STAT 151 A1

GROUP # 65

LAB 2

HWANG, Jinny

GRAD, Brianna

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

1. Suppose the amount of cola dispensed by a filling machine follows a normal distribution with a mean (μ) and a standard deviation (σ). Select the *Calculators* option in the Stat menu and then the *Normal* option. This applet contains a graph of the normal density function and a calculator that enables you to calculate normal probabilities when the parameters (μ and σ) are provided . Use the applet to answer the following questions:

    a. Assume that the mean amount dispensed by the machine is set at μ = 300 ml. Describe what happens to the percentage of underfilled bottles (the bottles containing less than 300 ml) when σ decreases or increases? In general, how does the magnitude of the standard deviation affect the filling process?

        ➛ When σ decreases or increases, the x-axis remains the same, however, the y-axis' range decreases by a quotient factor of the standard deviation value. As for the percentage of unfilled bottles, it remains unaffected by the changes of σ. Because the mean is set at 300, and we are looking for a percentage of unfilled bottles that are less than 300 in a normal graph (the median value), it will always calculate all the unfilled bottles on the left half side of the graph, which stays constant at 50%.

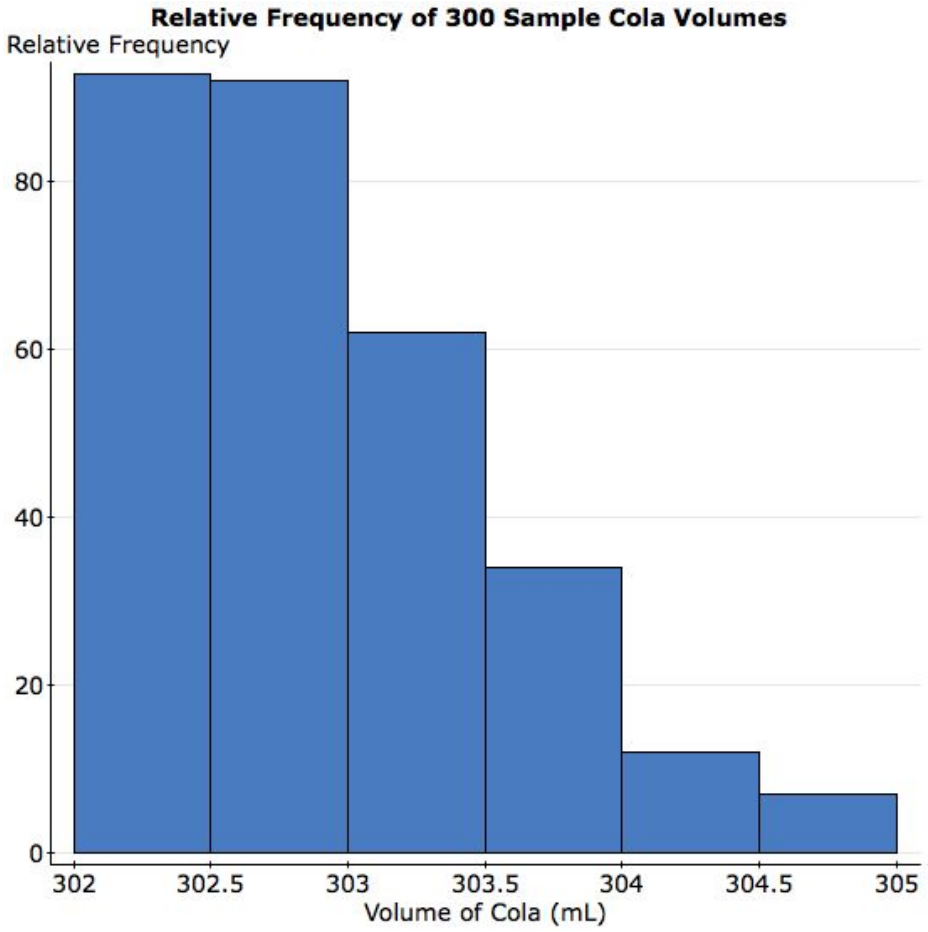
    b. Now assume that the mean amount dispensed by the machine is set at μ = 302 ml. Enter the value of σ as 2 ml. Calculate the percentage of underfilled bottles (the bottles containing less than 300 ml) in this case. What is the percentage of underfilled bottles if σ were 1 ml and 0.5 ml? In general, what is the effect of decreasing σ on the percentage of underfilled bottles?

        P (X < 300) = 0.15865525; there are 15.9% of underfilled bottles in this case.
        P (X < 300) when σ is 1mL, is 0.02275013, or 2.28% of underfilled bottles
        P (X < 300) when σ 0.5mL, is 0.00003167, or 0.00317%
        ➛ When σ decreases, the percentage of underfilled bottles decreases as well, since there is a smaller range of deviation. Judging by this equation, z= (y-μ)/σ, we know that σ and z-score have an inverse relationship, and so, when σ decreases, the z-score value increases. However, we must also note that the z score will be negative as the y value in this situation is 300 and the mean is 302, and 300 - 302 = -2. Connecting these two together, when σ decreases it will have a z-score number that is smaller (the number itself will be greater, however since it is a negative, essentially it is a smaller value), and vice versa. A smaller z-score value near the top of the *APPENDIX D - Tables and Selected Formulas* chart shows a smaller percentage in comparison to the bigger values near the bottom of the first chart.

2. Consider a random sample of 300 bottles obtained from the population of all bottles filled by the machine over a specific short time period. The volume amount of cola in each bottle is determined. The 300 observations recorded in the column volume are available in the data file *lab2a.txt* in eClass. Given the very large sample size, we may assume that the distribution of the volume amount of cola in the sample is close enough to the population distribution and its mean and standard deviation are close to the population parameters (μ and σ).

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

a. Obtain a relative frequency histogram of the 300 observations with the bins starting at 302 and using a width of 0.5. Paste the histogram into your report. The format of the histogram should be the same as the format of the histogram in *Lab 1 Instructions* (labels at the axes, title).

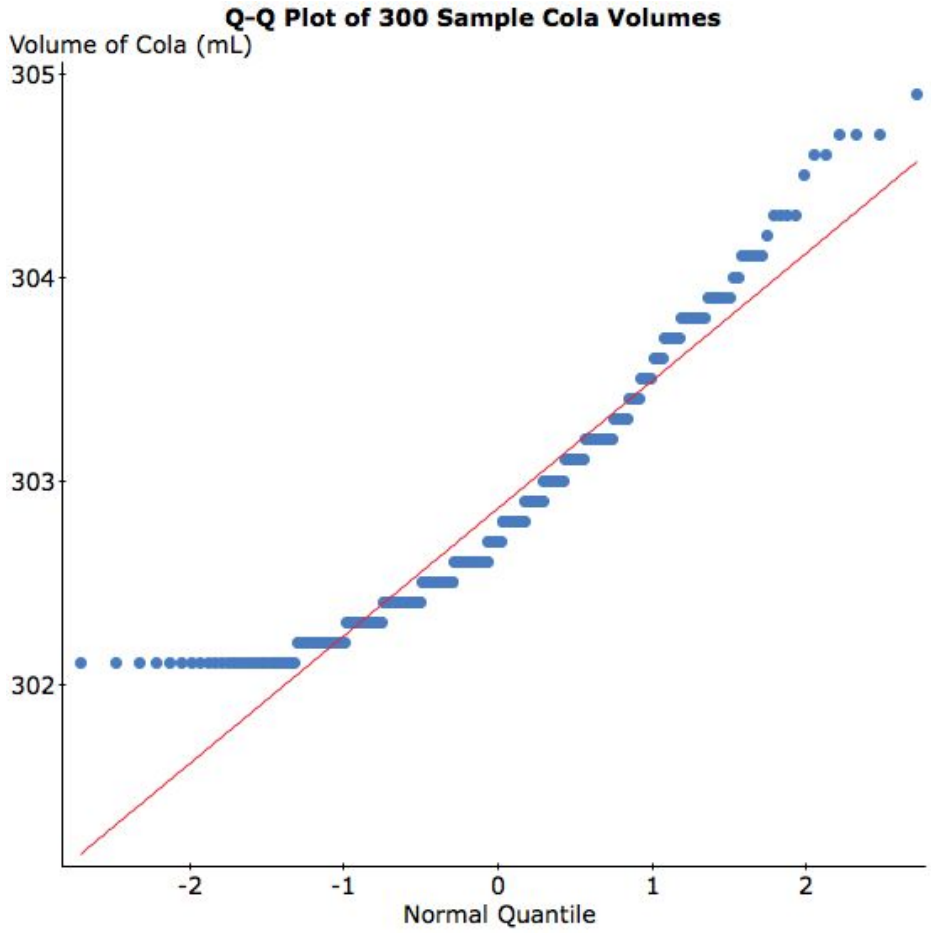


b. Describe the shape of the histogram obtained in part (a). Does the histogram support the claim of the company that the bottles are slightly overfilled?

➛ The histogram is heavily right-skewed, which means that there are less cola bottles that are super overfilled (roughly > 303 mL) than there are ones that are slightly overfilled (300 ml < Volume of Cola < 303 mL). The histogram fully supports the claim that company bottles are overfilled, as all the bottles are shown to be over the volume of 300mL.

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

c. Obtain the Q-Q plot for the 300 observations. Add a title to the plot. Paste the plot into your report. Does the plot confirm your findings in part (b) about the shape of the distribution?



➛ Since there aren't any bottles with a volume less than 302mL, it formed a horizontal line on the bottom left corner, rather than dots along the normal distribution line. As for the shape of the distribution, it is slightly right skewed, therefore, it matches the findings in part (b). The majority of the graph is aligned with the normal line, but at the top right corner it slightly goes over the line, which represents the right-skewness.

d. Use the *Summary Statistics (Columns)* feature to obtain the summary statistics (use the default options) for the 300 observations. Paste the summaries into your report. Is the relationship between the mean and median, as well as the relationship between the three quartiles, consistent with the observed shape of the histogram in part (b)?

Column	n	Mean	Variance	Std. dev.	Std. err.	Median	Range	Min	Max	Q1	Q3

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

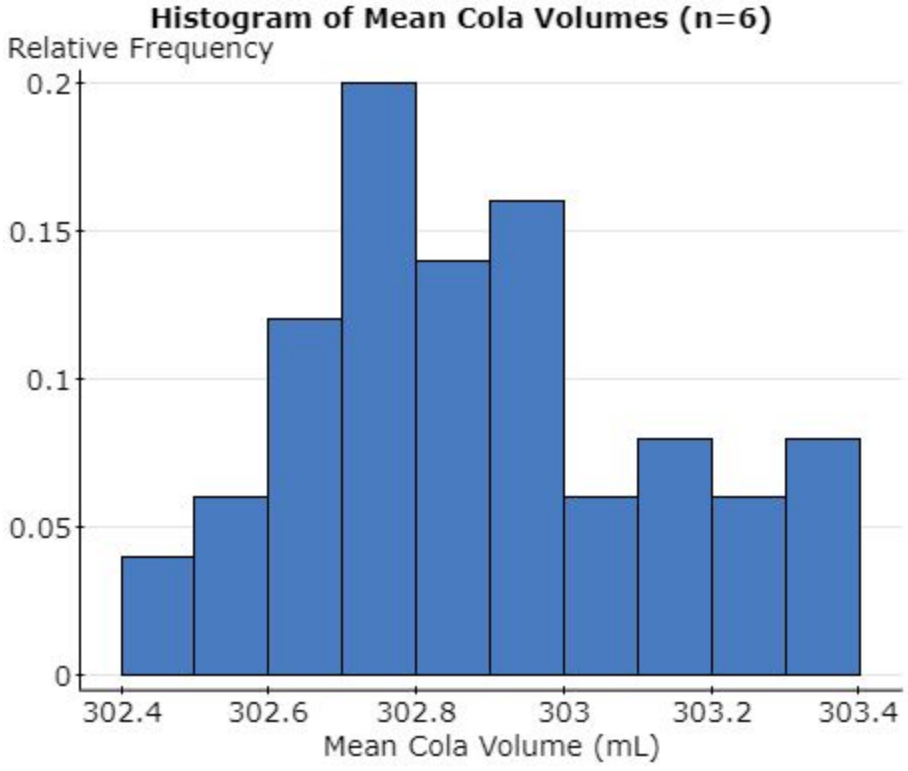
volume	300	302.866	0.392	0.626	0.036	302.7	2.8	302.1	304.9	302.4	303.2

➛ Due to no signs of outliers, the mean, median, and the three quartiles are consistent with the observed shape of the histogram in part (b). The mean and median values are close to the value of 302 mL, as the majority of the cola bottles were filled up to that volume. For the quartiles, they were also not very far from 302 mL. This is due to the fact that the standard deviation isn't very high, therefore, the range is also not very wide.

Suppose that 50 boxes are randomly selected, each consisting of 6 bottles of cola obtained from the population of all bottles filled over a certain short time period. The amount of cola in each bottle is determined. The measurements are saved in a table consisting of 6 rows (sample size) and 50 columns (number of random samples) that occupies the columns *Sample1(volume) – Sample50(volume)* in the StatCrunch file *lab2a.txt.*

3. Obtain the mean amount of cola for each sample consisting of 6 bottles with the Summary Stats (Columns) feature and save the results in a column. Make sure that all 50 columns are included in the right panel of the Column Statistics dialog box.

    a. Obtain a relative frequency histogram of the 50 means with the bins starting at 302.4 and using a width of 0.1. Paste the histogram into your report. The format of the histogram should be the same as the format of the histogram in Lab 1 Instructions (labels at the axes, title).

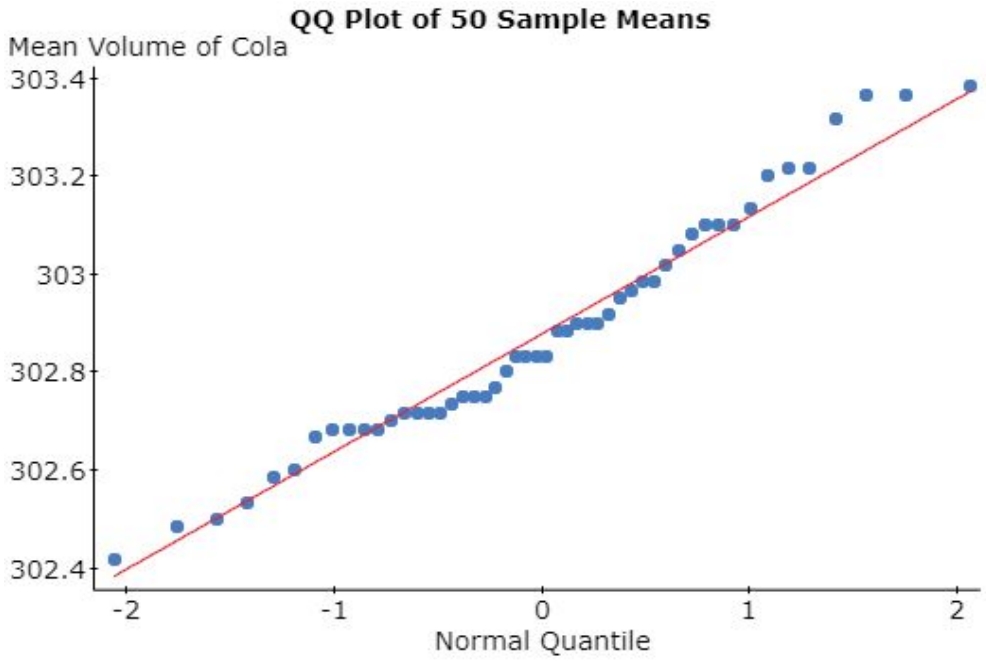


Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

b. Refer to the histogram obtained in part (a). Do the data appear to be normally distributed? Compare the distribution of the means to the distribution of individual observations studied in Question 2 in terms of their spread and degree of skewness.

➛The data appears to be slightly skewed to the right. Aside from that, the distribution is relatively even in terms of a "normal" curve. The data from the histogram in part (a) appears to be consistent with the observations made in question 2. Very few bottles of cola have volumes of 304 mL or more. Most of the bottles are of volumes equal to or less than 302.9 mL. This supports the company's claim that the bottles will be slightly overfilled.

c. Obtain the Q-Q plot for the 50 means. Add a title to the plot. Paste the plot into your report. Does the plot confirm your findings in part (b)? Compare the plot with the one in part (c) of Question 2.



➛ Yes, the data does appear to confirm the findings from part (b). The data points are mostly centered in the graph, with fewer points near the ends. They do appear to deviate quite a bit from the line between the first and second quantiles.
Because this is a QQ plot for the means of sample groups, it appears more even than the one plotting each of the 300 observations. The larger values used in the calculations would have evened out the means so they weren't as concentrated along the bottom left corner of the plot.

d. Use the Summary Statistics (Columns) tool to obtain the mean, and standard deviation of the 50 means. Paste the summaries into your report. Compare the values with the mean and the standard deviation of the sampling distribution of the sample mean predicted by the theory of sampling distributions. What does the standard deviation mean here?

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

Summary statistics:

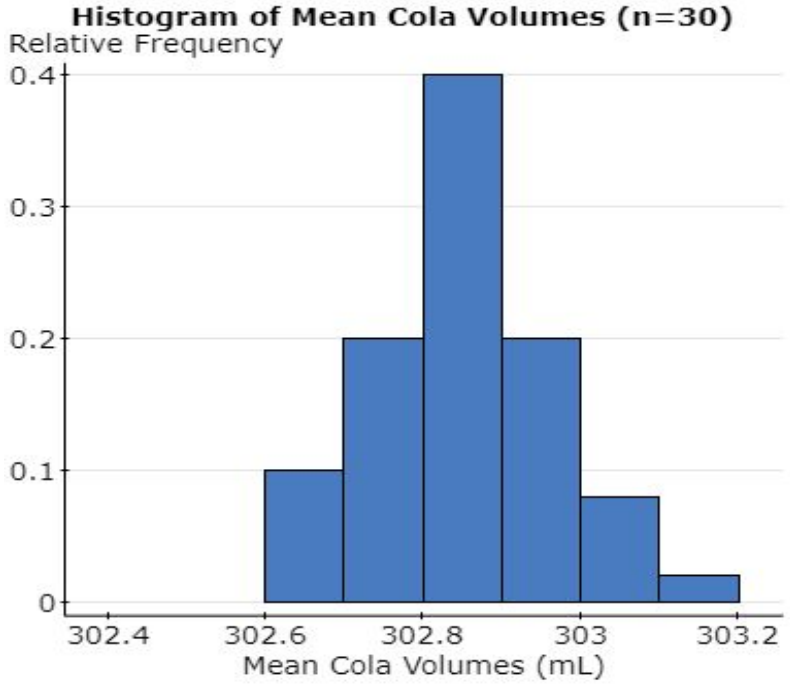
Column	Mean	Std. dev.
Mean	302.87767	0.24005408

➛ In this case, the standard deviation represents how many millilitres there are in each deviation from the center value, or the mean. This would be ± x0.2401 mL for each deviation. According to the theory of sampling distributions, a normal distribution should be a bell shape, with 66% of the distribution lying between the first standard deviation. The summary stats and the relative frequency histogram are consistent with this theory. Each normal quantile on the histogram appears to be around 0.24mL difference, which is the standard deviation. Approximately 50% of the data is on either side of the very center and the mean is identical for both graphs.

Now suppose 50 boxes are randomly selected, each consisting of 30 bottles of cola obtained from the population of all bottles filled over the same short time period. The amount of cola in each bottle is determined and the measurements are saved in the StatCrunch file lab2b.txt in the form of a table of 50 columns, each consisting of 30 rows.

4. Obtain the mean amount of cola for each sample consisting of 6 bottles with the *Summary Stats (Columns)* feature and save the results in a column. Make sure that all 50 columns are included in the right panel of the *Column Statistics* dialog box.

    a. Obtain a relative histogram of the 50 means with the bins starting at 302.4 and using a width of 0.1. Paste the histogram into your report. The format of the histogram should be the same as the format of the histogram in *Lab 1 Instructions* (labels at the axes, title).

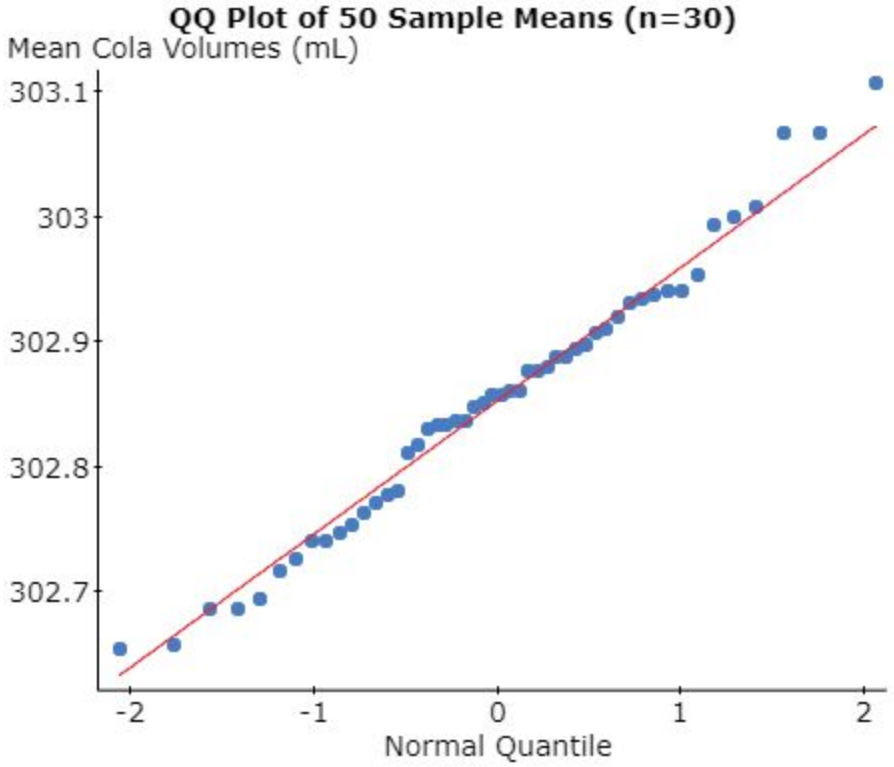


Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

b. Describe the shape of the histogram in part (a). Do the data appear to be approximately normally distributed? Compare the histogram with the histogram obtained in part (a) of Question 2 and the one in part (a) of Question 3. In particular, comment about differences in spread and degree of skewness between the two distributions.

- The data would be approximately normal distributed if there were more bars on the left side of the graph. The graph is slightly right skewed. Compared to the histogram of the (n=6) means from question 3.a, the spread of the data is not as large. It is more confined in the area between 302.8-303 mL. The "bell" is narrower than the histogram in question 3.
- A comparison of this histogram with that of question 2.a shows that 4.a is much less dramatic in terms of both spread and tail. The graph in 2.a is very right skewed, and has a wider range than that of 4.a. It is also not a bell shape, but a steady decreasing curved line, like that of a logarithmic equation.

c. Obtain the Q-Q plot for the 50 means. Add a title to the plot. Paste the plot into your report. Does it look that the sample means come from a normal distribution? Explain. Compare the Q-Q plot with the Q-Q plot obtained in part (c) of Question 3. What do you conclude?



- No, the sample means don't appear to come from a normal distribution. It's evident that the data points are slightly right skewed, which suggests that fewer of the cans of cola are extra overfilled. The points between the first and second positive quartile are far and few between. The others are condensed between negative one and zero. The QQ plot from 3.c has a more even distribution of data points. This suggests that larger data sets are more representative of the variations within a sample. It makes sense that fewer bottles would be super overfilled, as that would cost more money to produce.

Jinny Hwang ID# 1541265
Brianna Grad ID# 1529904

d. Use the *Summary Statistics (Columns)* feature to obtain the mean and standard deviation of the 50 means. Paste the summaries into your report. Compare the value of the standard deviation of the sample mean for n = 30 with the standard deviation of the sample mean in part (d) of Question 3 (for n = 6). Compare the values with the mean and the standard deviation of the sampling distribution of the sample mean predicted by the theory of sampling distributions. Which sample mean tends to be a more accurate estimate of the population mean?

Summary statistics:

Column	Mean	Std. dev.
Mean	302.85247	0.1064492

➛ The means from (n=30) and (n=6) are very close in value. There is only a difference of 0.0252mL. The standard deviations however, have a much bigger difference of 0.1336mL. The standard deviation from (n=6) is larger because there are fewer values used to calculate the sample means for each of the 50 samples. This proves that larger sample sizes give more accurate values for the standard deviation and mean. Theory states that when n increases, the values obtained will approach the theoretical values. This is due to the fact that there are more data gathered, therefore, there will be more precise results.