LAB 2 ASSIGNMENT

SAMPLING DISTRIBUTIONS, CENTRAL LIMIT THEOREM

In this lab assignment, you will explore important properties of the sampling distribution of a sample mean in the context of a filling process. In particular, you will use some sampling procedures in StatCrunch to demonstrate the validity of the Central Limit Theorem. You will see that the distribution of the sample mean for samples drawn from a highly skewed distribution becomes approximately normal as the sample size increases. Moreover, you will investigate how the spread of the sampling distribution of the sample mean is affected by sample size.

How Much Cola in the Bottle?

These days, soft drink dispensing soft drink (such as cola) is performed by filling machines. These are set to deliver a certain amount of drink, which we will call the target amount, and the contents of bottles will vary around this mean value. The amount of variation will depend on the efficiency of the machine itself as well as certain properties of the drink, such as its density. The bottler may be able to reduce this variation, but no amount of expertise or effort could lead to its complete removal.

A company uses a filling machine to fill plastic bottles with a popular cola. The bottles are supposed to contain 300 milliliters (ml) of the drink. However, when we buy a bottle of cola which bears a stamp claiming that the amount of the drink is 300 ml, would we expect to obtain exactly 300 ml of cola? We would probably expect some amount close but not exactly equal to 300 ml.

If the amount of drink dispensed by the filling machine follows a symmetric distribution and the mean target value is set equal to the claimed amount of 300 ml, half of the bottles would be underfilled and half would be overfilled. This may seem perfectly reasonable to the bottler but consumers may feel differently, particularly if they happen to buy the underfilled bottles. To make the customer happy, the bottler may decide to overfill the bottles slightly so that the target fill of the machine is more than the claimed amount. However, even a small increase in the target fill represents a loss of many thousands of dollars to the bottler.

The bottles are shipped in boxes containing either 6 or 30 bottles. How does the amount of drink vary from bottle to bottle? How does the average amount of drink vary from box to box containing the same number of bottles? How does the number of bottles in a box affect the distribution of the means? You will obtain the answers to all these questions in this lab.

Answer the following questions:

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 $\mu = 300$ 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?

   (b) Now assume that the mean amount dispensed by the machine is set at $\mu = 302$ 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?

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).

   (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?

   (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?

   (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)?

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).

   (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.

   (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.

   (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?

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 30 observations 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).

   (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.

   (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?

   (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?

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