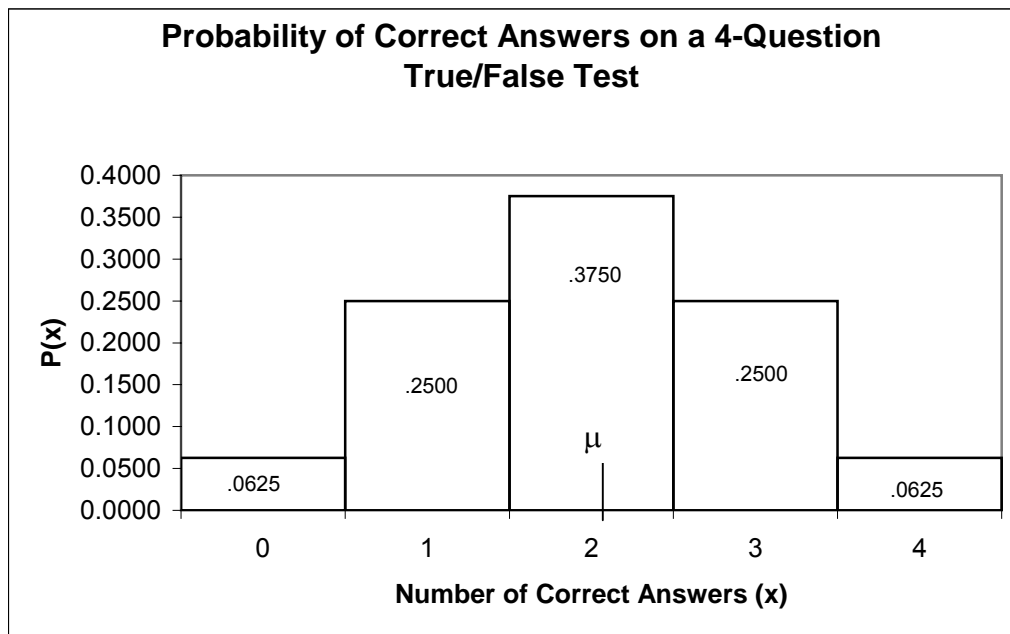


Chapter 7 Continuous Probability Distributions

Remember our example about the test with 4 True/False questions. The random variable “ x ” could have values of only 0, 1, 2, 3, or 4 correct answers. The probability distribution and histogram that resulted was:

x	$P(x)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625



We had also calculated the mean number of correct answers $\mu = 2$ and the standard deviation of correct answers $\sigma = 1$.

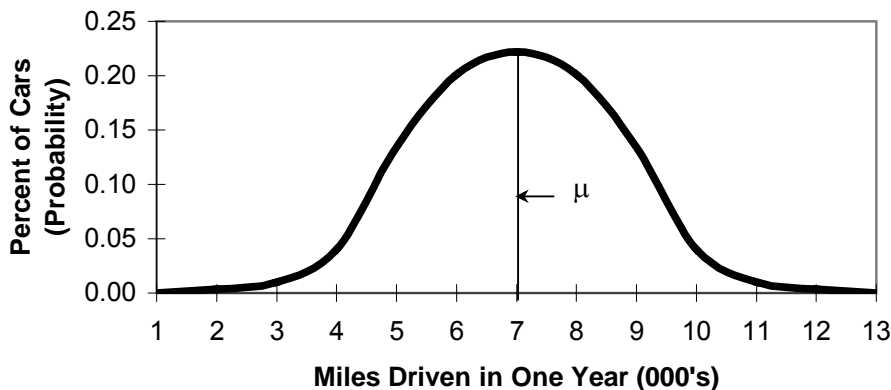
This is an example of a discrete probability distribution because the random variable x represents a finite, countable number of outcomes. That is, it can only be 0, 1, 2, 3, or 4. No other values satisfy this situation.

In addition, we could find the probability that a randomly selected exam would have between 1 and 3 correct answers, inclusive, by adding the areas of the bars for $x = 1$, $x = 2$, and $x = 3$: $.2500 + .3750 + .2500 = .875$. Hence, $P(1 \leq x \leq 3) = 0.875$.

Now assume we have a random variable that is not discrete, but continuous, like the heights of individuals in the population. In such a case, there is not a finite, countable number of heights, but an infinite number of possible heights that exist between 0 feet and 10 feet. That is, **ANY** height " x " might be a possible outcome. The number of possible heights for individuals cannot be counted. The probability histogram of this case is represented by a continuous curved line showing the existence of an outcome for each and every x on the x -axis.

When you renew your auto insurance you are always asked to provide the annual miles driven on the renewal form. The responses given, x , can be any number of miles and fractions of miles. Therefore, the responses come from a continuous distribution of outcomes, and the resultant probability distribution (% of possible miles driven) would be a continuous line as shown below.

Probability Distribution (Continuous Case)

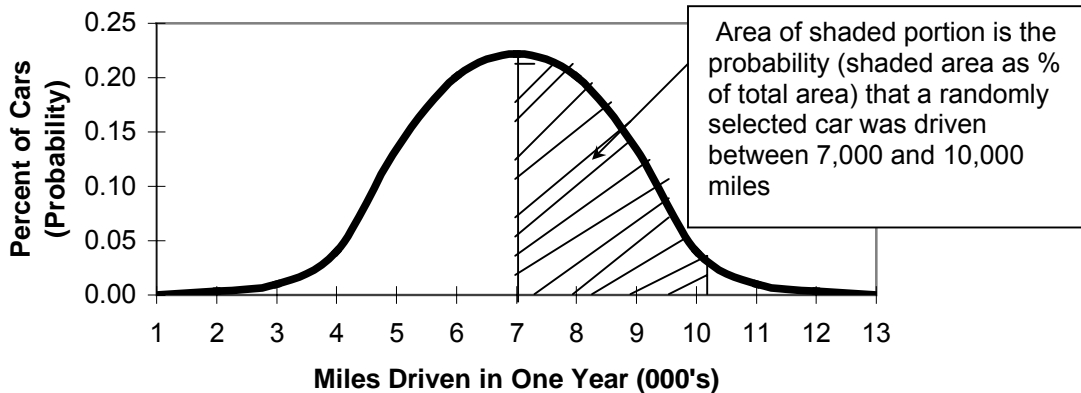


Notice that for any value of " x ", say 11,523 miles, there is a corresponding value of " y " representing the percentage of all cars reported to have driven that mileage - and it would represent the probability that a randomly selected car would be driven that many miles.

Remember that in the discrete case the areas under the bars represented the probability of that outcome. That is also the case for continuous probability distributions. In the discrete case above, we wanted to know $P(1 \leq x \leq 3)$, and we found it by adding the areas of the bars. Since each bar was 1 unit wide, that gave the same result as just adding the probabilities.

But if you look at the continuous graph above, how would we calculate the probability that a randomly selected car drove between 7,000 and 10,000 miles? By extension of the discrete case, we could just measure the area of the graph between 7,000 and 10,000! The question is, how do we do this given that we are no longer dealing with rectangles (bars) but with a curved line of varying height?

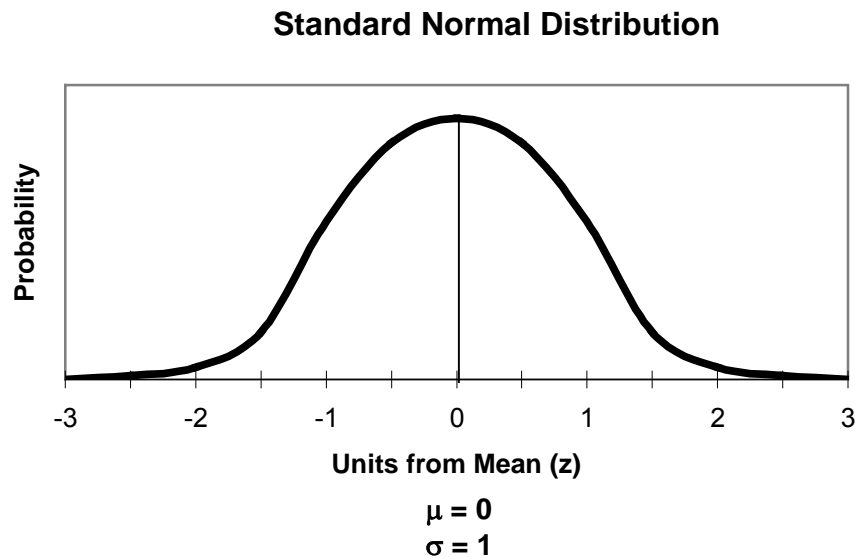
Probability Distribution (Continuous Case)



The answer is integral calculus. And since each distribution we might want to measure is different from every other one, this would result in a lot of higher mathematics in the field of statistics. Fortunately, statisticians and mathematicians have had mercy and come up with tools to avoid this math.

First, we start by limiting ourselves to a special case, one called "**normal**". A **normal distribution** is one that is symmetrical about the mean - that is, with the mean as the center point, the left half of the distribution is the mirror image of the right half. Under such a definition, then, **the area to the left of the mean must be 1/2 of the total area (probability) and the area to the right of the mean must be 1/2 of the total area (probability), with the mean exactly in the middle of the distribution.** This is the case in our car mileage example.

To make the math even simpler, we next select one special normal distribution - one where $\mu = 0$ and $\sigma = 1$. We will call this distribution the **Standard Normal Distribution**. *It is a normal distribution that has a mean of zero units and a standard deviation of 1 unit.* Hence, the standard normal distribution looks like this (see next page):



In order not to confuse the standard normal distribution with any other normal distribution, we no longer call the random variable x , but z . Therefore, on the standard normal distribution, we determine $P(z)$, not $P(x)$, the only difference being that z denotes x for this one special distribution.

Notice that, since the standard deviation of this distribution is 1, when $z = 1$ unit it is one standard deviation away from the mean. Likewise, $z = 2$ is two standard deviations from the mean, and $z = 1.5$ is one and one-half standard deviations away from the mean. Because of the way the standard normal distribution was created, then, z measures the number of standard deviations a value is away from the mean. On any normal distribution, if we measure the number of standard deviations x is from the mean, then we are stating x in terms of z , and we call the number of standard deviations that x is from its mean the z -score.

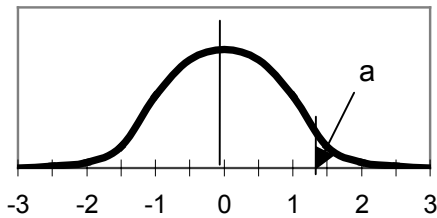
When the probability problem is limited to a standard normal distribution, the math involved in calculating the area between two numbers (z) is highly simplified. These areas are calculated for us and presented in the z -distribution table (Appendix D). Note in the table that, for a further simplification, **the table measures only the area (that is, probability) between the mean of 0 and some value of z , say $z=a$. Therefore, it is necessary that the problem be formulated in such a way as to arrive at the probability required by utilizing the z -table.**

To find the probability that an outcome in a standard normal distribution is between " a " standard deviations and " b " standard deviations of the mean, we would write $P(a < z < b)$. Likewise, $P(z < a)$ is the probability that z is less than " a " standard deviations from the mean and $P(z > a)$ is the probability that z is more than " a " standard deviations from the mean.

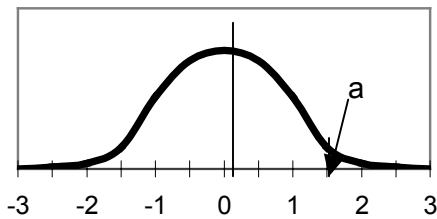
Because the z-table measures only the probability for $0 < z < a$, $a =$ some positive number, it is necessary to structure each problem individually to extract the correct probabilities. Note that, in order to look up a probability in this table, we must formulate the problem in terms of $P(0 < z < a)$ using our knowledge of normal probability distributions: $\Sigma P(x) = \text{total area under graph} = 1.0$, with the mean of the distribution exactly in the middle.

Most probabilities of this type can be solved using one of three “basic” cases, or a combination of them.

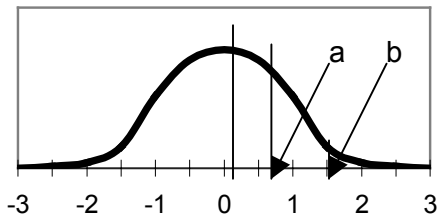
Basic Solutions Using the z-table



$P(z < a) = 0.5 + P(0 < z < a)$
For $a = 1.75$, $P(z < 1.75) = .5 + .4599 = .9599$

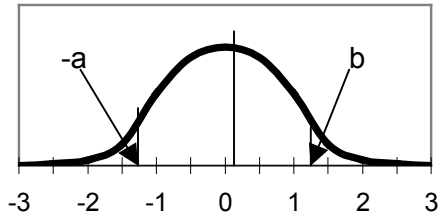


$P(z > a) = 0.5 - P(0 < z < a)$
For $a = 2.4$, $P(z < 2.4) = 0.5 - .4918 = .0082$



$P(a < z < b) = P(0 < z < b) - P(0 < z < a)$
For $b = 3.0$ and $a = 2.5$, $P(2.5 < z < 3.0) = .4987 - .4938 = .0049$

An example that combines two of the three “basic” solutions is shown here:



$P(z < a \text{ or } > b) = [0.5 - P(0 < z < a)] + [0.5 - P(0 < z < b)]$
<p>For $a = -1.96$ and $b = 1.96$</p> $P(z < -1.96 \text{ or } z > 1.96) = [0.5 - .4750] + [0.5 - .4750] = .0250 + .0250 = .0500$

Notice how the z-distribution table derives the empirical rule:

showing that 68% of the outcomes are within one standard deviation of the mean (as measured by $z = 1$ standard deviation and $z = -1$ standard deviation).

Homework: p. 196, Exercises 8, 9, 10, 12

Converting normal distributions to their standard normal equivalents.

It is undoubtedly obvious that very few distributions are going to be standard normal (i.e., $\mu = 0$ and $\sigma = 1$). Therefore, it seems that its usefulness would be limited except for one thing - **it measures probabilities in terms of how many standard deviations an outcome is from the mean of the distribution**, i.e. in terms of z . It follows that, if we can state an outcome " x " from any normal distribution (with mean μ and standard deviation σ) in terms of how many standard deviations x is from the mean, we will be stating it in terms of z and can look up the result in the z -table. To do this we measure how many units x is from μ by calculating $x - \mu$. Then we divide by the standard deviation σ to determine how far x is from the mean in terms of number of standard deviations. The above can be written as a formula:

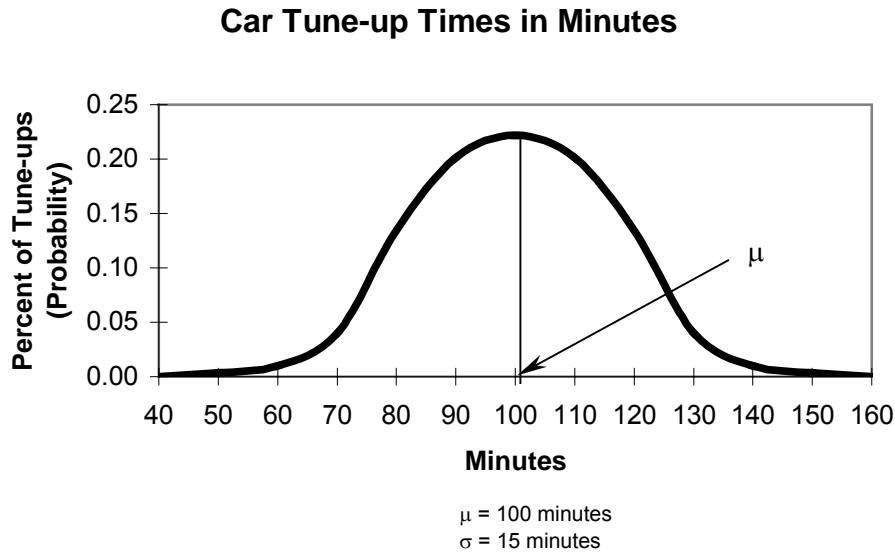
$$z = (x - \mu) / \sigma$$

The calculated value for z is entered into the z -table to determine the area under the curve (probability) between the mean and x .

Note that if a distribution is a standard normal distribution, then $\mu = 0$ and $\sigma = 1$, so that the equation above becomes

$$z = (x - \mu) / \sigma = (x - 0) / 1 = x \text{ or } z = x.$$

Example: Suppose a normally distributed population of car tune-up times have a mean of 100 min. and a standard deviation of 15 min. Then the graph of the distribution looks like this:



We can convert this distribution into its standard normal equivalent using the above formula as follows:

Note that $x = 115$ minutes is $z = 1$ standard deviations from the mean of 100 minutes. Therefore, the normally distributed population with mean = 100 minutes and standard deviation = 15 minutes is equivalent to a standard normal distribution with mean = 0 and standard deviation = 1 through use of the above formula. Probabilities on any normal distribution can therefore be obtained through a combination of the formula $z = (x - \mu) / \sigma$ and the z-table.

Therefore, if we want to know the probability that a give tune-up will take between 100 min. (μ) and 130 min. (x), we have only to calculate the z equivalent of the x's using the formula:

What is the probability that a given tune-up will take less than 100 minutes?
Less than 130 minutes? Less than 85 minutes?

What is the probability that a given tune-up will take between 85 and 130 minutes?

Algebraic properties of the z-formula.

Note that the formula $z = (x - \mu) / \sigma$ is an algebraic expression of 4 values. If we know any three of these values, we can always find the remaining unknown. For example, if we know z and want to find x , we could multiply both sides by σ and add μ to both sides to come up with the formula for x :

$$x = z\sigma + \mu$$

This very important concept will get much use as we go through the course. It says that, for any normal distribution with known mean and standard deviation, we can determine any value for x as long as we know how many standard deviations (z) that x is from the mean. If we want to know what values of x will result in selected probabilities, we can look up those probabilities in the z-table, find the associated z-score, and plug into the above formula. Thus, we can determine what values of x have any desired probability of occurring.

In the example above on wait times for a car tune-up, there is a 90% probability that a random car tune-up will take longer than how many minutes?

Here we know that $P(x > ?) = .90$. We also know that $\mu = 100$ minutes and $\sigma = 15$ minutes. To solve the formula $x = z\sigma + \mu$ for x , we must know the z-score associated with $P(z > ?) = .90$. Or, $P(z > ?) = 0.5 + P(0 < z < ?) = .90$. Therefore, $P(0 < z < ?) = .40$. Looking .40 up in the z-table gives a z-score of 1.28. In this case it must be -1.28, since the 90% of the distribution can only be above a number to the left of the mean. Therefore, $x = (-1.28)(15) + 100 = 80.8$ min.

Homework: p. 199-200, Exercises 13, 14, 15
p. 202, Exercises 17, 19, 21
p. 204, Exercises 23, 25, 27
pp. 206-208, Exercises 33, 34, 37, 39, 41, 43, 45, 47, 52

Chapter 8 The Central Limit Theorem (Calculating Probabilities about the Means of Samples)

In order to make use of sample data, we must be sure that we have a properly constructed **random sample**. Even so, the sample will still be subjected to **sampling errors (the difference between a sample statistic and its corresponding population parameter - how close is it?)** that result from the actual sampling process and **nonsampling errors** caused by things not related to the sampling, such as poorly worded questions and data entry errors. These are to be avoided to the maximum extent possible by proper experimental design.

The most common source of error, and the one that is hardest to avoid, is to have some amount of non-randomness in the sample. EACH member of the population must have the same chance of being included in the sample. If such is not the case, the value of the sample statistic is reduced or worse, misleading.

It has been found that the sample mean \bar{X} is the best possible **estimator** of the population mean μ , surpassing such other measures of central tendency as mode or median as an unbiased estimator. Therefore, the sample mean, \bar{X} , is the statistic used to make **inferences** about the true population mean, μ . We call \bar{X} a **point estimate** of the population mean.

Suppose we want to make a determination about the **mean** number of years of formal education from samples drawn from the adult population of the Columbia area. That is, instead of selecting a random individual and determining the probability that he has x years of education, we want to take a sample from the population and **determine probabilities about the mean \bar{X} of the sample**. We don't know any of this information about the population, but we can collect data from samples drawn from the population.

Suppose that in Columbia, an adult can have had between 8 and 16 years of formal education so as to simplify the problem. Additionally, assume that the distribution of years of education is normal. We take 10 random samples of size 5 from the adult population (experiment design) and get the following information.

<u>Sample n_s</u>	<u>Data (# of year of formal education)</u> <u>($n=5$, random variable x)</u>					<u>Total</u> <u>$\sum x$</u>	<u>Mean</u> <u>$\bar{x} = \sum x/n$</u>
1	10	11	12	16	9	58	11.6
2	15	16	10	8	12	61	12.2
3	12	12	13	14	11	62	12.4
4	16	14	10	12	12	64	12.8
5	8	10	12	14	16	60	12.0
6	9	16	12	13	14	64	12.8
7	15	16	16	10	12	69	13.8
8	14	8	9	12	16	59	11.8
9	15	8	16	14	10	63	12.6
10	12	11	9	9	16	57	<u>11.4</u>
							123.4

We now have a distribution of the mean number of years of formal education for samples of size 5 from the adult Columbia population. This distribution, the right column above, is like any other distribution. Therefore, it must have a mean and standard deviation.

$$\text{mean of the sample means} = \mu_x = \sum \bar{x} / n_s = 123.4 / 10 = 12.34$$

$$\text{std. dev. of the sample means} = \sigma_x = \sqrt{\frac{\sum (\bar{x} - \mu_x)^2}{n-1}} = \sqrt{4.4840 / 9} = .7058$$

as calculated from the table below.

Note that the mean and standard deviation of the sample (μ_x and σ_x) will not exactly equal the true mean and standard deviation of the population. The difference between a sample's statistics and the population's parameters is called **sampling error** (see above).

<u>\bar{x}</u>	<u>μ_x</u>	<u>$\bar{x} - \mu_x$</u>	<u>$(\bar{x} - \mu_x)^2$</u>
11.6	12.34	-0.74	0.5476
12.2	12.34	-0.14	0.0196
12.4	12.34	0.06	0.0036
12.8	12.34	0.46	0.2116
12.0	12.34	-0.34	0.1156
12.8	12.34	0.46	0.2116
13.8	12.34	1.46	2.1316
11.8	12.34	-0.54	0.2916
12.6	12.34	0.26	0.0676
11.4	12.34	<u>-0.94</u>	<u>0.8836</u>
		0	4.4840

The **Central Limit Theorem** states that

If $n > 30$ or the original population is normal, the distribution of the sample means approximates a normal distribution. The larger n , the more normal the approximation.

The mean of the sample means is equal to the population mean.

$$\mu_x = \mu$$

The standard deviation of the sample means is equal to the population standard deviation divided by the square root of the sample size. This value is known as the **Standard Error of the Mean**:

$$\sigma_x = \sigma/\sqrt{n}$$

Thus, the larger the sample size, the smaller the std. deviation of the sample means.

When faced with a problem concerning the means of samples, therefore, the only change that must be made is to replace σ with σ_x . That is, we adjust the population standard deviation by dividing it by the square root of the sample size. Then, using the new notation in the z-formula, we get

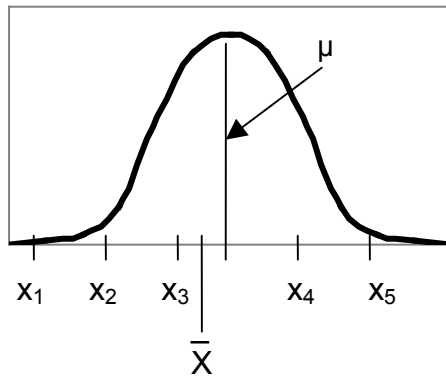
$$z = (\bar{x} - \mu_x) / \sigma_x$$

or
$$z = (\bar{x} - \mu) / (\sigma/\sqrt{n})$$
 since $\mu_x = \mu$ and $\sigma_x = \sigma/\sqrt{n}$

Note that x is replaced by the sample mean of interest, \bar{x} , because we are now using the **distribution of means of samples of size n** , not of individual outcomes. This determines the probability that, given a sample of size n , the mean of that sample will be between two numbers "a" and "b": $P(a < \bar{x} < b)$. Note the distinction between the formulation of this probability and that for the probability that an individual outcome will be between "a" and "b": $P(a < x < b)$.

All other calculations associated with a sample means problem is identical to that used with any other normal distribution.

The above results are better visualized if you consider the following: Suppose we take a sample of size "n" from a distribution of "x's". The x's that we select in taking the sample might fall anywhere on the x-axis as shown here:



Now, if we calculate the mean of our sample, \bar{X} , that mean will be much closer to the population mean than any of the individual x 's (see drawing). If we repeatedly take samples of the same size, this would hold true for each of the samples. By considering all the sample means we calculated to be their own distribution, you can see that the mean of all the sample means (μ_x) would be very close to the true population mean, μ , and that it would be much more concentrated about the mean than the original x 's. That is to say, it would have a smaller standard deviation than the original population. How much smaller? σ/\sqrt{n} .

Example: Forest Department wildlife surveys indicate that the mean weight of robins in South Carolina is 2.0 pounds with a standard deviation of 0.25 pounds. Following 3 years of drought, the department desires to see if the dry weather has affected the mean weight of robins. A sample of 40 robins is trapped and their weights are measured. What is the probability that the mean weight of the sample of robins would be less than 1.95 pounds if they had not lost weight due to the drought?

To summarize, the formal **Central Limit Theorem** consists of the following:

Given	a population of continuous random variables x mean of distribution of $x = \mu$ standard deviation of distribution of $x = \sigma$ samples of size n are randomly selected from the population
Then	<u>The</u> distribution made up of all possible samples means \bar{x} will be approximately normal for $n > 30$ The mean of the sample means $\mu_{\bar{x}}$ will be the same as the population mean μ. The standard deviation of the sample means $\sigma_{\bar{x}}$ will be the population std. dev. divided by the square root of the sample size σ / \sqrt{n}.
Homework:	p. 219, Exercise 4 p. 225, Exercise 5 p. 233, Exercise 12 p. 237, Exercises 15, 16, 17, 18 pp. 239-241, Exercises 21, 27, 31, 35, 37