

Lab 1: Basics of R

due date: Tuesday, September 29

Make sure RStudio is installed on your computer. If not, visit <https://www.rstudio.com/products/rstudio/download/> and install the free, open source, desktop version.

Start RStudio, and on the “Packages” tab, make sure “datasets” is checked.

Enter
`data()`
to see the list of datasets available and
`data(rivers)`
to load a dataset of lengths of major American rivers.
Enter
`rivers`
to see the list of lengths.

Questions

1. Plot a histogram using `hist(rivers)`. The vertical scale is labelled “frequency”, which could mean relative frequency or counts. Which is it?
2. Plot a histogram using `hist(rivers, freq=False)`. The vertical scale is labelled “density”, which is *not* relative frequency or counts. What is it that adds to 1 in this figure?
3. Find the min and max length in the dataset using `min(rivers)` and `max(rivers)`.
4. What is the mean of the dataset? (Type `? mean` to learn about the mean command.)
5. What is the variance?, the standard deviation? (look up the commands `var` and `sd` and use them)
6. Explain everything that is produced by `summary(rivers)`.
7. Remove the largest element of the data set with `r=rivers[rivers<max(rivers)]`, and compute the mean and standard deviation of `r`. How did they change from those for the full data set? Is that as you would have expected?
8. Remove the smallest element of the data set with `s=rivers[rivers>min(rivers)]`, and compute the mean and standard deviation of `s`. How did they change from those for the full data set? Is that as you would have expected?
9. Did removing the smallest or largest length have a greater effect on the mean? on the standard deviation? Referring to the histogram for the full data set, explain why you might have expected this.
10. Make four new histograms of `rivers` using 20, 40 60, and 80 equally sized bars. Compare them, and explain why you might prefer one over another to better understand this data.