

Stat 218 - Day 8
Sampling distribution of sample proportion

Recall the important terms **population** and **sample**.

- A numerical characteristic of a population is a **parameter**; its value is fixed but typically not known.
- A numerical characteristic of a sample is a **statistic**; it can vary from sample to sample and is often used to estimate the value of an unknown parameter.

Practice: Identify each of the following as a parameter or a statistic (you may need to decide for yourself what the population of interest in a given situation is):

- (a) the 56% of callers who said that Elvis is still alive
- (b) the 63% of voters who voted for President Roosevelt in 1936
- (c) the 57% of *Literary Digest* respondents who said that they would vote for Landon in 1936
- (d) the average number of letters in the ten words that you circled in the Gettysburg Address
- (e) the 4.295 average number of letters per word in the Gettysburg Address
- (f) the average number of points scored in a Super Bowl game
- (g) the proportion of voters who voted for President Bush in the 2004 election
- (h) the proportion of voters surveyed by CNN who voted for John Kerry in 2004
- (i) the proportion of people at the next party you attend who voted for Ralph Nader in the 2004 election

Example: Candy Colors

(a) Take a random sample of 25 candies and record the number and proportion of each color:

	orange	yellow	brown
number			
proportion			

- (b) Is the candy's color a quantitative or a categorical variable?
- (c) Is the proportion of orange candies among the 25 that you selected a parameter or a statistic?
- (d) Is the proportion of orange candies manufactured by Hershey a parameter or a statistic?
- (e) Do you *know* the value of the proportion of orange candies manufactured by Hershey?

- (f) Do you know the value of the proportion of orange candies among the 25 that you selected?
- (g) Did every student obtain the same proportion of orange candies in his/her sample?
- (h) If every student was to estimate the population proportion of orange candies by the proportion of orange candies in his/her sample, would everyone arrive at the same estimate?

The values of statistics vary from sample to sample. This phenomenon is called **sampling variability**. Fortunately, if we look at the results of many samples, there is a predictable pattern to this variability.

- (i) Add your sample proportion of orange candies to the graph on the board. Around what value (roughly) are the sample proportions centered?

Since random sampling is unbiased, the actual value of the population proportion should be close to the center of these sample proportions.

- (j) If every student was to estimate the population proportion of orange candies by the proportion of orange candies in his/her sample, would most estimates be reasonably close to the true parameter value? Would some estimates be way off? Explain.

We need to take more samples to see the pattern of how sample statistics vary more clearly. For this we can turn to an applet called “Reese’s Pieces.” For now we will suppose that 45% of the population is orange.

- (k) Use the applet to draw 500 samples of 25 candies each, assuming that the population proportion of orange is .45. (Pretend that this is really 500 students, each taking 25 candies and counting the number of orange ones.) Sketch and describe a graph of the sample proportions of orange obtained.

- (l) Is there an obvious pattern to the distribution of the sample proportions of orange candies? Is it approximately normal?

Even though the sample proportion of orange candies varies from sample to sample, there is a recognizable long-term pattern to that variation. This pattern is called the **sampling distribution** of the statistic.

(m) What are the mean and standard deviation of the sample proportions of orange candies?

(n) Now assume that the population proportion of orange candies is .55. Again use the computer to draw 500 samples of 25 candies each. How has the distribution changed?

- shape:
- center:
- spread:

(o) Now use the computer to draw 500 samples of 100 candies each (so these samples are four times larger than the ones you gathered in class). How has the distribution of sample proportions changed (or not changed) from when the sample size was only 25 candies?

- shape:
- center:
- spread:

Key Result 1: Suppose that the proportion of a population having some characteristic is denoted by p , and suppose that a random sample of size n is taken from the population. Then the sampling distribution of the sample proportion \hat{p} is approximately normal with mean p and

standard deviation $\sqrt{\frac{p(1-p)}{n}}$. This approximation is generally considered to be valid as long as $np \geq 10$ and $n(1-p) \geq 10$.

(p) Use this result to calculate the standard deviation of \hat{p} for the $n=25$ and $n=100$ cases, still assuming that 45% of all Reese's Pieces are orange. Are the calculated values close to what the simulation results revealed?

(q) Continue to suppose that 45% of all Reese's Pieces are orange. Use Key Result 1 to determine the (approximate) probability that more than half of the candies are orange in a sample of 75 candies.

(r) Use the applet to simulated 500 samples of size 75 candies each. What percentage of these samples produced more than half orange candies? Is this close to your answer to (q)?

(s) How would you expect your answer to (q) to change, if the sample size were 500 rather than 75? Provide an intuitive explanation.

(t) Calculate the probability in (s) to see if your intuition is correct.

Example: Suppose that you flip a fair coin 200 times. Determine the (approximate) probability that between 45% and 55% of the tosses would land heads.