Exercise 7

Riccardo De Bin and Vindi Jurinovic

Exercise 1:

1. Categorize the variable `bmi` into an underweight ($BMI < 18.5$), normal weight ($18.5 \leq BMI < 25$), overweight ($25 \leq BMI < 30$) and obese ($BMI \geq 30$) group. Turn the variable into a factor. What is the proportion of overweight or obese people according to the categorized BMI? What is the proportion of people ever diagnosed with being overweight (variable `ovrwght_ever`)? How many overweight people were actually ever diagnosed with being overweight?

2. Is there a difference in diabetes prevalence between obese people diagnosed with overweight and those who were never diagnosed? What about self-rated health? How do you explain the results?

3. With a function `sample`, you can create a random subsample of a data set. For example, the command `sample(1:1000, 500)` creates a subsample of size 500 from a vector containing numbers between 1 and 1000. Create a subsample of size 500 using your data set. In this subsample, test the relationship between heart diseases and gender. Repeat the analysis 10 times (with 10 different subsamples) and note the results. Compare them with the result for the whole data set. How do you explain the differences?

4. We have seen in a previous lecture that in our data set, smoking seems to be negatively associated with cancer: cancer prevalence in non-smokers is higher than in smokers. Further analysis revealed age as a possible confounder: older people get cancer more often, and smokers tend to be younger. When we adjust for age, we can see that young smokers actually do tend to develop more cancers than young non-smokers.
   For the following exercise, create a binary age (younger: age $\leq$ 50, and older: age>50) and a binary smoking variable.

   You are working for a tobacco company and want to show that smoking is good for you regardless of your age. Your employer provides you with our data set and asks you to produce a significant association between smoking and cancer prevalence in every age group (of course, a desirable association). Your task is to produce a subsample where in the end, smoking turns out to be protective against cancer, even when you adjust for age. Can you satisfy your employer and create such data set?

   Having in mind this last exercise, do you trust every result with a significant p-value?