**Activity: Central Limit Theorem Theory and Computations**

**Concepts:** The Central Limit Theorem; computations using the Central Limit Theorem.

**Prerequisites:** The student should be familiar with the ideas of the Central Limit Theorem; expected value; statistics such as the sample mean and sample variance; and using the normal distribution to find probabilities.

**Recap:** In our last activity (*Sampling Distributions and Introduction to the Central Limit Theorem*) we used simulation to examine the sampling distribution for the sample mean statistic $\overline{X}$. First we saw that the sample mean $\overline{X}$ is a random variable. Our investigation of the empirical probability distribution (aka sampling distribution) of $\overline{X}$ by taking many samples of the same size, *n*, from the same population resulted in the following observations about the sampling distribution of $\overline{X}$:

*Population Parameters*: mean = μ, standard deviation = σ;
*Sample Statistics*: mean = $\bar{x}$, standard deviation = *s*
*Observations about the sampling distribution of* $\overline{X}$:

- shape: Bell shaped (i.e. normal shaped) distribution for "large enough" sample sizes, *n*.
- center: Distribution of $\overline{X}$ centered at the population mean μ.
- spread: Spread of $\overline{X}$ depends on sample size, *n*. Spread decreases as *n* increases (actually spread is σ/sqrt(*n*)).

*Conclusion*: The distribution of the sample mean, $\overline{X}$, will be centered at the population mean and shaped like a normal distribution if *n* is large or the population is normal to begin with.

Our simulation results to compute the sampling distributions for the sample mean statistic $\overline{X}$ illustrated the ***Central Limit Theorem***. This theorem says the following about the sampling distribution of the sample mean $\overline{X}$:

- The mean of the sampling distribution of $\overline{X}$ equals the population mean μ, regardless of the sample size or the population distribution.
- The standard deviation of the sampling distribution of $\overline{X}$ 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 $\overline{X}$ 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.

In this activity sheet, we are going to see why parts of the *Central Limit Theorem* are true and learn how to use this theorem in computations.

First let's recall what we mean by a **random sample from a distribution (population)**: If $X_i$, $i$=1,..,$n$ are $n$ independent observations from the same distribution (population), then $X_i$, $i$=1,..,$n$ is a random sample of size $n$ from that common distribution (population).

**Examples of Random Samples:**

**Coin Flipping:** Suppose we flip a fair coin 10 times and let the random variable X denote the number of heads. Then X is b(10,0.5) with E(X)=_____ and Var(X)=_____. Now suppose we run this experiment 20 times and observe the value of X on each run of the experiment, letting $X_i$, denote the number of heads on the *i*th run of the experiment. *Then $X_i$ , i=1,..,20 is a random sample of size 20 from the b(10, 0.5) distribution.* Note that the sample mean of this random sample is given by $(1/n)\Sigma(X_i)$ (where the summation is from *i*=1,..,20 and *n*=20).

**Polling:** Suppose we randomly select 1000 Americans and ask them if they approve of the job the President is doing. Let $X_i$ =1 if the *i*th American selected approves, zero otherwise. *Then $X_i$, i=1,..,1000 is a random sample of size 1000 from the Bernoulli distribution where the parameter p is the proportion of all Americans that approve.*
What is the expected value of this Bernoulli distribution?____
What is the standard deviation of this Bernoulli distribution?____
How is the sample mean of this random sample defined? _____
What does the sample mean of the sample represent?_____
If we let the random variable $X$ be the number of successes (i.e. number who approve) out of the 1000 samples, then how is $X$ distributed? (Hint: Think Bernoulli Trial!)_____

**Penny Ages:** In part (c) of the Penny Ages scenario of *Sampling Distributions and Introduction to the Central Limit Theorem* activity, we repeatedly (i.e. 500 times) got random samples of size *n*=5 from a population with mean _____ and standard deviation _____. For each of these random samples we computed the sample mean.

**Professor Lectures Overtime:** In part (h) of the Professor Lectures Overtime scenario of *Sampling Distributions and Introduction to the Central Limit Theorem* activity, we repeatedly got random samples of size ____ from a population with a _____ distribution with distribution mean _____ and distribution standard deviation _____. Again, for each of these random samples we computed the sample mean statistic.

**Theory:**
Now, to see why the first two bullets of the *Central Limit Theorem* are true let's recall some results for expected value:

Let $X$ be a random variable, then we have the following rules (proven on the bottom of page 125 of your text by using Theorem 3.2-1 on page 121 of your text). Note: Make sure you can reproduce these rules if you are given Theorem 3.2-1.

- Rules for Expected Value: $E(aX+b) = aE(X)+b$
- Rules for Variance: $V(aX+b) = a^2V(X)$

A generalization of these facts for more than one random variable can be found on page 294 of your text in Theorem 6.2-3:

**Theorem 6.2-3:** With more than one random variable,

- $E(a_1X_1+a_2X_2+\ldots+a_nX_n) = a_1E(X_1) +a_2E(X_2)+\ldots+a_nE(X_n) = \Sigma a_iE(X_i)$ (NOTE: You do not need the random variables, $X_i$ , to be independent for this result to hold.)
- *If the random variables are independent*, then $V(a_1X_1+a_2X_2+\ldots+a_nX_n) = a_1^2V(X_1) +a_2^2V(X_2)+\ldots+a_n^2V(X_n) = \Sigma a_i^2V(X_i)$ (since we will be working with random samples (i.e. each $X_i$ can be considered to be an independent observation from the same distribution!) then the $X_i$ can be considered independent!).

(a) Use the facts above to show that $E(\bar{X}) = \mu$ when the $X_i$ are a random sample from a distribution with mean $\mu$. (Hint: Use the definition of $\bar{X}$ and Theorem 6.2-3. This is shown after Example 6.2-4 on page 295 of your text. Try to show it before looking at the answer!).

(b) Use the facts above to show that expression for $Var(\bar{X})$ in terms of the population standard deviation $\sigma$. (Hint: This is also shown after Example 6.2-4 on page 295 of your text. Try to prove it before looking at the answer!). Does this expression support your observation that the standard deviation of the sample mean decreases as the sample size *n* increases?

(c) How did the above derivations depend on the population size? On the shape of the population?

**Applying the Central Limit Theorem:**

Let's examine the third bullet in our statement of the *Central Limit Theorem* above. First note that if the distribution from which you are sampling (i.e. the population distribution) is normal, say with mean = $\mu$ and standard deviation = $\sigma$, i.e. the population is $N(\mu , \sigma^2)$, then no matter how small the sample size *n*, the distribution of the sample mean, $\bar{X}$, is given by $N(\mu , \sigma^2 / n)$. This is Theorem 6.3-1 on page 299 of your text. Note this says that $E(\bar{X})=\mu$ and the standard deviation of $\bar{X}$ is $\sigma / \sqrt{n}$. Use this result to find the distribution of $\bar{X}$ for the Professor Lectures Overtime example above:

You should get that $\overline{X}$ has the distribution: N($5$ , $(1.804)^2 / 25$) (i.e. it is normal with mean 5 and standard deviation 1.804/sqrt(5)).

***The Consequence of All This*:** You can standardize and use normal distribution tables in the back of your textbook to calculate probabilities for the sample mean!

**A Worked Example:**

(a) For the "Professor Lectures Overtime" example above, find the probability that the amount of time the professor will lecture overtime is less than 5.5 minutes. Carefully define your random variables.

**Answer:** Let $X$ be the amount of time the professor lectures after class should have ended. We are given that $X$ is normally distributed: $N(5,(1.804)^2)$. Thus

$P(X < 5.5) = P\left(\frac{X-5}{1.804} < \frac{5.5-5}{1.804}\right) = P(Z < .2772) = .6092$ (where Z is standard normal).

(b) Now suppose you observe the professor for five days and record her overtime amount on each day. Note: We are assuming the amount of time the professor lectures overtime is independent from day to day. What is the probability that the average of these times is less than 5.5 minutes? Carefully define your random variables.

**Answer:** We have taken a random sample of size 5 from the $N(5,(1.804)^2)$ distribution and have computed the sample mean of that sample, say $\overline{x}$. From the Central Limit Theorem, since we are sampling from a normal distribution, then we know that $\overline{X}$ is $N(5 , (1.804)^2 / 25)$.

Thus $P\left(\overline{X} < 5.5\right) = P\left(\frac{\overline{X}-5}{\frac{1.804}{\sqrt{25}}} < \frac{5.5-5}{\frac{1.804}{\sqrt{25}}}\right) = P(Z < 1.386) = 0.9171$ where $Z$ is standard normal (Computation done via Minitab). Note this is the probability that the <u>average</u> amount of time the professor lectures overtime in five independent lectures is less than 5.5 minutes.

(c) Compare the answers to (a) and (b). Which is larger? Why does this make sense? Now suppose you randomly observe the professor for 40 days and record her overtime amount on each day. How will the probability that the average of these times is less than 5.5 minutes compare to the probabilities found in (a) and (b)? Explain why.

(d) Now consider the case where the random sample comes from a population with a distribution that is not normal but has finite mean μ and standard deviation σ. By the first two bullets of the *Central Limit Theorem* above, we know that the mean of the sampling distribution of $\overline{X}$ equals the population mean μ and the standard deviation of the sampling distribution of $\overline{X}$ equals the population standard deviation σ divided by the square root of the sample size. The third bullet of the *Central Limit Theorem* above says that as the sample size increases, i.e. as $n \to \infty$, then the distribution of $\overline{X}$ "approaches" a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. This is the same thing as saying that the random variable $\dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ becomes standard normal as $n \to \infty$. This is essentially the statement of the *Central Limit Theorem* on page 308 of your text (Theorem 6.4-1). How "large" does *n* need to be before we can use the normal distribution to *approximate* the distribution of $\overline{X}$? (See the paragraph in the center of page 309 of your text for an answer to this question.)

(e) Use the Central Limit Theorem to determine an approximate distribution of the sample mean for the Polling example above.

**Answer:** $\overline{X}$ is approximately normal with mean *p* and standard deviation $\sqrt{\dfrac{p(1-p)}{1000}}$.

(f) Recall that the sample mean in the Polling example above represented the proportion of people in the sample that approved of the President's performance. Let's call that proportion $\hat{p}$, i.e. $\hat{p} = \overline{X} = \dfrac{1}{1000}\sum_{i=1}^{1000} X_i$. Then find the approximate probability that $P\left(.58 < \hat{p} < .62\right)$ if $p = 0.60$. (**Answer:** Hint – convert $\hat{p}$ to a z-score and use the normal distribution. Answers: 0.8032 via Minitab, 0.8030 via tables)

**Note:** Examples 6.4-1 through 6.4-3 on page 308-9 of your text are also examples of using the CLT for computations

**Scenario: Selling Aircraft Communication Units**

Suppose a communications company sells aircraft communication units to civilian markets. Each month's sales depend on market conditions that cannot be predicted exactly, but the company executives predict their sales through the following probability estimates:

| $x$ | 25 | 40 | 65 |
|---|---|---|---|
| p($x$) | .4 | .5 | .1 |

where $x$ number of units sold.

(a) What is the expected number of units sold in one month = $\mu$ = E(X)?

(b) Determine the variance, $\sigma^2$ , of the number of units sold per month.

(c) Suppose we wanted to examine the average number of units sold per month, say $\overline{X}$ , for 3 years ($n$=36 months). Based on the central limit theorem (and assuming the number of units sold from month to month is independent), what can you say about the sampling distribution of $\overline{X}$ ? Also draw as sketch of this sampling distribution and be sure to indicate a label and numerical scale on the horizontal axis.

(d) Use the above to approximate the probability that the average number of units sold per month in 36 months is 40 or higher. You can first use the above mean and standard deviation to *standardize* 40 and use the tables in the back of your book. Or use Minitab and choose Calc > Probability Distributions > Normal. Use Cumulative probability and specify the appropriate mean and standard deviation for the sampling distribution, entering 40 as the input constant. Be sure to use proper notation to express this probability as well (P( $\overline{X} \geq 40$)) and shade the corresponding area in the above graph of the distribution of $\overline{X}$ .

(f) Would this probability increase or decrease (or stay the same) if the number of months were to increase? Explain.

(g) Use the CLT to approximate the probability that the mean number of units sold in 36 months is between 35 and 40.