Introduction to Data

Chapter 1, Lab 1

OpenIntro Biostatistics

Topics

- Dataset manipulation in R
- Numerical summaries: mean, SD, median, IQR
- Graphical summaries: boxplots, histograms, scatterplots

The first two sections of this lab introduce basic tools for working with data matrices, as well as the commands for producing numerical and graphical summaries. The last section focuses on data interpretation and reinforces the statistical concepts presented in the text. The material in this lab corresponds to Sections 1.1 - 1.2 and 1.4 - 1.6 of *OpenIntro Biostatistics*.

Section 1: BRFSS.

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. 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, weekly exercise, possible tobacco use, and healthcare coverage.

1. Use the following command to download the dataset `cdc` from a URL. This dataset is a sample of 20,000 people from the survey conducted in 2000, and contains responses from a subset of the questions asked on the survey.

   ```
   source("http://www.openintro.org/stat/data/cdc.R")
   ```

2. Take a look at the Environment tab, where `cdc` should now be visible. Click the blue button next to the dataset name to view a summary of the 9 variables contained in the data matrix. To view the dataset itself, click on the name of the dataset; alternatively, enter the command

   ```
   View(cdc)
   ```

   Each row of the data matrix represents a case and each column represents a variable. Each variable corresponds to a question that was asked in the survey. For `genhlth`, respondents were asked to evaluate their general health as either "excellent", "very good", "good", "fair", or "poor". The variables `exerany`, `hlthplan`, and `smoke100` are binary variables, with responses recorded as either 0 for "no" and 1 for "yes": whether the respondent exercised in the past month, has health coverage, or has smoked at least 100 cigarettes in their lifetime. The other variables record the respondents' height in inches, weight in pounds, their desired weight (`wtdesire`), age in years, and gender.

3. The $ operator in R is used to access variables within a dataset; for example, `cdc$height` tells R to look in the `cdc` dataframe for the `weight` variable. Make a scatterplot of `height` and `weight` using the `plot( )` command:

```
plot(cdc$weight ~ cdc$height)
```

Do `height` and `weight` appear to be associated?

4. The conversion from inches to meters is 1 in = .0254 m. Create a new variable `height.m` that records height in meters. Similarly, the conversion from pounds to kilograms is 1 lb = .454 kg. Create a new variable `weight.kg` that records weight in kilograms.

5. BMI is calculated as weight in kilograms divided by height squared. Create a new variable `bmi` and make a scatterplot of `height` and `BMI`. Do `height` and `BMI` seem to be associated?

   A BMI of 30 or above is considered overweight. Why might health agencies choose to use BMI as a measure of obesity, rather than weight?

6. Row-and-column notation in combination with square brackets can be used to access a subset of the data. For example, to access the sixth variable (`weight`) of the 567th respondent, use the command:

   ```
   cdc[567, 6]
   ```

   To see the weight for the first ten respondents, use:

   ```
   cdc[1:10, 6]
   ```

   If the column number is omitted, then all the columns will be returned for rows 1 through 10:

   ```
   cdc[1:10, ]
   ```

   Likewise, omit the range for the rows to access all observations for column 6. The following will return the weight for all 20,000 respondents:

   ```
   cdc[ ,6]
   ```

7. Use bracket notation to make a scatterplot of `height` and `weight` for the first 100 respondents. There are multiple ways to do this—find one that works!

Section 2: Gene Transcript Lengths.

Before genes can be translated into proteins, DNA must first be transcribed into RNA. The dataset `coding.mrna` contains the length of known protein-coding transcripts (measured in base pairs). Load the dataset from the `oibiostat` package.

1. How many transcripts are represented in this dataset? Use the `nrow()` command to return the number of rows in the dataset.

```
nrow(coding.mrna)
```

2. Calculate the 5-number summary for the transcript lengths using the `summary()` command. What striking feature do you notice in the summary?

3. Draw a histogram and a boxplot of the distribution of transcript lengths. When you see them, you will notice that the plots are not particularly informative. Explain why you think that is the case.

4. For a data item `x`, the notation `x < a` is used to reference the subset of values of `x` that are less than the value `a`. Pick a reasonable length `a` and use the `subset()` command to create a trimmed version of `coding.mrna` called `lengths.subset` that only contains data for transcripts with length less than `a`. This is one simple strategy for making the structure of the data easier to view in the plot.

```
lengths.subset = subset(coding.mrna, coding.mrna$transcript_length < a)
```

   With the trimmed data, draw a histogram and boxplot, and calculate summary statistics. Now describe the shape of the data. Explain your choice of where to trim the data.

5. Use R to find out how many transcripts you have trimmed from the dataset. Hint: this might involve notation used in Questions 1 and 4.

6. One way of manipulating a large dataset is to take a random sample and construct numerical and graphical summaries of the sample. Use the following code to construct a random sample that consists of 10% of the original number of transcripts; the sampling is done without replacement, such that a single transcript cannot be chosen more than once.

   Using the `set.seed()` function allows for pseudo-random sampling; that is, a random sample that is reproducible. Replace `xxxx` in the function with four numbers of your choice, then run the code to create `transcript.sample`, a vector of transcript lengths.

```
set.seed(xxxx)
sample.size = 0.1 * nrow(coding.mrna)
transcript.sample = sample(coding.mrna$transcript_length, size = sample.size,
                           replace = FALSE)
```

   Now with `transcript.sample`, calculate the number of transcripts in the dataset, the five-number summary, and draw a histogram and boxplot. Does the sample data more closely resemble the complete version of the data or the trimmed version from Question 4?

7. Make side-by-side boxplots of transcript lengths by chromosome. Use the command:

```
boxplot(coding.mrna$transcript_length ~ coding.mrna$chromosome_name)
```

Select "Show in New Window" above the plot and expand the window to be able to see all the chromosome numbers displayed. Where are the longest transcripts located?

8. Subset `coding.mrna` to only include values from chromosome 2. Repeat for the Y chromosome. Hint: the notation is similar to that used in Question 4.

   Use `nrow()` to compare the number of transcripts on chromosome 2 and the Y chromosome. Are the results what you might expect, based on what you know about the inheritance of human sex chromosomes? Why or why not?

Section 3: NHANES.

The National Health and Nutrition Examination Survey (NHANES) is a survey conducted annually by the US National Center for Health Statistics (NCHS). While the original data uses a survey design that oversamples certain subpopulations, the data have been reweighted to undo oversampling effects and can be treated as if it were a simple random sample from the American population.

The following questions will be explored with the NHANES data:

1. At what age do Americans seem to reach full adult height?
2. What proportion of Americans age 25 or older have a college degree?
3. What is the relationship between education level and income?
4. How much more likely is it that someone *not* physically active has diabetes, compared to someone who is active?

The reweighted NHANES data are available from the NHANES package. To view the complete list of study variables and their descriptions, access the NHANES documentation page with `?NHANES`.

For convenience, descriptions of the variables used in this lab exercise are included below.

- `Age`: age in years at screening. Subjects 80 years or older were recorded as 80 years of age.
- `Education`: highest educational level of study participant, reported for participants aged 20 years or older. Recorded as either `8th Grade`, `9 - 11th Grade`, `High School`, `Some College`, or `College Grad`.
- `Poverty`: a ratio of family income to poverty guidelines. Smaller numbers indicate more poverty; i.e., a number below 1 indicates income below the poverty level.
- `Weight`: weight, measured in kilograms.
- `Height`: standing height, measured in centimeters.
- `Diabetes`: `Yes` if the participant was told by a health professional that they have diabetes, `No` otherwise.
- `PhysActive`: coded `Yes` if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities; `No` otherwise. Reported for participants 12 years or older.

Question 1.

a) Describe in words the distribution of ages for the study participants.

b) Using numerical and graphical summaries, describe the distribution of heights among study participants in terms of inches. Note that 1 centimeter is approximately 0.39 inches.

c) Use the following code to draw a random sample of 200 participants from the entire dataset. Using the random sample, `nhanes.samp`, investigate at which age people generally reach their adult height. Is it possible to do the same for weight; why or why not?

```
#draw a random sample
set.seed(5011)
row.num = sample(1:nrow(NHANES), 200, replace = FALSE)
nhanes.samp = NHANES[row.num, ]
```

Question 2.

a) What proportion of Americans at least 25 years of age are college graduates?

b) What proportion of Americans with a high school degree are college graduates?

Question 3.

a) Calculate the median and interquartile range of the distribution of the variable `Poverty`. Write a sentence explaining the median in the context of these data.

b) Compare the distribution of `Poverty` across each group in `Education` among adults (defined as individuals 25 years of age or older). Describe any trends or interesting observations.

Question 4.

a) Construct a two-way table, with `PhysActive` as the row variable and `Diabetes` as the column variable. Among participants who are not physically active, what proportion have diabetes? What proportion of physically active participants have diabetes?

```
addmargins(table(PhysActive=NHANES$PhysActive, Diabetes=NHANES$Diabetes))
```

b) In this context, relative risk is the ratio of the proportion of participants who have diabetes among those who are not physically active to the proportion of participants with diabetes among those physically active. Relative risks greater than 1 indicate that people who are not physically active seem to be at a higher risk for diabetes than physically active people. Calculate the relative risk of diabetes for the participants.

From these calculations, is it possible to conclude that being physically active reduces one's chance of becoming diabetic?