Math 2310
Lab 1 – Visualizing Data

In this lab assignment we will be using R to generate histograms.

As we have discussed, histograms are a common tool for visualizing numerical data. They allow us to see what sorts of values are more or less common, they let us identify whether a distribution is skewed or symmetric, and they let us identify if a distribution is unimodal, bimodal, or multimodal. You will get practice with these applications in Activity 1.

Histograms can also be used to compare the distributions of a numerical variable between two or more groups. You will get practice with this application in Activity 2.

Goals for this assignment:

- Learn how to load and manage data in R
- Learn how to make plots in R
- Understand what bimodal graphs indicate
- See how changing interval widths in a histogram impacts what is communicated
- Use histograms to make comparisons between groups
- Learn how to apply transformations to data to obtain symmetric distributions

The **skill objectives** and **analysis objectives** in each activity specify what you should be submitting in your lab report.

Grading: there are two possible points for each skill objective and for each analysis objective.

Activity 1

A group of botanists from the University of Toronto conducted a study looking at the feeding habits of baby snow geese. The geese were fed one of two types of food – either plants, or Purina Duck Chow. The researchers then measured their change in weight after feeding (as a percentage of initial weight), digestion efficiency (as percentage of their food that was digested), and the amount of acid-detergent fiber in their digestive tract (also measures as a percentage).

The data is included in the file snowgeese.xls. The columns are not labeled in this data set. The first variable is just an index of the trial number. The second variable is a description of the type of diet fed to the goose. The third variable is their percent weight change after being allowed to feed for 2.5 hours. The fourth variable is their digestion efficiency. The fifth variable is the amount of acid-detergent fiber in their digestive tract.

a) We will use a histogram to examine the amount of acid-detergent fiber for all geese.

Skill Objective: Using R, construct a histogram for acid-detergent fiber for all of the geese in our data set.

Hint: Refer to the Lecture 2 RMarkdown demo for example code that reads in data from an Excel file

Analysis Objective: You should notice a very bimodal histogram. Why is this? Write one or two sentences explaining what you can identify in the data set that is different about the geese whose fiber values are below 10 versus those whose fiber values are above 15, that could explain the bimodality of this graph.

b) Next, we will use a histogram to examine the weight change values, restricting our attention only to the geese fed a plant diet.

Skill Objective: In R, find a way to create a histogram displaying weight change values only for the geese fed a plant diet (there are multiple ways this could be accomplished, some easier and some more complicated).

Hint: Refer to the Lecture 2 RMarkdown demo for example code that filters a dataframe based on the values in one of the columns

Analysis Objective: Describe the shape of the histogram in one or two sentences. Then, in one or two sentences, discuss what the histogram tells us about how many of the geese tend to lose weight, gain weight, or stay at about the same weight.

c) By default, most statistical software will decide for you what widths of intervals to use for a histogram. But you can also manually specify the interval widths.

Skill Objective: Construct two new histograms for the same data as in part b) – one with wider intervals than used in the graph in part b, and one with narrower intervals.

Hint: Refer to the documentation to learn about how to change the histogram intervals.

Analysis Objective: Compare both of these to the histogram constructed in part b). In one or two sentences, explain which one of the three histograms you feel gives you the most useful summary of the data? Why?

Activity 2

A research paper from 1975 by Joanne Simpson, Anthony Olsen, and Jane C. Eden entitled "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification" examined the effectiveness of cloud seeding using silver nitrate. Between 1968 and 1972, experiments were conducted in which cumulus clouds were seeded with silver nitrate in the hopes of increasing cloud growth and prolonging cloud life, with the intention of leading to increased rainfall. Total rainfall (in acre-feet) was measured for 26 seeded clouds and 26 unseeded clouds. The data collected is in the file clouds.xlsx

a) We will use histograms to assess the impact of cloud seeding.

Skill Objective: Construct a single figure that includes two separate histograms, one for rainfall amounts for seeded clouds, and one for rainfall amounts for unseeded clouds. The two histograms should be arranged vertically, like the example we saw in class, with the same scale for the x-axis.

Hint: Refer to this page to learn how to stack plots on top of each other with R

Analysis Objective: Based on this pair of histograms, write one or two sentences discussing what you can tell about how cloud seeding impacts rainfall.

b) Note that both of the histograms are skewed. Many of the common statistical techniques that we will be learning later this quarter can only be used with data from symmetrical histograms (specifically, normal or Gaussian histograms, which we will discuss later). When we have data like this cloud seeding data with skewed histograms, we sometimes perform some sort of transformation to the data so that the result will have a more symmetric histogram. Common types of transformations include taking the logarithms of the data values, or taking various roots or powers.

Skill Objective: Using R, create three new variables (i.e., columns) in your data set, representing three different transformations of the data. You can choose which transformations to use. For each of the three new columns, create a pair of histograms similar to what you did in part a of this activity. Label what transformation was used for each new pair of histograms.

Analysis Objective: Of the three transformations you considered, which one gave the most symmetric histograms?