Activity: Sampling Distributions and Introduction to the Central Limit Theorem

Concepts: Random samples from populations, parameters, statistics, sampling distributions, empirical sampling distributions, the Central Limit Theorem.

Prerequisites: The student should be familiar with random variables, distributions (probability and empirical probability), expected value, and statistics such as the sample mean and sample variance.

In this activity we are going to take random samples from populations, compute statistics from those random samples, and then examine the probability distribution of the statistics. These probability distributions are known as *sampling distributions*. First we need a few definitions:

Def: Simple random sample = observations from a population that are independent and identically distributed (i.e. we will take random samples of the same size from the same population – either with replacement or from a very large population (N>20*n* where N is the population size and *n* is the sample size)).
Example: If each member of our class selects 20 students at random from the ASU student population, then each member of our class is taking a simple random sample of size 20 from the ASU student population.

Def: Parameter = numerical characteristic of the population
Examples: Mean GPA of all ASU students; proportion of all ASU students who prefer Coke
Symbols: population mean = μ, population standard deviation = σ, population proportion = p

Def: Statistic = numerical characteristic of a sample
Examples: Average GPA of 20 randomly chosen ASU students; proportion of 20 randomly chosen ASU students who prefer Coke
Symbols: sample mean = $\bar{x}$, sample standard deviation = s, sample proportion = $\hat{p}$
NOTE that if Suzie and Joey each take a random sample of size 20 from the ASU population, then the two sample means for average GPA that they compute will not necessarily be the same value!

Scenario: Penny Ages

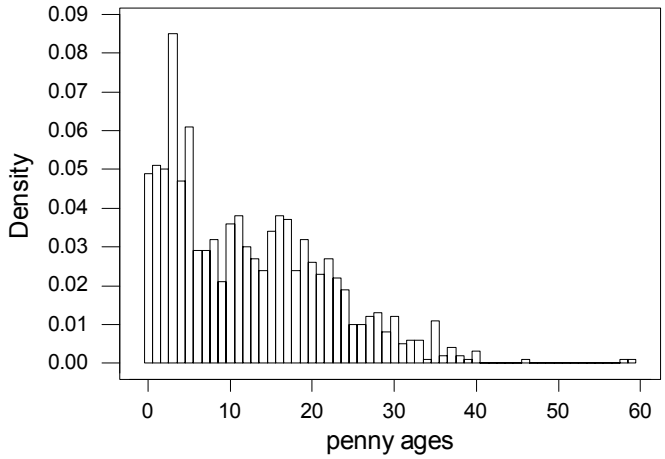
We will use Minitab to sample from a population and to examine the distribution of the sample mean statistic. First make sure you have a Minitab prompt (i.e. MTB>) in your session window. If you don't have a prompt then go to Edit|Preferences. The Preferences window should appear. Click on Session Window, then click Select. The Session Window Preferences window should appear. Click the Enable button for Command Language and click on the OK button. Now click the Save button in the Preferences window. A prompt should now appear in your session window.

Open the worksheet (File|Open Worksheet) `PENNIES_Student.MTW`. This file is located on the shared drive. The data in column one corresponds to penny ages for 1000 pennies. We will consider the data in column one to be the population data.

(a) Calculate the population mean and standard deviation of the 1000 penny ages (`MTB> describe c1`) and create a histogram of the penny ages (`MTB> histogram c1; SUBC> density; SUBC> bar; SUBC> connect.`):

mean μ = standard deviation σ =

Your histogram should look similar to the one below except it will have a curve showing the basic shape of the distribution.



(b) Comment on the shape of this distribution (mean, standard deviation, skewness, etc.). Show μ on the plot above.

(c) Now we are going to take a sample of size n=5 (with replacement) from this population. Note that we are taking a random sample from the 1000 pennies that comprise this population. Use Minitab to compute the sample mean, and place the sample mean into column `c3`:

```
MTB> sample 5 c1 c2;
SUBC> replace.
MTB> let c3(1)=mean(c2)
```

Now repeat this again but store the sample mean in the next row of column `c3`:

```
MTB> sample 5 c1 c2;
SUBC> replace.
MTB> let c3(2)=mean(c2)
```

Note that statistics such as sample mean and sample standard deviation change from sample to sample. Are your two sample means above the same value?

The Key Idea: Statistics are random variables!
Statistics have unknown numerical outcomes that follow a predictable pattern.
So statistics have probability distributions, expected value, variance, etc.
Our goal is to be able to describe the shape, center, and spread of the distribution of this statistic (the sample mean when the sample size is n=5).

We are now going to write a macro to get 500 random samples of size n=5 and compute the sample mean for each of these random samples. First enter the following command into Minitab:

```
MTB> let k1=1
```

Then copy into a text file (open Notepad under "Accessories") the following commands:

```
sample 5 c1 c2
let c3(k1)=mean(c2)
let k1=k1+1
```

Save the text file as "penny.mtb". Be sure to use the "mtb" extension and put quotes around the file name, and remember where you put it. (Save it to your disk if you have one.) Then within Minitab, select `File> Other Files> Run an Exec...`, tell it to execute 500 times, click on "Select file" and choose the "penny.mtb" file that you just created.

Name column `c3` and examine a histogram of the distribution of these 500 sample means, and calculate their mean and standard deviation:

```
MTB> name c3 'n=5 Sample Means'
MTB> histogram c3
MTB> describe c3
```

Comment on the shape and record these values:

shape:
mean of sample means:
standard deviation of sample means:

Lastly, generate a plot with the population distribution and the empirical sample mean distribution for sample size n=5:

```
MTB> histogram c1 c3;
SUBC> density;
SUBC> connect;
SUBC> overlay.
```

In order to clearly see the two distributions on one graph we are using the "connect" option instead of the "bar" option for the histogram. Make sure you know which curve belongs to which distribution! Note: The "connect" option connects the values at the center of each bar of the histogram with a line. Note that you can also use the pull down menu under Graph|Histogram to create this histogram. The Frame button will take you to options for graphing multiple graphs on the same sheet. Feel free to experiment with the Histogram options to get a better-looking histogram.

(d) Repeat (c) with samples of size n=10 and n=20. [*Hints*: Again use `c2` to store each sample generated and use `c4` to store the 500 sample means for n=10 and use `c5` to store the 500 sample means for n=20. You will need to edit the file penny.mtb to reflect the new sample sizes and where to store the sample means. Be sure to reset your counter (`k1`) to one on each run!]. When you are finished comment on the shape of the distribution of the sample means and record these values:

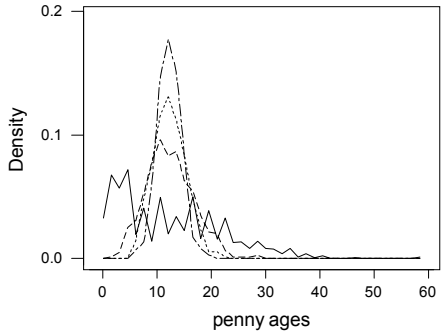
Shape for sample size n=10:
mean of sample means for sample size n=10:

standard deviation of sample means for sample size *n*=10:
Shape for sample size *n*=20:
mean of sample means for sample size *n*=20:
standard deviation of sample means for sample size *n*=20:

(e) Now create a histogram with the population distribution and the empirical sample mean distributions for *n*=5, 10, and 20:

```
MTB> histogram c1 c3 c4 c5;
SUBC> density;
SUBC> connect;
SUBC> overlay.
```

You should get something that looks like the graph to the right. Be sure that you are able to identify which curve belongs to which distribution. Also fill in the first three columns of the table below (Note to quickly view these data together type `MTB> describe c1,c3,c4,c5`).



Sample Size (*n*)	Mean of Sample Means	Standard Deviation of Sample Means	$\frac{\sigma}{\sqrt{n}}$
5			
10			
20			

(f) What do these histograms and simulation results reveal about the *shape* of the sampling distribution of $\overline{X}$ as the sample size increases?

(g) What do these histograms and simulation results reveal about the *mean* of the sampling distribution of $\overline{X}$ as the sample size increases?

(h) What do these histograms and simulation results reveal about the *standard deviation* of the sampling distribution of $\overline{X}$ as the sample size increases?

The above histograms (c3, c4, and c4 on your Minitab worksheet) show the *empirical sampling distributions of the* $\overline{X}$ *statistic based on samples of size 5, 10, and 20, respectively, from the population of the 1000 pennies.*

Def: Sampling Distribution = probability distribution of values the statistic can assume for all possible random samples of size *n* from the population.

Def: Empirical Sampling Distribution = distribution of observed values of the statistic for many, many samples of size *n* from the population.

(i) Now each for each *n*=5, 10, 20, compute σ/sqrt(*n*) (where σ is the population standard deviation) and add it to the table above. Is each value of σ/sqrt(*n*) close to the to the standard deviation for the sample mean for that value of *n*?

Your simulation results illustrate the famous *Central Limit Theorem* in probability and statistics. This theorem says the following about the sampling distribution of a sample mean $\bar{X}$:

- The mean of the sampling distribution of $\bar{X}$ equals the population mean μ, regardless of the sample size or the population distribution.
- The standard deviation of the sampling distribution of $\bar{X}$ equals the population standard deviation σ divided by the square root of the sample size, regardless of the population distribution.
- The shape of the sampling distribution of $\bar{X}$ is approximately normal for large sample sizes, regardless of the population distribution, and it is normal for any sample size when the population distribution is normal.

Soon we will see a mathematical justification for these results. First let's devote more time to understanding the results.

Scenario: Professor Lectures Overtime

Let X = amount of time a professor lectures after class should have ended. Suppose these times follow a Normal distribution with mean μ = 5 min and standard dev σ = 1.804 min.

(a) Draw a rough sketch (and label) this distribution.

(b) Is μ a parameter or a statistic?

(c) Suppose you record these times for 5 days $x_1, x_2, \ldots, x_5$ and calculate the sample mean $\bar{x}$. Is $\bar{x}$ a parameter or a statistic?

To investigate the sampling distribution of these $\bar{x}$ values, we will take many samples from this population and calculate the $\bar{x}$ value for each sample. Open the program *Sampling SIM* by double clicking on it in the shared drive (or desktop).

- Click the DISTRIBUTION button and select "Normal" from the list. You should see a sketch similar to what you drew in (a).
- From the Window menu, select "Samples."
- Click Draw Samples and one observation from the population is selected at random (Note: the program may be very slow the first couple of times you click this button). This is one realization of the random variable X.

(d) How long did the professor run over this time?

(e) Click Draw Samples again, did you observe the same time?

(f) Change the value in the Sample Size box from 1 to 5 and click Draw Samples. How does this distribution compare (roughly) to the population distribution?

(g) Click Draw Samples again. Did the distribution of your 5 sample values change?

(h) Change the sample size from 5 to 25 and click Draw Samples. Describe how this distribution differs from the ones in (f) and (g). How does the shape, center, and spread of this distribution compare to that of the population (roughly)? (The mean of this distribution is represented by $\bar{x}$, the standard deviation of this distribution is represented by *s*. Compare these values to μ and σ.)

(j) Click Draw Samples again. Did you get the same distribution? The same $\bar{x}$ and *s* values?

The main point here is that results vary from sample to sample. In particular, statistics such as $\bar{x}$ and s change from sample to sample. You will now look at the distribution of these statistics.

From the Windows menu, select Sampling Distribution. Move this window to the right so you can see all three windows at once. You should see one green dot in this window (it will be small and on the x-axis). This is the $\bar{x}$ value from the sample you generated in (i). In the Sampling Distribution window, click on "New Series" so it reads "Add More." Click the Draw Samples button. A new sample appears in the Sample Window and a second green dot appears in the Sampling Distribution window for this new sample mean. Click the Draw Samples button until you have 10 sample means displayed in the Sampling Distribution window. Note: You can click the F button in the Samples Window to speed up the animation. Record the values displayed in the "Mean of Sample Means" box and in the "Standard Dev. of Samples Means" box.

Mean of Sample Means _____

Standard Dev. of Sample Means _____

<u>Be very clear you understand</u> what these numbers represent. If not, ask an instructor or TA!

In the Population window, click on NORMAL to change the population to the one of the following which will be assigned to your group: Bimodal, Skew-, Skew+, Trimodal, U-Shaped, Uniform. Note, this changes the population mean μ and standard deviation σ as well. Record which population distribution you have been assigned and also the population mean and standard

deviation. Change Sample Size to 1 and number of samples to 500 (you definitely want to make sure you have the F button pressed in your Samples window to speed up the animation!). Click the Draw Samples button.

(j) Describe the shape, center, and spread of the Sampling Distribution of the $\bar{x}$ values. In particular, how do the shape, center, and spread compare to the population? You can click the purple population outline (upper left corner of Sampling Distribution window) for easier visual comparison. Sketch the distribution below. Record the mean and standard deviation of the sample means.

(k) Change the sample size to 5 and click the Draw Samples button. Answer again the questions from (j) and sketch the distribution below. This time try both the purple population outline and the blue normal outlines in the Sampling Distribution window.

(l) Change the sample size to 25, click the Draw Samples button, and answer the same questions.

(m) Does the blue normal distribution outline appear to be a better description of the sampling distribution of the sample mean $\bar{x}$ values than the purple population outline?

(n) Compare your answers to (j)-(m) with student groups who were assigned two different population distributions. What similarities do you find??

(o) Complete the table below twice (once each for two of the non-normal population distributions – clearly indicate which population distributions you use!). Are the theoretical values predicted by the Central Limit Theorem (CLT) close to the empirical values you got when you ran the simulations above?

Sample Size (*n*)	Population Mean	Empirical Mean of Sample Means	Theoretical Mean of Sample Means (via the CLT)	Population Standard Deviation	Empirical Standard Deviation of Sample Means	Theoretical Standard Dev. Of Sample Means (via the CLT)
1						
5						
25						