Chapter 7

The Central Limit Theorem

7.1 The Central Limit Theorem[1]

7.1.1 Student Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Averages.
- Apply and interpret the Central Limit Theorem for Sums.

7.1.2 Introduction

What does it mean to be average? Why are we so concerned with averages? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study averages and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size n from a population with a known mean, μ, and a known standard deviation, σ. The first alternative says that if we collect samples of size n and n is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means (averages) and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, n, that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more observations are needed for the sample averages or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

[1]This content is available online at <http://http://cnx.org/content/m16953/1.14/>.

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times. (The 8, 7, 9, and 11 were randomly chosen.)

Each time a person rolls more than one die, he/she calculates the **average** of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The average is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one average when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more averages for a total of 10 averages.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the average. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "average" when you roll one die, is just the face on the die, what distribution do these "averages" appear to be representing?

Draw the graph for the averages using 2 dice. Do the averages show any kind of pattern?

Draw the graph for the averages using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the averages using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The average of the averages remains approximately the same.
2. The spread of the averages (the standard deviation of the averages) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means (averages) tend toward a normal distribution (the sampling distribution).**

7.2 The Central Limit Theorem for Sample Means (Averages)[2]

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable $\overline{X}$ which consists of sample means, tends to be **normally distributed** and

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

The Central Limit Theorem for Sample Means (Averages) says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means (averages) form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n, the sample size. n is the number of values that are averaged together not the number of times the experiment is done.

[2]This content is available online at <http://http://cnx.org/content/m16947/1.20/>.

The random variable $\overline{X}$ has a different z-score associated with it than the random variable X. $\overline{x}$ is the value of $\overline{X}$ in one sample.

$$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \tag{7.1}$$

μ_X is both the average of X and of $\overline{X}$.

$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$ = standard deviation of $\overline{X}$ and is called the **standard error of the mean.**

Example 7.1

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size n = 25 are drawn randomly from the population.

Problem 1

Find the probability that the **sample mean** is between 85 and 92.

Solution

Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean (or average)**.

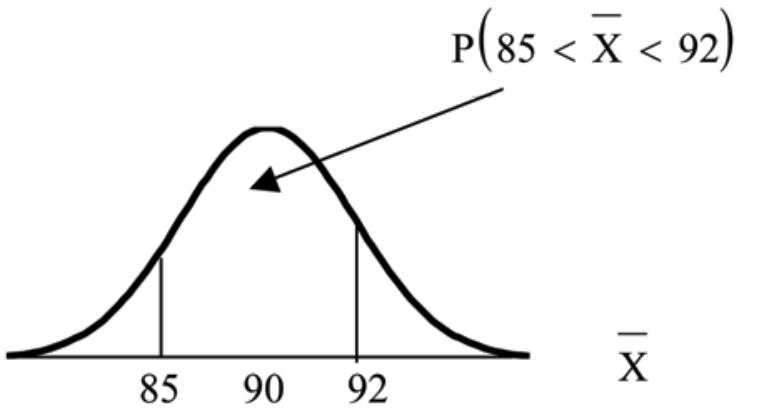
Let $\overline{X}$ = the mean or average of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$;

then $\overline{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$

Find $P\left(85 < \overline{X} < 92\right)$ Draw a graph.

$P\left(85 < \overline{X} < 92\right) = 0.6997$

The probability that the sample mean is between 85 and 92 is 0.6997.



TI-83 or 84: `normalcdf`(lower value, upper value, mean for averages, `stdev` for averages)

`stdev` = standard deviation

The parameter list is abbreviated (lower, upper, μ, $\frac{\sigma}{\sqrt{n}}$)

`normalcdf`$\left(85, 92, 90, \frac{15}{\sqrt{25}}\right) = 0.6997$

Problem 2
Find the average value that is 2 standard deviations above the the mean of the averages.

Solution
To find the average value that is 2 standard deviations above the mean of the averages, use the formula

value = μ_X + (#ofSTDEVs) $\left(\frac{\sigma_X}{\sqrt{n}}\right)$

value = $90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$

So, the average value that is 2 standard deviations above the mean of the averages is 96.

Example 7.2
The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size n = 50** is drawn randomly from the population.

Problem
Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution
Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean or average time, in hours**, it takes to play one soccer match.

Let $\overline{X}$ = the **average** time, in hours, it takes to play one soccer match.

If μ_X = _____, σ_X = _____, and n = _____, then $\overline{X} \sim N$(_____, _____) by the Central Limit Theorem for Averages of Sample Means.

$\mu_X = \mathbf{2}$, $\sigma_X = \mathbf{0.5}$, $n = \mathbf{50}$, and $X \sim N\left(2\ ,\ \frac{0.5}{\sqrt{50}}\right)$

Find $P\left(1.8 < \overline{X} < 2.3\right)$. Draw a graph.

$P\left(1.8 < \overline{X} < 2.3\right) = 0.9977$

$\texttt{normalcdf}\left(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}\right) = 0.9977$

The probability that the sample mean is between 1.8 hours and 2.3 hours is _____.

7.3 The Central Limit Theorem for Sums[3]

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

[3]This content is available online at <http://http://cnx.org/content/m16948/1.14/>.

a. μ_X = the mean of X
b. σ_X = the standard deviation of X

If you draw random samples of size n, then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

$$\Sigma X \sim N\left(n \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right)$$

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution). **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

a. Σx is one sum.
b. $z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$

a. $n \cdot \mu_X$ = the mean of ΣX
b. $\sqrt{n} \cdot \sigma_X$ = standard deviation of ΣX

Example 7.3
An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

Problem

a. Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
b. Find the sum that is 1.5 standard deviations below the mean of the sums.

Solution
Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values.**

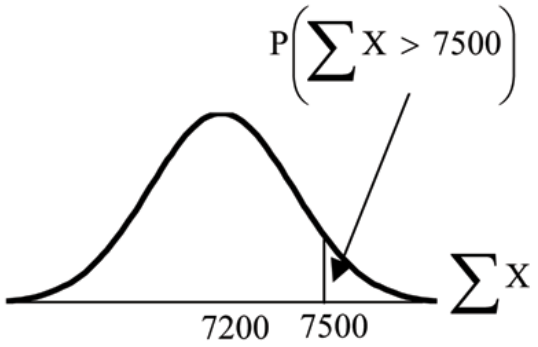
ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then

$$\Sigma X \sim N\left(80 \cdot 90, \sqrt{80} \cdot 15\right)$$

a. mean of the sums = $n \cdot \mu_X = (80)(90) = 7200$
b. standard deviation of the sums = $\sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
c. sum of 80 values = $\Sigma x = 7500$

Find $P(\Sigma X > 7500)$ Draw a graph.

$P(\Sigma X > 7500) = 0.0127$



`normalcdf`(lower value, upper value, mean of sums, `stdev` of sums)

The parameter list is abbreviated (lower, upper, $n \cdot \mu_X$, $\sqrt{n} \cdot \sigma_X$)

`normalcdf`(7500,1E99, $80 \cdot 90$, $\sqrt{80} \cdot 15 = 0.0127$

Reminder: $1E99 = 10^{99}$. Press the `EE` key for E.

7.4 Using the Central Limit Theorem[4]

It is important for you to understand when to use the **CLT**. If you are being asked to find the probability of an average or mean, use the CLT for means or averages. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for averages and sums.

NOTE: If you are being asked to find the probability of an **individual** value, do **not** use the CLT. **Use the distribution of its random variable.**

7.4.1 Examples of the Central Limit Theorem

Law of Large Numbers

The **Law of Large Numbers** says that if you take samples of larger and larger size from any population, then the mean $\overline{x}$ of the sample gets closer and closer to μ. From the Central Limit Theorem, we know that as n gets larger and larger, the sample averages follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for $\overline{X}$ is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean $\overline{x}$ must be close to the population mean μ. We can say that μ is the value that the sample averages approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean (Average) and Sum Examples

Example 7.4
A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

[4]This content is available online at <http://http://cnx.org/content/m16958/1.20/>.

1. The probability that the **average stress score** for the 75 students is less than 2.
2. The 90th percentile for the **average stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.

Problems 1. and 2. ask you to find a probability or a percentile for an **average** or **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n, is equal to 75.

Since the individual stress scores follow a uniform distribution, $X \sim U(1,5)$ where $a = 1$ and $b = 5$ (See Continuous Random Variables (Section 5.1) for the uniform).

$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$

$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$

For problems 1. and 2., let $\overline{X}$ = the average stress score for the 75 students. Then,

$\overline{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right)$ where $n = 75$.

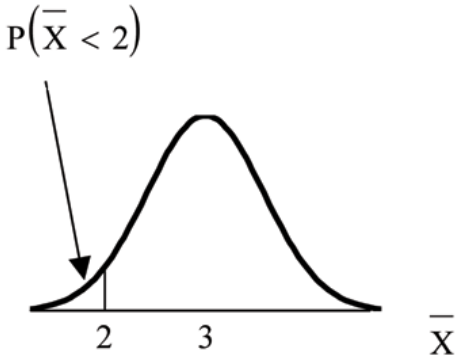
Problem 1

Find $P(\overline{X} < 2)$. Draw the graph.

Solution

$P(\overline{X} < 2) = 0$

The probability that the average stress score is less than 2 is about 0.



normalcdf $\left(1, 2, 3, \frac{1.15}{\sqrt{75}}\right) = 0$

REMINDER: The smallest stress score is 1. Therefore, the smallest average for 75 stress scores is 1.

Problem 2

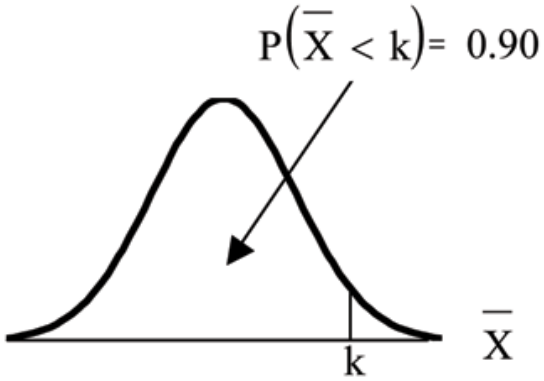
Find the 90th percentile for the average of 75 stress scores. Draw a graph.

Solution

Let k = the 90th precentile.

Find k where $P\left(\overline{X} < k\right) = 0.90$.

$k = 3.2$



The 90th percentile for the average of 75 scores is about 3.2. This means that 90% of all the averages of 75 stress scores are at most 3.2 and 10% are at least 3.2.

$\texttt{invNorm}\left(.90, 3, \frac{1.15}{\sqrt{75}}\right) = 3.2$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N\left[(75)\cdot(3), \sqrt{75}\cdot 1.15\right]$

Problem 3

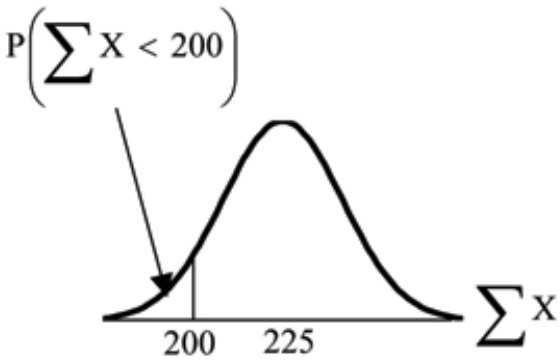
Find $P(\Sigma X < 200)$. Draw the graph.

Solution

The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75} \cdot 1.15 = 9.96$

$P(\Sigma X < 200) = 0$



The probability that the total of 75 scores is less than 200 is about 0.

$\texttt{normalcdf}\left(75, 200, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 0.$

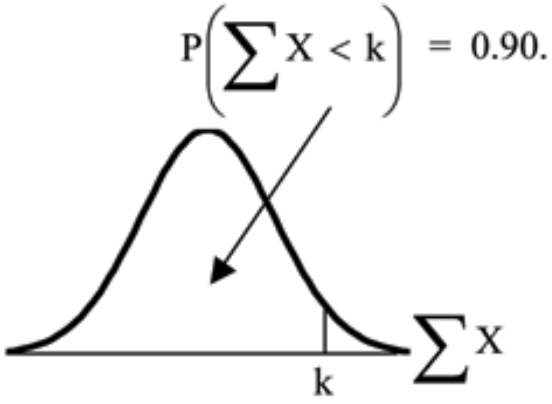
REMINDER: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Problem 4
Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution
Let k = the 90th percentile.

Find k where $P(\Sigma X < k) = 0.90$.

$k = 237.8$



The 90th percentile for the sum of 75 scores is about 237.8. This means that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

$\texttt{invNorm}\left(.90, 75 \cdot 3, \sqrt{75} \cdot 1.15\right) = 237.8$

Example 7.5
Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an **exponential distribution** with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim Exp\left(\frac{1}{22}\right)$ From Chapter 5, we know that $\mu = 22$ and $\sigma = 22$.

Let $\overline{X}$ = the AVERAGE excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$\overline{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means or Averages

Problem 1

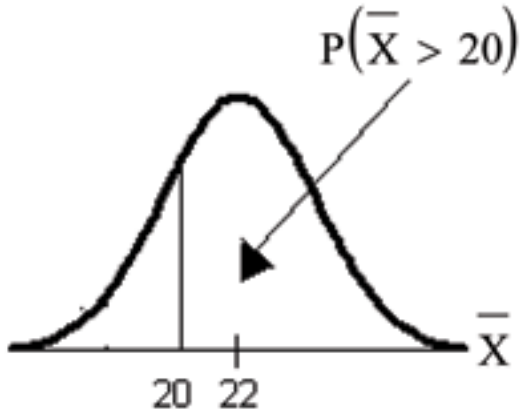
Using the CLT to find Probability:

a. Find the probability that the average excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P\left(\overline{X} > 20\right)$ Draw the graph.
b. Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P\left(X > 20\right)$
c. Explain why the probabilities in (a) and (b) are different.

Solution
Part a.
Find: $P\left(\overline{X} > 20\right)$

$P\left(\overline{X} > 20\right) = 0.7919$ using `normalcdf` $\left(20, 1E99, 22, \frac{22}{\sqrt{80}}\right)$

The probability is 0.7919 that the average excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



REMINDER: $1E99 = 10^{99}$ and $-1E99 = -10^{99}$. Press the EE key for E. Or just use 10^99 instead of 1E99.

Part b.
Find P(X>20) . Remember to use the exponential distribution for an **individual: X∼Exp(1/22)**.

P(X>20) = e^(–(1/22)*20) or e^(–.04545*20) = 0.4029

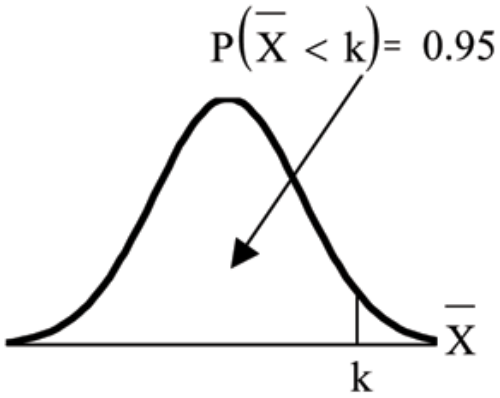
Part c. Explain why the probabilities in (a) and (b) are different.

$P\left(X > 20\right) = 0.4029$ but $P\left(\overline{X} > 20\right) = 0.7919$
The probabilities are not equal because we use different distributions to calculate the probability for individuals and for averages.
When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the CLT. Use the CLT with the normal distribution when you are being asked to find the probability for an average.

Problem 2
Using the CLT to find Percentiles:
Find the 95th percentile for the **sample average excess time** for samples of 80 customers who exceed their basic contract time allowances. Draw a graph.

Solution
Let k = the 95th percentile. Find k where $P\left(\overline{X} < k\right) = 0.95$

$k = 26.0$ using $\texttt{invNorm}\left(.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$



The 95th percentile for the **sample average excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

95% of such samples would have averages under 26 minutes; only 5% of such samples would have averages above 26 minutes.

NOTE: **(HISTORICAL): Normal Approximation to the Binomial**

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for n (say, 20). To calculate the probabilities with large values of n, you had to use the binomial formula which could be very complicated. Using the **Normal Approximation to the Binomial** simplified the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- . there are a certain number n of independent trials
- . the outcomes of any trial are success or failure
- . each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to X or subtract 0.5 from X (use $X + 0.5$ or $X - 0.5$. The number 0.5 is called the **continuity correction factor**.

Example 7.6
Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K - 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.

3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let $X =$ the number that favor a charter school for grades K - 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y. $Y \sim N(159, 8.6447)$. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so $P(X \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

`normalcdf` $(149.5, 10\text{^}99, 159, 8.6447) = 0.8641$.

For Problem 2., you **include 160** so $P(X \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

`normalcdf` $(0, 160.5, 159, 8.6447) = 0.5689$

For Problem 3., you **exclude 155** so $P(X > 155)$ has normal approximation $P(Y > 155.5) = 0.6572$.

`normalcdf` $(155.5, 10\text{^}99, 159, 8.6447) = 0.6572$

For Problem 4., you **exclude 147** so $P(X < 147)$ has normal approximation $P(Y < 146.5) = 0.0741$.

`normalcdf` $(0, 146.5, 159, 8.6447) = 0.0741$

For Problem 5., $P(X = 175)$ has normal approximation $P(174.5 < Y < 175.5) = 0.0083$.

`normalcdf` $(174.5, 175.5, 159, 8.6447) = 0.0083$

Because of calculators and computer software that easily let you calculate binomial probabilities for large values of n, it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial. In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.

For **Example 3**, the probabilities are calculated using the binomial ($n = 300$ and $p = 0.53$) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.

$P(X \geq 150)$: `1 - binomialcdf` $(300, 0.53, 149) = 0.8641$

$P(X \leq 160)$: `binomialcdf` $(300, 0.53, 160) = 0.5684$

$P(X > 155)$: `1 - binomialcdf` $(300, 0.53, 155) = 0.6576$

$P(X < 147)$: `binomialcdf` $(300, 0.53, 146) = 0.0742$

$P(X = 175)$: (You use the binomial pdf.) `binomialpdf` $(175, 0.53, 146) = 0.0083$

**Contributions made to Example 2 by Roberta Bloom

7.5 Summary of Formulas[5]

Rule 7.1: Central Limit Theorem for Sample Means (Averages)

$\overline{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ **Mean for Averages** $\left(\overline{X}\right)$**:** μ_X

Rule 7.2: Central Limit Theorem for Sample Means (Averages) Z-Score and Standard Error of the Mean

$z = \frac{\overline{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$ **Standard Error of the Mean (Standard Deviation for Averages** $\left(\overline{X}\right)$**):** $\frac{\sigma_X}{\sqrt{n}}$

Rule 7.3: Central Limit Theorem for Sums

$\Sigma X \sim N\left[(n) \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right]$ **Mean for Sums** (ΣX)**:** $n \cdot \mu_X$

Rule 7.4: Central Limit Theorem for Sums Z-Score and Standard Deviation for Sums

$z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$ **Standard Deviation for Sums** (ΣX)**:** $\sqrt{n} \cdot \sigma_X$

[5]This content is available online at <http://http://cnx.org/content/m16956/1.6/>.