that rise on a linear, however, step-wise scale. Furthermore, we can see that the red curve, the smokers curve, lies below the green, non smokers curve, which implies that the FEV1

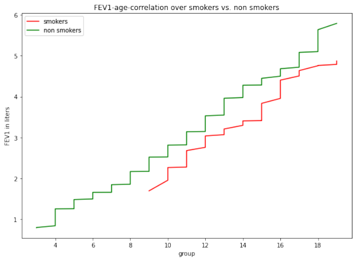

Figure 2: 2D plot of age versus FEV1

score for smokers is lower than for non smokers. This is in such way interesting, as we have shown in the exercises above that the mean FEV1 score for smokers is significantly higher than for non smokers and shows how much influence the distribution of the values has to the overall outcome.

5 Exercise 5

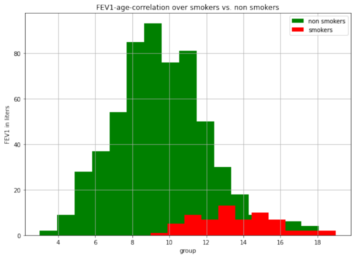

Figure 3: Histogram over the age of subjects in each of the two groups

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.