Introduction to Statistical Ideas and Methods

Summarizing Data: One Variable

Visualizing the Skeleton Data

The Skeleton Data were collected in order to assess the accuracy of methods used by anthropologists to estimate age at death. The Visualizing Skeleton Data Shiny app lets us quickly inspect the data by producing a number of helpful visualizations and summary statistics. In this exercise, you will use this app to apply the what you learned in the module *Summarizing Data: One Variable* to carry out the first stages of an exploratory analysis of these data.

1. To familiarize yourself with the how the app works, make a histogram of the age variable by selecting “histogram” under Plot Type and “Age” under Variable of Interest. Note that you are given summary statistics in addition to the plot. What does selecting a Subgroup option do?

2. Change the plot type to “pie chart”. What are the possible variables of interest that can now be plotted? Are they the same as for the “box plot” plot type? Why or why not?

3. Real world data often come with mistakes or mislabellings. The skeleton dataset includes BMI as both a quantitative variable and as a categorical variable. Try several different ways of plotting the quantitative BMIquant separated by subgroup BMIcat. Do you notice anything unusual?

4. If the data were overwhelmingly from males or overwhelmingly from obese people we might miss seeing effects due to BMI or age. This is because if one group is strongly overrepresented then there are limited data that can be used to differentiate an effect specific to that group. Try producing pie charts and bar charts for the categorical variables. Does it seem like any group is strongly overrepresented or underrepresented?

5. Pie charts are popular in some groups but not amongst statisticians. This question might help you understand why statisticians often avoid using them. Produce pie charts of Sex and BMIcat and use the pie charts to estimate the relative frequencies of each category visually. The true relative frequencies are shown in the Data Summary under the chart. Were your estimates accurate?

6. Observational data with many potential contributing factors can sometimes lead us to confuse the source of observed effects. For example, if almost all of the obese people in the dataset were very old then we might not be able to tell whether inaccurate age estimates are due to age or due to obesity (that is, we might attribute age affects to obesity and vice versa). Use the plots to explore whether this particular phenomenon for age and obesity is present in our dataset. Are there comparisons you would like to be able to make that are not possible with the plotting available in the app?

7. What is the average BMI for females in this data set? What is the average BMI for males?

8. The dataset includes age estimates from two different estimation methods: the method of Di Gangi et al. (DGestimate) and the Suchey-Brooks method (SBestimate). The errors in age estimation (True Age − Estimated Age) are given in the data as the variables DGerror and SBerror. For each of the age estimation methods, decide whether the method tends to overestimate or underestimate age at death. Which plots are most useful for assessing this? If you were an anthropologist, how might you make use of this information?

9. For each of the age estimation methods (Di Gangi and Suchey-Brooks), does it seem that body mass index or sex substantially impacts the size of the error in age estimation? Does BMI or sex substantially impact the variability of the error in age estimation? What kinds of plots are useful for assessing this?

10. Anthropologists would like some guidance about which of these age estimation methods they should use. What additional exploratory analysis (or analyses) might be useful to carry out, beyond what is available in this app, before making a preliminary recommendation? Are there any additional visualizations or plots that you think should be produced?