Mini-Project: Histograms

"Work is the best antidote to sorrow."
—SHERLOCK HOLMES

Overview
You will use R to make two histogram plots. The first plot will show the distribution of values for everyone in the data set, along with the average and SD. The second plot will show two histograms on a common scale to compare the values for two different groups within the data set, along with the two averages. Finally, you will briefly discuss the similarities and differences between the two groups.

The bulk of the work will be figuring out R. Use the many examples provided in Dr. Whalen's notes *Lesson 02: Shape, Center, and Spread*. Everything you have to do for this project has example code in the lesson notes.

The data files required for this project are posted in elearning. Download them to your RStudio project folder and add them to your workspace in RStudio.

Project Options
Choose one of the following two studies for your project that you find most interesting.

Option 1: The Cherry Blossom Ten Mile Race: Net Racing Times for Women and Men
Every year in April there is a 10-mile road race in Washington, D.C., among the famous cherry blossom trees. The results of the race are published. The data file `TenMileRace` has the numbers from the 2005 race. The file is part of the `mosaicData` R package, but it is also posted in elearning.

`state`: in which of the 50 states the runner lives
`time`: official time, in seconds, from the starting gun to the runner crossing the finish line
`net`: the recorded time, in seconds, from when the runner crossed the starting line to when the runner crossed the finish line. (There are a lot of runners in the race, so it takes some time for many runners just to get to the starting line. The `net` value is smaller than the `time` value.)
`age`: the runner's age in years
`sex`: whether the runner was female (`F`) or male (`M`)

Make a histogram to see the distribution of the `net` values, then make a plot of two histograms controlling for `sex`. Compare and contrast the distributions of the `net` values for the women and for the men.

Option 2: The Survey of Consumer Finance: Incomes and College Degrees
The Federal Reserve System carries out a nationwide survey using a complex sampling method to measure a variety of variables about income, debt, family, and education, among many others. The data from the 2013 survey are available in the R data file `scf2013` posted in elearning. There are many variables, but for this project only focus on these:
`EDCL`: "education level" of the head of household with these indicators—
1.No high school diploma, 2. High school diploma, 3. Some college, 4. College degree
`INCOME`: total amount of income of the household

For this project, consider only the households in the survey with a total household income less than $500,000. Make a histogram of the `INCOME` distribution. Then make a plot with two histograms, one for the incomes of households where the head of household has a college degree (`EDCL==4`), and another for incomes of households where the head of household does not have a college degree (`EDCL<4`). Compare and contrast the distributions of household incomes for those with and without college degrees.

Project Requirements

1. The single-histogram plot.
    a. Use the `hist` command to make a histogram in R (do not use any graphics package for this project).
    b. Choose your own `breaks` to make a readable histogram; make sure you have plenty of bins to see the shape.
    c. Clearly label the horizontal axis `xlab`, including units.
    d. Customize the `main` title as you see fit.
    e. Make a density histogram `prob=TRUE` and remove the y-axis and y-label with `yaxt` and `ylab` set to `'n'` or `NULL`.
    f. Mark the average with a solid vertical line using `abline`.
    g. Put dashed vertical lines at a distance of one SD on either side of the average.
    h. Optional: you can change the color of the histogram and lines with `col`. Pick what you want, as long as it is readable. Run `colors()` to see the hundreds of choices.
2. The two-histogram plot.
    a. Use `hist` to make the histograms in R. In the second, use `add=TRUE` to put the second histogram on the same plot as the first.
    b. Choose your own `breaks` to make a readable histogram. Make sure you have plenty of bins to see the shape. Use the same vector for both histograms.
    c. Customize the color of the second histogram (recommended: make the second histogram transparent, using `col=NULL`).
    d. Clearly label the horizontal axis `xlab`, including units.
    e. Customize the `main` title to include an indication of which histogram is which.
    f. Make them both density histograms `prob=TRUE` and remove the y-axis and y-label with `yaxt` and `ylab`.
    g. Mark the two averages with a solid vertical line using `abline`.
3. Brief discussion.
    a. How do the shapes of the two histograms compare? Are they the same? Can you roughly get from one to the other by shifting (adding) or scaling (stretching)?
    b. How do the SDs of the two groups compare? Are they similar or different?
    c. How do the averages of the two groups compare? What is the difference? How is it interpreted? Based on your answers to the previous parts, how adequate a summary would the difference between the averages be?

Project Submission

1. Fit the project requirements, items 1-3, on ONE PAGE. Images copied from R may not display with the default paste—in Microsoft Word, choose Paste > Picture (from the top-left menu area under the Home tab).
2. On the second page, list the names (full elearning name) of *anyone* that you worked with and what sources, if any, you used (except for Dr. Whalen's notes and the textbooks mentioned in the syllabus).
3. Starting on the third page, copy and paste your final R script. Make sure that the grader can copy your script, run it line by line, and recreate your work.
4. Save the document as a PDF.
5. Upload to the Mini-Project Histogram Submission assignment in elearning.

Warnings

- You are expected to submit your own individual work.
- Working together is great, but you must list anyone you worked with (by full elearning name) and be careful not simply to copy someone else's code and work.
- Beware of using online sources. If you do use any, include them by name and web address in your source list and mention what code you referenced.
- Do not use any R graphics packages (such as `ggplot`) for this project. (The higher the "data-to-ink" ratio, the better. Plots from `ggplot` can get too "busy." Of course, you're welcome to experiment with packages on your own.)