

JMP INTRO® Lab Activities

Lab Activity Sampling Variability and Law of Large Numbers

Data Set: Pollen

Using JMP INTRO open the data set called **Pollen**. You will notice you have 5 columns of data and 3848 rows of data. You will be writing a report that compares the distributions of the sample mean from sample of size 50 and samples of size 10 then investigate how the distribution is related to the entire population.

Situation:

Cornelius Growbetter has tried a new fertilizer for his corn. He has 3848 stalks of corn in his field. He is investigating if this product will produce better kernels on the corn. He would like to measure the number of kernels on the nub, edge, and crack of the corncob for a random sample of ears of corn. You have access to all the corn stalks through this data set.

Part 1 Using the entire data set.

- After the data set is open click the **blue triangle** in the left corner (*This will display the complete table showing Pollen, Columns and Rows*)
 - **Analyze® Distribution** highlight **edge** and **nub** highlight **Y column** and click **OK**
 - Click the **red triangle left of Distribution® Uniform Scaling** (*Scroll down the table and notice both graphs and Quantile and Moment Table*)
1. Cut and paste a histogram and box plot into your report. Also include the Quantiles and Moments tables. Describe and contrast the distributions (nub and edge) using the histogram and box plot. Looking at this information, how would you describe the distributions of the number of kernels found on the edge of the corn as compared to the nub.

Part 2 Using a subset of the sample.

- Take a sample size of 50 for both the edge and nub data.
 - **Tables® Subset** and list **50** as sample size then **OK**. This will pick a random sample size of 50 from the 3848 and store the resulting data in a new data table.
2. Cut and paste a histogram and boxplot into your report. Include the Quantiles and Moments tables. Describe and contrast the distribution (nub and edge) using the histogram and box plot. Looking at this information, how would you describe the distribution of the number of kernels found on the edge of the corn as compared to the nub.

3. Compare and contrast the distribution for the sample of size 50 with the population distributions in problem 1. Compare and contrast the histogram and boxplot from the population with those from the sample. How close is the sample mean to the population mean?
 - Look at the graphs and tables obtained by two other students. How do their distributions compare to yours? Describe the shape of their distributions and how they compare to yours and the original distributions in your report.
4. Now take a sample size of 10 and compare the shape of your distribution to the distribution from two other students.
 - How is the distribution for the sample of size 10 different from the population distribution?
 - How is the distribution for the sample size of 10 different from the sample size of 50?
5. How is variability related to the sampling size?

Your report should include:

- a histogram of the population distribution for both edge and nub,
- a histogram of the values from your sample of size 50 and 10 distribution for both edge and nub,
- a description of both sample distributions using numerical and graphical information,
- and a comparison of both sample distributions with the corresponding population.
- Be sure to write in the context to the situation.

JMP INTRO® Lab Activities

Teacher Notes

Lab Activity - Sampling Variability and Law of Large Numbers

Time Required: 45 minutes

Objectives:

- In this activity the student will investigate sampling variability and how it relates to the size of the sample.

Materials:

- Sampling Variability and Law of Large Numbers student activity directions
- Pollen data set.

Concept Notes:

- This data set is used because it has over 3000 observations and is in JMP INTRO, which is easily accessible. This is not a good data set to use to investigate the Central Limit Theorem because all of the variables have a distribution that is approximately normal. It is ideal for visualizing sampling variability because of the large number of observations.
- We have described the data in Pollen as the entire population. In a realistic setting, we most often do not have access to the entire population. Explain this to the students because they may not see the relevance of taking a sample when the entire population is available. Cornelius does not have access to the entire population. He does not want to sample the entire cornfield only a random sample.
- When the students are investigating the data and are concerned with negative values explain that often data is measured using an estimated ideal value and then rescaled using zero as the center. All values that fall below this ideal become negative and data that falls above are positive. Such is the case in this data set.
- It is important that this activity is done before investigation the sampling distributions of the means. The sampling distribution of the means will always be approximately normal regardless of what the population looks like providing the sample size is large enough. The sample of individual observations will generally follow the shape of the population, which may not be normal. In this activity we are looking at a random sample of individual observations.
- **Problem #4**
This activity may be done as a group activity or individually. The class may do problem 4 together as a class this would be more the ideal situation for students to see how their individual samples vary. There may be some students that have the same mean for their sample as in the population but by viewing all the samples together they will visually see the variability from sample to sample.
- In the future, after you have discussed the sampling distribution of the mean you may want to look at the distributions of the means.

- **Problem #5**

The students should visualize that when the sample size is 10 the histogram should be choppier and may not necessarily look like a normal distribution. This should reinforce why checking for normality with a small set of data is often difficult.

- **Problem #6**

The student should see that when the sample size is changed you are more likely to see that that appears similar to the population. Expand on this by using the data from the entire class to visualize this variability.

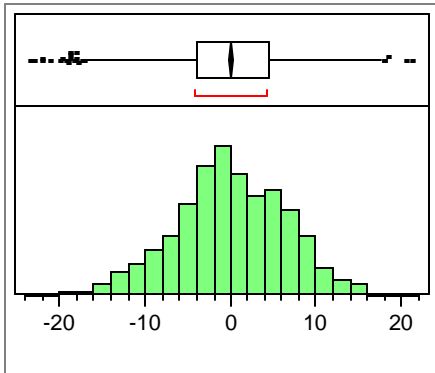
- You can easily connect this activity to the Law of Large Numbers and explain how the sample size is related to the mean of the population.

JMP INTRO® Lab Activities

Suggested answers for Sampling Variability:

edge

Question 1



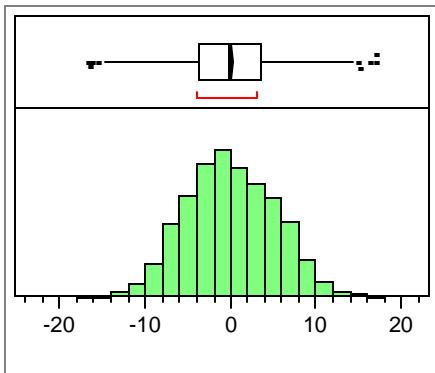
Quantiles

100.0%	maximum	21.41
99.5%		15.19
97.5%		12.15
90.0%		8.17
75.0%	quartile	4.65
50.0%	median	-0.16
25.0%	quartile	-3.99
10.0%		-8.49
2.5%		-13.00
0.5%		-16.76
0.0%	minimum	-23.28

Moments

Mean	-0.003637
Std Dev	6.3982366
Std Err Mean	0.1031437
upper 95% Mean	0.1985849
lower 95% Mean	-0.205858
N	3848

nub



Quantiles

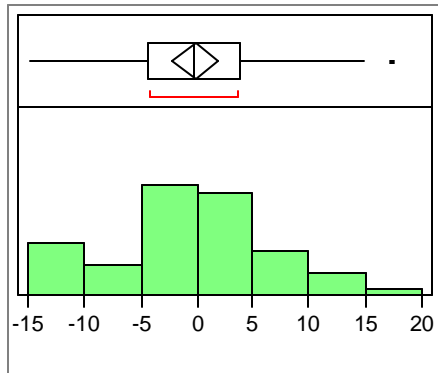
100.0%	maximum	17.26
99.5%		13.38
97.5%		9.93
90.0%		6.90
75.0%	quartile	3.76
50.0%	median	-0.23
25.0%	quartile	-3.76
10.0%		-6.65
2.5%		-9.68
0.5%		-12.73
0.0%	minimum	-16.39

Moments

Mean	0.0001597
Std Dev	5.1863106
Std Err Mean	0.0836067
upper 95% Mean	0.1640773
lower 95% Mean	-0.163758
N	3848

2. Answers will vary for each data set. This is a sample of a data set. The description should include a numerical summary of their data and a general description of the shape.

edge



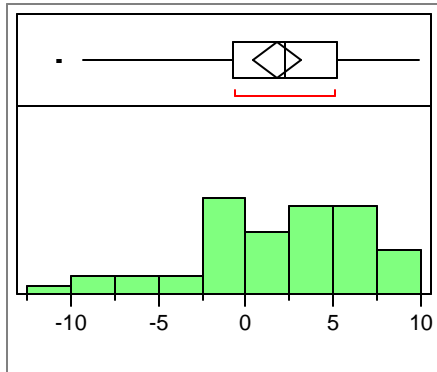
Quantiles

100.0%	maximum	17.49
99.5%		17.49
97.5%		16.78
90.0%		9.85
75.0%	quartile	3.88
50.0%	median	-0.31
25.0%	quartile	-4.39
10.0%		-11.41
2.5%		-14.27
0.5%		-14.77
0.0%	minimum	-14.77

Moments

Mean	-0.19823
Std Dev	7.1879047
Std Err Mean	1.0165232
upper 95% Mean	1.8445499
lower 95% Mean	-2.24101
N	50

nub



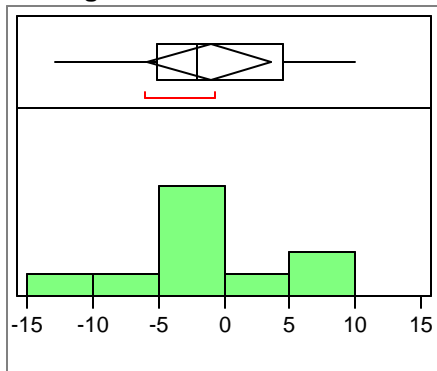
Quantiles

100.0%	maximum	9.89
99.5%		9.89
97.5%		9.72
90.0%		8.10
75.0%	quartile	5.24
50.0%	median	2.33
25.0%	quartile	-0.79
10.0%		-5.35
2.5%		-10.29
0.5%		-10.67
0.0%	minimum	-10.67

Moments

Mean	1.845276
Std Dev	4.8188894
Std Err Mean	0.6814939
upper 95% Mean	3.2147892
lower 95% Mean	0.4757628
N	50

- The description could include that the histogram of 50 is more choppy than the histogram for the entire population. They should also use numerical summary statistics when making their comparison.
- Answers will vary. Students should have data from two other students. It is important that the students note there is variability between their summary statistics and histograms.
- Answers may vary. Below is a sample
Edge- Distribution



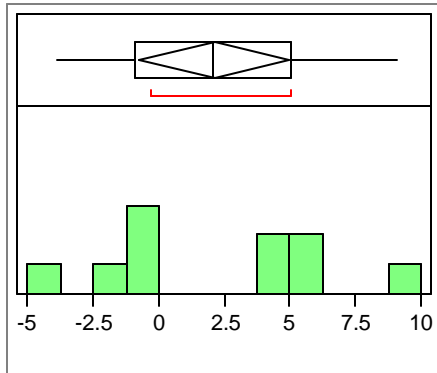
Quantiles

100.0%	maximum	9.96
99.5%		9.96
97.5%		9.96
90.0%		9.68
75.0%	Quartile	4.51
50.0%	Median	-2.06
25.0%	Quartile	-5.08
10.0%		-12.19
2.5%		-12.87
0.5%		-12.87
0.0%	Minimum	-12.87

Moments

Mean	-1.11457
Std Dev	6.6440217
Std Err Mean	2.1010241
upper 95% Mean	3.6382768
lower 95% Mean	-5.867417
N	10

nub



Quantiles

100.0%	maximum	9.072
99.5%		9.072
97.5%		9.072
90.0%		8.671
75.0%	quartile	5.060
50.0%	median	2.069
25.0%	quartile	-0.911
10.0%		-3.647
2.5%		-3.827
0.5%		-3.827
0.0%	minimum	-3.827

Moments

Mean	2.11456
Std Dev	4.0454924
Std Err Mean	1.279297
upper 95% Mean	5.0085309
lower 95% Mean	-0.779411
N	10

6. Answers may vary but the student should address the fact that the sample size of 10 produces a choppy less defined graph than a sample size of 50. The distribution of values for a sample of size 50 is more similar to the population distribution than for samples of size 10.