

NAME: Owen Sullivan

OBJECTIVES: Upon successful completion of this activity, you will be able to...

- Identify the type of variable in the context of a problem
- Use software to obtain a random sample
- Create histograms, stem and leaf plots, and dot plots for quantitative variables
- Calculate several statistics for quantitative variables
- Use the output provided by JMP to answer relevant questions

Delectable Delights is a large consumer food manufacturer and supplier selling its products in retail stores nationwide. You have landed your first job after graduation from Clemson in their advertising division. Since you took statistics as a part of your coursework, you are often called upon to perform data analysis for the advertising division, as well as other divisions of the company.

DIRECTIONS: Answer the following questions using complete sentences as though you were presenting your analysis to the employees of Delectable Delights. Please provide any appropriate output and/or screenshots from JMP. Instructions for creating several types of graphs or tables and statistics can be found on Canvas in the file **JMP Instructions.docx**. Paste your answers and any output into this document.

In the file **Candy Bars** is data on several variables for the most popular candy bars supplied by Delectable Delights. You will use the information from the file to practice graphical techniques as well as calculating and interpreting various statistics used with quantitative data. You will make your calculations by hand and with the assistance of JMP.

Calories

1. Open the file candy bars in JMP and look at the data in the column **Calories**. Since the values of calories are obtained from measuring instead of counting, they are usually considered to be **continuous** quantitative data. However, given how the data is presented, could you categorize the quantitative variable Calories another way? Explain your answer. (5 pts)

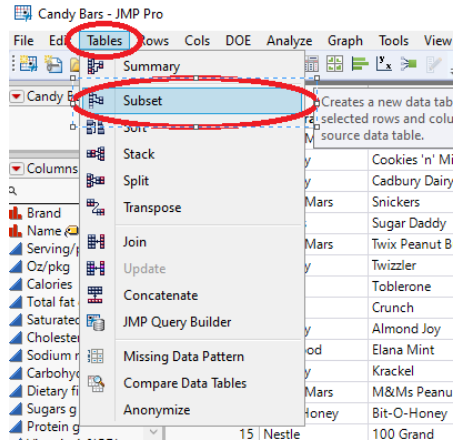
Calories are typically categorized as continuous because you can measure them and you can have, for example, half a calorie instead of just whole number of calories. However, based on how the data is presented for the quantitative variable Calories (which is in intervals of 10 calories), one could argue that the data is discrete because only integer values with intervals of 10 are presented.

2. Is **Calories** a nominal, or ordinal, or interval, or ratio variable? Explain your answer. (5 pts)

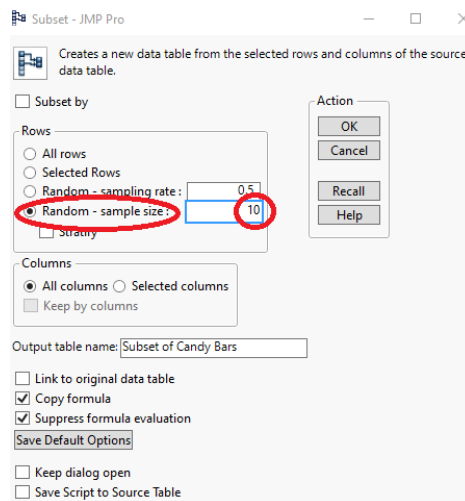
Calories is an interval variable because the data can be categorized, ranked, and evenly space but it does not have a natural zero.

3. So that the calculations you make by hand are simpler, take a random sample of size 10 from the data.

Step 1: In JMP, go to **Tables >> Subset**.



Step 2: In the section for **Rows**, select **Random-sample size**: enter **10** for the Random sample size. Select **OK**.



Step 3: Take a screen shot of your data table showing at least the first 5 columns. Paste your data table below. (5 pts)

	Brand	Name	Serving/pkg	Oz/pkg	Calories
1	M&M/Mars	Snickers Peanut ...	1	2	310
2	Hershey	Cadbury Dairy Milk	3.5	5	220
3	Hershey	Twizzler	1	2.2	190
4	Sherwood	Elana Mocca	1	1.6	230
5	M&M/Mars	Skittles	1	2.17	250
6	Tootsie	Charleston Chew	1	1.875	230
7	Nabisco	Planters Original ...	1	1.6	230
8	Hershey	Special Dark	1	2.6	410
9	Hershey	Reese's Pieces	1	1.63	230
10	Hershey	Cadbury Caramello	3.5	5	190

4. What is the mode of the variable **calories** from your sample? If there is no mode, simply state that. Remember to include **units**. (5 pts)

The mode of calories from my sample is 230 calories.

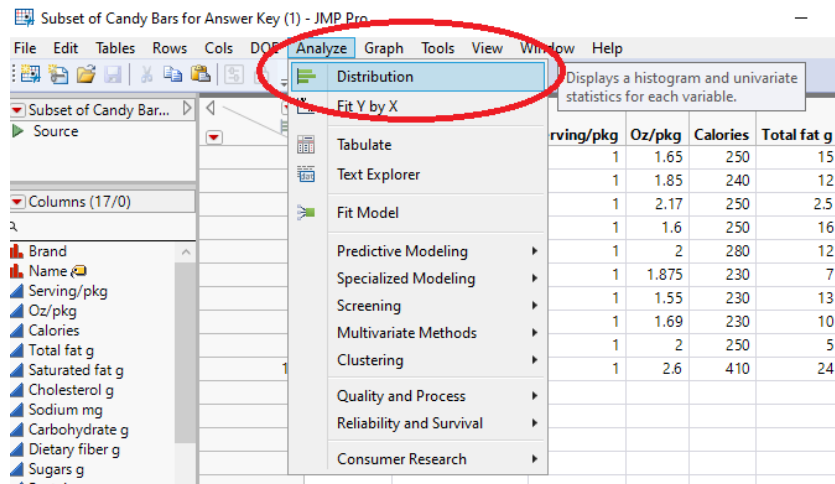
5. Using your sample, calculate the mean number of **calories** in the candy bars by hand. Include the **symbol** for a **sample mean** and **units** in your answer. Show your work by **filling in the formula** for the mean. You may type your formula or handwrite it and insert a picture of it into this document. (5 pts)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{310 + 220 + 190 + 230 + 250 + 230 + 230 + 410 + 230 + 190}{10} = 249 \text{ calories}$$

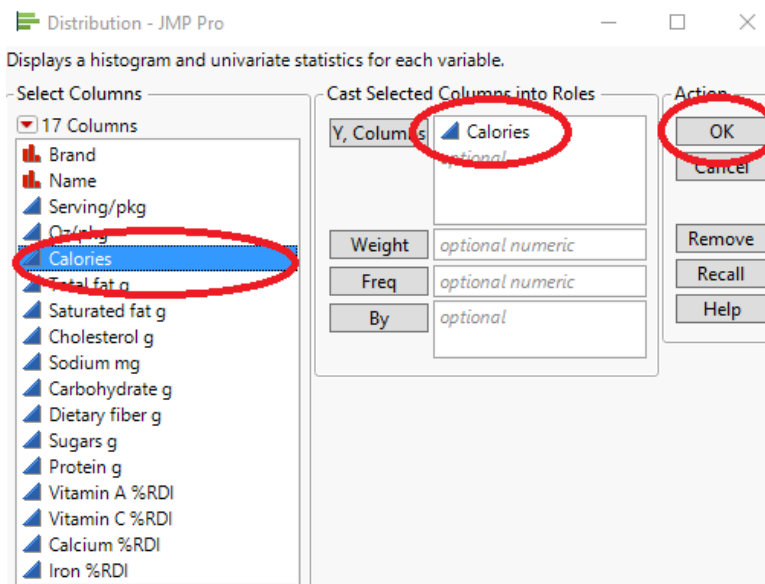
6. Show how to **fill in the formula** to calculate the variance of the variable calories using your sample data. Include the **symbol** for a **sample variance**. You do not need to calculate this value by hand. (5 pts)

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{10} (x_i - 249)^2}{10-1}$$

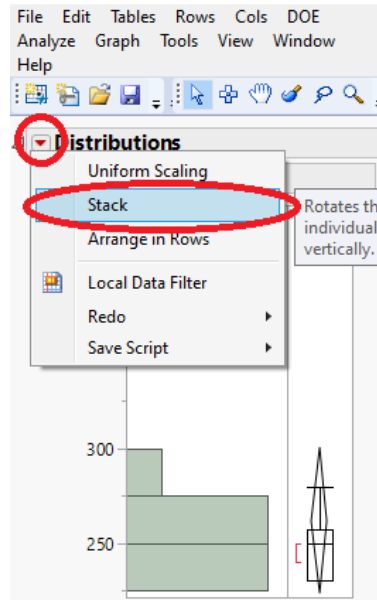
7. Use JMP to calculate basic summary statistics for the **calories** variable from your **sample**. Step 1: Go to **Analyze >> Distribution**. Make sure to do it on your sample, not the entire dataset.



Step 2: Put the column for **Calories** into **Y, Columns** and select **OK**.



Step 3: Change the orientation of the output to a horizontal format. To do this, select the triangle to the left of the word **Distributions** in your output and choose **Stack**.



Step 4, To paste a copy of the results into this document you can either take a screenshot of the Summary Statistics table, or right click on the table and copy it here. Also, you can double check your work for question 5 to be sure you found the mean correctly. (5 pts)

Summary Statistics	
Mean	249
Std Dev	65.735159
Std Err Mean	20.787282
Upper 95% Mean	296.0241
Lower 95% Mean	201.9759
N	10

8. Using the information in the Summary Statistics table, report the sample variance and the sample standard deviation of the variable **calories** from your sample. You may round your answers to 2 decimal places. Include the correct **symbols** and **units** in your answers. (5 pts)

Sample variance: 4321.11 calories

Sample standard deviation: 65.74 calories

9. What is the meaning of the standard deviation in the context of your data set? (5 pts)

Each value in the sample is, on average, 65.74 calories away from the mean of 249 calories.

10. Make the following changes to the default histogram that accompanies the Summary Statistics for the variable **calories** from your sample and copy it into this document. (5 pts)

Step 1, create a histogram: We have a sketch of histogram in the same JMP window where you find the mean and standard deviation. Make sure the histogram is horizontal, otherwise check the steps in question 7.

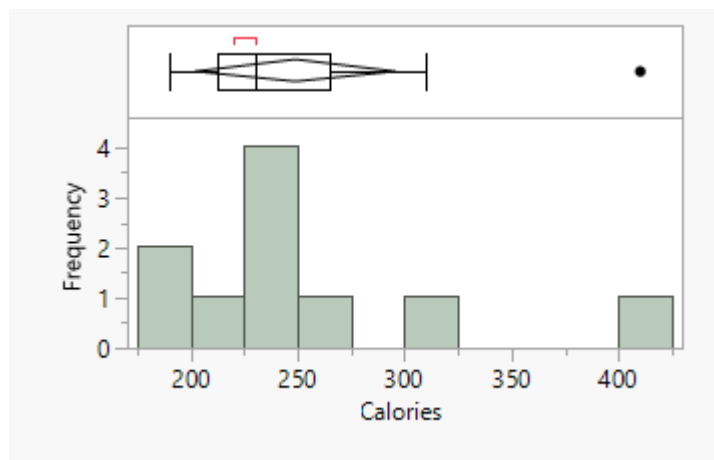
Step 2, modify axis/labels:

For y-axis, click the red triangle to the left of the column heading (in this case it is Calories). Select **Histogram Options >> Count Axis** or **Prob Axis**. This will place either the frequency or relative frequency axis on the histogram depends on what you want to plot. Once more click the red triangle and select **Display Options >> Axes on Left** to change the location of y-axis. Change the label of the y-axis to Frequency instead of Count, or Relative Frequency instead of Probability by single clicking on the default label.

For x-axis, right click on the values on x-axis and select **Add Axis Label**, then enter the name "Calories".

(Optional) Step 3, aesthetics: If you would like to change the color of the bars, right click in the white space around the bars and select **Histogram Color**. Note, we cannot separate bars in histogram for quantitative data as what we did in bar chart for qualitative data, since x-axis is a continuous axis now.

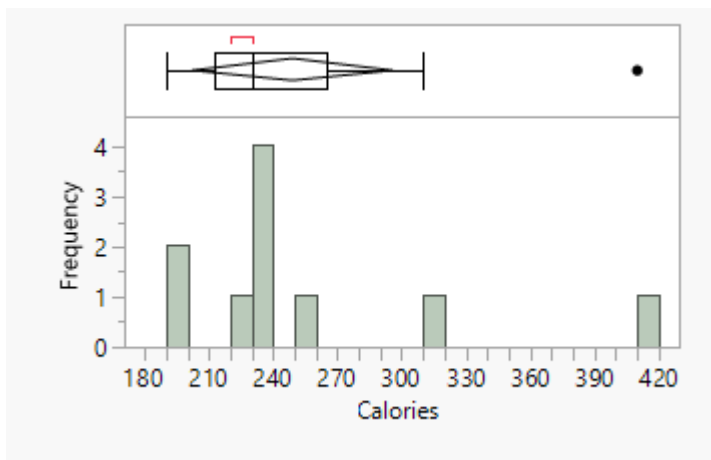
Step 4, copy/paste your result: Right click on the white space around the bars and select **Edit >> Copy Graph**, then paste it here.



11. JMP has selected a default bin width for the histogram. What is it for your sample? (If you hover your cursor over one of the bars in the histogram JMP will display the endpoints of the bin.) (5 pts)

The bin width is 25 calories.

12. Change the bin width to another value that you think shows the data nicely. Which bin width you want to try will depend on your sample. To change the bin width, select the **triangle** to the left of the column name of **Calories** above the histogram. Select **Histogram Options** and choose **Set Bin Width**. Paste the new histogram below. (5 pts)



This is with bin width 10, which I chose because the values in the sample for the variable calories are all differing by 10 calories.

13. The mean number of calories for the entire data set is 243.03 and the standard deviation of calories for the entire data set is 62.00. Using the **first** data point of the variable **calories** in your **SAMPLE**, calculate its z-score. Show your work by **filling in the formula** for a z-score. After you have the z-score, **interpret** this z-score using context of the problem. (5 pts)

$$z = \frac{310 - 243.03}{62} = 1.11616667$$

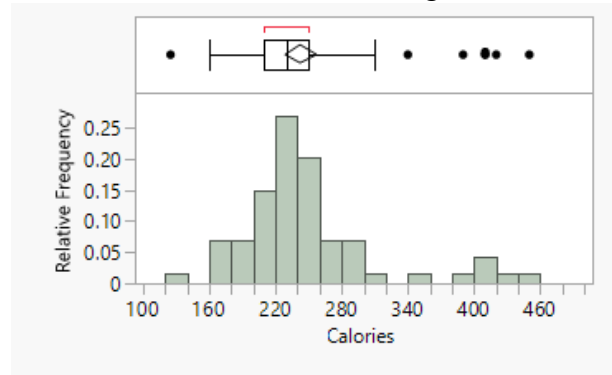
The first data point in my randomly generated sample, which represents a random candy bar from Delectable Delights, is more than one standard deviation away from the average number of calories for a candy bar from Delectable Delights. In other words, it has a relatively higher amount of calories compared to the average candy bar from Delectable Delights.

14. Again, we know the mean number of calories for the entire data set is 243.03 and standard deviation of calories for the entire data set is 62.00. Use Chebyshev's theorem to fill the following blanks: (9 pts)

At least 75% of calories data in the entire data set should lie between 119.03 and 367.03.

At least 84% of calories data in the entire data set should lie between 88.03 and 398.03.

15. Suppose the histogram of calories for the entire data is given:



Can we use the Empirical rule to estimate the proportion of calories data in the entire data set that will fall within 2 standard deviations from the mean? (5 pts)

- If your answer is no, explain the reason;
- If your answer is yes, explain the reason, and calculate the proportion by the Empirical rule.

No, because there are too many outliers resulting in the histogram not being shaped closely enough to a bell curve.

Sugars

16. Back to our SAMPLE and focus on the variable **Sugars** now. Make a stem and leaf plot for the variable **Sugars** column from your SAMPLE by hand. To order the values, go to the data table that contains your sample. Select **Tables >> Sort** and put the **Sugars** column into the section labeled **By**. Remember to include a key along with units to illustrate how to read your plot. (5 pts)

Stem and Leaf		
Stem	Leaf	Count
4	7	1
4		
3		
3	04	2
2	6	1
2	1344	4
1	9	1
1	3	1

1|3 represents 13

17. Using the **Sugars** column in your SAMPLE, calculate the range, include units. (5 pts)

$$47 - 13 = 34 \text{ grams of sugar}$$

18. For the variable **Sugars** in your SAMPLE, by hand calculate:

(a) 37th percentile

(b) 60th percentile

Explain how you arrived at each answer. (6 pts)

$$a) \quad l = 10 \left(\frac{37}{100} \right) = 3.7$$

$$b) \quad l = 10 \left(\frac{60}{100} \right) = 6$$

For each percentile answer, I used the Pth percentile formula. The sample size n is 10 for my sample, and P is equal to the Pth percentile being calculated. The 37th percentile is 3.7 grams of sugar, and the 60th percentile is 6 grams of sugar.

19. Interpret the 37th percentile you obtained from Question 18 based on the context. (5 pts)

Approximately 37% of the candy bars in the sample of size 10 contain less than or equal to 3.7 grams of sugar.