StatR 101: Fall 2012
Homework 2
Eli Gurarie
Due Saturday, October 6, 11:55 pm

Instructions: All homework must by typed. For these and subsequent homeworks, you do not need to provide code unless specifically requested. Export plots as needed from R and incorporate them into your document. Upload the completed homework assignment into the course webpage drop-box. Much of this homework builds on the code in the in-class computer lab: `lab2.r`.

Reading: Read (or skim) chapters 1, 2, 3 in the Braun Murdoch book. Much (though not all) of this material I covered in the lectures, but it is often helpful to "triangulate" by exposing yourself to other peoples' presentation.

1. **Analysis of table tennis appeal:** Download and load into `R` the `StudentSurvey.csv` data from the website. Among the columns in the undergraduate student data are two categorical columns: `Sex` and `Pingpong`. The latter is a response to the question: *"How much do you enjoy playing table tennis, on a scale of 1 (not at all) to 5 (it is basically an obsession)?"*

   (a) Formulate a prediction regarding the appeal of table tennis between male and female students.

   (b) Present a table summarizing the total number of responses for male, female and total number of students in each of the five categories.

   (c) Make side by side pie charts of pingpong enjoyment, one for males and one for females. Label each pie. Export this graphic to a file (e.g. `pdf`, `png`, `bmp`, etc) with high resolution. Hint: Use the `par(mfrow=c(nrow,ncol))` to place multiple plots on a single figure.

   (d) Produce a 2×5 matrix (call it `M1`) summarizing the proportional distribution of male and female students in each category such that $\sum_{i=1}^{5} P_{male,i} = 1$ and $\sum_{i=1}^{5} P_{female,i} = 1$ where i is the response category. This is equivalent to `sum(M1[1,])` and `sum(M1[2,])` both being equal to 1. (Hint: You can do this by dividing separate male and female vectors by the total count of males or females, or more efficiently by dividing the matrix by the output of the `rowSums(M1)` function.)

   (e) Produce a 5×2 matrix (call it `M2`) summarizing the proportion for each response of male and female respondents. $P_{male,i} + P_{female,i} = 1$ for each category i.

   (f) Produce two barplots using the following commands: `barplot(M1, beside = TRUE)` and `barplot(t(M2))`. Add a label the x-axis and customize the colors of the columns so they are not the (boring) grey default. Use the `legend()` command to add a legend identifying your unique colors with different sexes. An example of a way to do this is:

```
legend("topright", fill = c(col1, col2), legend = c("Male", "Female"))
```

(g) What conclusions do you draw from these tables and plots with respect to your initial prediction? Which of the four output plots do you feel is most informative? Why?

2. **Analysis of global patterns:** Download the CountryData.csv file and load it into R. This data file contains information on population (×1000), area (1000 km^2), literacy rate, per capita GDP ($1000), birth rate (number of births per 1000 people per year), percentage of land covered by water, and a classification by continental region.[1]

(a) Load the data into an object called CountryData. Create a separate vector for each of the columns in the data. (Note: There is no need to present anything for this problem).

(b) The following line of code uses order() to extract a list of the ten poorest countries:

```
Poorest <- Country[order(GDP)][1:10]
```

Using this code as a template, create a data frame of the 10 countries with the lowest and highest GDP per capita, the highest and lowest birth rates, and the lowest literacy. Present this as a table in your document. Comment on any patterns that you identify in these columns. Note that in order to make a data frame, you can use the data.frame() command as follows:

```
Names <- c("Alice", "Bruno", "Cassie")
Age <- c(8, 16, 32)
BirthMonth <- c("May", "June", "July")
MyDataFrame <- data.frame(Names, Age, BirthMonth)
```

(c) Identify the 10 countries with the highest and lowest densities, respectively, and present two tables that include their population, area and percentage of water coverage. (Hint: One way to do this is using the match() function.)

(d) Using the in-class lab as a model, create an overlapping *frequency* histogram of birth rates in Europe, Asia, and Africa in three different, transparent colors. Add a legend to the plot identifying the continents. Make sure that the axes are appropriately labeled and the plot has a meaningful title. Experiment with the bin widths to find one that you feel best illustrates the patterns.

(e) Create a *density* histogram of the same data, and add fitted density lines. Note that unlike a frequency histogram, in a density histogram, the bin widths can be tuned for each individual data set.

(f) Summarize the patterns in these distribution, commenting on the center, the spread, and the modality (i.e. number of humps).

[1] The source of these data are Wikipedia, e.g.: List of countries by birth rate, List of countries by area, List of countries by GDP. See sources within these articles for more details.