<u>Unit 1, Task 1 – Sample Proportions with Simulation</u>

Name(s): ____ Period: _______

I. Reviewing some basic definitions:

a. Think about a single bag of M&M's. Does this single bag represent a *sample* of M&M's or the *population* of M&M's?

____

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 62 M&M's in your bag, is 62 a statistic or a parameter? ___________

If Hershey claims that 24% of all M&M's are blue, is 24% a statistic or a parameter? ____________

II. How many blue candies should I expect in a bag of M&M's?

First, **random samples of 10 candies** were taken last year from bags of M&M's. The count and proportion of each color in the samples were recorded by each Math IV student in a table like this:

	Blue	**Not Blue**
Count		
Proportion		

The average of the sample proportions obtained by each class for sample size 10 was recorded and complied.

a. Do you think their averages are closer or farther from the true proportion of blue M&Ms than each individual student found in his/her sample? _____________ Why do you think that? ____

Next, **random sample of 25 candies** were taken and recorded by each Math IV student. The average of the sample proportions obtained by each class for sample size 25 was recorded and complied.

Finally, **random samples of 40 candies** were taken and recorded by each Math IV student. The averages of the sample proportions obtained by each class for sample size 40 was recorded and compiled. :

b. Do you think the averages of the 40 candies are closer or farther from the true proportion of blue M&Ms than each individual student found in his/her sample of size 10 and 25? _________ Explain: _______________________________________

Here is the data obtained in one of the class periods:

6th period

TOTALS:	**n= 10**	**n= 25**	**n= 40**
count:	46	105	165
# of samples:	170	425	680
sample proportion:	27.06%	24.71%	24.26%

c. What is your initial reaction to the three average sample proportions, i.e, which do you think is more accurate? Why? __________

____

III. Sampling Distribution of $\hat{p}$

We have been looking a number of different *sampling distributions* of $\hat{p}$, but we have seen that there is great variability in the distributions. We would like to know that $\hat{p}$ is a good estimate for the true proportion of blue M&M's. However, there are guidelines for when we can use the statistic to estimate the parameter. This is what we will investigate in the next section.

First, however, we need to understand the center, shape, and spread of the sampling distribution of $\hat{p}$.

We know that if we are counting the number of M&M's that are blue and comparing with those that are not blue, then the counts of blues follow a binomial distribution (given that the population is much larger than our sample size). Formulas for the mean and standard deviation of a binomial distribution were discussed in Math III.

Given that $\hat{p} = \frac{X}{n}$, where X is the count of blues and *n* is the total in the sample, how might we find $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$? Find formulas for each statistic. ____

This leads to the following 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 $\hat{p}$ be the proportion of the sample having that characteristic. Then:

- The mean of the sampling distribution is ____.
- The standard deviation of the sampling distribution is ___________.

a. Let's look at the standard deviation a bit more. What happens to the standard deviation as the sample size increases?

____

b. **Show** a few examples to verify your conclusion. Then use the formula to **explain** why your conjecture is true:

c. If we wanted to cut the standard deviation in half, thus decreasing the variability of $\hat{p}$, what would we need to do in terms of our sample size? Use the formula to explain.

***Caution**: We can only use the formula for the standard deviation of $\hat{p}$ when the population is at least 10 times as large as the *n*.

d. For each of the samples taken in part 3, determine what the population of M&M's 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.

IV. Simulating the Selection of Blue M&M's

There is variation in the distributions of M&M colors depending on the size of your sample. 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 blue candies. Let's assume $p = 0.24$.

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

   Type *randBin(50,0.24)* in your calculator. (*randBin* is found in the following way: Math→PROB→7.) What number did you get? __________ Compare with a neighbor. What do you think this command does?
   ____

   How could you obtain the proportion that are blue rather than the count? _________________________

b. Now, we want to generate 100 samples of size 50. This time, input *randBin(50,0.24,100)/**50** →L_1*. The latter part (store in L_1) puts all of the outputs into List 1.

   - Using your Stat Plots, create a histogram or stem-and-leaf plot of the proportions of blue candies. Sketch the graph below.

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

c. 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 from part 3?

   Mean: ______________ Standard Deviation: ______________

d. Use the TRACE button on the calculator to count how many of the 100 sample proportions are within ± 0.06 of 0.24. Note: 0.06 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 ± 0.12 and for within ± 0.18. Record the results below:

	Number of the 100 Sample Proportions	Percentage of the 100 Sample Proportions
Within ± 0.06 of 0.24		
Within ± 0.12 of 0.24		
Within ± 0.18 of 0.24		

e. If each of the 100 students who sampled M&M's were to estimate the population proportion of blue candies by going a distance of 0.12 on either side of his or her sample proportion, what percentage of the 100 students would capture the actual proportion (0.24) within this interval? _____________________

f. If you did not know the actual proportion of blues, would the simulation above provide you with a definitive way of knowing whether your sample was within 0.12 of the mean? ____________________ Explain:

   ____

g. Simulate drawing out 200 M&M's 100 times. 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.

   Mean: ______________ Standard Deviation: ______________

h. How does the plot of the sampling distribution different from the above plot? How do the mean and standard deviation compare?

Mean: _______________ Standard Deviation: _______________

i. What percentage of the 200 sample proportions fall within 0.06 of 0.24 (or approximately 2 standard deviations)? ______

How does this compare with the answer to part e? _________________________

j. You should notice that these distributions follow an approximately normal distribution, a topic you learned about in Math 2. You also learned the Empirical Rule that states how much of the data will fall within 1, 2, and 3 standard deviations of the mean. Restate the 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).
- ____% of the observations fall within 3 standard deviation (3σ) of the mean (μ).

k. Do your answers to parts e and i agree with the Empirical Rule? Explain.

This leads us to an important result in statistics: **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 $\hat{p}$ is approximately normal with $\mu = p$ and $\sigma = \sqrt{\frac{p(1-p)}{n}}$.

NOTE: **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. $np \geq 10$ and $n(1-p) \geq 10$.**

l. How might this theorem be helpful? ____

m. What advantage does this theorem provide in determining the likelihood of events? _______________

____