Lab Assignment 3: Descriptive Statistics

Creating and Analyzing New Variables.
Let’s continue using the CerealJMP data set from last time. We are going to create a new variable by "standardizing Complex Carbs".

1. First we need to create a new column, Cols → New Column, and call it Standardized Complex Carbs. Click on New Property and select Formula. Select ”Complex Carbs” under Table Columns, then use the keypad to subtract the mean of ”Complex Carbs”. We do this by first selecting ”Complex Carbs” again and then selecting Statistical → Col Mean under Functions. Click on the entire formula, so there is a red box around the entire quantity. Then use the keypad to divide by standard deviation of ”Complex Carbs”. It does not matter whether which order you select first - Statistical → Col Std Dev or ”Complex Carbs.” After the formula is complete, click on Apply. Directly after selecting Apply you should see ”Evaluations done” at the bottom left of the screen, then click OK. Once you get back to the New Column window, click on Apply → OK again. Look at the left hand side of the screen where the variables are listed. The ”Standardized Complex Carbo” variable should have a yellow box with a black plus sign in it, this indicates the variable was created using a formula.

2. Next, make a histogram of the new variable and make the graph orientation horizontal. Click on the red arrow next to variable name and select Fit Distribution → Normal. This will overlay a normal curve over the histogram. How well does the normal curve approximate the data?

The normal curve fits the data well on the right half of the histogram, but not the greatest on the left side. A Normal Quantile plot also illustrates this fit.

3. Now let’s focus on the elements of the box plot. The rectangle represents the IQR and the vertical line the median. The diamond within the IQR gives a 95% confidence interval for the mean, which is indicated by the vertical points of the diamond. (We will discuss 95% confidence intervals later in the course.)

4. To construct a Normal Quantile Plot click on the red arrow next to the variable name and select Normal Quantile Plot. If the normal curve is a good approximation of the data, then the points should fall within the two red confidence bands. Based on this plot, would you say the normal curve is a good approximation for this variable?

As in the previous question, the fit is good for half of the data, but not good for the other half.

5. Create four new columns: ”Calories” + 25, ”Calories” - 25, ”Calories”*2, ”Calories”*0.5. Draw a histogram of ”calories” and the four transformations of the ”calories”. Specify all 5 histograms with a single use of Analyze → Distribution of Y. You can do this by highlighting the desired variables, then selecting Y, Columns. Repeat until all five have been selected. Now select the red arrow to the left of Distribution and click on Stack.

- How does adding 25 to calories affect the histogram?
  Shifts the histogram to the right
• How does subtracting 25 from calories affect the histogram?
  \textbf{Shifts the histogram to the left}

• How does multiplying calories by 2 affect the histogram?
  \textbf{Expands the histogram by a factor of 2}

• How does multiplying calories by 0.5 affect the histogram?
  \textbf{Compresses the histogram by a factor of 2}

\textbf{Identifying Observations within a Variable.}
Draw histograms for ”calories” and ”manufacturer” in a single plot. Click on the bar representing Kellogg. Notice what happens in the histograms of calories. Alternate clicking on the histogram for different manufacturers. Watch what happens in the histogram of calories. This gives a preliminary look at the calorie distribution within each group.

\textbf{Importance of Plotting Histograms.}
It is always important to plot the histogram of variable to check for irregularities that may occur. Plot the histogram of ”Total Carbs”, and explain what you see. What happens when you overlay a normal curve over the histogram? Normal Quantile plot?

The histogram for Total Carbs is bimodal, meaning that it has two different peaks. When a normal curve is drawn over the histogram is does not approximate that data well because it is very flat (compensating for both peaks) and not a good fit. The normal quantile plot tells the same picture.

\textbf{Other Options.}
Select the \textbf{brush} (next to the hand) and investigate the affects the tool has on a histogram.