MOD 15 DISTRIBUTION OF SAMPLE PROPORTIONS (EXPLORING WITH SKITTLES)

Learning Goals

- Distinguish between a sample statistic and a population parameter
- Describe the sampling distribution for sample proportions and use it to identify unusual (and more common) sample results
- Use a z-score and the standard normal model to estimate probabilities of specified events

Activity 1

For this activity we define the population to be all Skittles in the world at this moment.

We need to summarize information about Skittle colors. However, since *color* is a categorical variable (as opposed to a quantitative variable), we cannot use the mean to summarize information about Skittle colors (**we can only use mean with a quantitative variable**). After all, how could we calculate the mean color or even worse the mean of a particular color such as orange? Instead, **we use proportions to summarize information about categorical variables**. We'll focus on orange skittles and try to determine the proportion of orange Skittles in the population of all Skittles.

Goal: using your bag of Skittles as a sample, predict the proportion of orange Skittles in the population of all Skittles.

Let *X* represent the color of a randomly selected Skittle.

1) Describe the population and sample(s), if any, for this activity.

2) Does *X* represent categorical or quantitative data? So, which sample statistic should we use, proportions, $\hat{p}$, or means, $\bar{x}$, to summarize the sample data (i.e. the data for each bag of skittles)?

Sidetrack: We use lower case letters to represent **sample statistics** and upper case letters to represent **population parameters**. For example the sample size is represented with a lower case ***n*** and the population size with an upper case ***N***, and the sample proportion is represented with lower case $\hat{p}$ (pronuounced p-hat) and the population proportion with an upper case ***P***.

3) Would you expect your sample proportion of orange Skittles $\hat{p}$ to be the same as the population proportion of orange Skittles P? Why or why not?

4) Would you expect all of the ***sample proportions*** of orange Skittles to be approximately the same? Why or why not? And more importantly, what is this question asking about anyway? In other words, what do we mean by "*the sample proportions of orange skittles*", and why is "*proportions*" plural in this question?

5) Open your bag of Skittles (do NOT eat any of them yet), dump them out on your paper plate, count them (and have multiple group members verify your count), and then wait patiently. To make sure each student has the same number of Skittles, the instructor will randomly remove Skittles from students with more than 60 and randomly distribute additional Skittles to students with fewer than 60. Once every student has the same number of Skittles, record each of the following for your sample.

n =

# red =	$\hat{p}_r =$	(proportion of red Skittles)
# green =	$\hat{p}_g =$	(proportion of green Skittles)
# purple=	$\hat{p}_p =$	(proportion of purple Skittles)
# orange=	$\hat{p}_o =$	(proportion of orange Skittles)
# yellow =	$\hat{p}_y =$	(proportion of yellow Skittles)

When your group has finished calculating these sample proportions, send a group member to the instructor station so your instructor can enter your data into Tinkerrplots. As your instructor enters the data from each group, please quietly work though number 6) on the next page but do not continue on to number 7). Note: You are not expected to know how to use Tinkerplots as the instructor is merely using it to display and summarize the Skittles data.

6) When we were working with quantitative data such height, we used standard deviation to measure variability and set up a normal density curve (3 standard deviations below the mean and three standard deviations above the mean). However, since "color of a randomly selected Skittle" is a categorical variable, we are not working with a mean, so we cannot calculate standard deviation to measure variability. Categorical variables require us to use proportions instead of means, so we must use standard error (SE) instead of standard deviation to measure spread. Here is the standard error (SE) formula.

$$SE = \sqrt{\frac{P(1-P)}{n}}$$

a) Identify what each letter represents in the standard error formula.

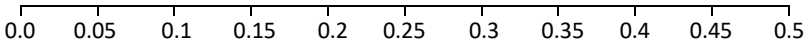
P:

n:

b) Since it is often impossible to KNOW the population proportion, P, we estimate the standard error using a sample proportion. Use your sample proportion for orange Skittles to estimate the standard error (i.e. the standard deviation) for the sampling distribution of orange Skittles (don't worry about this terminology right now, we'll chat about it in a few minutes).

$SE \approx$

7) Your instructor has displayed a histogram for the ***sampling distribution*** (i.e. the distribution of the sample proportions of orange Skittles) from this class along with the sample proportions from many other classes. Sketch the histogram for this distribution. Indicate your sample proportion $\hat{p}$ of orange Skittles on the horizontal axis. Also, mark the mean of the sample proportions, $\hat{p}$ on the horizontal axis.



a) If the histogram were converted to a dot-plot, what would each dot represent?

b) Looking at the histogram for the ***sampling distribution*** (i.e. the ***sample proportions of orange Skittles***), what approximate shape does the data have?

c) As it turns out, the Skittles folks promise that they produce an equal number of each color. So what is the population proportion P? What do you notice about the population proportion P and the mean of all the $\hat{p}$?

$P =$ Mean of the sample proportions rounded to two decimal places $=$

d) Do you think the mean of the sample proportions would approximately equal the population proportion if we only used the data from five randomly selected bags of 60 Skittles to calculate the mean of all the sample proportions (i.e. the mean of all the $\hat{p}$)? Explain.

e) What do you think the mean of the sample proportions, $\hat{p}$, would be if we could use the data from all possible samples of 60 Skittles?

8) Let's summarize what we know about the histogram for the sampling distribution of orange Skittles. Complete each of the following statements.

Center: The mean of the sample proportions is _____.

Spread: The standard deviation of the sample proportions is

_____ (formula). It's also called the _____.

Shape: The shape of the sampling distribution is approximately _____.

Determining if a sampling distribution is approximately normal

9) Typically we have only one sample or just a few samples as opposed to more than 100 samples (bags of Skittles) as we have in our *sampling distribution*. With only one or just a few samples, we cannot create a histogram for the sampling distribution to ***see*** if the shape is approximately normal. As it turns out a normal model is a good fit for a sampling distribution if the expected number of successes is at least 10, AND the expected number of failures is at least 10. Let's explore what we mean by this.

   a) With regard to randomly selecting a Skittle and testing to see if the Skittle is orange, what do we mean by "success"? What do we mean by "failure"?

   b) Since we know the population proportion of orange Skittles is $P = 0.2$, and since there are 60 Skittles in your sample, what is the ***expected number of successes***? What is the ***expected number of failures***? Show how you calculated these values.

      Expected number of orange Skittles (successes) =

      Expected number of NOT orange Skittles (failures) =

   c) So if the expected number of successes is at least 10 **AND** the expected number of failures is at least 10, we can claim that, "For all possible samples of a particular size n from the population, the sampling distribution is normal" – even if we actually only have one sample. Write formulas for these two tests.

      Expected number of successes is at least 10:

      Expected number of failures is at least 10:

Determining if a sample proportion is usual or unusual

10) Look at the graph you sketched in number 7) above and locate your sample proportion on the horizontal axis.

    a) Do you think your sample proportion is usual or unusual? Explain your reasoning.

    b) Based on your answer to part a), if we did not know the population proportion $P = 0.2$, would it be appropriate to infer the population proportion P from your sample proportion $\hat{p}$. Explain.

11) Recall that we say a data value is usual if it is within two standard deviations of the mean, and we say that a data value is unusual if it lies in one of the tails beyond two standard deviations from the mean.

    Here are two z-score formulas.

$z = \frac{x - \mu}{\sigma}$	$z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}}$
Define each of the following from this formula. z : x : μ : σ : When do we use this z-score formula?	Define each of the following from this formula. z : $\hat{p}$: P : $\sqrt{\frac{P(1-P)}{n}}$: When do we use this z-score formula?

12) Use the steps below to provide convincing evidence that your sample proportion, $\hat{p}$, of orange Skittles is usual or unusual.

Step 1: Determine whether we can assume the sampling distribution of all possible samples of 60 Skittles is approximately normal. Hint: see 9c) above.

Step 2: If the sampling distribution of all possible samples of 60 Skittles is approximately normal, calculate the z-score for your sample proportion, $\hat{p}$, of orange Skittles.

Step 3: State whether your sample proportion is usual or unusual and state why based on the z-score.

Activity 2

13) According to the March 7, 2011, article, *Annual Sleep in America Poll Exploring Connections with Communications Technology Use and Sleep*, nearly 60% of adults polled in the United States indicate they have a sleep problem every night or almost every night. Suppose that we select random samples from this population of US adults.

a) Further suppose we then calculate the proportion of the sample that has a sleep problem every night or almost every night. What sample size would we need for the sampling distribution to be approximately normal? Show your work.

b) If we repeatedly obtain random samples of 30 adults, what will be the mean and standard deviation of the sampling distribution of sample proportions? Show your work.

c) If we randomly select a sample of 30 US adults, what is the probability that between 50% and 70% will have a sleep problem every night or almost every night? Show your work.

d) If we randomly select a sample of 30 US adults, what is the probability that at most 65% will have a sleep problem every night or almost every night?