GSE Algebra II/Advanced Algebra
Unit 7: Inferences and Conclusions from Data
Colors of Skittles Learning Task (Task 9)

Name: ____

Date: ______________________ Period: ____________

STANDARDS ADDRESSED IN THIS TASK:
MGSE9-12.S.IC. 1 Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
MGSE9-12.S.IC. 2 Decide if a specified model is consistent with results from a given data-generating process, e.g. using simulation.
MGSE9-12.S.IC. 4 Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling

Part 1:

1. Reviewing some basics.

a) Think about a single bag of Skittles. Does this single bag represent a ***sample*** of Skittles candies or the ***population*** of Skittles candies?

b) We use the term ***statistic*** to refer to measures based on samples and the term ***parameter*** to refer to measures of the entire population. If there are 50 Skittles in your bag, is 50 a statistic or a parameter? If Mars claims that 20% of all Skittles are yellow, is 20% a statistic or a parameter?

c) What is **sampling variability**?

2. How many orange candies should I expect in a bag of Skittles?

a) From your bag of Skittles, take a random sample of 10 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (p)					

b) Do you think that every student in the class obtained the same proportion of orange candies in his or her sample? Why or why not?

c) Combine your results with the rest of the class and produce a *dot plot* for the distribution of sample proportions of **orange** candies (out of a sample of **10 candies**) obtained by the class members.

Make sure you label the axes of your dot plot correctly.

d) What is the average of the sample proportions obtained by your class?

e) Put the Skittles back in the bag and take a random sample of 25 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (p)					

f) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **25 candies**) obtained by the class members. Is there more or less **variability** than when you sampled 10 candies? Is this what you expected? Explain.

g) What is the average of the sample proportions (from the **samples of 25**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the value you found in part *d*? Explain.

h) This time, take a random sample of 40 candies. Record the count and proportion of each color in your sample.

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (*p*)					

i) Combine your results with the rest of the class and produce a dot plot for the distribution of sample proportions of **orange** candies (out of a sample of **40 candies**) obtained by the class members. Is there more or less **variability** than the previous two samples? Is this what you expected? Explain.

j) What is the average of the sample proportions (from the **samples of 40**) obtained by your class? Do you think this is closer or farther from the true proportion of **oranges** than the values you found in parts *d* and g? Explain.

Names of group members:____

COLORS OF SKITTLES CANDIES LEARNING TASK – PART 1 ANSWER SHEET

1a. ________________________

1b. 50: ________________________ 20%: ________________________

1c. ____

2a. Random Sample of 10 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (*p*)					

2b. ____

2c.

2d. ________________________

2e. Random Sample of 25 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (*p*)					

2f.

Variability? ________________________ What you expected? ___________________

Explain: ____

____

2g. Class average: ________________________ True proportion? _________________

Explain:____

____

2h. Random Sample of 40 Skittles

	Orange	**Yellow**	**Red**	**Green**	**Purple**
Count					
Proportion (*p*)					

2i.

Variability? ________________________ What you expected? ________________________

Explain:____

____

2j. Class average: ________________________ True proportion? _________________

Explain:____

____

Part 2

We have been looking a number of different ***sampling distributions*** of p (the distribution of the statistic for *all possible samples* of a given size), but we have seen that there is **variability** in the distributions.

We would like to know that p is a good estimate for the true proportion of orange Skittles. However, there are guidelines for when we can use the statistic to estimate the parameter (in other words using sample data to make assumption about the entire population).

First, however, we need to understand the center, shape, and spread of the sampling distribution of p .

We know that if we are counting the number of Skittles that are orange and comparing with those that are not orange, then the counts of orange follow a binomial distribution (given that the population is much larger than our sample size). There are two outcomes of Skittles, either orange or not-orange and each sample selected of Skittles was independent from the others.

Conditions for Binomial Distribution are:

- The experiment consists of *n* repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success, probability *p* (ranging from 0 to 1), and the other, a failure, probability, (1 – *p*).
- The *n* trials are independent; that is, the outcome on one trial does not affect the outcome on other trials. (Think of replacement with your skittles.)

The formulas for the mean and standard deviation of a **binomial distribution:**

$$\mu_X = np \qquad \sigma_X = \sqrt{np(1-p)}$$

where *n* = numbers of trials and *p* = proportion

a) Given that $p = \frac{x}{n}$, where *x* is the count of oranges and *n* is the total in the sample, we find μ_p and σ_p by dividing each by *n* also. Find the formulas for each statistic.

b) This leads us to the statement of the characteristics of the sampling distribution of a sample proportion.

The Sampling Distribution of a Sample Proportion:

Choose a simple random sample of size *n* from a large population with population parameter *p* having some characteristic of interest. Let p be the proportion of the sample having that characteristic. Then

- ○ The mean of the sampling distribution $\left(\mu_p\right)$ is ____ _______; and
- ○ The standard deviation of the sampling distribution (σ_p) is ________________.

c) Let's look at the standard deviation a bit more. What happens to the standard deviation as the sample size increases? Try a few examples to verify your conclusion. Then use the formula to explain why your conjecture is true.

Sample Size (n)	Let p = 0.2	Let p = 0.7
1		
5		
10		
25		
50		
100		
1000		

If we wanted to cut the standard deviation in **half**, thus decreasing the **variability** of p , what would we need to do in terms of our sample size? (Hint: multiply the formula for standard deviation by ½).

Caution: We can only use the formula for the standard deviation of p when the population is at least **10** times as large as the sample.

d) For each of the samples taken in Part 1 of the Skittles Task, determine what the population of Skittles must be for us to use the standard deviation formula derived above. Is it safe to assume that the population is at least as large as these amounts? Explain.

Part 3

Simulating the Selection of Orange Skittles

As we have seen, there is variation in the distributions depending on the size of your sample and which sample is chosen. To better investigate the distribution of the sample proportions, we need more samples and we need samples of **larger size**. We will turn to technology to help with this sampling. For this simulation, we need to assume a value for the true proportion of orange candies. **Let's assume $p = 0.20$.**

a) First, let's imagine that there are 100 students in the class and each takes a sample of 50 Skittles. We can simulate this situation with your calculator.

Type *randBin(50,0.20)* in your calculator. The command *randBin* is found in the following way.

- TI-83/84: Math → PROB → 7:randBin(

What number did you get? Compare with a neighbor. What do you think this command does?

How could you obtain the proportion that are orange?

b) Now, we want to generate 100 samples of size 50. This time, input *randBin(50,0.20,100)/50*→L_1. [The → is found by using the STO button on the bottom left of the calculator (TI-83/84).] The latter part (store in L_1) puts all of the outputs into List 1. **(Be patient, this takes a while.)**

Using your Stat Plots, create a histogram or stem-and-leaf plot of the proportions of orange candies. (Plot 1 → On → Type: third selection, top row will give you a histogram. Zoom → 9: ZoomStat → Graph)

Sketch the graph below.

Do you notice a pattern in the distribution of the sample proportions? Explain.

c) Find the sample mean and sample standard deviation of the theoretical sample proportion of orange Skittles with sample size 50 and proportion 0.20. Next, find the mean and standard deviation of the output using 1-Var Stats. How do these compare with the theoretical mean and standard deviation for a sampling distribution of a sample proportion?

Theoretical Sample: Mean: ____________ Standard Deviation: _____________

Your Sample: Mean: ______________ Standard Deviation: ______________

d) Use the TRACE button on the calculator to count how many of the 100 sample proportions are within ± 0.057 of 0.20. Note: 0.057 is close to the standard deviation you found above, so we are going about one standard deviation on each side of the mean. Then repeat for within ± 0.114 and for within ± 0.171. Record the results below:

	Number of the 100 Sample Proportions	Percentage of the 100 Sample Proportions
Within ± 0.057 of 0.20		
Within ± 0.114 of 0.20		
Within ± 0.171 of 0.20		

e) If each of the 100 students from your theoretical sample who sampled Skittles were to estimate the population proportion of orange candies by going a distance of 0.114 on either side of his or her sample proportion, what percentage of the 100 students would capture the actual proportion (0.20) within this interval?

f) Simulate drawing out 200 Skittles 100 times [randBin(200, .20, 100)/200→L1]. Find the mean and standard deviation of the set of sample proportions in this simulation. Compare with the theoretical mean and standard deviation of the sampling distribution with sample size 200. Create and study the dot plot for this data set.

Theoretical Sample: Mean: ______________ Standard Deviation: _____________

Your Sample: Mean: ______________ Standard Deviation: ______________

g) How is the plot of the sampling distribution from part f different from the plot in part b? How do the mean and standard deviation compare?

h) What percentage of the 200 sample proportions fall within 0.114 of 0.20 (or approximately 2 standard deviations)? How does this compare with the answer to part e?

i) You should notice that these distributions follow an approximately normal distribution. The Empirical Rule states how much of the data will fall within 1, 2, and 3 standard deviations of the mean. Restate the rule:

Empirical Rule: In a normal distribution with mean μ and standard deviation σ

- ○ ____% of the observations fall within 1 standard deviation (1σ) of the mean (μ),
- ○ ____% of the observations fall within 2 standard deviations (2σ) of the mean (μ), and
- ○ ____% of the observations fall within 3 standard deviation (3σ) of the mean (μ).

j) Do your answers to parts e and h agree with the Empirical Rule? Explain.

This leads us to an important result in statistics: the **Central Limit Theorem (CLT) for a Sample Proportion**:

Choose a simple random sample of size *n* from a large population with population parameter *p* having some characteristic of interest. Then the sampling distribution of the sample proportion p is approximately normal with mean *p* and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. This approximation becomes more and more accurate as the sample size *n* increases, and it is generally considered valid if the population is much larger than the sample, i.e. *np ≥ 10* and *n(1 – p) ≥ 10.*

k) How might this theorem be helpful? What advantage does this theorem provide in determining the likelihood of events?