Data Analysis in Physics and Astronomy

Exercise 1

Author:
Tom Carron

April 10, 2022

1 Sampling

- (a)The data set was reduced from 159989 entries to a simple random sample of 30000 entries. Each time the script runs, a new simple random sample is produced. This is achieved by randomly selecting 159989-30000 rows and excluding them from a copy of the dataset. This simple random sampling is a "good sample" and represents the larger population well as both the population size and sample size are large.
- (b) The sample was then further reduced by selecting only 9 variables of interest; "GENHLTH", "EXERANY", "HTF", "HTI", "SMOKE100", "WEIGHT", "WTDESIRE", "AGE" and "SEX".
- (c) The remaining data set has 30,000 cases with 9 variables.
- (d)The variable **genhlth** is categorical, ordinal.
- (e) The variable **weight** is numerical, continuous.
- (f) The variable **smoke100** is categorical.

2 One bar chart

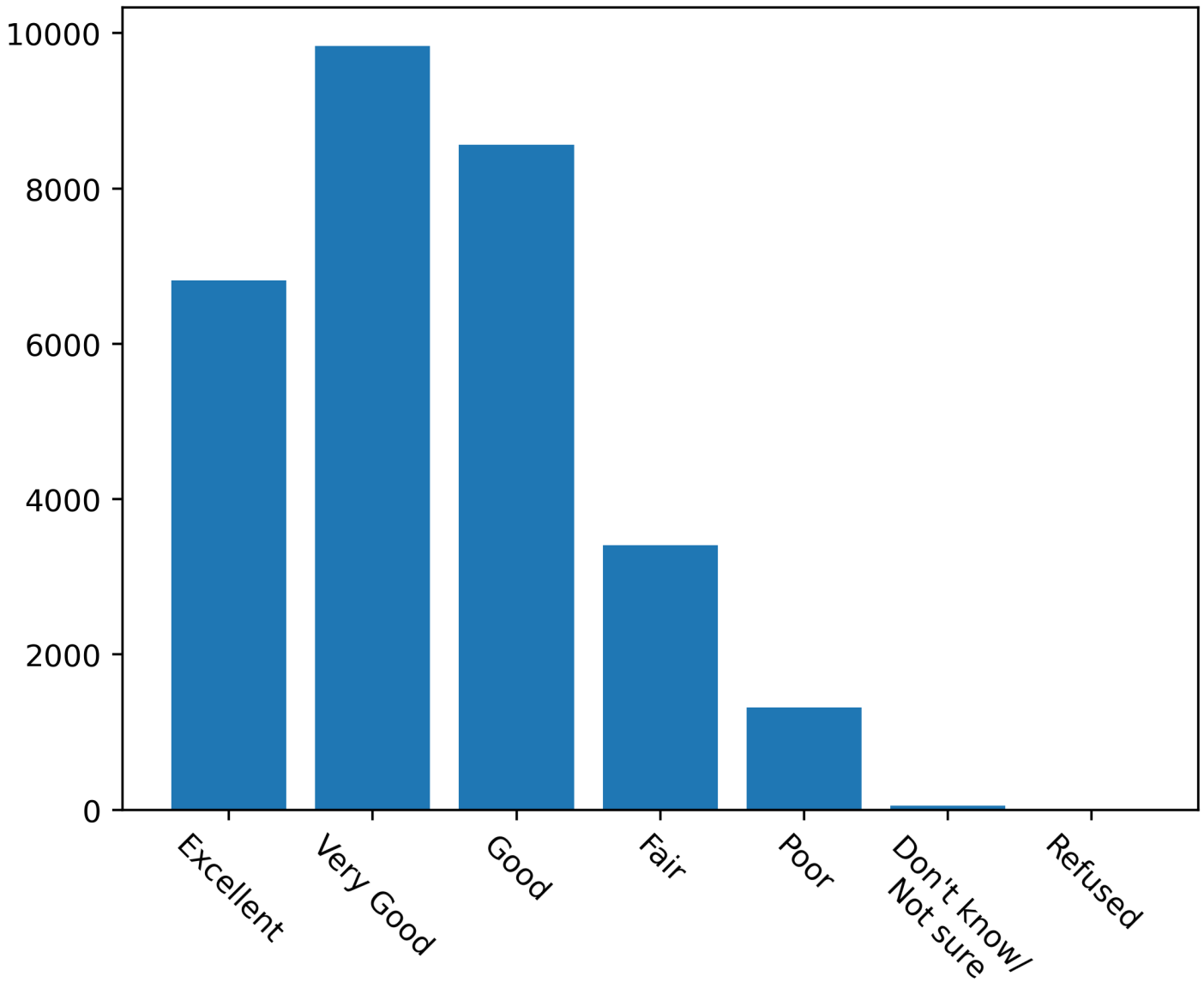

Figure 1: Bar chart visualizing the general health of the population sample.

3 Two Bar Charts

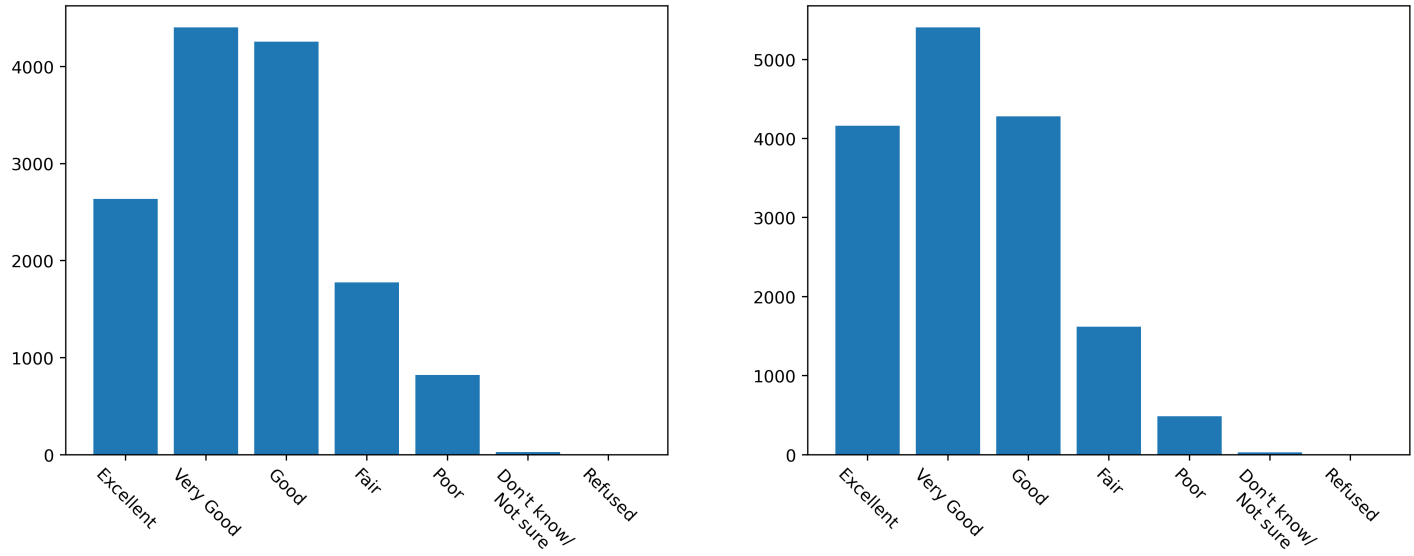

Figure 2: *Left:* Bar chart visualizing the general health of the portion of the sampled population who had smoked at least 100 cigarettes in their lives. *Right:* Bar chart visualizing the general health of the portion of the population that had not smoked at least 100 cigarettes in their lives.

4 BMI

Here we consider the Body Mass Index (BMI), a variable which does not show up directly in our data. The BMI is calculated using the **hti**, **htf** and **weight** variables using equation 1, where the factor of 703 is the approximate conversion to change from imperial to metric units.

$$\text{BMI} = \frac{703 * \text{weight(lb)}}{\text{height(in)}^2} \tag{1}$$

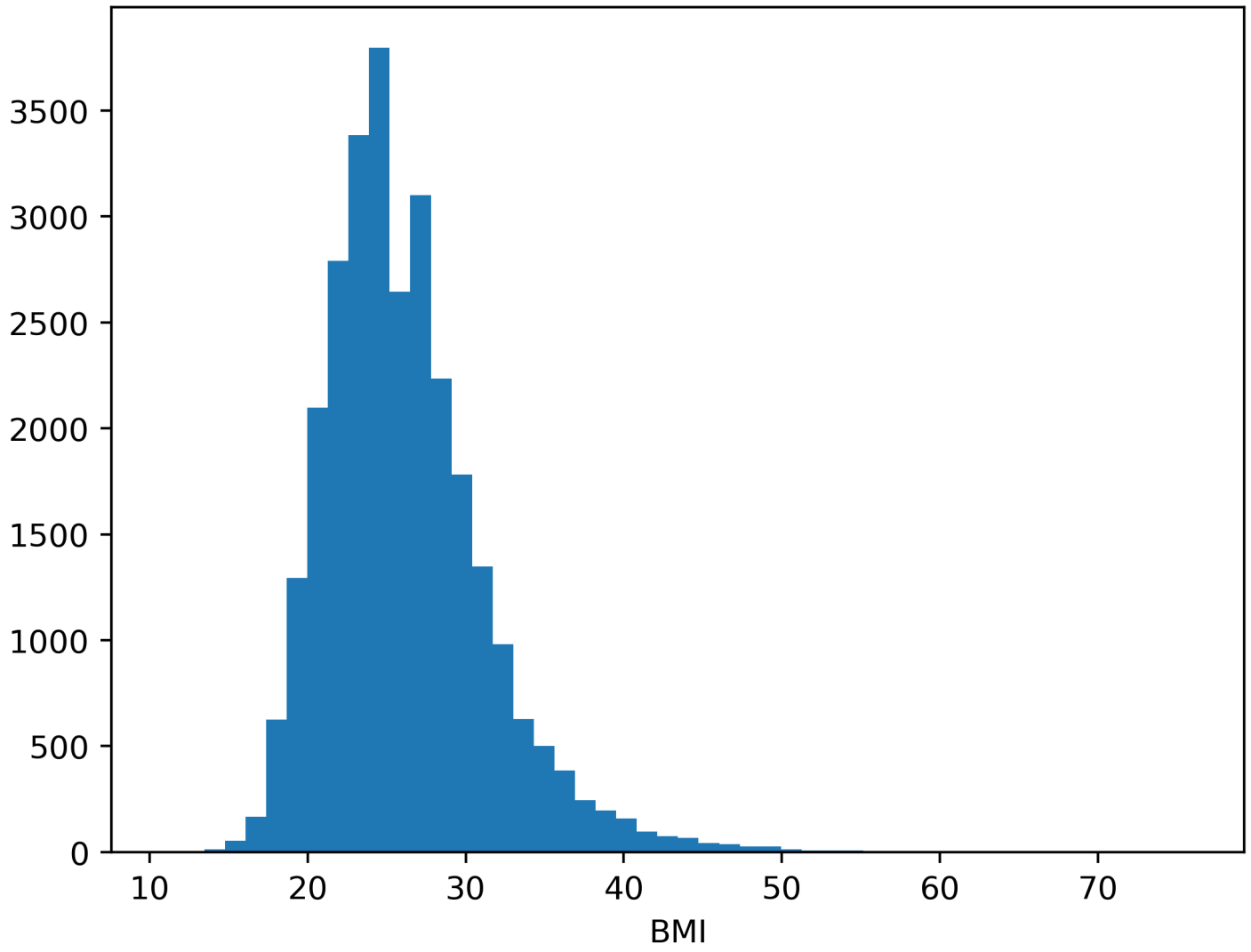

Figure 3: Visualization of the distribution of BMI in the population sample. 50 equal sized bins were used for the histogram.