Exploring Fundamental Definitions with a Study of Fingerprint Types in the US

Investigative Cycle: When conducting any research, you will need to follow an investigative cycle. In general, an investigative cycle includes *asking a question*, *collecting data* to answer that question, *analyzing the data* and then *making conclusions*. This activity will follow this format.

Directions: With your partner(s), discuss and record answers to the questions that follow. Assign someone the responsibility of sharing your answers with the class. Do not reference any notes or technology. You will have 5 minutes to discuss.

a. What does “randomness” mean?

b. How would you take a sample from a population?

Question

Fingerprints are often used by law enforcement to fight crime and apprehend wrong doers. Each person has a unique fingerprint. If we look at the fingerprint of the thumb, there are three common characteristics: loops (lines that bend back on themselves), whorls (such as circles or ellipses) and arches (wave like patterns). In the United States, 60% of the population has loops, 35% whorls, and 5% arches.

Loop	Whorl	Arch
		

Here are some actual prints. Can you correctly identify the shape of each?

___whorl_________

___arches________

____loop___________

In this activity, we will address the following questions:

- *If you were to take several random samples of American adults, would the sample proportion of individuals with <u>arches</u> change from sample to sample? Why is this?*
- *What would be the most typical value of the sample proportions? Why does this make sense?*
- *How do the shape and center of a sampling distribution for a proportion change as the sample size is increased?*

We will use a bin filled with beads to represent the American population and explore these questions.

Data

Collecting a sample of beads:

1. Why might it be reasonable to take a random sample of beads rather than look at all the beads?

   Answers will vary – too many beads to count, not enough time, etc.

In the bin of beads, each bead represents an individual with a certain type of pattern on their thumb: White for loops, Green for whorls, and Black for arches. We are going to collect a sample of 20 beads and 100 beads using the given paddles.

Draw a sample of size n=20 and a sample of size n=100. Complete the table provided below.

	n = 20	**n = 100**
x = Number of "Arches"		
n = Sample Size	**20**	**100**
Sample Proportion of "Arches" from your sample ($\hat{p} = x/n$)		

Now construct a bar graph to display the results of each sample.

Figure 1. *Results for n=20*

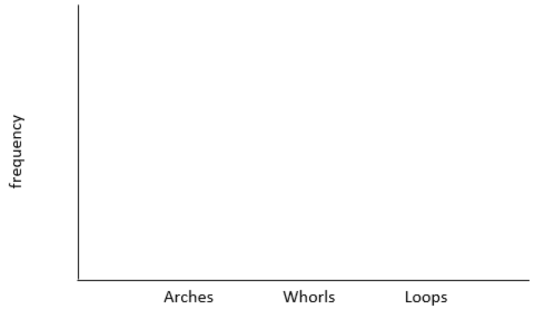
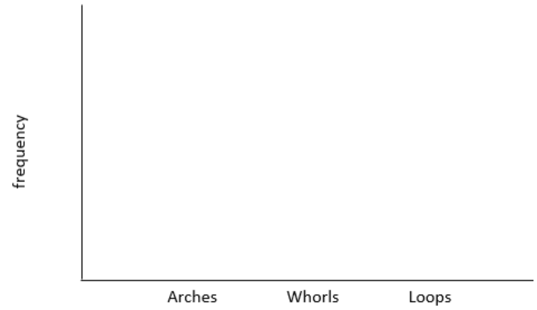


Figure 2. *Results for n=100*



Let's continue to summarize our findings with some descriptive statistics. Complete the third row of the above table.

2. Draw three more samples for n=20 and calculate their associated p-hats. Do you get the same value every time? Why is this?

   No, because of the randomness of the selection process

3. If you were to take several random samples of American adults, would the sample proportion of individuals with arches change from sample to sample? Why is this?

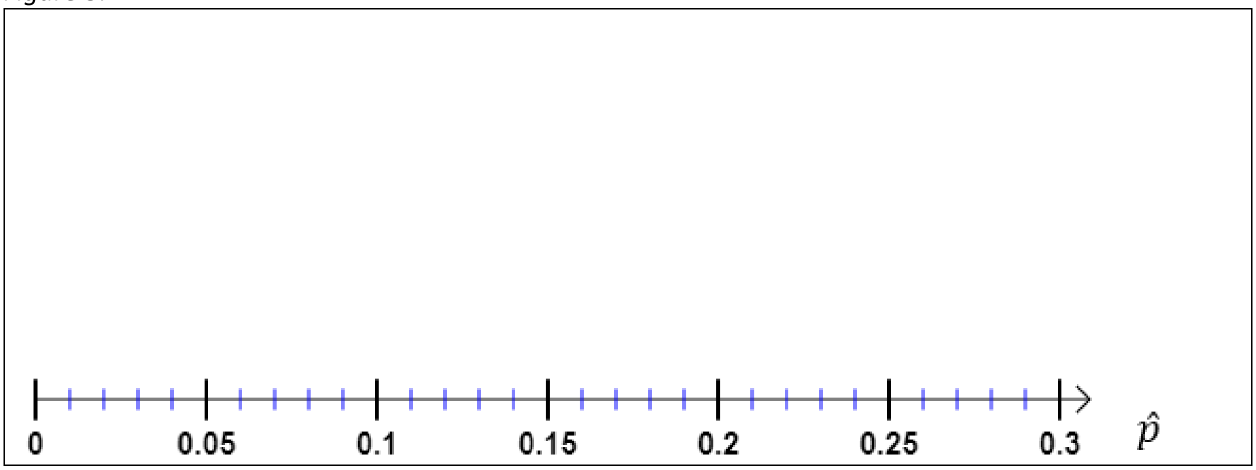
   Yes, because we have a new group of people. Randomness of the selection process makes it possible to have a slightly different proportion each time.

Discuss questions 1-3 as a class.

Analysis
Let's look at the results of everyone in the class. Your instructor will put two axes on the board. Make a dot for each proportion of arches from your samples of size n = 20 and n = 100. Recreate the plots below.

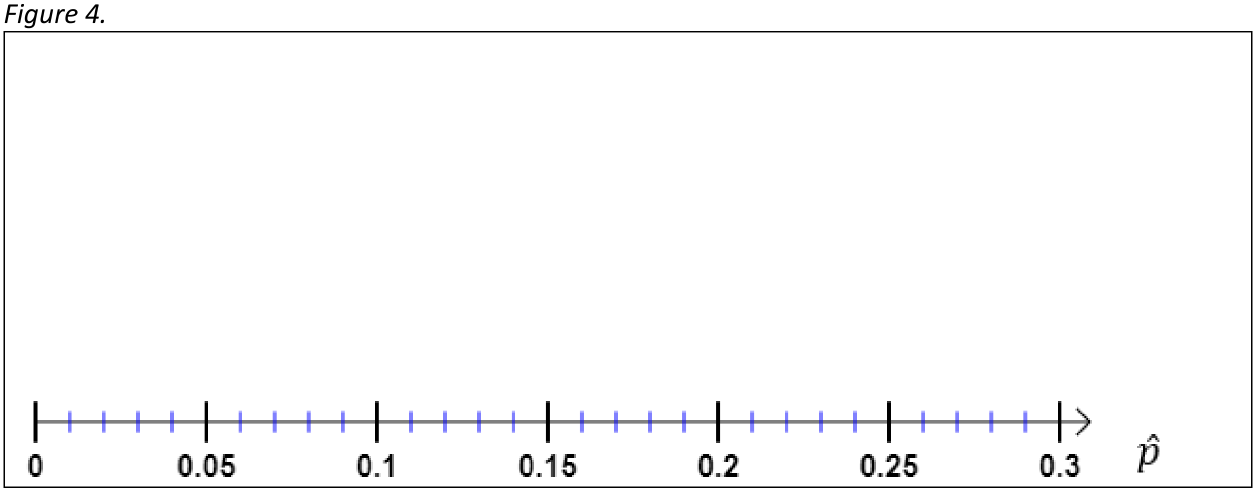
n = 20

Figure 3.



n = 100

Figure 4.



Let's now compare the features of the graphs in Figures 1 & 3.

For Figure 1:

4. What does the x-axis represent?

   The variable X = fingerprint type

5. What does the y-axis represent?

The number of people with each type of fingerprint (frequency)

6. Does this graph show the result of one sample or results from many samples?

One sample

For Figure 3:

7. What does the x-axis represent?

The values of the sample proportions (p-hat's) for the samples

8. Does this graph show the result of one sample or results from many samples?

Many samples

9. What does each dot on the graph represent?

Each dot represents the sample proportion for <u>one</u> sample.

<u>Discuss questions 4-9 as a class.</u>

10. What values of $\hat{p}$ should you reasonably expect when sampling? [Consider Figures 3 & 4.]

For n=20 at a time: [0, 0.10]

For n=100 at a time: [0.03, 0.07] – closer to the actual value of the parameter

11. Why do you think the sample proportions we calculated for each iteration (draw) were rarely (if ever) above 0.20?

The sample proportions were rarely above 0.20 because the value of the population parameter is 0.05.

Discuss questions 10-11 as a class.

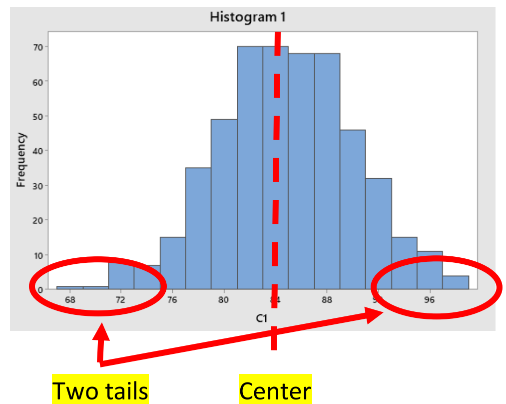
Let’s establish some definitions:

Distribution –

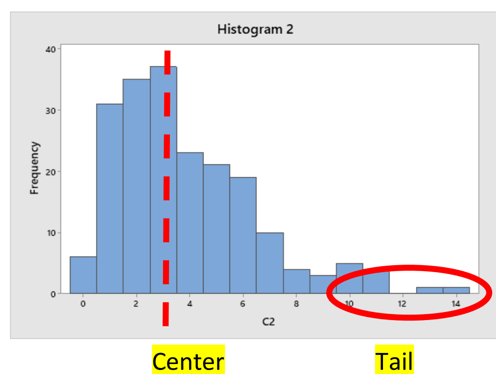
A way to organize the observed values in a sample or samples, along with the frequency of each value

Certain distributions tend to show up often. Here are a few:

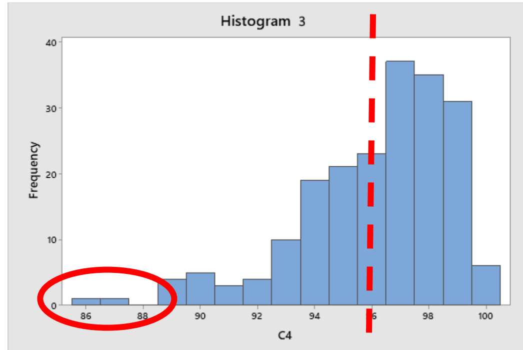
Bell-shaped
Symmetric



Skewed-right (positively skewed)
Not symmetric
One tail

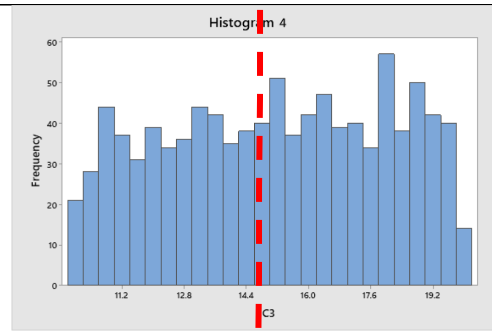


Skewed-left (negatively skewed)
Not symmetric
One tail



Tail Center

Uniform
Symmetric
No tails



Center

Let's define the following properties and identify them on each of the above graphs:
Center (graphically) –

Value on the x axis at which half of the data is to the left and the other half is to the right, graphical estimation of the median

Symmetry –

A graph in which there is a point on the x axis at which the left- and right sides are mirror images of each other)

Tail(s) –

Where a severe drop-off happens in the frequency of the data, will be on the far left and/or far right side of the graph

Sampling distribution of a statistic –

Distribution of a statistic from repeated sampling – the two graphs we created in figures 3 & 4 display sampling distributions for p-hat

Important Note: The above properties are not used to describe graphs that display distributions of categorical or qualitative data (x-axis shows categories), as shown in the bar graphs we created in Figures 1 & 2.

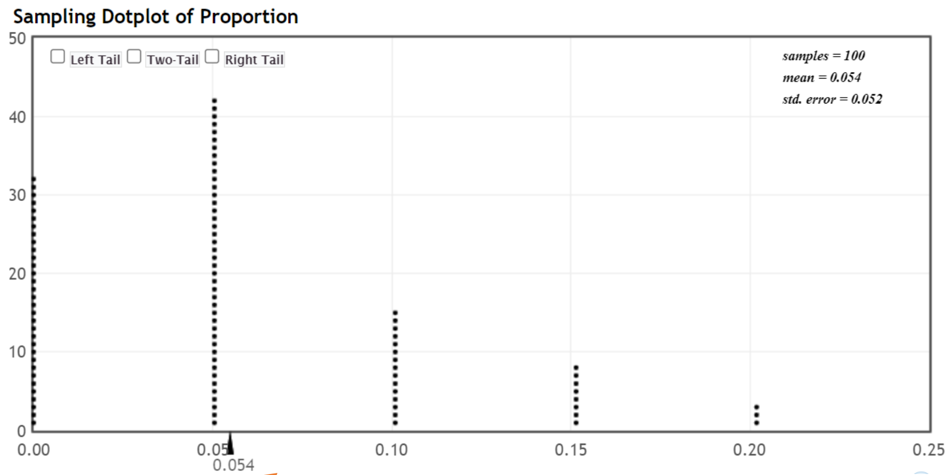
These properties are only used to describe graphs that display distributions of quantitative data (x-axis is a number line). Examples of acceptable graphs include the four histograms provided above and the dotplots we created in Figures 3 through 6 to display our sampling distributions.

Below are the results from using *StatKey* to run 100 samples for the two sample sizes, n=20 and n=100.

Comparing these graphs to the ones we constructed for our earlier samples (Figures 3 & 4), we see that as the number of samples is increased, the shape becomes more clear in each sampling distribution.

Figure 5.

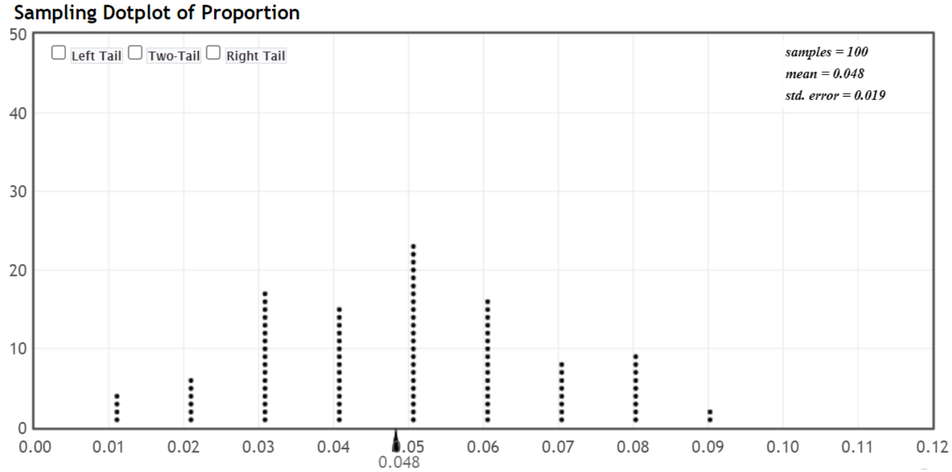
n = 20:



NOTE: The 0.054 shown below the x-axis is the mean of the 100 p-hats in this sampling distribution.

Figure 6.

n = 100:



NOTE: When comparing Figures 5 & 6 to our sampling distributions in Figures 3 & 4, we see that as the number of runs is increased, the shape becomes more pronounced in each sampling distribution.

12. Describe the shape, location of center, and symmetry (if any) of the sampling distribution for n=20 (Figure 5).

    The sampling distribution for n=20 is skewed right with a center at 0.05 and is not symmetric.

13. Describe the shape, location of center, and symmetry (if any) of the sampling distribution for n=100 (Figure 6).

    The sampling distribution for n=100 is bell-shaped and symmetric, with a center at 0.05.

Discuss questions 12-13 as a class.

Conclusion

Let's return to the questions posed at the start of this activity.

14. If you were to take several random samples of American adults, would the sample proportion of individuals with arches change from sample to sample? Why is this?

    Yes, the same people will not be selected from sample to sample due to the random nature of the sampling method; therefore, the sample proportion will likely be different each time.

15. What would be the most typical value of the sample proportions? Why does this make sense?

If the sample is representative of the population, the 0.05 should be most typical, because 0.05 is the value of the population parameter.

16. What property was the same in the n = 20 and n = 100 sampling distributions?

The center was roughly 0.05 for both sampling distributions. Since this is a simulation and we don't have every possible sample, the center for the simulated sampling distribution may be slightly off from the theoretical mean of 0.05.

17. What properties changed in the n = 20 and n = 100 sampling distributions?

The shape and associated symmetry changed.

Discuss questions 14-17 as a class.

18. Describe each of these terms **in the context of this investigation**.

a. *random* sampling

the use of a chance process to ensure each person had an equal chance of being selected

b. sample

n=20 (or n=100) people in the United States

c. *statistic* (give a written description and the associated symbol)

proportion of people from the sample whose fingerprint shapes are arches: $\hat{p}$ "p-hat"

d. population

All people in the United States

e. *parameter* (give a written description, the associated symbol, and the value)

proportion of people from the population whose fingerprint shapes are arches
p = 0.05