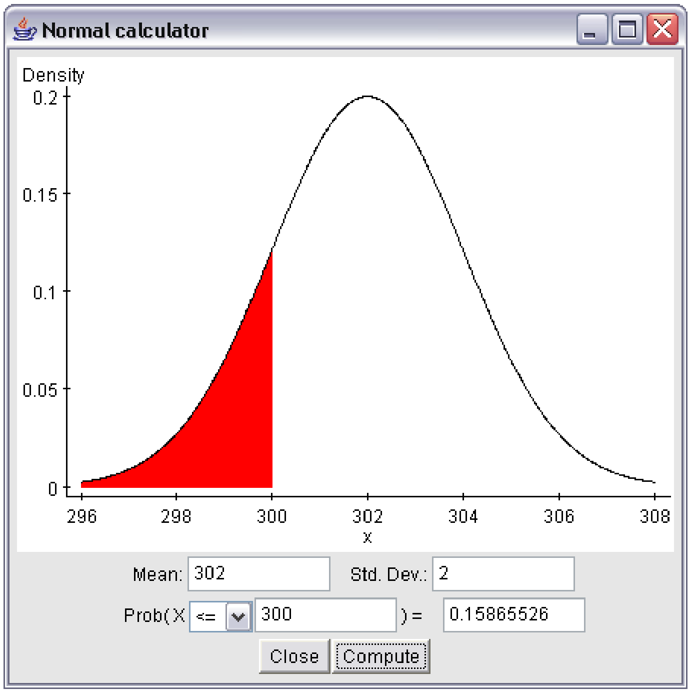
SOLUTIONS TO THE LAB 2 ASSIGNMENT

Question 1

(a) Regardless of the value of σ, the percentage of underfilled bottles (less than 300 ml) and overfilled bottles (more than 300 ml) would be always the same (50%).

The standard deviation controls the spread of a normal curve. Increasing the standard deviation makes the curve lower and flatter. Bottles selected from a distribution with a smaller σ are more likely to have amounts near the mean of 300 ml than bottles selected from a distribution with a larger σ.

(b) In this part, you compare the percentages of underfilled bottles as the value of σ decreases while the target mean amount is kept at the constant level of 302 ml.



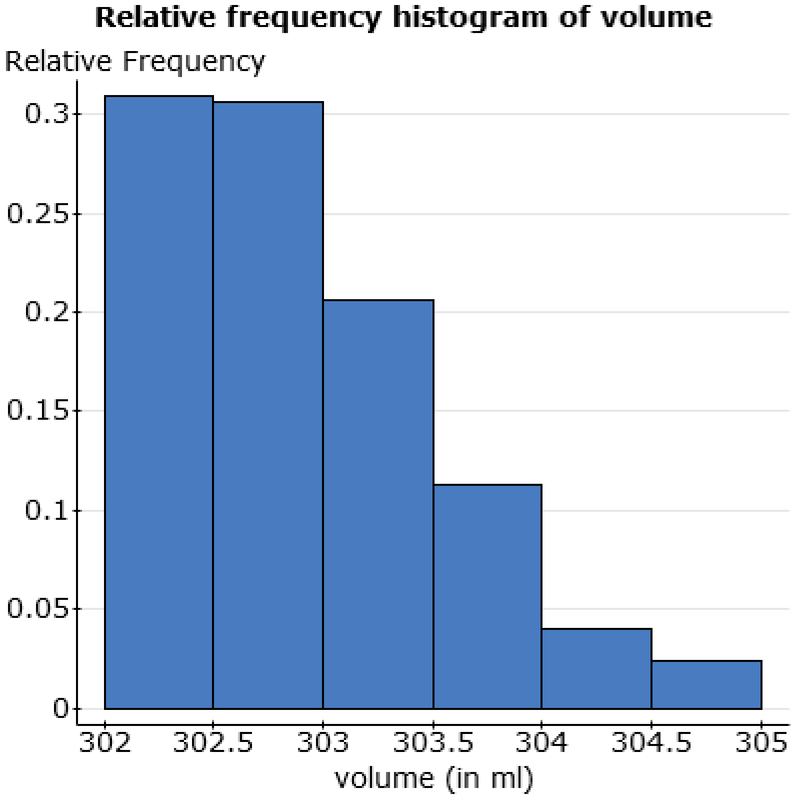
Mean: 302 Std. Dev.: 2

Prob(X <= 300) = 0.15865526

According to the above output, approximately 15.865% of bottles are underfilled when σ = 2 ml. Only 2.275% of bottles are underfilled when σ = 1 ml and almost no bottles are underfilled when σ = 0.5 ml (probability of 0.0000317). In general, as σ decreases, the percentage of underfilled bottles decreases (smaller and smaller areas lie below the curve to the left of 300).

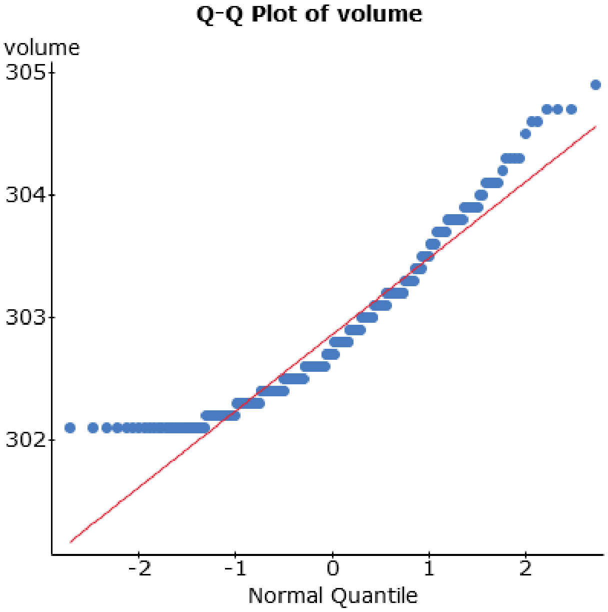
Question 2

(a) The relative frequency histogram with the bins starting at 302 and and using a width of 0.5 is shown below:



(b) The above relative frequency histogram shows that the distribution of volume is extremely skewed to the right. The vast majority of bottles filled by the filling machine have an amount of cola between 302 and 303 ml. The histogram supports the bottler's claim that the bottles are indeed overfilled (by at least 2 ml). The vast majority of the bottles are overfilled by 2-3 ml.

(c) The Q-Q plot for the 300 observations is shown below:



As the points in the plot clearly deviate from the straight-line pattern, there is evidence that the data do not follow a normal distribution. This conclusion is consistent with the conclusions obtained in part (b) for the histogram. Moreover, the shape of the plot (concave upward) is consistent with the right skewness in the data.

(d) The summary statistics are shown below:

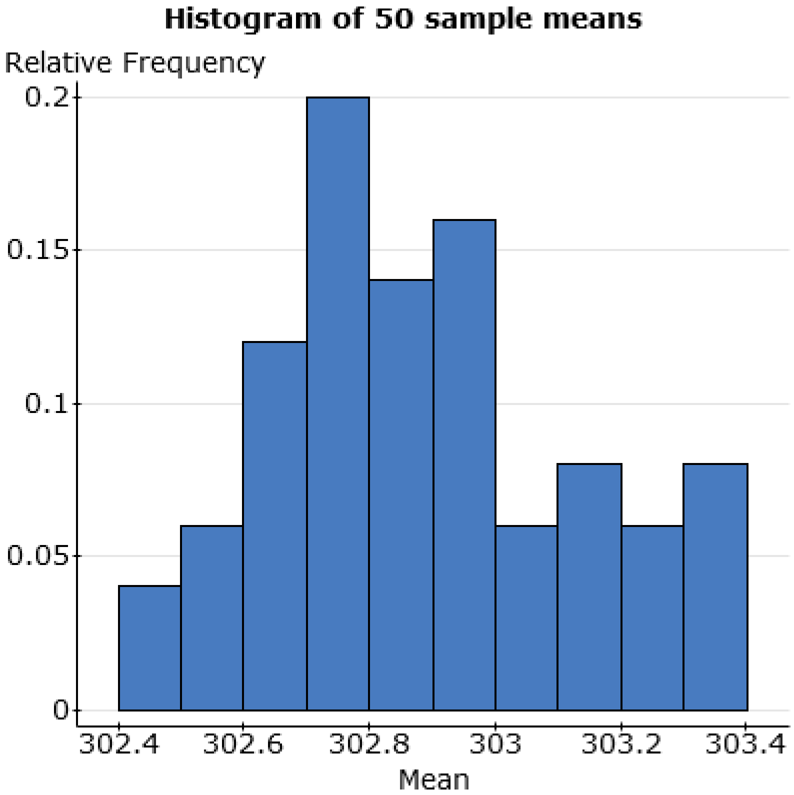
Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
volume	300	302.86566	0.39236242	0.62638843	0.036164552	302.7	2.8	302.1	304.9	302.4	303.2

For skewed distributions, the mean lies toward the direction of skew relative to the median. This is why the mean is larger than the median in this case. Notice that the positions of the three quartiles are also consistent with the right skewness of the volume distribution. Indeed, Q3 lies farther from the median (Q2) than Q1. These conclusions are consistent with the histogram.

Question 3

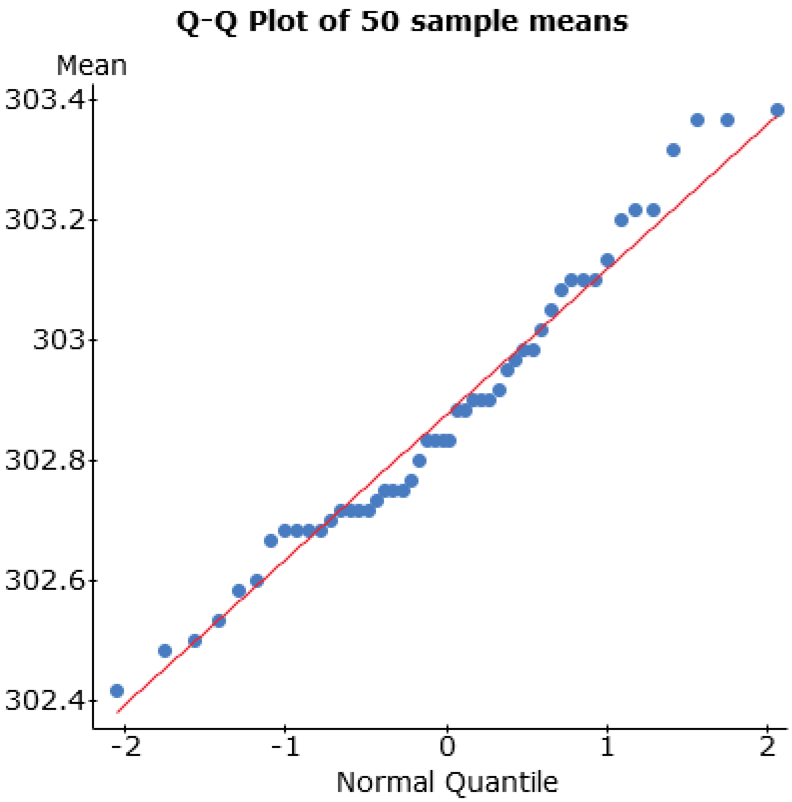
(a) In order to obtain the relative frequency histogram of the 50 sample means, we need to obtain the 50 sample means with the *Summary Statistics* (*Columns*) tool in StatCrunch. In this case, it is necessary to remove all other summaries listed in the right panel in the *Column Statistics* dialog box by clicking on each summary that is to be removed in the left panel. Also the option "*Store output in data table*" should be checked. The relative frequency histogram of the 50 sample means (n = 6) is shown below:



(b) From the histogram, the data do not appear to be normally distributed. More precisely, the distribution of sample means for n = 6 is right-skewed. The right-skewness is inherited from the parent distribution. It looks that the sample size n = 6 is not large enough to make the sampling distribution of the sample mean approximately normal due to the Central Limit Theorem. Nevertheless, the above distribution does not exhibit the same level of right-skewness as observed in the parent distribution in Question 2. The skewness should be smaller if the sample size were larger.

The spread is much smaller than the one for the parent distribution (the range is about 1 compared to the range of 3 for the parent distribution).

(c) The Q-Q plot for the 50 sample means (n = 6) is shown below:



There are some deviations in the plot from the straight-line pattern, but those deviations are not as extreme as those in part (c) of Question 2. The distribution of the 50 sample means for the sample size n = 6 does not strictly follow a normal distribution, but it is closer to a normal pattern than the distribution discussed in part (c) of Question 2. Similarly, this confirms the pattern of the histogram in part (b).

(d) The mean and standard deviation for the mean amount of cola in each of 50 boxes are shown below:

Summary statistics:

Column	n	Mean	Std. dev.
Mean	50	302.87767	0.24005408

Based on part (c) of Question 2, the population mean μ and standard deviation σ are approximately

$$\mu \simeq 302.867, \sigma \simeq 0.626$$

According to the properties of the sampling distribution of a sample mean, if a population has a mean μ and standard deviation σ, then the distribution of the sample mean has a mean and standard deviation defined by the formulas:

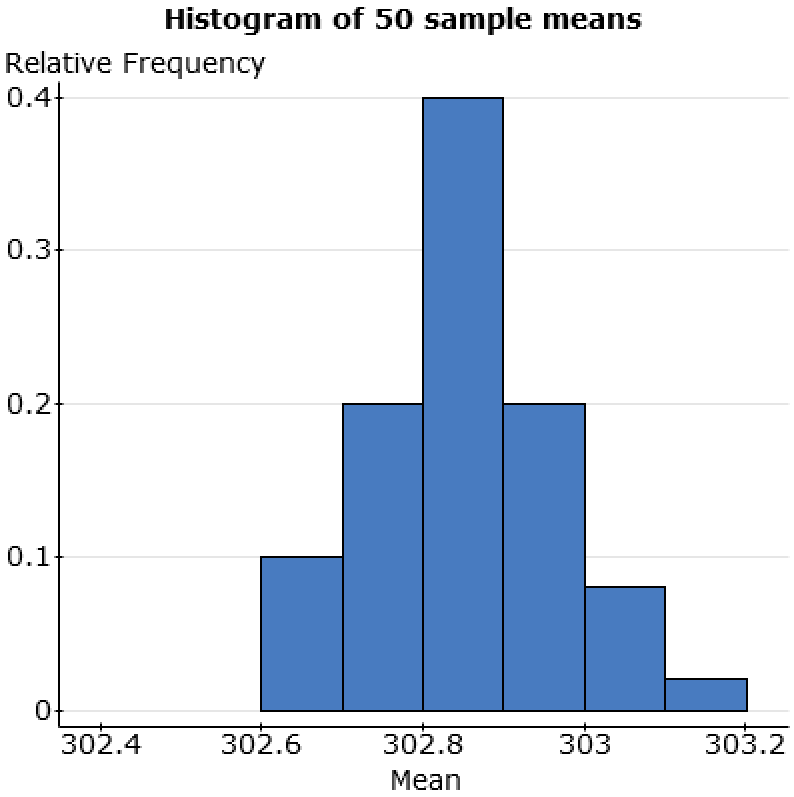
$$\mu_{\bar{Y}} = \mu_Y,\ \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

Notice that the mean of the 50 sample means is equal to 302.878 is very close to the approximate population mean $\mu \approx 302.867$ and the standard deviation of the 50 sample means (each based on a random sample of 6 observations) is equal to 0.240 is very close to the value $\sigma/\sqrt{n} = 0.62638843/\sqrt{6} = 0.256$ predicted by the theory.

In general, the standard deviation here is an estimate of the standard deviation of the sampling distribution of a sample mean. In our case, the sample mean is the mean of 50 sample means and the standard deviation of 0.240 reported in the output measures the variability of the 50 sample means.

Question 4

(a) The relative frequency histogram of the 50 sample means (n = 30) is shown below:

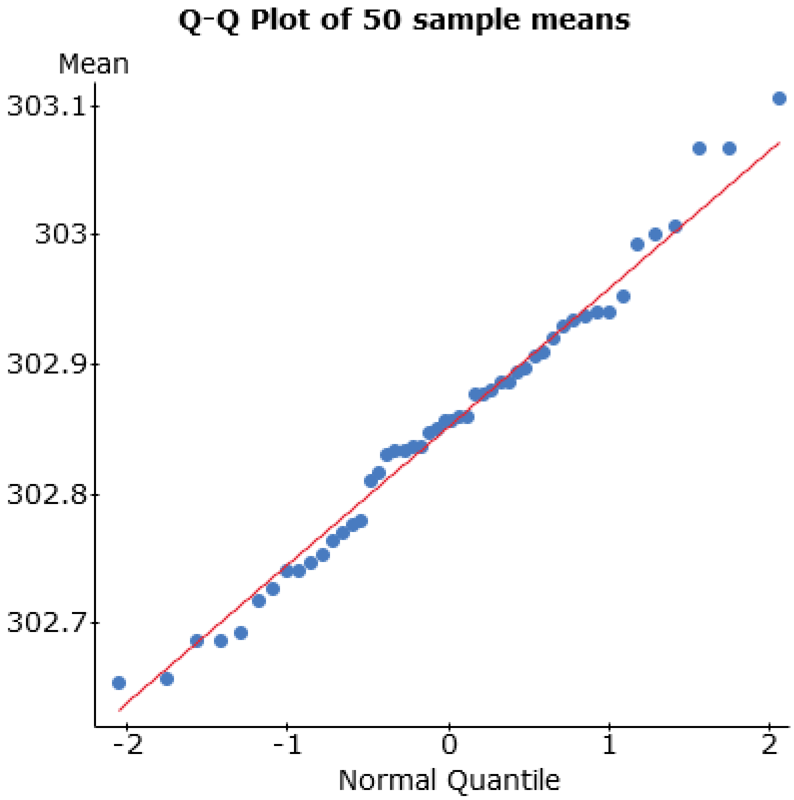


(b) From the histogram, the distribution of sample means is approximately normal (some minor skewness is still exhibited in the right tail).

Notice a considerable difference in the spread of the parent distribution shown in the histogram in Question 2 (a) and the one displayed above. The range of the observations in the parent distribution is about 3 compared to the range of 0.6 for the 50 sample means (n = 30). Moreover, the right-skewness in the histogram of the means for (n = 30) is very minor (compared to a considerable right-skewness in the histogram of the means for n = 6).

Now we compare the histogram in Question 4 with the histogram of sample averages (n = 6) in Question 3. The centers of the two distributions are similar (close to 302.8). However, the spread of sample means for n = 30 is much smaller than the spread of means for n = 6. Also the right-skewness in the histogram of means for n = 6 is more extreme than the minor skewness observed in the histogram in Question 4.

(c) The Q-Q plot for the 50 sample means (*n* = 30) is shown below:



The points in the above Q-Q plot are reasonably close to a straight line. The distribution of the 50 sample means approximately follows a normal distribution. The points in the above Q-Q plot lie closer to a straight line than those in Q-Q plot in part (c) of Question 3. Notice that the smaller range on the vertical axis in the above Q-Q plot may obfuscate a proper comparison with the Q-Q plot in part (c) of Question 3; the distances between the points and the line are here much smaller than the ones in part (c) of Question 3.

(d) The mean and standard deviation for the mean amount of cola in each of 50 boxes are shown below:

Summary statistics:

Column	n	Mean	Std. dev.
Mean	50	302.85247	0.1064492

The standard deviation for *n* = 30 is 0.106, which is smaller than the standard deviation of 0.240 reported in part (d) of Question 3 (for *n* = 6).

Notice that the mean of the 50 sample means is equal to 302.852 is very close to the approximate population mean μ ≈ 302.867 and the standard deviation of the 50 sample means (each based on a random sample of 30 observations) is equal to 0.106 is very close to the value $\sigma/\sqrt{n}$ = $0.62638843/\sqrt{30}$ = 0.114 predicted by the theory.

The estimate of the population mean μ with a smaller standard deviation tends to produce a more accurate estimate of the population mean.

LAB 2 ASSIGNMENT MARKING SCHEMA

Header and Appearance: 10 points

Question 1 (12)

(a) Percentage of underfilled bottles when the standard deviation decreases or increases: 2 points
How the magnitude of the standard deviation affects the filling process: 2 points
(b) Percentage of underfilled bottles when $\mu = 302$ and $\sigma = 2$ ml: 2 points
Percentage of underfilled bottles when $\mu = 302$ and $\sigma = 1$ ml: 2 points
Percentage of underfilled bottles when $\mu = 302$ and $\sigma = 0.5$ ml: 2 points
Effect of decreasing σ on the percentage of underfilled bottles: 2 points

Question 2 (22)

(a) Properly formatted histogram of the 300 observations: 4 points
(b) Shape of the histogram in part (a): 2 points
Conclusion about histogram support of company’s claim: 2 points
(c) Q-Q plot with a title: 4 points
Consistency with the conclusions in part (b): 2 points
(d) Summary statistics output: 2 points
Relationship between mean and median: 2 points
Relationship among the three quartiles: 2 points
Consistency with the conclusions in part (b): 2 points

Question 3 (25)

(a) Properly formatted histogram of the 50 sample means ($n = 6$): 4 points
(b) Shape of the histogram in part (a), normality: 2 points
Comparison with parent distribution (spread, degree of skewness): 4 points (2 points each feature)
(c) Q-Q plot with a title: 4 points
Comparison with conclusions in part (b): 2 points
Comparison with Q-Q plot in Question 2: 2 points
(d) Summary statistics output: 2 points
Comparison with the values predicted by the theory: 3 points
Standard deviation: 2 points

Question 4 (31)

(a) Properly formatted histogram of the 50 sample means ($n = 30$): 4 points
(b) Shape of the histogram in part (a), normality: 2 points
Comparison with graph from Question 2 (spread, skewness): 4 points (2 points for each feature)
Comparison with graph from Question 3 (spread, skewness): 4 points (2 points for each feature)
(c) Q-Q plot with a title: 4 points
Normality: 2 points
Comparison with graph from Question 3 and conclusion: 2 points
(d) Summary statistics output: 2 points
Comparison of the standard deviations: 2 points
Comparison with the values predicted by the theory: 3 points
Sample mean which is more accurate estimate of the population mean: 2 points

TOTAL = 100

Created by: Henryk Kolacz; Edited by: Paul Cartledge
University of Alberta
September 25, 2017