Exercise 4 (NHANES)

Today we will start to analyze the NHANES data from your R Data Project. Choose one subsample to work with. There are 5 data sets. You may also want to look at the **documentation** of the data.

Task 1 (General)

Generate the results as a Markdown file

1. Load the data in R.
2. What is the dimension of the data set? How many rows (samples), and how many columns (variables) does the data set contain? What are the variable names of the data set?
3. All the variables in the data set are either of a class integer, numeric or boolean (i.e., logical). However, some of the variables should be factors rather than numerical variables. Change the class of these variables to factor.
4. How many women and how many men are there in your data set?
5. What is the mean BMI in the overall population? What is the mean BMI for men and women?
6. Who has an higher mercury level in blood: men or women? People with chronic bronchitis or people without it? ‘Hispanic’, ‘White’, ‘Black’ or ‘Other/Mixed’ people?
7. Use the function summary to get the summarized information on all the variables in the data set.

Task 2 (Visualization)

Generate the results as a Markdown file

1. Plot the variable `rr_sys` as a function of `bmi`.
2. Now we want to plot the variable `rr_sys` against `diab_lft`. Which plot should we use here?
3. Plot the BMI against `educ` and give a short interpretation.
4. Plot the histogram of the high-density lipoprotein (HDL) cholesterol levels. How does the distribution of HDL look like? Can you convert the variable HDL so that its distribution looks more normal? Create such a variable and add it to your data set.

Task 3 (Confidence intervals)

1. Is the variable height normally distributed in males and females? Check this by analyzing graphically the two sub-samples. Use qqplots: `qqnorm()` and `qqline()`.
2. Suppose that the value of the variance for the male height is known and equal to 64. Construct a 95% confidence interval for the mean.
3. Repeat the procedure for the female population, with unknown variance and using a confidence level of 90%.

Task 4 (Tests)

1. Is the variance the same in both populations? Perform an appropriate test. What is the null hypothesis?
2. Which test can we use, if, instead, we want to check if males are on average taller than females? Set an adequate alternative hypothesis.
3. Analyze the confidence interval obtained in the previous point. Why doesn't it have an upper bound?
4. Now look at the distribution of the variable weight: can we graphically state its normality? Perform a transformation in order to recover it.
5. Test if the mean of the variable `weight` is 80 kg, testing $H_0 : log(weight) = log(80)$ versus $H_1 : log(weight) \neq log(80)$. Can you state an interval estimate with a level of 0.99?

6. Provide a punctual and an interval estimate for the prevalence of heart diseases and lung pathology.

7. Test if the prevalence is statistically different between men and women.

8. Turn the variable smokstat into a binary variable (put the non-smokers and the people who almost never smoked into one group).

9. Do you expect smokers to have a higher cancer prevalence than non-smokers? Test if there is a significant difference in cancer prevalence between smokers and non-smokers (using the variable from before). If so, which group has a higher prevalence? Did you get the result you expected?

10. Do the same test in the subgroup of people who are between 20 and 49 years old. What do you see now? How can you explain the results?

Task 5 (Discrete and non-parametric tests)

1. Use a chi-square test in order to test whether the presence of chronic bronchitis and the current smoking status are independent.

2. Use a Fisher test to verify the independence between sex and the presence of any kind of liver disease.

3. Perform a sign test both on hdl and on log-hdl to test the hypothesis that the median of the cholesterol level is 1.30. Is the median significantly different from 1.30? Do you obtain the same results using hdl and logHdl?

4. Use a Mann-Whitney test to test the null hypothesis $H_0 : maleweight = femaleweight$.

5. It has been shown that there is a 'social gradient' in health such that the richer you are, the more likely you are to have better health. Plot general self-rated health against relative income so that you can get an impression whether this is confirmed by our data. Which kind of plot is reasonable? Consider using a mosaic plot. E.g. function `mosaicplot()`.

6. Test the relation for statistical significance using an appropriate test.

7. Categorize the variable `bmi` into an underweight (BMI<18.5), normal weight (18.5 BMI<25), overweight (25 BMI<30) and obese (BMI 30) group. Turn the variable into a factor. You may use the function `cut()`.

8. 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?

9. 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?