NATURAL RESOURCES BIOMETRICS

Diane Kiernan
SUNY College of Environmental Science
and Forestry

LibreTexts™

SUNY College of Environmental Science and Forestry

Natural Resources Biometrics

Diane Kiernan

TABLE OF CONTENTS

9: Modeling Growth, Yield, and Site Index

10: Quantitative Measures of Diversity, Site Similarity, and Habitat Suitability

11: Biometric Labs

Index

Glossary

Glossary

Detailed Licensing

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing****.*

CHAPTER OVERVIEW

1: Descriptive Statistics and the Normal Distribution

Topic hierarchy

1.1: Descriptive Statistics

https://milnepublishing.geneseo.edu/...017/02/958.png

https://milnepublishing.geneseo.edu/...e35759_fmt.png

https://web.archive.org/web/20230328...e35759_fmt.png

A **population** is the group to be studied, and population data is a collection of **all** elements in the population. For example:

- All the fish in Long Lake.
- All the lakes in the Adirondack Park.
- All the grizzly bears in Yellowstone National Park.

A **sample** is a subset of data drawn from the population of interest. For example:

- 100 fish randomly sampled from Long Lake.
- 25 lakes randomly selected from the Adirondack Park.
- 60 grizzly bears with a home range in Yellowstone National Park.

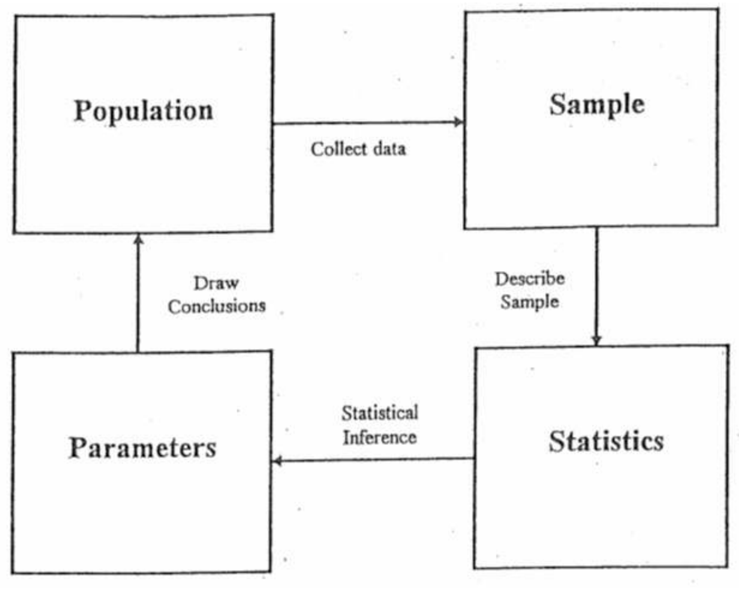


Figure 1*. Using sample statistics to estimate population parameters.*

Populations are characterized by descriptive measures called parameters. Inferences about **parameters** are based on sample **statistics**. For example, the population mean (μ) is estimated by the sample mean ($\bar{x}$). The population variance (σ^2) is estimated by the sample variance (s^2).

Variables are the characteristics we are interested in. For example:

- The length of fish in Long Lake.
- The pH of lakes in the Adirondack Park.
- The weight of grizzly bears in Yellowstone National Park.

Variables are divided into two major groups: qualitative and quantitative. Qualitative variables have values that are attributes or categories. Mathematical operations cannot be applied to qualitative variables. Examples of qualitative variables are gender, race, and petal color. **Quantitative** variables have values that are typically numeric, such as measurements. Mathematical operations can be applied to these data. Examples of quantitative variables are age, height, and length. **Quantitative** variables can be broken down further into two more categories: discrete and continuous variables. Discrete variables have a finite or countable number of possible values. Think of discrete variables as "hens". Hens can lay 1 egg, or 2 eggs, or 13 eggs… There are a limited, definable number of values that the variable could take on.

Continuous variables have an infinite number of possible values. Think of continuous variables as “cows”. Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk … There are an almost infinite number of values that a continuous variable could take on.

✓ Example 1.1.1:

Is the variable qualitative or quantitative?

a. Species
b. Weight
c. Diameter
d. Zip Code

Solution

(qualitative quantitative, quantitative, qualitative)

Descriptive Measures

Descriptive measures of populations are called parameters and are typically written using Greek letters. The population mean is μ (mu). The population variance is σ^2 (sigma squared) and population standard deviation is σ (sigma). Descriptive measures of samples are called statistics and are typically written using Roman letters. The sample mean is $\bar{x}$ (x-bar). The sample variance is s^2 and the sample standard deviation is s. Sample statistics are used to estimate unknown population parameters. In this section, we will examine descriptive statistics in terms of measures of center and measures of dispersion. These descriptive statistics help us to identify the center and spread of the data.

Measures of Center

Mean

The arithmetic mean of a variable, often called the average, is computed by adding up all the values and dividing by the total number of values. The population mean is represented by the Greek letter μ (mu). The sample mean is represented by $\bar{x}$ (x-bar). The sample mean is usually the best, unbiased estimate of the population mean. However, the mean is influenced by extreme values (outliers) and may not be the best measure of center with strongly skewed data. The following equations compute the population mean and sample mean.

$$\mu = \frac{\sum x_i}{N} \tag{1.1.1}$$

$$\bar{x} = \frac{\sum x_i}{n} \tag{1.1.2}$$

where x_i is an element in the data set, N is the number of elements in the population, and n is the number of elements in the sample data set.

✓ Example 1.1.2: mean

Find the mean for the following sample data set:

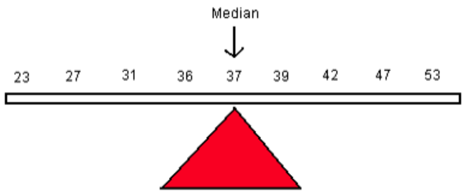
6.4, 5.2, 7.9, 3.4

Solution

$$\bar{x} = \frac{6.4 + 5.2 + 7.9 + 3.4}{4} = 5.725$$

Median

The median of a variable is the middle value of the data set when the data are sorted in order from least to greatest. It splits the data into two equal halves with 50% of the data below the median and 50% above the median. The median is resistant to the influence of outliers, and may be a better measure of center with strongly skewed data.



The calculation of the median depends on the number of observations in the data set.

To calculate the median with an odd number of values (*n* is odd), first sort the data from smallest to largest.

✓ Example 1.1.3: Calculating Median with Odd number of values

Find the median for the following sample data set:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47, 51$$

Solution

The median is 39. It is the middle value that separates the lower 50% of the data from the upper 50% of the data.

To calculate the median with an even number of values (*n* is even), first sort the data from smallest to largest and take the average of the two middle values.

✓ Example 1.1.4: Calculating Median with even number of values

Find the median for the following sample data set:

$$23, 27, 29, 31, 35, 39, 40, 42, 44, 47$$

Solution

$$M = \frac{35 + 39}{2} = 37$$

Mode

The mode is the most frequently occurring value and is commonly used with qualitative data as the values are categorical. Categorical data cannot be added, subtracted, multiplied or divided, so the mean and median cannot be computed. The mode is less commonly used with quantitative data as a measure of center. Sometimes each value occurs only once and the mode will not be meaningful.

Understanding the relationship between the mean and median is important. It gives us insight into the distribution of the variable. For example, if the distribution is skewed right (positively skewed), the mean will increase to account for the few larger observations that pull the distribution to the right. The median will be less affected by these extreme large values, so in this situation, the mean will be larger than the median. In a symmetric distribution, the mean, median, and mode will all be similar in value. If the distribution is skewed left (negatively skewed), the mean will decrease to account for the few smaller observations that pull the distribution to the left. Again, the median will be less affected by these extreme small observations, and in this situation, the mean will be less than the median.

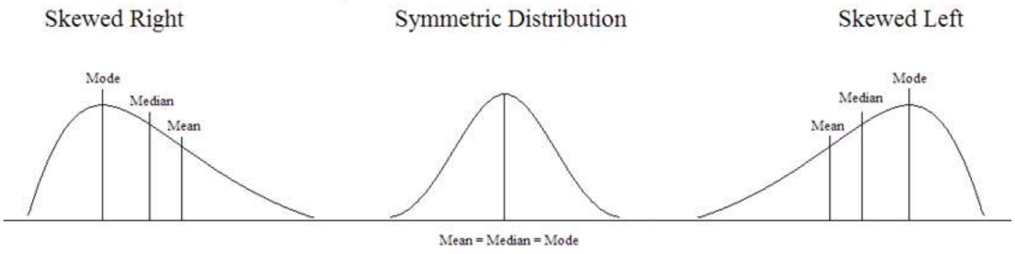


Figure 2. *Illustration of skewed and symmetric distributions.*

Measures of Dispersion

Measures of center look at the average or middle values of a data set. Measures of dispersion look at the spread or variation of the data. Variation refers to the amount that the values vary among themselves. Values in a data set that are relatively close to each other have lower measures of variation. Values that are spread farther apart have higher measures of variation.

Examine the two histograms below. Both groups have the same mean weight, but the values of Group A are more spread out compared to the values in Group B. Both groups have an average weight of 267 lb. but the weights of Group A are more variable.

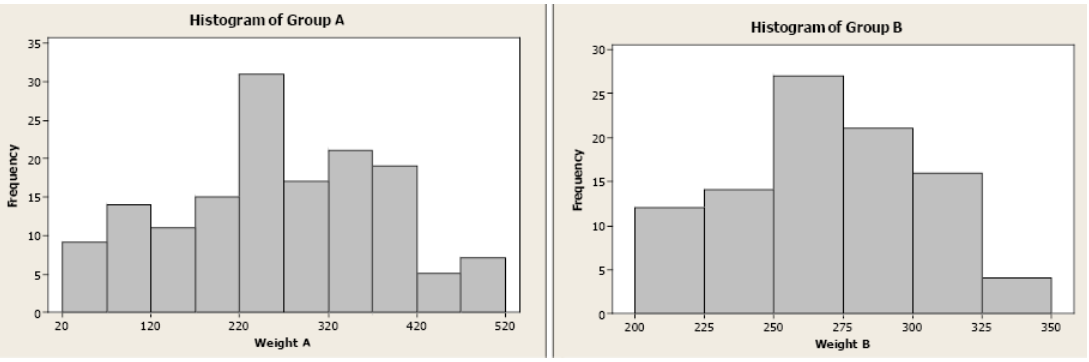


Figure 3. Histograms of Group A and Group B.

This section will examine five measures of dispersion: range, variance, standard deviation, standard error, and coefficient of variation.

Range

The range of a variable is the largest value minus the smallest value. It is the simplest measure and uses only these two values in a quantitative data set.

✓ Example 1.1.5: Computing Range

Find the range for the given data set.

$$12, 29, 32, 34, 38, 49, 57$$

$$Range = 57 - 12 = 45$$

Variance

The variance uses the difference between each value and its arithmetic mean. The differences are squared to deal with positive and negative differences. The sample variance (s^2) is an unbiased estimator of the population variance (σ^2), with n-1 degrees of freedom.

Degrees of freedom: In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question.

The sample variance is unbiased due to the difference in the denominator. If we used “n” in the denominator instead of “n – 1”, we would consistently underestimate the true population variance. To correct this bias, the denominator is modified to “n – 1”.

Definition: population variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \tag{1.1.3}$$

Definition: sample variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} \tag{1.1.4}$$

Example 1.1.6: Computing Variance

Compute the variance of the sample data: 3, 5, 7.

Solution

The sample mean ($\bar{x}$) is 5. Then use Equation 1.1.4

$$s^2 = \frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation

The standard deviation is the square root of the variance (both population and sample). While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared. The standard deviation is a common method for numerically describing the distribution of a variable. The population standard deviation is σ (sigma) and sample standard deviation is *s*.

Definition:SAMPLE STANDARD DEVIATION

$$s = \sqrt{s^2} \tag{1.1.5}$$

Definition:POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\sigma^2} \tag{1.1.6}$$

Example 1.1.7:

Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

Solution

The sample mean ($\bar{x}$) is 5, using the definition of standard deviation

$$s = \sqrt{\frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1}} = \sqrt{4} = 2$$

Standard Error of Mean

Commonly, we use the sample mean $\bar{x}$ to estimate the population mean μ. For example, if we want to estimate the heights of eighty-year-old cherry trees, we can proceed as follows:

- Randomly select 100 trees
- Compute the sample mean of the 100 heights
- Use that as our estimate

We want to use this sample mean to estimate the true but unknown population mean. But our sample of 100 trees is just one of many possible samples (of the same size) that could have been randomly selected. Imagine if we take a series of different random samples from the same population and all the same size:

- Sample 1—we compute sample mean $\bar{x}$
- Sample 2—we compute sample mean $\bar{x}$
- Sample 3—we compute sample mean $\bar{x}$
- Etc.

Each time we sample, we may get a different result as we are using a different subset of data to compute the sample mean. This shows us that the sample mean is a random variable!

The sample mean ($\bar{x}$) is a random variable with its own probability distribution called the sampling distribution of the sample mean. The distribution of the sample mean will have a mean equal to µ and a standard deviation equal to $\frac{s}{\sqrt{n}}$

Note

The standard error $\frac{s}{\sqrt{n}}$ is the standard deviation of all possible sample means.

In reality, we would only take one sample, but we need to understand and quantify the sample to sample variability that occurs in the sampling process.

The standard error is the standard deviation of the sample means and can be expressed in different ways.

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}} \tag{1.1.7}$$

Note

\(s^2\) is the sample variance and *s* is the sample standard deviation

Example 1.1.8:

Describe the distribution of the sample mean.

A population of fish has weights that are normally distributed with µ = 8 lb. and s = 2.6 lb. If you take a sample of size n=6, the sample mean will have a normal distribution with a mean of 8 and a standard deviation (standard error) of $\frac{2.6}{\sqrt{6}}$= 1.061 lb.

If you increase the sample size to 10, the sample mean will be normally distributed with a mean of 8 lb. and a standard deviation (standard error) of $\frac{2.6}{\sqrt{10}}$ = 0.822 lb.

Notice how the standard error decreases as the sample size increases.

The **Central Limit Theorem** (CLT) states that the sampling distribution of the sample means will approach a normal distribution as the sample size increases. If we do not have a normal distribution, or know nothing about our distribution of our random variable, the CLT tells us that the distribution of the $\bar{X}$'s will become normal as *n*increases. How large does *n* have to be? A general rule of thumb tells us that $n \geq 30$.

Note

The Central Limit Theorem tells us that regardless of the shape of our population, the sampling distribution of the sample mean will be normal as the sample size increases.

Coefficient of Variation

To compare standard deviations between different populations or samples is difficult because the standard deviation depends on units of measure. The coefficient of variation expresses the standard deviation as a percentage of the sample or population mean. It is a unitless measure.

Definition: CV of Population

$$CV = \frac{\sigma}{\mu} \times 100 \tag{1.1.8}$$

Definition: cv of sample

$$CV = \frac{s}{\bar{x}} \times 100 \tag{1.1.9}$$

Example 1.1.9:

Fisheries biologists were studying the length and weight of Pacific salmon. They took a random sample and computed the mean and standard deviation for length and weight (given below). While the standard deviations are similar, the differences in units between lengths and weights make it difficult to compare the variability. Computing the coefficient of variation for each variable allows the biologists to determine which variable has the greater standard deviation.

	Sample mean	Sample standard deviation
Length	63 cm	19.97 cm
Weight	37.6 kg	19.39 kg
	685.png	678.png

There is greater variability in Pacific salmon weight compared to length.

Variability

Variability is described in many different ways. Standard deviation measures point to point variability **within a sample**, i.e., variation among individual sampling units. Coefficient of variation also measures point to point variability but on a relative basis (relative to the mean), and is not influenced by measurement units. Standard error measures the **sample to sample variability**, i.e. variation among repeated samples in the sampling process. Typically, we only have one sample and standard error allows us to quantify the uncertainty in our sampling process.

Basic Statistics Example using Excel and Minitab Software

Consider the following tally from 11 sample plots on Heiburg Forest, where Xi is the number of downed logs per acre. Compute basic statistics for the sample plots.

ID					Order
1	25	625	-7.27	52.8529	4
2	35	1225	2.73	7.4529	6
4	55	3025	22.73	516.6529	10
5	15	225	-17.25	298.2529	2
6	40	1600	7.73	59.7529	8
7	25	625	-7.27	52.8529	5
8	55	3025	22.73	516.6529	11

ID					Order
9	35	1225	2.73	7.4529	7
10	45	2025	12.73	162.0529	9
11	5	25	-27.27	743.6529	1
Sum	20	400	-12.27	150.1819	3
	$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} X_i^2$	$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)$	$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$	

Table 1. Sample data on number of downed logs per acre from Heiburg Forest.

(1) Sample mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{355}{11}32.27 \tag{1.1.10}$$

(2) Median $= 35$

(3) Variance:

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} = \frac{2568.1519}{11-1} = 256.82 \\ &= \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1} = \frac{14025 - \frac{(355)^2}{11}}{11-1} = 256.82 \end{aligned}$$

(4) Standard deviation: $S = \sqrt{S^2} = \sqrt{256.82} = 16.0256$

(5) Range: $55 - 5 = 50$

(6) Coefficient of variation:

$$CV = \frac{S}{\bar{X}} \cdot 100 = \frac{16.0256}{32.27} \cdot 100 = 49.66\% \tag{1.1.11}$$

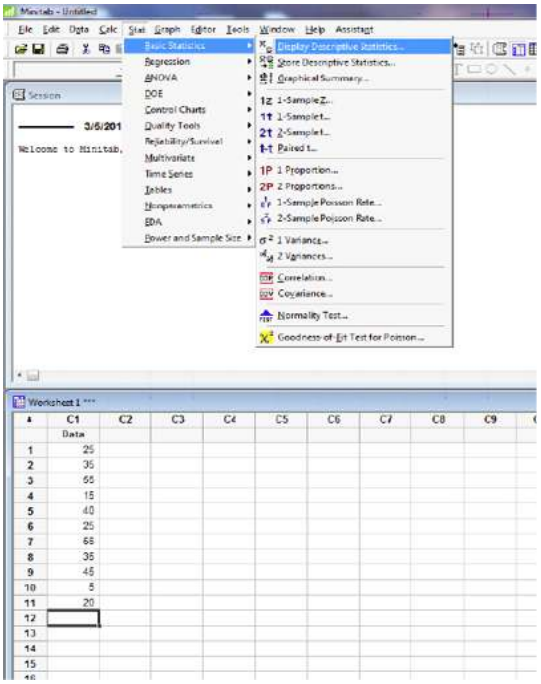
(7) Standard error of the mean:

$$\begin{aligned} S_{\bar{X}} &= \sqrt{\frac{S^2}{n}} = \sqrt{\frac{256.82}{11}} = 4.8319 \\ &= \frac{S}{\sqrt{n}} = \frac{16.0256}{\sqrt{11}} = 4.8319 \end{aligned}$$

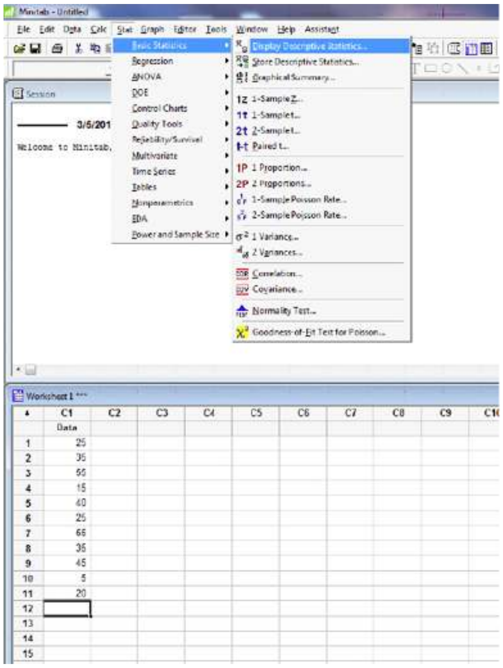
Software Solutions

Minitab

Open Minitab and enter data in the spreadsheet. Select STAT>Descriptive stats and check all statistics required.



↓	C1	C2	C3	C4	C5	C6	C7	C8	C9
	Data								
1	25								
2	35								
3	55								
4	15								
5	40								
6	25								
7	55								
8	35								
9	45								
10	5								
11	20								
12									
13									
14									
15									



↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
	Data									
1	25									
2	35									
3	55									
4	15									
5	40									
6	25									
7	55									
8	35									
9	45									
10	5									
11	20									
12										
13										
14										
15										

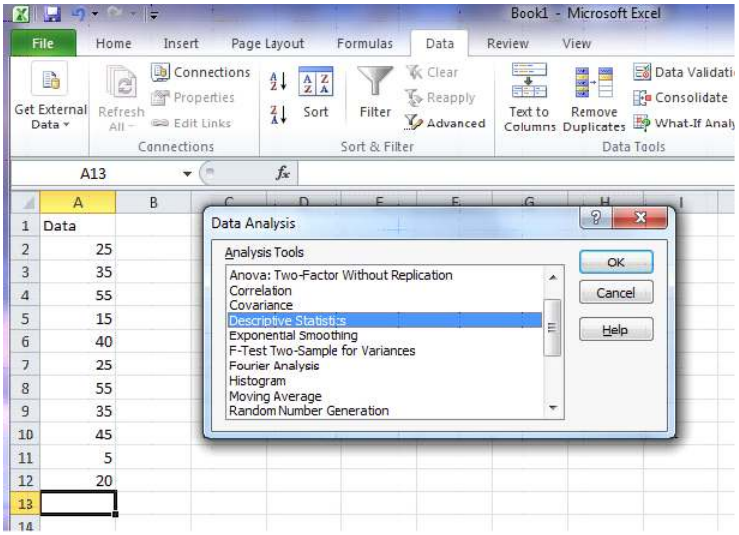
Descriptive Statistics: Data

Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q1
Data	11	0	32.27	4.83	16.03	256.82	49.66	5.00	20.00

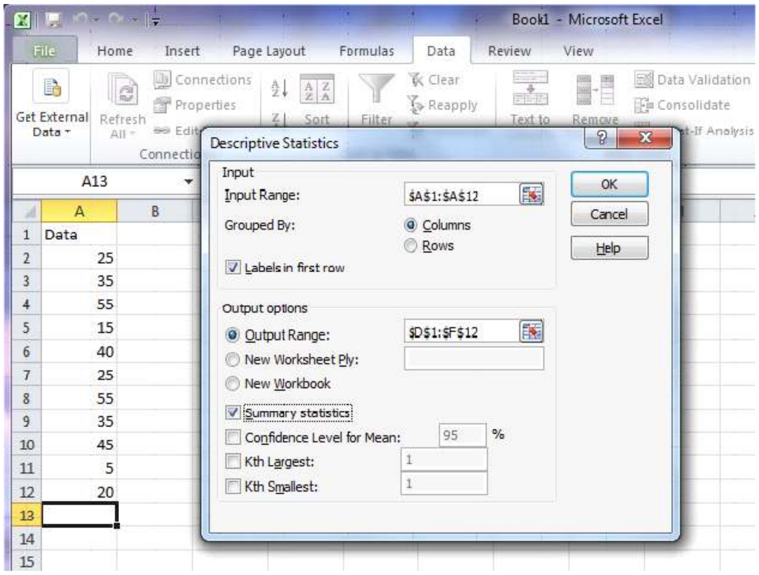
Variable	Median	Q3	Maximum	IQR
Data	35.00	45.00	55.00	25.00

Excel

Open up Excel and enter the data in the first column of the spreadsheet. Select DATA>Data Analysis>Descriptive Statistics. For the Input Range, select data in column A. Check “Labels in First Row” and “Summary Statistics”. Also check “Output Range” and select location for output.



	A
1	Data
2	25
3	35
4	55
5	15
6	40
7	25
8	55
9	35
10	45
11	5
12	20



Input

Input Range: $A$1:$A$12

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $D$1:$F$12

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

☐ Confidence Level for Mean: 95 %

☐ Kth Largest: 1

☐ Kth Smallest: 1

Data	
Mean	32.27273
Standard Error	4.831884
Median	35
Mode	25
Standard Deviation	16.02555
Sample Variance	256.8182
Kurtosis	-0.73643
Skewness	-0.05982

Data	
Range	50
Minimum	5
Maximum	55
Sum	355
Count	11

Graphic Representation

Data organization and summarization can be done graphically, as well as numerically. Tables and graphs allow for a quick overview of the information collected and support the presentation of the data used in the project. While there are a multitude of available graphics, this chapter will focus on a specific few commonly used tools.

Pie Charts

Pie charts are a good visual tool allowing the reader to quickly see the relationship between categories. It is important to clearly label each category, and adding the frequency or relative frequency is often helpful. However, too many categories can be confusing. Be careful of putting too much information in a pie chart. The first pie chart gives a clear idea of the representation of fish types relative to the whole sample. The second pie chart is more difficult to interpret, with too many categories. It is important to select the best graphic when presenting the information to the reader.

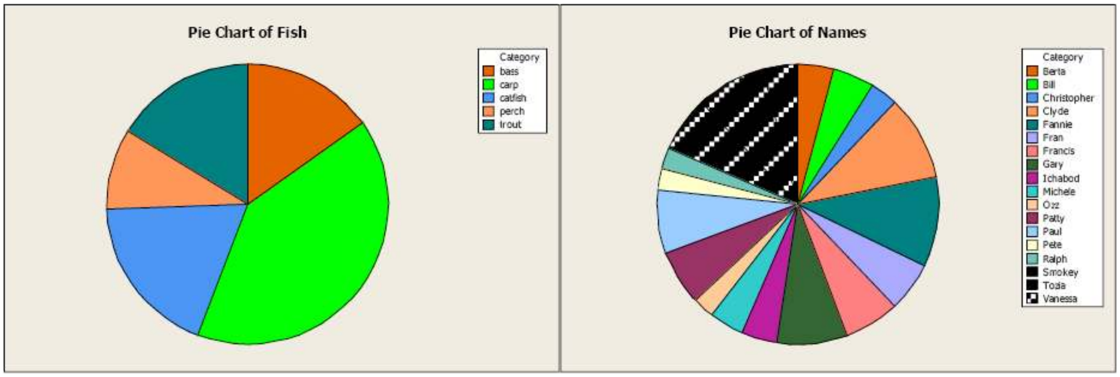


Figure 1.1.4:
Comparison of pie charts.
(Copyright; author via source)

Bar Charts and Histograms

Bar charts graphically describe the distribution of a qualitative variable (fish type) while histograms describe the distribution of a quantitative variable discrete or continuous variables (bear weight).

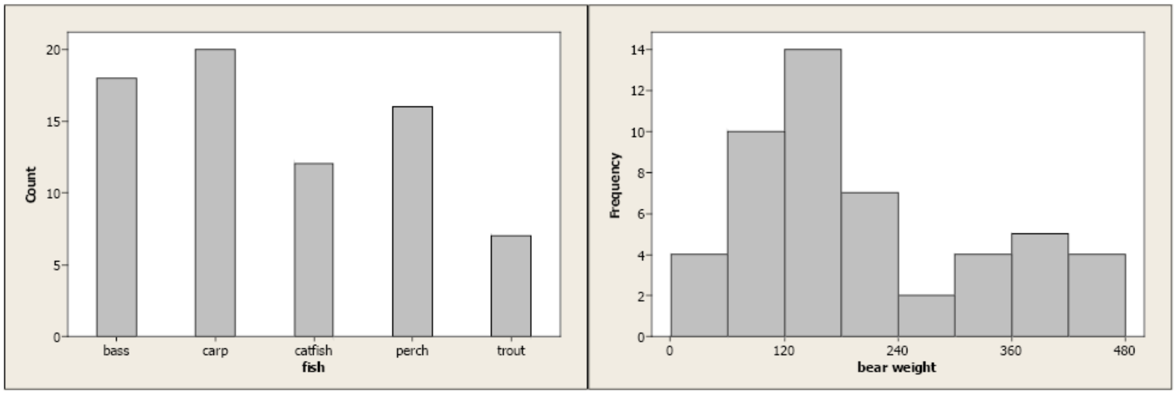


Figure 1.1.5:
Comparison of a bar chart for qualitative data and a histogram for quantitative data.
(Copyright; author via source)

In both cases, the bars' equal width and the y-axis are clearly defined. With qualitative data, each category is represented by a specific bar. With continuous data, lower and upper class limits must be defined with equal class widths. There should be no gaps between classes and each observation should fall into one, and only one, class.

Boxplots

Boxplots use the 5-number summary (minimum and maximum values with the three quartiles) to illustrate the center, spread, and distribution of your data. When paired with histograms, they give an excellent description, both numerically and graphically, of the data.

With symmetric data, the distribution is bell-shaped and somewhat symmetric. In the boxplot, we see that Q1 and Q3 are approximately equidistant from the median, as are the minimum and maximum values. Also, both whiskers (lines extending from the boxes) are approximately equal in length.

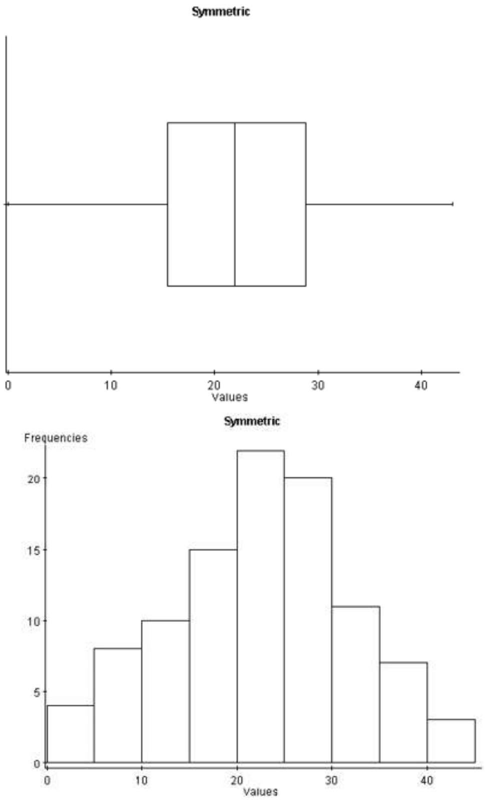


Figure 1.1.6: A histogram and boxplot of a normal distribution. (Copyright; author via source)

With skewed left distributions, we see that the histogram looks “pulled” to the left. In the boxplot, Q1 is farther away from the median as are the minimum values, and the left whisker is longer than the right whisker.

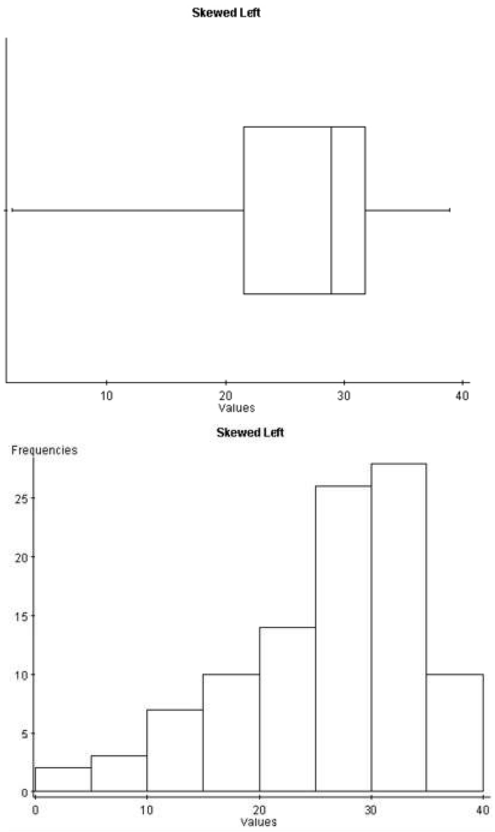


Figure 1.1.7:*A histogram and boxplot of a skewed left distribution.* (Copyright; author via source)

With skewed right distributions, we see that the histogram looks “pulled” to the right. In the boxplot, Q3 is farther away from the median, as is the maximum value, and the right whisker is longer than the left whisker.

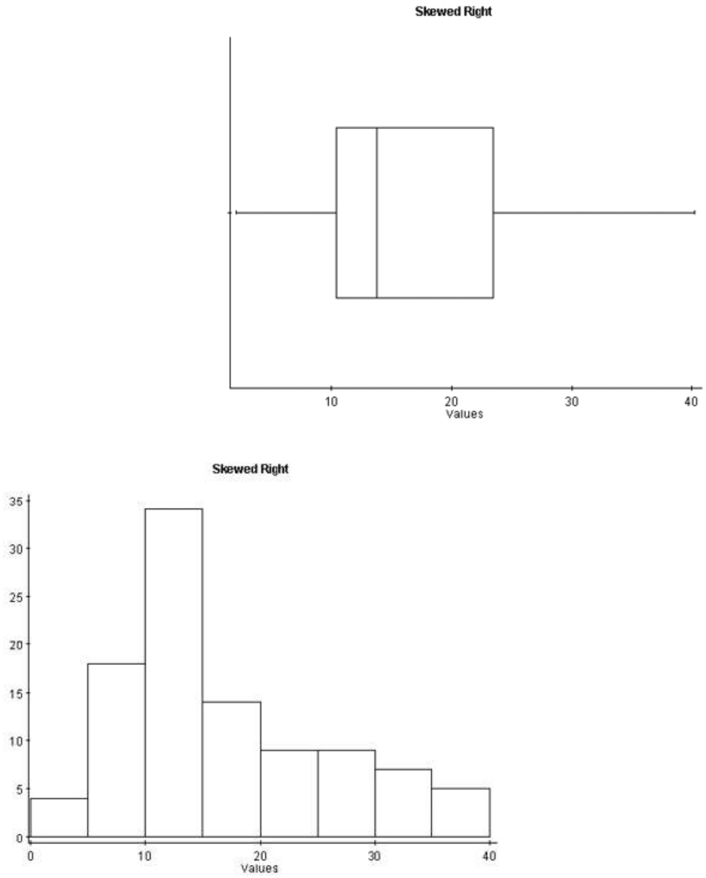


Figure 1.1.8: A *histogram and boxplot of a skewed right distribution.*(Copyright; author via source)

1.2: Probability Distribution

Once we have organized and summarized your sample data, the next step is to identify the underlying distribution of our random variable. Computing probabilities for continuous random variables are complicated by the fact that there are an infinite number of possible values that our random variable can take on, so the probability of observing a particular value for a random variable is zero. Therefore, to find the probabilities associated with a continuous random variable, we use a probability density function (PDF).

A PDF is an equation used to find probabilities for continuous random variables. The PDF must satisfy the following two rules:

1. The area under the curve must equal one (over all possible values of the random variable).
2. The probabilities must be equal to or greater than zero for all possible values of the random variable.

> The area under the curve of the probability density function over some interval represents the probability of observing those values of the random variable in that interval.

The Normal Distribution

Many continuous random variables have a bell-shaped or somewhat symmetric distribution. This is a normal distribution. In other words, the probability distribution of its relative frequency histogram follows a normal curve. The curve is bell-shaped, symmetric about the mean, and defined by µ and σ (the mean and standard deviation).

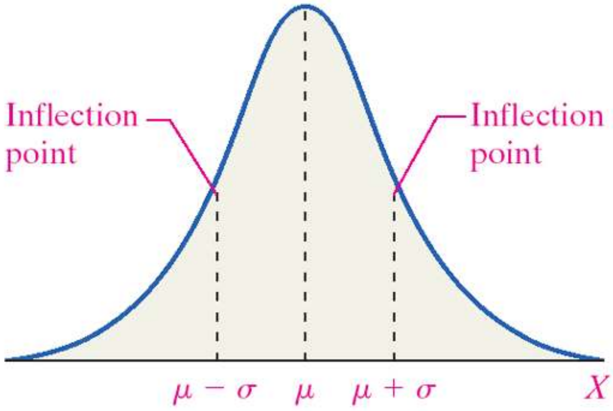


Figure 1.2.1: A nromal distribution. (Copyright; author via source)

There are normal curves for every combination of µ and σ. The mean (µ) shifts the curve to the left or right. The standard deviation (σ) alters the spread of the curve. The first pair of curves have different means but the same standard deviation. The second pair of curves share the same mean (µ) but have different standard deviations. The pink curve has a smaller standard deviation. It is narrower and taller, and the probability is spread over a smaller range of values. The blue curve has a larger standard deviation. The curve is flatter and the tails are thicker. The probability is spread over a larger range of values.

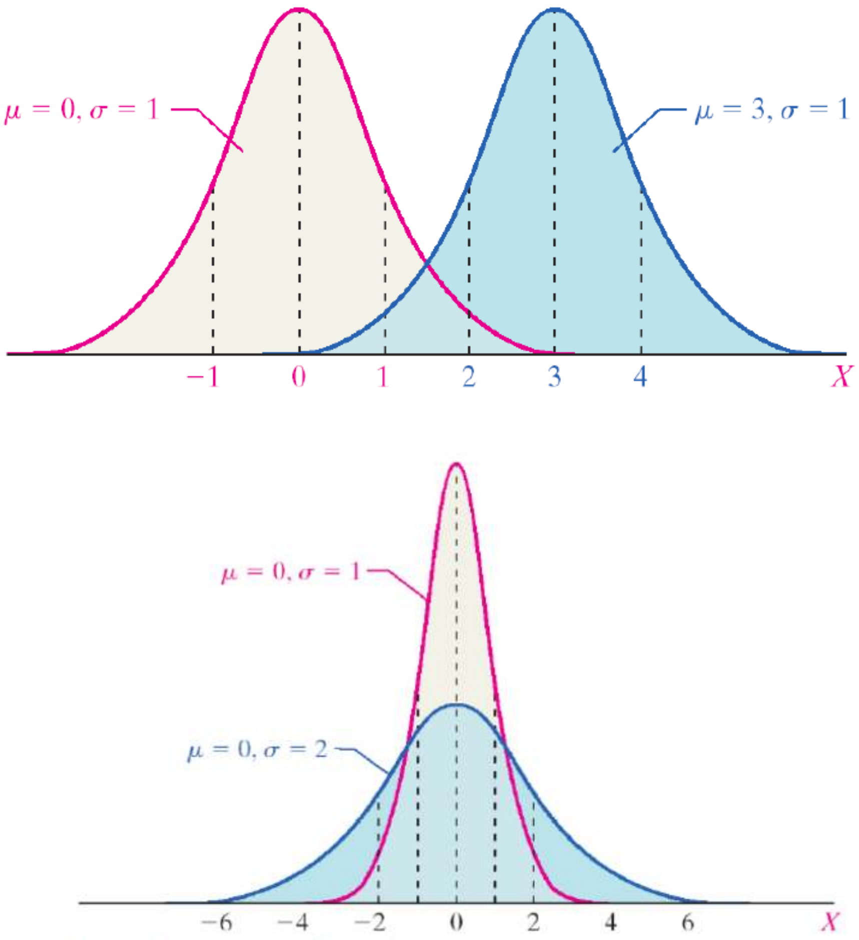


Figure 1.2.2: A comparison of normal curves.

Properties of the normal curve:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to one.
- As *x* increases and decreases, the curve goes to zero but never touches.
- The PDF of a normal curve is

$$y=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.2.1}$$

- A normal curve can be used to estimate probabilities.
- A normal curve can be used to estimate proportions of a population that have certain x-values.

The Standard Normal Distribution

There are millions of possible combinations of means and standard deviations for continuous random variables. Finding probabilities associated with these variables would require us to integrate the PDF over the range of values we are interested in. To avoid this, we can rely on the standard normal distribution. The standard normal distribution is a special normal distribution with a μ = 0 and **σ** = 1. We can use the Z-score to standardize any normal random variable, converting the x-values to Z-scores, thus allowing us to use probabilities from the standard normal table. So how do we find area under the curve associated with a Z-score?

Standard Normal Table

- The standard normal table gives probabilities associated with specific Z-scores.
- The table we use is cumulative from the left.
- The negative side is for all Z-scores less than zero (all values less than the mean).

- The positive side is for all Z-scores greater than zero (all values greater than the mean).
- Not all standard normal tables work the same way.

Example 1.2.1:

What is the area associated with the Z-score 1.62?

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

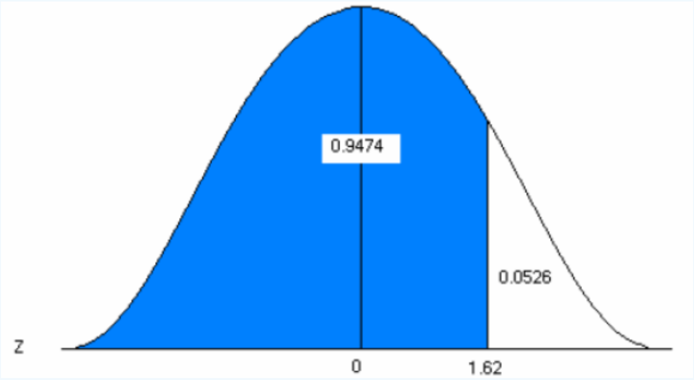


Figure 1.2.3: The standard normal table and associated area for z = 1.62.

Answer

The area is 0.9474.

Reading the Standard Normal Table

- Read down the Z-column to get the first part of the Z-score (1.6).
- Read across the top row to get the second decimal place in the Z-score (0.02).
- The intersection of this row and column gives the area under the curve to the left of the Z-score.

Finding Z-scores for a Given Area

- What if we have an area and we want to find the Z-score associated with that area?
- Instead of Z-score → area, we want area → Z-score.
- We can use the standard normal table to find the area in the body of values and read backwards to find the associated Z-score.
- Using the table, search the probabilities to find an area that is closest to the probability you are interested in.

Example 1.2.2:

To find a Z-score for which the area to the right is 5%:

Since the table is cumulative from the left, you must use the complement of 5%.

$$1.000 - 0.05 = 0.9500$$

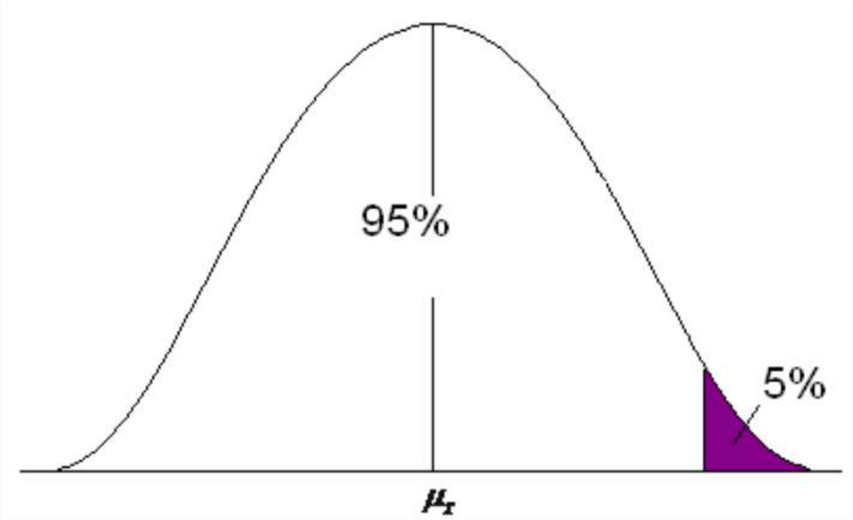


Figure 1.2.4: The upper 5% of the area under a normal curve.

- Find the Z-score for the area of 0.9500.
- Look at the probabilities and find a value as close to 0.9500 as possible.

The standard normal table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Answer

The Z-score for the 95th percentile is 1.64.

Area in between Two Z-scores

Example 1.2.3

To find Z-scores that limit the middle 95%:

- The middle 95% has 2.5% on the right and 2.5% on the left.
- Use the symmetry of the curve.

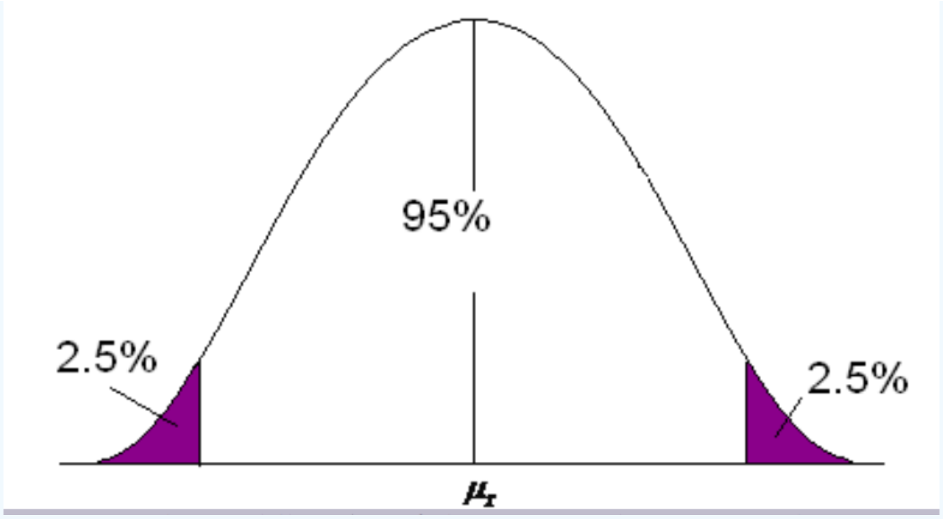


Figure 1.2.5: The middle 95% of the area under a normal curve.

Solution

- Look at your standard normal table. Since the table is cumulative from the left, it is easier to find the area to the left first.
- Find the area of 0.025 on the negative side of the table.
- The Z-score for the area to the left is -1.96.
- Since the curve is symmetric, the Z-score for the area to the right is 1.96.

Common Z-scores

There are many commonly used Z-scores:

- $Z_{.05}$ = 1.645 and the area between -1.645 and 1.645 is 90%
- $Z_{.025}$ = 1.96 and the area between -1.96 and 1.96 is 95%
- $Z_{.005}$ = 2.575 and the area between -2.575 and 2.575 is 99%

Applications of the Normal Distribution

Typically, our normally distributed data do not have μ = 0 and **σ** = 1, but we can relate any normal distribution to the standard normal distributions using the Z-score. We can transform values of x to values of z.

$$z=\frac{x-\mu}{\sigma} \tag{1.2.2}$$

For example, if a normally distributed random variable has a μ = 6 and **σ** = 2, then a value of x = 7 corresponds to a Z-score of 0.5.

$$Z=\frac{7-6}{2}=0.5 \tag{1.2.3}$$

This tells you that 7 is one-half a standard deviation above its mean. We can use this relationship to find probabilities for any normal random variable.

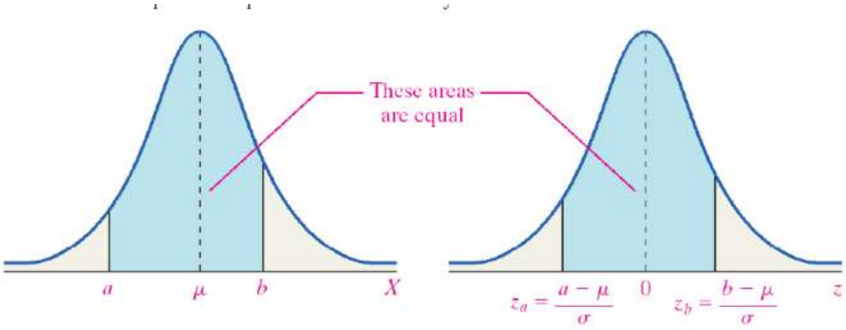


Figure 1.2.6: A normal and standard normal curve.

To find the area for values of X, a normal random variable, draw a picture of the area of interest, convert the x-values to Z-scores using the Z-score and then use the standard normal table to find areas to the left, to the right, or in between.

$$z=\frac{x-\mu}{\sigma} \tag{1.2.4}$$

Example 1.2.4:

Adult deer population weights are normally distributed with µ = 110 lb. and **σ** = 29.7 lb. As a biologist you determine that a weight less than 82 lb. is unhealthy and you want to know what proportion of your population is unhealthy.

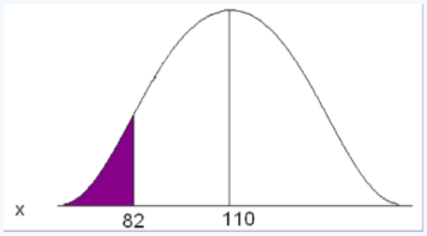


Figure 1.2.7: The area under a normal curve for P(x<82).

Convert 82 to a Z-score

$$z=\frac{82-110}{29.7}=-0.94$$

The *x* value of 82 is 0.94 standard deviations below the mean.

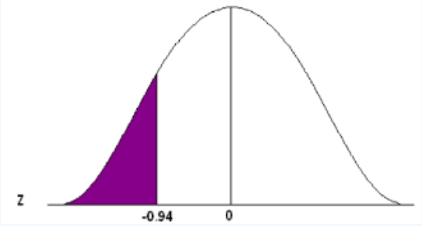


Figure 1.2.8: Area under a standard normal curve for P(z<-0.94).

Go to the standard normal table (negative side) and find the area associated with a Z-score of -0.94.

This is an “area to the left” problem so you can read directly from the table to get the probability.

$$P(x<82)=0.1736$$

Approximately 17.36% of the population of adult deer is underweight, OR one deer chosen at random will have a 17.36% chance of weighing less than 82 lb.

Example 1.2.5:

Statistics from the Midwest Regional Climate Center indicate that Jones City, which has a large wildlife refuge, gets an average of 36.7 in. of rain each year with a standard deviation of 5.1 in. The amount of rain is normally distributed. During what percent of the years does Jones City get more than 40 in. of rain?

$$P(x>40)$$

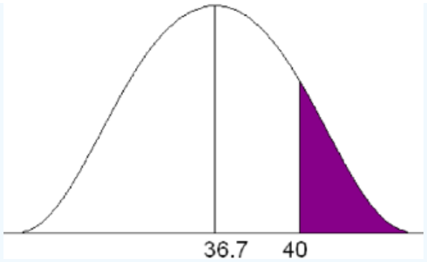


Figure 1.2.9: Area under a normal curve for P(x>40).

Solution

$$z = \frac{40 - 36.7}{5.1} = 0.65$$

$$P(x > 40) = (1 - 0.7422) = 0.2578$$

For approximately 25.78% of the years, Jones City will get more than 40 in. of rain.

Assessing Normality

If the distribution is unknown and the sample size is not greater than 30 (Central Limit Theorem), we have to assess the assumption of normality. Our primary method is the normal probability plot. This plot graphs the observed data, ranked in ascending order, against the “expected” Z-score of that rank. If the sample data were taken from a normally distributed random variable, then the plot would be approximately linear.

Examine the following probability plot. The center line is the relationship we would expect to see if the data were drawn from a perfectly normal distribution. Notice how the observed data (red dots) loosely follow this linear relationship. Minitab also computes an Anderson-Darling test to assess normality. The null hypothesis for this test is that the sample data have been drawn from a normally distributed population. A p-value greater than 0.05 supports the assumption of normality.

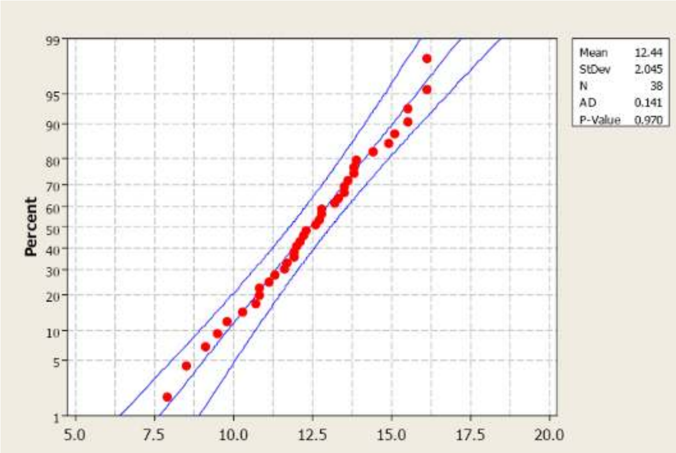


Figure 1.2.10: A normal probability plot generated using Minitab 16.

Compare the histogram and the normal probability plot in this next example. The histogram indicates a skewed right distribution.

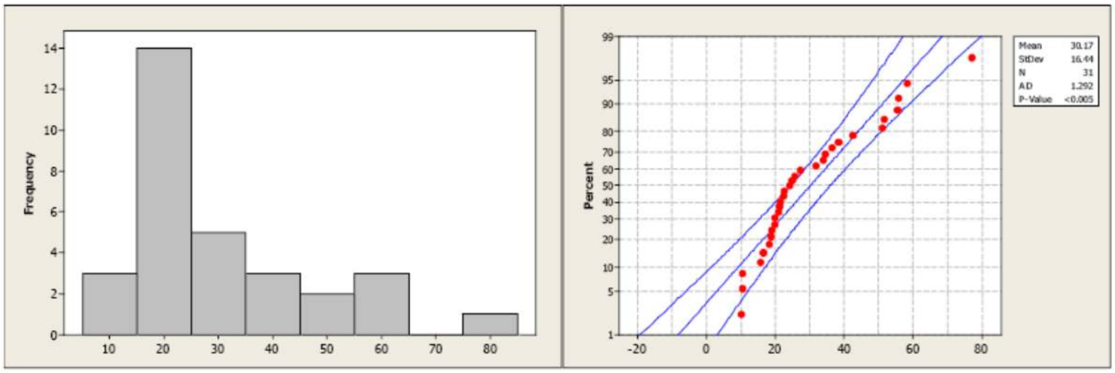


Figure 1.2.11: Histogram and normal probability plot for skewed right data.

The observed data do not follow a linear pattern and the p-value for the A-D test is less than 0.005 indicating a non-normal population distribution.

Normality cannot be assumed. You must always verify this assumption. Remember, the probabilities we are finding come from the standard NORMAL table. If our data are NOT normally distributed, then these probabilities DO NOT APPLY.

- Do you know if the population is normally distributed?
- Do you have a large enough sample size (n≥30)? Remember the Central Limit Theorem?
- Did you construct a normal probability plot?

CHAPTER OVERVIEW

2: Sampling Distributions and Confidence Intervals

2.1: Sampling Distribution of the Sample Mean

Inferential testing uses the sample mean ($\bar{x}$) to estimate the population mean (μ). Typically, we use the data from a single sample, but there are many possible samples of the same size that could be drawn from that population. As we saw in the previous chapter, the sample mean ($\bar{x}$) is a *random* variable with its own distribution.

- The distribution of the sample mean will have a mean equal to µ.
- It will have a standard deviation (standard error) equal to $\frac{\sigma}{\sqrt{n}}$

Because our inferences about the population mean rely on the sample mean, we focus on the distribution of the sample mean. Is it normal? What if our population is not normally distributed or we don't know anything about the distribution of our population?

The Central Limit Theorem (CLT)

The Central Limit Theorem states that the sampling distribution of the sample means will approach a normal distribution as the sample size increases.

So if we do not have a normal distribution, or know nothing about our distribution, the CLT tells us that the distribution of the sample means ($\bar{X}$) will become normal distributed as *n* (sample size) increases. How large does n have to be? A general rule of thumb tells us that n ≥ 30.

The Central Limit Theorem tells us that regardless of the shape of our population, the sampling distribution of the sample mean will be normal as the sample size increases.

Sampling Distribution of the Sample Proportion

The population proportion (p) is a parameter that is as commonly estimated as the mean. It is just as important to understand the distribution of the sample proportion, as the mean. With proportions, the element either has the characteristic you are interested in or the element does not have the characteristic. The sample proportion ($\hat{p}$) is calculated by

$$\hat{p} = \frac{x}{n} \tag{2.1.1}$$

where x is the number of elements in your population with the characteristic and n is the sample size.

Example 2.1.1: sample proportion

You are studying the number of cavity trees in the Monongahela National Forest for wildlife habitat. You have a sample size of n = 950 trees and, of those trees, x = 238 trees with cavities. Calculate the sample proportion.

A naturally formed tree hollow at the base of the tree. (CC BY 2.0; Lauren "Lolly" Weinhold).

Solution

This is a simple application of Equation 2.1.1:

$$\hat{p} = \frac{238}{950} = 0.25$$

The distribution of the sample proportion has a mean of

$$\mu_{\hat{p}} = p \tag{2.1.2}$$

and has a standard deviation of

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}. \tag{2.1.3}$$

The sample proportion is normally distributed if n is very large and $\hat{p}$ is not close to 0 or 1. We can also use the following relationship to assess normality when the parameter being estimated is *p*, the population proportion:

$$n\hat{p}(1-\hat{p}) \geq 10 \tag{2.1.4}$$

2.2: Confidence Intervals

In the preceding chapter we learned that populations are characterized by descriptive measures called parameters. Inferences about parameters are based on sample statistics. We now want to estimate population parameters and assess the reliability of our estimates based on our knowledge of the sampling distributions of these statistics.

Point Estimates

We start with a point estimate. This is a single value computed from the sample data that is used to estimate the population parameter of interest.

- The sample mean ($\bar{x}$) is a point estimate of the population mean (μ).
- The sample proportion ($\hat{p}$) is the point estimate of the population proportion (p).

We use point estimates to construct confidence intervals for unknown parameters.

- A confidence interval is an interval of values instead of a single point estimate.
- The level of confidence corresponds to the expected proportion of intervals that will contain the parameter if many confidence intervals are constructed of the same sample size from the same population.
- Our uncertainty is about whether our particular confidence interval is one of those that truly contains the true value of the parameter.

✓ Example 2.2.1: bear weight

We are 95% confident that our interval contains the population mean bear weight.

If we created 100 confidence intervals of the same size from the same population, we would expect 95 of them to contain the true parameter (the population mean weight). We also expect five of the intervals would not contain the parameter.

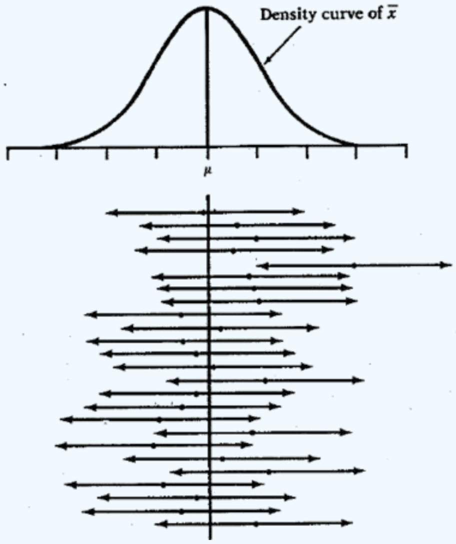


Figure* $2.2.1$*: *Confidence intervals from twenty-five different samples.*

In this example, twenty-five samples from the same population gave these 95% confidence intervals. In the long term, 95% of all samples give an interval that contains μ, the true (but unknown) population mean.

Level of confidence is expressed as a percent.

- The compliment to the level of confidence is α (alpha), the level of significance.
- The level of confidence is described as $(1-\alpha)\times 100$.

What does this really mean?

- We use a point estimate (e.g., sample mean) to estimate the population mean.

- We attach a level of confidence to this interval to describe how certain we are that this interval actually contains the unknown population parameter.
- We want to estimate the population parameter, such as the mean (μ) or proportion (p).

$$\bar{x} - E < \mu < \bar{x} + E \tag{2.2.1}$$

or

$$\hat{p} - E < p < \hat{p} + E \tag{2.2.2}$$

where E is the margin of error.

The confidence is based on area under a normal curve. So the assumption of normality must be met (Chapter 1).

Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Known

A confidence interval takes the form of: **point estimate $\pm$ margin of error.**

The point estimate

- The point estimate comes from the sample data.
- To estimate the population mean (μ), use the sample mean ($\bar{x}$) as the point estimate.

The margin of error

- Depends on the level of confidence, the sample size and the population standard deviation.
- It is computed as $E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ where $Z_{\frac{\alpha}{2}}$ is the critical value from the standard normal table associated with α (the level of significance).

The critical value $Z_{\frac{\alpha}{2}}$

- This is a Z-score that bounds the level of confidence.
- Confidence intervals are ALWAYS two-sided and the Z-scores are the limits of the area associated with the level of confidence.

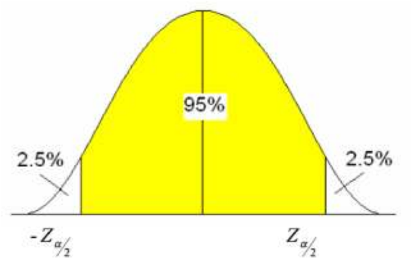


Figure 2.2.1: The middle 95% area under a standard normal curve.

- The level of significance (α) is divided into halves because we are looking at the middle 95% of the area under the curve.
- Go to your standard normal table and find the area of 0.025 in the body of values.
- What is the Z-score for that area?

Table 2.2.1: Common critical values (Z-scores).

Confidence Level	(level of significance)	$Z_{\alpha/2}$
99%	1%	2.575
95%	5%	1.96
90%	10%	1.645

Steps

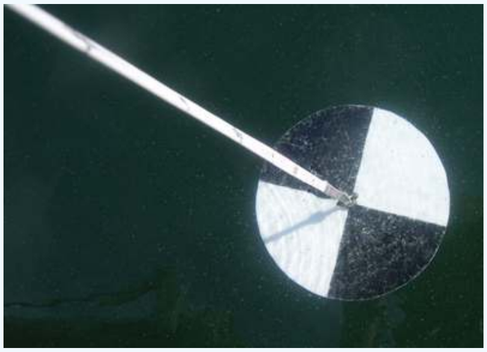
Construction of a confidence interval about μ when σ is known:

1. $Z_{\frac{\alpha}{2}}$ (critical value)

2. $E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ (margin of error)
3. $\bar{x} \pm E$ (point estimate ± margin of error)

✓ Example 2.2.3: Construct a confidence interval about the population mean

Researchers have been studying p-loading in Jones Lake for many years. It is known that mean water clarity (using a Secchi disk) is normally distributed with a population standard deviation of **σ** = 15.4 in. A random sample of 22 measurements was taken at various points on the lake with a sample mean of $\bar{x}$ = 57.8 in. The researchers want you to construct a 95% confidence interval for μ, the mean water clarity.



A secchi disk to measure turbidly of water. (CC SA; publiclab.org)

Solution

1) $Z_{\frac{\alpha}{2}}$ = 1.96

2) $E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ = $1.96 \times \frac{15.4}{\sqrt{22}}$ = 6.435

3) $\bar{x} \pm E$ = 57.8 ± 6.435

95% confidence interval for the mean water clarity is (51.36, 64.24).

We can be 95% confident that this interval contains the population mean water clarity for Jones Lake.

Now construct a 99% confidence interval for μ, the mean water clarity, and interpret.

1) $Z_{\frac{\alpha}{2}}$ = 2.575

2) $E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ = $2.575 \times \frac{15.4}{\sqrt{22}}$ = 8.454

3) $\bar{x} \pm E$ = 57.8± 8.454

99% confidence interval for the mean water clarity is (49.35, 66.25).

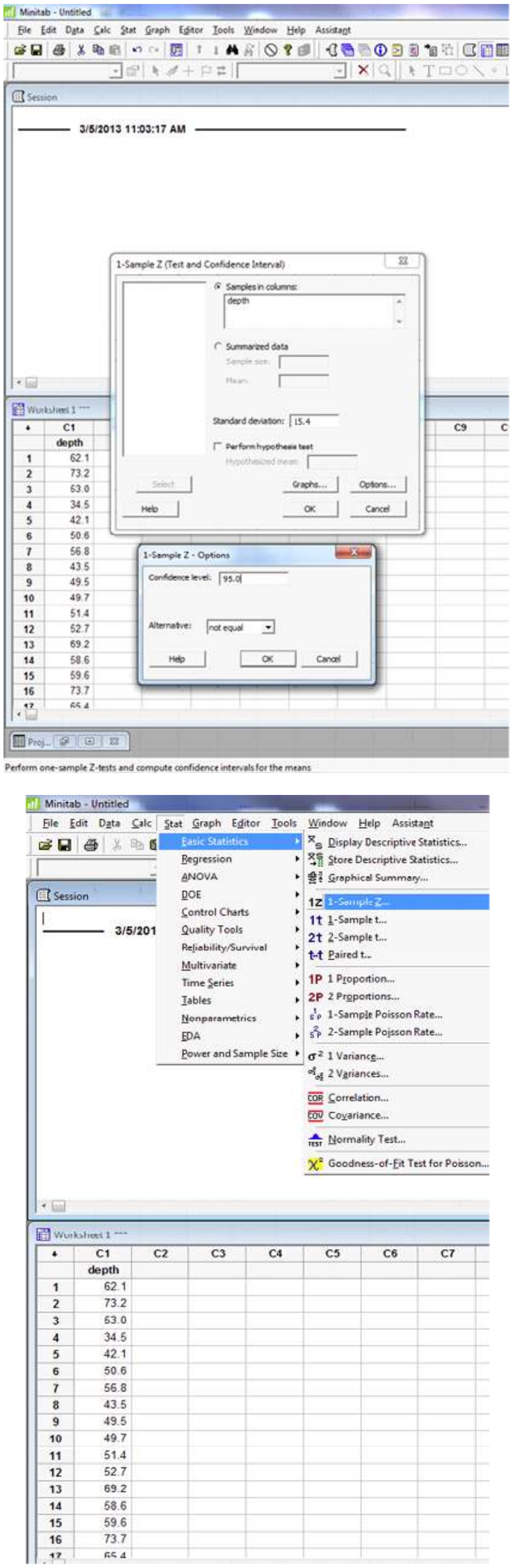
We can be 99% confident that this interval contains the population mean water clarity for Jones Lake.

As the level of confidence increased from 95% to 99%, the width of the interval increased. As the probability (area under the normal curve) increased, the critical value increased resulting in a wider interval.

Software Solutions

Minitab

You can use Minitab to construct this 95% confidence interval (Excel does not construct confidence intervals about the mean when the population standard deviation is known). Select Basic Statistics>1-sample Z. Enter the known population standard deviation and select the required level of confidence.



3/5/2013 11:03:17 AM

● Samples in columns: depth

○ Summarized data

Sample size:

Mean:

Standard deviation: 15.4

☐ Perform hypothesis test

Hypothesized mean:

Confidence level: 95.0

Alternative: not equal

Perform one-sample Z-tests and compute confidence intervals for the means

	C1
	depth
1	62.1
2	73.2
3	63.0
4	34.5
5	42.1
6	50.6
7	56.8
8	43.5
9	49.5
10	49.7
11	51.4
12	52.7
13	69.2
14	58.6
15	59.6
16	73.7

Minitab screen shots for constructing a confidence interval.

One-Sample Z: depth

The assumed standard deviation = 15.4

Variable	N	Mean	StDev	SE Mean	95% CI

depth	22	57.80	11.60	3.28	(51.36, 64.24)

Confidence Intervals about the Mean (μ) when the Population Standard Deviation (σ) is Unknown

Typically, in real life we often don't know the population standard deviation (σ). We can use the sample standard deviation (s) in place of σ. However, because of this change, we can't use the standard normal distribution to find the critical values necessary for constructing a confidence interval.

The Student's t-distribution was created for situations when σ was unknown. Gosset worked as a quality control engineer for Guinness Brewery in Dublin. He found errors in his testing and he knew it was due to the use of s instead of σ. He created this distribution to deal with the problem of an unknown population standard deviation and small sample sizes. A portion of the t-table is shown below.

Table 2.2.2: Portion of the student's t-table.

df	0.10	0.05	0.025	0.02	0.01	0.005
1	3.078	6.314	12.706	15.894	31.821	63.657
2	1.886	2.920	4.303	4.849	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.999	3.747	4.604
5	1.476	2.015	2.571	2.757	3.365	4.

Example 2.2.4

Find the critical value $t_{\frac{\alpha}{2}}$ for a 95% confidence interval with a sample size of n=13.

Solution

- Degrees of freedom (down the left-hand column) is equal to n-1 = 12
- α = 0.05 and α/2 = 0.025
- Go down the 0.025 column to 12 df
- $t_{\frac{\alpha}{2}} = 2.179$

The critical values from the students' t-distribution approach the critical values from the standard normal distribution as the sample size (n) increases.

Table 2.2.3: *Critical values from the student's t-table.*

n	Degrees of Freedom	$t_{0.25}$
11	10	2.228
51	50	2.009
101	100	1.984
1001	1000	1.962

Using the standard normal curve, the critical value for a 95% confidence interval is **1.96**. You can see how different samples sizes will change the critical value and thus the confidence interval, especially when the sample size is small.

Construction of a Confidence Interval

When **σ** is Unknown

1. $t_{\frac{\alpha}{2}}$ critical value with n-1 df

2. $E = t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$
3. $\bar{x} \pm E$

Example 2.2.5:

Researchers studying the effects of acid rain in the Adirondack Mountains collected water samples from 22 lakes. They measured the pH (acidity) of the water and want to construct a 99% confidence interval about the mean lake pH for this region. The sample mean is 6.4438 with a sample standard deviation of 0.7120. They do not know anything about the distribution of the pH of this population, and the sample is small (n<30), so they look at a normal probability plot.

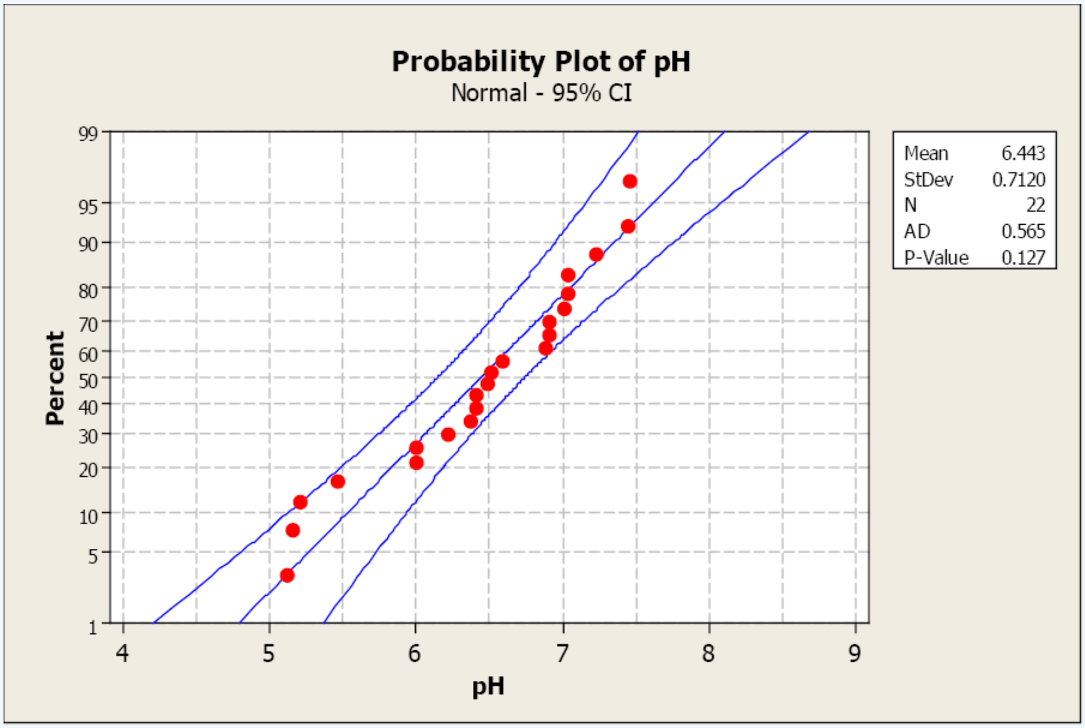


Figure 4. Normal probability plot.

Solution

The data is normally distributed. Now construct the 99% confidence interval about the mean pH.

1) $t_{\frac{\alpha}{2}}$ = 2.831

2) $E = t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$ = $2.831 \times \frac{0.7120}{\sqrt{22}}$ = 0.4297

3) $\bar{x} \pm E$ = 6.443 ± 0.4297

The 99% confidence interval about the mean pH is (6.013, 6.863).

We are 99% confident that this interval contains the mean lake pH for this lake population.

Now construct a 90% confidence interval about the mean pH for these lakes.

1) $t_{\frac{\alpha}{2}}$ = 1.721

2) $E = t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$ = $1.71221 \times \frac{0.7120}{\sqrt{22}}$0.2612

3) $\bar{x} \pm E$ = 6.443 ± 0.2612

The 90% confidence interval about the mean pH is (6.182, 6.704).

We are 90% confident that this interval contains the mean lake pH for this lake population.

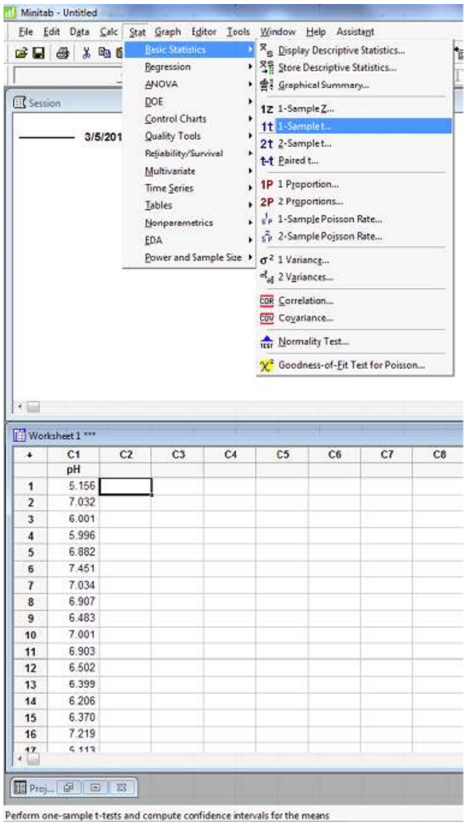
Notice how the width of the interval decreased as the level of confidence decreased from 99 to 90%.

> Construct a 90% confidence interval about the mean lake pH using Excel and Minitab.

Software Solutions

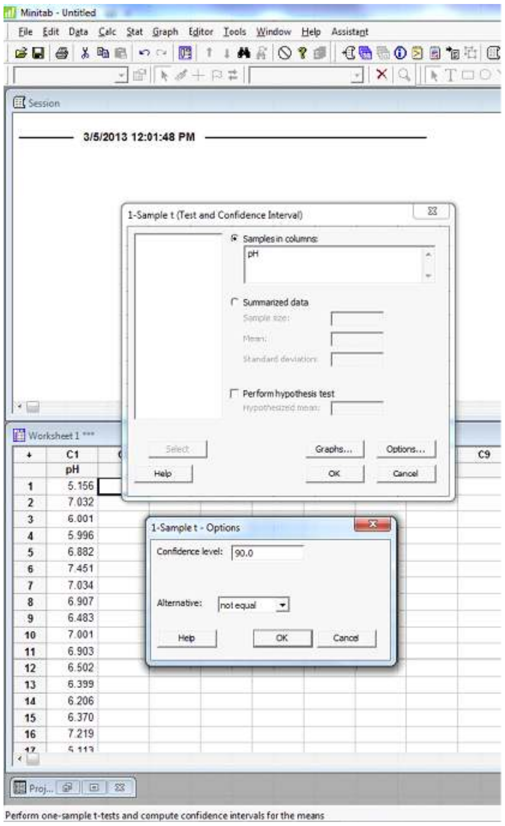
Minitab

For Minitab, enter the data in the spreadsheet and select Basic statistics and 1-sample t-test.



↓	C1	C2	C3	C4	C5	C6	C7	C8
	pH							
1	5.156							
2	7.032							
3	6.001							
4	5.996							
5	6.882							
6	7.451							
7	7.034							
8	6.907							
9	6.483							
10	7.001							
11	6.903							
12	6.502							
13	6.399							
14	6.206							
15	6.370							
16	7.219							
17	5.113							

Perform one-sample t-tests and compute confidence intervals for the means



3/5/2013 12:01:48 PM

1-Sample t (Test and Confidence Interval)

● Samples in columns: pH

○ Summarized data

Sample size:

Mean:

Standard deviation:

☐ Perform hypothesis test

Hypothesized mean:

1-Sample t - Options

Confidence level: 90.0

Alternative: not equal

	C1 pH
1	5.156
2	7.032
3	6.001
4	5.996
5	6.882
6	7.451
7	7.034
8	6.907
9	6.483
10	7.001
11	6.903
12	6.502
13	6.399
14	6.206
15	6.370
16	7.219
17	5.113

Perform one-sample t-tests and compute confidence intervals for the means

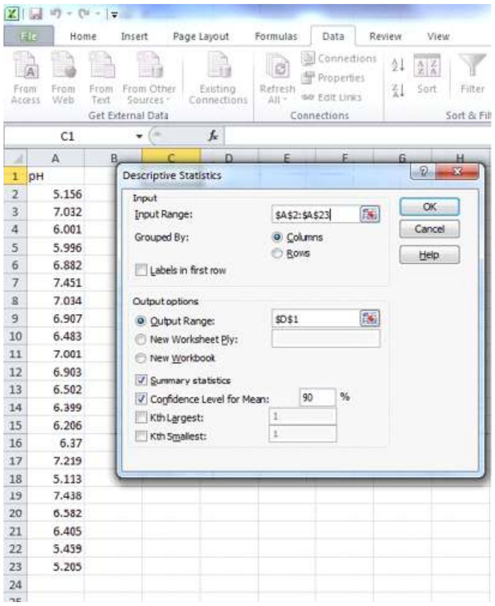
One-Sample T: pH

Variable	N	Mean	StDev	SE Mean	90% CI
pH	22	6.443	0.712	0.152	(6.182, 6.704)

Additional example:

Excel

For Excel, enter the data in the spreadsheet and select descriptive statistics. Check Summary Statistics and select the level and confidence.



	A
1	pH
2	5.156
3	7.032
4	6.001
5	5.996
6	6.882
7	7.451
8	7.034
9	6.907
10	6.483
11	7.001
12	6.903
13	6.502
14	6.399
15	6.206
16	6.37
17	7.219
18	5.113
19	7.438
20	6.582
21	6.405
22	5.459
23	5.205

Descriptive Statistics

Input

Input Range: $A$2:$A$23

Grouped By: ● Columns ○ Rows

☐ Labels in first row

Output options

● Output Range: $D$1

○ New Worksheet Ply:

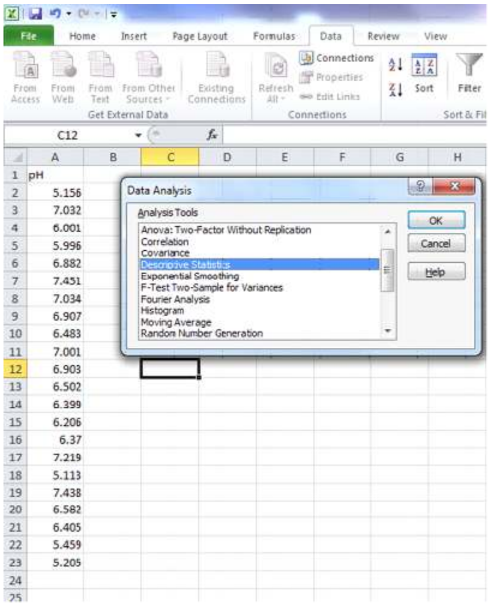
○ New Workbook

☑ Summary statistics

☑ Confidence Level for Mean: 90 %

☐ Kth Largest: 1

☐ Kth Smallest: 1



Data Analysis

Analysis Tools: Descriptive Statistics

Mean	6.442909
Standard Error	0.151801
Median	6.4925
Mode	#N/A
Standard Deviation	0.712008
Sample Variance	0.506956
Kurtosis	-0.5007
Skewness	-0.60591

Range	2.338
Minimum	5.113
Maximum	7.451
Sum	141.744
Count	22
Confidence Level(90.0%)	0.26121

Excel gives you the sample mean in the first line (6.442909) and the margin of error in the last line (0.26121). You must complete the computation yourself to obtain the interval (6.442909±0.26121).

Confidence Intervals about the Population Proportion (*p*)

Frequently, we are interested in estimating the population proportion (p), instead of the population mean (**μ**). For example, you may need to estimate the proportion of trees infected with beech bark disease, or the proportion of people who support "green" products. The parameter p can be estimated in the same ways as we estimated **μ**, the population mean.

The Sample Proportion

- The sample proportion is the best point estimate for the true population proportion.
- Sample proportion $\hat{p} = \frac{x}{n}$ where *x* is the number of elements in the sample with the characteristic you are interested in, and *n* is the sample size.

The Assumption of Normality when Estimating Proportions

- The assumption of a normally distributed population is still important, even though the parameter has changed.
- Normality can be verified if:

$$n \times \hat{p} \times (1-\hat{p}) \geq 10 \tag{2.2.3}$$

Constructing a Confidence Interval about the Population Proportion

Constructing a confidence interval about the proportion follows the same three steps we have used in previous examples.

1. $Z_{\frac{\alpha}{2}}$ (critical value from the standard normal table)
2. $E = Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (margin of error)
3. $\hat{p} \pm E$ (point estimate ± margin of error)

Example 2.2.6:

A botanist has produced a new variety of hybrid soybean that is better able to withstand drought. She wants to construct a 95% confidence interval about the germination rate (percent germination). She randomly selected 500 seeds and found that 421 have germinated.

Solution

First, compute the point estimate

$$\hat{p} = \frac{x}{n} = \frac{421}{500} = 0.842 \tag{2.2.4}$$

Check normality:

$$n \times \hat{p} \times (1-\hat{p}) \geq 10 = 500 \times 0.842 \times (1-0.842) = 66.5 \tag{2.2.5}$$

You can assume a normal distribution.

Now construct the confidence interval:

1) $Z_{\frac{\alpha}{2}}$ = 1.96

2) $E = Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $=1.96 \times \sqrt{\frac{0.842(1-0.842)}{500}}$ = 0.032

3) $\hat{p} \pm E = 0.842 \pm 0.0032$

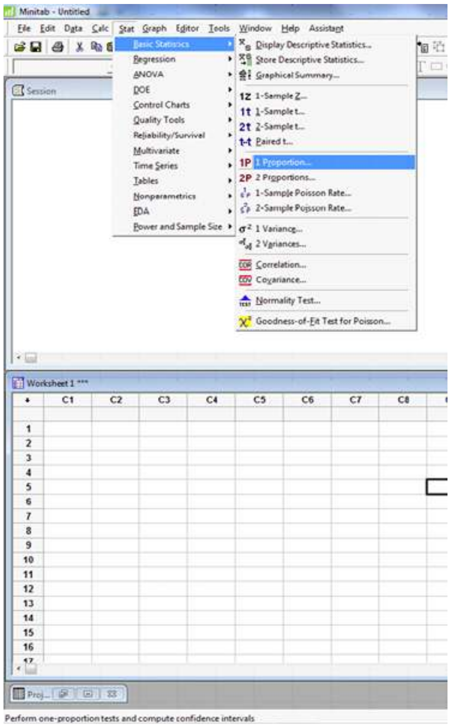
The 95% confidence interval for the germination rate is (81.0%, 87.4%).

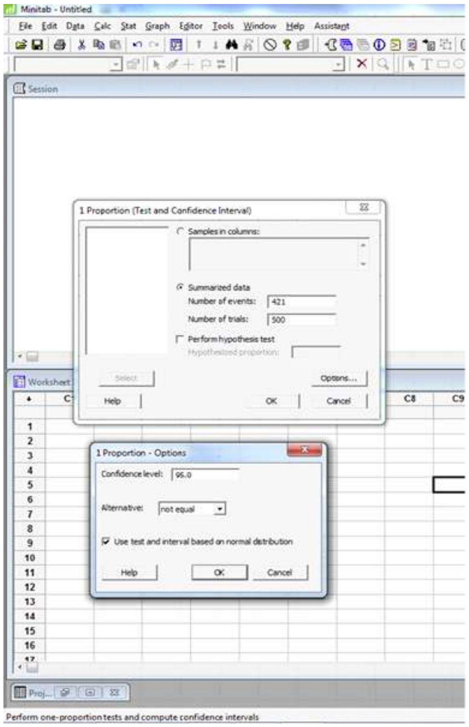
We can be 95% confident that this interval contains the true germination rate for this population.

Software Solutions

Minitab

You can use Minitab to compute the confidence interval. Select STAT>Basic stats>1-proportion. Select summarized data and enter the number of events (421) and the number of trials (500). Click Options and select the correct confidence level. Check “test and interval based on normal distribution” if the assumption of normality has been verified.




Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Worksheet
C8
C9
Select
Options...
Help
OK
Cancel
Proj...
Perform one-proportion tests and compute confidence intervals

1 Proportion (Test and Confidence Interval)

○ Samples in columns:

● Summarized data

Number of events: 421

Number of trials: 500

☐ Perform hypothesis test

Hypothesized proportion:

1 Proportion - Options

Confidence level: 95.0

Alternative: not equal

☑ Use test and interval based on normal distribution

Test and CI for One Proportion

Sample	X	N	Sample p	95% CI
1	421	500	0.842000	(0.810030, 0.873970)

Using the normal approximation.

Excel

Excel does not compute confidence intervals for estimating the population proportion.

Confidence Interval Summary

Which method do I use?

The first question to ask yourself is: **Which parameter are you trying to estimate?** If it is the mean (μ), then ask yourself: **Is the population standard deviation (σ) known?** If yes, then follow the next 3 steps:

Confidence Interval about the Population Mean (μ) when **σ** is Known

1. $Z_{\frac{\alpha}{2}}$ critical value (from the standard normal table)
2. $E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$
3. $\bar{x} \pm E$

If no, follow these 3 steps:

Confidence Interval about the Population Mean (μ) when **σ** is Unknown

1. $t_{\frac{\alpha}{2}}$ critical value with n-1 df from the student t-distribution
2. $E = t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$
3. $\bar{x} \pm E$

If you want to construct a confidence interval about the population proportion, follow these 3 steps:

Confidence Interval about the Proportion

1. $Z_{\frac{\alpha}{2}}$ critical value from the standard normal table
2. $E = Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
3. $\hat{p} \pm E$

Remember that the assumption of normality must be verified.

CHAPTER OVERVIEW

3: Hypothesis Testing

3.1: The Fundamentals of Hypothesis Testing

The previous two chapters introduced methods for organizing and summarizing sample data, and using sample statistics to estimate population parameters. This chapter introduces the next major topic of inferential statistics: hypothesis testing.

Note

A hypothesis is a statement or claim about a property of a population.

The Fundamentals of Hypothesis Testing

When conducting scientific research, typically there is some known information, perhaps from some past work or from a long accepted idea. We want to test whether this claim is believable. This is the basic idea behind a hypothesis test:

- State what we think is true.
- Quantify how confident we are about our claim.
- Use sample statistics to make inferences about population parameters.

For example, past research tells us that the average life span for a hummingbird is about four years. You have been studying the hummingbirds in the southeastern United States and find a sample mean lifespan of 4.8 years. Should you reject the known or accepted information in favor of your results? How confident are you in your estimate? At what point would you say that there is enough evidence to reject the known information and support your alternative claim? How far from the known mean of four years can the sample mean be before we reject the idea that the average lifespan of a hummingbird is four years?

Definition: hypothesis testing

Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a characteristic of a population.

A hypothesis is a claim or statement about a characteristic of a population of interest to us. A hypothesis test is a way for us to use our sample statistics to test a specific claim.

Example 3.1.1:

The population mean weight is known to be 157 lb. We want to test the claim that the mean weight has increased.

Example 3.1.2:

Two years ago, the proportion of infected plants was 37%. We believe that a treatment has helped, and we want to test the claim that there has been a reduction in the proportion of infected plants.

Components of a Formal Hypothesis Test

The null hypothesis is a statement about the value of a population parameter, such as the population mean (μ) or the population proportion (*p*). It contains the condition of equality and is denoted as H_0 (H-naught).

$$H_0 : \mu = 157 \text{ or } H0 : p = 0.37$$

The alternative hypothesis is the claim to be tested, the opposite of the null hypothesis. It contains the value of the parameter that we consider plausible and is denoted as H_1 .

$$H_1 : \mu > 157 \text{ or } H1 : p \neq 0.37$$

The test statistic is a value computed from the sample data that is used in making a decision about the rejection of the null hypothesis. The test statistic converts the sample mean ($\bar{x}$) or sample proportion ($\hat{p}$) to a Z- or t-score under the assumption that the null hypothesis is true. It is used to decide whether the difference between the sample statistic and the hypothesized claim is significant.

The p-value is the area under the curve to the left or right of the test statistic. It is compared to the level of significance (α).

The critical value is the value that defines the rejection zone (the test statistic values that would lead to rejection of the null hypothesis). It is defined by the level of significance.

The level of significance (α) is the probability that the test statistic will fall into the critical region when the null hypothesis is true. This level is set by the researcher.

The conclusion is the final decision of the hypothesis test. The conclusion must always be clearly stated, communicating the decision based on the components of the test. It is important to realize that we never prove or accept the null hypothesis. We are merely saying that the sample evidence is not strong enough to warrant the rejection of the null hypothesis. The conclusion is made up of two parts:

1) Reject or fail to reject the null hypothesis, and 2) there is or is not enough evidence to support the alternative claim.

Option 1) Reject the null hypothesis (H0). This means that you have enough statistical evidence to support the alternative claim (H_1).

Option 2) Fail to reject the null hypothesis (H0). This means that you do NOT have enough evidence to support the alternative claim (H_1).

Another way to think about hypothesis testing is to compare it to the US justice system. A defendant is innocent until proven guilty (Null hypothesis—innocent). The prosecuting attorney tries to prove that the defendant is guilty (Alternative hypothesis—guilty). There are two possible conclusions that the jury can reach. First, the defendant is guilty (Reject the null hypothesis). Second, the defendant is not guilty (Fail to reject the null hypothesis). This is NOT the same thing as saying the defendant is innocent! In the first case, the prosecutor had enough evidence to reject the null hypothesis (innocent) and support the alternative claim (guilty). In the second case, the prosecutor did NOT have enough evidence to reject the null hypothesis (innocent) and support the alternative claim of guilty.

The Null and Alternative Hypotheses

There are three different pairs of null and alternative hypotheses:

Table $PageIndex1$: The rejection zone for a two-sided hypothesis test.

Two-sided	Left-sided	Right-sided
$\mathrm{H_O} : \mu = \mathrm{c}$	$\mathbf{H_0} : \boldsymbol{\mu} = \mathbf{C}$	$\mathbf{H_0} : \boldsymbol{\mu} = \mathbf{C}$
$\mathbf{H_1} : \boldsymbol{\mu} \neq \mathbf{C}$	$\mathbf{H_1} : \boldsymbol{\mu} < \mathbf{C}$	$\mathbf{H_1} : \boldsymbol{\mu} > \mathbf{C}$

where *c* is some known value.

A Two-sided Test

This tests whether the population parameter is equal to, versus not equal to, some specific value.

$$\mathbf{Ho: \mu = 12 \text{ vs. } H_1: \mu \neq 12}$$

The critical region is divided equally into the two tails and the critical values are ± values that define the rejection zones.

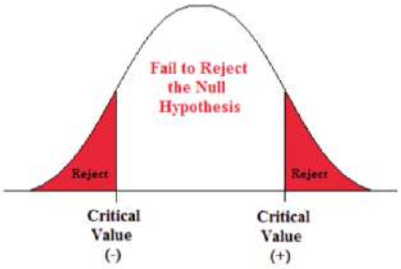


Figure 3.1.1: The rejection zone for a two-sided hypothesis test.

✓ Example 3.1.3:

A forester studying diameter growth of red pine believes that the mean diameter growth will be different if a fertilization treatment is applied to the stand.

- Ho: μ = 1.2 in./ year
- H_1: μ ≠ 1.2 in./ year

This is a two-sided question, as the forester doesn't state whether population mean diameter growth will increase or decrease.

A Right-sided Test

This tests whether the population parameter is equal to, versus greater than, some specific value.

$$\textbf{Ho: } \mu = 12 \textbf{ vs. } H_1\textbf{: } \mu > 12$$

The critical region is in the right tail and the critical value is a positive value that defines the rejection zone.

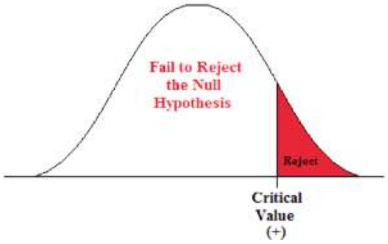


Figure 3.1.2: The rejection zone for a right-sided hypothesis test.

✓ Example 3.1.4:

A biologist believes that there has been an increase in the mean number of lakes infected with milfoil, an invasive species, since the last study five years ago.

- Ho: μ = 15 lakes
- H1: μ >15 lakes

This is a right-sided question, as the biologist believes that there has been an increase in population mean number of infected lakes.

A Left-sided Test

This tests whether the population parameter is equal to, versus less than, some specific value.

$$\textbf{Ho: } \mu = 12 \textbf{ vs. } H_1\textbf{: } \mu < 12$$

The critical region is in the left tail and the critical value is a negative value that defines the rejection zone.

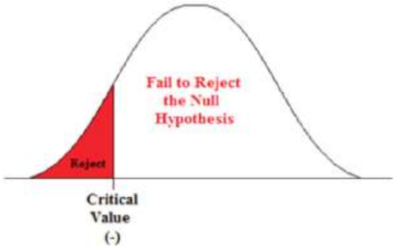


Figure 3.1.3: The rejection zone for a left-sided hypothesis test.

✓ Example 3.1.5:

A scientist's research indicates that there has been a change in the proportion of people who support certain environmental policies. He wants to test the claim that there has been a reduction in the proportion of people who support these policies.

- Ho: p = 0.57

- **H_1**: $p < 0.57$

> This is a left-sided question, as the scientist believes that there has been a reduction in the true population proportion.

Statistically Significant

When the observed results (the sample statistics) are unlikely (a low probability) under the assumption that the null hypothesis is true, we say that the result is statistically significant, and we reject the null hypothesis. This result depends on the level of significance, the sample statistic, sample size, and whether it is a one- or two-sided alternative hypothesis.

Types of Errors

When testing, we arrive at a conclusion of rejecting the null hypothesis or failing to reject the null hypothesis. Such conclusions are sometimes correct and sometimes incorrect (even when we have followed all the correct procedures). We use incomplete sample data to reach a conclusion and there is always the possibility of reaching the wrong conclusion. There are four possible conclusions to reach from hypothesis testing. Of the four possible outcomes, two are correct and two are NOT correct.

Table $3.1.2$. Possible outcomes from a hypothesis test.

	H_0 is True	H_1 is True
Do Not Reject H_0	Correct Conclusion	Type II Error
Reject H_0	Type I Error	Correct Conclusion

A Type I error is when we reject the null hypothesis when it is true. The symbol **α** (alpha) is used to represent Type I errors. This is the same alpha we use as the level of significance. By setting alpha as low as reasonably possible, we try to control the Type I error through the level of significance.

A Type II error is when we fail to reject the null hypothesis when it is false. The symbol β(beta) is used to represent Type II errors.

In general, Type I errors are considered more serious. One step in the hypothesis test procedure involves selecting the significance level (**α**), which is the probability of rejecting the null hypothesis when it is correct. So the researcher can select the level of significance that minimizes Type I errors. However, there is a mathematical relationship between α, β, and n (sample size).

- As α increases, β decreases
- As α decreases, β increases
- As sample size increases (n), both **α and β** decrease

The natural inclination is to select the smallest possible value for α, thinking to minimize the possibility of causing a Type I error. Unfortunately, this forces an increase in Type II errors. By making the rejection zone too small, you may fail to reject the null hypothesis, when, in fact, it is false. Typically, we select the best sample size and level of significance, automatically setting β.

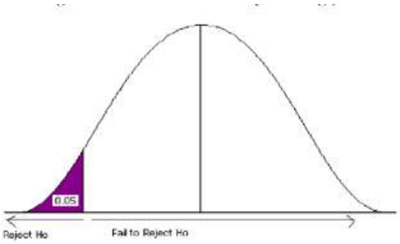


Figure $3.1.4$: Type 1 error.

Figure 4.

Power of the Test

A Type II error (β) is the probability of failing to reject a false null hypothesis. It follows that 1-β is the probability of rejecting a false null hypothesis. This probability is identified as the **power** of the test, and is often used to gauge the test's effectiveness in recognizing that a null hypothesis is false.

Definition: power of the test

The probability that at a fixed level α significance test will reject H0, when a particular alternative value of the parameter is true is called the power of the test.

Power is also directly linked to sample size. For example, suppose the null hypothesis is that the mean fish weight is 8.7 lb. Given sample data, a level of significance of 5%, and an alternative weight of 9.2 lb., we can compute the power of the test to reject μ = 8.7 lb. If we have a small sample size, the power will be low. However, increasing the sample size will increase the power of the test. Increasing the level of significance will also increase power. A 5% test of significance will have a greater chance of rejecting the null hypothesis than a 1% test because the strength of evidence required for the rejection is less. Decreasing the standard deviation has the same effect as increasing the sample size: there is more information about μ.

3.2: Hypothesis Test about the Population Mean when the Population Standard Deviation is Known

Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Known

We are going to examine two equivalent ways to perform a hypothesis test: the classical approach and the p-value approach. The **classical approach** is based on standard deviations. This method compares the test statistic (Z-score) to a critical value (Z-score) from the standard normal table. If the test statistic falls in the rejection zone, you reject the null hypothesis. The **p-value approach** is based on area under the normal curve. This method compares the area associated with the test statistic to alpha (α), the level of significance (which is also area under the normal curve). If the p-value is less than alpha, you would reject the null hypothesis.

> *As a past student poetically said: If the p-value is a wee value, Reject Ho*

Both methods must have:

- Data from a random sample.
- Verification of the assumption of normality.
- A null and alternative hypothesis.
- A criterion that determines if we reject or fail to reject the null hypothesis.
- A conclusion that answers the question.

There are four steps required for a hypothesis test:

1. State the null and alternative hypotheses.
2. State the level of significance and the critical value.
3. Compute the test statistic.
4. State a conclusion.

The Classical Method for Testing a Claim about the Population Mean (μ) when the Population Standard Deviation (σ) is Known

✓ Example 3.2.1: A Two-sided Test

A forester studying diameter growth of red pine believes that the mean diameter growth will be different from the known mean growth of 1.35 inches/year if a fertilization treatment is applied to the stand. He conducts his experiment, collects data from a sample of 32 plots, and gets a sample mean diameter growth of 1.6 in./year. The population standard deviation for this stand is known to be 0.46 in./year. Does he have enough evidence to support his claim?

Solution

Step 1) State the null and alternative hypotheses.

- Ho: μ = 1.35 in./year
- H_1: μ ≠ 1.35 in./year

Step 2) State the level of significance and the critical value.

- We will choose a level of significance of 5% (α = 0.05).
- For a two-sided question, we need a two-sided critical value – Z α/2 and + Z α/2.
- The level of significance is divided by 2 (since we are only testing “not equal”). We must have two rejection zones that can deal with either a greater than or less than outcome (to the right (+) or to the left (-)).
- We need to find the Z-score associated with the area of 0.025. The red areas are equal to α/2 = 0.05/2 = 0.025 or 2.5% of the area under the normal curve.
- Go into the body of values and find the negative Z-score associated with the area 0.025.

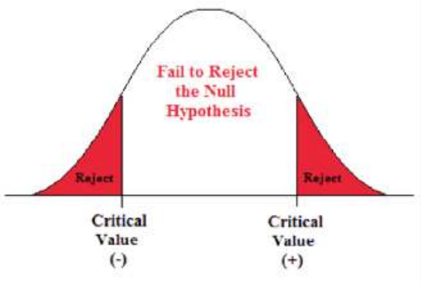


Figure 3.2.1: The rejection zone for a two-sided test

- The negative critical value is -1.96. Since the curve is symmetric, we know that the positive critical value is 1.96.
- ±1.96 are the critical values. These values set up the rejection zone. If the test statistic falls within these red rejection zones, we reject the null hypothesis.

Step 3) Compute the test statistic.

- The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.
- $$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.2.1}$$
- For this problem, the test statistic is

$$z=\frac{1.6-1.35}{\frac{0.46}{\sqrt{32}}}=3.07$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value. If the test statistic falls into the rejection zones, reject the null hypothesis. In other words, if the test statistic is greater than +1.96 or less than -1.96, reject the null hypothesis.

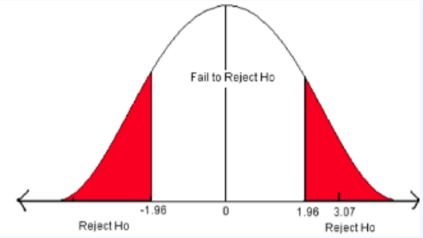


Figure 3.2.2: The critical values for a two-sided test when α = 0.05.

In this problem, the test statistic falls in the red rejection zone. The test statistic of 3.07 is greater than the critical value of 1.96.We will reject the null hypothesis. We have enough evidence to support the claim that the mean diameter growth is different from (not equal to) 1.35 in./year.

✓ Example 3.2.2: A Right-sided Test

A researcher believes that there has been an increase in the average farm size in his state since the last study five years ago. The previous study reported a mean size of 450 acres with a population standard deviation (**σ**) of 167 acres. He samples 45 farms and gets a sample mean of 485.8 acres. Is there enough information to support his claim?

Solution

Step 1) State the null and alternative hypotheses.

- Ho: μ = 450 acres
- H_1: μ >450 acres

Step 2) State the level of significance and the critical value.

- We will choose a level of significance of 5% (α = 0.05).

- For a one-sided question, we need a one-sided positive critical value Zα.
- The level of significance is all in the right side (the rejection zone is just on the right side).
- We need to find the Z-score associated with the 5% area in the right tail.

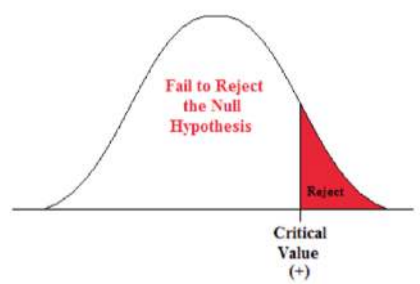


Figure 3.2.3: Rejection zone for a right-sided hypothesis test.

- Go into the body of values in the standard normal table and find the Z-score that separates the lower 95% from the upper 5%.
- The critical value is 1.645. This value sets up the rejection zone.

Step 3) Compute the test statistic.

- The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.
- $$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.2.2}$$
- For this problem, the test statistic is

$$z=\frac{485.8-450}{\frac{167}{\sqrt{45}}}=1.44$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value.

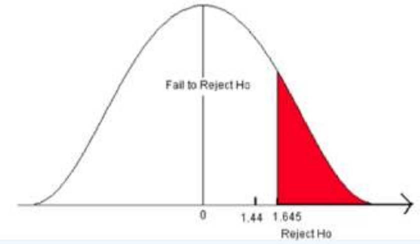


Figure 3.2.4: The critical value for a right-sided test when α = 0.05.

- The test statistic does not fall in the rejection zone. It is less than the critical value.

We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean farm size has increased from 450 acres.

✓ Example 3.2.3:A Left-sided Test

A researcher believes that there has been a reduction in the mean number of hours that college students spend preparing for final exams. A national study stated that students at a 4-year college spend an average of 23 hours preparing for 5 final exams each semester with a population standard deviation of 7.3 hours. The researcher sampled 227 students and found a sample mean study time of 19.6 hours. Does this indicate that the average study time for final exams has decreased? Use a 1% level of significance to test this claim.

Solution

Step 1) State the null and alternative hypotheses.

- Ho: μ = 23 hours
- H_1: μ < 23 hours

Step 2) State the level of significance and the critical value.

- This is a left-sided test so alpha (0.01) is all in the left tail.

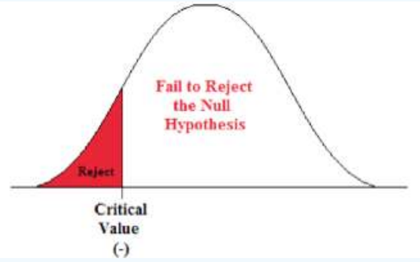


Figure 3.2.5: The rejection zone for a left-sided hypothesis test.

- Go into the body of values in the standard normal table and find the Z-score that defines the lower 1% of the area.
- The critical value is -2.33. This value sets up the rejection zone.

Step 3) Compute the test statistic.

- The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.
- $$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.2.3}$$
- For this problem, the test statistic is

$$z = \frac{19.6 - 23}{\frac{7.3}{\sqrt{277}}}$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value.

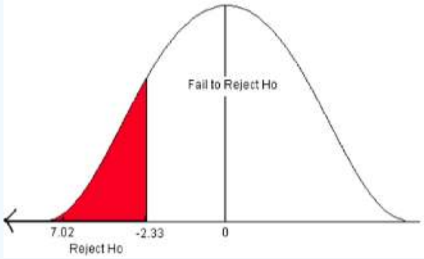


Figure 3.2.6: The critical value for a left-sided test when α = 0.01.

- The test statistic falls in the rejection zone. The test statistic of -7.02 is less than the critical value of -2.33.

We reject the null hypothesis. We have sufficient evidence to support the claim that the mean final exam study time has decreased below 23 hours.

Testing a Hypothesis using P-values

The p-value is the probability of observing our sample mean given that the null hypothesis is true. It is the area under the curve to the left or right of the test statistic. If the probability of observing such a sample mean is very small (less than the level of significance), we would reject the null hypothesis. Computations for the p-value depend on whether it is a one- or two-sided test.

Steps for a hypothesis test using p-values:

- State the null and alternative hypotheses.
- State the level of significance.
- Compute the test statistic and find the area associated with it (this is the p-value).

- Compare the p-value to alpha (α) and state a conclusion.

Instead of comparing Z-score test statistic to Z-score critical value, as in the classical method, we compare area of the test statistic to area of the level of significance.

Note:The Decision Rule

If the p-value is less than alpha, we reject the null hypothesis.

Computing P-values

If it is a two-sided test (the alternative claim is ≠), the p-value is equal to two times the probability of the absolute value of the test statistic. If the test is a left-sided test (the alternative claim is “<”), then the p-value is equal to the area to the left of the test statistic. If the test is a right-sided test (the alternative claim is “>”), then the p-value is equal to the area to the right of the test statistic.

Let’s look at Example 6 again.

A forester studying diameter growth of red pine believes that the mean diameter growth will be different from the known mean growth of 1.35 in./year if a fertilization treatment is applied to the stand. He conducts his experiment, collects data from a sample of 32 plots, and gets a sample mean diameter growth of 1.6 in./year. The population standard deviation for this stand is known to be 0.46 in./year. Does he have enough evidence to support his claim?

Step 1) State the null and alternative hypotheses.

- Ho: μ = 1.35 in./year
- H_1: μ ≠ 1.35 in./year

Step 2) State the level of significance.

- We will choose a level of significance of 5% (α = 0.05).

Step 3) Compute the test statistic.

- For this problem, the test statistic is:

$$z=\frac{1.6-1.35}{\frac{0.46}{\sqrt{32}}}=3.07$$

The p-value is two times the area of the absolute value of the test statistic (because the alternative claim is “not equal”).

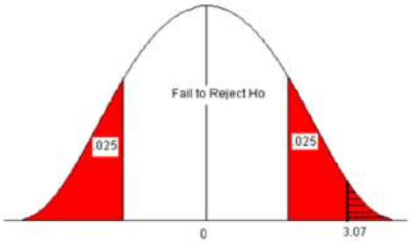


Figure 3.2.7: The p-value compared to the level of significanc

- Look up the area for the Z-score 3.07 in the standard normal table. The area (probability) is equal to 1 – 0.9989 = 0.0011.
- Multiply this by 2 to get the p-value = 2 * 0.0011 = 0.0022.

Step 4) Compare the p-value to alpha and state a conclusion.

- Use the Decision Rule (if the p-value is less than α, reject H_0).
- In this problem, the p-value (0.0022) is less than alpha (0.05).
- We reject the H_0. We have enough evidence to support the claim that the mean diameter growth is different from 1.35 inches/year.

Let’s look at Example 7 again.

A researcher believes that there has been an increase in the average farm size in his state since the last study five years ago. The previous study reported a mean size of 450 acres with a population standard deviation (**σ**) of 167 acres. He samples 45 farms and gets a sample mean of 485.8 acres. Is there enough information to support his claim?

Step 1) State the null and alternative hypotheses.

- Ho: μ = 450 acres
- H_1: μ >450 acres

Step 2) State the level of significance.

- We will choose a level of significance of 5% (α = 0.05).

Step 3) Compute the test statistic.

- For this problem, the test statistic is

$$z=\frac{485.8-450}{\frac{167}{\sqrt{45}}}=1.44$$

The p-value is the area to the right of the Z-score 1.44 (the hatched area).

- This is equal to 1 – 0.9251 = 0.0749.
- The p-value is 0.0749.

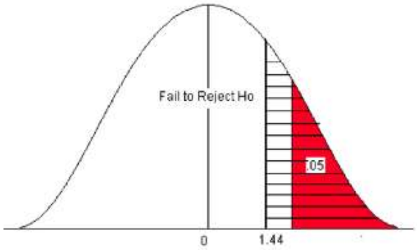


Figure 3.2.8: The p-value compared to the level of significance for a right-sided test.

Step 4) Compare the p-value to alpha and state a conclusion.

- Use the Decision Rule.
- In this problem, the p-value (0.0749) is greater than alpha (0.05), so we Fail to Reject the H_0.
- The area of the test statistic is greater than the area of alpha (α).

We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean farm size has increased.

Let’s look at Example 8 again.

A researcher believes that there has been a reduction in the mean number of hours that college students spend preparing for final exams. A national study stated that students at a 4-year college spend an average of 23 hours preparing for 5 final exams each semester with a population standard deviation of 7.3 hours. The researcher sampled 227 students and found a sample mean study time of 19.6 hours. Does this indicate that the average study time for final exams has decreased? Use a 1% level of significance to test this claim.

Step 1) State the null and alternative hypotheses.

- H_0: μ = 23 hours
- H_1: μ < 23 hours

Step 2) State the level of significance.

- This is a left-sided test so alpha (0.01) is all in the left tail.

Step 3) Compute the test statistic.

- For this problem, the test statistic is

$$z=\frac{19.6-23}{\frac{7.3}{\sqrt{227}}}=-7.02$$

The p-value is the area to the left of the test statistic (the little black area to the left of -7.02). The Z-score of -7.02 is not on the standard normal table. The smallest probability on the table is 0.0002. We know that the area for the Z-score -7.02 is smaller than this area (probability). Therefore, the p-value is <0.0002.

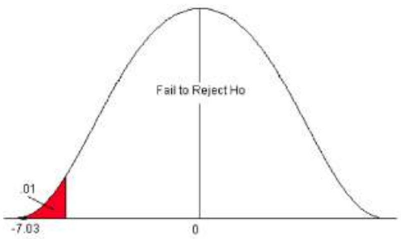


Figure 3.2.9: The p-value compared to the level of significance for a left-sided test.

Step 4) Compare the p-value to alpha and state a conclusion.

- Use the Decision Rule.
- In this problem, the p-value (p<0.0002) is less than alpha (0.01), so we Reject the H_0.
- The area of the test statistic is much less than the area of alpha (α).

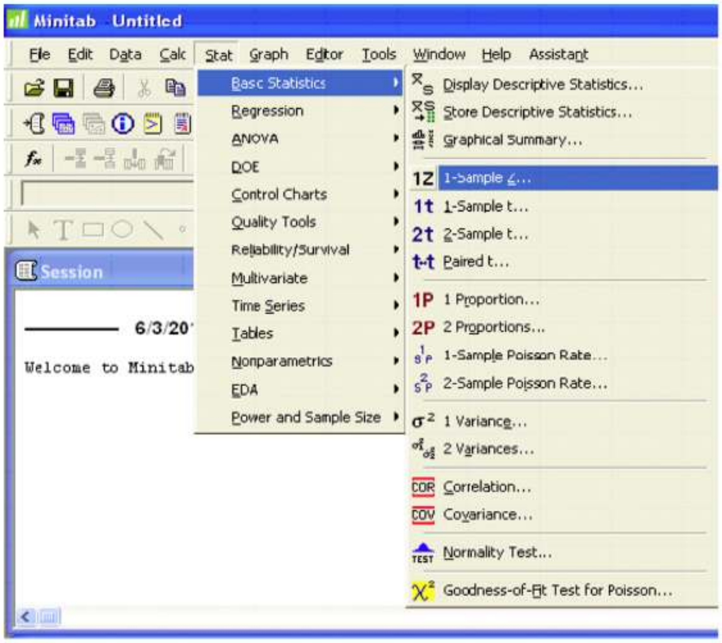
We reject the null hypothesis. We have enough evidence to support the claim that the mean final exam study time has decreased below 23 hours.

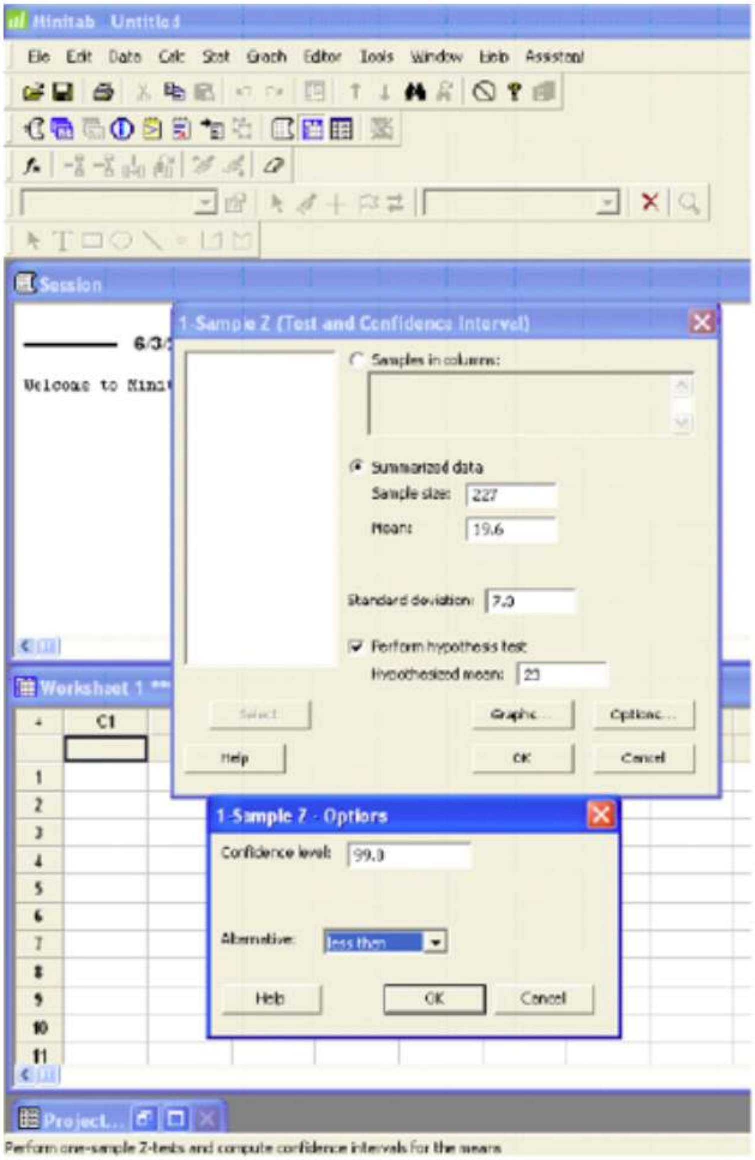
Both the classical method and p-value method for testing a hypothesis will arrive at the same conclusion. In the classical method, the critical Z-score is the number on the z-axis that defines the level of significance (α). The test statistic converts the sample mean to units of standard deviation (a Z-score). If the test statistic falls in the rejection zone defined by the critical value, we will reject the null hypothesis. In this approach, two Z-scores, which are numbers on the z-axis, are compared. In the p-value approach, the p-value is the area associated with the test statistic. In this method, we compare α (which is also area under the curve) to the p-value. If the p-value is less than α, we reject the null hypothesis. The p-value is the probability of observing such a sample mean when the null hypothesis is true. If the probability is too small (less than the level of significance), then we believe we have enough statistical evidence to reject the null hypothesis and support the alternative claim.

Software Solutions

Minitab

(referring to Ex. 8)


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
6/3/20
Welcome to Minitab
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Time Series
Tables
Nonparametrics
EDA
Power and Sample Size
Display Descriptive Statistics...
Store Descriptive Statistics...
Graphical Summary...
1Z 1-Sample Z...
1t 1-Sample t...
2t 2-Sample t...
t-t Paired t...
1P 1 Proportion...
2P 2 Proportions...
1-Sample Poisson Rate...
2-Sample Poisson Rate...
σ2 1 Variance...
2 Variances...
COR Correlation...
COV Covariance...
TEST Normality Test...
χ2 Goodness-of-Fit Test for Poisson...



○ Samples in columns:

● Summarized data

Sample size: 227

Mean: 19.6

Standard deviation: 7.3

☑ Perform hypothesis test

Hypothesized mean: 23

Confidence level: 99.0

Alternative: less than

Perform one-sample Z-tests and compute confidence intervals for the means

One-Sample Z

Test of mu = 23 vs. < 23
The assumed standard deviation = 7.3

			99% Upper		
N	Mean	SE Mean	Bound	Z	P
227	19.600	0.485	20.727	-7.02	0.000

Excel

Excel does not offer 1-sample hypothesis testing.

3.3: Hypothesis Test about the Population Mean when the Population Standard Deviation is Unknown

Hypothesis Test about the Population Mean (μ) when the Population Standard Deviation (σ) is Unknown

Frequently, the population standard deviation (σ) is not known. We can estimate the population standard deviation (σ) with the sample standard deviation (s). However, the test statistic will no longer follow the standard normal distribution. We must rely on the student's t-distribution with n-1 degrees of freedom. Because we use the sample standard deviation (s), the test statistic will change from a Z-score to a t-score.

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \tag{3.3.1}$$

Steps for a hypothesis test are the same that we covered in Section 2.

- State the null and alternative hypotheses.
- State the level of significance and the critical value.
- Compute the test statistic.
- State a conclusion.

Just as with the hypothesis test from the previous section, the data for this test must be from a random sample and requires either that the population from which the sample was drawn be normal or that the sample size is sufficiently large (n≥30). A t-test is robust, so small departures from normality will not adversely affect the results of the test. That being said, if the sample size is smaller than 30, it is always good to verify the assumption of normality through a normal probability plot.

We will still have the same three pairs of null and alternative hypotheses and we can still use either the classical approach or the p-value approach.

Table $PageIndex1$: The rejection zone for a two-sided hypothesis test.

Two-sided	Left-sided	Right-sided
$\mathrm{H}_{\mathrm{O}} : \boldsymbol{\mu} = \mathrm{c}$	$\mathbf{H_0} : \boldsymbol{\mu} = \mathbf{C}$	$\mathbf{H_0} : \boldsymbol{\mu} = \mathbf{C}$
$\mathbf{H_1} : \boldsymbol{\mu} \neq \mathbf{C}$	$\mathbf{H_1} : \boldsymbol{\mu} < \mathbf{C}$	$\mathbf{H_1} : \boldsymbol{\mu} > \mathbf{C}$

Selecting the correct critical value from the student's t-distribution table depends on three factors: the type of test (one-sided or two-sided alternative hypothesis), the sample size, and the level of significance.

For a two-sided test ("not equal" alternative hypothesis), the critical value (tα/2), is determined by alpha (α), the level of significance, divided by two, to deal with the possibility that the result could be less than OR greater than the known value.

- If your level of significance was 0.05, you would use the 0.025 column to find the correct critical value (0.05/2 = 0.025).
- If your level of significance was 0.01, you would use the 0.005 column to find the correct critical value (0.01/2 = 0.005).

For a one-sided test ("a less than" or "greater than" alternative hypothesis), the critical value (tα) , is determined by alpha (α), the level of significance, being all in the one side.

- If your level of significance was 0.05, you would use the 0.05 column to find the correct critical value for either a left or right-side question. If you are asking a "less than" (left-sided question, your critical value will be negative. If you are asking a "greater than" (right-sided question), your critical value will be positive.

✓ Example 3.3.1

Find the critical value you would use to test the claim that μ ≠ 112 with a sample size of 18 and a 5% level of significance.

Solution

In this case, the critical value ($t_{\alpha/2}$) would be 2.110. This is a two-sided question (≠) so you would divide alpha by 2 (0.05/2 = 0.025) and go down the 0.025 column to 17 degrees of freedom.

✓ Example 3.3.2

What would the critical value be if you wanted to test that $\mu < 112$ for the same data?

Solution

In this case, the critical value would be 1.740. This is a one-sided question (<) so alpha would be divided by 1 (0.05/1 = 0.05). You would go down the 0.05 column with 17 degrees of freedom to get the correct critical value.

✓ Example 3.3.3:A Two-sided Test

In 2005, the mean pH level of rain in a county in northern New York was 5.41. A biologist believes that the rain acidity has changed. He takes a random sample of 11 rain dates in 2010 and obtains the following data. Use a 1% level of significance to test his claim.

4.70, 5.63, 5.02, 5.78, 4.99, 5.91, 5.76, 5.54, 5.25, 5.18, 5.01

The sample size is small and we don't know anything about the distribution of the population, so we examine a normal probability plot. The distribution looks normal so we will continue with our test.

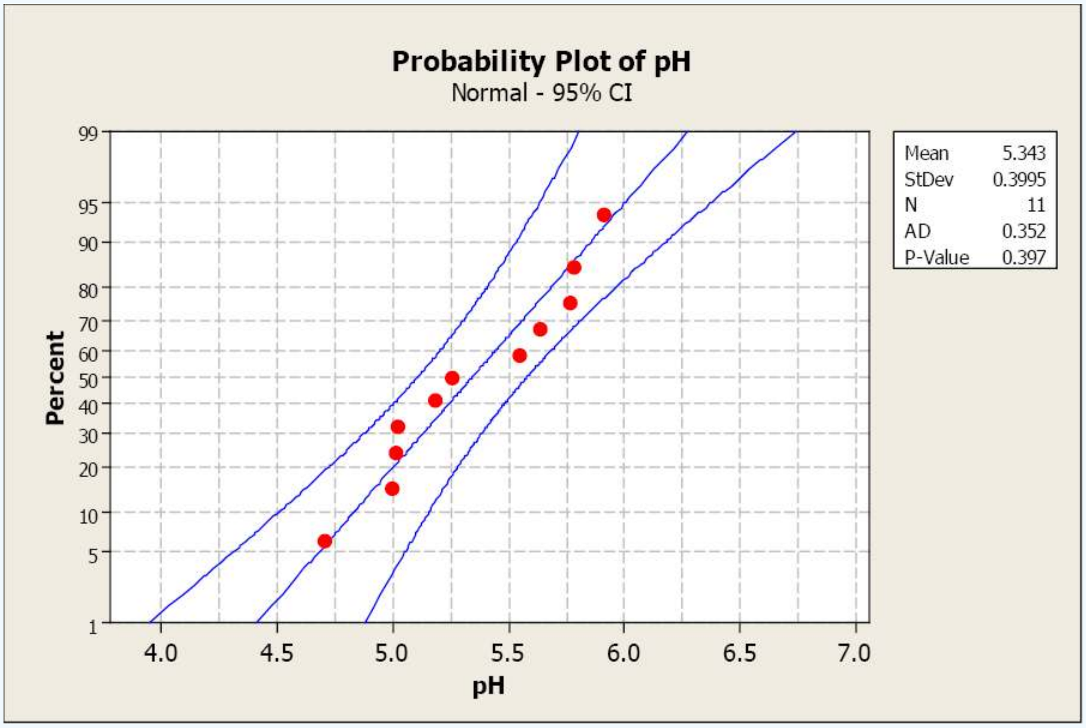


Figure 3.3.1: A normal probability plot for Example 3.3.3

The sample mean is 5.343 with a sample standard deviation of 0.397.

Solution

Step 1) State the null and alternative hypotheses.

- Ho: $\mu = 5.41$
- H1: $\mu \neq 5.41$

Step 2) State the level of significance and the critical value.

- This is a two-sided question so alpha is divided by two.

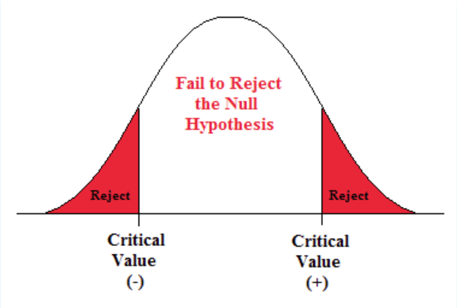


Figure 3.3.2: The rejection zones for a two-sided test.

- t α/2 is found by going down the 0.005 column with 14 degrees of freedom.
- t α/2 = ±3.169.

Step 3) Compute the test statistic.

- The test statistic is a t-score.

$$t=\frac{\bar{x}-\mu}{\frac{s}{sqrtn}} \tag{3.3.2}$$

- For this problem, the test statistic is

$$t=\frac{5.343-5.41}{\frac{0.397}{\sqrt{11}}}=-0.560$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value.

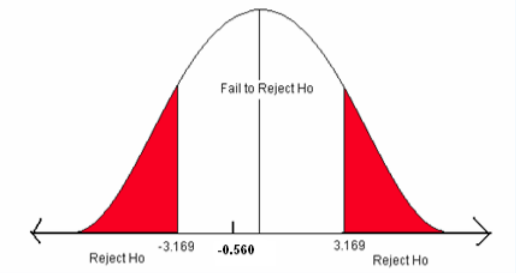


Figure 3.3.3: The critical values for a two-sided test when α = 0.01.

- The test statistic does not fall in the rejection zone.

We will fail to reject the null hypothesis. We do not have enough evidence to support the claim that the mean rain pH has changed.

✓ Example 3.3.4:A One-sided Test

Cadmium, a heavy metal, is toxic to animals. Mushrooms, however, are able to absorb and accumulate cadmium at high concentrations. The government has set safety limits for cadmium in dry vegetables at 0.5 ppm. Biologists believe that the mean level of cadmium in mushrooms growing near strip mines is greater than the recommended limit of 0.5 ppm, negatively impacting the animals that live in this ecosystem. A random sample of 51 mushrooms gave a sample mean of 0.59 ppm with a sample standard deviation of 0.29 ppm. Use a 5% level of significance to test the claim that the mean cadmium level is greater than the acceptable limit of 0.5 ppm.

The sample size is greater than 30 so we are assured of a normal distribution of the means.

Solution

Step 1) State the null and alternative hypotheses.

- Ho: **μ** = 0.5 ppm

- H1: **μ** > 0.5 ppm

Step 2) State the level of significance and the critical value.

- This is a right-sided question so alpha is all in the right tail.

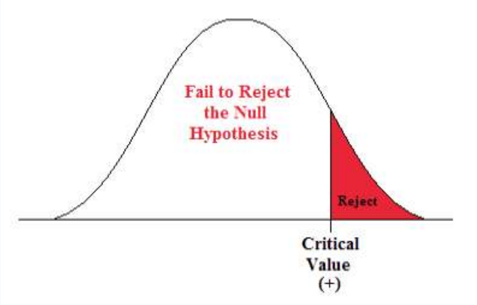


Figure 3.3.4: Rejection zone for a right-sided test.

- t α is found by going down the 0.05 column with 50 degrees of freedom.
- t α = 1.676

Step 3) Compute the test statistic.

- The test statistic is a t-score.

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \tag{3.3.3}$$

- For this problem, the test statistic is

$$t=\frac{0.59-0.50}{\frac{0.29}{\sqrt{51}}}=2.216$$

Step 4) State a Conclusion.

- Compare the test statistic to the critical value.

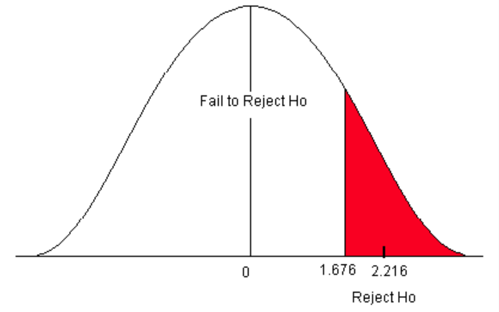


Figure 3.3.5: Critical value for a right-sided test when α = 0.05.

The test statistic falls in the rejection zone. We will reject the null hypothesis. We have enough evidence to support the claim that the mean cadmium level is greater than the acceptable safe limit.

BUT, what happens if the significance level changes to 1%?

The critical value is now found by going down the 0.01 column with 50 degrees of freedom. The critical value is 2.403. The test statistic is now LESS THAN the critical value. The test statistic does not fall in the rejection zone. The conclusion will change. We do NOT have enough evidence to support the claim that the mean cadmium level is greater than the acceptable safe limit of 0.5 ppm.

Note

The level of significance is the probability that you, as the researcher, set to decide if there is enough statistical evidence to support the alternative claim. It should be set before the experiment begins.

P-value Approach

We can also use the p-value approach for a hypothesis test about the mean when the population standard deviation (**σ**) is unknown. However, when using a student's t-table, we can only estimate the range of the p-value, not a specific value as when using the standard normal table. The student's t-table has area (probability) across the top row in the table, with t-scores in the body of the table.

- To find the p-value (the area associated with the test statistic), you would go to the row with the number of degrees of freedom.
- Go across that row until you find the two values that your test statistic is between, then go up those columns to find the estimated range for the p-value.

✓ Example $3.3.5$

Estimating P-value from a Student's T-table

Table $PageIndex2$. Portion of the student's t-table. t- distribution, Area in right tail.

df	.05	.025	.02	.01	.005
1	6.314	12.706	15.894	31.821	63.657
2	2.920	4.303	4.849	6.965	9.925
3	2.353	3.182	3.482	4.541	5.841
4	2.132	2.776	2.999	3.747	4.604
5	2.015	2.571	2.757	3.365	4.032

Solution

If your test statistic is 3.789 with 3 degrees of freedom, you would go across the 3 df row. The value 3.789 falls between the values 3.482 and 4.541 in that row. Therefore, the p-value is between 0.02 and 0.01. The p-value will be greater than 0.01 but less than 0.02 (0.01<p<0.02).

Conclusion

If your level of significance is 5%, you would reject the null hypothesis as the p-value (0.01-0.02) is less than alpha (α) of 0.05.

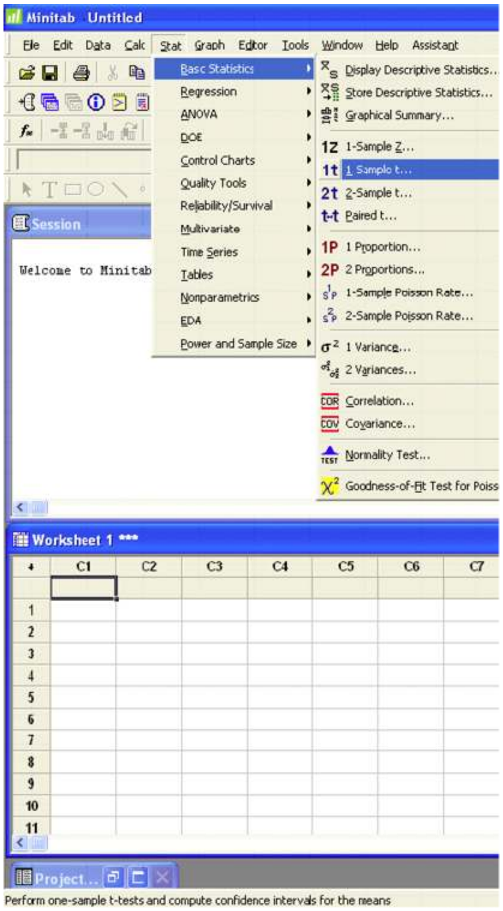
If your level of significance is 1%, you would fail to reject the null hypothesis as the p-value (0.01-0.02) is greater than alpha (α) of 0.01.

Software packages typically output p-values. It is easy to use the Decision Rule to answer your research question by the p-value method.

Software Solutions

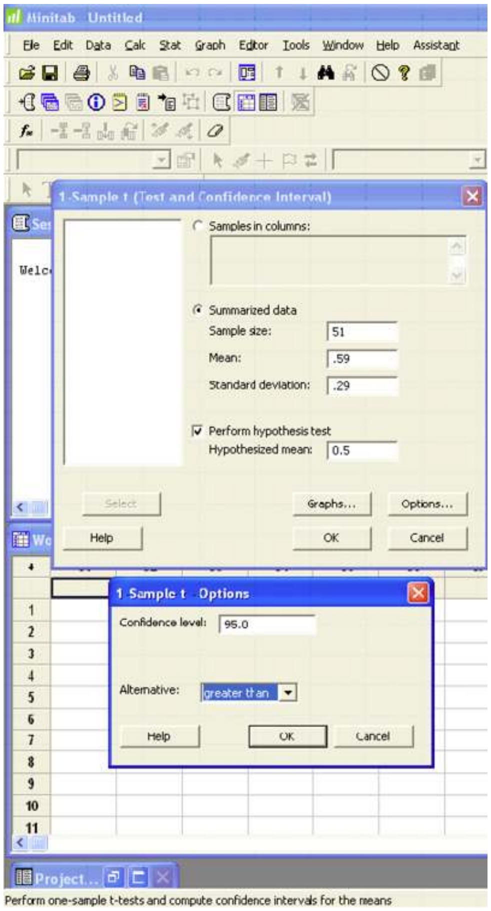
Minitab

(referring to Ex. 5)



```
Welcome to Minitab
```

Perform one-sample t-tests and compute confidence intervals for the means


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
1-Sample t (Test and Confidence Interval)
Select
Graphs...
Options...
Help
OK
Cancel
1 Sample t - Options
Help
OK
Cancel
Project...

○ Samples in columns:

● Summarized data

Sample size: 51

Mean: .59

Standard deviation: .29

☑ Perform hypothesis test

Hypothesized mean: 0.5

Confidence level: 95.0

Alternative: greater than

Perform one-sample t-tests and compute confidence intervals for the means

One-Sample T

Test of mu = 0.5 vs. > 0.5

95% Lower

N	Mean	StDev	SE Mean	Bound	T	P
51	0.5900	0.2900	0.0406	0.5219	2.22	0.016

Additional example: www.youtube.com/watch?v=WwdSjO4VUsg.

Excel

Excel does not offer 1-sample hypothesis testing.

3.4: Hypothesis Test for a Population Proportion

Hypothesis Test for a Population Proportion (*p*)

Frequently, the parameter we are testing is the population proportion.

- We are studying the proportion of trees with cavities for wildlife habitat.
- We need to know if the proportion of people who support green building materials has changed.
- Has the proportion of wolves that died last year in Yellowstone increased from the year before?

Recall that the best point estimate of *p*, the population proportion, is given by

$$\hat{p} = \frac{x}{n} \tag{3.4.1}$$

where *x* is the number of individuals in the sample with the characteristic studied and *n* is the sample size. The sampling distribution of $\hat{p}$ is approximately normal with a mean $\mu_{\hat{p}} = p$ and a standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{3.4.2}$$

when $np(1 – p) \geq 10$. We can use both the classical approach and the p-value approach for testing.

The steps for a hypothesis test are the same that we covered in Section 2.

- State the null and alternative hypotheses.
- State the level of significance and the critical value.
- Compute the test statistic.
- State a conclusion.

The test statistic follows the standard normal distribution. Notice that the standard error (the denominator) uses *p* instead of $\hat{p}$, which was used when constructing a confidence interval about the population proportion. In a hypothesis test, the null hypothesis is assumed to be true, so the known proportion is used.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.4.3}$$

- The critical value comes from the standard normal table, just as in Section 2. We will still use the same three pairs of null and alternative hypotheses as we used in the previous sections, but the parameter is now *p* instead of μ:

Table $PageIndex1$: The rejection zone for a two-sided hypothesis test.

Two-sided	Left-sided	Right-sided
$\mathrm{H_O} : \boldsymbol{p} = \mathrm{c}$	$\mathbf{H_0 : \boldsymbol{p} = C}$	$\mathbf{H_0 : \boldsymbol{p} = C}$
$\mathbf{H_1 : \boldsymbol{p} \neq C}$	$\mathbf{H_1 : \boldsymbol{p} < C}$	$\mathbf{H_1 : \boldsymbol{p} > C}$

- For a two-sided test, alpha will be divided by 2 giving a ± Zα/2 critical value.
- For a left-sided test, alpha will be all in the left tail giving a – Zα critical value.
- For a right-sided test, alpha will be all in the right tail giving a Zα critical value.

✓ Example 3.4.1

botanist has produced a new variety of hybrid soy plant that is better able to withstand drought than other varieties. The botanist knows the seed germination for the parent plants is 75%, but does not know the seed germination for the new hybrid. He tests the claim that it is different from the parent plants. To test this claim, 450 seeds from the hybrid plant are tested and 321 have germinated. Use a 5% level of significance to test this claim that the germination rate is different from 75%.

Solution

Step 1) State the null and alternative hypotheses.

- Ho: p = 0.75
- H1: p ≠ 0.75

Step 2) State the level of significance and the critical value.

This is a two-sided question so alpha is divided by 2.

- Alpha is 0.05 so the critical values are ± Zα/2 = ± Z.025.
- Look on the negative side of the standard normal table, in the body of values for 0.025.
- The critical values are ± 1.96.

Step 3) Compute the test statistic.

- The test statistic is the number of standard deviations the sample mean is from the known mean. It is also a Z-score, just like the critical value.

$$z=\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.4.4}$$

- For this problem, the test statistic is

$$z=\frac{0.713-0.75}{\sqrt{\frac{0.75(1-0.75)}{450}}}=-1.81$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value.

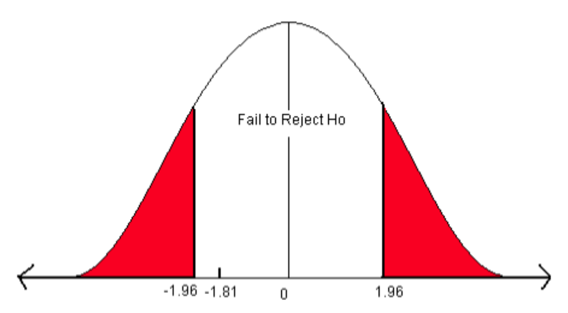


Figure 3.4.1: Critical values for a two-sided test when α = 0.05.

The test statistic does not fall in the rejection zone. We fail to reject the null hypothesis. We do not have enough evidence to support the claim that the germination rate of the hybrid plant is different from the parent plants.

Let’s answer this question using the p-value approach. Remember, for a two-sided alternative hypothesis (“not equal”), the p-value is two times the area of the test statistic. The test statistic is -1.81 and we want to find the area to the left of -1.81 from the standard normal table.

- On the negative page, find the Z-score -1.81. Find the area associated with this Z-score.
- The area = 0.0351.
- This is a two-sided test so multiply the area times 2 to get the p-value = 0.0351 x 2 = 0.0702.

Now compare the p-value to alpha. The Decision Rule states that if the p-value is less than alpha, reject the H0. In this case, the p-value (0.0702) is greater than alpha (0.05) so we will fail to reject H0. We do not have enough evidence to support the claim that the germination rate of the hybrid plant is different from the parent plants.

Example 3.4.2:

You are a biologist studying the wildlife habitat in the Monongahela National Forest. Cavities in older trees provide excellent habitat for a variety of birds and small mammals. A study five years ago stated that 32% of the trees in this forest had suitable cavities for this type of wildlife. You believe that the proportion of cavity trees has increased. You sample 196 trees and find that 79 trees have cavities. Does this evidence support your claim that there has been an increase in the proportion of cavity trees?

Use a 10% level of significance to test this claim.

Solution

Step 1) State the null and alternative hypotheses.

- Ho: p = 0.32
- H1: p > 0.32

Step 2) State the level of significance and the critical value.

This is a one-sided question so alpha is divided by 1.

- Alpha is 0.10 so the critical value is Zα = Z .10
- Look on the positive side of the standard normal table, in the body of values for 0.90.
- The critical value is 1.28.

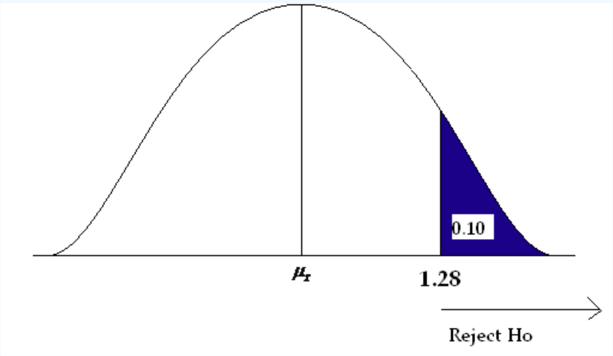


Figure 3.4.2: Critical value for a right-sided test where α = 0.10.

Step 3) Compute the test statistic.

- The test statistic is the number of standard deviations the sample proportion is from the known proportion. It is also a Z-score, just like the critical value.

$$z=\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.4.5}$$

- For this problem, the test statistic is:

$$z=\frac{0.403-0.32}{\sqrt{\frac{0.32(1-0.32)}{196}}}=2.49$$

Step 4) State a conclusion.

- Compare the test statistic to the critical value.

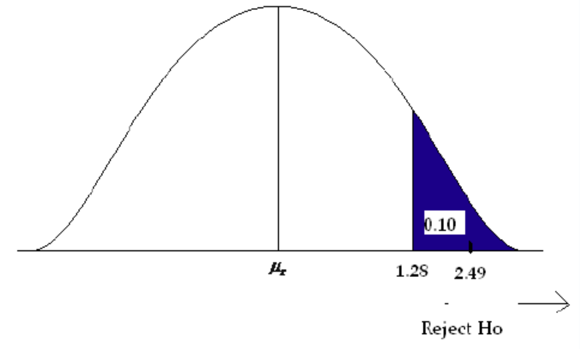


Figure 3.4.3: Comparison of the test statistic and the critical value.

The test statistic is larger than the critical value (it falls in the rejection zone). We will reject the null hypothesis. We have enough evidence to support the claim that there has been an increase in the proportion of cavity trees.

Now use the p-value approach to answer the question. This is a right-sided question ("greater than"), so the p-value is equal to the area to the right of the test statistic. Go to the positive side of the standard normal table and find the area associated with the Z-score of 2.49. The area is 0.9936. Remember that this table is cumulative from the left. To find the area to the right of 2.49, we subtract from one.

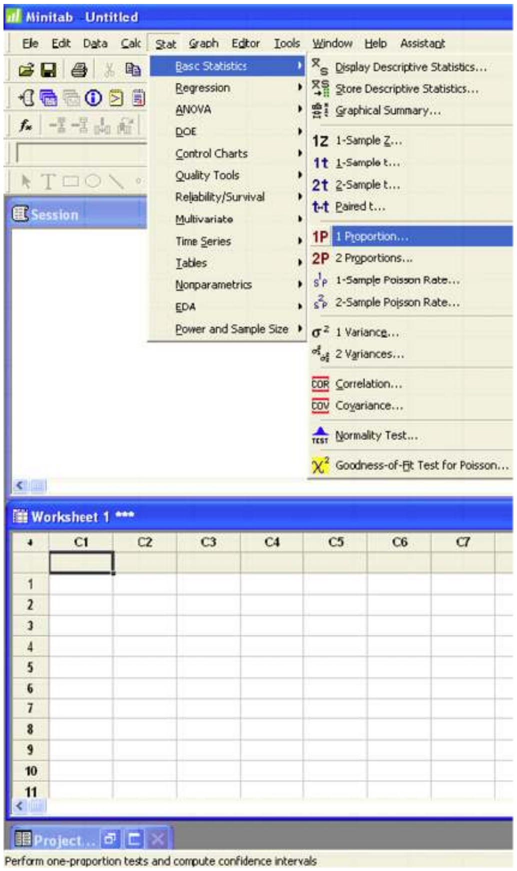
p-value = (1 – 0.9936) = 0.0064

The p-value is less than the level of significance (0.10), so we reject the null hypothesis. We have enough evidence to support the claim that the proportion of cavity trees has increased.

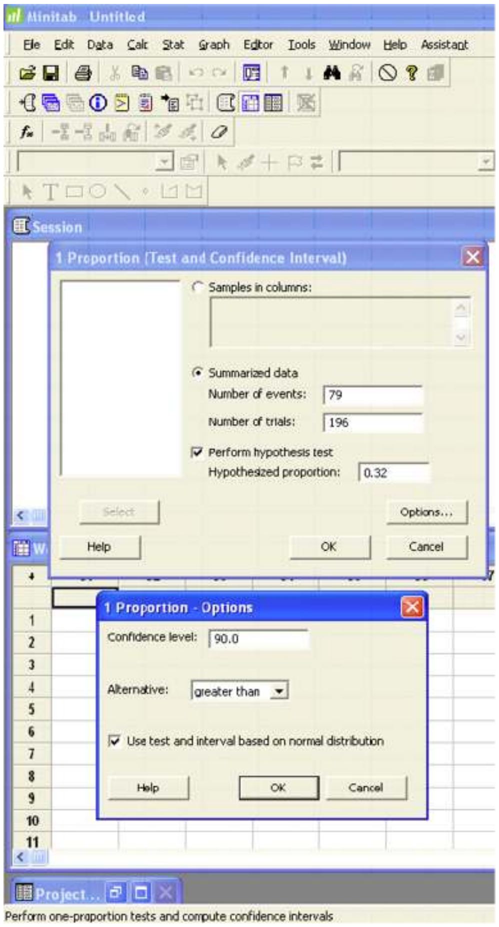
Software Solutions

Minitab

(referring to Ex. 15)



Perform one-proportion tests and compute confidence intervals


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Select
Options...
Help
OK
Cancel
Project...

1 Proportion (Test and Confidence Interval)

○ Samples in columns:

● Summarized data

Number of events: 79

Number of trials: 196

☑ Perform hypothesis test

Hypothesized proportion: 0.32

1 Proportion - Options

Confidence level: 90.0

Alternative: greater than

☑ Use test and interval based on normal distribution

Perform one-proportion tests and compute confidence intervals

Test and CI for One Proportion

Test of p = 0.32 vs. p > 0.32

90% Lower

Sample	X	N	Sample p	Bound	Z-Value	p-Value
1	79	196	0.403061	0.358160	2.49	0.006
Using the normal approximation.						

Excel

Excel does not offer 1-sample hypothesis testing.

3.5: Hypothesis Test about a Variance

Hypothesis Test about a Variance

When people think of statistical inference, they usually think of inferences involving population means or proportions. However, the particular population parameter needed to answer an experimenter's practical questions varies from one situation to another, and sometimes a population's variability is more important than its mean. Thus, product quality is often defined in terms of low variability.

Sample variance $\boldsymbol{s^2}$ can be used for inferences concerning a population variance $\boldsymbol{\sigma^2}$. For a random sample of n measurements drawn from a normal population with mean **μ** and variance $\boldsymbol{\sigma^2}$, the value $\boldsymbol{s^2}$ provides a **point estimate for** $\boldsymbol{\sigma^2}$. In addition, the quantity $\frac{(n-1)s^2}{\sigma^2}$ follows a **Chi-square**(χ^2) **distribution,** with $df = n-1$.

The properties of **Chi-square (**χ^2**) distribution** are:

- Unlike Z and t distributions, the values in a chi-square distribution are all positive.
- The chi-square distribution is asymmetric, unlike the Z and t distributions.
- There are many chi-square distributions. We obtain a particular one by specifying the degrees of freedom $(df = n-1)$ associated with the sample variances $\boldsymbol{s^2}$.

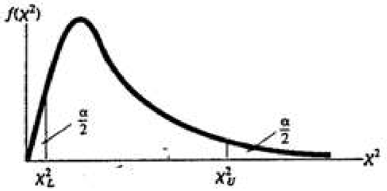


Figure 3.5.1: The chi-square distribution.

One-sample (χ^2) test for testing the hypotheses:

Null hypothesis: $H_0 : \sigma^2 = \sigma_0^2$ (constant)

Alternative hypothesis:

- $H_a : \sigma^2 > \sigma_0^2$ (one-tailed), reject H_0 if the observed $\chi^2 > \chi^2_U$ (upper-tail value at α).
- $H_a : \sigma^2 < \sigma_0^2$ (one-tailed), reject H_0 if the observed $\chi^2 < \chi^2_L$ (lower-tail value at α).
- $H_a : \sigma^2 \neq \sigma_0^2$ (two-tailed), reject H_0 if the observed $\chi^2 > \chi^2_U$ or $\chi^2 < \chi^2_L$ at α/2.

where the $\boldsymbol{\chi^2}$ critical value in the rejection region is based on degrees of freedom $df = n-1$ and a specified significance level of **α.**

Test statistic:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{3.5.1}$$

As with previous sections, if the test statistic falls in the rejection zone set by the critical value, you will reject the null hypothesis.

✓ Example 3.5.1:

A forester wants to control a dense understory of striped maple that is interfering with desirable hardwood regeneration using a mist blower to apply an herbicide treatment. She wants to make sure that treatment has a consistent application rate, in other words, low variability not exceeding 0.25 gal./acre (0.06 gal.2). She collects sample data (n = 11) on this type of mist blower and gets a sample variance of 0.064 gal.2 Using a 5% level of significance, test the claim that the variance is significantly greater than 0.06 gal.2

$H_0 : \sigma^2 = 0.06$

$H_1 : \sigma^2 > 0.06$

The critical value is 18.307. Any test statistic greater than this value will cause you to reject the null hypothesis.

The test statistic is

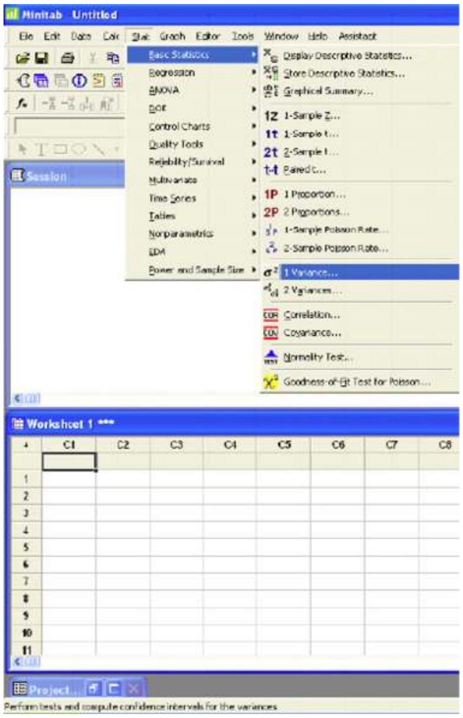
$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(11-1)0.064}{0.06} = 10.667$$

We fail to reject the null hypothesis. The forester does NOT have enough evidence to support the claim that the variance is greater than 0.06 gal.2 You can also estimate the p-value using the same method as for the student t-table. Go across the row for degrees of freedom until you find the two values that your test statistic falls between. In this case going across the row 10, the two table values are 4.865 and 15.987. Now go up those two columns to the top row to estimate the p-value (0.1-0.9). The p-value is greater than 0.1 and less than 0.9. Both are greater than the level of significance (0.05) causing us to fail to reject the null hypothesis.

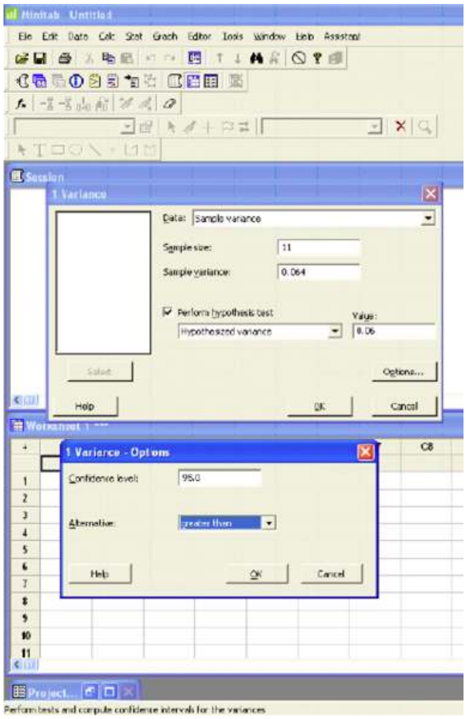
Software Solutions

Minitab

(referring to Ex. 3.5.1)



Perform tests and compute confidence intervals for the variances



Data: Sample variance

Sample size: 11

Sample variance: 0.064

☑ Perform hypothesis test

Hypothesized variance

Value: 0.06

Confidence level: 95.0

Alternative: greater than

Perform tests and compute confidence intervals for the variances

Test and CI for One Variance

Method		
Null hypothesis	Sigma-squared	= 0.06
Alternative hypothesis	Sigma-squared	> 0.06

The chi-square method is only for the normal distribution.

Tests

Test			
Method	Statistic	DF	P-Value
Chi-Square	10.67	10	0.384

Excel

Excel does not offer 1-sample χ^2 testing.

3.6: Putting it all Together Using the Classical Method

Putting it all Together Using the Classical Method

To Test a Claim about **μ** when **σ** is Known

- Write the null and alternative hypotheses.
- State the level of significance and get the critical value from the standard normal table.
- Compute the test statistic.

$$z=\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \tag{3.6.1}$$

- Compare the test statistic to the critical value (Z-score) and write the conclusion.

To Test a Claim about **μ** When **σ** is Unknown

- Write the null and alternative hypotheses.
- State the level of significance and get the critical value from the student's t-table with n-1 degrees of freedom.
- Compute the test statistic.

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \tag{3.6.2}$$

- Compare the test statistic to the critical value (t-score) and write the conclusion.

To Test a Claim about p

- Write the null and alternative hypotheses.
- State the level of significance and get the critical value from the standard normal distribution.
- Compute the test statistic.

$$z=\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \tag{3.6.3}$$

- Compare the test statistic to the critical value (Z-score) and write the conclusion.

Table $PageIndex1$. A summary table for critical Z-scores.

	Two-sided Test	One-sided Test
Alpha (á)	á	Z á
0.01	2.575	2.33
0.05	1.96	1.645
0.10	1.645	1.28

To Test a Claim about Variance

- Write the null and alternative hypotheses.
- State the level of significance and get the critical value from the chi-square table using n-1 degrees of freedom.
- Compute the test statistic.

$$\chi^2=\frac{(n-1)S^2}{\sigma_0^2} \tag{3.6.4}$$

- Compare the test statistic to the critical value and write the conclusion.

CHAPTER OVERVIEW

4: Inferences about the Differences of Two Populations

Topic hierarchy

4.1: Inferences about Two Means with Independent Samples (Assuming Unequal Variances)

Up to this point, we have discussed inferences regarding a single population parameter (e.g., μ, p, σ^2). We have used sample data to construct confidence intervals to estimate the population mean or proportion and to test hypotheses about the population mean and proportion. In both of these chapters, all the examples involved the use of one sample to form an inference about one population. Frequently, we need to compare two sets of data, and make inferences about two populations. This chapter deals with inferences about two means, proportions, or variances. For example:

- You are studying turkey habitat and want to see if the mean number of brood hens is different in New York compared to Pennsylvania.
- You want to determine if the treatment used in Skaneateles Lake has reduced the number of milfoil plants over the last three years.
- Is the proportion of people who support alternative energy in California greater compared to New York?
- Is the variability in application different between two mist blowers?

These questions can be answered by comparing the differences of:

- Mean number of hens in NY to the mean number of hens in PA.
- Number of plants in 2007 to the number of plants in 2010.
- Proportion of people in CA to the proportion of people in NY.
- Variances between the mist blowers.

This chapter is comprised of five sections. The first and second sections examine inferences about two means with two independent samples. The third section examines inferences about means with two dependent samples, the fourth section examines inferences about two proportions, and the fifth section examines inferences between two variances.

Inferences about Two Means with Independent Samples (Assuming Unequal Variances)

Using independent samples means that there is no relationship between the groups. The values in one sample have no association with the values in the other sample. For example, we want to see if the mean life span for hummingbirds in South Carolina is different from the mean life span in North Carolina. These populations are not related, and the samples are independent. We look at the difference of the independent means.

In Chapter 3, we did a one-sample t-test where we compared the sample mean ($\bar{x}$) to the hypothesized mean (μ). We expect that $\bar{x}$ would be close to μ. We use the sample mean, the sample standard deviation, and the sample size for the one-sample test.

With a two-sample t-test, we compare the population means to each other and again look at the difference. We expect that $\bar{x_1} - \bar{x_2}$ would be close to $\mu_1 - \mu_2$. The test statistic will use both sample means, sample standard deviations, and sample sizes for the test.

- For a one-sample t-test we used $\frac{s}{\sqrt{n}}$ as a measure of the standard deviation (the standard error).
- We can rewrite $\frac{s}{\sqrt{n}} \rightarrow \sqrt{\frac{s^2}{n}}$
- The numerator of the test statistic will be $(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)$
- This has a standard deviation of $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

A two-sample t-test follows the same four steps we saw in Chapter 3.

- Write the null and alternative hypotheses.
- State the level of significance and find the critical value. The critical value, from the student's t-distribution, has the lesser of n1-1 and n2 -1 degrees of freedom.
- Compute the test statistic.
- Compare the test statistic to the critical value and state a conclusion.

The assumptions we saw in Chapter 3 still must be met. Both samples come from independent random samples. The populations must be normally distributed, or both have large enough sample sizes (n1 and n2 ≥ 30). We will also use the same three pairs of null and alternative hypotheses.

Table 4.1.1. Null and alternative hypotheses.

Two-sided	Left-sided	Right=sided
$\mathrm{H}_0 : \mu_1 = \mu_2$	$\mathrm{H}_0 : \mu_1 = \mu_2$	(\mathrm{H}_{\mathrm{0}}: \mu_1=\mu_2\)
$\mathrm{H}_1 : \mu_1 \neq \mu_2$	$\mathrm{H}_1 : \mu_1 < \mu_2$	$\mathrm{H}_1 : \mu_1 > \mu_2$

Rewriting the null hypothesis of μ1 = μ2 to μ1 – μ2 = 0, simplifies the numerator. The test statistic is Welch's approximation (Satterthwaite Adjustment) under the assumption that the independent population variances are not equal.

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.1.1}$$

This test statistic follows the student's t-distribution with the degrees of freedom *adjusted* by

$$df = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{S_2^2}{n_2})^2} \tag{4.1.2}$$

A simpler alternative to determining degrees of freedom when working a problem long-hand is to use the lesser of n1-1 or n2-1 as the degrees of freedom. This method results in a smaller value for degrees of freedom and therefore a larger critical value. This makes the test more conservative, requiring more evidence to reject the null hypothesis.

✓ Example 4.1.1:

A forester is studying the number of cavity trees in old growth stands in Adirondack Park in northern New York. He wants to know if there is a significant difference between the mean number of cavity trees in the Adirondack Park and the old growth stands in the Monongahela National Forest. He collects two independent random samples from each forest. Use a 5% level of significance to test this claim.

Adirondack Park	Monongahela Forest
n_1 = 51 stands	n_2 = 56 stands
$\bar{x}_1$= 39.6	$\bar{x}_2$= 43.9
s_1 = 9.4	s_2 = 10.7

1) $H_0 : \mu_1 = \mu_2 or \mu_1 - \mu_2 = 0$ There is no difference between the two population means.

$H_1 : \mu_1 \neq \mu_2$ There is a difference between the two population means.

2) The level of significance is 5%. This is a two-sided test so alpha is split into two sides. Computing degrees of freedom using the equation above gives 105 degrees of freedom.

$$df = \frac{(\frac{9.4^2}{51} + \frac{10.7^2}{56})^2}{\frac{1}{51-1}(\frac{9.4^2}{51})^2 + \frac{1}{56-1}(\frac{10.7^2}{56})^2} = 104.9$$

The critical value ($(t_{\frac{\alpha}{2}}$, based on 100 degrees of freedom (closest value in the t-table), is ±1.984. Using 50 degrees of freedom, the critical value is ±2.009.

3) The test statistic is

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(39.6 - 43.9) - (0)}{\sqrt{\frac{9.4^2}{51} + \frac{10.7^2}{56}}} = -2.213$$

4) The test statistic falls in the rejection zone.

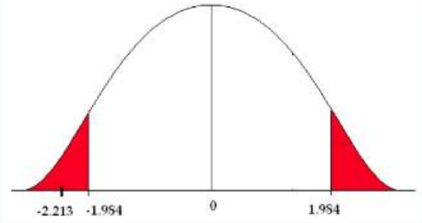


Figure 4.1.1: A comparison of the critical values and test statistic.

We reject the null hypothesis. We have enough evidence to support the claim that there is a difference in the mean number of cavity trees between the Adirondack Park and the Monongahela National Forest.

Construct and Interpret a Confidence Interval about the Difference of Two Independent Means

A hypothesis test will answer the question about the difference of the means. BUT, we can answer the same question by constructing a confidence interval about the difference of the means. This process is just like the confidence intervals from Chapter 2.

1. Find the critical value.
2. Compute the margin of error.
3. Point estimate ± margin of error.

Because we are working with two samples, we must modify the components of the confidence interval to incorporate the information from the two populations.

- The point estimate is $\bar{x}_1 - \bar{x}_2$.
- The standard error comes from the test statistic $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- The critical value $t_{\frac{\alpha}{2}}$ comes from the student's t-table.

The confidence interval takes the form of the point estimate plus or minus the standard error of the differences.

$$\bar{x}_1 - \bar{x}_2 \pm t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{4.1.3}$$

We will use the same three steps to construct a confidence interval about the difference of the means.

1. critical value $t_{\frac{\alpha}{2}}$
2. $E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
3. $\bar{x}_1 - \bar{x}_2 \pm E$

✓ Example 4.1.2:

Let's look at the mean number of cavity trees in old growth stands again. The forester wants to know if there is a difference between the mean number of cavity trees in old growth stands in the Adirondack forests and in the Monongahela Forest. We can answer this question by constructing a confidence interval about the difference of the means.

1) $t_{\frac{\alpha}{2}}$ = 2.009

2) $E = t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 2.009 \sqrt{\frac{9.4^2}{51} + \frac{10.7^2}{56}} = 3.904$

3) $\bar{x}_1 - \bar{x}_2 \pm 3.904$

The 95% confidence interval for the difference of the means is (-8.204, -0.396).

We can be 95% confident that this interval contains the mean difference in number of cavity trees between the two locations. BUT, this doesn't answer the question the forester asked. Is there a difference in the mean number of cavity trees between the Adirondack and Monongahela forests? To answer this, we must look at the confidence interval interpretations.

Confidence Interval Interpretations

- If the confidence interval contains all positive values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly greater than the mean of the second group.
- If the confidence interval contains all negative values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly less than the mean of the second group.
- If the confidence interval contains zero (it goes from negative to positive values), we find NO significant difference between the groups.

In this problem, the confidence interval is (-8.204, -0.396). We have all negative values, so we can conclude that there is a significant difference in the mean number of cavity trees AND that the mean number of cavity trees in the Adirondack forests is significantly less than the mean number of cavity trees in the Monongahela Forest. The confidence interval gives an estimate of the mean difference in number of cavity trees between the two forests. There are, on average, 0.396 to 8.204 fewer cavity trees in the Adirondack Park than the Monongahela Forest.

P-value Approach

We can also use the p-value approach to answer the question. Remember, the p-value is the area under the normal curve associated with the test statistic. This example is a two-sided test (H1: $\mu1 \neq \mu2$) so the p-value, when computed by hand, will be multiplied by two.

The test statistic equals -2.213, so the p-value is two times the area to the left of -2.213. We can only estimate the p-value using the student's t-table. Using the lesser of n1– 1 or n2– 1 as the degrees of freedom, we have 50 degrees of freedom. Go across the 50 row in the student's t-table until you find the absolute value of the test statistic. In this case, 2.213 falls between 2.109 and 2.403. Going up to the top of each of those columns gives you the estimate of the p-value (between 0.02 and 0.01).

Table 4.1.2. Student t-Distribution Area in Right Tail

df	.05	.025	.02	.01	.005
39	1.686	2.024	2.127	2.429	2.712
40	1.684	2.021	2.123	2.423	2.704
50	1.676	2.009	2.109	2.403	2.678
60	1.671	2.000	2.099	2.390	2.660
70	1.667	1.994	2.093	2.381	2.648

The p-value is 2x(0.01 – 0.02) = ($0.02 < p < 0.04$). The p-value is greater than 0.02 but less than 0.04. This is less than the level of significance (0.05), so we reject the null hypothesis. There is enough evidence to support the claim that there is a significant difference in the mean number of cavity trees between the areas.

✓ Example 4.1.3:

Researchers are studying the relationship between logging activities in the northern forests and amphibian habitats. They were comparing moisture levels between old-growth and post-harvest habitats. The researchers believe that post-harvest habitat has a lower moisture level. They collected data on moisture levels from two independent random samples. Test their claim using a 5% level of significance.

Old Growth	**Post Harvest**
n1 = 26	n2 = 31
6313.png=0.62 g/cm3	6320.png= 0.56 g/cm3
s1 = 0.12 g/cm3	s2 = 0.17 g/cm3

H0: $\mu1 = \mu2$ or $\mu1 – \mu2 = 0$. There is no difference between the two population means.

H1: $\mu1 > \mu2$. Mean moisture level in old growth forests is greater than post-harvest levels.

We will use the critical value based on the lesser of n1– 1 or n2– 1 degrees of freedom. In this problem, there are 25 degrees of freedom and the critical value is 1.708. Now compute the test statistic.

$$t=\frac{(0.62-0.56)-0}{\sqrt{\frac{0.12^2}{26}+\frac{0.17^2}{31}}}=1.556 \tag{4.1.4}$$

The test statistic does not fall in the rejection zone. We fail to reject the null hypothesis. There is not enough evidence to support the claim that the moisture level is significantly lower in the post-harvest habitat.

Now answer this question by constructing a 90% confidence interval about the difference of the means.

1) $t_{\frac{\alpha}{2}}$ = 1.708

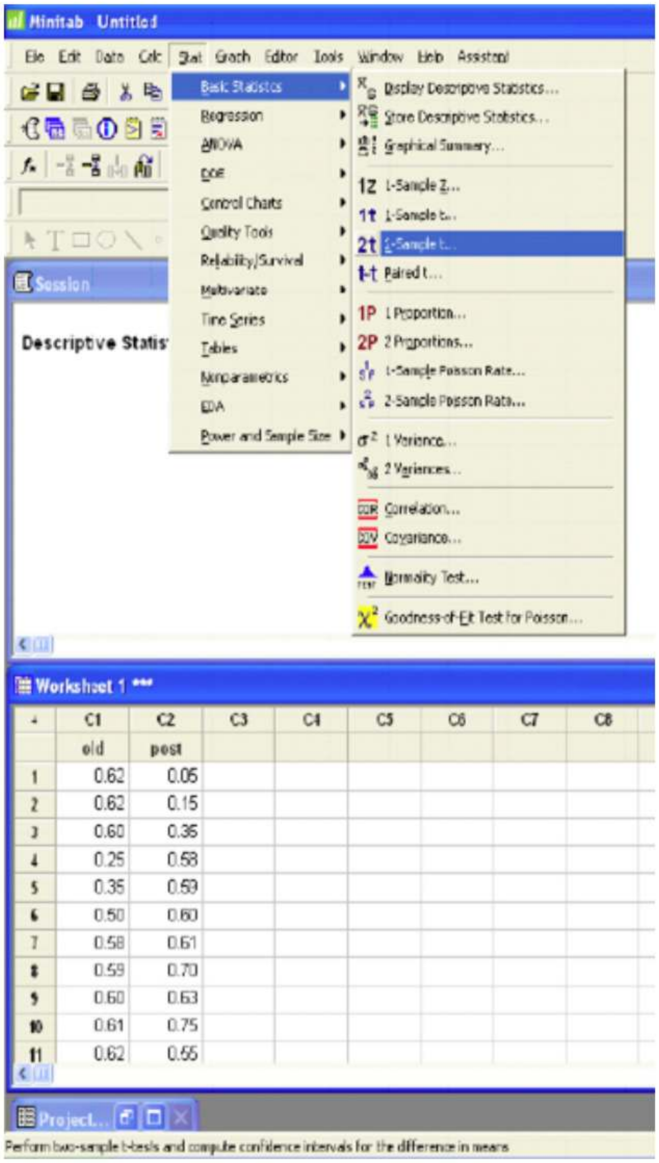
2) E = $t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=1.708\sqrt{\frac{0.12^2}{26}+\frac{0.17^2}{31}}=0.0658$

3) $\bar{x_1}-\bar{x_2}\pm E=(0.62-0.56)\pm 0.0658$

The 90% confidence interval for the difference of the means is (-0.0058, 0.1258). The values in the confidence interval run from negative to positive indicating that there is no significant different in the mean moisture levels between old growth and post-harvest stands.

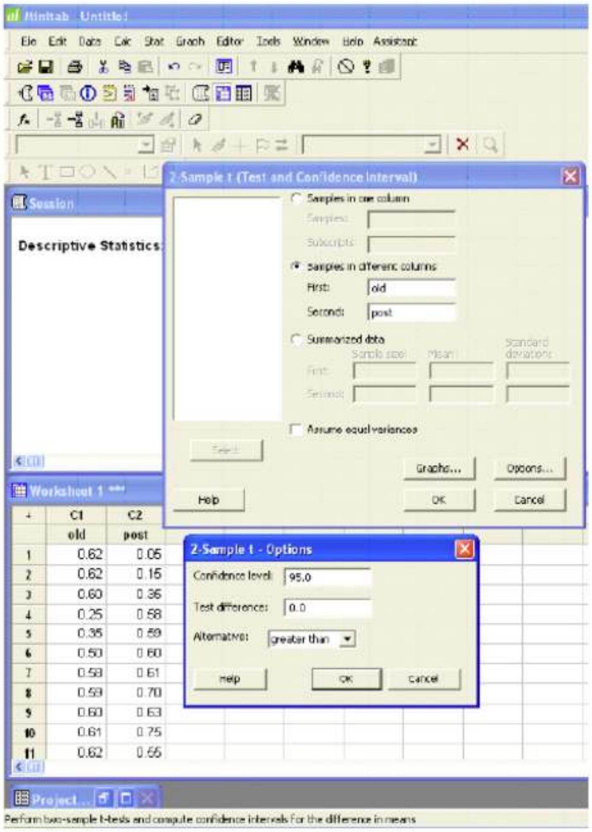
Software Solutions

Minitab



↓	C1	C2	C3	C4	C5	C6	C7	C8
	old	post						
1	0.62	0.05						
2	0.62	0.15						
3	0.60	0.35						
4	0.25	0.58						
5	0.35	0.59						
6	0.50	0.60						
7	0.58	0.61						
8	0.59	0.70						
9	0.60	0.63						
10	0.61	0.75						
11	0.62	0.55						

Perform two-sample t-tests and compute confidence intervals for the difference in means



○ Samples in one column

● Samples in different columns

First: old

Second: post

○ Summarized data

☐ Assume equal variances

Confidence level: 95.0

Test difference: 0.0

Alternative: greater than

+	C1	C2
	old	post
1	0.62	0.05
2	0.62	0.15
3	0.60	0.36
4	0.25	0.58
5	0.35	0.59
6	0.50	0.60
7	0.58	0.61
8	0.59	0.70
9	0.60	0.63
10	0.61	0.75
11	0.62	0.65

Perform two-sample t-tests and compute confidence intervals for the difference in means

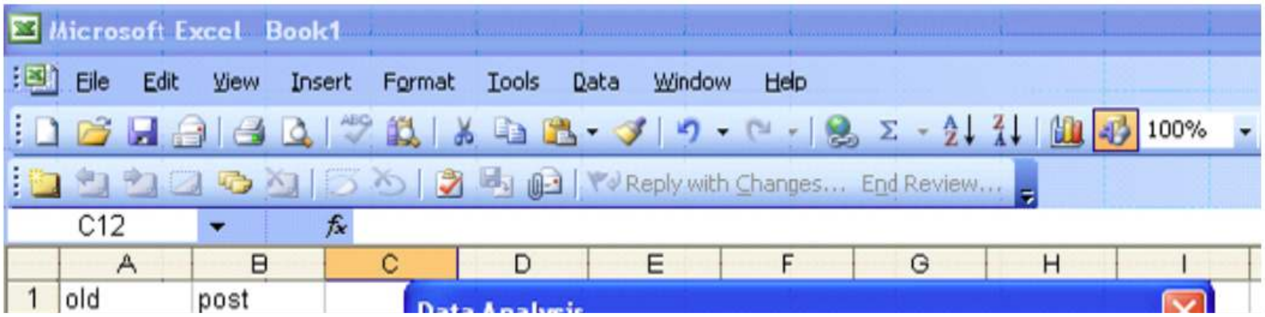
Two-Sample T-Test and CI: for old vs. post

	N	Mean	StDev	SE Mean
old	26	0.620	0.121	0.024
post	31	0.559	0.172	0.031
Difference = $\mu_{(old)} - \mu_{(post)}$				
Estimate for difference: 0.0603				
95% lower bound for difference: -0.0049				
T-Test of difference = 0 (vs >): T-Value = 1.55 p-Value = 0.064 DF = 53				

The p-value (0.064) is greater than the level of confidence so we fail to reject the null hypothesis.

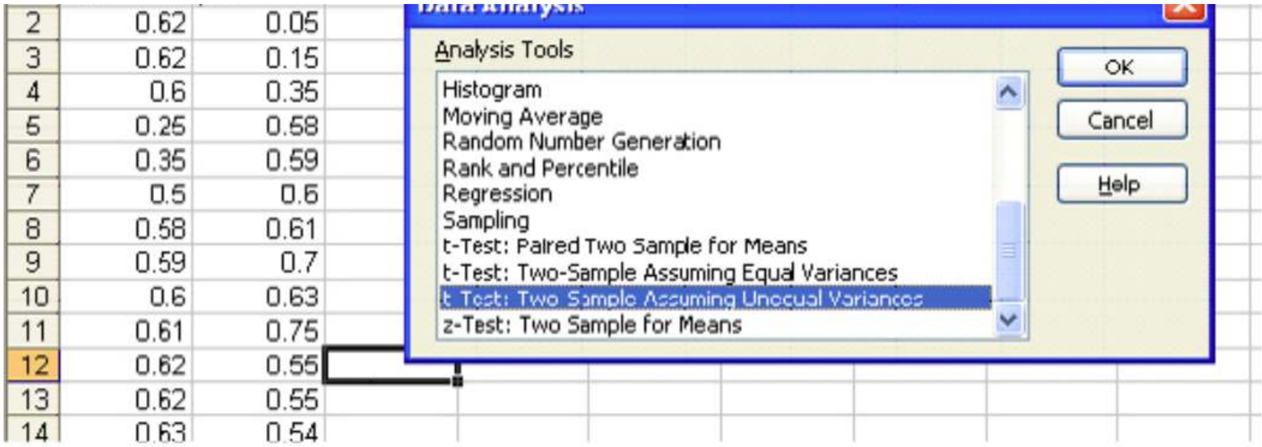
Additional example: www.youtube.com/watch?v=7pIb-GVixFo.

Excel

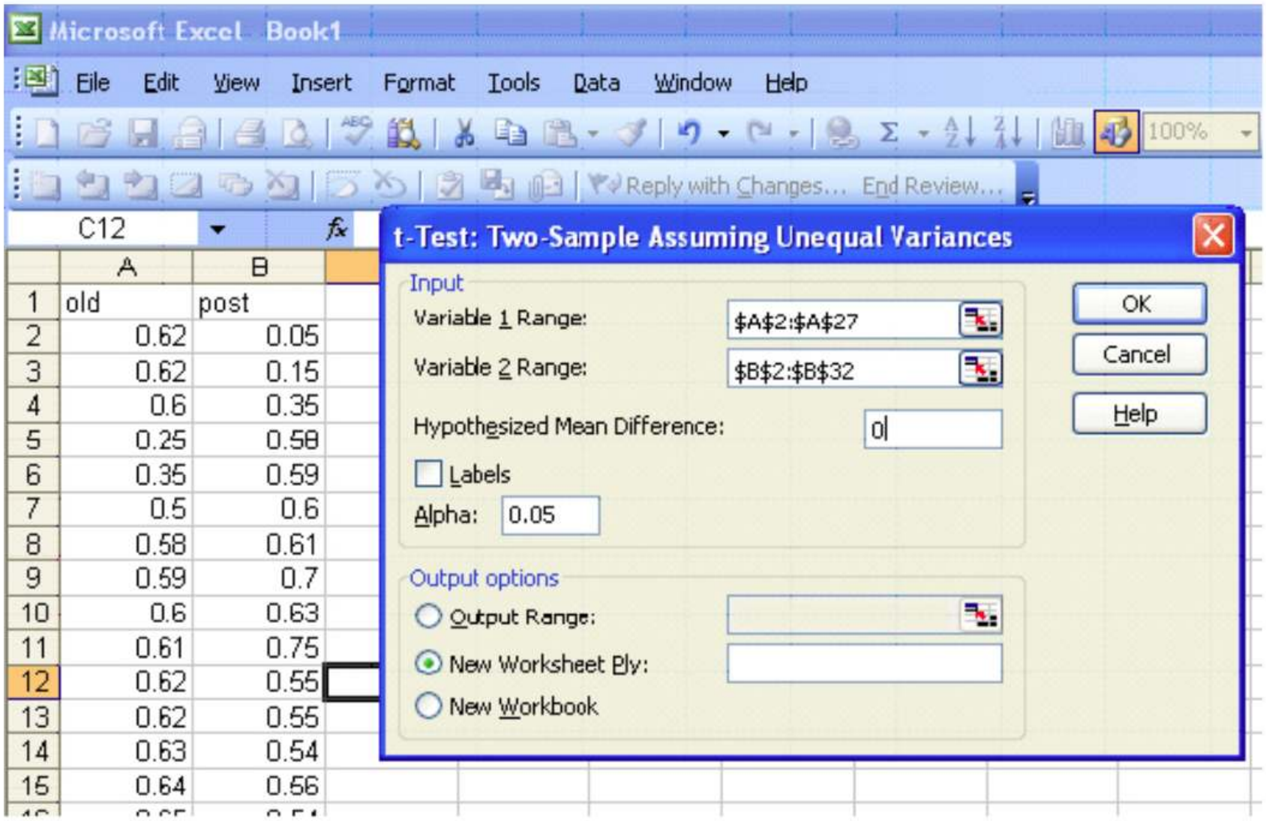

Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
100%
Reply with Changes... End Review...
A B C D E F G H I
1
Data Analysis

C12

	A	B
1	old	post


Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

2	0.62	0.05
3	0.62	0.15
4	0.6	0.35
5	0.25	0.58
6	0.35	0.59
7	0.5	0.6
8	0.58	0.61
9	0.59	0.7
10	0.6	0.63
11	0.61	0.75
12	0.62	0.55
13	0.62	0.55
14	0.63	0.54



	A	B
1	old	post
2	0.62	0.05
3	0.62	0.15
4	0.6	0.35
5	0.25	0.58
6	0.35	0.59
7	0.5	0.6
8	0.58	0.61
9	0.59	0.7
10	0.6	0.63
11	0.61	0.75
12	0.62	0.55
13	0.62	0.55
14	0.63	0.54
15	0.64	0.56

Variable 1 Range: $A$2:$A$27

Variable 2 Range: $B$2:$B$32

Hypothesized Mean Difference: 0

☐ Labels

Alpha: 0.05

○ Output Range:

● New Worksheet Ply:

○ New Workbook

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	Variable 2
Mean	0.619615	0.559355
Variance	0.014708	0.02948
Observations	26	31
Hypothesized Mean Difference	0	

	Variable 1	Variable 2
df	54	
t Stat	1.557361	
$P(T \leq t)$ one-tail	0.063809	
t Critical one-tail	1.673565	
$P(T \leq t)$ two-tail	0.127617	
t Critical two-tail	2.004879	

The one-tail p-value (0.063809) is greater than the level of significance, therefore, we fail to reject the null hypothesis.

4.2: Pooled Two-sampled t-test (Assuming Equal Variances)

Pooled Two-sampled t-test (Assuming Equal Variances)

In the previous section, we made the assumption of unequal variances between our two populations. Welch's t-test statistic does not assume that the population variances are equal and can be used whether the population variances are equal or not. The test that assumes equal population variances is referred to as the *pooled t-test*. Pooling refers to finding a weighted average of the two independent sample variances.

The pooled test statistic uses a weighted average of the two sample variances.

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = (\frac{n_1-1}{n_1+n_2-2})S_1^2+(\frac{n_2-1}{n_1+n_2-2})S_2^2 \tag{4.2.1}$$

If $n_1 = n_2$, then $S_p^2 = (\frac{1}{2}s_1^2 + (1/2)s_2^2$, the average of the two sample variances. But whenever n1≠n2, the s^2 based on the larger sample size will receive more weight than the other s^2.

The advantage of this test statistic is that it exactly follows the student's t-distribution with n1+ n2– 2 degrees of freedom.

$$t = \frac{\bar{X}_1-\bar{X}_2}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{\bar{X}_1-\bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{4.2.2}$$

The hypothesis test procedure will follow the same steps as the previous section.

It may be difficult to verify that two population variances might be equal based on sample data. The F-test is commonly used to test variances but is not robust. Small departures from normality greatly impact the outcome making the results of the F-test unreliable. It can be difficult to decide if a significant outcome from an F-test is due to the differences in variances or non-normality. Because of this, many researchers rely on Welch's *t* when comparing two means.

✓ Example 4.2.1:

Growth of pine seedlings in two different substrates was measured. We want to know if growth was better in substrate 2. Growth (in cm/yr) was measured and included in the table below. α = 0.05

Substrate 1	Substrate 2
3.2	4.5
4.5	6.2
3.8	5.8
4.0	6.0
3.7	7.1
3.2	6.8
4.1	7.2

Solution

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

$$S_p^2 = \frac{(7-1)0.474^2+(7-1)0.936^2}{7+7-2} = 0.55 \tag{4.2.3}$$

$$t = \frac{3.79-6.23}{\sqrt{0.55(\frac{1}{7}+\frac{1}{7})}} = \frac{-2.44}{0.396} = -6.16 \tag{4.2.4}$$

This is a one-sided test with $n_1 + n_2 - 2 = 12$ degrees of freedom. The critical value is -1.782. The test statistic is less than the critical value so we will reject the null hypothesis.

There is enough evidence to support the claim that the mean growth is less in substrate 1. Growth in substrate 2 is greater.

The confidence interval approach also uses the pooled variance and takes the form:

$$(\bar{x_1} - \bar{x_2}) \pm t_{\frac{\alpha}{2}} \sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})} \tag{4.2.5}$$

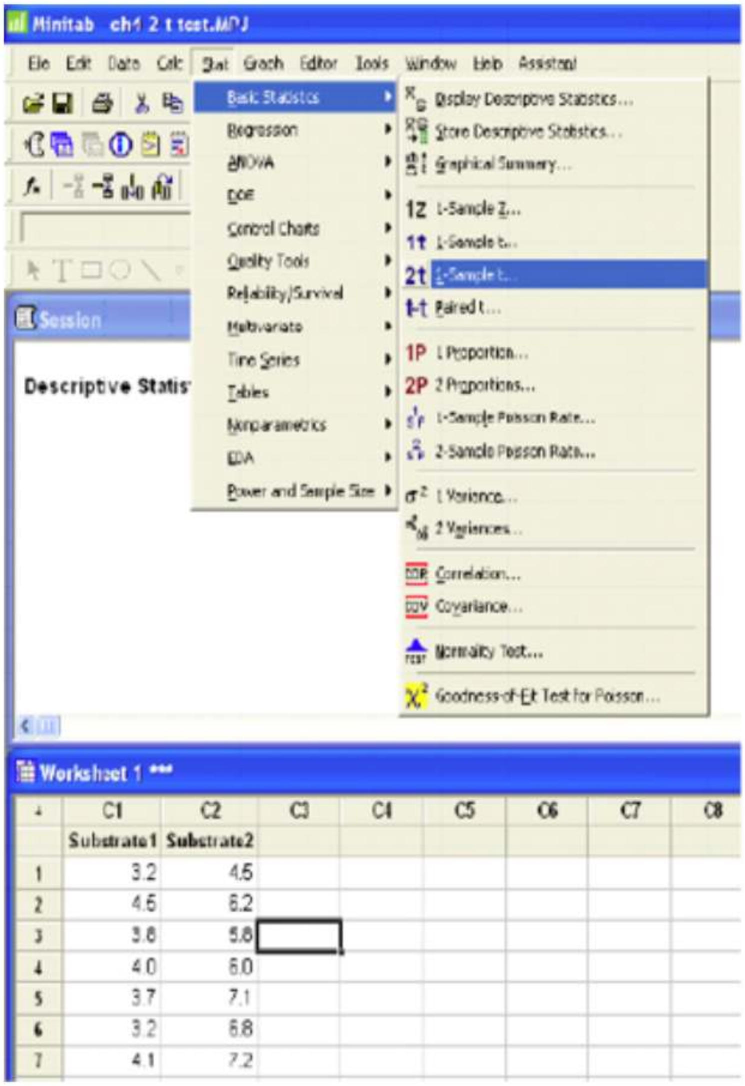
using $n_1 + n_2 - 2$ degrees of freedom. So let's answer the same question with a 90% confidence interval.

$$(3.79 - 6.23) \pm 1.782 \sqrt{0.55 (\frac{1}{7} + \frac{1}{7})} = (-2.44 \pm 0.7064) = (-3.146, -1.734) \tag{4.2.6}$$

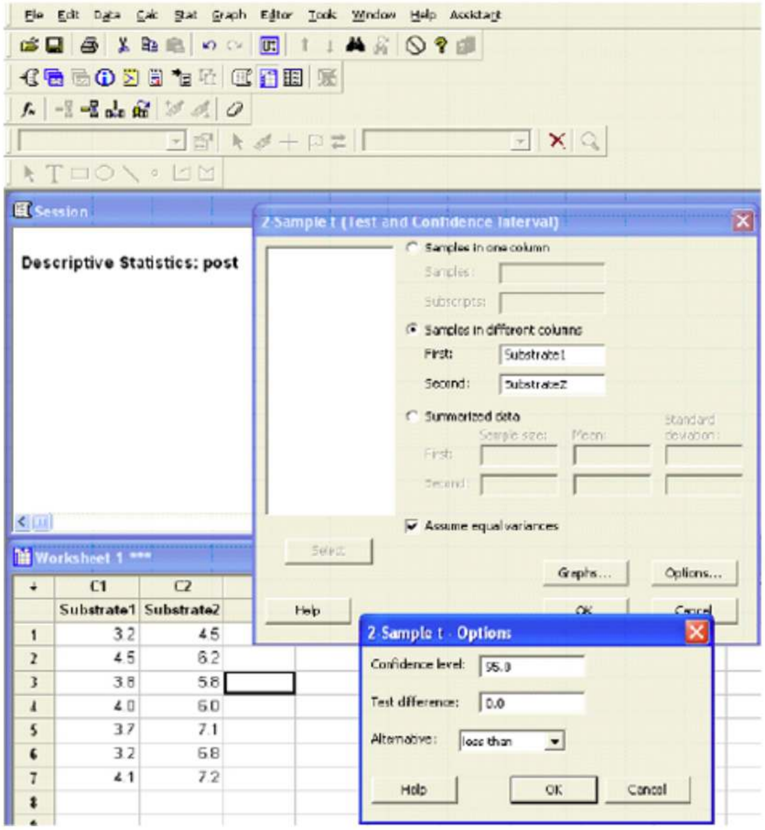
All negative values tell you that there is a significant difference between the mean growth for the two substrates and that the growth in substrate 1 is significantly lower than the growth in substrate 2 with reduction in growth ranging from 1.734 to 3.146 cm/yr.

Software Solutions

Minitab



↓	C1	C2	C3	C4	C5	C6	C7	C8
	Substrate1	Substrate2						
1	3.2	4.5						
2	4.5	6.2						
3	3.8	5.8						
4	4.0	6.0						
5	3.7	7.1						
6	3.2	6.8						
7	4.1	7.2						


File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Worksheet 1 ***
2-Sample t (Test and Confidence Interval)
Select
Graphs...
Options...
Help
OK
Cancel
2-Sample t - Options
Help
OK
Cancel

Descriptive Statistics: post

↓	C1	C2
	Substrate1	Substrate2
1	3.2	4.5
2	4.5	6.2
3	3.8	5.8
4	4.0	6.0
5	3.7	7.1
6	3.2	6.8
7	4.1	7.2
8		

○ Samples in one column
Samples:
Subscripts:
● Samples in different columns
First: Substrate1
Second: Substrate2
○ Summarized data
Sample size: Mean: Standard deviation:
First:
Second:
☑ Assume equal variances

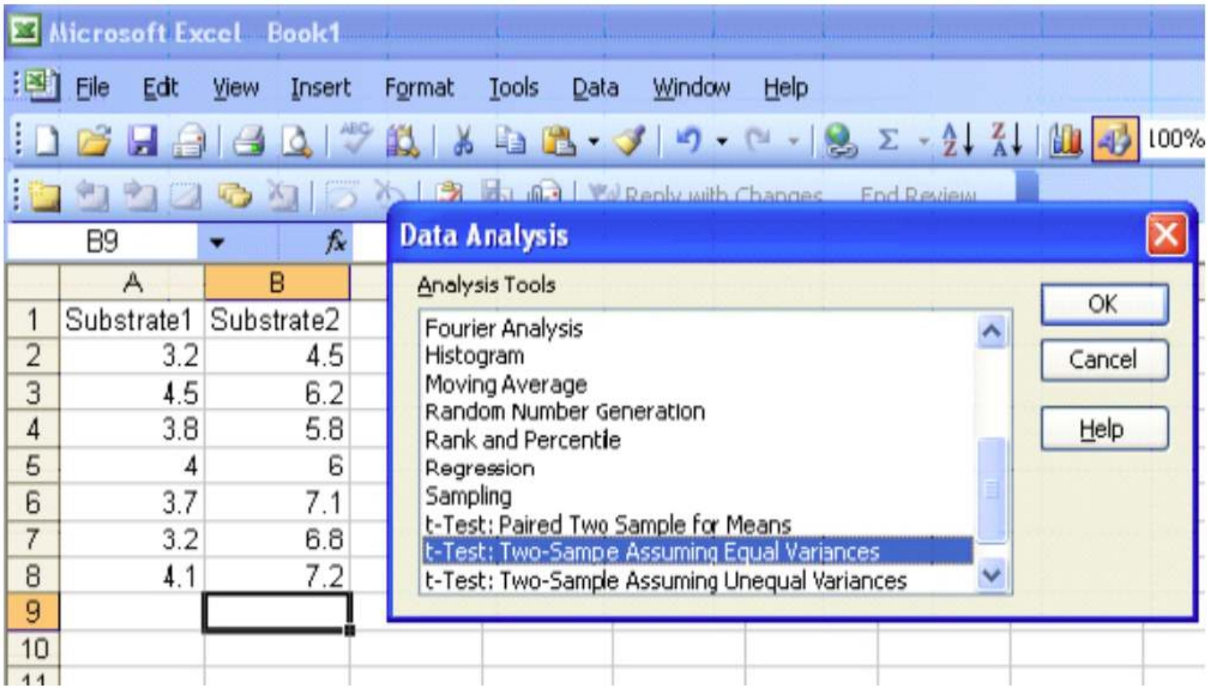
Confidence level: 95.0
Test difference: 0.0
Alternative: less than

Two-Sample T-Test and CI: Substrate1, Substrate2

Two-sample T for Substrate1 vs. Substrate2				
	N	Mean	StDev	SE Mean
Substrate1	7	3.786	0.474	0.18
Substrate2	7	6.229	0.936	0.35
Difference = mu (Substrate1) – mu (Substrate2)				
Estimate for difference: -2.443				
95% upper bound for difference: -1.736				
T-Test of difference = 0 (vs <): T-Value = -6.16 p-value = 0.000 DF = 12				
Both use Pooled StDev = 0.7418				

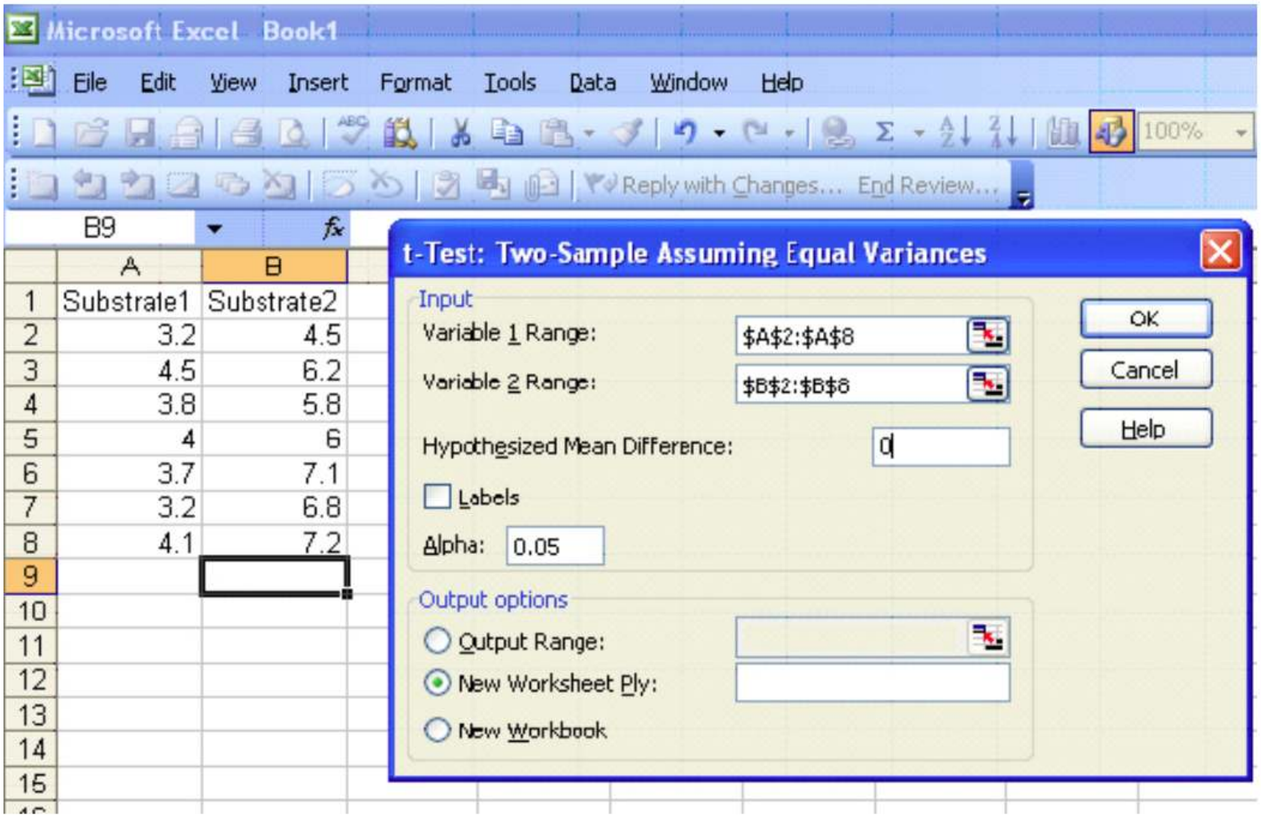
The p-value (0.000) is less than the level of significance (0.05). We will reject the null hypothesis.

Excel



	A	B
1	Substrate1	Substrate2
2	3.2	4.5
3	4.5	6.2
4	3.8	5.8
5	4	6
6	3.7	7.1
7	3.2	6.8
8	4.1	7.2
9		
10		

Analysis Tools: t-Test: Two-Sampe Assuming Equal Variances


Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
100%
Reply with Changes... End Review...
B9
t-Test: Two-Sample Assuming Equal Variances
Input
Output options
Output Range:
New Workbook
OK
Cancel
Help

	A	B
1	Substrate1	Substrate2
2	3.2	4.5
3	4.5	6.2
4	3.8	5.8
5	4	6
6	3.7	7.1
7	3.2	6.8
8	4.1	7.2
9		
10		
11		
12		
13		
14		
15		

Variable 1 Range: $A$2:$A$8

Variable 2 Range: $B$2:$B$8

Hypothesized Mean Difference: 0

☐ Labels

Alpha: 0.05

● New Worksheet Ply:

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	Variable 2
Mean	3.785714	6.228571
Variance	0.224762	0.875714
Observations	7	7
Pooled Variance	0.550238	
Hypothesized Mean Difference	0	
df	12	
t Stat	-6.16108	
P$(T \leq t)$ one-tail	2.43E-05	
t Critical one-tail	1.782288	
P$(T \leq t)$ two-tail	4.86E-05	
t Critical two-tail	2.178813	

This is a one-sided test (greater than) so use the P$(T \leq t)$ one-tail value 2.43E-05. The p-value (0.0000243) is less than the level of significance (0.05). We will reject the null hypothesis.

4.3: Inferences about Two Means with Dependent Samples—Matched Pairs

Inferences about Two Means with Dependent Samples—Matched Pairs

Dependent samples occur when there is a relationship between the samples. The data consists of matched pairs from random samples. A sampling method is dependent when the values selected for one sample are used to determine the values in the second sample. Before and after measurements on a population, such as people, lakes, or animals are an example of dependent samples. The objects in your sample are measured twice; measurements are taken at a certain point in time, and then re-taken at a later date. Dependency also occurs when the objects are related, such as eyes or tires on a car. Pairing isn't a problem; it's an opportunity to use the information that occurs with both measurements.

Before you begin your work, you must decide if your samples are dependent. If they are, you can take advantage of this fact. You can use this matching to better answer your research questions. Pairing data reduces measurement variability, which increases the accuracy of our statistical conclusions.

We use the difference (the subtraction) of the pairs of data in our analysis. For each pair, we subtract the values:

- d_1 = before1 – after 1
- d_2 = before 2 – after 2
- d_3 = before 3 – after 3
- …

We are creating a new random variable d (differences), and it is important to keep the sign, whether positive or negative. We can compute $\bar{d}$, the sample mean of the differences, and sd, the sample standard deviation of the differences as follows:

$$\bar{d} = \frac{\sum d_i}{n} \tag{4.3.1}$$

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} \tag{4.3.2}$$

Just as we used the sample mean and the sample standard deviation in a one-sample t-test, we will use the sample mean and sample standard deviation of the differences to test for matched pairs. The assumption of normality must still be verified. The differences must be normally distributed or the sample size must be large enough (n ≥ 30).

We can do a hypothesis test using matched pairs data following the same methods we used in the previous chapter.

- Write the null and alternative hypotheses.
- State the level of significance and find the critical value.
- Compute a test statistic.
- Compare the test statistic to the critical value and state a conclusion.

Since we are using the differences between the pairs of data, we identify this in our null and alternative hypotheses: $H_0 : \mu d = 0$. The mean of the differences is equal to zero; there is no difference in "before and after" values.

We'll use the same three pairs of null and alternative hypotheses we used in the previous chapter.

Table 4.3.1. Null and alternative hypotheses.

Two-sided	Left-sided	Right=sided
H$_0$: $\mu_d = c$	H$_0$: $\mu_d =$ \c	(\mathrm{H}_{\mathrm{0}}: \mu_d=c \)
H$_1$: $\mu_d \neq c$	H$_1$: $\mu_d < c$	H$_1$: $\mu_d > c$

Table 3. Null and alternative hypotheses.

The critical value comes from the student's t-distribution table with n – 1 degrees of freedom, where n = number of matched pairs. The test statistic follows the student's t-distribution

$$t=\frac{\bar{d}-\mu_d}{(s_d/\sqrt{n})} \tag{4.3.3}$$

The conclusion must always answer the question you are asking in the alternative hypothesis.

- Reject the H_0. There is enough evidence to support the alternative claim.
- Fail to reject the H_0. There is not enough evidence to support the alternative claim.

✓ Example 4.3.1:

An environmental biologist wants to know if the water clarity in Owasco Lake is improving. Using a Secchi disk, she takes measurements in specific locations at specific dates during the course of the year. She then repeats the measurements in the same locations and on the same dates five years later. She obtains the following results:

Date	Initial Depth	5-year Depth	Difference
5/11	38	52	-14
6/7	58	60	-2
6/24	65	72	-7
7/8	74	72	2
7/27	56	54	2
8/31	36	48	-12
9/30	56	58	-2
10/12	52	60	-8

Using a 5% level of significance, test the biologist's claim that water clarity is improving.

Solution

The data are paired by date with two measurements taken at each point five years apart. We will use the differences (right column) to see if there has been a significant improvement in water clarity. Using your calculator, Minitab, or Excel, compute the descriptive statistics on the differences to get the sample mean and the sample standard deviation of the differences.

$$\bar{d}=-5.125 \tag{4.3.4}$$

$$s_d=6.081 \tag{4.3.5}$$

1) The null and alternative hypotheses:

$H_0: \mu_d=0$ (The mean of the differences is equal to zero- no difference in water clarity over time.)

$H_1: \mu_d<0$ (The water clarity is improving.)

We test "less than" because of how we computed the differences and the question we are asking.

In this case, we hope to see greater depth (better water clarity) at the five-year measurements. By calculating Initial – 5-year we hope to see negative values, values less than zero, indicating greater depth and clarity at the 5-year mark. Think of it like this:

Initial Depth < 5-year depth

This gives you the direction of the test!

2) The critical value tα.

The critical value comes from the student's t-distribution table with n – 1 degrees of freedom. In this problem, we have eight pairs of data (n = 8) with 7 degrees of freedom. This is a one-sided test (less than), so alpha is all in the left tail. Go down the 0.05 column with 7 df to find the correct critical value (tα) of -1.895.

3) The test statistic

$$t=\frac{\bar{d}-\mu_d}{s_d/\sqrt{n}}=\frac{-5.125-0}{6.081/\sqrt{8}}=-2.38 \tag{4.3.6}$$

We subtract zero from d-bar because of our null hypothesis. Our null hypothesis is that the difference of the before and after values are statistically equal to zero. In other words, there has been no change in water clarity.

4) Compare the test statistic to the critical value and state a conclusion.

The test statistic (-2.38) is less than the critical value (-1.895). It falls in the rejection zone.

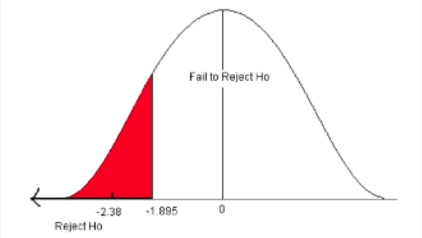


Figure 4.3.1: Comparison of the critical value and the test statistic.

We reject the null hypothesis. We have enough evidence to support the claim that the mean water clarity has improved.

P-value Approach

We can also use the p-value approach to answer the question. To estimate p-value using the student's t-table, go across the row for 7 degrees of freedom until you find the two values that the absolute value of your test statistic falls between.

Table 4. Student t-Distribution.

df	.05	.025	.02	.01	.005
5	2.015	2.571	2.757	3.365	4.032
6	1.943	2.447	2.612	3.143	3.707
7	1.895	2.365	2.517	2.998	3.499
8	1.860	2.306	2.449	2.896	3.355
9	1.833	2.262	2.398	2.821	3.250

The p-value for this test statistic is greater than 0.02 and just less than 0.025. Compare this to the level of significance (alpha). The Decision Rule says that if the p-value is less than α, reject the null hypothesis. In this case, the p-value estimate (0.02 – 0.025) is less than the level of significance (0.05). Reject the null hypothesis. We have enough evidence to support the claim that the mean water clarity has improved.

BUT, what if you used a 1% level of significance? In this case, the p-value is NOT less than the level of significance ((0.02 – 0.025)>0.01). We would fail to reject the null hypothesis. There is NOT enough evidence to support the claim that the water clarity has improved. It is important to set the level of significance at the start of your research and report the p-value. Another researcher may interpret your findings differently, based on your reported p-value and their own selected level of significance.

Construct and Interpret a Confidence Interval about the Differences of the Data for Matched Pairs

A hypothesis test for matched pairs data is very similar to a one-sample t-test. BUT, we can answer the same question by constructing a confidence interval about the mean of the differences. This process is just like the confidence intervals from Chapter 2.

1. Find the critical value.
2. Compute the margin of error.
3. Point estimate ± margin of error.

For matched pairs data, the critical value comes from the student's t-distribution with n – 1 degrees of freedom. The margin of error uses the sample standard deviation of the differences (sd) and the point estimate is $\bar{d}$, the mean of the differences.

For a (1 – α)*100% confidence interval for the mean of the differences

$$\bar{d} \pm t_{\frac{\alpha}{2}}(\frac{s_d}{\sqrt{n}}) \tag{4.3.7}$$

Where $t_{\frac{\alpha}{2}}$ is used because confidence intervals are always two-sided.

✓ Example 4.3.2:

Let's look at the biologist studying water clarity in Owasco Lake again. She wants to test the claim that water clarity has improved. We can answer this question by constructing a confidence interval about the mean of the differences.

$\bar{d}$ = -5.125	*sd* = 6.081	α = 0.05	n = 8

Solution

1) $t_{\frac{\alpha}{2}} = 2.365$

2) $E = t_{\frac{\alpha}{2}}(\frac{s_d}{\sqrt{n}}) = 2.365(\frac{6.081}{\sqrt{8}}) = 5.085$

3) $\bar{d} \pm E = -5.125 \pm 5.085$

The 95% confidence interval about the mean of the differences is

$(-10.21, -0.04)$

$(-10.21 \leq \mu_d \leq -0.04)$

We can be 95% confident that this interval contains the true mean of the differences in water clarity between the two time periods. BUT, this doesn't directly answer the question about improved water clarity. To do this, we use the interpretations given below.

Confidence Interval Interpretations

1. If the confidence interval contains all positive values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly greater than the mean of the second group.
2. If the confidence interval contains all negative values, we find a significant difference between the groups, AND we can conclude that the mean of the first group is significantly less than the mean of the second group.
3. If the confidence interval contains zero (it goes from negative to positive values), we find NO significant difference between the groups.

In this problem, the confidence interval is (-10.21, -0.04). We have all negative values, so we can conclude that there is a significant difference in the mean water clarity between the years AND…

- The mean water clarity for the initial time was significantly less than at the five-year re-measurement.
- Water clarity has improved during the five-year period. The confidence interval estimates the mean improvement.

✓ Example 4.3.3:

Biologists are studying elk migration in the western US and want to know if the four-lane interstate that was built ten years ago has disturbed elk migration to the winter feeding area. A random sample was gathered from nine wilderness districts in the winter feeding areas. These data were compared to a random sample collected from the same nine areas before the highway was built. Use a 1% level of significance to test this claim.

District	1	2	3	4	5	6	7	8	9
Before	11.6	18.7	15.9	20.6	10.1	17.4	7.2	12.2	11.7
After	10.0	21.6	13.9	22.8	11.5	16.2	8.1	10.8	9.6
d	1.6	-2.9	2.0	-2.2	-1.4	1.2	-0.9	1.4	2.1

$$\bar{d} = 0.100 \tag{4.3.8}$$

$$s_d = 1.946 \tag{4.3.9}$$

$H_0 : \mu_d = 0$

$H_1 : \mu_d \neq 0$

Determine the critical values: This is a two-sided question (alternative ≠) so the critical values are ±3.355.

Compute the test statistic:

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{0.100 - 0}{1.946/\sqrt{9}} = 0.1542 \tag{4.3.10}$$

Now compare the critical value to the test statistic and state a conclusion. The test statistic is NOT greater than 3.355 or less than -3.355 (it doesn't fall in the rejection zones). We fail to reject the null hypothesis. There is not enough evidence to support the claim that the highway has interfered with the elk migration (no difference before or after the highway).

Now construct a 99% confidence interval and answer the question.

1) $t_{\frac{\alpha}{2}}$ = 3.355

2) $E = t_{\frac{\alpha}{2}}(\frac{s_d}{\sqrt{n}}) = 3.355(1.946/\sqrt{9}) = 2.176$

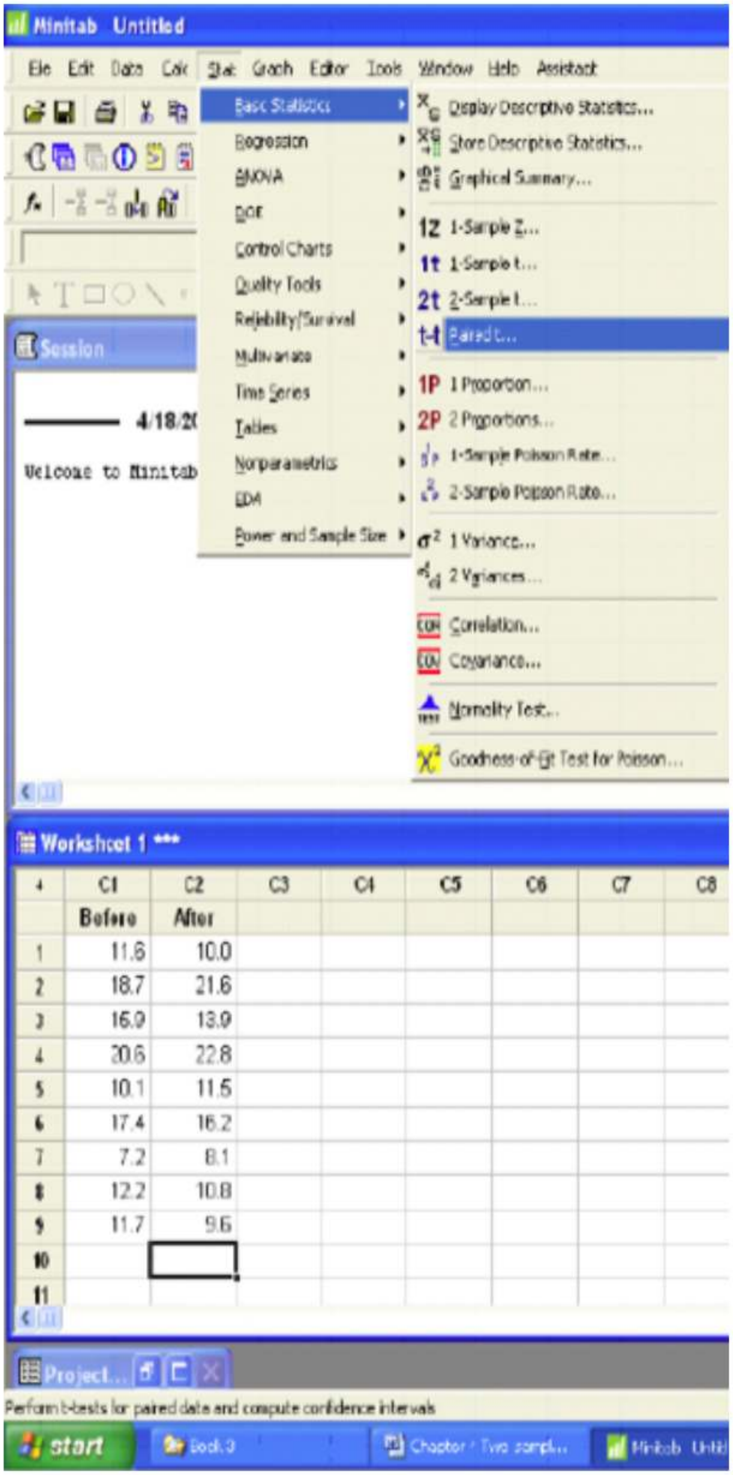
3) $\bar{d} \pm E = 0.100 \pm 2.176$

The 99% confidence interval about the difference of the means is: (-2.076, 2.276)

This confidence interval contains zero. The null hypothesis is that there is zero difference before and after the highway way was created. Therefore, we fail to reject the null hypothesis. There is not enough evidence to support the claim that the highway has interfered with the elk migration (no difference before or after the highway).

Software Solutions

Minitab

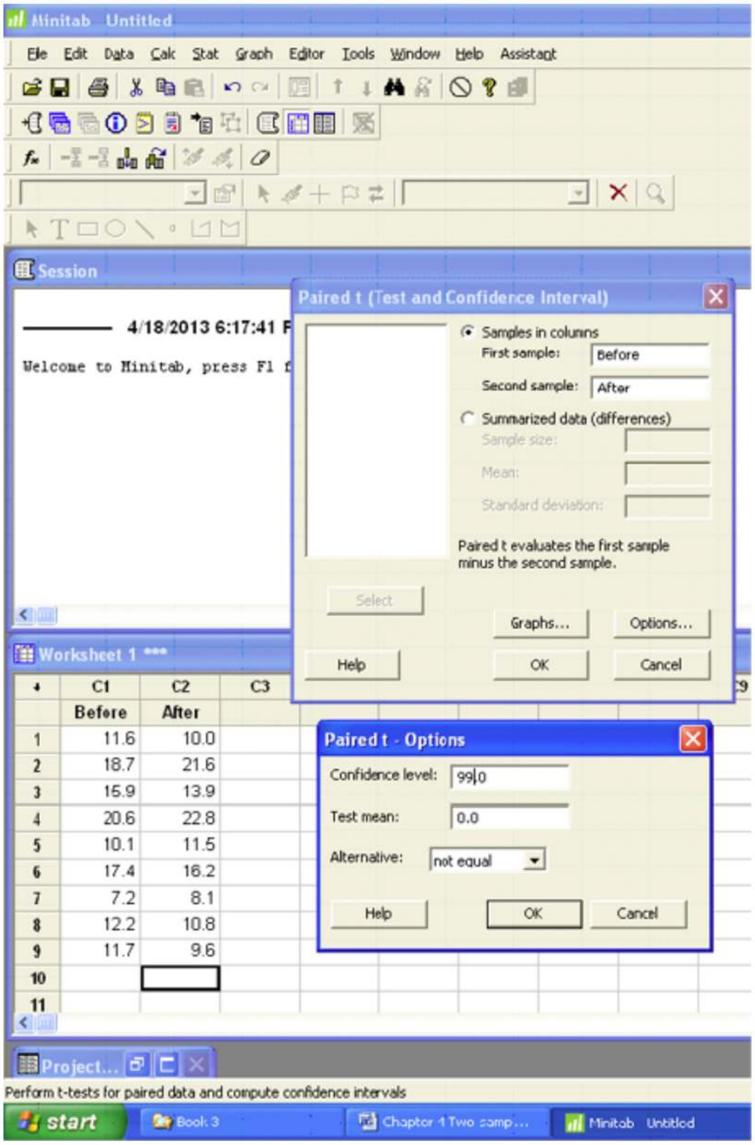


Welcome to Minitab

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7	C8
	Before	After						
1	11.6	10.0						
2	18.7	21.6						
3	16.9	13.9						
4	20.6	22.8						
5	10.1	11.5						
6	17.4	16.2						
7	7.2	8.1						
8	12.2	10.8						
9	11.7	9.6						
10								
11								

Perform t-tests for paired data and compute confidence intervals



4/18/2013 6:17:41 F

Welcome to Minitab, press F1 f

● Samples in columns
First sample: Before
Second sample: After
○ Summarized data (differences)
Sample size:
Mean:
Standard deviation:

Paired t evaluates the first sample minus the second sample.

Confidence level: 99.0
Test mean: 0.0
Alternative: not equal

	C1	C2	C3
	Before	After	
1	11.6	10.0	
2	18.7	21.6	
3	15.9	13.9	
4	20.6	22.8	
5	10.1	11.5	
6	17.4	16.2	
7	7.2	8.1	
8	12.2	10.8	
9	11.7	9.6	
10			
11			

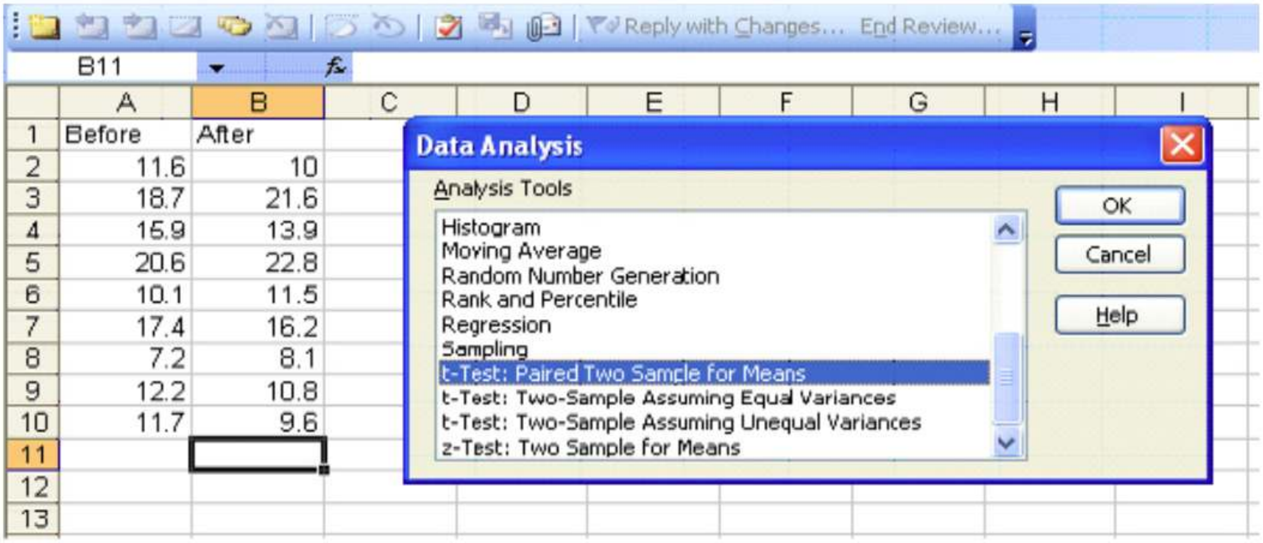
Perform t-tests for paired data and compute confidence intervals

Paired T-Test and CI: Before, After

	N	Mean	StDev	SE Mean
Before	9	13.93	4.42	1.47
After	9	13.83	5.32	1.77
Difference	9	0.100	1.946	0.649
99% CI for mean difference: (-2.077, 2.277)				
T-Test of mean difference = 0 (vs not = 0): T-Value = 0.15 p-value = 0.881				

Minitab gives the test statistic of 0.15 and the p-value of 0.881. It also gives a 99% confidence interval for the difference of the means (-2.077, 2.277). All results support failing to reject the null hypothesis.

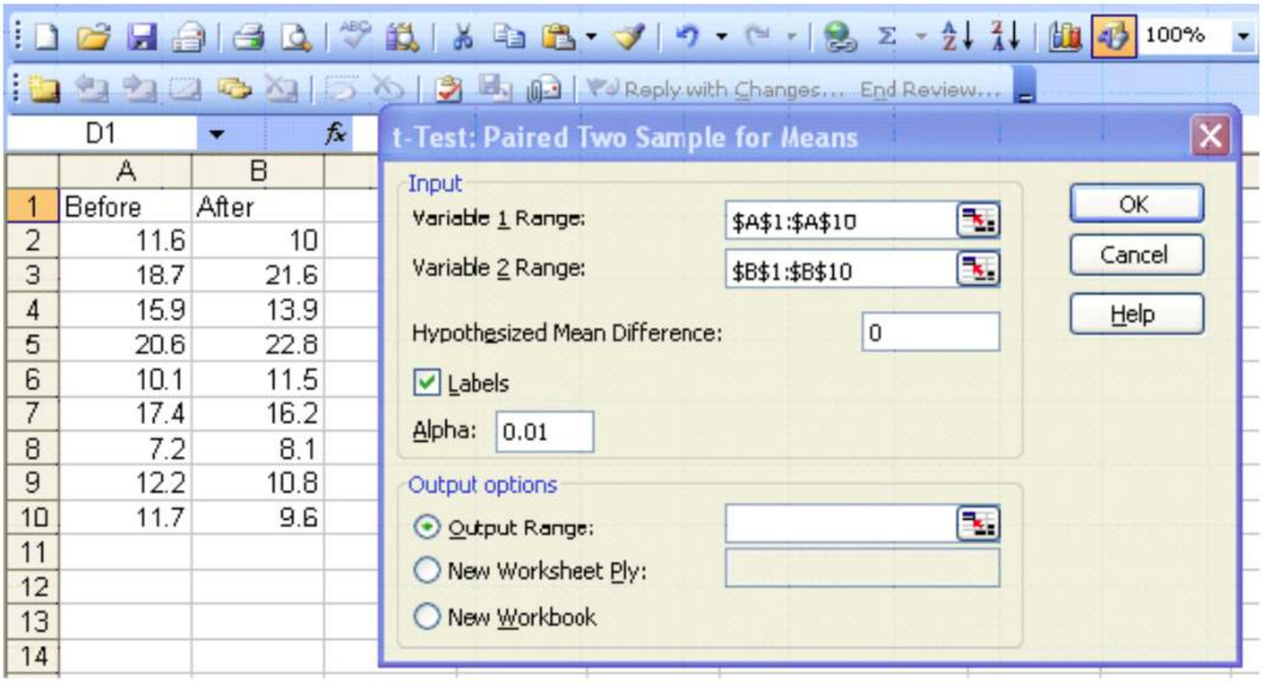
Excel



	A	B
1	Before	After
2	11.6	10
3	18.7	21.6
4	15.9	13.9
5	20.6	22.8
6	10.1	11.5
7	17.4	16.2
8	7.2	8.1
9	12.2	10.8
10	11.7	9.6

Analysis Tools

- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances
- z-Test: Two Sample for Means


Reply with Changes... End Review...
100%
D1
t-Test: Paired Two Sample for Means
OK
Cancel
Help

	A	B
1	Before	After
2	11.6	10
3	18.7	21.6
4	15.9	13.9
5	20.6	22.8
6	10.1	11.5
7	17.4	16.2
8	7.2	8.1
9	12.2	10.8
10	11.7	9.6

Input

Variable 1 Range: $A$1:$A$10

Variable 2 Range: $B$1:$B$10

Hypothesized Mean Difference: 0

☑ Labels

Alpha: 0.01

Output options

● Output Range:

○ New Worksheet Ply:

○ New Workbook

t-Test: Paired Two Sample for Means

	Before	After
Mean	13.93333	13.83333333
Variance	19.565	28.3075
Observations	9	9
Pearson Correlation	0.936635	

	Before	After
Hypothesized Mean Difference	0	
df	8	
t Stat	0.15415	
$P(T \leq t)$ one-tail	0.440654	
t Critical one-tail	2.896459	
$P(T \leq t)$ two-tail	0.881309	
t Critical two-tail	3.355387	

The test statistic is 0.15415. This is a two-sided question so we can use $P(T \leq t)$ two-tail = 0.881309. The p-value is NOT less than the 1% level of significance so we will fail to reject the null hypothesis.

4.4: Inferences about Two Population Proportions

Inferences about Two Population Proportions

We can apply the same methods we just learned with means to our two-sample proportion problems. We have two populations with two samples and we want to compare the population proportions.

- Is the proportion of lakes in New York with invasive species different from the proportion of lakes in Michigan with invasive species?
- Is the proportion of construction companies using certified lumber greater in the northeast than in the southeast?

A test of two population proportions is very similar to a test of two means, except that the parameter of interest is now "*p*" instead of "μ". With a one-sample proportion test, we used $\hat{p} = \frac{x}{n}$ as the point estimate of *p*. We expect that $\hat{p}$ would be close to *p*. With a test of two proportions, we will have two $\hat{p}$'s, and we expect that $(\hat{p}1 – \hat{p}2)$ will be close to $(p1 – p2)$. The test statistic accounts for both samples.

- With a one-sample proportion test, the test statistic is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{4.4.1}$$

and it has an approximate standard normal distribution.

- For a two-sample proportion test, we would expect the test statistic to be

$$z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \tag{4.4.2}$$

HOWEVER, the null hypothesis will be that $p1 = p2$. Because the H0 is assumed to be true, the test assumes that $p1 = p2$. We can then assume that $p1 = p2$ equals *p*, a common population proportion. We must compute a pooled estimate of *p* (its unknown) using our sample data.

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{4.4.3}$$

The test statistic then takes the form of

$$z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \tag{4.4.4}$$

The hypothesis test follows the same steps that we have seen in previous sections:

- State the null and alternative hypotheses
- State the level of significance and determine the critical value
- Compute the test statistic
- Compare the critical value and the test statistic and state a conclusion

The assumptions that we set for a one-sample proportion test still hold true for both samples. Both must be random samples from normally distributed populations satisfying the following statements:

- $n(p)(1-p) \geq 10$
- Each sample size is no more than 5% of the population size.

We can again use the same three pairs of null and alternative hypotheses. Notice that we are working with population proportions so the parameter is *p*.

Table 4.4.1. Null and alternative hypotheses.

Two-sided	Left-sided	Right=sided
$\mathrm{H}_0 : p_1 = p_2$	$\mathrm{H}_0 : p_1 = p_2$	(\mathrm{H}_{\mathrm{0}}: p_1=p_2 \)

H$_1$: \p$_1 \neq p_2$	H$_1$: \p$_1 < p_2$	H$_1$: \p$_1 > p_2$

The critical value comes from the standard normal table and depends on the alternative hypothesis (is the question one- or two-sided?). As usual, you must state a conclusion. You must always answer the question that is asked in the alternative hypothesis.

✓ Example 4.4.1:

A researcher believes that a greater proportion of construction companies in the northeast are using certified lumber in home construction projects compared to companies in the southeast. She collected a random sample of 173 companies in the southeast and found that 86 used at least 30% certified lumber. She collected another random sample of 115 companies from the northeast and found that 68 used at least 30% certified lumber. Test the researcher's claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to the southeast. α = 0.05.

Southeast	Northeast
$n_1 = 173$	$n_2 = 115$
$x_1 = 86$	$x_2 = 68$

Solution

Write the null and alternative hypotheses:

$H_0 : p_1 = p_2$ or $p_1 - p_2 = 0$

$H_1 : p_1 < p_2$

The critical value comes from the standard normal table. It is a one-sided test, so alpha is all in the left tail. The critical value is -1.645.

Compute the point estimates

$$\hat{p_1} = \frac{86}{173} = 0.497 \tag{4.4.5}$$

$$\hat{p_2} = \frac{68}{115} = 0.591 \tag{4.4.6}$$

Now compute $\bar{p}$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{86 + 68}{173 + 115} = 0.535 \tag{4.4.7}$$

The test statistic is

$$z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.497 - 0.591) - 0}{\sqrt{\frac{0.535(1-0.535)}{173} + \frac{0.535(1-0.535)}{115}}} = -1.57 \tag{4.4.8}$$

Now compare the critical value to the test statistic and state a conclusion.

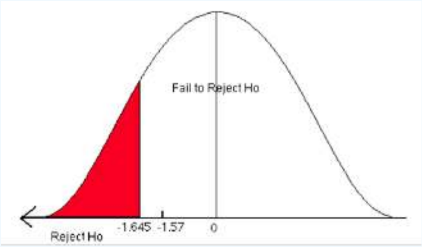


Figure 4.4.1: A comparison of the critical value and the test statistic.

We fail to reject the null hypothesis. There is not enough evidence to support the claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast.

Using the P-Value Approach

We can also answer this question using the p-value approach. The p-value is the area associated with the test statistic. This is a left-tailed problem with a test statistic of -1.57 so the p-value is the area to the left of -1.57. Look up the area associated with the Z-score -1.57 in the standard normal table.

The p-value is 0.0582.

The hatched area (p-value) is greater than the 5% level of significance (red area). We fail to reject the null hypothesis. There is not enough statistical evidence to support the claim that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast.

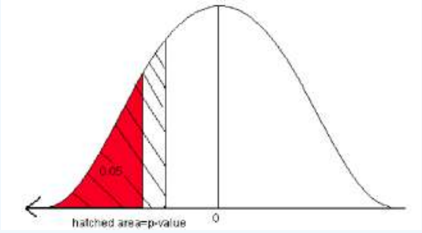


Figure 4.4.2: Comparison of p-value and the level of significance.

Construct and Interpret a Confidence Interval about the Difference of Two Proportions

Just like a two-sample t-test about the means, we can answer this question by constructing a confidence interval about the difference of the proportions. The point estimate is $\hat{p_1} - \hat{p_2}$. The standard error is $\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$ and the critical value $z_{\alpha/2}$ comes from the standard normal table.

The confidence interval takes the form of the point estimate ± the margin of error.

$$(\hat{p_1} - \hat{p_2}) \pm z_{\alpha/2} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \tag{4.4.9}$$

We will use the same three steps to construct a confidence interval about the difference of the proportions. Notice the estimate of the standard error of the differences. We do not rely on the pooled estimate of *p* when constructing confidence intervals to estimate the difference in proportions. This is because we are not making any assumptions regarding the equality of *p*1 and *p*2, as we did in the hypothesis test.

1) critical value $z_{\alpha/2}$

2) $E = z_{\alpha/2} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2}}{n_2}}$

3) $(\hat{p_1} - \hat{p_2}) \pm E$

Let’s revisit Ex. 6 again, but this time we will construct a confidence interval about the difference between the two proportions.

✓ Example 4.4.2:

The researcher claims that a greater proportion of companies in the northeast use at least 30% certified lumber compared to companies in the southeast. We can test this claim by constructing a 90% confidence interval about the difference of the proportions.

1) critical value $z_{\alpha/2} = 1.645$

2) $E = z_{\alpha/2} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2}}{n_2}} = 1.645 \sqrt{\frac{0.497(1-0.497)}{173} + \frac{0.591(1-0.591)}{115}} = 0.098$

3) $(\hat{p_1} - \hat{p_2}) \pm E = (0.497 - 0.591) \pm 0.098$

The 90% confidence interval about the difference of the proportions is (-0.192, 0.004).

BUT, this doesn’t answer the question the researcher asked. We must use one of the three interpretations seen in the previous section. In this problem, the confidence interval contains zero. Therefore we can conclude that there is no significant difference between the proportions of companies using certified lumber in the northeast and in the southeast.

Example 4.4.3:

A hydrologist is studying the use of Best Management Plans (BMP) in managed forest stands to protect riparian zones. He collects information from 62 stands that had a management plan by a forester and finds that 47 stands had correctly implemented BMPs to protect the riparian zones. He collected information from 58 stands that had no management plan and found that 26 of them had correctly implemented BMPs for riparian zones. Do these data suggest that there is a significant difference in the proportion of stands with and without management plans that had correct BMPs for riparian zones? α = 0.05.

Plan	No Plan
$x_1 = 47$	$x_2 = 26$
$n_1 = 62$	$n_2 = 58$

Let's answer this question both ways by first using a hypothesis test and then by constructing a confidence interval about the difference of the proportions.

$H_0 : p_1 = p_2$ or $p_1 - p_2 = 0$

$H_1 : p_1 \neq p_2$

Critical value: ±1.96

Test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.758 - 0.448) - 0}{\sqrt{\frac{0.608(1-0.608)}{62} + \frac{0.608(1-0.608)}{58}}} = 3.48$$

The test statistic is greater than 1.96 and falls in the rejection zone. There is enough evidence to support the claim that there is a significant difference in the proportion of correctly implemented BMPs with and without management plans.

Now compute the p-value and compare it to the level of significance. The p-value is two times the area under the curve to the right of 3.48. Look for the area (in the standard normal table) associated with a Z-score of 3.48. The area to the right of 3.48 is 1 – 0.9997 = 0.0003. The p-value is 2 x 0.0003 = 0.0006.

The p-value is less than 0.05. We will reject the null hypothesis and support the claim that the proportions are different.

Now, answer this question using a confidence interval.

1) critical value $z_{\alpha/2} = 1.96$

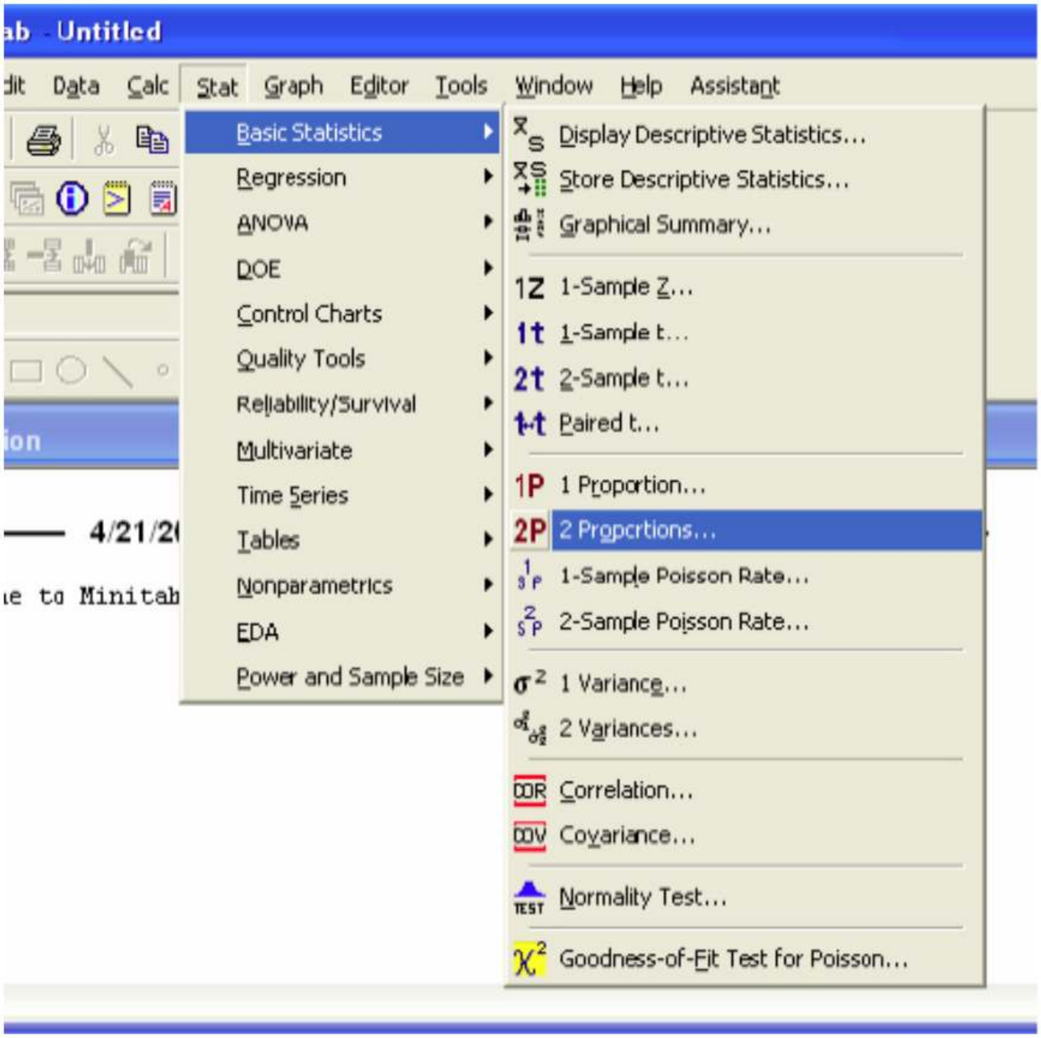
2) $E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 1.96\sqrt{\frac{0.758(1-0.758)}{62} + \frac{0.448(1-0.448)}{58}} = 0.1666$

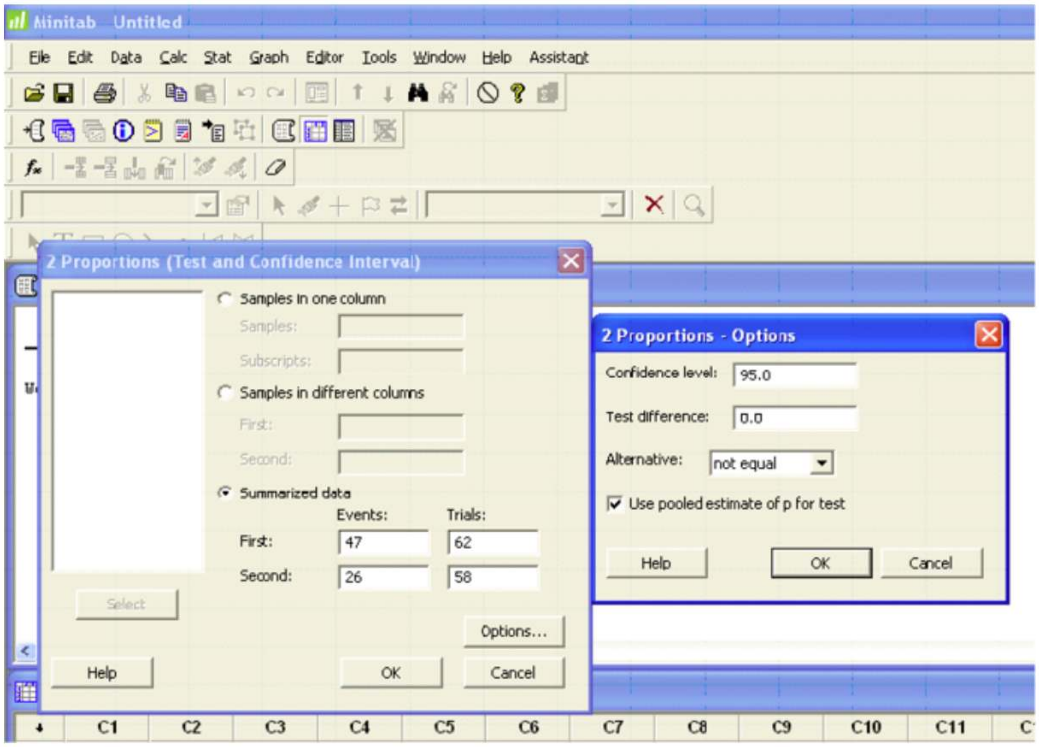
3) $\hat{p}_1 - \hat{p}_2 \pm E = (0.758, -0.448) \pm 0.1666$

The 95% confidence interval about the difference of the proportions is (0.143, 0.477). The confidence interval contains all positive values, telling you that there is a significant difference between the proportions AND the first group (BMPs used with management plans) is significantly greater than the second group (BMPs with no plans). This confidence interval estimates the difference in proportions. For this problem, we can say that correctly implemented BMPs with a plan occur in a greater proportion (14.3% to 44.7%) compared to those implemented without a management plan.

Software Solutions

Minitab


ab - Untitled
dit
Data
Calc
Stat
Graph
Editor
Tools
Window
Help
Assistant
Basic Statistics
Regression
ANOVA
DOE
Control Charts
Quality Tools
Reliability/Survival
Multivariate
Time Series
Tables
Nonparametrics
EDA
Power and Sample Size
Display Descriptive Statistics...
Store Descriptive Statistics...
Graphical Summary...
1-Sample Z...
1-Sample t...
2-Sample t...
Paired t...
1 Proportion...
2 Proportions...
1-Sample Poisson Rate...
2-Sample Poisson Rate...
1 Variance...
2 Variances...
Correlation...
Covariance...
Normality Test...
Goodness-of-Fit Test for Poisson...
ion
4/21/2
e to Minitab


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
2 Proportions (Test and Confidence Interval)
Select
Help
OK
Cancel
Options...
2 Proportions - Options
Help
OK
Cancel
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11

○ Samples in one column
Samples:
Subscripts:
○ Samples in different columns
First:
Second:
● Summarized data

	Events:	Trials:
First:	47	62
Second:	26	58

Confidence level: 95.0
Test difference: 0.0
Alternative: not equal
☑ Use pooled estimate of p for test

Test and CI for Two Proportions

Sample	X	N	Sample p
1	47	62	0.758065
2	26	58	0.448276

Difference = p (1) – p (2)

Estimate for difference: 0.309789

95% CI for difference: (0.143223, 0.476355)

Test for difference = 0 (vs. not = 0): Z = 3.47 p-value = 0.001

Fisher's exact test: p-value = 0.001

The p-value equals 0.001 which tells us to reject the null hypothesis. There is a significant difference in the proportion of correctly implemented BMPs with and without management plans. The confidence interval for the difference in proportions is also given (0.143223, 0.476355) which allows us to estimate the difference.

Excel

Excel does not analyze data from proportions.

4.5: F-Test for Comparing Two Population Variances

F-Test for Comparing Two Population Variances

One major application of a test for the equality of two population variances is for **checking** the validity of the equal variance assumption $(\sigma_1^2 = \sigma_2^2)$ for a two-sample t-test. First we hypothesize two populations of measurements that are normally distributed. We label these populations as 1 and 2, respectively. We are interested in comparing the variance of population 1 (σ_1^2) to the variance of population 2 (σ_2^2).

When independent random samples have been drawn from the respective populations, the ratio

$$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \tag{4.5.1}$$

possesses a probability distribution in repeated sampling that is referred to as an **F** distribution and its properties are:

- Unlike Z and t, but like χ^2, F can assume only positive values.
- The **F** distribution, unlike the Z and t distributions, but like the (\chi^2\) distribution, is non-symmetrical.
- There are many **F** distributions, and each one has a different shape. We specify a particular one by designating the degrees of freedom associated with S_1^2 and S_2^2. We denote these quantities by df_1 and df_2, respectively.

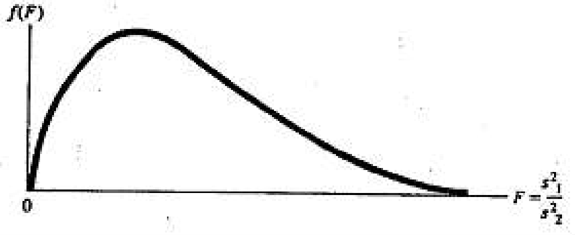


Figure 4.5.1: The F-distribution.

Note: A statistical test of the null hypothesis $\sigma_1^2 = \sigma_2^2$ utilizes the test statistic S_1^2/S_2^2. It may require either upper tail or lower tail rejection region, depending on which sample variance is larger. To alleviate this situation, we are at liberty to designate the population with the larger sample variance as **population 1** (i.e., used as the numerator of the ratio S_1^2/S_2^2). By this convention, the rejection region is only located in the **upper tail of the F distribution**.

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Alternative hypothesis:

- $H_a : \sigma_1^2 > \sigma_2^2$ (one-tailed), reject H_0 if the observed F > Fα
- $H_a : \sigma_1^2 \neq \sigma_2^2$ (two-tailed), reject H_0 if the observed F > Fα/2.

Test statistic: $F = \frac{S_1^2}{S_2^2}$ **assuming** $S_1^2 > S_2^2$,

where the **F** critical value in the rejection region is based on 2 degrees of freedom $df_1 = n_1 - 1$ (associated with numerator S_1^2) and $df_2 = n_2 - 1$ (associated with denominator S_2^2).

✓ Example 4.5.1:

A forester wants to compare two different mist blowers for consistent application. She wants to use the mist blower with the smaller variance, which means more consistent application. She wants to test that the variance of Type A (0.087 gal.2) is significantly greater than the variance of Type B (0.073 gal.2) using α = 0.05.

Type A	Type B
S_1^2 = 0.087	S_2^2=0.073
n_1= 16	n_2 = 21

Solution

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 > \sigma_2^2$

The critical value $(df_1 = 15 \text{ and } df_2 = 20)$ is 2.20.

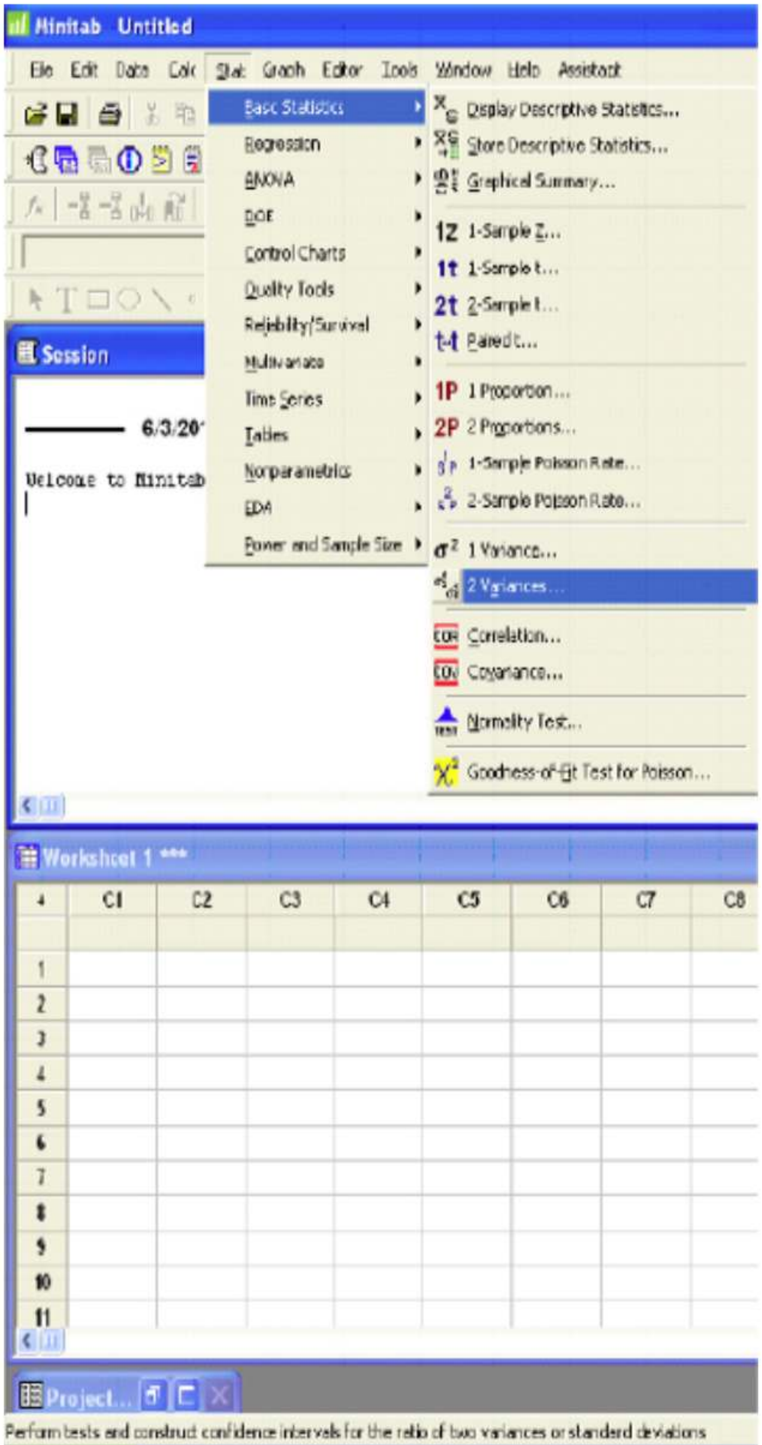
The test statistic is:

$$F = \frac{S_1^2}{S_2^2} = \frac{0.087}{0.073} = 1.192 \tag{4.5.2}$$

The test statistic is not larger than the critical value (it does not fall in the rejection zone) so we fail to reject the null hypothesis. While the variance of Type B is mathematically smaller than the variance of Type A, it is not statistically smaller. There is not enough statistical evidence to support the claim that the variance of Type A is significantly greater than the variance of Type B. Both mist blowers will deliver the chemical with equal consistency.

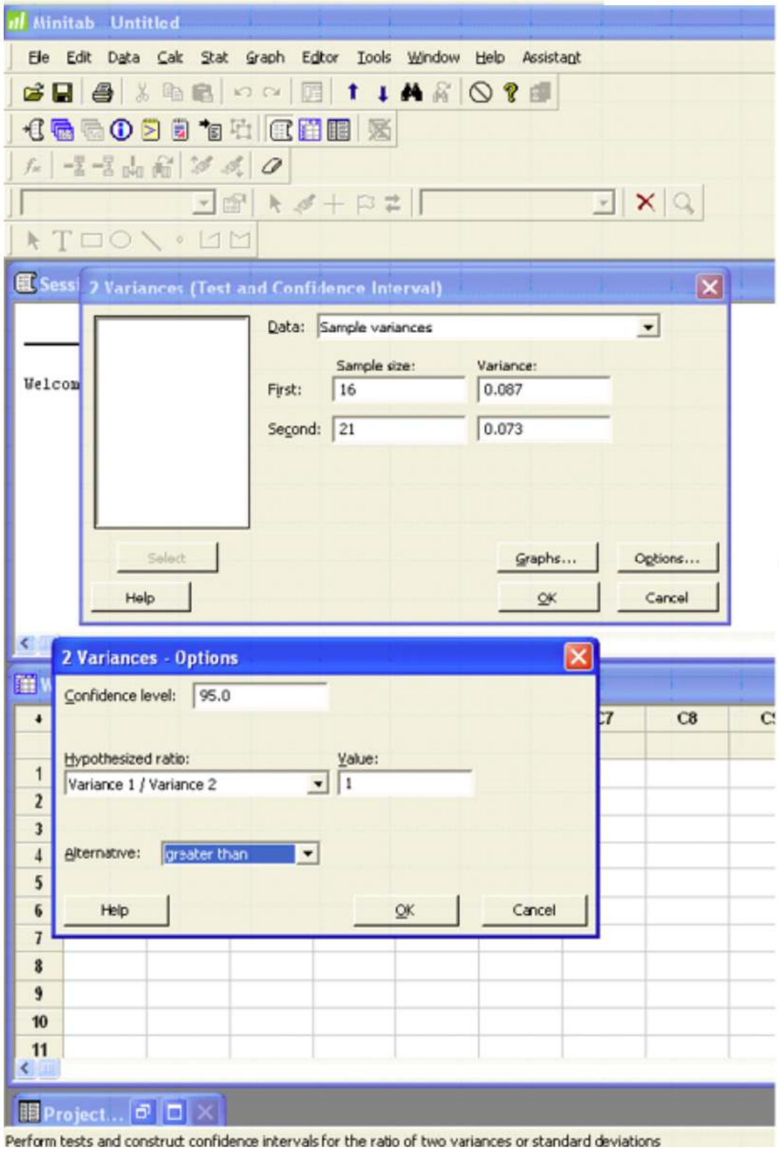
Software Solutions

Minitab



```
Welcome to Minitab
```

Perform tests and construct confidence intervals for the ratio of two variances or standard deviations



2 Variances (Test and Confidence Interval)

Data: Sample variances

	Sample size:	Variance:
First:	16	0.087
Second:	21	0.073

2 Variances - Options

Confidence level: 95.0

Hypothesized ratio: Variance 1 / Variance 2

Value: 1

Alternative: greater than

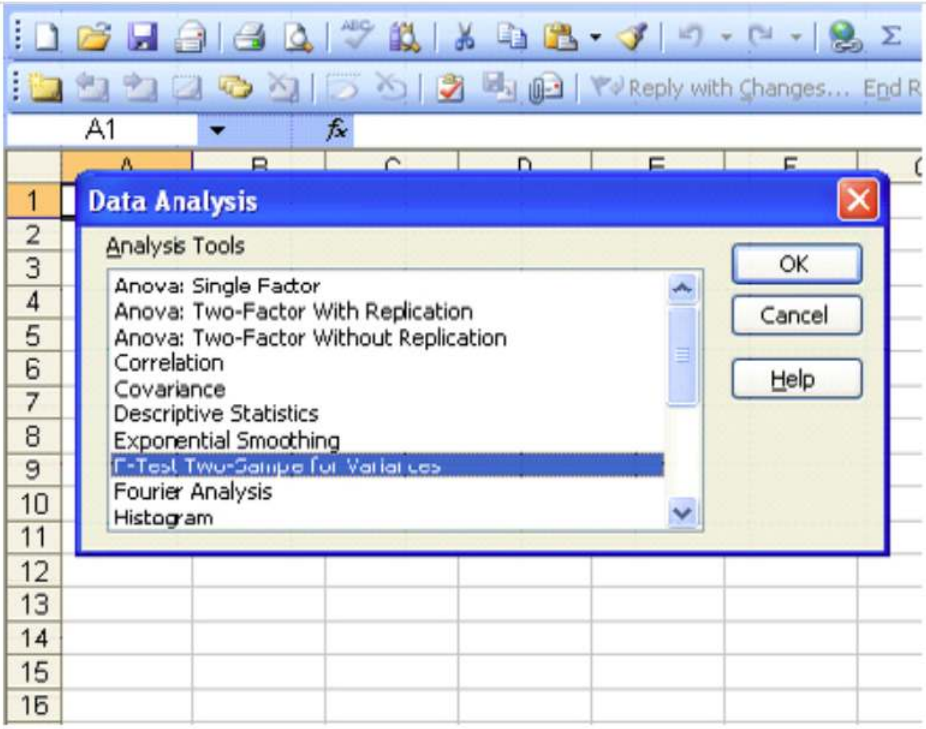
Perform tests and construct confidence intervals for the ratio of two variances or standard deviations

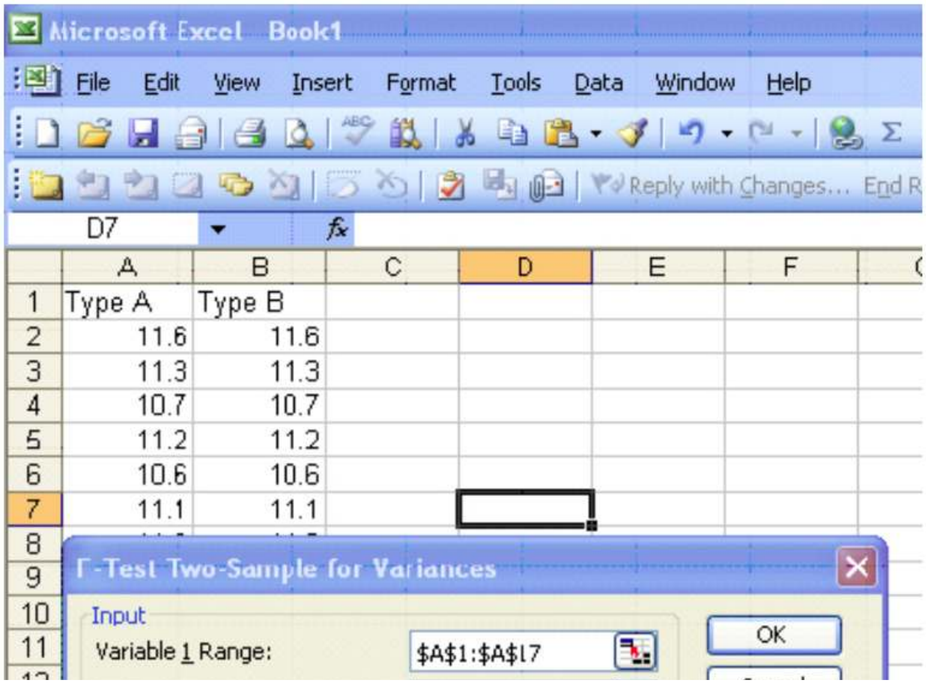
Test and CI for Two Variances - Methods

<table>
<tr><td>Null hypothesis</td><td colspan="3">Variance(1) / Variance(2) = 1</td></tr>
<tr><td>Alternative hypothesis</td><td colspan="3">Variance(1) / Variance(2) > 1</td></tr>
<tr><td>Significance level</td><td colspan="3">Alpha = 0.05</td></tr>
<tr><td colspan="4">Statistics</td></tr>
<tr><td>Sample</td><td>N</td><td>StDev</td><td>Variance</td></tr>
<tr><td>1</td><td>16</td><td>0.295</td><td>0.087</td></tr>
<tr><td>2</td><td>21</td><td>0.270</td><td>0.073</td></tr>
<tr><td colspan="4">Ratio of standard deviations = 1.092</td></tr>
<tr><td colspan="4">Ratio of variances = 1.192</td></tr>
<tr><td colspan="4">Tests</td></tr>
<tr><td></td><td colspan="3">Test</td></tr>
</table>

Null hypothesis	Variance(1) / Variance(2) = 1			
Method	DF1	DF2	Statistic	p-value
F Test (normal)	15	20	1.19	0.351

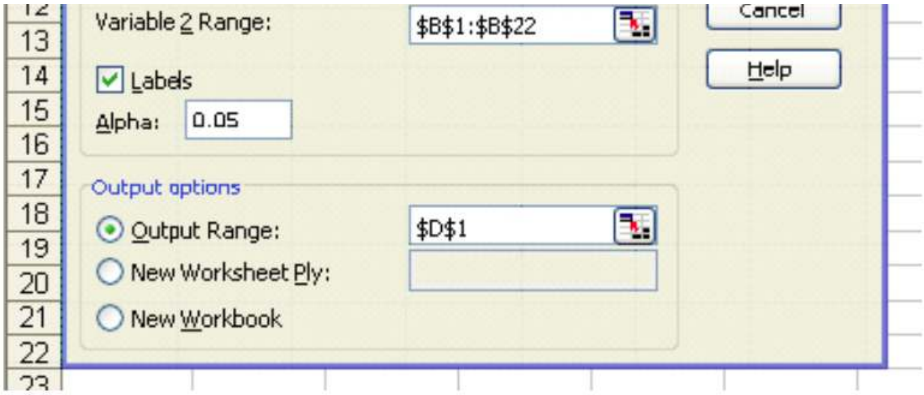
Excel





	A	B
1	Type A	Type B
2	11.6	11.6
3	11.3	11.3
4	10.7	10.7
5	11.2	11.2
6	10.6	10.6
7	11.1	11.1

Variable 1 Range: $A$1:$A$17



Variable 2 Range: $B$1:$B$22

☑ Labels

Alpha: 0.05

Output options

● Output Range: $D$1

○ New Worksheet Ply:

○ New Workbook

F-Test Two-Sample for Variances

	Type A	Type B
Mean	11.07188	11.10595
Variance	0.08699	0.073379
Observations	16	21
df	15	20
F	1.185483	
$P(F \leq f)$ one-tail	0.355098	
F Critical one-tail	2.203274	

4.6: Summary

Questions about the differences between two samples can be answered in several ways: hypothesis test, p-value approach, or confidence interval approach. In all cases, you must clearly state your question, the selected level of significance and the conclusion.

If you choose the hypothesis test approach, you need to compare the critical value to the test statistic. If the test statistic falls in the rejection zone set by the critical value, then you will reject the null hypothesis and support the alternative claim.

If you use the p-value approach, you must compute the test statistic and find the area associated with that value. For a two-sided test, the p-value is two times the area of the absolute value of the test statistic. For a one-sided test, the p-value is the area to the left or right of the test statistic. The decision rule states: If the p-value is less than α(level of significance), reject the null hypothesis and support the alternative claim.

The confidence interval approach constructs an interval about the difference of the means or proportions. If the interval contains zero, then you can conclude that there is no difference between the two groups. If the interval contains all positive values, you can conclude that group 1 is significantly greater than group 2. If the interval contains all negative numbers, you can conclude that group 2 is significantly greater than group 1.

In all approaches, a clear and concise conclusion is required. You MUST answer the question being asked by stating the results of your approach.

CHAPTER OVERVIEW

5: One-Way Analysis of Variance

5.1: Analysis of Variance

Variance Analysis

Previously, we have tested hypotheses about two population means. This chapter examines methods for comparing more than two means. Analysis of variance (ANOVA) is an inferential method used to test the equality of three or more population means.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdot = \mu_k$

This method is also referred to as single-factor ANOVA because we use a single property, or characteristic, for categorizing the populations. This characteristic is sometimes referred to as a treatment or factor.

Note

A treatment (or factor) is a property, or characteristic, that allows us to distinguish the different populations from one another.

The objects of ANOVA are (1) estimate treatment means, and the differences of treatment means; (2) test hypotheses for statistical significance of comparisons of treatment means, where “treatment” or “factor” is the characteristic that distinguishes the populations.

For example, a biologist might compare the effect that three different herbicides may have on seed production of an invasive species in a forest environment. The biologist would want to estimate the mean annual seed production under the three different treatments, while also testing to see which treatment results in the lowest annual seed production. The null and alternative hypotheses are:

$H_0 : \mu_1 = \mu_2 = \mu_3$	H_1: at least one of the means is significantly different from the others

It would be tempting to test this null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ by comparing the population means two at a time. If we continue this way, we would need to test three different pairs of hypotheses:

$H_0 : \mu_1 = \mu_2$	AND	$H_0 : \mu_1 = \mu_3$	AND	$H_0 : \mu_2 = \mu_3$
$H_1 : \mu_1 \neq \mu_2$		$H_1 : \mu_1 \neq \mu_3$		$H_1 : \mu_2 \neq \mu_3$

If we used a 5% level of significance, each test would have a probability of a Type I error (rejecting the null hypothesis when it is true) of α = 0.05. Each test would have a 95% probability of correctly not rejecting the null hypothesis. The probability that all three tests correctly do not reject the null hypothesis is 0.953 = 0.86. There is a 1 – 0.953 = 0.14 (14%) probability that at least one test will lead to an incorrect rejection of the null hypothesis. A 14% probability of a Type I error is much higher than the desired alpha of 5% (remember: α is the same as Type I error). As the number of populations increases, the probability of making a Type I error using multiple t-tests also increases. Analysis of variance allows us to test the null hypothesis (all means are equal) against the alternative hypothesis (at least one mean is different) with a specified value of α.

The assumptions for ANOVA are (1) observations in each treatment group represents a random sample from that population; (2) each of the populations is normally distributed; (3) population variances for each treatment group are homogeneous (i.e., Image37184.PNG). We can easily test the normality of the samples by creating a normal probability plot, however, verifying homogeneous variances can be more difficult. A general rule of thumb is as follows: *One-way ANOVA may be used if the largest sample standard deviation is no more than twice the smallest sample standard deviation.*

In the previous chapter, we used a two-sample t-test to compare the means from two independent samples with a common variance. The sample data are used to compute the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ where } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

is the pooled estimate of the common population variance σ2. To test more than two populations, we must extend this idea of pooled variance to include all samples as shown below:

$$s_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1+n_2+\ldots+n_k-k} \tag{5.1.1}$$

where s_w^2 represents the pooled estimate of the common variance σ^2, and it measures the variability of the observations within the different populations **whether or not H0 is true**. This is often referred to as the variance within samples (variation due to error).

If the null hypothesis IS true (all the means are equal), then all the populations are the same, with a common mean μ and variance σ^2. Instead of randomly selecting different samples from different populations, we are actually drawing *k* different samples from one population. We know that the sampling distribution for *k* means based on *n* observations will have mean $\mu\bar{x}$ and variance $\frac{\sigma^2}{n}$ (squared standard error). Since we have drawn *k* samples of *n* observations each, we can estimate the variance of the k sample means ($\frac{\sigma^2}{n}$) by

$$\frac{\sum(\bar{x_1}-\mu_{\bar{x}})^2}{k-1} = \frac{\sum \bar{x_i}^1 - \frac{[\sum \bar{x_i}]^2}{k}}{k-1} = \frac{\sigma^2}{n} \tag{5.1.2}$$

Consequently, *n* times the sample variance of the means estimates σ2. We designate this quantity as SB2 such that

$$S_B^2 = n * \frac{\sum(\bar{x_i}-\mu_{\bar{x}})^2}{k-1} = n * \frac{\sum \bar{x_i}^2 - \frac{[\bar{x_i}]^2}{k}}{k-1} \tag{5.1.3}$$

where S_B^2 is also an unbiased estimate of the common variance σ^2, IF H_0 IS TRUE. This is often referred to as the variance between samples (variation due to treatment).

Under the null hypothesis that all *k* populations are identical, we have two estimates of σ_2 ($S_W^2$and S_B^2). We can use the ratio of S_B^2/S_W^2 as a test statistic to test the null hypothesis that $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$, which follows an F-distribution with degrees of freedom $df_1 = k\text{–}1$ and \(df_2= N –k\) (where *k* is the number of populations and N is the total number of observations ($N = n_1 + n_2 + \ldots + n_k$). The numerator of the test statistic measures the variation between sample means. The estimate of the variance in the denominator depends only on the sample variances and is not affected by the differences among the sample means.

When the null hypothesis is true, the ratio of S_B^2 and S_W^2 will be close to 1. When the null hypothesis is false, S_B^2 will tend to be larger than S_W^2 due to the differences among the populations. We will reject the null hypothesis if the F test statistic is larger than the F critical value at a given level of significance (or if the p-value is less than the level of significance).

Tables are a convenient format for summarizing the key results in ANOVA calculations. The following one-way ANOVA table illustrates the required computations and the relationships between the various ANOVA table elements.

Table 1. One-way ANOVA table.

Source of Variation	df	Sum of Squares (MSS)	F-Test	p-value
Treatment	k-1	SSTr	MSTr=SSTr/(k-1)	
Error	N-k	SSE	MSE=SSE/(N-k)	
Total	N-1	SST0		

The sum of squares for the ANOVA table has the relationship of SSTo = SSTr + SSE where:

$$SSTo = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{\bar{x}})^2 \tag{5.1.4}$$

$$SSTr = \sum_{i=1}^{k} n_i(\bar{x_i}-\bar{\bar{x}})^2 \tag{5.1.5}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x_i})^2 \tag{5.1.6}$$

Total variation (SSTo) = explained variation (SSTr) + unexplained variation (SSE)

The degrees of freedom also have a similar relationship: df(SSTo) = df(SSTr) + df(SSE)

The Mean Sum of Squares for the treatment and error are found by dividing the Sums of Squares by the degrees of freedom for each. While the Sums of Squares are additive, the Mean Sums of Squares are not. The F-statistic is then found by dividing the Mean Sum of Squares for the treatment (MSTr) by the Mean Sum of Squares for the error(MSE). The MSTr is the S_B^2 and the MSE is the S_W^2.

$$F = \frac{S_B^2}{S_W^2} = \frac{MSTr}{MSE} \tag{5.1.7}$$

✓ Example 5.1.1:

An environmentalist wanted to determine if the mean acidity of rain differed among Alaska, Florida, and Texas. He randomly selected six rain dates at each site obtained the following data:

Table 2. Data for Alaska, Florida, and Texas.

Alaska	*Florida*	*Texas*
5.11	4.87	5.46
5.01	4.18	6.29
4.90	4.40	5.57
5.14	4.67	5.15
4.80	4.89	5.45
5.24	4.09	5.30

Solution

$H_0 : \mu_A = \mu_F = \mu_T$

H_1: at least one of the means is different

State	Sample size	Sample total	Sample mean	Sample variance
Alaska	n1 = 6	30.2	5.033	0.0265
Florida	n2 = 6	27.1	4.517	0.1193
Texas	n3 = 6	33.22	5.537	0.1575

Table 3. Summary Table.

Notice that there are differences among the sample means. Are the differences small enough to be explained solely by sampling variability? Or are they of sufficient magnitude so that a more reasonable explanation is that the μ's are not all equal? The conclusion depends on how much variation among the sample means (based on their deviations from the grand mean) compares to the variation within the three samples.

The grand mean is equal to the sum of all observations divided by the total sample size:

$\bar{\bar{x}}$= grand total/N = 90.52/18 = 5.0289

$$\begin{gathered} SSTo = (5.11-5.0289)^2 + (5.01-5.0289)^2 + \ldots + (5.24-5.0289)^2 + (4.87-5.0289)^2 + (4.18-5.0289)^2 + \\ \ldots + (4.09-5.0289)^2 + (5.46-5.0289)^2 + (6.29-5.0289)^2 + \ldots + (5.30-5.0289)^2 = 4.6384 \end{gathered} \tag{5.1.8}$$

$$SSTr = 6(5.033-5.0289)^2 + 6(4.517-5.0289)^2 + 6(5.537-5.0289)^2 = 3.1214 \tag{5.1.9}$$

$$SSE = SSTo – SSTr = 4.6384 – 3.1214 = 1.5170 \tag{5.1.10}$$

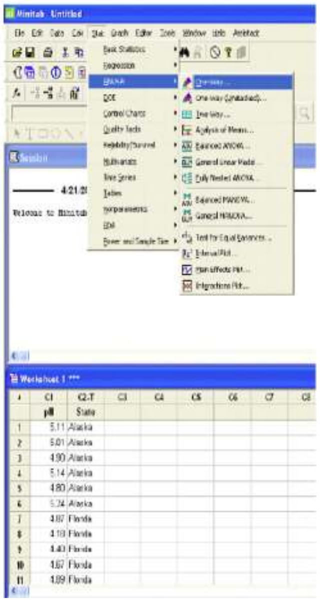
Table 4. One-way ANOVA Table.

Source of Variation	df	Sum of Squares (SS)	Mean Sum of Squares (MSS)	F-Test
Treatment	3-1	3.1214	3.1214/2=1.5607	1.5607/0.1011=15.4372
Error	18-3	1.5170	1.5170/15=0.1011	
Total	18-1	4.6384		

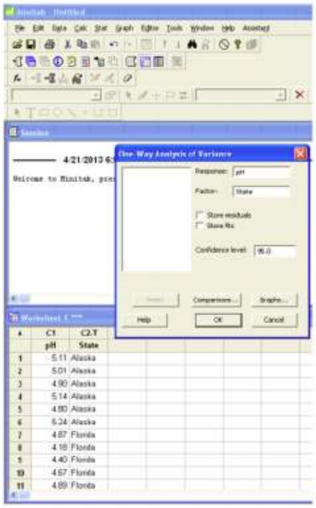
This test is based on $df_1 = k – 1 = 2$ and $df_2 = N – k = 15$. For α = 0.05, the F critical value is 3.68. Since the observed F = 15.4372 is greater than the F critical value of 3.68, we reject the null hypothesis. There is enough evidence to state that at least one of the means is different.

Software Solutions

Minitab



↓	C1	C2-T
	pH	State
1	5.11	Alaska
2	5.01	Alaska
3	4.90	Alaska
4	5.14	Alaska
5	4.80	Alaska
6	5.24	Alaska
7	4.87	Florida
8	4.18	Florida
9	4.40	Florida
10	4.67	Florida
11	4.89	Florida


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
One-Way Analysis of Variance
Select
Comparisons...
Graphs...
Help
OK
Cancel
Worksheet 1 ***

Response: pH

Factor: State

☐ Store residuals

☐ Store fits

Confidence level: 95.0

One-way ANOVA: pH vs. State

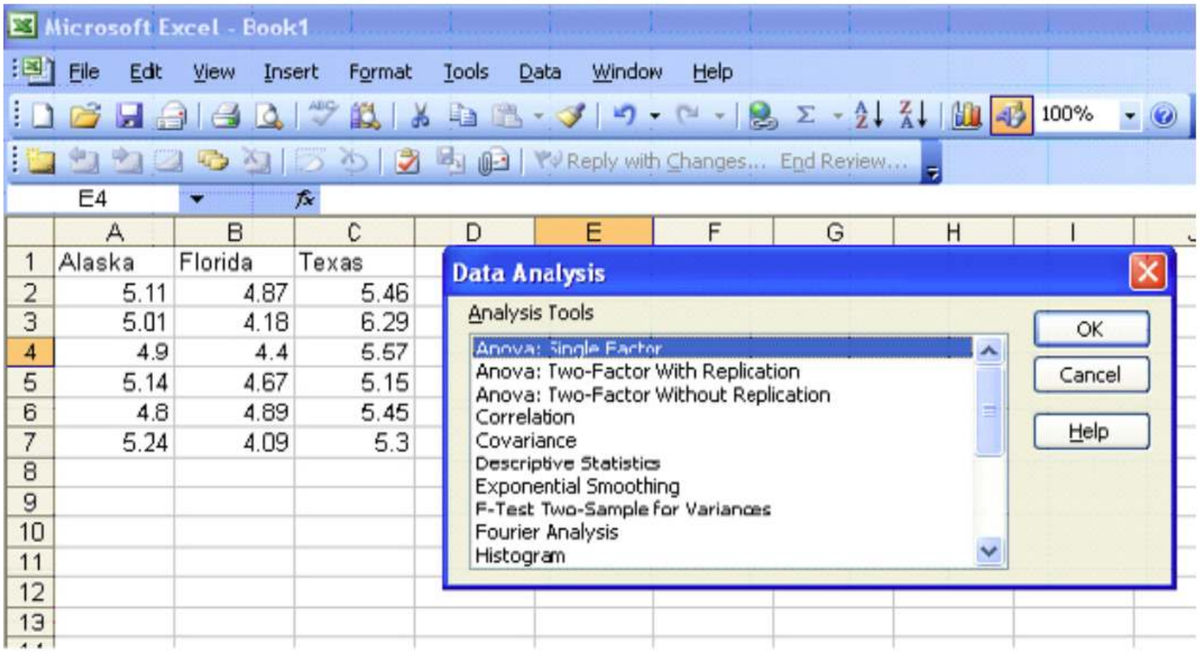
Source	DF	SS	MS	F	P
State	2	3.121	1.561	15.43	0.000
Error	15	1.517	0.101		
Total	17 4.638				
S = 0.3180 R-Sq = 67.29% R-Sq(adj) = 62.93%					

				Individual 95% CIs For Mean Based on Pooled StDev
Level	N	Mean	StDev	—-+———+———+———+——
Alaska	6	5.0333	0.1629	(——*——)
Florida	6	4.5167	0.3455	(——*——)

			Individual 95% CIs For Mean Based on Pooled StDev					
Texas	6	5.5367	0.3969			(——*——)		
				—-+———+———+———+——				
				4.40	4.80	5.20	5.60	
Pooled StDev = 0.3180								

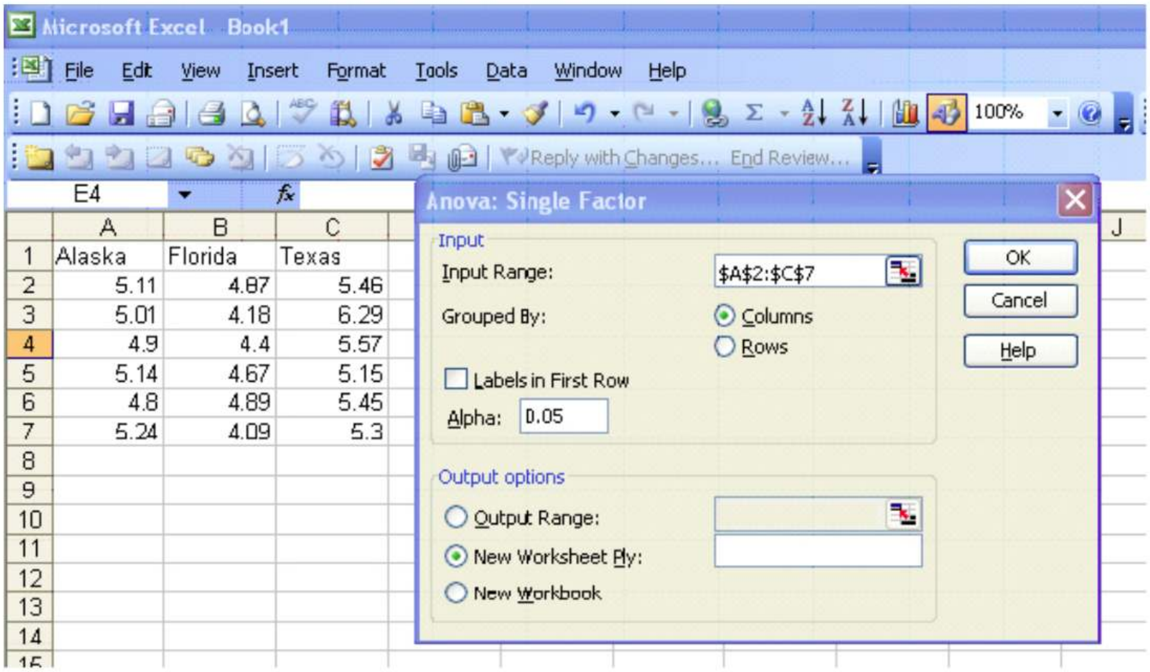
The p-value (0.000) is less than the level of significance (0.05) so we will reject the null hypothesis.

Excel



	A	B	C
1	Alaska	Florida	Texas
2	5.11	4.87	5.46
3	5.01	4.18	6.29
4	4.9	4.4	5.57
5	5.14	4.67	5.15
6	4.8	4.89	5.45
7	5.24	4.09	5.3

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram



Microsoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
100%
Reply with Changes... End Review...
E4
Anova: Single Factor
OK
Cancel
Help

	A	B	C
1	Alaska	Florida	Texas
2	5.11	4.87	5.46
3	5.01	4.18	6.29
4	4.9	4.4	5.57
5	5.14	4.67	5.15
6	4.8	4.89	5.45
7	5.24	4.09	5.3

Input
Input Range: $A$2:$C$7
Grouped By: ● Columns ○ Rows
☐ Labels in First Row
Alpha: 0.05

Output options
○ Output Range:
● New Worksheet Ply:
○ New Workbook

ANOVA: Single Factor

ANOVA: Single Factor

SUMMARY				
Groups	Count	Sum	Average	Variance
Column 1	6	30.2	5.033333	0.026547
Column 2	6	27.1	4.516667	0.119347
Column 3	6	33.22	5.536667	0.157507

ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	3.121378	2	1.560689	15.43199	0.000229	3.68232
Within Groups	1.517	15	0.101133			
Total	4.638378	17				

The p-value (0.000229) is less than alpha (0.05) so we reject the null hypothesis. There is enough evidence to support the claim that at least one of the means is different.

Once we have rejected the null hypothesis and found that at least one of the treatment means is different, the next step is to identify those differences. There are two approaches that can be used to answer this type of question: contrasts and multiple comparisons.

Contrasts can be used only when there are clear expectations BEFORE starting an experiment, and these are reflected in the experimental design. Contrasts are **planned comparisons**. For example, mule deer are treated with drug A, drug B, or a placebo to treat an infection. The three treatments are not symmetrical. The placebo is meant to provide a baseline against which the other drugs can be compared. Contrasts are more powerful than multiple comparisons because they are more specific. They are more able to pick up a significant difference. Contrasts are not always readily available in statistical software packages (when they are, you often need to assign the coefficients), or may be limited to comparing each sample to a control.

Multiple comparisons should be used when there are no justified expectations. They are *aposteriori*, **pair-wise tests** of significance. For example, we compare the gas mileage for six brands of all-terrain vehicles. We have no prior knowledge to expect any vehicle to perform differently from the rest. Pair-wise comparisons should be performed here, but only if an ANOVA test on all six vehicles rejected the null hypothesis first.

It is NOT appropriate to use a contrast test when suggested comparisons appear only after the data have been collected. We are going to focus on multiple comparisons instead of planned contrasts.

5.2: Multiple Comparisons

When the null hypothesis is rejected by the F-test, we believe that there are significant differences among the *k* population means. So, which ones are different? Multiple comparison method is the way to identify which of the means are different while controlling the experiment-wise error (the accumulated risk associated with a family of comparisons). There are many multiple comparison methods available.

In **The Least Significant Difference Test**, each individual hypothesis is tested with the student t-statistic. When the Type I error probability is set at some value and the variance s2 has *v* degrees of freedom, the null hypothesis is rejected for any observed value such that |to|>tα/2, v. It is an abbreviated version of conducting all possible pair-wise t-tests. This method has weak experiment-wise error rate. Fisher's Protected LSD is somewhat better at controlling this problem.

Bonferroni inequality is a conservative alternative when software is not available. When conducting n comparisons, αe≤ n αc therefore αc = αe/n. In other words, divide the experiment-wise level of significance by the number of multiple comparisons to get the comparison-wise level of significance. The Bonferroni procedure is based on computing confidence intervals for the differences between each possible pair of μ's. The critical value for the confidence intervals comes from a table with (N – *k*) degrees of freedom and *k*(*k* – 1)/2 number of intervals. If a particular interval does not contain zero, the two means are declared to be significantly different from one another. An interval that contains zero indicates that the two means are NOT significantly different.

Dunnett's procedure was created for studies where one of the treatments acts as a control treatment for some or all of the remaining treatments. It is primarily used if the interest of the study is determining whether the mean responses for the treatments differ from that of the control. Like Bonferroni, confidence intervals are created to estimate the difference between two treatment means with a specific table of critical values used to control the experiment-wise error rate. The standard error of the difference is Image37382.PNG.

Scheffe's test is also a conservative method for all possible simultaneous comparisons suggested by the data. This test equates the F statistic of ANOVA with the t-test statistic. Since t2 = F then t = √F, we can substitute √F(αe, v1, v2) for t(αe, v2) for Scheffe's statistic.

Tukey's test provides a strong sense of experiment-wise error rate for all pair-wise comparison of treatment means. This test is also known as the *Honestly Significant Difference*. This test orders the treatments from smallest to largest and uses the studentized range statistic

$$q=\frac{\bar{y}(largest)-\bar{y}(smallest)}{\sqrt{MSE/r}} \tag{5.2.1}$$

The absolute difference of the two means is used because the location of the two means in the calculated difference is arbitrary, with the sign of the difference depending on which mean is used first. For unequal replications, the Tukey-Kramer approximation is used instead.

Student-Newman-Keuls (SNK) test is a multiple range test based on the studentized range statistic like Tukey's. The critical value is based on a particular pair of means being tested within the entire set of ordered means. Two or more ranges among means are used for test criteria. While it is similar to Tukey's in terms of a test statistic, it has weak experiment-wise error rates.

Bonferroni, Dunnett's, and Scheffe's tests are the most conservative, meaning that the difference between the two means must be greater before concluding a significant difference. The LSD and SNK tests are the least conservative. Tukey's test is in the middle. Robert Kuehl, author of *Design of Experiments: Statistical Principles of Research Design and Analysis* (2000), states that the Tukey method provides the best protection against decision errors, along with a strong inference about magnitude and direction of differences.

Let's go back to our question on mean rain acidity in Alaska, Florida, and Texas. The null and alternative hypotheses were as follows:

H_0: μA = μF = μT	H_1: at least one of the means is different

The p-value for the F-test was 0.000229, which is less than our 5% level of significance. We rejected the null hypothesis and had enough evidence to support the claim that at least one of the means was significantly different from another. We will use Bonferroni and Tukey's methods for multiple comparisons in order to determine which mean(s) is different.

Bonferroni Multiple Comparison Method

A Bonferroni confidence interval is computed for each pair-wise comparison. For *k* populations, there will be *k*(*k*-1)/2 multiple comparisons. The confidence interval takes the form of:

$$For\ \mu_1-\mu_2:(\bar{x_1}-\bar{x_2})\pm(Bonferronit\ critical\ value)\sqrt{\frac{MSE}{n_1}+\frac{MSE}{n_2}}$$

$$For\ \mu_{k-1}-\mu_k:(\bar{x_{k-1}}-\bar{x_k})\pm(Bonferronit\ critical\ value)\sqrt{\frac{MSE}{n_{k-1}}+\frac{MSE}{n_k}}$$

Where MSE is from the analysis of variance table and the Bonferroni *t* critical value comes from the Bonferroni Table given below. The Bonferroni *t* critical value, instead of the student *t* critical value, combined with the use of the MSE is used to achieve a simultaneous confidence level of at least 95% for all intervals computed. The two means are judged to be significantly different if the corresponding interval does not include zero.

Table 5. Bonferroni t-critical values.

df	2	3	4	5	6	10
2	6.21	7.65	8.86	9.92	10.89	14.09
3	4.18	4.86	5.39	5.84	6.23	7.45
4	3.50	3.96	4.31	4.60	4.85	5.60
5	3.16	3.53	3.81	4.03	4.22	4.77
6	2.97	3.29	3.52	3.71	3.86	4.32
7	2.84	3.13	3.34	3.50	3.64	4.03
8	2.75	3.02	3.21	3.36	3.48	3.83
9	2.69	2.93	3.11	3.25	3.36	3.69
10	2.63	2.87	3.04	3.17	3.28	3.58
11	2.59	2.82	2.98	3.11	3.21	3.50
12	2.56	2.78	2.93	3.05	3.15	3.43
13	2.53	2.75	2.90	3.01	3.11	3.37
14	2.51	2.72	2.86	2.98	3.07	3.33
15	2.49	2.69	2.84	2.95	3.04	3.29
16	2.47	2.67	2.81	2.92	3.01	3.25
17	2.46	2.66	2.79	2.90	2.98	3.22
18	2.45	2.64	2.77	2.88	2.96	3.20
19	2.43	2.63	2.76	2.86	2.94	3.17
20	2.42	2.61	2.74	2.85	2.93	3.15
21	2.41	2.60	2.73	2.83	2.91	3.14
22	2.41	2.59	2.72	2.82	2.90	3.12
23	2.40	2.58	2.71	2.81	2.89	3.10
24	2.39	2.57	2.70	2.80	2.88	3.09
25	2.38	2.57	2.69	2.79	2.86	3.08

df	2	3	4	5	6	10
26	2.38	2.56	2.68	2.78	2.86	3.07
27	2.37	2.55	2.68	2.77	2.85	3.06
28	2.37	2.55	2.67	2.76	2.84	3.05
29	2.36	2.54	2.66	2.76	2.83	3.04
30	2.36	2.54	2.66	2.75	2.82	3.03
40	2.33	2.50	2.62	2.70	2.78	2.97
60	2.30	2.46	2.58	2.66	2.73	2.91
120	2.27	2.43	2.54	2.62	2.68	2.86

For this problem, $k = 3$ so there are $k(k - 1)/2= 3(3 - 1)/2 = 3$ multiple comparisons. The degrees of freedom are equal to N – k = 18 – 3 = 15. The Bonferroni critical value is 2.69.

$$For \mu_A - \mu_F : (5.033 - 4.517) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (0.0222, 1.0098)$$

$$For \mu_A - \mu_T : (5.033 - 5.537) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (-0.9978, -0.0102)$$

$$For \mu_F - \mu_T : (4.517 - 5.537) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (-1.5138, 0.5262)$$

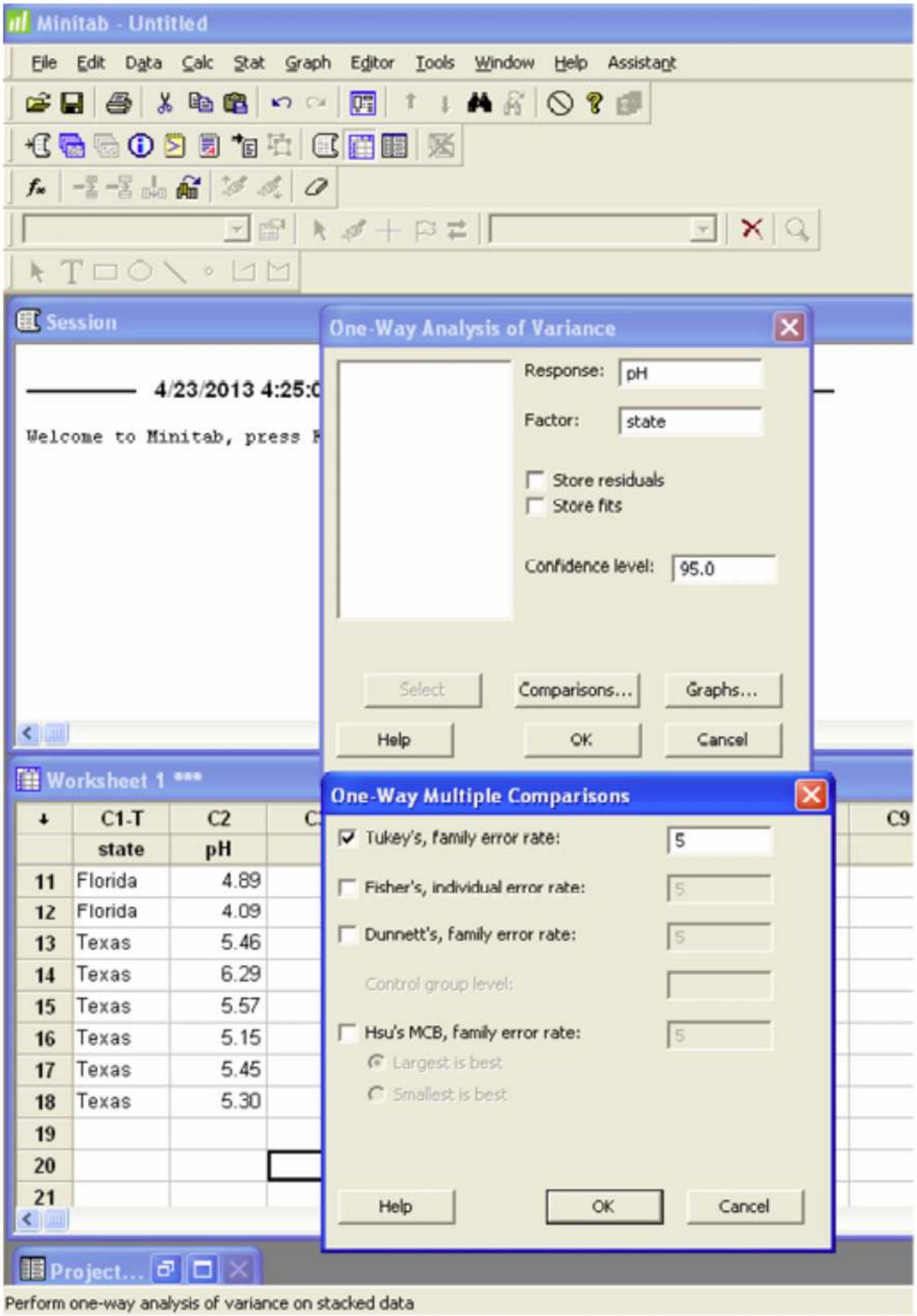
The first confidence interval contains all positive values. This tells you that there is a significant difference between the two means and that the mean rain pH for Alaska is significantly greater than the mean rain pH for Florida.

The second confidence interval contains all negative values. This tells you that there is a significant difference between the two means and that the mean rain pH of Alaska is significantly lower than the mean rain pH of Texas.

The third confidence interval also contains all negative values. This tells you that there is a significant difference between the two means and that the mean rain pH of Florida is significantly lower than the mean rain pH of Texas.

All three states have significantly different levels of rain pH. Texas has the highest rain pH, then Alaska followed by Florida, which has the lowest mean rain pH level. You can use the confidence intervals to estimate the mean difference between the states. For example, the average rain pH in Texas ranges from 0.5262 to 1.5138 higher than the average rain pH in Florida.

Now let’s use the Tukey method for multiple comparisons. We are going to let software compute the values for us. Excel doesn’t do multiple comparisons so we are going to rely on Minitab output.


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Worksheet 1 ***
One-Way Analysis of Variance
Select
Comparisons...
Graphs...
Help
OK
Cancel
One-Way Multiple Comparisons
Help
OK
Cancel
Project...

4/23/2013 4:25:0

Welcome to Minitab, press F

↓	C1-T state	C2 pH
11	Florida	4.89
12	Florida	4.09
13	Texas	5.46
14	Texas	6.29
15	Texas	5.57
16	Texas	5.15
17	Texas	5.45
18	Texas	5.30
19		
20		
21		

Response: pH

Factor: state

☐ Store residuals

☐ Store fits

Confidence level: 95.0

☑ Tukey's, family error rate: 5

☐ Fisher's, individual error rate: 5

☐ Dunnett's, family error rate: 5

Control group level:

☐ Hsu's MCB, family error rate: 5

● Largest is best

○ Smallest is best

Perform one-way analysis of variance on stacked data

One-way ANOVA: pH vs. state

Source	DF	SS	MS	F	P
state	2	3.121	1.561	15.4	0.000
Error	15	1.517	0.101		

Source	DF	SS	MS	F	P
Total	17	4.638			
S = 0.3180		R-Sq = 67.29%		R-Sq(adj) = 62.93%	

We have seen this part of the output before. We now want to focus on the *Grouping Information Using Tukey Method.* All three states have different letters indicating that the mean rain pH for each state is significantly different. They are also listed from highest to lowest. It is easy to see that Texas has the highest mean rain pH while Florida has the lowest.

Grouping Information Using Tukey Method

state	N	Mean	Grouping
Texas	6	5.5367	A
Alaska	6	5.0333	B
Florida	6	4.516	C
Means that do not share a letter are significantly different.			

This next set of confidence intervals is similar to the Bonferroni confidence intervals. They estimate the difference of each pair of means. The individual confidence interval level is set at 97.97% instead of 95% thus controlling the experiment-wise error rate.

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of state
Individual confidence level = **97.97%**

state = Alaska subtracted from:							
state	Lower	Center	Upper	———+———+———+———+			
Florida	-0.9931	-0.5167	-0.0402		(——*——)		
Texas	0.0269	0.5033	0.9798			(——*——)	
				———+———+———+———+			
				-0.80	0.00	0.80	1.60

state = Florida subtracted from:							
state	Lower	Center	Upper	———+———+———+———+			
Texas	0.5435	1.0200	1.4965			(——*——)	
				———+———+———+———+			
				-0.80	0.00	0.80	1.60

The first pairing is Florida – Alaska, which results in an interval of (-0.9931, -0.0402). The interval has all negative values indicating that Florida is significantly lower than Alaska. The second pairing is Texas – Alaska, which results in an interval of (0.0269, 0.9798). The interval has all positive values indicating that Texas is greater than Alaska. The third pairing is Texas – Florida, which results in an interval from (0.5435, 1.4965). All positive values indicate that Texas is greater than Florida.

The intervals are similar to the Bonferroni intervals with differences in width due to methods used. In both cases, the same conclusions are reached.

When we use one-way ANOVA and conclude that the differences among the means are significant, we can't be absolutely sure that the given factor is responsible for the differences. It is possible that the variation of some other unknown factor is responsible. One way to reduce the effect of extraneous factors is to design an experiment so that it has a completely randomized design. This means that each element has an equal probability of receiving any treatment or belonging to any different group. In general good results require that the experiment be carefully designed and executed.

Additional Example:

https://youtu.be/BMyYXc8cWHs

CHAPTER OVERVIEW

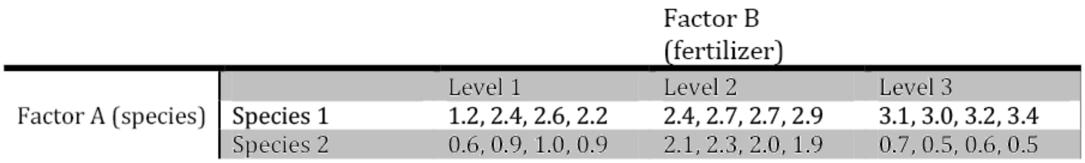
6: Two-way Analysis of Variance

6.1: Main Effects and Interaction Effect

In the previous chapter we used one-way ANOVA to analyze data from three or more populations using the null hypothesis that all means were the same (no treatment effect). For example, a biologist wants to compare mean growth for three different levels of fertilizer. A one-way ANOVA tests to see if at least one of the treatment means is significantly different from the others. If the null hypothesis is rejected, a multiple comparison method, such as Tukey's, can be used to identify which means are different, and the confidence interval can be used to estimate the difference between the different means.

Suppose the biologist wants to ask this same question but with two different species of plants while still testing the three different levels of fertilizer. The biologist needs to investigate not only the average growth between the two species (main effect A) and the average growth for the three levels of fertilizer (main effect B), but also the **interaction** or relationship between the two factors of species and fertilizer. Two-way analysis of variance allows the biologist to answer the question about growth affected by species and levels of fertilizer, and to account for the variation due to both factors simultaneously.

Our examination of one-way ANOVA was done in the context of a completely randomized design where the treatments are assigned randomly to each subject (or experimental unit). We now consider analysis in which two factors can explain variability in the response variable. Remember that we can deal with factors by controlling them, by fixing them at specific levels, and randomly applying the treatments so the effect of uncontrolled variables on the response variable is minimized. With two factors, we need a factorial experiment.



			Factor B (fertilizer)	
		Level 1	Level 2	Level 3
Factor A (species)	Species 1	1.2, 2.4, 2.6, 2.2	2.4, 2.7, 2.7, 2.9	3.1, 3.0, 3.2, 3.4
	Species 2	0.6, 0.9, 1.0, 0.9	2.1, 2.3, 2.0, 1.9	0.7, 0.5, 0.6, 0.5

Table 1. Observed data for two species at three levels of fertilizer.

This is an example of a factorial experiment in which there are a total of 2 x 3 = 6 possible combinations of the levels for the two different factors (species and level of fertilizer). These six combinations are referred to as treatments and the experiment is called a **2 x 3 factorial experiment**. We use this type of experiment to investigate the effect of multiple factors on a response and the interaction between the factors. Each of the *n* observations of the response variable for the different levels of the factors exists within a cell. In this example, there are six cells and each cell corresponds to a specific treatment.

When you compare treatment means for a factorial experiment (or for any other experiment), multiple observations are required for each treatment. These are called replicates. For example, if you have four observations for each of the six treatments, you have four **replications** of the experiment. Replication demonstrates the results to be reproducible and provides the means to estimate experimental error variance. Replication also provides the capacity to increase the precision for estimates of treatment means. Increasing replication decreases $s_{\bar{y}}^{2} = \frac{s^2}{r}$ thereby increasing the precision of $\bar{y}$.

Notation

- *k* = number of levels of factor A
- *l* = number of levels of factor B
- *kl* = number of treatments (each one a combination of a factor A level and a factor B level)
- *m* = number of observations on each treatment

Main Effects and Interaction Effect

Main effects deal with each factor separately. In the previous example we have two factors, A and B. The main effect of Factor A (species) is the difference between the mean growth for Species 1 and Species 2, averaged across the three levels of fertilizer. The main effect of Factor B (fertilizer) is the difference in mean growth for levels 1, 2, and 3 averaged across the two species. The interaction is the simultaneous changes in the levels of both factors. If the changes in the level of Factor A result in different changes in the value of the response variable for the different levels of Factor B, we say that there is an interaction effect between the factors. Consider the following example to help clarify this idea of interaction.

✓ Example 6.1.1:

Factor A has two levels and Factor B has two levels. In the left box, when Factor A is at level 1, Factor B changes by 3 units. When Factor A is at level 2, Factor B again changes by 3 units. Similarly, when Factor B is at level 1, Factor A changes by 2 units. When Factor B is at level 2, Factor A again changes by 2 units. There is no interaction. The change in the true average response when the level of either factor changes from 1 to 2 is the same for each level of the other factor. In this case, changes in levels of the two factors affect the true average response separately, or in an additive manner.

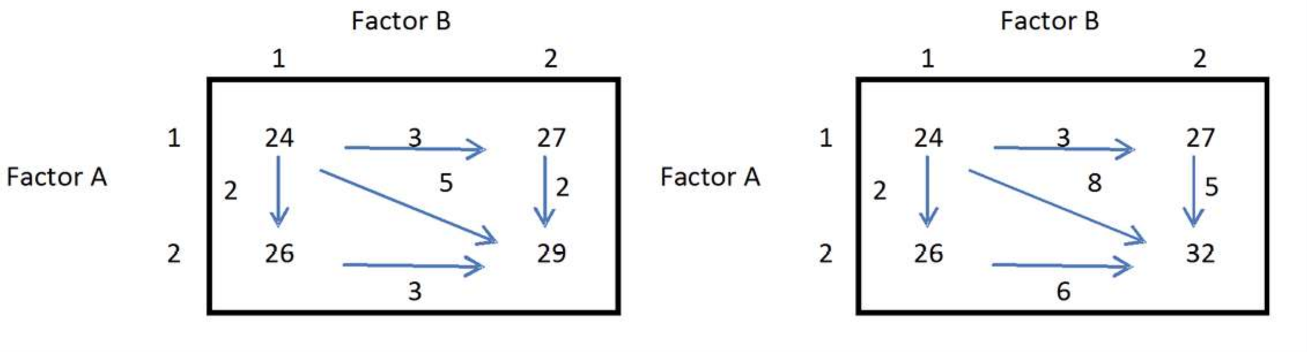


Figure 1. Illustration of interaction effect.

Solution

The right box illustrates the idea of interaction. When Factor A is at level 1, Factor B changes by 3 units but when Factor A is at level 2, Factor B changes by 6 units. When Factor B is at level 1, Factor A changes by 2 units but when Factor B is at level 2, Factor A changes by 5 units. The change in the true average response when the levels of both factors change simultaneously from level 1 to level 2 is 8 units, which is much larger than the separate changes suggest. In this case, there is an interaction between the two factors, so the effect of simultaneous changes cannot be determined from the individual effects of the separate changes. Change in the true average response when the level of one factor changes depends on the level of the other factor. You cannot determine the separate effect of Factor A or Factor B on the response because of the interaction.

Assumptions

Note: Basic Assumption

The observations on any particular treatment are independently selected from a normal distribution with variance σ2 (the same variance for each treatment), and samples from different treatments are independent of one another.

We can use normal probability plots to satisfy the assumption of normality for each treatment. The requirement for equal variances is more difficult to confirm, but we can generally check by making sure that the largest sample standard deviation is no more than twice the smallest sample standard deviation.

Although not a requirement for two-way ANOVA, having an equal number of observations in each treatment, referred to as a balance design, increases the power of the test. However, unequal replications (an unbalanced design), are very common. Some statistical software packages (such as Excel) will only work with balanced designs. Minitab will provide the correct analysis for both balanced and unbalanced designs in the General Linear Model component under ANOVA statistical analysis. However, for the sake of simplicity, we will focus on balanced designs in this chapter.

Sums of Squares and the ANOVA Table

In the previous chapter, the idea of sums of squares was introduced to partition the variation due to treatment and random variation. The relationship is as follows:

$$SSTo = SSTr + SSE \tag{6.1.1}$$

We now partition the variation even more to reflect the main effects (Factor A and Factor B) and the interaction term:

$$SSTo = SSA + SSB + SSAB + SSE \tag{6.1.2}$$

where

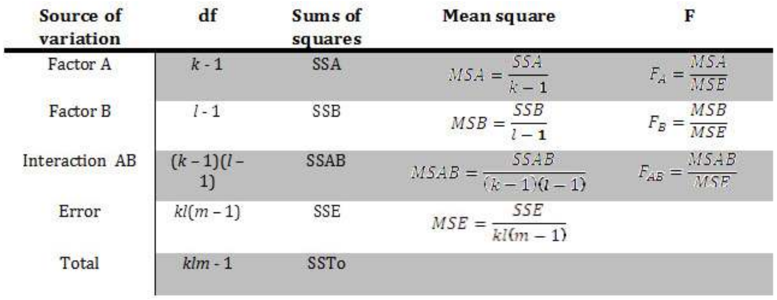
1. SSTo is the total sums of squares, with the associated degrees of freedom *klm* – 1
2. SSA is the factor A main effect sums of squares, with associated degrees of freedom *k* – 1
3. SSB is the factor B main effect sums of squares, with associated degrees of freedom *l* – 1
4. SSAB is the interaction sum of squares, with associated degrees of freedom (*k* – 1)(*l* – 1)
5. SSE is the error sum of squares, with associated degrees of freedom *kl*(*m* – 1)

As we saw in the previous chapter, the magnitude of the SSE is related entirely to the amount of underlying variability in the distributions being sampled. It has nothing to do with values of the various true average responses. SSAB reflects in part underlying variability, but its value is also affected by whether or not there is an interaction between the factors; the greater the interaction, the greater the value of SSAB.

The following ANOVA table illustrates the relationship between the sums of squares for each component and the resulting F-statistic for testing the three null and alternative hypotheses for a two-way ANOVA.

1. H_0: There is no interaction between factors
   H_1: There is a significant interaction between factors
2. H_0: There is no effect of Factor A on the response variable
   H_1: There is an effect of Factor A on the response variable
3. H_0: There is no effect of Factor B on the response variable
   H_1: There is an effect of Factor B on the response variable

If there is a significant interaction, then ignore the following two sets of hypotheses for the main effects. A significant interaction tells you that the change in the true average response for a level of Factor A depends on the level of Factor B. The effect of simultaneous changes cannot be determined by examining the main effects separately. If there is NOT a significant interaction, then proceed to test the main effects. The Factor A sums of squares will reflect random variation and any differences between the true average responses for different levels of Factor A. Similarly, Factor B sums of squares will reflect random variation and the true average responses for the different levels of Factor B.



Source of variation	df	Sums of squares	Mean square	F
Factor A	$k - 1$	SSA	$MSA = \frac{SSA}{k-1}$	$F_A = \frac{MSA}{MSE}$
Factor B	$l - 1$	SSB	$MSB = \frac{SSB}{l-1}$	$F_B = \frac{MSB}{MSE}$
Interaction AB	$(k-1)(l-1)$	SSAB	$MSAB = \frac{SSAB}{(k-1)(l-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Error	$kl(m-1)$	SSE	$MSE = \frac{SSE}{kl(m-1)}$	
Total	$klm - 1$	SSTo		

Table 2. Two-way ANOVA table.

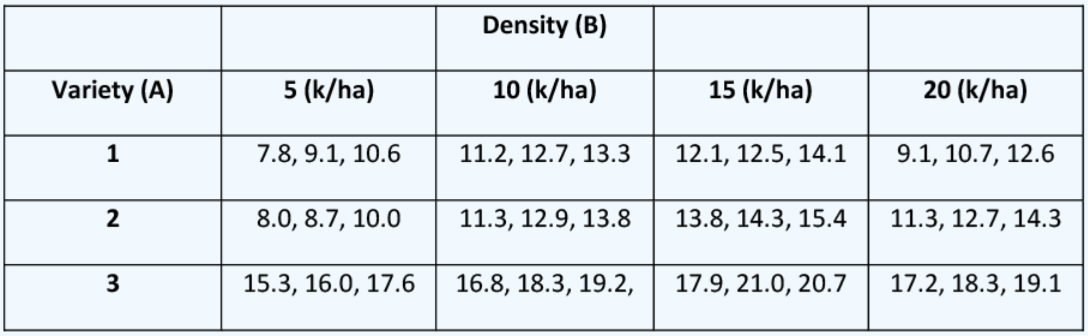
Each of the five sources of variation, when divided by the appropriate degrees of freedom (df), provides an estimate of the variation in the experiment. The estimates are called **mean squares** and are displayed along with their respective sums of squares and df in the analysis of variance table. In one-way ANOVA, the mean square error (MSE) is the best estimate of σ^2 (the population variance) and is the denominator in the F-statistic. In a two-way ANOVA, it is still the best estimate of σ^2. Notice that in each case, the MSE is the denominator in the test statistic and the numerator is the mean sum of squares for each main factor and interaction term. The F-statistic is found in the final column of this table and is used to answer the three alternative hypotheses. Typically, the p-values associated with each F-statistic are also presented in an ANOVA table. You will use the **Decision Rule** to determine the outcome for each of the three pairs of hypotheses.

If the p-value is smaller than α (level of significance), you will reject the null hypothesis.

When we conduct a two-way ANOVA, we always first test the hypothesis regarding the interaction effect. If the null hypothesis of no interaction is rejected, we do NOT interpret the results of the hypotheses involving the main effects. If the interaction term is NOT significant, then we examine the two main effects separately. Let’s look at an example.

Example 6.1.2:

An experiment was carried out to assess the effects of soy plant variety (factor A, with $k = 3$ levels) and planting density (factor B, with $l = 4$ levels – 5, 10, 15, and 20 thousand plants per hectare) on yield. Each of the 12 treatments ($k * l$) was randomly applied to $m = 3$ plots ($klm = 36$ total observations). Use a two-way ANOVA to assess the effects at a 5% level of significance.



		Density (B)		
Variety (A)	**5 (k/ha)**	**10 (k/ha)**	**15 (k/ha)**	**20 (k/ha)**
1	7.8, 9.1, 10.6	11.2, 12.7, 13.3	12.1, 12.5, 14.1	9.1, 10.7, 12.6
2	8.0, 8.7, 10.0	11.3, 12.9, 13.8	13.8, 14.3, 15.4	11.3, 12.7, 14.3
3	15.3, 16.0, 17.6	16.8, 18.3, 19.2,	17.9, 21.0, 20.7	17.2, 18.3, 19.1

Table 3. Observed data for three varieties of soy plants at four densities.

It is always important to look at the sample average yields for each treatment, each level of factor A, and each level of factor B.

Table 4. Summary table.

Density					
Variety	5	10	15	20	Sample average yield for each level of factor A
1	9.17	12.40	12.90	10.80	11.32
2	8.90	12.67	14.50	12.77	12.21
3	16.30	18.10	19.87	18.20	18.12
Sample average yield for each level of factor B	11.46	14.39	15.77	13.92	13.88

For example, 11.32 is the average yield for variety #1 over all levels of planting densities. The value 11.46 is the average yield for plots planted with 5,000 plants across all varieties. The grand mean is 13.88. The ANOVA table is presented next.

Table 5. Two-way ANOVA table.

Source	DF	SS	MSS	F	P
variety	2	327.774	163.887	100.48	<0.001
density	3	86.908	28.969	17.76	<0.001
variety*density	6	8.068	1.345	0.82	0.562
error	24	39.147	1.631		
total	35				

You begin with the following null and alternative hypotheses:

- H_0: There is no interaction between factors
- H_1: There is a significant interaction between factors

The F-statistic:

$$F_{AB} = \frac{MSAB}{MSE} = \frac{1.345}{1.631} = 0.82 \tag{6.1.3}$$

The p-value for the test for a significant interaction between factors is 0.562. This p-value is greater than 5% (α), therefore we fail to reject the null hypothesis. There is no evidence of a significant interaction between variety and density. So it is appropriate to carry out further tests concerning the presence of the main effects.

H_0: There is no effect of Factor A (variety) on the response variable

H_1: There is an effect of Factor A on the response variable

The F-statistic:

$$F_A = \frac{MSA}{MSE} = \frac{163.887}{1.631} = 100.48 \tag{6.1.4}$$

The p-value (<0.001) is less than 0.05 so we will reject the null hypothesis. There is a significant difference in yield between the three varieties.

- H_0: There is no effect of Factor B (density) on the response variable
- H_1: There is an effect of Factor B on the response variable

The F-statistic:

$$F_A = \frac{MSB}{MSE} = \frac{28.969}{1.631} = 17.76 \tag{6.1.5}$$

The p-value (<0.001) is less than 0.05 so we will reject the null hypothesis. There is a significant difference in yield between the four planting densities.

6.2: Multiple Comparisons

The next step is to examine the multiple comparisons for each main effect to determine the differences. We will proceed as we did with one-way ANOVA multiple comparisons by examining the Tukey's Grouping for each main effect. For factor A, variety, the sample means, and grouping letters are presented to identify those varieties that are significantly different from other varieties. Varieties 1 and 2 are not significantly different from each other, both producing similar yields. Variety 3 produced significantly greater yields than both variety 1 and 2.

Grouping Information Using Tukey Method and 95.0% Confidence

variety	N	Mean	Grouping	
3	12	18.117	A	
2	12	12.208		B
1	12	11.317		B

Means that do not share a letter are significantly different.

Some of the densities are also significantly different. We will follow the same procedure to determine the differences.

Grouping Information Using Tukey Method and 95.0% Confidence

density	N	Mean	Grouping		
15	9	15.756	A		
10	9	14.389	A	B	
20	9	13.922		B	
5	9	11.456			C

Means that do not share a letter are significantly different.

The Grouping Information shows us that a planting density of 15,000 plants/plot results in the greatest yield. However, there is no significant difference in yield between 10,000 and 15,000 plants/plot or between 10,000 and 20,000 plants/plot. The plots with 5,000 plants/plot result in the lowest yields and these yields are significantly lower than all other densities tested.

The main effects plots also illustrate the differences in yield across the three varieties and four densities.

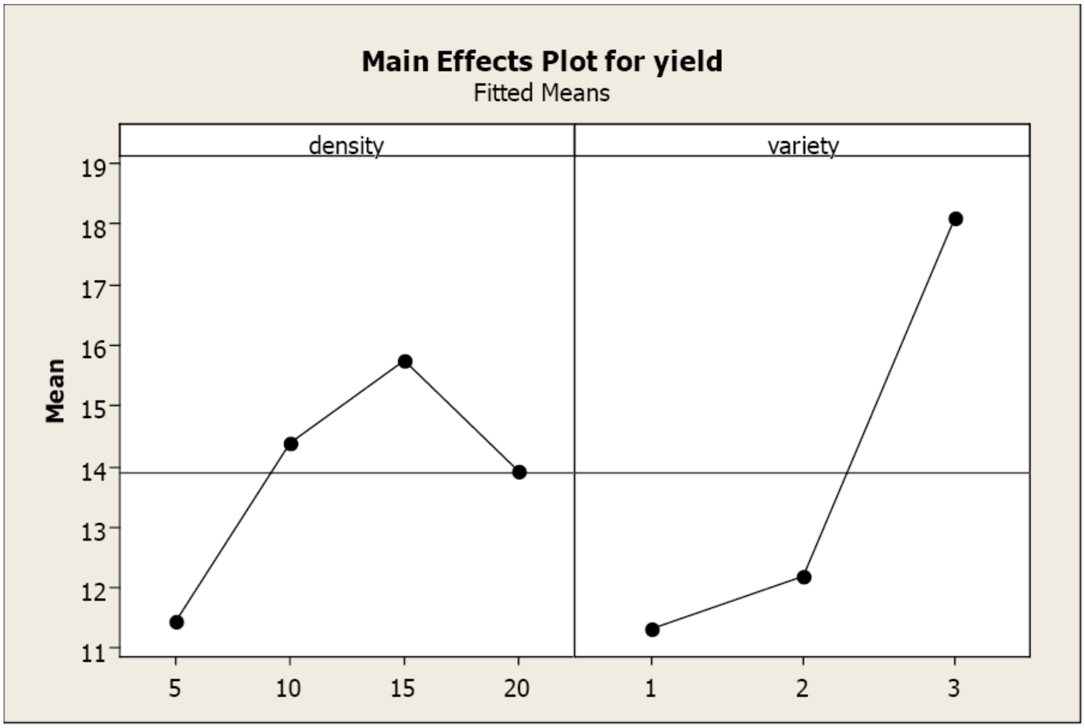


Figure 6.2.1: Main effects plots.

But what happens if there is a significant interaction between the main effects? This next example will demonstrate how a significant interaction alters the interpretation of a 2-way ANOVA.

✓ Example 6.2.1:

A researcher was interested in the effects of four levels of fertilization (control, 100 lb., 150 lb., and 200 lb.) and four levels of irrigation (A, B, C, and D) on biomass yield. The sixteen possible treatment combinations were randomly assigned to 80 plots (5 plots for each treatment). The total biomass yields for each treatment are listed below.

		Fertilizer		
Irrigation	Control	100 lb.	150 lb.	200 lb.
A	2700,2801,2720, 2390, 2890	3250, 3151, 3170, 3300, 3290	3300, 3235, 3025, 3165, 3120	3500, 3455, 3100, 3600, 3250
B	3101, 3035, 3205, 3007, 3100	2700, 2935, 2250, 2495, 2850	3050, 3110, 3033, 3195, 4250	3100, 3235, 3005, 3095, 3050
C	101, 97, 106, 142, 99	400, 302, 296, 315, 390	630, 624, 595, 675, 595	400, 325, 200, 375, 390
D	121, 174, 88, 100, 76	100, 125, 91, 222, 219	60, 28, 112, 89, 67	201, 223, 195, 120, 180

Table 6. Observed data for four irrigation levels and four fertilizer levels.

Factor A (irrigation level) has $k = 4$ levels and factor B (fertilizer) has $l = 4$ levels. There are $m = 5$ replicates and 80 total observations. This is a balanced design as the number of replicates is equal. The ANOVA table is presented next.

Two-way ANOVA table.

Source	DF	SS	MSS	F	P
fertilizer	3	1128272	376091	12.76	<0.001
irrigation	3	161776127	53925376	1830.16	<0.001

fert*irrigation	9	2088667	232074	7.88	<0.001
error	64	1885746	29465		
total	79	166878812			

We again begin with testing the interaction term. Remember, if the interaction term is significant, we ignore the main effects.

H_0: There is no interaction between factors

H_1: There is a significant interaction between factors

The F-statistic:

$$F_{AB} = \frac{MSAB}{MSE} = \frac{232074}{29465} = 7.88$$

The p-value for the test for a significant interaction between factors is <0.001. This p-value is less than 5%, therefore we reject the null hypothesis. There is evidence of a significant interaction between fertilizer and irrigation. Since the interaction term is significant, we do not investigate the presence of the main effects. We must now examine multiple comparisons for all 16 treatments (each combination of fertilizer and irrigation level) to determine the differences in yield, aided by the factor plot.

Grouping Information Using Tukey Method and 95.0% Confidence

fert	irrigation	N	Mean	Grouping			
200	A	5	3381.00	A			
150	B	5	3327.60	A			
100	A	5	3232.20	A			
150	A	5	3169.00	A			
200	B	5	3097.00	A			
C	B	5	3089.60	A			
C	A	5	2700.20		B		
100	B	5	2646.00		B		
150	C	5	623.80			C	
100	C	5	340.60			C	D
200	C	5	338.00			C	D
200	D	5	183.80				D
100	D	5	151.40				D
C	D	5	111.80				D
C	C	5	109.00				D
150	D	5	71.20				D

Means that do not share a letter are significantly different.

The factor plot allows you to visualize the differences between the 16 treatments. Factor plots can present the information two ways, each with a different factor on the x-axis. In the first plot, fertilizer level is on the x-axis. There is a clear distinction in average yields for the different treatments. Irrigation levels A and B appear to be producing greater yields across all levels of fertilizers compared to irrigation levels C and D. In the second plot, irrigation level is on the x-axis. All levels of fertilizer seem to result in greater yields for irrigation levels A and B compared to C and D.

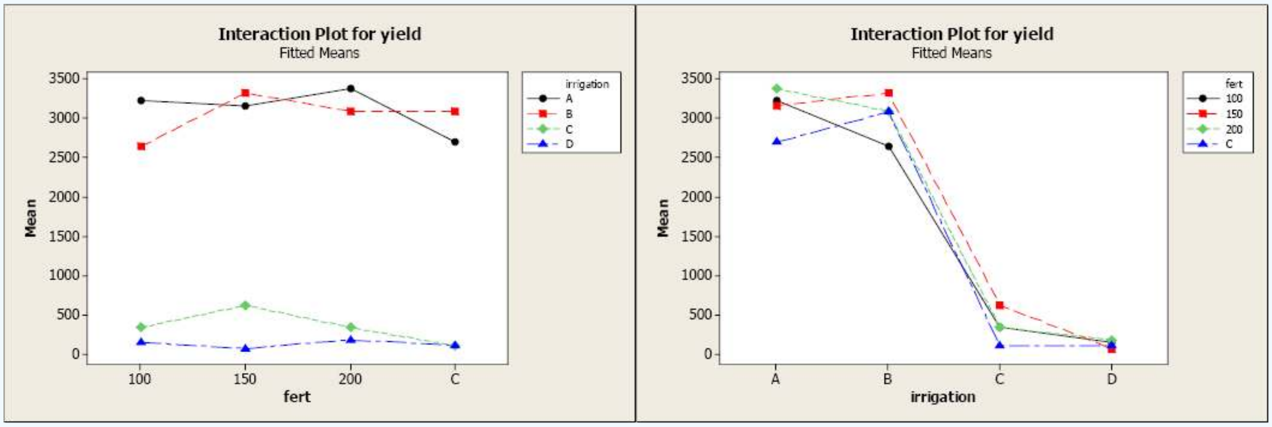


Figure 6.2.2: Interaction plots.

The next step is to use the multiple comparison output to determine where there are SIGNIFICANT differences. Let's focus on the first factor plot to do this.

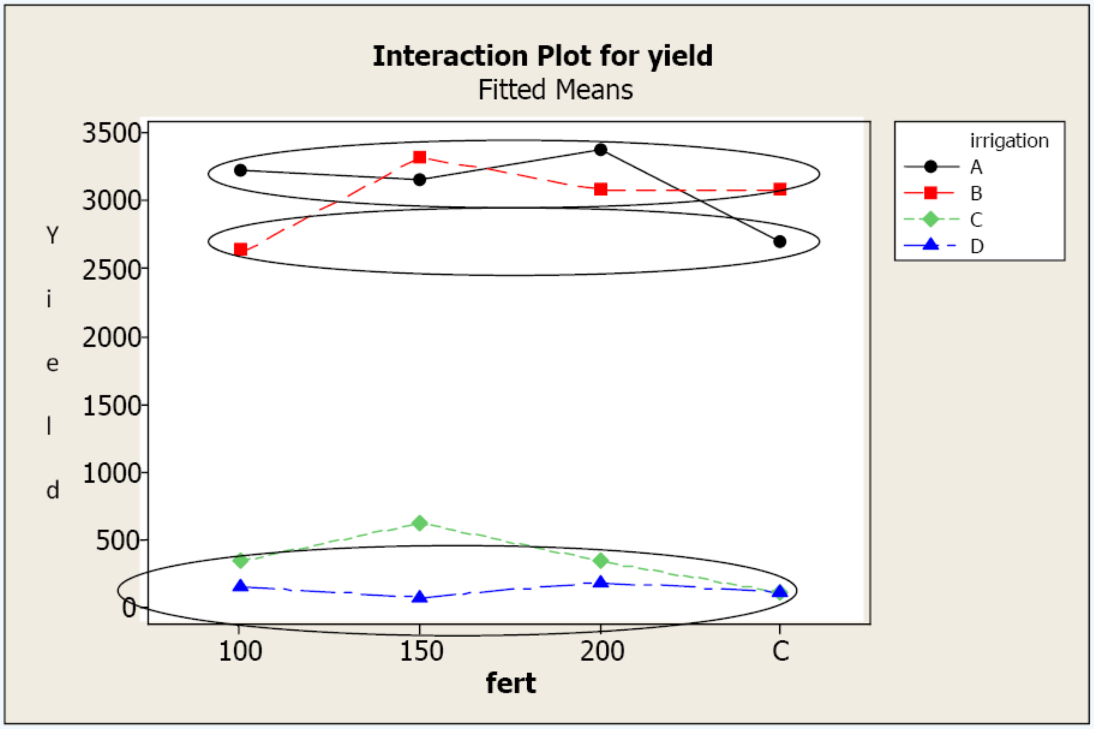


Figure 6.2.3: Interaction plot.

The Grouping Information tells us that while irrigation levels A and B look similar across all levels of fertilizer, only treatments A-100, A-150, A-200, B-control, B-150, and B-200 are statistically similar (upper circle). Treatment B-100 and A-control also result in similar yields (middle circle) and both have significantly lower yields than the first group.

Irrigation levels C and D result in the lowest yields across the fertilizer levels. We again refer to the Grouping Information to identify the differences. There is no significant difference in yield for irrigation level D over any level of fertilizer. Yields for D are also similar to yields for irrigation level C at 100, 200, and control levels for fertilizer (lowest circle). Irrigation level C at 150 level fertilizer results in significantly higher yields than any yield from irrigation level D for any fertilizer level, however, this yield is still significantly smaller than the first group using irrigation levels A and B.

Interpreting Factor Plots

When the interaction term is significant the analysis focuses solely on the treatments, not the main effects. The factor plot and grouping information allow the researcher to identify similarities and differences, along with any trends or patterns. The following series of factor plots illustrate some true average responses in terms of interactions and main effects.

This first plot clearly shows a significant interaction between the factors. The change in response when level B changes, depends on level A.

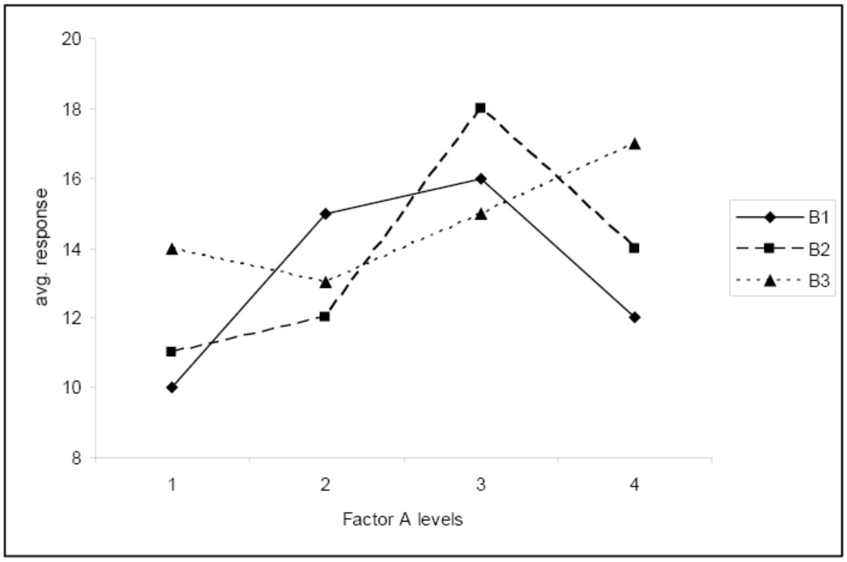


Figure 6.2.4: Interaction plot.

9609.png

Figure 5.

The second plot shows no significant interaction. The change in response for the level of factor A is the same for each level of factor B.

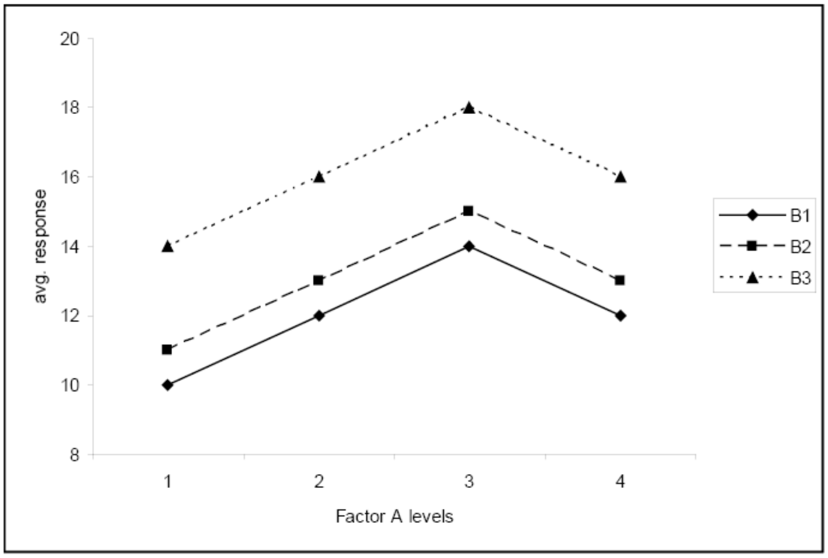


Figure 6.2.5: Figure 6.2.4: Interaction plot.

The third plot shows no significant interaction and shows that the average response does not depend on the level of factor A.

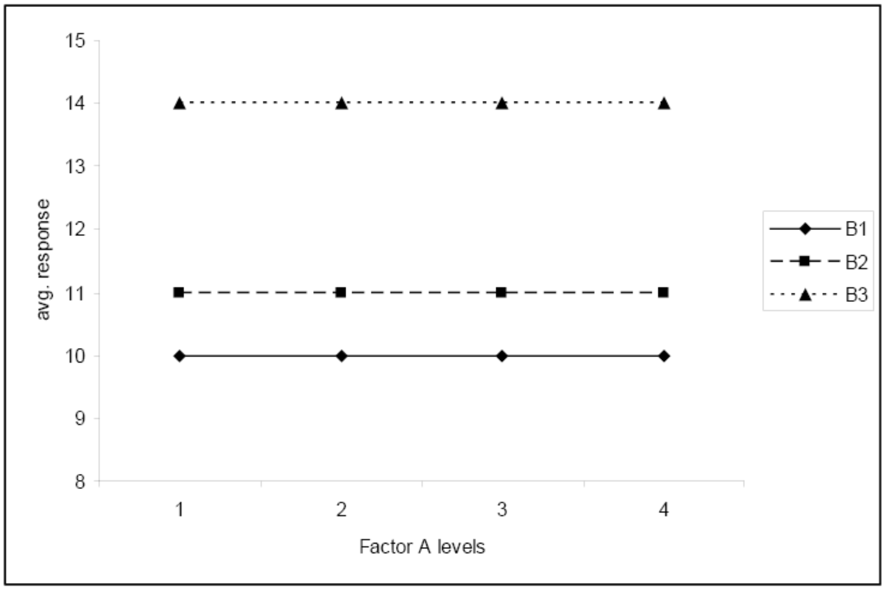


Figure 6.2.6: Figure 6.2.4: Interaction plot.

This fourth plot again shows no significant interaction and shows that the average response does not depend on the level of factor B.

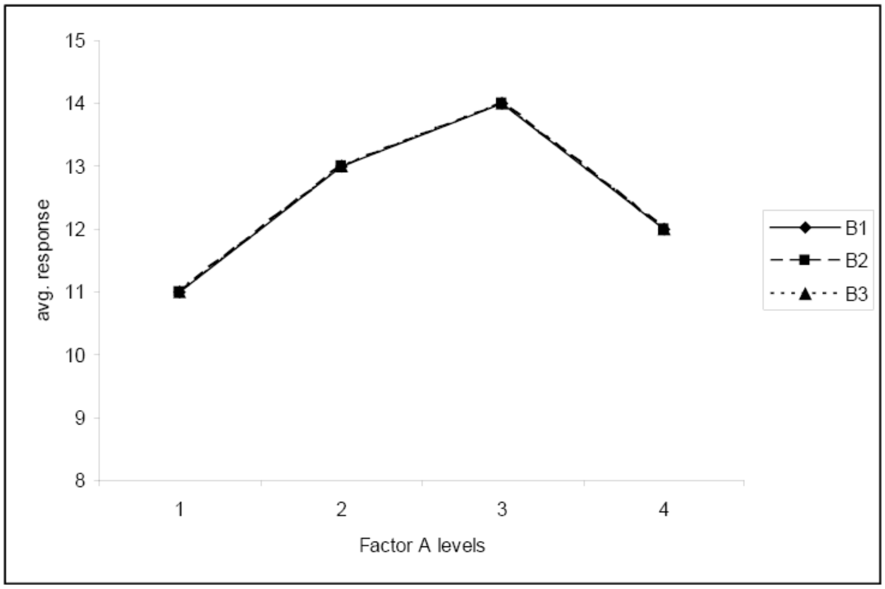


Figure 6.2.7: Figure 6.2.4: Interaction plot.

This final plot illustrates no interaction and neither factor has any effect on the response.

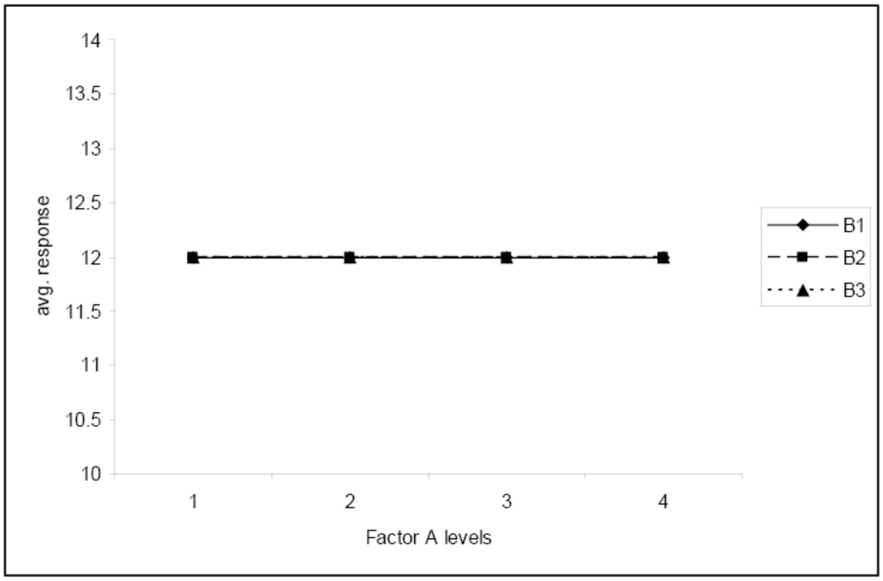


Figure 6.2.8: Figure 6.2.4: Interaction plot.

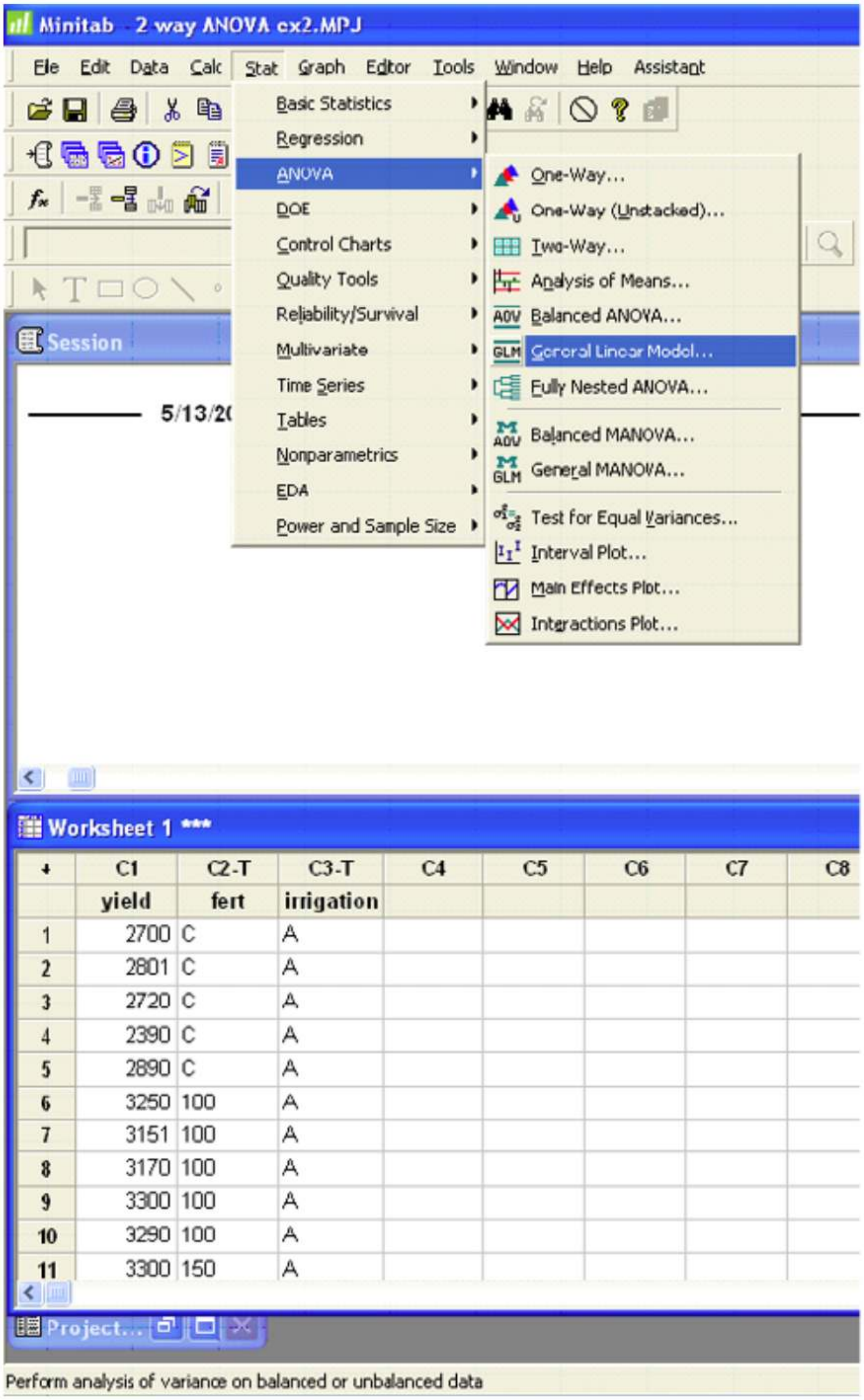
6.3: Summary And Software Solution

Summary

Two-way analysis of variance allows you to examine the effect of two factors simultaneously on the average response. The interaction of these two factors is always the starting point for two-way ANOVA. If the interaction term is significant, then you will ignore the main effects and focus solely on the unique treatments (combinations of the different levels of the two factors). If the interaction term is not significant, then it is appropriate to investigate the presence of the main effect of the response variable separately.

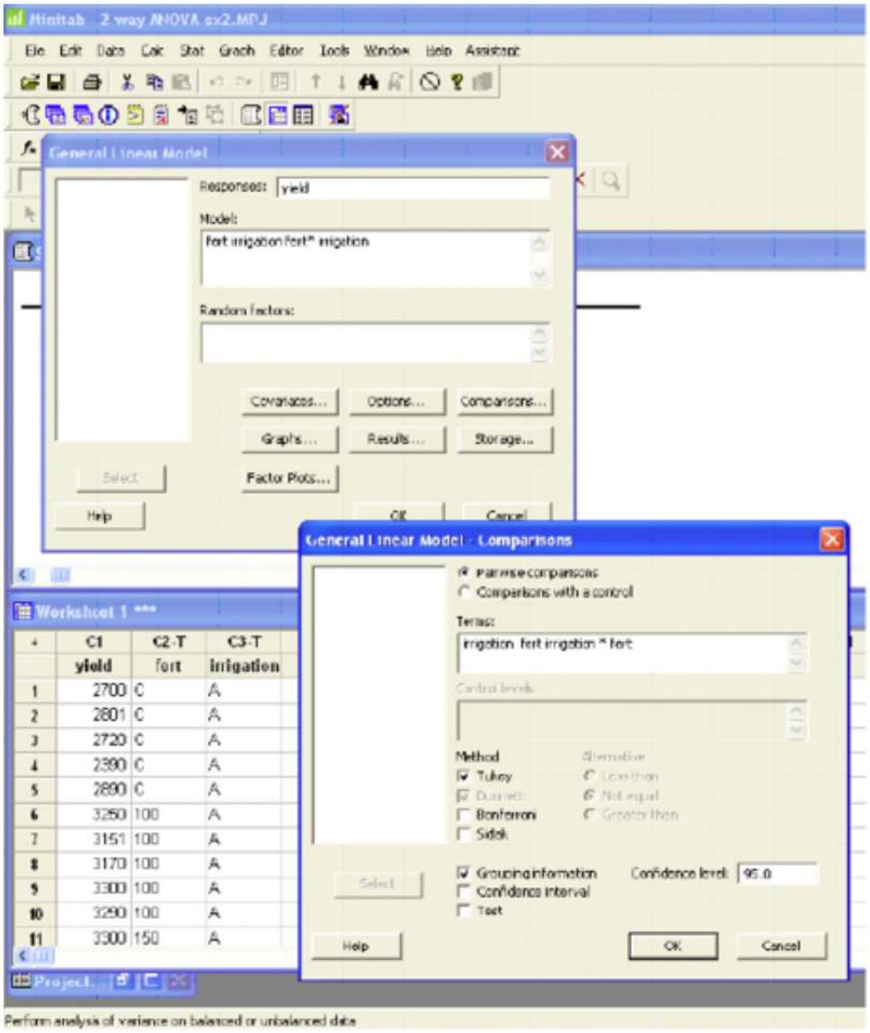
Software Solutions

Minitab



↓	C1	C2-T	C3-T	C4	C5	C6	C7	C8
	yield	fert	irrigation					
1	2700	C	A					
2	2801	C	A					
3	2720	C	A					
4	2390	C	A					
5	2890	C	A					
6	3250	100	A					
7	3151	100	A					
8	3170	100	A					
9	3300	100	A					
10	3290	100	A					
11	3300	150	A					

Perform analysis of variance on balanced or unbalanced data


Minitab - 2 way ANOVA ex2.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
General Linear Model
Covariates... Options... Comparisons...
Graphs... Results... Storage...
Factor Plots...
Select
Help
OK
Cancel
General Linear Model - Comparisons
Select
Help
OK
Cancel
Worksheet 1 ***
Project...

Responses: yield

Model: fert irrigation fert* irrigation

Random factors:

● Pairwise comparisons
○ Comparisons with a control

Terms: irrigation fert irrigation * fert

Control levels:

Method
☑ Tukey
☑ Dunnett
☐ Bonferroni
☐ Sidak

Alternative
○ Less than
● Not equal
○ Greater than

☑ Grouping information
☐ Confidence interval
☐ Test

Confidence level: 95.0

	C1	C2-T	C3-T
	yield	fert	irrigation
1	2700	C	A
2	2801	C	A
3	2720	C	A
4	2390	C	A
5	2890	C	A
6	3250	100	A
7	3151	100	A
8	3170	100	A
9	3300	100	A
10	3290	100	A
11	3300	150	A

Perform analysis of variance on balanced or unbalanced data

General Linear Model: yield vs. fert, irrigation

Factor	Type	Levels	Values			
fert	fixed	4	100,	150,	200,	C
irrigation	fixed	4	A,	B,	C,	D

Analysis of Variance for Yield, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
fert	3	1128272	1128272	376091	12.76	0.000
irrigation	3	161776127	161776127	53925376	1830.16	0.000
fert*irrigation	9	2088667	2088667	232074	7.88	0.000
Error	64	1885746	1885746	29465		
Total	79	166878812				
S = 171.653 R-Sq = 98.87% R-Sq(adj) = 98.61%						

Unusual Observations for yield

Obs	yield	Fit	SE	Fit	Residual	St	Resid
	4	2390.00	2700.20	76.77	-310.20	-2.02	R
	28	2250.00	2646.00	76.77	-396.00	-2.58	R
	35	4250.00	3327.60	76.77	922.40	6.01	R

R denotes an observation with a large standardized residual.

Grouping Information Using Tukey Method and 95.0% Confidence

irrigation	N	Mean	Grouping			
A	20	3120.60	A			
B	20	3040.05	A			
C	20	352.85		B		
D	20	129.55			C	

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

fert	N	Mean	Grouping			
150	20	1797.90	A			
200	20	1749.95	A			
100	20	1592.55		B		
C	20	1502.65		B		

Means that do not share a letter are significantly different.

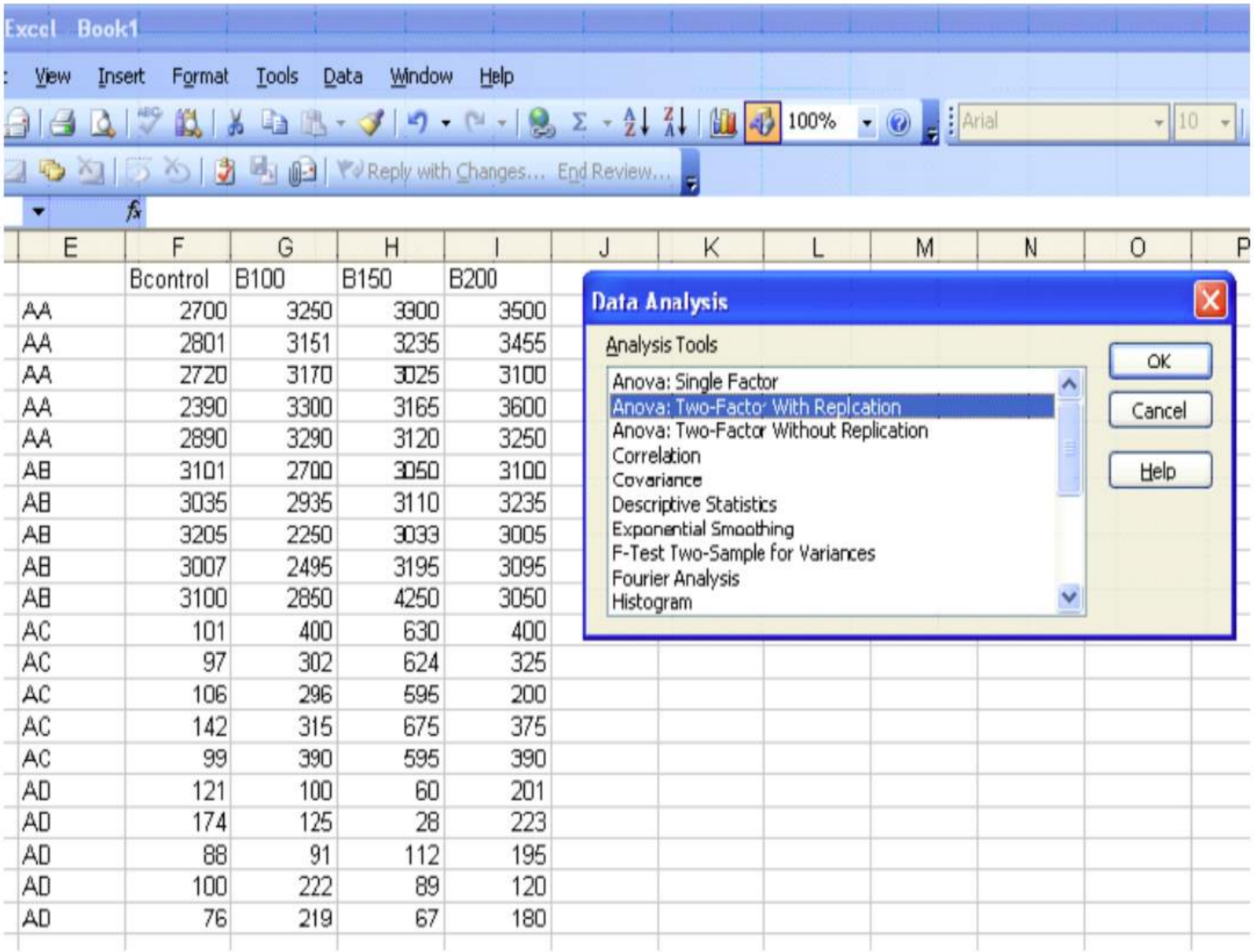
Grouping Information Using Tukey Method and 95.0% Confidence

fert	irrigation	N	Mean	Grouping			
200	A	5	3381.00	A			
150	B	5	3327.60	A			
100	A	5	3232.20	A			
150	A	5	3169.00	A			
200	B	5	3097.00	A			
C	B	5	3089.60	A			
C	A	5	2700.20		B		
100	B	5	2646.00		B		
150	C	5	623.80			C	
100	C	5	340.60			C	D
200	C	5	338.00			C	D

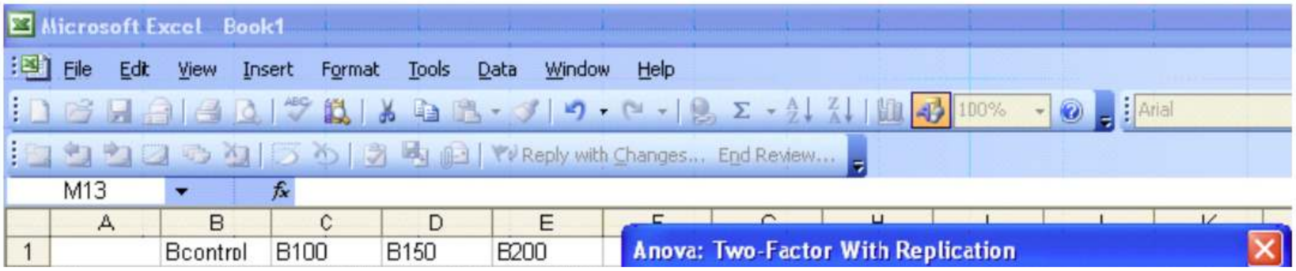
200	D	5	183.80				D
100	D	5	151.40				D
C	D	5	111.80				D
C	C	5	109.00				D
150	D	5	71.20				D

Means that do not share a letter are significantly different.

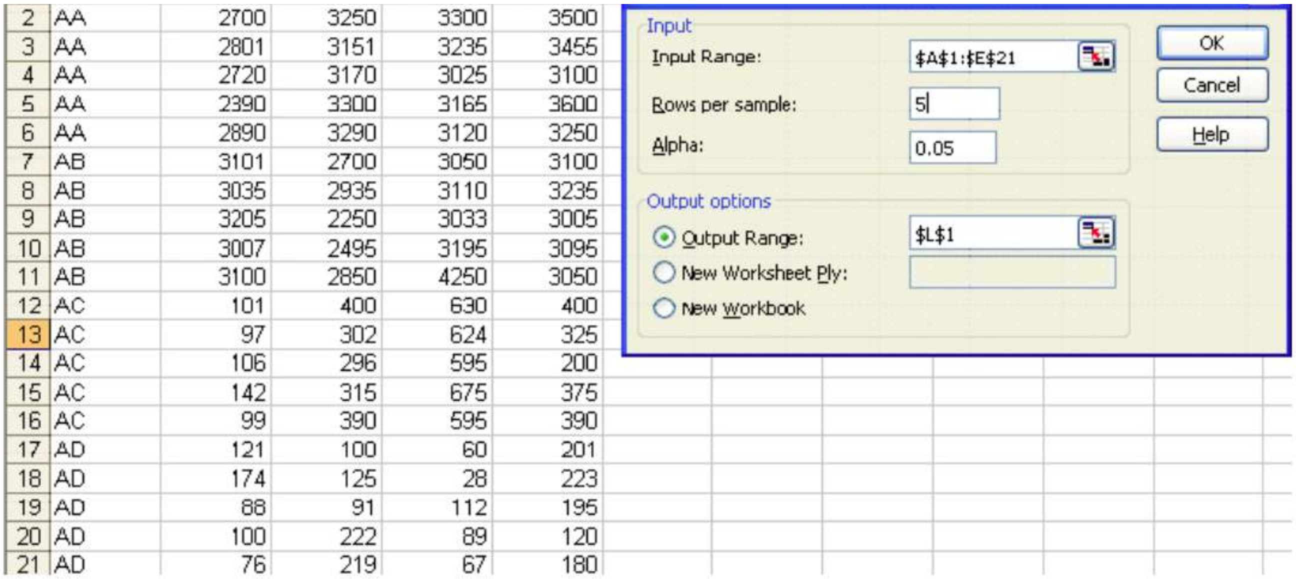
Excel


Excel Book1
View Insert Format Tools Data Window Help
100%
Arial
10
Reply with Changes... End Review...
Data Analysis
Analysis Tools
Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
OK
Cancel
Help

E	F	G	H	I
	Bcontrol	B100	B150	B200
AA	2700	3250	3300	3500
AA	2801	3151	3235	3455
AA	2720	3170	3025	3100
AA	2390	3300	3165	3600
AA	2890	3290	3120	3250
AB	3101	2700	3050	3100
AB	3035	2935	3110	3235
AB	3205	2250	3033	3005
AB	3007	2495	3195	3095
AB	3100	2850	4250	3050
AC	101	400	630	400
AC	97	302	624	325
AC	106	296	595	200
AC	142	315	675	375
AC	99	390	595	390
AD	121	100	60	201
AD	174	125	28	223
AD	88	91	112	195
AD	100	222	89	120
AD	76	219	67	180


Microsoft Excel Book1
File Edit View Insert Format Tools Data Window Help
100%
Arial
Reply with Changes... End Review...
M13
Anova: Two-Factor With Replication

	A	B	C	D	E
1		Bcontrol	B100	B150	B200



2	AA	2700	3250	3300	3500
3	AA	2801	3151	3235	3455
4	AA	2720	3170	3025	3100
5	AA	2390	3300	3165	3600
6	AA	2890	3290	3120	3250
7	AB	3101	2700	3050	3100
8	AB	3035	2935	3110	3235
9	AB	3205	2250	3033	3005
10	AB	3007	2495	3195	3095
11	AB	3100	2850	4250	3050
12	AC	101	400	630	400
13	AC	97	302	624	325
14	AC	106	296	595	200
15	AC	142	315	675	375
16	AC	99	390	595	390
17	AD	121	100	60	201
18	AD	174	125	28	223
19	AD	88	91	112	195
20	AD	100	222	89	120
21	AD	76	219	67	180

Input

Input Range: $A$1:$E$21

Rows per sample: 5

Alpha: 0.05

Output options

● Output Range: $L$1

○ New Worksheet Ply:

○ New Workbook

Anova: Two-Factor With Replication

SUMMARY	Bcontrol	B100	B150	B200	Total
AA					
Count	5	5	5	5	20
Sum	13501	16161	15845	16905	62412
Average	2700.2	3232.2	3169	3381	3120.6
Variance	35700.2	4679.2	11167.5	40930	87716.57
AB					
Count	5	5	5	5	20
Sum	15448	13230	16638	15485	60801
Average	3089.6	2646	3327.6	3097	3040.05
Variance	5839.8	76917.5	269901.3	7432.5	139929.4
AC					
Count	5	5	5	5	20
Sum	545	1703	3119	1690	7057
Average	109	340.6	623.8	338	352.85
Variance	351.5	2525.8	1079.7	6782.5	37326.03
AD					
Count	5	5	5	5	20
Sum	559	757	356	919	2591

Average	111.8	151.4	71.2	183.8	129.55	
Variance	1485.2	4135.3	997.7	1510.7	3590.366	
Total						
Count	20	20	20	20		
Sum	30053	31851	35958	34999		
Average	1502.65	1592.55	1797.9	1749.95		
Variance	2069464	1977134	2317478	2359637		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Sample	1.62E+08	3	53925376	1830.164	5.98E-62	2.748191
Columns	1128272	3	376090.7	12.76408	1.23E-06	2.748191
Interaction	2088667	9	232074.2	7.876325	1.02E-07	2.029792
Within	1885746	64	29464.78			
Total	1.67E+08	79				

CHAPTER OVERVIEW

7: Correlation and Simple Linear Regression

7.1: Correlation

In many studies, we measure more than one variable for each individual. For example, we measure precipitation and plant growth, or number of young with nesting habitat, or soil erosion and volume of water. We collect pairs of data and instead of examining each variable separately (univariate data), we want to find ways to describe **bivariate data**, in which two variables are measured on each subject in our sample. Given such data, we begin by determining if there is a relationship between these two variables. As the values of one variable change, do we see corresponding changes in the other variable?

We can describe the relationship between these two variables graphically and numerically. We begin by considering the concept of correlation.

Definition: Correlation

Correlation is defined as the statistical association between two variables.

A correlation exists between two variables when one of them is related to the other in some way. A scatterplot is the best place to start. A scatterplot (or scatter diagram) is a graph of the paired (x, y) sample data with a horizontal x-axis and a vertical y-axis. Each individual (x, y) pair is plotted as a single point.

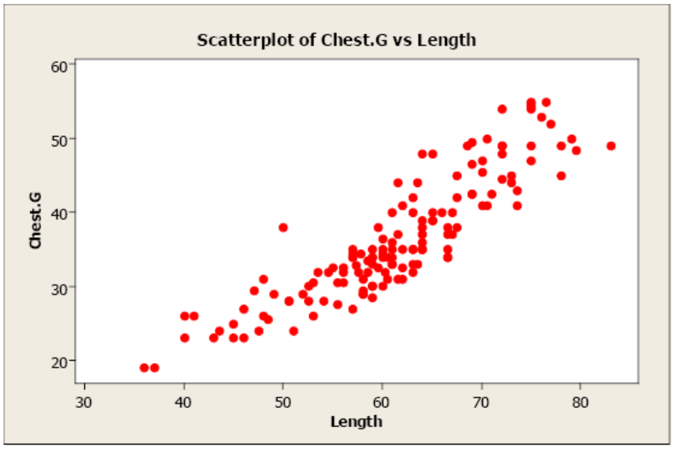


Figure 7.1.1. *Scatterplot of chest girth versus length.*

In this example, we plot bear chest girth (y) against bear length (x). When examining a scatterplot, we should study the overall pattern of the plotted points. In this example, we see that the value for chest girth does tend to increase as the value of length increases. We can see an upward slope and a straight-line pattern in the plotted data points.

A scatterplot can identify several different types of relationships between two variables.

- A relationship has **no correlation** when the points on a scatterplot do not show any pattern.
- A relationship is **non-linear** when the points on a scatterplot follow a pattern but not a straight line.
- A relationship is **linear** when the points on a scatterplot follow a somewhat straight line pattern. This is the relationship that we will examine.

Linear relationships can be either positive or negative. Positive relationships have points that incline upwards to the right. As x values increase, y values increase. As x values decrease, y values decrease. For example, when studying plants, height typically increases as diameter increases.

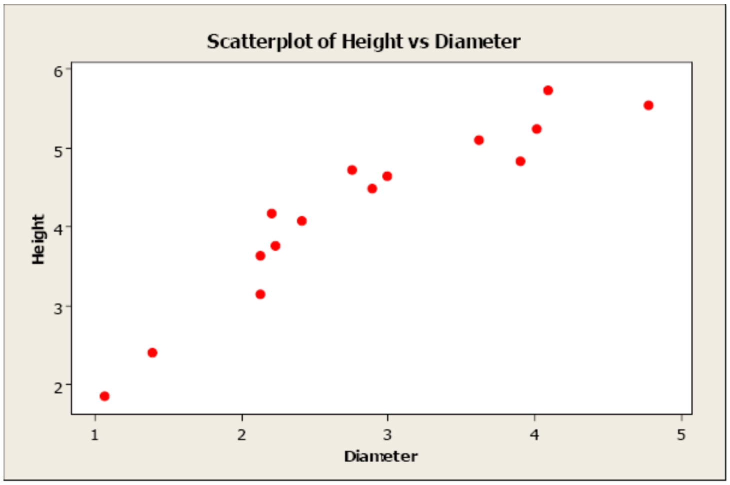


Figure 7.1.2. *Scatterplot of height versus diameter.*

Negative relationships have points that decline downward to the right. As *x* values increase, *y*values decrease. As *x* values decrease, *y* values increase. For example, as wind speed increases, wind chill temperature decreases.

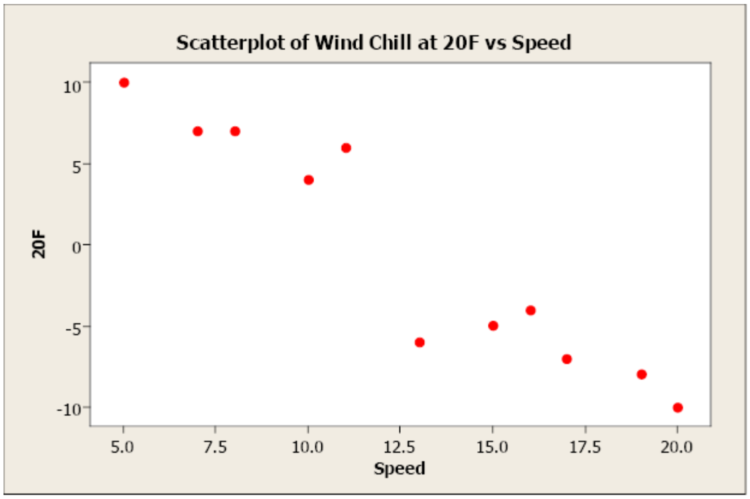


Figure 7.1.3. *Scatterplot of temperature versus wind speed.*

Non-linear relationships have an apparent pattern, just not linear. For example, as age increases height increases up to a point then levels off after reaching a maximum height.

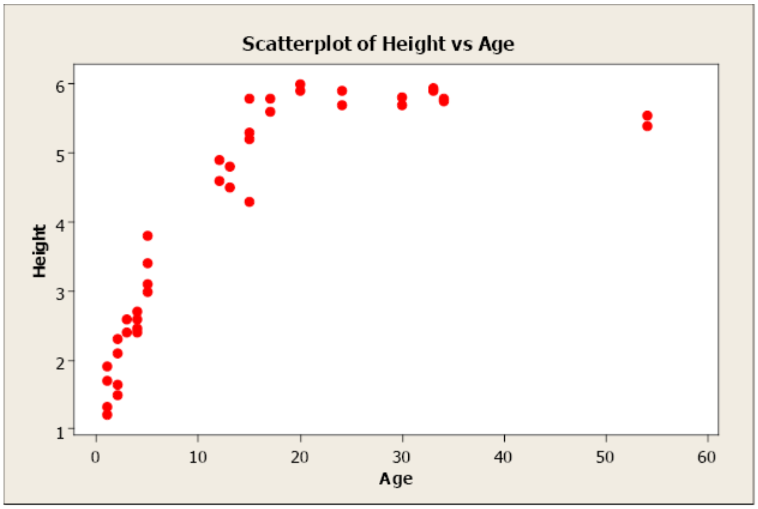


Figure 7.1.4. Scatterplot of height versus age.

When two variables have no relationship, there is no straight-line relationship or non-linear relationship. When one variable changes, it does not influence the other variable.

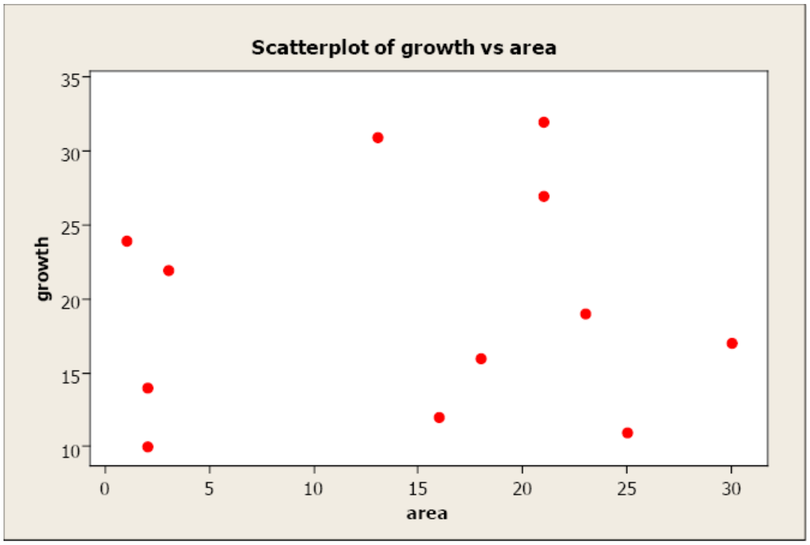


Figure 7.1.5. Scatterplot of growth versus area.

Linear Correlation Coefficient

Because visual examinations are largely subjective, we need a more precise and objective measure to define the correlation between the two variables. To quantify the strength and direction of the relationship between two variables, we use the linear correlation coefficient:

$$r=\frac{\sum \frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}}{n-1} \tag{7.1.1}$$

where $\bar{x}$ and s_x are the sample mean and sample standard deviation of the *x*'s, and $\bar{y}$ and s_y are the mean and standard deviation of the *y*'s. The sample size is *n*.

An alternate computation of the correlation coefficient is:

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{7.1.2}$$

where

$$S_{xx}=\sum x^2-\frac{(\sum x)^2}{n} \tag{7.1.3}$$

$$S_{xy}=\sum xy-\frac{(\sum x)(\sum y)}{n} \tag{7.1.4}$$

$$S_{yy}=\sum y^2-\frac{(\sum x)^2}{n} \tag{7.1.5}$$

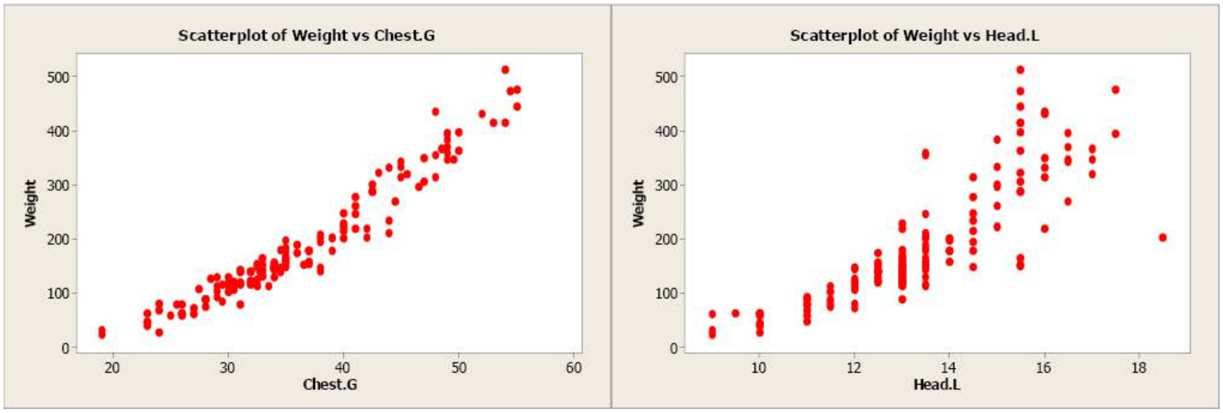
The linear correlation coefficient is also referred to as Pearson's product moment correlation coefficient in honor of Karl Pearson, who originally developed it. This statistic numerically describes how strong the straight-line or linear relationship is between the two variables and the direction, positive or negative.

The properties of "r":

- It is always between -1 and +1.

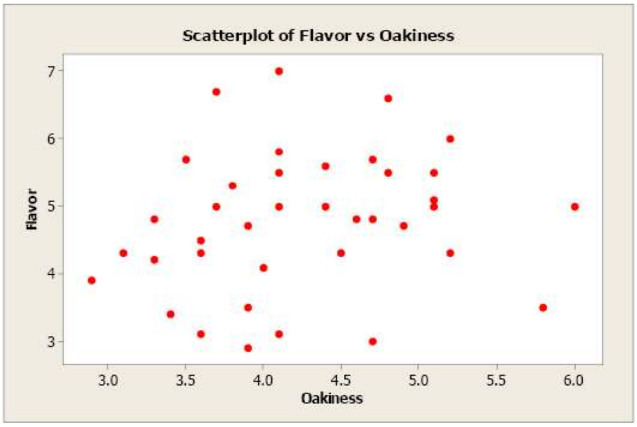
- It is a unitless measure so “r” would be the same value whether you measured the two variables in pounds and inches or in grams and centimeters.
- Positive values of “r” are associated with positive relationships.
- Negative values of “r” are associated with negative relationships.

Examples of Positive Correlation



Strong positive relationship
r = 0.96

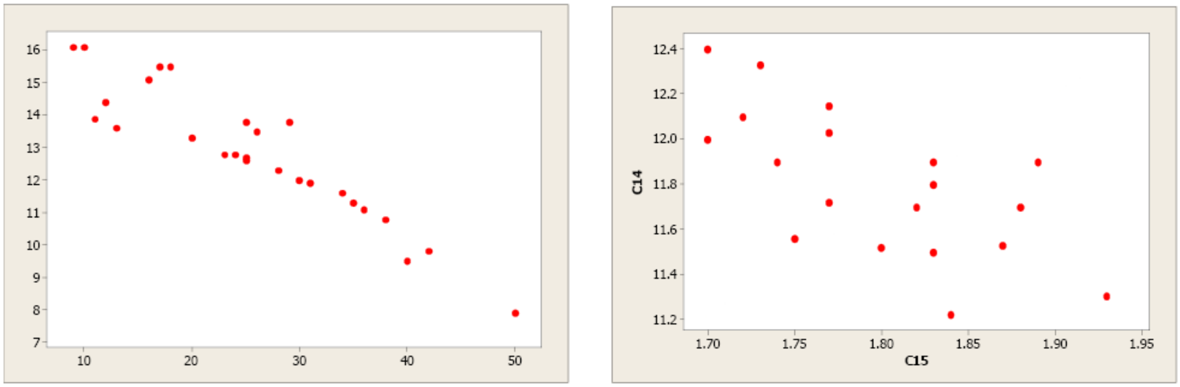
Moderate positive relationship
r = 0.67



Very weak positive relationship
r = 0.07

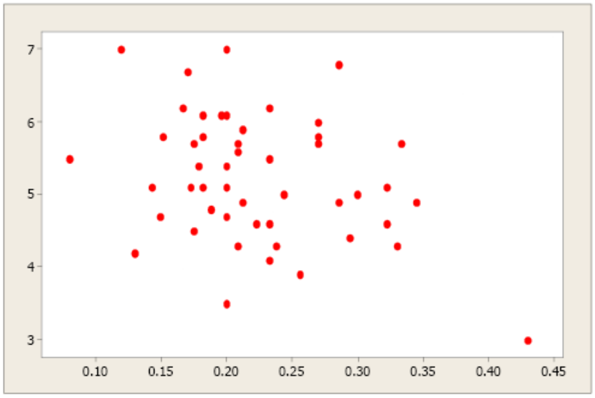
Figure $7.1.6$. *Examples of positive correlation.*

Examples of Negative Correlation



Very strong negative relationship
r = -0.93

Moderately strong negative relationship
r = -0.67

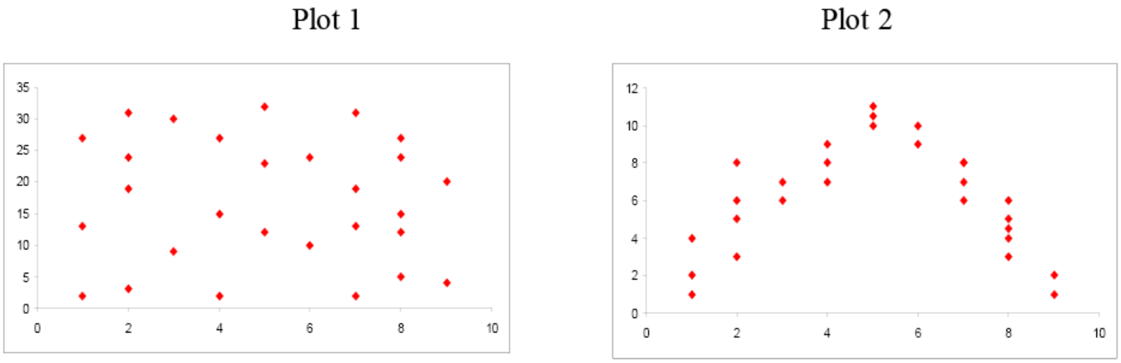


Very weak negative relationship
r = -0.13

Figure 7.1.7 *. Examples of negative correlation.*

> **Note**
>
> Correlation is **not** causation!!! Just because two variables are correlated does not mean that one variable causes another variable to change.

Examine these next two scatterplots. Both of these data sets have an r = 0.01, but they are very different. Plot 1 shows little linear relationship between *x* and *y* variables. Plot 2 shows a strong non-linear relationship. Pearson’s linear correlation coefficient only measures the strength and direction of a linear relationship. Ignoring the scatterplot could result in a serious mistake when describing the relationship between two variables.


Plot 1

Plot 2

Figure 7.1.8 *. Comparison of scatterplots.*

When you investigate the relationship between two variables, always begin with a scatterplot. This graph allows you to look for patterns (both linear and non-linear). The next step is to quantitatively describe the strength and direction of the linear relationship using “r”. Once you have established that a linear relationship exists, you can take the next step in model building.

7.2: Simple Linear Regression

Once we have identified two variables that are correlated, we would like to model this relationship. We want to use one variable as a **predictor** or **explanatory** variable to explain the other variable, the **response** or **dependent** variable. In order to do this, we need a good relationship between our two variables. The model can then be used to predict changes in our response variable. A strong relationship between the predictor variable and the response variable leads to a good model.

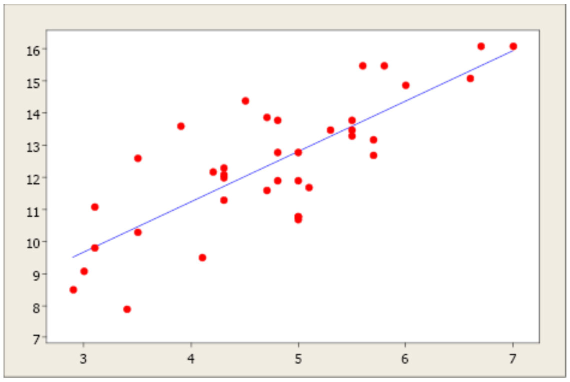


Figure 7.2.1*. Scatterplot with regression model.*

Definition: simple linear regression

A simple linear regression model is a mathematical equation that allows us to predict a response for a given predictor value.

Our model will take the form of $\hat{y} = b_0 + b_1 x$ where *b*0 is the y-intercept, *b*1 is the slope, *x* is the predictor variable, and $\hat{y}$ an estimate of the mean value of the response variable for any value of the predictor variable.

The y-intercept is the predicted value for the response (*y*) when *x* = 0. The slope describes the change in *y* for each one unit change in *x*. Let's look at this example to clarify the interpretation of the slope and intercept.

Example 7.2.1:

A hydrologist creates a model to predict the volume flow for a stream at a bridge crossing with a predictor variable of daily rainfall in inches.

Answer

$$\hat{y} = 1.6 + 29x$$

The y-intercept of 1.6 can be interpreted this way: On a day with no rainfall, there will be 1.6 gal. of water/min. flowing in the stream at that bridge crossing. The slope tells us that if it rained one inch that day the flow in the stream would increase by an additional 29 gal./min. If it rained 2 inches that day, the flow would increase by an additional 58 gal./min.

Example 7.2.2:

What would be the average stream flow if it rained 0.45 inches that day?

Answer

$$\hat{y} = 1.6 + 29x = 1.6 + 29(0.45) = 14.65 gal./min$$

The Least-Squares Regression Line (shortcut equations)

The equation is given by

$$\hat{y} = b_0 + b_1 x \tag{7.2.1}$$

where $b_1 = r\left(\frac{s_y}{s_x}\right)$ is the slope and $b_0 = \hat{y} - b_1\bar{x}$ is the y-intercept of the regression line.

An alternate computational equation for slope is:

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{S_{xy}}{S_{xx}} \tag{7.2.2}$$

This simple model is the line of best fit for our sample data. The regression line does not go through every point; instead it balances the difference between all data points and the straight-line model. The difference between the observed data value and the predicted value (the value on the straight line) is the error or **residual.** The criterion to determine the line that best describes the relation between two variables is based on the residuals.

$$Residual = Observed – Predicted \tag{7.2.3}$$

For example, if you wanted to predict the chest girth of a black bear given its weight, you could use the following model.

Chest girth = 13.2 +0.43 weight

The predicted chest girth of a bear that weighed 120 lb. is 64.8 in.

Chest girth = 13.2 + 0.43(120) = 64.8 in.

But a measured bear chest girth (observed value) for a bear that weighed 120 lb. was actually 62.1 in.

The residual would be 62.1 – 64.8 = -2.7 in.

A negative residual indicates that the model is over-predicting. A positive residual indicates that the model is under-predicting. In this instance, the model over-predicted the chest girth of a bear that actually weighed 120 lb.

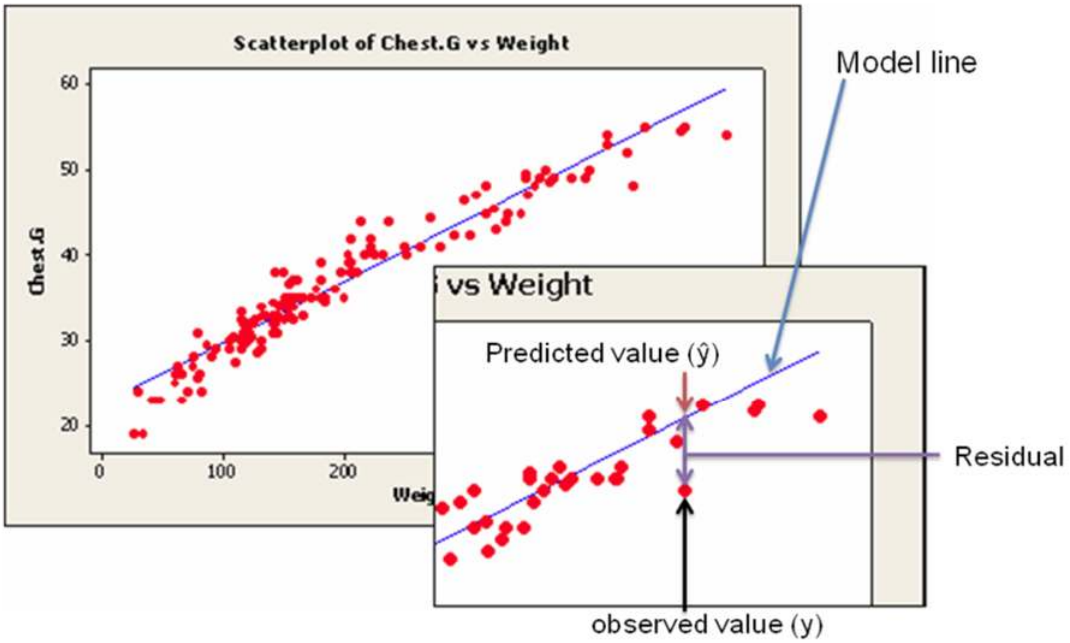


Figure 7.2.2. Scatterplot with regression model illustrating a residual value.

This random error (residual) takes into account all unpredictable and unknown factors that are not included in the model. An ordinary least squares regression line minimizes the sum of the squared errors between the observed and predicted values to create a best fitting line. The differences between the observed and predicted values are squared to deal with the positive and negative differences.

Coefficient of Determination

After we fit our regression line (compute $b0$ and $b1$), we usually wish to know how well the model fits our data. To determine this, we need to think back to the idea of analysis of variance. In ANOVA, we partitioned the variation using sums of squares so we could identify a treatment effect opposed to random variation that occurred in our data. The idea is the same for regression. We want to partition the total variability into two parts: the variation due to the regression and the variation due to random error. And we are again going to compute sums of squares to help us do this.

Suppose the total variability in the sample measurements about the sample mean is denoted by $\sum(y_i - \bar{y})^2$, called the **sums of squares of total variability about the mean (SST)**. The squared difference between the predicted value $\hat{y}$ and the sample mean is denoted by $\sum(\hat{y}_i - \bar{y})^2$, called the **sums of squares due to regression (SSR)**. The SSR represents the variability explained by the regression line. Finally, the variability which cannot be explained by the regression line is called the **sums of squares due to error (SSE)** and is denoted by $\sum(y_i - \hat{y})^2$. SSE is actually the squared residual.

SST	= SSR	+ SSE
$\sum(y_i - \bar{y})^2$	= $\sum(\hat{y}_i - \bar{y})^2$	+$\sum(\hat{y}_i - \bar{y})^2$

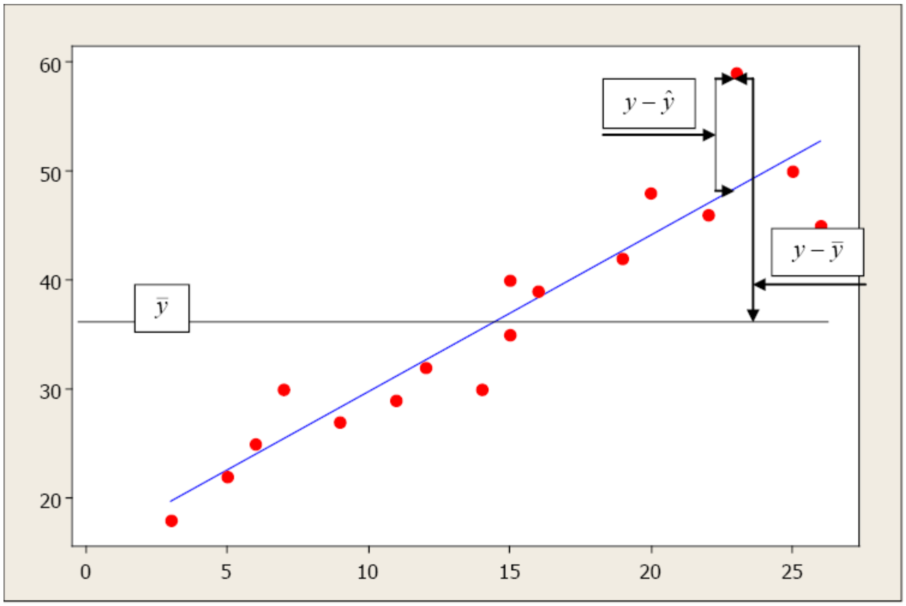


Figure $7.2.3$. An illustration of the relationship between the mean of the y's and the predicted and observed value of a specific y.

The sums of squares and mean sums of squares (just like ANOVA) are typically presented in the regression analysis of variance table. The ratio of the mean sums of squares for the regression (MSR) and mean sums of squares for error (MSE) form an F-test statistic used to test the regression model.

The relationship between these sums of square is defined as

$$Total\ Variation = Explained\ Variation + Unexplained\ Variation \tag{7.2.4}$$

The larger the explained variation, the better the model is at prediction. The larger the unexplained variation, the worse the model is at prediction. A quantitative measure of the explanatory power of a model is R^2, the Coefficient of Determination:

$$R^2 = \frac{Explained\ Variation}{Total\ Variation} \tag{7.2.5}$$

The Coefficient of Determination measures the percent variation in the response variable (*y*) that is explained by the model.

- Values range from 0 to 1.
- An R^2 close to zero indicates a model with very little explanatory power.
- An R^2 close to one indicates a model with more explanatory power.

The Coefficient of Determination and the linear correlation coefficient are related mathematically.

$$R^2 = r^2 \tag{7.2.6}$$

However, they have two very different meanings: *r* is a measure of the strength and direction of a linear relationship between two variables; *R*2 describes the percent variation in "*y*" that is explained by the model.

Residual and Normal Probability Plots

Even though you have determined, using a scatterplot, correlation coefficient and R2, that *x* is useful in predicting the value of *y*, the results of a regression analysis are valid only when the data satisfy the necessary regression assumptions.

1. The response variable (y) is a random variable while the predictor variable (x) is assumed non-random or fixed and measured without error.
2. The relationship between *y* and *x* must be linear, given by the model $\hat{y} = b_0 + b_1 x$.
3. The error of random term the values ε are independent, have a mean of 0 and a common variance σ^2, independent of *x*, and are normally distributed.

We can use **residual plots** to check for a constant variance, as well as to make sure that the linear model is in fact adequate. A residual plot is a scatterplot of the residual (= observed – predicted values) versus the predicted or fitted (as used in the residual plot) value. The center horizontal axis is set at zero. One property of the residuals is that they sum to zero and have a mean of zero. A residual plot should be free of any patterns and the residuals should appear as a random scatter of points about zero.

A residual plot with no appearance of any patterns indicates that the model assumptions are satisfied for these data.

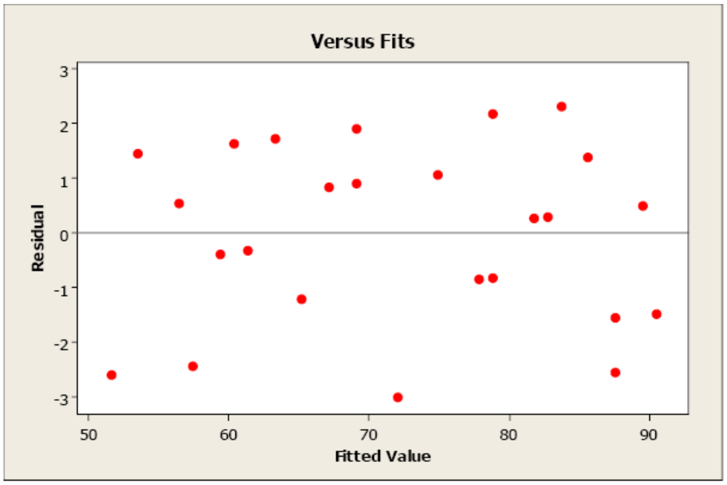


Figure $7.2.4$*. A residual plot.*

A residual plot that has a "fan shape" indicates a heterogeneous variance (non-constant variance). The residuals tend to fan out or fan in as error variance increases or decreases.

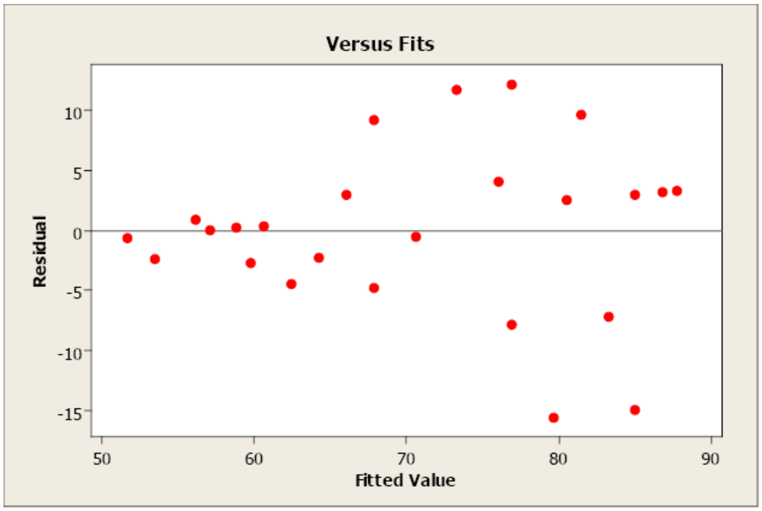


Figure 7.2.5. A residual plot that indicates a non-constant variance.

A residual plot that tends to “swoop” indicates that a linear model may not be appropriate. The model may need higher-order terms of *x*, or a non-linear model may be needed to better describe the relationship between *y* and *x*. Transformations on *x* or *y* may also be considered.

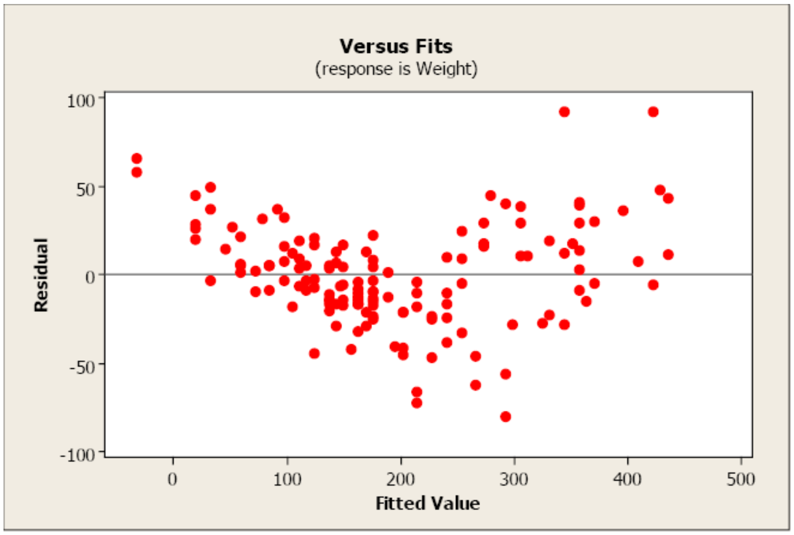


Figure 7.2.6. A residual plot that indicates the need for a higher order model.

A **normal probability plot** allows us to check that the errors are normally distributed. It plots the residuals against the expected value of the residual as if it had come from a normal distribution. Recall that when the residuals are normally distributed, they will follow a straight-line pattern, sloping upward.

This plot is not unusual and does not indicate any non-normality with the residuals.

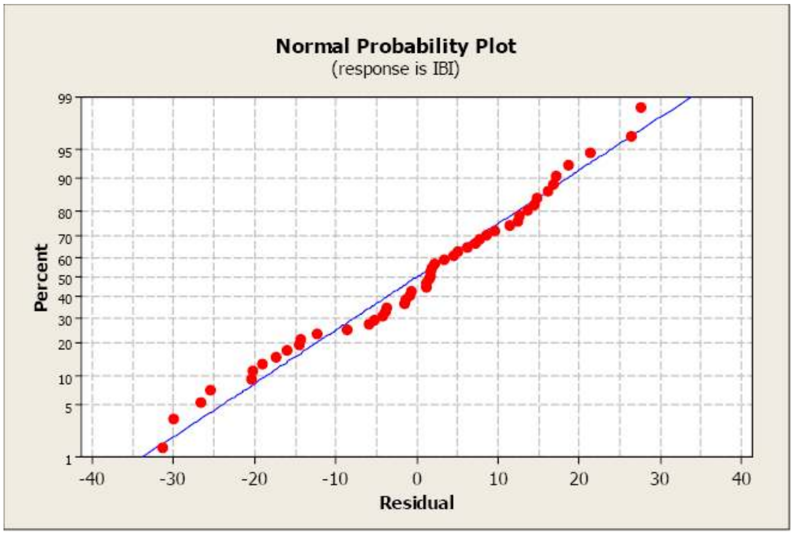


Figure 7.2.7. A normal probability plot.

This next plot clearly illustrates a non-normal distribution of the residuals.

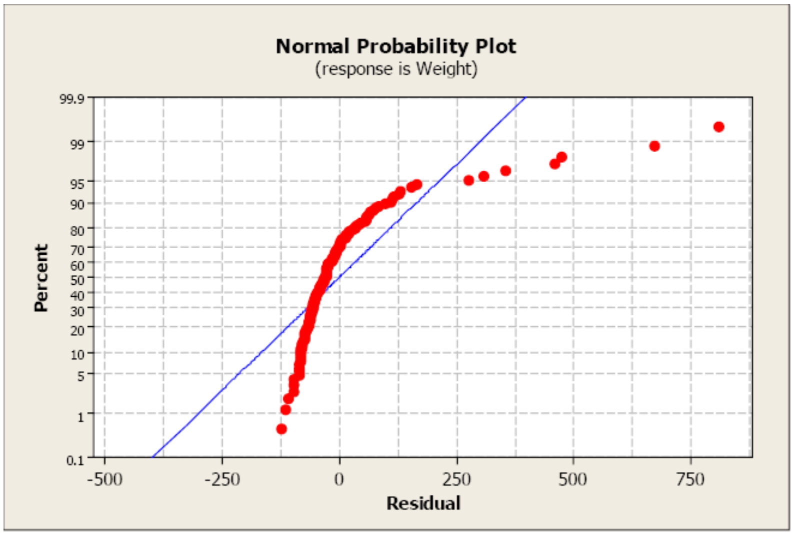


Figure $7.2.8$. A normal probability plot, which illustrates non-normal distribution.

The most serious violations of normality usually appear in the tails of the distribution because this is where the normal distribution differs most from other types of distributions with a similar mean and spread. Curvature in either or both ends of a normal probability plot is indicative of nonnormality.

7.3: Population Model

Our regression model is based on a sample of *n* bivariate observations drawn from a larger population of measurements.

$$\hat{y} = b_0 + b_1 x \tag{7.3.1}$$

We use the means and standard deviations of our sample data to compute the slope (*b*1) and y-intercept (*b*0) in order to create an ordinary least-squares regression line. But we want to describe the relationship between *y* and *x* in the population, not just within our sample data. We want to construct a **population model**. Now we will think of the least-squares line computed from a sample as an estimate of the true regression line for the population.

Definition: The Population Model

$\mu_y = \beta_0 + \beta_1 x$, where μ_y is the population mean response, β_0 is the y-intercept, and $beta_1$ is the slope for the population model.

In our population, there could be many different responses for a value of *x*. In simple linear regression, the model assumes that for each value of *x* the observed values of the response variable *y* are normally distributed with a mean that depends on *x*. We use μy to represent these means. We also assume that these means all lie on a straight line when plotted against *x* (a line of means).

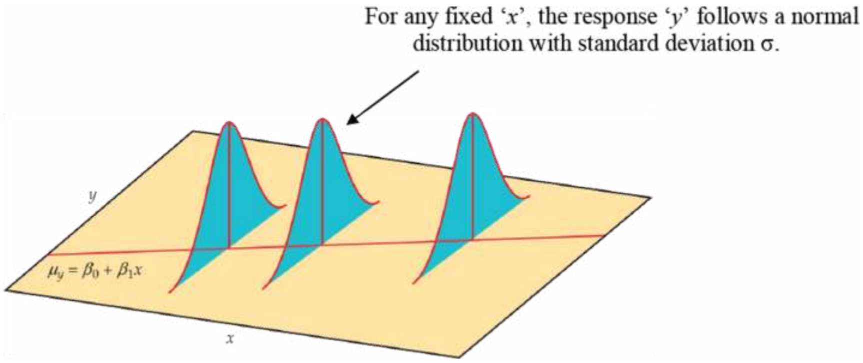

For any fixed 'x', the response 'y' follows a normal distribution with standard deviation σ.
y
$\mu_y = \beta_0 + \beta_1 x$
x

Figure $7.3.1$. *. The statistical model for linear regression; the mean response is a straight-line function of the predictor variable.*

The sample data then fit the statistical model:

Data = fit + residual

$$y_i = (\beta_0 + \beta_1 x_i) + \epsilon_i \tag{7.3.2}$$

where the errors (εi) are independent and normally distributed *N* (0, σ). Linear regression also assumes equal variance of *y* (σ is the same for all values of *x*). We use ε (Greek epsilon) to stand for the residual part of the statistical model. A response *y* is the sum of its mean and chance deviation εfrom the mean. The deviations ε represents the “noise” in the data. In other words, the noise is the variation in *y* due to other causes that prevent the observed (*x, y*) from forming a perfectly straight line.

The sample data used for regression are the observed values of *y* and *x*. The response *y* to a given *x*is a random variable, and the regression model describes the mean and standard deviation of this random variable *y*. The intercept β0, slope β1, and standard deviation σ of *y* are the unknown parameters of the regression model and must be estimated from the sample data.

- The value of ŷ from the least squares regression line is really a prediction of the mean value of *y* (μy) for a given value of *x*.
- The least squares regression line ($\hat{y} = b_0 + b_1 x$) obtained from sample data is the best estimate of the true population regression line
  ($\mu_y = \beta_0 + \beta_1 x$).

ŷ is an unbiased estimate for the mean response μy
***b*0 is an unbiased estimate for the intercept β0**
***b*1 is an unbiased estimate for the slope β1**

Parameter Estimation

Once we have estimates of β0 and β1 (from our sample data *b*0 and *b*1), the linear relationship determines the estimates of μy for all values of *x* in our population, not just for the observed values of *x*. We now want to use the least-squares line as a basis for inference about a population from which our sample was drawn.

Model assumptions tell us that *b*0 and *b*1 are normally distributed with means β0 and β1 with standard deviations that can be estimated from the data. Procedures for inference about the population regression line will be similar to those described in the previous chapter for means. As always, it is important to examine the data for outliers and influential observations.

In order to do this, we need to estimate σ, the regression standard error. This is the standard deviation of the model errors. It measures the variation of *y* about the population regression line. We will use the residuals to compute this value. Remember, the predicted value of *y* ($\hat{p}$) for a specific *x* is the point on the regression line. It is the unbiased estimate of the mean response (μy) for that *x*. The residual is:

residual = observed – predicted

$$\epsilon_i = y_i - \hat{y} = y_i - (b_0 + b_1 x) \tag{7.3.3}$$

The residual *e*i corresponds to model deviation ϵ_i where $\sum \epsilon_i = 0$ with a mean of 0. The regression standard error *s* is an unbiased estimate of σ.

$$s = \sqrt{\frac{\sum residual^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \tag{7.3.4}$$

The quantity *s* is the estimate of the regression standard error (σ) and s^2 is often called the mean square error (MSE). A small value of *s* suggests that observed values of *y* fall close to the true regression line and the line $\hat{y} = b_0 + b_1 x$ should provide accurate estimates and predictions.

Confidence Intervals and Significance Tests for Model Parameters

In an earlier chapter, we constructed confidence intervals and did significance tests for the population parameter μ (the population mean). We relied on sample statistics such as the mean and standard deviation for point estimates, margins of errors, and test statistics. Inference for the population parameters β0 (slope) and β1 (y-intercept) is very similar.

Inference for the slope and intercept are based on the normal distribution using the estimates *b*0 and *b*1. The standard deviations of these estimates are multiples of σ, the population regression standard error. Remember, we estimate σ with *s* (the variability of the data about the regression line). Because we use *s*, we rely on the student t-distribution with (*n* – 2) degrees of freedom.

$$\sigma_{\hat{\beta}_0} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \tag{7.3.5}$$

The standard error for estimate of β_0

$$\sigma_{\hat{\beta}_1} = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \tag{7.3.6}$$

The standard error for estimate of β_1

We can construct confidence intervals for the regression slope and intercept in much the same way as we did when estimating the population mean.

A **confidence interval** for $\beta_0 : b_0 \pm t_{\alpha/2} SE_{b_0}$

A **confidence interval** for $\beta_1 : b_1 \pm t_{\alpha/2} SE_{b_1}$

where SE_{b_0} and SE_{b_1} are the standard errors for the y-intercept and slope, respectively.

We can also test the hypothesis $H_0 : \beta_1 = 0$. When we substitute $\beta_1 = 0$ in the model, the x-term drops out and we are left with $\mu_y = \beta_0$. This tells us that the mean of *y* does NOT vary with *x*. In other words, there is no straight line relationship between *x* and *y* and the regression of *y* on *x* is of no value for predicting *y*.

Hypothesis test for β_1

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

The test statistic is $t = b_1 / SE_{b_1}$

We can also use the F-statistic (MSR/MSE) in the regression ANOVA table*

*Recall that t2 = F

So let’s pull all of this together in an example.

✓ Example 7.3.1:

The index of biotic integrity (IBI) is a measure of water quality in streams. As a manager for the natural resources in this region, you must monitor, track, and predict changes in water quality. You want to create a simple linear regression model that will allow you to predict changes in IBI in forested area. The following table conveys sample data from a coastal forest region and gives the data for IBI and forested area in square kilometers. Let forest area be the predictor variable (x) and IBI be the response variable (y).

IBI	Forest Area	IBI	Forest Area	IBI	Forest Area	IBI	Forest Area	IBI
47	38	41	22	61	43	71	79	84
72	9	33	25	62	47	33	79	83
21	10	23	31	18	49	59	80	82
19	10	32	32	44	49	81	86	82
72	52	80	33	30	52	71	89	86
56	14	31	33	65	52	75	90	79
49	66	78	33	78	59	64	95	67
89	17	21	39	71	63	41	95	56
43	18	43	41	60	68	82	100	85
66	21	45	43	58	75	60	100	91

Table 7.3.1. Observed data of biotic integrity and forest area.

Solution

We begin with a computing descriptive statistics and a scatterplot of IBI against Forest Area.

$\bar{x}$ = 47.42; *sx* 27.37; $\bar{y}$ = 58.80; *sy* = 21.38; r = 0.735

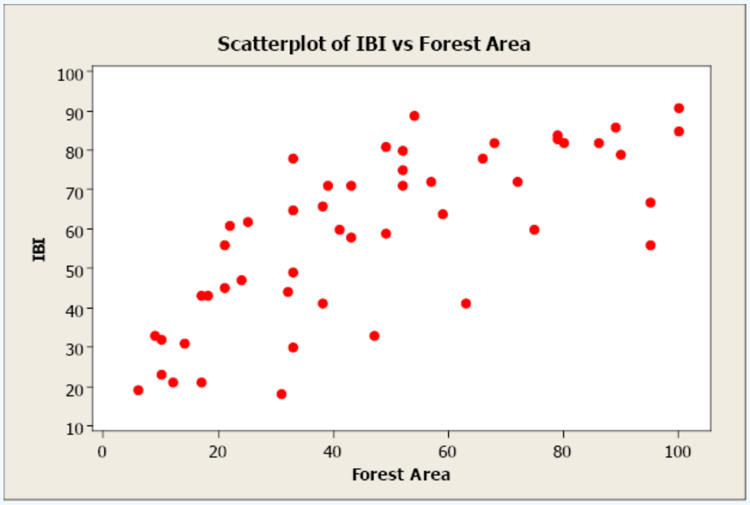


Figure 7.3.2. . Scatterplot of IBI vs. Forest Area.

There appears to be a positive linear relationship between the two variables. The linear correlation coefficient is r = 0.735. This indicates a strong, positive, linear relationship. In other words, forest area is a good predictor of IBI. Now let's create a simple linear regression model using forest area to predict IBI (response).

First, we will compute $b0$ and $b1$ using the shortcut equations.

$$b_1 = r(\frac{s_y}{s_x}) = 0.735(\frac{21.38}{27.37}) = 0.574$$

$$b_0 = \bar{y} - b_1\bar{x} = 58.80 - 0.574 \times 47.42 = 31.581$$

The regression equation is

$$\hat{y} = 31.58 + 0.574x$$

Now let's use Minitab to compute the regression model. The output appears below.

Regression Analysis: IBI versus Forest Area

The regression equation is IBI = 31.6 + 0.574 Forest Area

Predictor	Coef	SE Coef	T	P
Constant	31.583	**4.177**	7.56	0.000
Forest Area	0.57396	**0.07648**	7.50	0.000
S = 14.6505	**R-Sq = 54.0%**	R-Sq(adj) = 53.0%		

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	12089	12089	56.32	0.000
Residual Error	48	10303	**215**		
Total	49	22392			

The estimates for β0 and β1 are 31.6 and 0.574, respectively. We can interpret the y-intercept to mean that when there is zero forested area, the IBI will equal 31.6. For each additional square kilometer of forested area added, the IBI will increase by 0.574 units.

The coefficient of determination, R2, is 54.0%. This means that 54% of the variation in IBI is explained by this model. Approximately 46% of the variation in IBI is due to other factors or random variation. We would like R2 to be as high as possible (maximum value of 100%).

The residual and normal probability plots do not indicate any problems.

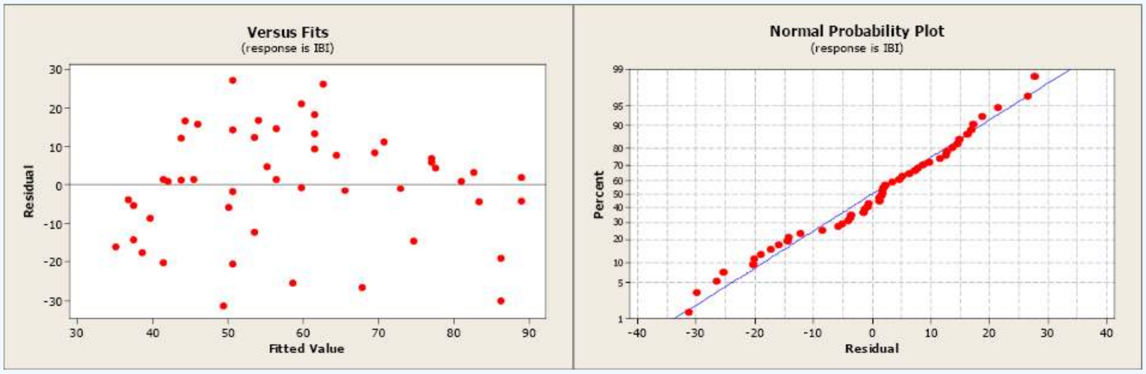


Figure 7.3.3. . A residual and normal probability plot.

The estimate of **σ**, the regression standard error, is s = 14.6505. This is a measure of the variation of the observed values about the population regression line. We would like this value to be as small as possible. The MSE is equal to 215. Remember, the $\sqrt{MSE} = s$. The standard errors for the coefficients are 4.177 for the y-intercept and 0.07648 for the slope.

We know that the values $b0$ = 31.6 and $b1$ = 0.574 are sample estimates of the true, but unknown, population parameters β0 and β1. We can construct 95% confidence intervals to better estimate these parameters. The critical value (tα/2) comes from the student t-distribution with (n – 2) degrees of freedom. Our sample size is 50 so we would have 48 degrees of freedom. The closest table value is 2.009.

95% confidence intervals for β0 and β1

$$b_0 \pm t_{\alpha/2} SE_{b_0} = 31.6 \pm 2.009(4.177) = (23.21, 39.99)$$

$$b_1 \pm t_{\alpha/2} SE_{b_1} = 0.574 \pm 2.009(0.07648) = (0.4204, 0.7277)$$

The next step is to test that the slope is significantly different from zero using a 5% level of significance.

H0: β1 =0	H1: β1 ≠0

$$t = \frac{b_1}{SE_{b_1}} = \frac{0.574}{0.07648} = 7.50523$$

We have 48 degrees of freedom and the closest critical value from the student t-distribution is 2.009. The test statistic is greater than the critical value, so we will reject the null hypothesis. The slope is significantly different from zero. We have found a statistically significant relationship between Forest Area and IBI.

The Minitab output also report the test statistic and p-value for this test.

The regression equation is IBI = 31.6 + 0.574 Forest Area				
Predictor	Coef	SE Coef	T	P
Constant	31.583	4.177	7.56	0.000
Forest Area	0.57396	0.07648	**7.50**	**0.000**
S = 14.6505	R-Sq = 54.0%	R-Sq(adj) = 53.0%		

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	12089	12089	**56.32**	**0.000**
Residual Error	48	10303	215		
Total	49	22392			

The t test statistic is 7.50 with an associated p-value of 0.000. The p-value is less than the level of significance (5%) so we will reject the null hypothesis. The slope is significantly different from zero. The same result can be found from the F-test statistic of 56.32 (7.5052 = 56.32). The p-value is the same (0.000) as the conclusion.

Confidence Interval for μ_y

Now that we have created a regression model built on a significant relationship between the predictor variable and the response variable, we are ready to use the model for

- estimating the average value of y for a given value of x
- predicting a particular value of y for a given value of x

Let's examine the first option. The sample data of n pairs that was drawn from a population was used to compute the regression coefficients $b0$ and $b1$ for our model, and gives us the average value of y for a specific value of x through our population model

$\mu_y = \beta_0 + \beta_1 x$

. For every specific value of x, there is an average y (*μy*), which falls on the straight line equation (a line of means). Remember, that there can be many different observed values of the *y* for a particular *x*, and these values are assumed to have a normal distribution with a mean equal to $\beta_0 + \beta_1 x$ and a variance of σ2. Since the computed values of *b*0 and *b*1 vary from sample to sample, each new sample may produce a slightly different regression equation. Each new model can be used to estimate a value of *y* for a value of *x*. How far will our estimator $\hat{y} = b_0 + b_1 x$ be from the true population mean for that value of *x*? This depends, as always, on the variability in our estimator, measured by the standard error.

It can be shown that the estimated value of *y* when *x* = *x*0 (some specified value of *x*), is an unbiased estimator of the population mean, and that $\hat{p}$ is normally distributed with a standard error of

(7.3.7)

SE_{\hat \mu} = s\sqrt {\frac {1}{n} + \frac {(x_0-\bar x)^2}{\sum (x_i - \bar x)^2}}\]

We can construct a confidence interval to better estimate this parameter (μy) following the same procedure illustrated previously in this chapter.

(7.3.8)

\hat {\mu_y} \pm t_{\alpha/2}SE_{\hat \mu}\]

where the critical value tα/2 comes from the student t-table with (*n* – 2) degrees of freedom.

Statistical software, such as Minitab, will compute the confidence intervals for you. Using the data from the previous example, we will use Minitab to compute the 95% confidence interval for the mean response for an average forested area of 32 km.

Predicted Values for New Observations			
New Obs Fit	SE Fit	95%	CI
1	49.9496	2.38400	(45.1562,54.7429)

If you sampled many areas that averaged 32 km. of forested area, your estimate of the average IBI would be from 45.1562 to 54.7429.

You can repeat this process many times for several different values of *x* and plot the confidence intervals for the mean response.

x	95% CI
20	(37.13, 48.88)
40	(50.22, 58.86)
60	(61.43, 70.61)
80	(70.98, 84.02)
100	(79.88, 98.07)

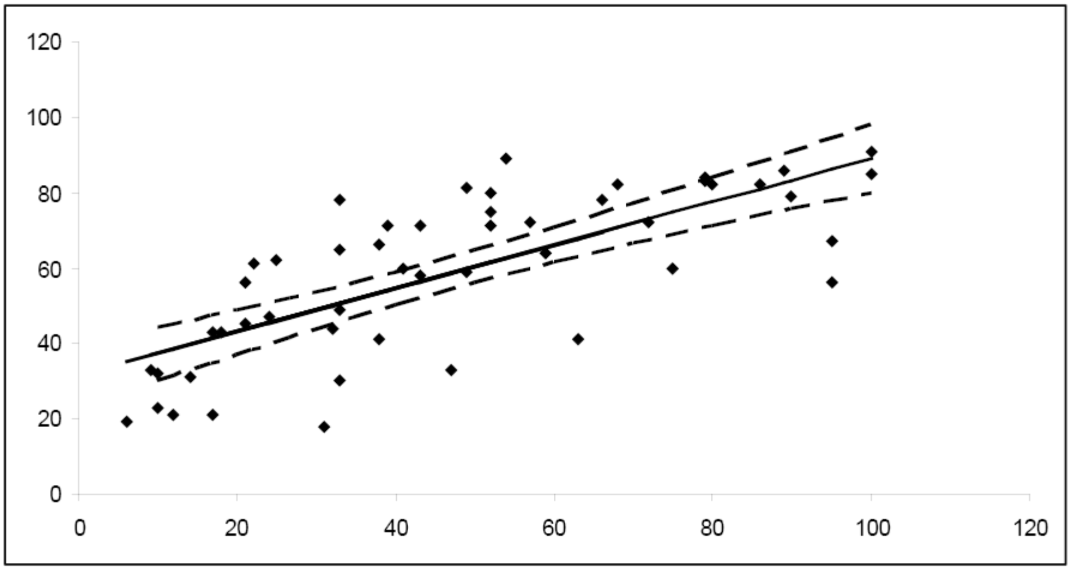


Figure $7.3.4$. 95% confidence intervals for the mean response.

Notice how the width of the 95% confidence interval varies for the different values of *x*. Since the confidence interval width is narrower for the central values of *x*, it follows that μy is estimated more precisely for values of *x* in this area. As you move towards the extreme limits of the data, the width of the intervals increases, indicating that it would be unwise to extrapolate beyond the limits of the data used to create this model.

Prediction Intervals

What if you want to predict a *particular* value of *y* when $x = x_0$? Or, perhaps you want to predict the next measurement for a given value of *x*? This problem differs from constructing a confidence interval for μ_y. Instead of constructing a confidence interval to estimate a population parameter, we need to construct a prediction interval. Choosing to predict a particular value of *y* incurs some additional error in the prediction because of the deviation of *y* from the line of means. Examine the figure below. You can see that the error in prediction has two components:

1. The error in using the fitted line to estimate the line of means
2. The error caused by the deviation of y from the line of means, measured by σ^2

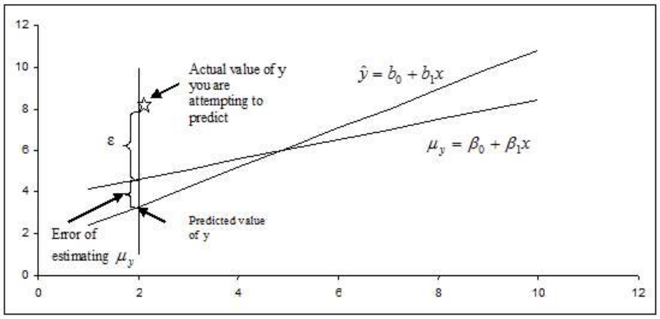

12
10
8
6
4
2
0
0
2
4
6
8
10
12
Actual value of y you are attempting to predict
$\hat{y} = b_0 + b_1 x$
ε
$\mu_y = \beta_0 + \beta_1 x$
Error of estimating μ_y
Predicted value of y

Figure $7.3.5$. Illustrating the two components in the error of prediction.

The variance of the difference between y and $\hat{y}$ is the sum of these two variances and forms the basis for the standard error of $(y - \hat{y})$ used for prediction. The resulting form of a prediction interval is as follows:

$$(7.3.9)$$

\hat y \pm t_{\alpha/2}s\sqrt {1+\frac {1}{n} + \frac {(x_0 - \bar x)^2}{\sum (x_i - \bar x)^2}} \]

where *x*0 is the given value for the predictor variable, *n* is the number of observations, and $t_{\alpha/2}$ is the critical value with (*n* – 2) degrees of freedom.

Software, such as Minitab, can compute the prediction intervals. Using the data from the previous example, we will use Minitab to compute the 95% prediction interval for the IBI of a specific forested area of 32 km.

Predicted Values for New Observations			
New Obs	Fit	SE Fit	95% PI
1	49.9496	2.38400	(20.1053, 79.7939)

You can repeat this process many times for several different values of *x* and plot the prediction intervals for the mean response.

x	95% PI
20	(13.01, 73.11)
40	(24.77, 84.31)
60	(36.21, 95.83)
80	(47.33, 107.67)
100	(58.15, 119.81)

Notice that the prediction interval bands are wider than the corresponding confidence interval bands, reflecting the fact that we are predicting the value of a random variable rather than estimating a population parameter. We would expect predictions for an individual value to be more variable than estimates of an average value.

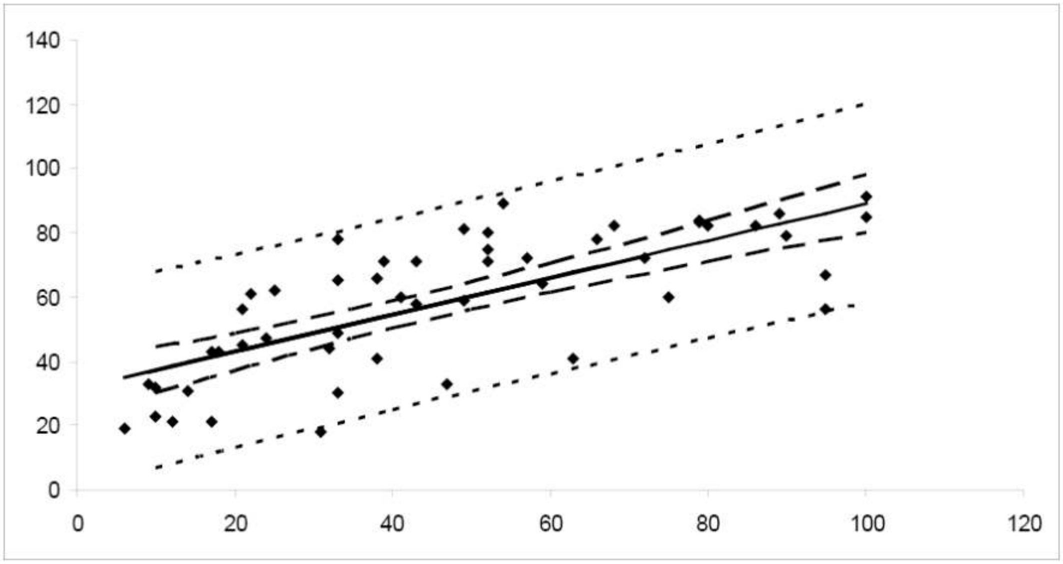


Figure $7.3.6$. A comparison of confidence and prediction intervals.

Transformations to Linearize Data Relationships

In many situations, the relationship between *x* and *y* is non-linear. In order to simplify the underlying model, we can transform or convert either *x* or *y* or both to result in a more linear relationship. There are many common transformations such as logarithmic and reciprocal. Including higher order terms on *x* may also help to linearize the relationship between *x* and *y*. Shown below are some common shapes of scatterplots and possible choices for transformations. However, the choice of transformation is frequently more a matter of trial and error than set rules.

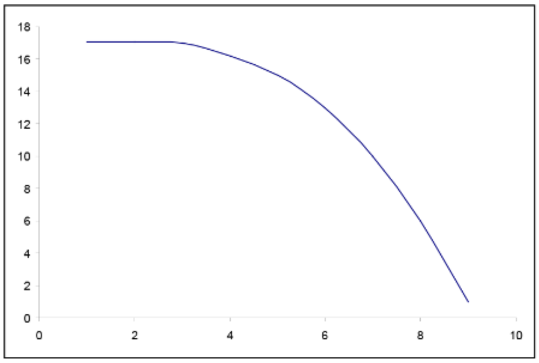
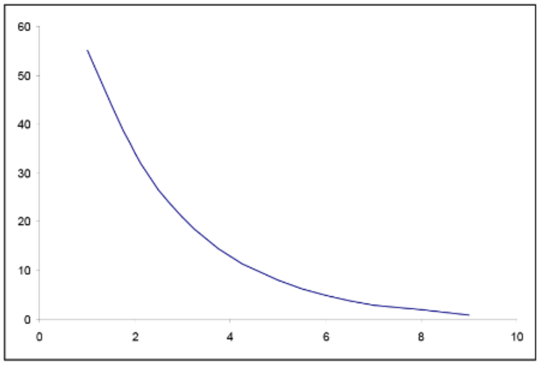
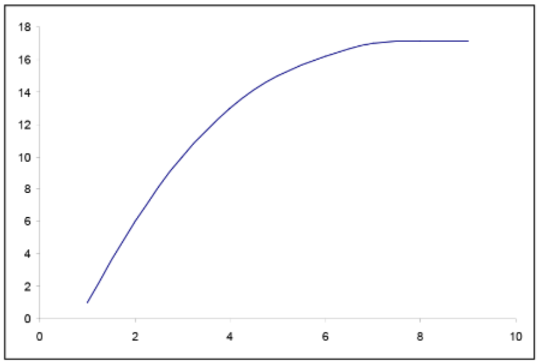
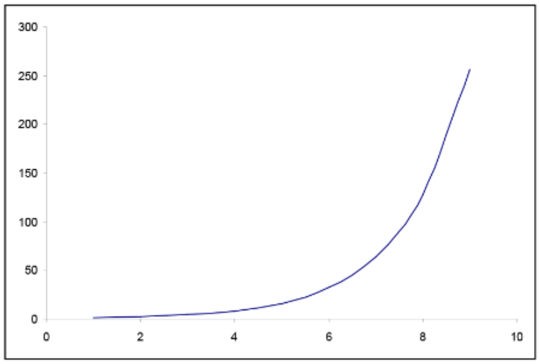
	**x** or **y** x^2 y^2 x^3 y^3
	**x** or **y** $\log x$ $\log y$ $-1/x$ $-1/y$
	**x** or **y** $\log x$ y^2 $-1/x$ y^3
Shape of scatterplot 	**Choice of transformation** **x** or **y** x^2 $\log y$ x^3 $-1/y$

Figure 7.3.7. Examples of possible transformations for x and y variables.

✓ Example 7.3.2:

A forester needs to create a simple linear regression model to predict tree volume using diameter-at-breast height (dbh) for sugar maple trees. He collects dbh and volume for 236 sugar maple trees and plots volume versus dbh. Given below is the scatterplot, correlation coefficient, and regression output from Minitab.

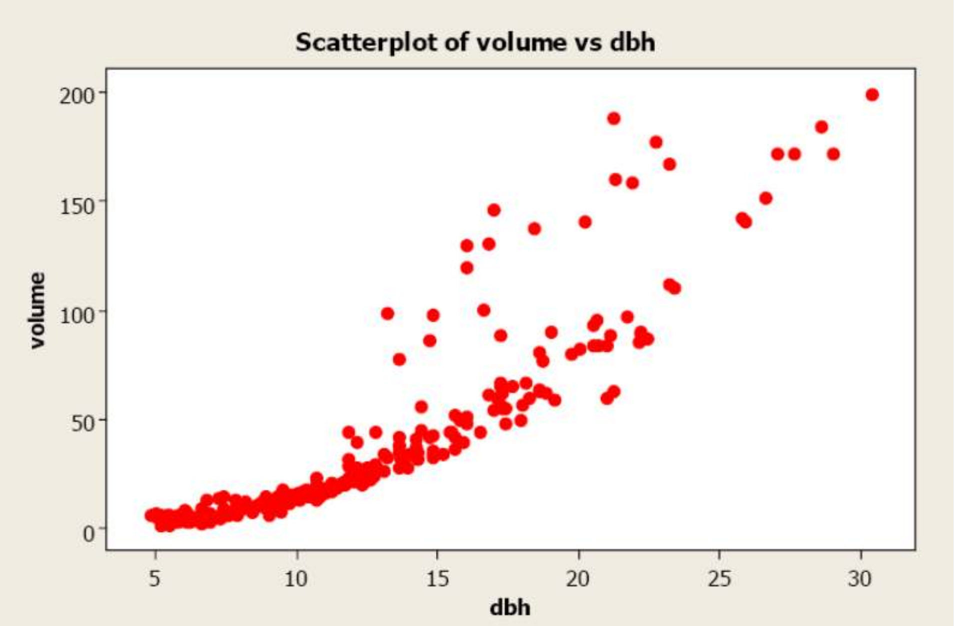


Figure 7.3.8 Scatterplot of volume versus dbh.

Pearson's linear correlation coefficient is 0.894, which indicates a strong, positive, linear relationship. However, the scatterplot shows a distinct nonlinear relationship.

Regression Analysis: volume versus dbh

The regression equation is volume = – 51.1 + 7.15 dbh				
Predictor	Coef	SE Coef	T	P
Constant	-51.097	3.271	-15.62	0.000
dbh	7.1500	0.2342	30.53	0.000
S = 19.5820	R-Sq = 79.9%		R-Sq(adj) = 79.8%	

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	357397	357397	932.04	0.000
Residual Error	234	89728	383		
Total	235	447125			

The R2 is 79.9% indicating a fairly strong model and the slope is significantly different from zero. However, both the residual plot and the residual normal probability plot indicate serious problems with this model. A transformation may help to create a more linear relationship between volume and dbh.

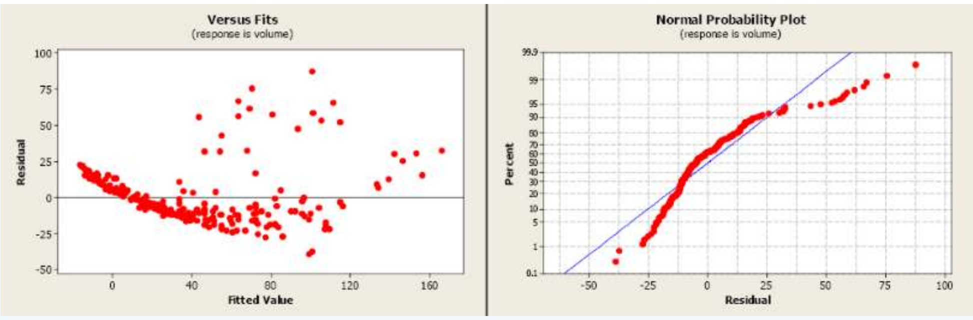


Figure 7.3.9. Residual and normal probability plots.

Volume was transformed to the natural log of volume and plotted against dbh (see scatterplot below). Unfortunately, this did little to improve the linearity of this relationship. The forester then took the natural log transformation of dbh. The scatterplot of the natural log of volume versus the natural log of dbh indicated a more linear relationship between these two variables. The linear correlation coefficient is 0.954.

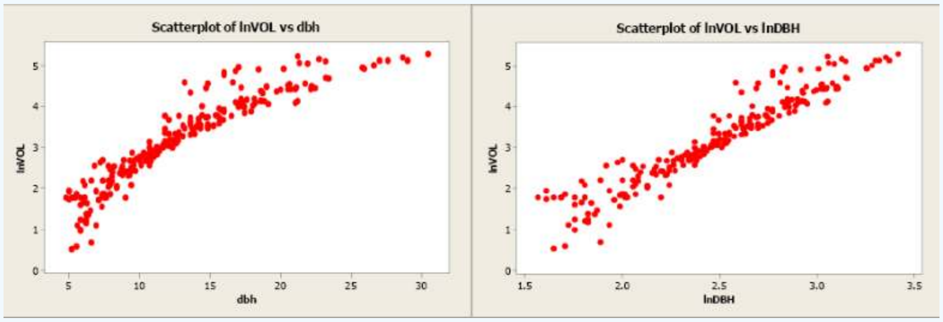


Figure 7.3.10. Scatterplots of natural log of volume versus dbh and natural log of volume versus natural log of dbh.

The regression analysis output from Minitab is given below.

Regression Analysis: lnVOL vs. lnDBH

The regression equation is lnVOL = – 2.86 + 2.44 lnDBH				
Predictor	Coef	SE Coef	T	P
Constant	-2.8571	0.1253	-22.80	0.000
lnDBH	2.44383	0.05007	48.80	0.000
S = 0.327327	R-Sq = 91.1%	R-Sq(adj) = 91.0%		

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	255.19	255.19	2381.78	0.000
Residual Error	234	25.07	0.11		
Total	235	280.26			

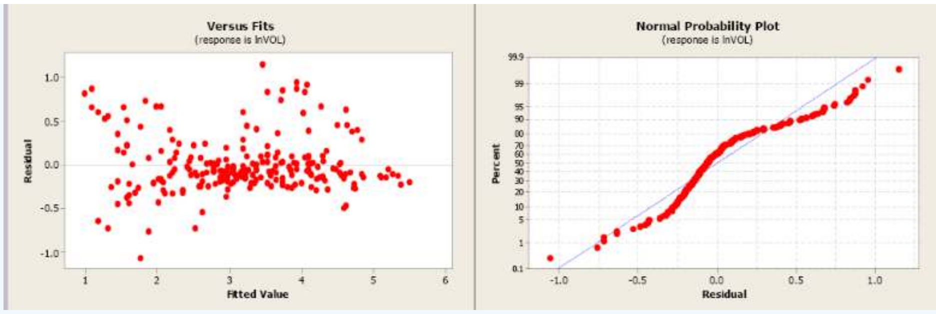


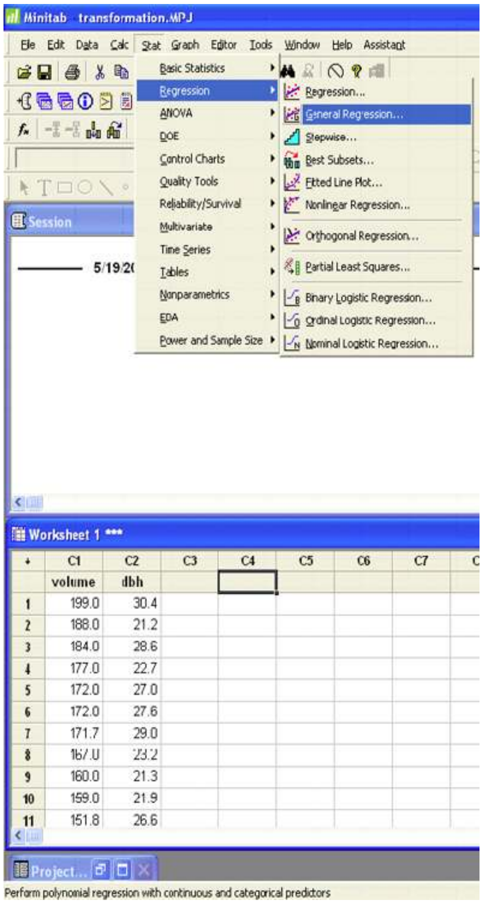
Figure 7.3.11 Residual and normal probability plots.

The model using the transformed values of volume and dbh has a more linear relationship and a more positive correlation coefficient. The slope is significantly different from zero and the R2 has increased from 79.9% to 91.1%. The residual plot shows a more random pattern and the normal probability plot shows some improvement.

There are many possible transformation combinations possible to linearize data. Each situation is unique and the user may need to try several alternatives before selecting the best transformation for x or y or both.

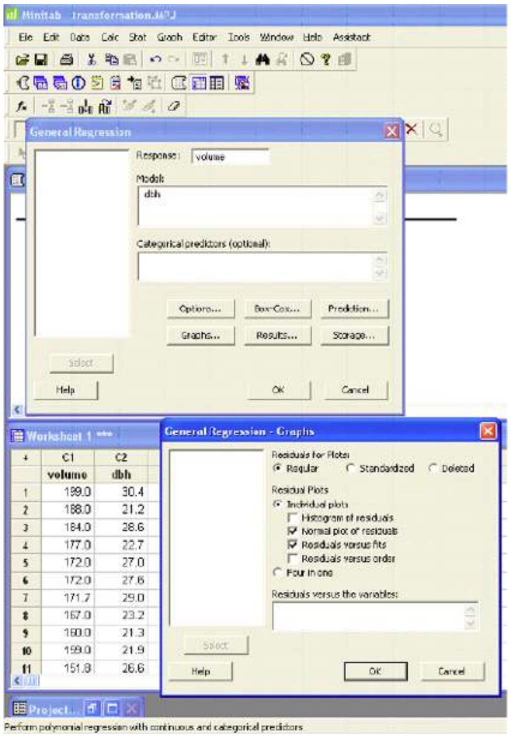
7.4: Software Solution

Minitab



+	C1	C2	C3	C4	C5	C6	C7
	volume	dbh					
1	199.0	30.4					
2	188.0	21.2					
3	184.0	28.6					
4	177.0	22.7					
5	172.0	27.0					
6	172.0	27.6					
7	171.7	29.0					
8	167.0	23.2					
9	160.0	21.3					
10	159.0	21.9					
11	151.8	26.6					

Perform polynomial regression with continuous and categorical predictors



Response: volume

Model: dbh

Categorical predictors (optional):

Residuals for Plots
● Regular ○ Standardized ○ Deleted

Residual Plots
● Individual plots
☐ Histogram of residuals
☑ Normal plot of residuals
☑ Residuals versus fits
☐ Residuals versus order
○ Four in one

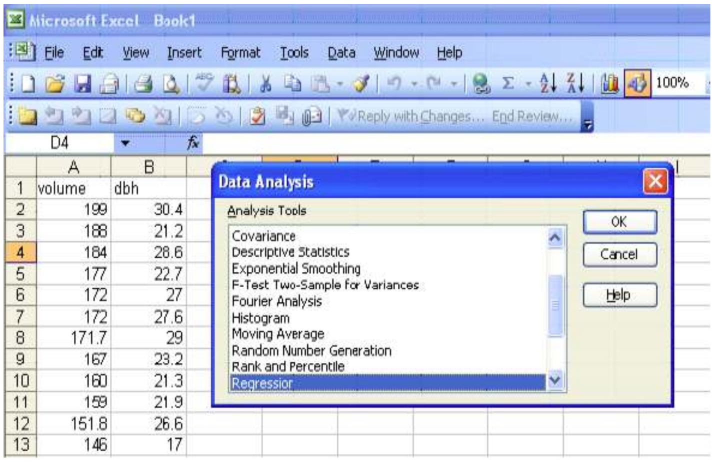
Residuals versus the variables:

+	C1	C2
	volume	dbh
1	199.0	30.4
2	188.0	21.2
3	184.0	28.6
4	177.0	22.7
5	172.0	27.0
6	172.0	27.6
7	171.7	29.0
8	167.0	23.2
9	160.0	21.3
10	159.0	21.9
11	151.8	26.6

Perform polynomial regression with continuous and categorical predictors

The Minitab output is shown above in Ex. 4.

Excel

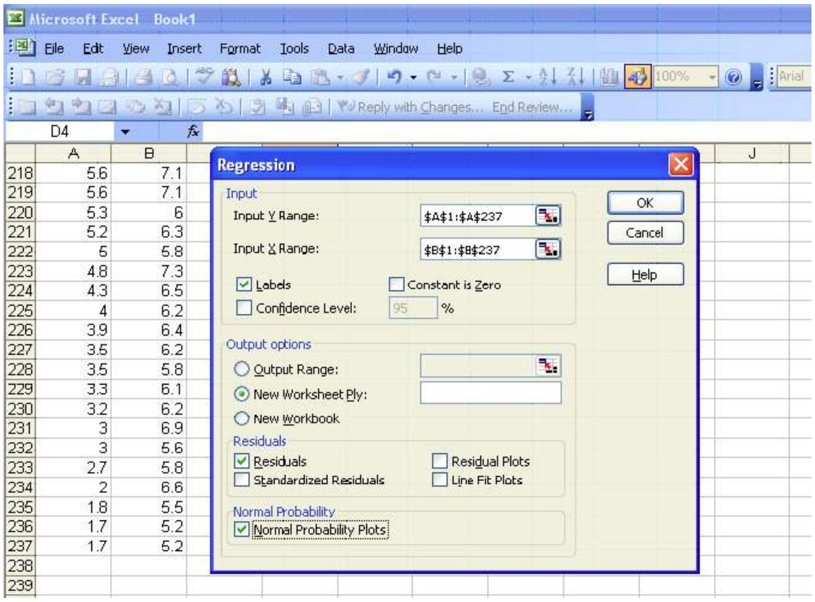

Microsoft Excel Book1
File Edit View Insert Format Tools Data Window Help
100%
Reply with Changes... End Review...
D4
OK
Cancel
Help

	A	B
1	volume	dbh
2	199	30.4
3	188	21.2
4	184	28.6
5	177	22.7
6	172	27
7	172	27.6
8	171.7	29
9	167	23.2
10	160	21.3
11	159	21.9
12	151.8	26.6
13	146	17

Data Analysis

Analysis Tools

- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression


Microsoft Excel Book1
File Edit View Insert Format Tools Data Window Help
100%
Arial
Reply with Changes... End Review...
D4
J
OK
Cancel
Help

	A	B
218	5.6	7.1
219	5.6	7.1
220	5.3	6
221	5.2	6.3
222	5	5.8
223	4.8	7.3
224	4.3	6.5
225	4	6.2
226	3.9	6.4
227	3.5	6.2
228	3.5	5.8
229	3.3	6.1
230	3.2	6.2
231	3	6.9
232	3	5.6
233	2.7	5.8
234	2	6.6
235	1.8	5.5
236	1.7	5.2
237	1.7	5.2
238		
239		

Regression

Input

Input Y Range: $A$1:$A$237

Input X Range: $B$1:$B$237

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☑ Normal Probability Plots

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.89404782
R Square	0.7993215
Adjusted R Square	0.7984639
Standard Error	19.5819962
Observations	236

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	357396.6011	357396.6	932.0442	1.44E-83
Residual	234	89728.37054	383.4546		
Total	235	447124.9717			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-51.096811	3.270935681	-15.6215	3.66E-38	-57.5411	-44.6526	-57.5411	-44.6526
dbh	7.14997446	0.234199651	30.5294	1.44E-83	6.688565	7.611384	6.688565	7.611384

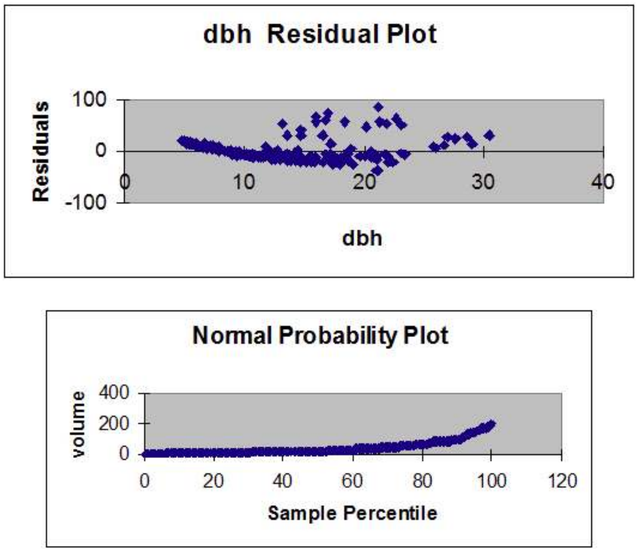


Figure 7.4.1. Residual and normal probability plots.

CHAPTER OVERVIEW

8: Multiple Linear Regression

8.1: Multiple Regressions

It frequently happens that a dependent variable (*y*) in which we are interested is related to more than one independent variable. If this relationship can be estimated, it may enable us to make more precise predictions of the dependent variable than would be possible by a simple linear regression. Regressions based on more than one independent variable are called **multiple regressions**.

Multiple linear regression is an extension of simple linear regression and many of the ideas we examined in simple linear regression carry over to the multiple regression setting. For example, scatterplots, correlation, and least squares method are still essential components for a multiple regression.

For example, a habitat suitability index (used to evaluate the impact on wildlife habitat from land use changes) for ruffed grouse might be related to three factors:

$\mathbf{x}_1$ = stem density
$\mathbf{x}_2$ = percent of conifers
$\mathbf{x}_3$ = amount of understory herbaceous matter

A researcher would collect data on these variables and use the sample data to construct a regression equation relating these three variables to the response. The researcher will have questions about his model similar to a simple linear regression model.

- How strong is the relationship between y and the three predictor variables?
- How well does the model fit?
- Have any important assumptions been violated?
- How good are the estimates and predictions?

The general linear regression model takes the form of

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{8.1.1}$$

with the mean value of y given as

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{8.1.2}$$

where:

- y is the random response variable and μy is the mean value of *y*,
- β0, β1, β2, and βk are the parameters to be estimated based on the sample data,
- *x*1, *x*2,…, *x*k are the predictor variables that are assumed to be non-random or fixed and measured without error, and k is the number of predictor variable,
- and ε is the random error, which allows each response to deviate from the average value of *y*. The errors are assumed to be independent, have a mean of zero and a common variance (σ2), and are normally distributed.

As you can see, the multiple regression model and assumptions are very similar to those for a simple linear regression model with one predictor variable. Examining residual plots and normal probability plots for the residuals is key to verifying the assumptions.

Correlation

As with simple linear regression, we should always begin with a scatterplot of the response variable versus each predictor variable. Linear correlation coefficients for each pair should also be computed. Instead of computing the correlation of each pair individually, we can create a correlation matrix, which shows the linear correlation between each pair of variables under consideration in a multiple linear regression model.

	y	**x1**	**x2**
x1	0.816		
	0.000		
x2	0.413	−0.144	
	0.029	0.466	
x3	0.768	0.588	0.406
	0.000	0.001	0.032

Table 8.1.1. *A correlation matrix.*

In this matrix, the upper value is the linear correlation coefficient and the lower value is the p-value for testing the null hypothesis that a correlation coefficient is equal to zero. This matrix allows us to see the strength and direction of the linear relationship between each predictor variable and the response variable, but also the relationship between the predictor variables. For example, *y* and *x1*have a strong, positive linear relationship with r = 0.816, which is statistically significant because p = 0.000. We can also see that predictor variables *x1* and *x3* have a moderately strong positive linear relationship (r = 0.588) that is significant (p = 0.001).

There are many different reasons for selecting which explanatory variables to include in our model (see Model Development and Selection), however, we frequently choose the ones that have a high linear correlation with the response variable, but we must be careful. We do not want to include explanatory variables that are highly correlated among themselves. We need to be aware of any multicollinearity between predictor variables.

Multicollinearity exists between two explanatory variables if they have a strong linear relationship.

For example, if we are trying to predict a person's blood pressure, one predictor variable would be weight and another predictor variable would be diet. Both predictor variables are highly correlated with blood pressure (as weight increases blood pressure typically increases, and as diet increases blood pressure also increases). But, both predictor variables are also highly correlated with each other. Both of these predictor variables are conveying essentially the same information when it comes to explaining blood pressure. Including both in the model may lead to problems when estimating the coefficients, as multicollinearity increases the standard errors of the coefficients. This means that coefficients for some variables may be found **not** to be significantly different from zero, whereas without multicollinearity and with lower standard errors, the same coefficients might have been found significant. Ways to test for multicollinearity are not covered in this text, however a general rule of thumb is to be wary of a linear correlation of less than -0.7 and greater than 0.7 between two predictor variables. Always examine the correlation matrix for relationships between predictor variables to avoid multicollinearity issues.

Estimation

Estimation and inference procedures are also very similar to simple linear regression. Just as we used our sample data to estimate β0 and β1 for our simple linear regression model, we are going to extend this process to estimate all the coefficients for our multiple regression models.

With the simpler population model

$$\mu_y = \beta_0 + \beta_1 x \tag{8.1.3}$$

β1 is the slope and tells the user what the change in the response would be as the predictor variable changes. With multiple predictor variables, and therefore multiple parameters to estimate, the coefficients β1, β2, β3 and so on are called partial slopes or partial regression coefficients. The partial slope βi measures the change in *y* for a one-unit change in *x*i when **all other independent variables are held constant.** These regression coefficients must be estimated from the sample data in order to obtain the general form of the estimated multiple regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k \tag{8.1.4}$$

and the population model

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \tag{8.1.5}$$

where *k* = the number of independent variables (also called predictor variables)

ŷ = the predicted value of the dependent variable (computed by using the multiple regression equation)

*x*1, *x*2, …, *x*k = the independent variables

β0 is the y-intercept (the value of y when all the predictor variables equal 0)

*b*0 is the estimate of β0 based on that sample data

β1, β2, β3,…βk are the coefficients of the independent variables *x*1, *x*2, …, *x*k

*b*1, *b*2, *b*3, …, *b*k are the sample estimates of the coefficients β1, β2, β3,…βk

The method of least-squares is still used to fit the model to the data. Remember that this method minimizes the sum of the squared deviations of the observed and predicted values (SSE).

The analysis of variance table for multiple regression has a similar appearance to that of a simple linear regression.

Source of variation	df	Seq sums of squares	Sums of squares	Mean sums of squares	F
Regression	k		SSR	SSR/k = MSR	MSR/MSE = F
Error	n - k - 1		SSE	SSE/(n - k - 1) = MSE	
Total	n -1		SST		

Table 8.1.2. *ANOVA table.*

Where k is the number of predictor variables and n is the number of observations.

The best estimate of the random variation σ^2—the variation that is unexplained by the predictor variables—is still s2, the MSE. The regression standard error, s, is the square root of the MSE.

A new column in the ANOVA table for multiple linear regression shows a decomposition of SSR, in which the conditional contribution of each predictor variable *given the variables already entered into the model* is shown for the order of entry that you specify in your regression. These conditional or **sequential sums of squares** each account for 1 regression degree of freedom, and allow the user to see the contribution of each predictor variable to the total variation explained by the regression model by using the ratio:

$$\frac{SeqSS}{SSR} \tag{8.1.6}$$

Adjusted R^2

In simple linear regression, we used the relationship between the explained and total variation as a measure of model fit:

$$R^2 = \frac{Explained\,Variation}{Total\,Variation} = \frac{SSR}{SSTo} = 1 - \frac{SSE}{SSTo} \tag{8.1.7}$$

Notice from this definition that the value of the coefficient of determination can never decrease with the addition of more variables into the regression model. Hence, R^2 can be artificially inflated as more variables (significant or not) are included in the model. An alternative measure of strength of the regression model is adjusted for degrees of freedom by using mean squares rather than sums of squares:

$$R^2(adj) = 1 - \frac{(n-1)(1-R^2)}{(n-p)} = (1 - \frac{MSE}{SSTo/(n-1)}) \tag{8.1.8}$$

The adjusted R^2 value represents the percentage of variation in the response variable explained by the independent variables, corrected for degrees of freedom. Unlike R^2, the adjusted R^2 will not tend to increase as variables are added and it will tend to stabilize around some upper limit as variables are added.

Tests of Significance

Recall in the previous chapter we tested to see if *y* and *x* were linearly related by testing

$H_0 : \beta_1 = 0$	$H_1 : \beta_1 \neq 0$

with the t-test (or the equivalent F-test). In multiple linear regression, there are several partial slopes and the t-test and F-test are no longer equivalent. Our question changes: Is the regression equation that uses information provided by the predictor variables x1, x2, x3, …, xk, better than the simple predictor 13615.png(the mean response value), which does not rely on any of these independent variables?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

$H1 : At\ least\ one\ of \beta_1, \beta_2, \beta_3, \ldots \beta_k \neq 0$

The F-test statistic is used to answer this question and is found in the ANOVA table.

$$F = \frac{MSR}{MSE} \tag{8.1.9}$$

This test statistic follows the F-distribution with $df_1 = k$ and $df_2 = (n-k-1)$. Since the exact p-value is given in the output, you can use the Decision Rule to answer the question.

If the p-value is less than the level of significance, reject the null hypothesis.

Rejecting the null hypothesis supports the claim that at least one of the predictor variables has a significant linear relationship with the response variable. The next step is to determine which predictor variables add important information for prediction in the presence of other predictors already in the model. To test the significance of the partial regression coefficients, you need to examine each relationship separately using individual t-tests.

$H_0 : \beta_i = 0$	$H_1 : \beta_i \neq 0$

$$t = \frac{b_i - \beta_o}{SE(b_i)} \; with \; df = (n-k-1) \tag{8.1.10}$$

where $SE(b_i)$ is the standard error of b_i. Exact p-values are also given for these tests. Examining specific p-values for each predictor variable will allow you to decide which variables are significantly related to the response variable. Typically, any insignificant variables are removed from the model, but remember these tests are done with other variables in the model. A good procedure is to remove the least significant variable and then refit the model with the reduced data set. With each new model, always check the regression standard error (lower is better), the adjusted R^2 (higher is better), the p-values for all predictor variables, and the residual and normal probability plots.

Because of the complexity of the calculations, we will rely on software to fit the model and give us the regression coefficients. Don’t forget… you always begin with scatterplots. Strong relationships between predictor and response variables make for a good model.

✓ Example 8.1.1:

A researcher collected data in a project to predict the annual growth per acre of upland boreal forests in southern Canada. They hypothesized that cubic foot volume growth (*y*) is a function of stand basal area per acre (*x*1), the percentage of that basal area in black spruce (*x*2), and the stand’s site index for black spruce (*x*3). α = 0.05.

CuFt	BA/ac	%BA Bspruce	SI		CuFt	BA/ac	%BA Bspruce	SI
55	51	79	45		71	65	93	35
68	100	48	53		67	87	68	41
60	63	67	44		73	108	51	54
40	52	52	31		87	105	82	51
45	67	52	29		80	100	70	45
49	42	82	43		77	103	61	43
62	81	80	42		64	55	96	51
56	70	65	36		60	60	80	47
93	108	96	63		65	70	76	40
76	90	81	60		65	78	74	46
94	110	78	56		83	85	96	55
82	111	59	48		67	92	58	50

CuFt	BA/ac	%BA Bspruce	SI		CuFt	BA/ac	%BA Bspruce	SI
86	94	84	53		61	82	58	38
55	82	48	40		51	56	69	35

Table 8.1.3. Observed data for cubic feet, stand basal area, percent basal area in black spruce, and site index.

Scatterplots of the response variable versus each predictor variable were created along with a correlation matrix.

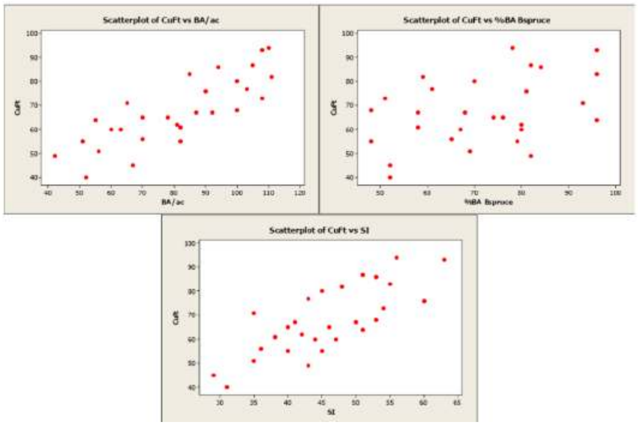


Figure 8.1.1. Scatterplots of cubic feet versus basal area, percent basal area in black spruce, and site index.

Correlations: CuFt, BA/ac, %BA Bspruce, SI

BA/ac	CuFt 0.816 0.000	BA/aC	$\frac{8}{8}$BA	Bspruce
○A Bspruce	0.413 0.029	−0.144 0.466		
SI	0.768 0.000	0.588 0.001		0.406 0.032

(8.1.11)

Table 8.1.4. Correlation matrix.

As you can see from the scatterplots and the correlation matrix, BA/ac has the strongest linear relationship with CuFt volume ($r = 0.816$) and %BA in black spruce has the weakest linear relationship ($r = 0.413$). Also of note is the moderately strong correlation between the two predictor variables, BA/ac and SI ($r = 0.588$). All three predictor variables have significant linear relationships with the response variable (volume) so we will begin by using all variables in our multiple linear regression model. The Minitab output is given below.

We begin by testing the following null and alternative hypotheses:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_1: At least one of β_1, β_2 , $\beta_3 \neq 0$

General Regression Analysis: CuFt versus BA/ac, SI, %BA Bspruce

Regression Equation: CuFt = -19.3858 + 0.591004 BA/ac + 0.0899883 SI + 0.489441 %BA Bspruce

Coefficients					
Term	Coef	SE Coef	T	P	95% CI
Constant	-19.3858	4.15332	-4.6675	0.000	(-27.9578, -10.8137)

BA/ac	0.5910	0.04294	13.7647	0.000	(0.5024, 0.6796)	
SI	0.0900	0.11262	0.7991	0.432	(-0.1424, 0.3224)	
%BA Bspruce	0.4894	0.05245	9.3311	0.000	(0.3812, 0.5977)	
Summary of Model						
S = 3.17736		R-Sq = 95.53%		R-Sq(adj) = 94.97%		
PRESS = 322.279		R-Sq(pred) = 94.05%				
Analysis of Variance						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	3	5176.56	5176.56	1725.52	**170.918**	**0.000000**
BA/ac	1	3611.17	1912.79	1912.79	189.467	0.000000
SI	1	686.37	6.45	6.45	0.638	0.432094
%BA Bspruce	1	879.02	879.02	879.02	87.069	0.000000
Error	24	242.30	242.30	10.10		
Total	27	5418.86				

The F-test statistic (and associated p-value) is used to answer this question and is found in the ANOVA table. For this example, F = 170.918 with a p-value of 0.00000. The p-value is smaller than our level of significance (0.0000<0.05) so we will reject the null hypothesis. At least one of the predictor variables significantly contributes to the prediction of volume.

The coefficients for the three predictor variables are all positive indicating that as they increase cubic foot volume will also increase. For example, if we hold values of SI and %BA Bspruce constant, this equation tells us that as basal area increases by 1 sq. ft., volume will increase an additional 0.591004 cu. ft. The signs of these coefficients are logical, and what we would expect. The adjusted R^2 is also very high at 94.97%.

The next step is to examine the individual t-tests for each predictor variable. The test statistics and associated p-values are found in the Minitab output and repeated below:

Coefficients					
Term	Coef	SE Coef	T	P	95% CI
Constant	-19.3858	4.15332	-4.6675	0.000	(-27.9578, -10.8137)
BA/ac	0.5910	0.04294	**13.7647**	**0.000**	(0.5024, 0.6796)
SI	0.0900	0.11262	**0.7991**	**0.432**	(-0.1424, 0.3224)
%BA Bspruce	0.4894	0.05245	**9.3311**	**0.000**	(0.3812, 0.5977)

The predictor variables BA/ac and %BA Bspruce have t-statistics of 13.7647 and 9.3311 and p-values of 0.0000, indicating that both are significantly contributing to the prediction of volume. However, SI has a t-statistic of 0.7991 with a p-value of 0.432. This variable does not significantly contribute to the prediction of cubic foot volume.

This result may surprise you as SI had the second strongest relationship with volume, but don’t forget about the correlation between SI and BA/ac (r = 0.588). The predictor variable BA/ac had the strongest linear relationship with volume, and using the sequential sums of squares, we can see that BA/ac is already accounting for 70% of the variation in cubic foot volume (3611.17/5176.56 = 0.6976). The information from SI may be too similar to the information in BA/ac, and SI only explains about 13% of the variation on volume (686.37/5176.56 = 0.1326) given that BA/ac is already in the model.

The next step is to examine the residual and normal probability plots. A single outlier is evident in the otherwise acceptable plots.

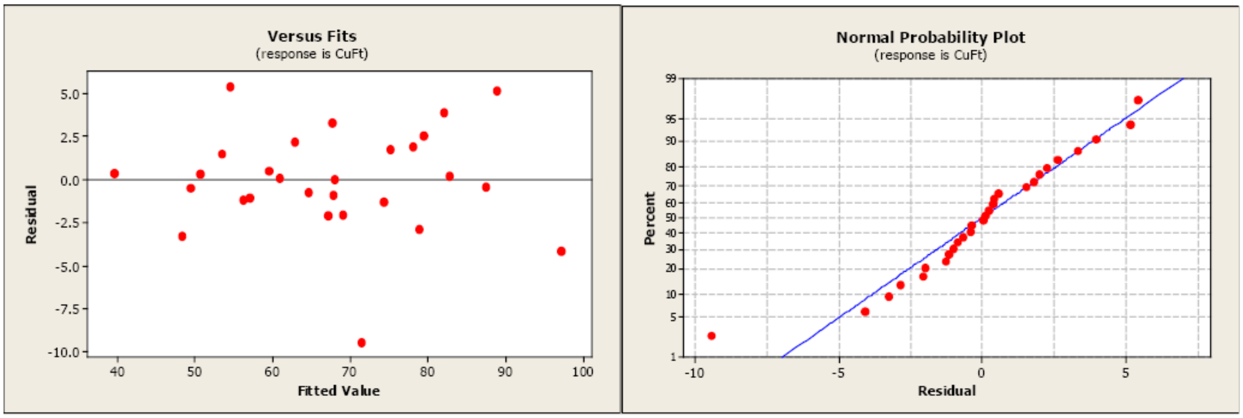


Figure 8.1.2. Residual and normal probability plots.

So where do we go from here?

We will remove the non-significant variable and re-fit the model excluding the data for SI in our model. The Minitab output is given below.

General Regression Analysis: CuFt versus BA/ac, %BA Bspruce

Regression Equation						
CuFt = -19.1142 + 0.615531 BA/ac + 0.515122 %BA Bspruce						
Coefficients						
Term	Coef	SE Coef	T	P	95% CI	
Constant	-19.1142	4.10936	-4.6514	0.000	(-27.5776, -10.6508)	
BA/ac	0.6155	0.02980	20.6523	0.000	(0.5541, 0.6769)	
%BA Bspruce	0.5151	0.04115	12.5173	0.000	(0.4304, 0.5999)	
Summary of Model						
S = 3.15431		R-Sq = 95.41%		R-Sq(adj) = 95.04%		
PRESS = 298.712		R-Sq(pred) = 94.49%				
Analysis of Variance						
Source	DF	SeqSS	AdjSS	AdjMS	F	P
Regression	2	5170.12	5170.12	2585.06	259.814	0.0000000
BA/ac	1	3611.17	4243.71	4243.71	426.519	0.0000000
%BA Bspruce	1	1558.95	1558.95	1558.95	156.684	0.0000000
Error	25	248.74	248.74	9.95		
Total	27	5418.86				

We will repeat the steps followed with our first model. We begin by again testing the following hypotheses:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_1 : At\ least\ one\ of\ \beta_1, \beta_2, \beta_3 \neq 0$

This reduced model has an F-statistic equal to 259.814 and a p-value of 0.0000. We will reject the null hypothesis. At least one of the predictor variables significantly contributes to the prediction of volume. The coefficients are still positive (as we expected) but the values have changed to account for the different model.

The individual t-tests for each coefficient (repeated below) show that both predictor variables are significantly different from zero and contribute to the prediction of volume.

Coefficients

Term	Coef	SE Coef	T	P	95% CI
Constant	-19.1142	4.10936	-4.6514	0.000	(-27.5776, -10.6508)
BA/ac	0.6155	0.02980	20.6523	0.000	(0.5541, 0.6769)
%BA Bspruce	0.5151	0.04115	12.5173	0.000	(0.4304, 0.5999)

Notice that the adjusted R2 has increased from 94.97% to 95.04% indicating a slightly better fit to the data. The regression standard error has also changed for the better, decreasing from 3.17736 to 3.15431 indicating slightly less variation of the observed data to the model.

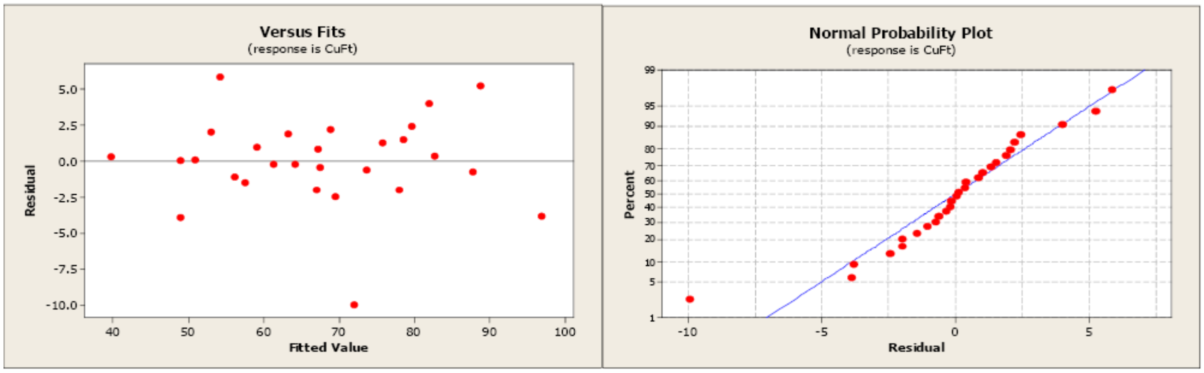


Figure 8.1.3. *Residual and normal probability plots.*

The residual and normal probability plots have changed little, still not indicating any issues with the regression assumption. By removing the non-significant variable, the model has improved.

8.2: Software Solution

Model Development and Selection

There are many different reasons for creating a multiple linear regression model and its purpose directly influences how the model is created. Listed below are several of the more commons uses for a regression model:

1. Describing the behavior of your response variable
2. Predicting a response or estimating the average response
3. Estimating the parameters (β0, β1, β2, …)
4. Developing an accurate model of the process

Depending on your objective for creating a regression model, your methodology may vary when it comes to variable selection, retention, and elimination.

When the object is simple description of your response variable, you are typically less concerned about eliminating non-significant variables. The best representation of the response variable, in terms of minimal residual sums of squares, is the full model, which includes all predictor variables available from the data set. It is less important that the variables are causally related or that the model is realistic.

A common reason for creating a regression model is for prediction and estimating. A researcher wants to be able to define events within the x-space of data that were collected for this model, and it is assumed that the system will continue to function as it did when the data were collected. Any measurable predictor variables that contain information on the response variable should be included. For this reason, non-significant variables may be retained in the model. However, regression equations with fewer variables are easier to use and have an economic advantage in terms of data collection. Additionally, there is a greater confidence attached to models that contain only significant variables.

If the objective is to estimate the model parameters, you will be more cautious when considering variable elimination. You want to avoid introducing a bias by removing a variable that has predictive information about the response. However, there is a statistical advantage in terms of reduced variance of the parameter estimates if variables truly unrelated to the response variable are removed.

Building a realistic model of the process you are studying is often a primary goal of much research. It is important to identify the variables that are linked to the response through some causal relationship. While you can identify which variables have a strong correlation with the response, this only serves as an indicator of which variables require further study. The principal objective is to develop a model whose functional form realistically reflects the behavior of a system.

The following figure is a strategy for building a regression model.

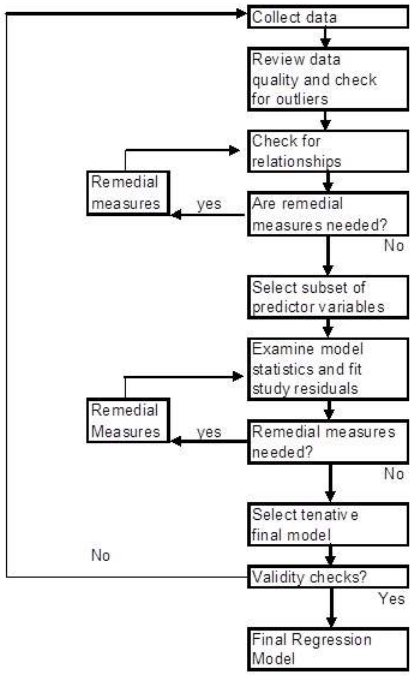
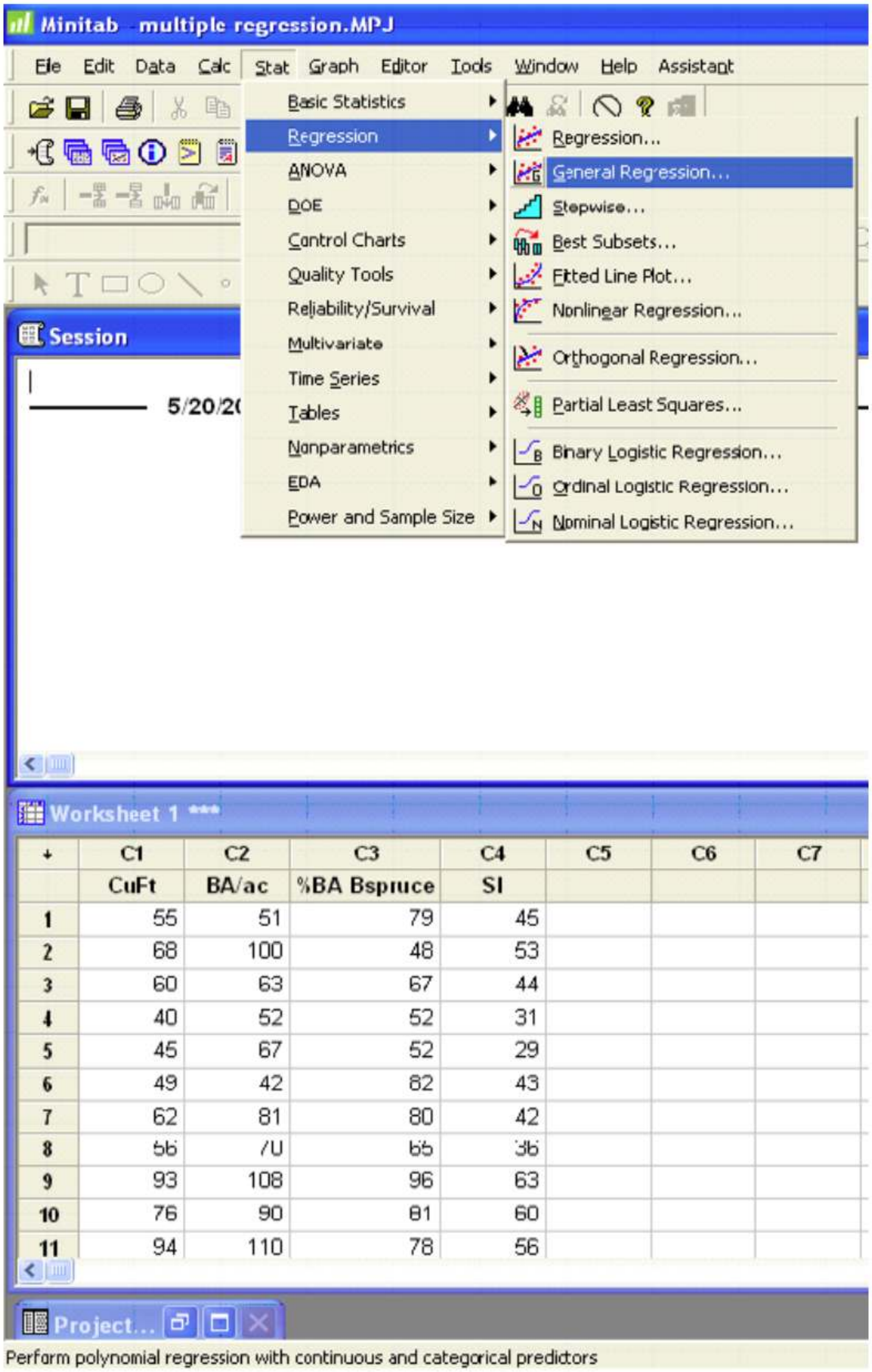

Collect data
Review data quality and check for outliers
Check for relationships
Remedial measures
yes
Are remedial measures needed?
No
Select subset of predictor variables
Examine model statistics and fit study residuals
Remedial Measures
yes
Remedial measures needed?
No
Select tentative final model
No
Validity checks?
Yes
Final Regression Model

Figure $8.2.1$. *Strategy for building a regression model.*

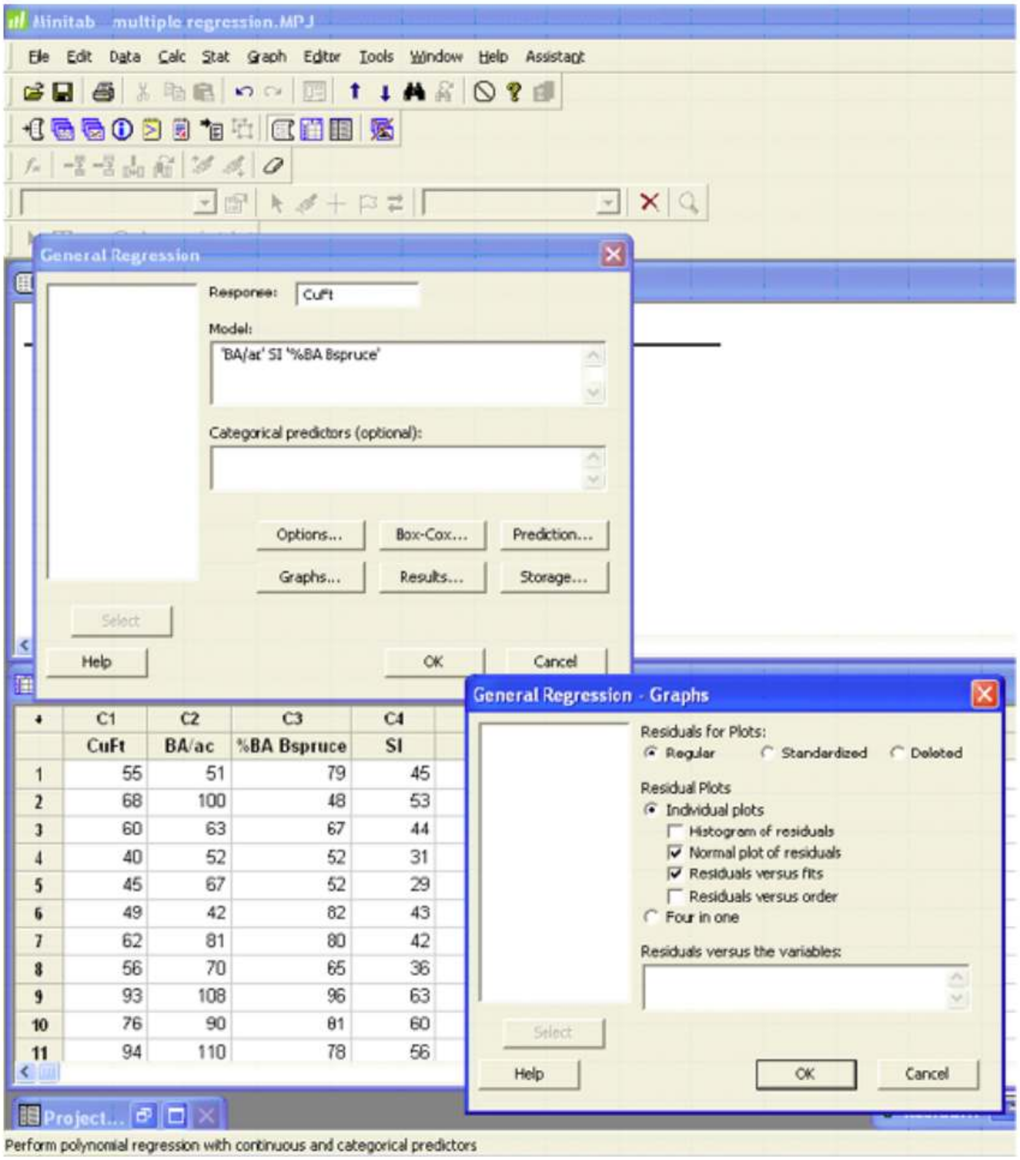
Software Solutions

Minitab



↓	C1	C2	C3	C4	C5	C6	C7
	CuFt	BA/ac	%BA Bspruce	SI			
1	55	51	79	45			
2	68	100	48	53			
3	60	63	67	44			
4	40	52	52	31			
5	45	67	52	29			
6	49	42	82	43			
7	62	81	80	42			
8	56	70	65	36			
9	93	108	96	63			
10	76	90	81	60			
11	94	110	78	56			

Perform polynomial regression with continuous and categorical predictors


Minitab - multiple regression.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
General Regression
Options... Box-Cox... Prediction...
Graphs... Results... Storage...
Select
Help OK Cancel
General Regression - Graphs
Select
Help OK Cancel
Project...

General Regression

Response: CuFt

Model: 'BA/ac' SI '%BA Bspruce'

Categorical predictors (optional):

	C1	C2	C3	C4
	CuFt	BA/ac	%BA Bspruce	SI
1	55	51	79	45
2	68	100	48	53
3	60	63	67	44
4	40	52	52	31
5	45	67	52	29
6	49	42	82	43
7	62	81	80	42
8	56	70	65	36
9	93	108	96	63
10	76	90	81	60
11	94	110	78	56

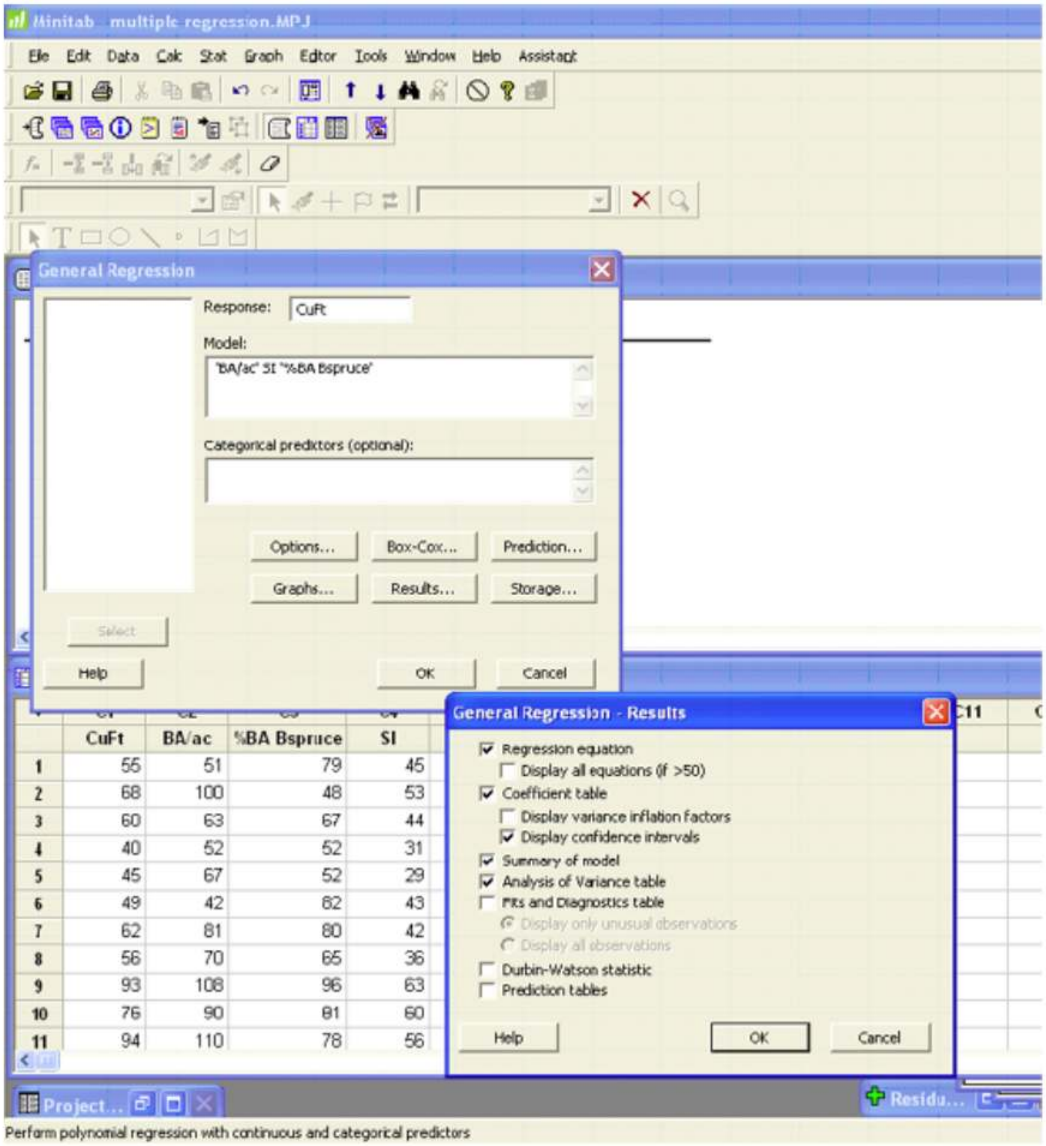
General Regression - Graphs

Residuals for Plots:
● Regular ○ Standardized ○ Deleted

Residual Plots
● Individual plots
☐ Histogram of residuals
☑ Normal plot of residuals
☑ Residuals versus fits
☐ Residuals versus order
○ Four in one

Residuals versus the variables:

Perform polynomial regression with continuous and categorical predictors


Minitab - multiple regression.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
General Regression
Select
Options... Box-Cox... Prediction...
Graphs... Results... Storage...
Help OK Cancel
General Regression - Results
Help OK Cancel
Project...
Residu...
C11

Response: CuFt

Model: 'BA/ac' SI '%BA Bspruce'

Categorical predictors (optional):

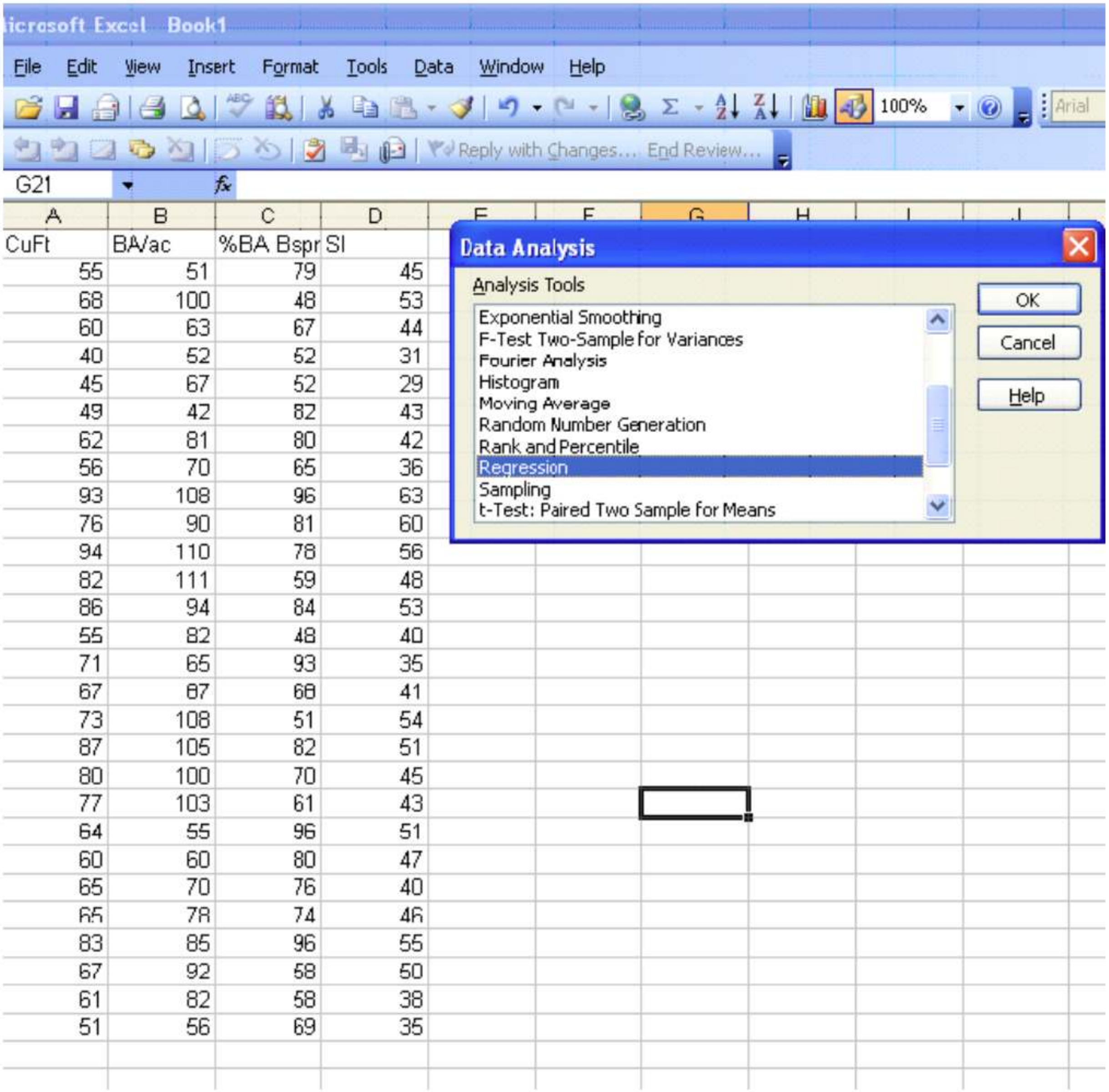
☑ Regression equation
  ☐ Display all equations (if >50)
☑ Coefficient table
  ☐ Display variance inflation factors
  ☑ Display confidence intervals
☑ Summary of model
☑ Analysis of Variance table
☐ Fits and Diagnostics table
  ● Display only unusual observations
  ○ Display all observations
☐ Durbin-Watson statistic
☐ Prediction tables

	CuFt	BA/ac	%BA Bspruce	SI
1	55	51	79	45
2	68	100	48	53
3	60	63	67	44
4	40	52	52	31
5	45	67	52	29
6	49	42	82	43
7	62	81	80	42
8	56	70	65	36
9	93	108	96	63
10	76	90	81	60
11	94	110	78	56

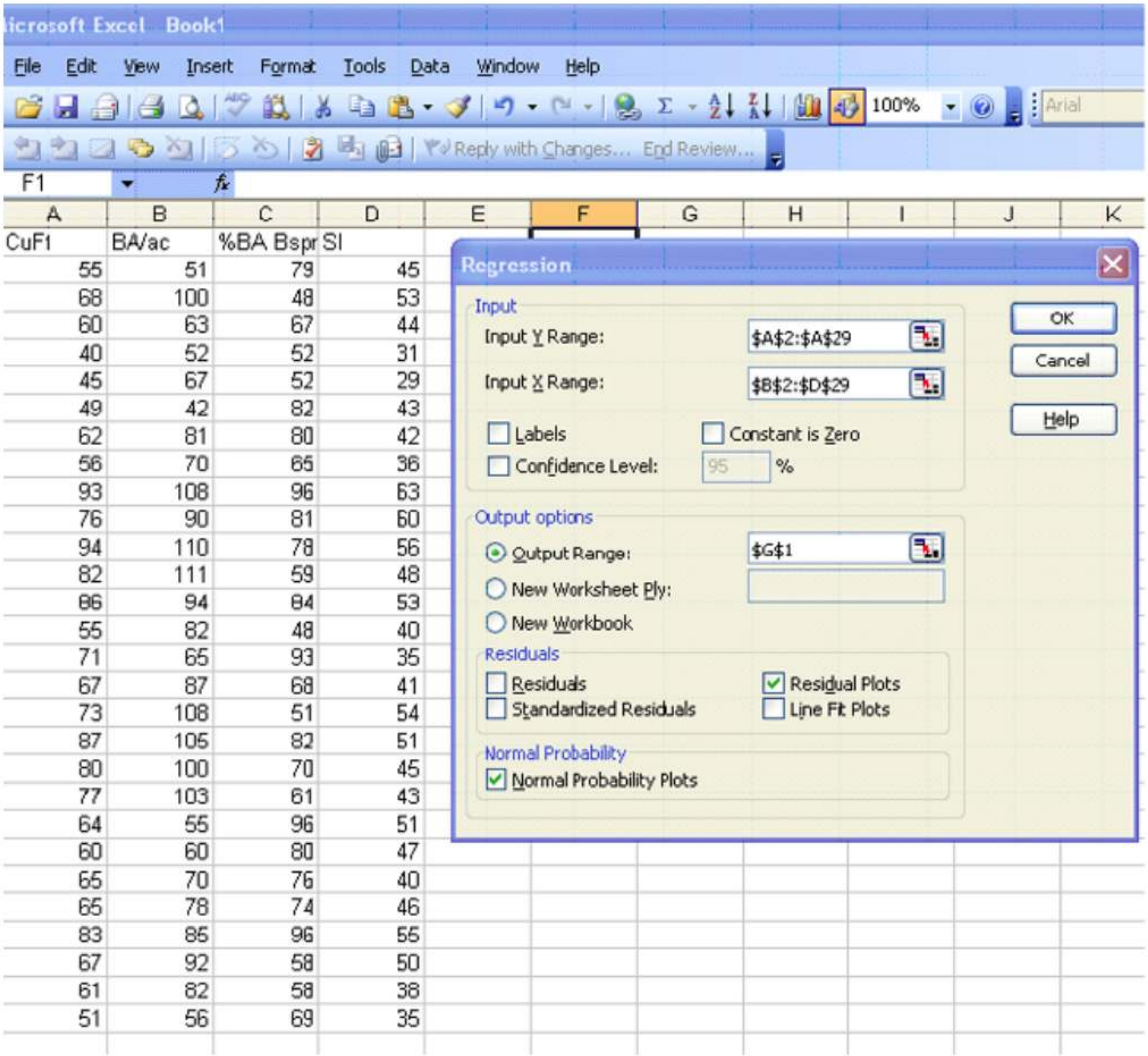
Perform polynomial regression with continuous and categorical predictors

The output and plots are given in the previous example.

Excel



CuFt	BA/ac	%BA Bspr	SI
55	51	79	45
68	100	48	53
60	63	67	44
40	52	52	31
45	67	52	29
49	42	82	43
62	81	80	42
56	70	65	36
93	108	96	63
76	90	81	60
94	110	78	56
82	111	59	48
86	94	84	53
55	82	48	40
71	65	93	35
67	87	68	41
73	108	51	54
87	105	82	51
80	100	70	45
77	103	61	43
64	55	96	51
60	60	80	47
65	70	76	40
65	78	74	46
83	85	96	55
67	92	58	50
61	82	58	38
51	56	69	35


icrosoft Excel - Book1
File Edit View Insert Format Tools Data Window Help
100%
Arial
Reply with Changes... End Review...
F1
A B C D E F G H I J K
Regression
OK
Cancel
Help

CuFt	BA/ac	%BA Bspr	SI
55	51	79	45
68	100	48	53
60	63	67	44
40	52	52	31
45	67	52	29
49	42	82	43
62	81	80	42
56	70	65	36
93	108	96	63
76	90	81	60
94	110	78	56
82	111	59	48
86	94	84	53
55	82	48	40
71	65	93	35
67	87	68	41
73	108	51	54
87	105	82	51
80	100	70	45
77	103	61	43
64	55	96	51
60	60	80	47
65	70	76	40
65	78	74	46
83	85	96	55
67	92	58	50
61	82	58	38
51	56	69	35

Input

Input Y Range: $A$2:$A$29

Input X Range: $B$2:$D$29

☐ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

● Output Range: $G$1

○ New Worksheet Ply:

○ New Workbook

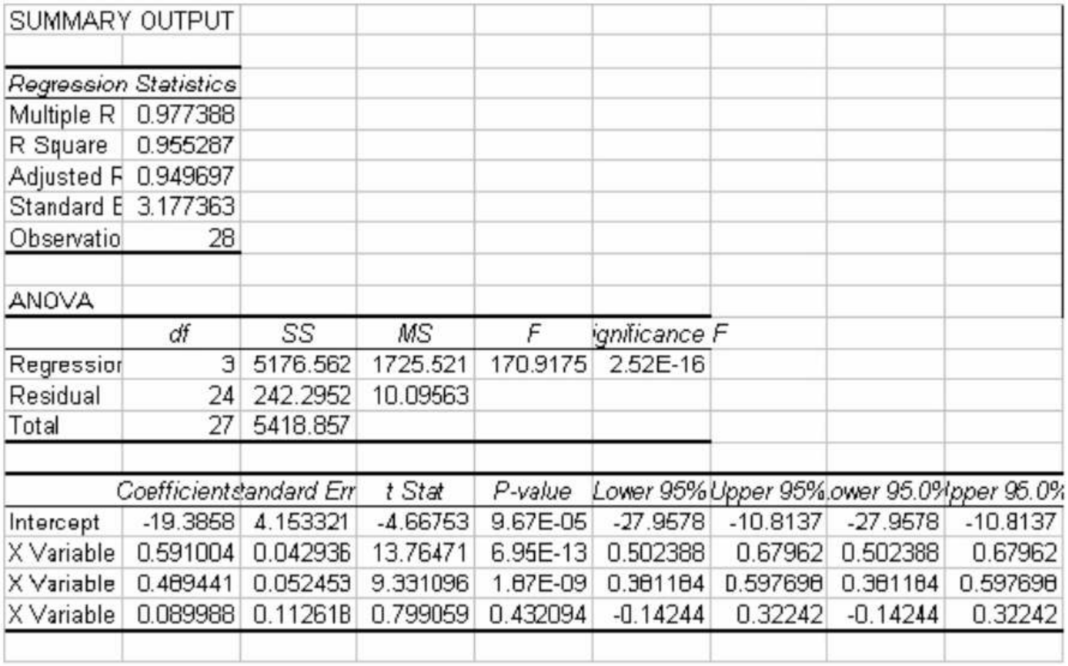
Residuals

☐ Residuals ☑ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☑ Normal Probability Plots



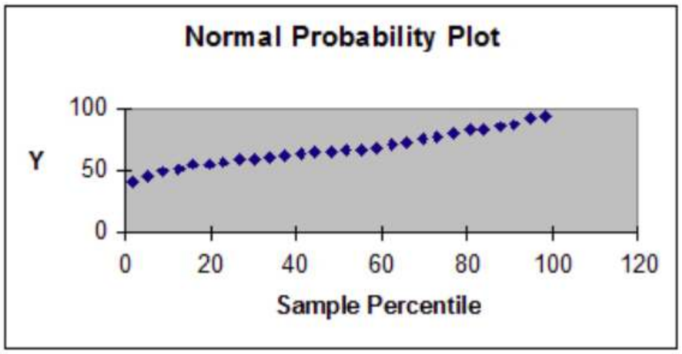
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.977388
R Square	0.955287
Adjusted F	0.949697
Standard E	3.177363
Observatio	28

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	3	5176.562	1725.521	170.9175	2.52E-16
Residual	24	242.2952	10.09563		
Total	27	5418.857			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	-19.3858	4.153321	-4.66753	9.67E-05	-27.9578	-10.8137	-27.9578	-10.8137
X Variable	0.591004	0.042936	13.76471	6.95E-13	0.502388	0.67962	0.502388	0.67962
X Variable	0.489441	0.052453	9.331096	1.87E-09	0.381184	0.597698	0.381184	0.597698
X Variable	0.089988	0.112618	0.799059	0.432094	-0.14244	0.32242	-0.14244	0.32242



This page titled 8.2: Software Solution is shared under a CC BY-NC-SA 3.0 license and was authored, remixed, and/or curated by Diane Kiernan (OpenSUNY) via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

9: Modeling Growth, Yield, and Site Index

9.1: Growth and Yield Models

Forest and natural resource management decisions are often based on information collected on past and present resource conditions. This information provides us with not only current details on the timber we manage (e.g., volume, diameter distribution) but also allows us to track changes in growth, mortality, and ingrowth over time. We use this information to make predictions of future growth and yield based on our management objectives. Techniques for forecasting stand dynamics are collectively referred to as growth and yield models. Growth and yield models are relationships between the amount of yield or growth and the many different factors that explain or predict this growth.

Before we continue our examination of growth and yield models, let's review some basic terms.

- **Yield**: total volume available for harvest at a given time
- **Growth**: difference in volume between the beginning and end of a specified period of time (V2 – V1)
- **Annual growth**: when growth is divided by number of years in the growing period
- **Model:** a mathematical function used to relate observed growth rates or yield to measured tree, stand, and site variables
- **Estimation:** a statistical process of obtaining coefficients for models that describe the growth rates or yield as a function of measured tree, stand, and site variables
- **Evaluation**: considering how, where, and by whom the model should be used, how the model and its components operate, and the quality of the system design and its biological reality
- **Verification**: the process of confirming that the model functions correctly with respect to the conceptual model. In other words, verification makes sure that there are no flaws in the programming logic or algorithms, and no bias in computation (systematic errors).
- **Validation**: checks the accuracy and consistency of the model and tests the model to see how well it reflects the real system, if possible, using an independent data set
- **Simulation**: using a computer program to simulate an abstract model of a particular system. We use a growth model to estimate stand development through time under alternative conditions or silvicultural practices.
- **Calibration**: the process of modifying the model to account for local conditions that may differ from those on which the model was based
- **Monitoring**: continually checking the simulation output of the system to identify any shortcomings of the model
- **Deterministic model**: a model in which the outcomes are determined through known relationships among states and events, without any room for random variation. In forestry, a deterministic model provides an estimate of average stand growth, and given the same initial conditions, a deterministic model will always predict the same result.
- **Stochastic model**: a model that attempts to illustrate the natural variation in a system by providing different predictions (each with a specific probability of occurrence) given the same initial conditions. A stochastic model requires multiple runs to provide estimates of the variability of predictions.
- **Process model**: a model that attempts to simulate biological processes that convert carbon dioxide, nutrients, and moisture into biomass through photosynthesis
- **Succession model**: a model that attempts to model species succession, but is generally unable to provide reliable information on timber yield

Models

Growth and yield models are typically stated as mathematical equations and can be implicit or explicit in form. An implicit model defines the variables in the equation but the specific relationship is not quantified. For example,

$$V = f(BA, H_t) \tag{9.1.1}$$

where *V* is volume (ft3/ac), *BA* is density (basal area in ft2), *Ht* is total tree height. This model says that volume is a function of (depends on) density and height, but it does not put a numerical value on the volume for specific values of basal area and height. This equation becomes explicit when we specify the relationship such as

\[ln(*V*) = -0.723 + 0.781*ln(*BA*)+ 0.922 ln(*H_t*)\]

Growth and yield models can be linear or nonlinear equations. In this linear model, all the independent variables of *X*1 and *X*2 are only raised to the first power.

$$y = 1.29 + 7.65X1 - 27.02X2 \tag{9.1.2}$$

A nonlinear model has independent variables with exponents different from one.

$$y = b_o e^{b_1 X} \tag{9.1.3}$$

In this example, *b*0 and *b*1 are parameters to be estimated and *X* is the independent variable.

Classification of Growth and Yield Models

Growth and yield models have long been part of forestry but development and use has greatly increased in the last 25 years due to the accessibility of computers. There are many different approaches to modeling, each with their own advantages and disadvantages. Selecting a specific type of modeling approach often depends on the type of data used. Growth and yield models are categorized depending on whether they model the whole stand, the diameter classes, or individual trees.

Whole Stand Models

Whole stand models may or may not contain density as an independent variable. Density-free whole stand models provide the basis for traditional normal yield tables since "normal" implies nature's maximum density, and empirical yield tables assume nature's average density. In both of these cases, stand volume at a specific age is typically a function of stand age and site index. Variable-density whole stand models use density as an explicit independent variable to predict current or future volume. Buckman (1962) published the first study in the United States that directly predicted growth from current stand variables, then integrated the growth function to obtain yield:

$$Y = 1.6689 + 0.041066BA – 0.00016303BA_2 – 0.076958A + 0.00022741A_2 + 0.06441S \tag{9.1.4}$$

where *Y* = periodic net annual basal area increment
BA = basal area, in square feet per acre
A = age, in years
S = site index

Diameter distribution models are a refinement of whole stand models. This type of model disaggregates the results at each age and then adds additional information about diameter class structure such as height and volume. The number of stems in each class is a function of the stand variables and all growth functions are for the stand. This type of whole stand model provides greater detail of the stand conditions in terms of volume, tree size, and value.

Diameter Class Models

Diameter class models (not to be confused with diameter distribution models) simulate growth and volume for each diameter class based on the average tree in each class. The number of trees in each class is empirically determined. The diameter class volumes are computed separately for each diameter class, then summed up to obtain stand values. Stand table projection is a common diameter class method used to predict short-term future conditions based on observable diameter growth for that stand. Mortality, harvest, and ingrowth must be computed separately. Differences in projection methods are based on the distribution of the number of stems in each class and how the growth rate is applied. For example, the simplest projection method is based on two assumptions: 1) that all tree diameters in a diameter class equal the midpoint diameter for that class, and 2) that they all grow at the same average rate. An improvement upon this method is to use a movement ratio that defines the proportion of trees which move into a higher DBH class.

$$m = \frac{g}{i} x100 \tag{9.1.5}$$

where *m* is the movement ratio, *g* is the average periodic diameter increment for that specific class, and *i* is the diameter class interval. Let's look at an example.

Assume for a specific DBH class that *g* is 1.2 in. and *i* (class interval) is 2.0 in.

$$m = \frac{1.2}{2.0} x1000 = 60 \tag{9.1.6}$$

This means that 60% of the trees in that diameter class will move up to the next diameter class, and 40% will remain in this class. If the diameter class interval was one inch, the movement ratio would be different.

$$m = \frac{1.2}{1.0} x100 = 120 \tag{9.1.7}$$

In this case, all the trees in this diameter class would move up at least one size class and 20% of them would move up two size classes.

Individual Tree Models

Individual tree models simulate the growth of each individual tree in the tree list. These models are more complex but have become more common as computing power has increased. Individual tree models typically simulate the height, diameter, and survival of each tree while calculating its growth. Individual tree data are aggregated *after* the model grows each tree, while stand models aggregate individual tree data into stand variables *before* the growth model is applied. Additionally, this type of model allows the user to include a measure of competition for each tree. Because of this, individual tree models are typically divided into two groups based on how competition is treated.

Distance-independent models define the competitive neighborhood for a subject tree by its own diameter, height, and condition to stand characteristics such as basal area, number of trees per area, and average diameter, however, the distances between trees are not required for computing the competition for each tree. Distance-dependent models include distance and bearing to all neighboring trees, along with their diameter. This way, the competitive neighborhood for each subject tree is precisely and uniquely defined. While this approach seems logically superior to distance-independent methods, there has not been any clear documented evidence to support the use of distance-dependent competition measures over distance-independent measures.

There are many growth and yield models and simulators available and it can be difficult to select the most appropriate model. There are advantages and disadvantages to many of these options and foresters must be concerned with the reliability of the estimates, the flexibility of the model to deal with management alternatives, the level of required detail, and the efficiency for providing information in a clear and useable fashion. Many models have been created using a broad range of available data. These models are best used for comparative purposes only. In other words, they are most appropriate when comparing the outcomes from different management options instead of predicting results for a specific stand. It is important to review and understand the foundations for any model or simulator before using it.

Forest Vegetation Simulator

The Forest Vegetation Simulator (FVS, Wykoff et al. 1982; Dixon 2002) is a distance-independent, individual-tree forest growth model commonly used in the United States to support forest management decisions. Projections are typically made at the stand level, but FVS has the ability to expand the spatial scope to much larger management units. FVS began as the Prognosis Model for Stand Development (Stage 1973) with the objective to predict stand dynamics in the mixed forests of Idaho and Montana. This model became the common modeling platform for the USDA Forest Service and was renamed FVS.

Stands are the basic unit of management and projections are dependent on the interactions among trees within stands using key variables such as density, species, diameter, height, crown ratio, diameter growth, and height growth. Values for slope, aspect, elevation, density, and a measure of site potential are included for each plot. There are 22 geographically specific versions of FVS called variants.

NE-TWIGS (Belcher 1982) is a common variant applicable to fourteen northeastern states. Stand growth projections are based on simulating the growth and mortality for trees in the 5-inch and larger DBH classes. Ingrowth can be manually entered or simulated using an automatic ingrowth function. The growth equation annually estimates a diameter for each sample tree and updates the crown ratio of the tree (Miner et al. 1988).

*Annual diameter growth = potential growth*competition modifier*

Potential growth is defined as the growth of the top 10% of the fastest growing trees and is predicted using the following equation:

$$Potential\ growth = b_1 * SI * [1.0 - e^{(-b_2 * D)}] \tag{9.1.8}$$

where,

potential growth is defined as the potential annual basal area growth of a tree (sq. ft./yr)
$b1$ and $b2$ are species specific coefficients
SI is site index (index age 50 years) and
D is current tree diameter in in.

The competition modifier is an index bounded from 0 to 1, and is found by:

Competition modifier = $e^{-b_3 * BA}$

where $b3$ is a species-specific coefficient and
BA is the current basal area (sq. ft./ac).

Tree mortality is calculated by estimating the probability of death of each tree in a given year:

Survival = 1-[1/(1+en)]

where $n = c_1 + c_2 * (D+1)^{c_3} * e^{c_4*D - c_5*BA - c_6*SI}$
$c1,\ldots,c6$ are species-specific coefficients
D is current tree diameter (inches)
BA is stand basal area (sq. ft./ac) and
SI is site index.

Inventory data and site information are entered into FVS, and a self-calibration process adjusts the growth models to match the rates present in the entered data. Harvests can be simulated with growth and mortality rates based on post-removal stand densities. Growth cycles run for 5-10 years and output includes a summary of current stand conditions, sampling statistics, and calibration results.

Applications of Regression Techniques

Regression models serve many purposes in the management of natural and forest resources. The following examples serve to highlight some of these applications.

Weight Scaling for Sawlogs

In 1962, Bower created the following equation for predicting loblolly pine sawlog volume based on truckload weights and the number of logs per truck:

$$Y = -3.954N + 0.0925W \tag{9.1.9}$$

where Y = total board-foot volume (International1/4- rule) for a truckload of logs
N = number of 16-ft logs on the truck
W = total load weight (lb.)

Notice that there is no y-intercept in the model. When there are no logs on the truck, there is no volume to be estimated.

Rates of Stem Taper

Kozak et al. (1969) developed a technique for estimating the fraction of volume per tree located in logs of any specified length and dib for any system of scaling (board feet, cubic feet, or weight). Their regression model also predicted taper curves and upper stem diameters (dib) for some conifer species.

$$\frac{d^2}{dbh^2} = b_0 + b_1(\frac{h}{H}) + b_2(\frac{h^2}{H^2}) \tag{9.1.10}$$

where d = stem diameter at any height h above ground
H = total tree height

This equation resolves to:

$$d = dbh\sqrt{b_0 + b_1(\frac{h}{H}) + b_2(\frac{h^2}{H^2})} \tag{9.1.11}$$

The predictor variables are the ratio, and squared ratio, of any height to total height.

Multiple Entry Volume Table that Allows for Variable Utilization Standards

Foresters commonly want to predict tree volume for various top diameters but many of the available volume equations were created for specific top limits. Burkhart (1977) created a regression model to predict volume (cubic feet) of loblolly pine to any desired merchantable top limit. His approach predicted total stem volume, then converted total volume to merchantable volume by applying predicted ratios of merchantable volume to total volume.

$$V = 0.34864 + 0.00232dbh^2H \tag{9.1.12}$$

$$R = 1 - 0.32354(\frac{d_t^{3.1579}}{dbh^{2.7115}}) \tag{9.1.13}$$

where *dbh* = diameter at breast height (in.)
H = total tree height (ft.)
V = total stem cubic-foot volume
R = merchantable cubic-foot volume to top diameter *d*t divided by total stem cubic-foot volume
*d*t = top dob (in.)

Weight Tables for Tree Boles

Belanger (1973) utilized a combined-variable approach to develop predictions of green-weight and dry weight of sycamore tree:

$$GBW = -32.35109 + 0.15544 dbh^2 H \tag{9.1.14}$$

$$DBW = -17.67910 + 0.06684 dbh^2 H \tag{9.1.15}$$

where *GBW* = green bole weight to 3-in.top (lb.)
DBW = dry bole weight to 3-in.top (lb.)
dbh = diameter at breast height (in.)
H = total tree height (ft.)

Biomass Prediction

A common approach to predicting tree biomass weight has been to use a logarithmic combined-variable formula (e.g. Edwards and McNab 1979). The observed relationship between these variables is typically non-linear, therefore a log or natural log transformation is needed to linearize the relationship.

$$log\, Y = b_0 + b_1 log\, dbh^2 H \tag{9.1.16}$$

where *Y* = total tree weight
dbh = diameter at breast height
H = total tree height

However, past studies (Tritton and Hornbeck 1982 and Wiant et al. 1979) indicated that there was little model improvement when height was added. Many dry-weight biomass models now follow this form:

$$ln\, wt = b_0 + b_1 ln\, dbh \tag{9.1.17}$$

$$wt = e^{b_0} dbh^{b_1} \tag{9.1.18}$$

where *wt* = total tree weight
dbh = diameter at breast height

Volume Predictions based on Stump Diameter

Bylin (1982) created a regression model to predict tree volume using stump diameter and stump height for species in Louisiana.

$$V = b_0 + b_1 S_{DIB}^2 + b_2 H_s \tag{9.1.19}$$

where *V* = tree volume (cu. ft.)
SDIB = stump diameters inside bark (in.)
HS = stump height (ft.)

Yield Estimation

MacKinney and Chaiken (1939) were the first to use multiple regression, with stand density as a predictor variable, to predict yield for loblolly pine trees.

$$log\, Y = b_0 + b_1 \frac{1}{A} + b_2 S + b_3 log\, SDI + b_4 C \tag{9.1.20}$$

where

- *Y* = yield (cu. ft./ac)

- *A* = stand age
- *S* = site index
- *SDI* = stand-density index
- *C* = composition index (loblolly pine BA/total BA)

Growth and Yield Prediction for Uneven-aged Stands

Moser and Hall (1969) developed a yield equation, expressed as a function of time, initial volume, and basal area, to predict volume in mixed northern hardwoods.

$$Y=[(Y_0)(8.3348BA_0^{-1.3175})]x[0.9348-(0.9348-1.0203BA_0^{-0.0125})e^{-0.0062t}]^{-105.5} \tag{9.1.21}$$

where

- *Y*0 = initial volume (cu. ft./ac)
- *BA*0 = initial basal area (sq. ft./ac)
- *t* = elapsed time interval (years from initial condition)
- *Y* = predicted volume (cu. ft./ac) *t* years after observation of initial conditions *Y*0 and *BA*0 at time *t*0

9.2: Site Index

Site Index

Site is defined by the Society of American Foresters (1971) as “an area considered in terms of its own environment, particularly as this determines the type and quality of the vegetation the area can carry.” Forest and natural resource managers use site measurement to identify the potential productivity of a forest stand and to provide a comparative frame of reference for management options. The productive potential or capacity of a site is often referred to as **site quality**.

Site quality can be measured directly or indirectly. Direct measurement of a stand’s productivity can be measured by analyzing the variables such as soil nutrients, moisture, temperature regimes, available light, slope, and aspect. A productivity-estimation method based on the permanent features of soil and topography can be used on any site and is suitable in areas where forest stands do not presently exist. Soil site index is an example of such an index. However, such indices are location specific and should not be used outside the geographic region in which they were developed. Unfortunately, environmental factor information is not always available and natural resource managers must use alternative methods.

Historical yield records also provide direct evidence of a site’s productivity by averaging the yields over multiple rotations or cutting cycles. Unfortunately, there are limited long-term data available, and yields may be affected by species composition, stand density, pests, rotation age, and genetics. Consequently, indirect methods of measuring site quality are frequently used, with the most common involving the relationship between tree height and tree age.

Using stand height data is an easy and reliable way to quantify site quality. Theoretically, height growth is sensitive to differences in site quality and height development of larger trees in an even-aged stand is seldom affected by stand density. Additionally, the volume-production potential is strongly correlated with height-growth rate. This measure of site quality is called site index and is the average total height of selected dominant-codominant trees on a site at a particular reference or index age. If you measure a stand that is at an index age, the average height of the dominant and codominant trees is the site index. It is the most widely accepted quantitative measure of site quality in the United States for even-aged stands (Avery and Burkhart 1994).

The objective of the site index method is to select the height development pattern that the stand can be expected to follow during the remainder of its life (not to predict stand height at the index age). Most height-based methods of site quality evaluation use site index curves. Site index curves are a family of height development patterns referenced by either age at breast height or total age. For example, site index curves for plantations are generally based on total age (years since planted), where age at breast height is frequently used for natural stands for the sake of convenience. If total age were to be used in this situation, the number of years required for a tree to grow from a seedling to DBH must be added in. Site index curves can either be anamorphic or polymorphic curves. Anamorphic curves (most common) are a family of curves with the same shape but different intercepts. Polymorphic curves are a family of curves with different shapes and intercepts.

The index age for this method is typically the culmination of mean annual growth. In the western part of the United States, 100 years is commonly used as the reference age with 50 years in the eastern part of this country. However, site index curves can be based on any index age that is needed. Coile and Schumacher (1964) created a family of anamorphic site index curves for plantation loblolly pine with an index age of 25 years. The following family of anamorphic site index curves for a southern pine is based on a reference age of 50 years.

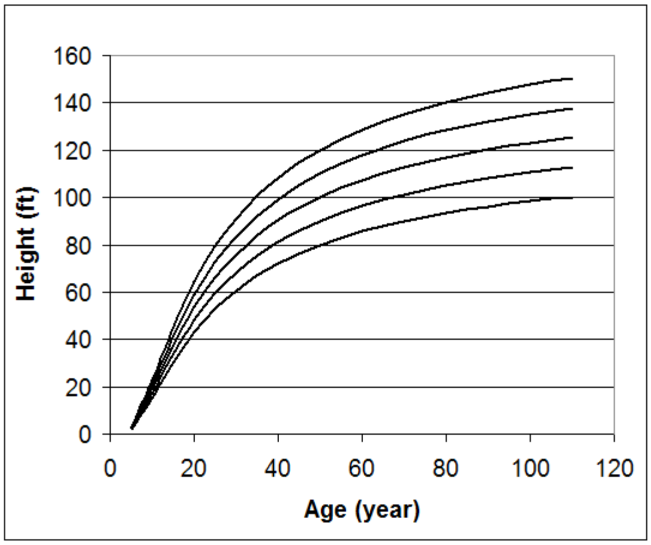


Figure 1. Site index curves with an index age of 50 years.

Creating a site index curve involves the random selection of dominant and codominant trees, measuring their total height, and statistically fitting the data to a mathematical equation. So, which equation do you use? Plotting height over age for single species, even-aged stands typically results in a sigmoid shaped pattern.

$$H_d = b_0 e^{(b_1 A^{-1})} \tag{9.2.1}$$

where *H*d is the height of dominant and codominant trees, *A* is stand age, and *b*0 and *b*1 are coefficients to be estimated. Variable transformation is needed if linear regression is to be used to fit the model. A common transformation is

$$ln\ H_d = b_0 + b_1 A^{-1} \tag{9.2.2}$$

Coile and Schumacher (1964) fit their data to the following model:

$$ln\ S = ln\ H + 5.190\left(\frac{1}{A} - \frac{1}{25}\right) \tag{9.2.3}$$

where *S* is site index, *H* is total tree height, and *A* is average age. The site index curve is created by fitting the model to data from stands of varying site qualities and ages, making sure that all necessary site index classes are equally represented at all ages. It is important not to bias the curve by using an incomplete range of data.

Data for the development of site index equations can come from measurement of tree or stand height and age from temporary or permanent inventory plots or from stem analysis. Inventory plot data are typically used for anamorphic curves only and sampling bias can occur when poor sites are over represented in older age classes. Stem analysis can be used for polymorphic curves but requires destructive sampling and it can be expensive to obtain such data.

We are going to examine three different methods for developing site index equations:

1. Guide curve method
2. Difference equation method
3. Parameter prediction method

Guide Curve Method

The guide curve method is commonly used to generate anamorphic site index equations. Let's begin with a commonly used model form:

$$ln\ H_d = b_0 + b_1 A^{-1} = b_0 + b_1 \frac{1}{A} \tag{9.2.4}$$

Parameterizing this model results in a "guide curve" (the average line for the sample data) that is used to create the individual height/age development curves that parallel the guide curve. For a particular site index the equation is:

$$ln\ H_d = b_{0i} + b_1 A^{-1} \tag{9.2.5}$$

where *b*oi is the unique y-intercept for that age. By definition, when *A* = *A*0 (index age), *H* is equal to site index *S*. Thus:

$$b_{0i} = ln\ S - b_1 A_0^{-1} \tag{9.2.6}$$

Substituting *b*oi into equation 9.2.5 gives:

which can be used to generate site index curves for given values of *S* and A_0 and a range of ages (*A*). The equation can be algebraically rearranged as:

$$ln\ S = ln\ H - b_1(A^{-1} - A_0^{-1}) = ln(H) - b_1(\frac{1}{A} - \frac{1}{A_0}) \tag{9.2.7}$$

This is the form to estimate site index (height at index age) when height and age data measurements are given. This process is sound only if the average site quality in the sample data is approximately the same for all age classes. If the average site quality varies systematically with age, the guide curve will be biased.

Difference Equation Method

This method requires either monumented plot, tree remeasurement data, or stem analysis data. The model is fit using differences of height and specific ages. This method is appropriate for anamorphic and polymorphic curves, especially for longer and/or multiple measurement periods. Schumacher (after Clutter et al. 1983) used this approach when estimating site index using the reciprocal of age and the natural log of height. He believed that there was a linear relationship between Point A (1/*A*1, ln*H*1) and Point B (1/*A*2, ln*H*2) and defined β1 (slope) as:

$$\beta_1 = \frac{ln(H_2) - ln(H_1)}{(1/A_2) - (1/A_1)} \tag{9.2.8}$$

where *H*1 and *A*1 were initial height and age, and *H*2 and *A*2 were height and age at the end of the remeasurement period. His height/age model became:

$$ln(H_2) = ln(H_1) + \beta_1(\frac{1}{A_2} - \frac{1}{A_1}) \tag{9.2.9}$$

Using remeasurement data, this equation would be fitted using linear regression procedures with the model

$$Y = \beta_1 X \tag{9.2.10}$$

where *Y* = ln(*H*2) – ln(*H*1)
X = (1/*A*2) – (1/*A*1)

After estimating β1, a site index equation is obtained from the height/age equation by letting *A*2equal *A*0 (the index age) so that *H*2 is, by definition, site index (*S*). The equation can then be written:

$$ln(S) = ln(H_1) + \beta_1(\frac{1}{A_0} - \frac{1}{A_1}) \tag{9.2.11}$$

Parameter Prediction Method

This method requires remeasurement or stem analysis data, and involves the following steps:

1. Fitting a linear or nonlinear height/age function to the data on a tree-by-tree (stem analysis data) or plot by plot (remeasurement data) basis
2. Using each fitted curve to assign a site index value to each tree or plot (put A0 in the equation to estimate site index)
3. Relating the parameters of the fitted curves to site index through linear or nonlinear regression procedures

Trousdell et al. (1974) used this approach to estimate site index for loblolly pine and it provides an example using the Chapman-Richards (Richards 1959) function for the height/age relationship. They collected stem analysis data on 44 dominant and codominant trees that had a minimum age of at least 50 years. The Chapman-Richards function was used to define the height/age relationship:

$$H = \theta_1[1 - e^{(-\theta_2 A)}]^{[(1-\theta_3)^{-1}]} \tag{9.2.12}$$

where H is height in feet at age A and θ1, θ2, and θ3 are parameters to be estimated. This equation was fitted separately to each tree. The fitted curves were all solved with $A = 50$ to obtain site index values (S) for each tree.

The parameters θ1, θ2, and θ3 were hypothesized to be functions of site index, where

$$\theta_1 = \beta_1 + \beta_2 S \tag{9.2.13}$$

$$\theta_2 = \beta_3 + \beta_4 S + \beta_5 S^2 \tag{9.2.14}$$

$$\theta_3 = \beta_6 + \beta_7 S + \beta_8 S^2 \tag{9.2.15}$$

The Chapman-Richards function was then expressed as:

$$H = (\beta_1 + \beta_2 S) 1 - e^{[-(\beta_3+\beta_4 S+\beta_5 S^2) A]^{[(1-\beta_6-\beta_7 S-\beta_8 S^2)^{-1}]}} \tag{9.2.16}$$

This function was then refitted to the data to estimate the parameters β1, β2, …β8. The estimating equations obtained for θ1, θ2, and θ3 were

$$\hat{\theta_1} = 63.1415 + 0.635080 S \tag{9.2.17}$$

$$\hat{\theta_2} = 0.00643041 + 0.000124189 S + 0.00000162545 S^2$$

$$\hat{\theta_3} = 0.0172714 - 0.00291877 S + 0.0000310915 S^2$$

For any given site index value, these equations can be solved to give a particular Chapman-Richards site index curve. By substituting various values of age into the equation and solving for H, we obtain height/age points that can be plotted for a site index curve. Since each site index curve has different parameter values, the curves are polymorphic.

Periodic Height Growth Data

An alternative to using current stand height as the surrogate for site quality is to use periodic height growth data, which is referred to as a growth intercept method. This method is practical only for species that display distinct annual branch whorls and is primarily used for juvenile stands because site index curves are less dependable for young stands.

This method requires the length measurement of a specified number of successive annual internodes or the length over a 5-year period. While the growth-intercept values can be used directly as measures of site quality, they are more commonly used to estimate site index.

Alban (1972) created a simple linear model to predict site index for red pine using 5-year growth intercept in feet beginning at 8 ft. above ground.

$SI = 32.54 + 3.43\ X$

where SI is site index at a base age of 50 years and X is 5-year growth intercept in feet.

Using periodic height growth data has the advantage of not requiring stand age or total tree height measurements, which can be difficult in young, dense stands. However, due to the short-term nature of the data, weather variation may strongly influence the internodal growth thereby rendering the results inaccurate.

Site index equations should be based on biological or mathematical theories, which will help the equation perform better. They should behave logically and not allow unreasonable values for predicted height, especially at very young or very old ages. The equations should also contain an asymptotic parameter to control unbounded height growth at old age. The asymptote should be some function of site index such that the asymptote increases with increases of site index.

When using site index, it is important to know the base age for the curve before use. It is also important to realize that site index based on one base age cannot be converted to another base age. Additionally, similar site indices for different species do not mean similar sites even when the same base age is used for both species. You have to understand how height and age were measured before you can safely interpret a site index curve. Site index is not a true measure of site quality; rather it is a measure of a tree growth component that is affected by site quality (top height is a measure of stand development, NOT site quality).

9.3: Reference

D.H. Alban, “An Improved Growth Intercept Method for Estimating Site Index of Red Pine,” *U.S. Forest Serv., North Central Forest Expt. Sta*., Res. Paper NC-80, 1972, p. 7.

T.E. Avery and H.E. Burkhart, *Forest Measurements*,. McGraw-Hill, 1994, p. 408.

R.P. Belanger, “Volume and Weight Tables for Plantation-grown Sycamore,” *U.S. Forest Serv. Southeast. Forest Expt. Sta*. Res. Paper SE-107, 1973, p. 8.

D.M. Belcher, “TWIGS: The Woodman’s Ideal Growth Projection System,” *Microcomputers, a New Tool for Foresters*, Purdue University Press, 1982, p. 70.

D.R. Bower, “Volume-weight Relationships for Loblolly Pine Sawlogs,” *J. Forestry* 60, 1962, pp. 411-412.

R.R. Buckman, “Growth and Yield of Red Pine in Minnesota,” *U.S. Department of Agriculture*,Technical Bulletin 1272, 1962.

H.E. Burkhart, “Cubic-foot Volume of Loblolly Pine to Any Merchantable Top Limit,” *So. J. Appl. For.* 1, 1977, pp. 7-9.

C.V. Bylin, “Volume Prediction from Stump Diameter and Stump Height of Selected Species in Louisiana,” *U.S. Forest Serv., Southern Forest. Expt. Sta*., Res. Paper SO-182, 1982, p. 11.

J.R. Clutter Et al., *Timber Management: A Quantitative Approach*, Wiley, 1983, p. 333.

T.S. Coile and F. X. Schumacher, *Soil-site Relations, Stand Structure, and Yields of Slash and Loblolly Pine Plantations in the Southern United States*, T.S. Coile, 1964.

G.E. Dixon (Comp.), “Essential FVS: A User’s Guide to the Forest Vegetation Simulator,” Internal Report. *U.S. Department of Agriculture, Forest Service, Forest Management Service Center*,2002, p. 189.

M.B. Edwards and W.H. McNab, “Biomass Prediction for Young Southern Pines,” *J. Forestry*, 77, 1979, pp. 291-292.

A.D. Kozak, D.D. Munro, and J.H.G. Smith, “Taper Functions and Their Application in Forest Inventory,” *Forestry Chronicle* 45, 1969, pp. 278-283.

A.L. MacKinney and L.E. Chaiken, “Volume, Yield, and Growth of Loblolly Pine in the Mid-Atlantic Coastal Region,” *U.S. Forest. Serv. Appalachian Forest Expt. Sta*., Tech. Note 33, 1939, p. 30.

C.L. Miner, N.R. Walters, and M.L. Belli, “A Guide to the TWIGS Program for the North Central United States,” *USDA Forest Serv., North Central Forest Exp.Sta.*, Gen. Tech. Rep. NC-125, 1988, p. 105.

J.W. Moser, Jr. and O.F. Hall, “Deriving Growth and Yield Functions for Uneven-aged Forest Stands,” *Forest Sci*. 15, 1969, pp. 183-188.

F.J. Richards, “A Flexible Growth Function for Empirical Use,” *J. Exp. Botany*, vol. 10, no. 2 1959, pp. 290-300.

Society of American Foresters, *Terminology of Forest Science, Technology, Practice, and Products*,Washington,D.C., Society of American Foresters, 1971, p. 349.

A.R. Stage, “Prognosis Model for Stand Development,” *U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Expt. Sta*., Res. Pa INT-137, 1973, p. 32.

L.M. Tritton and J.W. Hornbeck, “Biomass Equations for Major Tree Species of the Northeast,” *USDA For. Serv.* Gen. Tech. Rep. NE-GTR-69, 1982.

K.B. Trousdell, D.E. Beck, and F.T. Lloyd, “Site Index for Loblolly Pine in the Atlantic Coastal Plain of the Carolinas and Virginia,” *U.S. Southeastern Forest Expt. Sta*., 1974, p. 115.

H.J. Wiant et al., “Equations for Predicting Weights of Some Appalachian Hardwoods,” *West Virginia Univ. Agric. and Forest Expt. Sta., Coll.. of Agric. and Forest. West Virginia* Forest. Notes, no. 7, 1979.

W.R. Wykoff, N.L. Crookston, and A.R. Stage, “User’s Guide to the Stand Prognosis Model,” *U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Expt. Sta*., Gen. Tech. Re INT-133, 1982, p. 112.

CHAPTER OVERVIEW

10: Quantitative Measures of Diversity, Site Similarity, and Habitat Suitability

---

10.1: Introduction, Simpson’s Index and Shannon-Weiner Index

As forest and natural resource managers, we must be aware of how our timber management practices impact the biological communities in which they occur. A silvicultural prescription is going to influence not only the timber we are growing but also the plant and wildlife communities that inhabit these stands. Landowners, both public an(18)}{d private, often require management of non-timber components, such as wildlife, along with meeting the financial objectives achieved through timber management. Resource managers must be cognizant of the effect management practices have on plant and wildlife communities. The primary interface between timber and wildlife is habitat, and habitat is simply an amalgam of environmental factors necessary for species survival (e.g., food or cover). The key component to habitat for most wildlife is vegetation, which provides food and structural cover. Creating prescriptions that combine timber and wildlife management objectives are crucial for sustainable, long-term balance in the system.

So how do we develop a plan that will encompass multiple land use objectives? Knowledge is the key. We need information on the habitat required by the wildlife species of interest and we need to be aware of how timber harvesting and subsequent regeneration will affect the vegetative characteristics of the system. In other words, we need to understand the diversity of organisms present in the community and appreciate the impact our management practices will have on this system.

Diversity of organisms and the measurement of diversity have long interested ecologists and natural resource managers. Diversity is variety and at its simplest level it involves counting or listing species. Biological communities vary in the number of species they contain (richness) and relative abundance of these species (evenness). Species richness, as a measure on its own, does not take into account the number of individuals of each species present. It gives equal weight to those species with few individuals as it does to a species with many individuals. Thus a single yellow birch has as much influence on the richness of an area as 100 sugar maple trees. Evenness is a measure of the relative abundance of the different species making up the richness of an area. Consider the following example.

✓ Example 10.1.1:

Number of Individuals

Tree Species	Sample 1	Sample 2
Sugar Maple	167	391
Beech	145	24
Yellow Birch	134	31

Both samples have the same richness (3 species) and the same number of individuals (446). However, the first sample has more evenness than the second. The number of individuals is more evenly distributed between the three species. In the second sample, most of the individuals are sugar maples with fewer beech and yellow birch trees. In this example, the first sample would be considered more diverse.

A diversity index is a quantitative measure that reflects the number of different species and how evenly the individuals are distributed among those species. Typically, the value of a diversity index increases when the number of types increases and the evenness increases. For example, communities with a large number of species that are evenly distributed are the most diverse and communities with few species that are dominated by one species are the least diverse. We are going to examine several common measures of species diversity.

Simpson’s Index

Simpson (1949) developed an index of diversity that is computed as:

$$D=\sum_{i=1}^{R}\left(\frac{n_i(n_i-1)}{N(N-1)}\right) \tag{10.1.1}$$

where *ni* is the number of individuals in species *i*, and *N* is the total number of species in the sample. An equivalent formula is:

$$D=\sum_{i=1}^{R} p_i^2 \tag{10.1.2}$$

where p_i is the proportional abundance for each species and R is the total number of species in the sample. Simpson's index is a weighted arithmetic mean of proportional abundance and measures the probability that two individuals randomly selected from a sample will belong to the same species. Since the mean of the proportional abundance of the species increases with decreasing number of species and increasing abundance of the most abundant species, the value of D obtains small values in data sets of high diversity and large values in data sets with low diversity. The value of Simpson's D ranges from 0 to 1, with 0 representing infinite diversity and 1 representing no diversity, so the larger the value of D, the lower the diversity. For this reason, Simpson's index is usually expressed as its inverse (1/*D*) or its compliment (1-*D*) which is also known as the Gini-Simpson index. Let's look at an example.

✓ Example 10.1.2:calculating Simpson's Index

We want to compute Simpson's D for this hypothetical community with three species.

Species	No. of individuals
Sugar Maple	35
Beech	19
Yellow Birch	11

First, calculate N.

$$N=35+19+11=65$$

Then compute the index using the number of individuals for each species:

$$D=\sum_{i=1}^{R}\left(\frac{n_i(n_i-1)}{N(N-1)}\right)=\left(\frac{35(34)}{65(64)}+\frac{19(18)}{65(64)}+\frac{11(10)}{65(64)}\right)=0.3947$$

The inverse is found to be:

$$\frac{1}{0.3947}=2.5336$$

Using the inverse, the value of this index starts with 1 as the lowest possible figure. The higher the value of this inverse index the greater the diversity. If we use the compliment to Simpson's D, the value is:

$$1-0.3947=0.6053$$

This version of the index has values ranging from 0 to 1, but now, the greater the value, the greater the diversity of your sample. This compliment represents the probability that two individuals randomly selected from a sample will belong to different species. It is very important to clearly state which version of Simpson's D you are using when comparing diversity.

Shannon-Weiner Index

The Shannon-Weiner index (Barnes et al. 1998) was developed from information theory and is based on measuring uncertainty. The degree of uncertainty of predicting the species of a random sample is related to the diversity of a community. If a community has low diversity (dominated by one species), the uncertainty of prediction is low; a randomly sampled species is most likely going to be the dominant species. However, if diversity is high, uncertainty is high. It is computed as:

$$H'=-\sum_{i=1}^{R} ln(p_i)=ln\left(\frac{1}{\prod_{i=1}^{R} p_i^{p_i}}\right) \tag{10.1.3}$$

where *pi* is the proportion of individuals that belong to species *i* and *R* is the number of species in the sample. Since the sum of the *pi*'s equals unity by definition, the denominator equals the weighted geometric mean of the *pi* values, with the *pi* values being used as weights. The term in the parenthesis equals true diversity *D* and H'=ln(*D*). When all species in the data set are equally common, all *p*i values = 1/*R* and the Shannon-Weiner index equals ln(*R*). The more unequal the abundance of species, the larger the weighted geometric mean of the *p*i values, the smaller the index. If abundance is primarily concentrated into one species, the index will be close to zero.

An equivalent and computationally easier formula is:

$$H' = \frac{N ln\, N - \sum(n_i ln\, n_i)}{N} \tag{10.1.4}$$

where *N* is the total number of species and *n*i is the number of individuals in species *i*. The Shannon-Weiner index is most sensitive to the number of species in a sample, so it is usually considered to be biased toward measuring species richness.

Let's compute the Shannon-Weiner diversity index for the same hypothetical community in the previous example.

✓ Example 10.1.3:Calculating Shannon-Weiner Index

Species	No. of individuals
Sugar Maple	35
Beech	19
Yellow Birch	11

We know that N = 65. Now let's compute the index:

$$H' = \frac{271.335 - (124.437 + 55.944 + 26.377)}{65} = 0.993$$

10.2: Rank Abundance Graphs and Habitat Suitability Index

Rank Abundance Graphs

Species abundance distribution can also be expressed through rank abundance graphs. A common approach is to plot some measure of species abundance against their rank order of abundance. Such a plot allows the user to compare not only relative richness but also evenness. Species abundance models (also called abundance curves) use all available community information to create a mathematical model that describes the number and relative abundance of all species in a community. These models include the log normal, geometric, logarithmic, and MacArthur's brokenstick model. Many ecologists use these models as a way to express resource partitioning where the abundance of a species is equivalent to the percentage of space it occupies (Magurran 1988). Abundance curves offer an alternative to single number diversity indices by graphically describing community structure.

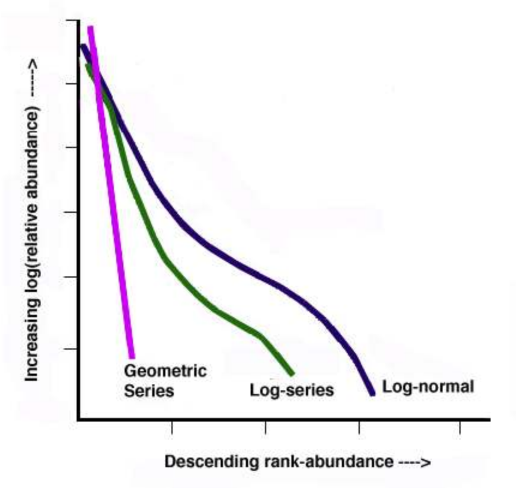


Figure $10.2.1$ *Generic Rank-abundance diagram of three common mathematical models used to fit species abundance distributions: Motomura's geometric series, Fisher's logseries, and Preston's log-normal series (modified from Magurran 1988) by Aedrake09.*

Let's compare the indices and a very simple abundance distribution in two different situations. Stand A and B both have the same number of species (same richness), but the number of individuals in each species is more similar in Stand A (greater evenness). In Stand B, species 1 has the most individuals, with the remaining nine species having a substantially smaller number of individuals per species. Richness, the compliment to Simpson's D, and Shannon's H' are computed for both stands. These two diversity indices incorporate both richness and evenness. In the abundance distribution graph, richness can be compared on the x-axis and evenness by the shape of the distribution. Because Stand A displays greater evenness it has greater overall diversity than Stand B. Notice that Stand A has higher values for both Simpson's and Shannon's indices compared to Stand B.

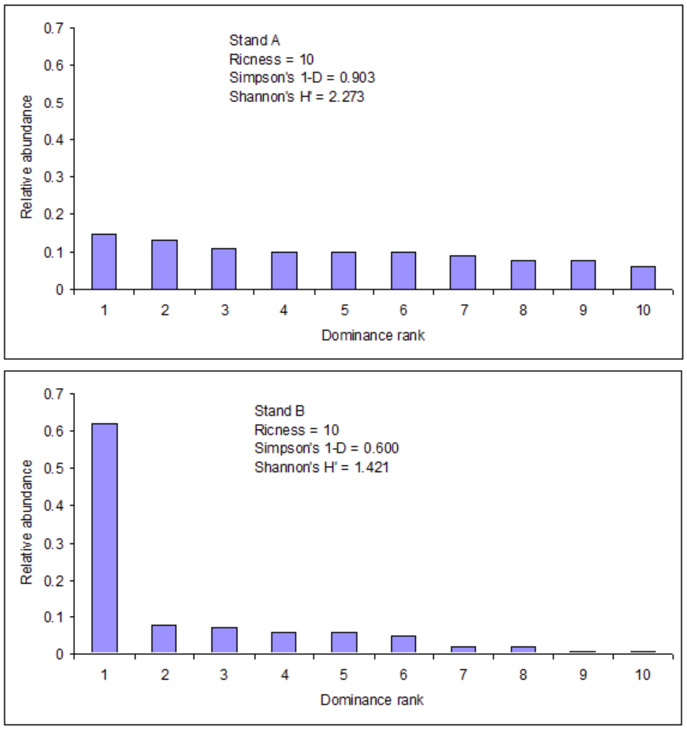


Figure $10.2.2$ *Two stands comparing richness, Simpson's D, and Shannon's index.*

Indices of diversity vary in computation and interpretation so it is important to make sure you understand which index is being used to measure diversity. It is unsuitable to compare diversity between two areas when different indices are computed for each area. However, when multiple indices are computed for each area, the sampled areas will rank similarly in diversity as measured by the different indices. Notice in this previous example both Simpson's and Shannon's index rank Stand A as more diverse and Stand B as less diverse.

Similarity between Sites

There are also indices that compare the similarity (and dissimilarity) between sites. The ideal objective is to express the ecological similarity of different sites; however, it is important to identify the aim or focus of the investigation in order to select the most appropriate index. While many indices are available, van Tongeren (1995) states that most of the indices do not have a firm theoretical basis and suggests that practical experience should guide the selection of available indices.

The Jaccard index (1912) compares two sites based on the presence or absence of species and is used with qualitative data (e.g., species lists). It is based on the idea that the more species both sites have in common, the more similar they are. The Jaccard index is the proportion of species out of the total species list of the two sites, which is common to both sites:

$$SJ = \frac{c}{(a+b+c)} \tag{10.2.1}$$

where SJ is the similarity index, c is the number of shared species between the two sites and a and b are the number of species unique to each site. Sørenson (1948) developed a similarity index that is frequently referred to as the coefficient of community (CC):

$$CC = \frac{2c}{(a+b+2c)} \tag{10.2.2}$$

As you can see, this index differs from Jaccard's in that the number of species shared between the two sites is divided by the average number of species instead of the total number of species for both sites. For both indices, the higher the value the more ecologically similar two sites are.

If quantitative data are available, a similarity ratio (Ball 1966) or a percentage similarity index, such as Gauch (1982), can be computed. Not only do these indices compare number of similar and dissimilar species present between two sites, but also incorporate abundance. The similarity ratio is:

$$SR_{ij}=\frac{\sum y_{ki}y_{kj}}{\sum y_{ki}^2+\sum y_{kj}^2-\sum (y_{ki}y_{kj})} \tag{10.2.3}$$

where *y*ki is the abundance of the *k*th species at site *i* (sites *i* and *j* are compared). Notice that this equation resolves to Jaccard's index when just presence or absence data is available. The percent similarity index is:

$$PS_{ij}=\frac{200\sum min(y_{ki},y_{kj})}{\sum y_{ki}+\sum y_{kj}} \tag{10.2.4}$$

Again, notice how this equation resolves to Sørenson's index with qualitative data only. So let's look at a simple example of how these indices allow us to compare similarity between three sites. The following example presents hypothetical data on species abundance from three different sites containing seven different species (A-G).

Table 10.2.0

	Site		
Species	**1**	**2**	**3**
A	4	0	1
B	0	1	0
C	0	0	0
D	1	0	1
E	1	4	0
F	3	1	1
G	1	0	3

Let's begin by computing Jaccard's and Sørenson's indices for the three comparisons (site 1 vs. site 2, site 1 vs. site 3, and site 2 vs. site 3).

$SJ1,2=\frac{2}{(3+1+2)}=0.33$ $SJ1,3=\frac{4}{(4+1+0)}=0.80$ $SJ2,3=\frac{1}{(1+2+3)}=0.17$

$CC1,2=\frac{2(2)}{(3+1+2(2))}=0.50$ $CC1,3=\frac{2(4)}{(1+0+2(4))}=0.89$ $CC2,3=\frac{2(1)}{(2+3+2(1))}=0.29$

Both of these qualitative indices declare that sites 1 and 3 are the most similar and sites 2 and 3 are the least similar. Now let's compute the similarity ratio and the percent similarity index for the same site comparisons.

$$SR1,2 = \frac{[(4\times 0)+(0\times 1)+(0\times 0)+(1\times 0)+(1\times 4)+\ (3\times 1)+(1\times 0)]}{(4^2+0^2+0^2+1^2+1^2+3^2+1^2)+(0^0+1^2+0^2+0^2+4^2+1^2+0^2)+(4\times 0)+(0\times 1)+(0\times 0)+(1\times 0)+(1\times 4)+\ (3\times 1)+(1\times 0)}$$

$$SR1,2=0.23 \tag{10.2.5}$$

$$SR1,3 = \frac{[(4\times 1)+(0\times 0)+(0\times 0)+(1\times 1)+(1\times 0)+(3\times 1)+(1\times 3)]}{(4^2+0^2+0^2+1^2+1^2+3^2+1^2)+(1^2+0^2+0^2+1^2+0^2+1^2+3^2)+(4\times 1)+(0\times 0)+(0\times 0)+(1\times 1)+(1\times 0)+(3\times 1)+(1\times 3)}$$

$$SR1,3=0.38 \tag{10.2.6}$$

$$SR2,3$$

$$=\frac{[(0\times 1)+(1\times 0)+(0\times 0)+(0\times 1)+(4\times 0)+(1\times 1)+(0\times 3)]}{(0^2+1^2+0^2+0^2+4^2+1^2+0^2)+(1^2+0^2+0^2+1^2+0^2+1^2+3^2)+(0\times 1)+(1\times 0)+(0\times 0)+(0\times 1)}$$

$$+(4\times 0)+(1\times 1)+(0\times 3)$$

$$SR1,3=0.03 \tag{10.2.7}$$

$$PS1,2=\frac{200(0+0+0+0+1+1+0)}{(4+0+0+1+1+3+1)+(0+1+0+0+4+1+0)}=25.0$$

$$PS1,3=\frac{200(1+0+0+1+0+1+1)}{(4+0+0+1+1+3+1)+(1+0+0+1+0+1+3)}=50.0$$

$$PS2,3=\frac{200(0+0+0+0+0+1+0)}{(0+1+0+0+4+1+0)+(1+0+0+1+0+1+3)}=16.7$$

A matrix of percent similarity values allows for easy interpretation (especially when comparing more than three sites).

	1	**2**
2	25.0	
3	50.0	16.7

Table 10.2.1 A matrix of percent similarity for three sites.

The quantitative indices return the same conclusions as the qualitative indices. Sites 1 and 3 are the most similar ecologically, and sites 2 and 3 are the least similar; and also site 2 is most unlike the other two sites.

Habitat Suitability Index (HSI)

In 1980, the U.S. Fish and Wildlife Service (USFWS) developed a procedure for documenting predicted impacts to fish and wildlife from proposed land and water resource development projects. The Habitat Evaluation Procedures (HEP) (Schamberger and Farmer 1978) were developed in response to the need to document the non-monetary value of fish and wildlife resources. HEP incorporates population and habitat theories for each species and is based on the assumption that habitat quality and quantity can be numerically described so that changes to the area could be assessed and compared. It is a species-habitat approach to impact assessment and habitat quality, for a specific species is quantified using a habitat suitability index (HSI).

Habitat suitability index (HSI) models provide a numerical index of habitat quality for a specific species (Schamberger et al. 1982) and in general assume a positive, linear relationship between carrying capacity (number of animals supported by some unit area) and HSI. Today's natural resource manager often faces economically and socially important decisions that will affect not only timber but wildlife and its habitat. HSI models provide managers with tools to investigate the requirements necessary for survival of a species. Understanding the relationships between animal habitat and forest management prescription is vital towards a more comprehensive management approach of our natural resources. An HSI model synthesizes habitat use information into a framework appropriate for fieldwork and is scaled to produce an index value between 0.0 (unsuitable habitat) to 1.0 (optimum habitat), with each increment of change being identical to another. For example, a change in HSI from 0.4 to 0.5 represents the same magnitude of change as from 0.7 to 0.8. The HSI values are multiplied by area of available habitat to obtain Habitat Units (HUs) for individual species. The U.S. Fish and Wildlife Service (USFWS) has documented a series of HSI models for a wide variety of species (FWS/OBS-82/10).

Let's examine a simple HSI model for the marten (*Martes americana*) which inhabits late successional forest communities in North America (Allen 1982). An HSI model must begin with habitat use information, understanding the species needs in terms of food, water, cover, reproduction, and range for this species. For this species, the winter cover requirements are more restrictive than cover requirements for any other season so it was assumed that if adequate winter cover was available, habitat requirements for the rest of the year would not be limiting. Additionally, all winter habitat requirements are satisfied in boreal evergreen forests. Given this, the research identified four crucial variables for winter cover that needed to be included in the model.

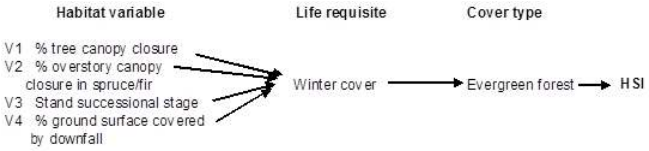

Habitat variable
Life requisite
Cover type
V1 % tree canopy closure
V2 % overstory canopy closure in spruce/fir
V3 Stand successional stage
V4 % ground surface covered by downfall
Winter cover
Evergreen forest
HSI

Figure 10.2.3 Habitat requirements for the marten.

For each of these four winter cover variables (V1, V2, V3, and V4), suitability index graphs were created to examine the relationship between various conditions of these variables and suitable habitat for the marten. A reproduction of the graph for % tree canopy closure is presented below.

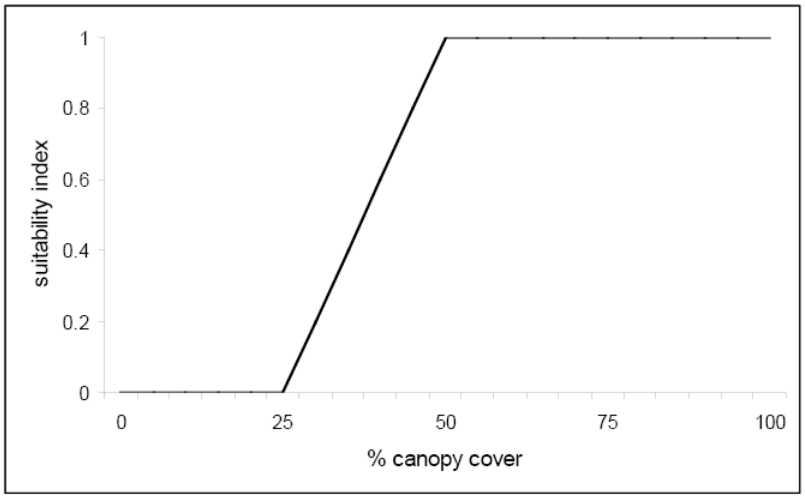


Figure 10.2.4 Suitability index graph for percent canopy cover.

Notice that any canopy cover less than 25% results in unacceptable habitat based on this variable alone. However, once 50% canopy cover is reached the suitability index reaches 1.0 and optimum habitat for this variable is achieved. The following equation was created that combined the life requisite values for the marten using these four variables:

$$\frac{(V_1 \times V_2 \times V_3 \times V_4)}{2}$$

Since winter cover was the only life requisite considered in this model, the HSI equals the winter cover value. As you can see, the more life requisites included in the model, the more complex the model becomes.

While HSI values identify the quality of the habitat for a specific species, wildlife diversity as a whole is a function of size and spatial arrangement of the treated stands (Porter 1986). Horizontal and structural diversity are important. Generally speaking, the more stands of different character an area contains, the greater the wildlife diversity. The spatial distribution of differing types of stands supports animals that need multiple cover types. In order to promote wildlife species diversity, a manager must develop forest management prescription that varies the spatial and temporal patterns of timber reproduction, thereby providing greater horizontal and vertical structural diversity.

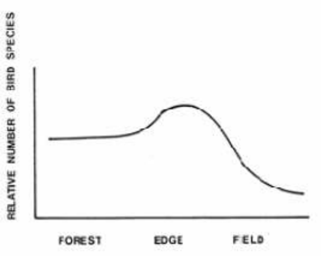


Figure 10.2.5: Bird species diversity nesting across a forest to field gradient (After Strelke and Dickson 1980).

Typically, even-aged management reduces vertical structural diversity, but options such as the shelterwood method tend to mitigate this problem. Selection system tends to promotes both horizontal and vertical diversity.

Integrated natural resource management can be a complicated process but not impossible. Vegetation response to silvicultural prescriptions provides the foundation for understanding the wildlife response. By examining the present characteristics of the managed stands, understanding the future response due to management, and comparing those with the requirements of specific species, we can achieve habitat manipulation together with timber management.

10.3: Reference

Aedrake09. “Modified logseries,” Wikipedia, en.Wikipedia.org/wiki/File:Co...eWhittaker.jpg, 2009.

A.W. Allen, “Habitat Suitability Index Models: Marten,” *U.S.D.I. Fish and Wildlife Service*.FWS/OBS-82/10.11., 1982,9 pp.

B.V.Barnes et al., *Forest Ecology* 4th *ed.* , Wiley, 1998.

P. Jacard, “The Distribution of the Flora of the Alpine Zone,” *New Phytologist* 11, 1912, pp. 37-50.

A.E. Magurran, *Ecological Diversity and Its Measurement*, Princeton Univ. Press, 1988.

W.F. Porter, “Integrating Wildlife Management with Even-aged Timber Systems,” *Managing Northern Hardwoods: Proceedings of a Silvicultural Symposium*, ed. R. Nyland, SUNY College of Environmental Science and Forestry, 23-25 June, 1986, pp. 319-337.

M. Schamberger and A. Farmer, “The Habitat Evaluation Procedures: Their Application in Project Planning and Impact Evaluation,” *Trans. N. A. Wildlife and Natural Resource Conf.* 43, 1978, pp. 274-283.

E.H. Simpson, “Measurement of Diversity,” *Nature* 163, 1949, p. 688.

T. Sørenson, “A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content,” *Det. Kong. Danske Vidensk. Selsk. Biol. Skr*. (Copenhagen) vol. 5, no. 4, 1948, pp. 1-34.

W.K. Strelke and J.G. Dickson, “Effect of Forest Clear-cut Edge on Breeding Birds in East Texas,” *J. Wildl. Manage*. vol, 44, no. 3, 1980, pp. 559-567.

U.S.D.I. Fish and Wildlife Service, “Habitat as a Basis for Environmental Assessment,” 101 ESM, 1980.

O.F.R. van Tongeren, “Cluster Analysis,” *Data Analysis in Community and Landscape Ecology*, Eds. R.H.G. Jongman, C.J.F. Ter Braak, and O.F.R. van Tongeren, 1995, pp. 174-212.

CHAPTER OVERVIEW

11: Biometric Labs

11.1: Biometrics Lab #1

Experiment 1

You are unhappy with the logging company you hired to thin a stand of red pine. You carefully laid out the skid trails leaving bumper trees to avoid excess damage to the remaining trees. In the contract, it is stated that the logging company would pay a penalty (3 times the stumpage rate) for trees damaged beyond the agreed amount of five or more damaged trees per acre. You want to estimate the number of damaged trees per acre to see if they exceeded this amount. You take 27 samples, from which you compute the sample mean, and then construct a 95% confidence interval about the mean number of damaged trees per acre.

2	4	0	3	5	0	0	1	3
2	7	4	8	10	0	2	1	1
5	3	5	6	4	9	5	3	6

Enter these data in the first column of the Minitab worksheet and label it “Trees.” Now calculate the sample mean and sample standard deviation. **Stat**>**Basic Statistics**>**Display Descriptive Statistics**. Select the column with your data in the variable box.

a) sample mean: ______________________________

sample standard deviation: ____________________

Examine the normal probability plot for this data set. Remember, for a sample size less than n = 30, we must verify the assumption of normality if we do not know that the random variable is normally distributed. Go to **GRAPH → PROBABILITY PLOT.** Enter the column with your data in the “**Graph variables**” box and click OK.

b) Would you say that this distribution is normal?

c) Calculate the 95% confidence interval by hand using $x \pm t_{\alpha/2}(\frac{s}{\backslash\text{squr}n})$ and the t-table.

95% CI for the mean number of damaged trees:______________________________________

Now find the 95% confidence interval for the mean using Minitab.

Go to **STAT> Basic Statistics> 1-sample t…**Enter data in **“Samples in columns.”** You do not have to enter the standard deviation but select **OPTIONS** and set the confidence level (make sure it is for 95%) and select “**Alternative:not equal.”**

d) 95% CI for the mean number of damaged trees: ____________________________________

e) Do you have enough statistical evidence to state that the logging company has exceeded the damage limit? Why?

Experiment 2

The amount of sewage and industrial pollution dumped into a body of water affects the health of the water by reducing the amount of dissolved oxygen available for aquatic life. If the population mean dissolved oxygen drops below five parts per million (ppm), a level some scientists think is marginal for supplying enough dissolved oxygen for fish, some remedial action will be attempted. Given the expense of remediation, a decision to take action will be made only if there is sufficient evidence to support the claim that the mean dissolved oxygen has DECREASED below 5 ppm. Below are weekly readings from the same location in a river over a two-month time period.

5.2, 4.9, 5.1, 4.2, 4.7, 4.5, 5.0, 5.2, 4.8, 4.6, 4.8

The population standard deviation is unknown and we have a small sample (n≤30). You must verify the assumption of normality. Go to **GRAPH→PROBABILITY PLOT**. Examine the normal probability plot. Does the distribution look normal?

Use DESCRIPTIVE STATISTICS (**Basic Statistics>Display Descriptive Statistics**) to get the mean and sample standard deviation.

Now test the claim that the mean dissolved oxygen is less than 5ppm using α = 0.05

a) First, state the null and alternative hypotheses

H0:____________________________________

H1: ____________________________________

b) Compute the test statistic by hand $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

c) Find the critical value from the t-table: ____

d) Do you reject the null hypothesis or fail to reject the null hypothesis?

Now use Minitab to do the hypothesis test. Go to **STAT > BASIC STAT > 1-SAMPLE t.** Check **PERFORM HYPOTHESIS TEST** and enter the hypothesized mean (5.00). Click **OPTIONS** and enter the confidence level (1-α) and select alternative hypothesis (H1). Click OK. Check to see that the null and alternative hypotheses shown in the session window are correct.

e) What is the p-value for this test?

f) Do you reject or fail to reject the null hypothesis?

g) State your conclusion:

Experiment 3

A forester believes that tent caterpillars are doing a significant amount of damage to the growth of the hardwood tree species in his stand. He has growth data from 21 plots before the infestation. Since then, he has re-measured those same plots and wants to know if there has been a significant reduction in the annual diameter growth.

Before	After
0.17	0.15
0.22	0.23
0.19	0.17
0.2	0.14
0.12	0.13
0.13	0.11
0.15	0.13
0.16	0.17
0.16	0.12
0.19	0.16
0.25	0.26
0.24	0.21
0.21	0.21
0.18	0.15
0.19	0.17
0.22	0.2
0.24	0.19
0.25	0.24
0.24	0.25
0.14	0.1
0.11	0.11

You need to compute the differences between the *before* values and the *after* values. To create a new variable (diff), type “diff” in the header of the column you want to use. Select **CALC>CALCULATOR**. In the “Expressions” box, type in the equation “**Before-After.**” In the box “**Store results in variable**” type “diff.” Click OK.

You now have a new data set of the *differences* with which you will complete your analyses. Compute basic descriptive statistics to get the sample mean $\bar{d}$ and sample standard deviation s_d of the differences. Use these statistics to test the claim that there has been a reduction in annual diameter growth. You can answer this question by using either a hypothesis test or confidence interval.

a) H0:_______________________________________

H1: _______________________________________

$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$ or $\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$

Do you reject or fail to reject the null hypothesis?

Now let Minitab do the work for you. Select **STAT> Basic Statistics> Paired t…** Select **SAMPLES IN COLUMNS**. Enter the *before* as the **First** sample and *after* data as the **Second** sample. Select **OPTIONS** to set the confidence level and alternative hypothesis. Make sure the Test mean is set to 0.0. Click OK.

b) Write the test statistic and p-value

c) Write a complete conclusion that answers the question.

Experiment 4

Alternative energy is an important topic these days and a researcher is studying a solar electric system. Each day at the same time he collected voltage readings from a meter connected to the system and the data are given below. Is there a significant difference in the mean voltage readings for the different types of days? First do an F-test to test for equal variances and then test the means using the appropriate 2-sample t-test based on the results from the F-test. Please state a complete conclusion for this problem. α = 0.05.

Sunny – 13.5, 15.8, 13.2, 13.9, 13.8, 14.0, 15.2, 12.1, 12.9, 14.9

Cloudy – 12.7, 12.5, 12.6, 12.7, 13.0, 13.0, 12.1, 12.2, 12.9, 12.7

F-Test

a) Write the null and alternative hypotheses to test the claim that the variances are not equal.

H0:_______________________________________ H1: _______________________________________

Select **STAT>BASIC STAT>2 Variances**. In the **Data** box select “Samples in different columns” and enter Sunny in the **First** box and Cloudy in the **Second** box. Click **OPTIONS** and in **Hypothesized Ratio** box select **Variance1/Variance2**. Make sure the **Alternative** is set at “**Not equal**.” Click OK. Look at the p-value for the F-test at the bottom of the output.

b) Do you reject for fail to reject the null hypothesis?

c) Can you assume equal variances?

Now conduct a 2-sample t-test (you should have rejected the null hypothesis in the F-test and assumed unequal variances). **STAT>BASIC STAT>2-Sample t**…Select the button for “**Samples in different columns**” and put Sunny in the **First** box and Cloudy in the **Second** box. Click **OPTIONS** and set the confidence level and select the correct alternative hypothesis. Set the **Test difference** at 0.0. Click OK.

d) What is the p-value for this test?

e) Do you reject or fail to reject the null hypothesis? State your conclusion.

11.2: Biometrics Lab #2

One-way ANOVA Computer Lab

Name: ____

Experiment 1

A forester working with uneven-aged northern hardwoods wants to know if there is a significant difference in total merchantable sawtimber volume (m3ha-1) produced from stands using three different methods of selection system and a 15-yr cutting cycle. The following data are the total merchantable volume from 7 sample plots for each method. If you find a significant difference (reject Ho), then test the multiple comparisons for significant differences. Report the findings using all available information. α=0.05.

Single Tree	Group Selection	Patch Strip
108.6	104.2	102.1
110.9	103.9	101.4
112.4	109.4	100.3
106.3	105.2	95.6
101.4	106.3	102.9
114.6	107.2	99.8
117	105.8	103.5

Write the null and alternative hypotheses.

H0: ___________________________________

H1: ___________________________________

Open Minitab and label the first column as Volume and the second column as Method. Enter all of the volumes in the first column and the methods in the second:

Volume	**Method**
108.6	Single
110.9…	Single…
104.2	Group
103.9…	Group…
102.1	Patch
101.4…	Patch…

Select **STAT>ANOVA>One-way**. In the **Response** box select Volume, and in the **Factor** box select Method. Click on the **Comparisons** box. Select Tukeys, family error rate "5." This tells Minitab that you want to control the experiment-wise error using Tukey's method while keeping the overall level of significance at 5% across all multiple comparisons. Click OK.

State the p-value from the ANOVA table ___________________________________

Write the value for the S2b ___________ and the S2w (MSE) ____________________

Do you reject or fail to reject the null hypothesis? ______________________________

Using the Grouping Information from the Tukey Method, describe the differences in volume produced using the three methods.

____

____

____

Now refer to the Tukey 95% Simultaneous Confidence intervals for the multiple comparisons. What is the Individual confidence interval level? ____________________ This is the adjusted level of significance used for all the multiple comparisons that keeps the 5% level of significance across the total experiment.

Using these confidence intervals, describe the estimated differences in sawtimber volume due to the three different treatments.

Example: The group method results in greater levels of sawtimber volume compared to patch. The group method yields, on average, 0.327 to 10.073 m3 more sawtimber volume per plot than the patch method.

Compare “Single” and “Patch,” and “Single” and “Group.”

____

____

____

____

____

Experiment 2

A plant physiologist is studying the rate of transpirational water loss (ml) of plants growing under five levels of soil moisture stress. This species is an important component to the wildlife habitat in this area and she wants to make sure it survives in an area that tends to be dry. She randomly assigns 18 pots to each treatment (N = 90). She is measuring total rate of water transpiring from the leaves (ml) per pot per unit area. Is there a significant difference in the transpiration rates between the levels of water stress (days)? $\alpha = 0.05$.

0 DAYS	5 DAYS	10 DAYS	20 DAYS	30 DAYS
7.78	7.15	9.1	4.72	1.05
8.09	9.12	5.86	3.53	1.29
7.27	7.67	9.45	4.96	1.11
11.35	10.82	7.14	5	0.83
11.94	12.31	6.87	3.82	1.08
10.89	9.76	8.72	4.36	1.09
10.93	8.46	8.58	2.91	0.75
9.16	11.01	9.93	4.91	0.99
7.83	7.54	9.28	4.99	0.71
8.6	9.48	6.65	4.95	1.02
9.32	9.47	10.55	3.28	1.01
6.46	10.2	7.93	3.53	1.08
8.12	6.04	7.68	5.37	1.99
10.47	7.99	5.42	6.54	3.01
5.98	8.05	4.99	5.51	2.61
6.9	7.42	5.29	4.24	2.99
7.57	5.76	7.65	4.39	2.62

0 DAYS	5 DAYS	10 DAYS	20 DAYS	30 DAYS
9.17	7.78	4.75	4.16	1.98

Write the null and alternative hypotheses.

H0: ____

H1: ____

State the p-value from the ANOVA table ____

Do you reject or fail to reject the null hypothesis? __________________________________

Using the Grouping Information using the Tukey Method, describe the differences in water loss between the five levels of water stress (0, 5, 10, 20, and 30).

____

____

____

____

Now refer to the Tukey 95% Simultaneous Confidence intervals for the multiple comparisons. What is the Individual confidence interval level? ____________________ This is the adjusted level of significance used for all the multiple comparisons that keeps the 5% level of significance across the total experiment.

Using these confidence intervals, describe the estimated differences in water loss between the five different treatments.

____

____

____

____

____

Experiment 3

A rifle club performed an experiment on a randomly selected group of first-time shooters. The purpose was to determine whether shooting accuracy is affected by method of sighting used: only the right eye open, only the left eye open, or both eyes open. Fifteen shooters were all given similar training except in the method of sighting. Their scores are recorded below. At the 0.05 level of significance, is there sufficient evidence to reject the claim that the three methods of sighting are equally effective? $\alpha = 0.05$.

Right	Left	Both
13	10	15
9	18	16
17	15	15
13	11	12
14	15	16

Write the null and alternative hypotheses.

H0: ____

H1: ____

State the p-value from the ANOVA table ____

Do you reject or fail to reject the null hypothesis? __________________________________

Give a complete conclusion.

____

____

____

____

____

Why do you think you were not able to identify any differences between the sighting methods?

____

____

____

____

____

11.3: Biometrics Lab #3

Name: ____

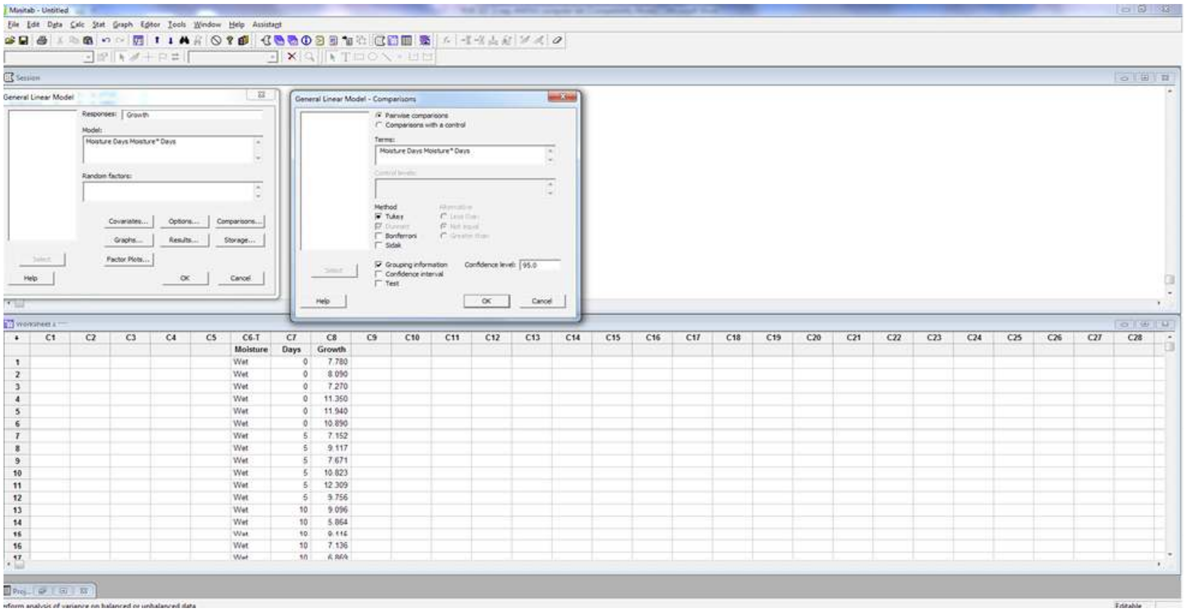
You are studying the growth of a hybrid species of Alaskan pine in three levels of soil moisture (wet, moderate, and dry) over a period of 30 days (0, 5, 10, 20, and 30). You want to determine if this species grows differently over time given different starting levels of soil moisture. Use the given data to test this claim (α = 0.05). If the interaction is significant, at what point does the difference in growth between the levels of soil moisture over time become significant? Use the factor plot and the Grouping information to specifically identify the difference in your conclusion.

Moisture	Days	Growth	Moisture	Days	Growth	Moisture	Days	Growth
Dry	0	7.78	Moderate	0	10.926	Wet	0	8.116
Dry	0	8.09	Moderate	0	9.162	Wet	0	10.473
Dry	0	7.27	Moderate	0	7.83	Wet	0	8.654
Dry	0	11.35	Moderate	0	8.604	Wet	0	6.901
Dry	0	11.94	Moderate	0	9.324	Wet	0	7.565
Dry	0	10.89	Moderate	0	6.462	Wet	0	9.169
Dry	5	7.152	Moderate	5	8.456	Wet	5	10.039
Dry	5	9.117	Moderate	5	11.012	Wet	5	9.994
Dry	5	7.671	Moderate	5	7.541	Wet	5	8.045
Dry	5	10.823	Moderate	5	9.482	Wet	5	9.445
Dry	5	12.309	Moderate	5	9.473	Wet	5	8.024
Dry	5	9.756	Moderate	5	10.2	Wet	5	7.783
Dry	10	9.096	Moderate	10	8.582	Wet	10	7.679
Dry	10	5.864	Moderate	10	9.934	Wet	10	11.671
Dry	10	9.445	Moderate	10	9.279	Wet	10	10.567
Dry	10	7.136	Moderate	10	6.651	Wet	10	9.66
Dry	10	6.869	Moderate	10	10.546	Wet	10	7.646
Dry	10	8.716	Moderate	10	7.927	Wet	10	8.953
Dry	20	4.716	Moderate	20	2.903	Wet	20	7.368
Dry	20	3.528	Moderate	20	4.91	Wet	20	6.539
Dry	20	4.964	Moderate	20	4.998	Wet	20	7.034
Dry	20	5.004	Moderate	20	4.954	Wet	20	7.258
Dry	20	3.824	Moderate	20	3.279	Wet	20	6.309
Dry	20	4.356	Moderate	20	3.528	Wet	20	7.223
Dry	30	1.053	Moderate	30	0.748	Wet	30	4.909
Dry	30	1.287	Moderate	30	0.997	Wet	30	5.891
Dry	30	1.11	Moderate	30	0.7	Wet	30	4.223
Dry	30	0.832	Moderate	30	1.018	Wet	30	3.997

Moisture	Days	Growth	Moisture	Days	Growth	Moisture	Days	Growth
Dry	30	1.082	Moderate	30	1.007	Wet	30	2.616
Dry	30	1.095	Moderate	30	1.083	Wet	30	3.995

Open Minitab and enter the data into a spreadsheet. Select **STAT>ANOVA>General Linear Model**.

Click in the **Response** box and select GROWTH for the Response box, and enter MOISTURE, DAYS, and MOISTURE*DAYS (interaction term) in the **Model** box, as shown.


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Covariates... Options... Comparisons...
Graphs... Results... Storage...
Factor Plots...
Select Help OK Cancel
Worksheet 1

General Linear Model

Responses: Growth

Model: Moisture Days Moisture* Days

Random factors:

General Linear Model - Comparisons

● Pairwise comparisons

○ Comparisons with a control

Terms: Moisture Days Moisture* Days

Method

☑ Tukey

☐ Bonferroni

☐ Sidak

☑ Grouping information

☐ Confidence interval

☐ Test

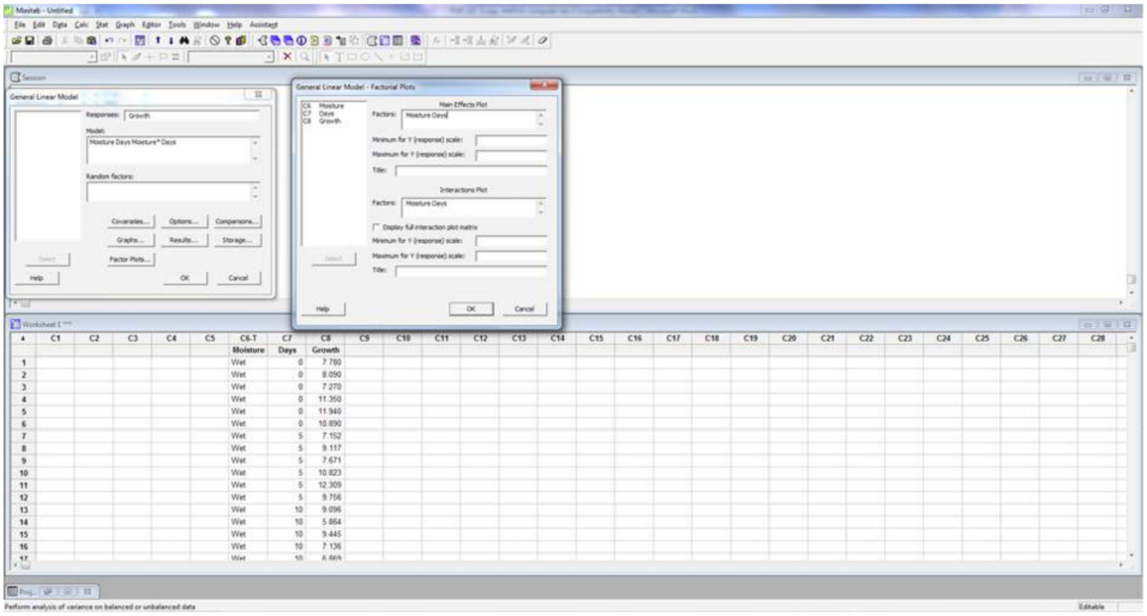
Confidence level: 95.0

	C6-T Moisture	C7 Days	C8 Growth
1	Wet	0	7.780
2	Wet	0	8.090
3	Wet	0	7.270
4	Wet	0	11.350
5	Wet	0	11.940
6	Wet	0	10.890
7	Wet	5	7.152
8	Wet	5	9.117
9	Wet	5	7.671
10	Wet	5	10.823
11	Wet	5	12.309
12	Wet	5	9.756
13	Wet	10	9.096
14	Wet	10	5.864
15	Wet	10	9.445
16	Wet	10	7.136

Under **OPTIONS**, select “Adjusted (Type III)” under **Sums of Squares**. Click OK.

Under **COMPARISONS**, select “Pairwise comparisons” using “**Tukey**” method and enter the two main effects and interaction (MOISTURE, DAYS, and MOISTURE*DAYS) in the terms box (click in the box first to select).

Check the **Grouping Information** box. Click OK.


Minitab - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Covariates... Options... Comparisons...
Graphs... Results... Storage...
Factor Plots...
Select Help OK Cancel
C6 Moisture
C7 Days
C8 Growth
Worksheet 1
Perform analysis of variance on balanced or unbalanced data
Editable

General Linear Model

Responses: Growth

Model: Moisture Days Moisture* Days

Random factors:

General Linear Model - Factorial Plots

Main Effects Plot

Factors: Moisture Days

Minimum for Y (response) scale:

Maximum for Y (response) scale:

Title:

Interactions Plot

Factors: Moisture Days

☐ Display full interaction plot matrix

Minimum for Y (response) scale:

Maximum for Y (response) scale:

Title:

	C6-T Moisture	C7 Days	C8 Growth
1	Wet	0	7.780
2	Wet	0	8.090
3	Wet	0	7.270
4	Wet	0	11.350
5	Wet	0	11.940
6	Wet	0	10.890
7	Wet	5	7.152
8	Wet	5	9.117
9	Wet	5	7.671
10	Wet	5	10.823
11	Wet	5	12.309
12	Wet	5	9.756
13	Wet	10	9.096
14	Wet	10	5.864
15	Wet	10	9.445
16	Wet	10	7.136

Under **RESULTS**, select “Analysis of Variance Table” for **Display of Results**. Click OK.

Under **FACTOR PLOTS**, enter MOISTURE and DAYS in both the main effects and interaction plot box. Click OK. Click OK.

Is the interaction term significant? ___________________

Write the p-value __________________________________

Use the third Grouping Information Using Tukey Method (for the interaction) and the Factor plot to determine where the differences are for each treatment.

Attach a complete conclusion describing the differences in growth for this species over the 30 days for the 3 different levels of soil moisture.

11.4: Biometrics Lab #4

Name: ____

Experiment 1

The following data were collected on Old Faithful geyser in Yellowstone Park. The **x**-variable is time between eruptions and the **y**-variable is length of eruptions.

X	Y
12.17	1.88
11.63	1.77
12.03	1.83
12.15	1.83
11.30	1.70
11.70	1.82
12.27	1.93
11.60	1.77
11.72	1.83
12.10	1.89
11.70	1.80
11.40	1.72
11.22	1.75
11.42	1.73
11.53	1.74
11.50	1.77
11.90	1.87
11.86	1.84

a) Determine if a relationship exists between the 2 variables using a scatterplot and the linear correlation coefficient. Select **Graph**> **Scatterplot.** Select the **Simple** plot and click OK. Enter the response variable (length of eruptions) in the **Y variables** box, and the predictor variable (time between eruptions) in the **X variables** box. Click OK. Describe the relationship that you see.

____

____

____

____

b) Calculate the linear correlation coefficient. **Statistics**> **Basic Stats**> **Correlation.** Enter the 2 variables in the **Variables** box and click OK.

r = ____________________________________

What two pieces of information about the relationship between these two variables does the linear correlation coefficient tell you?

____

____

____

____

____

c) Find a least squares regression line treating “time between eruptions” as the predictor variable (x) and “length of eruptions” as the response variable (y). **Stat**>**Regression**> **General Regression.**Enter “length of eruptions” in the **Response** box. Enter “time between eruptions” in the **Model** box. Click on **Options** and make sure that 95% is selected for all confidence intervals. Click on **Graphs**and select the **Residual plot** “Residual versus fits.” Click **Results** and make sure the Regression equation, Coefficient table, Display confidence intervals, Summary of model, Analysis of Variance table, and prediction tables are checked. Click OK.

Write the regression equation ____________________

What is the value of R2? _________________________

What does this mean?

____

____

____

Examine the residual model. Do you see any problems?

____

____

____

What is the value of the regression standard error? ________________________________

Write the confidence intervals for the y-intercept _________________________________

and slope ____

Use the output to test if the slope is significantly different from zero. Write the null and alternative hypotheses for this test.

H0:____

H1: ____

Using the test statistic and p-value from the Minitab output to test this claim.

Test statistic__________________________________ p-value __________________________________

Conclusion:

____

____

____

____

d) Using the regression equation, what would be the length of the eruption if the time between eruptions is 11.42 min.?

Experiment 2

The index of biotic integrity (IBI) is a measure of water quality in streams. The sample data given in the table below comes from the Piedmont forest region. The table gives the data for IBI and forested area in square kilometers. Let Forest Area be the predictor variable (x) and IBI be the response variable (y).

15108.png

Create a scatterplot and describe the relationship between these variables. Compute the linear correlation coefficient.

r = ____

Create a regression model for this data set following the steps from the first example. Write the regression model.

____

Is there significant evidence to support the claim that IBI increases with Forest Area? Write the test statistic/p-value used for this slope test along with your answer.

____

____

____

____

The researcher wants to estimate the population mean IBI for streams that have an average forested area of 48 sq. km. Click **STAT>REGRESSION> GENERAL REGRESSION**. Making sure that IBI is in the Response box and Forest Area is in the Model box, click on **Prediction** and enter 48 in the **New observation for continuous predictors** box and check Confidence limits. Click OK. Write the 95% confidence interval for mean IBI for streams in an average forested area of 48 sq. km. ____

You are working with a stream in an area with 19 sq. km. of forested area. Your management plan includes an afforestation project that will increase the forested area to 23 sq. km. You need to predict what the specific IBI would be for this stream when the forested area is increased. Create a prediction interval to estimate this IBI if the forested area increased to 23 sq. km.

Click **STAT>REGRESSION>GENERAL REGRESSION**. Making sure that IBI is in the Response box and Forest Area is in the Model box, click on **Prediction** and enter 23 in the **New observation for continuous predictors** box and check **Prediction limits**. Click OK. Write the 95% prediction interval for the IBI for this stream when the forested area is increased to 23 sq. km. ____

Explain the difference between the confidence and prediction intervals you just computed.

____

____

____

____

____

11.5: Biometrics Lab #5

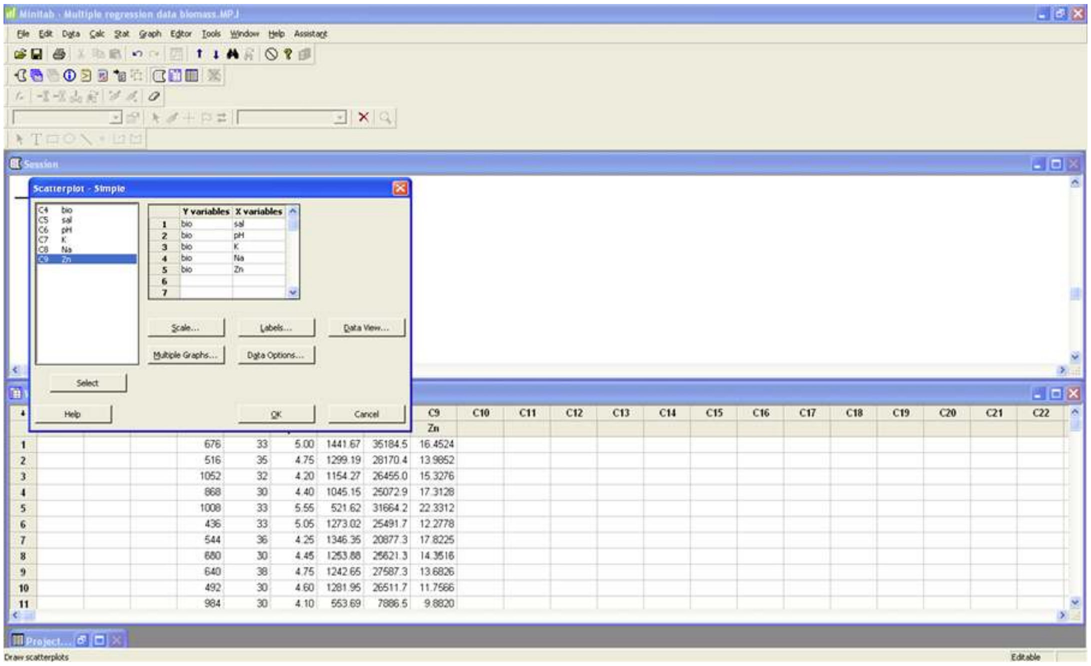
Name: ____

You are working on an alternative energy source and biomass is a key component. You want to predict above-ground biomass for this region, and you believe that biomass is related to substrate (subsoil) variables of salinity, water acidity, potassium, sodium, and zinc. Your crew collects information on biomass and these five variables for 45 plots.

Experiment 1

Before you create this regression model, you must examine the relationships between each of the five predictor variables and biomass (the response variable). Create five scatterplots using biomass as the response variable (y) and each of the predictor variables (x). Compute the linear correlation coefficient for each pair. Describe the relationships.

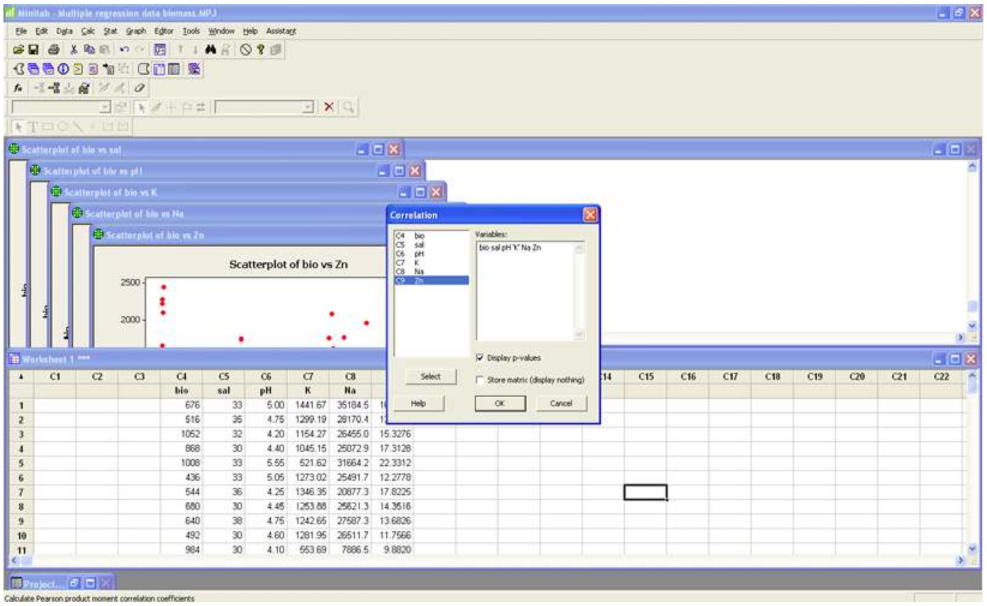
GRAPH>Scatterplot>Simple>OK. The response variable (y-variable) is Bio and the five predictor variables are the x-variables. Look at the scatterplots and describe each relationship below. Next compute the correlation coefficient for each pair and write the r-value below. **STAT>Basic Statistics>Correlation**. You can easily do all correlations at once by creating a correlation matrix. Put all predictor variables in the **Variables** box together.



C4	bio
C5	sal
C6	pH
C7	K
C8	Na
C9	Zn

	Y variables	X variables
1	bio	sal
2	bio	pH
3	bio	K
4	bio	Na
5	bio	Zn
6		
7		

									C9 Zn
1				676	33	5.00	1441.67	35184.5	16.4524
2				516	35	4.75	1299.19	28170.4	13.9852
3				1052	32	4.20	1154.27	26455.0	15.3276
4				868	30	4.40	1045.15	25072.9	17.3128
5				1008	33	5.55	521.62	31664.2	22.3312
6				436	33	5.05	1273.02	25491.7	12.2778
7				544	36	4.25	1346.35	20877.3	17.8225
8				680	30	4.45	1253.88	25621.3	14.3516
9				640	38	4.75	1242.65	27587.3	13.6826
10				492	30	4.60	1281.95	26511.7	11.7566
11				984	30	4.10	553.69	7886.5	9.8820


Minitab - Multiple regression data biomass.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Scatterplot of bio vs sal
Scatterplot of bio vs pH
Scatterplot of bio vs K
Scatterplot of bio vs Na
Scatterplot of bio vs Zn
2500
2000
Worksheet 1 ***
Project...
Calculate Pearson product moment correlation coefficients

Correlation

C4 bio
C5 sal
C6 pH
C7 K
C8 Na
C9 Zn

Variables: bio sal pH 'K' Na Zn

☑ Display p-values

☐ Store matrix (display nothing)

	C1	C2	C3	C4 bio	C5 sal	C6 pH	C7 K	C8 Na	
1				676	33	5.00	1441.67	35184.5	
2				516	35	4.75	1299.19	28170.4	
3				1052	32	4.20	1154.27	26455.0	15.3276
4				868	30	4.40	1045.15	25072.9	17.3128
5				1008	33	5.55	521.62	31664.2	22.3312
6				436	33	5.05	1273.02	25491.7	12.2778
7				544	36	4.25	1346.35	20877.3	17.8225
8				680	30	4.45	1253.88	25621.3	14.3516
9				640	38	4.75	1242.65	27587.3	13.6826
10				492	30	4.60	1281.95	26511.7	11.7566
11				984	30	4.10	553.69	7886.5	9.8820

Correlation (r) Description

Bio v. sal ____

Bio v.pH ____

Bio v. K ____

Bio v. Na ____

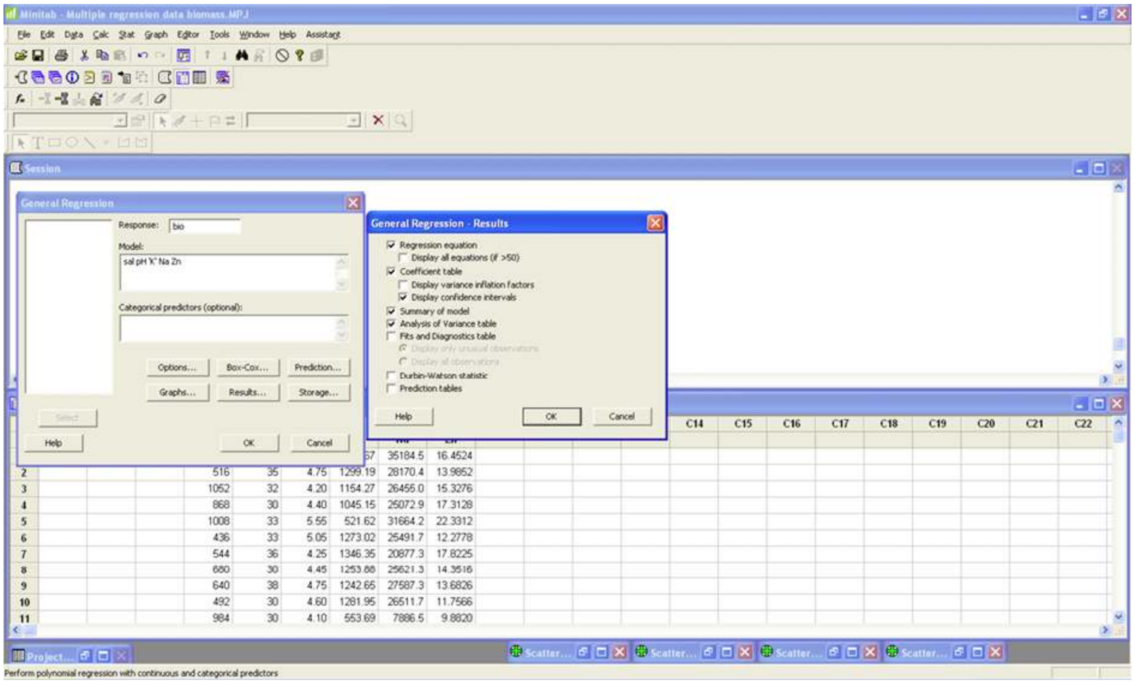
Bio v. Zn ____

Circle the above pair that has the strongest linear relationship.

Experiment 2

You are now going to create four regression models using the predictor variables. You will compare the adjusted R2, regression standard error, p-values for each coefficient, and the residuals for each model. Using this information, you will select the best model and state your reasons for this choice.

Begin with the full model using all five predictor variables. **STAT>Regression>General Regression**. Put Bio in the **Response** box and all five predictor variables in the **Model** box (see image). Click **Results** and make sure that the Regression equation, coefficient table, Display confidence intervals, Summary of Model, and Analysis of Variance Table are checked (see image). Click OK. Click **Graphs** and make sure that under **Residual Plots** that Individual plots and Residual versus Fits are selected (see image). Click OK.


Minitab - Multiple regression data biomass.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Options... Box-Cox... Prediction...
Graphs... Results... Storage...
Select Help OK Cancel
Project...
Scatter...
Perform polynomial regression with continuous and categorical predictors

General Regression

Response: bio

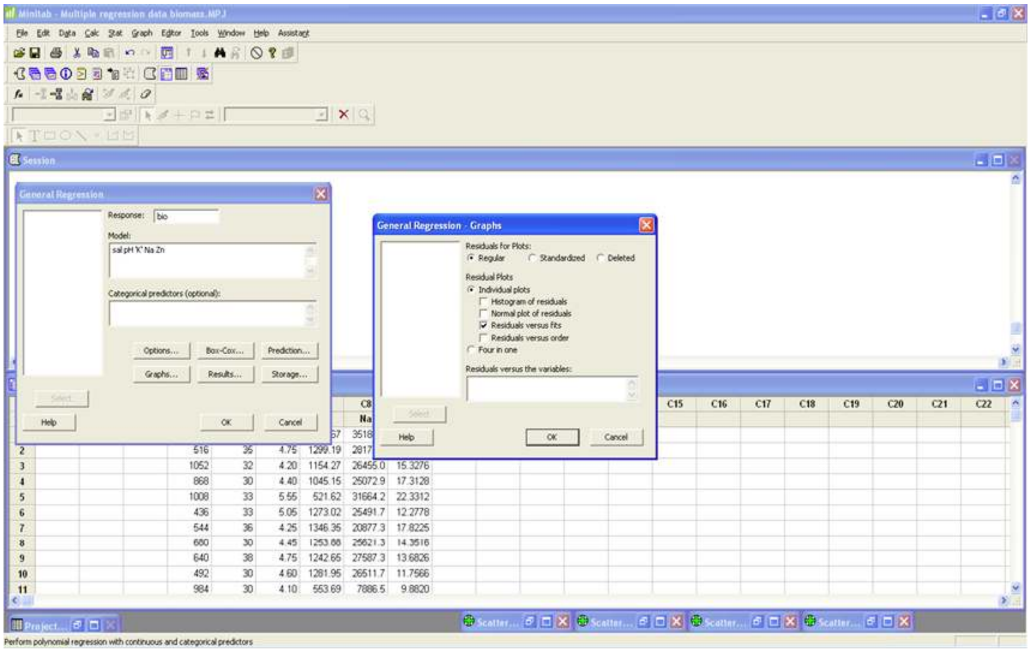
Model: sal pH 'K' Na Zn

Categorical predictors (optional):

General Regression - Results

☑ Regression equation
☐ Display all equations (if >50)
☑ Coefficient table
☐ Display variance inflation factors
☑ Display confidence intervals
☑ Summary of model
☑ Analysis of Variance table
☐ Fits and Diagnostics table
● Display only unusual observations
○ Display all observations
☐ Durbin-Watson statistic
☐ Prediction tables

2	516	35	4.75	1299.19	28170.4	13.9852
3	1052	32	4.20	1154.27	26455.0	15.3276
4	868	30	4.40	1045.15	25072.9	17.3128
5	1008	33	5.55	521.62	31664.2	22.3312
6	436	33	5.05	1273.02	25491.7	12.2778
7	544	36	4.25	1346.35	20877.3	17.8225
8	680	30	4.45	1253.88	25621.3	14.3516
9	640	38	4.75	1242.65	27587.3	13.6826
10	492	30	4.60	1281.95	26511.7	11.7566
11	984	30	4.10	553.69	7886.5	9.8820


Minitab - Multiple regression data biomass.MPJ
File Edit Data Calc Stat Graph Editor Tools Window Help Assistant
Session
Options... Box-Cox... Prediction...
Graphs... Results... Storage...
Select Help OK Cancel
Project...
Scatter...
Perform polynomial regression with continuous and categorical predictors

General Regression

Response: bio

Model: sal pH 'K' Na Zn

Categorical predictors (optional):

General Regression - Graphs

Residuals for Plots:
● Regular ○ Standardized ○ Deleted

Residual Plots
● Individual plots
☐ Histogram of residuals
☐ Normal plot of residuals
☑ Residuals versus fits
☐ Residuals versus order
○ Four in one

Residuals versus the variables:

MODEL 1

Write the regression model ____

Write the adj. R2 ____

Write the regression standard error ____

Examine the residual plot. Are there any problems? ____________________________

Write the variables which are NOT significant _________________________________

MODEL 2

Now remove the LEAST significant variable (highest p-value) and repeat the steps using only the remaining variables.

Write the regression model ____

Write the adj. R2 ____

Write the regression standard error ____

Examine the residual plot. Are there any problems? ____________________________

Write the variables which are NOT significant ________________________________

MODEL 3

Now remove the LEAST significant variable (highest p-value) and repeat the steps using only the remaining variables.

Write the regression model ____

Write the adj. R2 ____

Write the regression standard error ____

Examine the residual plot. Are there any problems? ____________________________

Write the variables which are NOT significant ________________________________

MODEL 4

Now remove the LEAST significant variable (highest p-value) and repeat the steps using only the remaining variables.

Write the regression model ____

Write the adj. R2 ____

Write the regression standard error ____

Examine the residual plot. Are there any problems? ____________________________

Write the variables which are NOT significant ________________________________

Experiment 3

Select the best model and state your reasons for selecting this model.

biomass	sal	pH	K	Na	Zn
676	33	5	1441.67	35184.5	16.4524
516	35	4.75	1299.19	28170.4	13.9852
1052	32	4.2	1154.27	26455	15.3276
868	30	4.4	1045.15	25072.9	17.3128
1008	33	5.55	521.62	31664.2	22.3312
436	33	5.05	1273.02	25491.7	12.2778
544	36	4.25	1346.35	20877.3	17.8225
680	30	4.45	1253.88	25621.3	14.3516
640	38	4.75	1242.65	27587.3	13.6826
492	30	4.6	1281.95	26511.7	11.7566
984	30	4.1	553.69	7886.5	9.882
1400	37	3.45	494.74	14596	16.6752
1276	33	3.45	525.97	9826.8	12.373
1736	36	4.1	571.14	11978.4	9.4058
1004	30	3.5	408.64	10368.6	14.9302
396	30	3.25	646.65	17307.4	31.2865
352	27	3.35	514.03	12822	30.1652
328	29	3.2	350.73	8582.6	28.5901
392	34	3.35	496.29	12369.5	19.8795
236	36	3.3	580.92	14731.9	18.5056
392	30	3.25	535.82	15060.6	22.1344
268	28	3.25	490.34	11056.3	28.6101
252	31	3.2	552.39	8118.9	23.1908
236	31	3.2	661.32	13009.5	24.6917
340	35	3.35	672.15	15003.7	22.6758
2436	29	7.1	528.65	10225	0.3729
2216	35	7.35	563.13	8024.2	0.2703
2096	35	7.45	497.96	10393	0.3205
1660	30	7.45	458.38	8711.6	0.2648
2272	30	7.4	498.25	10239.6	0.2105
824	26	4.85	936.26	20436	18.9875
1196	29	4.6	894.79	12519.9	20.9687
1960	25	5.2	941.36	18979	23.9841
2080	26	4.75	1038.79	22986.1	19.9727
1764	26	5.2	898.05	11704.5	21.3864
412	25	4.55	989.87	17721	23.7063
416	26	3.95	951.28	16485.2	30.5589
504	26	3.7	939.83	17101.3	26.8415
492	27	3.75	925.42	17849	27.7292
636	27	4.15	954.11	16949.6	21.5699
1756	24	5.6	720.72	11344.6	19.6531
1232	27	5.35	782.09	14752.4	20.3295
1400	26	5.5	773.3	13649.8	19.588
1620	28	5.5	829.26	14533	20.1328
1560	28	5.4	856.96	16892.2	19.242

Index

Glossary

Sample Word 1	Sample Definition 1

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Natural Resources Biometrics (Kiernan)

Webpages: 60

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 3.0: 83.3% (50 pages)
- Undeclared: 16.7% (10 pages)

By Page

- Natural Resources Biometrics (Kiernan) - *CC BY-NC-SA 3.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Descriptive Statistics and the Normal Distribution - *CC BY-NC-SA 3.0*
    - 1.1: Descriptive Statistics - *CC BY-NC-SA 3.0*
    - 1.2: Probability Distribution - *CC BY-NC-SA 3.0*
  - 2: Sampling Distributions and Confidence Intervals - *CC BY-NC-SA 3.0*
    - 2.1: Sampling Distribution of the Sample Mean - *CC BY-NC-SA 3.0*
    - 2.2: Confidence Intervals - *CC BY-NC-SA 3.0*
  - 3: Hypothesis Testing - *CC BY-NC-SA 3.0*
    - 3.1: The Fundamentals of Hypothesis Testing - *CC BY-NC-SA 3.0*
    - 3.2: Hypothesis Test about the Population Mean when the Population Standard Deviation is Known - *CC BY-NC-SA 3.0*
    - 3.3: Hypothesis Test about the Population Mean when the Population Standard Deviation is Unknown - *CC BY-NC-SA 3.0*
    - 3.4: Hypothesis Test for a Population Proportion - *CC BY-NC-SA 3.0*
    - 3.5: Hypothesis Test about a Variance - *CC BY-NC-SA 3.0*
    - 3.6: Putting it all Together Using the Classical Method - *CC BY-NC-SA 3.0*
  - 4: Inferences about the Differences of Two Populations - *CC BY-NC-SA 3.0*
    - 4.1: Inferences about Two Means with Independent Samples (Assuming Unequal Variances) - *CC BY-NC-SA 3.0*
    - 4.2: Pooled Two-sampled t-test (Assuming Equal Variances) - *CC BY-NC-SA 3.0*
    - 4.3: Inferences about Two Means with Dependent Samples—Matched Pairs - *CC BY-NC-SA 3.0*
    - 4.4: Inferences about Two Population Proportions - *CC BY-NC-SA 3.0*
    - 4.5: F-Test for Comparing Two Population Variances - *CC BY-NC-SA 3.0*
    - 4.6: Summary - *CC BY-NC-SA 3.0*
  - 5: One-Way Analysis of Variance - *CC BY-NC-SA 3.0*
    - 5.1: Analysis of Variance - *CC BY-NC-SA 3.0*
    - 5.2: Multiple Comparisons - *CC BY-NC-SA 3.0*
  - 6: Two-way Analysis of Variance - *CC BY-NC-SA 3.0*
    - 6.1: Main Effects and Interaction Effect - *CC BY-NC-SA 3.0*
    - 6.2: Multiple Comparisons - *CC BY-NC-SA 3.0*
    - 6.3: Summary And Software Solution - *CC BY-NC-SA 3.0*
  - 7: Correlation and Simple Linear Regression - *CC BY-NC-SA 3.0*
    - 7.1: Correlation - *CC BY-NC-SA 3.0*
    - 7.2: Simple Linear Regression - *CC BY-NC-SA 3.0*
    - 7.3: Population Model - *CC BY-NC-SA 3.0*
    - 7.4: Software Solution - *CC BY-NC-SA 3.0*
  - 8: Multiple Linear Regression - *CC BY-NC-SA 3.0*
    - 8.1: Multiple Regressions - *CC BY-NC-SA 3.0*
    - 8.2: Software Solution - *CC BY-NC-SA 3.0*
  - 9: Modeling Growth, Yield, and Site Index - *CC BY-NC-SA 3.0*
    - 9.1: Growth and Yield Models - *CC BY-NC-SA 3.0*
    - 9.2: Site Index - *CC BY-NC-SA 3.0*
    - 9.3: Reference - *CC BY-NC-SA 3.0*
  - 10: Quantitative Measures of Diversity, Site Similarity, and Habitat Suitability - *CC BY-NC-SA 3.0*
    - 10.1: Introduction, Simpson’s Index and Shannon-Weiner Index - *CC BY-NC-SA 3.0*

    - 10.2: Rank Abundance Graphs and Habitat Suitability Index - *CC BY-NC-SA 3.0*
    - 10.3: Reference - *CC BY-NC-SA 3.0*
  - 11: Biometric Labs - *CC BY-NC-SA 3.0*
    - 11.1: Biometrics Lab #1 - *CC BY-NC-SA 3.0*
    - 11.2: Biometrics Lab #2 - *CC BY-NC-SA 3.0*
    - 11.3: Biometrics Lab #3 - *CC BY-NC-SA 3.0*
    - 11.4: Biometrics Lab #4 - *CC BY-NC-SA 3.0*
    - 11.5: Biometrics Lab #5 - *CC BY-NC-SA 3.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*