appropriate numbers in the formula, and we get the probability of specific events. However, for these formulas to work they must be applied only to cases for which they were designed.

Constructing a Probability Distribution for Random Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-13

Probability with Discrete Random Variable Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-14

Check Your Understanding: Random Variables

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-23

Mean (expected value) of a Discrete Random Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-15

Expected Value

Example: Suppose we have a six-sided die marked with five 3's and one 6. What would you expect the average of 6000 rolls to be?

Answer: If we knew the value of each roll, we could compute the average by

summing the 6000 values and dividing by 6000. Without knowing the values, we can compute the **expected average** as follows.

Since there are five 3's and one 6, we expect roughly 5/6 of the rolls will give 3 and 1/6 will give 6. Assuming this to be exactly true, we have the following table of values and counts (Table 3):

Table 3

Value:	3	6
Expected counts:	5000	1000

The average of these 6000 values is then

$$\frac{5000 \cdot 3 + 1000 \cdot 6}{6000} = \frac{5}{6} \cdot 3 + \frac{1}{6} \cdot 6 = 3.5$$

We consider this the expected average in the sense that we "expect" each of the possible values to occur with the given frequencies.

Definition: Suppose X is a discrete random variable that takes values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_1), p(x_2), \ldots, p(x_n)$. The **expected value** of X is denoted $E(X)$ and defined by

$$E(X) = \sum_{j=1}^{n} p(x_j)\, x_j = p(x_1)\, x_1 + p(x_2)\, x_2 + \cdots + p(x_n)\, x_n$$

Notes:

The expected value is also called the **mean** or average of X and is often denoted by μ ("mu").

As seen in the above example, the expected value need not be a possible value of the random variable. Rather it is a weighted average of the possible values.

Expected value is a summary statistic, providing a measure of the location or central tendency of a random variable.

If all the values are equally probable, then the expected value is just the usual average of the values.

Probability Mass Function and Cumulative Distribution Function

It gets tiring and hard to read and write $P(X = a)$ for the probability that $X = a$. When we know we are talking about X we will simply write $p(a)$. If we want to make X explicit we will write $p_X(a)$.

Definition: The probability mass function (pmf) of a discrete random variable is the function $p(a) = P(X = a)$.

Note:

1. We always have $0 \leq p(a) \leq 1$.
2. We allow *a* to be any number. If *a* is a value that *X* never takes, then $p(a) = 0$.

Mean and Center of Mass

You may have wondered why we use the name "probability mass function." Here is the reason: if we place an object of mass $p(x_j)$ at position $x_1 = 3$ for each *j*, then *E(X)* is the position of the center of mass. Let us recall the latter notion via an example.

Example: Suppose we have two masses along the x-axis, mass $m_1 = 500$ at position $x_1 = 3$ and mass $m_2 = 100$ at position $x_2 = 6$. Where is the center of mass?

Answer: Intuitively we know that the center of mass is closer to the larger mass.

Figure 4

From physics, we know the center of mass is

$$\bar{x} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2} = \frac{500 \cdot 3 + 100 \cdot 6}{600} = 3.5$$

We call this formula a "weighted" average of the x_1 and x_2. Here x_1 is weighted more heavily because it has more mass.

Now look at the definition of expected value *E(X)*. It is a weighted average of the values of *X* with the weights being probabilities $p(x_i)$ rather than masses! We might say that "The expected value is the point at which the distribution would balance." Note the similarity between the physics example and the previous dice example.

Check Your Understanding: Mean of a Discrete Random Variable

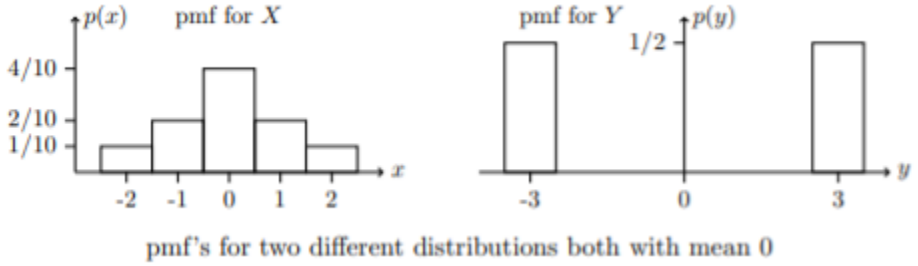
An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-24

Variance and Standard Deviation for Discrete Random Variables

The expected value (mean) of a random variable is a measure of location or central tendency. If you had to summarize a random variable with a single number, the mean would be a desirable choice. Still, the mean leaves out a good deal of information. For example, the random variables X and Y below both have mean 0, but their probability mass is spread out about the mean quite differently.

values X	-2	-1	0	1	2	values Y	-3	3
pmf $p(x)$	$1/10$	$2/10$	$4/10$	$2/10$	$1/10$	pmf $p(y)$	$1/2$	$1/2$

It is probably a little easier to see the different spreads in plots of the probability mass functions. We use bars instead of dots to give a better sense of the mass.



pmf's for two different distributions both with mean 0

Figure 5

Variance and Standard Deviation

Taking the mean as the center of a random variable's probability distribution, the variance is a measure of how much the probability mass is spread out around this center.

Definition: If X is a random variable with mean $E(X) = \mu$, then the variance of X is defined by

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right)$$

The **standard deviation** σ of X is defined by

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

If the relevant random variable is clear from context, then the variance and standard deviation are often denoted by σ^2 and σ ("sigma"), just as the mean is μ ("mu").

What does this mean? First, let us rewrite the definition explicitly as a sum. If X takes values $x_1, x_2, \ldots, x_n$ with probability mass function, $p\left(x_i\right)$ then

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \textstyle\sum_{i=1}^{n} p\left(x_i\right)\left(x_i-\mu\right)^2.$$

In words, the formula for *Var(X)* says to take a weighted average of the squared distance to the mean. By squaring, we make sure we are averaging only non-negative values, so that the spread to the right of the mean will not cancel that to the left.

Note on units:

1. σ has the same units as
2. *Var(X)* has the same units as the square of X. So, if X is in meters, then *Var(X)* is in meters squared.

Because σ and X have the same units, the standard deviation is a natural measure of spread.

Variance and Standard Deviation of a Discrete Random Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-16

Check Your Understanding: Variance and Standard Deviation for Discrete Random Variables

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-25

Bernoulli

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-17

Bernoulli Distributions

Model: The Bernoulli distribution models one trial in an experiment that can result in either a success or failure. This is the most important distribution and is also the simplest. A random variable X has a Bernoulli distribution with parameter p if:

1. X takes the values 0 and 1.

2. $P(X = 1) = p$ and $P(X = 0) = 1 - p$

We will write X ~ Bernoulli (p) or Ber (p), which is read "X follows a Bernoulli distribution with parameter " or " is drawn from a Bernoulli distribution with parameter p."

A simple model for the Bernoulli distribution is to flip a coin with probability p of heads, with $X = 1$ on heads and $X = 0$ on tails. The general terminology is to say X is 1 on success and 0 on failure, with success and failure defined by the context.

Many decisions can be modeled as a binary choice, such as votes for or against a proposal. If p is the proportion of the voting population that favors the proposal, then the vote of a random individual is modeled by a Bernoulli (p).

Here are the table and graphs of the pmf for the Bernoulli (1/2) distribution and below that for the general Bernoulli (p) distribution.

value a:	0	1
pmf $p(a)$:	1/2	1/2
cdf $F(a)$:	1/2	1

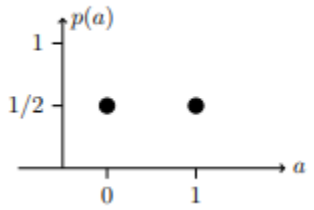


Figure 6: pmf for the Bernoulli(1/2) distribution

values a:	0	1
pmf $p(a)$:	1-p	p
cdf $F(a)$:	1-p	1

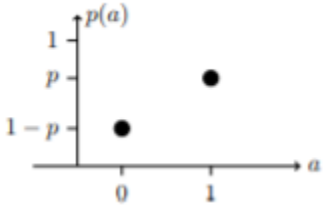


Figure 7: pmf for the Bernoulli(p) distribution

Mean and Variance of Bernoulli Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-18

Bernoulli Distribution Mean and Variance

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-19

The Variance of a Bernoulli Random Variable

Bernoulli random variables are fundamental, so we should know their variance.

If X~Bernoulli(p) then0

$$\mathrm{Var}(X) = p(1-p)$$

Proof: We know that $E(X) = p$. We compute $Var(X)$ using a table (Table 4).

Table 4

values X	**0**	**1**
pmf $p(x)$	$1-p$	p
$(X-\mu)^2$	$(0-p)^2$	$(1-p)^2$

$$\mathrm{Var}(X) = (1-p)p^2 + p(1-p)^2 = (1-p)p(1-p+p) = (1-p)p$$

Binomial Distributions

Binomial Variables

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-20

Check Your Understanding: Binomial Variables

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-26

Binomial Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-21

Visualizing a Binomial Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-22

Binomial Distributions

The binomial distribution Binomial (*n,p*), or Bin (*n,p*), models the number of successes in *n* independent Bernoulli(*p*) trials.

There is a hierarchy here. A single Bernoulli trial is, say, a toss of a coin. A single binomial trial consists of *n* Bernoulli trials. For coin flips the sample space for a Bernoulli trial is $\{H, T\}$. The sample space for a binomial trial is all *sequences* of heads and tails of length *n*. Likewise, a Bernoulli random variable takes values 0 and 1 and a binomial random variable takes values 0, 1, 2, …, *n*.

Example: The number of heads in *n* flips of a coin with probability *p* of heads follows a Binomial(*n,p*) distribution.

We describe *X*~Binomial(*n,p*) by giving its values and probabilities. For notation we will use *k* to mean an arbitrary number between 0 and *n*.

We remind you that '*n* choose *k*' $= \begin{pmatrix} n \\ k \end{pmatrix} = {}_nC_x$ is the number of ways to choose *k* things out of a collection of *n* things and it has the formula

$$\begin{pmatrix} n \\ x \end{pmatrix} = {}_nC_x = \frac{n!}{k!(n-k)!}$$

(It is also called a binomial coefficient). Table 5 is a table for the pmf of a Binomial(*n,k*) random variable:

Table 5

values *a*:	0	1	2	...	*k*	...	*n*
pmf *p*(*a*):	$(1-p)^n$	$\binom{n}{1}p^1(1-p)^{n-1}$	$\binom{n}{2}p^2(1-p)^{n-2}$	...	$\binom{n}{k}p^k(1-p)^{n-k}$	...	p^n

Example: What is the probability of 3 or more heads in 5 tosses of a fair coin?

Answer: The binomial coefficients associated with $n = 5$ are

$$\binom{5}{0} = 1, \quad \binom{5}{1} = \frac{5!}{1!4!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{4\cdot3\cdot2\cdot1} = 5, \binom{5}{2} = \frac{5!}{2!3!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{2\cdot1\cdot3\cdot2\cdot1} = \frac{5\cdot4}{2} = 10$$

And similarly

$$\binom{5}{3} = 10, \quad \binom{5}{4} = 5, \quad \binom{5}{5} = 1$$

Using these values, we get Table 6 for X~Binomial(5,*p*).

Table 6

values *a*:	0	1	2	3	4	5
pmf *p*(*a*):	$(1-p)^5$	$5p(1-p)^4$	$10p^2(1-p)^3$	$10p^3(1-p)^2$	$5p^4(1-p)$	p^5

We were told $p = 1/2$ so

$$P(X \geq 3) = 10\left(\tfrac{1}{2}\right)^3\left(\tfrac{1}{2}\right)^2 + 5\left(\tfrac{1}{2}\right)^4\left(\tfrac{1}{2}\right)^1 + \left(\tfrac{1}{2}\right)^5 = \frac{16}{32} = \frac{1}{2}$$

Binomial Probability Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-23

Explanation of the Binomial Probabilities

For concreteness, let $n = 5$ and $k = 2$ (the argument for arbitrary n and k is identical). So X~Binomial(5,p) and we want to compute $p(2)$. The long way to compute $p(2)$ is to list all the ways to get exactly 2 heads in 5-coin flips and add up their probabilities. The list has 10 entries: HHTTT, HTHTT, HTTHT, HTTTH, THHTT, THTHT, THTTH, TTHHT, TTHTH, TTTHH.

Each entry has the same probability of occurring, namely

$$p^2(1-p)^3$$

This is because each of the two heads has probability p and each of the 3 tails has

probability 1 – p. Because the individual tosses are independent, we can multiply probabilities. Therefore, the total probability of exactly 2 heads is the sum of 10 identical probabilities, i.e., $p(2) = 10p^2(1-p)^3$, as shown in Table 6.

This guides us to the shorter way to do the computation. We have to count the number of sequences with exactly 2 heads. To do this we need to choose 2 of the tosses to be heads and the remaining 3 to be tails. The number of such sequences is the number of ways to choose 2 out of the 5 things, which is $\begin{pmatrix} 5 \\ 2 \end{pmatrix}$. Since each such sequence has the same probability,$p^2(1-p)^3$, we get the probability of exactly 2 heads $p(2) = \begin{pmatrix} 5 \\ 2 \end{pmatrix} p^2(1-p)^3$.

Here are some binomial probability mass functions (here, frequency is the same as probability).

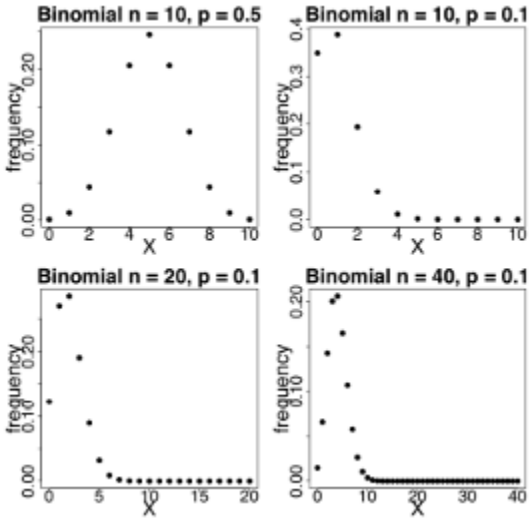


Figure 8

Check Your Understanding: Binomial Probability

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-27

*Characteristics of a **Binomial Experiment***

1. There are a fixed number of trials. Think of trials as repetitions of an

experiment. The letter n denotes the number of trials.

2. The random variable, x, number of successes, is discrete.
3. There are only two possible outcomes, called "success" and "failure," for each trial. The letter p denotes the probability of a success on any one trial, and q denotes the probability of a failure on any one trial. $p + q = 1$
4. The n trials are independent and are repeated using identical conditions. Think of this as drawing WITH replacement. Because the n trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, p, of a success and probability, q, of a failure remain the same. For example, randomly guessing at a true-false statistics question has only two outcomes. If a success is guessing correctly, then a failure is guessing incorrectly. Suppose Jow always guesses correctly on any statistics true-false question with a probability $p = 0.6$. Then, $q = 0.4$. This means that for every true-false statistics question Joe answers, his probability of success ($p = 0.6$) and his probability of failure ($q = 0.4$) remain the same.

Any experiment that has characteristics three and four and where $n = 1$ is called the Bernoulli Trial.

A word about independence

So far, we have been using the notion of an independent random variable without ever carefully defining it. For example, a binomial distribution is the sum of **independent** Bernoulli trials. This may (should?) have bothered you. Of course, we have an intuitive sense of what independence means for experimental trials. We also have the probabilistic sense that random variables X and Y are independent if knowing the value of X gives you no information about the value of Y.

Definition: The discrete random variables X and Y are independent if

$$P(X = a, Y = b) = P(X = a)P(Y = b)$$

For any values a,b. That is, the probabilities multiply.

Expected Value of a Binomial Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-24

Variance of a Binomial Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-25

Finding the Mean and Standard Deviation of a Binomial Random Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-26

Variance of Binomial (n,p)

Suppose X~Binomial(n,p). Since X is the sum of *independent* Bernoulli(p) variables and each Bernoulli variable has variance $p(1-p)$ we have

$$X \sim \text{Binomial}(n,p) \Rightarrow \text{Var}(X) = np(1-p)$$

Check Your Understanding: Expected Value of a Binomial Variable

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-28

Geometric Distributions

Geometric Random Variables Introduction

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-27

Check Your Understanding: Binomial vs. Geometric Random Variables

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-29

Geometric Distributions

A **geometric distribution** models the number of tails before the first head in a sequence of coin flips (Bernoulli trials).

Example: (a) Flip a coin repeatedly. Let X be the number of tails before the first heads. So, X can equal 0, i.e., the first flip is heads, 1, 2, …. In principle, it takes any nonnegative integer value.

(b) Give a flip of tails the value 0 and heads the value 1. In this case, X is the number of 0's before the first 1.

(c) Give a flip of tails the value 1 and heads the value 0. In this case, X is the number of 1's before the first 0.

(d) Call a flip of tails a success and heads a failure. So, X is the number of successes before the first failure.

(e) Call a flip of tails a failure and heads a success. So, X is the number of failures before the first success.

You can see this models many different scenarios of this type. The most neutral language is the number of tails before the first head.

Formal definition: The random variable X follows a geometric distribution with parameter p if

- X takes the values 0, 1, 2, 3, …
- Its pmf is given by $p(k) = P(X = k) = (1-p)^k p$

We denote this by $X \sim \text{geometric}(p)$ or $\text{geo}(p)$. In Table 7 we have:

Unchanged:

Table 7: $X \sim$ geometric (p) : $X =$ the number of 0 s before the first 1.

value	*a*:	*a*:	0	1	2	3	...	*k*	. . .
pmf	*p(a):*	*p*	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	...		$(1-p)^kp$	...

The geometric distribution is an example of a discrete distribution that takes an infinite number of possible values. Things can get confusing when we work with successes and failures since we might want to model the number of successes before the first failure, or we might want the number of failures before the first success. To keep things straight you can translate to the neutral language of the number of tails before the first heads.

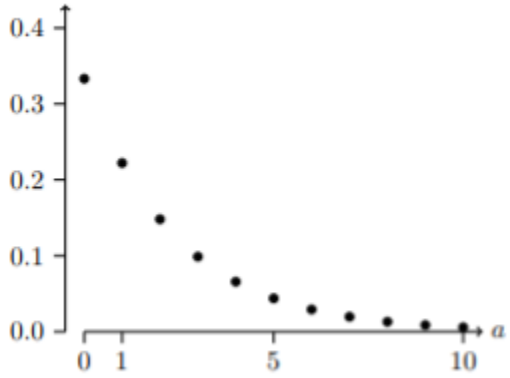


Figure 9: Pmf for the geometric(1/3) distribution

Probability for a Geometric Random Variable

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-28

Characteristics of a Geometric Experiment

The geometric probability density function builds upon what we have learned from the binomial distribution. In this case, the experiment continues until either a success or a failure occurs rather than for a set number of trials. Here are the main characteristics of a **geometric experiment**:

1. There are one or more Bernoulli trials with all failures except the last one, which is a success. In other words, you keep repeating what you are doing until the first success. Then you stop. For example, you throw a dart at a bullseye until you hit the bullseye. The first time you hit the bullseye is a "success," so you stop throwing the dart. It might take six tries until you

hit the bullseye. You can think of the trials as failure, failure, failure, failure, failure, success, STOP.

2. In theory, the number of trials could go on forever.
3. The probability, p, of a success and the probability, q, of a failure is the same for each trial. $p + q = 1$ and $q = 1 - p$. For example, the probability of rolling a three when you throw one fair die is $\frac{1}{6}$. This is true no matter how many times you roll the die. Suppose you want to know the probability of getting the first three on the fifth roll. On rolls one through four, you do not get a face with a three. The probability for each of the rolls is $q = \frac{5}{6}$, the probability of a failure. The probability of getting a three on the fifth roll is $\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) = 0.0804$
4. X = the number of independent trials until the first success.

Check Your Understanding: Probability for a Geometric Random Variable

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-30

Cumulative Geometric Probability (greater than a value)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-29

Cumulative Geometric Probability (less than a value)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-30

Check Your Understanding: Cumulative Geometric Probability

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-31

Poisson Distributions

Poisson Process 1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-31

Poisson Process 2

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-32

Poisson Distribution

Another useful probability distribution is the **Poisson distribution** or waiting time distribution. This distribution is used to determine how many checkout clerks are needed to keep the waiting time in line to specified levels, how many telephone lines are needed to keep the system from overloading, and many other practical applications. The distribution gets its name from Simeon Poisson who presented it in 1837 as an extension of the binomial distribution.

Here are the main characteristics of a Poisson experiment:

1. The **Poisson probability distribution** gives the probability of a number of events occurring in a fixed interval of time or space if these events happen with a known average rate.
2. The events are independent of the time since the last event. For example, a book

editor might be interested in the number of words spelled incorrectly in a particular book. It might be that, on the average, there are five words spelled incorrectly in 100 pages. The interval is the 100 pages, and it is assumed that there is no relationship between when misspellings occur.

3. The random variable X = the number of occurrences in the interval of interest.

Example: You notice that a news reporter says "uh," on average, two times per broadcast. What is the probability that the news reporter says "uh" more than two times per broadcast?

This is a Poisson problem because you are interested in knowing the number of times the news reporter says "uh" during a broadcast.

(a) What is the interval of interest?

a. one broadcast measured in minutes

(b) What is the average number of times the news reporter says "uh" during one broadcast?

a. 2

(c) Let X = _____. What values does X take on?

a. Let X = the number of times the news reporter says "uh" during one broadcast.
$x = 0, 1, 2, 3, \ldots$

d. The probability question is P(__).

a. $P(x > 2)$

Notation for the Poisson: P = Poisson Probability Distribution Function

$$X \sim P(\mu)$$

Read this as "X is a random variable with a Poisson distribution." The parameter is $\mu(\text{ or } \lambda); \mu(\text{ or } \lambda) =$ the mean for the interval of interest. The mean is the number of occurrences that occur on average during the interval period.

The formula for computing probabilities that are from a Poisson process is:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

where $P(X)$ is the probability of X successes, μ is the expected number of successes based upon historical data, e is the natural logarithm approximately equal to 2.718, and X is the number of successes per unit, usually per unit of time.

In order to use the Poisson distribution, certain assumptions must hold. These are: the probability of a success, μ, is unchanged within the interval, there cannot be simultaneous successes within the interval, and finally, that the probability of a success among intervals is independent, the same assumption of the binomial distribution.

In a way, the Poisson distribution can be thought of as a clever way to convert a continuous random variable, usually time, into a discrete random variable by breaking up time into discrete independent intervals. This way of thinking about the Poisson helps us understand why it can be used to estimate the probability for the discrete random variable from the binomial distribution. The Poisson is asking for the probability of a number of successes during a period of time while the binomial is asking for the probability of a certain number of successes for a given number of trials.

Example: Leah's answering machine receives about six telephone calls between 8 a.m. and 10 a.m. What is the probability that Leah receives more than one call in the next 15 minutes?

Let X = the number of calls Leah receives in 15 minutes (the interval of interest is 15 minutes or $\frac{1}{4}$ hour)

$x = 0, 1, 2, 3, \ldots$

If Leah receives, on average, six telephone calls in two hours, and there are eight 15 minutes intervals in two hours, then Leah receives $\left(\frac{1}{8}\right)(6) = 0.75$ calls in 15 minutes, on average. So, $\mu = 0.75$ for this problem.

$$X \sim P(0.75)$$

$$P(x > 1) = 0.1734$$

Probability that Leah receives more than one telephone call in the next 15 minutes is about 0.1734.

The graph of $X \sim P(0.75)$ is:

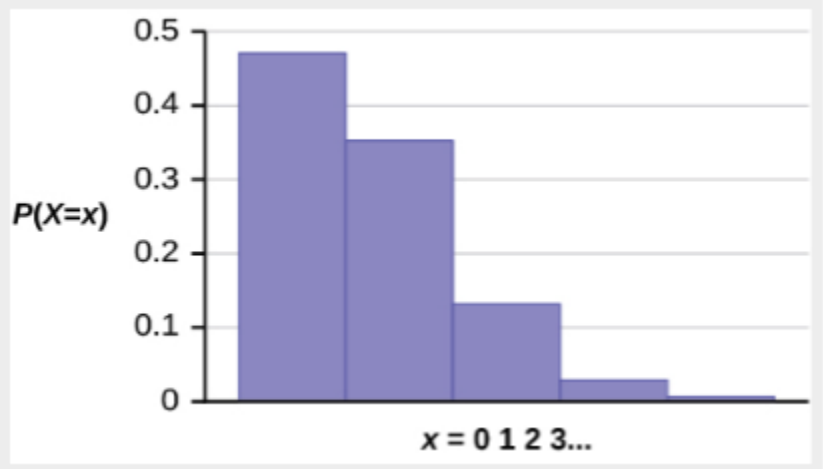


Figure 10

The y-axis contains the probability of x where X = the number of calls in 15 minutes.

Poisson Probability Distribution – Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-33

Check Your Understanding: Poisson Distributions

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-32

Tables of Distributions and Properties

Table 8

Distribution	Range	pmf	Mean	Variance
Bernoulli(p)	0,1	$p(0) = 1 - p, \quad p(1) = p$	p	$p(1-p)$
Binomial(n,p)	0, 1, … , n	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric(p)	0, 1, 2, …	$p(k) = p(1-p)^k$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

Let X be a discrete random variable with range $x_1, x_2, \ldots$ and pmf $p(x_j)$

Table 9

Expected Value:		**Variance:**
Synonyms:	mean, average	
Notation:	$E(X), \mu$	$\mathrm{Var}(X), \sigma^2$
Definition:	$E(X) = \sum_j p(x_j)\, x_j$	$E((X-\mu)^2) = \sum_j p(x_j)\,(x_j-\mu)^2$

SET UP AND WORK WITH DISTRIBUTIONS FOR CONTINUOUS RANDOM VARIABLES

This section will explain how to work with distributions for continuous random variables. It will start with some definition and a calculus warm-up to help prepare you. The videos help explain the concepts while there are problems after each short section to help check your understanding.

Continuous Random Variables

We now turn to continuous random variables. All random variables assign a number to each outcome in a sample space. Whereas discrete random variables take on a discrete set of possible values, continuous random variables have a continuous set of values.

Computationally, to go from discrete to continuous we simply replace sums by integrals. It will help you to keep in mind that (informally) an integral is just a continuous sum.

Calculus Warmup

Conceptually, you should be comfortable with two views of a definite integral.

1. $\int_a^b f(x)dx =$ area under the curve $y = f(x)$
2. $\int_a^b f(x)dx =$ 'sum of $f(x)dx$'

The connection between the two is seen below in Figure 11:

$$\text{area} \approx \text{sum of rectangle areas} = f(x_1)\Delta x + f(x_2)\Delta x + \ldots + f(x_n)\Delta x = \sum_1^n f(x_i)\Delta x.$$

Figure 11

As the width Δx of the intervals gets smaller the approximation becomes better.

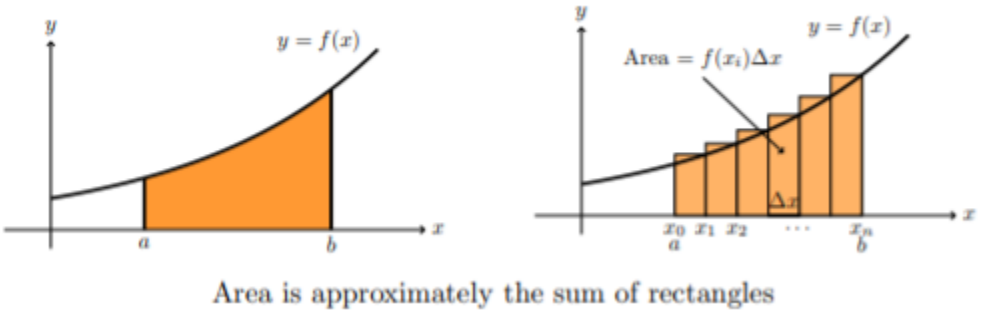


Figure 12

Note: In calculus, you learned to compute integrals by finding antiderivatives. This is important for calculations, but do not confuse this method for the reason we use integrals. Our interest in integrals comes primarily from its interpretation as a "sum" and to a much lesser extent its interpretation as area.

Probability Density Functions

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-34

Continuous Random Variables and Probability Density Functions

A continuous random variable takes a range of values, which may be finite in extent. Here are a few examples of ranges: $[0, 1]$, $[0, \infty)$, $(-\infty, \infty)$, $[a, b]$.

Definition: A random variable X is continuous if there is a function $f(x)$ such that for any $c \leq d$ we have

$$P(c \leq X \leq d) = \int_c^d f(x)dx$$

The function $f(x)$ is called the **probability density function (pdf).**

The pdf always satisfies the following properties:

1. $f(x) \geq 0$ (f is nonnegative).
2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (This is equivalent to: $P(-\infty < X < \infty) = 1$

The probability density function $f(x)$ of a continuous random variable is the analogue of the probability mass function $p(x)$ of a discrete random variable. Here are two significant differences:

1. Unlike $p(x)$, the pdf $f(x)$ is *not* a probability. You have to integrate it to get probability.

2. Since $f(x)$ is not a probability, there is no restriction that $f(x)$ be less than or equal to 1.

Note: In property 2, we integrated over $(-\infty, \infty)$ since we did not know the range of values taken by X. Formally, this makes sense because we just define $f(x)$ to be 0 outside of the range of X. In practice, we would integrate between bounds given by the range of X.

The Terms "Probability Mass" and "Probability Density"

Why do we use the terms mass and density to describe the pmf and pdf? What is the difference between the two? The simple answer is that these terms are completely analogous to the mass and density you saw in physics and calculus. We will review this first for the probability mass function and then discuss the probability density function. in extent.

Mass as a sum:

If masses $m_1, m_2, m_3,$ and m_4 and are set in a row at positions $x_1, x_2, x_3,$ and x_4, then the total mass is $m_1 + m_2 + m_3 + m_4$

Figure 13

We can define a 'mass function' $p(x)$ with $p(x_j) = m_j$ for $j = 1, 2, 3, 4$ and $p(x) = 0$ otherwise. In this notation the total mass is

$$p(x_1) + p(x_2) + p(x_3) + p(x_4)$$

The **probability mass function** behaves in exactly the same way, except it has the dimension of probability instead of mass.

Mass as an integral of density:

Suppose you have a rod of length L meters with varying density $f(x)$ kg/m. (Note the units are mass/length).

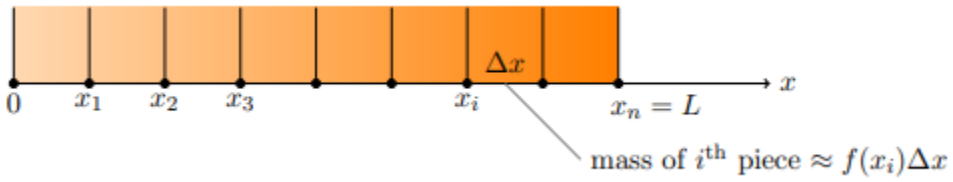


Figure 14

If the density varies continuously, we must find the total mass of the rod by integration: **total mass** = $\int_0^L f(x)dx$.

This formula comes from dividing the rod into small pieces and 'summing' up the mass of each piece. That is: $\text{total mass} \approx \sum_{i=1}^{n} f(x_i)\,\Delta x$

In the limit as Δx goes to zero the sum becomes the integral.

The **probability density function** behaves exactly the same way, except it has units of probability/ (unit *x*) instead of kg/m. Indeed, the equation $P(c \leq X \leq d) = \int_c^d f(x)dx$ is exactly analogous to the above integral for total mass.

While we are on a physics kick, note that for both discrete and continuous random variables, the expected value is simply the center of mass or balance point.

Properties of Continuous Probability Density Functions

The graph of a continuous probability distribution is a curve. Probability is represented by area under the curve. The curve is called the **probability density function** (pdf). We use the symbol *f* (*x*) to represent the curve. *f* (*x*) is the function that corresponds to the graph; we use the density function *f* (*x*) to draw the graph of the probability distribution.

Area under the curve is given by a different function called the **cumulative distribution function** (cdf). The cumulative distribution function is used to evaluate probability as area. Mathematically, the cumulative probability density function is the integral of the pdf, and the probability between two values of a continuous random variable will be the integral of the pdf between these two values: the area under the curve between these values. Remember that the area under the pdf for all possible values of the random variable is one, certainty. Probability thus can be seen as the relative percent of certainty between the two values of interest.

- The outcomes are measured, not counted.
- The entire area under the curve and above the x-axis is equal to one.
- Probability is found for intervals of x values rather than for individual x values.
- $P(c < x < d)$ is the probability that the random variable *X* is in the interval between the values *c* and *d*. $P(c < x < d)$ is the area under the curve, above the x-axis, to the right of *c* and the left of *d*.
- $P(x = c) = 0$ The probability that x takes on any single individual value is zero. The area below the curve, above the x-axis, and between *x* = *c* and *x* = *c* has no width, and therefore no area (area= 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$ is the same as $P(c \leq x \leq d)$ because probability is equal to area.

We will find the area that represents probability by using geometry, formulas, technology,

or probability tables. In general, integral calculus is needed to find the area under the curve for many probability density functions.

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way.

In this section, we will study the uniform distribution, the exponential distribution, and the normal distribution. The following graphs illustrate these distributions.

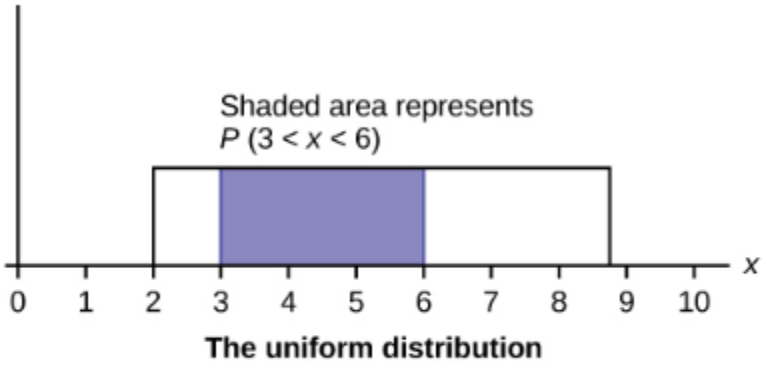


Figure 5.2 The graph shows a Uniform Distribution with the area between $x = 3$ and $x = 6$ shaded to represent the probability that the value of the random variable X is in the interval between three and six.

Figure 15

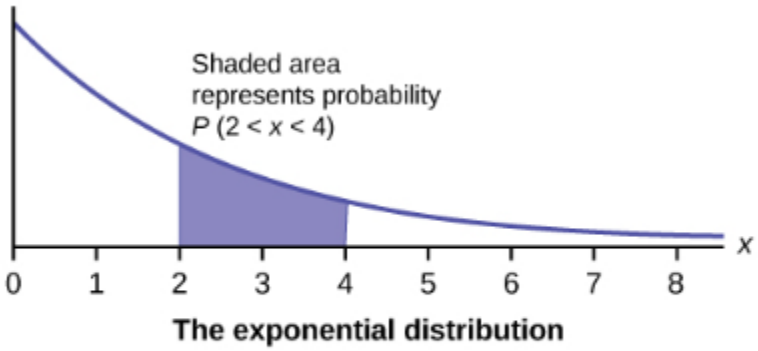


Figure 5.3 The graph shows an Exponential Distribution with the area between $x = 2$ and $x = 4$ shaded to represent the probability that the value of the random variable X is in the interval between two and four.

Figure 16

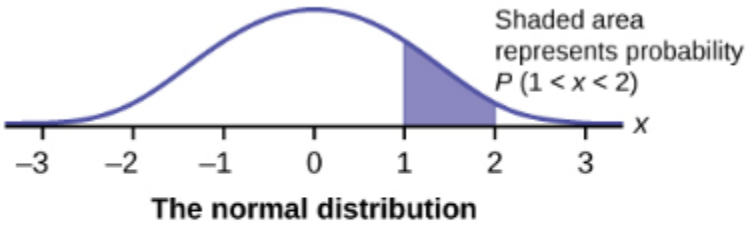


Figure 5.4 The graph shows the Standard Normal Distribution with the area between $x = 1$ and $x = 2$ shaded to represent the probability that the value of the random variable X is in the interval between one and two.

Figure 17

Cumulative Distribution Function

The cumulative distribution function (cdf) of a continuous random variable X is defined in exactly the same way as the cdf of a discrete random variable.

$$F(b) = P(X \leq b)$$

Note well that the definition is about probability. When using the cdf you should first think of it as a probability. Then when you go to calculate it, you can use

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x), \text{ where } f(x) \text{ is the pdf of } X$$

Notes:

1. For discrete random variables, there was not much occasion to use the cumulative distribution function. The cdf plays a far more prominent role for continuous random variables.
2. As before, we started the integral at $-\infty$ because we did not know the precise range of X. Formally, this still makes sense since $f(x) = 0$ outside the range of X. In practice, we will know the range and start the integral at the start of the range.
3. In practice, we often say "X has distribution $F(x)$" rather than "X has cumulative distribution function $F(x)$."

Check Your Understanding: Continuous Random Variables

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-33

Uniform Distributions

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-35

The Uniform Distribution

The **uniform distribution** is a continuous probability distribution and is concerned with events that are equally likely to occur. When working out problems that have a uniform distribution, be careful to not if the data is inclusive or exclusive of endpoints.

The mathematical statement of the uniform distribution is

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

Where a = the lowest value of x and b = the highest value of x.

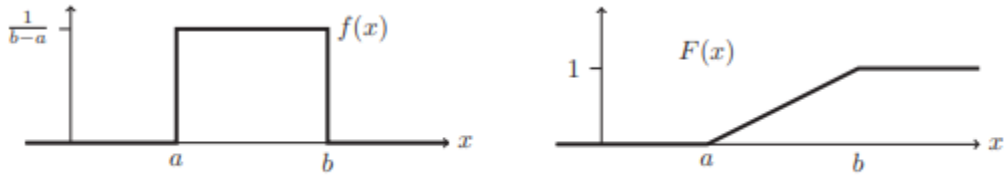
Formulas for the theoretical mean and standard deviation are

$$\mu = \frac{a+b}{2} \text{ and } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Uniform Distribution Properties

1. Parameters: a, b
2. Range: $[a, b]$
3. Notation: uniform (a, b) or $U(a, b)$
4. Density: $f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
5. Distribution: $F(x) = (x - a)/(b - a)$ for $a \leq x \leq b$
6. Models: All outcomes in the range have equal probability (more precisely all outcomes have the same probability density).

Graphs:



pdf and cdf for uniform(a,b) distribution.

Figure 18

Examples:

1. Suppose we have a tape measure with markings at each millimeter. If we measure (to the nearest marking) the length of items that are roughly a meter long, the rounding error will uniformly distributed between -0.5 and 0.5 millimeters.
2. Many boardgames use spinning arrows (spinners) to introduce randomness. When spun, the arrow stops at an angle that is uniformly distributed between 0 and 2π radians.
3. In most pseudo-random number generators, the basic generator simulates a uniform distribution, and all other distributions are constructed by transforming the basic generator.

Check Your Understanding: Uniform Distributions

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-34

Exponential Distributions

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-36

The Exponential Distribution

The **exponential distribution** is often concerned with the amount of time until some specific occurs. For example, the amount of time (beginning now) until an earthquake

occurs has an exponential distribution. Other examples include the length of time, in minutes, of long-distance business telephone calls, and the amount of time, in months, a car battery lasts. It can be shown, too, that the value of the change that you have in your pocket or purse approximately follows an exponential distribution.

Values for an exponential random variable occur in the following way. There are fewer large values and more small values. For example, marketing studies have shown that the amount of money customers spend in one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money.

Exponential distributions are commonly used in calculations of product reliability, or length of time a produce lasts.

The random variable for the exponential distribution is continuous and often measures a passage of time, although it can be used in other applications. Typical questions may be, "what is the probability that some event will occur within the next x hours or days, or what is the probability that some event will occur between x_1 hours and x_2 hours, or what is the probability that the event will take more than x_1 hours to perform?" In short, the random variable X equals (a) the time between events or (b) the passage of time to complete an action, e.g., wait on a customer. The probability density function is given by:

$$f(x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)x}$$

Where μ is the historical average waiting time and has a mean and standard deviation of $1/\mu$.

In order to calculate probabilities for specific probability density functions, the cumulative density function is used. The cumulative density function (cdf) is simply the integral of the pdf and is:

$$F(x) = \int_0^{\infty} \left[\frac{1}{\mu} e^{-\frac{x}{\mu}}\right] = 1 - e^{-\frac{x}{\mu}}$$

Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-37

Check Your Understanding: Exponential Probability Distribution

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-35

Integration by Parts – Exponential Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-38

Memorylessness of the Exponential Distribution

Say that the amount of time between customers for a postal clerk is exponentially distributed with a mean of two minutes. Suppose that five minutes have elapsed since the last customer arrived. Since an unusually long amount of time has now elapsed, it would seem to be more likely for a customer to arrive within the next minute. With the exponential distribution, this is not the case – the additional time spend waiting for the next customer does not depend on how much time has already elapsed since the last customer. This is referred to as the memoryless property. The exponential and geometric probability density functions are the only probability functions that have the memoryless**memoryless property.** property. Specifically, the memoryless property says that

$$P(X > r + t \mid X > r) = P(X > t) \text{ for all } r \geq 0 \text{ and } t \geq 0$$

For example, if five minutes have elapsed since the last customer arrives, then the probability that more than one minute will elapse before the next customer arrives is computed by using $r = 5$ and $t = 1$ in the foregoing equation.

$$P(X > 5 + 1 \mid X > 5) = P(X > 1) = e^{(-0.5)(1)} = 0.6065$$

This is the same probability as tat of waiting more than one minute for a customer to arrive after the previous arrival.

Check Your Understanding: Memorylessness of the Exponential Distribution

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-36

Relationship Between the Poisson and the Exponential Distribution

There is an interesting relationship between the exponential distribution and the Poisson distribution. Suppose that the time that elapses between two successive events follows the exponential distribution with a mean of μ units of time. Also assume that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit of time follows a Poisson distribution with mean μ. Recall that if X has the Poisson distribution with mean μ, then

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

The formula for the exponential distribution: $P(X = x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)(x)}$. Where $\mu =$ average time between occurrences.

We see that the exponential is the cousin of the Poisson distribution, and they are linked through this formula. There are important differences that make each distribution relevant for different types of probability problems.

First, the Poisson has a discrete random variable, x =, where time; a **continuous variable** is artificially broken into discrete pieces. We saw that the number of occurrences of an event in a given time interval, x, follows the Poisson distribution.

For example, the **number** of times the telephone rings per hour. By contrast, the time **between** occurrences follows the exponential distribution. For example, the telephone just rang, how long will it be until it rings again? We are measuring length of time of the interval, a continuous random variable, exponential, not events during an interval, Poisson.

The Exponential Distribution v. the Poisson Distribution

A visual way to show both the similarities and differences between these two distributions is with a timeline.

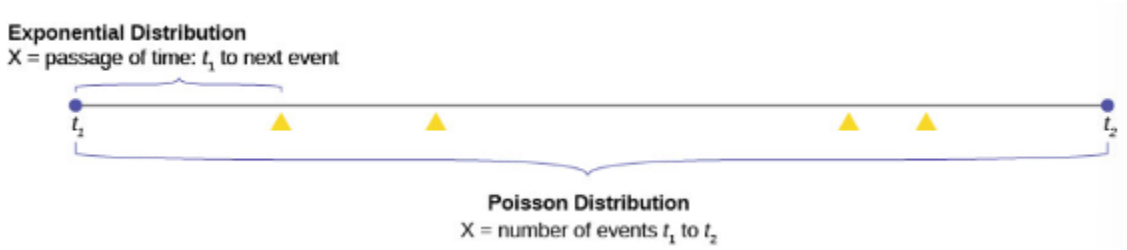


Figure 19

The random variable for the Poisson distribution is discrete and thus counts events during a given time period, t_1 to t_2 on Figure 19 above and calculates the probability of that number occurring. The number of events, four in the graph is measured in counting numbers; therefore, the random variable of the Poisson is a discrete random variable.

The exponential probability distribution calculates probabilities of the passage of time, a continuous random variable. In Figure 19, this is shown as the bracket from t_1 to the next occurrence of the event marked with a triangle.

Classic Poisson distribution questions are "how many people will arrive at my checkout window in the next hour?"

Classic exponential distribution questions are "how long it will be until the next person arrives," or a variant, "how long will the person remain here once they have arrived?"

Again, the formula for the exponential distribution is:

$$f(x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)(x)}$$

We see immediately the similarity between the exponential formula and the Poisson formula.

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Both probability density functions are based upon the relationship between time and exponential growth or decay. The "e" in the formula is a constant with the approximate value of 2.71828 and is the base of the natural logarithmic exponential growth formula. When people say that something has grown exponentially this is what they are talking about.

An example of the exponential and the Poisson will make clear the differences been the two. It will also show the interesting applications they have.

Poisson Distribution

Suppose that historically 10 customers arrive at the checkout lines each hour. Remember that this is still probability, so we have to be told these historical values. We see this is a Poisson probability problem.

We can put this information into the Poisson probability density function and get a general formula that will calculate the probability of any specific number of customers arriving in the next hour.

The formula is for any value of the random variable we chose, and so the x is put into the formula. This is the formula:

$$f(x) = \frac{10^x e^{-10}}{x!}$$

As an example, the probability of 15 people arriving at the checkout counter in the next hour would be

$$P(x = 15) = \frac{10^{15} e^{-10}}{15!} = 0.0611$$

Here we have inserted x = 15 and calculated the probability that in the next hour 15 people will arrive is .061.

Exponential Distribution

If we keep the same historical facts that 10 customers arrive each hour, but we now are interested in the service time a person spends at the counter, then we would use the exponential distribution. The exponential probability function for any value of x, the random variable, for this particular checkout counter historical data is:

$$f(x) = \frac{1}{.1} e^{-\frac{x}{1}} = 10e^{-10x}$$

To calculate μ, the historical average service time, we simply divide the number of people that arrive per hour, 10, into the time period, one hour, and have $\mu = 0.1$. Historically, people spend 0.1 of an hour at the checkout counter, or 6 minutes. This explains the .1 in the formula.

There is a natural confusion with μ in both the Poisson and exponential formulas. They have different meanings, although they have the same symbol. The mean of the exponential is one divided by the mean of the Poisson. If you are given the historical number of arrivals, you have the mean of the Poisson. If you are given an historical length of time between events, you have the mean of an exponential.

Continuing with our example at the checkout clerk; if we wanted to know the probability that a person would spend 9 minutes or less checking out, then we use this formula. First, we convert to the same time units which are parts of one hour. Nine minutes is 0.15 of one hour. Next, we note that we are asking for a range of values. This is always the case for a continuous random variable. We write the probability question as:

$$p(x \leq 9) = 1 - 10e^{-10x}$$

We can now put the numbers into the formula, and we have our result

$$p(x = .15) = 1 - 10e^{(-10)(.15)} = 0.7769$$

The probability that a customer will spend 9 minutes or less checking out is 0.7769.

We see that we have a high probability of getting out in less than nine minutes and a tiny probability of having 15 customers arriving in the next hour.

Exponential Distribution Properties

1. Parameter: $\lambda, \left(\frac{1}{\mu}\right)$
2. Range: $[0, \infty)$
3. Notation: exponential (λ) or $\exp(\lambda)$
4. Density: $f(x) = \lambda e^{-\lambda x}$ or $f(x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)(x)}$ for $0 \leq x$
5. Distribution: *(easy integral)* $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$
6. *Right tail distribution:* $P(X > x) = 1 - F(x) = e^{-\lambda x}$
7. Models: The waiting time for a continuous process to change state.

Examples:

1. If I step out to 77 Mass Ave after class and wait for the next taxi, m waiting time in minutes is exponentially distributed. We will see that in this case λ is given by one over the average number of taxis that pass per minute (on weekday afternoons).
2. The exponential distribution models the waiting time until an unstable isotope undergoes nuclear decay. In this case, the value of λ is related to the half-life of the isotope.

Graphs:

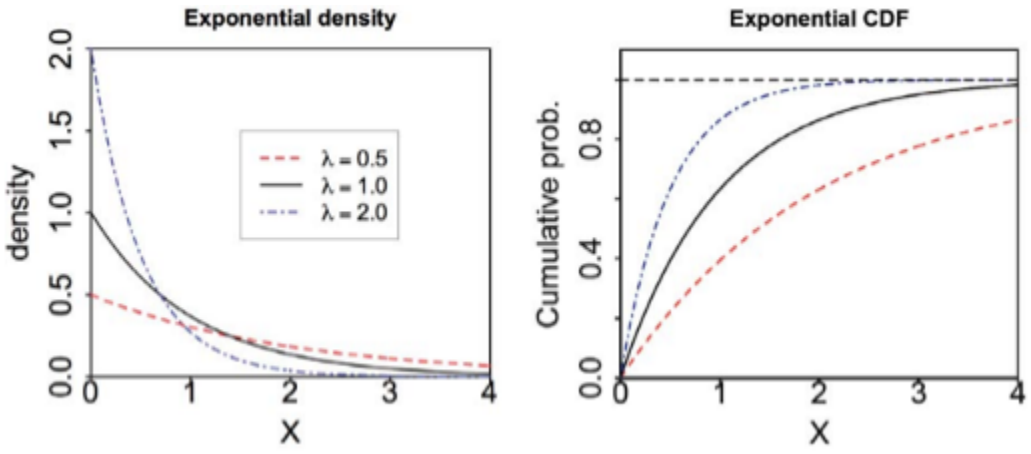


Figure 20

The Normal Distribution

The normal probability density function, a continuous distribution, is the most important of

all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

The normal distribution has two parameters (two numerical descriptive measures): the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing the following formula of the normal probability density function:

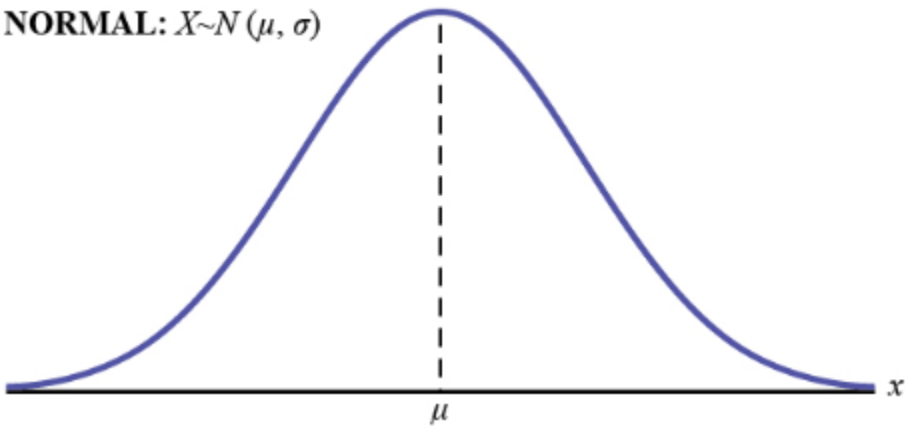


Figure 21

The probability density function is a rather complicated function. Do not memorize it. It is not necessary.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The curve is symmetric about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Note that this is unlike several probability density functions we have already studies, such as the Poisson, where the mean is equal to μ and the standard deviation simply the square root of the mean, or the binomial, where p is used to determine both the mean and standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the normal curve; the curve becomes fatter and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of the special interests is called the **standard normal distribution.**

The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z—scores. A z-score is measure in units of the standard deviation.**

The mean for the standard normal distribution is zero, and the standard deviation is one. What this does is dramatically simplify the mathematical calculation of probabilities. Take a moment and substitute zero and one in the appropriate places in the above formula and you can see that the equation collapses into one that can be much more easily solved using integral calculus. The transformation $z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0,1)$. The value x in the given equation comes from a known normal distribution with known mean μ and known standard deviation σ. The z-score tells how many standard deviations a particular *x* is away from the mean.

Z-Scores

If *X* is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score for a particular *x* is:

$z = \frac{x-\mu}{\sigma}$

The z-score tells you how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean, μ. Values of *x* that are larger than the mean have positive z-scores, and values of *x* that are smaller than the mean have negative z-scores. If *x* equals the mean, then *x* has a z-score of zero.

Example: Suppose $X \sim N(5, 6)$. This says that *X* is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then:

$z = \frac{x-\mu}{\sigma} = \frac{17-5}{6} = 2$

This means that $x = 17$ is **two standard deviations** (2σ) above or to the right of the mean $\mu = 5$.

Now suppose *x* = 1. Then: $z = \frac{x-\mu}{\sigma} = \frac{1-5}{6} = -0.67$ (rounded to two decimal places).

This means that *x* = 1 is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$.

The Empirical Rule

If *X* is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** states the following:

- About 68% of the *x* values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95% of the *x* values lie between -2σ and $+2\sigma$ of the mean μ(within two standard deviations of the mean).

- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are $+1$ and -1 respectively
- The z-scores for $+2\sigma$ and -2σ are $+2$ and -2 respectively.
- The z-scores for $+3\sigma$ and -3σ are $+3$ and -3 respectively.

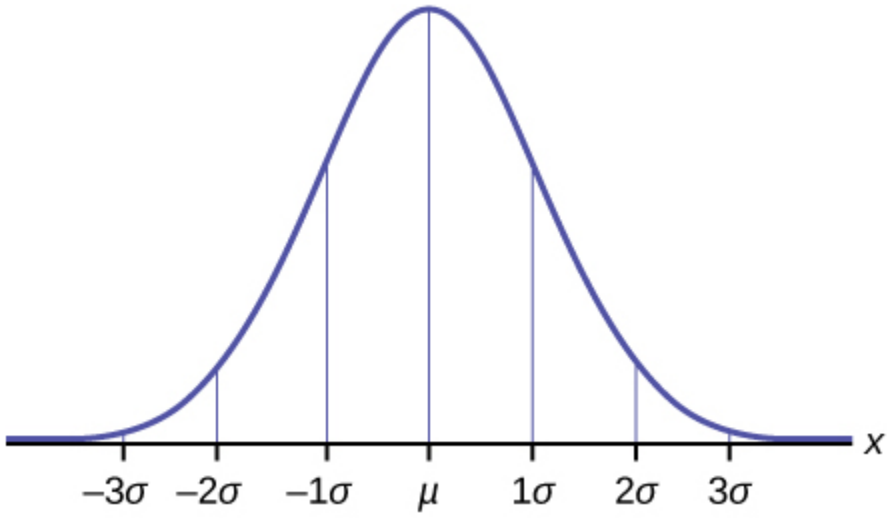


Figure 22

Normal Distribution Problems: Empirical Rule

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-39

Check Your Understanding: Normal Distribution Problems

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-37

Normal Distribution Properties

In 1809, Carl Friedrich Gauss published a monograph introducing several notions that have become fundamental to statistics: the normal distribution, maximum likelihood estimation, and the method of least squares. For this reason, the normal distribution is also called the *Gaussian* distribution, and it is the most important continuous distribution.

1. Parameters: μ, σ
2. Range: $(-\infty, \infty)$
3. Notation: normal $\left(\mu, \sigma^2\right)$ or $N\left(\mu, \sigma^2\right)$
4. Density: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$
5. Distribution: $F(x)$ has no formula, so use tables or software such as "pnorm" in R to compute F(x). (Will be discussed later in the module)
6. Models: Measurement error, intelligence/ability, height, averages of lots of data.

Here are some graphs of normal distributions. Note they are shaped like a *bell curve.* Note also that as σ increases as they become more spread out.

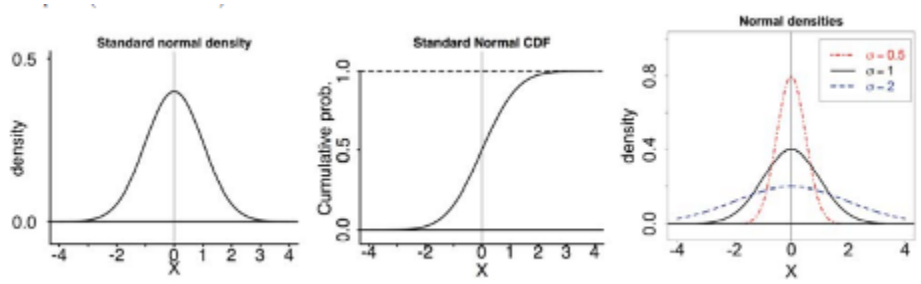


Figure 23

Using the Normal Distribution

The shaded area in Figure 24 indicates the area to the right of *x*. This area is represented by the probability $P(X > x)$. Normal tables provide the probability between the mean, zero for the standard normal distribution, and a specific value such as x_1. This is the unshaded part of the graph from the mean to x_1.

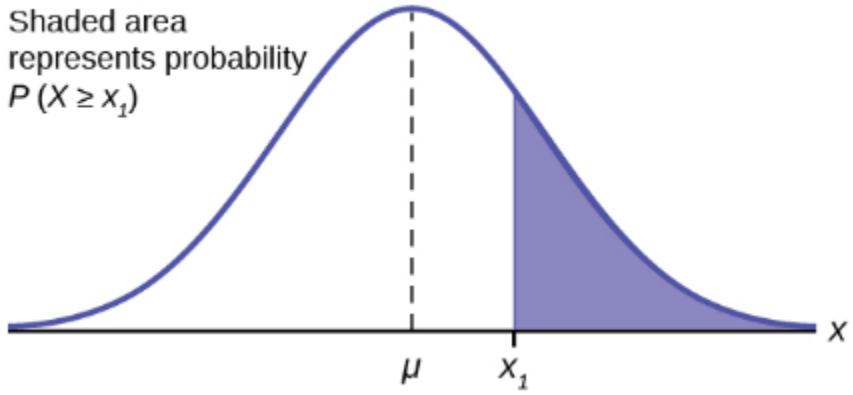


Figure 24

Because the normal distribution is symmetrical, if x_1 were the same distance to the left of the mean the area, probability, in the left tail, would be the same as the shaded area in the right tail. Also, bear in mind that because of the symmetry of this distribution, one-half of the probability is to the right of the mean and one-half is to the left of the mean.

Calculations of Probabilities

To find the probability for probability density functions with a continuous random variable we need to calculate the area under the function across the values of *X* we are interested in. For the normal distribution this seems a difficult task given the complexity of the formula. There is, however, a simple way to get what we want. Here again is the formula for the normal distribution:

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

Looking at the formula for the normal distribution it is not clear just how we are going to solve for the probability doing it the same way we did it with the previous probability functions. There we put the data into the formula and did the math.

To solve this puzzle, we start knowing that the area under a probability density function is the probability.

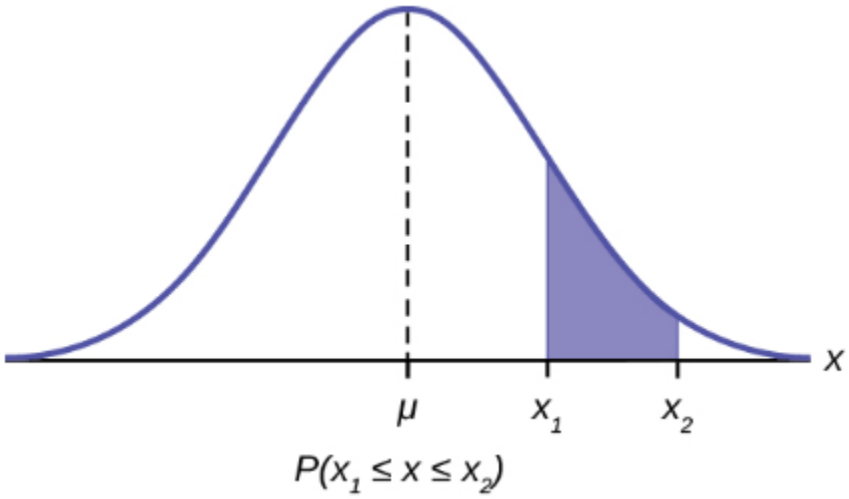


Figure 25: Normal distribution with a section shaded between x_1 and x_2.

This shows that the area between X_1 and X_2 is the probability as stated in the formula: $P\left(X_1 \leq x \leq X_2\right)$

The mathematical tool needed to find the area under a curve is integral calculus. The integral of the normal probability density function between the two points x_1 and x_2 is the area under the curve between these two points and is the probability between these two points.

Doing these integrals is no fun and can be very time consuming. But now, remembering that there are an infinite number of normal distributions out there, we can consider the one with a mean of zero and a standard deviation of 1. This particular normal distribution is given the name Standard Normal Distribution. Putting these values into the formula it reduces to a very simple equation. We can now quite easily calculate all probabilities for any value of x, for this particular normal distribution, that has a mean of zero and a standard deviation of 1. They are presented in numerous ways. The table is the most common presentation and is set up with probabilities for one-half the distribution beginning with zero, the mean, and moving outward. The shaded area in the graph at the top of the table in Statistical Tables represents the probability from zero to the specific Z value noted on the horizontal axis, Z.

The only problem is that even with this table, it would be a ridiculous coincidence that our data had a mean of zero and a standard deviation of one. The solution is to convert the distribution we have with its mean and standard deviation to this new Standard Normal Distribution. The Standard Normal has a random variable called Z.

Using the standard normal table, typically called the normal table, to find the probability of one standard deviation, go to the Z column, reading down to 1.0 and then read at column 0. That number, 0.3413 is the probability from zero to 1 standard deviation. At the top of the table is the shaded area in the distribution which is the probability for one standard deviation. The table has solved our integral calculus problem. But only if our data has a mean of zero and a standard deviation of 1.

However, the essential point here is, the probability for one standard deviation on one normal distribution is the same on every normal distribution. If the population data set has a mean of 10 and a standard deviation of 5 then the probability from 10 to 15, one standard deviation, is the same as from zero to 1, one standard deviation on the standard normal distribution. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the standardizing formula:

$$Z = \frac{x-\mu}{\sigma}$$

where Z is the value on the standard normal distribution, X is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability, so we always are dealing with the population, with known parameter values and a known distribution. It is also important to note that because the normal distribution is symmetrical it does not matter if the z-score is positive or negative when calculating a probability. One standard deviation to the left (negative Z-score) covers the same area as one standard deviation to the right (positive Z-score). This fact is why the Standard Normal tables do not provide areas for the left side of the distribution. Because of this symmetry, the Z-score formula is sometimes written as:

$$Z = \frac{|x-\mu|}{\sigma}$$

Where the vertical lines in the equation means the absolute value of the number. What the standardizing formula is really doing is computing the number of standard deviations X is from the mean of its own distribution.

Check Your Understanding: Using the Normal Distribution

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-38

Probabilities from Density Curves

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-40

Check Your Understanding: Probabilities from Density Curves

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-39

Standard Normal Table for Proportion Below

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-41

Standard Normal Table for Proportion Above

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-42

Check Your Understanding: Standard Normal Table for Proportion Below and Above

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-40

Standard Normal Table for Proportion Between Values

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-43

Check Your Understanding: Standard Normal Table for Proportion Between Values

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-41

Finding z-score for a Percentile

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-44

Check Your Understanding: Finding z-score for a Percentile

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#h5p-42

Expectation, Variance and Standard Deviation for Continuous Random Variables

So far, we have looked at expected value, standard deviation, and variance for discrete random variables. These summary statistics have the same meaning for continuous random variables:

- The expected value $\mu = E(X)$ is a measure of location or central tendency.
- The standard deviation σ is a measure of the spread or scale.
- The variance $\sigma^2 = \mathrm{Var}(X)$ is the square of the standard deviation.

To move from discrete to continuous, we will simply replace the sums in the formulas by integrals.

Expected Value of a Continuous Random Variable

Definition: Let X be a continuous random variable with range $[a, b]$ and probability density function $f(x)$. The expected value of X is defined by

$$E(X) = \int_a^b x f(x) dx$$

Let us see how this compares with the formula for a discrete random variable:

$$E(X) = \sum_{i=1}^{n} x_i p\left(x_i\right)$$

The discrete formula says to take a weighted sum of the values x_i of X, where the weights are the probabilities $p\left(x_i\right)$. Recall that $f(x)$ is a probability density. Its units are prob/(unit of X).

So, $f(x)dx$ represents the probability that X is in an infinitesimal range of width dx around x. Thus, we can interpret the formula for $E(X)$ as a weighted integral of the values x of X, where the weights are the probabilities $f(x)dx$.

As before, the expected value is also called the mean or average.

Example 1: Let $X \sim$ uniform $(0, 1)$. Find $E(X)$.

Answer: X has range [0,1] and density $f(x) = 1$. Therefore,

$$E(X) = \int_0^1 x dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$$

Not surprisingly the mean is at the midpoint of the range.

Example 2: Let X have a range $[0, 2]$ and density $\frac{3}{8}x^2$. Find $E(X)$..

Answer: $E(X) = \int_0^2 x f(x) dx = \int_0^2 \frac{3}{8}x^2 dx = \frac{3x^4}{32}\Big|_0^2 = \frac{3}{2}$

Does it make sense that this X has mean in the right half of its range?

Yes. Since the probability density increases as x increases over the range, the average value of x should be in the right half of the range.

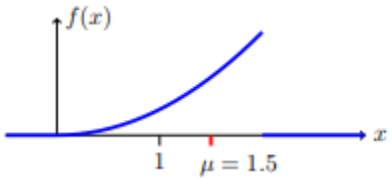


Figure 26

μ is "pulled" to the right of the midpoint 1 because there is more mass to the right.

Example 3: Let $X \sim \exp(\lambda)$. Find $E(X)$.

Answer: The range of X is $[0, \infty)$ and its pdf is $f(x) = \lambda e^{-\lambda x}$.

$$E(X) = \int_0^\infty \lambda e^{-\lambda x} dx = -\lambda e^{-\lambda x} - \frac{e^{-\lambda x}}{\lambda}\Big|_0^\infty = \frac{1}{\lambda}$$

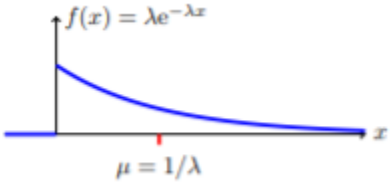


Figure 27: Mean of an exponential random variable

Example 4: Let $Z \sim N(0, 1)$. Find $E(Z)$.

Answer: The range of Z is $(-\infty, \infty)$ and its pdf is $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

$$E(Z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-\frac{z^2}{2}} dz = -\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\Big|_{-\infty}^{\infty} = 0$$

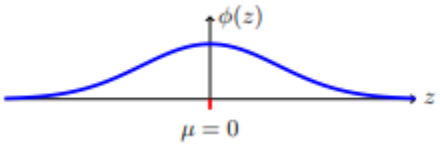


Figure 28: The standard normal distribution is symmetric and has mean 0.

Properties of E(X)

The properties of $E(X)$ for continuous random variables are the same as for discrete ones:

1. If X and Y are random variables on a sample space, Ω then $E(X+Y) = E(X) + E(Y)$
2. If a and b are constants, then $E(aX+B) = aE(X) + b$

Example 5: In this example we verify that for $X \sim N(\mu, \sigma)$ we have $E(X) = \mu$.

Answer: Example (4) showed that for standard normal Z, $E(Z) = 0$. We could mimic the calculation there to show that $E(X) = \mu$. Instead, we will use the linearity properties of $E(X)$. $X \sim N\left(\mu, \sigma^2\right)$ is a normal random variable we can standardize it:

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Inverting this formula, we have $X = \sigma Z + \mu$. The linearity of expected value now gives

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu$$

Expectation of Functions of *X*

This works exactly the same as the discrete case. If $h(x)$ is a function, then $Y = h(X)$ is a random variable and

$$E(Y) = E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

Example 6: Let $X \sim \exp(\lambda)$. Find $E\left(X^2\right)$

Answer: Using integration by parts we have

$$E\left(X^2\right) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x} - \frac{2x}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right]_0^{\infty}$$
$$= \frac{2}{\lambda^2}$$

Variance

Now that we have defined expectation for continuous random variable, the definition of **variance** is identical to that of discrete random variables.

Definition: Let X be a continuous random variable with mean μ. The variance of X is

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right)$$

Properties of Variance

These are exactly the same as in the discrete case.

1. If X and Y are independent, then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
2. For constants a and b, $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$.
3. Theorem: $\mathrm{Var}(X) = E\left(X^2\right) - E(X)^2 = E\left(X^2\right) - \mu^2$

For Property 1, note carefully the requirement that X and Y are independent.

Property 3 gives a formula for $\mathrm{Var}(X)$ that is often easier to use in hand calculations.

Example 7: Let $X \sim$ uniform $(0,1)$. Find $\mathrm{Var}(X)$ and σ_X

Answer: In Example 1 we found $\mu = \frac{1}{2}$. Next, we compute

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \int_0^1 \left(x-\tfrac{1}{2}\right)^2 dx = \tfrac{1}{12}$$

Example 8: Let $X \sim \exp(\lambda)$. Find $\mathrm{Var}(X)$ and σ_X

Answer: In Examples 3 and 6 we computed

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x}dx = \tfrac{1}{\lambda} \text{ and } E\left(X^2\right) = \int_0^\infty x^2\lambda e^{-\lambda x} = \tfrac{2}{\lambda^2}$$

So, by Property 3,

$$\mathrm{Var}(X) = E\left(X^2\right) - E(X)^2 = \tfrac{2}{\lambda^2} - \tfrac{1}{\lambda^2} = \tfrac{1}{\lambda^2} \text{ and } \sigma_X = \tfrac{1}{\lambda}$$

We could have skipped Property 3 and computed this directly from

$$\mathrm{Var}(X) = \int_0^\infty \left(x-\tfrac{1}{\lambda}\right)^2 \lambda e^{-\lambda x}dx$$

Example 9: Let $Z \sim N(0,1)$. Show $\mathrm{Var}(Z) = 1$

Note: The notation for normal variables is $X \sim N\left(\mu, \sigma^2\right)$. This is certainly suggestive, but as mathematicians we need to prove that $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Above we showed $E(X) = \mu$. This example shows that $\mathrm{Var}(Z) = 1$, just as the notation suggests. In the next example we will show $\mathrm{Var}(X) = \sigma^2$

Answer: Since $E(Z) = 0$, we have

$$\mathrm{Var}(Z) = E\left(Z^2\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz$$

(Using integration by parts with $u = z, v' = ze^{-\frac{z^2}{2}}$

$$\Rightarrow u' = 1, v = -e^{-\frac{z^2}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \left(-ze^{-\frac{z^2}{2}} \mid \begin{matrix} \infty \\ -\infty \end{matrix} \right) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz$$

The first term equals 0 because the exponential goes to zero much faster than *z* grows at both $\pm\infty$. The second term equals 1 because it is exactly the total probability integral of the pdf $\varphi(z)$ for $N(0,1)$. So $\mathrm{Var}(X) = 1$.

Example 10: Let $X \sim N\left(\mu, \sigma^2\right)$. Show $\mathrm{Var}(X) = \sigma^2$

Answer: This is an exercise in change of variables. Letting $z = (x - \mu)/\sigma$, we have

$$\mathrm{Var}(X) = E\left((X-\mu)^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx$$
$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz = \sigma^2$$

The integral in the last line is the same one we computed for $\mathrm{Var}(Z)$.

Quantiles

Definition: The median of *X* is the value *x* for which $P(X \le x) = 0.5$, i.e. the value of *x* such that $P(X \le X) = P(X \ge x)$. In other words, *X* has equal probability of being above or below the median, and each probability is therefore ½. In terms of the cdf $F(x) = P(X \le x)$, we can equivalently define the median as the value *x* satisfying $F(x) = 0.5$.

Think: What is the median of *Z*?

Answer: By symmetry, the median is 0.

Example 11: Find the median of $X \sim \exp(\lambda)$.

Answer: The cdf of X is $F(x) = 1 - e^{-\lambda x}$. So, the median is the value of *x* for which $F(x) = 1 - e^{-\lambda x} = 0.5$. Solving for *x* we find: $x = (\ln 2)/\lambda$.

Think: In this case the median does not equal the mean of $\mu = 1/\lambda$. Based on the graph of the pdf of *X* can you argue why the median is to the left of the mean.

Definition: The p^{th} quantile of X is the value q_p such that $P(X \leq q_p) = p$

Notes:

1. In this notation the median is $q_{0.5}$
2. We will usually write this in terms of the cdf: $F(q_p) = p$

With respect to the pdf $f(x)$, the quantile q_p is the value such that there is an area of p to the left of q_p and an area of $1 - p$ to the right of q_p. In the examples below, note how we can represent the quartile graphically using either area of the pdf or height of the cdf.

Example 12: Find the 0.6 quantile for $X \sim U(0, 1)$.

Answer: The cdf for X is $F(x) = x$ on the range $[0, 1]$. So $q_{0.6} = 0.6$

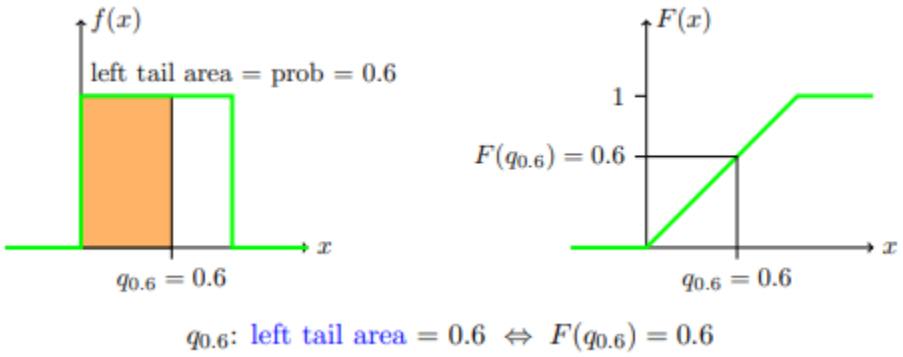


Figure 29

Quantile give a useful measure of location for a random variable.

Percentiles, decile, quartiles

For convenience, quantiles are often described in terms of percentile, decile, or quantiles. The 60th percentile is the same as the 0.6 quantile. For example, you are in the 60th percentile for height if you are taller than 60 percent of the population, i.e., the probability that you are taller than a randomly chosen person is 60 percent.

Likewise, deciles represent steps of 1/10. The third decile is the 0.3 quantile. **Quartiles** are in steps of ¼. Third quartile is the 0.75 quantile and the 75th percentile.

RELEVANCE TO TRANSPORTATION ENGINEERING COURSEWORK

This section will relate the fundamentals of transportation engineering by discussion and showing you queuing theory and models to help your understanding.

Queuing Theory: Exponential Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=271#oembed-45

Poisson Model of Vehicle Arrivals

The Poisson distribution is used to represent arrival patterns at a roadway location to conduct traffic queue and delay analysis. The distribution helps estimate the likelihood of n vehicles arriving in a specific time frame and the average arrival rate using data collected from roadway locations. For more information on the Poisson distribution, the reader is referred to in the above section labeled "Poisson Distributions."

Distribution of Time-Mean Speed

Spot speed studies during relatively congestion-free durations are conducted to estimate mean speed, modal speed, pace, standard deviation, and different percentile of speeds at a roadway location. 85th percentile speed estimated from a spot speed study is one of the crucial measures for determining the appropriate speed limit for the roadway section. Speed data for automobiles on roadway locations collected as part of the spot speed study typically follow the normal distribution discussed in "The Normal Distribution" section above.

Traffic Safety

Traffic safety analysis involves statistical methods that are used to process records of historical collisions or crash data. The negative binomial distribution is used to describe crash occurrences on roadway segments and intersections. The negative binomial model's formulation is a variant of Poisson's distribution (see "Poisson Distributions" above) that allows the standard deviation of the sample to exceed the mean.

Key Takeaways

- Probability distribution functions (PDFs) discussed in the chapter are used in several transportation engineering modeling applications. For example, arrival patterns at a roadway location critical to queue analysis follow a Poisson's distribution while the time interval between arrivals follows a negative binomial distribution.
- Negative binomial distribution and its variants are also used to model collision frequency on road segments and intersections. These models, known as Safety Performance Functions, are used to estimate expected crash frequency on roadway entities.
- Data from spot speed studies are used to ensure that the roadway design context aligns with the posted speed limit. The speeds on the roadway segments with no congestion follow a normal

distribution, and the data are used to estimate the mean, median, and 85th percentile of the normal distribution.

GLOSSARY: KEY TERMS

Bayes' Theorem[2] – a useful tool for calculating conditional probability. Bayes'**Bernoulli trials.** theorem can be expressed as: ◈1, ◈2,…, ◈◈ be a set of mutually exclusive events that together form the sample space S. Let B be any event from the same sample space, such that ◈(◈)>0. Then, $P(A \mid B) = \frac{P(A)P(B|A)}{P(B)}$

Bernoulli Trials[1] – an experiment with the following characteristics: (1) There are only two possible outcomes called "success" and "failure" for each trial. (2) The probability p of a success is the same for any trial (so the probability ◈=1–◈ of a failure is the same for any trial).

Binomial Experiment[1] – a statistical experiment that satisfies the following three conditions: (1) There are a fixed number of trials, n. (2) There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes, the probability of a failure on one trial. (3) The n trials are independent and are repeated using identical conditions.

Binomial Probability Distribution[1] – a discrete random variable that arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial random variable X is defined as the number of successes in n trials. The mean is ◈=◈◈ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$

Combination[2]– a combination is a selection of all or part of a set of objects, without regard to the order in which objects are selected

Conditional Probability[1] – the likelihood that an event will occur given that another event has already occurred.

Discrete Variable / continuous variable[2]– if a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise it is called a discrete variable

Expected Value[2] – the mean of the discrete random variable X is also called the expected value of X. Notationally, the expected value of X is denoted by ◈(◈).

Exponential Distribution[1] – a continuous random variable that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)x}, x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{\left(-\frac{1}{\mu}\right)x}$

Geometric Distribution[1] – a discrete random variable that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials

until the first success. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: ◈(◈=◈)=◈(1–◈)$^{◈-1}$ where one wants to know probability for the number of trials until the first success: the xth trial is the first success. An alternative formulation of the geometric distribution asks the question: what is the probability of x failures until the first success? In this formulation the trial that resulted in the first success is not counted. The formula for this presentation of the geometric is: ◈(◈=◈)=◈(1–◈)$^{◈}$. The expected value in this form of the geometric distribution is $\mu = \frac{1-p}{p}$. The easiest way to keep these two forms of the geometric distribution straight is to remember that p is the probability of success and (1–◈) is the probability of failure. In the formula the exponents simply count the number of successes and number of failures of the desired outcome of the experiment. Of course the sum of these two numbers must add to the number of trials in the experiment.

Geometric Experiment[1] – a statistical experiment with the following properties: (1) There are one or more Bernoulli trials with all failures except the last one, which is a success. (2) In theory, the number of trials could go on forever. There must be at least one trial. (3) The probability, p, of a success and the probability, q, of a failure do not change from trial to trial.

Independent[2] – two events are independent when the occurrence of one does not affect the probability of the occurrence of the other

Mean[2]– a mean score is an average score, often denoted by $\bar{X}$. It is the sum of individual scores divided by the number of individuals.

Memoryless Property[1] – For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds ◈+◈, given that it has exceeded x, is the same as the probability that X would exceed t if we had no knowledge about it. In symbols we say that $P(X > x + t \mid X > x) = P(X > t)$.

Multiplication Rule[2] – If events A and B come from the same sample space, the probability that both A and B occur is equal to the probability the event A occurs time the probability that B occurs, given that A has occurred. $P(A \cap B) = P(A)P(B \mid A)$.

Normal Distribution[1] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where ◈ is the mean of the distribution and ◈ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$,, the random variable, Z, is called the standard normal distribution

Permutation[2]– an arrangement of all or part of a set of objects, with regard to the order of the arrangement.

Poisson Distribution[1] – If there is a known average of ◈ events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of x events occurring in one unit time is equal to $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$

Poisson Probability Distribution[1] – a discrete random variable that counts the number of times a certain event will occur in a specific interval; characteristics of the variable: (1) the probability that the event occurs in a given interval is the same for all intervals. (2) the events occur with a known mean and independently of the time since the last event. The distribution is defined by the mean ◈ of the event in the interval. The mean is ◈=◈◈. The standard deviation is $\sigma = \sqrt{\mu}$. The probability

of having exactly x successes in r trials is $P(x) = \frac{\mu^x e^{-\mu}}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that np should be greater than or equal to 25 and p should be less than or equal to 0.01).

Probability Distribution Function (PDF)[1]– a mathematical description of a discrete random variable, given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Quartile[2] – Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by
◈$_1$, ◈$_2$, and ◈$_3$, respectively.

Random Variable[1] – a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z...; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3,… Variables in statistics differ from variables in intermediate algebra in the two following ways: (1) The domain of the random variable is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then domain is {black, blonde, gray, green, orange}. (2) We can tell what specific value x the random variable X takes only after performing the experiment.

Standard Deviation[2] – The standard deviation is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the standard deviation is big; and vice versa. It is important to distinguish between the standard deviation of a population and the standard deviation of a sample. They have different notation, and they are computed differently. The standard deviation of a population is denoted by ◈ and the standard deviation of a sample, by ◈

Standard Normal Distribution[1] – a continuous random variable ◈~◈(0,1); when X follows the standard normal distribution, it is often noted as ◈~◈(0,1).

Uniform Distribution[1] – a continuous random variable that has equally likely outcomes over the domain, ◈<◈<◈; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x.Thecumulativedistributionis[latex]P(X \leq x) = \frac{x-a}{b-a}$

Variance[2] – the variance is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the variance is big; and vice versa. It is important to distinguish between the variance of a population and the variance of a sample. They have different notation, and they are computed differently. The variance of a population is denoted by ◈^{2}; and the variance of a sample, by ◈^{2}

Z-Score[1] – the linear transformation of the form $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{|x-\mu|}{\sigma}$; if this transformation is applied to any normal distribution ◈~◈(◈,◈) the result is the standard normal distribution ◈~◈(0,1). If this transformation is applied to any specific value x of the random variable with mean ◈ and standard deviation ◈, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

[1] "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on

OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction

[2] Berman H.B., "Statistics Dictionary", [online] Available at: https://stattrek.com/statistics/dictionary?definition=select-term URL [Accessed Date: 8/16/2022].

MEDIA ATTRIBUTIONS

Note: All Khan Academy content is available for free at (www.khanacademy.org).

Videos

- Video 1: Count Outcomes Using Tree Diagram by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 2: Permutation Formula by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 3: Zero Factorial by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 4: Ways to Pick Officers by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 5: Introduction to Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 6: Combination Formula by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 7: Combination Example: 9 Card Hands by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 8: Probability Using Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 9: Probability and Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 10: Conditional Probability and Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 11: Random Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 12: Discrete and Continuous Random Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 13: Constructing a Probability Distribution for Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- Video 14: Probability with Discrete Random Variable Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 15: Mean (expected value) of a Discrete Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 16: Variance and Standard Deviation of a Discrete Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 17: Bernoulli by Paul Hewson is licensed by Attribution 3.0 Unported (CC BY 3.0)
- Video 18: Mean and Variance of Bernoulli Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 19: Bernoulli Distribution Mean and Variance by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 20: Binomial Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 21: Binomial Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 22: Visualizing a Binomial Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 23: Binomial Probability Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 24: Expected Value of a Binomial Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 25: Variance of a Binomial Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 26: Finding the Mean and Standard Deviation of a Binomial Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 27: Geometric Random Variables Introduction by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
-
- Video 28: Probability for a Geometric Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 29: Cumulative Geometric Probability (greater than a value) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- Video 30: Cumulative Geometric Probability (less than a value) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 31: Poisson Process 1 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 32: Poisson Process 2 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 33: Poisson Probability Distribution – Example by Peter Dalley is licensed under the Attribution 3.0 Unported (CC BY 3.0)
- Video 34: Probability Density Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 35: Uniform Distributions by Lee Wilsonwithers is licensed under Attribution 3.0 Unported (CC BY 3.0)
- Video 36: Exponential Distributions by Significant Statistics is licensed under Attribution 3.0 Unported (CC BY 3.0)
- Video 37: "Statistics Module 6 – Exponential Probability Distribution – Problem 6-4B" by Peter Dalley is licensed under Attribution 3.0 Unported (CC BY 3.0)
- Video 38: Integration by Parts – Exponential Distribution by Maurice Koster is licensed under Attribution 3.0 Unported (CC BY 3.0)
- Video 39: Normal Distribution Problems: Empirical Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 40: Probabilities from Density Curves by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 41: Standard Normal Table for Proportion Below by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 42: Standard Normal Table for Proportion Above by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 43: Standard Normal Table for Proportion Between Values by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 44: Finding z-score for a Percentile by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 45: Queuing Theory: Exponential Distribution by MOOC at IMT is licensed under Attribution 3.0 Unported (CC BY 3.0)

Figures

- Figure 1 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- Figure 2: by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Figure 3 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 4 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 5 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 6 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 7 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 8 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 9 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 10 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 11 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 12 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 13 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 14 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 15 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 16 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 17 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 18 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 19 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 20 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

- Figure 21 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 22 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 23 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 24 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 25 byAlexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 26 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 27 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 28 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 29 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

REFERENCES

- "Unit: Counting, Permutations, and Combinations" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/counting-permutations-and-combinations
- "Unit: Probability" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/probability-library
- "Unit: Random Variables" by Khan Academy https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library
- "Unit: Modeling Data Distributions" by Khan Academy https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data
- Mannering, F., and Washburn, S. (2013). Chapter 5: Fundamentals of Traffic Flow and Queuing Theory. In: Principles of Highway Engineering and Traffic Analysis 5th Edition. John Wiley & Sons, Inc., Hoboken, NJ. pp. 135-174.
- Farid, A. (2022). *Spot Speed Study*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- American Association of State Highway and Transportation Officials (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials, Washington,

D.C.

CHAPTER 9: DATA ANALYSIS - HYPOTHESIS TESTING, ESTIMATING SAMPLE SIZE, AND MODELING

This chapter provides the foundational concepts and tools for analyzing data commonly seen in the transportation profession. The concepts include hypothesis testing, assessing the adequacy of the sample sizes, and estimating the least square model fit for the data. These applications are useful in collecting and analyzing travel speed data, conducting before-after comparisons, and studying the association between variables, e.g., travel speed and congestion as measured by traffic density on the road.

Learning Objectives

At the end of the chapter, the reader should be able to do the following:

- Estimate the required sample size for testing.
- Use specific significance tests including, z-test, t-test (one and two samples), chi-squared test.
- Compute corresponding p-value for the tests.
- Compute and interpret simple linear regression between two variables.
- Estimate a least-squares fit of data.
- Find confidence intervals for parameter estimates.
- Use of spreadsheet tools (e.g., MS Excel) and basic programming (e.g., R or SPSS) to calculate complex and repetitive mathematical problems similar to earthwork estimates (cut, fill, area, etc.), trip generation and distribution, and linear optimization.
- Use of spreadsheet tools (e.g., MS Excel) and basic programming (e.g., R or SPSS) to create relevant graphs and charts from data points.
- Identify topics in the introductory transportation engineering courses that build on the concepts discussed in this chapter.

CENTRAL LIMIT THEOREM

In this section, you will learn about the the central limit theorem by reading each description along with watching the videos. Also, short problems to check your understanding are included.

Central Limit Theorem

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-1

The Central Limit theorem for Sample Means

The **sampling distribution** is a theoretical distribution. It is created by taking many samples of size n from a population. Each sample **mean** is then treated like a single observation of this new distribution, the sampling distribution. The genius of thinking this way is that it recognizes that when we sample, we are creating an observation and that observation must come from some particular distribution. The Central Limit Theorem answers the question: from what distribution dis a sample mean come? If this is discovered, then we can treat a sample mean just like any other observation and calculate probabilities about what values it might take on. We have effectively moved from the world of statistics where we know only what we have from the sample to the world of probability where we know the distribution from which the same mean came and the **parameters** of that distribution.

The reasons that one samples a population are obvious. The time and expense of checking every invoice to determine its validity or every shipment to see if it contains all the items may well exceed the cost of errors in billing or shipping. For some products, sampling would require destroying them, called destructive sampling. One such example is measuring the ability of a metal to withstand saltwater corrosion for parts on ocean going vessels.

Sampling thus raises an important question; just which sample was drawn. Even if the sample were randomly drawn, there are theoretically an almost infinite number of samples. With just 100 items, there are more than 75 million unique samples of size five that can be drawn. If six are in the sample, the number of possible samples increases to just more than one billion. Of the 75 million possible samples, then, which one did you get? If there is variation in the items to be sampled, there will be variation in the samples. One could draw an "unlucky" sample and make very wrong conclusions concerning the population. This recognition that any sample we draw is really only one from a distribution of samples provides us with what is probably the single most important theorem is statistics: the Central Limit Theorem. Without the Central Limit Theorem, it would be impossible to proceed to inferential statistics from simple probability theory. In its most basic form, the Central Limit Theorem states that regardless of the underlying probability density function of the population data, the theoretical distribution of the means of samples from the population will be normally distributed. In essence, this says that the mean of a sample should be treated like an observation drawn from a normal distribution. The Central Limit Theorem only holds if the sample size is "large enough" which has been shown to be only 30 observations or more.

Figure 1 graphically displays this very important proposition.

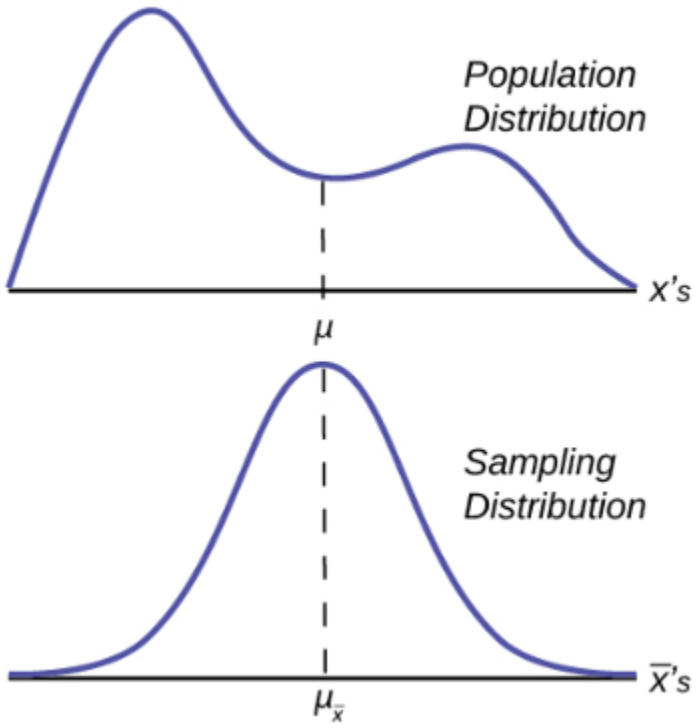


Figure 1

Notice that the horizontal axis in the top panel is labeled X. These are the individual observations of the population. This is the unknown distribution of the population values. The graph is purposefully drawn all squiggly to show that it does not matter just how odd ball it really is. Remember, we will never know what this distribution looks like, or its mean or **standard deviation** for that matter.

The horizontal axis in the bottom panel is labeled $\overline{X's}$. This is the theoretical distribution called the sampling distribution of the means. Each observation on this distribution is a sample mean. All these sample means were calculated from individual samples with the same sample size. The theoretical sampling distribution contains all of the sample mean values from all the possible samples that could have been taken from the population. Of course, no one would ever actually take all of these samples, but if they did this is how they would look. And the Central Limit Theorem says that they will be normally distributed.

The Central Limit Theorem goes even further and tells us the mean and standard deviation of this theoretical distribution.

Table 1

Parameter	**Population distribution**	**Sample**	**Sampling distribution of $\bar{X}$'s**
Mean	μ	$\bar{X}$	$\mu_{\bar{x}} \quad \text{and } \mathrm{E}(\mu_{\bar{x}}) = \mu$
Standard deviation	σ	s	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

The practical significant of The Central Limit Theorem is that now we can compute probabilities for drawing a sample mean, $\overline{X'}$, in just the same way as we did for drawing specific observations, $X^{-1}s$, when we knew the population mean and standard deviation and that the population data were normally distributed. The standardizing formula has to be amended to recognize that the mean and standard deviation of the sampling distribution, sometimes, called the standard error of the mean, are different from those of the population distribution, but otherwise nothing has changed. The new standardizing formula is

$$Z = \frac{\bar{X} - \mu_{-}}{\sigma_{-}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Notice that $\mu\bar{X}$ in the first formula has been changed to simply μ in the second version. The reason is that mathematically it can be shown that the expected value of $\mu\bar{X}$ is equal to μ.

Sampling Distribution of the Sample Mean

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-2

Sampling Distribution of the Sample Mean (Part 2)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-3

Sampling Distributions: Sampling Distribution of the Mean

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-4

Using 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 of the sampling distribution, $\mu\bar{X}$ tends to get close and close to the true population mean, μ. From the Central Limit Theorem, we know that as n gets larger

and larger, the sample means follow a normal distribution. The larger *n* gets, the smaller the standard deviation of the sampling distribution gets. (Remember that the standard deviation for sampling distribution of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$. This means that the sample mean $\bar{x}$ must be closer to the population mean μ as *n* increases. We can say that μ is the value that the sample means approach as *n* gets larger. The Central Limit Theorem illustrates the law of large numbers.

Indeed, there are two critical issues that flow from the Central Limit Theorem and the application of the Law of Large numbers to it. These are listed below.

1. The probability density function of the sampling distribution of means is normally distributed regardless of the underlying distribution of the population observations and
2. Standard deviation of the sampling distribution decreases as the size of the samples that were used to calculate the means for the sampling distribution increases.

Taking these in order. It would seem counterintuitive that the population may have any distribution and the distribution of means coming from it would be normally distributed. With the use of computers, experiments can be simulated that show the process by which the sampling distribution changes as the sample size is increased. These simulations show visually the results of the mathematical proof of the Central Limit Theorem.

Figure 2 shows a sampling distribution. The mean has been marked on the horizontal axis of the $\bar{x}'s$ and the standard deviation has been written to the right above the distribution. Notice that the standard deviation of the sampling distribution is the original standard deviation of the population, divided by the sample size. We have already seen that as the sample size increases the sampling distribution becomes closer and closer to the normal distribution. As this happens, the standard deviation of the sampling distribution changes in another way; the standard deviation decreases as n increases. At very large n, the standard deviation of the sampling distribution becomes very small and at infinity it collapses on top of the population mean. This is what it means that the expected value of $\mu_{\bar{x}}$ is the population mean, μ.

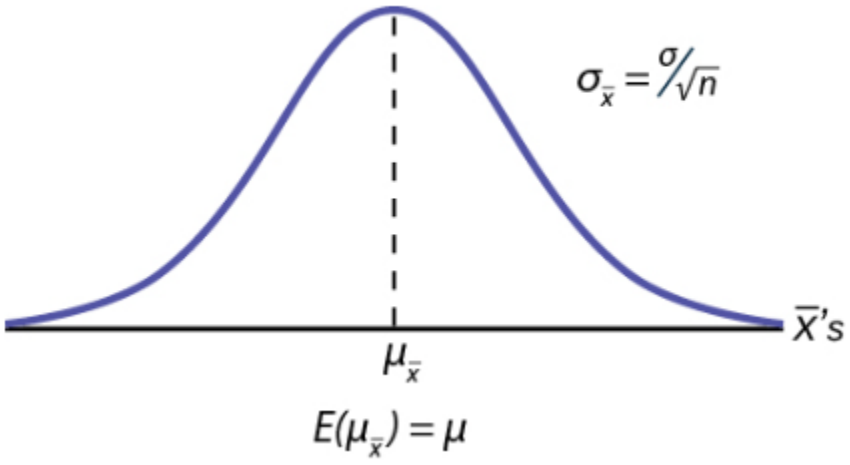


Figure 2

At non-extreme values of n, this relationship between the standard deviation of the sampling distribution and the sample size plays a very important part in our ability to estimate the parameters in which we are interested.

Figure 3 shows three sampling distributions. The only change that was made is the sample size that was used to get the sample means for each distribution. As the sample size increases, n goes from 10 to 30 to 50, the standard deviations of the respective sampling distributions decrease because the sample size is in the denominator of the standard deviations of the sampling distributions.

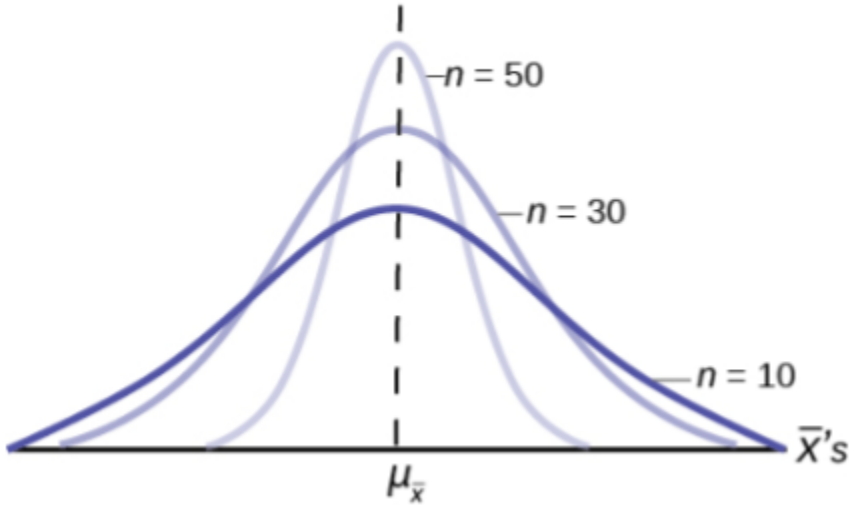


Figure 3

The implications for this are very important. Figure 4 shows the effect of the sample size on the confidence we will have in our estimates. These are two sampling distributions from the same population. One sampling distribution was created with samples of size 10 and the other with samples of size 50. All other things constant, the sampling distribution with sample size 50 has a smaller standard deviation that causes the graph to be higher and narrower. The important effect of this is that for the same probability of one standard deviation from the mean, this distribution covers much less of a range of possible values than the other distribution. One standard deviation is marked on the $\bar{X}$ axis for each distribution. This is shown by the two arrows that are plus or minus one standard deviation for each distribution. If the probability that the true mean is one standard deviation away from the mean, then for the sampling distribution with the smaller sample size, the possible range of values is much greater. A simple question is, would you rather have a sample mean from the narrow, tight distribution, or the flat, wide distribution as the estimate of the population mean? Your answer tells us why people intuitively will always choose data from a large sample rather than a small sample. The sample mean they are getting is coming from a more compact distribution.

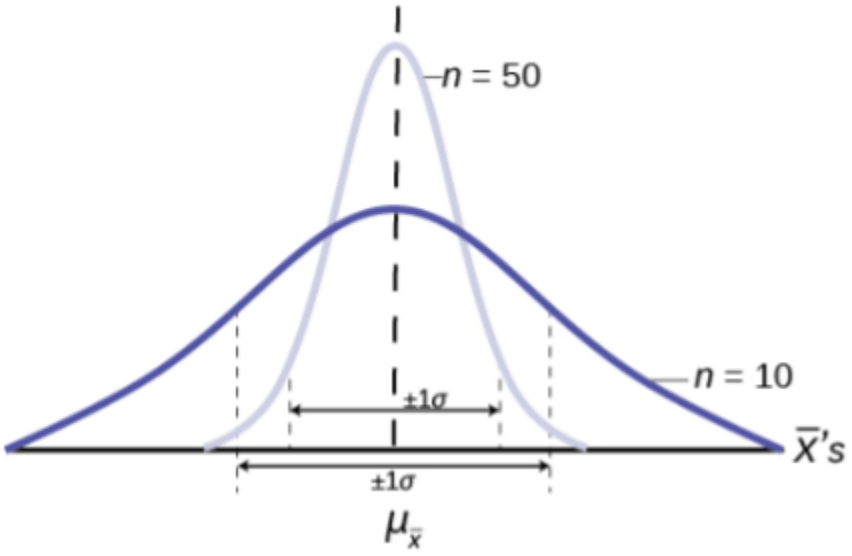


Figure 4

The Central Limit Theorem for Proportions

The Central Limit Theorem tells us that the **point estimate** for the sample mean, $\bar{x}$, comes from a normal distribution of $\bar{x}'s$. This theoretical distribution is called the sampling distribution of $\overline{x'}s$. We now investigate the sampling distribution for another important parameter we wish to estimate, p from the binomial probability density function.

If the random variable is discrete, such as for categorical data, then the parameter we wish to estimate is the population proportion. This is, of course, the probability of drawing a success in any one random draw. Unlike the case just discussed for a continuous random variable where we did not know the population distribution of X's, here we actually know the underlying probability density function for these data; it is the binomial. The random variable is X = the number of successes and the parameter we wish to know is p, the probability of drawing a success which is of course the proportion of successes in the population. The question at issue is: from what distribution was the sample proportion, $p' = \frac{x}{n}$ drawn? The sample size is n and X is the number of successes found in that sample. This is a parallel question that was just answered by the Central Limit Theorem: from what distribution was the sample mean, $\bar{x}$ drawn? We saw that once we knew that the distribution was the Normal distribution then we were able to create confidence intervals for the population parameter, μ. We will also use this same information to test hypotheses about the population mean later. We wish now to be able to develop confidence intervals for the population parameter "p" from the binomial probability density function.

In order to find the distribution from which sample proportions come we need to develop the sampling distribution of sample proportions just as we did for sample means. So again, imagine that we randomly sample say 50 people and ask them if they support the new school bond issue. From this we find a sample proportion, p', and graph it on the axis of p's. We do this again and again etc., etc. until we have the theoretical distribution of p's. Some sample proportions will show high favorability toward the bond issue and others will show low favorability because

random sampling will reflect the variation of views within the population. What we have done can be seen in Figure 5. The top panel is the population distributions of probabilities for each possible value of the random variable X. While we do not know what the specific distribution looks like because we do not know p, the population parameter, we do know that it must look something like this. In reality, we do not know either the mean or the standard deviation of this population distribution, the same difficulty we faced when analyzing the X's previously.

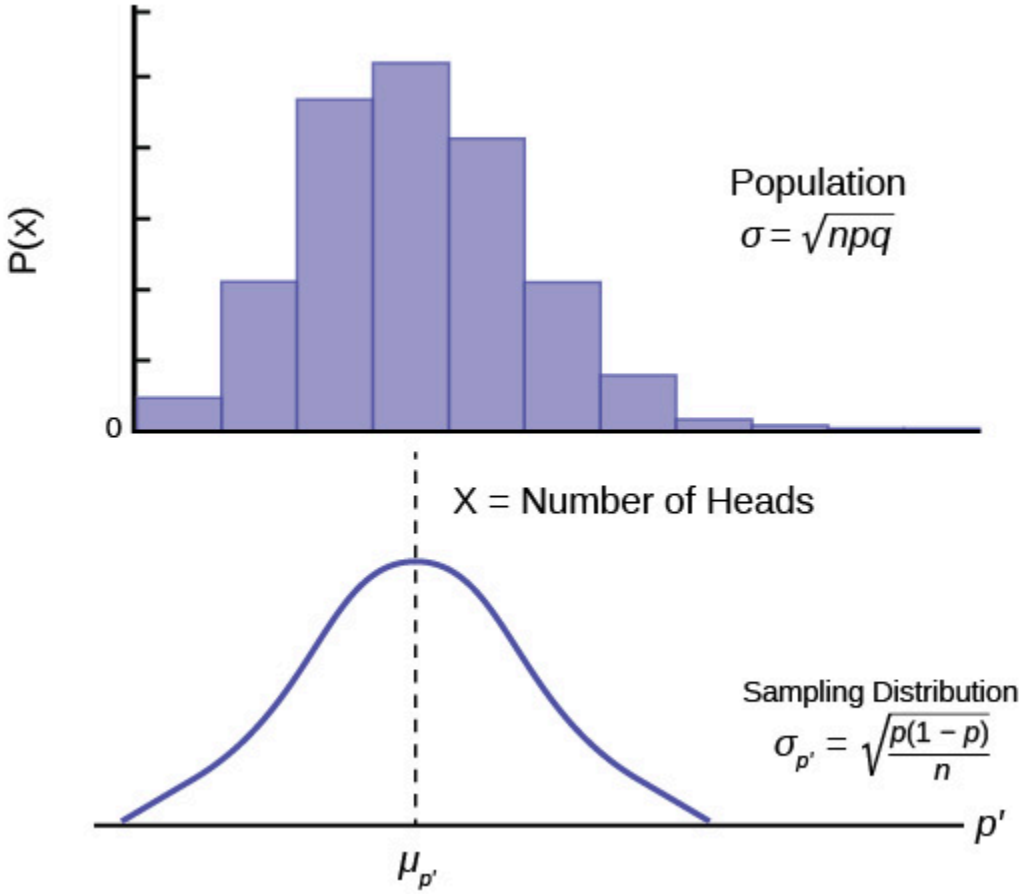


Figure 5

Figure 5 places the mean on the distribution of population probabilities as $\mu = np$ but of course we do not actually know the population mean because we do not know the population probability of success, *p*. Below the distribution of the population values is the sampling distribution of *p*'s. Again, the Central Limit Theorem tells us that this distribution is normally distributed just like the case of the sampling distribution for $\bar{x}'s$. This sampling distribution also has a mean, the mean of the *p*'s, and a standard deviation, $\sigma_{p'}$.

Importantly, in the case of the analysis of the distribution of sample means, the Central Limit Theorem told us the expected value of the mean of the sample means in the sampling distribution, and the standard deviation of the sampling distribution. Again, the Central Limit Theorem provides this information for the sampling distribution for proportions. The answers are:

1. The expected value of the mean of sampling distribution of sample proportions, $\mu_{p'}$, is

the population proportion, p.

2. The standard deviation of the sampling distribution of sample proportions, $\sigma_{p'}$, is the population standard deviation divided by the square root of the sample size, n.

Both these conclusions are the same as we found for the sampling distribution for sample means. However, in this case, because mean and standard deviation of the binomial distribution both rely upon *p*, the formula for the standard deviation of the sampling distribution requires algebraic manipulation to be useful. The standard deviation of the sampling distribution for proportions is thus:

$$\sigma_{p'} = \sqrt{\frac{p(1-P)}{n}}$$

Table 2

Parameter	Population distribution	Sample	Sampling distribution of *p*'s
Mean	$\mu = np$	$p' = \frac{x}{n}$	p' and $E(p') = p$
Standard deviation	$\sigma = \sqrt{npq}$		$\sigma_{\mathrm{p}}, = \sqrt{\frac{p(1-p)}{n}}$

Table 2 summarizes these results and shows the relationship between the population, sample, and sampling distribution.

Reviewing the formula for the standard deviation of the sampling distribution for proportions we see that as n increases the standard deviation decreases. This is the same observation we made for the standard deviation for the sampling distribution for means. Again, as the sample size increases, the point estimate for either μ or p is found to come from a distribution with a narrower and narrower distribution. We concluded that with a given level of probability, the range from which the point estimate comes is smaller as the sample size, n, increases.

Find Confidence Intervals for Parameter Estimates

In this section, you will learn how to find and estimate confidence intervals by reading each description along with watching the videos included. Also, short problems to check your understanding are included.

Confidence Intervals & Estimation: Point Estimates Explained

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-5

Introduction to Confidence Intervals

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed,

and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion the parameter p in the binomial probability density function.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals. What statistics provides us beyond a simple **average**, or point estimate, is an estimate to which we can attach a probability of accuracy, what we will call a confidence level. We make inferences with a known level of probability.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from iTunes. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, *s*. You would use $\bar{x}$ to estimate the population mean and *s* to estimate the population standard deviation. The same mean, $\bar{x}$, is the point estimate for the population mean, μ. The sample standard deviation, *s*, is the point estimate for the population standard deviation, σ.

$\bar{x}$ and *s* are each called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include the unknown population parameter.

Suppose for the iTunes examples, we do not know the population mean μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the Central Limit Theorem, the standard deviation of the sampling distribution of the sample means is $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$.

The **Empirical Rule**, which applies to the normal distribution, says that in approximately 95% of the samples, the same mean, $\bar{x}$, will be within two standard deviations of the population mean $\boldsymbol{\mu}$. For our iTunes example, two standard deviations is $(2)(0.1) = 0.2$. The sample mean $\bar{x}$ is likely to be within 0.2 units of $\boldsymbol{\mu}$.

Because $\bar{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\bar{x}$ with 95% probability. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $(2)(0.1)$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 955 of all the samples.

For the iTunes example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then with 95% probability the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \text{ and } \bar{x} + 0.2 = 2 + 0.2 = 2.2$$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from iTunes per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).** Please note that we talked in terms of 95% confidence using the empirical rule. The empirical rule for two standard deviations is only approximately 95% of the probability under the normal distribution. To be precise, two standard deviations under a normal distribution is actually 95.44% of the probability. To calculate the exact 95% confidence level, we would use 1.96 standard deviations.

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean ◈">◈, or our sample produce an $\bar{x}$ that is not within 0.2 units of the true mean ◈">◈. The first possibility happens for 95% of well-chosen samples. It is important to remember that the second possibility happens for 5% of samples, even though correct procedures are followed.

Remember that a confidence interval is created for an unknown population parameter like the population mean, ◈">◈.

For the confidence interval for a mean the formula would be:

$$\mu = \bar{X} \pm Z_\alpha \sigma / \sqrt{n}$$

Or written another way as:

$$\bar{X} - Z_\alpha \sigma / \sqrt{n} \leq \mu \leq \bar{X} + Z_\alpha \sigma / \sqrt{n}$$

Where $\bar{X}$s is the sample mean. $Z_{-}a$ is determined by the level of confidence desired by the analyst, and $\sigma / \sqrt{n}$ is the standard deviation of the sampling distribution for means given to us by the Central Limit Theorem.

A Confidence Interval for a Population Standard Deviation, Known or Large Sample Size

A confidence interval for a population mean, when the population standard deviation is known based on the conclusion of the Central Limit Theorem that the sampling distribution of the sample means follow an approximately normal distribution.

Calculating the Confidence Interval

Consider the standardizing formula for the sampling distribution developed in the discussion of the Central Limit Theorem:

$$Z_1 = \frac{\bar{X} - \mu_{-}}{\sigma_{-}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Notice that μ is substituted for $\mu_{-}x-$ because we know that the expected value of $\mu_{-}x-$ is from the Central Limit Theorem and $\sigma_{-}x-$ is replaced with $\sigma / \sqrt{n}$, also from the Central Limit Theorem.

In this formula we know $X, \sigma_{-}x-$ and n, the sample size. (In actuality we do not know the population standard deviation, but we do have a point estimate for it, s, from the sample we

took. More on this later.) What we do not know is μ of Z_1. We can solve for either one of these in terms of the other. Solving for ◈">◈ in terms of Z_1 gives:

$$\mu = \bar{X} \pm Z_1 \sigma / \sqrt{n}$$

Remembering that the Central Limit Theorem tells us that the distribution of the $\bar{X}'s$, the sampling distribution for means, is normal, and that the normal distribution is symmetrical, we can rearrange terms thus:

$$\bar{X} - Z_\alpha(\sigma/\sqrt{n}) \leq \mu \leq \bar{X} + Z_\alpha(\sigma/\sqrt{n})$$

This is the formula for a confidence interval for the mean of a population.

Notice that Z_α has been substituted for Z_1 in this equation. This is where the statistician must make a choice. The analyst must decide the level of confidence they wish to impose on the confidence interval. α is the probability that the interval will not contain the true population mean. The confidence level is defined as $(1-\alpha)$.Z_α is the number of standard deviations $\bar{X}$ lies from the mean with a certain probability. If we chose, $Z_\alpha = 1.96$ we are asking for the 95% confidence interval because we are setting the probability that the true mean lies within the range at 0.95. If we set Z_α at 1.64 we are asking for the 90% confidence interval because we have set the probability at 0.90. These numbers can be verified by consulting the Standard Normal table. Divide either 0.95 or 0.90 in half and find that probability inside the body of the table. Then read on the top and left margins the number of standard deviations it takes to get this level of probability.

Table 3

Confidence Level	
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.58

In reality, we can set whatever level of confidence we desire simply by changing the Z_α value in the formula. It is the analyst's choice. Common convention in Economics and most social sciences sets confidence intervals at either 90, 95, or 99 percent levels. Levels less than 90% are considered of little value. The level of confidence of a particular interval estimate is called b $(1-\alpha)$.

Let us say we know that the actual population mean number of iTunes downloads is 2.1. The true population mean falls within the range of the 95% confidence interval. There is absolutely nothing to guarantee that this will happen**. Further, if the true mean falls outside of the interval, we will never know it. We must always remember that we will never ever know the true mean.** Statistics simply allows us, with a given level of probability (confidence), to say that the true mean is within the range calculated.

Changing the Confidence Level or Sample Size

Here again is the formula for a confidence interval for an unknown population mean assuming we know the population standard deviation:

$$\bar{X} - Z_\alpha(\alpha/\sqrt{n}) \leq \mu \leq \bar{X} + Z_\alpha(\alpha/\sqrt{n})$$

It is clear that the confidence interval is driven by two things, the chosen level of confidence, Z_α, and the standard deviation of the sampling distribution. The standard deviation of the sampling distribution is further affected by two things, the standard deviation of the population and sample size we chose for our data. Here we wish to examine the effects of each of the choices we have made on the calculated confidence interval, the confidence level, and the sample size.

For a moment we should ask just what we desire in a confidence interval. Our goal was to estimate the population mean from a sample. We have forsaken the hope that we will ever find the true population mean, and population standard deviation for that matter, for any case except where we have an extremely small population and the cost of gathering the data of interest is very small. In all other cases we must rely on samples. With the Central Limit Theorem, we have the tools to provide a meaningful confidence interval with a given level of confidence, meaning a known probability of being wrong. By meaningful confidence interval we mean one that is useful. Imagine that you are asked for a confidence interval for the ages of your classmates. You have taken a sample and find a mean of 19.8 years. You wish to be very confident, so you report an interval between 9.8 years and 29.8 years. This interval would certainly contain the true population mean and have a very high confidence level. However, it hardly qualifies as meaningful. The very best confidence interval is narrow while having high confidence. There is a natural tension between these two goals. The higher the level of confidence the wider the confidence interval as the case of the students' ages above. We can see this tension in the equation for the confidence interval.

$$\mu = \bar{x} \pm Z_\alpha \left(\frac{\sigma}{\sqrt{n}}\right)$$

The confidence interval will increase in width as Z_α increases, Z_α increases as the level of confidence increases. There is a tradeoff between the level of confidence and the width of the interval. Now let us look at the formula again and we see that the sample size also plays an important role in the width of the confidence interval. The sample size, *n*, shows up in the denominator of the standard deviation of the sampling distribution. As the sample size increases, the standard deviation of the sampling distribution decreases and thus the width of the confidence interval, while holding constant the level of confidence. Again, we see the importance of having large samples for our analysis although we then face a second constraint, the cost of gathering data.

Calculating the Confidence Interval: An Alternative Approach

Another way to approach confidence intervals is through the use of something called the Error Bound (margin of error). The Error Bound gets its name from the recognition that it provides the boundary of the interval derived from the standard error of the sampling distribution. In the equations above it is seen that the interval is simply the estimated mean, sample mean, plus or minus something. That something is the Error Bound and is driven by the probability we desire

to maintain in our estimate, Z_α, times the standard deviation of the sampling distribution. The Error Bound for a mean is given the name, **Error Bound Mean**, or EBM (or margin of error, M.O.E.).

To construct a confidence interval for a single unknown population, mean ◈">◈, **where the population standard deviation is known**, we need $\bar{x}$ as an estimate for ◈">◈ and we need the margin of error. Here, the margin of error (EBM) is called the **error bound for a population mean**. The sample mean $\bar{x}$ is the **point estimate** of the unknown population mean ◈">◈.

The confidence interval estimate will have the form:

(Point estimate – error bound, point estimate + error bound) or, in symbols,

$$(\bar{x} - EBM, \bar{x} + EBM).$$

The mathematical formula for this confidence interval is:

$$\bar{X} - Z_\alpha(\sigma/\sqrt{n}) \leq \mu \leq \bar{X} + Z_\alpha(\sigma/\sqrt{n})$$

The margin of error (EBM) depends on the **confidence level** (abbreviated *CL*). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha $(\alpha).\alpha$ is related to the confidence level, *CL*. α is the probability that the interval does not contain the unknown population parameter. Mathematically, $1 - \alpha = CL$.

A confidence interval for a population mean with a known standard deviation is based on the fact that the sampling distribution of the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90% confidence interval (5, 15) where EMB = 5.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10\%$ in both tails, or 5% in each tail, of the normal distribution.

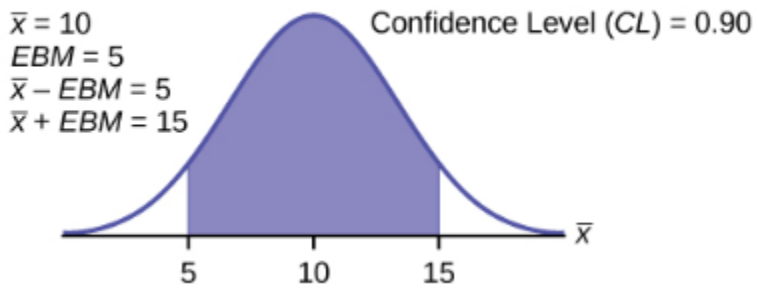


Figure 6

To capture the central 90%, we must go out 1.645 standard deviations on either side of the calculated sample mean. The value 1.645 is the z-score from a standard normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far-left tail, and an area of 0.05 in the far-right tail.

It is important that the standard deviation used must be appropriate for the parameter we are estimating, so in this section we need to use the standard deviation that applies to the sampling distribution for mean which we studied with the Central Limit Theorem and is, $\frac{\sigma}{\sqrt{n}}$.

Calculating the Confidence Interval Using EBM

To construct a confidence interval, estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are listed below.

- Calculate the sample mean $\bar{x}$ from the sample data. Remember, in this section we know the population standard deviation σ.
- Find the z-score from the standard normal table that corresponds to the confidence level desired.
- Calculate the error bound EBM.
- Construct the confidence interval.
- Write a sentence that interprets the estimate in the context of the situation in the problem.

Finding the z-score for the Stated Confidence Level

When we know the population standard deviation σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, *CL*, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $Z_{\frac{\alpha}{2}}$

For example, when $CL = 0.95, \alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $Z_{\frac{\alpha}{2}} = Z_{0.025}$

The area to the right of $Z_{0.025}$ is 0.025 and the area to the left of $Z_{0.025}$ is $1 - 0.025 = 0.975$.

$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$, using a standard normal probability table. We will see later that we can use a different probability table, the **Student's t-distribution**, for finding the number of standard deviations of commonly used levels of confidence.

Calculating the Error Bound (EBM)

The error bound formula for an unknown population mean μ when the population standard deviation σ is known as

$$EBM = \left(Z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

Constructing the Confidence Interval

The confidence interval estimate has the format $(\bar{x} - EMB, \bar{x} + EBM)$ or the formula:

$$\bar{X} - Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$$

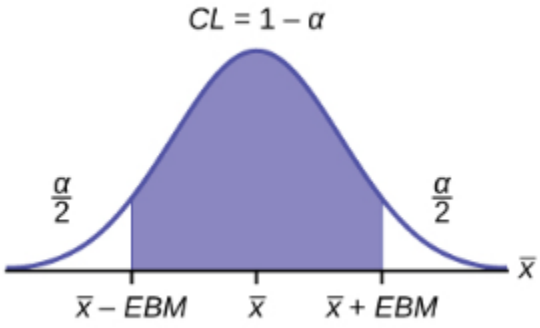


Figure 7

Confidence Interval for Mean: 1 Sample Z Test (Using Formula)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-6

Check Your Understanding: Confidence Interval for Mean

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-44

A Confidence Interval for a Population Standard Deviation Unknown, Small Sample Case

In practice, we rarely known the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation *s* as an estimate for σ and proceeded as before to calculate a confidence interval with close enough results. Statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876-1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barely produced very few samples. Just replacing σ with *s* did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution.** The name comes from the fact that Gosset wrote under the pen name "A Student."

Up until the mid-1970s, some statisticians used **the normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30 observations.

If you draw a simple random sample of size *n* from a population with mean ◈">◈ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student's t-distribution with *n* – 1 degrees of freedom.** The t-score has the same interpretation as the z-score. It measures how far in standard deviation units $\bar{x}$ is from its mean μ. For each sample size *n*, there is a different Student's t-distribution.

The **degrees of freedom, *n* – 1**, come from the calculation of the sample standard deviation *s***.** Remember when we first calculated a sample standard deviation, we divided the sum of the squared deviations by $n-1$, but we used *n* deviations $(x-\bar{x} \text{ values})$ to calculate *s*. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n-1$ deviations. The other $n-1$ deviations can change or vary freely. We call the number $n-1$ the **degrees of freedom** (df) in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the t-value increases, and the confidence interval increases in width.

Properties of the Student's t-distribution

- The graph for the Student's t-distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution. You can confirm this by reading the bottom line at infinite degrees of freedom for a familiar level of confidence, e.g., at column 0.05, 95% level of confidence, we find the t-value of 1.96 at infinite degrees of freedom.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero, again like the standard normal distribution.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So, the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation ◈. This assumption comes from the Central Limit Theorem because the individual observations in this case are the $\bar{x}'s$ of the sample distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal, then the assumption is met and does not need discussion.

A probability table for the Student's t-distribution is used to calculate t-values at various commonly used levels of confidence. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that at the bottom the table will show the t-value for infinite degrees of freedom. Mathematically, as the degrees of freedom increase, the t-distribution approaches the standard normal distribution. You can find familiar Z-values by looking in the relevant alpha column and reading value in the last row.

A Student's t-table gives t-scores given the degrees of freedom and the right-tailed probability.

The Student's t-distribution has one of the most desirable properties of the normal: it is symmetrical. What the Student's t-distribution does is spread out the horizontal axis, so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's t-distributions, one for each adjustment to the sample size. As the sample size increases, the Student's t-distribution become more and more like the normal distribution. When the sample size reaches 30 the normal distribution is usually substituted for the Student's t because they are so much alike. This relationship between the Student's t-distribution and the normal distribution is shown in Figure 8.

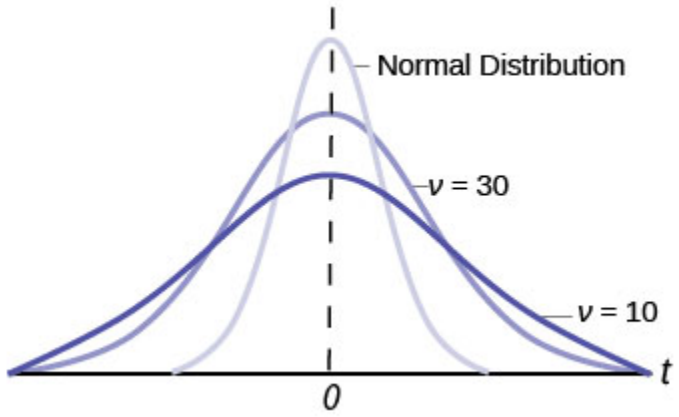


Figure 8

Here is the confidence interval for the mean for cases when the sample size is smaller than 30 and we do not know the population standard deviation, σ:

$$\bar{x} - t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{v,\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

Here the point estimate of the population standard deviation, *s* has been substituted for the population standard deviation, σ, and $t_{\nu,\alpha}$ has been substituted for Z_a. The Greek letter $\boldsymbol{V}$ (pronounced nu) is places in the general formula in recognition that there are many Student t_ν distributions, one for each sample size. V is the symbol for the degrees of freedom of the distribution and depends on the size of the sample. Often df is used to abbreviate degrees of freedom. For this type of problem, the degrees of freedom is $v = n - 1$, where *n* is the sample size. To look up a probability in the Student's t table we have to know the degrees of freedom in the problem.

Confidence Intervals: Using the t Distribution

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-7

Check Your Understanding: Confidence Intervals

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-45

A Confidence Interval for a Population Proportion

During an election year, we see articles in the newspaper that state confidence intervals in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43.

The procedure to find the confidence interval for a population proportion is similar to that for the population mean, but the formulas are a bit different although conceptually identical. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

How do you know you are dealing with a proportion problem? First, the underlying **distribution has a binary random variable and therefore is a binomial distribution.** (There is no mention of a mean or average). If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a sample proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable $\boldsymbol{P'}$(read "P Prime") is the sample proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\widehat{P}$, read "P hat.")

p' = the **estimated proportion** of successes or sample proportion of successes (p' is a **point estimate** for p, the true population proportion, and thus q is the probability of a failure in any one trial.)

x = the **number** of successes in the sample

n = the size of the sample

The formula for the confidence interval for a population proportion follows the same format as that for an estimate of a population mean. Remembering the sampling distribution for the proportion, the standard deviation was found to be:

$$\sigma_{p'} = \sqrt{\frac{p(1-p)}{n}}$$

The confidence interval for a population proportion, therefore, becomes:

$$p = p' \pm \left[Z_{\left(\frac{\alpha}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

$Z_{\left(\frac{\alpha}{2}\right)}$ is set according to our desired degree of confidence and $\sqrt{\frac{p'(1-p')}{n}}$ is the standard deviation of the sampling distribution.

The **sample proportions *p'* and *q'* are estimates of the unknown population proportions *p* and *q*.** The estimated proportions *p'* and *q'* are used because *p* and *q* are not known.

Remember that as p moves further from 0.5 the binomial distribution becomes less symmetrical. Because we are estimating the binomial with the symmetrical normal distribution the further away from symmetrical the binomial becomes the less confidence we have in the estimate.

This conclusion can be demonstrated through the following analysis. Proportions are based upon the binomial probability distribution. The possible outcomes are binary, either "success" or "failure." This gives rise to a proportion, meaning the percentage of the outcomes that are "successes." It was shown that the binomial distribution could be fully understood if we knew only the probability of a success in any one trial, called p. The mean and the standard deviation of the binomial were found to be:

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

It was also shown that the binomial could be estimated by the normal distribution if BOTH np AND nq were greater than 5. From the discussion above, it was found that the standardizing formula for the binomial distribution is:

$$Z = \frac{p'-p}{\sqrt{\left(\frac{pq}{n}\right)}}$$

Which is nothing more than a restatement of the general standardizing formula with appropriate substitutions for μ and σ from the binomial. We can use the standard normal distribution, the reason Z is in the equation, because the normal distribution is the limiting distribution of the binomial. This is another example of the Central Limit Theorem.

We can now manipulate this formula in just the same way we did for finding the confidence intervals for a mean, but to find the confidence interval for the binomial population parameter, *p*.

$$p' - Z_\alpha \sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'q'}{n}}$$

Where $p' = x/n$, the point estimate of *p* taken from the sample. Notice that *p'* has replaced *p* in the formula. This is because we do not know *p*, indeed, this is just what we are trying to estimate.

x = number of successes.

n = the number in the sample.

$q' = (1 - p')$, the failure in any trial

Unfortunately, there is no correction factor for cases where the sample size is small so *np'* and *nq'* must always be greater than 5 to develop an interval estimate for *p*.

Also written as:

$$p' - Z_\alpha \sqrt{\frac{p'(1-p')}{n}} \leq p \leq p' + Z_\alpha \sqrt{\frac{p'(1-p')}{n}}$$

How to Construct a Confidence Interval for Population Proportion

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-8

Check Your Understanding: How to Construct a Confidence Interval for Population Proportion

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-46

ESTIMATE THE REQUIRED SAMPLE SIZE FOR TESTING

In this section, you will learn how to calculate sample size with continuous and binary random

samples. by reading each description along with watching the videos included. Also, short problems to check your understanding are included.

Calculating the Sample Size n: Continuous and Binary Random Variables

Continuous Random Variables

Usually, we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly in both time and product. Simple telephone surveys will cost approximately $30.00 each, for example, and some sampling requires the destruction of the product.

If we go back to our standardizing formula for the sampling distribution for means, we can see that it is possible to solve it for *n*. If we do this, we have $(\bar{X} - \mu)$ in the denominator.

$$n = \frac{Z_\alpha^2 \sigma^2}{(\bar{X} - \mu)^2} = \frac{Z_\alpha^2 \sigma^2}{e^2}$$

Because we have not taken a sample, yet we do not know any of the variables in the formula except that we can set Z_α to the level of confidence we desire just as we did when determining confidence intervals. If we set a predetermined acceptable error, or tolerance, for the difference between $\bar{X} \text{ and } \mu, \text{ called } e$ in the formula, we are much further in solving for the sample size *n*. We still do not know the population standard deviation, σ. In practice, a pre-survey is usually done which allows for fine tuning the questionnaire and will give a sample standard deviation that can be used. In other cases, previous information from other surveys may be used for σ in the formula. While crude, this method of determining the sample size may help in reducing cost significantly. If will be the actual data gathered that determines the inferences about the population, so caution in the sample size is appropriate calling for high levels of confidence and small sampling errors.

Binary Random Variables

What was done in cases when looking for the mean of a distribution can also be done when sampling to determine the population parameter *p* for proportions. Manipulation of the standardizing formula for proportions gives:

$$n = \frac{Z_\alpha^2 pq}{e^2}$$

Where $e = (p' - p)$, and is the acceptable sampling error, or tolerance, for this application. This will be measured in percentage points.

In this case the very object of our search is in the formula, *p*, and of course *q* because $q = 1 - p$. This result occurs because the binomial distribution is a one parameter distribution. If we know *p* then we know the mean and the standard deviation. Therefore, *p* shows up in the standard deviation of the sampling distribution which is where we got this formula. If, in an abundance of caution, we substitute 0.5 for *p* we will draw the largest required sample size that will provided the level of confidence specified by Z_α and the tolerance we have selected. This is

true because of all combinations of two fractions that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p, this is the common practice. This may result in oversampling, but certainly not under sampling, thus, this is a cautious approach.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. Table 4 shows the appropriate sample size at different levels of confidence and different level of the acceptable error, or tolerance.

Table 4

Required sample size (90%)	Required sample size (95%)	Tolerance level
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

Table 4 is designed to show the maximum sample size required at different levels of confidence given an assumed $p = 0.5$ and $q = 0.5$ as discussed above.

The acceptable error, called tolerance in the table, is measured in plus or minus values from the actual proportion. For example, an acceptable error of 5% means that if the sample proportion was found to be 26 percent, the conclusion would be that the actual population proportion is between 21 and 31 percent with a 90 percent level of confidence if a sample of 271 had been taken. Likewise, if the acceptable error was set at 2%, then the population proportion would be between 24 and 28 percent with a 90 percent level of confidence but would require that the sample size be increased from 271 to 1,691. If we wished a higher level of confidence, we would require a larger sample size. Moving from a 90 percent level of confidence to a 95 percent level at a plus or minus 5% tolerance requires changing the sample size from 271 to 384. A very common sample size often seen reported in political surveys is 384. With the survey results it is frequently stated that the results are good to a plus or minus 5% level of "accuracy".

Example: Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

Solution: From the problem, we know that the acceptable error, *e*, is 0.03 (3% = 0.03) and $z_{\frac{\alpha}{2}}, z_{0.05} = 1.645,$ because the confidence level is 90%. The acceptable error, *e*, is the difference between the actual population proportion *p*, and the sample proportion we expect to get from the sample.

However, in order to find *n*, we need to know the estimated (sample) proportion *p'*. Remember that $q' = 1 - p'$. But we do not know *p'* yet. Since we multiply *p'* and *q'* together, we make them both equal to 0.5 because $p'q' = (0.5)(0.5) = 0.25$ results in the largest possible product. (Try other products: $(0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16$ and so on).

The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{e^2} \text{ gives } n = \frac{1.645^2(0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers aged 50+ who use text messaging on their cell phones.

Estimation and Confidence Intervals: Calculate Sample Size

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-9

Calculating Sample size to Predict a Population Proportion

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-10

USE SPECIFIC SIGNIFICANCE TESTS INCLUDING, Z-TEST, T-TEST (ONE AND TWO SAMPLES), CHI-SQUARED TEST

In this section, you will learn the fundamentals of hypothesis testing along with hypothesis testing with errors by reading each description along with watching the videos. Also, short problems to check your understanding are included.

Hypothesis Testing with One Sample

Statistical testing is part of a much larger process known as the scientific method. The scientific method, briefly, states that only by following a careful and specific process can some assertion be included in the accepted body of knowledge. This process begins with a set of assumptions upon which a theory, sometimes called a model, is built. This theory, if it has any validity, will lead to predictions; what we call hypotheses.

Statistics and statisticians are not necessarily in the business of developing theories, but in the business of testing others' theories. Hypotheses come from these theories based upon an explicit set of assumptions and sound logic. The **hypothesis** comes first, before any data are gathered.

Data do not create hypotheses; they are used to test them. If we bear this in mind as we study this section, the process of forming and testing hypotheses will make more sense.

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. Confidence intervals are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about the value of a specific parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called "**hypothesis testing.**" A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

Hypothesis Testing: The Fundamentals

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-11

Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0 : **The null hypothesis:** It is a statement of no difference between the variables–they are not related. This can often be considered the status quo and as a result if you cannot accept the null, it requires some action.

H_a : **The alternative hypothesis:** It is a claim about the population that is contradictory to H_0 and what we conclude when we cannot accept H_0. This is usually what the researcher is trying to prove. The alternative hypothesis is the contender and must win with significant evidence to overthrow the status quo. This concept is sometimes referred to the tyranny of the status quo because as we will see later, to overthrow the null hypothesis takes usually 90 or greater confidence that this is the proper decision.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are "cannot accept H_0" if the sample information favors the alternative hypothesis or "do not reject H_0" or "decline to reject H_0 " if the sample information is insufficient to reject the null hypothesis. These conclusions are all based upon a level of probability, a significance level, that is set by the analyst.

Table 5 presents the various hypotheses in the relevant pairs. For example, if the null hypothesis is equal to some value, the alternative has to be not equal to that value.

Table 5

H_0	$\boldsymbol{H}_a$
Equal (=)	Not equal $(\neq)$
Greater than or equal to $(\geq)$	Less than (<)
Less than or equal to $(\leq)$	More than (>)

Note: As a mathematical convention $\boldsymbol{H}_0$ always has a symbol with an equal in it. $\boldsymbol{H}_a$ never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test.

Example 1:

$\boldsymbol{H}_0$: No more than 30% of the registered voters in Santa Clara County voted in the primary election. $p \leq .30$

$\boldsymbol{H}_a$: More than 30% of the registered voters in Santa Clara County voted in the primary election. $p > .30$

Example 2: We wants to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are:

$H_0 : \mu = 2.0$

$H_a : \mu \neq 2.0$

Example 3: We want to test if college students take less than five years to graduate from college, on the average. The null and alternative hypotheses are:

$H_0 : \mu \geq 5$

$H_a : \mu < 5$

Hypothesis Testing: Setting up the Null and Alternative Hypothesis Statements

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-12

Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in Table 6:

Table 6

Statistical Decision	**H_0 is actually...**	
	True	False
Cannot reject H_0	Correct outcome	Type II error
Cannot accept H_0	Type I error	Correct outcome

The four possible outcomes in the table are:

1. The decision is **cannot reject** H_0 when H_0 **is true (correct decision).**
2. The decision is **cannot accept** H_0 when H_0 **is true** (incorrect decision known as a **Type I error**). This case is described as "rejecting a good null". As we will see later, it is this type of error that we will guard against by setting the probability of making such an error. The goal is to NOT take an action that is an error.
3. The decision is **cannot reject** H_0 when, in fact, H_0 **is false** (incorrect decision known as a **Type II error**). This is called "accepting a false null". In this situation you have allowed the status quo to remain in force when it should be overturned. As we will see, the null hypothesis has the advantage in competition with the alternative.
4. The decision is **cannot accept** H_0 when H_0 **is false (correct decision).**

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

$\alpha =$ probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true: rejecting a good null.

$\beta =$probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false. $(1 - \beta)$ is called the **Power of the Test.**

α and β should be as small as possible because they are probabilities of errors.

Statistics allows us to set the probability that we are making a Type I error. The probability of making a Type I error is α. Recall that the confidence intervals in the last section were set by choosing a value called Z_α (or t_α) and the alpha value determined the confidence level of the estimate because it was the probability of the interval failing to capture the true mean (or proportion parameter p). This alpha and that one are the same.

The easiest way to see the relationship between the alpha error and the level of confidence is in Figure 9.

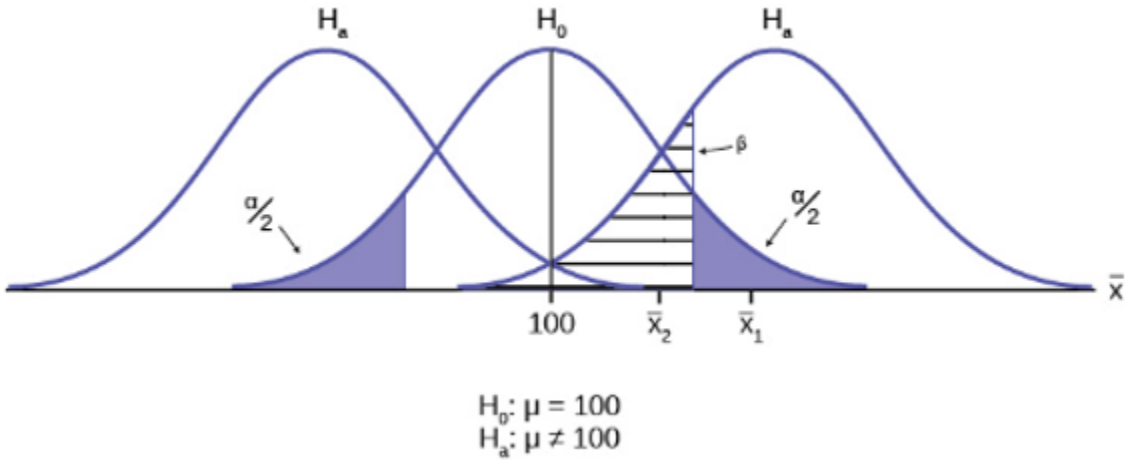


Figure 9

In the center of Figure 9 is a normally distributed sampling distribution marked H_0. This is a sampling distribution of $\bar{X}$ and by the Central Limit Theorem it is normally distributed. The distribution in the center is marked H_0 and represents the distribution for the null hypotheses $H_0 : \mu = 100$. This is the value that is being tested. The formal statements of the null and alternative hypotheses are listed below the figure.

The distributions on either side of the $\boldsymbol{H}_0$ distribution represent distributions that would be true if $\boldsymbol{H}_0$ is false, under the alternative hypothesis listed as $\boldsymbol{H}_a$. We do not know which is true, and will never know. There are, in fact, an infinite number of distributions from which the data could have been drawn if $\boldsymbol{H}_a$ is true, but only two of them are on Figure 9 representing all of the others.

To test a hypothesis, we take a sample from the population and determine if it could have come from the hypothesized distribution with an acceptable level of significance. This level of significance is the alpha error and is marked on Figure 9 as the shaded areas in each tail of the $\boldsymbol{H}_0$ distribution. (Each are actually $\alpha/2$ because the distribution is symmetrical, and the alternative hypothesis allows for the possibility for the value to be either greater than or less than the hypothesized value–called a two-tailed test).

If the sample mean marked as $\bar{X}_1$ is in the tail of the distribution of H_0, we conclude that the probability that it could have come from the H_0 distribution is less than alpha. We consequently state, "the null hypothesis cannot be accepted with (α) level of significance." The truth **may** be that this $\bar{X}_1$ did come from the $\boldsymbol{H}_0$ distribution, but from out in the tail. If this is so, then we have falsely rejected a true null hypothesis and have made a Type I error. What statistics has done is provide an estimate about what we know, and what we control, and that is the probability of us being wrong, α.

We can also see in Figure 9 that the sample mean could be really from an H_a distribution, but within the boundary set by the alpha level. Such a case is marked as $\bar{X}_2$. There is a probability that $\bar{X}_2$ actually came from $\boldsymbol{H}_a$ but shows up in the range of H_0 between the two tails. This probability is the beta error, the probability of accepting a false null.

Our problem is that we can only set the alpha error because there are an infinite number of alternative distributions from which the mean could have come that are not equal to $\boldsymbol{H}_0$. As a result, the statistician places the burden of proof on the alternative hypothesis. That is, we will not reject a null hypothesis unless there is a greater than 90, or 95, or even 99 percent probability that the null is false: the burden of proof lies with the alternative hypothesis. This is why we called this the tyranny of the status quo earlier.

By way of example, the American judicial system begins with the concept that a defendant is "presumed innocent". This is the status quo and is the null hypothesis. The judge will tell the jury that they cannot find the defendant guilty unless the evidence indicates guilt beyond a "reasonable doubt" which is usually defined in criminal cases as 95% certainty of guilt. If the jury cannot accept the null, innocent, then action will be taken, jail time. The burden of proof always lies with the alternative hypothesis. (In civil cases, the jury needs only to be more than 50% certain of wrongdoing to find culpability, called "a preponderance of the evidence").

The example above was for a test of a mean, but the same logic applies to tests of hypotheses for all statistical parameters one may wish to test.

Example: Suppose the null hypothesis, H_0, is Frank's rock climbing is safe.

Type I error: Frank thinks that his rock-climbing equipment may not be safe when, in fact, it really is safe.

Type II error: Frank thinks that his rock-climbing equipment may be safe when, in fact, it is not safe.

$\alpha =$**probability** that Frank thinks his rock-climbing equipment may not be safe when, in fact, it really is safe. $\boldsymbol{\beta} =$ **probability** that Frank thinks his rock-climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock-climbing equipment is safe, he will go ahead and use it.)

This is a situation described as "accepting a false null".

Hypothesis Testing: Type I and Type II Errors

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-13

Check Your Understanding: Hypothesis Testing: Type I and Type II Errors

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-47

Distribution Needed for Hypothesis Testing

Earlier, we discussed sampling distributions. Particular distributions are associated with hypothesis testing. We will perform hypotheses tests of a population mean using a normal distribution or a Student's t-distribution. (Remember, use a Student's t-distribution when the population standard deviation is unknown and the sample size is small, where small is considered to be less than 30 observations.) We perform tests of a population proportion using a normal distribution when we can assume that the distribution is normally distributed. We consider this to be true if the sample proportion, p', times the sample size is greater than 5 and $1 - p'$ times the sample size is also greater than 5. This is the same rule of thumb we used when developing the formula for the confidence interval for a population proportion.

Hypothesis Test for the Mean

Going back to the standardizing formula we can derive the **test statistic** for testing hypotheses concerning means.

$$Z_c = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

The standardizing formula cannot be solved as it is because we do not have μ, the population mean. However, if we substituted in the hypothesized value of the mean, μ_0

in the formula as above, we can compute a Z value. This is the test statistic for a test of hypothesis for a mean and is presented in Figure 10. We interpret this Z value as the associated probability that a sample with a sample mean of $\bar{X}$ could have come from a distribution with a population mean of H_0 and we call this Z value Z_c for "calculated." Figure 10 shows this process.

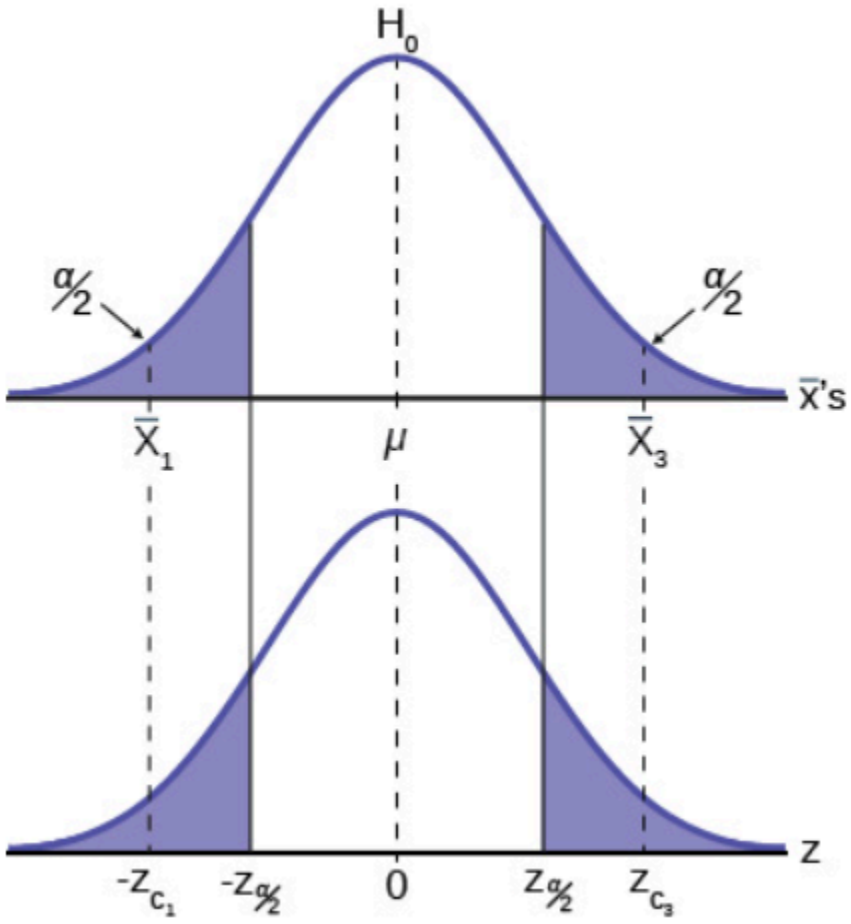


Figure 10

In Figure 10, two of the three possible outcomes are presented. $\bar{X}_1$ and $\bar{X}_3$ are in the tails of the hypothesized distribution of H_0. Notice that the horizontal axis in the top panel is labeled $\overline{X'}$, ¡em¿s. This is the same theoretical distribution of $\overline{X'}s$, the sampling distribution, that the Central Limit Theorem tells us is normally distributed. This is why we can draw it with this shape. The horizontal axis of the bottom panel is labeled Z and is the standard normal distribution. $Z_{\frac{\alpha}{2}}$ and $-Z_{\frac{\alpha}{2}}$, called the **critical values**, are marked on the bottom panel as the Z values associated with the probability the analyst has set as the level of significance in test, (α). The probabilities in the tails of both panels are, therefore, the same.

Notice that for each $\bar{X}$ there is an associated Z_c, called the calculated Z, that comes from solving the equation above. This calculated Z is nothing more than the number of standard deviations that the **hypothesized** mean is from the sample mean. If the sample

mean falls "too many" standard deviations from the hypothesized mean we conclude that the **sample** mean could not have come from the distribution with the hypothesized mean, given our pre-set required level of significance. It **could** have come from $\boldsymbol{H}_0$, but it is deemed just too unlikely. In Figure 10, both $\bar{X}_1$ and $\bar{X}_3$ are in the tails of the distribution. They are deemed "too far" from the hypothesized value of the mean given the chosen level of alpha. If in fact this sample mean it did come from H_0, but from in the tail, we have made a Type I error: we have rejected a good null. Our only real comfort is that we know the probability of making such an error, α, and we can control the size of α.

Figure 11 shows the third possibility for the location of the sample mean, $\bar{x}$. Here the sample mean is within the two critical values. That is, within the probability of $(1-\alpha)$ and we cannot reject the null hypothesis.

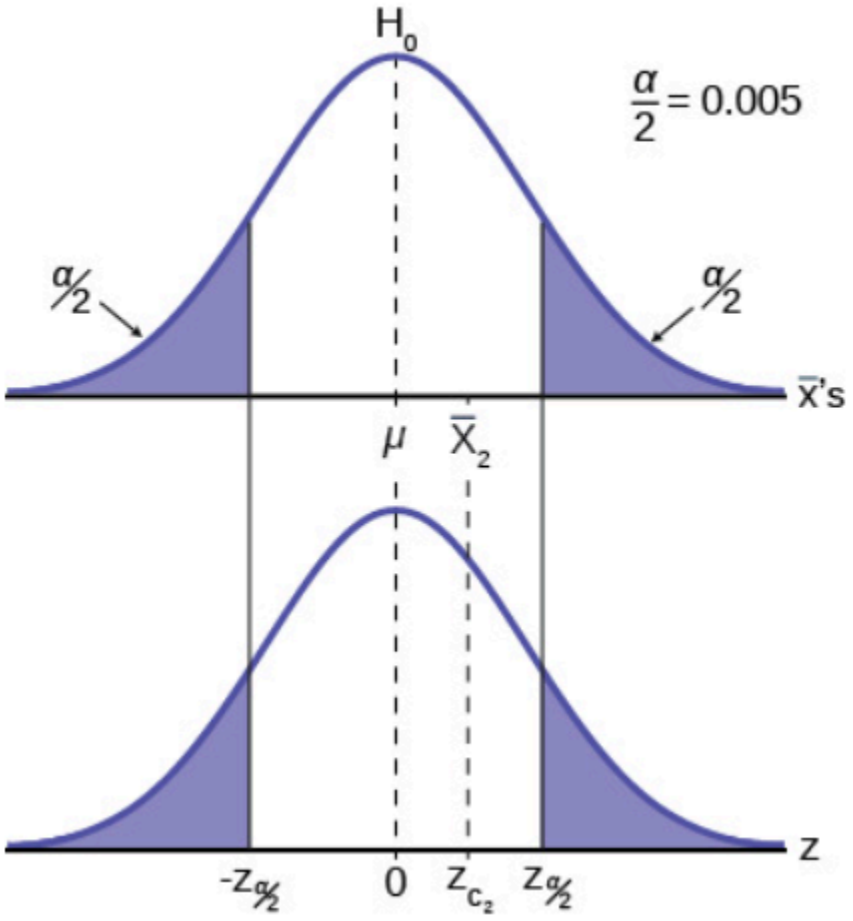


Figure 11

This gives us the decision rule for testing a hypothesis for a two-tailed test:

Table 7

Decision rule: two-tail test
If $\lvert Z_c \rvert < Z_{\frac{\alpha}{2}}$: then DO NOT REJECT H_0
If $\lvert Z_c \rvert > Z_{\frac{\alpha}{2}}$: then REJECT H_0

This rule will always be the same no matter what hypothesis we are testing or what formulas we are using to make the test. The only change will be change the Z_c to the appropriate symbol for the test statistic for the parameter being tested. Stating the decision rule another way: if the sample mean is unlikely to have come from the distribution with the hypothesized mean we cannot accept the null hypothesis. Here we define "unlikely" as having a probability less than alpha of occurring.

Hypothesis testing: Finding Critical Values

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-14

Normal Distribution: Finding Critical Values of Z

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-15

P-Value Approach

An alternative decision rule can be developed by calculating the probability that a sample mean could be found that would give a test statistic larger than the test statistic found from the current sample data assuming that the null hypothesis is true. Here the notion of "likely" and "unlikely" is defined by the probability of drawing a sample with a mean from a population with the hypothesized mean that is either larger or smaller than that found in the sample data. Simply stated, the p-value approach compares the desired significance level, α, to the p-value which is the probability of drawing a sample mean further from the hypothesized value than the actual sample mean. A large p-value calculated from the data indicates that we should not reject the null hypothesis. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it. The relationship between the decision rule of comparing the calculated test statistics, Z_c, and the Critical Value, Z_α, and using the p-value can be seen in Figure 12.

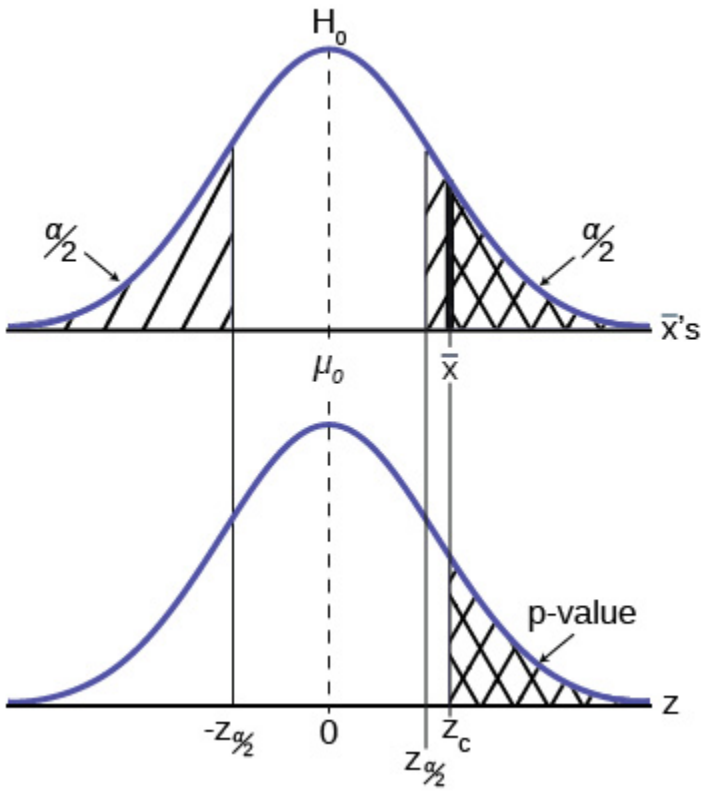


Figure 12

The calculated value of the test Z_c in this example and is marked on the bottom graph of the standard normal distribution because it is a Z value. In this case the calculated value is in the tail and thus we cannot accept the null hypothesis, the associated $\bar{X}$ is just too unusually large to believe that it came from the distribution with a mean of μ_0 with a significance level of α.

If we use the p-value decision rule, we need one more step. We need to find in the standard normal table the probability associated with the calculated test statistic, Z_c. We then compare that to the α associated with our selected level of confidence. In Figure 12 we see that the p-value is less than α and therefore we cannot accept the null. We know that the p-value is less than α because the area under the p-value is smaller than $\alpha/2$. It is important to note that two researchers drawing randomly from the same population may find two different P-values from their samples. This occurs because the P-value is calculated as the probability in the tail beyond the sample mean assuming that the null hypothesis is correct. Because the sample means will in all likelihood be different this will create two different P-values. Nevertheless, the conclusions as to the null hypothesis should be different with only the level of probability of α.

Here is a systematic way to make a decision of whether you cannot accept or cannot reject a null **hypothesis** if using the **p-value** and a **preset or preconceived α** (the "**significance level**"). A preset α is the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the

problem. In any case, the value of α is the decision of the analyst. When you make a decision to reject or not reject $\boldsymbol{H}_0$, do as follows:

- If $\alpha >$p-value, cannot accept $\boldsymbol{H}_0$. The results of the sample data are significant. There is sufficient evidence to conclude that $\boldsymbol{H}_0$ is an incorrect belief and that the **alternative hypothesis,** $\boldsymbol{H}_a$, may be correct.
- If $\alpha \leq$ p-value, cannot reject H_0. The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, $\boldsymbol{H}_a$, may be correct. In this case the status quo stands.
- When you "cannot reject H_0," it does not mean that you should believe that H_0 is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_0. Remember that the null is the status quo, and it takes high probability to overthrow the status quo. This bias in favor of the null hypothesis is what gives rise to the statement "tyranny of the status quo" when discussing hypothesis testing and the scientific method.

Both decision rules will result in the same decision, and it is a matter of preference which one is used.

What is a "P-Value?"

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-16

One and Two-Tailed Tests

The discussion with the figures above was based on the null and alternative hypothesis presented in the first figure. This was called a two-tailed test because the alternative hypothesis allowed that the mean could have come from a population which was either larger or smaller than the hypothesized mean in the null hypothesis. This could be seen by the statement of the alternative hypothesis as $\mu \neq 100$, in this example.

It may be that the analyst has no concern about the value being "too" high or "too" low from the hypothesized value. If this is the case, it becomes a one-tailed test, and all of the alpha probability is placed in just one tail and not split into $\alpha/2$ as in the above case of a two-tailed test. Any test of a claim will be a one-tailed test. For example, a car manufacturer claims that their Model 17B provides gas mileage of greater than 25 miles per gallon. The null and alternative hypothesis would be:

$H_0 : \mu \leq 25$

$H_a : \mu > 25$

The claim would be in the alternative hypothesis. The burden of proof in hypothesis testing is carried in the alternative. This is because failing to reject the null, the status quo, must be accomplished with 90 or 95 percent confidence that it cannot be maintained. Said another way, we want to have only a 5 or 10 percent probability of making a Type I error, rejecting a good null; overthrowing the status quo.

This is a one-tailed test, and all of the alpha probability is placed in just one tail and not split into $\alpha/2$ as in the above case of a two-tailed test.

Figure 13 shows two possible cases and the form of the null and alternative hypothesis that give rise to them.

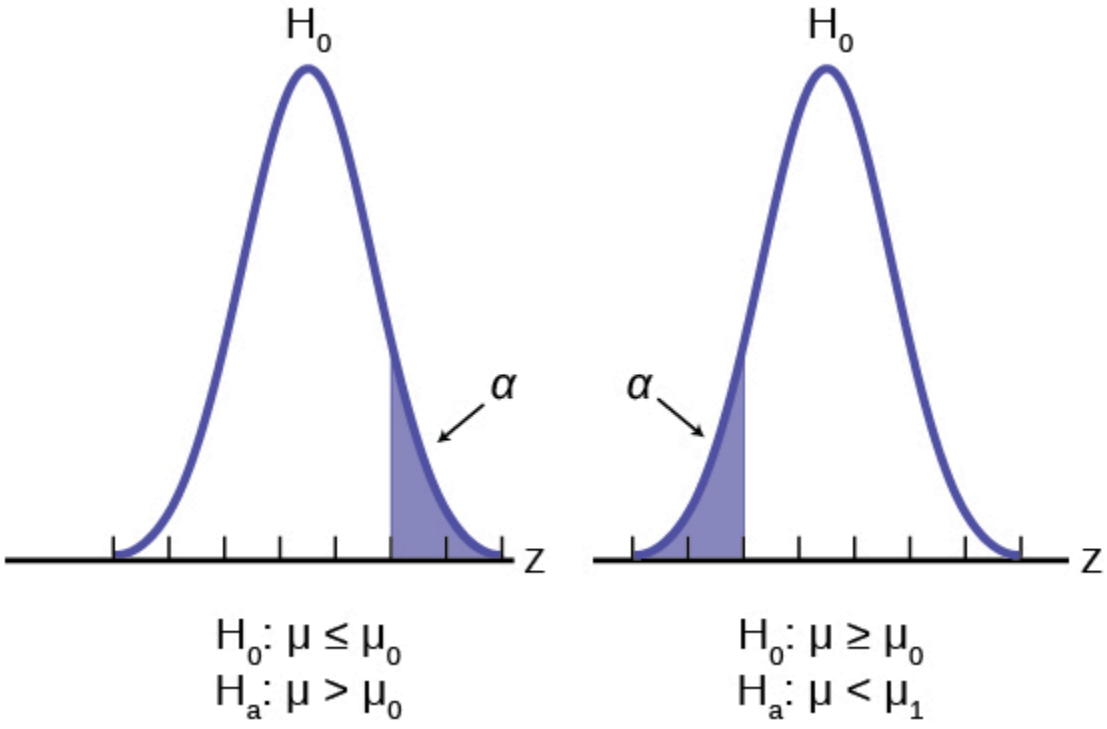


Figure 13

Where μ_0 is the hypothesized value of the population mean.

Table 8: Test Statistics for Test of Means, Varying Sample Size, Population Standard Deviation Known or Unknown

Sample size	Test statistic
$< 30\,(\sigma$ unknown)	$t_c = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
$< 30\,(\sigma$ known)	$Z_c = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$
$> 30\,(\ \sigma$ unknown)	$Z_c = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
$> 30\,(\ \sigma$ known)	$Z_c = \frac{X-\mu_0}{\sigma/\sqrt{n}}$

Effects of Sample Size on Test Statistic

In developing the confidence intervals for the mean from a sample, we found that most often we would not have the population standard deviation, σ. If the sample size were less than 30, we could simply substitute the point estimate for σ, the sample standard deviation, s, and use the student's t-distribution to correct for this lack of information.

When testing hypotheses, we are faced with this same problem and the solution is exactly the same. Namely: If the population standard deviation is unknown, and the sample size is less than 30, substitute s, the point estimate for the population standard deviation, σ, in the formula for the test statistic and use the student's t-distribution. All the formulas and figures above are unchanged except for this substitution and changing the Z distribution to the student's t-distribution on the graph. Remember that the student's t-distribution can only be computed knowing the proper degrees of freedom for the problem. In this case, the degrees of freedom is computed as before with confidence intervals: $df = (n-1)$. The calculated t-value is compared to the t-value associated with the pre-set level of confidence required in the test, $t_{\alpha,df}$ found in the student's t tables. If we do not know σ, but the sample size is 30 or more, we simply substitute s for σ and use the normal distribution.

Table 8 summarizes these rules.

A Systematic Approach for Testing a Hypothesis

A systematic approach to hypothesis testing follows the following steps and in this order. This template will work for all hypotheses that you will ever test.

- Set up the null and alternative hypothesis. This is typically the hardest part of the process. Here the question being asked is reviewed. What parameter is being tested, a mean, a proportion, differences in means, etc. Is this a one-tailed test or two-tailed test?

- Decide the level of significance required for this particular case and determine the critical value. These can be found in the appropriate statistical table. The levels of confidence typical for businesses are 80, 90, 95, 98, and 99. However, the level of significance is a policy decision and should be based upon the risk of making a Type I error, rejecting a good null. Consider the consequences of making a Type I error.
  Next, on the basis of the hypotheses and sample size, select the appropriate test statistic and find the relevant critical value: Z_α, t_α, etc. Drawing the relevant probability distribution and marking the critical value is always big help. Be sure to match the graph with the hypothesis, especially if it is a one-tailed test.
- Take a sample(s) and calculate the relevant parameters: sample mean, standard deviation, or proportion. Using the formula for the test statistic from above in step 2, now calculate the test statistic for this particular case using the parameters you have just calculated.
- Compare the calculated test statistic and the critical value. Marking these on the graph will give a good visual picture of the situation. There are now only two situations:
  - The test statistic is in the tail: Cannot Accept the null, the probability that this sample mean (proportion) came from the hypothesized distribution is too small to believe that it is the real home of these sample data.
  - The test statistic is not in the tail: Cannot Reject the null, the sample data are compatible with the hypothesized population parameter.
- Reach a conclusion. It is best to articulate the conclusion two different ways. First a formal statistical conclusion such as "With a 5 % level of significance we cannot accept the null hypotheses that the population mean is equal to XX (units of measurement)". The second statement of the conclusion is less formal and states the action, or lack of action, required. If the formal conclusion was that above, then the informal one might be, "The machine is broken, and we need to shut it down and call for repairs."

All hypotheses tested will go through this same process. The only changes are the relevant formulas and those are determined by the hypothesis required to answer the original question.

Hypothesis Testing: One Sample Z Test of the Mean (Critical Value Approach)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-17

Hypothesis Testing: t Test for the Mean (Critical Value Approach)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-18

Hypothesis Testing: 1 Sample Z Test of the Mean (Confidence Interval Approach)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-19

Hypothesis Testing: 1 Sample Z Test for Mean (P-Value Approach)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-20

Hypothesis Test for Proportions

Just as there were confidence intervals for proportions, or more formally, the population parameter p of the binomial distribution, there is the ability to test hypotheses concerning p.

The population parameter for the binomial is p. The estimated value (point estimate) for p is p' where $p' = x/n, x$ is the number of successes in the sample and n is the sample size.

When you perform a hypothesis test of a population proportion p, you take a simple random sample from the population. The conditions for a **binomial distribution** must be met, which are: there are a certain number n of independent trials meaning random sampling, the outcomes of any trial are binary, success or failure, and each trial has the same probability of a success 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 $(np' > 5 \text{ and } nq' > 5)$. In this case the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = np \text{ and } \sigma = \sqrt{npq}$. Remember that $q = 1 - p$. There is no distribution that can correct for this small sample bias and thus if these conditions are not met, we simply

cannot test the hypothesis with the data available at that time. We met this condition when we first were estimating confidence intervals for *p*.

Again, we begin with the standardizing formula modified because this is the distribution of a binomial.

$$Z = \frac{p' - p}{\sqrt{\frac{pq}{n}}}$$

Substituting p_0, the hypothesized value of *p*, we have:

$$Z_c = \frac{p' - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

This is the test statistic for testing hypothesized values of *p*, where the null and alternative hypotheses take one of the following forms:

Table 9

Two-tailed test	One-tailed test	One-tailed test
$H_0 : p = p_0$	$H_0 : p \leq p_0$	$H_0 : p \geq p_0$
$\boldsymbol{H}_a : p \neq p_0$	$H_a : p > p_0$	$\boldsymbol{H}_a : p < p_0$

The decision rule stated above applies here also: if the calculated value of Z_c shows that the sample proportion is "too many" standard deviations from the hypothesized proportion, the null hypothesis cannot be accepted. The decision as to what is "too many" is pre-determined by the analyst depending on the level of significance required in the test.

Hypothesis Testing: 1 Proportion using the Critical Value Approach

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-21

Hypothesis Testing with Two Samples

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether

SAT or GRE preparatory courses really help raise their scores. Many business applications require comparing two groups. It may be the investment returns of two different investment strategies, or the differences in production efficiency of different management styles.

To compare two means or two proportions, you work with two groups. The groups are classified either as independent or **matched pairs**. **Independent groups** consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Matched pairs consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions of each group.

Comparing Two Independent Population Means

The comparison of two independent population means is very common and provides a way to test the hypothesis that the two groups differ from each other. Is the night shift less productive than the day shift, are the rates of return from fixed asset investments different from those from common stock investments, and so on? An observed difference between two sample means depends on both the means and the sample standard deviations. Very different means can occur by chance if there is great variation among the individual samples. The test statistic will have to account for this fact. The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula we will see later was developed by Aspin-Welch.

When we developed the hypothesis test for the mean and proportions, we began with the Central Limit Theorem. We recognized that a sample mean came from a distribution of sample means, and sample proportions came from the sampling distribution of sample proportions. This made our sample parameters, the sample means and sample proportions, into random variables. It was important for us to know the distribution that these random variables came from. The Central Limit Theorem gave us the answer: the normal distribution. Our Z and t statistics came from this theorem. This provided us with the solution to our question of how to measure the probability that a sample mean came from a distribution with a particular hypothesized value of the mean or proportion. In both cases that was the question: what is the probability that the mean (or proportion) from our sample data came from a population distribution with the hypothesized value we are interested in?

Now we are interested in whether or not two samples have the same mean. Our question has not changed: Do these two samples come from the same population distribution? We recognize that we have two sample means, one from each set of data, and thus we have two random variables coming from two unknown distributions. To solve the problem, we create a new random variable, the difference between the sample means. This new random variable also has a distribution and, again, the Central Limit Theorem tells us that this new distribution is normally distributed, regardless of the underlying distributions of the original data. A graph may help to understand this concept.

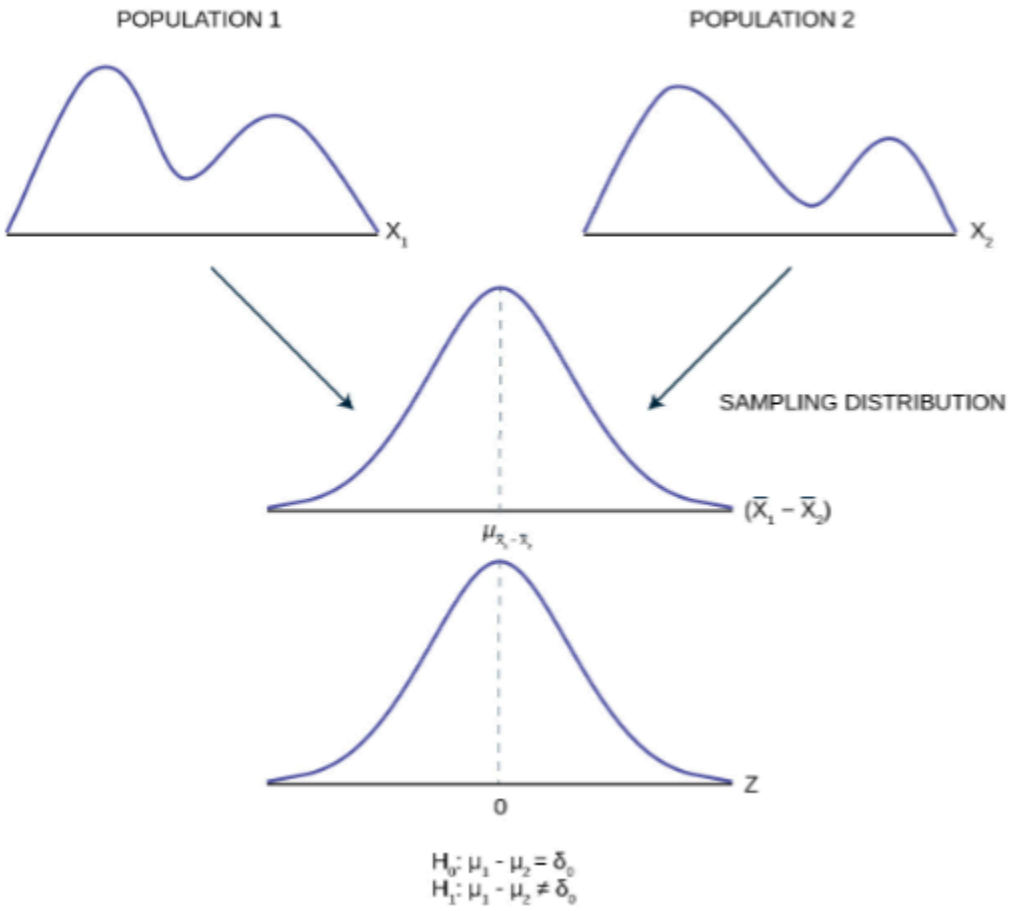


Figure 14

Pictured in Figure 14 are two distributions of data, X_1 and X_2, with unknown means and standard deviations. The second panel shows the sampling distribution of the newly created random variable $\left(\bar{X}_1 - \bar{X}_2\right)$. This distribution is the theoretical distribution of many samples means from population 1 minus sample means from population 2. The Central Limit Theorem tells us that this theoretical sampling distribution of differences in sample means is normally distributed, regardless of the distribution of the actual population data shown in the top panel. Because the sampling distribution is normally distributed, we can develop a standardizing formula and calculate probabilities from the standard normal distribution in the bottom panel, the Z distribution.

The Central Limit Theorem, as before, provides us with the standard deviation of the sampling distribution, and further, that the expected value of the mean of the distribution of differences in sample means is equal to the differences in the population means. Mathematically this can be stated:

$$E\left(\mu_{x_1} - \mu_{x_2}\right) = \mu_1 - \mu_2$$

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test,

we calculate the estimated standard deviation, or **standard error**, of the **difference in sample means,** $\bar{X}_1 - \bar{X}_2$.

The standard error is:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

We remember that substituting the sample variance for the population variance when we did not have the population variance was the technique we used when building the confidence interval and the test statistic for the test of hypothesis for a single mean back in Confidence Intervals and Hypothesis Testing with One Sample. **The test statistic (t-score) is calculated as follows:**

$$t_c = \frac{(\bar{x} - \bar{x}_2) - \delta_0}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Where:

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively
- σ_1 and σ_2 are the unknown population standard deviations
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means. μ_1 and μ_2 are the unknown population means.

The number of **degrees of freedom (df)** requires a somewhat complicated calculation. The *df* are not always a whole number. The test statistic above is approximated by the Student's t-distribution with *df* as follows:

$$df = \frac{\left(\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{(s_1)^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{(s_2)^2}{n_2}\right)^2}$$

When both sample sizes n_1 and n_2 are 30 or larger, the Student's t approximation is very good. If each sample has more than 30 observations, then the degrees of freedom can be calculation as $n_1 + n_2 - 2$.

The format of the sampling distribution, differences in sample means, specifies that the format of the null and alternative hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

$$H_a : \mu_1 - \mu_2 \neq \delta_0$$

Where δ_0 is the hypothesized difference between the two means. If the question is simply

"is there any difference between the means?" then $\delta_0 = 0$ and the null and alternative hypotheses becomes:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

An example of when δ_0 might not be zero is when the comparison of the two groups requires a specific difference for the decision to be meaningful. Imagine that you are making a capital investment. You are considering changing from your current model machine to another. You measure the productivity of your machines by the speed they produce the product. It may be that a contender to replace the old model is faster in terms of product throughout but is also more expensive. The second machine may also have more maintenance costs, setup costs, etc. The null hypothesis would be set up so that the new machine would have to be better than the old one by enough to cover these extra costs in terms of speed and cost of production. This form of the null and alternative hypothesis shows how valuable this particular hypothesis test can be. For most of our work we will be testing simple hypotheses asking if there is any difference between the two-distribution means.

Hypothesis Testing – Two Population Means

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-22

Two Population Means, One Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-23

Two Population Means, Two Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-24

Check Your Understanding: Hypothesis Testing (Two Population Means)

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-48

Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's d is a measure of "effect size" based on the differences between two means. Cohen's d, named for United States statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Table 10

Size of effect	**d**
Small	0.2
Medium	0.5
Large	0.8

Cohen's d is the measure of the difference between two means divided by the pooled standard deviation:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \text{ where } s_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

It is important to note that Cohen's d does not provide a level of confidence as to the magnitude of the size of the effect comparable to the other tests of hypothesis we have studied. The sizes of the effects are simply indicative.

Effect Size for a Significant Difference of Two Sample Means

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-25

Test for Differences in Means: Assuming Equal Population Variances

Typically, we can never expect to know any of the population parameters, mean,

proportion, or standard deviation. When testing hypotheses concerning differences in means we are faced with the difficulty of two unknown **variances** that play a critical role in the test statistic. We have been substituting the sample variances just as we did when testing hypotheses for a single mean. And as we did before, we used a Student's t to compensate for this lack of information on the population variance. There may be situations, however, when we do not know the population variances, but we can assume that the two populations have the same variance. If this is true, then the pooled sample variance will be smaller than the individual sample variances. This will give more precise estimates and reduce the probability of discarding a good null. The null and alternative hypotheses remain the same, but the test statistic changes to:

$$t_c = \frac{(\overline{x_1} - \bar{x}_2) - \delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where ${S_p}^2$ is the pooled variance given by the formula:

$$S_p^2 = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Example: A drug trial is attempted using a real drug and a pill made of just sugar. 18 people are given the real drug in hopes of increasing the production of endorphins. The increase in endorphins is found to be on average 8 micrograms per person, and the sample standard deviation is 5.4 micrograms. 11 people are given the sugar pill, and their average endorphin increase is 4 micrograms with a standard deviation of 2.4. From previous research on endorphins, it is determined that it can be assumed that the variances within the two samples can be assumed to be the same. Test at 5% to see if the population mean for the real drug had a significantly greater impact on the endorphins than the population mean with the sugar pill.

Solution:

First, we begin by designating one of the two groups Group 1 and the other Group 2. This will be needed to keep track of the null and alternative hypotheses. Let us set Group 1 as those who received the actual new medicine being tested and therefore Group 2 is those who received the sugar pill. We can now set up the null and alternative hypothesis as:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

This is set up as a one-tailed test with the claim in the alternative hypothesis that the medicine will produce more endorphins than the sugar pill. We now calculate the test statistic which requires us to calculate the pooled variance, ${S_p}^2$ using the formula above.

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(8 - 4) - 0}{\sqrt{20.4933\left(\frac{1}{18} + \frac{1}{11}\right)}} = 2.31$$

t_α, allows us to compare the test statistic and the critical value.

$$t_\alpha = 1.703 \text{ at } df = n_1 + n_2 - 2 = 18 + 11 - 2 = 27$$

The test statistic is clearly in the tail, 2.31 is larger than the critical value of 1.703, and therefore we cannot maintain the null hypothesis. Thus, we conclude that there is significant evidence at the 95% level of confidence that the new medicine produces the effect desired.

Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations is very unlikely), the following example illustrates hypothesis testing for independent means with known population standard deviations. The sampling distribution for the difference between the means is normal in accordance with the central limit theorem. The random variable is $\bar{X}_1 - \bar{X}_2$. The normal distribution has the following format:

The standard deviation is:

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$$

The test statistic (z-score) is:

$$Z_c = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}$$

Two Population Means, One Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-26

Two Population Means, Two Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-27

Check Your Understanding: Two Population Means

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-49

Matched or Paired Samples

In most cases of economic or business data we have little or no control over the process of how the data are gathered. In this sense the data are not the result of a planned controlled experiment. In some cases, however, we can develop data that are part of a controlled experiment. This situation occurs frequently in quality control situations. Imagine that the production rates of two machines built to the same design, but at different manufacturing plants, are being tested for differences in some production metric such as speed of output or meeting some production specification such as strength of the product. The test is the same in format to what we have been testing, but here we can have matched pairs for which we can test if differences exist. Each observation has its matched pair against which differences are calculated. First, the differences in the metric to be tested between the two lists of observations must be calculated, and this is typically labeled with the letter "d." Then, the average of these matched differences, $\bar{X}_d$ is calculated as is its standard deviation, S_d. We expect that the standard deviation of the differences of the matched pairs will be smaller than unmatched pairs because presumably fewer differences should exist because of the correlation between the two groups.

When using a hypothesis test for matched or paired samples, the following characteristics may be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's t-test for a single population mean with

$n - 1$ degrees of freedom, where n is the number of differences, that is, the number of pairs not the number of observations.

The null and alternative hypotheses for this test are:

$H_0 : \mu_d = 0$

$H_a : \mu_d \neq 0$

The test statistic is:

$t_c = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

Two Population Means, One Tail Test, Matched Sample

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-28

Two Population Means, One Tail test, Matched Sample (Hypothesized Test Different from Zero)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-29

Matched Sample, Two Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-30

Check Your Understanding: Matched Sample, Two Tail Test

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-50

Comparing Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are random samples are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or even perhaps 20 times the size of the sample. This keeps each population from being over-sampled and causing biased results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance in the sampling. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the two population proportions.

Like the case of differences in sample means, we construct a sampling distribution for differences in sample proportions: $(p'_A - p'_B)$ where $p'_A = X_{\frac{A}{n_A}}$ and $p'_B = X_{\frac{B}{n_B}}$ are the sample proportions for the two sets of data in question. X_A and X_B are the number of successes in each sample group respectively, and n_A and n_B are the respective sample sizes from the two groups. Again, we go the Central Limit Theorem to find the distribution of this sampling distribution for the differences in sample proportions. And again, we find that this sampling distribution, like the one's past, are normally distributed as proved by the Central Limit Theorem, as see in Figure 15.

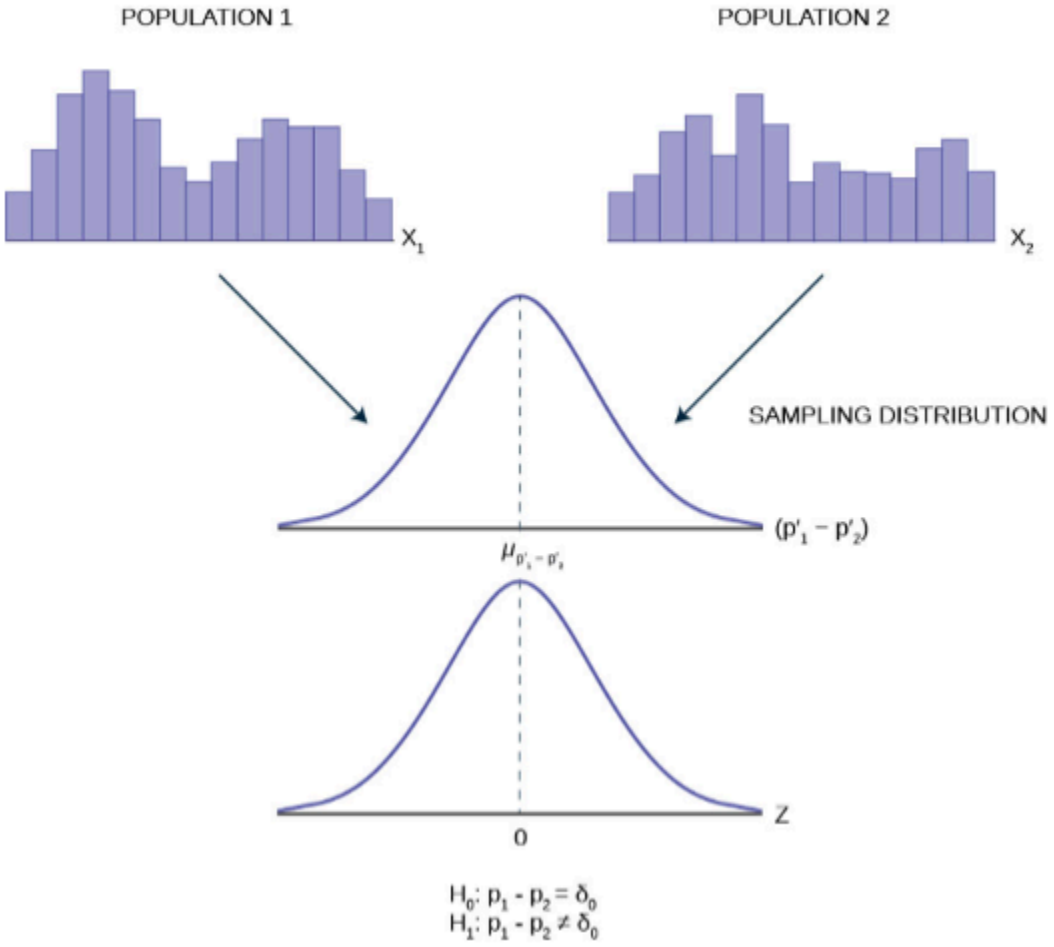


Figure 15

Generally, the null hypothesis allows for the test of a difference of a particular value, δ_0, just as we did for the case of difference in means.

$$H_0 : p_1 - p_2 = \delta_0$$

$$H_1 : p_1 - p_2 \neq \delta_0$$

Most common, however, is the test that the two proportions are the same. That is,

$$H_0 : p_A = p_B$$

$$H_a : p_A \neq p_B$$

To conduct the test, we use a pooled proportion, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

The test statistic (z-score) is:

$$Z_c = \frac{(p'_A - p'_B) - \delta_0}{\sqrt{p_c(1-p_c)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$$

Where δ_0 is the hypothesized differences between the two proportions and p_c is the pooled variance from the formula above.

Two Population Proportions, One Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-31

Two Population Proportion, Two Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-32

Check Your Understanding: Comparing Two Independent Population Proportions

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-51

The Chi-Square Distribution

Have you ever wondered if lottery winning numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.

In this section, you will learn the three major applications of the chi-square distribution:

1. The test of a single variance, which tests variability, such as in the coffee example
2. The goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
3. the test of independence, which determines if events are independent, such as in the movie example

Facts About the Chi-Square Distribution

The notation for the **chi-square distribution** is:

$$\chi \sim X^2_{df}$$

Where df = degrees of freedom which depends on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{d(df)}$.

The random variable is shown as χ^2.

The random variable for a chi-square distribution with *k* degrees of freedom is the sum of *k* independent, squared standard normal variables.

$$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$$

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df.

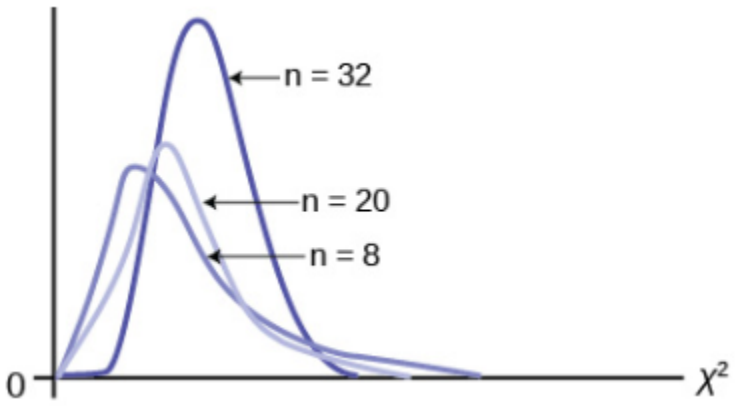


Figure 16

3. The test statistic for any test is always greater than or equal to zero.

4. When $df > 90$, the chi-square curve approximates the normal distribution. For

$\chi \sim \chi^2_{1,000}$ the mean, $\mu = df = 1,000$ and the standard deviation, $\sigma = \sqrt{2(1,000)} = 44.7$. Therefore, $\chi \sim N(1,000, 44.7)$, approximately.

5. The mean, μ, is located just to the right of the peak.

Test of a Single Variance

Thus far our interest has been exclusively on the population parameter μ or it is counterpart in the binomial, *p*. Surely the mean of a population is the most critical piece of information to have, but in some cases, we are interested in the variability of the outcomes of some distribution. In almost all production processes quality is measured not only by how closely the machine matches the target, but also the variability of the process. If one were filling bags with potato chips not only would there be interest in the average weight of the bag, but also how much variation there was in the weights. No one wants to be assured that the average weight is accurate when their bag has no chips. Electricity voltage may meet some average level, but great variability, spikes, can cause serious damage to electrical machines, especially computers. I would not only like to have a high mean grade in my classes, but also low variation about this mean. In short, statistical tests concerning the variance of a distribution have great value and many applications.

A **test of a single variance** assumes that the underlying distribution is **normal.** The null and alternative hypotheses are stated in terms of the **population variance**. The test statistic is:

$$\chi^2_c = \frac{(n-1)s^2}{\sigma^2_0}$$

Where:

- $n =$ the total number of observations in the sample data
- $s^2 =$ sample variance
- $\sigma^2_0 =$ hypothesized value of the population variance
- $H_0 : \sigma^2 = \sigma^2_0$
- $H_a : \sigma^2 \neq \sigma^2_0$

You may think of *s* as the random variable in this test. The number of degrees of freedom is $df = n - 1$. A test of a single variance may be right-tailed, left-tailed, or two-tailed. The example below will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

Example: Math instructors are not only interested in how their students do on exams,

on average, but how the exam scores vary. To many instructors, the variance (or standard deviation) may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Solution: Even though we are given the population standard deviation, we can set up the test using the population variance as follows:

$H_0 : \sigma^2 \leq 5^2$

$H_a : \sigma^2 > 5^2$

Single Population Variances, One-Tail Test

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-33

Check Your Understanding: Test of a Single Variance

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-52

Goodness-Of-Fit Test

In this type of hypothesis test, you determine whether the data "**fit**" a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of -fit test is:

$\sum_k \frac{(O-E)^2}{E}$

Where:

- O = **observed values** (data)
- E = **expected values** (from theory)
- k = the number of different data cells or categories

The observed values are the data values, and the expected values are the values you would expect to get it the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is $df = (\text{ number of categories } - 1)$.

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note: The number of expected values inside each cell needs to be at least five in order to use this test.

Chi-Square Statistic for Hypothesis Testing

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-34

Chi-Square Goodness-of-Fit Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-35

Check Your Understanding: Goodness-of-Fit Test

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-53

Test of Independence

Tests of independence involve using a **contingency table** of observed (data) values.

The test statistic of a **test of independence** is similar to that of a goodness-of-fit test:

$$\sum_{i \bullet j} \frac{(O-E)^2}{E}$$

Where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$

A test of independence determines whether two factors are independent or not.

Note: The expected value inside each cell needs to be at least five in order for you to use this test.

The test of independence is always right tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is:

$$df = (\text{ Number of columns } - 1\ )\ (\text{number of rows } - 1\ )$$

The following formula calculates the expected number (E):

$$E = \frac{(\text{ row total })(\text{ column total })}{\text{total number surveyed}}$$

Simple Explanation of Chi-Squared

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-36

Chi-Square Test for Association (independence)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-37

Check Your Understanding: Test of Independence

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-54

Test for Homogeneity

The goodness-of-fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the **test of homogeneity**, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

Note: The expected value inside each cell needs to be at least five in order for you to use this test.

Hypotheses

H_0 :: The distribution of the two populations is the same.

H_a : The distributions of the two population are not the same.

Test Statistic

Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of Freedom (df)

$$df = \text{number of columns} - 1$$

Requirements

All values in the table must be greater than or equal to five

Common uses

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Introduction to the Chi-Square Test for Homogeneity

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-38

Check Your Understanding: Test for Homogeneity

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-55

Comparison of the Chi-Square Tests

Above the χ^2 test statistic was used in three different circumstances. The following bulleted list is a summary of which χ^2 test is the appropriate one to use in different circumstances.

Goodness-of-fit: Use the goodness-of-fit test to decide whether a population with an unknown distribution "fits" a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single

population. Goodness-of-Fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are:

H_0 : The population fits given distribution

$\boldsymbol{H}_a$: The population does not fit the given distribution

Independence: Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated (independent) or related (dependent). The null and alternative hypotheses are:

H_0 :The two variables (factors) are independent

$\boldsymbol{H}_a$: The two variables (factors) are dependent

Homogeneity: Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are:

H_0 : The two populations follow the same distribution

$\boldsymbol{H}_a$: The two populations have different distributions

F Distribution and One-Way ANOVA

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies in several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her upbringing. A consumer looking for a new car might compared the average gas mileage of several models.

For hypothesis tests comparing averages among more than two groups, statisticians have developed a method called "**Analysis of Variance**" (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or **one-way ANOVA**. You will also study the *F* distribution, used for one-way ANOVA, and the test for differences between two variances. This is just a very brief overview of one-way ANOVA. One-Way ANOVA, as it is presented here, relies heavily on a calculator or computer.

Test of Two Variances

This chapter introduces a new probability density function, the F distribution. This distribution is used for many applications including ANOVA and for testing equality across multiple means. We begin with the F distribution and the test of hypothesis of

differences in variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. In order for a lid to fit a container, the variation in the lid and the container should be approximately the same. A supermarket might be interested in the variability of check-out times for two checkers. In finance, the variance is a measure of risk and thus an interesting question would be to test the hypothesis that two different investment portfolios have the same variance, the volatility.

In order to perform a F test of two variances, it is important that the following are true:

1. The populations from which the two samples are drawn are approximately normally distributed.
2. The two populations are independent of each other.

Unlike most other hypothesis tests in this Module, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, or close, the test can give a biased result for the test statistic.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the unknown population variances and s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio:

$$\boldsymbol{F} = \frac{\left[\frac{s_1^2}{\sigma_1^2}\right]}{\left[\frac{s_2^2}{\sigma_2^2}\right]}$$

F has the distribution $F \sim F\left(n_1 - 1, n_2 - 1\right)$

Where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F Ratio, test statistic, becomes

$$F_c = \frac{\left[\frac{s_1^2}{\sigma_1^2}\right]}{\left[\frac{s_2^2}{\sigma_2^2}\right]} = \frac{s_1^2}{s_2^2}$$

The various forms of the hypotheses tested are:

Table 11

Two-Tailed Test	One-Tailed Test	One-Tailed Test
$H_0 : {\sigma_1}^2 = {\sigma_2}^2$	$\mathrm{H}_0 : \sigma_1^2 \leq \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$\mathrm{H}_1 : \sigma_1^2 > \sigma_2^2$	$\mathrm{H}_1 : \sigma_1^2 < \sigma_2^2$

A more general form of the null and alternative hypothesis for a two tailed test would be:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \delta_0$$

$$H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq \delta_0$$

Where if $\delta_0 = 1$ it is a simple test of the hypothesis that the two variances are equal. This form of the hypothesis does have the benefit of allowing for tests that are more than for simple differences and can accommodate tests for specific differences as we did for differences in means and differences in proportions. This form of the hypothesis also shows the relationship between the F distribution and the χ^2: the F is a ratio of two chi squared distributions a distribution we saw in the last chapter. This is helpful in determining the degrees of freedom of the resultant F distribution.

If the two populations have equal variances, then $s_1^2 \text{ and } s_2^2$ are close in value and the test statistic, $F_c = \frac{s_1^2}{s_2^2}$ is close to one. But if the two population variances are very different, $s_1^2 \text{ and } s_2^2$ tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{s_1^2}{s_2^2}$ to be greater than one. If $s_1^2 \text{ and } s_2^2$ are far apart, then $F_c = \frac{s_1^2}{s_2^2}$ is a large number.

Therefore, if F is close to one, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than one, then the evidence is against the null hypothesis. In essence, we are asking if the calculated F statistic, test statistic, is significantly different from one.

To determine the critical points, we have to find $F_{\alpha,df1,df2}$. This F table has values for various levels of significance from 0.1 to 0.001 designated as "p" in the first column. To find the critical value choose the desired significance level and follow down and across to find the critical value at the intersection of the two different degrees of freedom. The F distribution has two different degrees of freedom, one associated with the numerator, $df1$, and one associated with the denominator, $df2$ and to complicate matters, the F distribution is not symmetrical and changes the degree of skewness as the degrees of freedom of change. The degrees of freedom in the numerator is $n_1 - 1, \text{ where } n_1$ is the sample size for group 1, and the degrees of freedom in the denominator is $n_2 - 1$, where n_2 is the sample size for group 2. $F_{\alpha,df1,df2}$ will give the critical value on the **upper** end of the F distribution.

To find the critical value for the lower end of the distribution, reverse the degrees of freedom and divide the F-value from the table into one.

- Upper tail critical value: $F_{\alpha,df1,df2}$

- Lower tail critical value: $1/F_{\alpha,df2,df1}$

When the calculated value of F is between the critical values, not in the tail, we cannot reject the null hypothesis that the two variances came from a population with the same variance. If the calculated F-value is in either tail we cannot accept the null hypothesis just as we have been doing for all of the previous tests of hypothesis.

An alternative way of finding the critical values of the F distribution makes the use of the F-table easier. We note in the F-table that all the values of F are greater than one therefore the critical F value for the left-hand tail will always be less than one because to find the critical value on the left tail we divide an F value into the number one as shown above. We also note that if the sample variance in the numerator of the test statistic is larger than the sample variance in the denominator, the resulting F value will be greater than one. The shorthand method for this test is thus to be sure that the larger of the two sample variances is placed in the numerator to calculate the test statistic. This will mean that only the right-hand tail critical value will have to be found in the F-table.

Hypothesis Test Two Population Variances

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-39

Check Your Understanding: F Distribution

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-56

One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test actually uses **variances** to

help determine if the means are equal or not. In order to perform a one-way ANOVA test, there are five basic **assumptions** to be fulfilled:

1. Each population from which a sample is taken is assumed to be normal.
2. All samples are randomly selected and independent.
3. The populations are assumed to have **equal standard deviations (or variances).**
4. The factor is a categorical variable.
5. The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is simply that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are *k* groups:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$

H_a : At least two of the group means $\mu_1, \mu_2, \mu_3, \ldots, \mu_k$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), $H_0 : \mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of the populations.

If the null hypothesis is false, then the variance of the combined data is larger which is caused by the different means as shown in the second graph (green box plots).

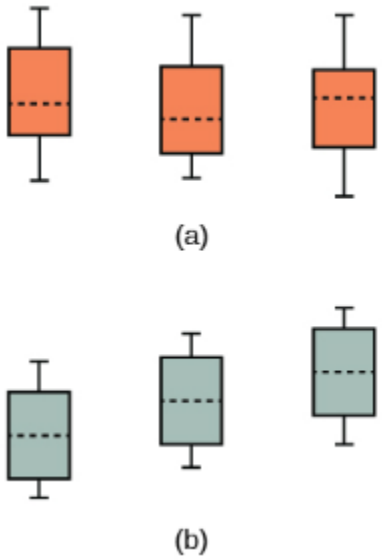


Figure 17: (a) H_0 is true. All means are the same; the differences are due to random variation. (b) H_0 is not true. All means are not the same; the differences are too large to be due to random variation.

ANOVA

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-40

The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the F distribution, invented by George Snedecor but named in honor of Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom: one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

To calculate the **F ratio**, two estimates of the variance are made.

1. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same). If the samples are different sizes, the variances between samples is weighted to account for the different sample sizes. The variances is also called **variance due to treatment or explained variation.**

2. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a **pooled variance**). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- SS_{between} = the **sum of squares** that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance.

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted.

MS means "**mean square.**" MS_{between} is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- $n_j =$ the size of the j_{th} group
- $s_j =$ the sum of the values in the j_{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum s = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{n}$
- Explains variation: sum of squares representing variation among the different samples: $SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j} - \frac{(\sum s_j)^2}{n} \right.$
- Unexplained variation: sum of squares represent variation within samples due to change: $SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$
- *d f*'s for different groups (*d f's* for the numerator): $df = k - 1$
- Equation for errors within samples (*d f's* for the denominator): $df_{\text{within}} = n - k$

- Mean square (variance estimate) explained by the different groups:
  $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$
- Mean square (variance estimate) that is due to chance (unexplained):
  $(\text{unexplained}):\ MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$

MS_{between} and MS_{within} can be written as follows:

- $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$
- $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{SS_{\text{within}}}{n-k}$

The one-way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

Note: The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ration or F Statistic

$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced form the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}. Then the

F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat, and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot s_{-}^2}{s_{\text{pooled}}^2}$$

Where:

- n = the sample size
- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- s_{pooled}^2 = the mean of the sample variances (pooled variance)
- $s_{\bar{x}}^2$ = the variance of the sample means.

Data are typically put into a table for easy viewing. One-Way ANOVA results are often displayed is this manner by computer software.

Table 12

Source of variation	**Sum of squares (*SS*)**	**Degrees of freedom (*df*)**	**Mean square (*MS*)**	***F***
Factor (Between)	*SS*(Factor)	$k - 1$	*MS*(Factor) = $\text{SS(Factory } /(k-1)$	F = $\text{MS(Factor)/MS(Error)}$
Error (Within)	*SS*(Error)	$n - k$	*MS*(Error) = $SS(\text{ Error}) /(n-k)$	
Total	*SS*(Total)	$n - 1$		

Example: Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in Table 13.

Table 13

Plan 1: $n_1 = 4$	**Plan 2:** $n_2 = 3$	**Plan 3:** $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

$s_1 = 16.5, s_2 = 15, s_3 = 15.5$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1+s_2+s_3)^2}{10}$$

Where $n_1 = 4, n_2 = 3, n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(15.5)^2}{3} - \frac{(16.5+15+15.5)^2}{10}$$

$$SS_{\text{between}} = 2.2458$$

$$S_{\text{total}} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right) - \frac{5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2}{10}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9$$

$$SS_{\text{total}} = 23.1$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$$= 23.1 - 2.2458$$

$$SS_{\text{within}} = 20.8542$$

Table 14

Source of variation	**Sum of squares (*SS*)**	**Degrees of freedom (*df*)**	**Mean square (*MS*)**	***F***
Factor (Between)	*SS*(Factor) =*SS*(Between) = 2.2458	*k* – 1 = 3 groups – 1 = 2	*MS*(Factor) = *SS*(Factor)/(*k* – 1) = 2.2458/2 = 1.1229	*F* = *MS*(Factor)/*MS*(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	*SS*(Error) = *SS*(Within) = 20.8542	*n* – *k* = 10 total data – 3 groups = 7	*MS* (Error) = *SS*(Error)/(*n* – *k*) = 20.8542/7 = 2.9792	
Total	*SS*(Total) = 2.2458 + 20.8542 = 23.1	*n* – *1* = 10 total data – 1 = 9		

The one-way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim F_{df(\text{ num }),df(\text{ denom })}$

Where $df(\text{ num }) = df_{\text{between}}$ and $df(\text{ denom }) = df_{\text{within}}$

The mean for the F distribution is $\mu = \frac{df(\text{ num })}{df(\text{ denom })-2}$

Calculating SST (total sum of squares)

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-41

Calculating SS_{within} and SS_{between}

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-42

Hypothesis Test with F-Statistic

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-43

Check Your Understanding: The F Distribution and the F-Ratio

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-57

Facts About the F Distribution

1. The curve is not symmetrical but skewed to the right.
2. There is a different curve for each set of degrees of freedom.
3. The F statistic is greater than or equal to zero.
4. As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal as can be seen in Figure 18. Remember that the F cannot ever be less than zero, so the distribution does not have a tail that goes to infinity on the left as the normal distribution does.
5. Other uses for the F distribution include comparing two variances and two-way Analysis of Variances. Two-Way Analysis is beyond the scope of this section.

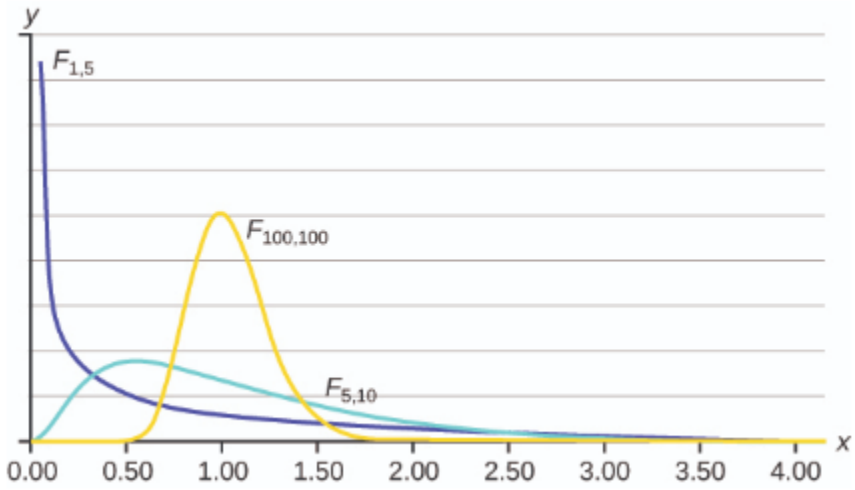


Figure 18

COMPUTE AND INTERPRET SIMPLE LINEAR REGRESSION BETWEEN TWO VARIABLES

Linear Regression and Correlation

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

This example may or may not be tied to a model, meaning that some theory suggested that a relationship exists. This link between a cause and an effect is the foundation of the scientific method and is the core of how we determine what we believe about how the world works. Beginning with a theory and developing a model of the theoretical relationship should result in a prediction, what we have called a hypothesis earlier. Now the hypothesis concerns a full set of relationships.

The foundation of all model building is perhaps the arrogant statement that we know what caused the result we see. This is embodied in the simple mathematical statement of the functional form that $y = f(x)$. The response, Y, is caused by the stimulus, X. Every model will

eventually come to this final place, and it will be here that the theory will live or die. Will the data support this hypothesis? If so, then fine, we shall believe this version of the world until a better theory comes to replace it. This is the process by which we moved from flat earth to round earth, from earth-center solar system to sun-center solar system, and on and on.

In this section we will begin with correlation, the investigation of relationships among variables that may or may not be founded on a cause-and-effect model. The variables simply move in the same, or opposite, direction. That is to say, they do not move randomly. Correlation provides a measure of the degree to which this is true. From there we develop a tool to measure cause and effect relationships, regression analysis. We will be able to formulate models and tests to determine if they are statistically sound. If they are found to be so, then we can use them to make predictions: if as a matter of policy, we changed the value of this variable what would happen to this other variable? If we imposed a gasoline tax of 50 cents per gallon how would that effect the carbon emissions, sales of Hummers/Hybrids, use of mass transit, etc.? The ability to provide answers to these types of questions is the value of regression as both a tool to help us understand our world and to make thoughtful policy decisions.

The Correlation Coefficient r

As we begin this section, we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable. It may be from two samples, but it is still a univariate variable. The type of data described for any model of cause and effect is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

Data can be classified into three broad categories: time series data, cross-section data, and panel data. Time series data measures a single unit of observation; say a person, or a company or a country, as time passes. What are measures will be at least two characteristics, say the person's income, the quantity of a particular good they buy and the price they paid. This would be three pieces of information in one time period, say 1985. If we followed that person across time we would have those same pieces of information for 1985, 1986, 1987, etc. This would constitute a time series data set.

A second type of data set is for cross-section data. Here the variation is not across time for a single unit of observation, but across units of observation during one point in time. For a particular period of time, we would gather the price paid, amount purchased, and income of many individual people.

A third type of data set is panel data. Here a panel of units of observation is followed across time. If we take our example from above, we might follow 500 people, the unit of observation, through time, ten years, and observe their income, price paid and quantity of the good purchased. If we had 500 people and data for ten years for price, income and quantity purchased we would have 15,000 pieces of information. These types of data sets are very expensive to construct and maintain. They do, however, provide a tremendous amount of information that can be used to answer very important questions. As an example, what is the effect on the labor force participation rate of women as their family of origin, mother and father, age? Or are

there differential effects on health outcomes depending upon the age at which a person started smoking? Only panel data can give answers to these and related questions because we must follow multiple people across time.

Beginning with a set of data with two independent variables we ask the question: are these related? One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimensions. Three dimensions are possible on a flat piece of paper but become very hard to fully conceptualize. Of course, more than three dimensions cannot be graphed although the relationships can be measured mathematically.

To provide mathematical precision to the measurement of what we see we use the correlation coefficient. The correlation tells us something about the co-movement of two variables, but nothing about why this movement occurred. Formally, correlation analysis assumes that both variables being analyzed are independent variables. This means that neither one causes the movement in the other. Further, it means that neither variable is dependent on the other, or for that matter, on any other variable. Even with these limitations, correlation analysis can yield some interesting results.

The **correlation coefficient**, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, the statistic, r, developed by Karl Pearson in the early 1900s, is an estimate of the population correlation and is defined mathematically as:

$$r = \frac{\frac{1}{n-1}\sum\left(X_{1i}-\bar{X}_1\right)\left(X_{2i}-\bar{X}_2\right)}{s_{x1}s_{x2}}$$

OR

$$r = \frac{\sum x_{1i}X_{2i}-n\bar{X}_1-\bar{X}_2}{\sqrt{\left(\sum X_{1i}^2-nX_1^2\right)\left(\sum X_{2i}^2-n\bar{X}_2^2\right.}}$$

Where S_{x1} and S_{x2} are the standard deviations of the two independent variables X_1 and $X_2, \bar{X}_1$ and $\bar{X}_2$ are the sample means of the two variables, and X_{1i} and X_{2i} are the individual observations of X_1 and X_2. The correlation coefficient r ranges in value from -1 to 1. The second equivalent formula is often used because it may be computationally easier. As scary as these formulas look, they are really just the ratio of the covariance between the two variables and the product of their two standard deviations. That is to say, it is a measure of relative variances.

In practice all correlation and regression analysis will be provided through computer software designed for these purposes. Anything more than perhaps one-half a dozen observations creates immense computational problems. It was because of this fact that correlation, and even more so, regression, were not widely used research tools until after the advent of "computing machines." Now the computing power required to analyze data using regression packages is deemed almost trivial by comparison to just a decade ago.

To visualize any **linear** relationship that may exist review the plot of a scatter diagrams of the

standardized data. Figure 19 presents several scatter diagrams and the calculated value of r. In panels (a) and (b) notice that the data generally trend together, (a) upward and (b) downward. Panel (a) is an example of a positive correlation and panel (b) is an example of a negative correlation, or relationship. The sign of the correlation coefficient tells us if the relationship is a positive or negative (inverse) one. If all the values of X_1 and X_2 are on a straight line the correlation coefficient will be either 1 or -1 depending on whether the line has a positive or negative slope and the closer to one or negative one the stronger the relationship between the two variables. BUT ALWAYS REMEMBER THAT THE CORRELATION COEFFICIENT DOES NOT TELL US THE SLOPE.

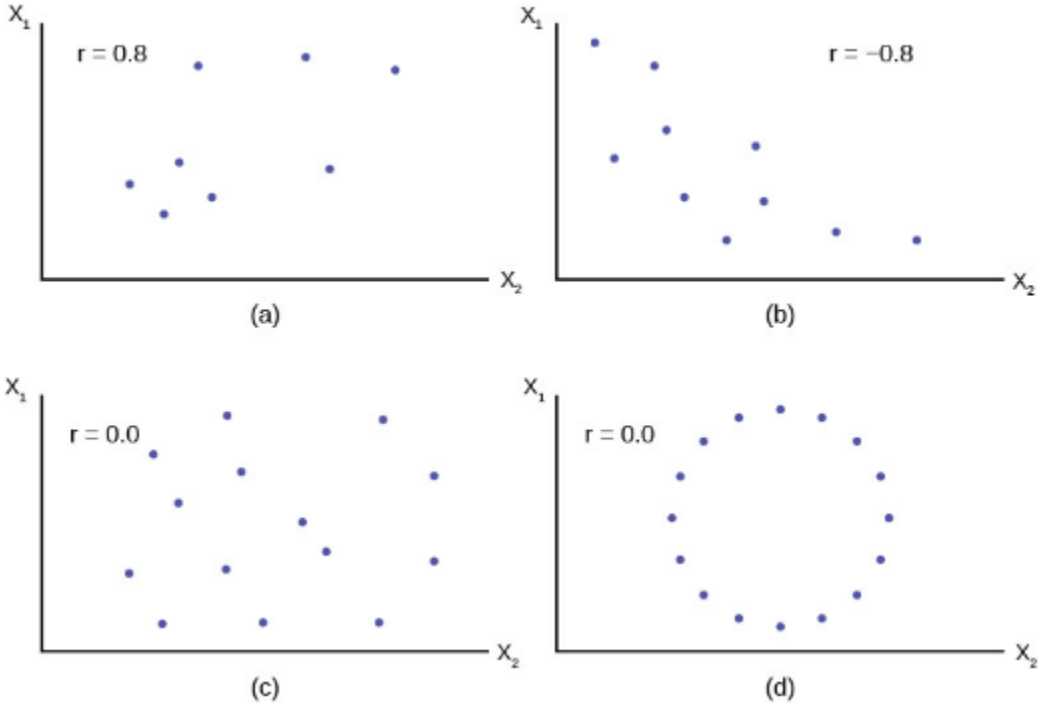


Figure 19

Remember, all the correlation coefficient tells us is whether or not the data are linearly related. In panel (d) the variables obviously have some type of very specific relationship to each other, but the correlation coefficient is zero, indicating no **linear** relationship exists.

If you suspect a linear relationship between X_1 and X_2 then r can measure how strong the linear relationship is.

What the VALUE of *r* tells us:

- The value of *r* is always between -1 and $+1$: $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates that strength of the **linear** relationship between

X_1 and X_2. Values of *r* close to -1 or $+1$ indicate a stronger linear relationship between X_1 and X_2.

- If $r = 0$ there is absolutely no linear relationship between X_1 and X_2 **(no linear correlation).**
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line: ANY straight line no matter what the slope. Of course, in the real world, this will not generally happen.

What the SIGN of *r* tells us

- A positive value of *r* means that when X_1 increases, X_2 tends to increase and when X_1 decreases, X_2 tend to decrease **(positive correlation).**
- A negative value of *r* means that when X_1 increases, X_2 tends to decrease and when X_1 decreases, X_2 tends to increase **(negative correlation).**

Note: Strong correlation does not suggest that X_1 causes X_2 or X_2 causes X_1. We say, "correlation does not imply causation."

Bivariate Relationship Linearity, Strength, and Direction

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-44

Check Your Understanding: The Correlation Coefficient r

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-58

Calculating Correlation Coefficient r

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-45

Example: Correlation Coefficient Intuition

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-46

Linear Equations

Linear regression to two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx$$

Where *a* and *b* are constant numbers.

The variable **x is the independent variable,** and **y is the dependent variable**. Another way to think about this equation is a statement of cause and effect. The X variable is the cause, and the Y variable is the hypothesized effect. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

The graph of a linear equation of the form $y = a + bx$ is a straight line. Any line that is not vertical can be described by this equation.

Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, *b* = slope and *a* = y-intercept. From algebra recall that the slop is a number that described the steepness of a line, and the y-intercept is the y coordinate of

the point $(0, a)$ where the line crosses the y-axis. From calculus the slope is the first derivative of the function. For a linear function, the slope is $\frac{dy}{dx} = b$ where we can read the mathematical expression as "the change in $y(dy)$ that results from a change in $x(dx) = b * dx$

Figure 20: Three possible graphs of $y = a + bx$. (a) if b > 0, the line slopes upward to the right. (b) if b=0, the line is horizontal. (c) if b < 0, the line slopes downward to the right.

The Regression Equation

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and econometricians. This presents us with some difficulties in economic analysis because many of our theoretical models are nonlinear. The marginal cost curve, for example, is decidedly nonlinear as is the total cost function, if we are to believe in the effect of specialization of labor and the Law of Diminishing Marginal Product. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Where β_0 is the intercept, β_i' 's are the slope between Y and the appropriate X_i, and ε (pronounced epsilon), is the error term that captures errors in measurement of Y and the effect on Y of any variables missing from the equation that would contribute to explaining variations in Y. This equation is the theoretical population equation and therefore uses Greek letters. The equation we will estimate will have the Roman equivalent symbols. This is parallel to how we kept track of the population parameters and sample parameters before. The symbol for the population mean was μ and for the sample mean $\bar{X}$ and for the population standard deviation was σ and for the sample standard deviation was *s*. The equation that will be estimated with a sample of data for two independent variables will thus be:

$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

As with our earlier work with probability distributions, this model works only if certain assumptions hold. These are that the Y is normally distributed, the errors are also normally distributed with a mean of zero and a constant standard deviation, and that the error terms are independent of the size of X and independent of each other.

Assumptions of the Ordinary Least Squares Regression Model

Each of these assumptions needs a bit more explanation. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The independent variables, x_i, are all measured without error, and are fixed numbers that are independent of the error term. This assumption is saying in effect that Y is deterministic, the result of a fixed component "X" and a random error component "ε."
2. The error term is a random variable with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased as an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. Figure 21 shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.
3. Error terms should be normally distributed. This can be seen in Figure 21 by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
4. The independent variables are independent of Y but are also assumed to be independent of the other X variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. Take the case of a simple supply curve where quantity supplied is theoretically related to the price of the product and the prices of inputs. There may be multiple inputs that may over time move together from general inflationary pressure. The input prices will therefore violate this assumption of regression analysis. This condition is called multicollinearity.
5. The error terms are uncorrelated with each other. This situation arises from an effect on one error term from another error term. While not exclusively a time series problem, it is here that we most often see this case. An X variable in time period one

has an effect on the Y variable, but this effect then has an effect in the next time period. This effect gives rise to a relationship among the error terms. This case is called autocorrelation, "self-correlated." The error terms are now not independent of each other, but rather have their own effect on subsequent error terms.

Figure 21 shows the case where the assumptions of the regression model are being satisfied. The estimated line is $\widehat{y} = a + bx$. Three values of X are shown. A normal distribution is placed at each point where X equals the estimated line and the associated error at each value of Y. Notice that the three distributions are normally distributed around the point on the line, and further, the variation, variance, around the predicted value is constant indicating homoscedasticity from assumption 2. Figure 21 does not show all the assumptions of the regression model, but it helps visualize these important ones.

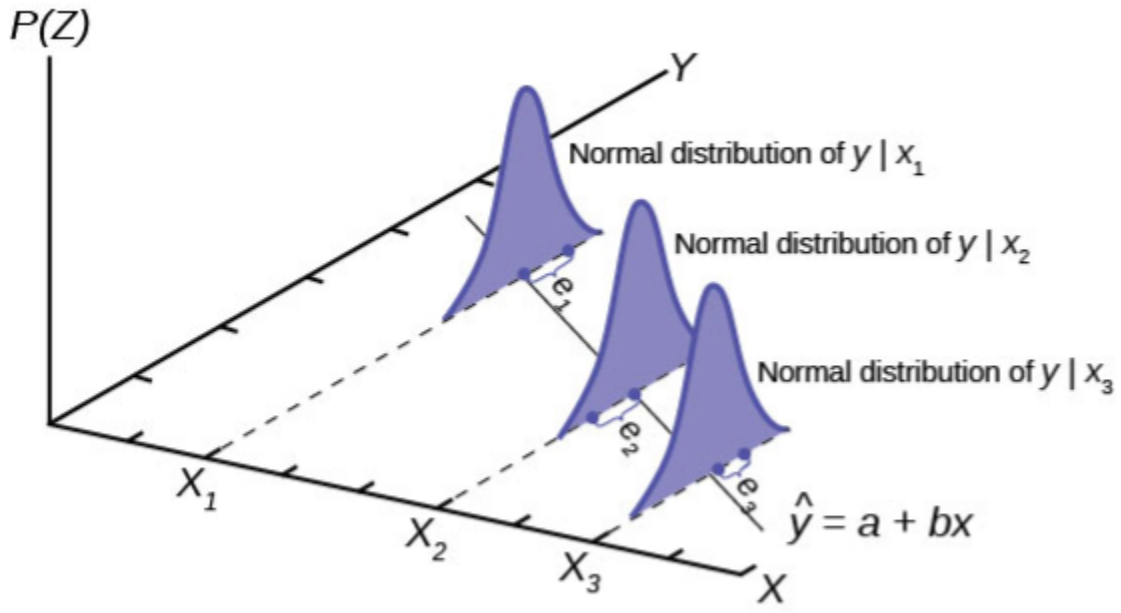
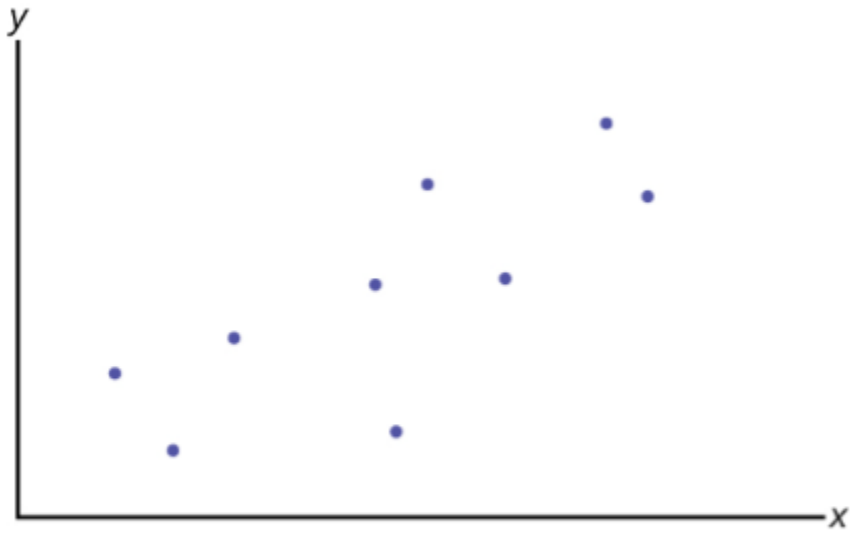


Figure 21



$$y = \beta_0 + \beta_1 X + \varepsilon$$

Figure 22

This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent (right-hand) variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. Figure 22 presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

The regression problem comes down to determining which straight line would best represent the data in Figure 23. Regression analysis is sometimes called "least squares" analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals of a line put through the data.

Figure 23: Population Equation: $C = \beta_0 + \beta_1 \text{ Income } + \varepsilon$
Estimated Equation: $C = b_0 + b_1 \text{ Income } + e$

Figure 23 shows the assumed relationship between consumption and income from macroeconomic theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e_1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person's savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph in Figure 24. The notation has returned to that for the more general model rather than the specific case of the Macroeconomic consumption function in our example.

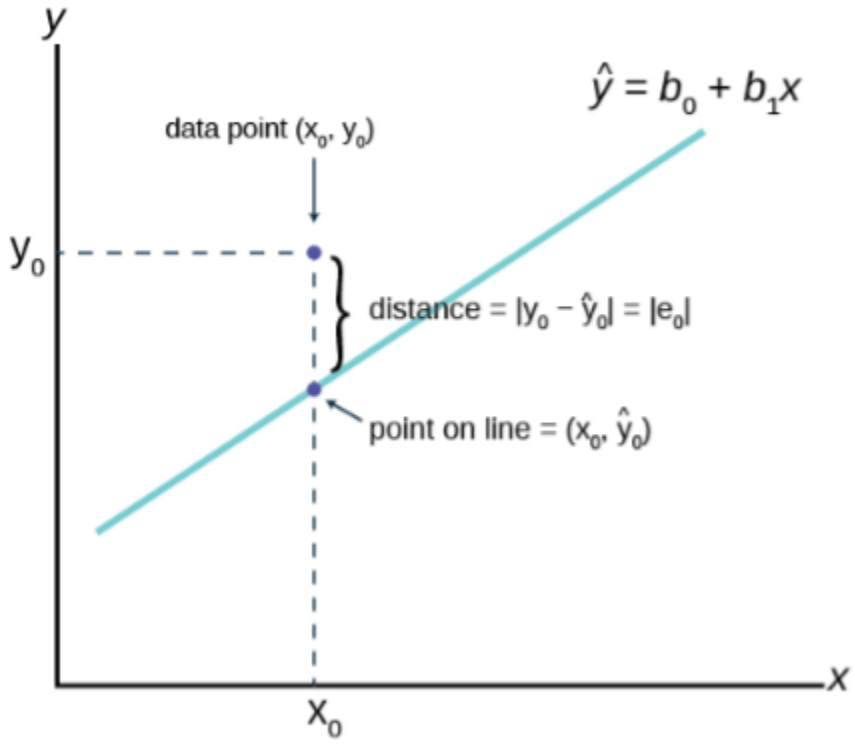


Figure 24

The $\hat{y}$ is read "**y hat**" and is the **estimated value of y.** (In Figure 23 $\widehat{C}$ represents the estimated value of consumption because it is on the estimated line.) It is the value of y obtained using the regression line. $\widehat{y}$ is not generally equal to y from the data.

The term $y_0 - \widehat{y}_0 = e_0$ is called the "**error" or residual.** It is not an error in the sense of a mistake. The error term was put into the estimating equation to capture missing variables and error in measurement that may have occurred in the dependent variables. The **absolute value of a residual** measure the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line as can be seen on the graph at point X_0.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y.

If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y.

In the graph, $y_0 - \hat{y}_0 = e_0$ is the residual for the point shown. Here the point lies above the line, and the residual is positive. For each data point the residuals, or errors, are calculated $y_i - \widehat{y}_i = e_i$ for $i = 1, 2, 3, \ldots, n$ where n is the sample size. Each $|e|$ is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SSE).**

Using calculus, you can determine the straight line that has the parameter values of b_0 and b_1

that minimizes the SSE. When you make the SSE a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = b_0 + b_1 x$$

$$\text{Where } b_0 = \bar{y} - b_1 \bar{x} \text{ and } b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2}$$

The sample means of the x values and the y values are $\bar{x}$ and $\bar{y}$, respectively. The best fit line always passes through the point $(\bar{x}, \bar{y})$ called the points of means.

The slope *b* can also be written as:

$$b_1 = r_{y,x}\left(\frac{s_y}{s_x}\right)$$

Where $s_y =$ the standard deviation of the *y* values and $s_x =$ the standard deviation of the *x* values and *r* is the correlation coefficient between *x* and *y*.

These equations are called the Normal Equations and come from another very important mathematical finding called the Gauss-Markov Theorem without which we could not do regression analysis. The Gauss-Markov Theorem tells us that the estimates we get from using the ordinary least squares (OLS) regression method will result in estimates that have some very important properties. In the Gauss-Markov Theorem it was proved that a least squares line is BLUE, which is, Best, Linear, Unbiased, Estimator. Best is the statistical property that an estimator is the one with the minimum variance. Linear refers to the property of the type of line being estimated. An unbiased estimator is one whose estimating function has an expected mean equal to the mean of the population. (You will remember that the expected value of $\frac{\mu_-}{x}$ was equal to the population mean μ in accordance with the Central Limit Theorem. This is exactly the same concept here).

Using the OLS method, we can now find the **estimate of the error variance,** which is the variance of the squared errors, e^2. This is sometimes called the **standard error of the estimate.** (Grammatically this is probably best said as the estimate of the **error's** variance). The formula for the estimate of the error variance is:

$$s_e^2 = \frac{\sum (y_i - \widehat{y}_i)^2}{n-k} = \frac{\sum e_i^2}{n-k}$$

Where $\hat{y}$ is the predicted value of y and y is the observed value, and thus the term $(y_i - \widehat{y}_i)^2$ is the squared errors that are to be minimized to find the estimates of the regression line parameters. This is really just the variance of the error terms and follows our regular variance formula. One important note is that here we are dividing by $(n - k)$, which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced b the number of estimated parameters, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how "tight" the dispersion is about the line. As we will see shortly, the greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the

hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistic and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of x and y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between x and y is affected by this difference between the estimate of the error variance.

Introduction to Residuals and Least-Squares Regression

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-47

Calculating Residual Example

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-48

Check Your Understanding: Linear Equations

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-59

Residual Plots

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-49

Check Your Understanding: Residual Plots

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-176

Calculating the Equation of a Regression Line

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-50

Check Your Understanding: Calculating the Equation of a Regression Line

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-61

Interpreting Slope of Regression Line

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-51

Interpreting y-intercept in Regression Model

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-52

Check Your Understanding: Interpreting Slope of Regression Line and Interpreting y-intercept in Regression Model

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-62

Using Least Squares Regression Output

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-53

Check Your Understanding: Using Least Squares Regression Output

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#h5p-63

How Good is the Equation?

The multiple correlation coefficient, also called the coefficient of multiple determination or the **coefficient of determination**, is given by the formula:

$$R^2 = \frac{SSR}{SST}$$

Where SSR is the regression sum of squares, the squared deviation of the predicted value of y from the mean value of y $(\hat{y} - \bar{y})$and SST is the total sum of squares which is the total squared deviation of the dependent variable, y, from its mean value, including the error term, SSE, the sum of squared errors. Figure 25 shows how the total deviation of the dependent variable, y, is partitioned into these two pieces.

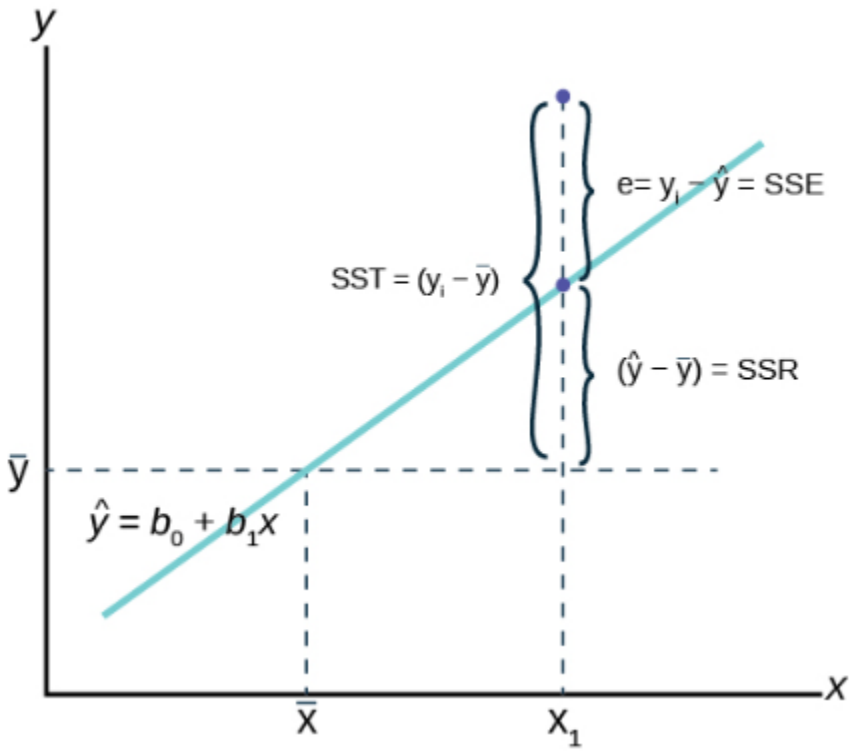


Figure 25

Figure 25 shows the estimated regression line and a single observation, x_1. Regression analysis tries to explain the variation of the data about the mean value of the dependent variable, y. The question is, why do the observations of y vary from the average level of y?

The value of y at observation x_1 varies from the mean of y by the difference $(y_i - \bar{y})$.) .). The sum of these differences squared is SST, the sum of squares total. The actual value of y at x_1 deviates from the estimated value, $\hat{y}$, by the difference between the estimate value and the actual value, $(y_i - \widehat{y})$. We recall that this is the error term, e, and the sum of these errors is SSE, sum of squared errors. The deviation of the predicted value of y, $\widehat{y}$, from the mean value of y is $(\hat{y} - \bar{y})$ and is the SSR, sum of squares regression. It is called "regression" because it is the deviation explained by the regression. (Sometimes the SSR is called SSM for sum of squares mean because it measures the deviation from the mean value of the dependent variable, y, as shown on the graph.).

Because the SST = SSR + SSE we see that the multiple correlation coefficient is the percent of the variance, or deviation in y from its mean value, that is explained by the equation when taken as a whole. R^2 will vary between zero and 1, with zero indicating that none of the variation in y was explained by the equation and a value of 1 indicating that 100% of the variation in y was explained by the equation. For time series studies expect a high R^2 and for cross-section data expect low R^2.

While a high R^2 is desirable, remember that it is the tests of the hypothesis concerning the existence of a relationship between a set of independent variables and a particular dependent variable that was the motivating factor in using the regression model. It is validating a cause-and-effect relationship developed by some theory that is the true reason that we chose the regression analysis. Increasing the number of independent variables will have the effect of increasing R^2. To account for this effect the proper measure of the coefficient of determination is $\bar{R}^2$, adjusted for degrees of freedom, to keep down mindless addition of independent variables.

There is no statistical test for the R^2 and thus little can be said about the model using R^2 with our characteristic confidence level. Two models that have the same size of SSE, that is sum of squared errors, may have very different R^2 if the competing models have different SST, total sum of squared deviations. The goodness of fit of the two models is the same; they both have the same sum of squares unexplained, errors squared, but because of the larger total sum of squares on one of the models the R^2 differs. Again, the real value of regression as a tool is to examine hypotheses developed from a model that predicts certain relationships among the variables. These are tests of hypotheses on the coefficients of the model and not a game of maximizing R^2.

R-Squared or Coefficient of Determination

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-54

DATA ANALYSIS TOOLS (SPREADSHEETS AND BASIC PROGRAMMING)

Descriptive Statistics

Using Microsoft Excel's "Descriptive Statistics" Tool

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-55

How to Run Descriptive Statistics in R

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-56

Descriptive Statistics in R Part 2

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-57

Regression Analysis

How to Use Microsoft Excel for Regression Analysis

Please read this text on how to use Microsoft Excel for regression analysis.

Simple Linear Regression in Excel

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-58

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://uta.pressbooks.pub/oert-mpsfundamentals/?p=355#oembed-59

RELEVANCE TO TRANSPORTATION ENGINEERING COURSEWORK

This section explains the relevance of the regression models for trip generation, mode choice, traffic flow-speed-density relationship, traffic safety, and appropriate sample size for spot speed study to transportation engineering coursework.

Regression Models for Trip Generation

The trip generation step is the first of the four-step process for estimating travel demand for infrastructure planning. It involves estimating the number of trips made to and from each traffic analysis zone (TAZ). Trip generation models are estimated based on land use and trip-making data. They use either linear regression or cross-tabulation of household characteristics. Simple linear regression is described in the section above titled "Compute and Interpret Simple Linear Regression Between Two Variables", and the tools to conduct the linear regression are discussed in "Data Analysis Tools (Spreadsheets and Basic Programming)".

Mode Choice

Estimation of Mode Choice is also part of the four-step process for estimating travel demand. It entails estimating the trip makers' transportation mode (drive alone, walk, take public transit, etc.) choices. The results of this step are the counts of trips categorized by mode. The most popular mode choice model is the discrete choice, multinomial logit model. Hypothesis tests are conducted for the estimated model parameters to assess whether they are "statistically significant." The section titled "Use Specific Significance Tests Including, Z-Test, T-Test (one and two samples), Chi-Squared Test" of this chapter provides extensive information on hypothesis testing.

Traffic Flow-Speed-Density Relationship

Greenshield's model is used to represent the traffic flow-speed-density relationship. Traffic speed and traffic density (number of vehicles per unit mile) are collected to estimate a linear regression model for speed as a function of density. "Compute and Interpret Simple Linear Regression Between Two Variables" above provides information on simple linear regression. "Data Analysis Tools (Spreadsheets and Basic Programming)" provides guidance for implementing the linear regression technique using tools available in Microsoft Excel and the programing language R.

Traffic Safety

Statistical analyses of traffic collisions are used to estimate traffic safety in roadway locations. A variety of regression models, most of which are beyond the scope of this book, are implemented to investigate the association between crashes and characteristics of the roadway locations. Hypothesis tests are conducted for the estimated model parameters to assess whether they are "statistically significant." "Find Confidence Intervals for Parameter Estimates" above describes the confidence intervals of the parameters, and "Use Specific Significance Tests Including, Z-Test, T-Test (one and two samples), Chi-Squared Test" discusses the different types of hypothesis tests. Programing language R includes statistical analysis toolkits or packages that may be used to estimate the regression models for crash data.

Appropriate Sample Size for Spot Speed Study

Spot speed studies during relatively congestion-free durations are conducted to estimate mean speed, modal speed, pace, standard deviation, and different percentile of speeds at a roadway location. An adequate number of data points (i.e., the number of vehicle speeds recorded) are required to ensure reliable results within an acceptable margin of error. "Estimate the Required Sample Size for Testing" in this chapter discusses the approach to assessing the adequacy of sample size.

Key Takeaways

- Estimating the total number of trips made to and from each traffic analysis zone (TAZ)is typically the first step in the four-step travel demand modeling process. Simple linear regression models and the tools (such as MS Excel and R) to estimate these models are used this step of the travel demand modeling process. The independent variables for these models are the total number of trips with demographic and employment data for the zones as independent variables.
- Hypothesis testing is used to determine the statistical significance of coefficients estimated for the discrete choice multinomial models. Such models are used for the mode choice step of the travel demand modeling process.
- Linear regression models are also used to estimate the relationship between speed and density on uninterrupted flow facilities such as freeway segments.
- Regression models are also implemented to investigate the association between crashes and characteristics of the roadway locations. Tools described in the chapter for linear regression may also be used for estimating these models, including negative binomial regression models.
- Statistical analysis is also used to ensure that an adequate number of data points (i.e., the number of vehicle speeds recorded) are used to obtain estimates of the mean, median, and 85th percentile of speed within an acceptable margin of error in a spot speed study.

GLOSSARY – KEY TERMS

– Coefficient of Determination[1] – this is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the

independent variable. Sometimes calculated by the equation $R^2 = \frac{SSR}{SST}$ where SSR is the "sum of squares regressions" and SST is the "sum of squares total." The appropriate coefficient of determination to be reported should always be adjusted for degrees of freedom first

A is the Symbol for the y-Intercept[1] – sometimes written as $\diamondsuit_0$, because when writing the theoretical linear model $\diamondsuit_0$ is used to represent a coefficient for a population

Analysis of Variance[1] – also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if: (1) all populations of interest are normally distributed (2) the populations have equal standard deviations (3) samples (not necessarily of the same size) are randomly and independently selected from each population (4) there is one independent variable and one dependent variable. The test statistic for analysis of variance is the F-ratio

Average[1] – a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

B is the Symbol for Slope[1] – the word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter "x." It will be written a $\diamondsuit_1$ when a sample is used, and $\diamondsuit_1$ will be used with a population or when writing the theoretical linear model.

Binomial Distribution[1] – a discrete random variable which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial random variable X is defined as the number of successes in n trials. The notation is: $\diamondsuit \sim \diamondsuit(\diamondsuit,\diamondsuit)$. The mean is $\diamondsuit = \diamondsuit\diamondsuit$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$

Bivariate[1] – two variables are present in the model where one is the "cause" or independent variable and the other is the "effect" of dependent variable.

Central Limit Theorem[1] – given a random variable with known mean $\diamondsuit$ and known standard deviation, $\diamondsuit$, we are sampling with size n, and we are interested in two new random variables: the sample mean, $\bar{X}$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size (n) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the **standard error of the mean**.

Cohen's d [1] – a measure of effect size based on the differences between two means. If d is between 0 and 0.2 then the effect is small. If d approaches is 0.5, then the effect is medium, and if d approaches 0.8, then it is a large effect.

Confidence Interval (CI)[1] – an interval estimate for an unknown population parameter. This depends on: (1) the desired confidence level (2) information that is known about the distribution (for example, known standard deviation) (3) the sample and its size

Confidence Level (CL)[1] – the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Contingency Table[1] – a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

Critical Value[1] – The t or Z value set by the researcher that measures the probability of a Type I error, ◈

Degrees of Freedom (df)[1] – the number of objects in a sample that are free to vary

Error Bound for a Population Mean (EBM)[1] – the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion (EBP)[1] – the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Goodness-of-Fit[1] – a hypothesis test that compares expected and observed values in order to look for significant differences within one non-parametric variable. The degrees of freedom used equals the (number of categories – 1).

Hypothesis[1] – a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation ◈$_0$) and the contradictory statement is called the alternative hypothesis (notation ◈$_◈$)

Hypothesis Testing[1] – Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Independent Groups[1] – two samples that are selected from two populations, and the values from one population are not related in any way to the values from the other population.

Inferential Statistics[1] – also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Linear[1] – a model that takes data and regresses it into a straight line equation.

Matched Pairs[1] – two samples that are dependent. Differences between a before and after scenario are tested by testing one population mean of differences.

Mean[1] – a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{sum of all values in the sample}}{\text{number of values in the sample}}$, and the mean for a population (denoted by ◈) is $\frac{\text{sum of all values in the population}}{\text{number of values in the population}}$

Multivariate[1] – a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.

Normal Distribution[1] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where ◈ is the mean of the distribution and ◈ is the standard deviation; notation: ◈~◈(◈,◈). If ◈=0 and ◈=1, the random variable, Z, is called the standard normal distribution.

Normal Distribution[1] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where ◈ is the mean of the distribution and ◈ is the standard deviation, notation: ◈~◈(◈,◈) and ◈=1, the random variable is called the standard normal distribution.

One-Way ANOVA[1] – a method of testing whether or not the means of three or more populations are equal; the method is applicable if: (1) all populations of interest are normally distributed (2) the populations have equal standard deviations (3) samples (not necessarily of the same size) are randomly and independently selected from each population. The test statistic for analysis of variance is the F-ratio

Parameter[1] – a numerical characteristic of a population
Point Estimate[1] – a single number computed from a sample and used to estimate a population parameter
Pooled Variance[1] – a weighted average of two variances that can then be used when calculating standard error.
R – Correlation Coefficient[1] – A number between –1 and 1 that represents the strength and direction of the relationship between "X" and "Y." The value for "r" will equal 1 or –1 only if all the plotted points form a perfectly straight line.
Residual or "Error"[1] – the value calculated from subtracting $y_0 - \hat{y}_0 = e_0$. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y that appears on the best-fit line.
Sampling Distribution[1] – Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.
Standard Deviation[1] – a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and ◈ for population standard deviation
Standard Error of the Mean[1] – the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$
Standard Error of the Proportion[1] – the standard deviation of the sampling distribution of proportions
Student's t-Distribution[1] – investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of this random variable are: (1) it is continuous and assumes any real values (2) the pdf is symmetrical about its mean of zero (3) it approaches the standard normal distribution as n gets larger (4) there is a "family" of t-distributions: each representative of the family is completely defined by the number of degrees of freedom, which depends upon the application for which the t is being used.
Sum of Squared Errors (SSE)[1] – the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.
Test for Homogeneity[1] – a test used to draw a conclusion about whether two populations have the same distribution. The degrees of freedom used equals the (number of columns – 1).
Test of Independence[1] – a hypothesis test that compares expected and observed values for contingency tables in order to test for independence between two variables. The degrees of freedom used equals the (number of columns – 1) multiplied by the (number of rows – 1).
Test Statistic[1] – the formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.
Type I Error[1] – the decision is to reject the null hypothesis when, in fact, the null hypothesis is true.
Type II Error[1] – the decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.
Variance[1] – mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.
X – the Independent Variable[1] – This will sometimes be referred to as the "predictor" variable,

because these values were measured in order to determine what possible outcomes could be predicted.

Y – the Dependent Variable[1] – also, using the letter "y" represents actual values while $\hat{y}$ represents predicted or estimated values. Predicted values will come from plugging in observed "x" values into a linear model

[1] "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction

MEDIA ATTRIBUTIONS

Note: All Khan Academy content is available for free at (www.khanacademy.org).

Videos

- Video 1: Central Limit Theorem by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 2: Sampling Distribution of the Sample Mean by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 3: Sampling Distribution of the Sample Mean (Part 2) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 4: Sampling Distributions: Sampling Distribution of the Mean by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 5: Confidence Intervals & Estimation: Point Estimates Explained by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 6: Confidence Interval for Mean: 1 Sample Z Test (Using Formula) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 7: Confidence Intervals: Using the t Distribution by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 8: How to Construct a Confidence Interval for Population Proportion by Simple Science and Maths is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 9: Estimation and Confidence Intervals: Calculate Sample Size by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 10: Calculating Sample size to Predict a Population Proportion by Matthew Simmons is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 11: Hypothesis Testing: The Fundamentals by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

- Video 12: Hypothesis Testing: Setting up the Null and Alternative Hypothesis Statements by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 13: Hypothesis Testing: Type I and Type II Errors by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 14: Hypothesis testing: Finding Critical Values by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 15: Normal Distribution: Finding Critical Values of Z by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 16: What is a "P-Value?" by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 17: Hypothesis Testing: One Sample Z Test of the Mean (Critical Value Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 18: Hypothesis Testing: t Test for the Mean (Critical Value Approach)by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 19: Hypothesis Testing: 1 Sample Z Test of the Mean (Confidence Interval Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 20: Hypothesis Testing: 1 Sample Z Test for Mean (P-Value Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 21: Hypothesis Testing: 1 Proportion using the Critical Value Approach by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 22: Hypothesis Testing – Two Population Means by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 23: Two Population Means, One Tail Test, unknown by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 24: Two Population Means, Two Tail Test, unknown by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 25: Effect Size for a Significant Difference of Two Sample Means by Dana Lee Ling is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 26: Two Population Means, One Tail Test, Known by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 27: Two Population Means, Two Tail Test, Known by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 28: Two Population Means, One Tail Test, Matched Sample by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 29: Two Population Means, One Tail test, Matched Sample by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 30: Matched Sample, Two Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

- Video 31: Two Population Proportions, One Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 32: Two Population Proportion, Two Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 33: Single Population Variances, One-Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 34: Chi-Square Statistic for Hypothesis Testing by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 35: Chi-Square Goodness-of-Fit Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 36: Simple Explanation of Chi-Squared by J David Eisenberg is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 37: Chi-Square Tet for Association (independence)by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 38: Introduction to the Chi-Square Test for Homogeneity by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 39: Hypothesis Test Two Population Variances by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 40: ANOVA by J David Eisenberg is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 41: Calculating SST (total sum of squares) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 42: Calculating by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 43: Hypothesis Test with F-Statistic by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 44: Bivariate Relationship Linearity, Strength, and Direction by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 45: Calculating Correlation Coefficient r by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 46: Example: Correlation Coefficient Intuition by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 47: Introduction to Residuals and Least-Squares Regression by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 48: Calculating Residual Example by Khan Academy is licensed by Creative

Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- Video 49: Residual Plots by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 50: Calculating the Equation of a Regression Line by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 51: Interpreting Slope of Regression Line by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 52: Interpreting y-intercept in Regression Model by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 53: Using Least Squares Regression Output by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 54: R-Squared or Coefficient of Determination by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Video 55: Using Microsoft Excel's "Descriptive Statistics" Tool by Linda Weiser Friedman is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 56: How to Run Descriptive Statistics in R by Learning Puree is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Video 57: Descriptive Statistics in R Part 2 by Learning Puree is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Video 58: Simple Linear Regression in Excel by Ramya Rachel is licensed by Creative Commons 3.0 Unported (CC BY 3.0)
- Video 59: Simple Linear Regression, fit and interpretations in Ris licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Figures

- Figure 1: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 2: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 3: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 4: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 5: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 6: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 7: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and

Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

- Figure 8: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 9: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 10: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 11: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 12: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 13: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 14: by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 15 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 16 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 17 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 18 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 19 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 20 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 21 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 22 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 23 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 24 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 25 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

REFERENCES

Note: All Khan Academy content is available for free at (www.khanacademy.org).

- "Unit: Sampling Distributions" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Inference for Categorical Data: Chi-Square" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Analysis of Variance (ANOVA)" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Exploring Two-Variable Quantitative Data" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Statistics" by Barbara Illosky and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- "Introduction to Probability and Statistics" by Jeremy Orloff and Jonathan Bloom is licensed by Creative Commons NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Farid, A. (2022). *Transportation Planning 1*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). *Transportation Planning 2*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). *Fundamentals of Traffic Flow Theory*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Washington, S., Karlaftis, M., Mannering, F., and Anastasopoulos, P. (2020). *Statistical and Econometric Methods for Transportation Data Analysis 3rd Edition*. Chemical Rubber Company (CRC) Press, Taylor and Francis Group, Boca Raton, FL.
- Farid, A. (2022). *Spot Speed Study.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

GLOSSARY

CHAPTER 1

Degree[1] – a unit of measure for angles equal to an angle with its vertex at the center of a circle and its sides cutting off 1/360 of the circumference
Law of Cosines[1] – a law in trigonometry: the square of a side of a plane triangle equals the sum of the squares of the remaining sides minus twice the product of those sides and the cosine of the angle between them
Law of Sines[1] – a law in trigonometry: the ratio of each side of a plane triangle to the sine of the opposite angle is the same for all three sides and angles
Radian[1] – a unit of plane angular measurement that is equal to the angle at the center of a circle subtended by an arc whose length equals the radius or approximately 57.3 degrees
Trigonometric Function[1] – a function (such as the sine, cosine, tangent, cotangent, secant, or cosecant) of an arc or angle most simply expressed in terms of the ratios of pairs of sides of a right-angled triangle.

CHAPTER 2

Binomial[1] – a mathematical expression consisting of two terms connected by a plus sign or minus sign
Common Factor[1] – (also called common divisor) a number or expression that divides two or more numbers or expressions without remainder
Exponent[1] – a symbol written above and to the right of a mathematical expression to indicate the operation of raising to a power
Exponential Function[1] – a mathematical function in which an independent variable appears in one of the exponents
Exponential Growth[2] – a process that increases quantity over time at an ever-increasing rate. It occurs when the instantaneous rate of change (that is, the derivative) of a quantity with respect to time is proportional to the quantity itself.
Logarithm[1] – the exponent that indicates the power to which a base number is raised to produce a given number
Logarithmic Function[1] – a function (such as $y = \log_a x$ or $y = \ln x$) that is the inverse of an exponential function (such as $y = a^x$ or $y = e^x$) so that the independent variable appears in a logarithm
Polynomial[1] – a mathematical expression of one or more algebraic terms each of which consists of a constant multiplied by one or more variables raised to a nonnegative integral power (such as $a + bx + cx2$)

CHAPTER 3

Augmented Matrix[1] – a matrix whose elements are the coefficients of a set of simultaneous linear equations with the constant terms of the equations entered in an added column

Matrix[1] – a rectangular array of mathematical elements of simultaneous linear equations that can be combined to form sums and products with similar arrays having an appropriate number of rows and columns

CHAPTER 4

Derivative[1] – the limit of the ratio of the change in a function to the corresponding change in its independent variable as the latter change approaches zero
Rate of Change[1] – a value that results from dividing the change in a function of a variable by the change in the variable
Slope[1] – the slope of the line tangent to a plane curve at a point
Tangent[1] – meeting a curve or surface in a single point if a sufficiently small interval is considered
Rational Function[1] – a function that is the quotient of two polynomials
Integration[1] – the operation of finding whose differential is known\
Indefinite[1] – having no exact limits
Definite[1] – having distinct or certain limits
Natural Logarithm[1] – a logarithm with e as a base
Absolute Value[1] – a nonnegative number equal in numerical value to a given real number
Piecewise[1] – with respect to a number of discrete intervals, sets, or pieces
Riemann Integral[1] – a definite integral defined as the limit of sums found by partitioning the interval comprising the domain of definition into subintervals, by finding the sum of products each of which consists of the width of a subinterval multiplied by the value of the function at some point in it, and by letting the maximum width of the subintervals approach zero
Fundamental Theorem of Calculus[3] – the theorem, central to the entire development of calculus, that establishes the relationship between differentiation and integration
Fundamental Theorem of Calculus, Part 1[3] – uses a definite integral to define an antiderivative of a function
Fundamental Theorem of Calculus, Part 2[3] – (also, evaluation theorem) we can evaluate a definite integral by evaluating the antiderivative of the integrand at the endpoints of the interval and subtracting

CHAPTER 5

Acceleration[4] – the rate at which an object's velocity changes over a period of time
Acceleration due to Gravity[4] – acceleration of an object as result of gravity
Accuracy[4] – the degree to which a measured value agrees with correct value for that measurement
Angular Velocity[4] – ⍰, the rate of change of the angle with which an object moves on a circular path
Arc Length[4] – Δ⍰, the distance traveled by an object along a circular path
Average Acceleration[4] – the change in velocity divided by the time over which it changes
Average Speed[4] – distance traveled divided by time during which motion occurs
Average Velocity[4] – displacement divided by the time over which displacement occurs
Banked Curve[4] – the curve in a road that is sloping in a manner that helps a vehicle negotiate the curve
Center of Mass[4] – the point where the entire mass of an object can be thought to be concentrated

Centrifugal Force[4] – a fictitious force that tends to throw an object off when the object is rotating in a non-inertial frame of reference
Centripetal Acceleration[4] – the acceleration of an object moving in a circle, directed toward the center
Centripetal Force[4] – any net force causing uniform circular motion
Classical Physics[4] – physics that was developed from the Renaissance to the end of the 19th century
Conversion Factor[4] – a ratio expressing how many of one unit are equal to another unit
Coriolis Force[4] – the fictitious force causing the apparent deflection of moving objects when viewed in a rotating frame of reference
Deceleration[4] – acceleration in the direction opposite to velocity; acceleration that results in a decrease in velocity
Displacement[4] – the change in position of an object
Distance[4] – the magnitude of displacement between two positions
Distance Traveled[4] – the total length of the path traveled between two positions
Elapsed Time[4] – the difference between the ending time and beginning time
English Units[4] – system of measurement used in the United States; includes units of measurement such as feet, gallons, and pounds
External Force[4] – a force acting on an object or system that originates outside of the object or system
Fictious Force[4] – a force having no physical origin
Force[4] – a push or pull on an object with a specific magnitude and direction; can be represented by vectors; can be expressed as a multiple of a standard force
Free-body Diagram[4] – a sketch showing all of the external forces acting on an object or system; the system is represented by a dot, and the forces are represented by vectors extending outward from the dot
Free-Fall[4] – a situation in which the only force acting on an object is the force due to gravity
Friction[4] – a force past each other of objects that are touching; examples include rough surfaces and air resistance
Gravitational Constant, G[4] – a proportionality factor used in the equation for Newton's universal law of gravitation; it is a universal constant—that is, it is thought to be the same everywhere in the universe
Ideal Banking[4] – the sloping of a curve in a road, where the angle of the slope allows the vehicle to negotiate the curve at a certain speed without the aid of friction between the tires and the road; the net external force on the vehicle equals the horizontal centripetal force in the absence of friction
Ideal Speed[4] – the maximum safe speed at which a vehicle can turn on a curve without the aid of friction between the tire and the road
Inertia[4] – the tendency of an object to remain at rest or remain in motion
Inertial Frame of Reference[4] – a coordinate system that is not accelerating; all forces acting in an inertial frame of reference are real forces, as opposed to fictitious forces that are observed due to an accelerating frame of reference
Instantaneous Acceleration[4] – acceleration at a specific point in time
Instantaneous Speed[4] – magnitude of the instantaneous velocity
Instantaneous Velocity[4] – velocity at a specific instant, or the average velocity over an infinitesimal time interval

Kilogram[4] – the SI unit for mass, abbreviated (kg)
Kinematics[4] – the study of motion without considering its causes
Kinetic Friction[4] – a force that opposes the motion of two systems that are in contact and moving relative to one another
Law[4] – a description, using concise language or a mathematical formula, a generalized pattern in nature that is supported by scientific evidence and repeated experiments
Law of Inertia[4] – see Newton's first law of motion
Magnitude of Kinetic Friction[4] – ◇◇=◇◇◇, where ◇◇ is the coefficient of kinetic friction
Magnitude of Static Friction[4] – ◇◇≤◇◇◇, where ◇◇ is the coefficient of static friction and ◇ is the magnitude of the normal force
Mass[4] – the quantity of matter in a substance; measured in kilograms
Meter[4] – the SI unit for length, abbreviated (m)
Metric System[4] – a system in which values can be calculated in factors of 10
Model[4] – simplified description that contains only those elements necessary to describe the physics of a physical situation
Net External Force[4] – the vector sum of all external forces acting on an object or system; causes a mass to accelerate
Newton's First Law of Motion[4] – in an inertial frame of reference, a body at rest remains at rest, or, if in motion, remains in motion at a constant velocity unless acted on by a net external force; also known as the law of inertia
Newton's Second Law of Motion[4] – the net external force ◇◇◇◇ on an object with mass ◇ is proportional to and in the same direction as the acceleration of the object, ◇, and inversely proportional to the mass; defined mathematically as $a = \frac{F_{net}}{m}$
Newton's Third Law of Motion[4] – whenever one body exerts a force on a second body, the first body experiences a force that is equal in magnitude and opposite in direction to the force that the first body exerts
Newton's Universal Law of Gravitation[4] – every particle in the universe attracts every other particle with a force along a line joining them; the force is directly proportional to the product of their masses and inversely proportional to the square of the distance between them
Non-Inertial Frame of Reference[4] – an accelerated frame of reference
Normal Force[4] – the force that a surface applies to an object to support the weight of the object; acts perpendicular to the surface on which the object rests
Physical Quantity[4] – a characteristic or property of an object that can be measured or calculated from other measurements
Physics[4] – the science concerned with describing the interactions of energy, matter, space, and time; it is especially interested in what fundamental mechanisms underlie every phenomenon
Pit[4] – a tiny indentation on the spiral track moulded into the top of the polycarbonate layer of CD
Position[4] – the location of an object at a particular time
Radians[4] – a unit of angle measurement
Radius of Curvature[4] – radius of a circular path
Rotation Angle[4] – the ratio of arc length to the radius of curvature on a circular path: $\Delta\theta = \frac{\Delta s}{r}$
Scalar[4] – a quantity that is described by magnitude, but not direction
Second[4] – the SI unit for time, abbreviated (s)

SI Units[4] – the international system of units that scientists in most countries have agreed to use; includes units such as meters, liters, and grams
Significant Figures[4] – express the precision of a measuring tool used to measure a value
Slope[4] – the difference in y-value (the rise) divided by the difference in x-value (the run) of two points on a straight line
Static Friction[4] – a force that opposes the motion of two systems that are in contact and are not moving relative to one another
System[4] – defined by the boundaries of an object or collection of objects being observed; all forces originating from outside of the system are considered external forces
Tension[4] – the pulling force that acts along a medium, especially a stretched flexible connector, such as a rope or cable; when a rope supports the weight of an object, the force on the object due to the rope is called a tension force
Thrust[4] – a reaction force that pushes a body forward in response to a backward force; rockets, airplanes, and cars are pushed forward by a thrust reaction force
Time[4] – change, or the interval over which change occurs
Ultracentrifuge[4] – a centrifuge optimized for spinning a rotor at very high speeds
Uniform Circular Motion[4] – the motion of an object in a circular path at constant speed
Units[4] – a standard used for expressing and comparing measurements
Vector[4] – a quantity that is described by both magnitude and direction
Weight[4] – the force ⃗ due to gravity acting on an object of mass ; defined mathematically as: ⃗=⃗, where ⃗ is the magnitude and direction of the acceleration due to gravity
Y-Intercept[4] – the y-value when =0, or when the graph crosses the y-axis

CHAPTER 6

Center of Gravity[4] – the point where the total weight of the body is assumed to be concentrated
Change in Momentum[4] – the difference between the final and initial momentum; the mass times the change in velocity
Conservation of Momentum Principle[4] – when the net external force is zero, the total momentum of the system is conserved or constant
Dynamic Equilibrium[4] – a state of equilibrium in which the net external force and torque on a system moving with constant velocity are zero
Elastic Collision[4] – a collision that also conserves internal kinetic energy
Impulse[4] – the average net external force times the time it acts; equal to the change in momentum
Inelastic Collision[4] – a collision in which internal kinetic energy is not conserved
Internal Kinetic Energy[4] – the sum of the kinetic energies of the objects in a system
Isolated System[4] – a system in which the net external force is zero
Linear Momentum[4] – the product of mass and velocity
Neutral Equilibrium[4] – a state of equilibrium that is independent of a system's displacements from its original position
Perfectly Inelastic Collision[4] – a collision in which the colliding objects stick together
Perpendicular Lever Arm[4] – the shortest distance from the pivot point to the line along which ⃗ lies
Point Masses[4] – structureless particles with no rotation or spin

Second Law of Motion[4] – physical law that states that the net external force equals the change in momentum of a system divided by the time over which it changes
SI Units of Torque[4] – newton times meters, usually written as ⍰·⍰
Stable Equilibrium[4] – a system, when displaced, experiences a net force or torque in a direction opposite to the direction of the displacement
Static Equilibrium[4] – a state of equilibrium in which the net external force and torque acting on a system is zero; equilibrium in which the acceleration of the system is zero and accelerated rotation does not occur
Torque[4] – turning or twisting effectiveness of a force
Unstable Equilibrium[4] – a system, when displaced, experiences a net force or torque in the same direction as the displacement from equilibrium

CHAPTER 7

Amplitude[4] – the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position
Doppler effect[4] – an alteration in the observed frequency of a sound due to motion of either the source or the observer
Doppler shift[4] – the actual change in frequency due to relative motion of source and observer
Frequency[4] – number of events per unit of time
Hearing[4] – the perception of sound
Intensity[4] – power per unit area; sounds below 20 Hz
Longitudinal Wave[4] – a wave in which the disturbance is parallel to the direction of propagation
Loudness[4] – the perception of sound intensity
Natural Frequency[4] – the frequency at which a system would oscillate if there were no driving and no damping forces
Period[4] – time it takes to complete one oscillation
Pitch[4] – the perception of the frequency of a sound
Sound[4] – a disturbance of matter that is transmitted from its source outward
Sound Intensity Level[4] – a unitless quantity telling you the level of the sound relative to a fixed standard
Sound Pressure Level[4] – the ratio of the pressure amplitude to a reference pressure
Tone[4] – number and relative intensity of multiple sound frequencies
Traverse Wave[4] – a wave in which the disturbance is perpendicular to the direction of propagation
Wave[4] – a disturbance that moves from its source and carries energy
Wave Velocity[4] – the speed at which the disturbance moves. Also called the propagation velocity or propagation speed
Wavelength[4] – the distance between adjacent identical parts of a wave

CHAPTER 8

Bayes Theorem[6] – a useful tool for calculating conditional probability. Bayes theorem can be expressed as: ⍰$_1$, ⍰$_2$,…, ⍰$_⍰$ be a set of mutually exclusive events that together form the sample space S. Let B be any event from the same sample space, such that ⍰(⍰)>0. Then, $P(A \mid B) = \frac{P(A)P(B|A)}{P(B)}$
Bernoulli Trials[5] – an experiment with the following characteristics: (1) There are only two

possible outcomes called "success" and "failure" for each trial. (2) The probability p of a success is the same for any trial (so the probability ◈=1–◈ of a failure is the same for any trial).

Binomial Experiment[5] – a statistical experiment that satisfies the following three conditions: (1) There are a fixed number of trials, n. (2) There are only two possible outcomes, called "success" and, "failure," for each trial. The letter p denotes the probability of a success on one trial, and q denotes, the probability of a failure on one trial. (3) The n trials are independent and are repeated using identical conditions.

Binomial Probability Distribution[5] – a discrete random variable that arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial one) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial random variable X is defined as the number of successes in n trials. The mean is ◈=◈◈ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$

Combination[6] – a combination is a selection of all or part of a set of objects, without regard to the order in which objects are selected

Conditional Probability[5] – the likelihood that an event will occur given that another event has already occurred.

Discrete Variable / continuous variable[6] – if a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise it is called a discrete variable

Expected Value[6] – the mean of the discrete random variable X is also called the expected value of X. Notationally, the expected value of X is denoted by ◈(◈).

Exponential Distribution[5] – a continuous random variable that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = \frac{1}{\mu} e^{\left(-\frac{1}{\mu}\right)x}, x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{\left(-\frac{1}{\mu}\right)x}$

Geometric Distribution[5] – a discrete random variable that arises from the Bernoulli trials; the trials are repeated until the first success. The geometric variable X is defined as the number of trials until the first success. The mean is $\mu = \frac{1}{p}$ and the standard deviation is $\sigma = \sqrt{\frac{1}{p}\left(\frac{1}{p} - 1\right)}$. The probability of exactly x failures before the first success is given by the formula: ◈(◈=◈)=◈(1–◈)$^{◈-1}$ where one wants to know probability for the number of trials until the first success: the xth trial is the first success. An alternative formulation of the geometric distribution asks the question: what is the probability of x failures until the first success? In this formulation the trial that resulted in the first success is not counted. The formula for this presentation of the geometric is: ◈(◈=◈)=◈(1–◈)$^{◈}$. The expected value in this form of the geometric distribution is $\mu = \frac{1-p}{p}$. The easiest way to keep these two forms of the geometric distribution straight is to remember that p is the probability of success and (1–◈) is the probability of failure. In the formula the exponents simply count the number of successes and number of failures of the desired outcome of the experiment. Of course the sum of these two numbers must add to the number of trials in the experiment.

Geometric Experiment[5] – a statistical experiment with the following properties: (1) There are one or more Bernoulli trials with all failures except the last one, which is a success. (2) In theory, the number of trials could go on forever. There must be at least one trial. (3) The probability, p, of a success and the probability, q, of a failure do not change from trial to trial.

Independent[6] – two events are independent when the occurrence of one does not affect the probability of the occurrence of the other

Mean[6] – a mean score is an average score, often denoted by $\bar{X}$. It is the sum of individual scores divided by the number of individuals.

Memoryless Property[5] – For an exponential random variable X, the memoryless property is the statement that knowledge of what has occurred in the past has no effect on future probabilities. This means that the probability that X exceeds ◈+◈, given that it has exceeded x, is the same as the probability that X would exceed t if we had no knowledge about it. In symbols we say that $P(X > x + t \mid X > x) = P(X > t)$.

Multiplication Rule[6] – If events A and B come from the same sample space, the probability that both A and B occur is equal to the probability the event A occurs time the probability that B occurs, given that A has occurred. $P(A \cap B) = P(A)P(B \mid A)$.

Normal Distribution[5] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where ◈ is the mean of the distribution and ◈ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$,, the random variable, Z, is called the standard normal distribution

Permutation[6] – an arrangement of all or part of a set of objects, with regard to the order of the arrangement.

Poisson Distribution[5] – If there is a known average of ◈ events occurring per unit time, and these events are independent of each other, then the number of events X occurring in one unit of time has the Poisson distribution. The probability of x events occurring in one unit time is equal to $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$

Poisson Probability Distribution[5] – a discrete random variable that counts the number of times a certain event will occur in a specific interval; characteristics of the variable: (1) the probability that the event occurs in a given interval is the same for all intervals. (2) the events occur with a known mean and independently of the time since the last event. The distribution is defined by the mean ◈ of the event in the interval. The mean is ◈=◈◈. The standard deviation is $\sigma = \sqrt{\mu}$. The probability of having exactly x successes in r trials is $P(x) = \frac{\mu^x e^{-\mu}}{x!}$. The Poisson distribution is often used to approximate the binomial distribution, when n is "large" and p is "small" (a general rule is that np should be greater than or equal to 25 and p should be less than or equal to 0.01).

Probability Distribution Function (PDF)[5] – a mathematical description of a discrete random variable, given either in the form of an equation (formula) or in the form of a table listing all the possible outcomes of an experiment and the probability associated with each outcome.

Quartile[6] – Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by
◈$_1$, ◈$_2$, and ◈$_3$, respectively.

Random Variable[5] – a characteristic of interest in a population being studied; common notation for variables are upper case Latin letters X, Y, Z..; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, and z. For example, if X is the

number of children in a family, then x represents a specific integer 0, 1, 2, 3,… Variables in statistics differ from variables in intermediate algebra in the two following ways: (1) The domain of the random variable is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then domain is {black, blonde, gray, green, orange}. (2) We can tell what specific value x the random variable X takes only after performing the experiment.

Standard Deviation[6] – The standard deviation is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the standard deviation is big; and vice versa. It is important to distinguish between the standard deviation of a population and the standard deviation of a sample. They have different notation, and they are computed differently. The standard deviation of a population is denoted by ◈ and the standard deviation of a sample, by ◈

Standard Normal Distribution[5] – a continuous random variable ◈~◈(0,1); when X follows the standard normal distribution, it is often noted as ◈~◈(0,1).

Uniform Distribution[5] – a continuous random variable that has equally likely outcomes over the domain, ◈<◈<◈; it is often referred as the rectangular distribution because the graph of the pdf has the form of a rectangle. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x.The cumulative distribution is [latex] P(X \leq x) = \frac{x-a}{b-a}$

Variance[6] – the variance is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the variance is big; and vice versa. It is important to distinguish between the variance of a population and the variance of a sample. They have different notation, and they are computed differently. The variance of a population is denoted by ◈^{2}; and the variance of a sample, by ◈^{2}

Z-Score[5] – the linear transformation of the form $z = \frac{x-\mu}{\sigma}$ or written as $z = \frac{|x-\mu|}{\sigma}$; if this transformation is applied to any normal distribution ◈~◈(◈,◈) the result is the standard normal distribution ◈~◈(0,1). If this transformation is applied to any specific value x of the random variable with mean ◈ and standard deviation ◈, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

CHAPTER 9

◈$^◈$ **– Coefficient of Determination[5] –** this is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the independent variable. Sometimes calculated by the equation $R^2 = \frac{SSR}{SST}$ where SSR is the "sum of squares regressions" and SST is the "sum of squares total." The appropriate coefficient of determination to be reported should always be adjusted for degrees of freedom first

A is the Symbol for the y-Intercept[5] – sometimes written as ◈$_0$, because when writing the theoretical linear model ◈$_0$ is used to represent a coefficient for a population

Analysis of Variance[5] – also referred to as ANOVA, is a method of testing whether or not the means of three or more populations are equal. The method is applicable if: (1) all populations of interest are normally distributed (2) the populations have equal standard deviations (3) samples (not necessarily of the same size) are randomly and independently selected from each population (4) there is one

independent variable and one dependent variable. The test statistic for analysis of variance is the F-ratio

Average[5] – a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

B is the Symbol for Slope[5] – the word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter "x." It will be written a $\diamondsuit_1$ when a sample is used, and $\diamondsuit_1$ will be used with a population or when writing the theoretical linear model.

Binomial Distribution[5] – a discrete random variable which arises from Bernoulli trials; there are a fixed number, n, of independent trials. "Independent" means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial random variable X is defined as the number of successes in n trials. The notation is: ◇~◇(◇,◇). The mean is ◇=◇◇ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$

Bivariate[5] – two variables are present in the model where one is the "cause" or independent variable and the other is the "effect" of dependent variable.

Central Limit Theorem[5] – given a random variable with known mean ◇ and known standard deviation, ◇, we are sampling with size n, and we are interested in two new random variables: the sample mean, $\bar{X}$. If the size (n) of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. If the size (n) of the sample is sufficiently large, then the distribution of the sample means will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Cohen's d [5] – a measure of effect size based on the differences between two means. If d is between 0 and 0.2 then the effect is small. If d approaches is 0.5, then the effect is medium, and if d approaches 0.8, then it is a large effect.

Confidence Interval (CI)[5] – an interval estimate for an unknown population parameter. This depends on: (1) the desired confidence level (2) information that is known about the distribution (for example, known standard deviation) (3) the sample and its size

Confidence Level (CL)[5] – the percent expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90%, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Contingency Table[5] – a table that displays sample values for two different factors that may be dependent or contingent on one another; it facilitates determining conditional probabilities.

Critical Value[5] – The t or Z value set by the researcher that measures the probability of a Type I error, ◇

Degrees of Freedom (df)[5] – the number of objects in a sample that are free to vary

Error Bound for a Population Mean (EBM)[5] – the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation.

Error Bound for a Population Proportion (EBP)[5] – the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes.

Goodness-of-Fit[5] – a hypothesis test that compares expected and observed values in order to look

for significant differences within one non-parametric variable. The degrees of freedom used equals the (number of categories – 1).

Hypothesis[5] – a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation ◈$_0$) and the contradictory statement is called the alternative hypothesis (notation ◈$_◈$)

Hypothesis Testing[5] – Based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

Independent Groups[5] – two samples that are selected from two populations, and the values from one population are not related in any way to the values from the other population.

Inferential Statistics[5] – also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if four out of the 100 calculators sampled are defective we might infer that four percent of the production is defective.

Linear[5] – a model that takes data and regresses it into a straight line equation.

Matched Pairs[5] – two samples that are dependent. Differences between a before and after scenario are tested by testing one population mean of differences.

Mean[5] – a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{sum of all values in the sample}}{\text{number of values in the sample}}$, and the mean for a population (denoted by ◈) is $\frac{\text{sum of all values in the population}}{\text{number of values in the population}}$

Multivariate[5] – a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.

Normal Distribution[5] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, where ◈ is the mean of the distribution and ◈ is the standard deviation; notation: ◈~◈(◈,◈). If ◈=0 and ◈=1, the random variable, Z, is called the standard normal distribution.

Normal Distribution[5] – a continuous random variable with pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where ◈ is the mean of the distribution and ◈ is the standard deviation, notation: ◈~◈(◈,◈) and ◈=1, the random variable is called the standard normal distribution.

One-Way ANOVA[5] – a method of testing whether or not the means of three or more populations are equal; the method is applicable if: (1) all populations of interest are normally distributed (2) the populations have equal standard deviations (3) samples (not necessarily of the same size) are randomly and independently selected from each population. The test statistic for analysis of variance is the F-ratio

Parameter[5] – a numerical characteristic of a population

Point Estimate[5] – a single number computed from a sample and used to estimate a population parameter

Pooled Variance[5] – a weighted average of two variances that can then be used when calculating standard error.

R – Correlation Coefficient[5] – A number between –1 and 1 that represents the strength and direction of the relationship between "X" and "Y." The value for "r" will equal 1 or –1 only if all the plotted points form a perfectly straight line.

Residual or "Error"[5] – the value calculated from subtracting $y_0 - \hat{y}_0 = e_0$. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y that appears on the best-fit line.
Sampling Distribution[5] – Given simple random samples of size n from a given population with a measured characteristic such as mean, proportion, or standard deviation for each sample, the probability distribution of all the measured characteristics is called a sampling distribution.
Standard Deviation[5] – a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and ◈ for population standard deviation
Standard Error of the Mean[5] -the standard deviation of the distribution of the sample means, or $\frac{\sigma}{\sqrt{n}}$
Standard Error of the Proportion[5] -the standard deviation of the sampling distribution of proportions
Student's t-Distribution[5] – investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student; the major characteristics of this random variable are: (1) it is continuous and assumes any real values (2) the pdf is symmetrical about its mean of zero (3) it approaches the standard normal distribution as n gets larger (4) there is a "family" of t-distributions: each representative of the family is completely defined by the number of degrees of freedom, which depends upon the application for which the t is being used.
Sum of Squared Errors (SSE)[5] – the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.
Test for Homogeneity[5] – a test used to draw a conclusion about whether two populations have the same distribution. The degrees of freedom used equals the (number of columns – 1).
Test of Independence[5] – a hypothesis test that compares expected and observed values for contingency tables in order to test for independence between two variables. The degrees of freedom used equals the (number of columns – 1) multiplied by the (number of rows – 1).
Test Statistic[5] – The formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.
Type I Error[5] -The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.
Type II Error[5] – The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.
Variance[5] – mean of the squared deviations from the mean; the square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.
X – the Independent Variable[5] – This will sometimes be referred to as the "predictor" variable, because these values were measured in order to determine what possible outcomes could be predicted.
Y – the Dependent Variable[5] – also, using the letter "y" represents actual values while $\hat{y}$ represents predicted or estimated values. Predicted values will come from plugging in observed "x" values into a linear model

GLOSSARY REFERENCES

[1] *Merriam-Webster.com.* Retrieved July 6, 2023, from https://www.merriam-webster.com/

[2] *Wikipedia.com.* Retrieved January 16, 2024, from https://en.wikipedia.org/wiki/Exponential_growth
[3] "Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax, Chapter 5.3: The Fundamental Theorem of Calculus https://openstax.org/books/calculus-volume-1/pages/5-3-the-fundamental-theorem-of-calculus
[4] "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses
[5] "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction
[6] Berman H.B., "Statistics Dictionary", [online] Available at: https://stattrek.com/statistics/dictionary?definition=select-term URL [Accessed Date: 8/16/2022].

LINKS BY CHAPTER

Below are the attributions for each chapter of the textbook. Included is the name of the source, the author of the source, and the Creative Commons license for the source.

Note: Text by Greg Wolfe, et al.: Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

Note: All Khan Academy content is available for free at (www.khanacademy.org).

CHAPTER 1

- "Algebra 2 Trigonometry" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Similarity" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Trigonometry" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Law of sines: solving for a side | Trigonometry by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Law of cosines: solving for a side | Trigonometry (video) by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Right triangles & trigonometry | High school geometry by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Trigonometric equations and identities by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Precalculus 2e: Right Triangle Trigonometry" by Jay Abramson is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Algebra and Trigonometry: Chapter 7: Review Exercises and Practice Test" by Jay Abramson is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Algebra and Trigonometry: Chapter 10: Review Exercises and Practice Test." by Jay Abramson is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Elements of Simple Circular Curves" by Mr. Shashikant B. Gosavi is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Design of Vertical Curves on Roads" by Mr. Ashok Kumar is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

- "Types of Transition Curves" by Mr. Ashok Kumar is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

CHAPTER 2

- "Chapter 10: Exponential and Logarithmic Functions (10.2 and 10.3): Intermediate Algebra" by Lynn Marecek is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Unit: Exponential and Logarithmic Functions" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Graphing Lab – Quadratics and Polynomials" by Dr. Mary Alice Smeal is licensed by Creative Commons Public Domain Mark 1.0
- "Graphing Polynomial Functions" by Michael Pemberton is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Graphs of Polynomials" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Polynomial Expressions, Equations, & Functions" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Solving Polynomial Equations by Factoring" by Keith Mann is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Solving Polynomial Equations by Using Synthetic Substitution" by Keith Mann is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Solve Polynomial Equations by Factoring" by LibreTexts is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Dividing Polynomials" by LibreTexts is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Chapter 5: Polynomials and Polynomial Functions Review Exercises: Intermediate Algebra" by Lynn Marecek is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- https://www.merriam-webster.com/ is available under the *Creative Commons* CC0 License
- https://www.lexico.com/

CHAPTER 3

- "Example 4.9: Intermediate Algebra" by Lynn Marecek is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Chapter 4: Chapter Review Exercises and Practice Test: Intermediate Algebra" by Lynn Marecek, OpenStax is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Algebra (all content). Unit: System of Equations" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- Solving Linear Systems with 3 Variables by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Chapter 4.4 Examples: Intermediate Algebra" by Lynn Marecek, OpenStax is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Chapter 4, Chapter Review Exercises and Practice Test: Intermediate Algebra" by Lynn Marecek, OpenStax is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Using Matrices to Solve Systems of Linear Equations" by Linda Green is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "4.5 Solve Systems of Equations Using Matrices: Intermediate Algebra 2e" by Lynn Marecek and Andrea Honeycutt Mathis, OpenStax, is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Solve Systems of Linear Equations in Excel Without Using VBA" by Roel Van de Paar is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Solving Sets of Linear Equations – Excel Solver" by Rebecca Ong is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- https://www.merriam-webster.com/ is available under the *Creative Commons* CC0 License

CHAPTER 4

- "Interpreting Derivatives" by Linda Green is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit: Taking Derivatives" by Khan Academy by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit 3.1: Introduction to Derivatives Practice Problems: Calculus I" by The Saylor Foundation is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Contemporary Calculus" by Dave Hoffman is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit 2.1: Instantaneous Rates of Change – The Derivative: Virginia Military Institute Calculus" by Gregory Hartman et al. is licensed by Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
- "Unit: Taking Derivatives" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Continuity & Differentiability" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Differentiation: Composite, Implicit, and Inverse Functions: The Chain Rule: Introduction" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Differentiation: Definition and Basic Derivative Rules: Applying the Power Rule" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike

3.0 United States (CC BY-NC-SA 3.0 US)

- "Derivatives of Trigonometric Functions" by Tyler Wallace is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Chapter 3.5: Derivatives of Trigonometric Functions: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- "Unit: Taking Derivatives" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Chapter 3: Derivatives: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Calculus 1 Unit: Integrals" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Chapter 5: Integration Review Exercises: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Chapter 5.4 Exercises: Integration Formulas and the Net Change Theorem: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Chapter 5.5 Exercises: Substitution: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Calculus 1 Unit: Integrals" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Finding Area Under a Curve Using Integration" by Learning Videos is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Chapter 6.1: Area Between Curves: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Chapter 5.1 Approximately Areas: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- "Math 144 – 5.4 – The Fundamental Theorem of Calculus" by Vincent Bouchard is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Chapter 5.3: The Fundamental Theorem of Calculus: Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). Access for free.
- https://www.merriam-webster.com/ is available under the *Creative Commons* CC0 License

- "Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax, Chapter 5.3: The Fundamental Theorem of Calculus: https://openstax.org/books/calculus-volume-1/pages/5-3-the-fundamental-theorem-of-calculus

CHAPTER 5

- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- "Interpreting Motion Graphs" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- PhET Interaction Simulations, University of Colorado Boulder is licensed by is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0).
- "Unit: One-dimensional Motion" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Motion Diagram" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Comparing Motion Diagrams" by Andrew Duffy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- "Unit: One-dimensional Motion" by Khan Academy by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Free Body Diagrams" by James Dann, Ph.D is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- "Unit: Forces and Newton's Law of Motion" by Khan Academy by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Normal Force" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Kinetic and Static Friction Forces" by North Carolina School of Science and Mathematics s licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Static Friction Equation" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Kinetic Friction Equation" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Static and Kinetic Friction Example" by Khan Academy by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Practice: Static and Kinetic Friction" by Khan Academy by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-

SA 3.0 US)

- "Newton's 3 Laws of Motion" by Engineering Technology Simulation Learning Videos is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- PhET Interaction Simulations, University of Colorado Boulder Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Laws of Gravitation" by Study Animated is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- Introduction to spatial interaction modelling by Andy Newing through National Centre for Research Methods online learning resource is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Angle Units" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Angular Position and Displacement" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Angular Velocity and Acceleration" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Uniform Circular Motion" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit: Centripetal Force and Gravitation" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Fictious Forces" by Camilla Tac is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

CHAPTER 6

- "Unit: Impacts and Linear Momentum" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Momentum Notes" by Jeff Batey is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- PhET Interaction Simulations, University of Colorado Boulder is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0).
- "Equilibrium Forces" by Jennifer Cash is licensed by Creative Commons Attribution 3.0

Unported (CC BY 3.0)

- "Static Equilibrium: Concept" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- PhET Interaction Simulations, University of Colorado Boulder is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0).
- "Center of Gravity" by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Dynamic Balance" by Alejandro Garcia is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

CHAPTER 7

- "Basic Properties of Waves" by Animations for Physics and Astronomy is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- "Wave on a String" by PhET Simulations is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0).
- "Wave Interference" by PhET Simulations is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0).
- "Acoustics and Your Environment: The Bais of Sound and Highway Traffic Noise" by PublicResourceOrg is licensed by Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)
- "Sound Waves and Their Sources (1933) Remastered" by Sterocanvas is licensed by Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)
- "What is the Doppler Effect?" by Engineering Technology Simulation Learning Videos is licensed by Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)
- "What is Noise Pollution? What does noise pollution mean? Noise pollution meaning and explanation?" is licensed by Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)
- "Adsorptive Sound Wall Systems" by SoundFighterWalls is licensed by Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)
- "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access for free at https://openstax.org/books/college-physics-ap-courses/pages/

1-connection-for-ap-r-courses

CHAPTER 8

- "Introduction to Probability and Statistics" by Jeremy Orloff and Jonathan Bloom from MIT OpenCourseWare is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).
- "Unit: Counting, Permutations, and Combinations" by Khan Academy.
- "Unit: Probability" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- "Unit: Random Variables" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Bernoulli" by Paul Hewson is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 5 – Poisson Probability Distribution – Problem 5-6A" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit: Random Variables" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Uniform Distribution Intro" by Lee Wilsonwithers is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "The Exponential Distribution: Concept" by Significant Statistics is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "M 14 04: Integration by Parts – Exponential Distribution (Probability Theory)" by Mathematics 1: Calculus – by Maurice Koster is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 6 – Exponential Probability Distribution – Problem 6-4B" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit: Modeling Data Distributions" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction
- Berman H.B., "Statistics Dictionary", [online] Available at: https://stattrek.com/statistics/dictionary?definition=select-term URL [Accessed Date: 8/16/2022].

CHAPTER 9

- "Unit: Sampling Distributions" by Khan Academy is licensed by Creative Commons

Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- "Sampling Distributions: Sampling Distribution of the Mean (including Central Limit Theorem)" by onlinestatbook is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Confidence Intervals & Estimation: Point Estimates Explained" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Confidence Interval for Mean: 1 Sample Z Test (Using Formula)" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Confidence Intervals: Using the t Distribution" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "How to Construct a Confidence Interval for Population Proportion" by Simple Science and Maths is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0).
- "Calculating Sample Size to Predict a Population Proportion" by Matthew Simmons is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Estimation and Confidence Intervals: Calculate Sample Size" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: The Fundamentals" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: Setting up the Null and Alternative Hypothesis Statements" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: Type I and Type II Errors" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: Finding Critical Values" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "What is a 'P-Value'?" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Normal Distribution: Finding Critical Values of Z" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: One Sample Z Test of the Mean (Critical Value Approach)" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: t Test for the Mean using Critical Value Approach" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: 1 Sample Z Test of the Mean (Confidence Interval Approach)" by Linda Willliams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Hypothesis Testing: 1 Sample Z Test for Mean (P Value Approach)" by Linda Williams is

licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

- "Hypothesis Testing: 1 Proportion using the Critical Value Approach" by Linda Williams is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics – Module 10 – Hypothesis Testing – Two Populations Means and Proportions" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 – Two Population Means, One Tail Test, Sigma Unknown – Problem 10-2A" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics – Module 10 V2 – Two Population Means, Two Tail Test, Sigma Unknown – Problem 10-2C" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Stats 11.3 Effect size for a significant difference of two sample means" by Dana Lee Ling is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Two Population Proportions, One Tail Test, Problem 10-4A" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Two Population Proportions, Two Tail Test, Problem 10-4C by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Two Population Means, One Tail Test, Sigma Known, Problem 10-1B" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Two Population Means, Two Tail Test, Sigma Known, Problem 10-1D" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Matched Sample, One Tail Test, Problem, 10-3A" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics Module 10 V2 – Matched Sample, One Tail Test, Problem 10-3B" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Matched Sample, Two Tail Test, Problem 10-3C" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Unit: Inference for Categorical Data: Chi-Square" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Simple Explanation of Chi-Squared" by J David Eisenberg is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics – Module 11 – Single Pop Variances, one-tail test – Problem 11-1C" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Statistics – Module 11 – Hypothesis Test Two Pop Variances – Problem 11-2A" by Peter Dalley is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Analysis of Variance (ANOVA)" by J David Eisenbeg is licensed by Creative Commons

Attribution 3.0 Unported (CC BY 3.0)

- "Unit: Analysis of Variance (ANOVA)" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Exploriing Two-Variable Quantitative Data" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Using MS Excel's 'Descriptive Statistics' Tool" by Linda Weiser Friedman is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "How to Run Descriptive Statistics in R …[Mean, Median, Mode – Measures of Central Tendency – Tips] by Learning Puree is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Descriptive Statistics in R [Range, IQR, Variance, SD, MAD, CV] – How to Run Descriptive Statistics in R? by Learning Puree is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- "Simple Linear Regression in Excel | with Interpretation of Regression Output by Ramya Rachel is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Simple Linear Regression, fit and interpretations" by Edward Roualdes is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)
- "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction

IMAGE CREDITS

This work contains copyrighted material the use of which has not always been specifically authorized by the copyright owners. The creators of this OER believe in good faith that reuse of these materials constitutes a "fair use" as described in Section 107 of the Copyright Act. Downstream users who wish to use copyrighted material from this resource for purposes beyond those authorized by the fair use doctrine should obtain permission from the copyright owner(s) of the original content.

Unless noted below, all figures used in the text are the intellectual property of the content creators and are licensed under a Creative Commons Attribution 4.0 International license.

NOTES

- All Khan Academy content is available for free at (www.khanacademy.org).
- Text by Greg Wolfe, et al.: Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

CHAPTER 1

- Figure 1: "Landscape, Aerial View, Mountain Road" by Freddy Dendoktoor under (CC0 1.0)
- Figure 2: "Curvy Road" no author by (CC0 1.0)
- Figure 3: "Sample intersection sight triangle" by Lin et al. under public domain

CHAPTER 5

- Figure 1: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 2: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 3:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 4: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 5: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

- Figure 6-8: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 9:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 10: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 11: "Unit: Forces and Newton's Law of Motion" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Figure 12: "Unit: Forces and Newton's Law of Motion" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Figure 13: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 14: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 15: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 16: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 17-18: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 19: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 20: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 21: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

- Figure 22: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 23: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 24: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

CHAPTER 6

- Figure 1: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 2: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 3: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 4: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 5: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 6:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 7: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 8: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya,

Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

- Figure 9: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 10: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 11: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 12: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 13: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 14: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 15: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 16: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.
- Figure 17 is licensed by Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

CHAPTER 7

- Figure 1: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0

International (CC BY 4.0)

- Figure 2: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 3: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 4: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 5: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 6: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 7: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 8: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 9: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 10: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)
- Figure 11: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

CHAPTER 8

- Figure 1 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Figure 2: by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- Figure 3 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 4 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 5 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 6 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 7 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 8 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 9 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 10 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 11 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 12 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 13 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 14 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 15 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 16 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 17 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 18 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

- Figure 19 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 20 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 21 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 22 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 23 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 24 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 25 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)
- Figure 26 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 27 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 28 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Figure 29 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

CHAPTER 9

- Figure 1: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 2: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 3: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 4: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 5: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 6: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 7: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

- Figure 8: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 9: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 10: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 11: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 12: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 13: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 14: by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 15 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 16 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 17 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 18 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 19 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 20 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 21 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 22 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 23 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 24 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- Figure 25 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

REFERENCES

Note: All Khan Academy content is available for free at (www.khanacademy.org).

CHAPTER 1

- Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). Considerations for Intersections. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

CHAPTER 2

- Farid, A. (2022). *Engineering Economics.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Mannering, F., and Washburn, S. (2013). Chapter 5: Fundamentals of Traffic Flow and Queuing Theory. In: *Principles of Highway Engineering and Traffic Analysis 5th Edition*. John Wiley & Sons, Inc., Hoboken, NJ. pp. 135-174.
- Farid, A. (2022). *Transportation Planning 1.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). *Transportation Planning 2.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). *Highway Design.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

CHAPTER 3

- Farid, A. (2022). Transportation Planning 2. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

CHAPTER 5

- Farid, A. (2021/2022). *Chapter 4: Modeling Transportation Demand and Supply Transportation Planning 1.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2021/2022). *Chapter 9Highway Design.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- RahimFarid, A. (2022). *Traffic Safety and Human Factor Considerations Superelevation.* Personal

Collection of Ashraf Rahim Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

- RahimFarid, A. (2022). *Stopping Sight Distance (SSD)Pavement Engineering.* Personal Collection of Ashraf Rahim Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2021/2022). *Chapter 2 Fundamentals of Traffic Flow Theory.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- "Traffic Flow" by David Levinson https://www.youtube.com/watch?v=NjxIhgG4cNw

CHAPTER 6

- Farid, A. (2022). Traffic Safety and Human Factor Considerations. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

CHAPTER 7

- "Traffic Noise and Transportation" by The Center for Environmental Excellence by the American Association of State Highway and Transportation Officials (AASHTO).

CHAPTER 8

- "Unit: Counting, Permutations, and Combinations" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/counting-permutations-and-combinations
- "Unit: Probability" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/probability-library
- "Unit: Random Variables" by Khan Academy https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library
- "Unit: Modeling Data Distributions" by Khan Academy https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data
- Mannering, F., and Washburn, S. (2013). Chapter 5: Fundamentals of Traffic Flow and Queuing Theory. In: Principles of Highway Engineering and Traffic Analysis 5th Edition. John Wiley & Sons, Inc., Hoboken, NJ. pp. 135-174.
- Farid, A. (2022). *Spot Speed Study*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- American Association of State Highway and Transportation Officials (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials, Washington, D.C.

CHAPTER 9

- "Unit: Sampling Distributions" by Khan Academy is licensed by Creative Commons

NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

- "Unit: Inference for Categorical Data: Chi-Square" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Analysis of Variance (ANOVA)" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Unit: Exploring Two-Variable Quantitative Data" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)
- "Statistics" by Barbara Illosky and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)
- "Introduction to Probability and Statistics" by Jeremy Orloff and Jonathan Bloom is licensed by Creative Commons NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
- Farid, A. (2022). Transportation Planning 1. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). Transportation Planning 2. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Farid, A. (2022). *Fundamentals of Traffic Flow Theory*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.
- Washington, S., Karlaftis, M., Mannering, F., and Anastasopoulos, P. (2020). *Statistical and Econometric Methods for Transportation Data Analysis 3rd Edition*. Chemical Rubber Company (CRC) Press, Taylor and Francis Group, Boca Raton, FL.
- Farid, A. (2022). *Spot Speed Study.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

DERIVATIVE NOTES

Minor revisions not noted below include editing language for clarity, length, and flow as well as corrections to and additions of hyperlinks and citations. Other revisions not listed below include removing references to sections of the text that were not included in the book/article. The following is a chapter list of source material used in the creation of the text.

Note: All Khan Academy content is available for free at (www.khanacademy.org).

CHAPTER 1: TRIGONOMETRY FUNCTIONS AND GEOMETRIC MEASUREMENTS

Videos

Video 1: Radians & Degrees by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 2: Solving Similar Triangles by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: Introduction to the Trigonometric Ratios by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: The Trig Functions & Right Triangle Trig Ratios by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: Solving for a Side in a Right Triangle Using the Trigonometric Ratios by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 6: Introduction to arcsine by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 7: Introduction to arctangent by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 8: Introduction to arccosine by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 9: Law of Sines by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 10: Law of Cosines by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 11: Elements of Simple Circular Curves by Mr. Shashikant B. Gosavi, Assistant Professor,

Department of Civil Engineering, Walchand Institute of Technology, Solapur is licensed by Attribution 3.0 Unported (CC BY 3.0)

Video 12: Elements of Parabolic Curves by Mr. Ashok Kumar, Assistant Professor, Department of Civil Engineering, Walchand Institute of Technology, Solapur is licensed by Attribution 3.0 Unported (CC BY 3.0)

Video 13: Elements of Spiral Curves by Mr. Ashok Kumar, Assistant Professor, Department of Civil Engineering, Walchand Institute of Technology, Solapur is licensed by Attribution 3.0 Unported (CC BY 3.0)

References

Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). Considerations for Intersections. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Figures

Figure 1: is licensed by Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

Figure 2 is licensed by Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

Figure 3 is licensed by Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

CHAPTER 2: POLYNOMIAL, EXPONENTIAL, AND LOGARITHMIC FUNCTIONS

Videos

Video 1: Introduction to Exponential Functions by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 2: Exponential Function Graph by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: Graphing Exponential Functions by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Graphs of Exponential Growth by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: Graphing Exponential Growth and Decay by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 6: Analyzing Graphs of Exponential Functions by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 7: Analyzing Tables of Exponential Functions by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 8: Introduction to Logarithms by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 9: Graphs of Logarithmic Functions by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 10: Relationship Between Exponentials and Logarithms: Graphs by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 11: Relationship Between Exponentials and Logarithms: Tables by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 12: Solving Exponential Equations Using Exponent Properties by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 13: Solving Exponential Equations Using Exponent Properties (Advanced) by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 14: Logarithmic Equations: Variable in the Argument by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 15: Logarithmic Equations: Variable in the Base by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 16: Graphing Polynomial Functions by Michael Pemberton is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 17: Solving Polynomial Equations by Factoring by Keith Mann is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 18: Dividing Polynomials: Long Division by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 19: Solving Polynomial Equations by using Synthetic Substitution by Keith Mann is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

References

Farid, A. (2022). *Engineering Economics.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Mannering, F., and Washburn, S. (2013). Chapter 5: Fundamentals of Traffic Flow and Queuing

Theory. In: *Principles of Highway Engineering and Traffic Analysis 5th Edition*. John Wiley & Sons, Inc., Hoboken, NJ. pp. 135-174.

Farid, A. (2022). *Transportation Planning 1.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Transportation Planning 2.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Highway Design.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA

CHAPTER 3: SYSTEMS OF LINEAR EQUATIONS

Videos

Video 1: "How to set up a system of equations" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 2: "How to solve systems of equations graphically" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: "Testing a solution to a system of equations" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Solving System of Equations with Substitution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: "Solving Systems of Equations with Elimination" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 6: "Introduction to Linear Systems with Three Variables" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 7: "Solving Linear Systems with Three Variables" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 8: "Solving Linear Systems with Three Variables: No Solution" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 9: "Using Matrices to Solve Systems of Linear Equations" by Linda Green is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 10: "Solving System of Equations Using Excel" by Roel Van de Paar is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 11: How to Use the Solver Tool in Excel to Solve Systems of Linear Equations in Algebra by Rebecca Ong is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

References

Farid, A. (2022). Transportation Planning 2. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

CHAPTER 4: CALCULUS – INTERPRETATION AND METHODS FOR INTEGRATION AND DIFFERENTIATION

Videos

Video 1: Interpreting Derivatives by Linda Green is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 2: Derivative as Slope of Curve by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: The Derivative as Slope of Tangent Line by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Tangent Slope as Instantaneous Rate of Change by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: Approximating Instantaneous Rate of Change with Average Rate of Change by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 6: Basic Derivative Rules by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 7: Basic Derivative Rules (Part 2) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 8: Product Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 9: Quotient Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 10: Chain Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 11: The Power Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 12: Derivatives of Trigonometric Functions by Tyler Wallace is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 13: Differentiability at a Point: Algebraic by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 14: Differentiating Polynomials by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 15: Fractional Powers Differentiation by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 16: Radical Functions Differentiation Introduction by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 17: Worked Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 18: Differentiating Products by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 19: Differentiate Quotients by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 20: Differentiating Rational Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 23: Derivatives of tan(x) and cot(x) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 24: Derivatives of sec(x) and csc(x) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 25: Differentiate Exponential Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 26: Differentiate Logarithmic Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 27: Indefinite Integral of 1/x by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 28: Indefinite Integrals of sin(x), cos(x), and by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 29: Reverse Power Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 30: Rational Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 31: Radical Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 32: Trig Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 33: Natural Logs by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 34: Absolute Value Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 35: Piecewise Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 36: Integrating with U-Substitution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 37: U-Substitution: Definite Integral of Exponential Function by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 38: Introduction to Integral Calculus by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 39: Introduction to Definite Integrals by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 40: Introduction to Riemann Approximation by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 41: Definite Integral as the Limit of a Riemann Sum by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 42: Finding Area Under a Curve Using Integration by Learning Videos is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 43: The Fundamental Theorem of Calculus by Vincent Bouchard is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

References

"Calculus Volume 1" by Gilbert Strang, Edwin "Jed" Herman on OpenStax, Chapter 5.3: The Fundamental Theorem of Calculus: https://openstax.org/books/calculus-volume-1/pages/5-3-the-fundamental-theorem-of-calculus

CHAPTER 5: MOTION AND FORCES

Videos

Video 1: Interpreting Motion Graphs by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 2: Motion Diagram by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 3: Choosing Kinematic Equations by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Free-body Diagrams by James Dann, Ph.D is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Video 5: Normal Force by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 6: Kinetic and Static Friction Forces by North Carolina School of Science and Mathematics is licensed by Creative Commons Attribution 3.0 Unported

Video 7: Static Friction Equation by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 8: Kinetic Friction Equation by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 9: Static and Kinetic Friction Example by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 10: Newton's 3 Laws of Motion by Engineering Technology Simulation Learning Videos is licensed by Creative Commons Attribution 3.0 Unported

Video 11: Laws of Gravitation by Study Animated is licensed by Creative Commons Attribution 3.0 Unported

Video 12: Introduction to Spatial Interaction Modelling by National Centre for Research Methods online learning resource is licensed by Creative Commons Attribution 3.0 Unported

Video 13: Angle Units by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 14: Angular Position and Displacement by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 15: Angular Velocity and Acceleration by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 16: Uniform Circular Motion by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported

Video 17: Centripetal Force and Acceleration Intuition by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 18: Visual Understanding of Centripetal Acceleration Formula by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 19: Centripetal Force Problem Solving by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 20: Fictitious Forces by by Camilla Tac is licensed by Creative Commons Attribution 3.0 Unported

Figures

Figure 1: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 2: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 3:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 4: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 5: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 6-8: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 9:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 10: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 11: "Unit: Forces and Newton's Law of Motion" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Figure 12: "Unit: Forces and Newton's Law of Motion" by Khan Academy is licensed by Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Figure 13: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 14: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 15: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 16: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 17-18: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 19: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 20: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 21: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 22: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 23: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 24: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0)

References

Farid, A. (2021/2022). *Chapter 4: Modeling Transportation Demand and Supply Transportation Planning 1.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2021/2022). *Chapter 9Highway Design.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Traffic Safety and Human Factor Considerations Superelevation.* Personal Collection of Ashraf Rahim Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Stopping Sight Distance (SSD)Pavement Engineering.* Personal Collection of Ashraf Rahim Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2021/2022). *Chapter 2 Fundamentals of Traffic Flow Theory.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

"Traffic Flow" by David Levinson https://www.youtube.com/watch?v=NjxIhgG4cNw

CHAPTER 6: BASIC DYNAMICS AND STATIC EQUILIBRIUM

Videos

Video 10: Deriving the Shortcut to Solve Elastic Collision Problems by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 11: How to Use the Shortcut for Solving Elastic Collisions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 12: Equilibrium Forces by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 13: Static Equilibrium by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 14: Center of Gravity by Jennifer Cash is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Video 15: Dynamic Balance by Alejandro Garcia is licensed by Creative Commons Attribution 3.0 Unported (CC BY 3.0)

Figures

Figure 1: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 2: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 3: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 4: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 5: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 6:"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 7: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 8: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 9: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 10: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 11: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 12: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 13: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 14: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 15: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 16: "College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram is licensed by Creative Commons Attribution 4.0 International (CC BY 4.0). Access for free.

Figure 17 is licensed by Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

References

Farid, A. (2022). Traffic Safety and Human Factor Considerations. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). Highway Design. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access

for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

CHAPTER 7: WAVES AND DOPPLER EFFECT

Videos

Video 1:"Basic Properties of Waves" by Animations for Physics and Astronomy is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)

Video 2: "Sound Waves and Their Sources (1933) Remastered" by Sterocanvas is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)

Video 3: "Acoustics and Your Environment: The Bais of Sound and Highway Traffic Noise" by PublicResourceOrg is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)

Video 4: "What is the Doppler Effect?" by Engineering Technology Simulation Learning Videos is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)

Video 5: "Adsorptive Sound Wall Systems" by SoundFighterWalls is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)

Figures

Figure 1: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 2: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 3: "Chapter 16: Oscillatory Motion and Waves" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 4: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 5: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 6: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 7: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 8: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 9: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 10: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Figure 11: "Chapter 17: Physics of Hearing" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, and Douglas Ingram is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Tables

Table 1: Garnett, A. (2022, April 26). Traffic noise overview. Center for Environmental Excellence | AASHTO. Retrieved July 1, 2022, from https://environment.transportation.org/education/environmental-topics/traffic-noise/traffic-noise-overview/

Simulations

Simulation 1: "Wave on a String" by PhET Simulations is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

Simulation 2: "Wave Interference" by PhET Simulations is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

References

"Traffic Noise and Transportation" by The Center for Environmental Excellence by the American Association of State Highway and Transportation Officials (AASHTO).

"College Physics for AP® Courses" by Greg Wolfe, Erika Gasper, John Stoke, Julie Kretchman, David Anderson, Nathan Czuba, Sudhi Oberoi, Liza Pujji, Irina Lyublinskaya, Douglas Ingram. Access for free at https://openstax.org/books/college-physics-ap-courses/pages/1-connection-for-ap-r-courses

CHAPTER 8: PROBABILITY: BASIC PRINCIPLES AND DISTRIBUTIONS

Videos

Video 1: Count Outcomes Using Tree Diagram by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 2: Permutation Formula by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: Zero Factorial by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Ways to Pick Officers by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: Introduction to Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 6: Combination Formula by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 7: Combination Example: 9 Card Hands by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 8: Probability Using Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 9: Probability and Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 10: Conditional Probability and Combinations by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 11: Random Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 12: Discrete and Continuous Random Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 13: Constructing a Probability Distribution for Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 14: Probability with Discrete Random Variable Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 15: Mean (expected value) of a Discrete Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 16: Variance and Standard Deviation of a Discrete Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 17: Bernoulli by Paul Hewson is licensed by Attribution 3.0 Unported (CC BY 3.0)

Video 18: Mean and Variance of Bernoulli Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 19: Bernoulli Distribution Mean and Variance by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 20: Binomial Variables by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 21: Binomial Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 22: Visualizing a Binomial Distribution by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 23: Binomial Probability Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 24: Expected Value of a Binomial Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 25: Variance of a Binomial Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 26: Finding the Mean and Standard Deviation of a Binomial Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 27: Geometric Random Variables Introduction by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 28: Probability for a Geometric Random Variable by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 29: Cumulative Geometric Probability (greater than a value) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 30: Cumulative Geometric Probability (less than a value) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 31: Poisson Process 1 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 32: Poisson Process 2 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 33: Poisson Probability Distribution – Example by Peter Dalley is licensed under the Attribution 3.0 Unported (CC BY 3.0)

Video 34: Probability Density Functions by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 35: Uniform Distributions by Lee Wilsonwithers is licensed under Attribution 3.0 Unported (CC BY 3.0)

Video 36: Exponential Distributions by Significant Statistics is licensed under Attribution 3.0 Unported (CC BY 3.0)

Video 37: "Statistics Module 6 – Exponential Probability Distribution – Problem 6-4B" by Peter Dalley is licensed under Attribution 3.0 Unported (CC BY 3.0)

Video 38: Integration by Parts – Exponential Distribution by Maurice Koster is licensed under Attribution 3.0 Unported (CC BY 3.0)

Video 39: Normal Distribution Problems: Empirical Rule by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 40: Probabilities from Density Curves by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 41: Standard Normal Table for Proportion Below by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 42: Standard Normal Table for Proportion Above by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 43: Standard Normal Table for Proportion Between Values by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 44: Finding z-score for a Percentile by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 45: Queuing Theory: Exponential Distribution by MOOC at IMT is licensed under Attribution 3.0 Unported (CC BY 3.0)

Figures

Figure 1 by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Figure 2: by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Figure 3 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 20 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 21 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)

Figure 22 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)

Figure 23 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 24 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)

Figure 25 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Attribution 4.0 International (CC BY 4.0)

Figure 26 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 27 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 28 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Figure 29 by Jeremy Orloff and Jonathan Bloom is licensed by NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

References

"Unit: Counting, Permutations, and Combinations" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/counting-permutations-and-combinations

"Unit: Probability" by Khan Academy. https://www.khanacademy.org/math/statistics-probability/probability-library

"Unit: Random Variables" by Khan Academy https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library

"Unit: Modeling Data Distributions" by Khan Academy https://www.khanacademy.org/math/statistics-probability/modeling-distributions-of-data

Mannering, F., and Washburn, S. (2013). Chapter 5: Fundamentals of Traffic Flow and Queuing Theory. In: Principles of Highway Engineering and Traffic Analysis 5th Edition. John Wiley & Sons, Inc., Hoboken, NJ. pp. 135-174.

Farid, A. (2022). *Spot Speed Study*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

American Association of State Highway and Transportation Officials (2010). *Highway Safety Manual*. American Association of State Highway and Transportation Officials,

"Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean on OpenStax. Access for free at https://openstax.org/books/introductory-business-statistics/pages/1-introduction

Berman H.B., "Statistics Dictionary", [online] Available at: https://stattrek.com/statistics/dictionary?definition=select-term URL [Accessed Date: 8/16/2022].

CHAPTER 9: DATA ANALYSIS – HYPOTHESIS TESTING, ESTIMATING SAMPLE SIZE, AND MODELING

Videos

Video 1: Central Limit Theorem by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 2: Sampling Distribution of the Sample Mean by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 3: Sampling Distribution of the Sample Mean (Part 2) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 4: Sampling Distributions: Sampling Distribution of the Mean by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 5: Confidence Intervals & Estimation: Point Estimates Explained by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 6: Confidence Interval for Mean: 1 Sample Z Test (Using Formula) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 7: Confidence Intervals: Using the t Distribution by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 8: How to Construct a Confidence Interval for Population Proportion by Simple Science and Maths is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 9: Estimation and Confidence Intervals: Calculate Sample Size by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 10: Calculating Sample size to Predict a Population Proportion by Matthew Simmons is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 11: Hypothesis Testing: The Fundamentals by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 12: Hypothesis Testing: Setting up the Null and Alternative Hypothesis Statements by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 13: Hypothesis Testing: Type I and Type II Errors by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 14: Hypothesis testing: Finding Critical Values by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 15: Normal Distribution: Finding Critical Values of Z by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 16: What is a "P-Value?" by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 17: Hypothesis Testing: One Sample Z Test of the Mean (Critical Value Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 18: Hypothesis Testing: t Test for the Mean (Critical Value Approach)by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 19: Hypothesis Testing: 1 Sample Z Test of the Mean (Confidence Interval Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 20: Hypothesis Testing: 1 Sample Z Test for Mean (P-Value Approach) by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 21: Hypothesis Testing: 1 Proportion using the Critical Value Approach by Linda Williams is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 22: Hypothesis Testing – Two Population Means by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 23: Two Population Means, One Tail Test, unknown by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 24: Two Population Means, Two Tail Test, unknown by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 25: Effect Size for a Significant Difference of Two Sample Means by Dana Lee Ling is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 26: Two Population Means, One Tail Test, Known by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 27: Two Population Means, Two Tail Test, Known by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 28: Two Population Means, One Tail Test, Matched Sample by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 29: Two Population Means, One Tail test, Matched Sample by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 30: Matched Sample, Two Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 31: Two Population Proportions, One Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 32: Two Population Proportion, Two Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 33: Single Population Variances, One-Tail Test by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 34: Chi-Square Statistic for Hypothesis Testing by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 35: Chi-Square Goodness-of-Fit Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 36: Simple Explanation of Chi-Squared by J David Eisenberg is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 37: Chi-Square Tet for Association (independence)by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 38: Introduction to the Chi-Square Test for Homogeneity by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 39: Hypothesis Test Two Population Variances by Peter Dalley is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 40: ANOVA by J David Eisenberg is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 41: Calculating SST (total sum of squares) by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 42: Calculating by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 43: Hypothesis Test with F-Statistic by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 44: Bivariate Relationship Linearity, Strength, and Direction by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 45: Calculating Correlation Coefficient r by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 46: Example: Correlation Coefficient Intuition by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 47: Introduction to Residuals and Least-Squares Regression by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 48: Calculating Residual Example by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 49: Residual Plots by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 50: Calculating the Equation of a Regression Line by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 51: Interpreting Slope of Regression Line by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 52: Interpreting y-intercept in Regression Model by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 53: Using Least Squares Regression Output by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 54: R-Squared or Coefficient of Determination by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

Video 55: Using Microsoft Excel's "Descriptive Statistics" Tool by Linda Weiser Friedman is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 56: How to Run Descriptive Statistics in R by Learning Puree is licensed by Creative Commons 4.0 International (CC BY 4.0)

Video 57: Descriptive Statistics in R Part 2 by Learning Puree is licensed by Creative Commons 4.0 International (CC BY 4.0)

Video 58: Simple Linear Regression in Excel by Ramya Rachel is licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Video 59: Simple Linear Regression, fit and interpretations in Ris licensed by Creative Commons 3.0 Unported (CC BY 3.0)

Figures

Figure 1: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 2: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 3: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 4: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 5: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 6: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 7: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 8: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 9: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 10: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 11: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 12: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 13: "Introductory Business Statistics" by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 14: by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 15 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 16 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 17 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 18 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 19 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 20 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 21 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 22 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 23 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 24 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

Figure 25 by Alexander Holmes, Barbara Illowsky, and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

References

"Unit: Sampling Distributions" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

"Unit: Inference for Categorical Data: Chi-Square" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

"Unit: Analysis of Variance (ANOVA)" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

"Unit: Exploring Two-Variable Quantitative Data" by Khan Academy is licensed by Creative Commons NonCommercial-ShareAlike 3.0 United States (CC BY-NC-SA 3.0 US)

"Statistics" by Barbara Illosky and Susan Dean is licensed by Creative Commons 4.0 International (CC BY 4.0)

"Introduction to Probability and Statistics" by Jeremy Orloff and Jonathan Bloom is licensed by Creative Commons NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Farid, A. (2022). *Transportation Planning 1*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Transportation Planning 2*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Farid, A. (2022). *Fundamentals of Traffic Flow Theory*. Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

Washington, S., Karlaftis, M., Mannering, F., and Anastasopoulos, P. (2020). *Statistical and Econometric Methods for Transportation Data Analysis 3rd Edition*. Chemical Rubber Company (CRC) Press, Taylor and Francis Group, Boca Raton, FL.

Farid, A. (2022). *Spot Speed Study.* Personal Collection of Ahmed Farid, California Polytechnic State University, San Luis Obispo, CA.

ERRATA AND VERSIONING HISTORY

MAVS OPEN PRESS

The peer review process for Mavs Open Press resources varies by publication and author. We strive for transparency and describe the creation process in the front matter of each text so readers may evaluate the work on its own merit. Because we offer web-based services, we can quickly and easily address any errors that arise after publication. This page provides a record of edits and changes made to the web version of this book since its initial publication.

Changes logged here are reflected in the web text only until a new version of the resource is released, at which point the file downloads are also updated. Version updates are noted in the Status field below.

If you have a correction to suggest, submit it to library-open@listserv.uta.edu. We will contact the author then make and log necessary changes here. Below is a list of Correction Types we currently update:

1. Typo
2. Broken link
3. Addition
4. Other factual inaccuracy in content
5. Incorrect calculation or solution
6. General/pedagogical suggestion or question

Date Submitted	Format	Correction Type	Location	Description