Lab 1: Describing Data with Graphs and Numbers

Objective: In this module, you will use some graphical and numerical tools to summarize the distribution for a quantitative variable or response – a histogram, a boxplot, mean, median, standard deviation, and interquartile range (IQR). You will also be introduced to side-by-side boxplots for comparing two or more distributions and bar charts for summarizing categorical data. These techniques can be very useful at the start of data analysis to get a feel for the data.

Overview: Two graphs that can be used to summarize the distribution for a single quantitative variable or response are a **histogram** and a **boxplot**. Each graph provides different information about the distribution. When used properly, graphs can be a very effective way to summarize data. Data on a single quantitative variable should first be examined graphically. The overall shape of the distribution and existence of outliers can generally be used to assess if the data appear to be coming from a relatively homogenous population. If so, then various numerical summaries may be used to characterize the center of the distribution (such as mean and median) and the spread of the distribution (such as the standard deviation and the IQR). For categorical variables, a **bar chart** can be used to display the number falling in each category (frequency distribution).

Histograms: A histogram displays the distribution of a quantitative variable by showing the frequency (count) or percent of the values that are in various classes. The classes are typically intervals of numbers that cover the full range of the variable. Histograms can be used to assess the **symmetry** and **modality** of a single distribution or for comparing the relative locations and shapes of several distributions.

Boxplots: One plot that can detect extreme observations or outliers is the **boxplot.** A boxplot is a graphical representation of the five-number summary, namely, the minimum, first quartile, median, third quartile, and maximum of the data. The centerline of the box marks the median or the 50^{th} percentile. The sides of the box show the first (lower) quartile, Q1, and the third (upper) quartile, Q3. Thus a boxplot shows the overall range (maximum – minimum) and the **interquartile range** (IQR = Q3 – Q1). A modified boxplot uses a rule for identifying values that are extraordinary compared to the others (**outliers** or **outside values**). Circles (o) are used to denote outliers and asterisks (*) to denote extreme outliers if any are present. Any point below Q1 – (1.5 x IQR) or above Q3 + (1.5 x IQR) is considered an outlier. Extreme outliers are points below Q1 – (2 x IQR) or above Q3 + (2 x IQR). *Box plots <u>cannot</u> tell you the <u>shape</u> of the distribution.*

Side-by-side Boxplots: These plots are helpful for comparing two or more distributions with respect to the five-number summary. For example, suppose you are interested in comparing the distribution of a variable, such as the salary of the employees of a certain company. If you have information on sex for the group, you might be interested in comparing the distribution of salary of females with respect to males. In this case, the side-by-side boxplot will be an important part of the descriptive analysis of the data set involved.

Bar Charts: One way to display the number or frequency distribution for a categorical variable is with a bar chart. A bar chart shows the percentage of items that fall into each **category** or value of a **categorical variable**. It displays a bar for each category with the height of each bar equal to the number, the proportion, or the percentage of items in that category**.** If the categories have no inherent order, we could rearrange the bars in the graph in any way we like. In such cases, the shape of the bar graph would have no bearing on its interpretation.

Measures of Center: Measures of center are numerical values that tend to report the middle of a set of data. The two that we will focus on are the mean and the median.

1. **Mean:** The mean of a set of *n* observations is simply the sum of the observations divided by the number of observations, *n*.

2. **Median:** The median of a set of observations, ordered from smallest to largest, is a value such that at least half of the observations are less than or equal to that value and at least half the observations are greater than or equal to that value.

Measures of Variation or Spread: Measures of variation include the IQR and standard deviation. These numerical summaries describe the amount of spread that is found among the data, with larger values indicating more variability.

a. **Standard Deviation:** Standard deviation is a measure of the spread of the observations from the mean. It is actually the square root of an average of the squared deviations of the observations from the mean. *We can think of the standard deviation as approximately an average distance of the observations from the mean.*

b. **IQR:** The IQR measures the spread of the middle 50% of the data. It is defined as the difference between the 3rd quartile (Q3) and the 1st quartile (Q1). These quartiles are also called the 75th and 25th percentiles, respectively. IQR = Q3 – Q1.

Warm-Up: Mean and Median

We are interested in analyzing Lebron James' scoring output by game. We have his scoring output for 5 games which we have arranged from lowest to highest: 6, 24, 28, 34, 36. There are two measures of center we could report.

Which measure would be better to report?	Median		Mean
Which is the most likely value for the mean?	8	12	26
What is the median?	24	28	34

ILP: Visualizing and Exploring a Data Set

In this activity, you will learn how to create graphs and obtain descriptive statistics for a data set using R.

Task: The data set **employee.Rdata** contains information on employees at a company. Explore possible questions this data could be used to address. Create appropriate graphs and obtain descriptive statistics for current salary, and discuss the results.

1. First, we must start up R Commander within R. To do this from within R, select **Packages > Load package** and from the list, scroll down and select **Rcmdr** and click OK. This will boot up the **R Commander window.**

2. To obtain the data set, go to Canvas, and find the "Data Sets" folder under the "Files" tool. Select **employee.Rdata** and save it to a directory of your choice. To open the **employee.Rdata** data set from within R Commander, click on the **Data** menu at the top, and then select **Load data set**. Open **employee.Rdata** from the directory you saved the file in.

3. We now see that our data is loaded, as we can see **employee** in the active data set box (next to the R logo). Before we begin analyzing the data, let's first view what the data set looks like – to do this, click on **View data set**. Here, you can see the variables in the data set and their values. The first variable you should see is ID.
   What is the second variable present in the data set?
   What type of variable is it?
   What is the eighth variable present in the data set?
   What type of variable is it?

4. Create a histogram for current salary. Use the graphs menu – **Graphs > Histogram** and select SALARY, then click OK.

Note: All homework and prelab assignments will require that students provide an appropriate title and their name on each R chart or output. For histograms, click on the **Options** tab and enter your title there.

   Draw a quick sketch of the histogram and describe what the histogram shows about the distribution of current salaries. A good description will have information about the shape, general center, variability, and if relevant, a comment about potential outliers.

5. You would like to compare the distribution of salary for minorities versus non-minorities. Generate histograms again, but this time, click the **Plot by groups** button and select MINORITY as the grouping variable. Note: Since the dataset contains far more non-minorities, it may be useful to go to **Options** and select **Percentages** for the axis scaling. You can also change the number of bins (or bars) if you are interested in seeing a finer detail. Compare and contrast the distribution of salary for the two groups – can we use the same descriptions for both histograms?

6. Obtain a boxplot for current salary. Use: **Graphs > Boxplot** and select the SALARY variable again. Make a quick sketch of this boxplot, and describe what the boxplot shows about the distribution of current salary. What do the various lines on the boxplot represent?

7. As we did with histograms, we can also use side-by-side boxplots to compare the distributions. Return to the dialog for boxplots, and again click the **Plot by groups** button and select MINORITY as the grouping variable.

   How does the distribution for current salary compare for minorities versus non-minorities?

8. Numerical summaries may also be obtained for any quantitative variable. Basic descriptive summaries can be obtained via **Statistics > Summaries > Numerical Summaries**. The summaries can then be found in the output box. Fill in the basic summary measures for current salary (some require hand calculation).

Mean:	Median:	Standard Deviation:
Q1:	Q3:	IQR: Q3-Q1 =
Min:	Max:	Range: Max-Min =

Cool-down: Which Measure of Center to Report?

Mark is a Stats 250 GSI who would like to report a measure of center for scores on the first exam. The mean score for his lab section was 77.46 points and the median was 84 points. One of Mark's students did not take the exam and received a zero. Since Mark knows this score will not count against the student, he removes the score of zero from his data.

How will the mean test score change if the grade of 0 is not included?

If there is an outlier test score of 20 points, which measure of center would you recommend that Mark report?