Data Analysis in Astronomy and Physics (SoSe22)

Instructor: PD Dr. Markus Röllig
TA: Dr. Christof Buchbender
TA: Craig Yanitski

Exercise Set 1

Due: **10:00 11 April 2022**
Discussion: **13:00 22 August 2022**
Online submission at via ILIAS in the directory Exercises / Übungen -> Submission of Exercises / Rückgabe des Übungsblätter

1. Sampling [100 points]

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage. The BRFSS Web site (http://www.cdc.gov/brfss) contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

Download the ASCII data file from the course web site (Name: `cdbrfss1999.zip`)

The file `Codebook99.rtf` contains a detailed explanation of the individual data entries. Use the Cookbook to understand the data columns that you are working on in this problem.

Large data

The dataset contains 159989 entries.
a. Take a sample of 30000 from this dataset and export it to an ASCII file. Make sure that your method allows to draw more than one sample from the population. [**5 points**]
b. Discuss your method to do so. Is your sampling a "good sample" in the sense that it is representative for the larger "population"? [**10 points**]

Large columns

Each case in the dataset can have up to 241 variables. Each one of these variables corresponds to a question that was asked in the survey. For example, for **genhlth**, respondents were asked to evaluate their general health, responding either excellent, very good, good, fair or poor. The **exerany** variable indicates whether the respondent exercised in the past month (1) or did not (0). Likewise, **hlthplan** indicates whether the respondent had some form of health coverage (1) or did not (0). The **smoke100** variable indicates whether the respondent had smoked at least 100 cigarettes in her lifetime. The other variables record the respondent's height in inches (**hti**) and feet (**htf**), **weight** in pounds as well as their desired weight, **wtdesire**, **age** in years, and **sex**.

a. Locate the columns corresponding to the variables **genhlth**, **exerany**, **htf**, **hti**, **smoke100**, **weight**, **wtdesire**, **age**, and **sex**.

b. Reduce your sample to include only these variables and export it to an ASCII file. **[5 points]**

c. How many cases and how many variables are there in your sample? **[5 points]**

- (a) 9 cases; 30,000 variables
- (b) 8 cases; 30,000 variables
- (c) 30,000 cases; 9 variables
- (d) 159,989 cases; 10 variables

d. What type of variable is **genhlth**? **[5 points]**

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical (not ordinal)
- (d) categorical, ordinal

e. What type of variable is **weight**? **[5 points]**

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical (not ordinal)
- (d) categorical, ordinal

f. What type of variable is **smoke100**? **[5 points]**

- (a) numerical, continuous
- (b) numerical, discrete
- (c) categorical (not ordinal)
- (d) categorical, ordinal

One Bar chart

Take all **genhlth** entries from your sample and draw a bar chart to visualize how the cases are distributed across the possible categories. [**10 points**]

Two Bar Charts

Combine the **smoke100** with the **genhlth** entries from your sample and draw *two* bar charts, one showing the health of the smokers and a second one showing the health of the non-smokers. [**20 points**]

BMI

Next let's consider a new variable **bmi** that doesn't show up directly in this data set: Body Mass Index (BMI). [**30 points**]
BMI is a weight to height ratio and can be calculated as.

$$\text{BMI} = \frac{\text{weight(lb)}}{\text{height(in)}^2} * 703$$

703 is the approximate conversion factor to change units to metric (meters and kilograms) from imperial (inches and pounds). Compute the bmi for each case in your sample and add it to the sample (e.g. as additional column). Visualize the distribution of the BMI in your sample.
ATTENTION: Remember, that the height was given in feet and inches separately. make sure to compute the total height in inches!