# QUANTITATIVE METHODS FOR PLANT BREEDING



*Walter Suza and Kendall Lamkey*
Iowa State University

LibreTexts™

# Quantitative Methods for Plant Breeding

Walter Suza and Kendall Lamkey

Iowa State University

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.




LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 10/02/2026

# TABLE OF CONTENTS


**Licensing**

**About the PBEA Series**

**Contributors**

**1: Chapters**

1.1: Basic Principles

1.2: Distributions and Probability

1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests

1.4: Categorical Data - Binary

1.5: Categorical Data Multivariate

1.6: Continuous Data

1.7: Linear Correlation, Regression and Prediction

1.8: The Analysis of Variance (ANOVA)

1.9: Two Factor ANOVAs

1.10: Mean Comparisons

1.11: Randomized Complete Block Design

1.12: Data Transformation

1.13: Multiple Regression

1.14: Nonlinear Regression

1.15: Multivariate Analysis

1.16: Algebra Review Guide

1.17: Applied Learning Activities

**Index**

**Glossary**

**Detailed Licensing**



This page titled Quantitative Methods for Plant Breeding (Suza and Lamkey) is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey via source content that was edited to the style and standards of the LibreTexts platform.

# Licensing

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

This page titled Licensing is shared under a license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# About the PBEA Series

1

## Background

The Plant Breeding E-Learning in Africa (PBEA) e-modules were originally developed as part of the Bill & Melinda Gates Foundation Contract No. 24576.

Building on Iowa State University's expertise with online plant breeding education, the PBEA e-modules were developed for use in curricula to train African students in the management of crop breeding programs for public, local, and international organizations. Collaborating with faculty at Makerere University in Uganda, University of KwaZulu-Natal in South Africa, and Kwame Nkrumah University of Science and Technology in Ghana, our team created several e-modules that hone essential capabilities with real-world challenges of cultivar development in Africa using Applied Learning Activities. Our collaboration embraces shared goals, sharing knowledge and building consensus. The pedagogical emphasis on application produces a coursework-intensive MSc program for Africa.

**PBEA Project Director:** Walter Suza

**Original Module Coordinators:** Thomas Lübberstedt, William Beavis

**Collaborating Faculty and Experts in Africa:** Richard Akromah, Stephen Amoah, Maxwell Asante, Ben Banful, John Derera, Richard Edema, Paul Gibson, Sadik Kassim, Rufaro Madakadze, Settumba Mukasa, Margaret Nabasirye, Daniel Nyadanu, Thomas Odong, Patrick Ongom, Joseph Sarkodie-Addo, Paul Shanahan, Husein Shimelis, Julia Sibiya, Pangirayi Tongoona, Phinehas Tukamuhabwa.

The authors of this textbook series adapted and built upon the PBEA modules to develop a series of textbooks covering individual topic areas. It is our hope that this project will facilitate wider dissemination and reuse of the PBEA modules' content.

## Explore the Series

- Crop Genetics
- Quantitative Methods for Plant Breeding
- Molecular Plant Breeding
- Quantitative Genetics for Plant Breeding
- Crop Improvement
- Cultivar Development

This page titled About the PBEA Series is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# Contributors

1

## Editors

### Walter Suza

Suza is an Adjunct Associate Professor at Iowa State University. He teaches courses on Genetics and Crop Physiology in the Department of Agronomy. In addition to co-developing courses for the ISU Distance MS in Plant Breeding Program, Suza also served as the director of Plant Breeding e-Learning in Africa Program (PBEA) for 8 years. With PBEA, Suza helped provide access to open educational resources on topics related to the genetic improvement of crops. His research is on the metabolism and physiology of plant sterols. Suza holds a Ph.D. in the plant sciences area (with emphasis in molecular physiology) from the University of Nebraska-Lincoln.

### Kendall Lamkey

Lamkey is the Associate Dean for Facilities and Operations for the College of Agriculture and Life Sciences at Iowa State University. He works in collaboration with the dean, associate deans, department chairs, college-level centers, and other unit leaders to ensure that operations directly advance the mission of the college and that resources are deployed wisely and efficiently. Previously, he served as the chair for the Department of Agronomy at Iowa State University, where, in addition to advocating for research and the PBEA program, he oversaw the Agronomy Department's educational direction, its faculty, and Agronomy Extension and Outreach. Dr. Lamkey is a corn breeder and quantitative geneticist and conducts research on the quantitative genetics of selection response, inbreeding depression, and heterosis. He holds a Ph.D. in plant breeding from Iowa State University and a master's in plant breeding from the University of Illinois. Lamkey is a fellow of the American Society of Agronomy and the Crop Science Society of America and has served as an associate editor, technical editor, and editor for *Crop Science*.

## Chapter Authors

William Beavis, Ursula Frei, M. L. Harbur, Reka Howard, Kendra Meade, Laura Merrick, Ken Moore, Ron Mowers, and Dennis Todey

## Contributors

Anthony A. Mahama, Gretchen Anderson, Todd Hartnell, Andy Rohrback, Tyler Price, Glenn Wiedenhoeft, and Abbey K. Elder

This page titled Contributors is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1: CHAPTERS


1.1: Basic Principles

1.2: Distributions and Probability

1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests

1.4: Categorical Data - Binary

1.5: Categorical Data Multivariate

1.6: Continuous Data

1.7: Linear Correlation, Regression and Prediction

1.8: The Analysis of Variance (ANOVA)

1.9: Two Factor ANOVAs

1.10: Mean Comparisons

1.11: Randomized Complete Block Design

1.12: Data Transformation

1.13: Multiple Regression

1.14: Nonlinear Regression

1.15: Multivariate Analysis

1.16: Algebra Review Guide

1.17: Applied Learning Activities



This page titled 1: Chapters is shared under a license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.1: Basic Principles

Most agricultural knowledge, including seed science and plant breeding, has been learned through the process of experimentation. Recommendations about crop varieties, seeding rates, and other management practices are all based on information acquired from experiments (Figure 1). This lesson introduces you to basic concepts of experimentation.

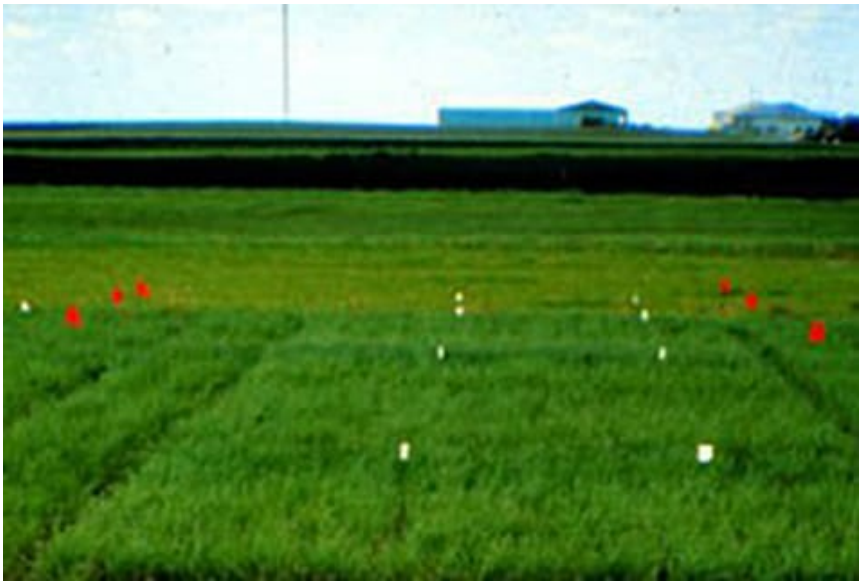


Figure 1 Legume research plots at a university research farm. Photo by Iowa State University.

**Learning Objectives**

- How the scientific method relates to agronomic research
- The different approaches to experimentation used in agronomy
- The roles of replication, randomization, and design control in experimentation
- The different types of data collected in agronomic experiments
- Measures of center and dispersion
- How to organize and summarize data with Excel

## Scientific Method

### Four Basic Steps of the Scientific Method

The scientific method helps lead to discovery of new knowledge. The goal of plant-related agricultural research is to develop new knowledge about crops and how they interact with the environment. The scientific method is a process used by researchers to discover new knowledge about their field of endeavor. It involves the systematic application of four procedures (Figure 2).




Figure 2 The Scientific Method is a cyclical process. New findings lead to new questions. Figure by Abbey Elder.

- **Observation**—recognition of the question
- **Hypothesis**—a tentative explanation of the observed phenomena
- **Experiment**—testing the hypothesis
- **Conclusion**—accept or reject the hypothesis

## Iterative Process

The scientific method is an iterative process. Often the results of one experiment lead to new questions and further experiments. The process can be viewed as iterative rather than linear. What is gained from one experiment can be used to refine knowledge gained from previous experiments. Then further experiments can follow, leading to findings in the same direction or leading to knowledge in a completely new direction of research. This method of research is termed inductive in that particular observations (Figure 3) are used to support a more general conclusion. A known problem exists, but no solution is apparent. The goal of research is to discover answers to some question or problem.



Figure 3 Conducting research at a field plot. Photo by Iowa State University.

# Statistical Science

Before the advent of modern scientific methods, people simply observed phenomena, and without experiments found them difficult to explain. Others who may not have observed exactly the same thing, could ask, “Is what you saw just a chance occurrence, with no true underlying cause?” By designing experiments to test the repeatability and explore causes, we have a better process for drawing conclusions (Figure 4).



Figure 4 A researcher walks through an experimental field plot. Photo by Iowa State University.

The key point is the question, “Did this just occur by chance?” This is where statistics and probability enter into the scientific method. Scientists want to rule out chance happenings, so they often agree that if there is less than 0.05 probability of occurrence by chance, we must be observing a real effect.

## Usage Example

An agronomist observes that all alfalfa varieties appear to grow best when planted on side slopes in pastures. Thinking about what might account for this observation, the agronomist develops the hypothesis that differences in soil characteristics between slopes and other landscape positions are responsible. An experiment is designed to test the hypothesis. Data are collected from several sites; it is concluded that there is indeed a relationship between soil type and alfalfa adaptation. However, the experiment was not conducted in such a way to establish a causal relationship. Therefore, a new hypothesis is developed which states that alfalfa adaptation is a function of soil pH, which differs for the soil types. This leads to a new experiment, and so on.

# Experimental Design

## Observational Experiments

Experimental design allows researchers to control factors influencing outcomes. Statistical analysis is a powerful approach to understanding collections of data. The analysis employed depends on the type of data and the manner in which it was collected. There are two broad categories or approaches to research that commonly are used: observational experiments and designed experiments.



Figure 5 Researchers gathering data. Photo by Iowa State University.

Observational experiments involve collecting data from a population of individuals to which no treatments have been applied (Figure 5). They are descriptive in nature and usually involve studying the relationships among two or more variables of interest. It is important to understand that the variables studied in an observational experiment occur naturally and are not manipulated by the researcher in any way. An example of an observational experiment would be a comparison of groundwater nitrate concentrations among several Iowa counties.

## Designed Experiments

Designed experiments differ from observational experiments in that data are collected from units that have been manipulated by the researcher in some way before the data are collected (Figure 6). This is often described as applying treatments to experimental units. Some good agricultural examples of treatments are the application of specific fertilizer rates and the planting of specific crop varieties for the purposes of comparison. In agronomic terms, the smallest entity to which treatments are applied is usually a field plot.



Figure 6 Crop varieties planted in experimental units. Photo by Iowa State University.

Study Questions 1 – Experiment Design

Query 1.1.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Key Characteristics

- Replication – treatments are repeated two or more times on different experimental units (plots)
- Randomization – treatments are randomly assigned to experimental units (plots)
- Design Control – how treatments are applied to various groupings and sizes of experimental units (subplots, plots, blocks, locations)

### Design Principles

**Replication allows scientists to increase the accuracy of their results.**

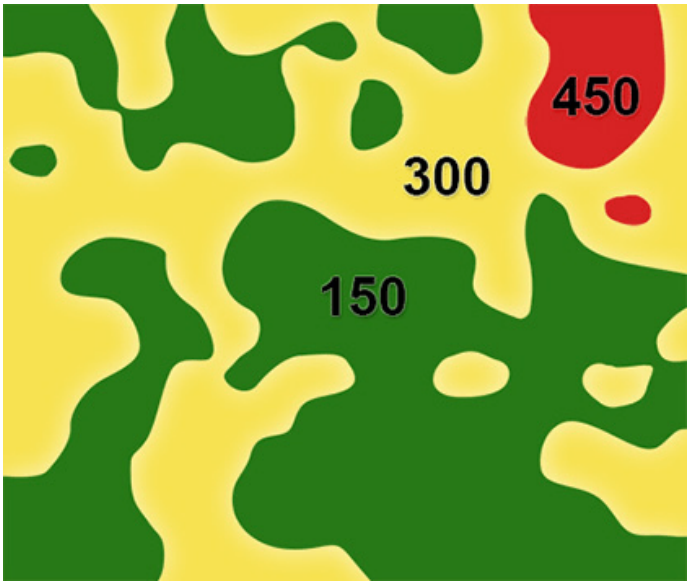



Figure 7 Variability in potassium levels (ppm) in a 480-acre field.

Why is replication important? Let's consider an example. A farmer wants to know if a new variety he has heard about is better than the one he has grown for the past few years. To find out, he might decide to plant a few acres of the new variety and compare its yield with his current one. This is a reasonable approach, but there are some potential problems. Soil properties vary considerably among and within fields. If the varieties are planted in separate fields or even areas within a field, any yield difference observed between the varieties may actually be caused by differences in soil properties (Figure 7).

## Separating Effects

To separate the effects of variety and soil properties, it is necessary to replicate or repeat the comparison over more fields or plots. By comparing the yield averaged over replications (fields), the farmer will get a clearer and truer picture of how the varieties will perform over his/her whole farm. The more times the comparison is replicated, the more likely that any observed difference

in yield is due to the variety planted. This is why replication is an essential part of designing field experiments to compare crop varieties.

Let's use an example to demonstrate. A farmer wants to compare his favorite corn hybrid with one recommended by his father-in-law. He decides to plant two of his smaller fields to compare the hybrids (Figure 8).



Figure 8 Two fields using different corn hybrids. Photo by Iowa State University.

## Study Questions 2

Here are the details on the two fields the farmer wishes to use. For simplicity, we assume each field has one soil type only (Tables 1 and 2).

*Table 1 Field 1 of size 116 hectares*

| **Soil characteristics** | **Amount of organic matter and minerals** |
|---|---|
| Sharpsburg – silty clay loam, 9-14% slope, moderately well-drained, surface layer depth is 8-18cm, subsoil layer depth is 122cm | moderate organic matter |
| | low subsoil P |
| | medium subsoil K |

*Table 2 Field 2 of size 216 hectares*

| **Soil characteristics** | **Amount of organic matter and minerals** |
|---|---|
| Macksburg – silty clay loam, 0-2% slope, somewhat poorly drained, surface layer depth is 61cm subsoil layer depth is 140cm | high organic matter |
| | low subsoil P |
| | medium subsoil K |

What are some possible ways of approaching this problem? Here are three options for performing this experiment, and the results are shown in Tables 3, 4, and 5.

**Option 1**

The farmer decides to plant Field 1 with the new variety and Field 2 with the old variety. When he harvests, he finds the following results (Table 3):

*Table 3 Results of experiment option 1*

| Field 1 | Field 2 |
|---|---|
| New yield variety: 8780 kg/ha | Old yield variety: 9410 kg/ha |

**Option 2**

The farmer decides to plant Field 1 with the old variety and Field 2 with the new variety. When he harvests, he finds the following results (Table 4):

*Table 4 Results of experiment option 2*

| Field 1 | Field 2 |
|---|---|
| Old yield variety: 7530 kg/ha | New yield variety: 10660 kg/ha |

**Option 3**

Realizing that there is a soil fertility difference between the fields, he decides to plant half of each field with each variety. This produces the following results (Table 5):

*Table 5 Results of experiment option 3*

| Field 1- Plot 1 | Field 1 – Plot 2 | Field 2 – Plot 1 | Field 2 – Plot 2 |
|---|---|---|---|
| Old yield variety: 7530 kg/ha | New yield variety: 8780 kg/ha | Old yield variety: 9410 kg/ha | New yield variety: 10660 kg/ha |

Query 1.1.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Increasing Precision

The replication of both corn hybrids on both fields allows us to better separate yield effects due to hybrids and due to fields. We estimate a 1250 kg/ha difference due to hybrids. Although we cannot be sure this difference would repeat in another year or on another pair of fields, the replication has provided a better insight into the variety of differences in these fields.

It is clear that which field the hybrids are planted in will have a large impact on the outcome of the experiment. When using Option 1 or 2, the effects of hybrid and field are confounded. The only reasonable solution is to replicate the experiment.

**Replication also increases precision and allows a measure of repeatability.**

We saw in the previous example that sampling more fields or areas (replication) improved accuracy for testing yields of two hybrids. Replication also allows us to see repeatability in our data, i.e. how consistent the treatment effects are and how much error the experiment has. More “reps” result in more precision in the measurement of hybrid yield differences.

## Randomization

Random assignment of treatments to experimental units is necessary to avoid unintentional bias in the results. Continuing the example from above, the farmer, for practical reasons, might consider planting replications of the same hybrid in adjacent fields. However, this may lead to bias in the results because fields located adjacent to one another tend to be more similar than those located farther apart. If the hybrids to be planted in each field are chosen at random, any bias that may occur due to soil

properties or other characteristics associated with the fields is left to chance.

Randomization ensures that the yields measured in the experiment are due only to the treatment (hybrid) and the random effect associated with the field where it was grown. Randomization gives an equal chance that each treatment (hybrid) will be assigned to each experimental unit (field). This equal-chance assignment provides a probability basis for the statistical tests of hybrid differences.

## Design Control

Design control helps reduce undesirable error variation. Design control refers to the way in which treatments are assigned to experimental units. In the ideal experiment, differences among experimental units treated alike will be small compared to those between units receiving different treatments. In this case, we say that the experimental units are homogeneous, and no design control is necessary. Many times, however, it is not possible to identify the required number of experimental units (plots) that are similar enough to compare all of the treatments in an experiment. In this case, we say that the experimental units are heterogeneous, and often we can improve the precision of the experiment by exercising some design control.

## Blocking

One form of design control is blocking experimental units into homogenous groups called blocks. When each block is large enough to accommodate the complete set of treatments of interest, we refer to the design as a randomized complete block design (RCBD). The RCBD is a very common design used in agricultural field experiments. There are many other types of design control that we will discuss in subsequent lessons.

In our example above, we could greatly improve the precision of the experiment by blocking treatments (hybrids) according to the field. In this case, we would make certain that each hybrid is planted the same number of times in each field. The field effect (1880 kg/ha) would be distributed evenly over each hybrid such that the differences in means between treatments should reflect the true difference between hybrids (1250 kg/ha). Not only does blocking treatments in this manner improve the estimates of the means, but we will see later that it greatly improves (reduces) the error variance used to test the significance of treatment effects.

### Study Questions 3: Experiment Design Elements

#### Query 1.1.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Types of Data

## Quantitative vs. Qualitative

Usually, when we conduct an experiment, we are interested in collecting information on certain characteristics of our experimental units. These characteristics are generally referred to as variables. Variables describe some measurable attribute such as yield or color.

Variables can either be qualitative or quantitative. Qualitative variables are those to which no meaningful numerical values can be assigned. Qualitative variables are often referred to as either classification or categorical variables because they can be used to group data. However, their order has no inherent meaning.

Quantitative variables, on the other hand, are numerical in nature. They can be ranked along some scale of measurement, which

has inherent meaning.

For example, variables in a variety trial might include seed yield and seed color. Yield is a numerical variable and the values collected for this variable would therefore be quantitative. We would be interested in knowing which variety has the highest value for seed yield. Seed color, such as white or yellow, is a description of the seed, but it cannot be ranked numerically. Therefore, color would be a qualitative variable.

Each variate has a value. For example, plant height is a variable and if we measure it five times we have five variates. The variates might have values of 79, 81, 82, 83, and 85.

## Discussion Topic

Write down short responses to the following questions, and then if possible, check your classmates' responses to see how the perception and use of statistics in agriculture professions varies, and its level of incorporation in agricultural businesses.

1. If you have had a job in a business related to agriculture, what was your job title and position, what were your job responsibilities, and did you personally use statistics to do your job?
2. How did statistical analysis (as an area of science) interact with your position, and your job or your company's decisions or policies?
3. How do you foresee statistical analysis and design affecting your professional decisions in the future?

# Measures of Center and Dispersion

## Nominal and Ordinal Scales

There are a number of measurement scales used in agronomic research, including nominal, ordinal, and continuous. A nominal scale, meaning "in name only", is a system of classifying or categorizing qualitative data. Examples of nominal scale classification schemes are sex, plant taxonomy, and soil type. The values are generally characters, such as M or F, rather than numeric data types. Sometimes you categorize data using numbers, even though they are just names, for example, block 1 and block 2, or varieties 1, 2, and 3. It is important to remember that even though the data type is numeric, the order of treatments has no intrinsic meaning.

Data collected on an ordinal scale can be arranged in order according to rank. However, the rank value contains no information about how similar or different two adjacent values are; all that can be said is that one is larger than the other. That is to say, equal differences between any two points on an ordinal scale may not have equal meaning. Agronomists often use arbitrary scales to numerically rank the characteristics of soils and plants. A plant breeder, for example, might devise a scale for ranking disease resistance in progeny plots. The scale is useful for ranking the disease resistance of the genotypes evaluated, but the rankings do not indicate how much actual disease is present.

## Continuous Scales

Continuous scales differ from ordinal scales in that they have a constant interval size. Therefore, the difference between two values has a known quantitative meaning. Examples are temperature, concentrations of phosphorus and potassium in soil solution and plant height and weight. Continuous variables theoretically can take any value in the range afforded by your measuring device, for example, 67.3528...°C (Figure 9).




Figure 9 Continuous variables theoretically can take any value in the range afforded by your measuring device, for example, 67.3528...°C.

The types of analysis which can be performed on a data set depend on the type of measurement scale used. You will learn more about this as we study various statistical procedures.

## Significant Digits

Use common sense when reporting the results of your experiments, and report results only with proper significant digits. Express numerical results only to the level of precision warranted by the measuring instruments. For example, if corn yields are measured in kg per plot, say 8.95 kg, and corrected for grain moisture, e.g. 28.3%, we might get a calculated yield equal to 11.48153 Mg/ha. What should we report? Since the original measurements are three digits, we should report 11.5 Mg/ha. The key principle is to do calculations with as high precision as possible but report only to the level of precision that data warrants.

When reporting results, do not report with more precision than the least precise measurement. For example, you may be trying to measure ear length of maize in an experiment. The ruler used can be read to the nearest millimeter. You might be able to approximate your measurement to the decimal fraction between millimeters. For this individual measurement, you would then be measuring to the 0.1 mm place. When reporting average ear lengths in your company report, how precise can you be in your averages?

### Study Question 5: Precision

#### Query 1.1.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Calculation Statistics

The concept of significant digits when performing calculations using software programs should be somewhat obvious. When calculating statistics in Excel and other statistical programs, the programs will typically output as many digits as you allow. When you have measured ear length to the 0.1-millimeter precision, reporting an average ear length of 225.39568 mm does not make any sense. Using the convention would lead to reporting 225.4 mm.

It is important, however, to keep as many digits as your calculator or computer can accommodate for intermediate calculations. For example, the statistical program SAS uses double precision in its statistical computations and is more accurate than Excel for some of these calculations. It is when you report your results that you need to remember to round off to the proper number of significant digits.

# Standard of Measurement

The standard of measurement in the sciences and most countries in the world besides the United States is the metric (System International; SI) system. In scientific reports, metric measurements are standard. The U.S. system is known as the Imperial system. Examples in this course will be given in the metric system.

The importance of having a measurement standard cannot be overstressed. A recent NASA (National Aeronautics and Space Administration, the federal agency) mission to Mars lost its satellite because of a miscommunication between the builder of the satellite (which used English units) and the user (which operated the satellite using metric measurements). Because of this error, a $125 million satellite was lost.

## Populations and Samples

A population is a set of individuals for which we draw inferences. The purpose of experiments is to draw conclusions about a population. For example, we might want to draw conclusions about all mid-season maturity corn plants on irrigated farms near Grand Island, Nebraska. The population is a theoretical concept. It is generally a very broad group of individuals to which we wish

to extend inferences from our experiment.

The way we draw conclusions about populations is to take **samples**. For example, our experiment may have been conducted on seven farms in the Grand Island area. We earlier saw randomization as a key idea in getting an unbiased sample, and we will explore this idea further in later chapters. From the sample in our experiment, we want to infer the properties of a population represented by that sample.

## Parameters

Parameters characterize a population and are estimated from sample statistics. There are certain descriptive measures for the population that define the **center**. Others describe how **dispersed** the population values are. We sample values from the population to get information, albeit incomplete, about the population and its parameters. We calculate sample statistics to estimate population parameters. For example, we calculate the **sample average** (i.e., **sample mean**) to estimate the true population mean.

Although it may seem at first to be a minor point, we need to distinguish the true population values or parameters from the estimates of them called **statistics**. An easy way to remember the difference in definitions is the mnemonic device: population and parameter begin with the letter **p**, and sample and statistics begin with **s**. The true average corn yield in an area is a population parameter. It is estimated with a sample average, our sample statistic. It is important to remember that the sample average is not the true average yield for the farms in the target area and can vary from sample to sample. We can get into trouble by assuming that the sample average is the true average instead of reporting a range of yields in which the true parameter is likely to be contained.

## Population and Sample Mean

Measures of the center are the mean, median, and mode. The sample mean (sample average) is a measure of the center of a population. It is calculated by averaging the values in the sample. The formula for the sample average is:

$$\bar{x} = \frac{\sum x}{n} \tag{1.1.1}$$

$$\bar{x} = \tag{1.1.2}$$

$$\sum x = x_1 + x_2 + ... + x_n \tag{1.1.3}$$

**where:**
$x_1$= $1^{st}$ variate,
$x_2$= $2^{nd}$ variate,
$x_n$= $n^{th}$ variate,
$n$= sample size (# of individuals measured or variates).

We use Greek letters for parameters and Latin letters for statistics. The sample mean $\bar{x}$ is used to estimate the **population mean μ**. For example, suppose our seven sample values are: 178, 170, 203, 185, 199, 178, 210 kilograms per hectare of a certain crop variety, corrected to 15.5% grain moisture. The sample mean is 189 kg/ha.

## Median and Mode

Other measures of center include the value which occurs most often in the sample, called the **mode**, and the **median**, the (n+1)/2 ranked number when the observations are sorted by value. The median can be thought of as the middle number in the series. 50% of the measurements are above the median; the other 50% are below. For our example, the sample mode is 178. The sample median is 185, seen by ordering the values: 170, 178, 178, 185, 199, 203, 210. If the sample contains an even number of observations, the median is the average of the two middle values.

Study Question 6: Measures of Center

Query 1.1.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Variance

Measures of dispersion include the difference between the highest and lowest values in the sample (the range), the variance, whose formula is listed below, and its square root, called the standard deviation. The variance is an average of squared deviations from the mean. Its formula is:

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

**where:**
$S^2$= sample variance,
$x$= each variate in the sample,
$\bar{x}$= sample mean,
$n$= sample size (# of individuals measured or variates).

This formula has two parts, the numerator, which is a sum of squared deviations, and the denominator (n-1), called degrees of freedom. The Greek letter, Sigma (Σ), in the formula means "sum the terms which follow". Any value not exactly coincident with the mean contributes positively to the variance through this sum of squares. In the earlier crop example, even the yield of 185, which is close to the sample mean of 189, has a small contribution to the variance $(185\text{-}189)^2 = (-4)^2 = 16$. The more distant from the mean, the higher will be the contribution of each value to this numerator sum of squares. The value 210 contributes $(210 - 189)^2 = 21^2 = 441$ to the sum of squares. For our example, the sample variance is 226 and the sample standard deviation is 15 kg/ha.

The reason for dividing by ($n$ – 1) is to make the sample variance an unbiased estimator for the population variance. This concept is called the degrees of freedom. This is an important concept, but equally difficult to explain! Basically, the deviations – the value of $(x - \bar{x})$ for each individual in the sample must sum to zero, by definition of the sample mean (try it!). Because of this, only $n$ – *1* of the individual values are free to vary. For example, if the $n$ – 1 deviations sum to 9, then we know the nth value must equal $(\bar{x} - 9)$. You can also think of the denominator as telling you that it is impossible to get a sample variance if your sample only has one value (division by zero is impossible).

## Standard Deviation

The standard deviation is simply the square root of the variance. It is recorded in the same units as the original measurements. It is very often used as the measure of spread or dispersion of a sample. In the chapter on Distributions and Probability, as we explore the normal, bell-shaped distribution, we will see that the standard deviation is very useful. (For a normally distributed random variable, 95% of the values of the population are within about two standard deviations of the mean.) We use the sample variance $s^2$ to estimate the population variance $\sigma^2$. The sample standard deviation s estimates the population standard deviation σ.

## Coefficient of Variation

Another measure of variation, which is independent of the units of measurement, is the ratio of standard deviation to sample

mean, called the coefficient of variation (CV). The formula is:

$$CV = \frac{S}{\bar{x}} 100\% \tag{1.1.5}$$

**where:**
$CV$= sample coefficient of variation,
$S$= sample standard deviation,
$\bar{x}$= sample mean.

For our example with crop yields, the *CV* is 15.03/189 = 7.95%. The *CV* is most often used when computed with the "error" standard deviation divided by the experiment mean. In this sense, it measures the random and unexplained variation in an experiment. You will see in future units how the experiment variation is partitioned into "explained" and error variance.

The *CV*, because of its independence from units, can be used to compare variation for different traits or even for different crops. We might contrast the variation in a soybean variety yield trial with a corn one, citing that the *CV* is 12% for the soybean trial compared with 8% for corn yields. However, care is needed in interpreting any of these statistics. Since the *CV* is the ratio of standard deviation to mean, traits with a low average, such as near-zero stalk lodging, may have extremely large *CV*s, often hundreds or even over a thousand percent.

### Study Question 7 – Measures of Center and Dispersion

#### Query 1.1.6

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Excel Exercises

## Introduction

Microsoft Excel is a powerful tool for organizing, analyzing, and displaying data. In this lesson, we will be using data from the Iowa Crop Performance Test program in central United States to demonstrate how to organize data in Excel. In this exercise, you will learn how to enter data and equations, sort data, filter data, create a pivot table, and make a graph in Excel.

The Excel Exercises in this course were designed under version 2010. If you are using a newer version of Excel, the way the program functions or your view of the results may differ from what is depicted in the lessons.

## Excel Exercises

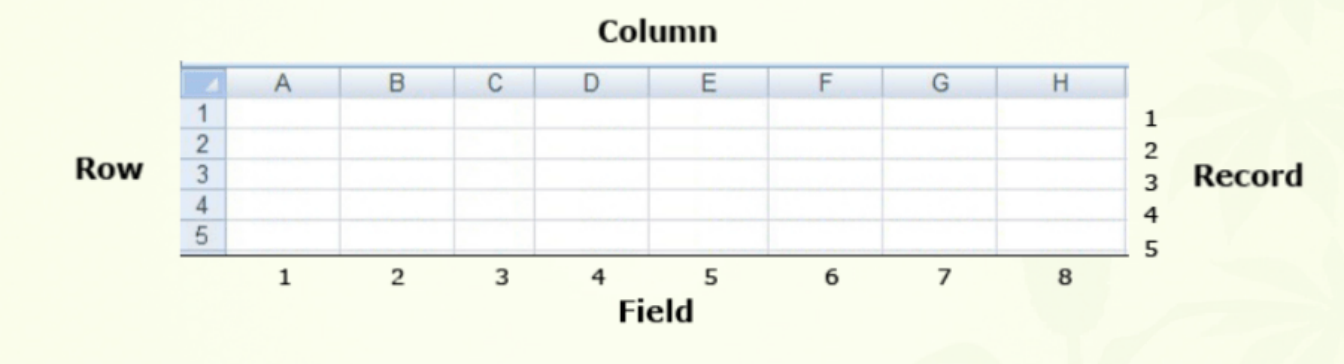
The data used to generate the examples below were collected during the 1995 and 1996 growing seasons as part of the Iowa Crop Performance Test program. Two representative hybrids grown near Oskaloosa, Iowa, were selected for the purpose of these exercises.

Each hybrid was planted at a seeding rate of 11,736 kernels per hectare. Plots consisted of four 5.49 m rows spaced 76.2 cm apart. Yield data were collected from only the center two rows. Plot yields are in units of kilograms per plot.

For Exercise 1, download the Excel file here.

## Exercise 1: Organizing Data in Excel

Data are organized in spreadsheets in rows and columns. In Excel, rows are designated by numerals and columns by capital letters. Each intersection of a row and column is called a cell and is labeled using the column letter and row number (e.g. B3).




The data in the Hybrid Test Data worksheet are arranged using this format. The first row contains headings that identify each of the columns (fields). Each subsequent row (record) contains all of the data pertaining to a specific field plot (Figure 10). At this point, each record consists of:

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| Year | Hybrid | Rep | Harvest Wt. (kg/plot) | Moisture % | DM Yield (kg/ha) | Adjusted Yield (kg/ha) |

Figure 10 Headings on an MS Excel document.

**Ex. 1: Data Columns**

Each plot is identified by the year and field replication in which it was grown together with the treatment (in this case, hybrid) that was applied to it. In statistical jargon, year, replication, and hybrid are independent class variables. The remaining two columns contain measurements that were made on the plots. These represent dependent variables and the goal of our analysis is to determine how they are affected by the independent variables.

The yield data in column D of the **Hybrid Test Data** worksheet is in units of kgs per plot, which is not very meaningful to most people. Columns F and G are labeled DM Yield (kg/ha), and Adjusted Yield (Kg/ha). However, the cells below these headings do not contain data.

Using the data in columns D and E together with the information described in the introduction section, calculate the missing values in columns F, and G.

HINTS:

- A hectare is 10,000 square meters.
- Each plot is two rows wide by one row long.
- Adjusted yields are calculated on a 15% moisture basis.

**Ex. 1: Column F Formula**

In the data worksheet, cells with a red triangle in the corner contain helpful comments (Figure 11). Place the cursor over the cell to read the comment. Note that cell F2 occurs at the intersection of Row 2 and Column F on the spreadsheet.

- Type the following formula into the cell F2: **=D2*1195*((100-E2)/100)**
- After entering the formula in cell F2, press “Enter”. Next, select the cell (F2) by clicking on it to highlight it. Once highlighted, grab the square in the lower right corner and drag it down to cell F17 to copy the formula into the remaining empty cells in the F column.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Hybrid | Rep | Harvest Wt. (kg/plot) | Moisture % | DM Yield (kg/ha) | Adjusted Yield (kg/ha) |
| 2 | 1995 | DK580 | 1 | 6.31 | 13.3 | | |
| 3 | 1995 | DK580 | 2 | 6.58 | 14.1 | | |
| 4 | 1995 | DK580 | 3 | 6.30 | 13.8 | | |
| 5 | 1995 | DK580 | 4 | 6.32 | 16.4 | | |
| 6 | 1996 | DK580 | 1 | 8.41 | 19.1 | | |
| 7 | 1996 | DK580 | 2 | 8.39 | 19.9 | | |
| 8 | 1996 | DK580 | 3 | 8.35 | 20.5 | | |
| 9 | 1996 | DK580 | 4 | 8.83 | 20.2 | | |
| 10 | 1995 | DK604 | 1 | 6.92 | 19.8 | | |
| 11 | 1995 | DK604 | 2 | 5.97 | 21.4 | | |
| 12 | 1995 | DK604 | 3 | 7.22 | 20.3 | | |
| 13 | 1995 | DK604 | 4 | 5.85 | 21.6 | | |
| 14 | 1996 | DK604 | 1 | 8.53 | 20.3 | | |
| 15 | 1996 | DK604 | 2 | 8.63 | 19.7 | | |
| 16 | 1996 | DK604 | 3 | 8.55 | 20.6 | | |
| 17 | 1996 | DK604 | 4 | 8.70 | 19.6 | | |

Figure 11 Harvest weight and moisture percent data of 17 rows of hybrid maize.

Here is the reasoning behind the formula in the F2 cell:

1. First, we convert from kilograms per plot to kilograms per hectare. So we multiply by 1195 to convert from a two-row, 5.49 m plot (for 76.2 cm rows, this would be 8.37 $m^2$) to a hectare. (Check the math: 10,000 $m^2$ (one hectare) divided by 8.37 $m^2$ = 1,195.)
2. Second, we multiply by (100 – E2)/100. This calculates the percentage dry weight, which is (100 – grain moisture)/100. Multiplying the kilograms per hectare by the percentage dry weight gives us the dry matter yield per hectare.

**Ex. 1: Column G Formula**

- Type the following formula into the cell G2: **=D2*1195*0.85**
- After entering the formula in cell G2, drag the cursor over it to select the cell. Once highlighted, grab the square in the lower right corner and drag it down to cell G17 to copy the formula into the remaining empty cells in the G column.

Here is the reasoning behind the formula in the G2 cell:

You are adjusting the yield to 15% (0.15) moisture by dividing by 0.85 (which equals 1-0.15). Doing that increases the yield by 15% to account for the standard moisture in reported grain yields

**Ex. 1: Completed Data Table**

- Excel formulas always begin with the equal sign.
- Spreadsheet formulas can be written with relative cell references so that once a formula is entered, it can be copied to other cells.

Your completed data table should look like Table 12.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Hybrid | Rep | Harvest Wt. (kg/plot) | Moisture % | DM Yield (kg/ha) | Adjusted Yield (kg/ha) |
| 2 | 1995 | DK580 | 1 | 6.31 | 13.3 | 6537.57 | 6,409.38 |
| 3 | 1995 | DK580 | 2 | 6.58 | 14.1 | 6754.403 | 6,683.64 |
| 4 | 1995 | DK580 | 3 | 6.30 | 13.8 | 6489.567 | 6,399.23 |
| 5 | 1995 | DK580 | 4 | 6.32 | 16.4 | 6313.806 | 6,419.54 |
| 6 | 1996 | DK580 | 1 | 8.41 | 19.1 | 8130.41 | 8,542.46 |
| 7 | 1996 | DK580 | 2 | 8.39 | 19.9 | 8030.866 | 8,522.14 |
| 8 | 1996 | DK580 | 3 | 8.35 | 20.5 | 7932.709 | 8,481.51 |
| 9 | 1996 | DK580 | 4 | 8.83 | 20.2 | 8420.376 | 8,969.07 |
| 10 | 1995 | DK604 | 1 | 6.92 | 19.8 | 6632.059 | 7,028.99 |
| 11 | 1995 | DK604 | 2 | 5.97 | 21.4 | 5607.442 | 6,064.03 |
| 12 | 1995 | DK604 | 3 | 7.22 | 20.3 | 6876.436 | 7,333.72 |
| 13 | 1995 | DK604 | 4 | 5.85 | 21.6 | 5480.748 | 5,942.14 |
| 14 | 1996 | DK604 | 1 | 8.53 | 20.3 | 8124.1 | 8,664.35 |
| 15 | 1996 | DK604 | 2 | 8.63 | 19.7 | 8281.219 | 8,765.92 |
| 16 | 1996 | DK604 | 3 | 8.55 | 20.6 | 8112.497 | 8,684.66 |
| 17 | 1996 | DK604 | 4 | 8.70 | 19.6 | 8358.786 | 8,837.03 |

Figure 12 Completed data with DM and adjusted yield.

## Exercise 2: Sorting Data in Excel

Sorting data in Excel is easy. You can quickly sort a data set using several fields at once by following the directions below.

**Steps:**

1. Using the mouse, select the data you wish to sort.
2. Make sure not to include the header row (Field Names) in your selection.
3. Select **Sort** from the **Data** menu at the very top of the screen.
4. Enter the field you wish to **sort by** in the space provided.
5. Select **Values** in the **Sort On** field.
6. Select whether you want ascending or descending sort order in the **Order** field.
7. To add another field to **Sort by**, select **Add Level** at the top of the dialog box.
8. Repeat the steps above to select the **Field**, **Values**, and **Order**.
9. Click **OK** to sort the data.

Sort the data in Hybrid Test Data spreadsheet by Hybrid, Year, and Replication.

Your completed data table should look like Table 13.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Hybrid | Rep | Harvest Wt. (kg/plot) | Moisture % | DM Yield (kg/ha) | Adjusted Yield (kg/ha) |
| 2 | 1995 | DK580 | 1 | 6.31 | 13.3 | 6537.57 | 6,409.38 |
| 3 | 1995 | DK580 | 2 | 6.58 | 14.1 | 6754.403 | 6,683.64 |
| 4 | 1995 | DK580 | 3 | 6.30 | 13.8 | 6489.567 | 6,399.23 |
| 5 | 1995 | DK580 | 4 | 6.32 | 16.4 | 6313.806 | 6,419.54 |
| 6 | 1996 | DK580 | 1 | 8.41 | 19.1 | 8130.41 | 8,542.46 |
| 7 | 1996 | DK580 | 2 | 8.39 | 19.9 | 8030.866 | 8,522.14 |
| 8 | 1996 | DK580 | 3 | 8.35 | 20.5 | 7932.709 | 8,481.51 |
| 9 | 1996 | DK580 | 4 | 8.83 | 20.2 | 8420.376 | 8,969.07 |
| 10 | 1995 | DK604 | 1 | 6.92 | 19.8 | 6632.059 | 7,028.99 |
| 11 | 1995 | DK604 | 2 | 5.97 | 21.4 | 5607.442 | 6,064.03 |
| 12 | 1995 | DK604 | 3 | 7.22 | 20.3 | 6876.436 | 7,333.72 |
| 13 | 1995 | DK604 | 4 | 5.85 | 21.6 | 5480.748 | 5,942.14 |
| 14 | 1996 | DK604 | 1 | 8.53 | 20.3 | 8124.1 | 8,664.35 |
| 15 | 1996 | DK604 | 2 | 8.63 | 19.7 | 8281.219 | 8,765.92 |
| 16 | 1996 | DK604 | 3 | 8.55 | 20.6 | 8112.497 | 8,684.66 |
| 17 | 1996 | DK604 | 4 | 8.70 | 19.6 | 8358.786 | 8,837.03 |

Figure 13 Sorted data set

## Ex. 3: Filtering Data in Excel

Often when you are working with a large data set you want to isolate a subset of the data for analysis. It is possible to do this simply by sorting the data so that the information you want is grouped together. In the sort example in Exercise 2 we regrouped the data by year making it easier to compare the two hybrids within each year. With larger data tables that use more than three sort fields, it is easier group data using a filter.

**Steps:**

1. Using the mouse, select the data you wish to filter. Be sure to select the top row which contains the data labels.
2. Select **Filter** from the **Data** menu.
3. A small box with an arrow in it will appear in the lower right corner of each cell in the header row.
4. Click on any of these boxes to filter data by that field.
5. A pull-down menu will appear containing the available filters for that field.
6. Select the one you want, and only data matching that criterion will be displayed.
7. You may use any combination of filters to select a specific subset of the data.

Filter the data in Hybrid Test Data spreadsheet to display data only for 1995 and DK604.

Your completed table should look like Figure 14.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Hybrid | Rep | Harvest Wt. (kg/plc | Moisture % | DM Yield (kg/ha | Adjusted Yield (kg/ha |
| 6 | 1995 | DK604 | 1 | 6.92 | 19.8 | 6632.059 | 7,028.99 |
| 7 | 1995 | DK604 | 2 | 5.97 | 21.4 | 5607.442 | 6,064.03 |
| 8 | 1995 | DK604 | 3 | 7.22 | 20.3 | 6876.436 | 7,333.72 |
| 9 | 1995 | DK604 | 4 | 5.85 | 21.6 | 5480.748 | 5,942.14 |

Figure 14 Completed, sorted data table.

## Ex. 4: Evaluating Data With a Pivot Table

A pivot table is used to summarize data contained in other tables. It is an extremely powerful tool for evaluating data. In our case we will use a pivot table to compare means of the two hybrids for each of the two years. Follow the steps outlined below to create a summary table of adjusted yields for the **Hybrid Test Data**.

**Steps:**

1. Using the mouse, select the data you want to summarize. Be sure to select the top row which contains the data labels.
2. Select **PivotTable** from the **Insert** menu.
3. A dialog box will open and the data you have selected will automatically appear in the box next to **Select a table or range**.
4. Click the circle next to **New Worksheet** and then **OK**.
5. The next screen will show an empty table with a panel on the right side titled **PivotTableFieldList**, which is used to format the table:
6. Drag the **Year** field into the **Row Labels** box in the panel.
7. Drag the **Hybrid** button into the **Column Labels** box in the panel.
8. Drag the **Adjusted Yield** button into the **Values** box in the panel.
9. Click on the **Sum of Adjusted Yield** field and select **Value Field Settings**... from the popup menu that appears.
10. Select Average from the list of options that appear, then click OK.
11. The 2 x 2 table you have created will be displayed in the new worksheet.

Your completed pivot table should look like Table 15.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| 3 | Average of Adjusted Yield (kg/ha) | Hybrid | | |
| 4 | Year | DK580 | DK604 | Grand Total |
| 5 | 1995 | 6477.945625 | 6592.2175 | 6535.081563 |
| 6 | 1996 | 8628.79625 | 8737.989375 | 8683.392813 |
| 7 | Grand Total | 7553.370938 | 7665.103438 | 7609.237188 |

Figure 15 Average of Adjusted Yield pivot table.

## Ex. 5: Graphing Data in Excel

Now that the data is summarized in a pivot table it is a simple process to graph the results. Since Hybrids are a qualitative variable, we will use a bar graph. Follow the steps outlined below to graph the adjusted yield means for the **Hybrid Test Data**.

**Steps:**

1. Using the mouse, select any cell within the pivot table.
2. Select **PivotChart** from the **Tools** toolbar.
3. Select **Column** from the **Insert Chart** menu on the left of the dialog box.
4. Select the side-by-side chart type (first item under **Column**)
5. A graph will appear within the worksheet.
6. You can rearrange the chart by dragging the field labels between the **Legend Fields**... and **Axis Fields (Categories)** boxes in the panel to the right of the screen.
7. To change the location of the graph:
    1. Right-click anywhere within the margins of the graph.
    2. Select Move Chart...
    3. Click the circle next to New sheet: and enter a label for the new worksheet.
    4. Click OK to finish.
8. To change the appearance of the graph:
    1. Right-click anywhere within the margins of the graph.

2. Select Format Chart Area... from the popup menu.
3. Change settings under each tab to alter the appearance of the graph.

Your completed graph should look like Figure 16.

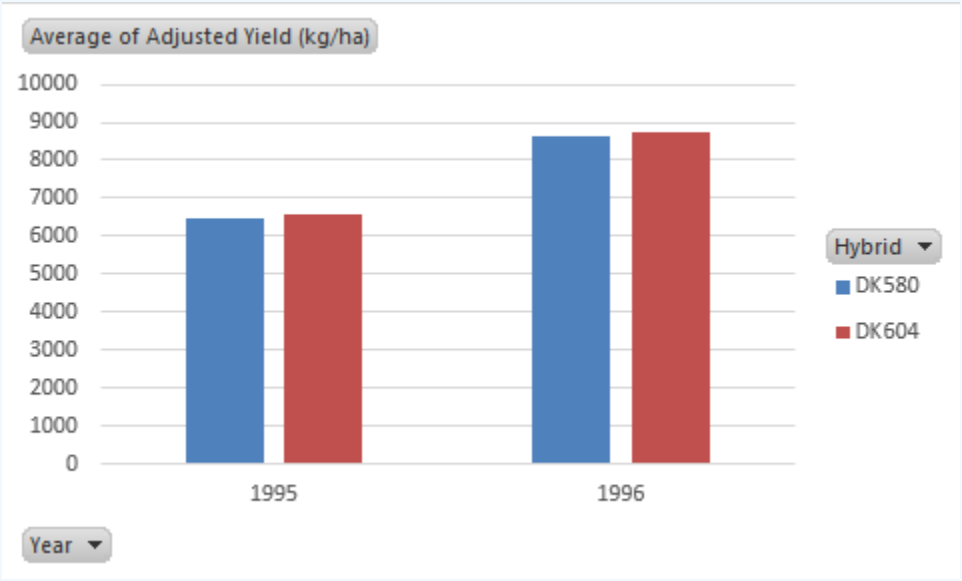



Figure 16 Bar graph of the complete data set.

## Ex. 6: Calculating Measures of Dispersion with a Pivot Table

We have already created a pivot table to summarize data from the Hybrid Test dataset. In Exercise 4, we used a pivot table to calculate means (or averages) of the two hybrids for each of the two years. In this exercise, we will modify that table to include a common measure of dispersion, the standard deviation (SD).

You should already have created a pivot table that displays means of the hybrids for each year averaged over reps. Follow the steps below to calculate the standard deviation for each hybrid by year treatment combination.

**Steps:**

1. Using the mouse, select any cell within the pivot table. This should pull up the **PivotTable Field List** on the right of your window.
2. Drag the **Adjusted Yield** button into the Values box in the panel. This will create another cell in the table that is labeled **Sum of Adjusted Yield.**
3. Click on the **Sum of Adjusted Yield** field and select **Value Field Settings**... from the popup menu that appears.
4. Select **StdDev** from the list of options that appear, then click **OK.** The resulting 2 x 2 table should now show the mean of each combination of hybrid and yield along with its standard deviation.

Your completed pivot table should look like this Figure 17):

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | Hybrid | | |
| 4 | Year | Data | DK580 | DK604 | Grand Total |
| 5 | 1995 | Average of Adjusted Yield (kg/ha) | 6477.945625 | 6592.2175 | 6535.081563 |
| 6 | | StdDev of Adjusted Yield (kg/ha) | 137.3768235 | 693.3439362 | 466.7380329 |
| 7 | 1996 | Average of Adjusted Yield (kg/ha) | 8628.79625 | 8737.989375 | 8683.392813 |
| 8 | | StdDev of Adjusted Yield (kg/ha) | 228.2614228 | 79.27840362 | 168.6125332 |
| 9 | Total Average of Adjusted Yield (kg/ha) | | 7553.370938 | 7665.103438 | 7609.237188 |
| 10 | Total StdDev of Adjusted Yield (kg/ha) | | 1162.831755 | 1234.602684 | 1160.025487 |

Figure 17 Pivot table of averages.

## Ex. 7: Using Excel Functions to Calculate Statistics

Excel has a number of embedded functions that can be used to calculate common statistics. In this example, we will use Excel functions to calculate the mean, median, variance and standard deviation of the adjusted yield values in the **Hybrid Test Data** worksheet.

In Excel, functions are entered as formulas in the cells where you want to display the results. Follow the steps below to calculate the mean, median, variance and standard deviation of the adjusted yield values in the **Hybrid Test Data** worksheet.

**Steps:**

1. Starting with the **Hybrid Test Data** worksheet where you calculated **Adjusted Yield** values in Part 1, enter the label "**Mean**" in cell **F19**.
2. Enter the formula **=AVERAGE($G$2:$G$17)** in cell **G19**. This calculates and displays the overall mean in cell **h29**. The dollar signs (**$**) in the formula indicate an absolute range relative to rows and columns, which will allow you to copy and paste the formula in other cells without losing reference to the correct range of data.
3. Adding additional calculations is now as easy as entering a new label, copying the formula you just entered into a new cell, and editing it to call a different function.
4. Using the approach just described, calculate the **median**, **variance,** and **standard deviation** (SD) for the adjusted mean values. The Excel functions for these three statistics are: **=Median(range)**, **=Var(range)**, and **=Stdev(range).**

The cells you entered into the Hybrid Test Data worksheet should look like this (see box in lower right corner) (Figure 18):

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Year | Hybrid | Rep | Harvest Wt. (kg/plot) | Moisture % | DM Yield (kg/ha) | Adjusted Yield (kg/ha) |
| 2 | 1995 | DK580 | 1 | 6.31 | 13.3 | 6537.57 | 6,409.38 |
| 3 | 1995 | DK580 | 2 | 6.58 | 14.1 | 6754.403 | 6,683.64 |
| 4 | 1995 | DK580 | 3 | 6.30 | 13.8 | 6489.567 | 6,399.23 |
| 5 | 1995 | DK580 | 4 | 6.32 | 16.4 | 6313.806 | 6,419.54 |
| 6 | 1996 | DK580 | 1 | 8.41 | 19.1 | 8130.41 | 8,542.46 |
| 7 | 1996 | DK580 | 2 | 8.39 | 19.9 | 8030.866 | 8,522.14 |
| 8 | 1996 | DK580 | 3 | 8.35 | 20.5 | 7932.709 | 8,481.51 |
| 9 | 1996 | DK580 | 4 | 8.83 | 20.2 | 8420.376 | 8,969.07 |
| 10 | 1995 | DK604 | 1 | 6.92 | 19.8 | 6632.059 | 7,028.99 |
| 11 | 1995 | DK604 | 2 | 5.97 | 21.4 | 5607.442 | 6,064.03 |
| 12 | 1995 | DK604 | 3 | 7.22 | 20.3 | 6876.436 | 7,333.72 |
| 13 | 1995 | DK604 | 4 | 5.85 | 21.6 | 5480.748 | 5,942.14 |
| 14 | 1996 | DK604 | 1 | 8.53 | 20.3 | 8124.1 | 8,664.35 |
| 15 | 1996 | DK604 | 2 | 8.63 | 19.7 | 8281.219 | 8,765.92 |
| 16 | 1996 | DK604 | 3 | 8.55 | 20.6 | 8112.497 | 8,684.66 |
| 17 | 1996 | DK604 | 4 | 8.70 | 19.6 | 8358.786 | 8,837.03 |
| 18 | | | | | | | |
| 19 | | | | | | | |
| 20 | | | | | | | |
| 21 | | | | | | Mean | 7,609.24 |
| 22 | | | | | | Median | 7,907.61 |
| 23 | | | | | | Variance | 1,345,659.13 |
| 24 | | | | | | SD | 1,160.03 |

Figure 18 Summary statistics (mean, median, variance, and standard deviation).

## Ex. 8: Calculate Descriptive Statistics

Excel comes with an add-in called Analysis Toolpak which contains a number of macros for calculating various statistics. The tools represented by the macros in Analysis Toolpak will have some limitations as your analyses become more complex, but they are useful for simple cases and for understanding how to interpret basic statistics. Your installation of Excel may not have the Analysis Toolpak installed and ready to use. If that is the case, follow these instructions to activate the Add-In (Figure 19).

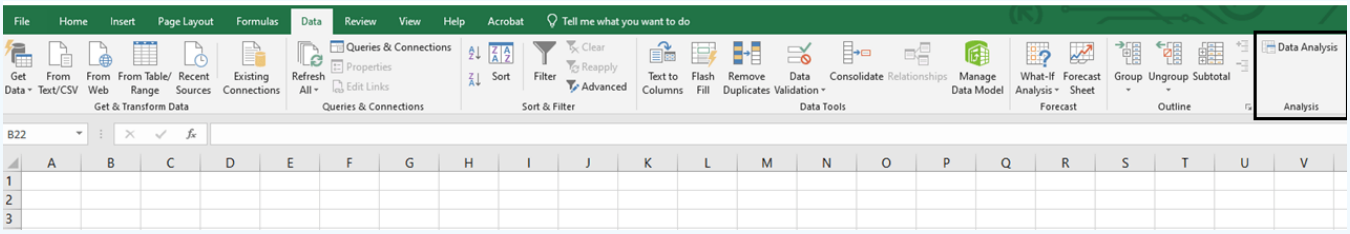



Figure 19 How to Access data analysis tools in Excel.

**Steps:**

1. Under the “Data” menu, select “Data Analysis” from the “Analysis” group.
2. Select Descriptive Statistics in the Data Analysis dialog box that appears and then click OK.
3. A new dialog box will appear labeled Descriptive Statistics. Click on the spreadsheet icon at the far right of the line labeled Input Range:
4. Use your mouse to select the range of data listed under Adjusted Yield (h3:h27).
5. Under Output Options select New Worksheet Ply: and Summary Statistics and then click OK. A table of descriptive statistics will be created and displayed in a new worksheet.

The table of descriptive statistics should look like Figure 20.

| | A | B |
|---|---|---|
| 1 | *Column1* | |
| 2 | | |
| 3 | Mean | 7609.237188 |
| 4 | Standard Error | 290.0063717 |
| 5 | Median | 7907.61375 |
| 6 | Mode | #N/A |
| 7 | Standard Deviation | 1160.025487 |
| 8 | Sample Variance | 1345659.13 |
| 9 | Kurtosis | -1.915710834 |
| 10 | Skewness | -0.192612817 |
| 11 | Range | 3026.935 |
| 12 | Minimum | 5942.1375 |
| 13 | Maximum | 8969.0725 |
| 14 | Sum | 121747.795 |
| 15 | Count | 16 |

Table 20 Descriptive statistics generated by Excel.

You should be familiar with all of these statistics except the Standard Error, Kurtosis, and Skewness. The standard error is another measure of dispersion. It is the square root of the variance divided by the number of observations (count). It will become more important later in the course when we learn about mean comparisons. The Kurtosis and Skewness statistics relate to the distribution of the data and are useful for evaluating whether or not a set of observations are distributed normally. We will learn more about them in future lessons as well.

## Summary

- Iterative process of discovery
- Observation, Hypothesis, Experiment, Conclusion
- Statistics answers “Did this occur simply by chance?”

## Replication

- Increases accuracy by better sampling
- Increases precision of treatment averages
- Gives a measure of repeatability

## Randomization

- Provides insurance against bias
- Gives statistical basis for hypothesis tests

## Design Control

- Reduces error from confounding factors (eg. blocking to remove soil variation)

## Measurement Scales

- Nominal (in name only) / Ordinal (can be ordered) / Continuous
- Report only significant digits

## Parameters

- Characterize the population

## Statistics

- Calculated from the sample to estimate parameters

## Measures of Center

- Mean, median, and mode

## Measure of Dispersion

- Standard deviation, variance, range, and coefficient of variation (CV)

**How to cite this chapter:** Moore, K., M.L. Harbur, R. Mowers, L. Merrick, and A. A. Mahama. 2023. Chapter 1. Basic Principles. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.1: Basic Principles is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ken Moore; M. L. Harbur; Ron Mowers; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.2: Distributions and Probability

In Chapter 1, we learned about the scientific method, principles in the design of experiments, and how descriptive statistics summarize data, especially the center and spread of a distribution. In this chapter, we extend these topics, looking further into how to draw samples (using appropriate instruments, e.g., Figure 1) from populations and some possible distributions of random variables.

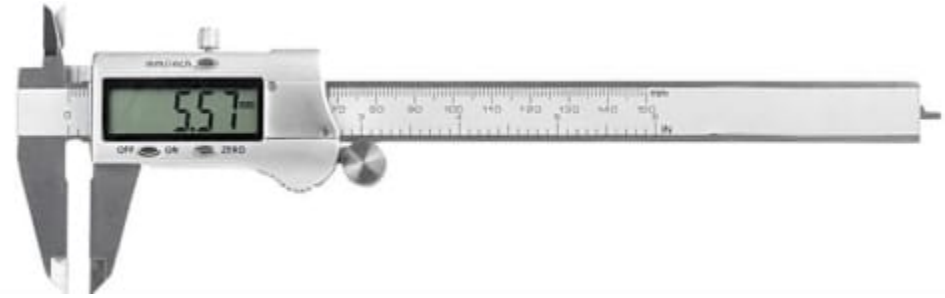



Figure 1 A digital caliper. Photo: Lior Streng. Licensed under GFDL via Wikimedia Commons.

An old saying warns that it is "sometimes difficult to see the forest for the trees." The same situation can arise with experimental data. Even with a well-designed experiment, special skill is needed to understand the mountains of data which are produced (experts, Figure 2). We must pay attention to the variability of the data and the distribution of values. By plotting data and looking for overall patterns in the data, we can improve our understanding of the results.

## Learning Objectives

- How sampling can affect your view of a population
- How to use Excel to produce a frequency histogram
- How to interpret frequency histograms and percentiles
- How to recognize the normal distribution
- How to determine the z-value for a member of the population



Figure 2 Soil sampling near Iowa City, IA. Photo by U.S. National Resources Conservation Service.

# Samples & Populations

## Sampling Conducted in a Designed Experiment

The way a population is sampled, or an experiment is designed and conducted affects the conclusions that we can draw.



Figure 3 Careful measurement and recording is key to achieving the objectives of any experiment.

One of the first considerations for any experiment is the objective of the experiment. In Chapter 1 on Basic Principles, we saw how the scientific method is an iterative process, and that hypotheses are formed and tested as part of the overall scientific goal. Once the scientist decides on an objective, he or she carefully considers what design to use for the experiment. We saw earlier that the principles of randomization, replication, and controlling extraneous variables are important in experimental design. Not only does the experiment need to be well-designed, but scientists must also take measurements very carefully (with appropriate instruments, e.g., Figure 3), and all aspects of the conduct of the experiment must be done as meticulously as possible, to ensure that the experiment can achieve the objectives.

To draw conclusions from the experiment, we need to understand the nature of the data and decide on proper statistical analysis. Data can be continuous, count data, or even form classes (categorical). It is from the nature of the data that we can make probability statements.

## Sampling to Characterize a Population

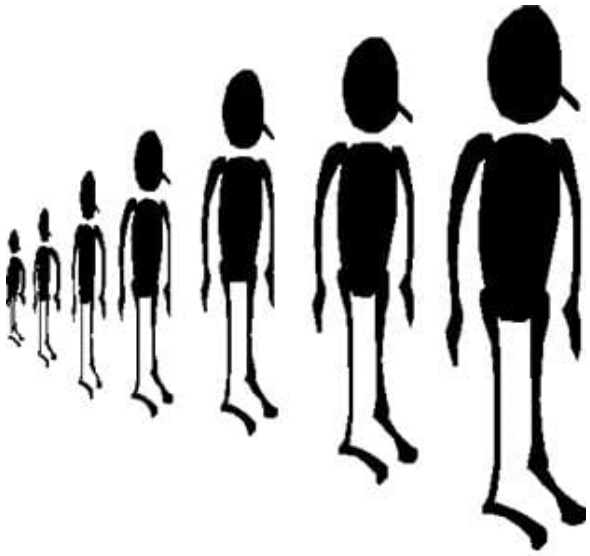


Figure 4 A population is any group of individuals about which data are sought.

An experiment is usually conducted in order to gain information about a population (Figure 4). What is a population? It can be many things: all farmers in the United States, the corn produced in Marshall County, Iowa, or the plants in a 20-acre (8-hectare) soybean field. It is important to understand the scope of your population because it defines the population to which your results apply. If, for instance, the samples you collected are from a single 20-acre (8 ha) soybean field, then the only inferences you can make apply to that single field. If, however, you sampled several fields within a watershed, you could extend your inferences to the entire area.

We seek to describe parameters, or characteristics of that population. Ideally, we would like to evaluate every unit within that population, but this is often impractical — i.e., it would be expensive and time-consuming to contact every farmer in the country, and the testing of all of the soil in a field would require the permanent removal of hundreds of tons of topsoil.

We often resort to taking a sample, or set of measurements, from the population. This sample is more likely to accurately represent the population if we randomize our samples — that is, if we take soil samples randomly from different sections of the field, rather than from one section. For example, would a measurement of soil pH represent a whole field if all samples were taken near the edge of a field, next to a limestone road?

## Randomization

Therefore, we often resort to taking a sample, or set of measurements, from the population. This sample is more likely to accurately represent the population if we randomize our samples — that is, if we take soil samples randomly from different sections of the field, rather than from one section. For example, would a measurement of soil pH represent a whole field if all samples were taken near the edge of a field, next to a limestone road?



Figure 5 Soil sampling is subject to error brought on by the locations of samples.

Even careful selection of sampling sites and using proper and careful procedures in taking the sample can produce values which do not represent the whole population (Figure 5). In the late spring of 1998 (a relatively wet spring in Central Iowa) a soil sample was taken from the Iowa State University Agronomy Farm to measure the soil moisture content. The 5 ft (1.5 m) soil core suggested that only 3 in. (7.6 cm) of water was available in the profile. This was judged as far too low; the site field capacity was around 10 in. (25.4 cm). Sampling at another point in the field indicated 8 in (20 cm) of water in the profile. Apparently, the sampler hit a core of sand which drained very quickly and produced non-representative results despite careful sampling methods. This is why replication and critical review of the data is necessary.

## Try This: Assess a Population by Sampling

Many issues become apparent when trying to assess the parameters of a population by sampling. In this hypothetical example, you will assess the mean and variability of potassium in a field.

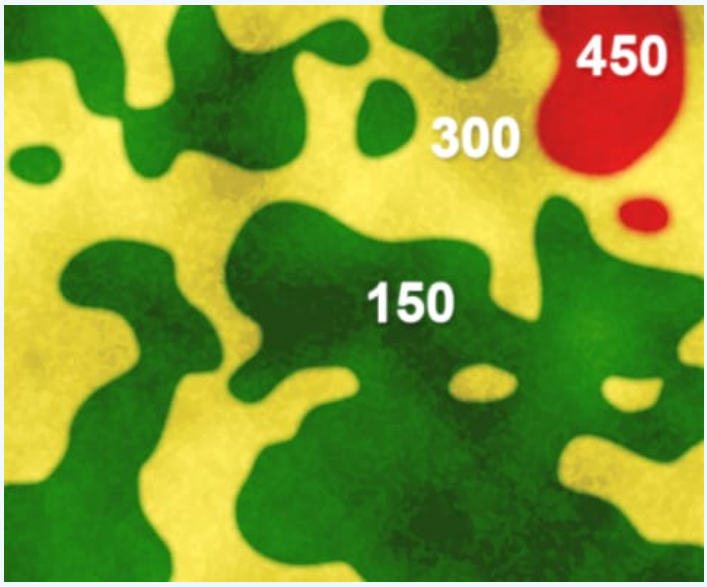



Figure 6 Color variation depicts variation in potassium levels depending on soil properties (measurements are in ppm).

The field in this example is the 480-acre (194 ha) field illustrated in Figure 6. In the figure, color variation depicts variation in potassium levels measured in parts per million (ppm), with red representing 450 ppm, yellow as 300 ppm, and green as 150 ppm. The figure illustrates that potassium levels vary depending on soil properties.

However, in the illustration below (Figure 7), we can see that different areas of the field vary much more widely than implied in Figure 6. Even within the broader pattern of major soil variants, here, the darker shades indicate higher levels of potassium, and lighter shades indicate lower ones.

## Try This: Assess a Population

Pick a set of three separate squares (locations) in the field to sample from (Figure 7). Select sets of locations to calculate an average potassium value for the field. For example, a set with 300, 150, and 450 would result in a mean of 300.

What does this suggest about how to design a sampling scheme to best represent the “population” mean value for the field? What happens when you increase the number of samples or vary their location? Would it be better to be random in your sampling scheme or systematic?

If you do this computation several times, between sets, you will notice that variation in the calculated average value occurs depending on where the samples are drawn.

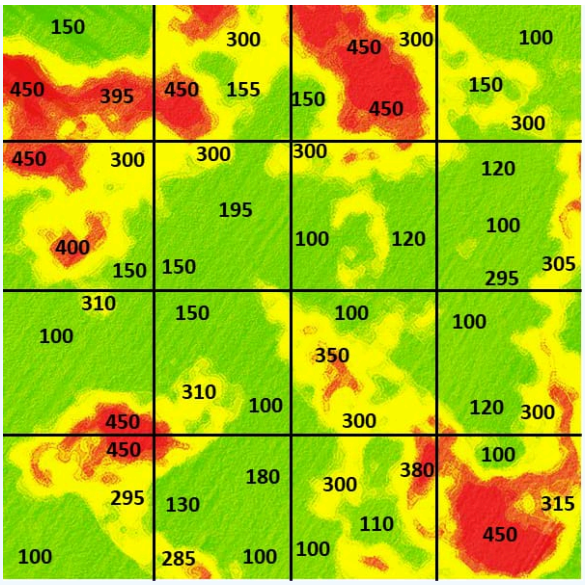



Figure 7 Potassium concentration variation across the field.

## Study Question 1 – A Sample Represents a Population

### Query 1.2.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



## Discussion

Where did you sample? How many samples did you need to understand the variability? What other issues are involved in sampling?

## Accurate Samples

Our sampling example illustrates that it can be hard to get a sample to accurately reflect the population (Figure 8).

We see from the example in the previous screens that it is difficult to get a representative sample for measuring an underlying parameter of the population, such as the average potassium concentration in an entire field. Taking three observations, and

calculating a mean from them, better represents the true population mean than would an individual value. More than three observations would be even better.

It is generally better to take a random sample than a systematic one. Random samples provide a method to get unbiased estimates of population values. W.G. Cochran gives an example illustrating this on page 121 of his book *Experimental Designs*, co-authored with Gertrude Cox. Even experts have a bias when trying to select a representative sample.

In an experiment to measure the heights of wheat plants in England, several expert samplers chose what they thought were eight representative plants from each of six small areas containing about 80 plants. Every expert ended up choosing samples taller than the average of all the plants in the area. Of the 36 total samples, only 3 had shorter average wheat height than the corresponding area. Samplers averaged from 1.2 cm to nearly 7 cm over the actual height for systematic samples compared with the actual average for the six plots. It is likely that their eyes were drawn to the taller plants. A properly conducted random sampling scheme would have avoided this bias.

# Histograms & Percentiles

## Purpose of Histograms

The frequency histogram gives a “picture” of the population. Once data are collected, we wish to understand the nature of the data better, and one method is to picture the data distribution with a histogram. The histogram is a diagram that gives frequencies of occurrence of data points on one axis and the values of the measurements on a second axis. In the frequency histogram, the value (height) of each bar is the number of variates (samples) that have that value or fall in that data range. An example is given here in Figure 9.

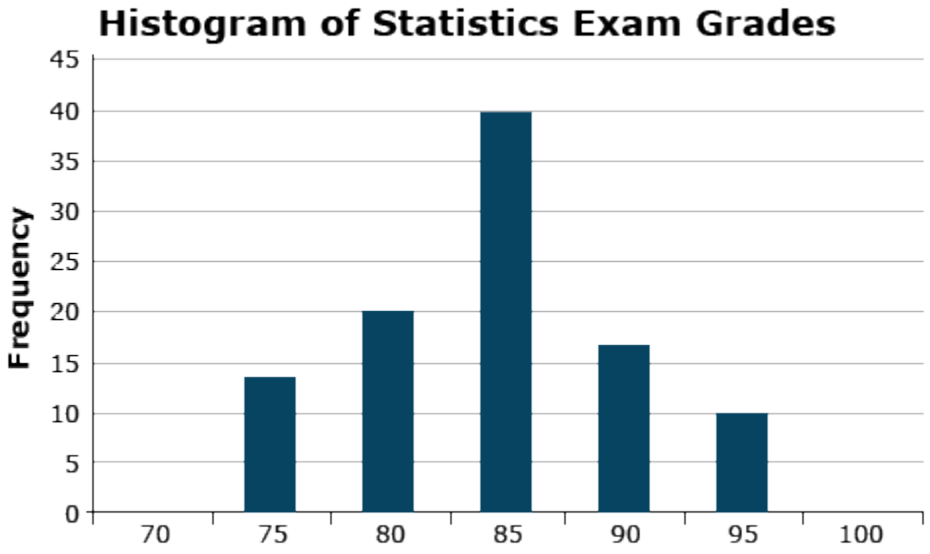



Figure 9 An example of a frequency histogram.

The histogram gives the overall pattern of the data. It shows how spread out the data is, and which values occur most frequently. It also shows potential outliers or unusual data values.

## Creating a Histogram

Proportions of datasets can be viewed in histograms.

One can view in the histogram the proportion of data less than or greater than a given value. We can also find values for which, say, 10% of the observations will be less than that value. This defines the “10th percentile” of the distribution. The median is the 50th percentile. Other useful percentiles are the first quartile (25th percentile) and third quartile (75th percentile).

The following exercise will use the Weldon, Illinois (USA) data provided in the Excel file named QM-mod2-ex1data.xls. For this exercise, we will use Weldon Root Pulls Set 4552 worksheet in the file. In the homework in the chapter on Central Limit Theorem,

Confidence Intervals, and Hypothesis Tests, we will use a second dataset located in this same Excel file, but in a different worksheet. For a portion of the homework assignment, instead of the Weldon data, we will use data from Slater, Iowa (USA) found in the worksheet titled **Slater Root Pulls Set 4552**.

## Exercises: Using Histograms

**Ex. 1: Using Histograms (1)**

An example of the use of histograms for exploratory data analysis

Our company entomologist called to ask me a question about data. One of his technicians, after using equipment to record root-pull data, thought values taken for the Slater, Iowa location were too high. Root-pull data are taken with a data recorder connected to an electronic load cell. A boom on a front-end loader of a tractor is hitched by cable to corn plants, and the force required to pull plants from the ground is recorded from the load cell. Entomologists can then select for corn hybrids with stronger root systems, sometimes even infesting with corn rootworm eggs, to select hybrids effective against these pests. But, to correctly select, we must have good measurements.

We decided to compare the measurements obtained from the Slater location with those taken the week before from a location near Weldon, Illinois. The entomologist asked for histograms of the distributions of the root-pull measurements from each location, means, standard deviations, quantiles, and medians for the data. In particular, he wanted to know if his group should stop additional root pulling at Slater because of poor quality data.

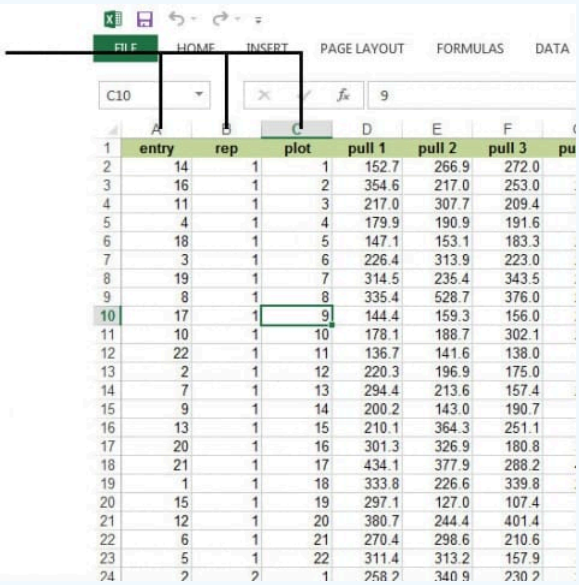
Typically, we need to do some work to get the data into the correct form to do statistical analysis.

The scientists provided data to me in an Excel file. My first step was to get the data for each location into a separate worksheet, with the titles of each variable in the first row. I then went through the following steps, which I request that you do in the following exercise.

**Ex. 1: Using Histograms (2)**

Open the data table in the Excel file named QM-mod2-ex1data.xls. Choose the Weldon worksheet. Because Weldon was the first location where data were taken, we want to examine this distribution and later compare it with our Slater data.

First, we need to inspect the data. For example, we need to ensure the data for any particular characteristic are not a mixture of character and numeric values. If there is a mixture, it may be necessary to modify data in the Excel worksheet. Also, for future use of SAS or R software or other statistical analyses, it is important to keep the names of the variables as in the first row in the Excel spreadsheet, with appropriate data arranged in the rows below each column. Other text or headings in the Excel spreadsheet can mess up your data analyses when transferring the data to another analysis program (Figure 10).




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | entry | rep | plot | pull 1 | pull 2 | pull 3 |
| 2 | 14 | 1 | 1 | 152.7 | 266.9 | 272.0 |
| 3 | 16 | 1 | 2 | 354.6 | 217.0 | 253.0 |
| 4 | 11 | 1 | 3 | 217.0 | 307.7 | 209.4 |
| 5 | 4 | 1 | 4 | 179.9 | 190.9 | 191.6 |
| 6 | 18 | 1 | 5 | 147.1 | 153.1 | 183.3 |
| 7 | 3 | 1 | 6 | 226.4 | 313.9 | 223.0 |
| 8 | 19 | 1 | 7 | 314.5 | 235.4 | 343.5 |
| 9 | 8 | 1 | 8 | 335.4 | 528.7 | 376.0 |
| 10 | 17 | 1 | 9 | 144.4 | 159.3 | 156.0 |
| 11 | 10 | 1 | 10 | 178.1 | 188.7 | 302.1 |
| 12 | 22 | 1 | 11 | 136.7 | 141.6 | 138.0 |
| 13 | 2 | 1 | 12 | 220.3 | 196.9 | 175.0 |
| 14 | 7 | 1 | 13 | 294.4 | 213.6 | 157.4 |
| 15 | 9 | 1 | 14 | 200.2 | 143.0 | 190.7 |
| 16 | 13 | 1 | 15 | 210.1 | 364.3 | 251.1 |
| 17 | 20 | 1 | 16 | 301.3 | 326.9 | 180.8 |
| 18 | 21 | 1 | 17 | 434.1 | 377.9 | 288.2 |
| 19 | 1 | 1 | 18 | 333.8 | 226.6 | 339.8 |
| 20 | 15 | 1 | 19 | 297.1 | 127.0 | 107.4 |
| 21 | 12 | 1 | 20 | 380.7 | 244.4 | 401.4 |
| 22 | 6 | 1 | 21 | 270.4 | 298.6 | 210.6 |
| 23 | 5 | 1 | 22 | 311.4 | 313.2 | 157.9 |
| 24 | 2 | 2 | 1 | 258.2 | 340.9 | 230.2 |

Figure 10. Sample worksheet of data.

Notice that there are columns for entry (corn hybrid number), rep, and plot within the rep. There are 22 corn hybrids in each of the 3 reps (blocks) of this experiment. Root pull measurements are recorded for 8 plants within each row for each plot.

For a detailed description of how to create a histogram in Excel, go to Excel Help under the File tab and search for histogram. The first option gives a step-by-step description.

**Ex. 1: Using Histograms (3)**

**Follow the steps in Figs 11, 12, and 13, to generate a histogram.**

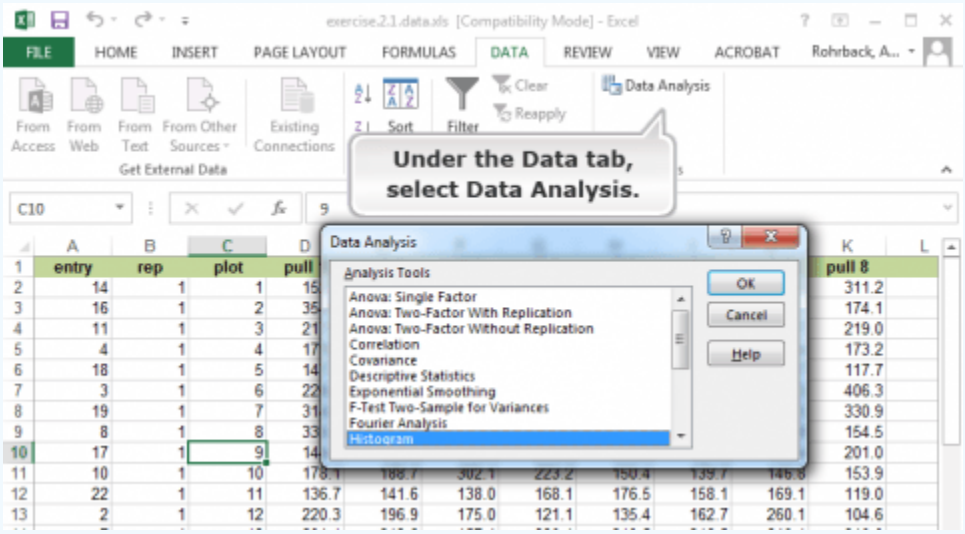



Figure 11 Finding and selecting the Histogram option.

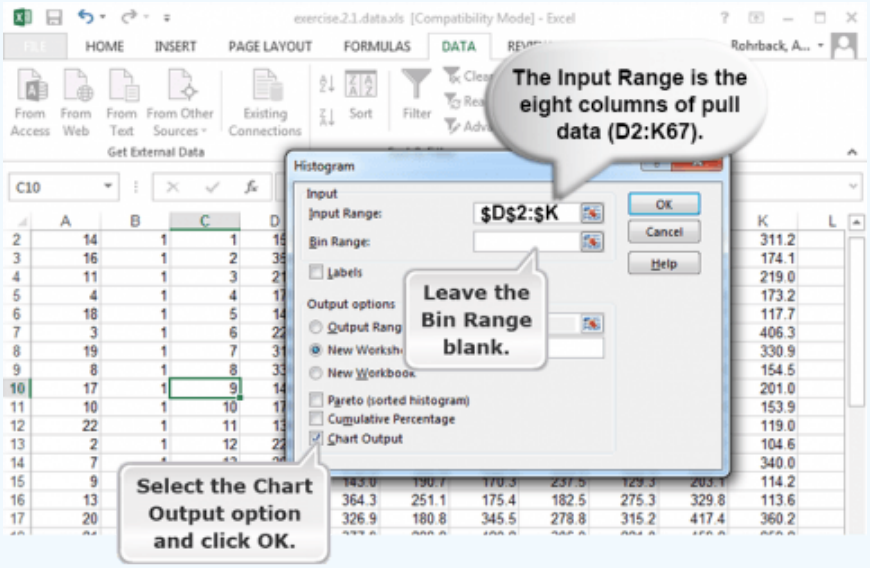



Figure 12 Enter the pull data, select the Chart Output option, and click OK

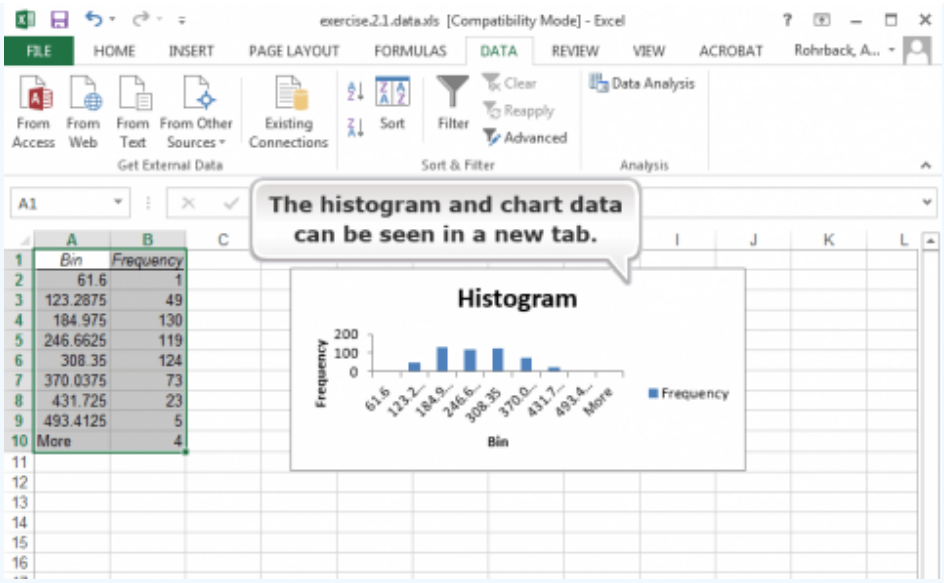



Figure 13 View the histogram and chart data

**Ex. 1: Using Histograms (4)**

**Now that we have the histogram, how do we interpret it?**

The histogram visually shows the distribution of the root pull values, ranging from low of 62 to high of 555 pounds (you wouldn't want to try to hand pull that one) (Figure 14). The first quartile, as the name suggests, is the value below which 25% of the distribution lies. For the Weldon root-pull distribution, 25% of the values are less than 165.50. The median is the second quartile (227.75 in this example) and the third quartile is the value below which 75% of the distribution lies (294.02).

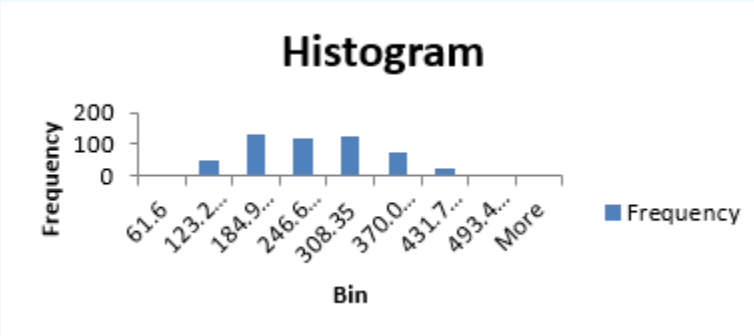



Figure 14 Histogram of root pull values.

Quartiles can be easily calculated in Excel. Select an empty cell and enter the formula "=Quartile(D2:K67, 1)" for the first quartile (25th percentile). The "1" in this function will give you the first quartile, a "2" will give you the second quartile, and a "3" will give you the third quartile. The median can be calculated using 2 in the formula instead of 1, and the 3rd quartile can be calculated using 3 in the formula.

In general, quantiles are the values below which a certain proportion of the distribution lies. For example, the 90th percentile is the value below which 90% of the distribution lies (347.92 in this case).

The maximum and minimum values of a sample can be found using the "Max" and "Min" formulas, respectively.

The distribution of root pulls tapers off toward each extreme, with only a few values higher than 400 or lower than 100 pounds. The mean is 233.78, very close to the median, and the standard deviation is 87.89. It is the shape of this distribution and its summary statistics which we will use to compare with the data from Slater.

**What about getting the output into a Word document?**

You can use copy and paste or cut and paste to get portions of the output into other programs such as Microsoft Word. Always place your analyses into Word documents before submitting them.

# Probability

## Taking a Random Sample

When an experiment is performed, there is frequently a chance that the value observed from a trial will differ from the previous observation, even when the environment is the same. Consider four soil samples taken throughout a field. The farmer wishes to apply the fertilizer evenly throughout the field, so all samples are considered to have been taken from the same environment. When the soil samples are analyzed, it is determined that the pH differs for each of the four samples. The differences in pH between soil samples are due to variables that cannot be controlled in the trial or experiment. When a trial or experiment is affected by variables that cannot be controlled, it is called random.

## Sample Space

Within every random experiment, there exists a sample space. This is a mathematical space; you could think of the sample space as the area on a graph from which all possible observations could be drawn. The pH measurements from the soil samples exist in a one-dimensional sample space of all possible pH measurements (Figure 15).

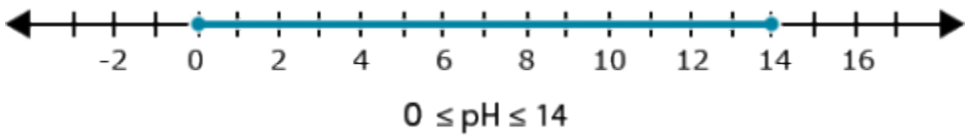



Figure 15 Sample space associated with pH observations

## Events

The sample space of an experiment contains subsets called **events**. In the soil sample experiment, each event is a measurement of a particular pH. Since pH is a **continuous** measurement and, therefore, more complex, let's use the example of a trait that differs between plants due to one gene. Each gene has slight changes within its makeup that cause differences between plants. Each of these forms of the gene can be differentiated and are called alleles. For this example, let's say that there are two alleles segregating in a population. In other words, *there are two possible outcomes to each trial*, and the sample space is made up of these two possible outcomes. The results of the test can be that the allele form is $A_1$ or $A_2$, and the sample space is called A. An event, in this case, is the observation of either $A_1$ or $A_2$ from sample space A.

## Defining Sample Spaces and Events of Interest

It is possible to imagine that there are more than two alleles of gene A in the segregating population. However, it's important to understand that this is an example of defining a sample space and events of interest. This decision is usually made from previous information about the observations being taken or the environment. The population of interest may be known to the researcher, and they know that there are only two alleles of gene A in the population. In reference to the soil sample experiment, there aren't any pH scores higher than 14. This results in a defined sample space of $0 \leq pH \leq 14$. These sorts of definitions are common throughout mathematics and statistics. It is necessary to make some limitations on possibilities to allow an experiment to be analyzed.



Figure 16 A dice roll is one of the classic illustrations of probability. The mean of all possible rolls is 7, and when tabulating a large number of rolls, the frequency distribution will center around the number 7.

In a random experiment, there is uncertainty about what the outcome will be. The outcome will be within the sample space, but where in the sample space remains uncertain until the observation is made. It is useful to be able to measure how likely an event is to occur during an experiment. The measure of the expectation that an event will occur during an **experiment** is called probability (Figure 16).

Probabilities have classically been calculated as the ratio of the number of times an event occurred to the total number of events. If an event occurs h times and the total number of observations is n, then the probability of the event occurring is h/n. The sum across all events in a sample space will be 1. This is a defined property of probabilities.

## Mathematical Symbols Used in Probability Equations

The symbol ∪ means the “union of” or “or”, so A∪B means the set of those elements that are either in A, or in B, or in both. Below we will use the symbol, ∩, which means “intersect” or “and”, so A∩B means the set that contains all those elements that A and B have in common (Figure 17).

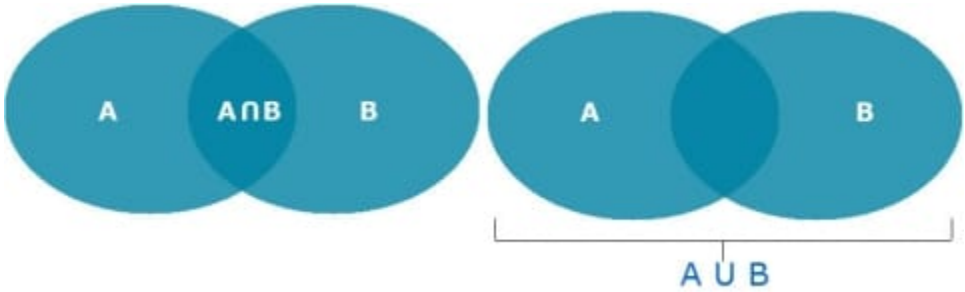



Figure 17 Illustrations of probability terms “intersection” (left) and “union” (right).

## Mutually Exclusive



Figure 18 A coin toss is often used to determine who plays offense first in a sporting event. Photo by the U.S. Navy.

To better understand when events are **mutually exclusive**, let's go back to the classic example of flipping a coin (Figure 18). When you flip a coin, it cannot be both heads and tails. The events of heads or tails on the same coin flip are mutually exclusive. It's very common to want to determine the probability of two mutually exclusive events.

The probability of a flip resulting in heads (h) or a flip resulting in tails (t) is defined as:

Pr(h)=0.50

Pr(t)=0.50

Pr(h or t) = Pr(h U t) = Pr(h)+Pr(t) = 1

## Calculating the Probability Associated with Two or More Events

The calculation of the probability of two or more events occurring follows some basic rules that depend on whether or not the events are mutually exclusive. For example, a chromosome with one copy of gene A cannot carry $A_1$ and $A_2$ at the same time (Table 1).

*Table 1 Probability of alleles of gene A on a chromosome.*

| Allele | Number of times allele was observed (n) | Total observations (n) | Probability of allele |
|---|---|---|---|
| $A_1$ | 120 | 200 | $\Pr(A_1) = \frac{120}{200} = 0.60$ |
| $A_2$ | 80 | 200 | $\Pr(A_1) = \frac{80}{200} = 0.40$ |

However, if two chromosomes are considered, there are two alleles. If these two chromosomes segregate independently, so the two events that occur are non-mutually exclusive. This is not usually immediately obvious — for more details, keep reading.

## Non-Mutually Exclusive Events

When two events can occur at the same, they are **not mutually exclusive**. Consider the example of two alleles in a diploid plant, i.e., there is an allele on each of two chromosomes. Gene A has two events, $A_1$ and $A_2$. The events that occur on each chromosome within a plant are inherited independently of each other and are not mutually exclusive. Calculating that two events will occur together (either simultaneously or sequentially) is done by calculating the two probabilities by each other.

Calculating that the two chromosomes will both carry allele $A_1$:

$$\Pr(A_1 \text{ and } A_2) = \Pr(A_1A_2) = 0.6 \times 0.6 = 36 \tag{1.2.1}$$

## Joint Probability

The probability associated with two independent events occurring is called the **joint probability.** It is calculated as the product of the probabilities of each of the events occurring. A common use of joint probability is a Punnett Square, which is a way of calculating the probabilities and possibilities associated different alleles of a gene on two chromosomes (Figure 19).

Figure illustrates the equations use to calculate the probability of two independent events occurring as the product of their individual probabilities.

Figure 19 Punnett Square in probability notation equations.

## Marginal Probability

Another term for the probability of an event occurring is the **marginal probability**. The joint probabilities associated with an event can be summed to the marginal probability (Figure 20).

Illustrations of the equations use to calculate the marginal probability of each of two independent events occurring as the sum of their individual probabilities.

Figure 20 Punnett Square and Marginal Probability

## Conditional Probability

A conditional probability can be calculated when it is known that one event has occurred or will occur. It is calculated as the ratio of both events occurring to the probability of the known event.

$$\text{Equation 2} \tag{1.2.2}$$

Events may not be independent of each other. In this case, the probability of both occurring is not the product. An example of this would be two linked genes, gene A and gene B. Gene A has the two alleles mentioned before. Gene B also has two alleles segregating in the population, $B_1$ and $B_2$. Since inheriting alleles at locus A is not independent of locus B, the probability of any two alleles occurring together on the same chromosome (the joint probability) has to be determined through the behavior of the two events. For this example, the joint probabilities have been given. In order to calculate the probability of allele $B_1$ occurring when it is known that allele $A_1$ is present, the following calculation can be used:

Assume that $\Pr(B_1 \cap A_1) = 0.3$. Then

$$\text{Equation 3} \tag{1.2.3}$$

This is exactly how bio-markers are used to determine genetic risk for many human diseases, and it is how markers are used in selection during plant breeding.

## Probability Distributions

Each set of events in a sample space has an associated probability distribution. A distribution is a means of depicting the frequency with which each event occurs. Along the x-axis are the events, and along the y-axis are the probabilities from 0 – 1. The probabilities across all events on the sample space equal 1 (or 100%) and the distribution describes how the probability is allotted across all events. The two types of observations, discrete and continuous, have different types of probability distributions.

## Discrete Distribution

Within a **discrete distribution**, each event type is clearly defined and has a probability assigned to it. The segregating gene A is an example of a discrete data type. There are two clearly defined events, $A_1$ and $A_2$. Each of these events has a probability assigned to it and those probabilities add to 1. Each of the events serves as a step in the total distribution. The probability

distribution can be seen in Figure 21.

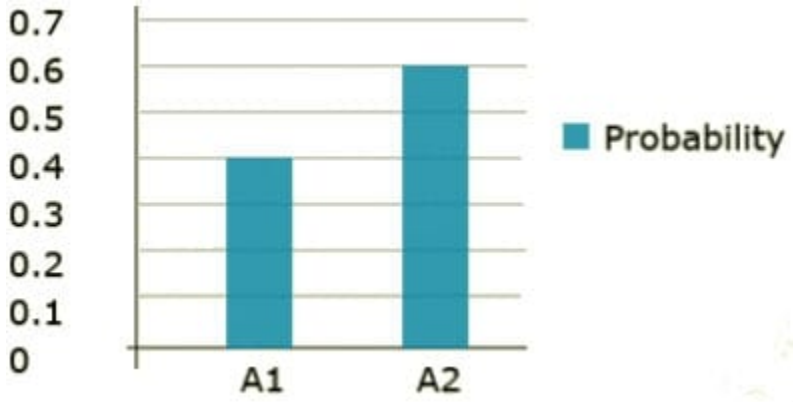



Figure 21 Probability distribution of alleles of Gene A.

## Continuous Distribution

Continuous data types have an infinite number of possible events in the sample space. For example, a pH measurement could theoretically be measured to an infinite number of significant digits. This results in a necessary change to the probability distribution. When there are an infinite number of possible events, no single event has a probability. This situation results in a continuous distribution rather than a distribution comprised of discrete events. The continuous distribution is derived from an infinite number of events. The probability of an event of interest is measured as the area under the distribution to the left or right of the event of interest. That probability is more commonly called a p-value. The probability distribution for the pH of the soil samples is given in Figure 22.

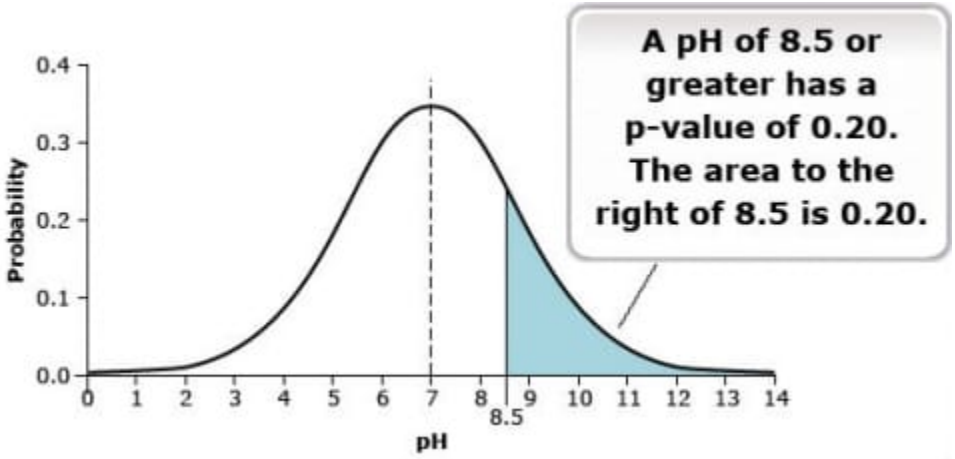



Figure 22 Probability distribution for pH of soil samples.

# Normal Distribution

## The Most Common, Best-Studied Statistical Distribution

In the root-pull examples, we saw histograms of a distribution which had most values concentrated toward the center and fewer at the extremes. Although not a perfect bell-shaped curve, the curve formed by connecting the tops of the rectangles in the frequency histogram approximates the shape of a bell. These bell-shaped distributions, referred to by statisticians as normal distributions (Figure 23), are typical for many types of data, including plant heights, grain yields, grain moistures, soil nutrient values, and many others.

Mathematicians and scientists first noticed and mathematically described the normal distribution in the early 1700s. At that time, many scientists were studying astronomy, and they saw a characteristic bell-shaped curve in histograms of the errors of their

measurements. Abraham De Moivre first described the distribution mathematically in 1733. The famous French mathematician Pierre LaPlace and German mathematical prodigy Karl Gauss both applied the normal curve to errors of measurement.

## What it Does

**Normal distribution allows us to calculate probabilities of events.**

By using this type of distribution, we can infer a great amount of information and make probability statements based on the distribution. The probability function for a normally distributed random variable, x, is described mathematically as:

Normal Distribution Equation (Equation 4)

Equation 4 (1.2.4)

**where:**
$f$= probability function
$x$= a variate with normal distribution
\sigma^2= variance of populaton
$\sigma$= standard deviation of populaton
$\mu$= population mean
$e$= base of natural logarithm

The probabilities for the occurrence of an event, such as corn yields between 150 and 170 bushels/acre (9.4 – 10.7 tons/ha) are found by summing the areas under the distribution, also called integrating the function. The area is approximated by computer programs which subdivide the area into very small rectangles and then sum their areas. Adhering to the rules of probability, we know the total area under the curve and above the horizontal axis sums to 1.

## Basic Parameters

**The normal distribution is based on knowledge of two main parameters: the mean and standard deviation.**

While the Normal Distribution equation may appear daunting, the main aspect of it is that you can calculate the probability of ranges of values in a normal distribution by knowing two parameters:

- **mean**
- **standard deviation**

Based on knowledge of these parameters, you can substitute their values and predict the outcomes. However, this assumes that you know the true values for the parameters describing the population. If we do not know the true values and instead estimate the values from a sample of the population we have to recognize that the sample could be biased and have some error.

## Reasons for Widespread Use

There are four main reasons for the widespread use of the normal distribution:

1. The normal distribution allows probabilities to be assigned to outcomes of an experiment, completing the experiment process from collecting data to interpreting results using these probabilities.
2. Many variables have distributions close to the theoretical normal distribution.
3. Even variables with other distributions, such as the binomial distribution (e.g., number of seeds germinating from 400 planted) are reasonably approximated by the normal when samples are large. Other variables can be transformed to closely follow the normal distribution.
4. Averages from any distribution approach the normal distribution as sample size gets large. This characteristic is very helpful because we often want to make probability statements about mean

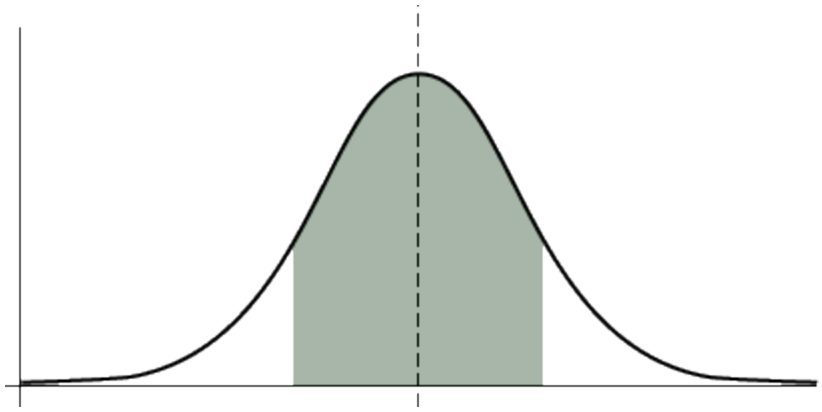


Figure 23 The normal distribution.

## Properties of Normal Distribution

**Properties of the normal distribution help us draw statistical inferences.**

Properties of the normal distribution include the following:

- The distribution is symmetric about its mean
- Roughly two-thirds of the values (68%) lie within one standard deviation of the mean
- About 95% of the population is within 2 standard deviations of the mean
- Probabilities for intervals of values can be computed

To compute probabilities, for example those in Appendix 1, we only need to use the normal distribution with mean zero and standard deviation one (μ=0, σ=1). We then convert any value in the population to its number of standard deviations above or below the mean. For example, if corn yields are normally distributed with mean 150 bushels/acre (9.4 tons/ha) and standard deviation 10, we know that a yield of 140 bushels/acre (8.8 tons/ha) is one standard deviation below the mean. Using property #3, we expect 95% of the values for this population to be between 130 and 170 bushels per acre (8.2 – 10.7 tons/ha).

### Study Question 2: Normal Distribution

#### Query 1.2.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Z-Scores

**Z-Scores are the number of standard deviations above or below the mean.**

Knowing the mean and standard deviation of a normally distributed variable completely describes the population distribution and allows comparison with other populations. It is also useful for evaluating where a specific value occurs in the distribution, how common or how extreme, i.e., how far it occurs from the mean. These comparisons are evaluated using the z-score (Equation 5). The z-score formula is defined as:

$$Z = \frac{x_i - \mu}{\sigma} \tag{1.2.5}$$

**where:**
$x_i$= value of observation i
$\mu$ = population mean
$\sigma$ = population standard deviation

The z-scores from a normal population have a standard normal distribution, i.e., a mean of 0 and a standard deviation of 1. They are dimensionless because the numerator and denominator of the fraction have the same units, and when doing the division, the units are removed. These facets of the z-score make it useful for comparing populations with known mean and standard deviation. Any value in a normally distributed population can be assigned a z-value.

## Calculation

Let's look at an example. On July 1 you hear on the radio that daily high temperatures during the month of June averaged 85.6°F (29.8°C), while the average June daily high temperature over many years is X = 81.6°F (27.6°C). The 4°F (2.2°C) difference doesn't seem that large, but how warm is this based on historical June data? The standard deviation for monthly average high temperatures for Ames, IA in June is about 3.75°F, using data since 1900. Meteorologists have seen that average monthly high or low daily temperatures are normally distributed (see Figure 24). What is the z-score for 85.6°F, and what is the probability of a value this high or higher?

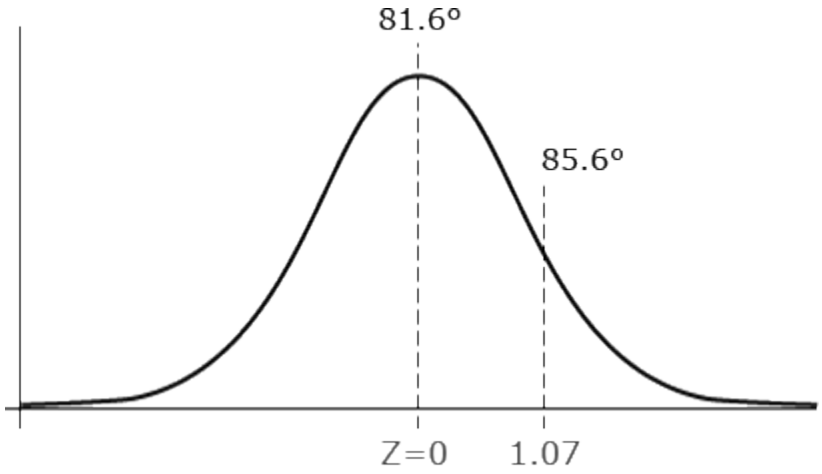



Figure 24 Comparing mean temperature in June of one year to the 100-year mean temperature for June by using Z-scores.

$$Z = \frac{x_i - \bar{x}}{\sigma} \tag{1.2.6}$$

**where:**
$x_i$= value of observation i
$\bar{x}$ = sample mean
$\sigma$ = population standard deviation

## Interpretation

The value of 85.6°F (29.8°C) is just a little more than 1 standard deviation larger than the mean. The probability for a z-score less than or equal to 1.07 is 0.8577. Since this is the probability less than z, or to the left of z, the probability of a higher value is 1 – 0.8577 = 0.1423. This means that there is a 14% chance of obtaining a value larger than 85.6°F June monthly average high. The 85.6°F is somewhat warm, but not extremely so. The standardization of values produced using the z-score allows comparison of different distributions which is sometimes difficult when looking only at raw numbers.

Suppose we want to compute the probability of finding a value less than 23.3 if it comes from a normal distribution with mean = 27.3 and standard deviation = 6.25. Note that here
z = (23.3 – 27.3) / 6.25 = -0.64.

By the symmetry of the normal distribution, the probability of a value less than z = -0.64 is the same as the probability of a value greater than z = 0.64. The probability of a value greater than z = 0.64 is 1 – P(z < 0.64) = 1 – 0.7389 = 0.2611. Therefore, the

probability of a value less than 23.3 is 26%.

### Study Question 3: Z-scores and Probability

#### Query 1.2.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## A Second Example of the Calculation of Z-Scores

Determining the probability of some value occurring between two values in normal distribution is also possible. By calculating the z-value of the two numbers, one can calculate the range between them and percentage chance of that occurrence. If we use the June temperature example in the previous screens, let's find out how often temperatures in June are between plus and minus 4°F (2.2°C) of the mean high temperature (81.6°F or 27.6°C). Also, we make use of the symmetric nature of the normal distribution (see Figure 25). We found above that 14% of the time temperatures are greater than 85.6 °F (29.8°C). We then know that 14% of the time temperatures will be less than 77.6°F (29.8°C). To complete the calculation:

$$P(avg \pm 4^\circ F) = 100\% - P(Z > 1.067) - P(Z < 1.067) = 72$$1.067)-P(Z<1.067)=72" title="P(avg\pm 4^{\circ} F)=100\%-P(Z>1.067)-P(Z<1.067)=72" class="latex mathjax">]

Equation 7 (1.2.7)

This same calculation can be done for any two values and their associated z-values. We have also illustrated the comparability of z-values above.

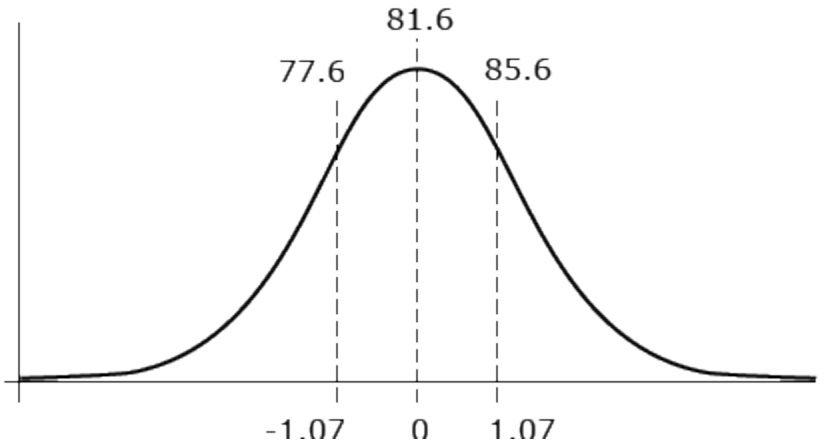



Figure 25 Normal distribution of z-values.

## Normal vs. Non-Normal Distribution

As we can see, for normally distributed data, it is relatively easy to compute probabilities. Not all distributions follow a normal distribution, however, and become more difficult to handle. One example is precipitation data (Figure 26).

Other methods of computing probabilities are needed for non-normal distributions.

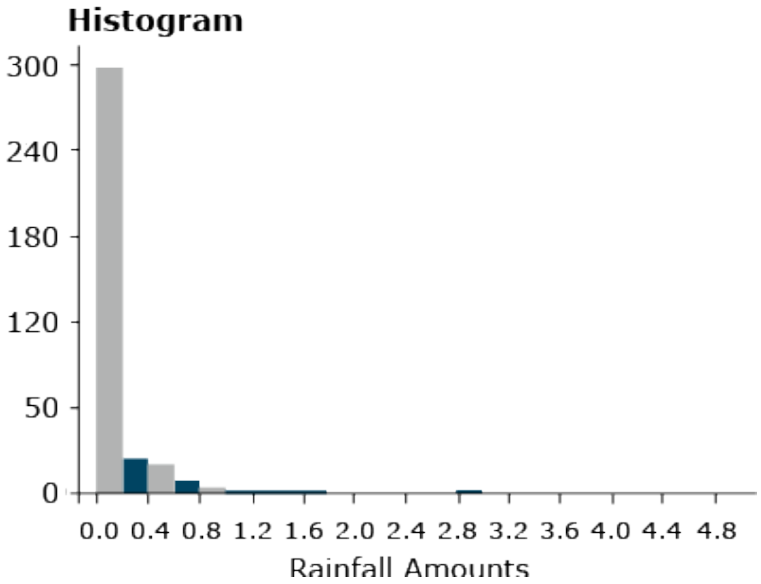



Figure 26 Daily rainfall values for a single year.

Study Question 4: Non-normal Distributions

Query 1.2.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Other Distributions

## Non-Normal Continuous Distributions

There are many other distributions which do not follow the normal, bell-shaped curve. The distribution of rainfall events (measured from a trace to several inches or centimeters) is an example of a non-normal continuous distribution. This distribution only takes values greater than zero. As shown in Figure 26, the histogram seems to decrease exponentially.

Other continuous distributions may be skewed rather than symmetric. Some data follow a distribution in which the logarithm of the variable is normally distributed. These data are skewed to varying degrees. Even symmetric data, such as from a uniform distribution, do not follow the bell-shaped histogram of a normal distribution.

## Non-Normal, Non-Continuous Distribution

Some data cannot follow a normal distribution because they are discrete and not continuous, for example, count data. We could have a uniform distribution of count data if, for example, corn plants in each 100 square-foot area (two 30 in. rows, 20 ft. long) planted with a precision planter have 60 plants in each plot.

Another example of count data is for rare events, such as the number of a certain type of weed in these same 100-square-foot plots. Counts of rare events often follow what is called a Poisson distribution. The formula for this distribution is in Equation 8:

$$P_k = \frac{\mu^k e^{-\mu}}{k!} \tag{1.2.8}$$

**where:**
$P_k$= probability of count being $k$
$\mu$= population mean
$e$= base of natural logarithms.

In this equation, $k!$ is the product of all integers less than or equal to $k$, $k!$ = $k(k-1)(k-2)...(2)(1)$. The symbol e is the base of the natural logarithm, approximately equal to 2.71828. It is used to describe exponential growth phenomena, such as compound interest.

## Poisson Distribution Example

An example of the Poisson distribution is given in the text of Robert Steel and James Torrie (1980, Principles and Procedures of Statistics, 2nd edition), in which the number of yeast cells is counted in each square of a grid of 400. It was rare to get even as many as four cells in a square, and for this sample there were never more than six. The observed frequencies are given in Figure 27.

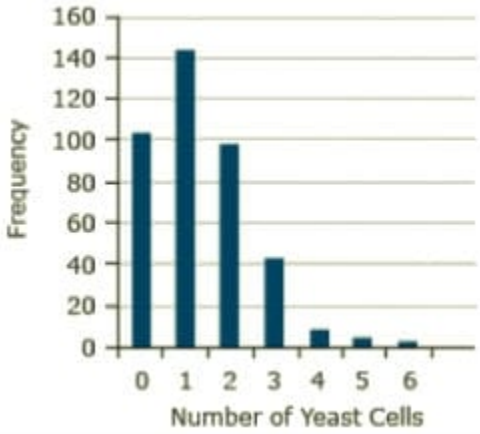



Figure 27 Distribution (Poisson) of yeast cells.

## Calculations

The total number of yeast cells for this sample is 529, and the mean or average per square is 529/400 = 1.3225. If this estimate is substituted for the true mean, the formula becomes

For example, the estimated probability for three cells computed from the formula is

$$P_k = \frac{\mu^k e^{-\mu}}{k!} = P_k = \frac{1.3225^3 e^{-1.3225}}{k!} \tag{1.2.9}$$

$$P_k = \frac{1.3225^3 e^{-1.3225}}{3!} \tag{1.2.10}$$

The expected frequency with 400 squares is 400(0.1027) = 41.09. The observed frequency of squares with three cells is 42. The Poisson distribution does a good job of describing the distribution in this sample.

We will see this distribution again when we explore transformations later in the course to make data more closely follow a normal distribution.

## Binomial Distribution

Another discrete distribution even more common in agronomic data is the binomial distribution. This is used, for example, in counting numbers of germinated seeds, seedlings which emerge, or diseased plants. In this distribution, we measure for each

experimental unit (plant or seed) one of two outcomes, for example, germinated or not, or alive or dead. We assume a true proportion p of seeds which would germinate and that each sample is of the same size, say n = 100 seeds. Each seed will germinate or not independently of any of the other seeds. The distribution of the number of seeds germinating follows what is called the binomial distribution.

Because of the importance of the binomial distribution, we will devote an entire unit to studying it.

Even though there are numerous examples of distributions other than the Normal, it is such an important distribution that we will concentrate in the next several units on analysis methods for normally distributed data.

# Summary

## Samples Represent the Population

- Sampling scheme or experiment design affect our ability to test hypotheses
- “Representative plants” generally have more bias than a random sample

## Histogram

- Gives a “picture” of the population

## Normal Distributions

- Bell-shaped curve
- Characteristic of many types of data (yields, plant heights, errors in a measurement)
- Provides a mechanism to assign probabilities

## Properties of Normal Distributions

- Bell-shaped curve
- Symmetric about the mean, $\mu$
- 68% of values are within 1 $\sigma$ and 95% are within 2 $\sigma$ of mean

## Z-scores

- Tell how many standard deviations above or below the mean
- Defined as $(Y - \mu)/\sigma$
- Allow computation of probabilities with the normal distribution

## Non-normal Distributions

- Many examples (daily rainfall)
- Some may be transformed to be normal

**How to cite this chapter:** Mowers, R., K. Moore, D., Todey, M.L. Harbur, K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Distributions and Probability. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.2: Distributions and Probability is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Ken Moore; Dennis Todey; M. L. Harbur; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests

In Chapter 2 on Distributions and Probability, we saw some examples of data distributions, especially the normal distribution, and learned to make probability statements from that distribution. In this unit, we see that sample means from a normal distribution are also distributed normally, but with variance reduced by a factor of n (the number of observations in the sample). We also tie these ideas to the scientific method of chapter 1 on Basic Principles by learning how to test hypotheses.

**Learning Objectives**

- Sample averages from a normal distribution are normally distributed.
- The central limit theorem.
- How to form confidence intervals for the mean of a normal distribution.
- How to create and test a hypothesis.

## Distribution of Sample Averages

### Averages of Samples

**Averages of samples from a normal distribution also follow a normal distribution.** Often we wish to estimate the mean of a population. We use sample averages to estimate the mean of a population of interest. What about these sample means — how can they be described? A set of sample means comprises its own population, which surprisingly approaches the normal distribution about the population mean.

The idea is this: suppose we start with a full normal distribution, take a sample of 10 observations, then compute the sample average. Then we take a second sample of 10 observations and again compute its sample mean. We repeat this process many times.

**Study Question 1: Distribution of Sample Averages**

**Query 1.3.1**

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Extreme Sample Mean Values

When we try to produce a distribution using sample means of just a few individuals, we notice that the curve is flatter and wider than if more individuals are in the sample. If an extreme (far from the mean) value is included in the sample, then it will have a large proportional effect in widening the distribution if the sample size is small. As we increase the number of individuals in the sample, the distribution for sample means grows narrower and taller. The effect of extreme values is diluted by the larger number of values closer to the population mean.

# Probability Distribution

Click here to download a file that examines the distribution of sample means: Distributions [XLS]

The probability distribution for a set of sample means, like the probability distribution for the actual population, can be described with only two measurements: the mean, and the standard error of the mean. The mean, $\mu_{\bar{x}}$, is the mean for the population of the sample means. The standard error of the mean, or standard deviation of sample means, is denoted by $\sigma_{\bar{x}}$. It describes the standard deviation of individual sample means around the population mean for the set, and is estimated by:

$$S_{\bar{x}}^2 \tag{1.3.1}$$

**where:**
$S_{\bar{x}}^2$ = variance of sample mean, $\bar{x}$
$S^2$ = variance of the sample
$n$ = number of values in the sample

The important thing to see from this formula is the relationship of the variance of sample means to the original variance of the sample — it is reduced by a factor of n. Sample means have less variance than do the original observations. The reduction factor is the sample size n.

Whenever we try to describe a population based on a sample, we must ask: how well does our sample mean represent the actual population mean, μ? We will answer this question when we study confidence intervals for the mean.

## Standard Error of the Mean

Just as the standard deviation of the population is calculated as the square root of the population variance, the standard error of the mean, SE, is calculated as the square root of the sample variance, divided by the number of observations:

$$SE = \sqrt{\frac{S}{n}} \tag{1.3.2}$$

**where:**
$SE$ = standard error of mean
$S^2$ = sample variance
$n$ = number of values in the sample

The standard error of the mean (SE) is commonly used to tell how much sample means fluctuate from sample to sample. Usually, we take a set of individuals from a population and calculate the mean of them to try to estimate the population mean. How that mean compares to the population mean is assessed using the standard error.

### Study Question 2: Distribution of Sample Averages

#### Query 1.3.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Central Limit Theorem

## Normal Distribution

Interestingly enough, means of samples from even a non-normal distribution can be normally distributed. This is the gist of what is called the "**Central Limit Theorem**." It states that when we draw simple random samples of size n from any distribution, the sample means become approximately normally distributed with mean μ and variance $\sigma^2/n$ (for large sample sizes).

Naturally, one question that comes to mind is, "How large must n be for the normal approximation to hold?" This depends on how similar to a normal distribution our original distribution is. If the original distribution is a normal one, the sample means are always normal, no matter how small the sample (even individual values are normally distributed). As the original distribution becomes less and less like the bell-shaped curve, sample sizes need to be larger to have sample averages nearly normally distributed.

### Study Question 3: Drawl Limit Theorem

#### Query 1.3.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Data Distributions

In the chapter 2 on Distributions and Probability, we looked at various types of data distributions including the normal distribution. Many groups of data are not normally distributed, but some closely approximate it. We will look at one of those and do some work with those data.

We have looked at monthly temperature data in chapter 2 and assumed that they were from a normal distribution. Now we will test whether that data is close to a normal distribution. This is both a review of some concepts from chapter 2 and some new methods for testing and working with the normal distribution.

The z-statistic in a normal distribution in essence standardizes the value. The standardized distribution has a mean of zero and a standard deviation of 1. This standardized value, z, allows you to compare different numbers and how different they are from the mean in their populations.

### Exercises

**Ex. 1: Evaluating a Distribution**

**Calculations**

Test some attributes of this data set on January average minimum temperatures to see if it follows a normal distribution.

- Open the file QM-mod3-ex1data.xlsx. The file holds data comprised of average January temperatures over a 100 year period.
- The first step is to determine the mean, median, and standard deviation.
  - In the empty cell next to the Mean label, enter the formula: "=average(B2:B102)". This will give the mean of the temperature data or the mean January temperature from 1900-2000.

- Next calculate the median using “=median(B2:B102)”. The median is the middle number after sorting the data by order of magnitude if there are an odd number of observations.
- Finally, the standard deviation is easily calculated using “=stdev.s(B2:B102)”; (Figure 1).

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | Year | January Minimum temperature | | | | | | |
| 2 | 1900 | -8.4 | | | | | | |
| 3 | 1901 | -9.9 | | Mean | -12.305 | | | |
| 4 | 1902 | -11.8 | | Median | -12.1 | | | |
| 5 | 1903 | -10.1 | | Std Dev | 3.39766 | | | |
| 6 | 1904 | -14.7 | | | | | | |
| 7 | 1905 | -17.1 | | | | | | |
| 8 | 1906 | -9.4 | | | | | | |
| 9 | 1907 | -12.3 | | | | | | |
| 10 | 1908 | -10.3 | | | | | | |
| 11 | 1909 | -10.8 | | | | | | |
| 12 | 1910 | -13.9 | | | | | | |
| 13 | 1911 | -13.1 | | | | | | |
| 14 | 1912 | -21.4 | | | | | | |
| 15 | 1913 | -12.7 | | | | | | |
| 16 | 1914 | -7.8 | | | | | | |
| 17 | 1915 | -13.1 | | | | | | |
| 18 | 1916 | -13.9 | | | | | | |
| 19 | 1917 | -14.9 | | | | | | |
| 20 | 1918 | -17.4 | | | | | | |
| 21 | 1919 | -8.2 | | | | | | |
| 22 | 1920 | -14 | | | | | | |
| 23 | 1921 | -6.4 | | | | | | |
| 24 | 1922 | -12.6 | | | | | | |

Figure 1 The Excel file.

**Create Histogram**

- In a normal distribution, the mean, median, and mode are identical.
- Create a histogram of the temperature data.
  - A histogram can be produced using the Data Analysis tool on the Data tab.
    - Select Data Analysis and then Histogram. Fill in the pop-up window in the manner in Figure 2 below:

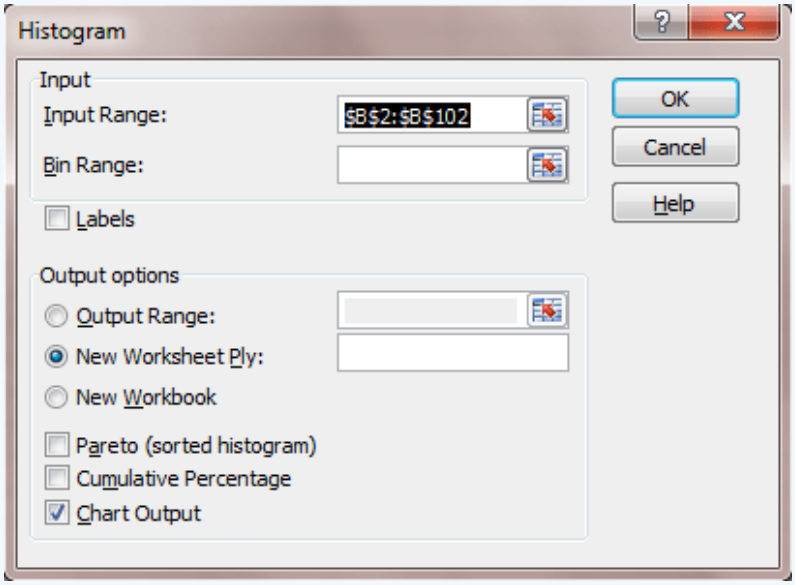



Figure 2 The Histogram settings.

**Finished Histogram**

- The histogram is given below (Figure 3):

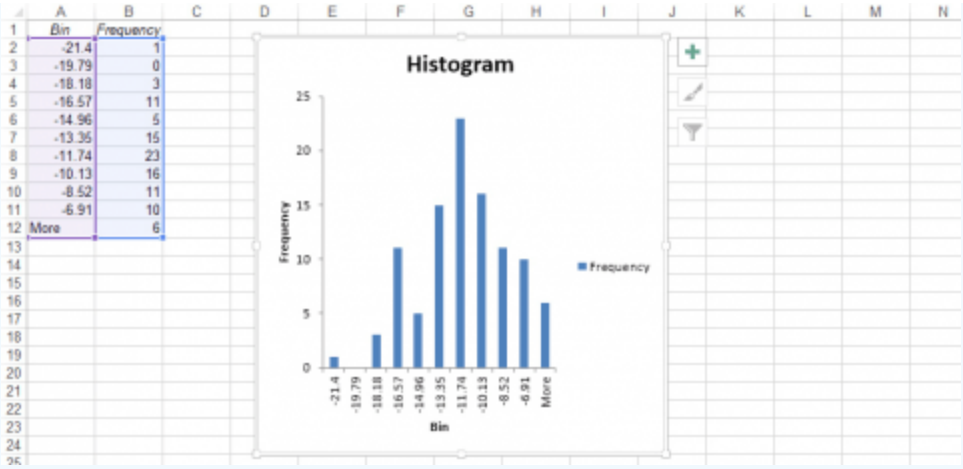



Figure 3 The Excel histogram.

**Probability Plot**

- We will now generate a normal probability plot
  - Go back to the sheet with the data set and select Data / Data Analysis / Rank and Percentile.
    - Fill in the pop-up window in the manner below (Figure 4):

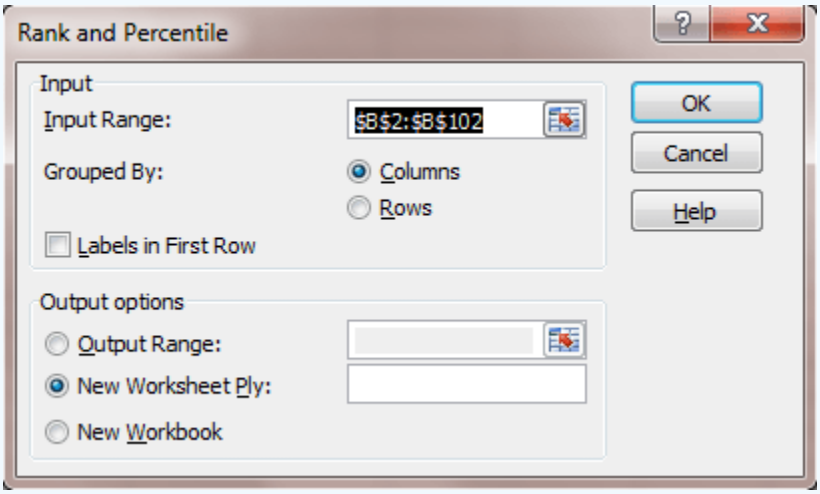
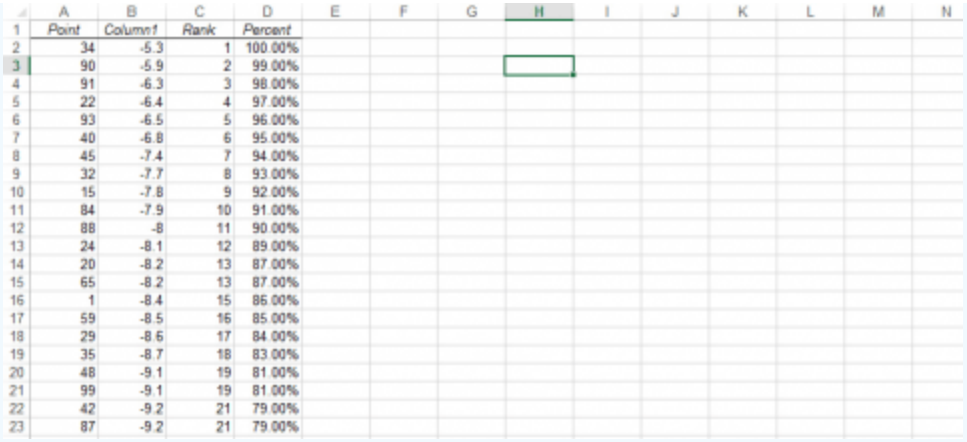



Figure 4 Rank and Percentile dialog box.

**Create Probability Plot**

- On the page with the rank and percentile data, click on a cell to the right of Column D (Figure 5). Select the Insert Tab and Scatter, then select the first type of scatter plot.



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Point | Column1 | Rank | Percent |
| 2 | 34 | -5.3 | 1 | 100.00% |
| 3 | 90 | -5.9 | 2 | 99.00% |
| 4 | 91 | -6.3 | 3 | 98.00% |
| 5 | 22 | -6.4 | 4 | 97.00% |
| 6 | 93 | -6.5 | 5 | 96.00% |
| 7 | 40 | -6.8 | 6 | 95.00% |
| 8 | 45 | -7.4 | 7 | 94.00% |
| 9 | 32 | -7.7 | 8 | 93.00% |
| 10 | 15 | -7.8 | 9 | 92.00% |
| 11 | 84 | -7.9 | 10 | 91.00% |
| 12 | 88 | -8 | 11 | 90.00% |
| 13 | 24 | -8.1 | 12 | 89.00% |
| 14 | 20 | -8.2 | 13 | 87.00% |
| 15 | 65 | -8.2 | 13 | 87.00% |
| 16 | 1 | -8.4 | 15 | 86.00% |
| 17 | 59 | -8.5 | 16 | 85.00% |
| 18 | 29 | -8.6 | 17 | 84.00% |
| 19 | 35 | -8.7 | 18 | 83.00% |
| 20 | 48 | -9.1 | 19 | 81.00% |
| 21 | 99 | -9.1 | 19 | 81.00% |
| 22 | 42 | -9.2 | 21 | 79.00% |
| 23 | 87 | -9.2 | 21 | 79.00% |

Figure 5 The Excel file.

**Finished Probability Plot**

- When the plot window opens and is selected, the chart tools will be available. Click Select Data and Add under Legend Entries (Series).
- The X-values are the Percent column, and the Y-values are in Column 1. Do not select the column headings in the x or y ranges, as that will cause you to create a plot that is the inverse of what you should have.
- Select OK twice, and the following plot will be visible (Figure 6).

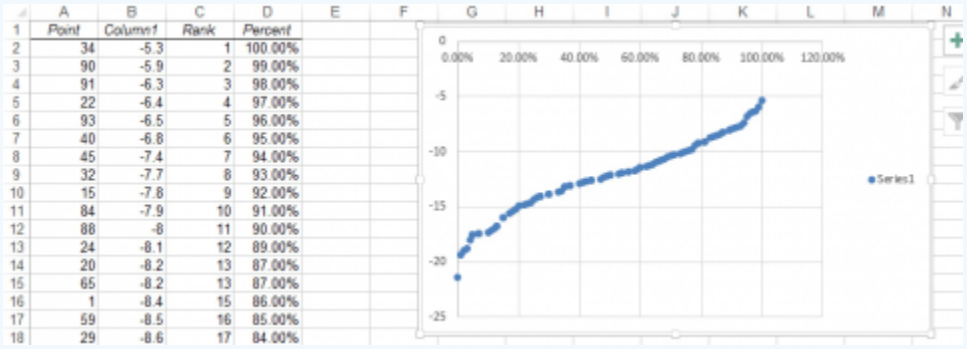



Figure 6 The completed scatter plot.

**Evaluating the Distribution**

Normally distributed data will fall along a straight line.

- A true Normal distribution is rare in a data set. It is up to the researcher to determine if the data are close to normally distributed. One informal test is to place a pen over the line of the Normal Probability Plot, and if there are few observations that can be seen then the observations are approximately normally distributed. There are also tests of Normality that can be used, but are not covered here.
- If the tails are above and below the straight line (as in this case), there is a little more variance than you might expect from a perfect normal distribution. If the tails are below the straight line, then the variance is less than expected from a normal distribution.
- This data are approximately normally distributed. While there are a few possible outliers, the normal probability plot and histogram suggest approximate normality.

**Ex. 2: Calculate the Z-Statistic**

**Using Excel Formulas**

Excel has bui .78 in January.

There are many ways to find the answer to this question, and here we illustrate two. In the following, we are using the mean = -12.305 and standard deviation = 3.398 found for the normal curve in exercise 1.

One way to answer the question is to calculate the z-value for a temperature of zero.

Calculate z for the temperature = -17.78. z = (-12.305 – (-17.78))/3.398, or z = -1.61. The probability of a value less than this is 1 – 0.9463, where 0.9463 is the table value for z = 1.61. Our answer is 0.0537.

A second way to calculate this probability is to use Excel.

- Open a new Excel workbook (Figure 7).
- Label the first column Original Temperature, the second Mean, the third Standard Deviation, a fourth column Z-Value, and the fifth column Probability.
- Enter -17.78 in the first row under temperature, -12.305 under the mean, and 3.398 under standard deviation.
- Enter the formula "=(A2 – B2)/C2" into the first row under z-value.
- Finally, enter the formula "=norm.s.dist(D2, TRUE)" in the cell below Probability.

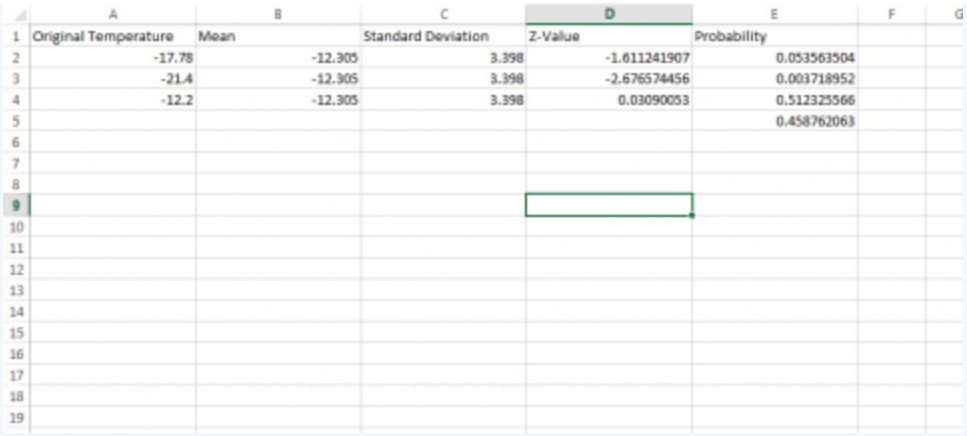
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Original Temperature | Mean | Standard Deviation | Z-Value | Probability | | |
| 2 | -17.78 | -12.305 | 3.398 | -1.611241907 | 0.053563504 | | |
| 3 | | | | | | | |
| 4 | | | | | | | |
| 5 | | | | | | | |
| 6 | | | | | | | |
| 7 | | | | | | | |
| 8 | | | | | | | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| 16 | | | | | | | |
| 17 | | | | | | | |
| 18 | | | | | | | |
| 19 | | | | | | | |

Figure 7 The Excel file showing the calculated values.

**Further Application**

As you can see, the probability of a value being below any z-value can be computed with this formula.

- How often does a temperature of -17.78°C occur? According to these data, about once every 20 years.
- Try the same for the extreme value of -21.4. You should see that it is extremely unlikely for this low a temperature to occur. This would imply that the average low temperature for the entire month is -21.4°C, an extremely cold month!
- Calculate this by entering -21.4 into the second cell under Temperature, then copy and paste the mean, standard deviation, and two formulas into the rows next to the new temperature.
- To find the probability of a value falling between two values we have to do the calculation twice.
- Try this for values between -17.78 and -12.2°C.
- We already have the probability of temperatures less than 0 in the first row.
- Get the probability for -12.2 in another row. Since these are just probabilities of values less than these temperatures or z-values, we can subtract to get the probabilities of being between these two limits, 0.510-0.054 = 0.456.
- Thus, there is a 45.6% probability of average January minimum temperatures between -17.78 and -12.2°C (Figure 8).



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Original Temperature | Mean | Standard Deviation | Z-Value | Probability | | |
| 2 | -17.78 | -12.305 | 3.398 | -1.611241907 | 0.053563504 | | |
| 3 | -21.4 | -12.305 | 3.398 | -2.676574456 | 0.003718952 | | |
| 4 | -12.2 | -12.305 | 3.398 | 0.03090053 | 0.512325566 | | |
| 5 | | | | | 0.458762063 | | |
| 6 | | | | | | | |
| 7 | | | | | | | |
| 8 | | | | | | | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| 16 | | | | | | | |
| 17 | | | | | | | |
| 18 | | | | | | | |
| 19 | | | | | | | |

Figure 8 The Excel file with different values applied.

# Confidence Interval for a Mean

## Computing a Confidence Interval

For a normal population with known Standard Deviation $\delta$, we can sample and compute a Confidence Interval for the Population Mean $\mu$.

Suppose we want to estimate the true yield difference for 2 corn hybrids. From data comparing many pairs of corn hybrids, we know that the differences are normally distributed, and we also have very good knowledge of the standard deviation of these differences. We have a sample of 16 locations where the 2 corn hybrids are both planted and take the difference in hybrid yields. Each difference is an individual from a population of all such differences. A 95% confidence interval for the true average difference in hybrid yields is centered on the average difference ($\bar{X}$) plus and minus 2 standard errors of the mean (σ/√n).

## Estimating a True Mean

We use properties of the normal distribution and sample averages to get this confidence interval. The reasoning is as follows.

Suppose we have a normal distribution with known standard deviation σ, but the true mean μ is not known, and we wish to estimate it.

1. We know from the properties of the normal distribution that 95% of the individuals differ from the true population mean by no more than 1.96 (about 2) standard deviations.
2. By the central limit theorem, 95% of sample averages will differ from the population mean by less than about two standard errors (σ/√n).
3. Therefore, we can estimate μ using the sample average, $\bar{Y}$, and an interval of 1.96 σ/√n above and below it.
4. We have 95% confidence that this procedure will give correct results, provided our assumptions are met.

### Study Question 4: Confidence Interval for a Mean

#### Query 1.3.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Null Hypothesis

## Testing a Hypothesis

Designing an experiment to test a hypothesis begins with a statement that can be falsified with experimental data, *i.e.*, the **Null Hypothesis**.

In the previous chapter, we used our knowledge of the normal distribution to examine how the high temperature in June of one year compared to the historical high temperature for that month. We wanted to know if the average high of 29.8°C differed from the historical mean of 27.6°C. To do this we calculated a z-value to determine how many standard deviations the observed value was from the historical mean. It is possible, however, to approach this question in another way. We could ask "is the high temperature observed for June of this year significantly different from the historical mean high temperature in June?" Using our knowledge of the normal distribution we can perform a statistical test to answer this question.

By convention, a statement stating no differences (null) is usually called a null hypothesis. In our example, the null hypothesis is that the observed high temperature is equal to the historical mean high temperature. This may be written symbolically as: $H_0$:$\mu = \mu_0$. The alternative hypothesis is that the means are not equal, or $H_\alpha$:$\mu \neq \mu_0$. Ultimately our experimental data will either reject the null hypothesis or support it.

## Errors Can Occur If the Null Is True or False

There are two types of errors that can occur when testing a null hypothesis (Table 1). The first type (I) is rejecting the null hypothesis when it is actually correct. The second type (II) is the opposite case; supporting the null hypothesis when it is incorrect. The probability of making a Type I error is called the significance level, denoted by the Greek letter alpha (α). A commonly used alpha level in agronomic experiments is 5%. This means that in one out of twenty tests a significant difference will be found where none actually exists. The probability of a Type II error is referred to as beta (ß) or the error of accepting the null hypothesis when it is not true. The errors are related; reducing the value of alpha increases the possibility of committing a Type II error. Choice of levels of alpha and beta depends on where you wish to assign the chance of error. In agronomic experiments, we are generally more concerned with controlling Type I errors than Type II errors because we consider the consequences of making a Type I error (saying there is a difference when there really is none) to be more serious. We will revisit this topic throughout the course.

*Table 1 Two types of errors in hypothesis testing.*

| | **$H_0$ is true** | **$H_0$ is false** |
|---|---|---|
| **$H_0$ rejected** | Type I | none |
| **$H_0$ not rejected** | None | Type II |

## Examining the Hypothesis

Returning to our example, we now need to determine the level of significance for our test. Since the consequences of our being wrong are not very great, let's use an alpha level of 20%. This means that we will declare any temperature value that occurs less than 20% of the time based upon the normal probability distribution to be different from the historical mean. The alternative hypothesis here is that the June high temperature differs from the mean high temperature. It could be above or below the true mean. This is called a "two-tailed" test because we reject $H_0$ if the June high temperature is either above or below the mean, *i.e.*, in either the upper tail or lower tail of the distribution.

Now, let us assume that we have a normal distribution of monthly high temperatures with a known standard deviation (1.6°C). We want to find z-values for which 20% of the observations will be in the tails. We need 10% in each tail. Using a table of z-values with P = 10% in one tail, we find z = 1.2816. Since our calculated z-value 1.05 is less than 1.2816 we fail to reject the null hypothesis and conclude that the observed June high is not significantly different from the historical mean.

## Study Question 5: Null Hypothesis

### Query 1.3.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Here are the details on the two fields the farmer wishes to use. For simplicity, we assume each field has one soil type only (Table 2).

*Table 2 Characteristics of two fields.*

| **Soil characteristics** | **Amount of organic matter and minerals** |
|---|---|
| **Field 1 16 hectares** | |
| Sharpsburg – silty clay loam | moderate organic matter |
| 9-14% slope | low subsoil P |
| moderately well drained | medium subsoil K |
| surface layer depth is 8-18cm | n/a |
| subsoil layer depth is 122cm | n/a |
| **Field 2 16 hectares** | |
| Macksburg – silty clay loam | high organic matter |
| 0-2% slope | low subsoil P |
| somewhat poorly drained | medium subsoil K |
| surface layer depth is 61cm | n/a |
| subsoil layer depth is 140cm | n/a |

**Option 1**

The farmer decides to plant Field 1 with the new variety and Field 2 with the old variety. When he harvests, he finds the following results (Table 3):

*Table 3 Results of experiment option 1.*

| **Field 1** | **Field 2** |
|---|---|
| New yield variety: 8780 kg/ha | Old yield variety: 9410 kg/ha |

**Option 2**

The farmer decides to plant Field 1 with the old variety and Field 2 with the new variety. When he harvests he finds the following results (Table 4):

*Table 4 Results of experiment option 2.*

| **Field 1** | **Field 2** |
|---|---|

| | |
|---|---|
| Old yield variety: 7530 kg/ha | New yield variety: 10660 kg/ha |

**Option 3**

Realizing that there is a soil fertility difference between the fields, he decides the plant half of each field with each variety. This produces the following results (Table 5):

*Table 5 Results of experiment option 3.*

| Field 1 – Plot 1 | Field 1 – Plot 2 | Field 2 – Plot 1 | Field 2 – Plot 2 |
|---|---|---|---|
| Old yield variety: 7530 kg/ha | New yield variety: 8780 kg/ha | Old yield variety: 9410 kg/ha | New yield variety: 10660 kg/ha |

**Experiment Conclusions**

Usually, it is difficult to limit both Type I and Type II errors. The decision must be made on which error is least detrimental in the decision-making process. This usually includes economic factors in the decision-making. You can limit both errors by increasing the number of units in a sample.

## What Does It Mean To "Accept The Null Hypothesis?"

It only means that our data do not provide enough evidence to reject the null. It doesn't mean that the null hypothesis is definitely true. For example, two people form null hypotheses; person A says his null is that the true corn yield difference is zero, person B says the true difference is 0.1. An experiment is run, and we fail to reject the hypothesis of A. Does this mean he is right and B is wrong? Of course not. The difference might be small and just not detected in the experiment.

The null hypothesis is a "straw man" set up to be torn down by experimental evidence. We may just not have enough evidence to reject $H_0$. So we fail to reject $H_0$ even though it may not be the true state of nature. If you see the words "*accept the null hypothesis*", you should mentally translate to "*fail to reject the null hypothesis*".

# Testing a Hypothesis

## Soil Sampling Example 1

Let's take a different example, fertilizer recommendations. Return to our soil sampling example (below) from Chapter 2 on Distributions and Probability. You are sampling here to determine if potassium fertilizer is necessary. ISU recommendations are for soil K to be between 70 and 130 ppm. Usually, you are checking to see if there is sufficient K. You decide to apply fertilizer if the soil K is less than 130 ppm. Your hypothesis here is:

$H_0$: sampled soil potassium is greater than or equal to 130 ppm (no action is taken)

$H_a$: sampled soil potassium is less than 130 ppm (fertilizer must be applied)

This is chosen because your action to be taken is to apply fertilizer. You want to be assured that you need fertilizer before applying. Economically, though, K is not very expensive to apply. So, your criterion for significance is not very high. You are willing to risk a larger chance of a Type I error for a smaller Type II error. You are more willing to reject your null hypothesis when it is true (to apply fertilizer when none is needed) than to accept your null hypothesis when it is false (to not apply K even though it is needed).

## Soil Sampling Example 2

Click on the field to sample that location (square) (Fig, 9). Select sets of locations to calculate an average potassium value for the field. For example, a set with 300, 150, and 450 would result in a mean of 300.

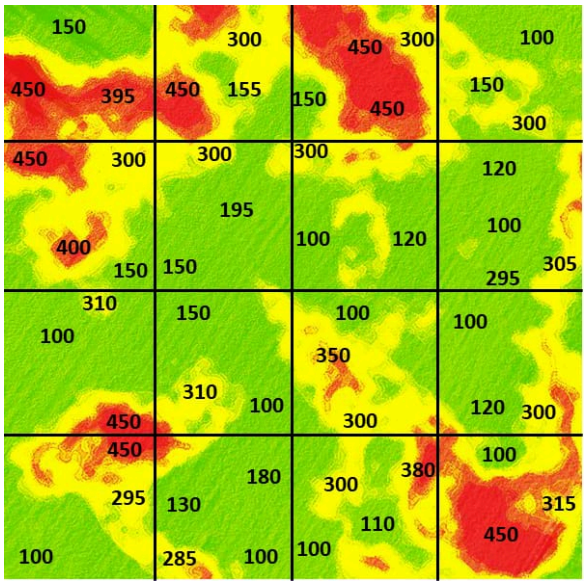


Figure 9 Potassium concentration variation across the field.

What does this suggest about how to design a sampling scheme to best represent the “population” mean value for the field? What happens when you increase the number of samples or vary their location? Would it be better to be random in your sampling scheme or systematic?

If you do this computation several times, pressing the Reset button between sets, you will notice that variation in the calculated average value occurs depending on where the samples are drawn.

## Sampling Exercise Application

What does this suggest about how to design a sampling scheme to best represent the “population” mean value for the field? What happens when you increase the number of samples or vary their location? Would it be better to be random in your sampling scheme or systematic?

If you do this computation several times, you will notice that variation in the calculated average value occurs depending on where the samples are drawn.

We see from this activity that it is difficult to get a representative sample for measuring an underlying parameter of the population, such as the average potassium concentration in an entire field. Taking three observations, and calculating a mean from them, better represents the true population mean than would an individual value. More than three observations would be even better.

It is generally better to take a random sample than a systematic one. Random samples provide a method to get unbiased estimates of population values. W.G. Cochran gives an example illustrating this on page 121 of his book Experimental Designs, co-authored with Gertrude Cox. Even experts have a bias when trying to select a representative sample.

In an experiment to measure the heights of wheat plants in England, several expert samplers chose what they thought were eight representative plants from each of six small areas containing about 80 plants. Every expert ended up choosing samples taller than the average of all the plants in the area. Of the 36 total samples, only 3 had shorter average wheat height than the corresponding area. Samplers averaged from 1.2 cm to nearly 7 cm over the actual height for systematic samples compared with the actual average for the six plots. It is likely that their eyes were drawn to the taller plants. A properly conducted random sampling scheme would have avoided this bias.

## Sampling Conclusions

Think about the samples you gathered when sampling the soil potassium situation in this example. What values did you get? The actual average of the field is somewhere near 225 ppm. However, at this point we do not know the variability in the samples to test for significance. If we had knowledge of the distribution and the true variances of soil K measurements in the field, we could use the normal distribution and test the hypothesis. But at this stage, we cannot test the value specifically because we do not

know whether we have a normal distribution with known standard deviation. We will discuss that more in a later unit.

## Hypothesis Testing

### Normal Distribution And Hypothesis Testing

Decisions on varieties to plant or whether to apply fertilizers are made many times by producers. They do not realize they are using a scientific method to make the decision, though. The actual method of making the decision proceeds using the steps below:

1. Clearly decide the population and set a definable null and alternative hypothesis.
2. Choose a significance level at which you are comfortable, i.e., where the possibility of random error will not cause you to make a poor choice. If you are reasonably comfortable that Type I error of 5% is acceptable, use it. If you wish to be more certain not to wrongly reject your null hypothesis, choose a smaller alpha. Be aware that doing this opens the door for accepting the null hypothesis when it is not true (Type II error).
3. Compute the statistic to be tested. How does the statistic compare with what you expect for the population if the null hypothesis is true?
4. Use the calculated statistic to accept or reject your null hypothesis.

When the population follows a normal distribution with known variance, you can test the hypothesis as in the June temperature example. If you are sampling from an approximately normal distribution, with unknown variance, then we will need to use a *t*-statistic, which is the subject of a later unit.

# Summary

### Sample averages

- From a normal distribution are also normally distributed.
- Have variance estimated by \(\dfrac
  ParseError: EOF expected (click for details)

  ```
  Callstack:
      at (Bookshelves/Agriculture_and_Horticulture/Quantitative_Methods_for_Plant_Breeding_(S
  ```

  {n}\).
- Variance of sample means is reduced by factor $\frac{1}{n}$.

### Central Limit Theorem

- Even sample means from non-normal populations become normally distributed as n gets large.
- This allows us to make probability statements (confidence intervals).

### Confidence Intervals

- A $95\%$ CI for $\mu$ is $Y \pm 1.96\frac{\sigma}{\sqrt{n}}$.
- A $99\%$ CI for $\mu$ is $Y \pm 2.58\frac{\sigma}{\sqrt{n}}$.

### Hypothesis Tests

- Start with null hypothesis (no treatment effect).
- Type I errors when true null is rejected.
- Type II errors when false null is accepted.
- Based on probability, we reject or we fail to reject $H_0$, and then draw conclusions.

## Reflection

The **Chapter Reflection** appears as the last "task" in each chapter. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the chapter and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this chapter.
2. What is the most valuable concept that you learned from the chapter? Why is this concept valuable to you?
3. What concepts in the chapter are still unclear/the least clear to you?

**How to cite this chapter:** Mowers, R, D. Todey, K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Central Limit Theorem, Confidence Intervals, and Hypothesis Tests. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Dennis Todey; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.4: Categorical Data - Binary

Some types of data are measured in discrete units. For example, we might count the number of insects on a plant or the number of plants in a segregating population with a transgene.

In this chapter, we will concentrate on binomial data, which are characterized by having only two possible states, for example, germinated or not germinated in a seed germination experiment. Other examples are cornstalks which are either stalk lodged or not, plants diseased or not, and survey data in which farmers either agree or disagree with an issue.

**Learning Objectives**

- To recognize the binomial situation.
- To be able to use the formula for binomial probability.
- To compute the mean and variance for a binomial distribution and use Excel to compute probabilities for events.
- To use the normal approximation to the binomial to compute probabilities for number of successes, and to form confidence intervals and hypothesis tests for a proportion.
- To estimate confidence intervals and test hypotheses for differences in proportions for independent samples from two populations.

## Definition of Binomial

A fixed number of independent trials, each with two outcomes and constant proportion of success, follow the **binomial distribution**. Not all data we wish to analyze are from a normal distribution. As we saw in unit one, some variables have discrete values. For this unit and the next, we analyze data which are discrete or categorical.

The first type we analyze are data which arise from a Bernoulli trial, *i.e.*, two possible outcomes. They are characterized by four requirements: two outcomes for each trial, a fixed number of trials, independent trials, and a constant probability of success on each trial. These types of trials give rise to the binomial distribution. The true probability of success, designated $p$ is one parameter of the binomial distribution. The probability of failure is $q = 1 - p$. For example, in a seed germination experiment, we only have two outcomes for each seed: either it germinates or fails to germinate.

The use of $p$ to represent the proportion of successes does not adhere to the usual convention of using Greek letters for parameters. Some texts use the symbol $\pi$ for the parameter, but we will use $p$ because it is widely used in population and quantitative genetics. Do not confuse this $p$ with the use of P as a probability statement. Your discernment will need to be based on context rather than memorization.

## Another Binomial Situation

For the binomial distribution, there is a fixed number ($n$) of independent trials for which we count the number of successes. $n$ represents the second parameter of the binomial distribution. In germination tests, we often use $n = 100$ seeds. These trials are assumed to be independent; where if one seed germinates, it does not affect whether another seed will germinate. We also assume there is a constant true germination proportion ($p$) for any seed in the seed lot.

Another example of a binomial situation is root lodging counts in corn yield trial plots. There are 60 plants per plot planted with a precision planter, and each plant can be considered an independent trial with two possible outcomes, root lodged or not. We assume that there is a constant genetic proportion of a root lodged plant (“success”) for any of the individual plants (trials). The inherent genetic susceptibility to root lodging is considered a characteristic of a corn hybrid, and we want to estimate the proportion, $p$, which will root lodge. We count the number of lodged plants per plot and use those data for determining differences among corn hybrids.

To determine if you are in a binomial situation, there are four requirements:

1. There should be two outcomes for each trial (S or F, 0 or 1, etc.).

2. There should be a fixed number of trials.
3. Each trial should be independent of the others.
4. There is a true proportion of success, $p$, which is the same for each observation or trial.

## Study Questions 1: Definition of Binomial

### Query 1.4.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



### Query 1.4.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



### Query 1.4.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

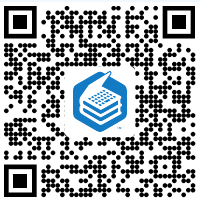
View this Interactive Content online



### Query 1.4.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



### Query 1.4.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



## Study Questions 2: Definition of Binomial

### Query 1.4.6

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Study Questions 3: Definition of Binomial

### Query 1.4.7

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Query 1.4.8

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Query 1.4.9

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Query 1.4.10

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.4.11

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Assumption of Independence

You might wonder how often the assumption of independence is violated in the same way as our example of drawing cards from a deck without replacement. For example, if a seed salesman wishes to survey 40 of his 650 customers with a yes-no question, does he violate the assumptions necessary for using the binomial distribution?

In general, if our sample is 10%, or less, than the size of the target population, and all other assumptions are met, we can use the binomial distribution results. Otherwise, we need to adjust the standard errors using a finite population correction factor, an adjustment found in some statistics textbooks, but not covered here.

Do not restrict your sample size just to be able to use the binomial distribution if, for example, it is necessary to sample, say 100 of the 650 salesmen to get precise results. Get the sample size adequate for the precision you need, and then, if needed, get help from a statistician to analyze the data.

## Discussion

Do you think that all the assumptions of a binomial situation are valid for the number of lodged plants per plot? Why or why not?

# The Binomial Probability Function

## Calculate Probabilities

A formula allows us to calculate probabilities for binomial data. Generally, we are interested in two types of questions for a binomial experiment. One is the number of successes in the n independent trials. We might want to know how often the count of the number of "successes" is greater than a given value. For example, if the count is the number of dead insects from a set of 25 corn rootworms fed on transgenic root tissue, we may want to know how often we observe fewer than 16 dead if the true proportion which die is $p$ = 0.9.

A second important question is how to estimate ***p***, the true proportion of successes. This parameter ***p*** is the true average number of successes divided by $n$. For example, if the true average for the entire population is that 21 of the 25 insects die when fed the tissue (in other words, for all possible sets of 25 test insects fed this tissue, an average of 21 will die), the true proportion of success is $p$ = 0.84. We will see in the next section how to estimate p from a sample.

The formula for answering the first question is as follows. If **s** is the number of successes in **n** independent trials for the binomial situation, the probability that **s** is a given value **k** is:

$$P_{(s=k)} = \frac{n!}{k!(n-k)!} p^k q^{(n-k)} \tag{1.4.1}$$

where

- $n$ = number of trials,
- $q$ = 1-p,
- $k$ = any integer between zero and n, representing the number of successes in the trial.

Recall that a factorial (denoted by !) is calculated as: n!=n*(n-1)*(n-2)*...2*1. How do we use this formula? Suppose we sample 10 plants, each with true probability 0.25 of containing a Bt gene. Leaf samples from each of the plants are evaluated with an error-free diagnostic test for the gene. Assume these plants are a random sample of a large number of plants. What is the probability that we get 2 of 10 samples that are positive for the gene?

First, think, "Is this a binomial situation?" There are two outcomes: either the gene is present in a plant or not. There are 10 independent trials because if any plant has the gene, that does not affect whether another plant does. We can also assume that there is constant genetic proportion of plants with the gene, p = 0.25. This does fit the binomial situation.

To calculate the probability that two plants will have the gene, substitute into the formula to find:

$$P_{(s=2)} = \frac{10!}{2!(10-2)!}0.25^2 * 0.75^{(10-2)} \tag{1.4.2}$$

$$P_{(s=2)} = \frac{10!}{2!(8)!}0.25^2 0.75^{(8)} \tag{1.4.3}$$

$$P_{(s=2)} = \frac{10*9*8*7*6*5*4*3*2*1}{2*1(8*7*6*5*4*3*2*1)}0.25^2 0.75^{(8)} \tag{1.4.4}$$

This is the probability of getting 2 positives.

## Study Questions 4 and 5: Binomial Probability Function

### Query 1.4.12

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Query 1.4.13

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Try This! Probability Exercises

**Ex. 1: Binomial Probabilities (1)**

In this exercise, we use Excel to get probabilities for the number of successes as shown in Equation 1. We will also illustrate the difference between the Binomial Distribution and Binomial Probability functions.

We will create a table of probabilities for an experiment with n=10 independent trials and p=0.25. We will get probabilities for each number of successes possible for this experiment, 0, 1, . . . , 10. We will also get the binomial distribution cumulative probabilities for numbers of successes less than or equal to each of these values.

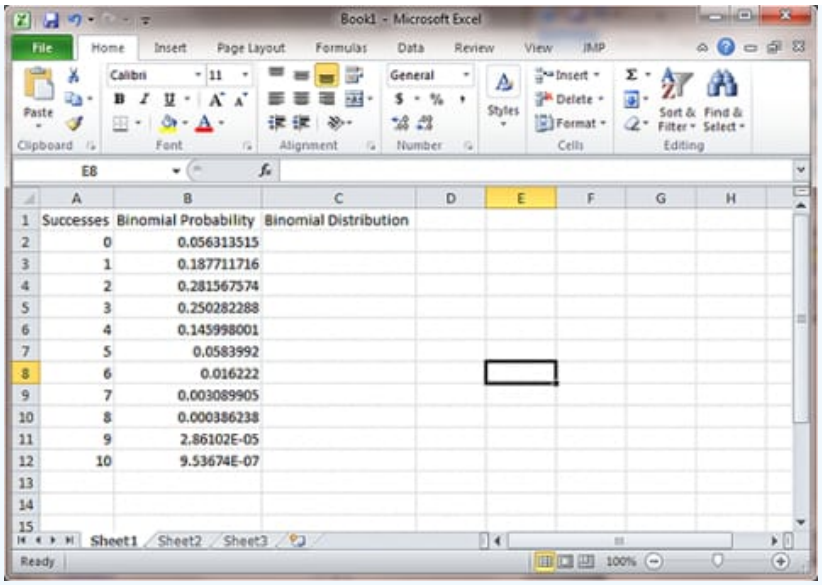
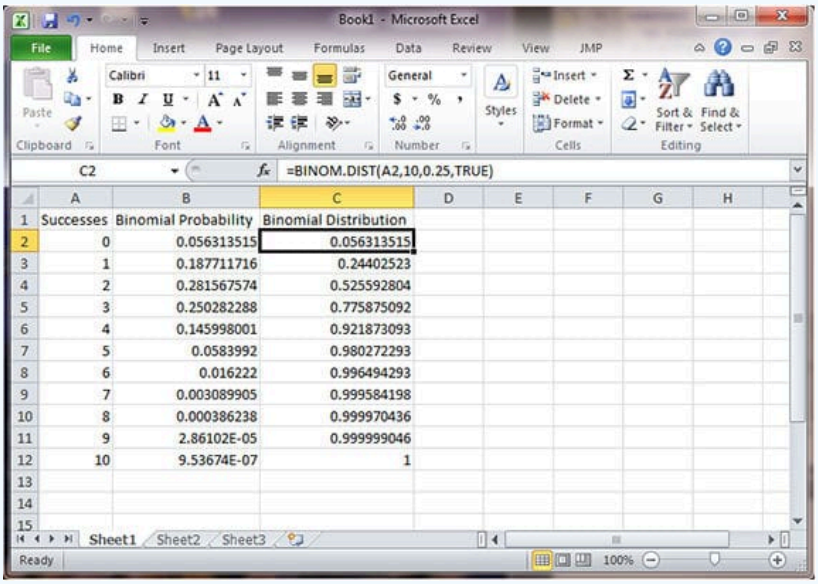



Figure 1 Excel table of binomial probabilities.

1. Open a new Excel workbook and label three columns: Successes, Binomial Probability and Binomial Distribution. Under Successes, fill in successive integers 0-10 (Figure 1).
2. Enter the formula "=Binom.dist(A2, 10, 0.25, False)" in the cell next to zero under Binomial Probability. A2 refers to the cell with the number of successes, 10 is the number of trials, p=0.25, and False indicates that the Probability Mass Function should be used. This means that Excel returns the probability of each success in Column A. If we instead set the last parameter of "=Binom.dist" to True, Excel calculates the combined probability of all successes of s or lower. For example, the probability of 2 successes would also include the probability of 1 or 0 successes. Now, copy that formula into the next ten cells in the Binomial Probability column.

**Ex. 1: Binomial Probabilities (2)**




3. This gives the probability for each possible number of successes (k). For example, k=2 successes has 0.2816 probability of occurring.
4. To get the cumulative probabilities, for the third column, enter the formula "=Binom.dist(A2, 10, 0.25, TRUE)".

This gives the table to the right. Notice that the probability for 2 or fewer successes is 0.5256, as we computed in Study Question 4.

### Ex. 2: Table of Probabilities

In this exercise we use Excel to reconstruct a column of the table of probabilities. For this we use p = 0.4, n=6, and the number of successes as shown in the middle column of the table (Figure 2).

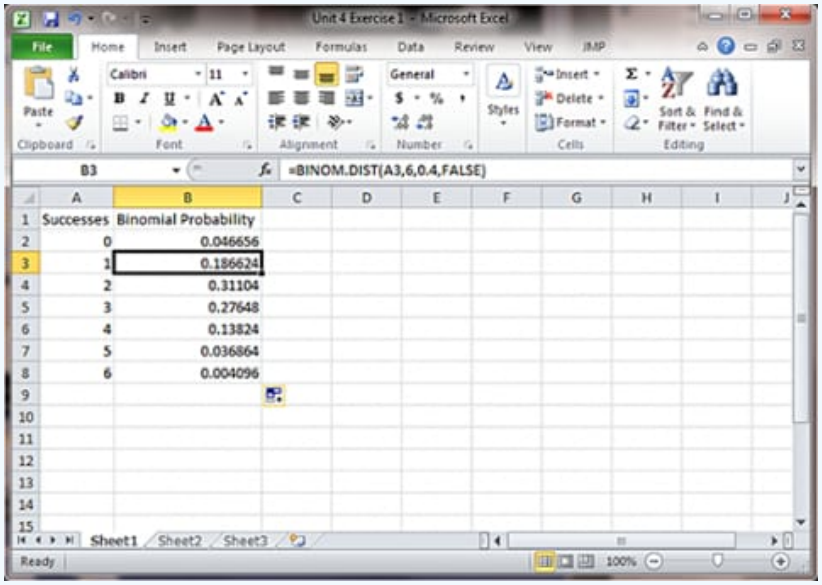



Figure 2 Excel table of binomial probabilities.

1. Open a new Excel workbook. Follow the same steps for Binomial Probability, but the number of successes is 6 and p=0.4. The formula to use is “=Binom.dist(A2, 6, 0.4, FALSE)”.
2. This gives the probability for each possible number of successes (k). For example, k=3 successes has 0.2765 probability of occurring.

### Ex. 3: Cumulative Binomial Probability

In this exercise, we use Excel to solve a probability question. The question asks for the probability of 7 or fewer successes in 15 trials with p=0.4. Because we need a cumulative probability, we use the Binomial Distribution function with p = 0.4, n=15, and k=7 (Figure 3).

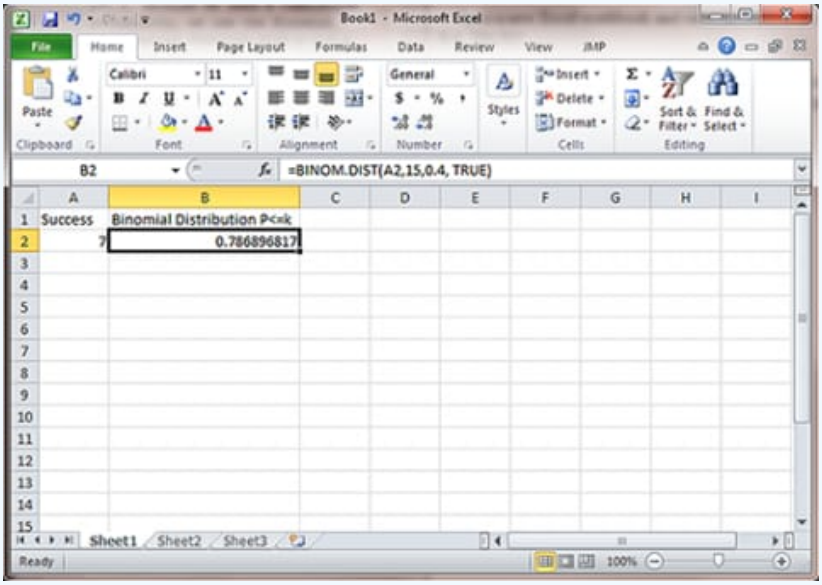



Figure 3 Excel workbook with number of successes entered as 7.

Open a new Excel workbook and enter the number of successes as 7. This exercise calculates the cumulative probability, so the formulas “=Binom.dist(A2, 15, 0.4, TRUE)”

This gives the probability for each possible number of successes (k). For our example, k=7 or fewer successes has 0.7869 probability of occurring. Notice that the Excel Binomial Distribution function gives the cumulative probability based on a binomial distribution rather than a normal approximation (0.7852).

**Ex. 4: Probability Computations**

In this exercise, we use Excel to exactly solve a more complicated probability question. This question asks for the probability of between 4 and 6 successes in 10 trials with p=0.25. The probability of between 4 and 6 successes is the probability of 6 or fewer successes minus the probability of 3 or fewer successes. Because we need to use cumulative probabilities, we use the Binomial Distribution function with p = 0.25, n=10, and k=3 or 6 (Figure 4).

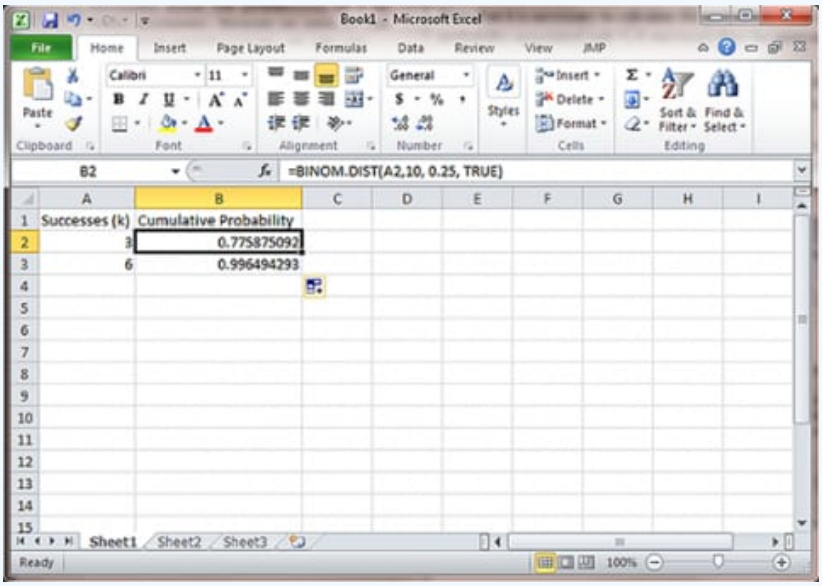



Figure 4 The finished Excel cumulative probability table.

This function gives, in the second column, the cumulative probability for each possible number of successes (k). For our problem, k=3 or fewer successes has 0.7759 probability of occurring, and k=6 or fewer successes has probability 0.9965 of occurring. We subtract these, P(k<=6) – P(k<=3) or 0.9965 – 0.7759 = 0.2206, the probability of between 4 and 6 successes occurring.

# The Mean and Variance

## Mean and Variance

The number of successes $s$ for a binomial distribution has mean $np$ and variance $npq$

For number of successes, s:
Mean = $np$
Variance = $npq$

The sample proportion in a binomial distribution, $\hat{p} = \frac{s}{n}$, has mean and variance:

For the sample proportion, $\hat{p}$:
Mean = $p$
Variance = $\frac{pq}{n}$.

Suppose we have binary data for root lodging with true proportion lodged, p = 0.08 and n = 60 plants per plot. The true average

number of root lodged plants in a plot is 4.8 and the variance is 4.4. In other words, we expect to count about five lodged plants per plot, and have standard deviation of about two plants per plot (std = sqrt(4.4) = 2.09).

## Standard Deviations

One problem not yet discussed with the example of root lodging is that although the variable itself may be considered to follow a binomial distribution with constant genetic proportion of lodged plants, there are other sources of variation. Field, disease, weather-related, and other variation will add to the variation of root lodging in a real situation. Our standard deviations are likely to be much greater when you consider mistakes in counting, the variation of soils in the yield trial field area, and the variations in thunderstorm wind speed throughout the yield trial location. The binomial variance gives us the inherent variation in lodging counts, but there are other sources of error that will inflate the estimated variance of a population.

The variance for the sample proportion can be computed if we know p, and the maximum variance is when $p$ = 0.50. As an example, suppose our objective is to estimate the germination percentage from a sample of 100 seeds. If the true germination proportion for the seed lot is 0.95, what is the variance for $\hat{p}$, our sample estimator of $\hat{p}$? The variance of $p$ is 0.95*0.05/100 = 0.000475. The standard deviation is 0.022, or 2.2% germination. If the true proportion is 0.80 and we have 100 seeds, the variance is 0.80*0.20/100 = 0.0016, and the standard deviation is 0.04, or 4% germination. The maximum variance of $p$ will occur when $\hat{p}$ = q = 0.50, and is 0.50*0.50/100 = 0.0025. The standard deviation is 0.05, or 5% germination.

### Study Questions 6: Mean and Variance for a Binomial Variable

#### Query 1.4.14

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

#### Query 1.4.15

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

#### Query 1.4.16

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Estimating Trial Numbers

The previous study question illustrates a method for finding the number of trials, $n$, needed to achieve a certain level of precision

for estimating $p$ in a binomial situation. We know the maximum variance will occur when $p$ = 0.5, and we can solve to get:

$$n = \frac{pq}{v} \tag{1.4.5}$$

**where:**
$n$ = minimum number of trials,
$p$ = proportion of successes,
$q$ = 1-p,
$v$ = variance.

This is a conservative estimate because we are using the maximum variance of our estimator. If we knew more about the true proportion p, for example, 0.30, we would use the formula $n$ = 0.3*0.7/ $v$.

# The Normal Approximation

## Approximate the Binomial

**If sample sizes are fairly large (np and nq $\geq$ 5), we can use the normal distribution to approximate the binomial.**

It is an interesting phenomenon that the histogram for the number of successes in a binomial distribution looks like the normal distribution, especially if $n$ is large and $p$ is not too close to either zero or one. In fact, the normal distribution can be used as an approximation to the binomial.

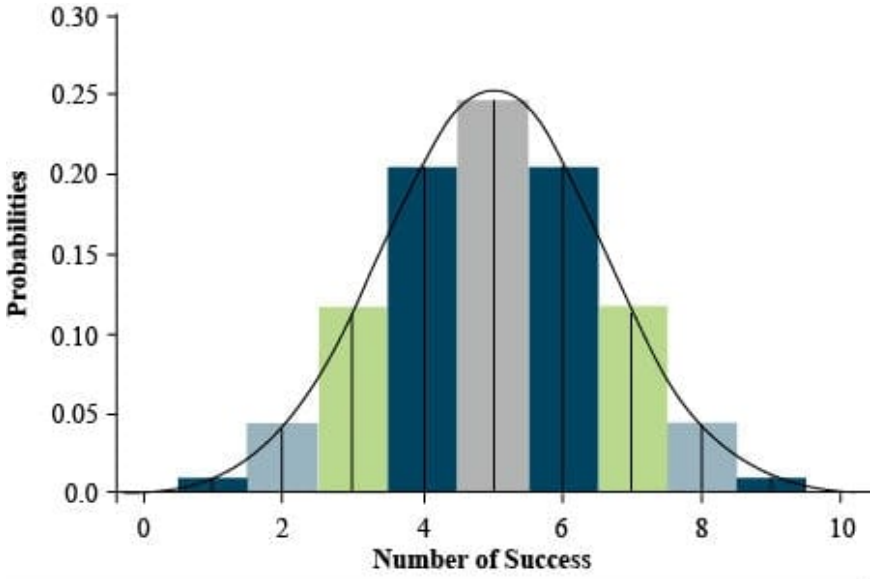



Figure 5 Illustration of correspondence of a binomial distribution (p = 0.5 and n = 10) to the normal.

Figure 5 shows the correspondence of the histogram of a binomial distribution to the frequency curve of the normal distribution. This is for a binomial distribution with n = 10 trials and p = 0.5. For this case, the normal curve passes very closely to the center of each bar of the binomial histogram. Even if p is not 0.5, if n is large, the normal curve can approximate the binomial.

## Normal Approximation

We see from Figure 5 that probabilities under the normal curve will better approximate the histogram for the binomial distribution if we measure the area for s values less than $k$ + 0.5. For example, to sum the area of 0, 1, and 2 successes, we would add the probabilities (areas of bars) for binomial, and this area is better approximated by the area under the normal curve less than s = 2.5. If we just use the probability for the normal less than k, we would miss half the bar representing the probability in the binomial histogram. Therefore, $P_{Binomial}(s \leq k)$ is better approximated by the normal distribution using $P_{Normal}(s < k + 0.5)$. This correction factor is used when the normal approximation can be used, but sample sizes are small.

The general rule to know when to use the normal approximation is to use it for the number of successes in a binomial distribution when the mean ($np$) is not too close to zero or one. Specifically, we use the normal approximation when ($np \geq 5$) and ($nq \geq 5$).

You might wonder why we should use the normal approximation at all when we can use computer programs to compute exact probabilities for the binomial distribution. With the normal approximation, we can sometimes do confidence intervals much more easily, for example using the mean plus or minus two standard deviations for an approximate 95% confidence interval. We will also see that we can employ a normal approximation for testing hypotheses about proportions or comparing proportions from independent samples.

### Study Question 8: Normal Approximation to the Binomial

#### Query 1.4.17

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Conclusions from the Example

In this example we have a case that requires us to use the exact binomial distribution. The value of $p$ is so small that $np < 5$, and it is necessary to use the exact binomial distribution.

What can we conclude from the situation in study question 8 if we ran our test and found the 600-seed sample to be negative (no transgenic event present)?

Using Equation 1, we calculate that the exact probability for no positives in a sample of 600 is:

$$P_0 = (1)(0.001)^0(0.999)^{600} = (0.999)^{600} = 0.549$$

The value is not at all unusual if our null hypothesis is $H_0$: $p = 0.001$. Our sample is in concert with this null hypothesis. However, if our null hypothesis is that $p$ is 0.005, the probability of observing no transgenic event in a sample of 600 is $(0.995)^{600} = 0.049$. Thus, it is unlikely that the amount of contamination is as high as 0.005 (a half percent).

We have used our rule of thumb ($np$ and $nq \geq 5$ for normal approximation) to dictate that we need the exact binomial distribution. However, we also would not use the normal approximation in problems where there are severe consequences if we get the probabilities wrong.

### Study Question 9: Normal Approximation to the Binomial

#### Query 1.4.18

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Computing a Probability

If we germinate 400 seeds and the true percent germination is 95%, we can use the book's Appendix 1 table to compute the probability of observing a sample with less than 93% germination. We only use this example as an illustration because Excel can compute the probability more accurately with the binomial distribution.

Compute the mean and standard deviation as

$$np = 400(0.95) = 380, and \sqrt{npq} = \sqrt{4000 * 950 * 0.05} = 436 \tag{1.4.6}$$

Observing 93% or fewer is observing 0.93*400 = 372 or less. Then, the probability of 93% or less in our sample has $Z = \frac{372+0.5-380}{4.36} = -1.72$, and using the

normal approximation and Appendix 1, $P(Z < -1.72) = P(Z > 1.72) = 1.00 - 0.9573 = 0.043$.

## Sample Exercises

This example is somewhat complicated, so we illustrate how to do the same example with Excel.

Another example of how to use the normal approximation to compute probabilities is as follows. Suppose we run a germination test using 200 seeds, and assume the true p is 0.91. What is the probability that between 174 and 190 of the seeds germinate?

We can use the normal approximation because *np* = 182 and *nq* = 18. Our general method for computing the probabilities is to first draw a curve with the mean and standard deviation of the normal distribution. The mean is *np* = 182, and the standard deviation is *√(npq)* = 4.047. From the normal approximation, about 68% of the values should be between 178 and 186, and about 95% are between 174 and 190. We see in the next 'Try This' how to get the probability, but we can just estimate it. Notice that 174 is about 2 std below the mean, 190 is 2 std above the mean, and so the probability is about 0.95.

### Try This: Use Excel to Compute the Binomial Probability

Type your exercises here.

- First
- Second

### **Try This:** Use Excel to Compute the Binomial Probability

**Ex. 5: Compute the Binomial Probability**

In this exercise we use Excel to exactly solve the probability of 93% or fewer seeds germinating in 400 trials with p=0.95. Because we need a cumulative probability, we use the Binomial Distribution function with p=0.95, n=400, and k=372. The value for k is from 93% of 400, or 0.93*400 = 372.

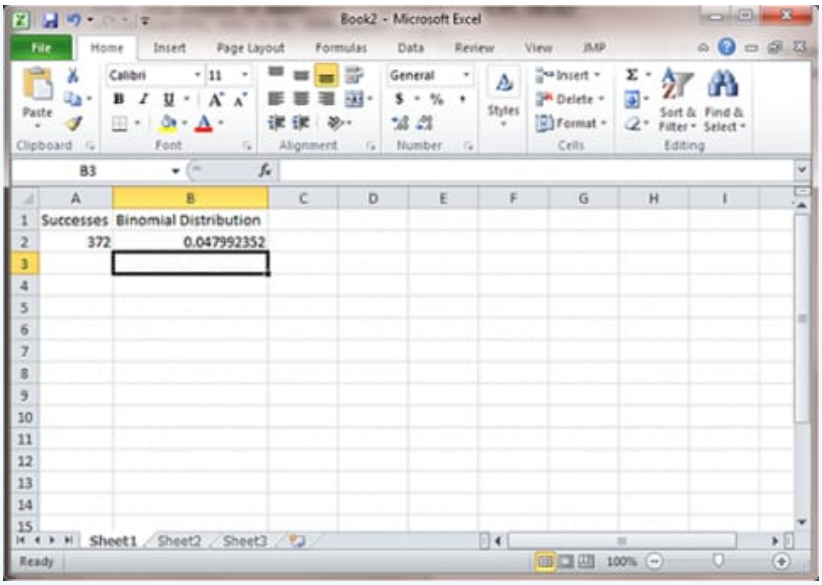



Figure 6 The Excel file with Successes at 372 and binomial distribution 0.047992352

- Open Excel and enter the formula for a cumulative probability since the statistic of interest is 93% or fewer seeds. This formula is “=binom.dist(372, 400, 0.95, TRUE)”, (Figure 6).

This gives the cumulative probability for each possible number of successes (k) from 0 to 372. For our example, k=372 or fewer successes has 0.048 probability of occurring .

**Ex. 6: Compute the Binomial Probability**

In this exercise, we use Excel to exactly solve a more complicated probability question. This question asks for the probability of between 174 and 190 successes in 200 trials with p=0.91. The probability of between 174 and 190 successes is the probability of 190 or fewer successes minus the probability of 173 or fewer successes. Because we need to use cumulative probabilities, we use the Binomial Distribution function with p = 0.91, n=200, and k = 173 or 190. Note the use of 173 rather than 174 to accommodate the entire interval (Figure 7).

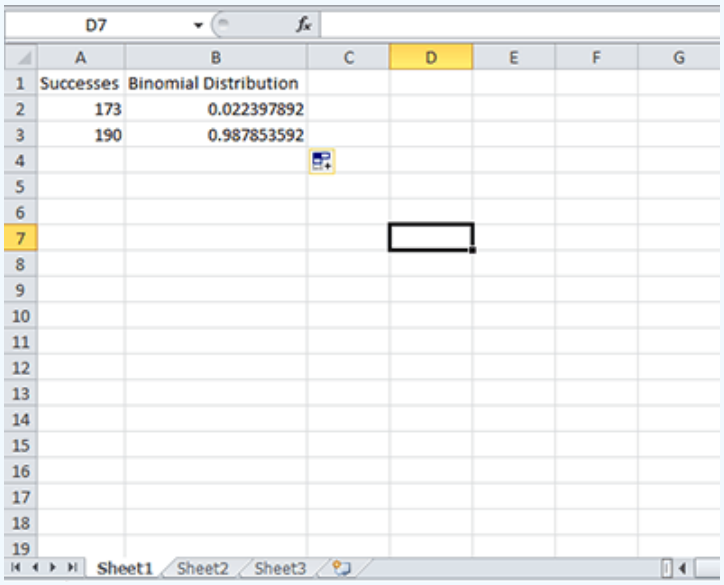



Figure 7 The Excel file with two columns: Successes and Binomial Distribution.

- This exercise uses the same technique as Exercise 4. Change the function so that it reflects p, n, and k for this problem.

This function gives, in the second column, the cumulative probability for each possible number of successes (k). For our problem, k=173 or fewer successes has 0.0224 probability of occurring, and k=190 or fewer successes has 0.9879 probability of occurring. We subtract these, P(k ≤ 190) – P(k ≤ 173) or 0.9879 – 0.0224 = 0.9655, the probability of between 174 and 190 successes occurring.

### Study Questions 10 and 11: Normal Approximation to the Binomial

#### Query 1.4.19

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

#### Query 1.4.20

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Reason to Use Normal Approximation

The reason we can use the normal approximation when $p$ = 0.5, but not when $p$ = 0.3 is that when $p$ = 0.3, the binomial distribution is skewed. Even though our sample size is fairly small (12), if $p$ = 0.5, the binomial distribution is symmetric and is better approximated by the symmetric normal distribution than when $p$ = 0.3.

# Confidence Intervals

## Sample Proportion

The normal approximation allows us to compute confidence intervals for $p$.

If we have a normal distribution, a 95% confidence interval will be centered on the calculated average plus and minus two standard deviations. Note: A confidence interval is not the same as a confidence limit. The confidence interval contains a specified proportion of the distribution (e.g. 95%). The confidence limits are the endpoints of the confidence interval.

The sample proportion has true mean $p$ and true variance $pq/n$. A 95% confidence interval for $p$ is:

$$\hat{p} = \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \tag{1.4.7}$$

**where:**
$\hat{p}$ = sample proportion,
$\hat{q}$ = (1-p),
$1.96$ = number of standard deviations for a 95% confidence interval.

## Study Questions 12 and 13: Confidence Intervals for P

### Query 1.4.21

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

### Query 1.4.22

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Confidence Interval Exercise

A confidence interval based on the approximation is easier to compute than the exact method, which is given in the 'Try This!' below. However, in some cases we need to use the method based on the exact binomial distribution, for example in computing a confidence interval for a proportion of plants contaminated with a genetically modified organism. Also, note that "exact" refers to the use of the binomial distribution rather than its normal approximation. We do not know "exactly" the value of the parameter *p*, but just provide an interval and have 95% confidence in the procedure to calculate it.

## Try This: Use Excel and the Exact Binomial Distribution to Compute the Confidence Interval for p

**Ex. 7: Confidence Interval for p**

In this exercise we use Excel to solve the problem of finding a confidence interval. The problem is to find a 95% confidence interval for p when the sample of 20 has 4 successes.

We will do this by creating a table of probabilities from which we will find 0.025 probability in each tail, then find the values of proportions (p) corresponding to each of these Binomial distribution probabilities. We know two parts of the binomial formula, n=20 and k=4. We need to find values for p. We start by creating a table with 1000 potential values for p, from 0.001 to 1.000.

- Open Excel and label 3 columns Proportion (p), Probability for p-upper, and Probability for p-lower.
- Under proportion fill in the proportions starting with 0.001 to 1 by thousandths (i.e. 0.001, 0.002, 0.003,.....,1).
- Under 'Probability for p-upper' enter the formula "=binom.dist(4, 20, A2, TRUE)". Under 'Probability for p-lower' enter the formula "=1-binom.dist(4, 20, A2, TRUE)".
- Fill in the columns so that there are two probabilities for all 1000 values of p.
- Find the Probability that is closest to 0.025 without going over. This is found by dividing alpha = 0.05 by two for a two-sided test. If a value that is larger than 0.025 is selected, the interval will be too small.
- You should find the interval (0.086, 0.437).
- Use the link on the right to check your work.

Exercise 7 Solution

# Testing Hypotheses

## Testing for a Proportion

We can test hypotheses for a proportion using the normal approximation.

We wish to test the null hypothesis $H_0 : p = p_0$. We can do this with the normal approximation to the binomial. We can also do this with a more exact method based on the binomial distribution itself.

The test statistic for a large sample test is:

$$\frac{\hat{p} - \hat{q}}{\sqrt{\frac{p_0 q_0}{n}}} \tag{1.4.8}$$

Here $p_0$ is the hypothesized value of $p$, $\hat{p}$ is the estimate from the sample, and $q_0$ is (1 – $p_0$).

### Try This 1: Use Excel to Test a Hypothesis for p

**Ex. 8: Test the Hypothesis for p**

For this example, we test whether 112 successes (purple-stemmed plants) of 158 total could fit the hypothesis of p=0.75. In the sample, there are 112 purple and 46 green to make the 158 total. We want to know if they could fit the ratio 3:1, or 0.75 purple. We can use Equation 8.

The hypothesis tested is:

- $H_0$: p = 0.75 and q = 0.25
- $H_\alpha$: p ≠ 0.75 and q ≠ 0.25
- This hypothesis can be tested using a confidence interval. If p=0.75 falls within the confidence interval, we fail to reject the null. If it does not fall within the CI, we reject the null hypothesis.

**Steps:**

- Open Excel and enter the formula “=binom.dist(112, 158, 0.75, TRUE)”.
- Equation 8 covered finding a 95% confidence interval with the upper and lower values being associated with a probability of at most 0.025. If the probability associated with 112 purple in a sample of 158 and p=0.75 is greater than 0.025, then p=0.75 is within the confidence interval and we fail to reject the null hypothesis.

The probability is > 0.13 and we fail to reject the null hypothesis (Figure 8).

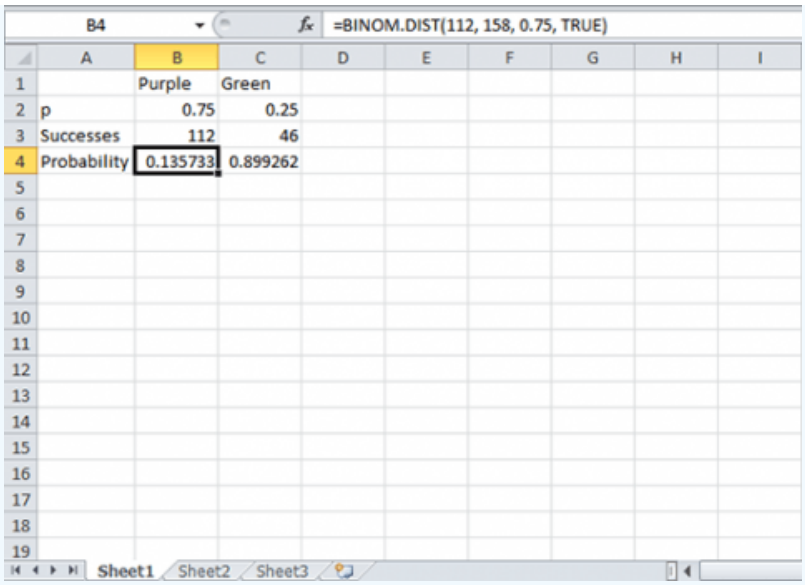



Figure 8 The Excel probability table with columns for purple and green seeds.

Notice that if you calculate the exact confidence interval for proportion purple (0.638, 0.779) it is the same as presented in the book. The confidence interval does include 0.75, and we fail to reject the null that p=0.75.

## Try This 2: Use Excel to Test a Hypothesis for p

**Ex. 9: Test the Hypothesis for p**

For this example, we test whether 7 successes (cuttings which root) of 10 total could fit the hypothesis of p ≥ 0.9 (Figure 9). We test the null that p ≥ 0.9 vs. the alternative that p is less than 0.9. This is a one-sided test, so the alpha level is 0.05, and is not divided by 2.

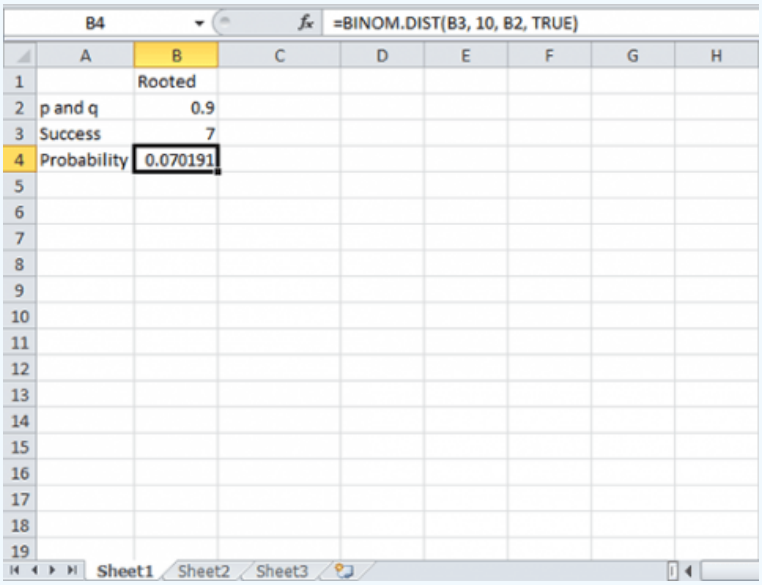



Figure 9 Excel file with values for p and q, success and probability.

This is a one-sided hypothesis.

- $H_0$: p ≥ 0.9
- $H_\alpha$: p < 0.9

- Open Excel and a new workbook.
- Enter the information from the image to the right.

Notice that the 90% confidence interval for proportion rooting is (0.493, 0.913). The probability value for the hypothesis test is P = 0.0702, which is not less than 0.05, so we fail to reject the null.

## Comparing Proportions

In the same way that we can use the normal distribution to approximate the binomial for a single population, we can also approximate the binomial for two different populations. Thus, we can use the approximation for differences of proportions when we have two independent samples.

We can use this approximation to compute confidence intervals and test hypotheses for differences in proportions. The variance of a difference in two proportions is $p_1q_1/n_1 + p_2q_2/n_2$, which is just the sum of the variances from the two independent populations. Consequently, we compute a 95% confidence interval for $p_1 - p_2$ as:

$$(\hat{p_1} - \hat{p_2}) \pm 1.96\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}} \tag{1.4.9}$$

Here, $(\hat{q})_1 = 1 - (\hat{p})_1$

and $(\hat{q})_2 = 1 - (\hat{p})_2$

This method of estimating a confidence interval for differences in proportions relies on large sample sizes and proportions not near 0 or 1. We do not give an example of hypothesis tests for differences in proportions because those are better done with contingency tables in the next unit.

## Summary

**Recognize the Binomial Situation**

- Two outcomes.
- Fixed number of trials.
- Independent trials.
- Constant proportion of success

**Binomial Probability**

- Formula allows calculation.
- Excel computes probability or cumulative (Binomial Distribution).

**Mean and Variance for Successes**

- Mean = np, Variance = npq

**Mean and Variance for Sample Proportion**

- Mean = p, Variance = pq/n

**Normal Approximation to Binomial**

- Can use when np and nq ≥ 5.
- Helpful for approximate 95% Confidence Interval.

**Estimate Confidence Intervals**

- Exact Binomial method uses Excel.

**Tests of Hypotheses**

- For $p = p_0$ using Excel.
- For differences in proportions, see next chapter.

## Reflection

The **Chapter Reflection** appears as the last "task" in each chapter. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the chapter and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this chapter.
2. What is the most valuable concept that you learned from the chapter? Why is this concept valuable to you?
3. What concepts in the chapter are still unclear/the least clear to you?

**How to cite this chapter:** Mowers, R., K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Categorical Data: Binary. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.4: Categorical Data - Binary is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.5: Categorical Data Multivariate

In the previous chapter, we saw examples of data from the binomial distribution, in which there are two categories for classification. In there, we develop a method for categorical data that occur in multiple categories. Discrete distributions which can have more than two categories (Figure 1) are referred to as multinomial distributions, and their analyses often employ the Chi-squared, $\chi^2$, method.



Figure 1 Marigolds exhibit differing coloration. Photo by Loz Pycock; licensed under CC-SA 2.0 via Wikimedia Commons.

**Learning Objectives**

- Understand how to use the $\chi^2$ test to evaluate discrete distributions of data.
- Be able to categorize data using contingency tables.
- Learn how to conduct statistical tests of differences among proportions, of independence, and of heterogeneity of discrete data sets.

## Chi-Square Testing

**Chi-square ($\chi^2$) tests may be used to analyze counts in categorical data.** Counts of data or data categorized on the basis of qualitative characteristics may be evaluated for significant differences using the chi-square test.

For example, during a recent growing season, 17 days had precipitation of 0-10mm, 9 days had precipitation of 10-20mm, 6 days had precipitation of 20-30mm, and 5 days had greater than 30mm of rain. Individually, these data may not be of much use. But using the chi-square test, you can determine if this distribution is significantly different from that expected based on the historical record.

A question you might ask would be: Did we have more days of heavier rain (> 30mm) or fewer days of light rain (0-10mm)? Often this is a more meaningful way of classifying rainfall than total rainfall during a season.

### Evaluation

Such data need to be evaluated to determine if differences in counts of the data are significant. Analyses of such data are done as analyses of counts. The test often applied to these is the $\chi^2$ test.

$$\chi^2 \tag{1.5.1}$$

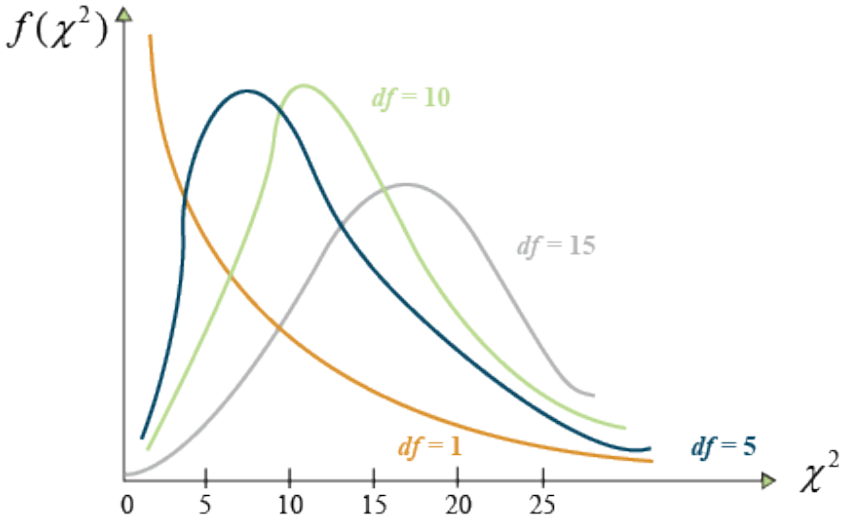



Figure 2 Values of the chi-distribution for various degrees of freedom.

In Figure 2, the number on the Y-axis is the probability that the$\chi^2$ value indicated on the X-axis occurs. The probability of significance is determined by the area under the graph, usually to the right of a certain number. For example, the probability of a $\chi^2$ value of 25 or larger of occurring.

Notice that as the number of df becomes larger, the $\chi^2$ distribution more closely approaches the normal distribution to which it is related. At lower degrees of freedom, the $\chi^2$ distribution is skewed, with lower values having a higher probability of occurrence. At higher degrees of freedom (>15), the distribution closely approximates a normal distribution.

## Formula

The chi-square method performs the analysis using counts. It estimates the difference in the observed number counted from the number expected. The value of the test on the $\chi^2$ distribution is determined by its closeness to the expected values.

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{1.5.2}$$

The summed terms comprise a $\chi^2$ value which occurs somewhere on the graph.

The farther to the right on the graph the value occurs, the more the data deviate from expected. A critical value is determined for a given probability level (Figure 2).

## Evaluating Hypotheses

For a significance level of 0.05, a $\chi^2$ value to the right of the critical $\chi^2$ value would have a 5% chance of occurring. Thus, the $\chi^2$ test would be significant at the 0.05 level. The interpretation is that the data deviated enough from expected to produce the large $\chi^2$ value in the analysis. Data to the left of critical $\chi^2$ value line are not significant (Figure 3).

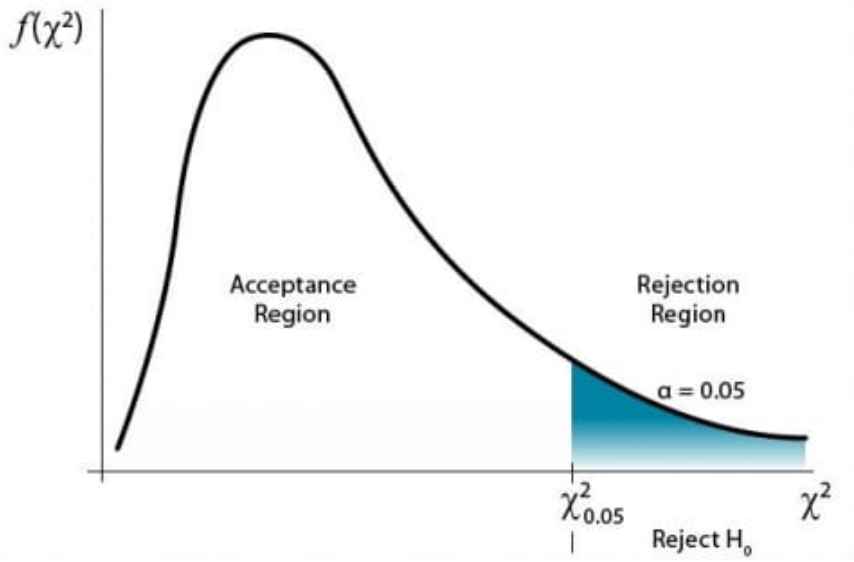



Figure 3 Acceptance and rejection regions for a chi-square distribution. Values greater than the significance value indicate that the data counted deviate significantly from the expected counts.

This is based on a null hypothesis that the observed data occur as expected. The alternate hypothesis would be that the data deviate from the expected values.

## Yates' Correction

Small numbers of counts should be avoided if possible, as well as situations with few degrees of freedom. In these cases, the power of the test is low and the chi-square approximation to the true distribution probabilities may be inaccurate. Notice how the curves above change from having a few degrees of freedom to many degrees of freedom. When possible, it is recommended to have each category contain at least five data points. In situations where you have one df, such as 2 x 2 contingency table, you should apply Yates' Correction for Continuity. This correction factor involves subtracting 0.5 from the absolute value of each category or cell:

Yates' correction for continuity:

$$\chi^2 = \sum \frac{(|observed - expected| - 0.5)^2}{expected} \tag{1.5.3}$$

The two straight vertical lines around "observed-expected," i.e. '||', mean that you should use the absolute value of that difference — you should convert any negative values to positive before subtracting 0.5.

# Testing Proportions

## Hypotheses

Let's look at a simple example. A plant breeder is examining the inheritance of chlorophyll in a new maize cultivar. It is hypothesized that three of every four (3:1 ratio) plants of the new population will be colored green while the others will be yellow. After conducting an emergence test, it is found that of 200 plants emerged, the ratio of green to yellow is 1:1, 100 green:100 yellow. Is this result different from what was expected? At what probability level is it significant?

The $\chi^2$ test will be used to test this example. We begin with the hypothesis:

- $H_0$: The ratio of green to yellow in the emerged plants is the same as expected.
- $H_\alpha$: The ratio of green to yellow in the emerged plants is different from expected
- $\alpha = 0.05$

## Calculation

The method of testing here divides the total emerged plants into their expected numbers. Of 200 plants it is found that 100 were green and 100 yellow. Compare this to the expected numbers. In the 3:1 expected ratio, 150 would be expected to be green and 50 expected to be yellow. You have the data necessary to calculate the $\chi^2$ statistic. Basically, there are two different cells or observations, the number of green and the number of yellow plants. We know the expected value of each. Use the $\chi^2$ formula to sum the differences over the two categories.

Since there are two observations and we are constraining the observations by one, the df are 2 – 1 = 1. We use the Yates correction for continuity accordingly and subtract 0.5 from the absolute difference for each category. An example is below.

$$\chi^2 = \sum \frac{(|observed - expected| - 0.5)^2}{expected} \tag{1.5.4}$$

$$\chi^2 = \sum \frac{(|100 - 150| - 0.5)^2}{150} + \frac{(|100 - 50| - 0.5)^2}{50} \tag{1.5.5}$$

$$\chi^2 = \frac{(49.5)^2}{150} + \frac{(49.5)^2}{50} = \frac{(2450.25)}{150} + \frac{(2450.25)}{50} \tag{1.5.6}$$

$$\chi^2 = 16.335 + 49.005 = 65.34 \tag{1.5.7}$$

$$\chi^2 = 65.34 \text{ with 1 df} \tag{1.5.8}$$

$$\chi^2 @ \text{ alpha} = 0.05 \text{ and 1 df is 3.84} \tag{1.5.9}$$

$$\chi^2 \text{ is therefore significant} \tag{1.5.10}$$

The $\chi^2$ significance value at 0.05 is 3.84. Even at a significance level of 0.001, the critical $\chi^2$ value is 10.827. What is observed is very different from what was expected. The testing indicates the null hypothesis IS NOT correct. We would reject the null hypothesis, i.e., the ratio of green to yellow plants is different from expected.

Note that the binomial distribution provides an exact test, which is better than the $\chi^2$ test, even with the Yates' correction (subtracting the 0.5 in the formula).

### Exercise: Calculating a Chi-Square Test (1)

**Calculating a Chi-Square Test (1)**

The discussion of the $\chi^2$ test has introduced several interpretations of the test. Its main premise is to test a set of counts, cells, or categories to determine if the numbers in each are significantly different from the numbers expected in each situation.

In this exercise, we will perform the calculation on the original simple $\chi^2$ test (without Yates' correction). In an emergence test of 200 plants, 100 were observed to be green, while 100 were found to be yellow (a 1:1 ratio). The expected ratio was 3:1 (150 green to 50 yellow in this case). Are the observed numbers different enough to be significant? In other words, test the hypothesis:

$H_0$: The true ratio is 3 green : 1 yellow

$H_A$: The true ratio is not 3 green : 1 yellow

**Steps**

Enter the data in Excel to get this table.

| Color | Count |
|---|---|
| Green | 100 |

| | |
|---|---|
| Yellow | 100 |
| Total | **=SUM(B2:B3)** |

The expected counts are 150 green and 50 yellow. The expected counts can be calculated from the total observations and the 3:1 ratio. If the expected ratio is 3:1 there are 4 total.

| Color | Count | Expected Ratio |
|---|---|---|
| Green | 100 | 3 |
| Yellow | 100 | 1 |
| Total | 200 | **=SUM(C2:C3)** |

Divide both sides of the ratio by 4 to get a ratio with a total of 1.

| Color | Count | Expected Ratio | Normalized Ratio |
|---|---|---|---|
| Green | 100 | 3 | **=C2/C4** |
| Yellow | 100 | 1 | **=C3/C4** |
| Total | 200 | 4 | **=SUM(D2:D3)** |

The expected counts can then be calculated by multiplying the total number of observations by the expected ratio.

| Color | Count | Expected Ratio | Normalized Ratio | Expected Count |
|---|---|---|---|---|
| Green | 100 | 3 | 0.75 | **=D2*B4** |
| Yellow | 100 | 1 | 0.25 | **=D3*B4** |
| Total | 200 | 4 | 1 | **=SUM(E2:E3)** |

Check your math by making sure that the expected counts sum to the same number as the observed counts.

| Color | Count | Expected Ratio | Normalized Ratio | Expected Count |
|---|---|---|---|---|
| Green | 100 | 3 | 0.75 | 150 |
| Yellow | 100 | 1 | 0.25 | 50 |
| Total | 200 | 4 | 1 | 200 |

The $\chi^2$ statistic is calculated using the observed and expected counts. There are two categories in this problem: green and yellow. Within each category, calculate ((O – E)2/ E) where O is an observed count and E is an expected count, then sum across the categories to find the statistic.

| Color | Count | Expected Ratio | Normalized Ratio | Expected Count | Chi-Squared |
|---|---|---|---|---|---|
| Green | 100 | 3 | 0.75 | 150 | **=((B2-E2)^2)/E2)** |
| Yellow | 100 | 1 | 0.25 | 50 | **=((B3-E3)^2)/E3)** |
| Total | 200 | 4 | 1 | 200 | **=SUM(F2:F3)** |

| Color | Count | Expected Ratio | Normalized Ratio | Expected Count | Chi-Squared |
|---|---|---|---|---|---|
| Green | 100 | 3 | 0.75 | 150 | 16.66666667 |
| Yellow | 100 | 1 | 0.25 | 50 | 50 |

| | | | | | |
|---|---|---|---|---|---|
| Total | 200 | 4 | 1 | 200 | 66.66666667 |

Find the degrees of freedom for this test by subtracting one from the number of rows. There are two rows, so there is one degree of freedom.

The p-value for this test can be found using the formula “CHISQ.DIST.RT(Chi, DF)”. Enter the calculated chi-squared statistic for Chi and the correct degrees of freedom.

| | | |
|---|---|---|
| **Chi-Squared** | =F4 | 66.66666667 |
| **Deg. of Freedom** | 1 | 1 |
| **p-value** | =CHISQ.DIST.RT(C8,C9) | 3.21526E-16 |

A p-value less than 0.05 means that the test is significant. In this case, the p-value is very small. This means that the test is highly significant, and there is little or no chance that the results would have occurred by chance, and the null hypothesis should be rejected.

## Testing a 9:3:3:1 genetic ratio

We use the same principles as in the first exercise to check other genetic ratios. Use Excel to analyze the data from Example 18.10 on page 282 in our textbook. We observe 150, 42, 50, and 8 in classes A, B, C, and D, respectively. From genetic theory, we hypothesize a 9:3:3:1 ratio. Should we reject this hypothesis?

- $H_0$: The true ratio is 9A : 3B : 3C : 1D
- $H_A$: The true ratio is not 9A : 3B : 3C : 1D

Open a new Excel workbook and enter this data set:

| Class | Count |
|---|---|
| A | 150 |
| B | 42 |
| C | 50 |
| D | 8 |

Follow the same steps used in Exercise 5.1. The ratio for this hypothesis is 9:3:3:1, with a total of 16. This ratio is used to calculate the expected counts.

| Class | Count | Expected Ratio | Normalized Ratio | Expected Count | Chi-Squared |
|---|---|---|---|---|---|
| A | 150 | 9 | 0.5625 | 140.625 | 0.625 |
| B | 42 | 3 | 0.1875 | 46.875 | 0.507 |
| C | 50 | 3 | 0.1875 | 46.875 | 0.2083333333 |
| D | 8 | 1 | 0.0625 | 15.625 | 3.721 |
| Total | 250 | 16 | 1 | 250 | 5.0613333333 |

| | |
|---|---|
| **Chi-Squared** | 5.0613333333 |
| **Deg. of Freedom** | 3 |
| **p-value** | 0.1674 |

The probability of a greater $\chi^2$ (the p-value) is 0.1674, and we fail to reject the null hypothesis. This can be determined from the

p-value, which is greater than the alpha level of 0.05.

## Observations vs. Expectations

**To test whether an observed proportion is different from the theoretical proportion.** A proportion measures what percentage of a population that has a certain characteristic or does not have a certain characteristic. These are measured as a proportion or percentage of the population (35% of the population will have a trait) or as ratios (3:1) ratio means 3 of every four members of a population contain a genetic allele) within the population. When sampling a population, you may want to know whether this sample population has a characteristic occurring in a different proportion as compared to the whole population. Or an experimental treatment may cause a population to have a different ratio of occurrence of a certain characteristic than expected. Did what happened in an experiment deviate from what was expected? How much did it deviate? These are questions which can be answered using the $\chi^2$ test.

We have already seen some proportion data in the binomial distribution in chapter 3 on Categorical Data—Binary. The chi-square analyses for the simple case of two categories agree with the results of the normal approximation to the binomial. However, the $\chi^2$ can be used for counts from more than just two categories.

# Contingency Tables

**Contingency tables** are tables of count data and can be analyzed with $\chi^2$. The simple proportion example shown earlier could have been analyzed with a contingency table. Often more complex experiments have interactions between two ways of categorizing the data. For example, flower color and leaf pubescence. The contingency table simplifies the comparison by breaking down the categories for each variant into a table format, which is designed for completing two-way analyses. It has a form that classifies the first set of data over the columns and the second set over the rows (Table 1).

*Table 1 Contingency table layout.*

| **Level** | **1** | **2** | **3** | **4** | n/a | **c** | **Total** |
|---|---|---|---|---|---|---|---|
| 1 | $O_{11}$ | $O_{12}$ | $O_{13}$ | $O_{14}$ | n/a | $O_{1c}$ | $r_1$ |
| 2 | $O_{21}$ | $O_{22}$ | $O_{23}$ | $O_{24}$ | n/a | $O_{2c}$ | $r_2$ |
| 3 | $O_{31}$ | $O_{32}$ | $O_{33}$ | $O_{34}$ | n/a | $O_{3c}$ | $r_3$ |
| n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| r | $O_{r1}$ | $O_{r2}$ | $O_{r3}$ | $O_{r4}$ | n/a | $O_{rc}$ | $r_k$ |
| Total | $c_1$ | $c_2$ | $c_3$ | $c_4$ | n/a | $c_m$ | n |

Notice that each row and column have a total, $r_k$ and $c_m$, respectively. These numbers are the counts for each cell defined by the row category and column category. Comparing these numbers to the total of the whole table establishes the total proportion break-down of each cell. The row totals give a breakdown of the row categories and the column totals for the column categories. Summing each of the row totals and column totals produces the grand total. The row and column totals are also integral to the $\chi^2$ test because they can be used to calculate an expected value in each cell for a two-way analysis. This expected value is then compared with the actual number counted by using the $\chi^2$ test.

## Expected Values

The expected value calculation assumes independence of the two criteria (which will be tested in the next section). This assumption states that the variables in the columns and the rows have no interaction; they are independent of each other. One property of independence is that the expected value of each cell should be the product of the row and column category proportions times total number. This results in the following formula (Equation 3):

$$\text{cell expected value} = \frac{(\text{row total}) \text{ x } (\text{column total})}{\text{grand total}} \tag{1.5.11}$$

The number calculated uses the relative proportion of each row and column to calculate the number each cell should contain. The calculation of the $\chi^2$ occurs by summing the deviations of each cell from its expected value.

## Degrees of Freedom

The degrees of freedom associated with this are calculated as the product of one less than the row and column categories.

$$\text{degrees of freedom} = (\text{number of rows - } 1) \text{ x } (\text{number of columns - } 1) \tag{1.5.12}$$

For example, a 2 x 2 table would have (2-1) x (2-1) = 1 df.

For better statistical inferences, each cell should contain at least a count of five. If smaller counts occur, combining of row or column categories is suggested to create cell counts larger than five. Recent studies suggest that even though the observed count in a cell is five, the expected count for cells need only be larger than six when significance at the 0.05 level is needed or an expected value larger than 10 when significance at the 0.01 level is desired.

The simplest case of contingency tables is a 2 x 2 analysis. But more detailed tables may be created, even into multiple dimensions. When using a 2 x 2 table, the correction for continuity should be used. Other corrections than Yates' Correction for Continuity exist for other situations.

# Test for Independence

## Two-way Example



Figure 4 Effect of Blackleg bacteria on potato tuber. Photo by the United Nations Economic Commission for Europe.

An example of this is the effect of different fertilizer treatments on the incidence of blackleg (*Bacterium phytotherum*) on numbers of potato seedlings and tubers (Figure 4).

Our objective is to test whether the occurrence of the disease has some relationship to nitrogen or manure fertilizer application.

The null hypothesis is that fertilizer and Blackleg occurrence have no relationship, i.e., fertilizer application and Blackleg are independent. Data are contained in Table 2.

We then compute a $\chi^2$ statistic to see if the value is high enough to reject the null.

*Table 2 Observed number of potato seedlings in response to fertilization and Blackleg bacteria.*

| **Observed frequencies** | **Blackleg** | **No blackleg** | **Total** |
|---|---|---|---|
| No fertilizer | 16 | 85 | 101 |
| Nitrogen only | 10 | 85 | 95 |
| Manure only | 4 | 109 | 113 |
| Nitrogen and manure | 14 | 127 | 141 |
| Total | 44 | 406 | 450 |

## Blackleg Example

These observed values (Table 2) are compared to the calculated expected values using the expected value equation (Equation 3) and set up in an expected value table (Table 3).

*Table 3 Expected number of potato seedlings in response to fertilization and Blackleg bacteria.*

| **Expected frequencies** | **Blackleg** | **No blackleg** | **Total** |
|---|---|---|---|
| No fertilizer | 9.9 | 91.1 | 101 |
| Nitrogen only | 9.3 | 85.7 | 95 |
| Manure only | 11.0 | 102.0 | 113 |
| Nitrogen and manure | 13.8 | 127.2 | 141 |
| Total | 44 | 406 | 450 |

The computed $\chi^2$ values for each cell using Equation 1 are shown in Table 4.

*Table 4 Computed Chi-square values.*

| $\frac{(\text{Observed - expected})^2}{\text{expected}}$ | **Blackleg** | **No blackleg** |
|---|---|---|
| No fertilizer | 3.76 | 0.41 |
| Nitrogen only | 0.05 | 0.01 |
| Manure only | 4.45 | 0.48 |
| Nitrogen and manure | 0.00 | 0.00 |
| Total | **8.26** | **0.90** |

## Calculating Differences

The calculated $\chi^2$ from summing over the table values is **9.16**. This is larger than the significance value at the 0.05 level (3 df), **7.82**. The degrees of freedom are (4–1)(2–1) = 3 because there are 4 rows and 2 columns. The deviations from expected in the cells are large. The expected value in each cell assumes independence of the conditions. Since the data deviate from those values significantly, we reject the null hypothesis of independence and conclude that the **fertilizer treatment affected the incidence of blackleg** in this experiment.

# Testing for Independence of Data



Figure 5 Germinating plants. Photo by Iowa State University.

We will duplicate the analysis of independence of data sets using the text example 18.13 with Excel. Five storage methods were tested for effects on the germination of peas (Figure 5). The data are from Table 18.5 in *Practical Statistics and Experimental Design for Plant and Crop Science*.

First, enter data into an Excel table with the variables 'Germinated' (Yes or No) and 'Count'. Your data should look like in Table 5.

*Table 5 Germination and counts of peas seedlings.*

| Storage Method | Germinated | Count |
|---|---|---|
| A | yes | 112 |
| A | no | 12 |
| B | yes | 76 |
| B | no | 14 |
| C | yes | 88 |
| C | no | 32 |
| D | yes | 43 |
| D | no | 7 |
| E | yes | 92 |
| E | no | 8 |

**The hypothesis tested here is:**

- $H_0$: Germination and Storage method are independent of each other
- $H_A$: Germination and storage method are not independent of each other

# Testing for Independence of Data (2)

Make the analysis easier by arranging the data into a table.

Sum across each column and row. Then sum the column or row totals to find the grand total. This is a good chance to check the

arithmetic by making sure that the column totals and row totals sum to the same value (Table 6).

*Table 6 Contingency table of germination and counts of peas seedlings.*

| **Observed** | A | B | C | D | E | **Row Total** |
|---|---|---|---|---|---|---|
| Yes | 112 | 76 | 88 | 43 | 92 | **411** |
| No | 12 | 14 | 32 | 7 | 8 | **73** |
| **Column Total** | **124** | **90** | **120** | **50** | **100** | n/a |
| **Grand Total** | n/a | n/a | n/a | n/a | n/a | **484** |

The expected counts can now be calculated using row, column, and grand total. Each expected count is calculated using the formula (Equation 5):

$$\frac{(\text{Col. Total x Row Total})}{\text{Grand Total}} \tag{1.5.13}$$

The expected values have been filled into Table 7. See if you can repeat the results in your own table.

*Table 7 Expected counts from germination and counts of peas seedlings values in table 6.*

| **Observed** | A | B | C | D | E | **Row Total** |
|---|---|---|---|---|---|---|
| Yes | 105.2975207 | 76.42561983 | 101.9008264 | 42.45867769 | 84.91735537 | 411 |
| No | 18.70247934 | 13.57438017 | 18.09917355 | 7.541322314 | 15.08264463 | 73 |
| **Column Total** | 124 | 90 | 120 | 50 | 100 | n/a |
| **Grand Total** | n/a | n/a | n/a | n/a | n/a | **484** |

We see that the Pearson $\chi^2$, computed from data in Table 8, is 19.379 (Table 9), and we, therefore, reject the hypothesis that storage methods are independent of germination. The degrees of freedom can be calculated as: (Rows – 1) * (Columns – 1). The p-value is less than 0.05, so we reject the null hypothesis.

*Table 8 Computed chi-square values.*

| $\chi^2$ | A | B | C | D | E |
|---|---|---|---|---|---|
| Yes | 0.426631406 | 0.002370308 | 1.896284678 | 0.00690153 | 0.59073767 |
| No | 2.401993258 | 0.013345158 | 10.6763425 | 0.038856561 | 3.325932299 |
| Column Total | 2.828624664 | 0.015715465 | 12.57262718 | 0.045758091 | 3.916669666 |

*Table 9 Computed chi-square, df, and p-values.*

| | |
|---|---|
| **Total** | 19.37939507 |
| **DF=(r-1)*(c-1)** | 4 |
| **p-value** | 0.000661886 |

## Testing for Independence of Data (3)

The test was for independence here. Independence would mean that the two categories have no effect on each other. The observed cell values, if independent, would not deviate much from the expected values. In this case, the data have deviated enough to be significant at the 0.001 level. Our conclusion then is that there is some effect of storage method on the viability of plants.

## Two-way Contingency Tables

We can test to see if two categorical variables are associated. The contingency table is applied in situations where we have a two–way (or higher) classification structure. We may wish to test to see if the two different bases for categories are independent of each other. For example, we might want to learn if two transgenes are segregating independently or if they are linked. In fact, the null–hypothesis assumption in calculating the expected values of cells assumes independence of the two categorizations. The two–way analysis sorts the data to compare the interaction between variables. If the data are independent, then the numbers in the cells should be similar to the expected values. If the data are not independent, or there is a significant amount of interaction between variables, the contingency table and $\chi^2$ test will indicate numbers different from expected in the cells.

The $\chi^2$ statistic can be applied to test for independence. The calculation is done as illustrated previously. The squared differences between the observed and expected, divided by expected, are summed over all cells of the table and tested via the $\chi^2$ statistic.

# Test for Heterogeneity

We also can test whether several samples are **homogeneous enough to be pooled together.** The test for heterogeneity is similar in method to the test for independence, which tests each cell for its difference from expected. But interpretation is distinctly different. In the test for heterogeneity, two or more different samples are identified. Samples are tested to see if they could have been drawn from the same population (i.e., the proportion for each sample is similar). If the samples are similar, they are considered homogeneous and from the same population. If they are not, then they are heterogeneous and considered from different populations. For example, we might want to test whether the genetic linkage between two transgenes is the same in two different genetic backgrounds. It is determined by testing the breakdown of numbers into categories from one sample to the next. In the test of heterogeneity, when the $\chi^2$ value is significant, the samples are heterogeneous.

## Pooling Data

If several samples are found to be homogeneous, the data can be pooled. The usefulness of pooling the samples is in order to create larger sample sizes with fewer categories. Each additional category adds a degree of freedom to the analysis. Each degree of freedom we remove by pooling categories allows for a smaller $\chi^2$ value to be significant, thus making the test more powerful. The pooled samples should have the same characteristics as the individual samples but allow better detection of real differences. The text “Agricultural Experimentation, Design and Analysis” by Thomas Little and F. Jackson Hills (1978, John Wiley and Sons) describes an example of the breakdown of eight progenies of marigolds into normal and virescent categories, which we will use to illustrate this test. Virescent means that chlorophyll is present in the petals of the flower.

*Table 10 Counts of normal and virescent marigold plants and associated chi-square values for eight progeny groups.*

| **Progeny** | **Normal** | **Virescent** | $\chi^2$**(3:1)** |
|---|---|---|---|
| 1 | 315 | 85 | 3.00 |
| 2 | 602 | 170 | 3.65 |
| 3 | 868 | 252 | 3.73 |
| 4 | 174 | 42 | 3.56 |
| 5 | 192 | 48 | 3.20 |
| 6 | 165 | 39 | 3.76 |
| 7 | 161 | 43 | 1.67 |
| 8 | 629 | 175 | 4.48 |
| Totals | 3106 | 854 | 27.05 |

In this example, we want to test whether the 8 samples, each of which can be tested for a 3:1 ratio, can be pooled together. To do this, we calculate the $\chi^2$ for each sample, sum these together, and subtract the $\chi^2$ for the pooled data. This gives a measure of interaction, which, if large, implies the samples are too heterogeneous to pool together.

Why would we want to pool in the first place? In the above table, we can see that nearly every sample (progeny) has $\chi^2$ value

above 3.0 for a 1 df test. This is not large enough to reject the null hypothesis (critical $\chi^2$ = 3.84) but is significant at the 10% level. With several progenies individually showing this trend, we want to combine the data to have sufficient evidence to reject the 3:1 ratio if it is not true. However, we must test for heterogeneity first to know whether the samples can be combined.

We do this test for heterogeneity in three steps:

1. Compute individual chi-square statistics for each of the individual samples and add them
2. Compute the chi-square for pooled samples
3. Subtract chi-square values and degrees of freedom to test for heterogeneity. If the $\chi^2$ from subtraction is small relative to the table value, we would fail to reject the null and conclude progeny ratios are homogeneous. If large, we reject the null and consider them heterogeneous.

## Chi-Square Values

The eight different progenies were tested for their difference in the normal versus virescent from an expected 3:1 ratio. Only progeny group 8 differed from that ratio significantly (Table 10). They seem to be similar in their sample makeup. To further test the data, we start by pooling the samples to test for heterogeneity (Table 11). The raw numbers can be summed, and the $\chi^2$ value is calculated for a 3:1 ratio. Note that the total number of plants is 3,960, and we expect 2,970 normal:990 virescent.

$$\chi^2 = \frac{(|100 - 150|)^2}{150} + \frac{(|100 - 50|)^2}{50} \tag{1.5.14}$$

*Table 11 Computed chi-squares for total, pooled, and heterogeneity.*

| **Source** | **df** | $\chi^2$ |
|---|---|---|
| Total | 8 | 27.05 |
| Pooled | 1 | 24.91 |
| Heterogeneity | 7 | 2.14 |

This value is highly significant, with 1 df. But are we justified in pooling? To find if the samples are heterogeneous, the second step is summing the $\chi^2$ values from each sample. The sum of those is 27.05. (This is a property of $\chi^2$ values; they may be added for independent groups, such as the 8 independent progeny.) The eight total degrees of freedom can be partitioned into the pooled $\chi^2$ with 1 df and the heterogeneity (non-homogeneity) with 7 df. The difference in the $\chi^2$ values gives the $\chi^2$ for the heterogeneity.

In this case, the heterogeneity is not significant. Therefore, the data are considered to be homogenous. We will work more with this example in the next Try This exercise.

One historical source of confusion in testing heterogeneity is this: from where do the heterogeneity degrees of freedom come? Part of this confusion is because earlier in this chapter, you were taught that the degrees of freedom for the chi-square distribution was equal to the (number of categories – 1) for a one-way (single factor) study and equal to (rows-1)*(columns-1) for a two-way study (see contingency tables).

For the test of heterogeneity, the degrees of freedom associated with the heterogeneity are calculated differently. In effect, we calculate the degrees of freedom for each population (“progeny” in the table) and then add those degrees of freedom. So for Progeny 1, there was 1 degree of freedom associated with the chi-square. Since there are 8 total progeny, there are 8 degrees of freedom associated with the heterogeneity chi-square.

## Testing for Hetero/Homogeneity

The chi-square analysis can be used to test for differences in the proportions of samples. When several repeated samples are gathered, they may be tested to determine if they may have come from the same population (are homogenous). If they have come from the same population, the samples may be pooled, strengthening the test by adding replicated measurements. The hypothesis tested is:

- $H_0$: The samples are homogenous and can be pooled.
- $H_a$: The samples are not homogeneous and should not be pooled.

A test of homogeneity is done by first testing each sample (referred to as progeny here), then adding each category together and calculating a chi-square statistic for the entire sample.

Download our data (QM-Example 4 Data [xls]) to test this hypothesis. The progeny 1 sample has been filled in. The same tools from previous exercises are used, but a different hypothesis is tested.

The P-value is much higher than 0.05, and it is appropriate to fail to reject the null hypothesis and conclude that the samples are homogeneous.

# Summary

**Chi-Square Test**

- Has degrees of freedom depending on the number of categories.
- Goodness of fit: $\chi^2$ = Sum (O-E)$^2$/E
- Yates' continuity correction for small df

**Tests of Proportions**

- Find the expected number in each class
- Use of the $\chi^2$ goodness-of-fit

**Contingency Tables**

- Tables of count data
- Tested with chi-square
- The expected value is (Row proportion x Col Proportion) / (Total)
- Degrees of freedom is (Rows-1) x (Cols-1)
- Comparison of two proportions is 2 x 2 contingency table

**Test for Independence**

- Bell-shaped curve
- Symmetric about the mean, μ
- 68% of values are within 1 σ and 95% are within 2 σ of mean

**Test for Heterogeneity**

- Tell how many standard deviations above or below the mean
- Defined as (Y – μ)/σ
- Allow computation of probabilities with the normal distribution

**How to cite this Chapter:** Mowers, R., K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Categorical Data: Multivariate. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.5: Categorical Data Multivariate is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.6: Continuous Data

In this chapter, we will learn that sample means from a population that has a Normal distribution are also distributed as Normal, but with a smaller variance that has a population. We will then extend the concept to samples from two populations. For example, sometimes, we want to identify the best management practice between two alternatives. Such as comparing the use of starter fertilizer to no starter fertilizer, or one corn hybrid against another, or a new insecticide against a common one. In all of these comparisons, the key question to be answered is, "Are the means of the populations representing the two treatments the same." In the language of statistics, this is called a two-sample hypothesis; the most common method for testing it is the t-test.

**Learning Objectives**

- How the distribution of a set of sample means around their population mean can be described in terms of t-values
- How to calculate the t-value for sample means as related to their population means
- How a confidence limit describes a range of possible values for the population mean, μ
- How a t-test can be used to determine whether the difference between two sample means is significant
- How to conduct t-tests using Excel

## The t-Distribution

**The *t*-distribution is used for sample mean when the variance is not known.** Statistical standards must apply to a wide variety of research, from corn yield trials to soybean disease studies to nitrate sampling in groundwater. Even within one of these categories, the amount of variability may differ from one month or year to the next. Therefore, we need to describe the distribution of our sample means according to the distribution curve itself, and in a way that is unitless. We cannot use the normal distribution described in Chapter 2 on Distributions and Probability because we do not know the variance. We define this distribution of sample means in terms of the *t*-distribution.

As we learned, the shape of the distribution curve for sample means is based on the number of observations per sample. Our values for *t* will also vary with the number of observations or replications, that occur. However, for normally distributed variables and a given number of degrees of freedom (df), we know that 95% of the sample means will lie within a particular *t*-value of the mean. This *t*-value remains the same, for the given df, whether we are studying swine, alfalfa, or soil moisture.

What are **degrees of freedom**? In general, the number of degrees of freedom associated with a set of sample means is the number of individuals in the sample used to calculate the mean minus 1.

## A Sample t-Value Table

For example, look at the following *t*-value table. For a sample size of 5, 95% of the sample means will be within plus or minus 2.776 standard errors (S/√n) of the mean:

*Table 1 t-values: Probability of obtaining a value as large or larger.*

| **Degrees of Freedom** | **Percentage Points in Top Tail** | | | | |
|---|---|---|---|---|---|
| | **5** | **2.5** | **1** | **0.5** | **1** |
| 1 | 6.314 | 12.706 | 31.821 | 63.657 | 318.309 |
| 2 | 2.920 | 4.303 | 6.965 | 9.925 | 22.327 |
| 3 | 2.353 | 3.182 | 4.541 | 5.841 | 10.215 |
| 4 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 2.015 | 2.571 | 3.365 | 4.032 | 5.893 |
| 6 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |

| | | | | | |
|---|---|---|---|---|---|
| 7 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 |
| 8 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 |
| 9 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 |
| 10 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 |
| 11 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 |
| 12 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 |
| 13 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |

In this example, a sample with four degrees of freedom will have 5% of the distribution 2.776 × *t* above the mean and another 5% of the distribution 2.776 × *t* below the mean.

**Study Question 1: Continuous Data**

**Query 1.6.1**

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Sample Means in Values of t

Now that we understand how *t* values relate to the distribution curve for a set of sample means, we must learn how to express our sample means in values of *t*. This is given by the following formula:

$$t = \frac{\hat{Y} - \mu}{S_{\hat{y}}} \tag{1.6.1}$$

**where:**
$\hat{Y}$ = sample mean,
$\mu$ = population mean,
$S_{\hat{y}}$ = standard error of the sample mean = $\frac{S}{\sqrt{n}}$.

We use the *t*-value rather than *z* because we are estimating $\sigma^2$ with $s^2$. The *t*-value represents the difference between a particular sample mean and the mean of the whole population of sample means as a function of the standard error for that population, $S_y = S/\sqrt{n}$. Stated more simply, the *t*-value is the number of standard errors a particular sample mean is from the average value for all of the sample means. This tells us whether our difference is large in relation to the variation of our sample.

## t-Value Scenarios

For example, a 500 kg/ha difference between the sample mean of a certain corn hybrid and the population mean of all corn hybrids would not be a large difference when we are dealing with corn (which might have a standard error of 750 kg/ha for a given number of plots) as when we are studying soybean (which could have a standard error of 250 kg/ha for the same number of plots). The *t*-values for each of these scenarios are given below:

**Corn:**

$$t = \frac{\hat{Y} - \mu}{S_{\hat{y}}} = \frac{250kg}{750kg} = 0.34 \tag{1.6.2}$$

**Soybean:**

$$t = \frac{\hat{Y} - \mu}{S_{\hat{y}}} = \frac{500kg}{250kg} = 2.00 \tag{1.6.3}$$

Equation 3 Soybean example of t value computation (1.6.4)

### Study Questions 2 & 3: Continuous Data

#### Query 1.6.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Confidence Limits & Intervals

**Confidence intervals (CI)** are used to describe a range of values for a set of sample means from multiple samples of the same population. A CI does not predict that the true population mean will be included in the interval.

We can set the ranges for our confidence interval using Excel. You will learn how to do this in the Excel lab for this chapter.

### Try This! Calculation Exercise

We use the term confidence to remind us that the variation associated with this procedure is the result of our own experimental and sampling techniques. Stated otherwise, the population mean in which we are interested stays the same — it is our confidence interval that shifts according to the number of and particular samples in our set. There is no probability or chance associated with the population mean but rather with the value of the sample means.

**Ex. 1: Calculating a Confidence Interval**

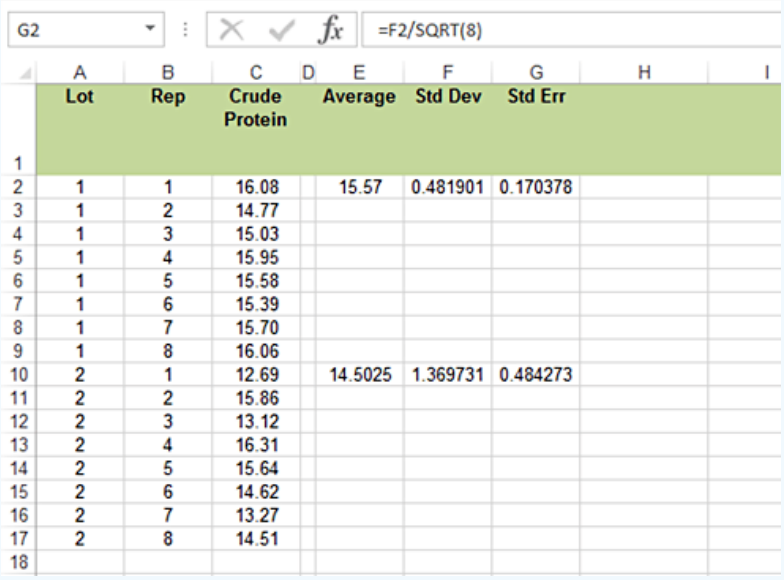
Often we are asked to make decisions about some characteristic of a population based on samples selected from it. In this example, we will consider the protein value of two lots of alfalfa hay.

The data in the Alfalfa Quality worksheet were collected from two lots of hay that were offered for sale at auction. Each lot was first-cutting hay and weighed approximately 30 tons. Based upon a one percentage unit difference in the protein analysis, lot 2 was discounted $10 per ton because the buyer claimed it was less than 15.5% protein. The buyer's method to see if a lot is less than 15.5% is to calculate a 95% confidence interval for protein and see if it includes 15.5%.

Our first question should be, "How confident are we that the sample means represent the true mean." One approach to answering this question is to calculate a confidence interval for the mean of each sample population.

**STEPS**

- Calculate the 95% confidence interval of the sample mean for each lot. Download and open the Excel file QM-mod6-ex1data.xls.
- There should be three columns: Lot, Rep, and Crude Protein.
- We will calculate a mean, standard deviation, and standard error for each lot. Use the Average and STDEV.S to calculate the mean and standard deviation of each lot sample. Then use the standard deviation to calculate the standard error by dividing the standard deviation by the square root of the number of samples in a lot (Figure 1).




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | | |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | | |
| 3 | 1 | 2 | 14.77 | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | | |
| 11 | 2 | 2 | 15.86 | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | |
| 18 | | | | | | | | | |

Figure 1 Calculation of 95 % confidence interval.

It is possible to use Excel to calculate a confidence interval from a t-distribution using the formula Confidence.T. The arguments are the alpha level, standard deviation, and the number of samples or replications.

- Now subtract and add the calculated value from the mean to find the confidence interval.
- Now, we will hand calculate the confidence interval to see that it matches that given by Excel. Use Equation 4 and the SE from Excel. Use the t-value for 5% probability in two tails (2.5% in one tail) from either Appendix 3 or the t-table given earlier in this unit. Remember, there are 7 df for 8 reps in each sample.
- Calculate the 95% confidence interval for each lot (Figure 2).
- Do the two confidence intervals overlap?
- Save the file to your desktop for use in the next exercises.

H2 =CONFIDENCE.T(0.05,F2,8)

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | Confidence Interval Calculation | Lower Limit of CI | Upper Limit of CI |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | 0.402879318 | 15.17 | 15.97 |
| 3 | 1 | 2 | 14.77 | | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | 1.145124162 | 13.36 | 15.65 |
| 11 | 2 | 2 | 15.86 | | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | | |
| 18 | | | | | | | | | | |

Figure 2 Calculation of two 95 % confidence intervals

### Study Question 4: Continuous Data

#### Query 1.6.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Discussion

Should Lot 2 have been discounted for a small difference? Support your assertions.

# t-Tests For Significance

***t*-tests can be used to compare two treatment means.**

Confidence intervals may work well for single samples compared to the mean. But what can we use to compare the two samples? We can also use *t*-values to determine whether the calculated mean values of two sample sets are significantly different. This difference can be expressed in terms of *t*:

$$t = \frac{\bar{d} - \mu_{\bar{d}}}{S_{\bar{d}}} \tag{1.6.5}$$

**where:**
$\bar{d}$ = difference between sample means,
$\mu_{\bar{d}}$ = difference between population means,
$S_{\bar{d}}$ = standard error of the difference.

In most cases, we are trying to identify any significant difference; the value we use for the mean difference is zero. A zero value for a mean difference, therefore, represents the null hypothesis that there is no treatment effect. Our *t*-value expresses the actual

difference between the two sample means as a function of the standard error of the difference. Again, it is this conversion that makes the difference meaningful regardless of what particular variable we are measuring.

## Paired *t*-Test

It is at this point that we must know more about the experiment or how the populations were sampled in order to compute $S_d$. If both treatments of an experiment are applied to units that are paired in blocks, we use a paired *t*-test method. If the populations are independent or treatments are applied completely at random to the experimental units, we use what is called an independent samples *t*-test. Each of these has a different standard error of the difference $S_d$.

Suppose first that the treatments are on units that are naturally paired (and thus not independent). An example is a test for differences in two wheat varieties, each grown in the same block or area in a field for 5 different blocks. There is a natural pairing of the varieties because they each appear in each block even though they have been randomly assigned to either the left or right side of the block. Another example is a test of side dressing of nitrogen fertilizer vs the check of no additional fertilizer, each applied to half a field on fields from 12 different farms.

For the paired *t*-test, the computation of the *t*-value is easy. Just take the differences of treatment 1 – treatment 2 for each pair. Then, use these differences as the data and compute the *t* value as done in equation 1. The d in equation 5 is just the average of the differences, and the $S_d$ is the square root of $S_d^2/n$. The variance of the differences is $S_d^2$, and number of pairs is n. The test of the null hypothesis that treatments have the same mean is just a test of differences being zero, $H_0$: $\mu_d = 0$. We will see in the Excel exercises how this is done.

## Independent Samples t-Test

The independent samples *t*-test, done for independent populations or samples, uses the $S_d$ given in Equation 5 below. One independent sample example is a set of 24 pots of soybeans in a growth chamber, half of which receive a zinc foliar supplement, while the other half receive no supplement. The pots which receive the treatment are randomly chosen, for example, pots 2, 3, 7, 9, 10, 11, 13, 14, 16, 18, 21, and 24. Sometimes samples are considered to be independent when they are drawn at random from two mutually-exclusive populations, for example, a poll on farmers' opinions on a land-use policy, with 100 respondents in Illinois having an average farm size of over 1000 acres and another 100 with farms less than 1000 acres. It is important that these samples be chosen at random for a valid comparison.

When populations are independent, as in the 2-sample case, the variance of a difference is the sum of the two variances. In the 2-sample (independent samples) case, the variance of a difference in sample means $(\bar{x}_1 + \bar{x}_2)$ is the sum of their variances $(S^2_{\bar{x}_1} + S^2_{\bar{x}_2})$, where subscripts indicate the population. In equation 5, the standard error used to divide the mean difference is the square root of the sum of the variances associated with $x_1$ and $x_2$:

$$S_{\bar{d}} = \sqrt{S_1^2 + S_2^2} \tag{1.6.6}$$

**where:**
$S_{\bar{d}}$ = standard error of the difference between sample means,
$S_1^2$ = variance from sample population $x_1$,
$S_2^2$ = variance from sample population $x_2$.

## Calculation Results

In many texts, the standard error of the difference is calculated as:

$$S_{\bar{d}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{1.6.7}$$

**where:**
$S_{\bar{d}}$ = standard error of the difference between sample means,
$S_1^2$ = variance from sample population $x_1$,
$S_2^2$ = variance from sample population $x_2$
$n_1$ = sample size for population 1,

$n_2$ = sample size for population 2.

This is because the variance of population 1 is calculated as:

$$S_1^2 = \frac{S_1^2}{n_1} \tag{1.6.8}$$

and likewise for population 2.

Since both samples have variability involved, both variables must be accounted for. The number of degrees of freedom associated with this difference is either calculated by the computer, or if done by hand, is the degrees of freedom associated with the smaller sample. For samples of equal size, and if we can assume $S_1^2 = S_2^2$, the number of degrees of freedom is 2(n-1).

After calculating the *t*-value of the difference in which we are interested, we then look up the critical table *t*-value, for which there is only a 5% chance of a larger value occurring. Generally, we have critical *t*-values of about 2, or slightly higher, for a 0.05-level two-tail test. If the *t*-value of our difference exceeds the critical *t*-value, then we say that we have a significant difference between the two means. Would a significant *t*-test indicate that the two populations are truly different?

## Try This: t-Test Exercises

There are three classes of t-tests that are commonly used in agronomic experiments (which can be calculated using Excel:

**Ex. 2: *t*–test assuming equal variances**

There is a more direct approach to comparing the means of two sample populations – the *t*-test. In our alfalfa example, suppose the populations have equal variances, and what we really want to know is whether or not the two sample means are different from one another. This question can be stated as a simple statistical hypothesis: $H_0$:$\mu_1$=$\mu_2$ vs. the alternative $H_a$: $\mu_1 \neq \mu_2$. In this exercise, we will use the *t*-test procedure in Excel to test this hypothesis.

**Exercise 2**

- Perform a *t*-test to compare mean protein concentrations of the two alfalfa lots using the following steps (Figure 3): Open the file created in QM-mod6-ex1data.xls.
- Label a new column, '*t*-test: Equal variance,' and in the first row under the title, enter the formula "=T.TEST(C2:C9,C10:C17,2,2)".

This dictates that the observations for the first sample are in C2:C9, the second set are in C10:C17, it is a two-tailed test, and that this is an independent sample *t*-test assuming equal variances.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | Confidence Interval Calculation | Lower Limit of CI | Upper Limit of CI | t-test: Equal Variances (Pr > t) | | | | |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | 0.402879316 | 15.17 | 15.97 | 0.056441163 | | | | |
| 3 | 1 | 2 | 14.77 | | | | | | | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | 1.145124162 | 13.36 | 15.65 | | | | | |
| 11 | 2 | 2 | 15.86 | | | | | | | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | | | | | | | |

Figure 3 Calculation of a t-test statistic.

- Read the results of the *t*-test (assuming equal variances) as Prob > |t|. This is the probability, assuming the null hypothesis is true, of observing the results in your data set, and is 0.0564.
- Note that the *t*-test is assuming equal variances and has a two-tailed alternative. To convert the probability to a one-tailed alternative, divide the two-tailed probability by 2. It is 0.0564/2 = 0.0282.
- Do the means differ? (Is the Lot 1 mean significantly lower than the Lot 2 mean based on these samples?)

**Ex. 3: t-Test Assuming Unequal Variances**

You may have noticed in the *t*-test output from the last exercise that the standard error for Lot 2 is about three times as great as that for Lot 1.

One of the assumptions we made for the *t*-test was that the two sample populations shared a common variance. Hence, a pooled estimate of the population variance was used in the test.

We can easily test the hypothesis that the two variances are equal by computing a simple F test (Figure 4):

- Open the QM-mod6-ex1data.xls file.
- Calculate the variance for each lot in a new column using the formula Var.S.
- Divide the larger of the two variances by the smaller one to calculate F.
- Determine the critical F value. (Numerator and denominator df are those from the two samples.) The alpha level is 0.05.
- If the calculated F is greater than the critical F, we conclude that the variances are different and use another *t*-test.
- If the calculated F is less than the critical F, we conclude that the variances are the same and that a pooled variance was appropriate for the *t*-test.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | Confidence Interval Calculation | Lower Limit of CI | Upper Limit of CI | t-test: Equal Variances (Pr > t) | Variance | F-value (ratio of variances) | Pr > F (probability of a value of F or greater by chance) | t-test: Unequal Variances |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | 0.402879316 | 15.17 | 15.97 | 0.056441163 | 0.23 | 8.078955 | 0.003911 | 0.068367 |
| 3 | 1 | 2 | 14.77 | | | | | | | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | 1.145124162 | 13.36 | 15.65 | | 1.88 | | | |
| 11 | 2 | 2 | 15.86 | | | | | | | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | | |

Figure 4 Calculation of t-value for unequal variances.

For our example, the calculated F is (1.88/.23) = 8.17 and the critical F = 3.79. Therefore, we conclude that the variances of the two sample populations are different and a different *t*-test is required (Figure 5).

- It is possible to find a p-value for this test using “=F.DIST.RT(M2,8,8)”. M2 is the cell with the F-ratio in it.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | Confidence Interval Calculation | Lower Limit of CI | Upper Limit of CI | t-test: Equal Variances (Pr > t) | Variance | F-value (ratio of variances) | Pr > F (probability of a value of F or greater by chance) | t-test: Unequal Variances |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | 0.402879316 | 15.17 | 15.97 | 0.056441163 | 0.23 | 8.078955 | 0.003911 | 0.068367 |
| 3 | 1 | 2 | 14.77 | | | | | | | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | 1.145124162 | 13.36 | 15.65 | | 1.88 | | | |
| 11 | 2 | 2 | 15.86 | | | | | | | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | | |

Figure 5 Calculation of t-value for one tail test.

Note, however, that this test and other tests for equality of variance are not robust procedures. We saw earlier that the *t*-test is reasonably robust. It returns fairly accurate probability statements for samples from similarly shaped distributions when sample sizes are nearly equal. The F-test for variances, however, is very sensitive to non-normal distributions.

**Exercise 3**

Perform a t-test to compare mean protein concentrations of the two alfalfa lots. This is the same hypothesis as Exercise 2: $H_0$:$\mu_1 = \mu_2$ vs. the alternative $H_\alpha$: $\mu_1 \neq \mu_2$ Use the following steps:

- Label a new column '*t*-test: Unequal Variances' (Figure 6).
- This time, we again look for the P>|t| for the test of the null hypothesis that the means for the two lots are equal.
- Use the same formula from Exercise 2, but dictate that the required test assumes unequal variances.
- For the test with unequal variance, P = 0.0684. This is again for the two-tailed alternative, by default.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Lot | Rep | Crude Protein | | Average | Std Dev | Std Err | Confidence Interval Calculation | Lower Limit of CI | Upper Limit of CI | t-test: Equal Variances | Variance | Bartlett's Test | F-test | t-test: Unequal Variances |
| 2 | 1 | 1 | 16.08 | | 15.57 | 0.481901 | 0.170378 | 0.402879316 | 15.17 | 15.97 | 0.056441163 | 0.23 | 8.078955 | 0.003911 | 0.068367 |
| 3 | 1 | 2 | 14.77 | | | | | | | | | | | | |
| 4 | 1 | 3 | 15.03 | | | | | | | | | | | | |
| 5 | 1 | 4 | 15.95 | | | | | | | | | | | | |
| 6 | 1 | 5 | 15.58 | | | | | | | | | | | | |
| 7 | 1 | 6 | 15.39 | | | | | | | | | | | | |
| 8 | 1 | 7 | 15.70 | | | | | | | | | | | | |
| 9 | 1 | 8 | 16.06 | | | | | | | | | | | | |
| 10 | 2 | 1 | 12.69 | | 14.5025 | 1.369731 | 0.484273 | 1.145124162 | 13.36 | 15.65 | | 1.88 | | | |
| 11 | 2 | 2 | 15.86 | | | | | | | | | | | | |
| 12 | 2 | 3 | 13.12 | | | | | | | | | | | | |
| 13 | 2 | 4 | 16.31 | | | | | | | | | | | | |
| 14 | 2 | 5 | 15.64 | | | | | | | | | | | | |
| 15 | 2 | 6 | 14.62 | | | | | | | | | | | | |
| 16 | 2 | 7 | 13.27 | | | | | | | | | | | | |
| 17 | 2 | 8 | 14.51 | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | | | |
| 19 | | | | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | | | | |

Figure 6 Calculation of Bartlett's test statistic.

**Ex. 4: Paired t-test**

Sometimes data collected from two sample populations are related in some way. When this occurs we say that the observations are paired. A good example of this is the strip design that is often used in on-farm trials to compare two treatments. Treatments are applied to strips that run across a field in such a way that every adjacent pair of strips contains both treatments. Treatments are randomly allotted to strips within each pair. For a more complete description of this methodology read the paper by **Exner and Thompson**.

**Exercise 4**

- Perform a paired t-test to determine if starter fertilizer had an effect on corn yield using the following steps:
- Download and open the Excel file QM-mod6-ex4data.xls.
- We will first calculate the paired t-test step-by-step (Figure 7).
- Calculate the difference between each pair of observations.
- Calculate the mean of the differences.
- Now calculate the standard deviation, using Stdev.S.
- The standard error is the standard deviation divided by the square root of the number of pairs.
- The t-statistic is the mean divided by the standard error.
- The calculated value is less than the tabular value, so we fail to reject the null. There is not a significant difference between the two treatments.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Pair | Starter Fert Yield (kg/ha) | No Fert Yield (kg/ha) | Difference | Mean Diff | StDev Diff | Std Err Diff | Tabular Statistic: 5 df, alpha = 0.05/2 | t-statistic | Paired t-test (Pr >t) |
| 2 | 1 | 5945.0 | 5895.0 | -50.0 | 280.8 | 436.9487 | 178.38356 | 2.571 | 1.57432296 | 0.176227685 |
| 3 | 2 | 5600.0 | 6255.0 | 660.0 | | | | | | |
| 4 | 3 | 5035.0 | 5995.0 | 960.0 | | | | | | |
| 5 | 4 | 5540.0 | 5745.0 | 205.0 | | | | | | |
| 6 | 5 | 5995.0 | 5845.0 | -150.0 | | | | | | |
| 7 | 6 | 5925.0 | 5985.0 | 60.0 | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | | | | | | | | | |
| 10 | | | | | | | | | | |
| 11 | | | | | | | | | | |
| 12 | | | | | | | | | | |
| 13 | | | | | | | | | | |
| 14 | | | | | | | | | | |
| 15 | | | | | | | | | | |

Figure 7 Calculation of paired t-test statistic.

- Excel also contains a formula to calculate a paired t-test. It is essentially the same as the one used in Exercises 2 and 3: “=T.TEST(B2:B7,C2:C7,2,1)” (Figure 8). However, the two arrays are the sets of observations from the two treatments sorted by pairs, and the last variable is one to indicate a paired test.

Did the use of starter fertilizer have any effect? No, since Prob > |t| = 0.18, we do not reject the null hypothesis and conclude there is no statistically significant effect.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Pair | Starter Fert Yield (kg/ha) | No Fert Yield (kg/ha) | Difference | Mean Diff | StDev Diff | Std Err Diff | Tabular Statistic: 5 df, alpha = 0.05/2 | t-statistic |
| 2 | 1 | 5945.0 | 5895.0 | -50.0 | 280.8 | 436.9487 | 178.38356 | 2.571 | 1.57432296 |
| 3 | 2 | 5600.0 | 6255.0 | 660.0 | | | | | |
| 4 | 3 | 5035.0 | 5995.0 | 960.0 | | | | | |
| 5 | 4 | 5540.0 | 5745.0 | 205.0 | | | | | |
| 6 | 5 | 5995.0 | 5845.0 | -150.0 | | | | | |
| 7 | 6 | 5925.0 | 5985.0 | 60.0 | | | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | |
| 17 | | | | | | | | | |
| 18 | | | | | | | | | |
| 19 | | | | | | | | | |
| 20 | | | | | | | | | |

Figure 8 Calculation of t-test value for paired t-test.

### Study Questions 5 and 6 : Continuous Data

#### Query 1.6.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Two-Sample Hypothesis Testing

There are three sets of hypotheses that can be tested with a t-test (Table 2).

*Table 2 Hypothesis tests*

| Set | Null Hypothesis | Alternative Hypothesis |
|---|---|---|
| 1 | $\mu_1 = \mu_2$ | $\mu_1 \neq \mu_2$ |
| 2 | $\mu_1 \geq \mu_2$ | $\mu_1 < \mu_2$ |
| 3 | $\mu_1 \leq \mu_2$ | $\mu_1 > \mu_2$ |

The first set simply states that there is either a difference between the two means or there is not. The alternative hypothesis gives no regard to how the two means may differ but simply states that they are different in some way. We use this hypothesis when we do not have a preconceived idea as to how the means will differ. For example, if we are comparing yields of two highly recommended varieties, we probably have no expectation of which will be greater. This type of comparison is called a two-tailed test because we want to know if the observed t lies within either tail of the distribution (Figure 9). Assuming a mean difference for the two varieties of 250 kg/ha, a standard error of 120, and 8 df the test would look like this:

\[t=\frac{\bar{d}}{S_{\bar{d}}} = \frac{250}{120}=2.083\lt {\text{}t_

ParseError: EOF expected (click for details)

```
Callstack:
    at (Bookshelves/Agriculture_and_Horticulture/Quantitative_Methods_for_Plant_Breeding_(Suza
```

= 2.306} \label{8}\]

## Test Conclusion

In this example, since our calculated *t* is less than the tabular *t*, we fail to reject the null hypothesis and conclude that the means are not significantly different at alpha 0.05.

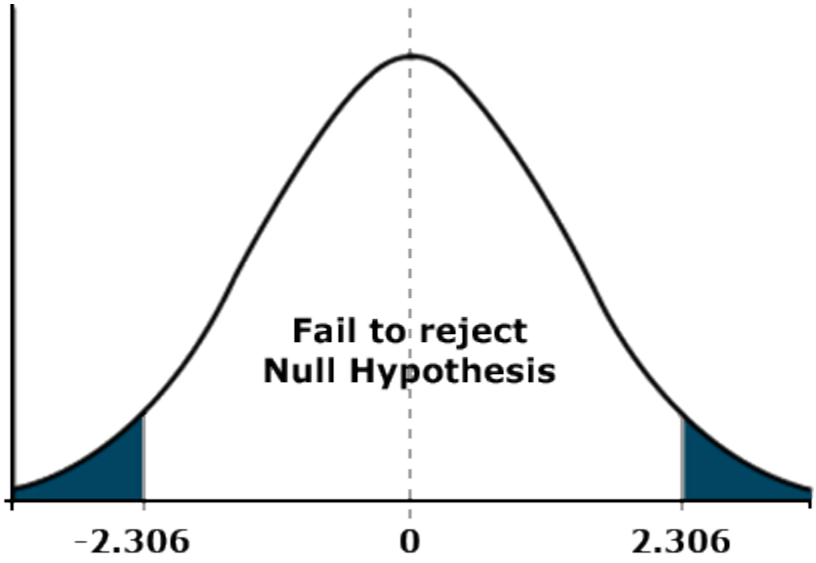



Figure 9 Two-tailed t-test acceptance region.

### Study Question 7: Continuous Data

#### Query 1.6.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Direction of the Mean Difference

With the other two sets of hypotheses, we consider the nature or direction of the mean difference. For the alternative hypothesis, we want to know if the mean for population 1 is significantly smaller (set 2) or larger (set 3) than the mean of population 2. We use one of these hypotheses when we have an expectation about how the means will differ. For example, if we are comparing the yield of a new variety with that of an older control variety, we expect the new one to outperform the old one. This type of comparison is called a one-tailed test (Figure 10) because we want to know if the observed *t* lies within one tail of the distribution. Assuming a mean difference for the two varieties of 250 kg/ha, a standard error of 120, and 8 df, the one-tailed test would look like this:

\[t=\frac{\bar{d}}{S_{\bar{d}}} = \frac{250}{120}=2.083\lt {\text{}t_

ParseError: EOF expected (click for details)

```
Callstack:
```

= 1.860} \label{9}\]

## Calculate t Results

In this case, the calculated *t* exceeds the tabular *t,* and we conclude that the mean difference of 5 bu/acre is significant. The tabular *t* for a one-tailed test can be obtained from Appendix 3 in your text. However, the probability values listed in Table 1 are for a two-tailed test. To convert these values to probabilities for a one-tail test, divide them by 2. For our example above, you would use the *t*-values listed under 0.100 because 0.100 divided by 2 equals 0.05. The reason *t*-values differ for one and two-tailed tests has to do with the shaded areas of the curve. In the two-tailed test above, each shaded area represents 2.5% of the total area under the curve. For the one-tailed test, the shaded area under the right tail represents 5% of the total area.

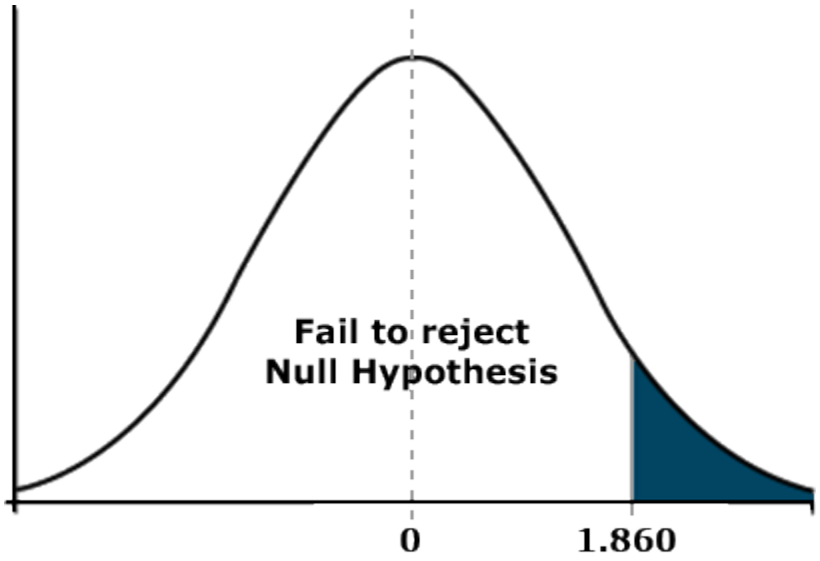



Figure 10 One-tailed t-test acceptance region.

## Linear Additive Model

Underlying the two sample experiment is a linear additive model. Linear models are mathematical representations of the variables contributing to the value of an observation. The observed response is partitioned into effects due to treatments and the experimental units. That is, each measured response is an additive function of the treatment effect and a random effect error associated with the experimental unit on which the measurement was made.

The linear model for the *t*-test is:

$$Y_{ij} = \mu + T_i + \varepsilon_{(i)j} \tag{1.6.9}$$

**where:**
$S_{\bar{d}}$ = standard error of the difference between sample means,
$Y_{ij}$ = response observed for the $ij^{th}$ experimental unit,
$\mu$ = overall population mean,
$T_i$ = effect of the $i^{th}$ treatment,
$\varepsilon_{(i)j}$ = effect associated with the $ij^{th}$ experimental unit; commonly referred to as error}.

## Assumptions for Linear Models

Now let's look at an example to see if we can make better sense of all this. A sales representative puts out a demonstration plot to compare a new herbicide product with an older one. Being a clever person, she has replicated each herbicide treatment four times. She, therefore, has eight plots and will have eight observations for each variable measured. The linear additive model Equation 11), for example, plot 3 of this experiment says the observed response (number of weeds) in the plot equals the overall

mean response for the population plus an effect due to the herbicide applied plus a random effect due to the plot itself. This latter effect acknowledges that there will be some differences in the measured response among plots receiving the same herbicide treatment.

$$\text{Number of weeds in plot 3} = \text{Average of all plots} + \text{Effect of Herbicide 2} + \text{Plot 3 error} \tag{1.6.10}$$

### Essential assumptions for linear models:

- Additivity – the effect of each term is additive.
- Treatment effects are constant for all experimental units (plots).
- The effects of plots are independent and not related to treatment effects.

The linear additive model for an experiment determines how the data are to be analyzed. We will revisit this topic in later lessons as we discuss the analysis of variance.

# Confidence Limits

## Least Significant Difference

The least significant difference (LSD) allows an easy test for the difference between two treatment means.

Earlier, we established confidence limits for our sample mean. The same procedure can be applied to a difference between two means, using the t-value of the difference between the two means. The confidence interval for the difference between two means is:

$$\text{CL} = \bar{d} \pm t_{S_{\bar{d}}} \tag{1.6.11}$$

**where:**
$\bar{d}$ = the difference between sample means,
$S_{\bar{d}}$ = standard error of the difference.

Confidence intervals (CI) for the true mean difference $\mu_d$ are obtained by computing the difference in means, d, and adding and subtracting t standard errors of d. To test the null hypothesis that $\mu_d$ is zero, we just see if zero is within the CL, and fail to reject the null if it is. This actually gives us an easy way to test the null hypothesis of zero difference: subtract the means and see if the difference is larger than $ts_d$.

The least significant difference (LSD) between two sample means is simply this difference:

$$\text{LSD} = t_{S_{\bar{d}}} \tag{1.6.12}$$

**where:**
$S_{\bar{d}}$ = the standard error of the difference,
$t$ = t-value for the appropriate df and probability.

## Comparing Several Means

But what if we are comparing not one or two means but several? For example, what if we were comparing the grain yield of five corn varieties? We would then have to calculate the differences among all varieties (Table 3).

*Table 3 Potential comparisons between five corn varieties.*

| | | | |
|---|---|---|---|
| Variety 1 vs. Variety 2 | Variety 1 vs. Variety 3 | Variety 1 vs. Variety 4 | Variety 1 vs. Variety 5 |
| Variety 2 vs. Variety 3 | Variety 2 vs. Variety 4 | Variety 2 vs. Variety 5 | n/a |
| Variety 3 vs. Variety 4 | Variety 3 vs. Variety 5 | n/a | n/a |
| Variety 4 vs. Variety 5 | n/a | n/a | n/a |

That is, ten different comparisons from a relatively simple experiment! Would it not be nice if we could know whether there were any significant differences to be found before we conducted this series of comparisons? As we will learn later, there is such a way.

## Try This: Determining the LSD

**Ex. 5: Determining the LSD**

Having reached the conclusion that the 280 kg/ha difference between strips receiving starter fertilizer and those that did not was nonsignificant, you might be wondering just how large a mean difference it would take to be declared statistically different. We can answer this question easily by calculating the least significant difference (LSD).

The LSD is the smallest mean difference that can be declared significantly different from zero. Its value depends on the standard error of the mean difference (SED) and the number of degrees of freedom associated with it. The formula for calculating the LSD is [$t$ x (SED)], where $t$ is the tabular $t$ for the error df at whatever level of significance is desired. In agronomic research, we often use the 0.05 probability level for this purpose.

**Exercise 5**

Calculate the LSD for the starter fertilizer experiment using the following steps (Figure 11):

- From the Exercise 4 Excel workbook of moments for the differences (QM-mod6-ex4data.xls), find the standard error of mean difference (SED = 3.57).
- From the $t$-table, either at the start of this unit or in Appendix 3, find the tabular $t$ for a two-tailed probability of 0.05. Remember, this is based on (6-1) = 5 df and the alpha level of 0.05 must be divided by two for a two-tailed test.
- Multiply the table $t$-value (2.571) and the SED to get the LSD = 9.18.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Pair | Starter Fert Yield (kg/ha) | No Fert Yield (kg/ha) | Difference | Mean Diff | StDev Diff | Std Err Diff | Tabular Statistic: 5 df, alpha = 0.05/2 | t-statistic | Excel Test | LSD |
| 2 | 1 | 5945.0 | 5895.0 | -50.0 | 280.8 | 436.9487 | 178.38356 | 2.571 | 1.57432296 | ### | 458.6241 |
| 3 | 2 | 5600.0 | 6255.0 | 660.0 | | | | | | | |
| 4 | 3 | 5035.0 | 5995.0 | 960.0 | | | | | | | |
| 5 | 4 | 5540.0 | 5745.0 | 205.0 | | | | | | | |
| 6 | 5 | 5995.0 | 5845.0 | -150.0 | | | | | | | |
| 7 | 6 | 5925.0 | 5985.0 | 60.0 | | | | | | | |
| 8 | | | | | | | | | | | |
| 9 | | | | | | | | | | | |
| 10 | | | | | | | | | | | |
| 11 | | | | | | | | | | | |
| 12 | | | | | | | | | | | |
| 13 | | | | | | | | | | | |
| 14 | | | | | | | | | | | |
| 15 | | | | | | | | | | | |
| 16 | | | | | | | | | | | |

Figure 11 LSD calculation.

# Summary

## T-Distribution

- Used for sample means when variance is not known, but estimated.
- Depends on degrees of freedom (df).
- As df gets large, $t$-distribution becomes same as normal.

## T-Values

- Like $z$-value, but used when we estimate variance.
- Tells how many standard errors from the mean.

## Confidence Intervals (CI)

- Give range for the population mean μ.
- Interval is Y ± *t* standard errors.
- *t* – value depends on df (usually n-1) and on probability in the tails.

## T-Tests

- Can test for difference in two treatment means.
- Test is $t = \frac{d-\mu_d}{S_d}$.
- Paired *t* – test has sd computed from differences.
- Independent samples t – test has $S_d^2 = (\frac{S_1^2}{n_1}) + (\frac{S_2^2}{n_2})$

## Least Significant Difference

- Allows easy test for difference in two treatment means.
- LSD = $t\sqrt{\frac{2S^2}{n}}$ for independent samples of size n.

**How to cite this chapter**: Mowers, R., K. Moore, M.L. Harbur, K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Continuous Data. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods.* Iowa State University Digital Press.

This page titled 1.6: Continuous Data is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Ken Moore; M. L. Harbur; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.7: Linear Correlation, Regression and Prediction

Determining the relationship between two continuous variables can help us to understand a response to an associated action. The concept of linear correlation can illustrate a possible relationship between two variables. For instance, the ideas that more rainfall and more fertilizer available to a crop produce greater yield are very plausible. To determine whether a relationship exists statistically, employ the use of linear models. Once a relationship is established, methods of linear regression can be used to quantify the amount of response and strength of the relationship, such as finding that 5 cm of additional precipitation produces a 30 kg $ha^{-1}$ yield increase or applying 10 kg $ha^{-1}$ less N reduces yields by 40 kg $ha^{-1}$. For students of plant breeding, the concepts of regression and prediction will be fundamental to understanding Quantitative Genetics and Breeding Values.

## Learning Objectives

- The proper use of and differences between correlation and regression
- How to estimate a correlation relationship from a scatter plot
- How to establish a linear relationship between a dependent variable and an independent variable using regression methods

## Correlation

**Correlation is a measure of the strength and direction of a linear relationship.**

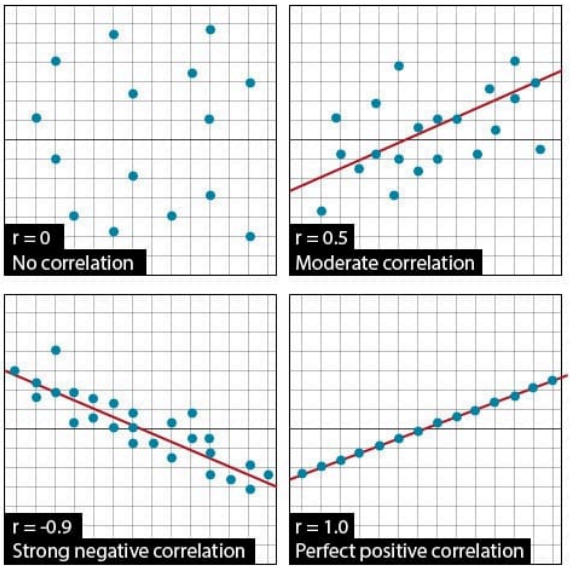



Figure 1 Example scatter plots and the r-values associated with them.

The Pearson Correlation Coefficient (r), or correlation coefficient for short, is a measure of the degree of linear relationship between two variables. The measure determines how close to linear is the change in one variable with respect to the other. The emphasis is on the degree to which they vary linearly. Later in this lesson, we will discuss regression, where the interest is in the rate of change, and how one variable is predicted by the other. In correlation, the strength of the relationship is of interest.

The correlation coefficient may take any value between 1 and –1.

The sign of the correlation coefficient (+, –) defines the direction of the relationship, either positive or negative. A positive relationship means that a positive change in one variable is related to a corresponding positive change in the other (e.g. more fertilizer produces more yield), while a negative relationship produces a negative result (e.g. increasing numbers of black cutworms decreases yields).

The absolute value of the correlation coefficient describes the strength of the relationship. A correlation coefficient of 0.50 indicates a stronger degree of linear relationship than one of r = 0.40. Likewise, a correlation coefficient of r = –0.50 indicates a greater degree of relationship than one of r = –0.40. Thus, a correlation coefficient of r = 0.0 indicates the absence of a linear relationship; correlation coefficients of r = +1.0 and r = –1.0 indicate perfect linear relationships (Figure 1).

## Scatter Plots

A straightforward and necessary way to visualize correlations is through the use of scatter plots. Usually, the dependent variable is plotted on the vertical axis of the plot while the other variable is plotted on the horizontal axis. Such a plot can provide evidence of a linear relationship between the variables. An example is shown below in Figure 2.

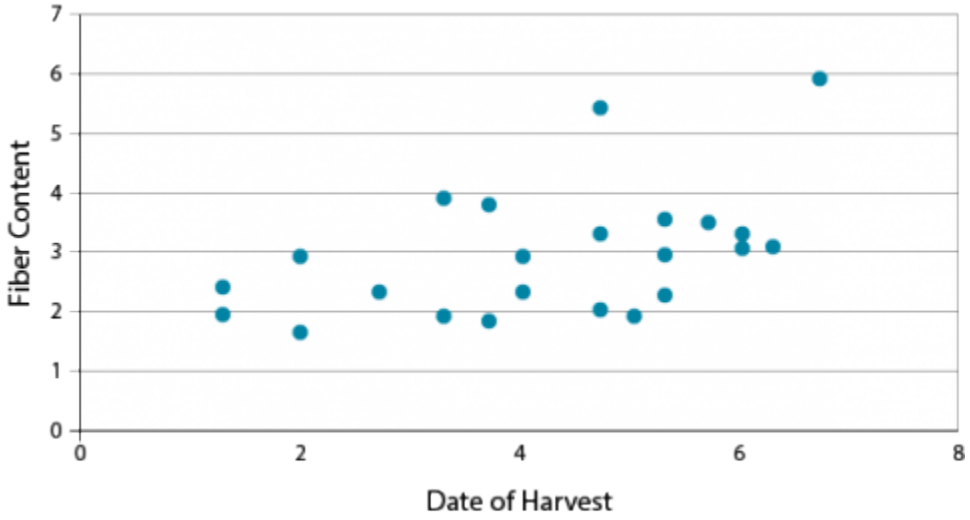



Figure 2 Fiber content plotted as compared to the date of harvest. The fiber content seems to increase linearly as the date of harvest increases.

### Try This: Correlation (animation video)

How well related are these two measurements? What is their correlation coefficient?

This animation estimates some correlations by drawing an approximate visual best-fit line (blue) using randomly generated data sets. Then the “Show” reveals the least-squares regression line (red).

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Video online

Study Question

Query 1.7.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Correlation: Calculating r

Correlation is an often misused concept and statistic. When two things are correlated, what does that really mean? Misconceptions about correlations between variables are common. Correlations can also be totally spurious. For example, a positive relationship between the number of sheep in the United States and the number of golf courses does not mean that sheep numbers have increased because there are more golf courses. Both variables are likely to be related to an underlying trend of increasing population in the U.S. Many things can be correlated, but it is the physical or biological relationship that gives a correlation relevance. Correlation only states the degree of linear association (not cause and effect) between the two variables.

Calculation of r involves estimating the covariance of two variables or how much they vary together. The correlation is defined in the equation below, where one of the variables is represented by x and the other by y.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}x}\sqrt{S_{yy}}} \tag{1.7.1}$$

**where:**
$S_{xy}$ = sum of products = $y = \sum xy - \frac{\sum x \sum y}{n}$,
$S_{xx}$ = sum of squares of $x = \sum x^2 - \frac{\sum x^2}{n}$ ,
$S_{yy}$ = sum of squares of $y = \sum y^2 - \frac{\sum y^2}{n}$.

The correlation equation may seem monstrous at first. Do not panic! Actually, the concept behind the equation is closely related to the z-scores we calculated earlier.

The numerator, the sum of squares of xy (Sxy), measures the combined distances of all points from the center of the plot $(\bar{x}, \bar{y})$. The more closely X and Y are related, the greater this value will be.

The denominator is the product of the square roots of the sums of squares of X and Y. The product of these two roots quantifies how much X and Y vary independently of each other.

Thus, r is the ratio of the amount that X and Y vary together to the amount X and Y vary in total. The more X and Y vary together, the greater the ratio will be. The maximum possible values (1 or -1) occur when all variation in X and Y is related.

How individual variables vary is of interest. If large Y's are associated with large X's, it would stand to reason that there would be a positive correlation between the variables.

## Correlation Example

Some measurements were taken on the amount of flow, Y ($m^3$/s), in a normally dry drainage ditch next to a field. These were measured from run-off after 30 minutes rainfalls. The total rainfalls were designated as X and measured in millimeters (mm). Hydrologists wanted to know how much water ran off under different conditions and how closely the two measurements were related in this field. The relevant sums are included along with the computational form of the *r* calculation (Table 1).

*Table 1 Total rainfall (X, mm) and amount of flow (Y, $m^3$/s ) in a drainage ditch.*

| x | $x^2$ | y | $y^2$ | xy |
|---|---|---|---|---|
| 2.6 | 6.76 | 0.1 | 0.01 | 0.26 |
| 12.2 | 148.84 | 1.3 | 1.69 | 15.86 |
| 14.1 | 198.81 | 2.5 | 6.25 | 35.25 |
| 14.6 | 213.04 | 3.5 | 12.25 | 51.1 |
| 15.2 | 231.04 | 9.1 | 82.81 | 138.32 |
| 15.6 | 243.36 | 9.3 | 86.49 | 145.08 |
| 15.9 | 252.81 | 12.2 | 148.84 | 193.98 |
| 17.4 | 302.76 | 13.2 | 174.24 | 229.68 |
| 18.8 | 353.44 | 15.9 | 252.81 | 298.92 |
| 19.0 | 361.0 | 19.3 | 372.49 | 366.7 |
| $\Sigma x$ = **145.4** | $\Sigma x^2$ = **2311.98** | $\Sigma y$ = **86.4** | $\Sigma y^2$ = **1137.88** | $\Sigma xy$ = **1475.15** |

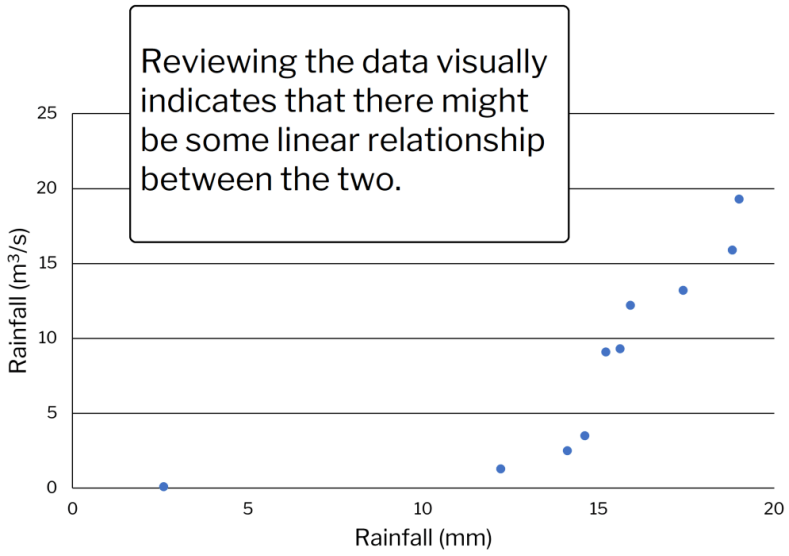



Figure 3 Runoff from a field as a function of rainfall.

## Correlation Example Calculations

Using Equation 1 and the data in Table 1, $r$ is calculated as shown below.

$$\text{Sum of products, } S_{xy} = \sum \text{xy} - \frac{\sum \text{x} \sum \text{y}}{\text{n}} = 1476.8 - \frac{(145.5)(86.5)}{10} = 218.2$$

$$\text{Sum of squares of } x, S_{xx} = \sum \text{x}^2 - \frac{(\sum \text{x}^2)}{\text{n}} = 2313.4 - \frac{145.5^2}{10} = 196.4$$

$$\text{Sum of squares of } y, S_{yy} = \sum \text{y}^2 - \frac{(\sum \text{y}^2)}{\text{n}} = 1139.4 - \frac{86.5^2}{10} = 391.2$$

$$r \tag{1.7.2}$$

The computed r-value is 0.79. This is a moderately large correlation. How large the correlation is, depends upon the variability of the data. Correlations can range well above 0.9 or below -0.9 in many cases. Physically, there would seem to be a cause and effect here (Figure 3). Heavier rainfall would produce more run-off, while light rainfall produces little or none. The best indicator of

that can be seen at the heaviest rainfall rates, where the magnitude of the run-off increases substantially. The one data point at lower rainfall levels is problematic. We assume it is real. Occasionally, a single outlier data point can slightly skew a relationship. Although not as linearly related to the other data, it does fit the plausible model: lighter rainfall, less run-off.

## Ex. 1: Calculating the Correlation in a Bivariate Set of Data

As shown in the examples displayed in the text, a good first guess in establishing a relationship between variables is to view the data on an X-Y scatter plot. Trends should start to appear. We will produce a scatterplot in Excel and determine the correlation.

- Open the QM-mod7-ex1data.xls workbook. It is July weather data with average temperature and precipitation data.

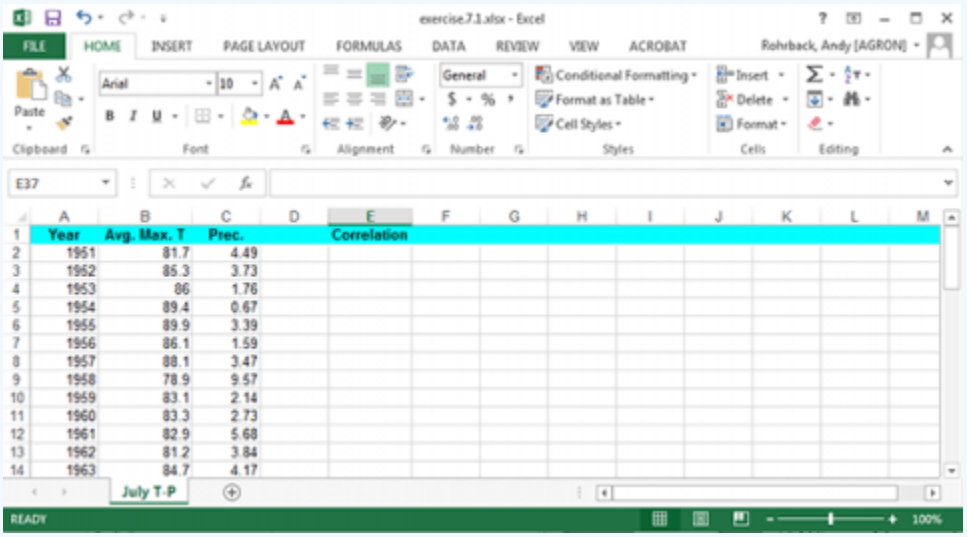



Figure 4 July weather data from.

**Ex. 1: Bivariate Set of Data (1)**

We will begin by producing a scatter plot to view the data. For this analysis, the Temperature will be assigned to the x-axis and Precipitation to the y-axis.

**Ex. 1: Bivariate Set of Data (2)**

Highlight the two columns with the precipitation and temperature data (Figure 5).

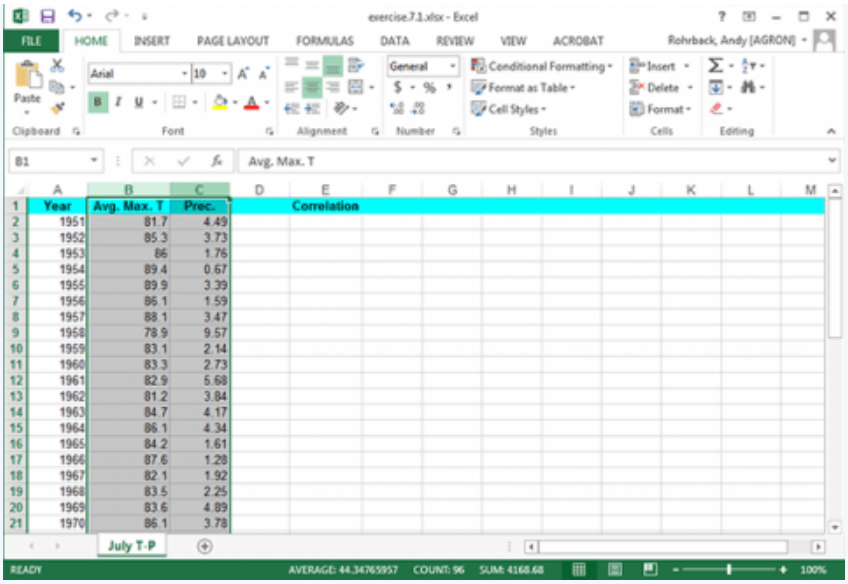



Figure 5 July weather data

Select the **Insert** tab and click the **Scatter Plot** tool. Select the first type (Figure 6).

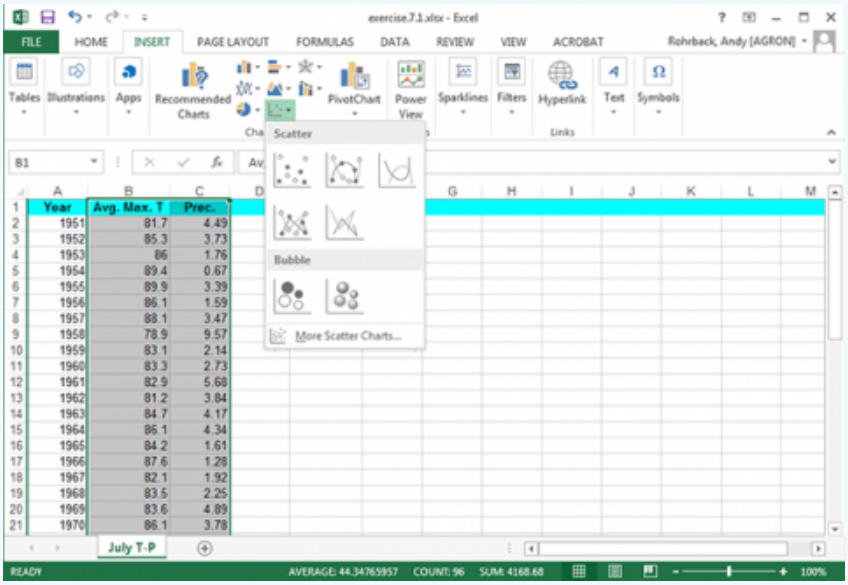



Figure 6 July weather data.

**Ex. 1: Bivariate Set of Data (3)**

Change the x-axis label to *Temperature*, the y-axis label to *Precipitation*, and the plot title to *Precipitation vs. Temperature* (Figure 7). Click on each text box in the plot to change it.

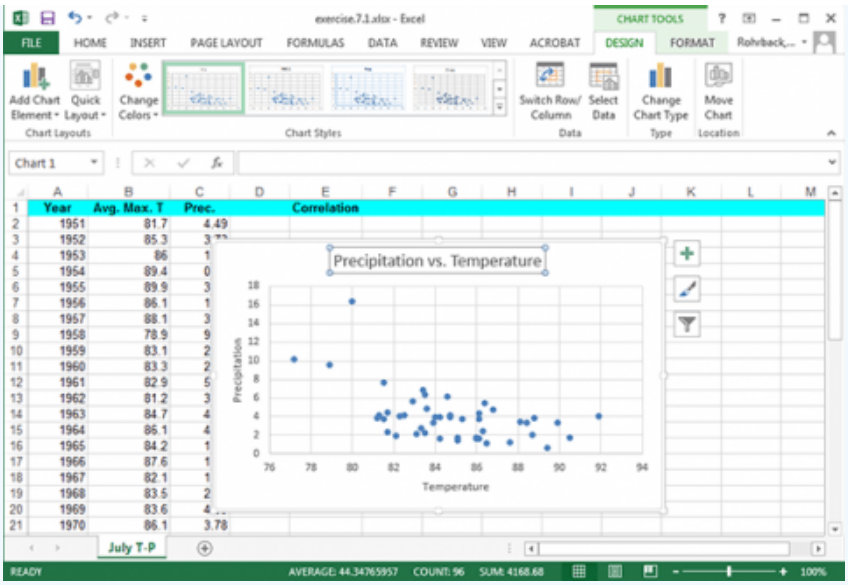



Figure 7 July weather data and scatter plot generated.

**Ex. 1: Bivariate Set of Data (4)**

The correlation between precipitation and correlation can be easily calculated. Label a fourth column "**Correlation**" and enter the formula "**=Correl(B2:B48, C2:C48)**" (Figure 8).

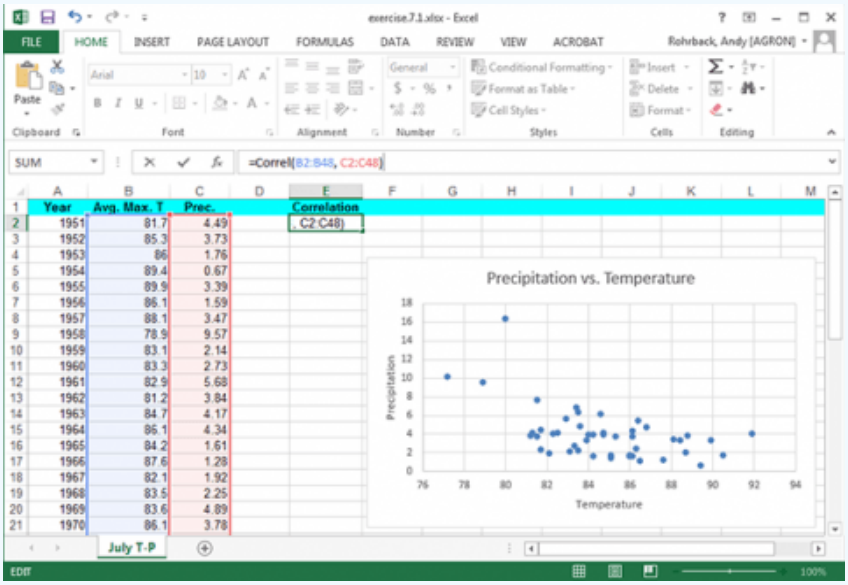



Figure 8 July weather data and scatter plot

**Ex. 1: Bivariate Set of Data (5)**

The correlation is moderately large (but not exceptionally high), with an r-value of -0.534. You may see the effects of an outlier here. The correlation between July temperatures and precipitation makes sense. This relationship follows a meteorological pattern.



Figure 9 Rain falls on crop fields. Photo by Malene, Wikimedia Commons.

More precipitation wets the soil surface, causing more latent heating and less warming of the air by sensible heating. More precipitation generally means more clouds Figure 9). Both are associated with lower temperatures. The single data value at the top may skew your view of the correlation while having a rather small effect on the total correlation. There does seem to be a qualitative relationship without a very strong correlation.

**Ex. 1: Bivariate Set of Data (6)**

Hold the cursor over the possible outlier in the Excel scatterplot Figure 10).

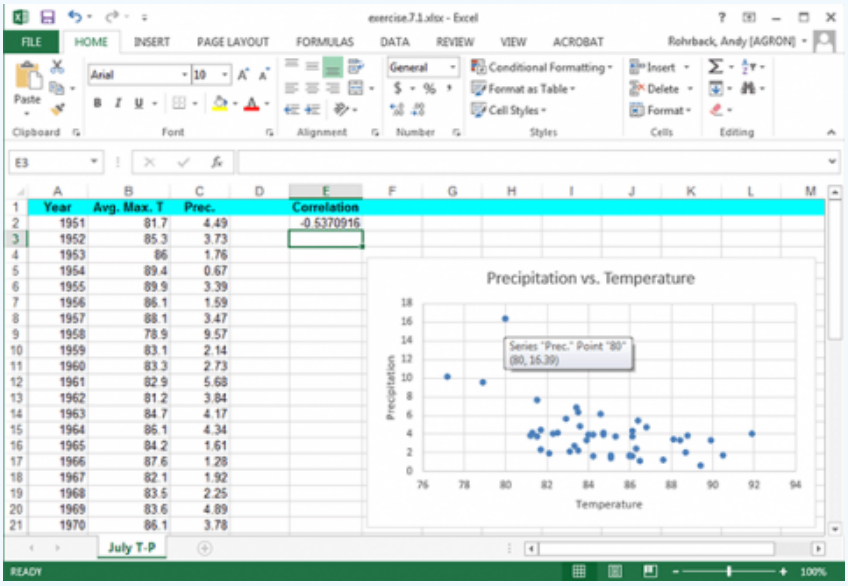



Figure 10 July weather data and scatter plot generated.

Select the row with 1993 data in the original data and delete it (Figure 11).

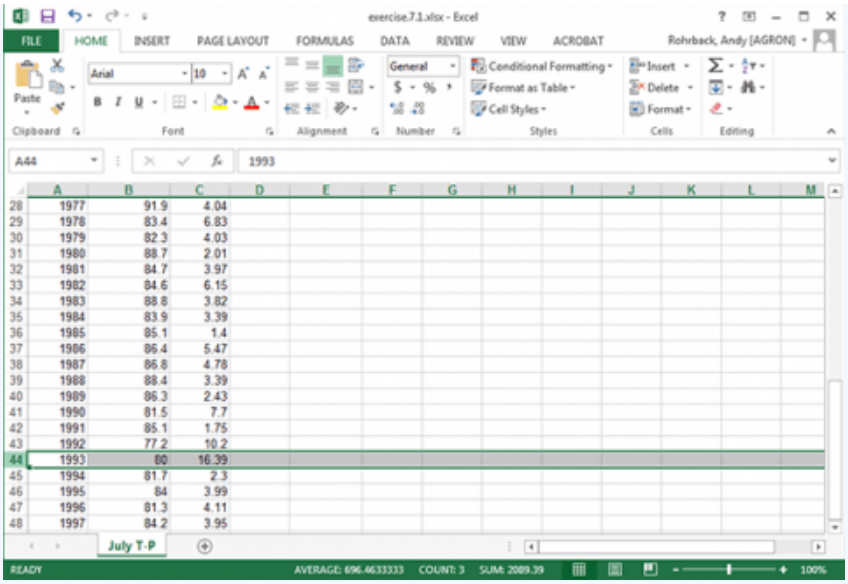



Figure 11 July weather data.

The plot and correlation will change (Figure 12). However, be careful when discarding what appears to be an outlier. It is a good idea to try and determine if there was an obvious experimental error that can be found. If there is not, it may not be a true outlier.

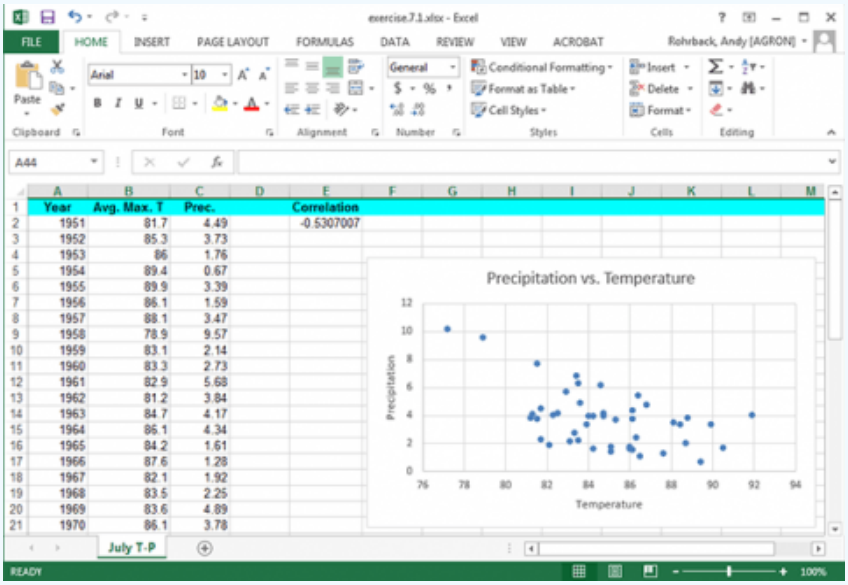



Figure 12 July weather data.

## Discussion: Correlation

How much did the correlation change by removing the 1993 data? What do you think about the results of this?

# Linear Regression

**Linear Regression** establishes a predictive relationship between two variables. While correlation attempts to establish a linear relationship between two variables, regression techniques try to determine a predictive relationship between the two. Translated, “Can values of one variable be used to predict values of the other?”

When someone wishes to apply fertilizer (Figure 13), the expected amount of yield gained for the amount of fertilizer applied is

needed. Sound economic and ecological choices may be based on regression and relationships between variables. Understanding the regression relationship allows the producer to use the amount of fertilizer that can give the best yield or financial return for the money invested.

Several other physical variables obviously are involved in translating the fertilizer into a yield result, such as rainfall, soil fertility, pest populations, etc.



Figure 13 Fertilizer application on a field. Photo by Iowa State University.

## Regression Lines

Referring to the scatter plot diagrams in the web exercise, one can estimate the magnitude of a correlation. When establishing a regression relationship, a single line delineating the relationship is necessary. One could use several methods to estimate the linear relationship that best fits the data.

Connecting the two endpoints in the data or eyeballing a resultant line are two examples. These will usually provide a qualitative result that lacks precision and accuracy.

In the animation above, we drew a line of “best fit” by eye and observed the least-squares regression line was different. The regression line is that line that minimizes the sum of squared vertical distances of points on the line. If you mentally determined the line minimizing the perpendicular distances, it would not be the same as the least-squares regression line. The regression line is “best” in the sense of least error for the line with fixed x-values.

# Sources of Variation

The preferred method to estimate a regression line is to use the data to numerically calculate the line which minimizes the error or the scatter of the points around the line. This is done using the least-squares method. (Figure 14)

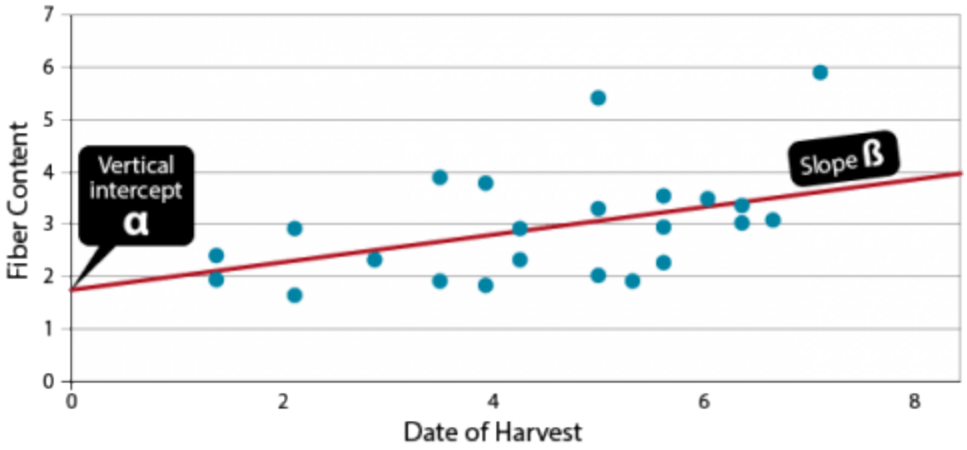



Figure 14 Regression line statistics.

The result of a linear regression is an equation of the form ($\hat{Y} = \hat{\alpha} + \hat{\beta}x$). The hats over Y, α, and β indicate these are estimates, not the actual regression line. This equation determines the relationship between the x and a predicted $\hat{Y}$ based on the estimated slope of the regression line and the vertical intercept $\hat{\alpha}$.

As we have discussed before, we do not know the actual relationship between the two variables. Therefore, we estimate it, based on gathered data. We assume there is a true regression line: y = α + βx + ε, and we estimate intercept with α and slope with $\hat{\beta}$.

## Point-Slope Formula

The point-slope formula to create the line can be found using sums of squares as calculated in the previous section. The slope of the line is determined using this equation.

$$\beta \tag{1.7.3}$$

**where:**
$S_{xy}$ = sum of products = $y = \sum xy - \frac{\sum x \sum y}{n}$,
$S_{xx}$ = sum of squares of $x = \sum x^2 - \frac{\sum x^2}{n}$.

Note the similarities to and distinct differences from the calculation of r. There are an infinite number of lines which can be described with this slope, thus another piece of information to describe a line is necessary.

## Y-Intercept Formula

A specific point on the line (usually the vertical-intercept) along with the slope fixes a single line to the data. The Y-intercept of the line is determined by the equation below.

$$\hat{\alpha} = \frac{\sum \mathrm{y} - \hat{\beta} \sum \mathrm{x}}{\mathrm{n}} \tag{1.7.4}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{1.7.5}$$

Another point which the line intercepts is the point ($\bar{\mathrm{X}}, \bar{\mathrm{Y}}$). Knowing a point on the line and its slope completely describes the regression line through the data.

## Example Calculation: Slope

Using the data from the previous section, we can calculate the regression slope using a hand computational formula in Equation 2.

$$\beta \tag{1.7.6}$$

$$\text{Sum of products, } S_{xy} = \sum \text{xy} - \frac{\sum \text{x} \sum \text{y}}{\text{n}} = 1476.8 - \frac{(145.5)(86.5)}{10} = 218.2$$

$$\text{Sum of squares of } x, S_{xx} = \sum \text{x}^2 - \frac{(\sum \text{x}^2)}{\text{n}} = 2313.4 - \frac{145.5^2}{10} = 196.4$$

$$\beta \tag{1.7.7}$$

The Y-intercept of the data can be calculated similarly.

$$\hat{\alpha} = \frac{\sum \text{y} - \hat{\beta} \sum \text{x}}{\text{n}} \tag{1.7.8}$$

$$\hat{\alpha} = \frac{86.5 - 1.11(145.5)}{10} \tag{1.7.9}$$

$$\hat{\alpha} = -7.4 \tag{1.7.10}$$

## Example Calculation: Interpretation

This line indicates that according to the measured data, the run-off will increase by 1.11 $m^3$/s for each additional mm of rainfall. The line created is the “best” in describing the linear response of run-off to the associated rainfalls (Figure 15).

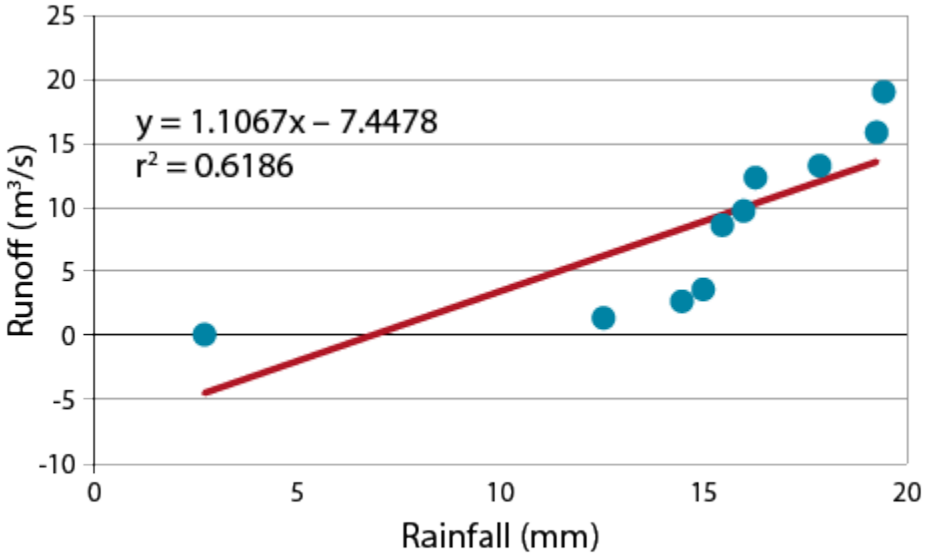



Figure 15 The best fit line for the rainfall-runoff data. The equation of the line and the r-squared value of the data are included.

The strength of the relationship is $r^2$, *i.e.*, the correlation coefficient squared. Note that the line fitted to the data intercepts the vertical axis at a negative value. An initial interpretation of “negative runoff” is clearly nonsense, but a little reflection on the nature of the problem suggests that up to a certain level of rainfall the water will infiltrate the soil before there is runoff. Thus the negative value can be interpreted as the “infiltration potential” of the soil. You may also notice that there is some bias in the way the data deviate from the regression line. The line overestimates the run-off for rainfalls from 10-15 mm and underestimates above 16 mm. This hints that a linear relationship may not be the best choice for this relationship.

### Ex. 2: Estimating Regression

**Exercise 2: Plotting Data to Estimate Regression**



Figure 16 Predictive analysis of corn yields are key to farm economies.
Photo by Iowa State University.

We will now use data to calculate a regression line. We could have calculated a regression line for the previous data, but since it is not obvious which is the cause and which the effect for July temperatures and precipitation, using one to predict the other is somewhat questionable (Figure 16).

Here we will use the relationship between water stress and corn yield reduction in Iowa.

Water is the independent variable with yield being the dependent (or predicted) variable. In this case, the researcher controlled the amount of water applied and measured the yield.

**Ex. 2: Plotting Data (1)**

Download and open the Excel file QM-mod7-ex2data.xls. Select the “Water stress” worksheet.

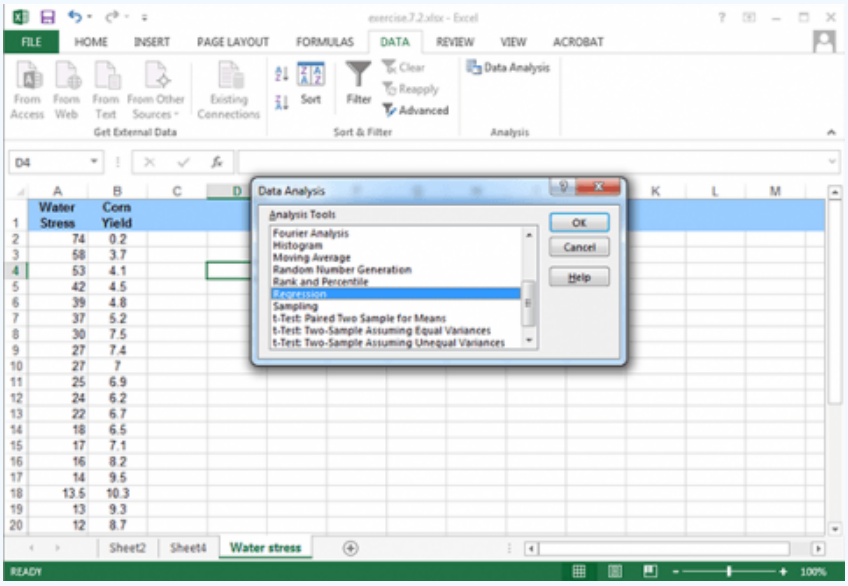



Figure 17 Water stress and corn yield data

**Ex. 2: Plotting Data (2)**

Select the Data tab and click on the Data Analysis tool.

Scroll down to Regression; highlight it and click OK (Figure 17).

**Ex. 2: Plotting Data (3)**

Fill in the options as shown (Figure 18):

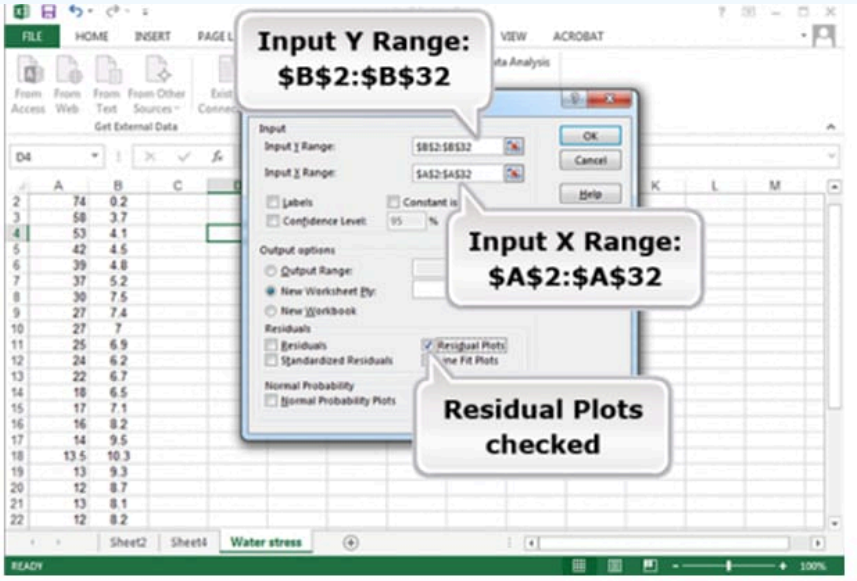



Figure 18 Water stress data for residual plot

**Ex. 2: Plotting Data (4)**

This gives the Linear Fit for the least squares regression line and an ANOVA for Regression. It also gives the residual plot (Figure 19).

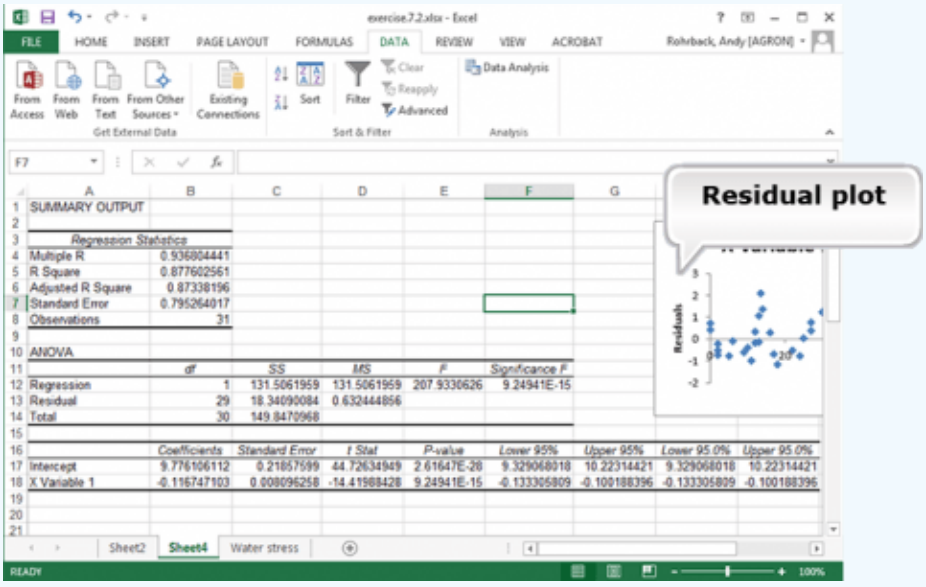



Figure 19 Regression ANOVA.

Notice that the prediction equation is : E(Yield (in 1000 kg/ha)) = 9.78 – 0.117 (Water Stress) (Figure 20). This can be determined from the coefficients for Intercept and X Variable 1. The regression equation is Y = 9.78 – 0.117(WS) + error.

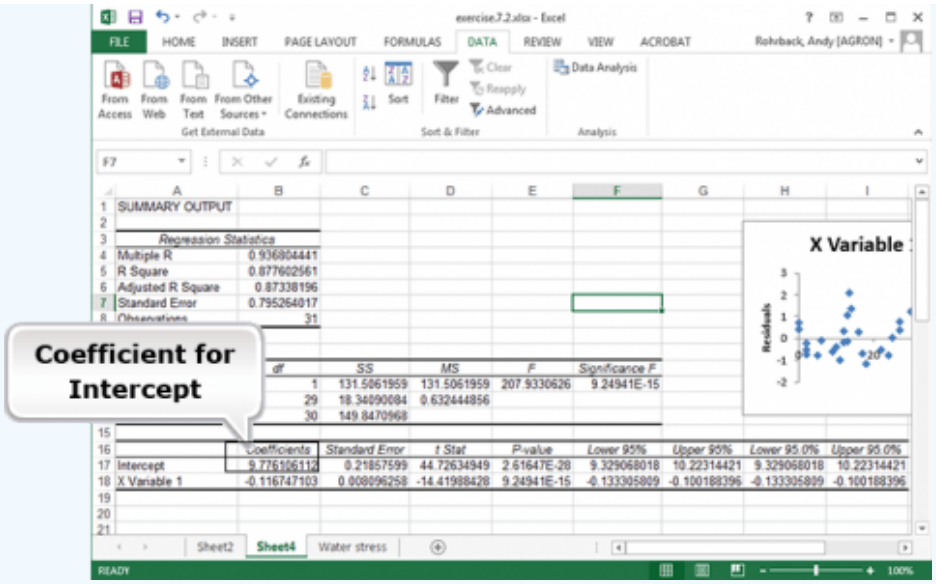



Figure 20 Regression ANOVA Table.

**Ex. 2: Plotting Data (5)**

Save this analysis in Figure 21 for Exercise 3.

## Estimation Formula

**Sources of Variation** include the line and deviation from the line. The line produced in linear regression is calculated to minimize the average distance of the Y-values from the line. Thus, summing the deviations of the actual Y-values from the regression-predicted values will equal 0. Measurement of observed data always has some variability associated with it due to the nature of error in experimental data. This variability can be accounted for and partitioned into its sources with an Analysis of Variance (ANOVA). Some variability of the Y's occurs because of their relationship with X. This is quantified by the squared correlation coefficient $(r)^2$, the proportion of variance in Y that can be accounted for by linear association with X. The rest of the variability around the line cannot be accounted for (at least in its relationship with the X variable). This is attributed to error. The linear model that accounts for this is depicted in this equation.

$$Y \tag{1.7.11}$$

**where:**

$$\alpha \tag{1.7.12}$$

$$\beta \tag{1.7.13}$$

$$\epsilon \tag{1.7.14}$$

## Errors

The true error is assumed to be in the measurement of the Y values only (Figure 21). It is assumed that the X's are fixed or that their measurement error is very small. The situation where the X's have error is termed a bivariate normal distribution, in which case the assumptions for regression are not valid. We saw the effect of measurement errors in the X-variable in an earlier "Try This". Some other assumptions are necessary for regression:

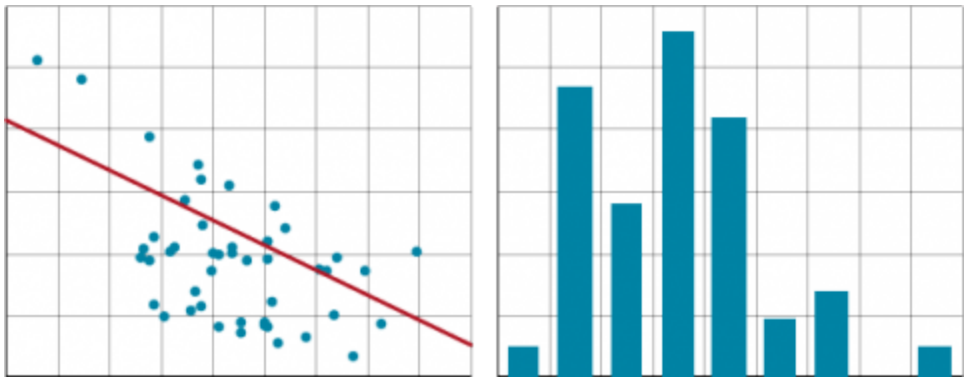


Figure 21 Scatter plots and histograms are visual representations of variation in data sets.

- for any value of X there is a normal distribution of errors
- the variances must be the same for all Y values
- the Y-values are randomly obtained and independent of each other
- the mean of the Y-values at a given X is on the regression line

Notice that these assumptions — independence of Y-value, normal distribution, constant variance, and adequacy of the model — will be essential throughout the remainder of the course.

## Sum of Squares

It can be shown that the total sum of squares for Y is the sum of that associated with the regression line and that from the errors, or sum of squares not related to the relationship with X.

The correlation coefficient squared, $r^2$, from the previous section describes the amount of variation attributable to the regression equation below. For example, if $r^2$ = 0.75, then 75% of the total variation in Y is accounted for by the linear regression.

The total Sums of Squared (SS) deviations in the response variable (Y) is given by Equation 5.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{1.7.15}$$

This represents the total variability about the average response variable and is used extensively throughout this and future units.

## Regression and Total SS

Because we have to estimate one parameter (the average response) in the total SS, we need to recognize that there are n-1 degrees of freedom (df) associated with this calculation.

$$\text{Regression SS} = \beta \times \text{Sxy} = \beta \times \left[\sum \text{xy} - \frac{\sum \text{x} \sum \text{y}}{\text{n}}\right] \tag{1.7.16}$$

**where:**

$$\text{Total SS = Sum of squares of } y, S_{yy} = \sum \text{y}^2 - \frac{(\sum \text{y}^2)}{\text{n}}$$

**or**

$$S_{yy} \tag{1.7.17}$$

## Partitioning Variation

The total sum of square deviations can be partitioned into regression and residual sums of squares based on these formulae.

$$\text{Regression SS} = (r^2 S_{yy})(\text{Residual or error}) = (1 - r^2) S_{yy} \tag{1.7.18}$$

This captures the essential relationship between the correlation coefficient, the variance of the Y values, and the partition of variation into that associated with the model (Regression SS) and that which is unexplained (residual). As the residual becomes large relative to the total variance, the correlation coefficient becomes smaller. Thus, the correlation coefficient is a function of both the residual and the total variance of Y.

# Statistical Significance

Statistical Significance of the regression relationship can be tested with an **F-test.** The calculated regression slope is based on gathered data, from a sample. Calculations based on these data estimate the actual regression relationship. Even though we have calculated a regression coefficient, it is merely an estimate of how the variables are related. The true relationship may be slightly different from the one calculated. This could result in stating that there is a relationship when none exists. Testing the significance of the slope of the regression is done in a manner similar to other types of hypothesis testing.

The null and alternative hypotheses for this test:

$$H_0 : \beta = 0 \tag{1.7.19}$$

$$\text{Equation} \tag{1.7.20}$$

## Formula for F

The test is used to determine if the slope of the regression line is different from 0. Two related statistical tests may be used to test this hypothesis. The first, shown below, uses the sums of squares to determine if the regression coefficient captures enough of the variance in the data using the F test.

$$\text{F= } \frac{\text{Regression MS}}{\text{Residual MS}} (\text{with 1 and n-2 degrees of freedom}).$$

$$\text{Equation 9} \tag{1.7.21}$$

If the slope explains a significant proportion of the variability in the regression, then the slope is considered different from 0. If not enough variation is explained at some level of significance, often 0.05, then the slope cannot be considered different from 0.

## ANOVA Table

As discussed, the variability can be partitioned. The linear regression sum of squares is calculated as shown in the Analysis of Variance (ANOVA) table or using the equation from the previous section. The ANOVA table for linear regression is shown in Table 3.

*Table 3 The ANOVA table for linear regression.*

| Source of Variaton | Sum of Squares | Df | Mean Square | F |
|---|---|---|---|---|
| Regression | $\hat{\beta} Sxy$ | 1 | $\frac{\text{Regression SS}}{\text{Regression Df}}$ | $\frac{\text{Regression MS}}{\text{Residual MS}}$ |
| Residuals | Syy - regression SS | n -2 | $\frac{\text{Residual SS}}{\text{Residual Df}}$ | n/a |

Notice the sums of squares and degrees of freedom for regression. The regression SS is written as a formula involving a ratio. This equates to the proportion $R^2$ of the Total SS of Y (Equation 6). The regression relationship has 1 df, because the test is for the slope being zero. In an ANOVA, mean squares are SS divided by df.

## Example: ANOVA

Let's test the regression calculated from the previous data set. Calculating the sum of squares for Y gives a value of 391.3. Knowing the $r^2$ value of 0.62, we can fill in the following ANOVA table:

*Table 4 Example on testing regression null hypothesis slope equal zero.*

| Source of Variation | Sum of Squares | Df | Mean Square | F | P < F |
|---|---|---|---|---|---|
| Regression | 242.1 | 1 | 242.1 | 12.95 | 0.007 |
| Residuals | 149.2 | 8 | 18.7 | | |

The F-test is used to compare the equality of variances. In this case, we are testing whether the variance associated with the estimated slope, $\hat{\beta}$, is greater than the residual variance.

A calculated F value greater than the critical F value indicates the slope is significantly different from zero. Alternatively, most statistical software will calculate the probability of a given F value.

## Ex. 3: Calculating a Regression Line and Testing the Slope (1)

In this exercise we will use Excel to find the regression equation. Calculations from the raw data are possible, but equations can be calculated easily using statistical software.

Return to the water stress computer output from the last exercise or re-run that analysis (Figure 22).

### Ex. 3: Calculating a Regression Line and Testing the Slope (2)

**A great deal of information is available from the Summary Output and Analysis of Variance (Figure 22).**

1. R (correlation between yield and water stress)
2. R-Squared (R2)
3. Adjusted R Square
   1-(Residual MS/Total MS)
   A better measure for “goodness of fit” in multiple regression and comparing regression lines with different numbers of replication than is R-squared.
4. Standard Error (square root of the residual mean square)

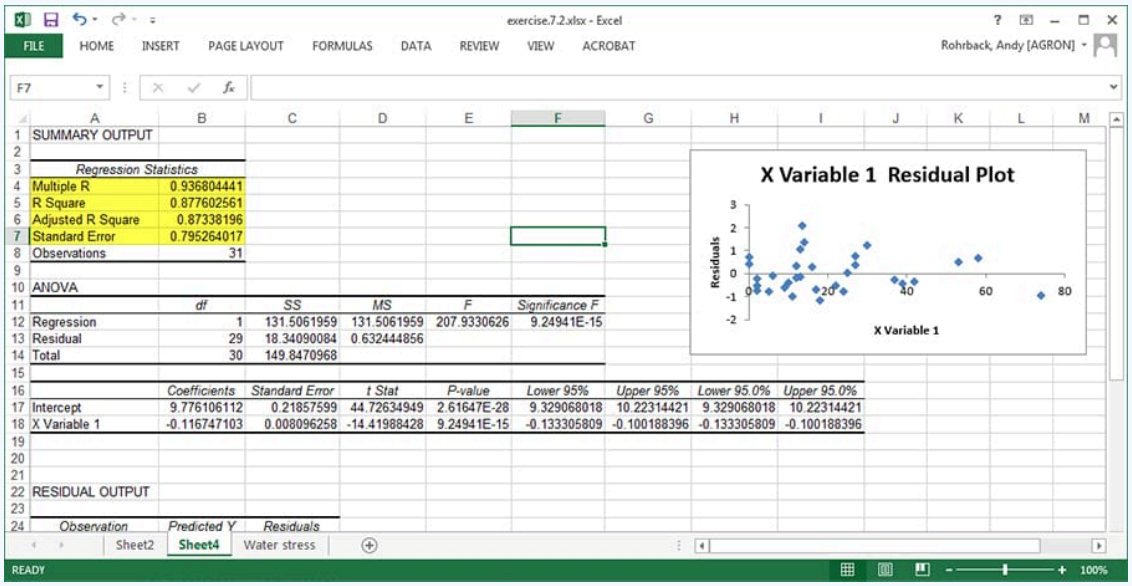



Figure 22 ANOVA table, summary statistics, and residual plot

- Again, notice the regression equation (Linear Fit). E(Y) = 9.78 – 0.117x Is the slope statistically significantly different from zero?
- The ANOVA table, which has the F-test for slope based on the residual mean square (0.632), supplies the answer.

- The Prob > F, which tests the null hypothesis of no linear regression relationship (i.e., slope =0), implies to reject H0 because the probability is < .0001.
- There is another test for the significance of water stress, as a t-ratio for water stress in the table beneath the ANOVA.
- The regression slope, estimated by -0.117, is significantly different from zero.

**Ex. 3: Calculating a Regression Line and Testing the Slope (2)**

We can get some information on how well the line fits the data by examining the Residual plot.

The sum of the residuals should be 0 (or very small due to rounding errors of the computer and software). The plot of residuals displays how the actual Y values deviate from the regression-predicted Y values at each X. These should be scattered randomly along the X-axis. If there is any regularity to the residuals, the data may not be fit well by linear regression, or one of the assumptions of linear regression may have been violated. This is a small data set, so it would be easy to think that there is a pattern there. However, given the small size of the sample, it does appear to be approximately randomly distributed around zero.

Study Question 2

Query 1.7.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Confidence Limits

**Confidence limits can be established for the regression slope.**

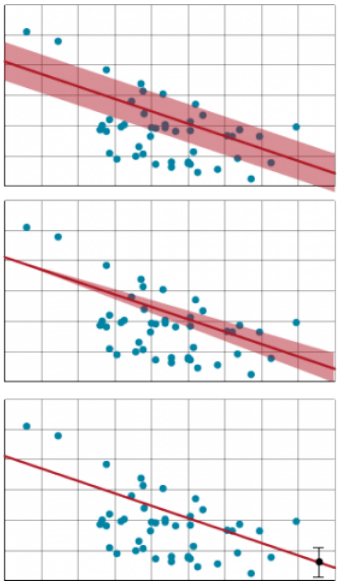


Figure 23 Display methods for confidence limits on the regression line of a set of data and the predictions made from it.

When you have calculated a regression line and tested the slope for significance, you can be reasonably assured that the sampled regression line approximates the slope, b, of the model. Testing whether the slope describes a significant amount of the total

variability is based on the F-test. From Table 24, we note that the calculated F value of 12.95 is a really large value relative to an F distribution where the slope is equal to zero. Indeed, the probability of getting such a value or larger (P>F), given the null hypothesis, is 0.007. Thus the data do not support the null hypothesis of no linear response, although we might be wrong with such a statement about 7 times out of a thousand.
Another test related to the F-test is the t-test. The t-test can be used to test the significance of the regression line or more commonly it can be used to set error limits of the regression line. Typically, these are displayed as error bars encompassing some percentage of the data based on the estimated variance, $s^2$. These limits come in three different types, error bars describing a confidence interval where the regression line occurs, error bars around the estimate of the average response, and bars encompassing an individual predicted value for a given X.

## Equation

Confidence limits are the lower and upper bounds of a confidence interval. In the context of regression line they can be used to test whether the slope of the regression line is different from 0. The null hypothesis is that the slope of the regression line is 0. The alternative hypothesis is that the slope of the line is not zero. If the confidence interval includes 0, the slope cannot be considered different from 0 at that level of significance. The error is merely that associated with the regression line. Confidence limits on a regression line are similar to those calculated for sample means (Equation 10).

$$\text{CL}=\hat{\beta} \pm t \times \text{SE}.$$

Equation 10 (1.7.22)

**where:**
$\beta$ = slope estimate
**t** = t – value for the given degrees of freedom and significance level
**SE** = standard error of $\hat{\beta} = \sqrt{\frac{\text{Residual Mean Square}}{\text{Sxx}}}$

## Using t-Test

Restructuring Equation 10 to solve for t allows us to use a t-test for testing whether the slope is different from 0. That test is equivalent to the F-test of regression in the ANOVA. The confidence limits bracket possible slopes of the regression line (Figure 24). A 95% confidence interval for the slope of a regression line means that this procedure will bracket the true regression slope 95% of the time.

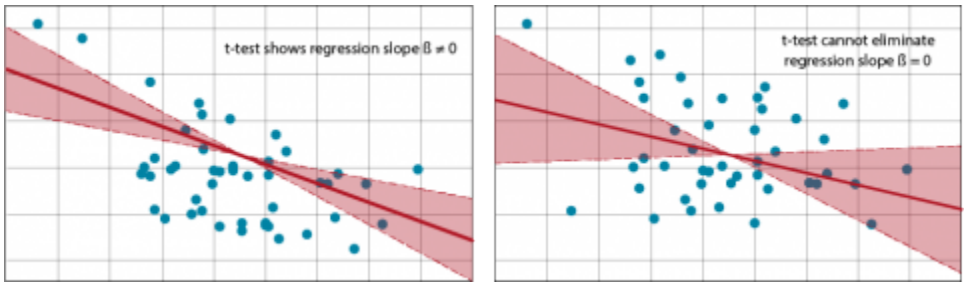



Figure 24 Regression line and confidence limits bracketing the slope.

When the standard error is large or the estimated slope of the regression line is small, a distribution will fail the t-test because a slope of 0 is possible under the given confidence level. Limits on the estimates of a specific Y from the equation, correspondingly, will have a wider limit. The estimate of the mean or predicted values includes not only the variance of the regression line but also that of individual means or values at each X-value.

### Ex. 4: Confidence Limits

Open the Excel water stress workbook you used earlier (Figure 22 or Figure 23).

Directly underneath the ANOVA table is a table with t-ratios and confidence intervals for the intercept and Water Stress coefficients.
A 95% confidence interval for $\hat{\beta}$ is between -0.133 and -0.100.
This is also easily computed from the formula

$$\hat{\beta} \pm t \times \mathrm{SE}$$

or
-0.117±(2.045)(0.0081)

where, with alpha = 0.05 in two tails and 29 error df, the table t-value is 2.045 and the standard error for b is 0.008096.
Use a calculator to show that the confidence limits for $\hat{\beta}$ match those in the table.

## Replicated Regression

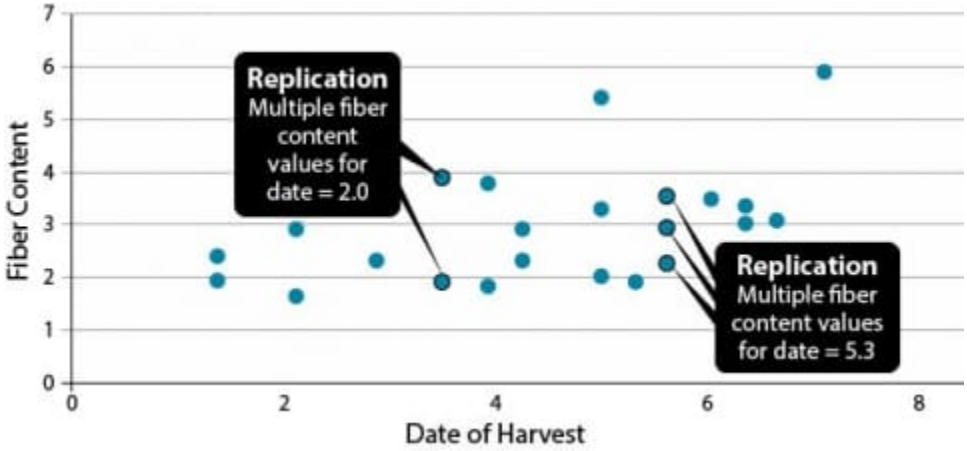



Figure 25 Replication effect on fiber content of multiple samples on same harvest date.

These data on fiber content in corn kernels related to harvest date shows replication: some samples harvested on the same date have different fiber contents (Figure 25).

**Regression in replicated data allows a goodness-of-fit test.** Agronomic experiments are usually replicated. In this situation, when the data are grouped, replication of Y values at X's will occur (Figure 26). Calculating a total regression includes the variability in Y replication at the X values.

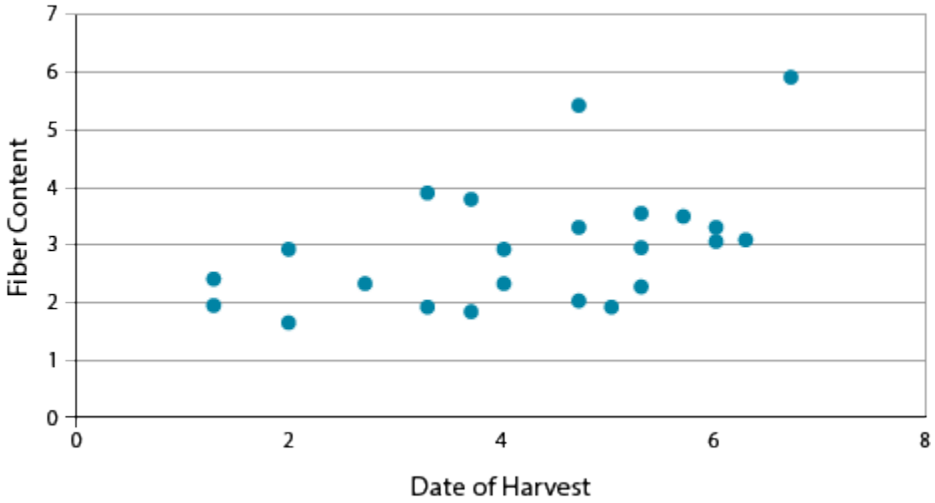



Figure 26 Scatter plot showing fiber content of corn by harvest date.

The initial ANOVA table provides the sum of squares and the test for the significance of the regression line. After the 1 df for

regression, 21 df remain. This variability can be partitioned into the two sources discussed, the lack of fit to the model and the pure error. The lack of fit variability comes from the difference between the actual means of the Y's at each X, and the regression line predicted Y at each X. This value describes how much error is associated with the regression line. This value can be tested to determine if the regression lack of fit is different from 0. The error which is left over is termed Pure error.

*Table 5 Regression ANOVA output*

| Source | SS | Df | MS | F | P < F |
|---|---|---|---|---|---|
| Regression | 6.32 | 1 | 6.32 | 6.26 | <.05 |
| Residual | 21.19 | 21 | 1.01 | | |

## Error Calculation

Pure error is the deviation sum of squares of each individual Y from the mean Y at each X. The degrees of freedom are the sum of one less than the number of replicated Y's at each X. The pure error is calculated and subtracted from the residual to find the lack of fit.

$$\text{Total Error} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \tag{1.7.23}$$

**where:**
**i** = level of X
**j** = each replicated Y at a given X
$\mathbf{Y_{ij}}$ = each observation at a given level of the x (independent variable)
$\mathbf{Y_i}$ = mean for each level of the x variable

In effect, you are calculating a new sum of squares that estimates the total variance of observations around the mean for each value of x. This tells us how scattered were the data points we tried to fit with the regression line. The pure error degrees of freedom are calculated as:

$$\text{Pure Error df} = \sum (n-1) \text{at each X} \tag{1.7.24}$$

## Example: ANOVA Table

These equations produce the ANOVA table (Table 6).

*Table 6 Output from ANOVA of data*

| Source | SS | Df | MS | F | P < F |
|---|---|---|---|---|---|
| Regression | 6.32 | 1 | 6.32 | 6.26 | < .05 |
| Regression | 21.19 | 21 | 1.01 | | |
| Lack of fit | 8.78 | 11 | 0.79 | 0.64 | |
| Pure error | 12.41 | 10 | 1.24 | | |

The second F test compares the MS-lack of fit to the MS-pure error. This tests the linearity of the regression line. Since it is not significant, we assume the regression is linear and we do not have to try another model.

### Ex. 5: ANOVA with Replicated Data

In this exercise, we will calculate the ANOVA presented in the lesson for replicated measurements of fiber content over different harvest dates. This will require calculating the ANOVA and partitioning the degrees of freedom and sums of squares.

- Download the QM-mod7-ex5data.xls file and save it.
- Run the regression analysis covered in Exercise 2 on this data set, then go back to the worksheet with the original data set.
- The lack of fit test is not calculated automatically in the regression analysis, so we will have to do it step-by-step.

On the same page as the data set, add two columns. Label one Date and the other Mean at Date. (Figure 27)

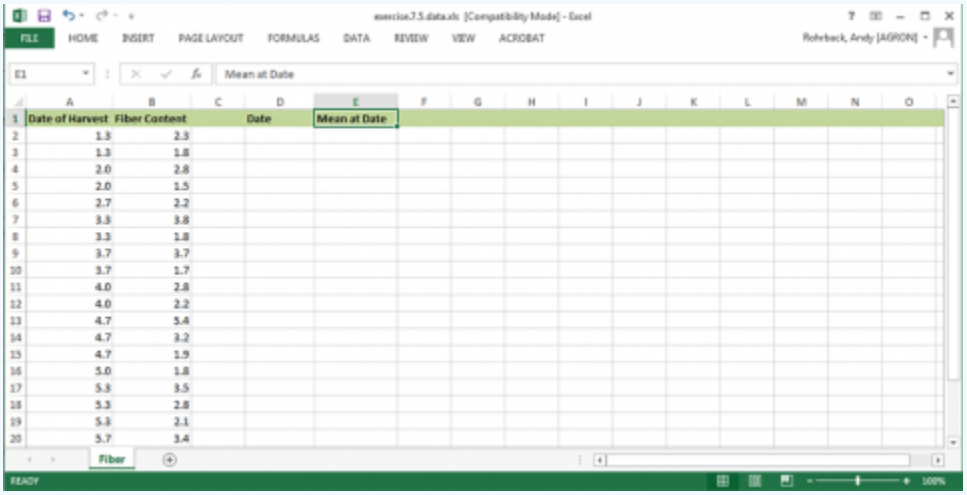



Figure 27 Date of harvest and fiber content data.

**Ex. 5: ANOVA with Replicated Data (2)**

Under date, copy each of the dates once (Figure 28).

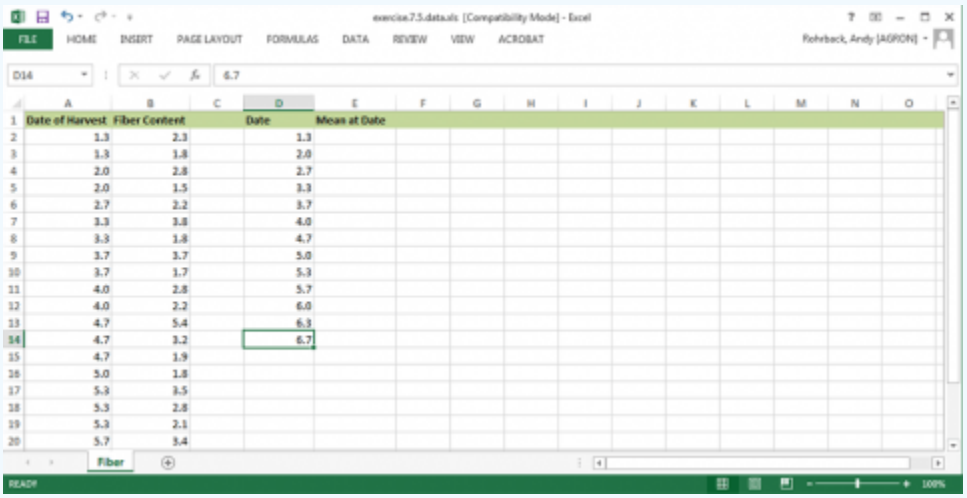



Figure 28 Date of harvest and fiber content with a column of single dates.

Under Mean at Date, calculate the average of the observations at the given date (Figure 29).

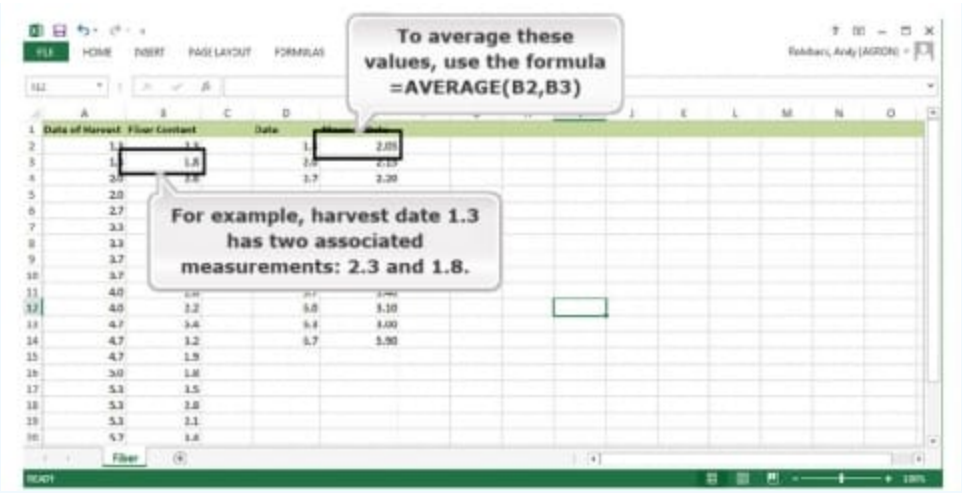



Figure 29 Date of harvest and fiber content with a column of single dates and averages.

**Ex. 5: ANOVA with Replicated Data (3)**

Now we will find the residuals associated with pure error using the means that were calculated (Figure 30).

1. Insert a new column next to the Fiber Content data.
2. Place the average for each date next to the associated date. Use this formula: =LOOKUP(A2,E$2:E$14, F$2:F$14)
3. Add another new column for the residuals.

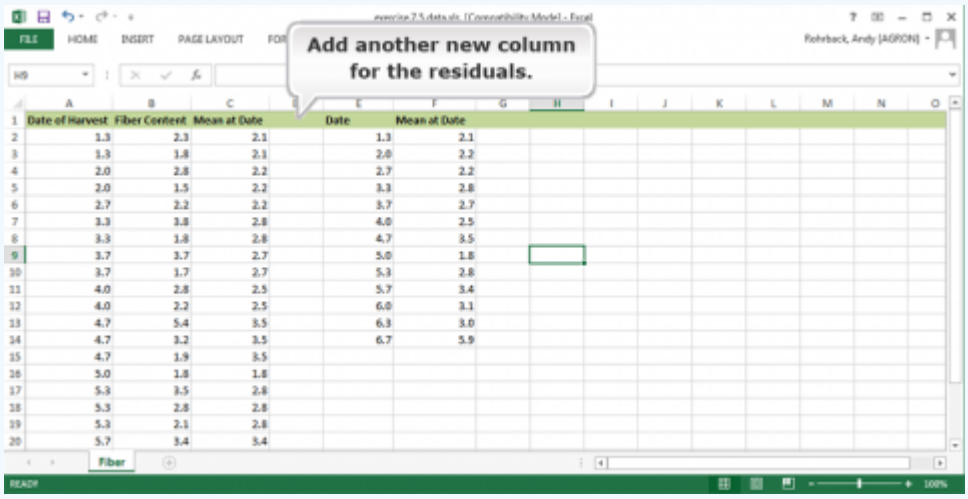



Figure 30 Date of harvest and fiber content.

4. The residuals can then be determined by subtracting the mean at each date from the observation at that date. =B2-C2
5. Add another column and insert the sums of squares (SS) associated with pure error by squaring each of the residuals. =POWER(D2,2)
6. Add another column and insert the sums of squares (SS) associated with pure error by squaring each of the residuals. =POWER(D2,2)
7. Then sum all the squares of residual pure error. =SUM(E2:E24)
8. The degrees of freedom for pure error are calculated by determining the number of observations at each date and subtracting one from the number of replications at each date. =IF(COUNTIF(A$2:INDIRECT("A"&ROW()),A2)>1, "", COUNTIF(A$2:A24,A2)-1)
9. These values are then summed. =SUM(F2:F24)
10. The mean squares for pure error are then calculated by dividing the pure error sums of squares by the degrees of freedom for pure error.
11. The lack of fit sums of squares and degrees of freedom can be found by subtracting the pure error sums of squares or degrees of freedom from the residual sums of squares.
12. The mean squares for lack of fit are found by dividing the lack of fit SS by the lack of fit DF.

13. Calculate the F-test for the lack of fit test by dividing the lack of fit MS by the pure error MS.
14. Finally, the p-value for the F-test can be found using the formula “=f.dist.rt(A,B,C)” where A is the F-statistic, B is the df for lack of fit, and C is the pure error DF.

**Ex. 5: ANOVA with Replicated Data (4)**

Fill in the values in the ANOVA table under the regression analysis.

Insert two rows for the “Lack of Fit” and “Pure Error” statistics (Figure 31).

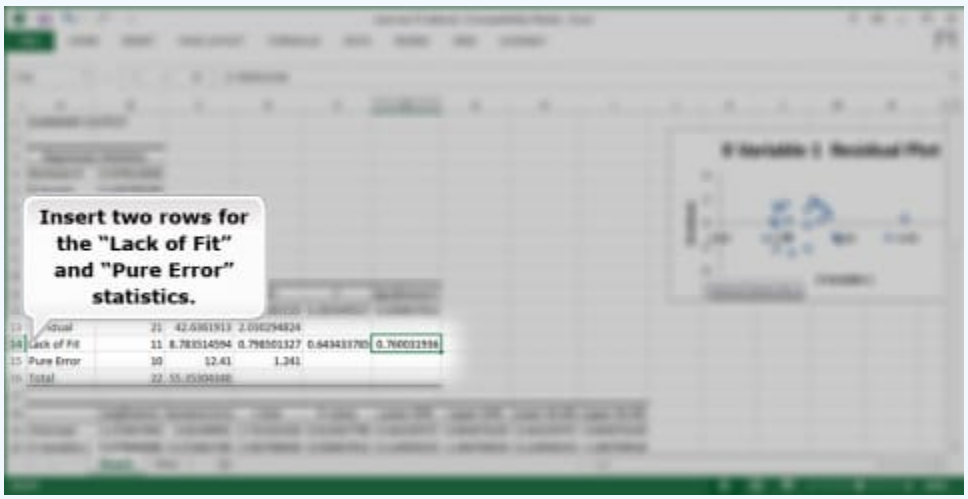



Figure 31 ANOVA table of output from Figure 21

Add the values you calculated to the ANOVA table under the regression analysis.

The completed analysis can be found in QM-mod7-ex5solved.xlsx.

# Summary

## Correlation (r)

- Measures degree or strength of linear relationship
- Tells direction of linear relationship; positive implies x and y increase or decrease together; negative (y decreases as x increases or vice versa)
- Ranges between -1 and +1, with 0 being no linear relationship
- The scatter plot is important to help interpret

## Linear regression

- Establishes a mathematical relationship between two variables
- Prediction equation is Y = a + bx
- Parameter estimates are intercept (a) and slope (b)
- The line of best fit minimizes squares of vertical deviations from the line

## ANOVA for Regression

- Has sources of variation for regression with 1 df and error with (n – 2) df
- $r^2$ = (square of correlation) is the proportion of variation attributed to linear regression
- Tests statistical significance of linear regression

## Confidence Limits

- Can be established for regression slope
- CL = b ± tsb
- Can also be computed for a mean of y-values or individual y given x.

## Regression with Replicated Data

- Allows a goodness-of-fit test of the model

**How to cite this chapter**: Mowers, R., D. Todey, K. Meade, W. Beavis, L. Merrick, and A. A. Mahama. 2023. Linear Correlation, Regression and Prediction. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.7: Linear Correlation, Regression and Prediction is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Dennis Todey; Kendra Meade; William Beavis; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.8: The Analysis of Variance (ANOVA)

A t-test is appropriate for comparing mean values of two treatments. Many times, however, we want to compare several treatments. For example, a plant breeder may want to compare 25 genotypes in a field trial, or a soil scientist, three sources of nitrogen fertilizer. What test or procedure should you use in cases like these? In this lesson we will learn how to use a procedure called the analysis of variance (ANOVA) to test multiple sample hypotheses such as these.

**Learning Objectives**

- The conceptual basis of the analysis of variance (ANOVA).
- How to construct an ANOVA table.
- How to calculate the sum of squares (SS) and mean squares (MS) associated with the sources of variation.
- About the linear additive model for the ANOVA.

## One-Factor ANOVA

**ANOVA helps determine if treatments are different.** Put most simply, an analysis of variance (ANOVA) compares the variance associated with treatments to that variance which occurs naturally between experimental units (usually plots). Consider the many potential sources of variation in field plot research: soil properties may vary among plots, insects or other pests may attack one plot more than another, or carry-over effects from previous crops may vary from plot to plot. What we are asking when we run an analysis, then, is whether the differences between our treatment means are greater than these "background" differences between our plots.

## Variance

Variance arising from many sources can make experimental results difficult to decipher. Using the analysis of variance can help solve this by partitioning the total variance into discrete variances associated with our treatments and by lumping all the unexplained variance into a single term we call error. The name is a little confusing because the error variance does not really reflect what we normally think of as errors or mistakes. The error mean square actually describes the natural and unexplained variation among our experimental units, which in agronomy are often field plots. Think of it this way, if you plant the same crop variety in four different plots, would you really expect the yields you measured on the plots to be exactly the same? It would be very unusual if they did. There are a whole lot of characteristics associated with the plots that affect yield in addition to the variety that was planted. The experimental error sequesters these effects and provides a way for us to compare the variation associated with varieties with that which occurs naturally among our plots. Your text and other references often refer to the error variance or variation as the residual.

## Example

Let's develop an example to see how an ANOVA is done. An experiment in northwest Iowa compared the yield of a corn hybrid planted at three plant densities to determine the optimum planting rate. The data are given in Table 1.

*Table 1 Yield Data (t/ha) for corn planted at three plant populations in northwest Iowa.*

| Population (plants/$m^{2)}$ | 1 (t/ha) | 2 (t/ha) | 3 (t/ha) | mean (t/ha) |
|---|---|---|---|---|
| 7.5 | 8.64 | 7.84 | 9.19 | 8.56 |
| 10 | 10.46 | 9.29 | 8.99 | 9.58 |
| 12.5 | 6.64 | 5.45 | 4.74 | 5.61 |

Over the next few pages, we will conduct an analysis of these data.

## Discussion

**Using the data given in Table 1 only (do not use any preconceived ideas or knowledge of plant populations), are there differences in the treatments (plant populations)? How do they differ?**

Here are some points germane to our discussion:

- We use ANOVA to find differences among treatment means. This example problem may be examined further to estimate the mathematical relationship between yield and plant population. However, in this unit, we just use this as an example for understanding how to tell if mean yields for each plant population level differ.
- We assume the plant populations are each randomly allocated to three of the nine plots. This is called a "completely randomized design" or CRD. You may anticipate that a better design would group all three treatments into the same area for each replication of the experiment. That is, in fact, how most of these experiments are designed, and we will later study such a design.
- This ANOVA for a CRD is an extension of the two-sample t-test, now with more than just two treatments. The samples are assumed to be independent because treatments (plant populations) are assigned to plots completely at random.

# ANOVA Table

**The ANOVA table is a tool to separate the variances.**

The table is often arranged so that the total variation is listed on the bottom line of the table. The corn population data (Table 1) can be reorganized by computation formulae to compare variances in an ANOVA table (Table 2).

*Table 2 ANOVA table for corn planted at three populations in northwest Iowa*

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | 12.769 | 19.205 | 5.14 |
| Error | 6 | 3.989 | 0.665 | n/a | n/a |
| Total | 8 | 29.526 | n/a | n/a | n/a |

Next, we will look at how to get to this point and the use and meaning of each part of the table.

## Sources of Variation

The ANOVA in Table 2 for a CRD separates variation into sources.

We separate these sources into:

- **treatment** (in this case, plant populations): the variation associated with the treatments we are comparing.
- **error** (residual): the variation that occurs within samples (between experimental units receiving the same treatments).
- **total**: the total variation in the experiment (the sum of other sources of variation).

## Degrees of Freedom

The ANOVA in Table 2 has degrees of freedom (df) for each source.

**The second column in the ANOVA table lists the degrees of freedom for each source of variation.** Degrees of freedom, as you learned in earlier units, are related to the number of samples that occur. The degrees of freedom associated with the sources of variation are calculated for an experiment with a completely randomized design using the formulae in Table 3.

*Table 3 Generic ANOVA table for sources of variation and degrees of freedom.*

| Source of Variation | Degrees of Freedom |
|---|---|
| Treatment | $\text{Levels of treatment } -1=t$ |

| Source of Variation | Degrees of Freedom |
|---|---|
| Error | $T - t$ |
| Total | Levels of treatment $\times$ Replications $- 1 = T$ |

### Study Question 1: Degrees of Freedom

Using the experiment given (Table 1), fill in the degrees of freedom for the ANOVA table.

| Population (plants/m$^2$) | 1 (t/ha) | 2 (1/ha) | 3 (t/ha) | mean (t/ha) |
|---|---|---|---|---|
| 7.5 | 8.64 | 7.84 | 9.19 | 8.56 |
| 10 | 10.46 | 9.29 | 8.99 | 9.58 |
| 12.5 | 6.64 | 5.45 | 4.74 | 5.61 |

#### Query 1.8.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Sum of Squares

The sum of squares for each source is equal to the numerator in the equation we used to calculate variance in Chapter 1 on Basic Principles .

$$\text{Treatment SS} = \sum(\frac{T^2}{r}) - \text{CF} \tag{1.8.1}$$

**where:**
**T** = each treatment total
**r** = number of replications
**CF**= correlation factor, explained below

The treatment sum of squares describes the distribution of the treatment means around the overall population mean μ (sometimes called the grand mean). The Correction Factor has been used in the class before. It is calculated as:

$$\text{CF} = \frac{(\sum x)^2}{n} \tag{1.8.2}$$

**where:**
**x** = each observation
**n** = number of observations

## Sum of Squares Calculations

If you review Chapter 2 on “Distributions and Probability” and Chapter 7 on “Linear Correlation, Regression and Prediction,” you will note that CF is the second half of the calculation of the sum of squares. We give it a special designation in this lesson because we use the same CF to calculate the Treat SS, and the Total SS. For our own corn experiment, the CF is:

$$\text{CF} = \frac{(\sum x)^2}{n} \tag{1.8.3}$$

$$\text{CF} = \frac{(8.64 + 7.84 + 9.19 + 10.46 + 9.29 + 8.99 + 6.64 + 9.29 + 4.74)^2}{n} \tag{1.8.4}$$

$$\text{CF} = \frac{(71.24)^2}{9} \tag{1.8.5}$$

$$\text{CF} = 563.90 \tag{1.8.6}$$

The Treatment SS is therefore calculated as:

$$\text{Treatment SS} = \sum(\frac{T^2}{r}) - \text{CF} \tag{1.8.7}$$

$$\text{Treatment SS} = (\frac{25.67^2}{3} + \frac{28.74^2}{3} + \frac{16.81^2}{3}) - \text{CF} \tag{1.8.8}$$

$$\text{Treatment SS} = (\frac{658.94}{3} + \frac{825.99}{3} + \frac{282.91}{3}) - 563.75 = 25.537 \tag{1.8.9}$$

## Sum of Squares: Total SS

The Total SS is calculated as:

$$\text{Total SS} = \sum x^2 - \text{CF} \tag{1.8.10}$$

**where:**
**x** = each observation
**CF** = correlation factor

For the corn example, the Total SS is:

$$\text{Total SS} = (8.64^2 + 7.84^2 + 9.19^2 + 10.46^2 + 9.29^2 + 8.99^2 + 6.64^2 + 5.45^2 + 4.74^2) - \text{CF} \tag{1.8.11}$$

$$\text{Total SS} = (593.27 - 563.50) = 29.526 \tag{1.8.12}$$

## Sum of Squares: Residual SS

The residual or error SS, Resid SS, is the difference between the Total SS and the Treatment SS:

$$\text{Resid SS} = \text{Total SS - Treatment SS} = 29.526 - 25.537 = 3.989 \tag{1.8.13}$$

The residual sum of squares represents the distribution of the observations around each treatment mean. This represents the "background" differences between experimental units.

### Study Question 2: Treatment or Error

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | n/a | n/a | n/a |
| Error | 6 | 3.989 | n/a | n/a | n/a |
| Total | 8 | 29.526 | n/a | n/a | n/a |

Query 1.8.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## One-Way ANOVA (1)

The analysis of variance is a method of separating the variability in an experiment into useful and more manageable parts. This separation helps to assess the significance of what is being tested. Different samples will give rise to different significance levels in ANOVA because summary statistics (mean, sums of squared differences, etc.) that go into the analysis differ among samples.

Altering samples (either by adding data points or changing values) results in corresponding changes in the summary statistics, the ANOVA table, and the requested charts of sums of squares. Some questions which usually need to be answered are as in the next section immediately below.

## One-Way ANOVA (2)

### Discussion: One-Way ANOVA

- **How does the change in mean between treatments affect the significance of their differences?**
- **How much does the variability within treatments affect the significance of their differences?**
- **How much effect do individual samples have in the final analysis?**
- **What is the effect of adding additional samples to the analysis?**

### Mean Squares

Mean squares are the variance estimates in the ANOVA table.
The mean square associated with the source of variation is calculated by dividing the sum of squares (the numerator in the variance formula) by the degrees of freedom (the denominator in the variance formula). This gives us the following values for our table:

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | 12.769 | 19.205 | n/a |
| Error | 6 | 3.989 | 0.665 | n/a | n/a |
| Total | 8 | 29.526 | n/a | n/a | n/a |

Obviously, the greater the number of treatment levels and the larger number of experimental units involved in an experiment, the more variation is introduced. An experiment with 50 experimental units will generally have a larger total variability than one with 10 units. The variation has to be partitioned or divided among the treatment and error sources of variation. This number is then called the treatment mean square (TMS) and residual or error mean square (RMS). These, respectively, are estimates of variance among treatments (TMS) and within treatments, or residual variance (RMS, also called $s^2$).

### Observed F-Ratio

The F-ratio tests whether treatments are different.
The last step in computing the analysis of variance is to compare the difference in variation between our treatment means and

the error by dividing the mean square associated with treatments (TMS) by the mean square associated with the residual (RMS). This mean square, or variance, ratio gives us an observed F-value:

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | 12.769 | 19.205 | n/a |
| Error | 6 | 3.989 | 0.665 | n/a | n/a |
| Total | 8 | 29.526 | n/a | n/a | n/a |

The F-ratio now explains how much of the variability can be attributed to the designed part of the experiment (treatments) as opposed to the error or residual. The larger the F-ratio is, the less likely that the treatments are the same. If treatments are the same, we expect the observed F-ratio to be about 1. It will be compared with a table F-value for determining significance (next section).

### Study Question 3: Different Plant Populations

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | 12.769 | 19.205 | n/a |
| Error | 6 | 3.989 | 0.665 | n/a | n/a |
| Total | 8 | 29.526 | n/a | n/a | n/a |

#### Query 1.8.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Testing Hypotheses

**The hypothesis we test is whether the true treatment means are the same.**The process of performing an ANOVA is to separate the designed and explainable variation from the random and unexplainable variation. Once this has been done, some test is necessary to assess what has occurred. A hypothesis test can be set up similar to what is done with *t*-tests. When using the *t*-test we were comparing two means.

The null hypothesis we are testing in the analysis of variance is that all treatment means are equal. In the case of our corn experiment the null hypothesis is: $H_0$: $\mu_1 = \mu_2 = \mu_3$. This null hypothesis states that average corn yields are the same for any of the three plant populations. To test this hypothesis, we determine the probability that our calculated value might have occurred by chance. A common approach to this is to look up a critical F-value in a table for the alpha level we are willing to accept and compare it to the calculated F-value.

## Comparing Values: The Critical F-value

Returning to the example, our F-value can be compared with those in the table below, in a manner similar to that used with the *t*-test. Unlike the *t*-test, however, the F-test considers the degrees of freedom associated with both the numerator and the denominator to calculate the F-ratio. In this case, these are the degrees of freedom associated with the treatment and error, respectively.

**Critical F-values are listed below, and in this downloadable PDF (F distribution table) for different probabilities**. Note

that for each combination of numerator and denominator degrees of freedom, the tables list four values. These values reflect the minimum F-value associated with the 5% (0.05), 2.5% (0.025), 1% (0.01), and 0.1% (0.001) probabilities of occurrence. The degrees of freedom for the numerator are listed across the top of the table. The degrees of freedom associated with the denominator are listed in the left column of the table. The table is duplicated (Table 8), but for the sake of clarity, only the values associated with P=0.05 are shown.

*Table 4 Critical values of F at the p=0.05 level of significance. Top row = Degrees of Freedom*

| F | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 161.45 | 199.5 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 | 241.88 |
| **2** | 18.51 | 19 | 19.16 | 19.25 | 19.3 | 19.33 | 19.35 | 19.37 | 19.39 | 19.4 |
| **3** | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 |
| **4** | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6 | 5.96 |
| **5** | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 |
| **6** | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.1 | 4.06 |
| **7** | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 |
| **8** | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.5 | 3.44 | 3.39 | 3.35 |
| **9** | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 |
| **10** | 4.97 | 4.1 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 |
| **11** | 4.84 | 3.98 | 3.59 | 3.36 | 3.2 | 3.1 | 3.01 | 2.95 | 2.9 | 2.85 |
| **12** | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3 | 2.91 | 2.85 | 2.8 | 2.75 |
| **13** | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 |
| **14** | 4.6 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.7 | 2.65 | 2.6 |
| **15** | 4.54 | 3.68 | 3.29 | 3.06 | 2.9 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 |
| **16** | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 |
| **17** | 4.45 | 3.59 | 3.2 | 2.97 | 2.81 | 2.7 | 2.61 | 2.55 | 2.49 | 2.45 |
| **18** | 4.41 | 3.56 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 |
| **19** | 4.38 | 3.52 | 3.13 | 2.9 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 |
| **20** | 4.35 | 3.49 | 3.1 | 2.87 | 2.71 | 2.6 | 2.51 | 2.45 | 2.39 | 2.35 |
| **21** | 4.33 | 3.47 | 3.07 | 2.84 | 2.69 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 |
| **22** | 4.3 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.4 | 2.34 | 2.3 |
| **23** | 4.28 | 3.42 | 3.03 | 2.8 | 2.64 | 2.53 | 2.44 | 2.38 | 2.32 | 2.28 |
| **24** | 4.26 | 3.4 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.3 | 2.26 |
| **25** | 4.24 | 3.39 | 2.99 | 2.76 | 2.6 | 2.49 | 2.41 | 2.34 | 2.28 | 2.24 |
| **26** | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 |
| **27** | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.2 |
| **28** | 4.2 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 |
| **29** | 4.18 | 3.33 | 2.93 | 2.7 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 |
| **30** | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.17 |
| **31** | 4.16 | 3.31 | 2.91 | 2.68 | 2.52 | 2.41 | 2.32 | 2.26 | 2.2 | 2.15 |
| **32** | 4.15 | 3.3 | 2.9 | 2.67 | 2.51 | 2.4 | 2.31 | 2.24 | 2.19 | 2.14 |

| F | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **33** | 4.14 | 3.29 | 2.89 | 2.66 | 2.5 | 2.39 | 2.3 | 2.24 | 2.18 | 2.13 |
| **34** | 4.13 | 3.28 | 2.88 | 2.65 | 2.49 | 2.38 | 2.29 | 2.23 | 2.17 | 2.12 |
| **35** | 4.12 | 3.27 | 2.87 | 2.64 | 2.49 | 2.37 | 2.29 | 2.22 | 2.16 | 2.11 |
| **36** | 4.11 | 3.26 | 2.87 | 2.63 | 2.48 | 2.36 | 2.28 | 2.21 | 2.15 | 2.11 |
| **37** | 4.11 | 3.25 | 2.86 | 2.63 | 2.47 | 2.36 | 2.27 | 2.2 | 2.15 | 2.1 |
| **38** | 4.1 | 3.25 | 2.85 | 2.62 | 2.46 | 2.35 | 2.26 | 2.19 | 2.14 | 2.09 |
| **39** | 4.09 | 3.24 | 2.85 | 2.61 | 2.46 | 2.34 | 2.26 | 2.19 | 2.13 | 2.08 |
| **40** | 4.09 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 |
| **41** | 4.08 | 3.23 | 2.83 | 2.6 | 2.44 | 2.33 | 2.24 | 2.17 | 2.12 | 2.07 |
| **42** | 4.07 | 3.22 | 2.83 | 2.59 | 2.44 | 2.32 | 2.24 | 2.17 | 2.11 | 2.07 |
| **43** | 4.07 | 3.21 | 2.82 | 2.59 | 2.43 | 2.32 | 2.23 | 2.16 | 2.11 | 2.06 |
| **44** | 4.06 | 3.21 | 2.82 | 2.58 | 2.43 | 2.31 | 2.23 | 2.16 | 2.1 | 2.05 |
| **45** | 4.06 | 3.2 | 2.81 | 2.58 | 2.42 | 2.31 | 2.22 | 2.15 | 2.1 | 2.05 |
| **46** | 4.05 | 3.2 | 2.81 | 2.57 | 2.42 | 2.3 | 2.22 | 2.15 | 2.09 | 2.04 |
| **47** | 4.05 | 3.2 | 2.8 | 2.57 | 2.41 | 2.3 | 2.21 | 2.14 | 2.09 | 2.04 |
| **48** | 4.04 | 3.19 | 2.8 | 2.57 | 2.41 | 2.3 | 2.21 | 2.14 | 2.08 | 2.04 |
| **49** | 4.04 | 3.19 | 2.79 | 2.56 | 2.4 | 2.29 | 2.2 | 2.13 | 2.08 | 2.03 |
| **50** | 4.03 | 3.18 | 2.79 | 2.56 | 2.4 | 2.29 | 2.2 | 2.13 | 2.07 | 2.03 |
| **51** | 4.03 | 3.18 | 2.79 | 2.55 | 2.4 | 2.28 | 2.2 | 2.13 | 2.07 | 2.02 |
| **52** | 4.03 | 3.18 | 2.78 | 2.55 | 2.39 | 2.28 | 2.19 | 2.12 | 2.07 | 2.02 |
| **53** | 4.02 | 3.17 | 2.78 | 2.55 | 2.39 | 2.28 | 2.19 | 2.12 | 2.06 | 2.02 |
| **54** | 4.02 | 3.17 | 2.78 | 2.54 | 2.39 | 2.27 | 2.19 | 2.12 | 2.06 | 2.01 |
| **55** | 4.02 | 3.17 | 2.77 | 2.54 | 2.38 | 2.27 | 2.18 | 2.11 | 2.06 | 2.01 |
| **56** | 4.01 | 3.16 | 2.77 | 2.54 | 2.38 | 2.27 | 2.18 | 2.11 | 2.05 | 2.01 |
| **57** | 4.01 | 3.16 | 2.77 | 2.53 | 2.38 | 2.26 | 2.18 | 2.11 | 2.05 | 2 |
| **58** | 4.01 | 3.16 | 2.76 | 2.53 | 2.37 | 2.26 | 2.17 | 2.1 | 2.05 | 2 |
| **59** | 4 | 3.15 | 2.76 | 2.53 | 2.37 | 2.26 | 2.17 | 2.1 | 2.04 | 2 |
| **60** | 4 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.1 | 2.04 | 1.99 |
| **61** | 4 | 3.15 | 2.76 | 2.52 | 2.37 | 2.25 | 2.16 | 2.09 | 2.04 | 1.99 |
| **62** | 4 | 3.15 | 2.75 | 2.52 | 2.36 | 2.25 | 2.16 | 2.09 | 2.04 | 1.99 |
| **63** | 3.99 | 3.14 | 2.75 | 2.52 | 2.36 | 2.25 | 2.16 | 2.09 | 2.03 | 1.99 |
| **64** | 3.99 | 3.14 | 2.75 | 2.52 | 2.36 | 2.24 | 2.16 | 2.09 | 2.03 | 1.98 |
| **65** | 3.99 | 3.14 | 2.75 | 2.51 | 2.36 | 2.24 | 2.15 | 2.08 | 2.03 | 1.98 |
| **66** | 3.99 | 3.14 | 2.74 | 2.51 | 2.35 | 2.24 | 2.15 | 2.08 | 2.03 | 1.98 |
| **67** | 3.98 | 3.13 | 2.74 | 2.51 | 2.35 | 2.24 | 2.15 | 2.08 | 2.02 | 1.98 |
| **68** | 3.98 | 3.13 | 2.74 | 2.51 | 2.35 | 2.24 | 2.15 | 2.08 | 2.02 | 1.97 |
| **69** | 3.98 | 3.13 | 2.74 | 2.51 | 2.35 | 2.23 | 2.15 | 2.08 | 2.02 | 1.97 |

| F | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **70** | 3.98 | 3.13 | 2.74 | 2.5 | 2.35 | 2.23 | 2.14 | 2.07 | 2.02 | 1.97 |
| **71** | 3.98 | 3.13 | 2.73 | 2.5 | 2.34 | 2.23 | 2.14 | 2.07 | 2.02 | 1.97 |
| **72** | 3.97 | 3.12 | 2.73 | 2.5 | 2.34 | 2.23 | 2.14 | 2.07 | 2.01 | 1.97 |
| **73** | 3.97 | 3.12 | 2.73 | 2.5 | 2.34 | 2.23 | 2.14 | 2.07 | 2.01 | 1.96 |
| **74** | 3.97 | 3.12 | 2.73 | 2.5 | 2.34 | 2.22 | 2.14 | 2.07 | 2.01 | 1.96 |
| **75** | 3.97 | 3.12 | 2.73 | 2.49 | 2.34 | 2.22 | 2.13 | 2.06 | 2.01 | 1.96 |
| **76** | 3.97 | 3.12 | 2.73 | 2.49 | 2.34 | 2.22 | 2.13 | 2.06 | 2.01 | 1.96 |
| **77** | 3.97 | 3.12 | 2.72 | 2.49 | 2.33 | 2.22 | 2.13 | 2.06 | 2 | 1.96 |
| **78** | 3.96 | 3.11 | 2.72 | 2.49 | 2.33 | 2.22 | 2.13 | 2.06 | 2 | 1.95 |
| **79** | 3.96 | 3.11 | 2.72 | 2.49 | 2.33 | 2.22 | 2.13 | 2.06 | 2 | 1.95 |
| **80** | 3.96 | 3.11 | 2.72 | 2.49 | 2.33 | 2.21 | 2.13 | 2.06 | 2 | 1.95 |
| **81** | 3.96 | 3.11 | 2.72 | 2.48 | 2.33 | 2.21 | 2.13 | 2.06 | 2 | 1.95 |
| **82** | 3.96 | 3.11 | 2.72 | 2.48 | 2.33 | 2.21 | 2.12 | 2.05 | 2 | 1.95 |
| **83** | 3.96 | 3.11 | 2.72 | 2.48 | 2.32 | 2.21 | 2.12 | 2.05 | 2 | 1.95 |
| **84** | 3.96 | 3.11 | 2.71 | 2.48 | 2.32 | 2.21 | 2.12 | 2.05 | 1.99 | 1.95 |
| **85** | 3.95 | 3.1 | 2.71 | 2.48 | 2.32 | 2.21 | 2.12 | 2.05 | 1.99 | 1.94 |
| **86** | 3.95 | 3.1 | 2.71 | 2.48 | 2.32 | 2.21 | 2.12 | 2.05 | 1.99 | 1.94 |
| **87** | 3.95 | 3.1 | 2.71 | 2.48 | 2.32 | 2.21 | 2.12 | 2.05 | 1.99 | 1.94 |
| **88** | 3.95 | 3.1 | 2.71 | 2.48 | 2.32 | 2.2 | 2.12 | 2.05 | 1.99 | 1.94 |
| **89** | 3.95 | 3.1 | 2.71 | 2.47 | 2.32 | 2.2 | 2.11 | 2.04 | 1.99 | 1.94 |
| **90** | 3.95 | 3.1 | 2.71 | 2.47 | 2.32 | 2.2 | 2.11 | 2.04 | 1.99 | 1.94 |
| **91** | 3.95 | 3.1 | 2.71 | 2.47 | 2.32 | 2.2 | 2.11 | 2.04 | 1.98 | 1.94 |
| **92** | 3.95 | 3.1 | 2.7 | 2.47 | 2.31 | 2.2 | 2.11 | 2.04 | 1.98 | 1.94 |
| **93** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.2 | 2.11 | 2.04 | 1.98 | 1.93 |
| **94** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.2 | 2.11 | 2.04 | 1.98 | 1.93 |
| **95** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.2 | 2.11 | 2.04 | 1.98 | 1.93 |
| **96** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.2 | 2.11 | 2.04 | 1.98 | 1.93 |
| **97** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.19 | 2.11 | 2.04 | 1.98 | 1.93 |
| **98** | 3.94 | 3.09 | 2.7 | 2.47 | 2.31 | 2.19 | 2.1 | 2.03 | 1.98 | 1.93 |
| **99** | 3.94 | 3.09 | 2.7 | 2.46 | 2.31 | 2.19 | 2.1 | 2.03 | 1.98 | 1.93 |
| **100** | 3.94 | 3.09 | 2.7 | 2.46 | 2.31 | 2.19 | 2.1 | 2.03 | 1.98 | 1.93 |

## Study Question 4: Accept or Reject

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Treatment | 2 | 25.537 | 12.769 | 19.205 | 5.14 |
| Error | 6 | 3.989 | 0.665 | n/a | n/a |

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Required F (5%) |
|---|---|---|---|---|---|
| Total | 8 | 29.526 | n/a | n/a | n/a |

Query 1.8.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Explanation

Many statistical packages including R will calculate the actual probability of getting the observed F-value by chance, given that the null hypothesis is true. This is listed as Prob > F. If this P-value is below 0.05, we know our calculated F-value would have exceeded the Table F-value and we reject the null hypothesis.
The reason we reject the null hypothesis when the P-value is small is that the P-value is the probability, assuming the null is true, of finding such a result by random chance. When the p-value is small, it is unlikely that the null is actually true. Therefore, we reject it.

### Ex. 1: One-Factor ANOVA of a CRD

**R Code Functions**

- getwd()
- rm(list=())
- str()
- setwd()
- hist()
- as.factor()
- read.csv()
- attach()
- aov()
- rm()
- boxplot()
- summary()

**Activity Objectives**

- Students will conduct exploratory data analyses (EDA) on data from a simple Completely Randomized Design (CRD).
- Assess whether students know how to interpret results from EDA.
- Students will conduct an Analysis of Variance (ANOVA) on data from a simple CRD.

**Source data**

The data can be downloaded here and saved it as CRD.1.data.csv file in the working directory.

**Ex. 1: Read the Data Set into R**

Before you can conduct any analysis on data from a text file or spreadsheet, you must first enter, or read, the data file into the R data frame. For this activity, our data is in the form of an excel comma separated values (or CSV) file; a commonly used file type for inputting and exporting data from R.
Make sure that the data file for this exercise is in the working directory folder on your desktop.

Note: We previously discussed how to set the working directory to a folder named on your desktop. For this activity, we will repeat the steps of setting the working directory to reinforce the concept.
In the Console window (lower left pane), enter getwd(). R will return the current working directory below the command you entered:

getwd()
[1] “C:/Users/[Name]/Documents”

Set the working directory to the folder on your computer. For a folder named ‘wd’ on our desktop, we enter:

setwd(“C:/Users/[Name]/Desktop/wd”)

Now, we want to read the CSV file from our working directory into RStudio. At this point, we learn an important operator: **<-**. This operator is used to name data that is being read into the R data frame. The name you give to the file goes on the left side of this operator, while the command ***read.csv*** goes to its right. The name of the CSV file from your working directory, in this case CRD.1.data.csv, is entered in the parenthesis and within quotations after the *read.csv* command. The command **header = T** is used in the function to tell R that the first row of the data file contains column names, and not data.
Read the file into R by entering into the **Console**:

data <- read.csv(“CRD.1.data.csv”, header=T)

**Tip:** If you are working out of the **Console** and received an error message because you typed something incorrectly, just press the ↑ key to bring up the line which you previously entered. You can then make corrections on the line of code without having to retype the entire line in the console window again. This can be an extremely useful and time saving tool when learning to use a new function. Try it out.
If the data was successfully read into R, you will see the name that you assigned the data in the **Workspace/History** window (top-right).

**Ex. 1: Exploratory Data Analysis**

Let’s do some preliminary exploring of the data.
Read the data set into the R data frame.

data <- read.csv(“CRD.1.data.csv”, header=T)

First, let’s look at a histogram of the yield data to see if they follow a normal distribution. We can accomplish this using the hist command.
Enter into the console:

hist(data$Yield, col=”blue”, main=”Histogram of Yield of 3 Synthetic Maize Populations”, xlab=”Yield (t/ha)”, vlab=”Frequency”)

R returns the histogram (Figure 1) in the **Files/Plots/Packages/Help** window (bottom-right).

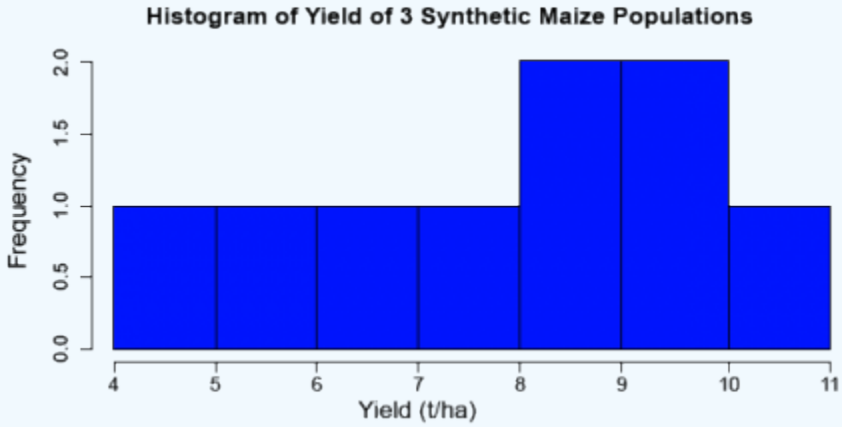


Figure 1 Histogram of yield of 3 synthetic maize populations

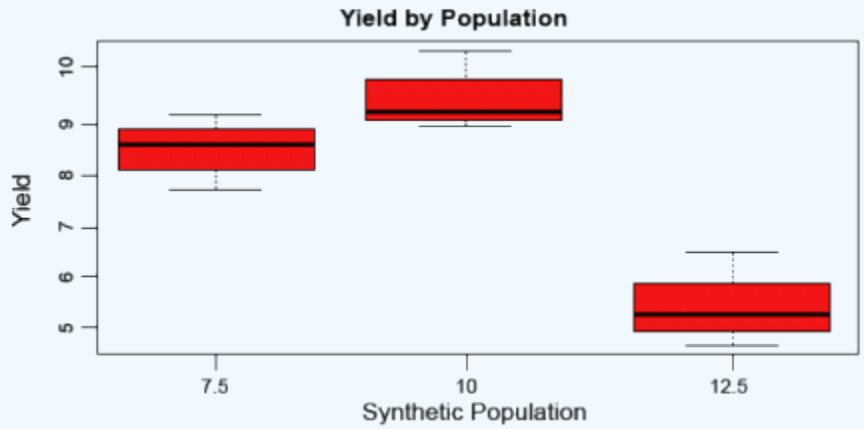
**Ex. 1: Create a Boxplot**

Let's go through the command we just entered: data$Yield specifies that we want to plot the values from the column Yield in the data, col="blue" indicates which color the histogram should be, the entry in quotations after main= indicates the title that you'd like to give the histogram, the entries afterxlab= and ylab= indicate how the x and y axes of the histogram should be labeled. The histogram appears in the bottom-right window in RStudio. The histogram can be saved to your current working directory by clicking 'export' on the toolbar at the top of the lower-right window, then clicking "save plot as PNG" or "save plot as a PDF". You may then select the size dimensions you would like applied to the saved histogram.
Let's now look at some boxplots of yield by population for this data. First, enter into the Console windowattach(data). The attach command specifies to R which data set we want to work with, and simplifies some of the coding by allowing us just to use the names of columns in the data, i.e. Yield vs.data$Yield. After we enter the attach command, we'll enter the boxplot command.

attach(data)

boxplot(Yield~Pop, col="red", main="Yield by Population", xlab="Synthetic Population", ylab="Yield"

R returns the boxplot in the bottom-right window.




Let's go through the boxplot command: Yield~Pop indicates that we want boxplots of the yield data for each of the 3 populations in our data, col= indicates the color that we want our boxplots to be, main= indicates the title we want to give the boxplots, and xlab= and ylab= indicate what we want the x and y axes labeled as.
*Note*: Yield is capitalized in our data file, thus it MUST also be capitalized in the **boxplot** command.

**Ex. 1: Calculate Coefficient of Variance**

The coefficient of variance can be calculated for each population in the data set. Looking at the data, we can see that lines 1 to 3 pertain to population 7.5. We know that the coefficient of variation for a sample is the mean of the sample divided by the standard deviation of the sample. By using the command mean(), we can calculate the mean for a sample. Remember that to specify a column from a data frame, we use the $ operator. If we want to calculate the mean of population 7.5 from the data (rows 1 to 3 in the data), we can enter

mean(data$Yield[1:3])

To calculate the standard deviation of the yield for population 7.5, enter

sd(data$Yield[1:3])

The coefficient of variance is therefore calculated by entering

mean(data$Yield[1:3])/sd(data$Yield[1:3])

## Ex. 1: Carry Out ANOVA

Now that we've gained some intuition about how the data behave, let's carry out an ANOVA with one factor (Pop) on the data. We first need to specify to R that we want Population to be a factor. Enter into the **Console**

Pop<-as.factor(Pop)

Let's go through the command above: as.factor(data$Pop) specifies that we want the Pop column in dataset data to be a factor, which we've called Pop.
Now that we have population as a factor, we're ready to conduct the ANOVA. The model that we are using for this one-factor ANOVA is Yield=Population. In the **Console**, enter

out <- summary(aov(Yield ~ Pop))

Let's look at the ANOVA table. Enter out in the **Console** window.

out

Df Sum Sq Mean Sq F value Pr(>F)
Pop 2 25.537 12.769 19.2 0.00247 **
Residuals 6 3.989 0.665
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Ex. 1: Interpreting the Results

In this ANOVA table, the error row is labeled *Residuals*. In the second and subsequent columns you see the degrees of freedom for *Pop* and *Residuals* (2 and 6), the treatment and error sums of squares (25.537 and 3.989), the treatment mean square of 12.769, the error variance = 0.665, the F ratio and the P value (19.2 and 0.0025). The double asterisks (**) next to the P value indicate that the difference between the yield means of the three populations is significant at 0.1% (i.e., we reject the null hypothesis that the yield means of each population are statistically equivalent). Notice that R does not print the bottom row of the ANOVA table showing the total sum of squares and total degrees of freedom.

### Ex. 2: Wheat Yield Example

**R Code Functions**

- getwd()
- hist()
- as.factor()
- setwd()
- attach()
- aov()
- read.csv()
- boxplot()
- summary()
- head()
- str()

**The Scenario**

You are a data analyst for the respected seed company "Vavilov's Varieties". In an effort to find a source of genetic resistance to a strain (**UG99**) of rust that is plaguing the company's current wheat lines, the company has acquired 300 genetically diverse wheat landraces from central Asia. A test for resistance to the rust strain was done on each of the landraces, and 100 out of the 300 landraces were found to be completely resistant. The company's plan is to introgress via backcrossing the resistance gene/genes from a single rust-resistant landrace into an elite, high yielding cultivar already being sold by the company. A preliminary yield test with each of the 100 resistant landraces planted in 2 reps at a single location was conducted. Your supervisor wants to know if there is a statistically significant difference among the yield results of the landraces, and if it is therefore possible to minimize yield drag by selecting the highest yielding resistant landrace for use in

converting the elite germplasm for resistance to the rust. There are no funds in your budget to acquire commercial statistical software, and in addition your supervisor tells you that he needs to make a decision as soon as possible (precluding you from doing the analysis by hand).

**Source Data**

You can download the data file here and save it as a .CSV file.

**Ex. 2: Enter the Data into R**

Do some exploratory analysis on the data (i.e., create a histogram of yield, boxplots for the yield of each landrace, calculate the mean, standard deviation, and coefficient of variation of yield for all of the landraces). Then, carry out an ANOVA with one factor (LR) on the yield data. Finally, explain the results of your analysis in the context of the problem (i.e. does the ANOVA lead you to accept or reject the null hypothesis that the yields of all of the landraces are statistically equivalent). Finally, make a decision as to which landrace should be entered into your company's wheat breeding program. The R code for this example is almost the same as the Corn Population Example in Activity 8.2. The response variable is still 'Yield' and the factor variable is 'LR' (landrace).
The following code will assist you in carrying out these analyses.
If you are working in the **Script** window of R studio, you can enter comments/descriptions on lines of code by first entering the # symbol, then entering your comment after. R will recognize this line as a comment, and not try to execute the line as command when you enter the code from the **Script** window into the **Console** using CTRL+ENTER. This can be very useful when working in a team setting where you may have to share code with other team members who might not be familiar with some types of analyses or commands R.

Read the CSV data file into R.

```
data<-read.csv("CRD.2.data.csv", header=T)
```

Look at the first few lines of the dataset.

```
head(data)
LR Rep Yield
1 1 1 1.854
2 1 2 1.895
3 2 1 2.157
4 2 2 2.250
5 3 1 1.595
6 3 2 1.777
```

Create a histogram of the yield data from the dataset.

```
hist(data$Yield, col="green", main="Histogram of yield of wheat landraces", xla b="Yield (t/ha)",ylab="Frequency")
```

Create a boxplot of the yield for each landrace.

```
boxplot(data$Yield~data$LR, col="yellow", main="Yield by Population", xlab="Landrace", ylab="Yield (t/ha)")
```

Display the structure of the data frame 'data'.

```
str(data)
'data.frame': 200 obs. of 3 variables:
$ LR : int 1 1 2 2 3 3 4 4 5 5 ...
$ Rep : int 1 2 1 2 1 2 1 2 1 2 ...
$ Yield: num 1.85 1.9 2.16 2.25 1.59 ...
```

Make Landrace (LR) a factor by which to separate the yield data

```
LR <- as.factor(data$LR)
```

```
Perform a one-factor ANOVA

out <- summary(aov(Yield ~ LR))

out
Df Sum Sq Mean Sq F value Pr(>F)
LR 99 9.721 0.09819 2.997 4.71e-08 ***
Residuals 100 3.276 0.03276
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

### Ex. 2: Interpret the ANOVA

Now, let’s interpret the results of the ANOVA.
The P value is given in the ANOVA table as 4.65e-08, or 0.0000000465. Looking at the significance codes at the bottom of the ANOVA table output in R, we can see that this p-value is significant at even the most stringent of significance level (0.001). This leads us to reject the null hypothesis that the yields of the different landraces are statistically the same. What does this mean for us as plant breeders? If the yields of the different landraces are not all statistically equivalent, then there must be some landraces that have higher mean yields than others. We can therefore select the top performing landrace for yield to minimize the effect of yield drag when introgressing the gene/s that confer resistance to rust strain **UG99** into an elite background.

## The Linear Additive Model

A way to conceptualize how treatments, such as plant density or fertilizer, affect crop yields is to write the yield (Y) as composed of an overall mean, a treatment effect, and error. The common way to express this is in the form of a linear model equation.

Later we will use this linear model equation method for understanding more complex situations and experiments. It is important to realize that every designed experiment can be described by a linear additive model and that this model determines the manner in which we apply the ANOVA to the experiment. The linear additive model lists all the sources of variation that are accounted for in the experiment relative to the overall mean. For the one-way analysis of variance, the ANOVA table contains two sources: treatments and error. In experiments containing additional treatments as factors, there will be more sources in the model and these will correspond to additional rows in the ANOVA table.

### The Linear Model Equation

As discussed in the previous sections, the ANOVA decomposes the variability for each part of the experiment. Each effect can be considered as a part of the variability seen in each experimental unit. Since these effects are additive, we can view the result from each unit as a combination of each effect. Symbolically, the linear additive model for ANOVA is:

$$Y_{ij} = \mu + T_i + \epsilon_{(i)j} \tag{1.8.14}$$

Equation 5 Linear model for ANOVA (1.8.15)

**where:**

**$Y_{ij}$** = response observed for the ij$^{th}$ experimental unit
$\mu$ = overall population mean
**$T_i$** = effect of the i$^{th}$ treatment
$\epsilon_{(i)j}$= effect associated with the ij$^{th}$ experimental unit, commonly referred to as error or residual

About these subscripts: “i” refers to the treatment, and “j” refers to the replication of that treatment. So an experiment will have “i*j” experimental units; each unit will be identified by a unique “ij” combination. For example, the response of experimental unit with the 4$^{th}$ replication of the 3$^{rd}$ treatment will be notated as $Y_{34}$.
This is the same as the linear model for the t-test, expanded to more than just two treatments or populations.

## Application

For the corn population experiment we can rewrite the model to indicate the true sources of variation as:

$$Yield_{ij} = \mu + POP_i + PLOT_{(i)j} \tag{1.8.16}$$

Equation 6 Linear model for ANOVA with true sources of variation (1.8.17)

**where:**

**Yield$_{ij}$** = corn yield for the ij$^{th}$ plot
$\mu$ = average corn yield for the experiment
**POP$_i$** = effect of the i$^{th}$ planting population
**PLOT$_{(i)j}$** = effect associated with the ij$^{th}$ plot (residual)

You can think of a linear model as a **ledger**. The dependent variable ($Y_{ij}$) is analogous to the current balance and the mean can be thought of as the initial balance. Each of the other terms (independent variables) are either a credit or a debit. So, in the model above, corn yield in a given plot is simply the value that remains after the average yield has been adjusted for the effects of population and error.

Visual Guide

Query 1.8.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Summary

## ANOVA

- Tests hypothesis that treatment means are same.
- Compares variance from treatment mean differences with that from error.
- Separates the variances in a table, with sources being Treatment, Error, and Total.
- Has degrees of freedom (df) for each source of variation.
- Has sum of squares (SS) for each source.
- Computes variance estimates as mean squares, ms = ss/df.

## F-test

- The null hypothesis (no difference in treatment means) is tested with the F-ratio,F = Treatment MS/Error MS.
- For each alpha level, the table F – value depends on numerator and denominator df.

## Linear Additive Model

- Expresses the partition of Y into overall mean, treatment, and error.
- These components are added to get Y.
- Provides basis for more complex linear models in later units.

**How to cite this chapter**: Moore, K., R. Mowers, M.L. Harbur, L. Merrick, and A. A. Mahama. 2023. The Analysis of Variance (ANOVA). In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.8: The Analysis of Variance (ANOVA) is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ken Moore; Ron Mowers; M. L. Harbur; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.9: Two Factor ANOVAs

In the previous chapter we discussed how one could analyze for differences in treatments, each replicated on several plots, using an analysis of variance. Often there is more than one type of treatment we wish to use in an experiment. For example, we may want to know the effect of different plant populations on yield, but also how different levels of nitrogen application affect the yield. This chapter will build on the principle you learned in the last unit and add another factor to have two factors in the ANOVA.

**Learning Objectives**

- How to analyze a second factor in an ANOVA
- The linear additive model for a two-factor ANOVA
- The advantages of factorial experiments
- How to completely randomize an experiment

## Factorial Experiments

**Factorial experiments involve more than one treatment factor.** In the corn population experiment analyzed in chapter 7 on The Analysis of Variance (ANOVA), we were interested only in the response of one hybrid. However, corn hybrids respond differently to changes in plant population. What if we wanted to compare the response of three different hybrids to population changes? What are our options in designing this experiment?

### Option 1

We could consider each hybrid in a separate experiment. For each hybrid, then, we would plant our plots using each of the three populations. We would need the following experimental units:

- hybrid A x 3 populations x 3 replications = 9 plots
- hybrid B x 3 populations x 3 replications = 9 plots
- hybrid C x 3 populations x 3 replications = 9 plots

*Table 1 ANOVA table for corn planted at three populations in northwest Iowa.*

| Source of Variation | Degrees of Freedom |
|---|---|
| Population | 2 |
| Error | 6 |
| Total | 8 |

The three experiments would require 27 (9+9+9) total experimental units, in this case plots. The results from this experiment would tell us the response of each hybrid to population changes. Each experiment would have an ANOVA table (Table 1) similar to that used previously.

### Combining Factors

Yet, since each hybrid test is conducted as a separate experiment, we could not compare the yields of each hybrid. In other words, we would know which population is most appropriate for each hybrid. But, we could not predict which hybrid would produce the greatest yield with any confidence. The experiments were simply not designed in a manner to allow this.

### Option 2

We could develop an experimental arrangement where these two factors are combined into a factorial experiment. In this case,

all combinations of population and hybrid would occur within the same experiment. The total number of treatments would be 9 (3 hybrids x 3 populations). In this case, with three replications, we would have: 3 hybrids x 3 populations x 3 replications = 27 plots.

*Table 2 ANOVA table for three corn varieties planted at three populations in Northwest Iowa.*

| Source of Variation | Degrees of Freedom |
|---|---|
| Population | 2 |
| Hybrid | 2 |
| Interaction | 4 |
| Error | 18 |
| Total | 26 |

However, this factorial experiment will produce more useful information than the first design. Specifically, this experiment will tell us whether the effect produced by changing plant population is the same regardless of hybrid, or if each hybrid reacts differently to population. If each hybrid reacts differently to population, we will say that there is an interaction between the hybrids and plant populations. To analyze the results from this experiment, the ANOVA must be expanded to include the additional effects: another main factor and its interaction with the original main factor (Table 2).

### Degrees of Freedom

The table is similar to the one used before, with the exception of these two new sources of variation. The degrees of freedom are calculated using the formulae in Table 3. Note that even though we have the same number of experimental units whether we evaluate the three hybrids separately (Table 3) or together (Table 3), our residual degrees of freedom and the sensitivity of our F-test increase dramatically with factorial design.

*Table 3 Degrees of freedom factorial experiment*

| Source of Variation | Degrees of Freedom |
|---|---|
| **Treatment A** | # of levels of treatment A-1 |
| **Treatment B** | # of levels of treatment B-1 |
| **Interaction** | (df for Trt A) x (df for Trt B) |
| **Error** | (# of levels of Trt A) x (# of levels of Trt B) x (# of replications – 1) |
| **Total** | (# of levels of Trt A) x (# of levels of Trt B) x (# of replications) – 1 |

The ANOVA for factorial experiments can be completed using a “cookbook” method similar to that described in the previous chapter. However, since most experiments of this type are analyzed using computer software, we will not continue the example here. Rather, we will use the computer to analyze a similar experiment using R to calculate the statistics we need.

## Interaction

**Interaction** is a differential response of one factor at different levels of another. In addition to maximizing the efficiency of an experiment by combining the treatments into one set of plots, a two-factor ANOVA introduces another effect; the interaction between the treatments. As introduced in the ANOVA Table 1, there is an additional source of variation in the analysis: the interaction between our two main treatment factors.

As is often realized, additional effects occur because of interaction between two treatments. Two factors cannot interact at all, interact positively, or interact negatively. These can be depicted well by graphs. Let’s consider two different varieties at three different levels of N.

## Example 1: No Interaction

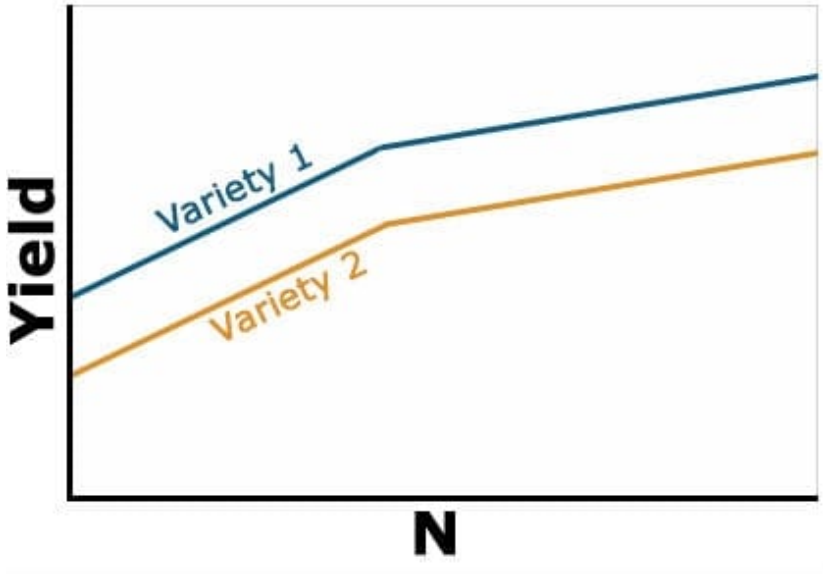



Figure 1 No Interaction

Here yield of the two varieties reacts similarly to N-rate. Thus, there is no interaction between N and variety (Figure 1). Notice that the lines are parallel, so the effect of the variety is constant as N rate increases. Variety 1 is uniformly superior at each N level and higher levels of N result in improved yield of each variety.

## Example 2: Positive Interaction

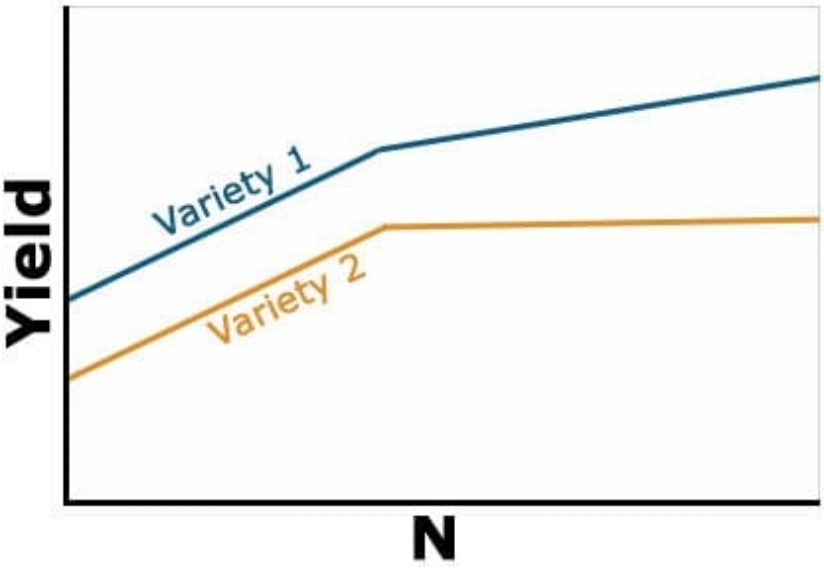



Figure 2 Positive Interaction

Yield of the two varieties differs based on level of N applied, but the variety 1 always yields better than variety 2 (Figure 2). The response of each variety depends on the level of N applied. It is no longer constant as in the previous example. The type of interaction is sometimes referred to as a change in the magnitude of the response interaction. Variety 1 is still the highest yielding variety, but the magnitude of its response to N depends on the amount of N applied. With this type of interaction, you can't evaluate the effect of N independent of variety because the response is different depending on which variety you are talking about. Whenever an interaction tests as significant in the ANOVA, it is important to ignore the main factor means and focus on the interaction means.

## Example 3: Negative Interaction

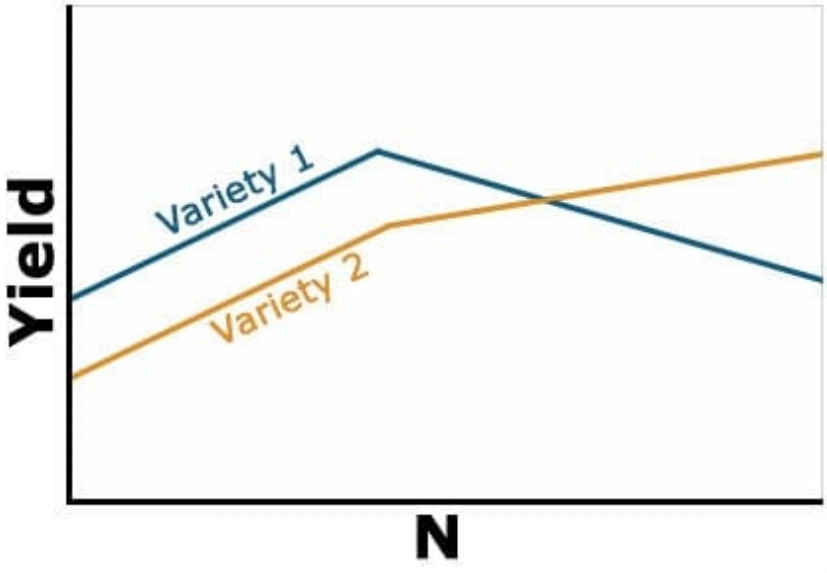



Figure 3 Negative Interaction

The last type of interaction occurs when the yield response is completely opposite based on the level of N (Figure 3). Note that the lines in the figure now cross each other. This changes the interpretation drastically because the best variety to grow now depends on the level of N that is applied. You would need to fertilize each variety differently based on the response. This is often called a crossover interaction and it is very important to pay attention when they occur because your recommendation about one factor depends heavily on the level of the other. Often when a crossover interaction occurs one or both main factors involved in the interaction tests as nonsignificant in the ANOVA. When this occurs, you should look at the results very carefully because failure to assess the interaction carefully can result in poor recommendations.

# Linear Additive Model for Two-Factor ANOVA

The linear model for a factorial is a simple extension of the linear model. The addition of factors and their possible interaction produce further complexity in the ANOVA. But the effects can be added just as the previous effects were to the linear model.

The linear model for a two-factor ANOVA is:

$$Y_{ijk} = \mu + A_i + \beta_j + A\beta_{ij} + \epsilon_{(ij)k} \tag{1.9.1}$$

**where:**
$Y_{ijk}$ = response observed for the $ijk^{th}$ experimental unit,
$\mu$ = overall mean,
$A_i$ = effect of the $i^{th}$ level of factor A,
$\beta_j$ = effect of the $j^{th}$ level of factor B,
$A\beta_{ij}$ = effect of the interaction between the $i^{th}$ level of factor A and the $j^{th}$ level of factor B,
$\epsilon_{(ij)k}$ = effect associated with the $ijk^{th}$ experimental unit; commonly referred to as error.

## True Sources of Variation

Notice that the only change in the linear model from equation 8 is that the treatment structure is modified. We now have sources for factor A, factor B and their interaction. The error structure is not changed; we still have only a single random error for the experiment. Factorial refers to the treatment structure, not the assignment of treatments to experimental units.

For the corn population x hybrid experiment we can rewrite the model to indicate the true sources of variation as:

$$\text{Yield}_{ijk} = \mu + POP_i + HYBRID_j + PH_{ij} + PLOT_{(ij)k} \tag{1.9.2}$$

**where:**
$\text{Yield}_{ijk}$ = corn yield observed for the $ijk^{th}$ plot,
$\mu$ = average yield for the experiment,
$POP_i$ = effect of the $i^{th}$ plant population,
$HYBRID_j$ = effect of the $j^{th}$ hybrid,
$PH_{ij}$ = effect of the interaction between the $i^{th}$ population and the $j^{th}$ hybrid,
$PLOT_{(ij)k}$ = random error effect of the $ijk^{th}$ plot.

**Try: Running an ANOVA for a Two-factor CRD in the next screens**

# Ex. 1: Running an ANOVA for a Two-Factor CRD

## R Code Functions

- setwd()
- aov()
- summary()
- <-
- attach()
- subset()
- read.csv()
- detach()
- pf()
- head()
- as.factor()
- interaction.plot()
- as.data.frame()
- aov()

## The Scenario

You are an employee for a maize development company in charge of developing new high-yielding maize hybrids for use in central Iowa. For the past 2 years you have been developing A-lines and now you want to test their general combining ability with the company's elite R-line. Unfortunately, (because it's Iowa), your plots are savaged by a tornado and bowling ball-sized hail and only 2 of your candidate inbreds remain. You cross these ridiculously fortunate inbreds to the R-line to produce seed for 2 hybrids which are planted the following spring in 3 locations across central Iowa utilizing a randomized complete design with 3 reps at each location. You also plant a standard maize hybrid along with the two you developed to serve as a check within your field (the check is Hybrid C). With the help of your intern, you harvest each plot and calculate the projected bushels/acre.

**Source data:** The yield data from the hybrids can be found in the file ANOVA 2factorCRD [XLSX]

## Ex. 1: Data Set

In this data set, you can see that we have columns for treatment, location, hybrid, replication, and the yield in t/ha (Figure 4). It's good to remember that while you have 3 hybrids and 3 locations, your total number of treatments is 9, not just the 3 hybrids or the 3 locations. In this activity you will learn how to run a 2-factor ANOVA that will help you choose which of the 3 hybrids you will select and move on to the next stage of testing in your program.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Treatment | Location | Hybrid | Rep | Yield |
| 2 | 1 | 1 | A | 1 | 7.446 |
| 3 | 1 | 1 | A | 2 | 9.844 |
| 4 | 1 | 1 | A | 3 | 11.178 |
| 5 | 2 | 1 | B | 1 | 9.13 |
| 6 | 2 | 1 | B | 2 | 9.269 |
| 7 | 2 | 1 | B | 3 | 9.864 |
| 8 | 3 | 1 | C | 1 | 8.638 |
| 9 | 3 | 1 | C | 2 | 7.835 |
| 10 | 3 | 1 | C | 3 | 9.194 |
| 11 | 4 | 2 | A | 1 | 8.252 |
| 12 | 4 | 2 | A | 2 | 5.476 |
| 13 | 4 | 2 | A | 3 | 11.98 |
| 14 | 5 | 2 | B | 1 | 8.844 |
| 15 | 5 | 2 | B | 2 | 9.822 |
| 16 | 5 | 2 | B | 3 | 11.38 |
| 17 | 6 | 2 | C | 1 | 10.467 |
| 18 | 6 | 2 | C | 2 | 9.288 |
| 19 | 6 | 2 | C | 3 | 8.992 |
| 20 | 7 | 3 | A | 1 | 10.363 |
| 21 | 7 | 3 | A | 2 | 10.875 |
| 22 | 7 | 3 | A | 3 | 10.722 |
| 23 | 8 | 3 | B | 1 | 6.191 |
| 24 | 8 | 3 | B | 2 | 8.842 |
| 25 | 8 | 3 | B | 3 | 8.496 |
| 26 | 9 | 3 | C | 1 | 6.638 |
| 27 | 9 | 3 | C | 2 | 5.45 |
| 28 | 9 | 3 | C | 3 | 4.724 |

Figure 4 Two factor experimental data file.

## Activity Objectives

- Build and run an ANOVA for a 2 factor model that accounts for population, hybrid, and the interaction between population and hybrid
- Assess whether each individual hybrid is significantly affected by location

### Ex. 1: Run the ANOVA

First you need to read in the data set.

Ex9.1<-read.csv(“ANOVA 2factorCRD.csv”, header=T)

head(Ex9.1)

Treatment Location Hybrid Rep Yield
1 1 1 A 1 7.446
2 1 1 A 2 9.844
3 1 1 A 3 11.178
4 2 1 B 1 9.130
5 2 1 B 2 9.269
6 2 1 B 3 9.864

This may not be necessary, but you can use this line of code to make sure that your data is read as a data frame.

Ex9.1<-as.data.frame(Ex9.1)

### Ex. 1: Make Adjustments

Check the structure of the data frame; notice that ‘Location’ is considered to be an integer because it’s listed as numbers in the .csv file.

str(Ex9.1)

‘data.frame:27 obs. of 5 variables:
$ Treatment: int 1 1 1 2 2 2 3 3 3 4 ...
$ Location : int 1 1 1 1 1 1 1 1 1 2 ...

$ Hybrid : Factor w/ 3 levels “A”,”B”,”C”: 1 1 1 2 2 2 3 3 3 1 ...
$ Rep : int 1 2 3 1 2 3 1 2 3 1 ...
$ Yield : num 7.45 9.84 11.18 9.13 9.27...

You will need to change the ‘Location’ to a factor in order for the analysis to work.

Ex9.1$Location<-as.factor(Ex9.1$Location)

Location<-as.factor(Ex9.1$Location)

Check the structure again and observe that ‘Location’ is now considered to be a factor.

attach(Ex9.1)

## Ex. 1: Two-Way ANOVA

Run the ANOVA using the aov() function, just like with one-way designs. The only difference is that now we want to include both factors and their interaction in our model. There are two equivalent ways to do this. The first way explicitly specifies each term; the second way is a shortcut.

str(Ex9.1)

‘data.frame’: 27 obs. of 5 variables:
$ Treatment: int 1 1 1 2 2 2 3 3 3 4 ...
$ Location : int 1 1 1 1 1 1 1 1 2 ...
$ Hybrid : Factor w/ 3 levels “A”,”B”,”C”: 1 1 1 2 2 2 3 3 3 1 ...
$ Rep : int 1 2 3 1 2 3 1 2 3 1 ...
$ Yield : num 7.45 9.84 11.18 9.13 9.27 ...

Or:

Ex9.1.outB = aov(Yield ~ Location*Hybrid, data = Ex9.1)

summary(Ex9.1.outB)

Df Sum Sq Mean Sq F value Pr(>F)
Location 2 9.45 4.726 2.109 0.1504
Hybrid 2 13.09 6.544 2.920 0.0797 .
Location:Hybrid 4 30.25 7.562 3.374 0.0316 *
Residuals 18 40.34 2.241
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

## Ex. 1: Run Individual ANOVAs

The significant interaction between Population and Hybrid type indicates the simple effects of one factor differ among levels of another factor. One research questions for this example might ask if the yields of each Hybrid type are the same across three Locations. To test the simple main effect of each Hybrid, across three levels of Location, we can perform the analysis as follows:

1. Separate the data into subsets, based on Hybrid type
2. Run an ANOVA testing the effect of Location in each subset

A = subset(Ex9.1, Hybrid == “A”)

B = subset(Ex9.1, Hybrid == “B”)

C = subset(Ex9.1, Hybrid == “C”)

A.out = summary(aov(Yield ~ Location, A))

A.out

Df Sum Sq Mean Sq F value Pr(>F)
Location 2 6.544 3.272 0.687 0.539
Residuals 6 28.593 4.765

B.out = summary(aov(Yield ~ Location, B))

B.out

Df Sum Sq Mean Sq F value Pr(>F)
Location 2 7.562 3.781 2.935 0.129
Residuals 6 7.729 1.288

C.out = summary(aov(Yield ~ Location, C))

C.out

Df Sum Sq Mean Sq F value Pr(>F)
Location 2 25.594 12.80 19.11 0.0025 **
Residuals 6 4.019 0.67
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Ex. 1: Simple Main Effects

Stop and think! Look at the MSE in these simple main effect ANOVAs. They are pretty different. Also, compare $df_{residual}$ to the full model ANOVA. The $df_{residual}$ (6) are lower in the simple main effect ANOVAs. Since we are comfortable that homogeneity of variance is a reasonable assumption for our data, it is probably safe to use the pooled error for these tests. We can do some calculations to use the pooled error from the overall ANOVA, and its larger number of df (18).

Use pooled error to calculate F and P values for the simple main effects of Location at Hybrid A, B, and C:

Fhybrid.A <- 3.272/2.241

Fhybrid.A

[1] 1.460062

pf(Fhybrid.A, 2, 18, lower.tail = F)

[1] 0.2584485

Fhybrid.B <- 3.272/2.241

Fhybrid.B

[1] 1.687193

pf(Fhybrid.B, 2, 18, lower.tail = F)

[1] 0.2130147

Fhybrid.C <- 12.80/2.241

Fhybrid.C

[1] 0.2130147

pf(Fhybrid.C, 2, 18, lower.tail = F)

[1] 0.2130147

Notice that the P value (0.012) of the test for Hybrid C is less than 0.05, so we would conclude that the yield of this type of Hybrid is affected by Location. The F tests of the Hybrid A and B are not significant at P = 0.05, so we conclude that the yields of these two Hybrids are not influenced by Location.

## Ex. 1: Plot the Interaction

attach(Ex9.1)

interaction.plot(Location, Hybrid, Yield, col = c("green", "red", "blue"), main="Interaction plot of Hybrids A, B, and C")

The attach() function can be used to make objects within data frames accessible in R with fewer keystrokes. The interaction plot shows that the mean response to Location depends upon the level of Hybrid. Also, we know from our previous tests that Location doesn't affect the yields of Hybrid A and B, but it has an impact on yield of Hybrid C.

## Ex. 1: Interaction Plot

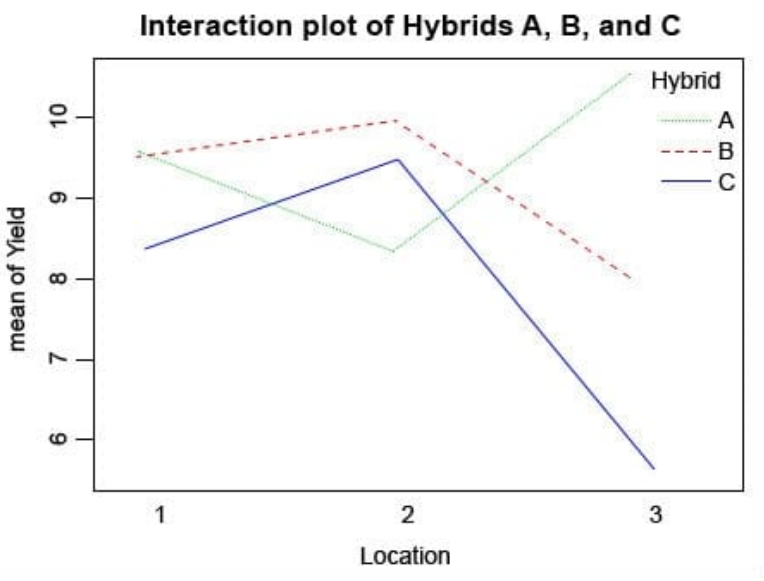



Figure 5 Interaction plot of hybrids A, B, and C

Looking at this interaction plot, you can see that the rank of each line changes depending on the location (Figure 5). This indicates that there is an interaction between location and environment, but is this interaction significant? Go back to the first ANOVA that you ran and check to see if the interaction is significant, and in this case it is. Based on this information you have gathered from the ANOVAs and your interaction plot, which hybrid would you advance to further testing? Would you choose any hybrid at all? What else do you need to know in order to make this decision?

## Ex. 1: ANOVA for 10 Hybrids

Now imagine that the following winter you find some residual seed and replant the inbred lines that were destroyed in the tornado. This time the weather is more cooperative and the following summer you are able to test 10 hybrids at 10 locations (hybrid 10 is the check this time). Use the file "9.1 larger data set" and run the same code as before for the full ANOVA and the interaction plot.

set2<-read.csv("lesson 9.1 larger set.csv", header=T)

attach(set2)

set2$Location<-as.factor(set2$Location)

Location<-as.factor(set2$Location)

```
set2$Hybrid<-as.factor(set2$Hybrid)

Hybrid<-as.factor(set2$Hybrid)

detach(set2)

attach(set2)

str(set2)

'data.frame': 300 obs. of 8 variables:
$ Hybrid : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
$ Location : Factor w/ 10 levels "1","2","3","4",..: 1 1 1 2 2 2 3 3 3 4 ...
$ Rep : int 1 2 3 1 2 3 1 2 3 1 ...
$ overall.Mean: int 170 170 170 170 170 170 170 170 170 170 ...
$ G : int 6 6 6 6 6 6 6 6 6 6 ...
$ E : num 3.2 11.84 8.05 -1.97 7.41 ...
$ error : num -0.209 -0.941 8.491 1.868 0.953 ...
$ Yield : num 179 187 193 176 184 ...

set2out <- aov(Yield ~ Location+Hybrid+Location:Hybrid, data = set2)

summary(Ex9.1.outA)

Df Sum Sq Mean Sq F value Pr(>F)
Location 9 822 91.3 1.430 0.177
Hybrid 9 18671
Location:Hybrid
Residuals
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

interaction.plot(Location, Hybrid, Yield, col = c("green", "red", "blue", "orange", "black"), main="Interaction plot of Hybrids 1-10")
```

## Ex. 1: Interaction Plot of 10 Hybrids

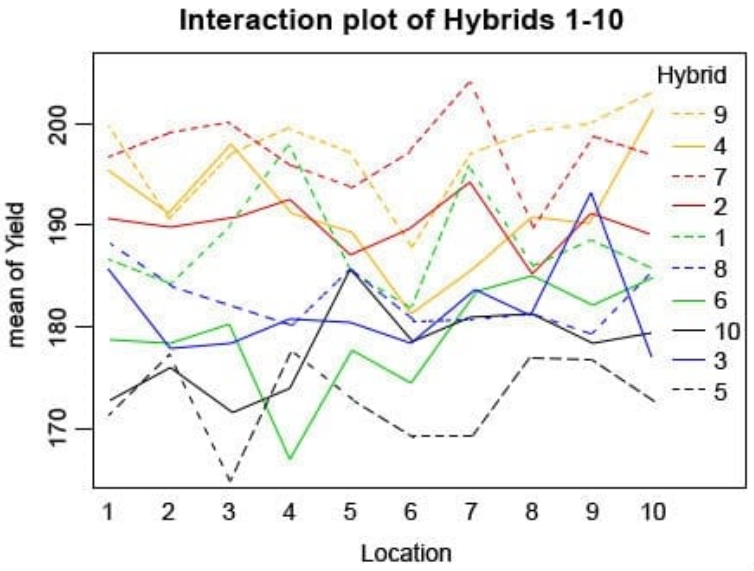



Figure 6 Interaction plot of hybrids 1 through 10

Looking at the plot, you can see that many of the hybrids change rank in different locations (Figure 6), but if you look at the ANOVA you can see that despite this the interaction is not significant. In real life, you will likely have even larger data sets with

many individuals in which case the interaction plot will look very messy and even more difficult to read. This is why it is important to always check the ANOVA and not just rely on the plot.

## Ex. 1: Review Questions

1. What information can you get from an ANOVA?
2. How can you use the ANOVA to make selection decisions?
3. What is an interaction?

*Table 4 R code glossary terms and the functions they perform.*

| **R Code Glossary** | |
|---|---|
| **setwd(“”)** | Set the working directory. Make sure to use the file path where you downloaded your data sets, and not the example I have included! |
| **<-** | Assignment Operator. Assign value to a variable. Ex: X<-1 means “X gets t”. An = sign does the same thing. |
| **read.csv(“”)** | Read in a .csv file. Make sure your file name ends in .csv and if you have column names, you need to specify that header=T |
| **head(mydataframe)** | Returns the top part of the data set specified in the () |
| **as.data.frame(mydataframe)** | Changes the format of an R object (i.e., a data set) to a data frame. This is one of the more common formats for working with a data set. |
| **str(mydataframe)** | Returns the structure of an R object |
| **attach(mydataframe)** | Attach an R object so that R knows this is what you want to work with at this moment. |
| **detach(mydataframe)** | Detaches the R object you were working with |
| **as.factor(mydataframe$variance)** | Changes a variable within an R object to a factor variable. An example is when you have variables designated with numbers but they are meant to be categorical variables so you use this function to tell R that. |
| **aov(y ~ A + B + A:B, data=mydataframe)** | Perform a 2-factor analysis of variance on an R object. |
| **summary()** | Returns the summary of an analysis. |
| **subset(mydataframe, variable == “variable value”)** | Subsets a variable within a data frame based on a particular variable value. |
| **pd(F-value, df1, df2, lower.trial = F)** | Calculate the P-value. Lower.tail=F means P[X > x] |
| **interaction.plot(x.factor, trace.factor, response,...)** | Create interaction plot. X.factor is the factor that forms the x-axis, trace.factor is another factor whose levels for the traces, response is the numeric variable giving the response. You can also add specific colors to the plot with col=c”color1”, “color2” and a title with main = “title” |

Study Questions: Linear Additive Model for Two-Factor Anova

Query 1.9.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.9.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.9.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.9.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# ANOVA and Experimental Design

## Experimental Design and Analysis

You will recall from Chapter 1 on Basic Principles that experimental design refers to the manner in which treatments are assigned to experimental units (plots). In this chapter, we have assumed that all treatments are applied at random to the entire set of experimental units used in the experiment. This is what is known as a Completely Random Design (CRD). In a CRD every experimental unit has the same chance of receiving any given treatment.

## Experimental Design

**Experimental design allows researchers to control factors influencing outcomes.**

Statistical analysis is a powerful approach to understanding collections of data. The analysis employed depends on the type of data and the manner in which it was collected. There are two broad categories or approaches to research that commonly are used: observational experiments and designed experiments.

Observational experiments involve collecting data from a population of individuals to which no treatments have been applied. They are descriptive in nature and usually involve studying the relationships among two or more variables of interest. It is important to understand that the variables studied in an observational experiment occur naturally and are not manipulated by the researcher in any way. An example of an observational experiment would be a comparison of groundwater nitrate concentrations among several Iowa counties.

Designed experdesigned experimentsiments differ from observational experiments in that data are collected from units that have been manipulated by the researcher in some way before the data are collected. This is often described as applying treatments to experimental units. Some good agricultural examples of treatments are the application of specific fertilizer rates and the planting of specific crop varieties for the purposes of comparison. In agronomic terms, the smallest entity to which treatments are applied is usually a field plot.

**Characteristics of designed experiments:**

- Replication – treatments are repeated two or more times on different experimental units (plots)
- Randomization – treatments are randomly assigned to experimental units (plots)
- Design Control – how treatments are applied to various groupings and sizes of experimental units (subplots, plots, blocks, locations)

We use a CRD when we expect the magnitude of the plot effect to be similar among all the plots used in the experiment. Statisticians refer to this condition as homogeneity, and it is assumed when we use a CRD. The CRD is a common design (Figure 7); there are many cases where its use is appropriate. For example, in a growth chamber, we might have 25 flats filled with sand, each planted with 30 wheat seeds, 5 reps of each of 5 varieties.

| Range \ Plot | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 100 | 2 | 1 | 3 | 5 | 4 |
| 200 | 5 | 4 | 4 | 4 | 5 |
| 300 | 3 | 1 | 2 | 1 | 1 |
| 400 | 3 | 3 | 2 | 5 | 2 |

Figure 7 Randomization for a CRD with 5 treatments and 4 replications.

However, whenever there are not enough plots with similar characteristics to accommodate all the treatments and replications in an experiment, alternative designs should be considered to improve the precision of the experiment. We will learn more about this as we study other designs in subsequent chapters.

## Randomly Assign Treatments

There are many ways to randomly assign treatments to experimental units. The use of a table of random numbers is described

in the text. This method involves listing all the treatments and their replications and assigning a random number from a table of random numbers to each one. This random order is then matched to a predetermined order of experimental units. A far easier approach is to use a computer to randomize treatments. You can use Excel or R to randomize an experiment.

# Ex. 2: Randomized Complete Design using R

## R Code Functions

- setwd()
- paste()
- sample()
- <-
- data.frame()
- write.csv()
- read.csv()
- write.table()
- as.data.frame()
- head()
- matrix()

## The Scenario

You are an employee for WinField Solutions interested in studying the impact of Japanese beetles (Popillia japonica) on soybeans and recently two new insecticides have come on the market. To test their effectiveness against Japanese beetles, you choose to test the insecticides on the three most commonly grown cultivars in Iowa. You want to design an experiment where each insecticide is paired with each cultivar and replicated four times. To account for differences within your field you want to test these pairs in a completely randomized design. Ultimately you wish to be able to show clients which insecticide/cultivar pairs are the most effective at preventing damage from Japanese beetles.

## Ex. 2: Activity Objectives

1. Randomize a list of all the insecticide/cultivar pairs.
2. Assemble the list into a rectangular field plot.

Start by setting the working directory. As always, you need to use your own chosen directory, this is just an example.

setwd(“C:/Users/UserName/Desktop/SAS to R”)

After you read in the data, be sure to check the head to make sure it was read in properly.

crd<-read.csv(“Randomization 2factor CRD.csv”, header=T)

head(crd)

Insecticide Cultivar

1 1 1

2 2 1

3 1 2

4 2 2

5 1 3

6 2 3

## Ex. 2: Randomize as Pairs

Because we have two types of treatments and we want them to be randomized as pairs, it makes sense to have insecticide and cultivar treatments be represented as a single value such as '1–2' to represent insecticide 1 and cultivar 2. It will be up to you to decide which insecticide and cultivar will be represented by each number. To do this in R, we can merge the two columns and separate them with '–' with this code:

IC <- as.factor(paste(crd$Insecticide, crd$Cultivar, sep = "–"))

IC

[1] 1–1 2–1 1–2 2–2 1–3 2–3 1–1 2–1 1–2 2–2 1–3 2–3 1–1 2–1 1–2 2–2 1–3 2–3 1–1 2–1 1–2 2–2

[23] 1–3 2–3

Levels: 1–1 1–2 1–3 2–1 2–2 2–3

Originally, I used just one '-' between the numbers but I found that when I imported this into Excel, it is read as a date and it is not easy to get Excel to display them properly. Now that we have made a vector of the treatment combinations, we need to create another vector of numbers that will list the order of the plots, so we need the list to be the same length as the total number of plots (24). This is important for when we randomize the order.

v<-1:24

v

[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

Now we combine the treatments with the vector we just created.

data <- data.frame(Treatment=IC, tempplot=v)

data

Treatment tempplot
1 1–1 1
2 2–1 2
3 1–2 3
4 2–2 4
5 1–3 5
6 2–3 6
7 1–1 7
8 2–1 8
9 1–2 9
10 2–2 10
11 1–3 11
12 2–3 12
13 1–1 13
14 2–1 14
15 1–2 15
16 2–2 16
17 1–3 17
18 2–3 18
19 1–1 19
20 2–1 20
21 1–2 21
22 2–2 22
23 1–3 23
24 2–3 24

## Ex. 2: Randomize the Order

Now that we have created the data set that we can work with, we can randomize the order of the treatments.

randdata <- data[sample(1:nrow(data)),]

randdata

Treatment tempplot
13 1–1 13
22 2–2 22
4 2–2 4
23 1–3 23
2 2–1 2
24 2–3 24
12 2–3 12
17 1–3 17
9 1–2 9
7 1–1 7
18 2–3 18
8 2–1 8
19 1–1 19
21 1–2 21
11 1–3 11
3 1–2 3
20 2–1 20
5 1–3 5
15 1–2 15
6 2–3 6
1 1–1 1
10 2–2 10
16 2–2 16
14 2–1 14

Now if you look at the order of the treatments, you can see that they have been completely randomized. However, because the list of 1-24 is also out of order, this can be a little confusing if you are trying to get a planting plan together. We can keep the treatments randomized but we can have the plot numbers be in order with this code:

treatment<-as.factor(randdata$Treatment)

treatment

[1] 1–1 2–2 2–2 1–3 2–1 2–3 2–3 1–3 1–2 1–1 2–3 2–1 1–1 1–2 1–3 1–2
[17] 2–1 1–3 1–2 2–3 1–1 2–2 2–2 2–1
Levels: 1–1 1–2 1–3 2–1 2–2 2–3

treatment<-as.factor(randdata$Treatment)

finaldata

Treatment tempplot
1 1–1 1
2 2–2 2
3 2–2 3
4 1–3 4
5 2–1 5
6 2–3 6
7 2–3 7
8 1–3 8

9 1–2 9
10 1–1 10
11 2–3 11
12 2–1 12
13 1–1 13
14 1–2 14
15 1–3 15
16 1–2 16
17 2–1 17
18 1–3 18
19 1–2 19
20 2–3 20
21 1–1 21
22 2–2 22
23 2–2 23
24 2–1 24

## Ex. 2: Matrix Form

Finally, if you know you want to plant these soybean plants in a rectangular field, you can use this code to convert your randomized treatments to a matrix form.

A = matrix((treatment), nrow=4, ncol=6)

A

[,1] [,2] [,3] [,4] [,5] [,6]
[1,] “1–1” “2–1” “1–2” “1–1” “2–1” “1–1”
[2,] “2–2” “2–3” “1–1” “1–2” “1–3” “2–2”
[3,] “2–2” “2–3” “2–3” “1–3” “1–2” “2–2”
[4,] “1–3” “1–3” “2–1” “1–2” “2–3” “2–1”

Remember, this is a completely randomized design, not a randomized complete block design. You may well have a field design where you have a section of the same cultivar or insecticide repeated several times, but it will still be random.

You can export your randomized list and field layout as a text file or .csv file.

write.table(finaldata, file = “Randomized list.txt”)

write.table(A, file = “field layout.txt”)

write.csv(finaldata, file = “Randomized list.csv”)

write.csv(A, file = “field layout.csv”)

## Ex. 2: Visualizing Results in Excel

This way you can save your results and you can do further visualization in programs like Excel.

For example (Figure 8):

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| | | V1 | V2 | V3 | V4 | V5 | V6 |
| 1 | | 1--2 | 2--3 | 2--2 | 1--3 | 1--2 | 2--2 |
| 2 | | 1--1 | 1--3 | 2--2 | 1--3 | 1--1 | 1--1 |
| 3 | | 1--3 | 2--1 | 1--2 | 2--3 | 2--2 | 2--1 |
| 4 | | 1--1 | 2--1 | 1--2 | 2--3 | 2--3 | 2--1 |

Figure 8 Treatment assignments in Excel.

You can start with this and add ranges and row labels to get a clearer idea of what the field layout will be like Figure 9.

| | | | | | | |
|---|---|---|---|---|---|---|
| | North | | | | | |
| Range4 | 1--2 | 2--3 | 2--2 | 1--3 | 1--2 | 2--2 |
| Range3 | 1--1 | 1--3 | 2--2 | 1--3 | 1--1 | 1--1 |
| Range2 | 1--3 | 2--1 | 1--2 | 2--3 | 2--2 | 2--1 |
| Range1 | 1--1 | 2--1 | 1--2 | 2--3 | 2--3 | 2--1 |
| | Row1 | Row2 | Row3 | Row4 | Row5 | Row6 |
| | South | | | | | |

Figure 9 Treatment arrangements with rows and ranges.

## Ex. 2: R Code Glossary

*Table 5 R code glossary terms and the functions they perform.*

| **R Code Glossary** | |
|---|---|
| **setwd("")** | Set the working directory. Make sure to use the file path where you downloaded your data sets, and not the example I have included! |
| **<-** | Assignment Operator. Assign value to a variable. Ex: X<-1means "X gets 1". An = sign does the same thing. |
| **read.csv("")** | Read in a .csv file. Make sure your file name ends in .csv and if you have a column names, you need to specify that header=T. |
| **head(mydataframe)** | Returns the top part of the data set specified in the (). |
| **as.factor(mydataframes$variable)** | Changes a variable within an R object to a factor variable. An example is when you have variables designated with numbers but they are meant to be categorical variables so you use this function to tell R that. |
| **paste(...,sep="")** | Concatenate strings after using sep string to seperate them.<br>paste("x",1:3,sep="") returns c("x1","x2","x3")<br>paste("x",1:3,sep="M") returns c("xM1","xM2","xM3")<br>paste("Today is", date()) |
| **data.frame()** | Combines variables into a single data frame. |
| **sample()** | Returns a random permutation of a vector. |

| | |
|---|---|
| **Matrix((variable), number of rows=, number of columns)** | Takes a vector and transforms it into a matrix with the specified dimensions of rows and columns. You need row and column numbers that multiply to be the total number of individuals, or you will get an error. |
| **write.csv(mydataframe, ".csv")** | Write your data frame to a .csv file. |
| **write.table(mydataframe, ".txt")** | Write your data frame to a .txt file. |

## Error Structure

Once the randomization is completed and the experiment properly conducted according to plan, the error structure for the experiment is, at least partially, determined. For the CRD, for example, we do not have a blocking factor because treatments are assigned to experimental units completely at random. In chapter 11 on Randomized Complete Block Design, we will introduce another type of design, the Randomized (Complete) Block Design, or RCBD, in which some restrictions are put on the randomization. For the RCBD, there is yet another recognizable source of variation in the ANOVA, that for blocks, as you will see in chapter 11.

# Summary

## Factorial Experiments

- Involve more than one treatment factor.
- Allow exploration of combinations of factors, interaction
- Refers to the treatment structure, not how treatments are assigned to plots.

## Interaction

- Differential response of one factor at different levels of another.

## Linear Model

- Factorial Linear Model is simple extension for more detail on treatment factors and interaction.

## Experimental Design

- Determines model and ANOVA.
- CRD has treatments assigned to experimental units completely at random.

**How to cite this chapter**: Mowers, R., M., L. Harbur, K. Moore, L. Merrick, A. A. Mahama. 2023. Two Factor ANOVAs. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.9: Two Factor ANOVAs is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; M. L. Harbur; Ken Moore; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.10: Mean Comparisons

The analysis of variance is useful for testing the hypothesis that one or more treatments applied in an experiment have a significant effect or response. If we declare the F-test to be significant, we can say, within the limits of probability allowed by our test, that at least one of the treatments tested is significantly different from the others. However, with the exception of the limited case where only two treatments are compared, the ANOVA does not indicate how treatment responses vary from one another. In order to answer this important question, we must employ other statistical tests commonly referred to as mean comparison procedures.

**Learning Objectives**

- About the various approaches used to compare treatment means
- How to use the least significant difference (LSD) to test the difference between adjacent means
- How to use HSD (honestly significant difference) to distinguish differences among several means
- The advantages of using contrasts to test differences
- How to choose contrasts to identify specific treatment effects
- How to use contrasts to analyze trends for quantitative variables

## Comparing Means

There are many ways to compare treatment means calculated from an experiment. A great deal of controversy exists about which ones to use. In this lesson, we will present three approaches to evaluating mean responses and discuss under what circumstances each should be used:

- Multiple comparison procedures
- Contrasts: Planned T-tests or F-tests
- Trend analysis

### Multiple Comparison Procedures

Pairwise comparison procedures such as the Least Significant Difference (LSD) and Tukey's Honestly Significant Difference (HSD) are useful for making comparisons among levels of qualitative factors. These tests are appropriate for experiments such as cultivar and herbicide trials where you are interested in comparing a large number of treatments. In these experiments, you typically want to identify the superior treatments while having little or no prior knowledge with which to develop planned comparisons between specific means or groups of means.

The LSD, HSD, and other multiple comparison techniques, such as Duncan's Multiple Range Test (DMRT), are commonly used and misused mean comparison procedures in Agronomy. The HSD is more conservative than the LSD. The LSD is the easiest to use and provides valid results as long as you limit the number of comparisons made to a reasonable number. Some statisticians recommend using the LSD only to compare adjacent means or for making preplanned comparisons. An example of a reasonable preplanned comparison would be comparing individual cultivar means against a common control cultivar. Even in this case, there is a test called Dunnett's procedure, which is somewhat better than LSD. However, we will concentrate on LSD and HSD in this unit. We recommend only using the LSD following a significant F-test in the ANOVA. This is known as an F-protected LSD, a more conservative approach than just comparing pairs of means without a significant F. However, the use of the LSD test is really a matter of preference, and unprotected LSD comparisons are often made. The LSD and HSD should both only be used when the other two approaches to comparing means described in the next screen are not possible.

### Planned t-tests or F-tests

#### Contrasts

In many experiments, the treatment structure itself suggests certain planned comparisons. For example, consider a fertility trial

in which urea, ammonium nitrate, ammonium sulfate, calcium nitrate, and potassium nitrate are compared as sources of fertilizer nitrogen. The treatment structure suggests at least three meaningful comparisons:

1. urea vs. nitrate sources,
2. urea vs. ammonium sources, and
3. nitrate vs. ammonium sources.

These comparisons can be easily made by doing a t-test for contrasts of means. In an ANOVA, this can also be done by partitioning the sum of squares for treatments into individual single-degree-of-freedom contrasts that can be tested against the error mean square. The use of planned F-tests does not require a significant F-test for treatments and generally results in more sensitive tests than multiple comparison procedures.

## Trend Analysis

For quantitative data such as fertilizer and herbicide application rates, trend analysis is more appropriate than the other mean comparison procedures. With quantitative variables, it is possible to examine a functional relationship between the dependent variable and the treatment (independent) variable. By describing the relationship, it is not only possible to predict the treatment response for the treatment rates applied in the experiment but for every possible value between the lowest and highest rates applied.

There are several approaches to trend analysis. A common one is to use orthogonal polynomial coefficients to determine the highest order polynomial that describes the treatment response. This approach is useful for detecting whether the response is linear or has curvature. Another approach to trend analysis is curve fitting using regression techniques. These topics will be covered in greater detail in chapters 13 and 14 on Multiple and Nonlinear Regression, respectively.

# Least Significant Difference

## Stated Level of Significance

Work by Cochran and Cox (1957) indicated that experimenters who looked at the data after completing the experiment would tend to choose the highest and lowest treatments and compare them using the LSD. Because of this, the chance of making a Type I error increases dramatically depending on the number of treatments.

*Table 1 Probability of Type I error with use of LSD to compare the highest and lowest performing of 20 varieties in a variety trial.*

| # of Treatments | Alpha |
|---|---|
| 3 | 0.13 |
| 6 | 0.40 |
| 10 | 0.60 |
| 20 | 0.90 |

In other words, the probability of making a Type I error when you use LSD to compare the highest and lowest yielding of 20 varieties in a variety trial is 90%! This amounts to a fishing expedition for variety differences!

There are ways of testing a mean (not originally slated to be compared) that appears to be different after gathering the data. This can be done using methods found in Cochran and Cox (1957).

## Definition

**The Least Significant Difference (LSD)** test is an easy-to-use and valuable test for comparison. However, it must be used with caution. The LSD test should be used only to compare adjacent means in an array (where the means are arranged from highest to lowest value). In addition, comparisons should be meaningful and pre-planned. If used indiscriminately to locate any chance difference, the test is reduced to a fishing expedition — save the fishing for a real lake. In other words, the LSD and any other means comparison test should not be used to locate any significant differences which may exist, but rather, to answer the

questions that interest you!

In addition, as you make more and more comparisons with an LSD, the probability of making a Type I error becomes higher, and the alpha level for all comparisons is no longer the stated level of significance. Instead, the chance of falsely declaring a significant difference across all the comparisons made is multiplied. The result is an increased likelihood of falsely declaring a significant treatment effect somewhere in the whole experiment! Another way of stating this is that the more decisions you make, the more likely you are to make an error so it makes sense to limit them to include only the most important ones. As mentioned earlier, a good way to lessen the risk of this occurring is to use the F-protected test. This is accomplished by not using the LSD unless an F-test has already demonstrated that a significant treatment effect exists.

**Herbicide Treatments**

**Query 1.10.1**

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Formulas

The LSD test is derived from the t-test we studied earlier. Specifically, it uses the t-test for differences between means to determine the minimum difference necessary for those two means to be significantly different. The numerator in the original equation for the t-value is replaced by the LSD:

$$t = \frac{\bar{y_1} - \bar{y_2}}{SED} \tag{1.10.1}$$

$$\text{Equation 1} \tag{1.10.2}$$

**where:**
$\bar{y_1} - \bar{y_2}$ = difference in means for treatments 1 and 2 experimental unit,
$SED$=standard error of difference,
$t$=t value appropriate df and significance level.

Solving for LSD gives:

$$\text{LSD} = \text{t x SED} \tag{1.10.3}$$

$$\text{Equation 2} \tag{1.10.4}$$

**where:**
$SED$=standard error of difference,
$t$=t value appropriate df and significance level.

## CRD and RCRD

For two treatments in a Completely Randomized Design (CRD) or Randomized Complete Block Design (RCRD), the standard error of the difference (SED) is the square root of the sum of variances of each mean, or ($S^2/n_1 + S^2/n_2$). When the two means have the same number of observations, $n_1 = n_2 = r$ replications each, the $S_d^2 = 2S^2/r$. The estimate $S^2$ is the residual (error) mean square from the ANOVA table.

$$\text{LSD} = \text{t} \times \sqrt{\frac{2S^2}{r}} \tag{1.10.5}$$

$$\text{Equation 3} \tag{1.10.6}$$

**where:**
$S^2$=residual mean square,
$r$=number of replications,
$t$=t value appropriate df and significance level.

## LSD Example

We will use a similar dataset (ANOVA 2factor CRD [XLSX]) as in Chapter 9 on Two Factor ANOVAs, where three hybrids were tested at different plant densities to illustrate several methods of means comparisons, even though trend analysis or orthogonal comparisons are the most suitable methods for this experiment. We start by showing the LSD method for testing for differences in means. The completely randomized design experiment produced the following means (Table 2).

*Table 2 Yield data (t/ha) for three corn varieties planted at three populations.*

| Population (plats/m$^2$) | A | B | C |
|---|---|---|---|
| 7.5 | 9.34 | 9.27 | 8.42 |
| 10 | 8.43 | 9.86 | 9.43 |
| 12.5 | 10.48 | 7.72 | 5.52 |

The treatment means are viewed more easily as a list of yields (Table 3):

*Table 3 Yield data (t/ha) three corn varieties planted at three populations.*

| Population (plants/m$^2$) | Variety | mean |
|---|---|---|
| 7.5 | A | 9.34 |
| 7.5 | B | 9.27 |
| 7.5 | C | 8.42 |
| 10 | A | 8.43 |
| 10 | B | 9.86 |
| 10 | C | 9.43 |
| 12.5 | A | 10.48 |
| 12.5 | B | 7.72 |
| 12.5 | C | 5.52 |

### Study Question 2: LSD

What is the LSD which would be used for comparison? (Hint: The error mean square is 0.669, based on 18 df, and there are 3 reps per treatment.)

Yield data (t/ha) for three corn varieties planted at three populations.

| Population (plats/m$^2$) | A | B | C |
|---|---|---|---|
| 7.5 | 9.34 | 9.27 | 8.42 |
| 10 | 8.43 | 9.86 | 9.43 |

| Population (plats/m$^2$) | A | B | C |
|---|---|---|---|
| 12.5 | 10.48 | 7.72 | 5.52 |

Yield data (t/ha) three corn varieties planted at three populations

| Population (plants/m$^2$) | Variety | mean |
|---|---|---|
| 7.5 | A | 9.34 |
| 7.5 | B | 9.27 |
| 7.5 | C | 8.42 |
| 10 | A | 8.43 |
| 10 | B | 9.86 |
| 10 | C | 9.43 |
| 12.5 | A | 10.48 |
| 12.5 | B | 7.72 |
| 12.5 | C | 5.52 |

## Query 1.10.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Study Question 3: Adjacent Means

How many significant differences could you find between adjacent means (note that you will have to reorder the treatment means from Table 3)? (Hint: Be sure to first arrange the means in ranked order before comparing adjacent means with LSD.)

Yield data (t/ha) for three corn varieties planted at three populations.

| Population (plants/m$^2$) | A | B | C |
|---|---|---|---|
| 7.5 | 9.34 | 9.27 | 8.42 |
| 10 | 8.43 | 9.86 | 9.43 |
| 12.5 | 10.48 | 7.72 | 5.52 |

Yield data (t/ha) three corn varieties planted at three populations

| Population (plants/m$^2$) | Variety | mean |
|---|---|---|
| 7.5 | A | 9.34 |
| 7.5 | B | 9.27 |
| 7.5 | C | 8.42 |
| 10 | A | 8.43 |

| Population (plants/m$^2$) | Variety | mean |
|---|---|---|
| 10 | B | 9.86 |
| 10 | C | 9.43 |
| 12.5 | A | 10.48 |
| 12.5 | B | 7.72 |
| 12.5 | C | 5.52 |

Query 1.10.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Conclusions

It is obvious that there are individual mean comparisons that exceed the LSD. If for example, variety C had been planted as a control with which we intended to compare the other two varieties at each population, then we could conclude that variety A was greater than the control for the 12.5 population level.

*Table 2 Yield data (t/ha) for three corn varieties planted at three populations.*

| Population (plants/m$^2$) | A | B | C |
|---|---|---|---|
| 7.5 | 9.34 | 9.27 | 8.42 |
| 10 | 8.43 | 9.86 | 9.43 |
| 12.5 | 10.48 | 7.72 | 5.52 |

*Table 3 Yield data (t/ha) three corn varieties planted at three populations.*

| Population (plants/m$^2$) | Variety | mean |
|---|---|---|
| 7.5 | A | 9.34 |
| 7.5 | B | 9.27 |
| 7.5 | C | 8.42 |
| 10 | A | 8.43 |
| 10 | B | 9.86 |
| 10 | C | 9.43 |
| 12.5 | A | 10.48 |
| 12.5 | B | 7.72 |
| 12.5 | C | 5.52 |

## Calculations

The least significant difference (LSD) is probably the most often used mean comparison procedure for interpreting agronomic research. Because only one value is required, it is easy to calculate and easy to apply.

The LSD often is used for variety trials and other experiments where a large number of qualitative treatments are compared. It is typically included at the bottom of a column of means for which its use is intended.

For the purpose of this exercise, we will use results from another corn experiment. This was a field experiment in which three corn hybrids were fertilized with three different rates of N fertilizer. The experiment was replicated three times. The objective of the experiment was to determine the effects of hybrid and N fertilization on the yield of corn planted in narrow strips. Hybrids and N Rates were in factorial combination, so there are a total of nine (three hybrids × three N Rates) treatments. The analysis of variance and summary statistics for the experiment are presented in the Excel file QM-mod10-ex1.xls.

## Steps and Results

Calculate an LSD appropriate for comparing the nine treatment means presented in the Summary table of the ANOVA worksheet.

### Steps

1. Open the Excel file QM-mod10-ex1.xls.
2. Calculate the standard error of the difference for treatments using the formula:

$$SED = \sqrt{\frac{2RMS}{r}} \tag{1.10.7}$$

Equation 4 (1.10.8)

**where:**
$SED$=standard error of the difference,
$RMS$=residual (or error) mean square,
$r$= number of replications.

3. Activate cell B10 by clicking on it.
4. Enter the Excel formula: =SQRT((2*D6/3)) to calculate the SED.
5. Or use a calculator to compute the standard error of the difference (SED).
6. Calculate the LSD using the formula: LSD = t × SED.
7. Enter the Excel formula: =TINV(0.05,18)*B10 to calculate the LSD.
8. If you do not wish to use Excel, find the 0.05 two-tailed t-value for 18df and calculate the LSD.
9. Sort the means in the data Summary table by yield.
10. Select all data in the Summary table (A14:E23).
11. Select Sort from the Data menu above.
12. Sort by the Average field.

### Results

Use the LSD to compare adjacent means. Are there any significant differences? Despite the warnings in the text and lecture notes, the LSD is often used to compare pairs of means which are not adjacent to one another. How many pairwise comparisons are possible with nine treatments? If you were to use this (non-recommended) means separation method, are there any pair of means with a difference > than the LSD that would be considered different (at the .05 alpha level).

## Exercise 1: Calculating LSD and Tukey's HSD

**Notes to Educators and Students**

This activity will focus on calculating LSDs and HSDs in R and then interpreting them; it will not focus heavily on the mathematics involved in these calculations. Additional materials on calculating LSDs and HSDs can be found at the end of this activity.

**R Code Functions**

- setwd()
- aov()
- LSD.test()
- install.packages()
- summary()
- HSD.test()
- read.csv()
- library()

**The Scenario**

You are a graduate student studying the effects of planting density on the top 3 corn hybrids currently grown in western Iowa and Nebraska and you wish to assess whether any of the hybrids are significantly affected by the planting density. You plant each of the 3 hybrids at 7.5, 10.0, and 12.5 plants/$m^2$ giving you a total of 9 treatments for your experiment, and each treatment is replicated 3 times. At harvest, you calculate the yield in t/ha for each hybrid at each planting density. Ultimately you want to make a recommendation to farmers in western Iowa and Nebraska for each hybrid at a given density, and one way that can help you make that decision is to calculate the Least Significant Differences.

**Activity Objectives**

- Calculate the Least Significant Differences for the data set.
- Calculate the Honestly Significant Differences for the data set.
- Understand how the two calculations differ and when to use them.

**Ex. 1: What are LSDs and HSDs?**

They are both methods of making pairwise comparisons between different levels of a qualitative factor. LSDs are an easy-to-use method for making these comparisons, but a certain level of caution is advised because the more comparisons you make the greater the likelihood of making a Type I error. That is why you should only calculate LSDs if it is backed by a significant F-value. For instance, we know from our ANOVA that Hybrid has a significant effect on yield, but Population does not. Therefore, it would be better to only calculate LSDs comparing different hybrids because we already know this factor is already significant. Performing LSDs on Population, which is not significant, could result in a Type I error.

HSDs are another method and are more conservative in making pairwise comparisons by being less likely to result in a Type I error because the test statistic controls for the Type I error rate so that it stays at 0.05%. HSDs are good for multiple comparisons, whereas LSDs are only good for a few specific comparisons. Remember, if you are only comparing two treatments, then LSD=HSD. For a more detailed look at how these values are calculated, see the supplementary materials at the end of this activity.

**Ex. 1: Getting Ready**

First, set your working directory and read in the data set.

```
setwd("C:/Users/dadykema/Desktop/SAS to R")
corn<-read.csv("exercise.10.2.data.csv",header=T)
```

You will also have to install a new package called "**agricolae**" before you can calculate LSDs and HSDs.

```
install.packages("agricolae")
```

Before we can calculate LSDs and HSDs, we need to run an ANOVA in order to see if any of the variables are significant. If we were to calculate LSDs without an ANOVA first, we would have a greater chance of making a Type I error. Also, make sure that Population is considered to be a factor, or you will have incorrect degrees of freedom in your ANOVA.

```
corn<-read.csv("exercise.10.2.data.csv",header=T)
corn$Population<-as.factor(corn$Population)
Population<-corn$Population
```

## Ex. 1: ANOVA Output

Take a look at the ANOVA output. Are any of the factors significant here? In this example, we can see that Hybrid is significant, so we will calculate LSDs for this factor and not for Population.

```
cornaov<- aov(Yield ~ Population*Hybrid, data=corn)

Summary(cornaov)

Df Sum Sq Mean Sq F value Pr(>F)
Population 2 9.18 4.588 2.114 0.1498
Hybrid 2 12.18 6.342 2.922 0.0796
Population:Hybrid 4 29.30 7.325 3.375 0.0316
Residuals 18 39.07 2.170
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Next, load the ‘agricolae’ package.

```
library(”agricolae”)
```

Once you do this, you can calculate your LSDs and HSDs for Hybrid.

## Ex. 1: LSD Test

Use the LSD.test() function and be sure to include the model that we ran and then the variable we wish to analyze:

```
LSD.Hybrid<- LSD.test(cornaov, “Hybrid”)

LSD.Hybrid

$statistics
Mean CV MSerror LSD
8.718148 16.89873 2.170485 1.45909

$parameters
Df ntr t.value
18 3 2.100922

$means
Yields std r LCL UCL Min Max
A 9.418889 2.061822 9 8.387156 10.45062 5.39 11.79
B 8.947778 1.362028 9 7.916045 9.97951 6.09 11.20
C 7.787778 1.893949 9 6.756045 8.81951 4.65 10.30

$comparison
NULL

$groups
trt means M
1 A 9.418889 a
2 B 8.947778 ab
3 C 7.787778 b
```

## Ex. 1: HSD Test

It is the same process for the HSDs:

```
HSD.Hybrid<- HSD.test(cornaov, “Hybrid”)

HSD.Hybrid

$statistics
Mean CV MSerror HSD
8.718148 16.89873 2.170485 1.772477

$parameters
Df ntr StudentizedRange
18 3 3.609304

$means
Yields std r Min Max
A 9.418889 2.061822 9 5.39 11.79
B 8.947778 1.362028 9 6.09 11.20
C 7.787778 1.893949 9 4.65 10.30

$comparison
NULL

$groups
trt means M
1 A 9.418889 a
2 B 8.947778 a
3 C 7.787778 a
```

## Ex. 1: LSD Output

From this output, you can see that the calculated LSD (23.53) is smaller than the HSD (28.59) because the HSD is more conservative (blue arrows). When the treatment means are compared, we can see with the LSD method we see that Hybrids A and C are different, but when we look at the HSD results, they are not considered different due to the LSD not controlling for the Type I error (red arrows).

Now let us take a look at Population. We already know from the ANOVA that this factor was not significant, but let us see if we can confirm this with LSDs and HSDs.

```
LSD.Population<- LSD.test(cornaov, “Population”)
LSD.Population

$statistics
Mean CV MSerror LSD
140.6185 16.89662 564.527 23.53131

$parameters
Df ntr t.value
18 3 2.100922
```

$means

| | Yield | std | r | LCL | UCL | Min | Max |
|---|---|---|---|---|---|---|---|
| 30 | 145.3222 | 17.75287 | 9 | 128.6831 | 161.9614 | 118.2 | 177.4 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 40 | 149.0222 | 30.27506 | 9 | 132.3831 | 165.6614 | 86.9 | 190.2 |
| 50 | 127.5111 | 37.45285 | 9 | 110.8720 | 144.1503 | 75.0 | 172.6 |

| $comparison |
|---|
| NULL |

| $groups | | | |
|---|---|---|---|
| | trt | means | M |
| 1 | 40 | 149.0222 | a |
| 2 | 30 | 145.3222 | a |
| 3 | 50 | 127.5111 | a |

### Ex. 1: HSD Output

HSD.Population<- HSD.test(coranaov, “Population”)
HSD.Population

| $statistics | | | |
|---|---|---|---|
| Mean | CV | MSerror | HSD |
| 140.6185 | 16.89662 | 564.527 | 28.58542 |

| $parameters | | |
|---|---|---|
| DF | ntr | StudentizedRange |

| | | |
|---|---|---|
| 18 | 3 | 3.609304 |

| $means | | | | | |
|---|---|---|---|---|---|
| | Yield | std | r | Min | Max |
| 30 | 145.3222 | 17.75287 | 9 | 118.2 | 177.4 |
| 40 | 149.0222 | 30.27506 | 9 | 86.9 | 190.2 |
| 50 | 127.5111 | 37.45285 | 9 | 75.0 | 172.6 |

| $groups | | | |
|---|---|---|---|
| | trt | means | M |
| 1 | 40 | 149.0222 | a |
| 2 | 30 | 145.3222 | a |
| 3 | 50 | 127.5111 | a |

## Ex. 1: Review

We can see that the LSD and HSD calculations are the same as when we calculated for hybrid, but when we compare the means, we see that there are no differences between the treatments. This isn't surprising considering that we didn't find a significant value in the ANOVA for Population.

## Review Questions

- What have we learned from this lesson?
- How do LSDs and HSDs help make selection decisions?

*Table 4 R codes and the outputs from their execution*

| R Code Glossary | |
|---|---|
| setwd("") | Set the working directory, be sure to use your own file path. |

| R Code Glossary | |
|---|---|
| install.packages("") | Install a new R package on your computer. You only need to install a package once, unless there is an update of which R should notify you. |
| read.csv("") | Read in a .csv file. Remember to include if it has a header or not. |
| aov(y ~ A + B + A:B, data=mydataframe) | Perform a 2-factor analysis of variance on an R object. Can also write as aov(y~A*B, data = mydataframe). |
| summary() | Results the summary of an analysis. |
| library("") | Loads a package you have already downloaded. |
| LSD.test(anova output, "variable") | Calculates LSD for an ANOVA you have already run for a particular variable in your data set. |
| HSD.test(anova output, "variable") | Calculates HSD for an ANOVA you have already run for a particular variable in your data set. |

## Ex. 1: Supplement – Calculate LSD

**How to Calculate Least Significant Difference**

*Remember*: you want to know the difference between the means of the different treatments, but how do you decide if that difference is significant? Using Fisher's least significant differences lets you calculate the smallest difference between means needed in order to still be a statistically significant difference. This formula is based on the t-test which allows you to calculate the difference between two means.

The formula:

$$LSD = \sqrt{\frac{2MSE}{n^*}} \tag{1.10.9}$$

Equation 5 (1.10.10)

**where:**
**MSE** = this comes from the ANOVA test which you must run prior to calculating the LSDs
**n*** = number of scores used to calculate the mean.

## Ex. 1: Supplement – LSD Calculation Steps

**Calculate the LSD:**

**Step 1:** Run the ANOVA. From this you will get the mean square and the degrees of freedom. Important!!!!! If you don't have a significant F-statistic for your variable, this test will increase the likelihood of a Type I error!!!!!

**Step 2:** Find the critical t-value. You need to choose your alpha level (ex: 0.05) and use the appropriate degrees of freedom.

**Step 3:** Plug your information into the given formula and solve.

**Ordering the means:**

**Step 1:** After you calculate the LSD, you need to rank the means of the variable from lowest to highest.

*Table 5 Ranking of means in descending order.*

| Treatment | Mean Yield |
|---|---|
| 3 | 115 |

| Treatment | Mean Yield |
|---|---|
| 1 | 117 |
| 5 | 121 |
| 4 | 124 |
| 2 | 128 |
| 6 | 135 |

**Step 2:** Calculate the differences between each mean to see if the difference is greater than the Least Significant Difference that you have already calculated. Hint: start with the highest and lowest means. If they are not significantly different, then none of the means between them are significantly different and then test ends here. If they are significant, continue by comparing the highest and second lowest and the second highest with the lowest. Continue until there are no more significantly different comparisons.

**Step 3:** You can visualize the differences between the means by writing the treatments horizontally.

3 1 5 4 2 6

**Step 4:** Underline the treatments that are not significantly different and ignore the lines that fall completely within the boundaries of other lines.

3 1 5 4 2 6

_____
_____
_____
____

**Ex. 1: Supplement – LSD Results**

**Step 5:** Use lowercase letters to label each line (Figure 1).

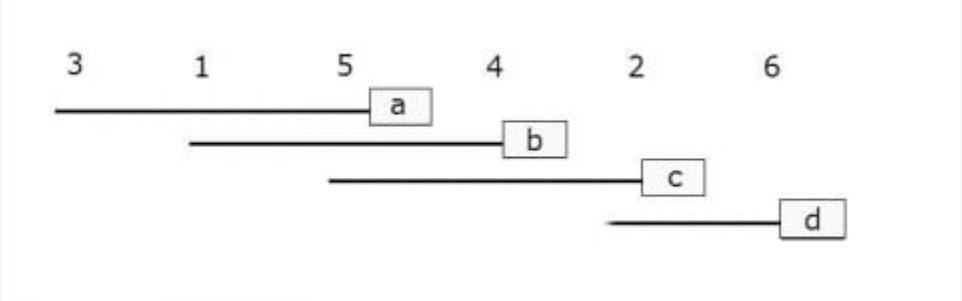



Figure 1 Illustrating how to easily show significant differences.

**Step 6:** You can do this in table format as well, and this is what you will see in the R output.

*Table 6 Use of letters to illustrate differences among treatment means.*

| Treatment | Mean Yield |
|---|---|
| 3 | 115 **a** |
| 1 | 117 **ab** |
| 5 | 121 **abc** |
| 4 | 124 **bc** |
| 2 | 128 **cd** |

| Treatment | Mean Yield |
|---|---|
| 6 | 135 **d** |

Basically, if any means share a common letter, they are not significantly different!

### Ex. 1: Supplement – HSD Test

**Tukey's Honestly Significant Difference**

This procedure is essentially the same as the LSD, but this test takes into consideration the number of treatment means and utilizes the studentized range statistic to control for the Type I error. This is because the test statistic already limits the Type I error to 0.05.

To find the studentized range statistic (q) you need to know two values:

- p: The number of treatments for your group
- f: the number of error degrees of freedom

You can look up the corresponding q value using these values in a table online (Tukey's test statistic). Once you have your q value, use the formula to calculate the HSD:

$$T_\alpha = q_\alpha(p, f)\sqrt{\frac{MSError}{r}} \tag{1.10.11}$$

Equation 6 (1.10.12)

Here, r is the number of replications.

Once you calculate the HSD, you can use the same procedure as the LSD to figure out which means are significantly different.

## Ex. 1: Supplement – LSD and HSD Resources

How to Calculate the Least Significant Difference (LSD).

- Statistics How To. Web. 25 July 2023.
- MEAN SEPARATION TESTS (Fisher's LSD AND Tukey's Procedure).

### Study Question 4: Differences Between Adjacent Means

#### Query 1.10.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Study Question 5

Query 1.10.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Multiple Range Tests

## Calculating Differences

The multiple range tests are extensions of the LSD. When there are several means, for example in the range between the highest and lowest, the LSD is multiplied by a factor depending on the number of means. For example, in the Duncan's MRT, the LSD might be multiplied by 1.14 if we have 9 means counting the lowest up to the highest, or 1.08 if there are 4 in the range tested (for example first to fourth or third to sixth). We only declare the highest and lowest different if their difference exceeds 1.14 times the LSD.

We have described the method for testing various ranges of means but will rely on the computer to actually carry out a multiple range test (MRT). We have studied the method to understand how computer programs assign the same letter to ranges that are not significantly different. We will not, however, be laboring through all the calculations done by a computer in a MRT.

The main point about output from a MRT is to realize that means which share a common letter are not significantly different. Suppose after a MRT, our mean yields in an experiment with 6 treatments are listed as:

*Table 7 Means separation using multiple range test.*

| Treatment | Mean Yield |
|---|---|
| 3 | 115 **a** |
| 1 | 117 **ab** |
| 5 | 121 **abc** |
| 4 | 124 **bc** |
| 2 | 128 **cd** |
| 6 | 135 **d** |

## Definition

**Multiple range tests protect better than LSD against Type I errors.** The LSD discussed in the previous section works well for comparing selected means, usually adjacent means, in an experiment. The method is not useful for numerous comparisons or for comparing all the means. In an experiment with nine treatments there are 36 [9(9 – 1)/2] pairwise comparisons that can be made. If you were to use an LSD at the P = 0.05, then you might expect to make up to two Type I errors (0.05 x 36 = 1.8) among all the comparisons assuming there are no real differences among the treatments. The more comparisons you make, the more likely you are to make a mistake using the LSD, so we naturally try to limit our comparisons to those that are most important to us.

What if we want to make comparisons between all means? For example, what if we are comparing the grain yields of multiple corn hybrids? In this case, we can employ a Multiple Range Test. Such a test conservatively adjusts the required difference for significance to adjust for the distance between means in an array. The Honestly Significant Difference (HSD) is one of the tests

for this purpose.

## HSD

The **HSD test**, like the LSD test, determines the minimum significant difference (MSD) between means arrayed by magnitude (value). This difference is MSD = Q $(S^2/r)^{1/2}$, which is essentially the same formula as that for the LSD, except that Q contains a modified t-value times √2. For an experiment comparing only two means, LSD = HSD. When there are more than two means to be compared, the HSD controls the "experimentwise" Type I error rate to be 0.05. In other words, the HSD ensures that even with, say 15 treatment means, you will only falsely declare a difference to be significant in 5% of such experiments. However, with the HSD, individual comparisons are made at a P somewhat less than the stated probability level, so it is much more conservative than the LSD, which controls "comparisonwise" Type I error rate.

Appendix Table 6 has a column for error df on the left and separate columns for numbers of means compared (2 up to 20). The studentized range (Q) is then read from the table. For an experiment with the broadest comparison of six means with 24 df for the error mean square, Q = 4.373.

## How to do HSD

The HSD multiple range test should be conducted in the following manner.

1. Arrange means from highest to lowest.
2. Compare the highest and lowest mean values. If the difference between the two values is not significant, draw a line beside the list connecting the two means or put the same letter next to each in the range. Conclude that there are no treatment differences.
3. If the difference is significant, then the test continues. Compare the highest mean to the second-lowest mean and the lowest mean to the second-highest mean. If either difference is not significant, draw a line between the two means or put the same letter next to each in the range.
4. When the difference between two means is declared not significant, then all means between the two compared are also declared not significant.
5. Continue until all means have been compared directly or shown to be not significantly different within another comparison.

### Study Question 6

#### Query 1.10.6

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Study Question 7

Query 1.10.7

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Contrasts

**Contrasts are comparisons among several means.** To this point, we have only been comparing pairs of treatment means, but often we need to compare several means. For example, suppose we have 3 nitrogen fertilizer treatments in a wheat experiment: manure, 25 kg N, and 50 kg N, the latter two being applied as urea. We would be interested in a comparison or contrast of average yield of the manure treatment vs. the two chemical fertilizer treatments. We want to test $H_0$: (mean of treatment 1) = (mean of treatments 2 and 3). Just as with a contrast of two treatment means, we can use a t-test for testing whether a linear combination of the three treatment means is zero.

We can analyze contrasts with either a t-test (the next slide) or an F-test (described later in this lesson). With either test, we are comparing the differences between means or groups of means to the residual variation described by the standard error (in the F-test).

## Test Equations

We test the linear combination:

$$L = \bar{Y} - \frac{\bar{Y}_1 + \bar{Y}_2}{2} \tag{1.10.13}$$

Equation 7 (1.10.14)

The t-test has form:

$$t = \frac{L}{S_L} \tag{1.10.15}$$

Equation 8 (1.10.16)

**where:**
**L** = the difference between the means of groups or groups of means,
**$S_L$** = the standard error of the contrast.

## Estimating Variance

The testing of contrasts of several means is somewhat more complicated than testing pairs of means because we need to estimate the variance of the linear combination, $S_L^2$. For our example,

$$S_L^2 = (1)^2(\frac{S^2}{n_1}) + (-0.5)^2(\frac{S^2}{n_2}) + (-0.5)^2(\frac{S^2}{n_3}) \tag{1.10.17}$$

Equation 9 (1.10.18)

**where:**

$S^2$= residual or error mean square,
$n_1, n_2, n_3$= numbers of observations in treatments 1, 2, and 3, respectively.

Where did the "1" and "-0.5"s come from? By multiplying the corresponding means by these numbers, we calculate the same value as we did for $L = \bar{Y}_1 - (\bar{Y}_2 + \bar{Y}_3)/2$. In effect, we are finding the difference between $\bar{Y}_1$ and the mean of $\bar{Y}_2$ and $\bar{Y}_3$.

# Linear Combination and Variance of Linear Contrast

## Linear Combination

A linear combination of the treatment means is:

$$L = c_1\bar{Y}_1 + c_2\bar{Y}_2 + ...c_t\bar{Y}_t \text{ or } \mathrm{L} = \sum_{i=1}^{t} c_i Y_i \tag{1.10.19}$$

$$\text{Equation 10} \tag{1.10.20}$$

**where:**
$c_t$= contrast coefficient for treatment t,
$\bar{Y}_t$= mean of treatment t.

### Variance of the Linear Contrast

The variance of the linear contrast, each of whose treatments has r replications, is:
is

$$S_L^2 = \sum_{i=1}^{t} c_i^2 \tag{1.10.21}$$

$$\text{Equation 11} \tag{1.10.22}$$

**where:**
$S^2$= residual or error mean square,
$c_2$= contrast coefficient for ith treatment mean
$r$= number replications.

This formula for the estimated variance of a contrast makes sense because we are comparing t independent means. They are independent because of randomization in the experimental design. Each mean has an estimated variance ($S^2/r$). Because the variance of a constant times a variable is the square of the constant times the variance of the variable, we have $c_i^2$ in the formula. In the simplest case, this reduces to: $S_d^2 = (S^2/r) + (S^2/r)$ for contrasting two treatment means [because $c_i^2 = (1)^2$ or $(-1)^2$]. The square root of this is the standard error for computing the LSD.

## Study Question 8

### Query 1.10.8

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Study Question 9

### Query 1.10.9

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Study Question 10

If in our wheat experiment, yields of the 3 treatments, each with 6 reps, are \bar{Y} = 70 bu/A, \bar{Y}_2 = 57 bu/A and \bar{Y}_3 = 64 bu/A, and the error mean square from the ANOVA is 78, what is the calculated t-value for the contrast?

### Query 1.10.10

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Testing

The idea of contrasts is a straightforward extension of the idea of comparing two treatment means. In general, we make the contrast (L) among the treatment means and test the null hypothesis that the mean of the contrast is zero. If $t = L/S_L$ is larger than the critical t-value (for degrees of freedom in $S^2$), we reject the null hypothesis. With 10 or more error df, the contrast is significant at the 0.05 probability level when L is over about 2 standard errors ($S_L$) away from zero.

One more point is pertinent: we can also test the contrast with an F-test. Earlier we saw that a two-tailed t-test with k error degrees of freedom is the same as an F-test with one numerator and k denominator degrees of freedom because $F = t^2$. F-tests for contrasts are often arranged into an ANOVA table so we can subdivide the treatment sum of squares into logical single-degree-of-freedom tests. We explore this further in the next section.

# Planned F-Tests

**Contrasts are planned comparisons and can be tested with F-tests.** Perhaps the most powerful and useful method for comparing means is through contrasts, outlined in the previous section. They go beyond simple mean comparisons to answering specific questions about the treatment effect. They are especially useful in factorial experiments, where they allow the effect of one factor to be isolated and studied. In addition, contrasts can be used with quantitative variables (i.e., levels of fertilizer) to detect trends in the response of the experimental units. We outline in detail in this section how contrasts are set up, some of their properties, how they fit into an ANOVA table, and how they are tested with F-tests.

The first step in designing a contrast is to determine the questions that we desire to answer. This is the beauty of a contrast — it allows us to cut through all of the numbers and get back to the concepts for which we conducted the experiment!

For example, what questions would we want to answer with regard to the corn hybrid and population factorial experiment? Of course, we want to know whether higher plant populations improve corn yield. We also want to know whether the hybrids produce different yields.

## Corn Example

Now let us further define these questions. We begin by specifying the population question. We ask two questions. First, how does the yield at 12.5 plants/$m^2$ compare to the mean yield of 7.5 and 10.0 plants/$m^2$? Secondly, does the yield for 7.5 plants/$m^2$ differ from the yield for 10.0 plants/$m^2$?

Among the corn hybrids we could compare the mean yield of hybrid C with the mean yield of hybrids A and B. We could then compare the yield of hybrid A with that of hybrid B.

These are examples of some of the logical contrasts which a researcher might want to test in his/her experiment. The contrasts chosen by a researcher will depend on the objectives for each experiment.

**Population:**

- 5.0 vs. 10.0 plants/$m^2$
- 7.5 and 10.0 vs. 12.5 plants/$m^2$

**Hybrid:**

- hybrid A vs. hybrid B
- hybrid A and B vs. hybrid C

## Assigning Contrast Coefficients (1)

**Contrast Coefficients are assigned +1 or -1 to compare equal-sized groups.** The nuts and bolts behind a contrast is the generation of a difference. For example, in comparing the populations previously, what we are really doing is examining the numerical difference between the mean yield at 7.5 and 10.0 plants/$m^2$. Therefore, the second step in a contrast is to generate this difference. This is accomplished by assigning different weights, called coefficients, to the nine treatment means produced by the experiment.

To determine the difference in mean yield between corn planted at 7.5 and 10.0 plants/$m^2$, we need to compare two groups: the three treatment means produced at 7.5 plants/$m^2$ and the three treatment means produced at 10.0 plants/$m^2$. We will assign a coefficient of +1 to every treatment that is produced at 7.5 plants/$m^2$ and a coefficient of -1 to every treatment produced at 10.0 plants/$m^2$. We assign a coefficient of 0 to every treatment produced at 12.5 plants/$m^2$, since we are not comparing that population in this contrast. As a reminder, in our numbering of treatments, the first three are 7.5 plants/$m^2$ for hybrids A, B and C, treatments 4 to 6 are 10.0 and 7 to 9 are 12.5. The coefficients for our contrast of the nine treatments are:

*Table 8 Coefficients for population contrasts.*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/m$^2$ | +1 | +1 | +1 | -1 | -1 | -1 | 0 | 0 | 0 |

## Assigning Contrast Coefficients (2)

What we are doing, then, is adding up all means produced at 7.5 plants/m$^2$ and then subtracting all means produced at 10 plants/m$^2$. The result is a difference which we will analyze. You might wonder why we use whole number coefficients instead of 1/3 and -1/3 since the contrast of interest is $(\frac{\bar{Y}_1+\bar{Y}_2+\bar{Y}_3}{3})-(\frac{\bar{Y}_4+\bar{Y}_5+\bar{Y}_6}{3})$. It turns out that the F-ratio or t-test is the same whether we use the fractions or whole numbers, and it is easier to just use whole numbers. However, when calculating the actual difference being compared, it is important to use the correct fraction. In the example above, leaving out the denominator would give you the difference in totals rather than the difference in means. We can use totals to evaluate the significance of the contrast, but we need to use actual means when we estimate a treatment difference.

To determine the difference in mean yield between hybrid A and hybrid B, we follow the same procedure. We assign a coefficient of +1 to every treatment which includes hybrid A, and a coefficient of -1 to every treatment which includes hybrid B. We will assign a coefficient of 0 to every treatment which includes hybrid C, since that hybrid is not a part of this comparison. We assign the following coefficients to our nine treatments:

*Table 9 Coefficients for variety contrasts.*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Hybrid A vs B | +1 | -1 | 0 | +1 | -1 | 0 | +1 | -1 | 0 |

Notice that it is imperative to pay close attention to how treatments are assigned.

## Assigning Contrast Coefficients – Sums

**Contrast coefficients are assigned weights which sum to zero.** Comparing the yield produced at a population of 12.5 plants/m$^2$ with the mean yield of 7.5 and 10.0 plants/m$^2$ is a little trickier. Again, we are comparing two groups. This time, however, one of the groups is composed of two populations — 7.5 and 10.0 plants/m$^2$ — while the other group is composed of only one population — 12.5 plants/m$^2$. This changes the coefficients which must be assigned.

At first, it might appear that we should assign a coefficient of +1 to all treatments containing 7.5 or 10.0 plants/m$^2$ and a coefficient of -1 to all treatments containing 12.5 plants/m$^2$. This, however, would be wrong, for we would be comparing the sum of six treatment means (7.5 and 10.0 plants/m$^2$) with the sum of only three treatment means (12.5 plants/m$^2$).

## Assigning Contrast Coefficients – Weighting

Instead, we assign a coefficient of +1 to all treatments containing 7.5 or 10.0 plants/m$^2$ and a coefficient of -2 to all treatments containing 12.5 plants/m$^2$. In doing this, we in effect weight the yields produced with 7.5 and 10.0 plants/m$^2$ before comparing them with the yields produced at 12.5 plants/m$^2$. We assign the following coefficients to our nine treatments:

*Table 10 Coefficients for population contrasts.*

| Contrasts | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/m$^2$ | +1 | +1 | +1 | -1 | -1 | -1 | 0 | 0 | 0 |
| 12.5 vs. 7.5 | +1 | +1 | +1 | +1 | +1 | +1 | -2 | -2 | -2 |

Notice that our contrast coefficients are balanced in that they sum to zero. This is a characteristic of contrast coefficients when treatments are equally replicated. Also notice that they are directly proportional to the fractional coefficients (+1/6, +1/6, +1/6, +1/

6, +1/6, +1/6, -1/3, -1/3, -1/3) which result from the contrast of the (mean of 7.5 or 10.0 plants/$m^2$) vs. (mean of 12.5 plants/$m^2$).

### Assigning Contrast Coefficients – Comparison

A good rule of thumb to remember when comparing groups containing different numbers of treatment means is this: assign each member of the first group a coefficient equal to the number of treatments in the second group. Then assign each member of the second group the negative value of the number of treatments in the first group.

We treat the comparison of the yield of hybrid C and the mean yield of hybrids A and B in the same way. We assign a coefficient of +1 to all treatments containing hybrids A or B and a coefficient of -2 to all treatments containing hybrid C.

*Table 11 Coefficients for variety contrasts*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Hybrid A vs. B | +1 | -1 | 0 | +1 | -1 | 0 | +1 | -1 | 0 |
| Hybrids C vs. A & B | +1 | +1 | -2 | +1 | +1 | -2 | +1 | +1 | -2 |

## Independence of Comparisons

**Orthogonal comparisons have the property of independence.**

Two rules must be followed in order for a set of contrasts to be independent of each other. These are for treatments with equal numbers of reps, which is the usual case when there is no missing data in an experiment.

### Rule 1

The sum of the coefficients in each contrast must equal zero.

*Table 12 Coefficients for variety contrasts.*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Hybrid A vs. B | +1 | -1 | 0 | +1 | -1 | 0 | +1 | -1 | 0 |

### Rule 2

The sum of the product of the corresponding coefficients of any two contrasts must equal zero.

*Table 13 Coefficients for variety contrasts*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Hybrid A vs. B | +1 | -1 | 0 | +1 | -1 | 0 | +1 | -1 | 0 |
| Hybrids C vs. A & B | +1 | +1 | -2 | +1 | +1 | -2 | +1 | +1 | -2 |
| Product of contrasts | +1 | -1 | 0 | +1 | -1 | 0 | +1 | -1 | 0 |

## Non-orthogonal Contrasts

It is possible and sometimes even desirable to make a set of contrasts that are not orthogonal. The advantage of using an orthogonal set is that if the sums of squares for each contrast are added together, the sum is equal to the sum of squares for treatments in the ANOVA. This characteristic can be useful when performing calculations. It also provides a way to efficiently use all the information available for treatment comparisons. Interpretation of a set of contrasts is also more straight forward when they are independent.

Study Question 11

For 4 treatments, are the following pairs orthogonal contrasts?

Query 1.10.11

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Contrast Sums of Squares

**Each contrast has a sum of squares in the ANOVA.**

Each contrast produces a sum of squares which is calculated using treatment means and coefficients. The F-ratio is the square of the calculated t-value $L/S_L$. The formula for the contrast sum of squares is

$$\text{Contrast SS} = \frac{r(\sum c\bar{Y})^2}{\sum c^2} \tag{1.10.23}$$

Equation 12 (1.10.24)

**where:**
$c$= contrast coefficient,
$r$= number replicates.

The contrast SS has 1 df, so the Contrast MS = Contrast SS. Basically, $t^2$ = F (the variance ratio). So, the square of the t-value you calculated earlier is equal to the Contrast MS/ Residual Mean Square (RMS).

## Calculating Contrast SS

The sum of squares for comparison of yields produced at 7.5 and 10.0 plants/$m^2$ is calculated as follows. Of course, the computer calculates this for you, but it is good to see this calculation to know how much work it saves you!

.

## Corn Population Example

The results can be arranged in the same manner as an ANOVA table. The sums of squares for all four contrasts are therefore:

*Table 14 ANOVA for contrasts.*

| Contrast | df | SS | MS | F | F crit (P = 0.05) |
|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/$m^2$ | n/a | 62.72 | n/a | n/a | n/a |
| 7.5 and 10.0 vs. 12.5 plants/$m^2$ | n/a | 107.925 | n/a | n/a | n/a |
| hybrid A vs. B | n/a | 5.021 | n/a | n/a | n/a |
| hybrid A and B vs. C | n/a | 1.564 | n/a | n/a | n/a |
| error | n/a | 563.611 | n/a | n/a | n/a |

The sum of squares for error is calculated as in the ANOVA with all treatment sums of squares, not just those for the 4 contrasts, removed.

## Mean Square

**Contrasts have 1 df, so contrast MS = contrast SS.**

The mean square for each contrast and the error is calculated by dividing the sum of squares by the degrees of freedom. Each contrast has one degree of freedom. Therefore, the mean square for each contrast is equal to its sum of squares:

*Table 15 ANOVA for contrasts.*

| Contrast | df | SS | MS | F | F crit (P = 0.05) |
|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/$m^2$ | 1 | 4.726 | 4.726 | n/a | n/a |
| 7.5 and 10.0 vs. 12.5 plants/$m^2$ | 1 | 107.925 | 107.925 | n/a | n/a |
| hybrid A vs. B | 1 | 5.021 | 5.021 | n/a | n/a |
| hybrid A and B vs. C | 1 | 1.564 | 1.564 | n/a | n/a |
| error | 18 | 563.611 | 31.312 | n/a | n/a |

Compare this result with the ANOVA table calculated in R Exercise 1 in Chapter 9 on Two Factor ANOVAs. Notice that the contrast df and SS for population and for hybrid sum to those found in the ANOVA table (with some small rounding errors). There is one portion of the ANOVA from Chapter 9 missing, though, the interaction between the treatments. Orthogonal contrasts could also be constructed for the interaction, but we will not do that here.

## F-Tests

To calculate the F-value for each contrast, the mean square for each contrast is divided by the mean square error:

*Table 16 ANOVA for contrasts.*

| Contrast | df | SS | MS | F | F crit (P = 0.05) |
|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/$m^2$ | 1 | 4.726 | 4.726 | 0.15 | n/a |
| 7.5 and 10.0 vs. 12.5 plants/$m^2$ | 1 | 107.925 | 107.925 | 3.45 | n/a |
| hybrid A vs. B | 1 | 5.021 | 5.021 | 0.16 | n/a |
| hybrid A and B vs. C | 1 | 1.564 | 1.564 | 0.05 | n/a |
| error | 18 | 563.611 | 31.312 | n/a | n/a |

## F-Test Critical Value

The contrast for each F-value is then compared with the critical F-value for the desired level of significance (F with 1 numerator and 18 denominator df).

*Table 17 ANOVA for contrasts.*

| Contrast | df | SS | MS | F | F crit (P = 0.05) |
|---|---|---|---|---|---|
| 7.5 vs. 10.0 plants/$m^2$ | 1 | 4.726 | 4.726 | 0.15 | 0.70 |
| 7.5 and 10.0 vs. 12.5 plants/$m^2$ | 1 | 107.925 | 107.925 | 3.45 | 0.08 |
| hybrid A vs. B | 1 | 5.021 | 5.021 | 0.16 | 0.69 |
| hybrid A and B vs. C | 1 | 1.564 | 1.564 | 0.05 | 0.83 |

| | | | | | |
|---|---|---|---|---|---|
| error | 18 | 563.611 | 31.312 | n/a | n/a |

## Study Question 12

### Query 1.10.12

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Exercise 2: Calculating Contrasts

**Notes To Instructors**

This lesson will focus on the calculating of contrasts using the R software, it is assumed that we already how to calculate contrast coefficients in order to make specific comparisons. A brief overview of how to assign contrast coefficients can be found in the supplementary materials at the end of the activity.

**Ex. 2: Getting Ready**

**R Code Functions**

- read.csv()
- as.factor()
- list()
- aov()
- matrix()
- split()
- summary()
- contrasts()
- interaction()

**The Premise**

The previous LSD and HSD test indicate that the means of Population are not significantly different. However, the LSD test shows the average yield of Hybrid A and B is different than that of Hybrid C. If we want to further to explore specific means or group of means comparisons for variables Population and Hybrid, contrast is the best way to do this. So, continuing from the same experiment of planting 3 hybrids at 3 planting densities, you now wish to make specific comparisons between hybrids and comparisons between populations.

**Activity Objectives**

Use R to calculate specific contrasts that you choose to run.

**Ex. 2: Read Data**

Suppose we want to create two contrasts for main effect of Population, we named:

- C1: compare the yields of two populations 7.5 and 10 plants/$m^2$.
- C2: compare the yield produced at a population of 12 plants/$m^2$ with the mean yield of populations 7.5 and 10 plants/m2.

If you have picked up from the previous activity, you will not need to run the ANOVA again, but if you are starting fresh with this assignment, use the following code to read in the data set and run the two-factor ANOVA. Be sure to have Population as a factor in R before you run your analysis or you will not have the appropriate number of degrees of freedom for Population.

corn<-read.csv("exercise.10.2.data.csv",header=T)

**Ex. 2: Contrast Coefficients**

corn$Population <- as. factor(corn$Population)
Population <- corn$Population
cornaov<- aov(Yield ~ Population*Hybrid, data=corn)
Summary(cornaov)

Df Sum Sq Mean Sq F value Pr(>F)
Population 2 9.18 4.588 2.114 0.1498

Hybrid 2 12.18 6.342 2.922 0.0796

Population:Hybrid 4 29.30 7.325 3.375 0.0316

Residuals 18 39.07 2.170

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Next, make a matrix with the two sets of contrast coefficients. Then set the "contrasts" attribute of factor Population. For more information on how to calculate the contrast coefficients, see the supplementary materials for this activity. The following code will spell out the desired contrast and calculate it for us.

contrasts(corn$Population) <- matrix(c(1,-1,0,-1,-1,2), nrow = 3)

contrasts(corn$Population)

[,1] [,2]
30 1 -1
40 -1 -1
50 0 2

## Ex. 2: Contrast Coefficients – Output

In the output, we have generated a matrix of contrast coefficients comparing 7.5 plants /$m^2$ to 10 plants/$m^2$ in column 1 and the mean of 7.5 and 10 compared to 12.5 plants/$m^2$ in column 2. Remember, always double check that your contrasts (columns here) add up to 0! Next, we run the ANOVA again and this time we use the 'split' and 'list' function to add the contrasts we are interested in to the ANOVA.

Pop.model <- aov(Yield ~ Population*Hybrid, data = corn)

summary.aov(Pop.model, split = list(Population = list("7.5 vs 10" = 1, "12.5 vs 7.5+10" =2)))

Df Sum Sq Mean Sq F value Pr(>F)
Population 2 9.18 4.588 2.114 0.14975
Population: 7.5 vs 10 1 0.24 0.238 0.110 0.74434
Population: 12.5 vs 7.5+10 1 8.94 8.939 4.118 0.05748 .
Hybrid 2 12.68 6.342 2.922 0.07962 .
Population:Hybrid 4 29.30 7.325 3.375 0.03156 *
Population:Hybrid: 7.5 vs 10 2 3.04 1.521 0.701 0.50931
Population:Hybrid: 12.5 vs 7.5+10 2 26.26 13.130 6.049 0.00978 **

Residuals 18 39.07 2.170
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The results show that the both contrasts “7.5 vs 10” and “12.5 vs 7.5 + 10” on the main effect of Population are not significant with P = 0.7444 and P = 0.057, respectively. And also that the second contrast on Population has a significant interaction with Hybrid, F(1,18) = 6.409, P = 0.01.

## Ex. 2: Contrasts for Hybrid Effect

Let’s switch to construct contrasts for main effect of Hybrid. Similarly, we named,

- C3: compare the mean yield between hybrid A and hybrid B,
- C4: compare the yield of hybrid C and the mean yield of hybrids A and B.

The R code for calculating above two contrasts would be the same as the ones of Population.

contrasts(corn$Hybrid) <- matrix(c(1,-1,0,-1,-1,2), nrow = 3)

contrasts(corn$Hybrid)

[,1] [,2]
A 1 -1
B -1 -1
C 0 2

Hybrid.model <- aov(Yield ~ Population*Hybrid, data = corn)

summary.aov(Hybrid.model, split = list(Hybrid = list(”A vs B” = 1, “C vs A+B” = 2)))

Df Sum Sq Mean Sq F value Pr(>F)
Population 2 9.18 4.588 2.114 0.1498
Hybrid 2 12.68 6.342 2.922 0.0796 .
Hybrid: A vs B 1 1.00 0.999 0.460 0.5062
Hybrid: C vs A+B 1 11.69 11.685 5.384 0.0323 *
Population:Hybrid 4 29.30 7.325 3.375 0.0316 *
Population:Hybrid: A vs B 2 13.53 6.765 3.117 0.0688 .
Population:Hybrid: C vs A+B 2 15.77 7.886 3.633 0.0473 *
Redsiduals 18 39.07 2.170
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

## Ex. 2: Interaction

We can see that the yield of Hybrid A is not different from that of Hybrid B with P = 0.5062. However, the average yield of Hybrid A and B is different than that of Hybrid C with P = 0.0323. Now we know that if say, Hybrid C was a standard check variety, it would be different than the other two varieties in your trial, but Hybrid A and Hybrid B are not really different in terms of yield. You’ll want to take information like this into consideration if you are trying to decide which Hybrids to keep in your breeding program.

Let’s say we think that a particular hybrid is affected by Population because we saw evidence of it in the ANOVA or the interaction plot. In this example, we can see that Hybrid C is affected by Population. To figure out which Populations differ in yield when growing Hybrid C, we need to make two contrasts: 7.5 vs. 10 and 10 vs. 12.5. To do this in R, we first have to compute a factor which represents the interaction of Population and Hybrid.

We named the new interaction factor as “P.H”, which has 9 levels, ordered from 7.5.A to 12.5.C.

corn$P.H <- interaction(corn$Population, corn$Hybrid)

```
corn$P.H

[1] 7.5.A 7.5.A 7.5.A 7.5.B 7.5.B 7.5.B 7.5.C 7.5.C 7.5.C 10.A 10.A 10.A 10.B 10.B 10.B 10.C
[17] 10.C 10.C 12.5.A 12.5.A 12.5.A 12.5.B 12.5.B 12.5.B 12.5.C 12.5.C 12.5.C
Levels: 7.5.A 10.A 12.5.A 7.5.B 10.B 12.5.B 7.5.C 10.C 12.5.C
```

## Ex. 2: Compare Yield – 7.5 vs 10

Next we set the contrast coefficients to compare the yield difference between 7.5 and 10 for Hybrid C.

```
contrasts(corn$P.H) <- c(0,0,0,0,0,0,1,-1,0)

contrasts(corn$P.H)

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
7.5.A 0 -0.33333333 -0.33333333 -0.33333333 -0.33333333 -0.09763107 -0.5690356 -0.33333333
10.A 0 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.73151455 -0.5648479 -0.08333333
12.5.A 0 0.91666667 -0.08333333 -0.08333333 -0.08333333 -0.02440777 -0.1422589 -0.08333333
7.5.B 0 -0.08333333 0.91666667 -0.08333333 -0.08333333 -0.02440777 -0.1422589 -0.08333333
10.B 0 -0.08333333 -0.08333333 0.91666667 -0.08333333 -0.02440777 -0.1422589 -0.08333333
12.5.B 0 -0.08333333 -0.08333333 -0.08333333 0.91666667 -0.02440777 -0.1422589 -0.08333333
7.5.C 0 -0.08333333 -0.08333333 -0.08333333 -0.08333333 0.47559223 0.3577411 -0.08333333
10.C 1 -0.08333333 -0.08333333 -0.08333333 -0.08333333 0.47559223 0.3577411 -0.08333333
12.5.C -1 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.02440777 -0.1422589 0.91666667
```

The contrasts get stored as attributes of the factor P.H. So when we run a new ANOVA they will get applied automatically. The contrast matrix has eight sets of contrasts. We are only interested in the first one and ignore the rest of them. Therefore, the argument of list is “label of contrast = 1”.

```
Interaction.model <- aov(Yield ~ P.H, data = corn)

summary.aov(Interaction.model, split = list(P.H = list("7.5 vs 10, Hybrid C" = 1)))

DF Sum Sq Mean Sq F value Pr(>F)
P.H 8 51.16 6.395 2.946 0.0271
P.H: 7.5 vs 10, Hybrid C 1 1.53 1.530 0.705 0.4121
Residuals 18 39.07 2.170
—
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

## Ex. 2: Compare Yield – 10 vs 12.5

The same R code can be applied for calculating the contrast 10 vs. 12.5 except changing the contrast coefficients of factor P.H and the label in *list*.

```
contrasts(corn$P.H) <- c(0,0,0,0,0,0,1,-1,0)

contrasts(corn$P.H)

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
7.5.A 0 -0.33333333 -0.33333333 -0.33333333 -0.33333333 -0.33333333 -0.09763107 -0.5690356
10.A 0 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.73151455 -0.5648479
12.5.A 0 0.91666667 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.02440777 -0.1422589
7.5.B 0 -0.08333333 0.91666667 -0.08333333 -0.08333333 -0.08333333 -0.02440777 -0.1422589
10.B 0 -0.08333333 -0.08333333 0.91666667 -0.08333333 -0.08333333 -0.02440777 -0.1422589
12.5.B 0 -0.08333333 -0.08333333 -0.08333333 0.91666667 -0.08333333 -0.02440777 -0.1422589
7.5.C 0 -0.08333333 -0.08333333 -0.08333333 -0.08333333 0.91666667 -0.02440777 -0.1422589
10.C 1 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.08333333 0.47559223 0.3577411
```

12.5.C -1 -0.08333333 -0.08333333 -0.08333333 -0.08333333 -0.08333333 0.47559223 0.3577411

Interaction.model <- aov(Yield ~ P.H, data = corn)

summary.aov(Interaction.model, split = list(P.H = list("7.5 vs 10, Hybrid C" = 1)))

DF Sum Sq Mean Sq F value Pr(>F)
P.H 8 51.16 6.395 2.946 0.0271
P.H: 7.5 vs 10, Hybrid C 1 1.53 1.530 0.705 0.4121
Residuals 18 39.07 2.170
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results of the two contrasts show that the yields of Hybrid C are the same for the 7.5 and 10 populations with F(1, 18) = 0.705, P = 0.4121 and differ from the yield at the 12.5 population with F(1, 18) = 10.602, P = 0.0044. Therefore it is reasonable to assume that Hybrid C is impacted by a higher planting density than the other two Hybrids.

## Ex. 2: Review

### Review Questions

- What have we learned from this lesson?
- How do LSDs and HSDs help make selection decisions?

*Table 18 R codes and the outputs from executing them.*

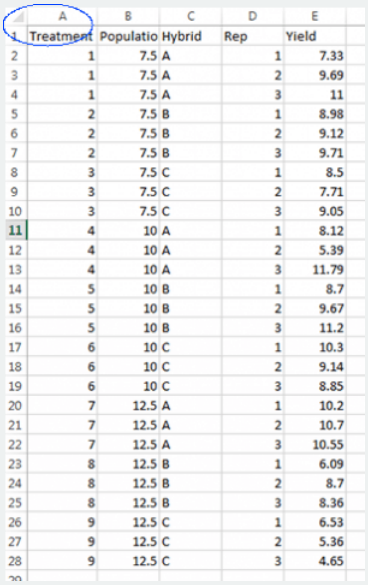
| R Code Glossary | |
|---|---|
| **read.csv("")** | Read in a .csv file. Remember to include if it has a header or not. |
| **aov(y ~ A + B + A:B, data=mydataframe)** | Perform a 2-factor analysis of variance on an R object. Can also write as aov (y~A*B, data = mydataframe). |
| **summary()** | Returns the summary of an analysis |
| **as.factor(mydataframe$variable)** | Changes a variable within an R object to a factor variable. An example is when you have variables designated with numbers, but they are meant to be categorical variables, so you use this function to tell R that. |
| **matrix(,nrow=)** | Creates a matrix of a given size. You can specify a vector you already have or enter your own. |
| **contrasts(x)** | Sets contrasts matrix, x is a factor. In this activity, we specify a factor within a data frame ex: (corn$Population). |
| **list()** | A generic vector containing other objects. |
| **split()** | Splits a character vector. |
| **interaction(...)** | Computes a factor representing the interaction of the given factors. |

### Exercise 2: Supplements

**Ex. 2: Supplement – Decide Comparison(s)**

When assigning contrast coefficients, you first need to decide what comparison(s) you wish to make. In the activity example, we compare planting densities of 7.5 and 10 plants/$m^{2,}$ and we make another comparison of 12.5 plants/m2 versus the mean of 7.5 and 10 plants/$m^2$. For this first comparison, we will assign a coefficient of +1 to every treatment of 7.5 and -1 for every treatment of 10. The treatments of 12.5 are not included in this comparison, so every treatment with 12.5 will have a coefficient of 0.

So, what are the treatments? If you go back to the datasheet, you will see that there is a column labeled 'treatment', and it lists treatments 1-9. Each treatment represents each Hybrid and Population combination, and each treatment is listed 3 times for the 3 replications (Figure 2). If you have to come up with your own contrast coefficients, it is a good idea to clearly label your treatments for easy reference when assigning coefficients.



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Treatment | Populatio | Hybrid | Rep | Yield |
| 2 | 1 | 7.5 | A | 1 | 7.33 |
| 3 | 1 | 7.5 | A | 2 | 9.69 |
| 4 | 1 | 7.5 | A | 3 | 11 |
| 5 | 2 | 7.5 | B | 1 | 8.98 |
| 6 | 2 | 7.5 | B | 2 | 9.12 |
| 7 | 2 | 7.5 | B | 3 | 9.71 |
| 8 | 3 | 7.5 | C | 1 | 8.5 |
| 9 | 3 | 7.5 | C | 2 | 7.71 |
| 10 | 3 | 7.5 | C | 3 | 9.05 |
| 11 | 4 | 10 | A | 1 | 8.12 |
| 12 | 4 | 10 | A | 2 | 5.39 |
| 13 | 4 | 10 | A | 3 | 11.79 |
| 14 | 5 | 10 | B | 1 | 8.7 |
| 15 | 5 | 10 | B | 2 | 9.67 |
| 16 | 5 | 10 | B | 3 | 11.2 |
| 17 | 6 | 10 | C | 1 | 10.3 |
| 18 | 6 | 10 | C | 2 | 9.14 |
| 19 | 6 | 10 | C | 3 | 8.85 |
| 20 | 7 | 12.5 | A | 1 | 10.2 |
| 21 | 7 | 12.5 | A | 2 | 10.7 |
| 22 | 7 | 12.5 | A | 3 | 10.55 |
| 23 | 8 | 12.5 | B | 1 | 6.09 |
| 24 | 8 | 12.5 | B | 2 | 8.7 |
| 25 | 8 | 12.5 | B | 3 | 8.36 |
| 26 | 9 | 12.5 | C | 1 | 6.53 |
| 27 | 9 | 12.5 | C | 2 | 5.36 |
| 28 | 9 | 12.5 | C | 3 | 4.65 |

Figure 2 Datasheet showing column labeled 'treatment'

**Ex. 2: Supplement – Assign Coefficients**

Now that we have defined the comparisons and labeled our treatments, it is time to assign the coefficients. For 7.5 vs. 10 plants/$m^2$ we assign a +1 for each treatment of 7.5 plants and -1 for each assignment of 10 plants. Remember, 12.5 is not included in this comparison so any 12.5 plant treatments are given a 0. What is happening here is we are subtracting the mean of the 10 plants/$m^2$ from the mean of the 7.5 plants per acre and we will analyze the difference between the two for this contrast. You could use a coefficient of +1/3 instead of +1 since you have 3 treatments and while the F-test will turn out the same you would have to use the correct fraction to get the correct answer. This can be tricky with more complicated contrasts so you may just want to stick to whole numbers for your coefficients.

*Table 19 Treatments and corresponding contrast values for populations, 7.5 versus 10.0 plants per meter squared.*

| Treatment # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Contrast 7.5 vs. 10.0 | +1 | +1 | +1 | -1 | -1 | -1 | 0 | 0 | 0 |

Following this same procedure, we can compare 12.5 plants/$m^2$ to the mean of 7.5 and 10 plants/$m^2$. This time we do not assign a coefficient of -1 to 7.5 and 10 and +1 to 12.5 because then we would be comparing the sum of 6 treatment means to only 3 treatment means. To deal with this we weight the means by assigning -1 to 7.5 and 10 and +2 to 12.5 plants/$m^2$ to make this an even comparison.

*Table 20 Treatments and corresponding contrast values for populations, 7.5 and 10.0 versus 12.5 plants per meter squared*

| Treatment # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Contrast 7.5 vs. 10.0 | +1 | +1 | +1 | -1 | -1 | -1 | 0 | 0 | 0 |
| 7.5 and 10 vs. 12.5 | +1 | +1 | +1 | -1 | -1 | -1 | 2 | 2 | 2 |

You can use the same procedure to design any comparison that you wish to make.

# Trend Comparisons

**Trend comparisons are contrasts among quantitative treatments.** The contrasts performed so far in this lesson have been class comparisons — they determine differences between different qualitative traits or levels of treatments. Class comparisons are by themselves adequate when we are working with qualitative data. For example, we would use qualitative comparisons when comparing different tillage implements or different kinds of fertilizer. The qualitative comparison is also appropriate for comparing corn hybrids, as we just did.

When we are working with quantitative data, however, we should not be content with class comparisons alone! For example, when comparing different rates of fertilizer, we want to know how the crop responds to extra fertilizer:

- Does the crop respond positively to every additional increment of fertilizer?
- Is there an optimum amount of fertilizer beyond which the crop actually responds negatively?
- If we can answer these questions, then we vastly expand our understanding of the yield response to applied fertilizer.

Whereas the class comparisons are useful when looking at discrete variables, the trend comparison can indicate responses of continuous variables. For instance, we found the response at 0, 25, and 50 kg/ha.

- What would happen if there were 35 kg N?

## Linear Trend

**A linear trend contrast is a comparison of high population with low**

The response of corn yield to increasing population can be described using trend comparisons. Yield may increase with every increase in population (Figure 3). The numbers along the left axis correspond to the coefficients that would be assigned to each population in a trend comparison.

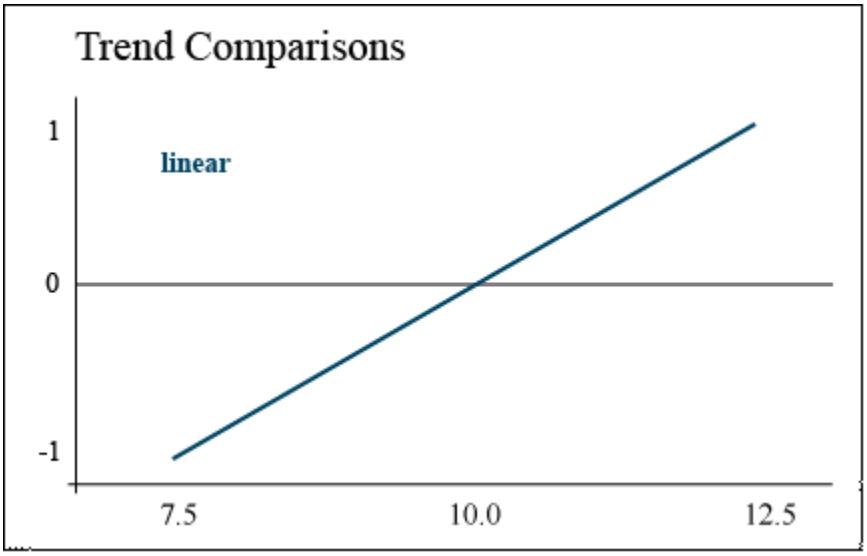



Figure 3 Graph of linear Comparison.

In this case, the contrast is comparing the difference between the highest and lowest populations, assuming a straight line between them. This is known as a **linear** trend.

## Quadratic Trend

**A quadratic trend comparison tests for curvature**

Alternately, corn yield may increase with population up to 10.0 plants/$m^2$, but then decrease when the population is increased to 12.5 plants/$m^2$. Corn is particularly sensitive to available sunlight — a population which is too high will cause excessive shading

and barrenness. This response trend is illustrated by the parabolic response curve (Figure 4).

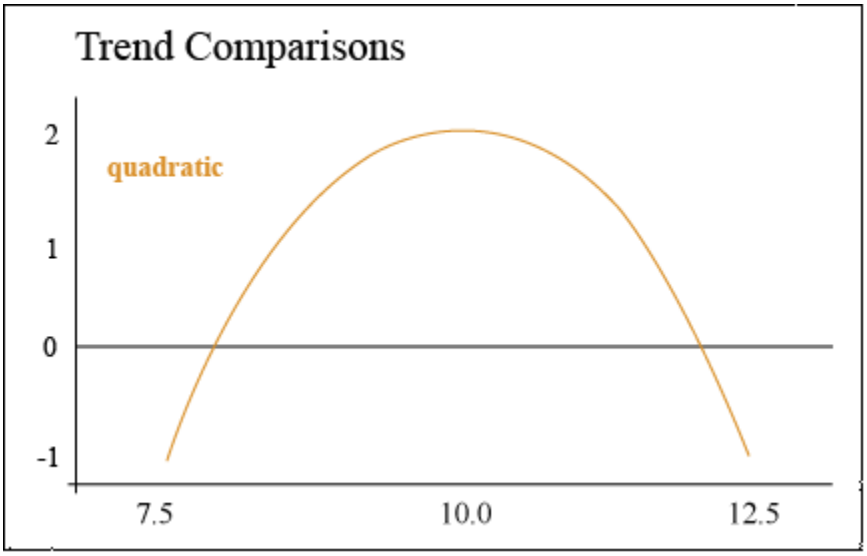



Figure 4 Quadratic trend of yield with increasing population.

In this case, the contrast compares the middle plant population to the sum of the high and low population to determine whether there is a significant peak in yield. In other words, we are determining whether the population is optimal at 10.0.

## Contrast Weights

Contrasts can be used to ascertain the order or the equation that best describes the relationship between a dependent variable such as yield and a quantitative variable such as population. You can use orthogonal polynomial contrasts for quantitative variables for polynomial models up to t – 1 terms; i.e. the maximum order you can explore with the contrasts is one less than the total number of levels for the quantitative treatment. In our example, we are constrained to and really only interested in computing linear (1$^{st}$ order) and quadratic (2$^{nd}$ order) contrasts. These will tell us if the response is a straight line or has some curvature (i.e. nonlinear).

The coefficients to be used with each comparison are shown (Table 21). In the case of both the linear and quadratic trend, these coefficients will test either an increasing or a decreasing trend. In other words, if yield decreased with population, or was actually lowest at 10 plants/m$^2$, the coefficients below would detect those trends as well.

*Table 21 Coefficient for population trend comparisons.*

| Contrast | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Linear | -1 | -1 | -1 | 0 | 0 | 0 | +1 | +1 | +1 |
| Quadratic | -1 | -1 | -1 | 2 | 2 | 2 | -1 | -1 | -1 |

Coefficients similar to these can be created for cubic, quartic, etc., trends if there are more plant population treatments. Tabulations of coefficients for different levels of treatments can be found under the Orthogonal tab in the Statistical Tables workbook.

## Data Analysis

The two comparisons of the corn population experiment are tested using the same procedure as for class comparisons. The sum of squares and mean square for each comparison is calculated, and then the mean square for each comparison is subjected to the F-test. The results are shown in Table 5:

*Table 22 ANOVA table to test contrasts*

| Contrast | df | SS | MS | F | F crit (p = 0.05) |
|---|---|---|---|---|---|
| linear | 1 | 4.726 | 39.15 | 1.25 | 4.41 |
| quadratic | 1 | 107.925 | 37.5015 | 1.20 | |
| error | 18 | 563.611 | 31.312 | | |

Viewing the data in Figure 5 helps to explain why the linear and quadratic trends did not show significance here. There are hints of both types of trend (linear and quadratic) for each population, but no clear trend for all hybrids is obvious.

Hybrids B and C show an optimal population at 10.0, while hybrid A actually performed worst at 10.0. This discussion on linear and quadratic contrasts for trend analysis completes our study of contrasts as mean separation techniques.

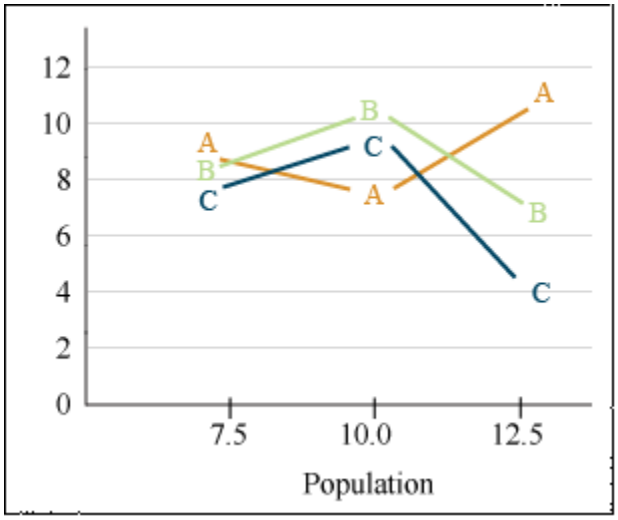



Figure 5 Yield trends for the three hybrids plotted as a function of population.

# Summary

## Comparing Means

- LSD for adjacent pairs or comparison with check cultivar
- HSD for better Type I Error control
- Multiple range tests also protect better than LSD

## LSD

- Equals t standard errors of a difference
- Is the most commonly used method for comparing means

## Contrasts

- Are planned comparisons
- Tested with t-test or F-test
- LSD is the simplest case
- Coefficients sum to zero if equal replication

## Orthogonal Contrasts

- Independent comparisons
- Partition the treatment sum of squares

## Trend Analysis

- Can be done with contrasts
- Useful for quantitative data
- Provides response curve
- Will be done with regression in this chapter and the one on Randomized Complete Block Design

**How to cite this chapter**: Moore, K., R. Mowers, M.L. Harbur, L. Merrick, and A. A. Mahama. 2023. Mean Comparisons. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.10: Mean Comparisons is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ken Moore; Ron Mowers; M. L. Harbur; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.11: Randomized Complete Block Design

It is important when conducting an experiment that the experimental units be as homogenous as possible. This ideal may be met without much difficulty in a lab or within a field with particularly uniform soils. In many field locations, however, the landscape can vary greatly over a short distance. How can we assure, then, that an observed agronomic difference is the result of a specific treatment, rather than the result of the experimental units to which it was allocated? In other words, how do we prevent our treatment results from being confounded with our experimental units? The Randomized Complete Block Design (RCBD) offers one solution.

**Learning Objectives**

- How heterogeneity of experimental units can reduce the sensitivity of an experiment
- How the Randomized Complete Block Design (RCBD) can be used to reduce the heterogeneity of experimental units
- How to conduct the analysis of variance (ANOVA) for an experiment that employs the RCBD
- How to test for the efficiency of the RCBD versus that of the Completely Randomized Design

## Blocking

**The rationale for blocking is to achieve homogeneous experimental units within blocks.** In Chapter 1 on Basic Principles, you learned about the importance of replication in designing a valid experiment. We applied those concepts in Chapter 8 on The Analysis of Variance (ANOVA) when we introduced the analysis of variance. Treatments and replications were assigned to experimental units through the process of randomization. The result of this effort is referred to as a Completely Random Design (CRD).

The CRD is an appropriate experimental design when all experimental units are assumed to be similar or homogeneous (as statisticians like to say). If this is the case, then any observed differences among treatments will cause us to conclude that there was a treatment effect. As mentioned in the introduction, however, homogeneity of experimental units can be difficult to achieve in field plot experiments.

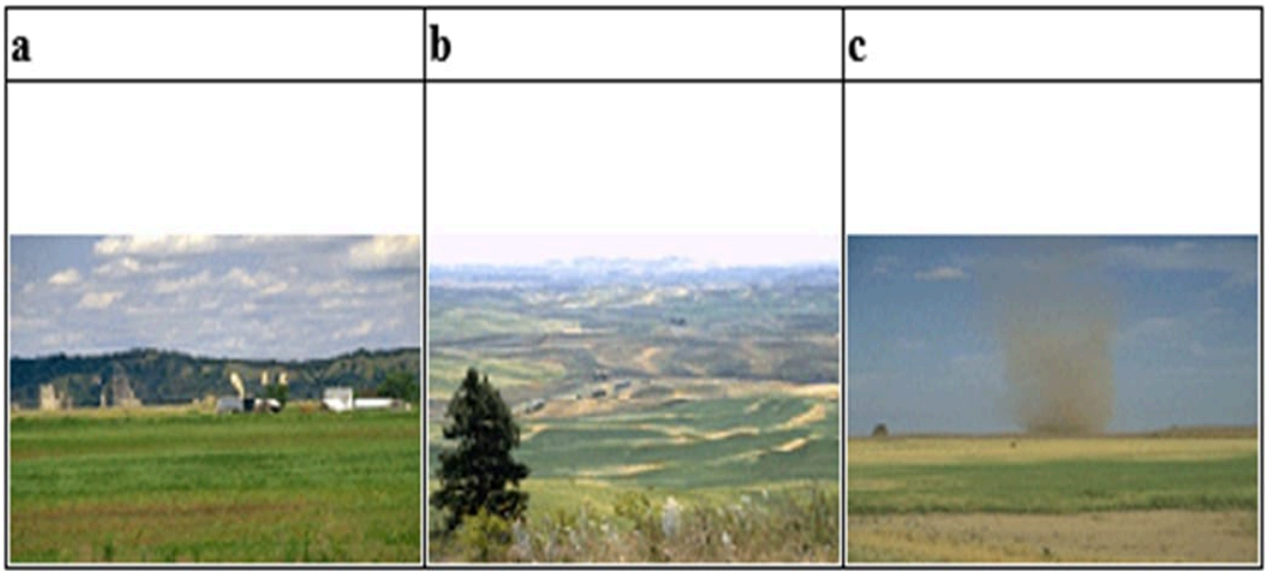



Figure 1 Image for Study Question 1. Different landscapes.

Study Question 1: Homogeneity (radio buttons below are labeled b, c, and a from top to bottom)

Query 1.11.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Heterogeneity

Heterogeneity of experimental units presents two problems. First, the failure to recognize differences between experimental units may lead us to conclude that differences in our variates are the result of the treatments applied, when they were actually caused by the pre-existing condition of the experimental units.

Study Question 2: Type of Error

Query 1.11.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Variance of the Error

Differences between plots (experimental units) not related to the treatments applied to them can inflate the variance of the error associated with the experiment. Recall from Chapter 8 on The Analysis of Variance (ANOVA) that we tested the significance of our treatment using an F-test (Equation 1 below).

If the residual mean square error is increased, then the treatment mean square must also increase in order to maintain the same F-value. In other words, the greater the pre-existing differences between our plots (experimental units), the greater and more profound the difference between treatments must be just to be recognized as statistically significant. Heterogeneity of experimental units thus reduces the sensitivity of our experiment.

$$F = \frac{TMS}{RMS} \tag{1.11.1}$$

Equation 1 (1.11.2)

**where:**
$TMS$= treatment mean square,
$RMS$= residual (error) mean square.

# How to Block

The first step in using the RCBD is to recognize the source(s) of potential heterogeneity among plots (experimental units). In field research, this potential most often exists between plots situated on different soil map units, for example, as the slope changes.

We viewed some of the potential differences in soils related to landscape in Study Question 1. There are many other possible sources of heterogeneity among plots, though. These map units often differ in their yield potential, with the result that some of the variation in the yield measurement is due to the plot location. One example of site heterogeneity is found in Figure 2.

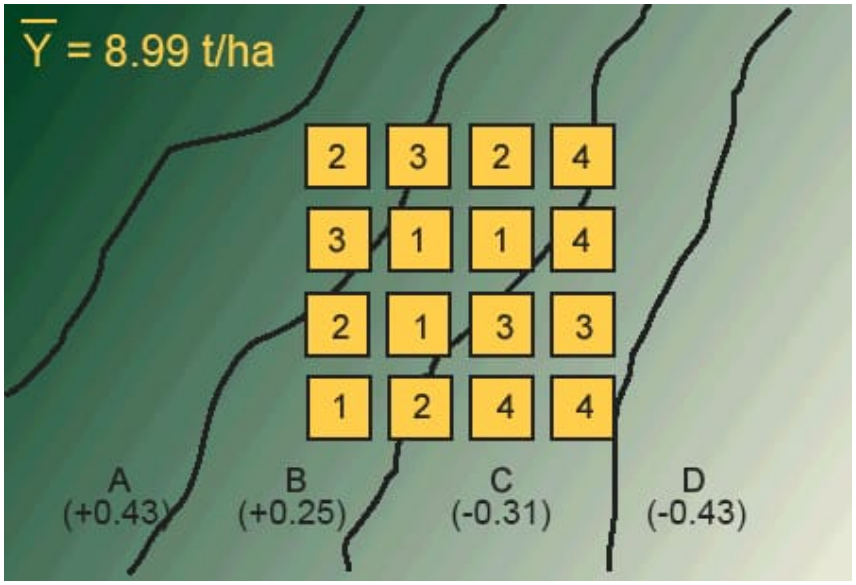



Figure 2 Completely randomized design with four treatments across a yield potential gradient.

In the map in Figure 1, we have four map units (A-D) labeled with the difference in yield between each map unit and the mean for the entire field. As we move from map unit A across to map unit D, the yield potential decreases. In such a case, we say that we have a "production gradient" across the map units. Now let's suppose that we are comparing four different levels of fertilizer (0, 50, 100, 150 kg/ha). If we used a completely randomized design (CRD) across these map units, we risk the possibility of placing all of the high-fertility treatments on extreme map units. This would lead to an unfair comparison of the treatments.

### Study Question 3: Treatment Mean Comparisons

Answer the following question using the treatment map above.

#### Query 1.11.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Treatments

In blocking, we generally place an equal-sized block on every map unit. Each block, in this case, contains four experimental units (plots). Each treatment is applied to one experimental unit within the block (Figure 3).

Each block can be thought of as a replication. Every treatment is forced to occur within each block. Treatments are randomly allocated within each block so that separate randomization is made for each block. In this way, the treatments are prevented from being affected by a second, unrecognized source of heterogeneity that could exist between experimental units within the blocks.

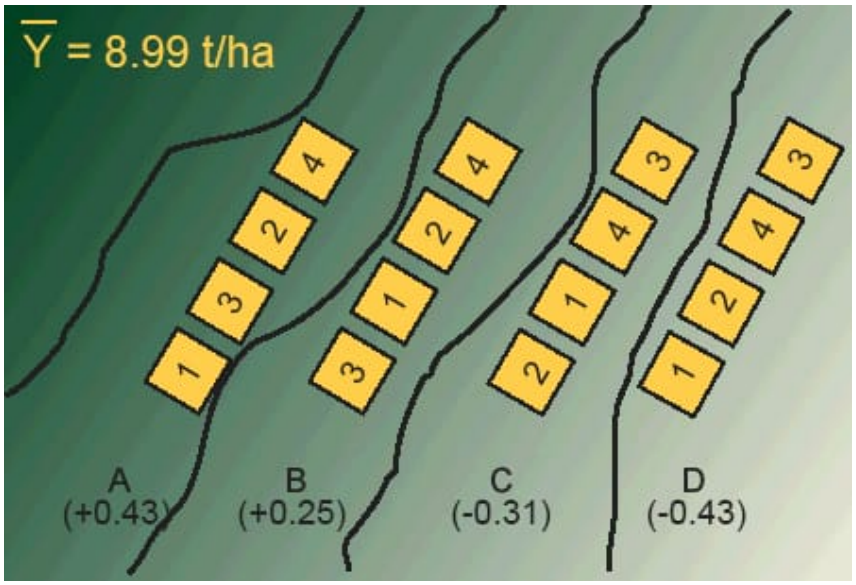


Figure 3 Blocking of randomized treatments to account for a known yield potential gradient.

## Design Control

Forcing each treatment to occur once in every block is sometimes referred to as a restriction on randomization. The restriction in the case of a RCBD is that every treatment must occur in every block. In a CRD, every plot would have the same chance of receiving any treatment, so there is no restriction on randomization; hence the name.

Blocking is a form of design control that was discussed in Chapter 1 on Basic Principles. It is one of the three characteristics of designed experiments (do you remember the other two?). Blocking in a field experiment amounts to grouping plots into more similar sets such that the variation associated with the blocks (whatever is causing them to differ) can be estimated and associated with the block. In a CRD, this variation would be associated with and show up in the Error MS, thus inflating the estimate and reducing the precision of the experiment. When blocking is effective (i.e. there is some variation in the measured response associated with the blocking criterion), this variation is removed from the Error MS and the ability to detect true treatment differences (i.e. precision of the experiment) is improved.

### Further Thought: Discussion

What other possible influences could affect each block beyond the known yield potential gradient?

## Randomization

Figure 4 below illustrates a RCBD randomization scheme. Notice that the Blocks are numbered 100, 200, 300, and 400 and that each of the 5 treatments is on one plot in each block.

Randomization should be done using some sort of randomization method, not just arbitrarily. This precludes any bias which may be unintentionally introduced due to the assignment of treatment. Check out the exercise in the next screens to learn how to randomize treatments for a RCBD using Excel.

Randomization for a RCBD with 5 Treatments and 4 Blocks

| Range | Plot 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 100 | 4 | 5 | 1 | 3 | 2 |
| 200 | 2 | 4 | 1 | 5 | 3 |
| 300 | 2 | 5 | 1 | 4 | 3 |
| 400 | 2 | 1 | 5 | 3 | 4 |

Figure 4 Randomization routine for RCBD.

## Exercise 1: Randomizing Treatments For a RCBD

**Ex. 1: Randomizing Treatments for a RCBD**

As we said before, experimental design is really about how treatments are assigned to experimental units (plots). In the case of the randomized complete block design (RCBD), treatments are blocked into groups of experimental units that are similar for some characteristic. In field experiments, treatments are usually blocked perpendicular to some perceived gradient present in the field. The gradient may be related to such characteristics as soil properties, previous crop history, or any number of other factors that can occur naturally or unnaturally in a field. In a RCBD, treatments are allotted to plots at random within each block of experimental units (plots) such that within any given block, each plot has the same probability of receiving a particular treatment as any other. The only restriction is that every treatment has to occur in every block.

For the purpose of this exercise, we will develop a plot plan for an oat variety trial experiment. The treatments consist of five cultivars replicated four times for a total of twenty plots. The treatments and plots are listed in the Treatments worksheet of the Excel file QM-mod11-ex1data.xls. A map of the field layout is presented in the Plot Plan worksheet. Our goal is to randomize the plot order within blocks and permanently associate it with the treatments.

**Ex. 1: Create a Random Assignment**

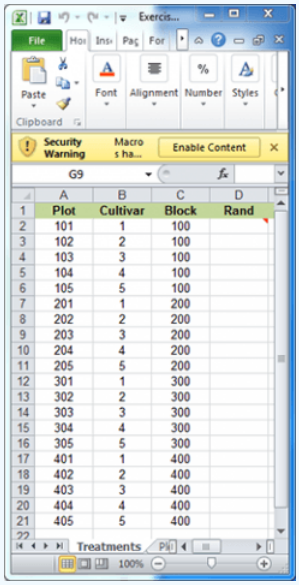
Randomly assign the five oat cultivars to each of the four blocks identified in the **Treatments** worksheet. The idea now is to create a random order of the treatments under the restriction that each **Cultivar** must occur once in each **Block**.

**Steps:**

1. Enter the formula **=RAND()** in cell D2.
2. Copy the formula in cell D2 to cells D3:D21. A fast way to do this is to double-click the square that appears in the lower right-hand corner of the cell when you select cell D2. The **RAND()** function will return a random number between 0 and 1 to each cell the formula is copied to.
3. Using your mouse, select all the cells in the Cultivar, Block, and Rand columns, including the column headings (B1:D21). Do not select the Plot column!
4. Select Sort from the Data menu at the top of the main window.
5. In the Sort by box in the Sort dialog box, select Block.
6. If not already present, add another sort level by clicking on Add Level.
7. In the second Sort by box, select Rand and then click OK.

**Ex. 1: Finished Random Assignment**

You should now have a random assignment of **Cultivars** in column **B,** and each cultivar should occur once in every block. Your worksheet should look something like Figure 5.

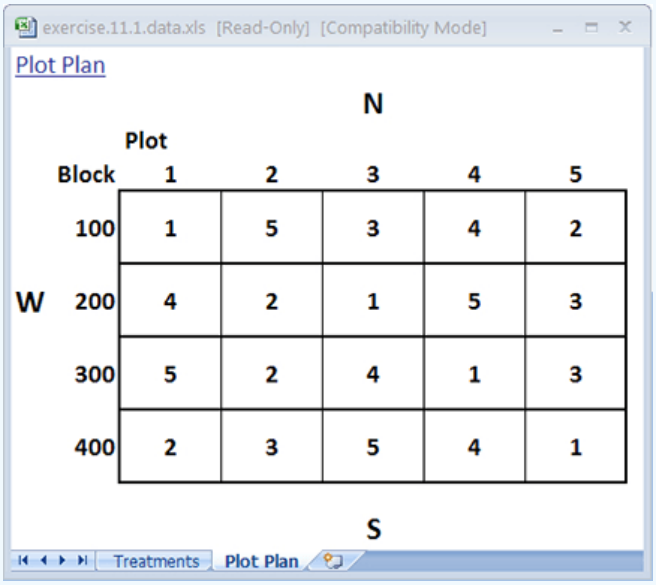


| | A | B | C | D |
|---|---|---|---|---|
| 1 | Plot | Cultivar | Block | Rand |
| 2 | 101 | 1 | 100 | |
| 3 | 102 | 2 | 100 | |
| 4 | 103 | 3 | 100 | |
| 5 | 104 | 4 | 100 | |
| 6 | 105 | 5 | 100 | |
| 7 | 201 | 1 | 200 | |
| 8 | 202 | 2 | 200 | |
| 9 | 203 | 3 | 200 | |
| 10 | 204 | 4 | 200 | |
| 11 | 205 | 5 | 200 | |
| 12 | 301 | 1 | 300 | |
| 13 | 302 | 2 | 300 | |
| 14 | 303 | 3 | 300 | |
| 15 | 304 | 4 | 300 | |
| 16 | 305 | 5 | 300 | |
| 17 | 401 | 1 | 400 | |
| 18 | 402 | 2 | 400 | |
| 19 | 403 | 3 | 400 | |
| 20 | 404 | 4 | 400 | |
| 21 | 405 | 5 | 400 | |

Figure 5 Assignment of cultivars to plots.

**Ex. 1: Plot Plan**

The treatments are now sorted according to plot order. Look at the Plot Plan worksheet to see the field layout of the plots according to the new randomization (Figure 6). Your plan should look something like the one below. Note that every cultivar occurs once in every block. However, the order of treatments within each block will vary with each randomization.



exercise.11.1.data.xls [Read-Only] [Compatibility Mode]

Plot Plan

| | | N | | | | |
|---|---|---|---|---|---|---|
| | Plot | | | | | |
| | Block | 1 | 2 | 3 | 4 | 5 |
| | 100 | 1 | 5 | 3 | 4 | 2 |
| W | 200 | 4 | 2 | 1 | 5 | 3 |
| | 300 | 5 | 2 | 4 | 1 | 3 |
| | 400 | 2 | 3 | 5 | 4 | 1 |
| | | S | | | | |

Treatments | Plot Plan

Figure 6 Plot plan layout for five cultivars.

**Ex. 1: RCBD vs. CRD Randomization**

You may have noticed that randomizing treatments for the CRD and RCBD follow a similar process in Excel. The main difference is in the number of sort levels you use. In the CRD, only one sort level was used which corresponded to the random number generated by the RAND() function. In the case of the RCBD, you used two sort levels: 1) the first one assigned to Blocks, and 2) the second assigned to the random number. You can compare and contrast the plot map from the previous page for an RCBD with one generated for a CRD by removing the first sort level so that the only one that remains is the random number.

The restriction on randomization has been removed so that any cultivar can be assigned to any plot (Figure 7).

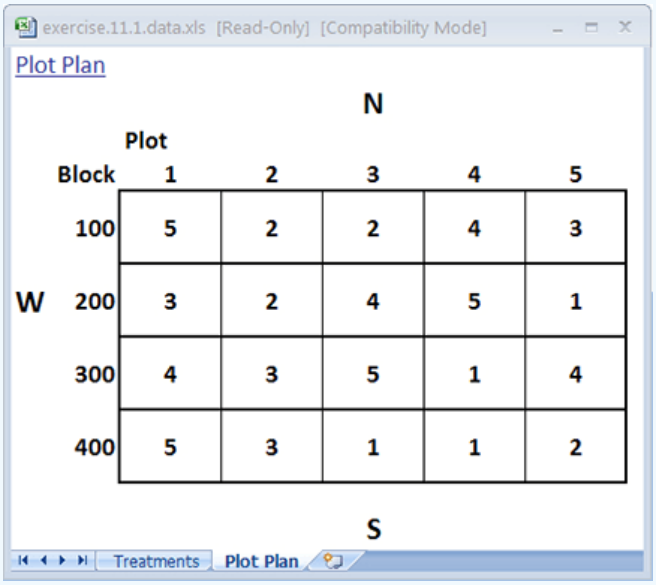



Figure 7 Assignment of cultivars in RCBD

**Ex. 1: R Code Functions**

- matrix
- for
- gl
- runif
- cbind
- _[order]

Experimental design is really about how treatments are assigned to experimental units (or plots). In the case of the randomized complete block design (RCBD), treatments are blocked into groups of experimental units that are similar for some characteristic. In field experiments, treatments are usually blocked perpendicular to some perceived gradient present in the field. The gradient may be related to such characteristics as soil properties, previous crop history, or any number of other factors that can occur naturally or unnaturally in a field. In a RCBD, treatments are allotted to plots at random within each block of experimental units (plots) such that within any given block, each plot has the same probability of receiving a particular treatment (location) as any other. The only restriction is that every treatment has to occur in every block.

**Ex. 1: Maize Yield Test**

You are a maize breeder in charge of designing a field for a yield test of 3 synthetic maize populations. The field has never been used before by your company, and the area surrounding the field is known to have very localized deposits of clay which inhibit root growth and thus negatively affect the yield of plants planted directly on top of the deposits. You want to account for the heterogeneity of the field by assigning each of the 3 maize populations to one of the 3 positions in each of 3 blocks.

In other words, you want to create a random order of the three populations (or treatments) within each block under the restriction that each population must occur once in each block. To make the coding a bit easier for this exercise, we will rename Population 7.5 to Population 1, Population 10.0 to Population 2, and Population 12.5 to Population 3.

**Exercise 1:**

We will learn two ways to do this. The first way will take a little longer than the second but involves less coding. The second way will be much faster than the first. However, it will involve at least a basic understanding of some coding tools, such as loops.

**Ex. 1: Creating a Field**

Let's start by creating a matrix of all zeros, where each entry represents a plot that will be assigned to a cultivar, and each column represents a block. We have 3 cultivars in each of the 3 blocks, thus the dimensions of this 'field' matrix should be 3×3. We will use the matrix command to create the field in R. In the Console window, enter in the parenthesis after matrix, the 0 indicates the type of element we want the matrix to be composed of (i.e., in this case, zeros because we are going to fill in the cultivar numbers, which will be randomly assigned to each plot in each block). The first 3 indicates the number of rows we want in the matrix; since we have 3 populations in each block, we need our matrix to have 3 rows. Finally, the second 3 indicates that we want 3 columns (blocks) in our matrix.

field<-matrix(0,3,3)

Have a look at the field you've created. Enter field into the Console.

field

[,1][,2][,3]

[1,] 0 0 0

[2,] 0 0 0

[3,] 0 0 0

**Ex. 1: Creating a Vector**

Now, let's create a vector representing the cultivars we have in our breeding program. We'll accomplish this by using the gl command. This command allows us to create a vector (or column) of factor variables. In the parenthesis after the gl command, the first entry (3) indicates the number of factor levels, and the second number (1) indicates the number of replications of each factor variable. You may be asking yourself why we are not entering a 3 for the number of replications, as we have 3 randomized complete blocks. The reason we do not enter a 3 for replications is due to the fact that we will create randomized complete blocks individually and enter them into each block (column) in the field matrix. This will become apparent in the following steps.

Enter into the Console:

pop<-gl(3,1)

pop

[1] 1 2 3

Levels: 1 2 3

**Ex. 1: Vector with 3 Entries**

Now let's create a vector with 3 entries, where each entry is a random number between 0 and 1. We'll use the runif command to do this. This command is used by entering the number of entries of numbers between 0 and 1 you want your vector to be composed of in parenthesis after the runif command (i.e. for this example, we will enter 3, since we have 3 maize populations). Let us call this vector with 3 entries of random numbers between 0 and 1 rand. Create the vector *rand*, then look at it by entering the vector name (*rand*).

rand<-runif(3)

rand

[1] 0.1165839 0.5730972

0.3469669

**Ex. 1: New Matrix with Block**

Now, let us create a new matrix called *block* by putting the *cultivar* and *rand* vectors together, so each random number in the rand vector corresponds to one of the 3 maize populations. The cbind command can be used to put two vectors together. The command is used by entering the cbind command followed by the two vectors you want to be put together (or concatenated) in parenthesis separated by a comma. Create the matrix called block with the cbind command, then look at the matrix by entering the name of the matrix (block).

block<-cbind(pop,rand)

block

pop rand

[1,] 1 0.1165839

[2,] 2 0.5730972

[3,] 3 0.3469669

**Ex. 1: Ordering the Population in Block**

We can now order the maize populations in the first column of the block matrix by each of the 3 population's corresponding random number in the rand vector. We will use the order command to sort the populations based on their corresponding number in the **rand** column, by the population with the smallest random number first to the population with the largest random number last. The population with the smallest random number is Population 1 (0.1165839), and the population with the largest random number is Population 2 (0.5730972). Thus, the order for this randomized complete block should be 1,3,2. We will now go through the order command.

Let us call this randomized complete block *randblock*. Create matrix *randblock* and take a look at it by entering the following in the Console.

block<-cbind(pop,rand)

block

pop rand

[1,] 1 0.1165839

[2,] 2 0.5730972

[3,] 3 0.3469669

**Ex. 1: Filling the Block**

To use the order command, we first write the matrix or vector that we'd like ordered (i.e., in this case, block), followed by brackets. Then, we write the command order followed by the column that we would like the ordering of the rows based off of (i.e., in this case the second column of the block matrix, which contains the random numbers between 0 and 1 for each population. Note: the row number is left blank in the brackets before the comma block[order(block[,2]),] to specify that we want all of the rows sorted based on the values of column 2 in the *block* matrix. Since the second column in the *block* matrix is the random number column, this is the column that we want to order the populations by; thus, we enter block[,2] to specify we want the rows ordered based on the values in the second column in the *block* matrix. The default of the order function is to sort in a ascending order, with the lowest value at the top of the column and the highest value at the bottom. We must now enter this randomized block into the first block (column) of the *field* matrix we created previously. We'll do this by setting the first column of the field matrix (specified by field[,1]) to equal the first column of the *randblock* matrix (specified by randblock[,1]). Carry out this operation, then look at the *field* matrix to make sure you've entered the first column from the *randblock* matrix into the *field* matrix.

field[1,]<-randblock[1,]

field

[,1] [,2] [,3]

[1,] 1 0 0

[2,] 2 0 0

[3,] 3 0 0

Again, we specify the first column of the field matrix with field[,1], and set it equal to the first column of the *randblock* matrix, specified by randblock[,1]. Note: The order of the cultivars in the first column of your field matrix may not be the same as in this lesson due to the randomization process.

To fill in the rest of the blocks, carry out all of the same steps you just did, but when entering the next block into the field matrix, change the
field[,1]<-randblock[,1] to field[,2]<-randblock[,1]
to specify that you want to enter the randomized order for the second block (or column) in the *field* matrix.

**Ex. 1: Review RCBD Method 1**

To summarize what we've just learned, the first method for creating 3 randomized complete blocks with 3 populations is reviewed succinctly below.

First, we create a field matrix of all zeros, with the dimensions of the number of entries in each block, and the number of blocks desired (i.e., in this case we have 3 populations or entries, and 3 blocks).

field<-matrix(0,3,3)

Then, go through steps 1-5 (presented below) 3 times. After each time through steps 1-5, add 1 to the column indicated in the *field* matrix in step 5. For example, the second time through steps 1-5, step 5 will be field[,2]<-randblock[,1], and the third time through step five will be field[,3]<-randblock[,1], etc.

cultivar<-gl(3,1)

rand<-runif(3)

block<-cbind(cultivar,rand)

randblock <- block[order(block[,2]),]

field[,1]<-randblock[,1]

Have a look at the field you've just created:

field

[,1] [,2] [,3]

[1,] 1 1 2

[2,] 3 3 1

[3,] 2 2 3

**Ex. 1: RCBD Method 2**

This method can save time in comparison with the first method, especially if you have many treatments that you want randomized in many blocks. The two methods are computationally equivalent, however the second method utilizes a loop command to repeat

the operations that we previously did for each block in Method 1.

We can use a for loop to go through a set of operations a specified number of times. Using the for loop, we must first assign an iteration variable, which corresponds to the number of times the set of operations has been completed. For example, if we assign i=1:3 as the iteration variable, the first time through the set of commands i=1, the second time i=2, the third time i=3. In this example, we'll use the letter i to indicate the iteration variable in our loop.

Let's clear the entire data frame before starting Method 2. Use the rm(list=ls()) command to clear the entire data frame/ environment. The upper-right window should now be clear of all variables and data.

```
rm(list=1s())
```

**Ex. 1: Creating a Field Matrix**

Great! Now create the *field* matrix in the same way as in Method 1.

```
rm(-matrix(0,3,3)
```

Now we need to enter the loop with the number of cycles, or iterations we want carried out. In this case, we want to create 3 randomized complete blocks, so we the total number of iterations is 3. Enter the for command into the Console with the iteration variable i ndicating that we want 3 iterations carried out

```
for (i in 1:3)
```

Good, we have indicated that we want i to be our iteration variable ranging from 1 to 3. Now, we need to enter the bracket { , then enter lines 1-5 from the method 1 code with line 5 ending in a } bracket.

```
{pop<-gl(3,1)

rand<-runif(3)

block<-cbind(pop,rand)

randblock <- block[order(block[,2]),]

field[,i]<-randblock[,1]}
```

**Ex. 1: Finished Field Matrix**

Look at line 5 for Method 2 directly above. Do you notice anything different than line 5 in Method 1? Instead of manually entering the block number in the field[,1]<-randblock[,1] command, we simply enter i, so that line 5 in Method 2 becomes field[,i]<-randblock[,1]. The value of i is 1 for the first iteration, 2 for the second iteration, and 3 for the third iteration. This can save us a lot of time if we are trying to create many randomized complete blocks for many treatments.

Let's now go through method 2 in its entirety.

```
for (i in 1:3)

{pop<-gl(3,1)

rand<-runif(3)

block<-cbind(pop,rand)

randblock <-

block[order(block[,2]),]

field[,i]<-randblock[,1]}
```

Look at the field you've just created.

field

[,1] [,2] [,3]

[1,] 1 1 2

[2,] 3 3 1

[3,] 2 2 3

**Ex. 1: Review RCBD Method 2**

To reiterate, Method 2 is accomplished by first creating the *field* matrix.

field<-matrix(0,3,3)

Then entering the for command, specifying the iteration variable (i.e., i in 1:3)

for (i in 1:3)

and finally entering a { bracket, lines 1 to 5 from Method 1, with a } after the final line (line 5).

{cultivar<-gl(3,1)

rand<-runif(3)

block<-cbind(cultivar,rand)

randblock <- block[order(block[,2]),]

field[,i]<-randblock[,1]}

# Linear Additive Model

**The Linear Additive Model also applies for the RCBD.** In Chapter 8 on The Analysis of Variance (ANOVA) we introduced the linear additive model in describing the analysis of variance. The model showed how the measured or dependent variable is affected by different factors in the model (independent or classification variables). According to the linear model, the observed result can be estimated by adding (summing) the effects of the model terms. By introducing blocking into the experimental design, another possible source of variation is included in the linear additive mode that is associated with blocks. Thus, the model for a RCBD includes an additional term for blocks. The error term can now be thought of as that variability among plots (experimental units) that cannot be accounted for by blocks or treatments.

The linear additive model for the RCBD must include the effects of blocking, treatment(s), and error (Equation 2). The model for a single-factor RCBD is:

$$Y_{ij} = \mu + \beta_i + T_j + \beta T_{ij} \tag{1.11.3}$$

$$\text{Equation 2} \tag{1.11.4}$$

**where:**
$Y_{ij}$ = response observed for the $ij^{th}$ experimental unit,
$\mu$ = grand mean,
$B$ = block effect,
$T$ = treatment effect,
$BT$ = block error x treatment interaction.

## Differences in Models

This model differs from the linear additive model for the CRD in two ways. First, it differs by including blocks as an effect. Blocks capture the effect related to the blocking criterion, which is often soil heterogeneity in field experiments. The RCBD model, by having a block effect, inherently includes a **restriction on randomization**. The RCBD is not completely randomized — instead, each level of the treatment is "forced" to occur in each block.

The second difference is that the block x treatment interaction is the error term. Why should we use this interaction to test the effect of treatment? The answer is that we test treatment differences to see if they remain relatively large compared with their random changes for different blocks. These random changes in treatments over blocks comprise the BT interaction.

Suppose that soybean yield means for three herbicide treatments are 2.7 t/ha for herbicide A, 3.0 for B, and 3.3 for C. If these differences remain fairly consistent over each block, the block x treatment interaction will be small relative to the mean yield differences, and the F-ratio of treatment MS to error (BT interaction) will be large. In effect, we are testing whether the yield differences remain relatively large in comparison with their (random) changes over blocks, i.e., block x treatment interaction. The block x treatment is the proper error term for testing for treatment differences and it is used for the F-Test in the ANOVA. A large F-value implies treatment differences are large relative to the error.

## Treatment Differences

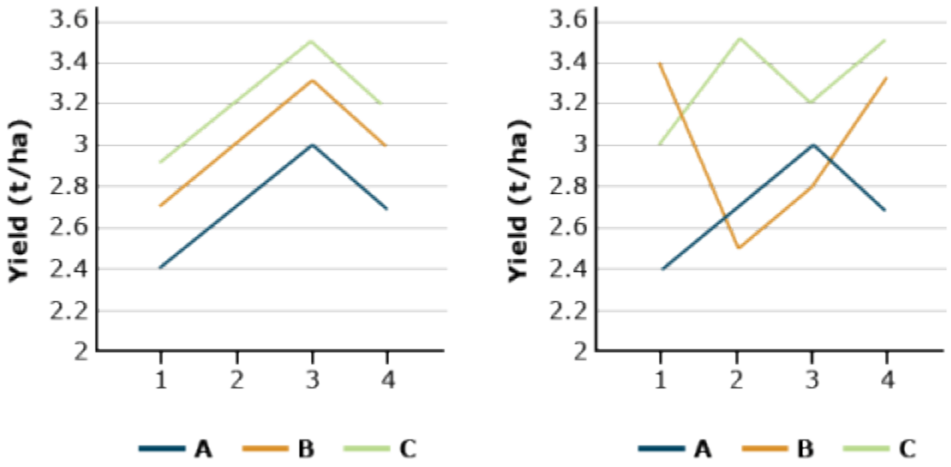



Figure 8 Illustration of variation of Treatments relative to the treatment-by-block interaction. (Treatment means are the same for both cases.)

In both graphs in Figure 8, the average soybean yields are the same (2.7, 3.0, and 3.3 t/ha). We trust the results more if these differences are consistent across blocks (left-hand side). When the B x T interaction is large, (right-hand side), the random error obscures the treatment differences.

## Estimate Effects Using ANOVA

Perhaps the linear model will be clearer if we use an example. Recall the corn population experiment from Chapter 5 on Categorical Data—Multivariate. In that experiment, corn was planted at three populations in order to determine the effect on grain yield. If we treat the repetitions as being blocks, the linear model for that specific experiment is (Equation 3):

$$Y_{ij} = \mu + Blk_i + POP_j + BlkxPOP_{ij} \tag{1.11.5}$$

$$\text{Equation 3} \tag{1.11.6}$$

**where:**
$\mu$ = grand mean,
$Blk$ = block effect,
$POP$ = population effect.

The linear model provides a convenient method of listing the effects which are to be estimated using an ANOVA. As you will see,

every effect from the linear model (with the exception of the mean) will be included in the ANOVA table. The arguments you list in the aov() function in R correspond directly to the terms that comprise the linear additive model.

## Exercise 2: Analyze an RCBD Experiment

**R Code Functions**

- attach()
- as.factor()
- summary()
- aov()

In this example, you will learn how to analyze data from an experiment that has a restriction on randomization. This exercise will give us an opportunity to evaluate how blocking affects the sum of squares for our model. More specifically, it will allow us to discern how Total SS are partitioned between the two designs (blocking vs. not blocking).

**Exercise 2: Analyze an RCBD experiment**

We return to the synthetic maize population scenario which we used in the previous exercise, where we created randomized complete blocks. You are again a maize breeder and have now been asked by your supervisor to analyze the yield data for the 3 synthetic maize populations that were planted in a yield trial in 3 randomized complete blocks. Conduct an ANOVA on this data with cultivar and block as factors.

**Ex. 2: Beginning Analysis**

**Reading the Data**To begin the analysis, first set the working directory and read the data into R ...steps presented in the CRD activity.

**Analyzing The Data**

The code required to analyze the experiment is similar to what we used to analyze the Maize Population Example in the CRD activity. The linear additive model for an RCBD includes an additional term to account for the linear effect of blocks.

Let's first do an analysis without incorporating block into the model. This is exactly what we did in the analysis of the CRD experiment, where the only factor in the model was population.

data<-read.csv("exercise.11.2.data.csv", header =T)

head(data, n=3)

Pop Block Yield

1 7.5 1 8.50

2 7.5 2 7.71

3 7.5 3 8.50

attach(data)

**Ex. 2: Running ANOVA**

Set population as a factor, equal to variable Pop.

Pop<- as.factor(data$Pop)

out<- summary(aov(Yield ~ Pop))

out

Df Sum Sq Mean Sq F value Pr(>F)

Pop 2 24.809 12.405 19.15 0.00249 **

Residuals 6 3.887 0.648

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ ‘ ‘ 1

Run the ANOVA as we did in the CRD activity. The one-factor model for the ANOVA is: Yield = Population. After you run the ANOVA, look at the ANOVA table.

Let’s run the ANOVA incorporating block as a second factor. Now we can run the ANOVA. The model we will use is: Yield = Population + Block.

Pop<- as.factor(data$Pop)

Block<- as.factor(data$Block)

out<- summary(aov(data$Yield ~ Pop+Block))

out

Df Sum Sq Mean Sq F value Pr(>F)

Pop 2 24.809 12.405 25.65 0.00523 **

Block 2 1.953 0.977 2.02 0.24756

Residuals 4 1.934 0.484

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ ‘ ‘ 1

**Ex. 2: Interpreting Results, Comparing The CRD Vs. RCBD ANOVA**

The information that most interests us is the last 2 columns of the first row in the ANOVA table; the F-value and P-value for factor variable pop. The F-value is 25.65, which corresponds to a P-value of 0.00523. We would therefore conclude that population had a significant effect at 0.1% on yield in this experiment.

Our next step in the analysis would normally be to evaluate the mean differences among the three populations. However, at this point, we will take the opportunity to compare and contrast the two designs.

The first question we want to answer is, “How did including blocks as a factor in the analysis affect the partitioning of df and SS”? You will notice right away that the total df and SS are exactly the same for the two analyses. What has changed is how these are partitioned between the Model and Residuals. The Model df in the RCBD analysis increased by 2, reflecting the two additional df from adding the block term. Note that these were partitioned out of the Error df, which was reduced by 2, from 6 to 4. Something changed with the SS from the one-factor ANOVA to the two-factor ANOVA. Some of the variation that was unexplained and associated with error has been partitioned into the model SS; i.e. it went from being unexplained to being explained by the model. By viewing the ANOVA at the bottom of the figure, you can see that this variation was attributed to blocks in the RCBD.

**Ex. 2: Conclusions**

This improves the precision of the F-test for population because the Error MS decreased as a result of partitioning some of the variation into blocks. Since the Error MS is smaller for the RCBD, there will also be improved precision for comparing means because the standard errors used to calculate these tests are calculated from the Error MS.

You may have noticed that R computed an F-test for Blocks and indicated that it was nonsignificant (with a P-value of 0.24756). Normally we are not interested in this test because our primary goal with blocking is to reduce the Error MS. In this case, blocking resulted in a modest reduction in the Error MS, so the test turned out to be nonsignificant. Regardless of the magnitude of the Block effect, we would use this analysis since our experimental treatments were arranged in an RCBD.

There is a cost to blocking that affects the precision of statistical tests in cases like this where df for error is marginal. If you have a look at a distribution of F-values, you can see that the critical F-value (the one that you have to exceed to demonstrate significance) decreases rapidly as you add error df up to about 10, after which it decreases much more slowly. The reason for this is that we are much less confident in variances estimated from a small number of samples than those estimated from larger numbers. In our example, even though the F-value was larger for the RCBD, the probability of it occurring at random was actually higher than the CRD (P > F = 0.0052 vs. 0.0025). This is a direct consequence of the reduction in df for Error.

# Analysis of Variance for RCBD

The analysis of variance used with RCBD is similar to that used with the Completely Randomized Design for a factorial experiment. The differences are the inclusion of the block effect and the replacement of the error term with the block x treatment effect. The ANOVA table is structured to account for these effects (Table 1).

*Table 1 ANOVA Table layout for RCBD.*

| Source of Variation | n/a |
|---|---|
| **Treatment** | n/a |
| **Block** | n/a |
| **Error** | n/a |
| **Total** | n/a |

Compare these to the linear model from the previous section.

## Example Using RCBD

For our example, we will use the same dataset as in Chapter 8 on The Analysis of Variance (ANOVA), corn planted at three populations. In this example, however, the three replicates are arranged in blocks, in contrast to the Completely Randomized Design in Chapter 9 on Two Factor ANOVAS. Three replications were included in the experiment for a total of 9 (3 pop x 3 rep) experimental units. The data are listed in Table 2.

*Table 2 Yield data (t/ha) for corn planted at three populations using RCBD.*

| Treatment | Blocks | | | |
|---|---|---|---|---|
| **Populations (plants/acre)** | **1** | **2** | **3** | **4** |
| **7.5** | 8.50 | 7.71 | 9.05 | 8.190 |
| **10** | 10.30 | 9.14 | 8.85 | 9.573 |
| **12.5** | 6.53 | 5.36 | 4.65 | 7.260 |
| $y_{*j}$ | 6.583 | 6.053 | 6.388 | n/a |

If the replications were blocked, then the ANOVA table for the corn population experiment example would include the following sources of variation: blocks, population (the treatment), the interaction (block x population), and total.

## Degrees of Freedom

The degrees of freedom are also calculated in a manner similar to the factorial experiment. The block and treatment degrees of freedom are simply the number of blocks or treatment levels minus one. The error (interaction) df is the product of the block and treatment degrees of freedom. These are summarized below:

*Table 3 Degrees of Freedom RCBD.*

| Source of Variation | Degrees of Freedom |
|---|---|
| **Treatment** | # of levels of treatment -1 |
| **Block** | # of blocks -1 |
| **Error** | (df for treatment x df for blocks) |
| **Total** | [(# of levels of treatment) x (numbers of blocks)] – 1 |

For the sample experiment, there are 9 total experimental units in the experiment, leaving eight total df. The df associated with blocks and treatment are, in each case, one less than the number of levels for each factor, or 2 = (3 – 1) df. The interaction df is the product of the block and treatment df, or 4 = (2 x 2).

## Sum of Squares

The sums of squares for the Randomized Complete Block Design are similar to those calculated in Chapter 8 on The Analysis of Variance (ANOVA). Recall we first calculate the correction factor (CF) (Equation 4):

$$\text{CF} = \frac{(\sum x)^2}{n} \tag{1.11.7}$$

$$\text{Equation 4} \tag{1.11.8}$$

**where:**
$x$= each observation,
$n$= number of observations.

Our correction factor is the same as in Chapter 8 on The Analysis of Variance (ANOVA):

$$\text{CF} = \frac{8.50 + 7.71 + 9.05 + 10.30 + 9.14 + 8.85 + 6.53 + 5.36 + 4.65)^2}{9} = \frac{4912.61}{9} = 545.85 \tag{1.11.9}$$

We also calculate the Treatment sum of squares as in Chapter 8 on The Analysis of Variance (ANOVA) (Equation 5):

$$\text{Treatment SS} = \sum(\frac{T^2}{r}) - \text{CF} \tag{1.11.10}$$

$$\text{Equation 5} \tag{1.11.11}$$

**where:**
$T$= each treatment level,
$r$= number of replications,
$CF$ = correction factor.

## Sum of Squares Example

In our example, the Treat SS is:

$$\text{Treatment SS} = (\frac{25.26^2}{3} + \frac{28.29^2}{3} + \frac{16.54^2}{3}) - \text{CF} = (212 - 69 + 266.77 + 91.19) - 545.85 = 24.81 \tag{1.11.12}$$

The total sum of squares is (Equation 6):

$$\text{Total SS} = \sum x^2 - \text{CF} \tag{1.11.13}$$

$$\text{Equation 6} \tag{1.11.14}$$

**where:**
$x$= each observation,

$CF$ = correction factor.

Thus, the total SS in our example is:

$$\text{Total SS} = (8.5^2 + 7.71^2 + 9.05^2 + 10.3^2 + 9.14^2 + 8.85^2 + 6.53^2 + 5.36^2 + 4.65^2) - 545.85 = 28.70 \quad (1.11.15)$$

## Difference in RCBD and CRD

So far, every source of variation in the Randomized Complete Block Design is exactly the same as the Completely Randomized Design. The RCBD differs from the CRD in that it includes Blocks as a source of variation (Equation 7).

$$\text{Block SS} = \sum(\frac{B^2}{t}) - \text{CF} \quad (1.11.16)$$

$$\text{Equation 7} \quad (1.11.17)$$

**where:**
$B$= each block total,
$t$= number of treatments,
$CF$ = correction factor.

For example, the Block SS is:

$$\text{Block SS} = \text{Block SS} = (\frac{25.33^2}{3} + \frac{22.21^2}{3} + \frac{22.55^2}{3}) - \text{CF} = (213.87 + 164.43 + 169.50) - 545.85 = 547.80 - 545.85 = 1.95 \quad (1.11.18)$$

We calculate the residual (error) sum of squares for the RCBD similar to how we did for the CRD, only now we subtract both the Treat SS and Block SS from the Total SS (Equation 8).

$$\text{Residual SS} = \text{Total SS - Block SS - Treatment SS} = 28.70 - 1.95 - 24.80 = 1.95 \quad (1.11.19)$$

$$\text{Equation 8} \quad (1.11.20)$$

You may see the Residual SS listed in some tables as the Block*Treatment interaction.

### Study Question 4: Observations

#### Query 1.11.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Sum of Squares Table

*Table 4 Sum of squares ANOVA table for corn planted at three populations in northwest Iowa.*

| Source of Variation | Degrees of Freedom | Sum of Squares |
|---|---|---|
| Treatment | 2 | 24.81 |
| Block | 2 | 1.95 |
| Error | 4 | 1.93 |
| Total | 8 | 28.70 |

The sums of squares for each source of variation in the experiment are shown below.

## Mean Squares

Mean squares are calculated in the same manner regardless of the design used: in each case, the mean square is equal to the sum of squares divided by the degrees of freedom for each source of variation. The mean squares for the sample experiment are shown below.

*Table 5 Mean squares ANOVA table for corn planted at three populations in northwest Iowa.*

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square |
|---|---|---|---|
| **Treatment** | 2 | 24.81 | 12.4 |
| **Block** | 2 | 1.95 | 0.977 |
| **Error** | 4 | 1.93 | 0.484 |
| **Total** | 8 | 28.70 | n/a |

## F-Values and F-Test

The observed F-value for the treatment is calculated for the RCBD experiment by dividing the treatment mean square (TMS) by the residual mean square (RMS). In other words, F = TMS / RMS (Equation 1).

The observed F-value for treatment must be compared with a critical F-value in order to test the significance of the treatment effect. This critical F-value is determined using the same procedure as for the CRD: the value is selected from Appendix 4a, using the treatment df to select the column and the error df to select the row. The desired significance level (P=0.05, 0.025, 0.01, or 0.001) determines which of the 4 numbers is chosen.

The observed and critical F-values for the sample experiment are shown below:

*Table 6 Observes F and critical F ANOVA table for corn planted at three populations in northwest Iowa.*

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Observed F(5%) |
|---|---|---|---|---|---|
| **Treatment** | 2 | 24.81 | 12.405 | 25.65 | 6.95 |
| **Block** | 2 | 1.95 | 0.977 | n/a | n/a |
| **Error** | 4 | 1.93 | 0.484 | n/a | n/a |
| **Total** | 8 | 28.70 | n/a | n/a | n/a |

## RCBD Analysis Exercises using R

Since the calculated F (25.72) exceeds the critical F (6.94), we reject the null hypothesis and conclude that there is a significant difference due to treatments.

We see next that R can be used to do an RCBD analysis.

### R Code Functions

- read.csv
- as.factor
- attach
- summary
- aov
- sqrt

## Exercise 3

**Exercise: Analyzing Another RCBD**

You are a forage breeder and have been asked by your supervisor to analyze data from a variety trial in which 10 cultivars of red clover were evaluated for dry yield. The experimental design was an RCBD with four replications (4 randomized complete blocks, each consisting of all 10 cultivars in a randomized order). The yield data represent seasonal totals in tons/acre. Carry out an ANOVA on the yield data, and determine the effect of blocking on the partitioning of residual SS.

The data are in a **.xls** file found here. Download it and name it **exercise.11.3.data.csv**.

**Ex. 3: Two-Factor ANOVA**

Run the analysis of variance using a two-factor model (with Cultivar and Block as factors).

cult<-as.factor(data$Cultivar)

block<-as.factor(data$Block)

out <- summary(aov(Yield ~ cult + block))

out

Df Sum Sq Mean Sq F value Pr(>F)

cult 9 0.04257 0.004730 5.326 0.00033 ***

block 3 0.07997 0.026657 30.013 9.5e-90 ***

Residuals 27 0.02398 0.000888

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Look at the ANOVA table.

**Ex. 3: Interpreting Results**

Looking first at the F-test for Cultivar, we see that the calculated F-value is 5.326, which is significant at P = 0.00033. The chance of making a Type I Error in declaring that there is a difference in yield among the ten cultivars is very small, so we conclude that such a difference exists. You can also see from the ANOVA that blocking was much more effective in this example than in the previous one. In fact, Block explained more variation among our plots than did Cultivar. If you were to analyze this data as a CRD, you would find that Cultivar did not affect red clover yield in this experiment. The SS associated with Cultivar would not differ between the two analyses. What differs is that some of the variation associated with plots in the CRD analysis has been partitioned into a Block effect in the RCBD. Therefore, the error MS is smaller for the RCBD giving you more precision for the F-test and subsequent mean comparisons.

**Ex. 3: Standard Error of the Mean (SEM)**

The **Standard Error of the Mean (SEM)** is often reported in association with the treatment means of an experiment. The SEM is the square root of the variance of the mean. It is a very useful statistic for comparing means, as the SEM can be used to calculate significant ranges in a number of multiple comparison procedures, such as the Fisher’s LSD mean comparison. The SEM is calculated as (Equation 9):

$$SEM = \sqrt{\frac{2RMS}{r}} \tag{1.11.21}$$

Equation 9 (1.11.22)

**where:**
$RMS$=residual (or error) mean square,
$r$= is the number of observations used to calculate the mean. Usually, r is equal to the number of replications or blocks.

**Ex. 3: RCBD – Red Clover Variety Trial**

Use the ANOVA table that you just made for the RCBD ***Red Clover*** variety trial to calculate the standard error of the mean for the experiment.

1. Look up the error mean square in the analysis of variance table from Exercise 3.

0.000888

2. Compute the SEM from the above formula and the error MS from the ANOVA table.

The mean square of the residuals is 0.0888. The number of blocks per treatment is 4. We can use the **sqrt** command in R to calculate the square root of RMS/r.

sqrt(0.000888/4)

[1] 0.01489966

The SEM is, therefore, 0.01489966.

## Exercise 4

**Ex. 4: Mean Comparisons with RCBD**

**R Code Functions**

- install.packages("")
- library
- LSD.test()
- sqrt
- abs(qt())
- order

The mean comparison procedure we'll use for the Red Clover variety trial is the least significant difference (LSD) comparison. This is because we are comparing a large number of qualitative treatments for which there are no obvious preplanned comparisons. The LSD tells us the minimum mean difference that we should consider between individuals in the sample population that we are analyzing.

**Ex. 4: Calculating LSD**

Formula 1:

A useful formula for calculating an LSD is (Equation 10)

$$\text{LSD} = t \times \sqrt{\frac{2 * MSE}{n}} \tag{1.11.23}$$

Equation 10 (1.11.24)

**where:**
$MSE$= error mean square,
$n$= the number of observations used to calculate each mean.

$$\text{SEM} = \sqrt{\frac{MSE}{n}} \tag{1.11.25}$$

Formula 2:

LSD can also be calculated as (Equation 11)

\[\textrm{LSD} = \textrm{t}\times\sqrt

UndefinedNameError: reference to undefined name 'MSE' (click for details)

```
Callstack:
    at (Bookshelves/Agriculture_and_Horticulture/Quantitative_Methods_for_Plant_Breeding_(Su
```

\]

Equation 11 (1.11.26)

Ex. 4: LSD Calculation Exercise

In this exercise, we will calculate the LSD for comparing means of the Red Clover data using the following steps:

1. Obtain the residual (or error) MS from the ANOVA table for the Red Clover variety trial. We did this in the last exercise. The ANOVA table is presented below.

cult<-as.factor(data$Cultivar)

block<-as.factor(data$Block)

out<- summary(aov(Yield ~ cult + block))

out

Df Sum Sq Mean Sq F value Pr(>F)

cult 9 0.04257 0.004730 5.326 0.00033 ***

block 3 0.07997 0.026657 30.013 9.5e-09 ***

Residuals 27 0.02398 0.000888

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

2. Compute the LSD using the appropriate t-value.
3. We can calculate the two-sided t-value at the 0.05 P level with 27 degrees of freedom and set it equal to variable t with the following command:

t<-abs(qt(0.5/2,27)) # Calculate the two-sided P-level of 0.05 with 27df

t

[1] 2.051831

Let us go through the command above: the P-level (or significance value) is specified after the second parenthesis (i.e. 0.05/2), and the residual degrees of freedom are entered to the right of the P-level after a comma (i.e. 27 in this example). The two-sided t value at the 0.05 P level with 27 degrees of freedom is thus 2.051831.

4. Calculate the LSD from the ANOVA using Formula 1.

```
LSD<-t*sqrt((2*0.000888)/4)

LSD

[1] 0.04323475
```

**Ex. 4: Second LSD Calculation**

Good, now let's go through the second LSD calculation using the SEM, which we calculated in the previous exercise. First, let us set the variable SEM equal to the SEM, then look at the answer R returns.

```
SEM <-sqrt(0.000888/4)

SEM

[1] 0.01489966
```

Great, now we can carry out the second LSD calculation by entering

```
LSD2<- t * SEM * sqrt(2)

LSD2

[1] 0.04323475
```

You can easily see that the same result is obtained using both equations.

**Ex. 4: Interpretation of LSD**

Again, the LSD tells us the minimum mean difference that we should consider between individuals in the sample population that we are analyzing. In this example, the minimum mean difference that we would consider among the red clover cultivars is 0.432 tons per hectare.

**Comparing RCBD Means**

To perform an LSD comparison in R with the red clover data that we've been working with, we first need to install the 'agricolae' package, if you have not done so before, use the use the **library** command to access the functions in the package.

Now we can use the **LSD.test** command. Let's set the output of the command equal to variable *out*

```
out<-LSD.test(data$Yield, data$Cultivar,27,0.000888, p.adj="bonferroni",group=TRUE)
```

The inputs in the parenthesis after the LDS.test function are as follows: data$Yield specifies the experimental unit (Yield), data$Cultivar indicates the treatment (Cultivar), 27 is the residual (or error) degrees of freedom from the ANOVA table, 0.000888 is the MSerror (also from the ANOVA table),p.adj="bonferroni" indicates that we are using the Bonferroni p-value correction method, and group=TRUE indicates that each treatment (cultivar) should be treated as a separate group (mean calculations should be done for each cultivar).

**Ex. 4: R Output**

Let's look at the output.

```
out

$statistics

Mean CV MSerror LSD
```

0.64345 4.63118 0.000888 0.07689099

$parameters

Df ntr bonferroni

27 10 3.649085

$mean

data$Yield std r LCL UCL Min Max

1 0.66175 0.05882956 4 0.6311784 0.6923216 0.607 0.737

2 0.61075 0.03508442 4 0.5801784 0.6413216 0.577 0.643

3 0.63425 0.01590335 4 0.6063784 0.6648216 0.620 0.657

4 0.57675 0.03398406 4 0.5461784 0.6073216 0.541 0.611

5 0.64775 0.04358421 4 0.6171784 0.6783216 0.601 0.705

6 0.63300 0.09809519 4 0.6024284 0.6635716 0.526 0.734

7 0.69925 0.05518076 4 0.6686784 0.7298216 0.634 0.766

8 0.67950 0.09113177 4 0.6489284 0.7100716 0.559 0.763

9 0.63700 0.06243397 4 0.6064284 0.6675716 0.580 0.725

10 0.65450 0.04219400 4 0.6239284 0.6850716 0.628 0.717

$comparison

NULL

$groups

trt means m

1 7 0.69925 a

2 8 0.67950 ab

3 1 0.66175 ab

4 10 0.65450 ab

5 5 0.64775 abc

6 9 0.63700 abc

7 3 0.63425 abc

8 6 0.63300 abc

9 2 0.61075 bc

10 4 0.57675 c

**Ex. 4: Interpret the Results/Make a Decision**

At the bottom of the results, you'll find the mean data for each cultivar in the **$groups** table. In the column labeled M at the far right, we are given a new piece of information; the means of the 10 cultivars fall into 3 distinct groups based on the LSD. Means that have the same letters are not statistically different from one another; i.e. the difference between them is less than the LSD (0.0432 tons/ha). Cultivars 7, 8, and 1 clearly outyielded the others and should be the ones selected for advancement in the breeding program.

Cultivars in the ANOVA

Query 1.11.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Cultivar Yield Average

Query 1.11.6

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

# Blocking Efficiency

Blocking efficiency can be tested vs. CRD. Blocking is not always beneficial. When it is not necessary or is done inappropriately, blocking can actually reduce the precision of an experiment. Let us compare the results of using an RCBD to those which we would have obtained using a CRD for our sample experiment. First, we have the results using RCBD:

*Table 7 ANOVA table for corn planted at three populations in northwest Iowa.*

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean Square | Observed F | Observed F(5%) |
|---|---|---|---|---|---|
| **Treatment** | 2 | 24.81 | 12.405 | 25.7 | 6.94 |
| **Block** | 2 | 1.95 | 0.977 | n/a | n/a |
| **Error** | 4 | 1.93 | 0.484 | n/a | n/a |
| **Total** | 8 | 28.70 | n/a | n/a | n/a |

Blocking is a tradeoff; We reduce the error variance by blocking, but we also reduce the degrees of freedom that we use to determine our critical F-value.

## Calculating Blocking Efficiency

Whenever we block, therefore, we must ask ourselves the following question: "Will the increase in our F-value for treatment be

large enough to offset the increase in the critical F-value?" The relative efficiency of blocking for an RCBD experiment can be calculated as (Equation 12):

$$\text{Block Efficiency} = \frac{(n_B+1)(n_C+3)MS_{eC}}{(n_C+a)(n_B+3)MS_{eB}} \times 100 \tag{1.11.27}$$

$$\text{Equation 12} \tag{1.11.28}$$

**where:**
$MS_eC$= error mean square for CRD,
$MS_eB$= error mean square for RCBD,
$n_C$= error df for CRD,
$n_B$= error df for RCBD.

Now you might expect that we can just estimate $MS_{eC}$ by running the analysis without blocks in the model as we did in Chapter 5 on Categorical Data—Multivariate. However, this does not give a proper estimate of the CRD error mean square over all possible randomizations, but rather just for the one having each treatment in each block. Cochran and Cox (1957 Experimental Designs, 2nd Edition, p.112) prove that $MS_{eC}$ should be estimated as (Equation 13):

$$MS_{eC} = \frac{df_B(MS_B) + (df_T + df_E)MS_E}{df_B + df_T + df_E} \tag{1.11.29}$$

$$\text{Equation 13} \tag{1.11.30}$$

**where:**
$MS_eC$= error mean square for CRD,
$df_B, df_T, df_E$ = degree of freedom for blocks, treatments, and error in the RCBD ANOVA,
$MS_B, MS_E$ = mean squares for blocks and for error in the RCBD ANOVA.

## Calculating Error Mean Square for CRD

For our experiment, this is (using Equation 13):

$$MS_{eC} = \frac{2(0.997) + (2+4)0.484}{2+2+4} = 0.607 \tag{1.11.31}$$

This error mean square estimate for a CRD is somewhat less than the error mean square computed by just re-running the model without blocks, which is 168.6.

A value for Blocking Efficiency greater than 1.00 suggests that we gained efficiency in our experiment by blocking. The blocking efficiency for the sample experiment is (using Equation 12):

$$\text{Block Efficiency} = \frac{(4+1)(6+3)0.607}{(6+1)(4+3)0.484} \times 100 = 115.2 \tag{1.11.32}$$

Thus, our sample experiment was improved by blocking. The efficiency is about 15.2% greater than what it would have been in a CRD.

It has been said that one can "never lose by blocking." While this is not always the case, it is true that blocking will generally improve an experiment whenever a production gradient is recognized and blocks are appropriately arranged across that gradient.

# Summary

## Reason for Blocking

- To achieve more homogeneous conditions for experimental units.
- Allows better separation of treatment effects and error.

## RCBD Linear Model

- Includes term for Blocks.
- Error term is the Block * Treatment interaction.

## Analysis of Variance for RCBD

- Has sources for Treatment, Block, and Error.
- Degrees of freedom are (t-1), (b-1) and (t-1)(b-1), respectively.
- Sums of squares are computed in the same manner (as in CRD).
- Mean Squares are SS/df.
- F = MST/MSE

## Relative Efficiency of Blocking

- Can be compared with no blocking as in CRD.

**How to cite this chapter**: Harbur, M.L., K. Moore, R. Mowers, L. Merrick, and A. A. Mahama. 2023. Randomized Complete Block Design. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.11: Randomized Complete Block Design is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by M. L. Harbur; Ken Moore; Ron Mowers; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.12: Data Transformation

The analyses of variance (ANOVAs) with which you have worked extensively in the past 4 lessons share basic assumptions about the data being analyzed. The analysis is invalid if these assumptions are incorrect. Fortunately, these assumptions can be quickly tested. Finally, if the assumptions regarding a particular data set are found to be false, there are methods which can be followed to properly modify the data before conducting an analysis.

**Learning Objectives**

- Know the assumptions made in conducting the analysis of variance.
- Know how to test for heterogeneity of variances.
- Describe experimental situations which may produce heterogeneity of variances.
- Know how to transform data so that it meets the assumptions of the analysis of variance

## Assumptions of ANOVA

**There are 3 main assumptions in Analysis of Variance.** For an analysis of variance to be valid, all of the following assumptions must apply:

- The error terms are normally, independently, and randomly distributed.
- The variances are homogeneous and not correlated with the means of different samples (treatment levels).
- The main effects and interactions are additive.

Each of these assumptions is further discussed on the following pages.

### Normality, Independence and Random Distribution of Errors

Normality, or following a normal distribution, is often assumed of data sets, but is infrequently realized in practice. This is because the number of observations in the data set is often too few for the set to resemble a normal distribution curve. Fortunately, the analysis of variance is a robust procedure, and is rarely seriously affected by deviations from normality.Independence assumes that there is no relationship between the size of error of a treatment group and the experimental units (plots) to which it is allocated. The same treatment effect should be apparent, then, regardless of the experimental units to which the treatment is applied. In other words, the assumption of independence implies that the error associated with each level of treatment should reflect the natural variation in experimental units.

| A | | | |
|---|---|---|---|
| 1 | 3 | 2 | 4 |
| 4 | 1 | 3 | 3 |
| 2 | 2 | 1 | 4 |
| 3 | 4 | 2 | 1 |

| B | | | |
|---|---|---|---|
| 1 | 1 | 4 | 4 |
| 1 | 1 | 3 | 4 |
| 2 | 2 | 4 | 3 |
| 2 | 2 | 3 | 3 |

Study question 1 outline/map experimental layouts plans.

### Study Question 1: Experimental Treatments

#### Query 1.12.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Homogeneity of Variances

Error variances should be constant. Refer to any of the ANOVAs from past lessons, and you will note that the difference between all levels of a treatment is tested using one error term. This error term is pooled, or the average of the variances associated with each level of the treatment. The variances associated with the different treatment levels must be homogeneous if the pooled error can be correctly used to test their differences.

What happens if this assumption of homogeneity is violated? The result will be that the error term used to test the difference between treatments will be too large for comparing treatment levels with small variances and too small for comparing the treatment levels with large variances.

If data are determined to have heterogeneity of variances, then the researcher has two options. She or he may want to arrange the treatment levels into groups with similar variances and analyze each group separately. Alternatively, she or he may decide to transform the data, as will be discussed later in this lesson.

Constant variance implies independence of Means and Variances. When heterogeneity of variances occurs, the variance often does not vary at random with the treatment level. Instead, the variance is dependent on the mean so that the two are correlated. The variance may increase with the treatment mean so that larger means have larger variances.

The assumption of constant variance is more likely to be violated in data sets in which the means vary widely or with certain types of data that we will learn about later in this lesson.

## Linear Additive Model

We assume the linear additive model holds. The linear additive model, as you have already seen, can be used to describe an experimental model. The additive model suggests that each treatment effect is constant across different levels of other treatments or blocks. Differences between observations receiving the same treatment level or combination of treatments are, therefore, entirely the result of variation between experimental units.

For example, the linear additive model for the RCBD is shown in Equation 1.

$$Y_{ij} = \mu + \beta_i + T_j + \beta T_{ij} \qquad \text{(Equation 1)}$$

**where:**

- $Y_{ij}$ = response observed for the $ij^{th}$ experimental unit,
- $\mu$ = grand mean,
- $B$ = block effect,
- $T$ = treatment effect,
- $BT$ = block error x treatment interaction.

## Additive Treatments

If we were comparing different rates of fertilizer, then the assumption of additivity is that the difference in the crop yield produced by fertilizer rates will be the same, regardless of the experimental unit. The data from our experiment might resemble that in the "additive" columns of the following table:

*Table 1 Additive and Multiplicative Effects*

| n/a | Additive | | Multiplicative | |
|---|---|---|---|---|
| Treatment | 1 | 2 | 1 | 2 |
| 1 | 10 | 20 | 10 | 20 |
| 2 | 30 | 40 | 30 | 60 |
| 3 | 50 | 60 | 50 | 100 |

This assumption is occasionally violated when there is an interaction between the plots (experimental units) themselves and the treatment. For example, the difference between the nitrogen rates may be much more profound in experimental units with adequate phosphorous and potassium fertility than in experimental units with poorer soils. Our data may then turn out to be multiplicative, as shown in the two right columns (Table 1).

Data which do not conform to the additivity assumption may be transformed using the log transformation discussed later in this lesson.

# Testing Heterogeneity

## Bartlett's Test

**Bartlett's Test is used to test for constant variance.** We saw in an earlier example that for two treatments, we could test the equality of variances with the ratio of larger sample variance to smaller, eg. $\mathrm{F}=\frac{S_1^2}{S_2^2}$ if $S_{12}$ is larger. We reject the null hypothesis of equal variances if the F value is high. But how do we do a similar test if there are three or more treatments? We can use a test developed by Maurice Bartlett.

A data set which is suspected of not meeting the homogeneity of variance assumption can be tested using Bartlett's Test for Homogeneity of Variance. We will go through the steps of Bartlett's test to illustrate the underlying formulas but will later use R in solving problems to test variance homogeneity.

## Example

*Table 2 Illustration of Bartlett's test for variance homogeneity.*

| Experiment | SS (error) | df | $S^2$ | $ln(S^2)$ |
|---|---|---|---|---|
| 1 | 157.8 | 18 | 8.77 | 2.171 |
| 2 | 134.5 | 18 | 7.47 | 2.011 |
| 3 | 325.5 | 18 | 18.08 | 2.895 |
| 4 | 308.4 | 18 | 17.13 | 2.841 |
| 5 | 111.3 | 18 | 6.18 | 1.822 |
| 6 | 214.2 | 18 | 11.90 | 2.477 |
| $S_P^2$ | n/a | n/a | 11.59 | n/a |
| $\sum lnS_i^2$ | n/a | n/a | n/a | 14.217 |

First, the error sum of squares associated with each mean is recorded in the table above.

Second, the degrees of freedom are listed, and the error sums of squares are divided by the degrees of freedom associated with each level of treatment. So far, this process is very similar to setting up an ANOVA table. In fact, $S^2$ for each treatment level is equivalent to the mean square error for that level.

Third, the ln, or natural log, of each variance is calculated and recorded in the fourth column.

The appropriate means and sums are calculated at the bottom of the table.

## Chi-Square Value

Finally, the values from the table are appropriately inserted into the following formula, where a chi-square value is calculated.

### Bartlett's Test for Homogeneity of Variance

$$\chi^2 = \frac{(df)\left[nlnS_p^2 - \sum lnS_i^2\right]}{1 + \frac{n+1}{3n(df)}} \tag{1.12.1}$$

Equation 2 (1.12.2)

**where:**
$df$ = degrees of freedom associated with each treatment,
$n$ = number of treatments,
$S_p^2$ = pooled error variance,
$S_i^2$ = error variance for treatment.

This value is used to test the null hypothesis $H_0$: all treatment variances are equal. The observed chi-square value is compared with the critical value, using the degrees of freedom associated with the number of treatments, or n-1.

## Interpretation

The observed chi-square value is then interpreted in the following manner:

- If P>0.01, then we accept that there is no significant difference between the variances and, therefore, no need to transform the data.
- If P<0.001, then we reject the null hypothesis and determine that the variances are indeed different. The data must therefore be transformed or analyzed in another appropriate manner.

- If 0.01>P>0.001, then we must try to determine whether there is any theoretical basis for why the data would not meet the heterogeneity of variance requirement; otherwise, we should not transform.

If necessary, the data can usually be transformed using one of the methods below.

## Study Question 2: Bartlett's Test for Homogeneity of Variance

$$\chi^2 = \frac{(18)\left[6(2.450) - 14.217\right]}{1 + \frac{6+1}{3(6)(18)}} = 8.521 \tag{1.12.3}$$

Equation 3 (1.12.4)

### Query 1.12.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Exercise 1 (part 1)

**Ex. 1: Evaluating the Variances**

Let us try this using R and a different set of data.

The data in the QM-mod12-ex1data.xls worksheet are from a growth chamber experiment in which a number of treatments were applied to eastern gamagrass seed in an attempt to break dormancy and thereby increase germination percentage. The treatments consisted of:

- control (no treatment)
- wet chilling for 2 weeks
- wet chilling for 4 weeks
- wet chilling for 2 weeks with scarification, and
- wet chilling for 4 weeks with scarification

The experimental design was a CRD with five replications.

The data are expressed as percentages. This should lead us to suspect that there may be a potential for problems with the assumptions for the ANOVA.

In this exercise, we will use Excel to create a table of treatment means and variances which we will examine to see if they conform to the homogeneity assumption.

**Ex. 1: Creating a Pivot Table**

1. Using the mouse, select all the data in the worksheet. Be sure to select the top row, which contains the data labels.

2. Select **PivotTable** from the **Insert** menu.

3. A dialog box will open, and the data you have selected will automatically appear in the box next to **Select a table or range**.

4. Click the circle next to **New Worksheet** and then **OK**.

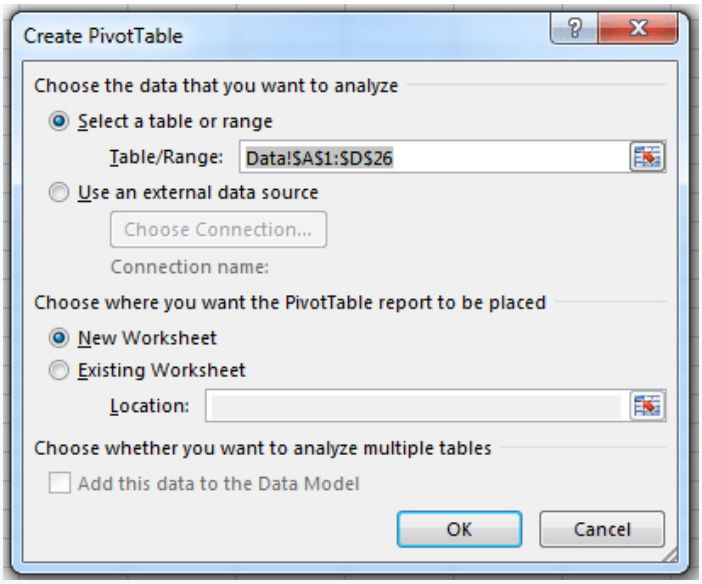



Figure 1 Creating a Pivot Table

5. The next screen will show an empty table with a panel on the right side titled **PivotTable Field List**, which is used to format the table.

6. Drag the **Treatment** field into the **Row Labels** box in the panel.

7. Drag the **Germination** button into the Values box in the panel.

8. Click or Double-click on the **Sum of Germination** field and select **Value Field Settings**... from the popup menu that appears.

9. Select **Average** from the list of options that appear, then click **OK** to calculate and display the five treatment means.

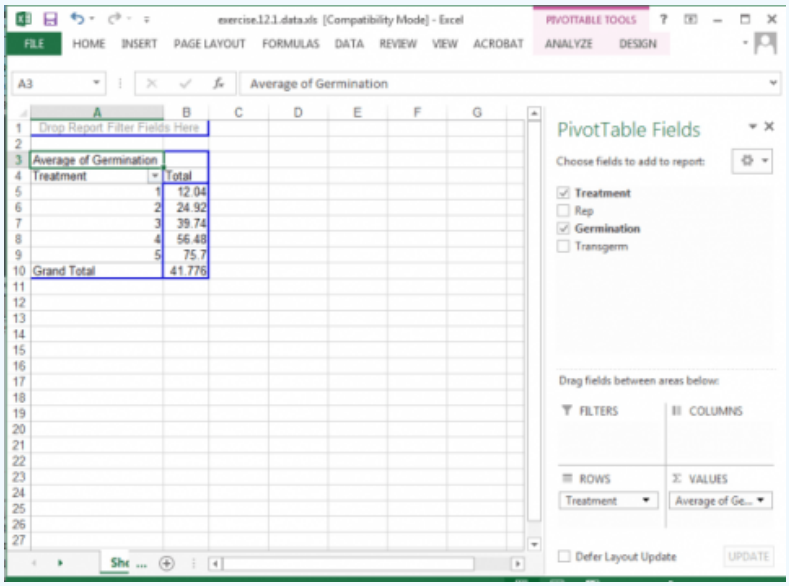



Figure 2 Displaying treatment means.

10. Once again, drag the **Germination** button into the **Values** box in the panel.

11. Click on the **Sum of Germination** field and select **Value Field Settings**... from the popup menu that appears.

12. This time, select **Var** from the list of options that appear, then click **OK** to calculate and display the five treatment variances.

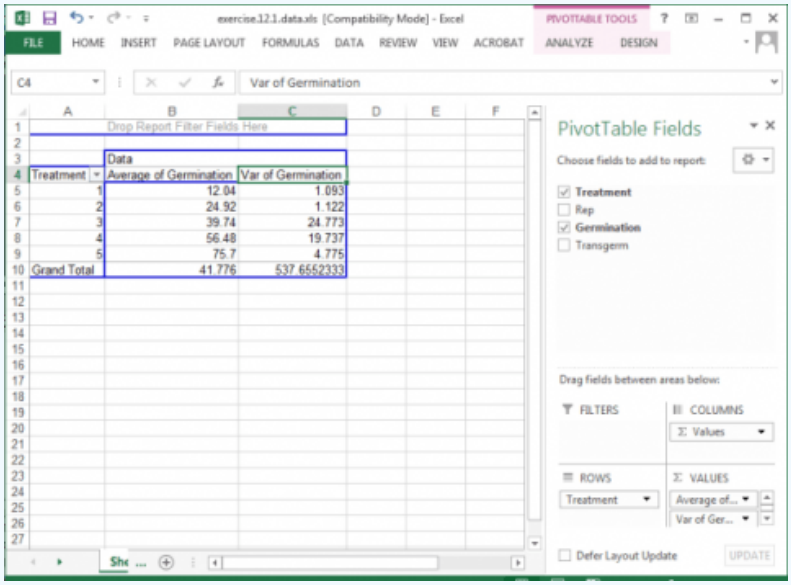



Figure 3 Completed Pivot Table

Your completed pivot table should look like Figure 3.

**Ex. 1: Examining Homogeneity Assumption**

In looking at the variances associated with the five treatments, there appears to be a potential problem with the homogeneity assumption (Figure 4). The variance of Treatment 1 is 1.093 compared with 24.773 for Treatment 3. The ratio of these two variances is 22.67, which is relatively large. This situation should lead us to examine more thoroughly the assumption of homogeneity.

| | | Data | |
|---|---|---|---|
| Treatment | | Average of Germination | Var of Germination |
| 1 | | 12.04 | 1.093 |
| 2 | | 24.92 | 1.122 |
| 3 | | 39.74 | 24.773 |
| 4 | | 56.48 | 19.737 |
| 5 | | 75.7 | 4.775 |
| Grand Total | | 41.776 | 537.6552333 |

Figure 4 Examining the two variances.

Note: Later in this unit, we will learn how to “transform” data so that the variances are more equal. Sometimes it will help us to decide what transformation function to use if we look at how both variance and standard deviation vary with treatment means.

To calculate the standard deviation for treatment means, simply choose “Std Dev” instead of “Var” in Step 12.

## Study Question 3: Variance

### Query 1.12.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Exercise 1 (part 2)

**Ex. 1: Plotting the Variance against Means**

A good way to visualize what is going on with the data is to plot the variances against their associated means.

In this exercise, we will use Excel to create a plot of the means and variances. The most expedient way to accomplish this is to create a Pivot Chart using the Pivot Table we just made in Exercise 1.

Create an XY graph of the means and variances using the following steps:

1. Using your mouse, place the cursor anywhere within in the Pivot Table you created for the last exercise.
2. At the very top of the Excel window to the right, a menu item labeled PivotTable Tools will appear. Select this menu item and then click on the PivotChart icon that appears.
3. Select the first chart style that appears in the Column template.
4. The graph that is created should look something like (Figure 5).

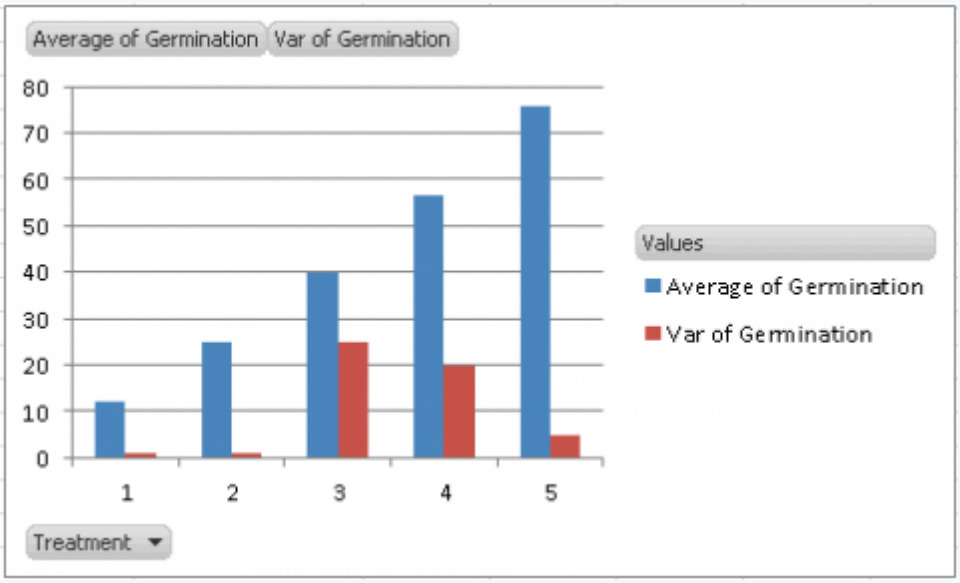



Figure 5 Excel Pivot Chart

Our assumption for the ANOVA is that the variances of the five treatments are equal. By looking at the graph, we see that the variances for Treatments 3 and 4 are considerably higher than the others, which should lead us to explore our assumption further.

Note that you can produce the same style chart if you have produced a pivot table with standard deviations instead of variances.

Sometimes, it is easier to visualize the relationship between the variance and treatment means if we plot them in a scatter plot. This is a little more tricky than the PivotChart you just made, but it allows us to better see how variance increases with treatment means. In addition, to choose which transformation function to use (see Table 12.5), sometimes you need to compare the relationship between variance and mean to that between standard deviation and mean to see which relationship is more linear.

**Ex. 1: Creating a Scatterplot**

To produce a scatterplot with variances on the Y-axis and treatment means on the X-axis:

1. Select the “Average” and “Var” columns from your PivotChart (Figure 6). Do not select the treatment column or the Grand Total row. Copy and paste these below your original chart.

| | Data | |
|---|---|---|
| Treatment | Average of Germination | Var of Germination |
| 1 | 12.04 | 1.093 |
| 2 | 24.92 | 1.122 |
| 3 | 39.74 | 24.773 |
| 4 | 56.48 | 19.737 |
| 5 | 75.7 | 4.775 |
| Grand Total | 41.776 | 537.6552333 |

| Average of Germination | Var of Germination |
|---|---|
| 12.04 | 1.093 |
| 24.92 | 1.122 |
| 39.74 | 24.773 |
| 56.48 | 19.737 |
| 75.7 | 4.775 |

Figure 6 Duplicate the “Average” and “Var” columns

2. Highlight your new table. Then choose Insert from the menu bar, then Scatter, then the first chart style (Figure 7).

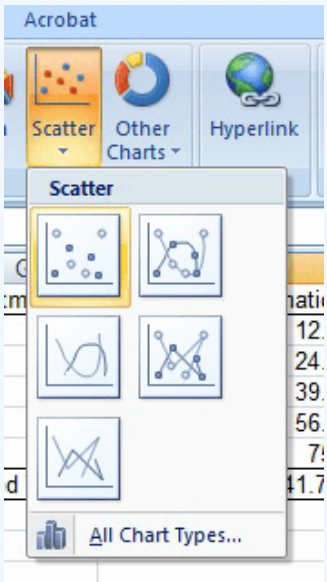



Figure 7 Select the Scatter chart style

Your chart should look like Figure 8. Again, the treatment means are on the X-axis, and the variances are on the Y-axis.

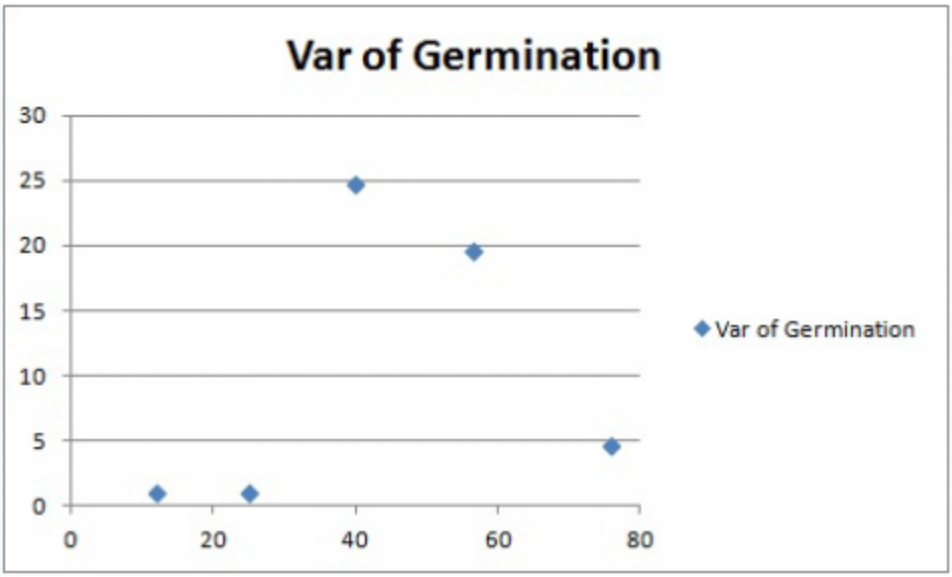



Figure 8 Var of Germination chart

In this case, there does not seem to be a linear relationship between variance and treatment mean. However, this chart again shows that the variances associated with two of the treatments are many times greater than those associated with the other. Data transformation may be required so that the variances meet the assumptions of the Analysis of Variance.

The same method can be used to create a plot with Standard Deviation on the Y-axis and Treatment means on the X-axis. Simply create a pivot table with treatment means and standard deviation, and then follow the two steps above.

## Notes to educators and students

There is some prerequisite knowledge required in order to understand everything in this lesson. Participants should be familiar with designing linear models and conducting ANOVAs and also understand how the least significant differences (LSDs) are calculated.

### Exercise 2: Test for Homogeneity of Variances – Using R

**Ex. 2: Introduction**

When conducting an analysis of data, there are numerous assumptions that are made about the data. Sadly, in the real world, these assumptions are often violated, leading to poor interpretation of the results. In this activity, we will explore ways to deal with these assumption violations when it comes to performing an analysis of variance (ANOVA).

**R Code Used In This Exercise**

**Packages:**

- plyr
- reshape2
- ggplot2
- agricolae

**Code:**

- setwd()
- detach()
- names()

- log()
- read.csv()
- as.factor()
- cbind()
- sqrt()
- head()
- aov()
- melt()
- asin()
- str()
- summary()
- qplot()
- LSD.test()
- attach()
- aggregate()
- bartlett.test()

**Ex. 2: Review of 3 ANOVA's Main Assumptions**

**Review – These Are The 3 Main Assumptions Of The Anova**

**The error terms are normally, independently, and randomly distributed**.
This means that the error terms follow a normal distribution, although this is difficult in smaller sample sizes. There should also be independence between the size of the error of a treatment group and the experimental units to which it is allocated.

**The variances are homogenous and not correlated with the means of different levels.**
The error variances should be constant. Remember that the difference between all levels of a treatment is tested using one error term (pooled), and if this assumption is violated, then the error term used to test the differences between treatments will be too large for comparing treatment levels with small variances and too small for comparing treatment levels with large variances.

**The main effects and interactions are additive.**
This means that we assume that the linear additive model we are using to analyze the data holds true. Sometimes this assumption is violated because there is an interaction between the plots and the treatments, but this data can be transformed using the log transformation.

In this lesson, we will explore how to determine if your data should be transformed and, if so, what kind of transformation is appropriate.

**Ex. 2: Exercise Introduction**

You are a student, and you wish to study the effects of 5 treatments on breaking gamagrass seed dormancy. The treatments were control (no treatment), wet chilling for 2 weeks, wet chilling for 4 weeks, wet chilling for 2 weeks plus scarification, and wet chilling for 4 weeks plus scarification. You decide to test the treatments on a single, common variety using a completely random design with 5 replications for each of the treatments. The data for these treatments are expressed as the percent germination in the file Set1.csv. In this activity, we will use R to check the second and third assumptions of the ANOVA by creating a table of treatment means and variances which we will examine to see if the homogeneity of the variances assumption was violated. Then, if the assumption is violated, we will explore ways to transform the data for a more accurate analysis.

Our first step is to start visualizing our data, and one way to do this is to create a table of means and variances of the means. First, you will need to read in the data and ensure that it was read in properly as well as make sure that the "Treatment" column is considered to be a factor.

**Ex. 2: Read the Data**

Read in the data and check the structure:

germdata<-read.csv("Set1.csv", header=T)

head(germdata)

Treatment Rep Germination

1 1 1 12.7

2 1 2 11.3

3 1 3 13.4

4 1 4 10.8

5 1 5 12.0

6 1 1 25.9

str(germdata)

'data.frame': 25 obs. of 3 variables:

$ Treatment : int 1 1 1 1 1 2 2 2 2 2 ...

$ Rep : int 1 2 3 4 5 1 2 3 4 5 ...

$ Germination : num 12.7 11.3 13.4 10.8 12 25.9 24.5 26.2 23.9 24.1...

Since "Treatment" is considered an integer, not a factor, we need to change that with the as.factor function:

germdata$Treatment<-as.factor(germdata$Treatment)

Treatment<- as.factor(germdata$Treatment)

detach(germdata)

attach(germdata)

str(germdata)

'data.frame': 25 obs. of 3 variables:

$ Treatment : Factor w/ 5 levels "1", "2", "3", "4",..: 1 1 1 1 1 2 2 2 2 2 ...

$ Rep : int 1 2 3 4 5 1 2 3 4 5 ...

$ Germination : num 12.7 11.3 13.4 10.8 12 25.9 24.5 26.2 23.9 24.1 ...

## Ex. 2: Visualize the Data

To help visualize the data, we will start by making a table of the treatment means and variances, and standard deviations. This can be done with the following code, which analyzes the data set using the specified function (mean, variance, and standard deviation, respectively) based on a specified variable in the data set (here, it is Treatment). The code also renames the new column that is made to better reflect which function was used. Otherwise, without renaming, the column head would just read "Germination," regardless of what you have actually calculated. This first code calculates the treatment means:

means <- aggregate(germdata["Germination"], by=germdata["Treatment"], FUN=mean)

names(means)[names(means)++"Germination"] <- "GermMean"

means

Treatment GermMean

1 1 12.04

2 2 24.92

3 3 39.74

4 4 56.48

5 5 75.70

The same code format can be used to calculate the treatment variances and standard deviations just by changing the “FUN=” and the name of the new column.

var <- aggregate(germdata[“Germination”], by=germdata[“Treatment”], FUN=var)

names(var)[names(var)==”Germination”] <- “GermVar”

var

Treatment GermVar

1 1 1.093

2 2 1.122

3 3 24.773

4 4 19.737

5 5 4.775

stdev <- aggregate(germdata[“Germination”], by=germdata[“Treatment”], FUN=sd)

names(stdev)[names(stdev)==”Germination”] <- “Stdev”

stdev

Treatment Stdev

1 1 1.093

2 2 1.122

3 3 24.773

4 4 19.737

5 5 4.775

## Ex. 2: Combine into a Single Table

Now we can combine these into a single table for easy reference.

Summary<-cbind(means, var$GermVar, std$Stdev)

Summary

Treatment GermVar

1 1 1.093

2 2 1.122

3 3 24.773

4 4 19.737

5 5 4.775

You can leave the column names as is, but there is a simple way of renaming them using the “plyr” package:

library(plyr)

Warning message:

package ‘plyr’ was built under R version 3.1.1

Summary<-rename(Summary, c(“var$GermVar”=”GermVar”, “stdev$Stdev”=”Stdev”))

Summary

Treatment GermMean GermVar Stdev

1 1 12.04 1.093 1.045466

2 2 24.92 1.122 1.059245

3 3 39.74 24.773 4.977248

4 4 56.48 19.737 4.442634

5 5 75.70 4.775 2.185177

## Ex. 2: Graph Means and Variance

For this next step, we want to graph the treatment means and the treatment variances on the same graph. Normally in R, making a bar graph is very straightforward, but in this case, we have two variables to plot at the same time, so the code becomes a little more complex. Before we can do anything, however, we have to manipulate the data set into a format that R can read in order to give us the output that we want. You will need to utilize the package “reshape2”.

library(reshape2)

From here, you can use the following code to melt your data from the wide format to the long format. We only want means and variances for this next bit, so we are only including those variables in our new melted data set.

Summary.melt <- melt(data = Summary, id.vars=c(‘Treatment’)

+ measure.vars=c(‘GermMean’,’GermVar’))

Summary

Treatment variable value

1 1 GermMean 12.040

2 2 GermMean 24.920

3 3 GermMean 39.740

4 4 GermMean 56.480

5 5 GermMean 75.700

6 1 GermVar 1.093

7 2 GermVar 1.122

8 3 GermVar 24.773

9 4 GermVar 19.737

10 5 GermVar 4.775

## Ex. 2: Reshape Data

Reshaping data can take some practice; the melt function is not always intuitive, and it can take a while to manipulate your data set into the desired format. There is another way of achieving the same results with another package called “reshapeGUI”. After you install the package, open the library and then use the code reshapeGUI() to open a more friendly graphic user interface. Here, you can actually see what is happening to your data as you select your data set and then choose the ID and Measure variables in the ‘melt’ tab.

library(reshapeGUI)

Loading required package: gwidgets

Attaching package: ‘gwidgets’

The following object is masked from ‘package:plyr’:

id

Loading required package: gwidgetRGtk2

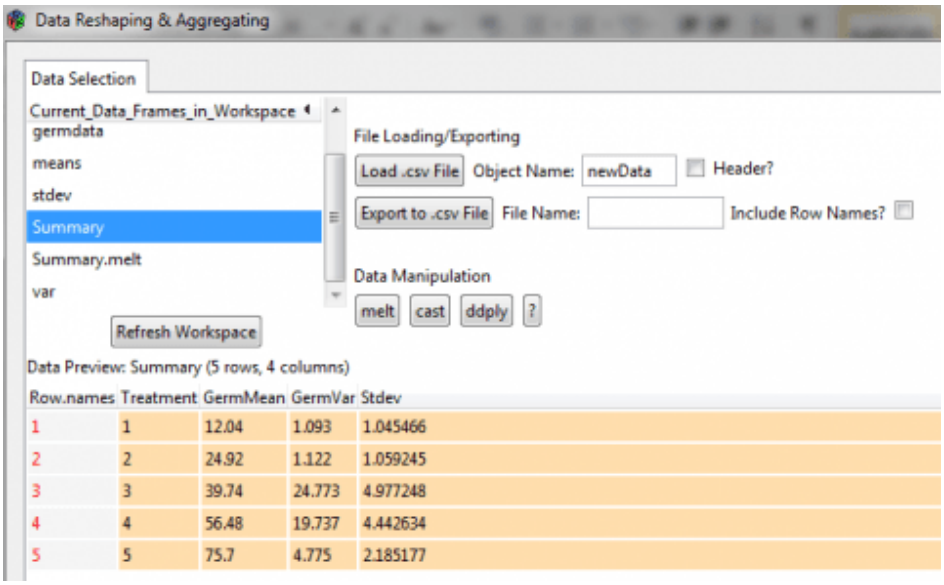
Loading required package: RGtk2

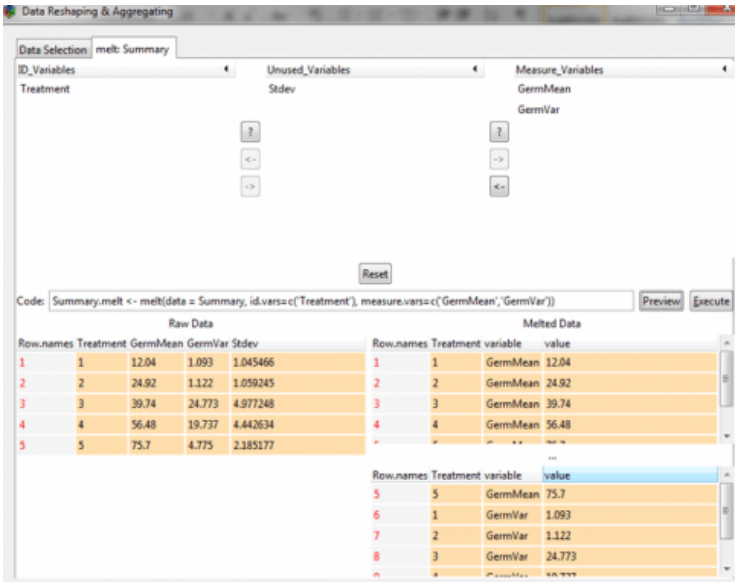
Warning message:

package ‘reshapeGUI’ was built under R version 3.1.1

## Ex. 2: Reshape GUI

reshapeGUI




## Ex. 2: Preview Result

You can hit the preview button to see what your data will look like. The Execute button will actually run the code you have generated, but if you want to save the code itself, you can copy and paste it into a script editor.

Now we can start plotting our means and variances. The package “ggplot2” has a lot of versatility in visualizing data, so we will use this package:

library(reshapeGUI)

Now use this code to create a bar plot that includes both the treatment means and the treatment variances and also includes some nice colors and a graph title:

means.var.varplot <- qplot(x-Treatment, y=value, fill=variable, data=Summary.melt, geom=”bar”, stat=”identity”, position=”dodge”, main=”Barplot of Means and Variance of germination for 5 treatments”)

means.var.barplot

This gives us this graph (Figure 9).

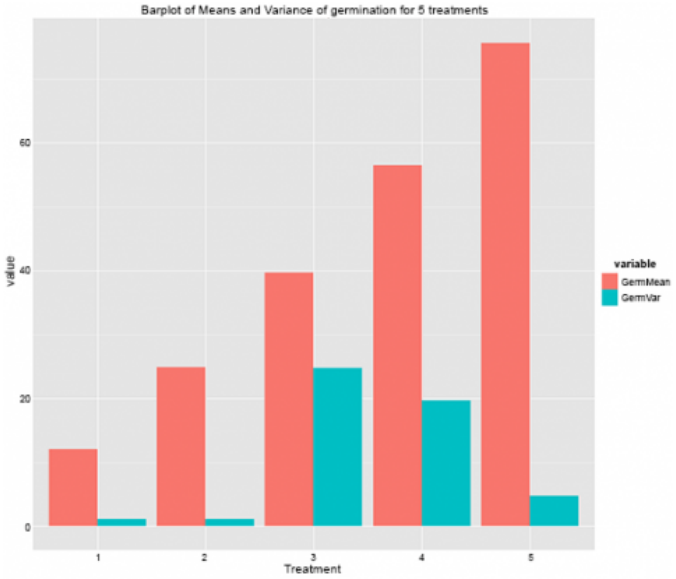



Figure 9 Barplot of Means and Variance for Germination for 5 Treatments

Remember how one of the assumptions of the ANOVA is equal variances across the treatments? Looking at this graph here, we can clearly see that this is not true of this data set. The variances for treatments 3 and 4 are much higher than the other treatments, and this should prompt you to explore this further.

## Ex. 2: Scatter Plot to Visualize Data

Another way to visualize this data is to plot the variances against the means in a scatterplot which can be done using ggplot2 with the following code:

```
means.var.scatterplot <- qplot(data=Summary, x=GermMean, y=GermVar,
+ main="Scatterplot of germination means against germination variances") +
+ geom_point(size = 5)

means.var.scatterplot
```

And the graph looks like this (Figure 10):

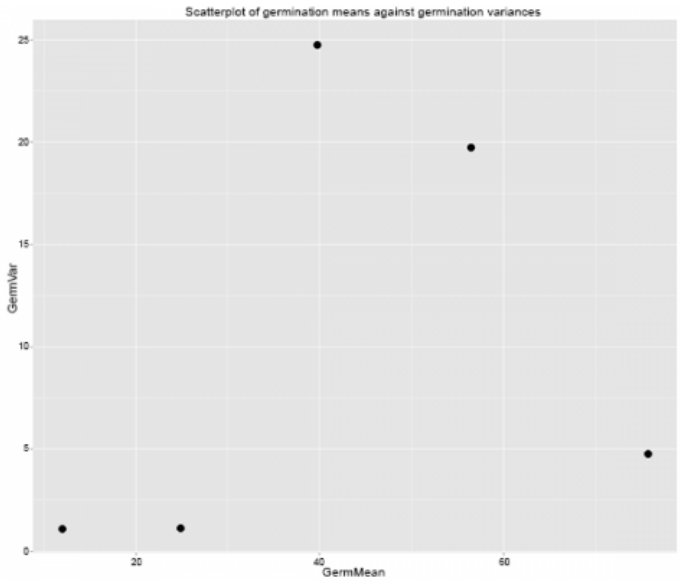



Figure 10 Scatterplot of Germination Means Against Germination Variances

## Ex. 2: Bartlett's Test

As we saw with the bar graph, treatments 3 and 4 clearly have greater variances than the other treatments in the data set. However, this is not enough evidence to prompt data transformation, and a different test should be run to be sure.

Running the Bartlett's test for homogeneity of variances:

This test lets us test to see if the variances between treatments are significantly different enough to warrant a data transformation. It uses the formula in Equation 2 above.

## Ex. 2: Conclusions

This tests the null hypothesis that all treatment variances are equal, and the calculated chi-square value is compared to the critical value, which is based on the degrees of freedom. When it comes to interpreting the test, there are several possible outcomes based on the p-value:

- If the p-value is >0.01, then we accept the null hypothesis and don't transform the data
- If P<0.001, the null is rejected, and we need to transform the data
- If 0.01>P>0.001, then you must try to figure out whether there is any reason why the data would not meet the requirement of heterogeneity of variances; otherwise, we don't transform the data.

Performing Bartlett's test in R is very easy and can be done with one line of code

bartlett.test(Germination~Treatment, germdata)

Bartlett test of homogeneity of variances

data: Germination by Treatment

Bartlett's K-squared = 13.4583, df = 4, p-value = 0.009241

Looking at these results, our p-value falls between 0.01 and 0.001, which is the range in which you must decide if there is a theoretical basis for transforming the data or not. Since our experimental data is expressed in percentages (which often leads to unequal variances), we can conclude that the variances are not homogenous, and we do have a reason for transforming this data set.

# Data Transformation

**There are 3 main transformations to achieve more constant variance.** Data which fail to conform to assumptions regarding independence of variance and mean, independence of standard deviation and mean, or additivity may be transformed using one of many transformations. The most common three methods are shown in Table 3.

*Table 3 Transformations types.*

| Condition | Transformation | Types of Data |
|---|---|---|
| Standard deviation proportional to mean | Natural log, ln(y) | growth data and counts with a wide range of values |
| Variance proportional to mean | Square root, sqrt(y) | small whole number data, counts of rare events |
| S2 from binomial data | Arcsine, arcsin*sqrt(y) | percentages, proportions |

Different conditions have different mathematical relationships, which can be used to produce constant variance after transformation. We will examine some of these further.

## The Natural Log Transformation

The natural log transformation is appropriately used to transform data where the standard deviation of treatments is roughly proportional to the means of those treatments. It is also appropriate where effects appear to be multiplicative rather than additive. This kind of transformation is most often necessary when dealing with growth data, where differences between plants become more obvious as their mean size increases.

For example, review the data set in Table 4.

*Table 4 Untransformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 10.35 | 16.80 | 28.86 | 38.84 | 36.85 |
| 2 | 12.14 | 16.68 | 26.68 | 39.14 | 49.04 |
| 3 | 12.04 | 19.75 | 29.07 | 35.39 | 57.72 |
| 4 | 10.26 | 16.18 | 28.35 | 32.36 | 50.47 |
| 5 | 9.14 | 18.68 | 25.06 | 38.55 | 40.07 |
| Mean | 10.79 | 17.62 | 27.60 | 36.46 | 46.83 |
| Std.Dev. | 1.28 | 1.52 | 1.70 | 2.72 | 8.40 |

### Data before Transformation

It is clear from looking at the data set that the standard deviation differs among treatments. A plot of the standard deviations by treatment means reveals that the standard deviation tends to increase with treatment mean. (Figure 11)

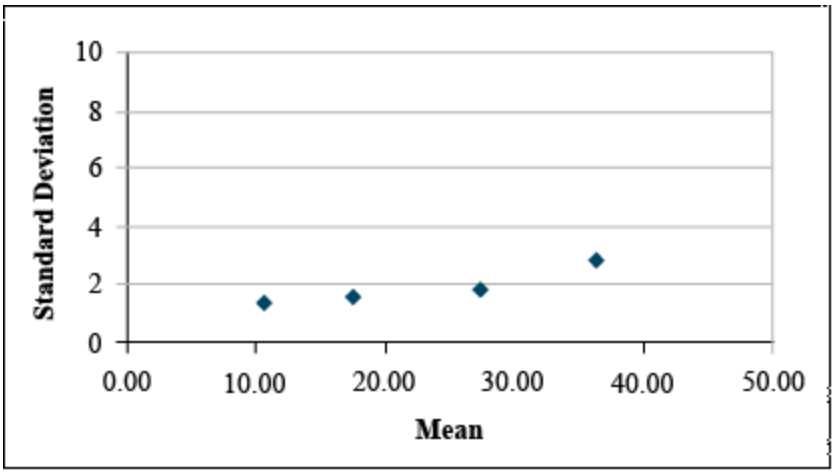



Figure 11 Untransformed data: The standard deviation of the individual treatments increases with the mean of the treatment.

## Transform using Natural Log

If Bartlett's test confirms that these differences in standard deviation are significant, then we should transform these data using the natural log transformation. This gives us the data set and plot below.

*Table 5 Transformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 2.34 | 2.82 | 3.36 | 3.61 | 3.61 |
| 2 | 2.50 | 2.81 | 3.28 | 3.67 | 3.89 |
| 3 | 2.49 | 2.98 | 3.37 | 3.57 | 4.06 |
| 4 | 2.33 | 2.78 | 3.34 | 3.48 | 3.92 |
| 5 | 2.21 | 2.93 | 3.22 | 3.65 | 3.69 |
| Mean | 2.37 | 2.87 | 3.32 | 3.59 | 3.83 |
| Std.Dev. | 0.12 | 0.09 | 0.06 | 0.08 | 0.18 |

The transformation, thus, can be used to reduce the apparent relationship between the standard deviation and the means of the treatments (Figure 12).

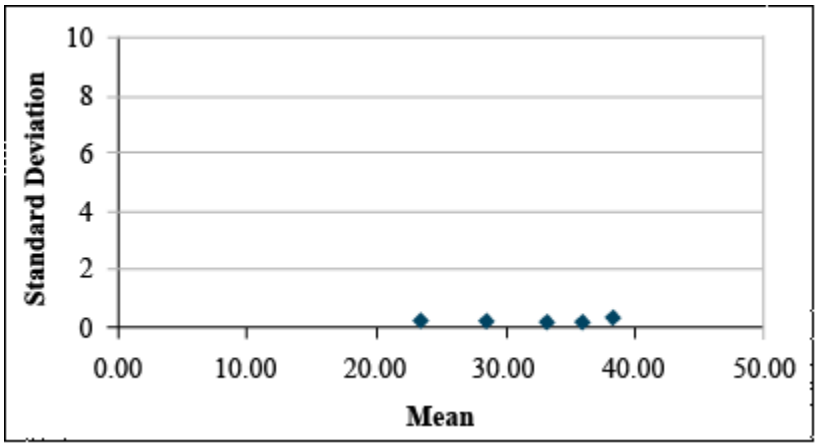



Figure 12 Transformed data: The standard deviation of the individual treatments is much better behaved after transformation.

# The Square Root Transformation

The square root transformation is most often necessary in dealing with counts of rare events — those with a very low probability of occurring in any one individual. Such counts tend to follow a Poisson distribution instead of a normal distribution. The mean tends to increase as the number of observations that are greater than zero increases. The result of such a distribution is that the variance tends to be proportional to the mean. This distribution might arise if we were dealing with insect counts, as your book suggests, or perhaps if we were sampling weed populations in a particularly “clean” field.

For example, review the data set in Table 6.

*Table 6 Untransformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 7.37 | 27.71 | 47.01 | 58.04 | 72.32 |
| 2 | 12.80 | 37.12 | 58.95 | 71.43 | 86.79 |
| 3 | 11.50 | 34.86 | 56.09 | 68.22 | 83.32 |
| 4 | 11.16 | 34.27 | 55.35 | 67.39 | 82.42 |
| 5 | 8.46 | 29.60 | 49.42 | 60.74 | 75.24 |
| Mean | 10.26 | 32.71 | 53.36 | 65.17 | 80.02 |
| Std.Dev. | 5.09 | 15.29 | 24.63 | 30.96 | 36.18 |

## Poisson Distribution

The Poisson Distribution has the equation form:

$$\mathrm{P(Y{=}k)} = \frac{e^{-\mu}\mu^{k}}{k!} \tag{1.12.5}$$

Equation 4 (1.12.6)

**where:**
$k$ = Poisson random variable or number of occurences, e.g., 0, 1, 2, ...
$e$ = Euler’s number,
$\mu$ = expected value,
! = factorial function,
$k!$ = k × k – 1 × k – 2 ... × 1.

For example: 4! = 4 × 3 × 2 × 1

Figure 13 shows a Poisson distribution curve.

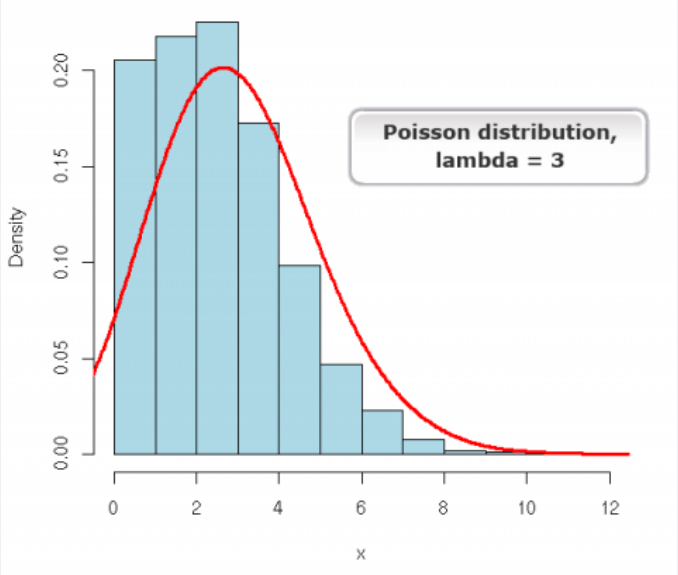



Figure 13 Poisson Distribution

## Data before Transformation

It is clear from looking at the data set that the variances differ greatly among treatments. A plot of the variances by treatment means reveals that variance tends to increase with treatment mean (Figure 14).

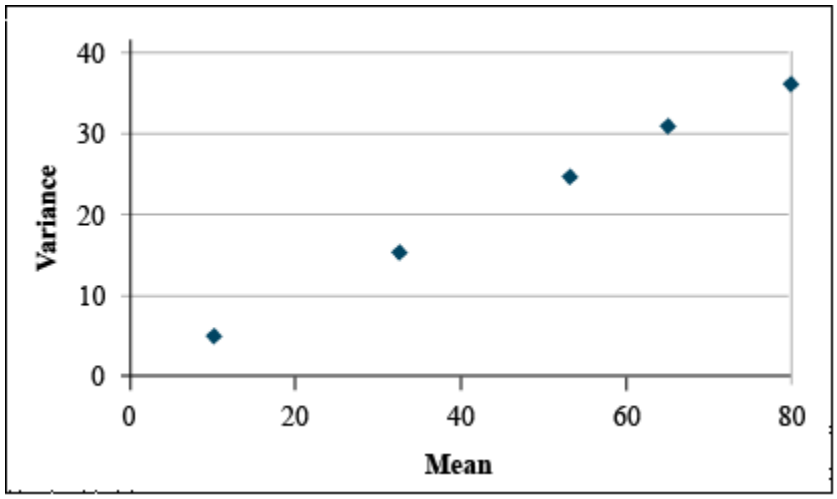



Figure 14 Untransformed data: the variance of the treatments seems to be proportional to the mean of the treatments.

## Transform using Square Root

If Bartlett's test confirms that these differences in variance are significant, then we should transform these data using the square root transformation. This gives us the data set and plot at below (Table 7 and Figure 15).

It is easy to see that the transformation has reduced the relationship between variance and mean.

*Table 7 Transformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 2.71 | 5.26 | 6.86 | 7.62 | 8.50 |
| 2 | 3.58 | 6.09 | 7.68 | 8.45 | 9.32 |

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 3 | 3.39 | 5.90 | 7.49 | 8.26 | 9.13 |
| 4 | 3.34 | 5.85 | 7.44 | 8.21 | 9.08 |
| 5 | 2.91 | 5.44 | 7.03 | 7.79 | 8.67 |
| Mean | 3.19 | 5.71 | 7.30 | 8.07 | 8.94 |
| Std.Dev | 0.13 | 0.12 | 0.12 | 0.12 | 0.11 |

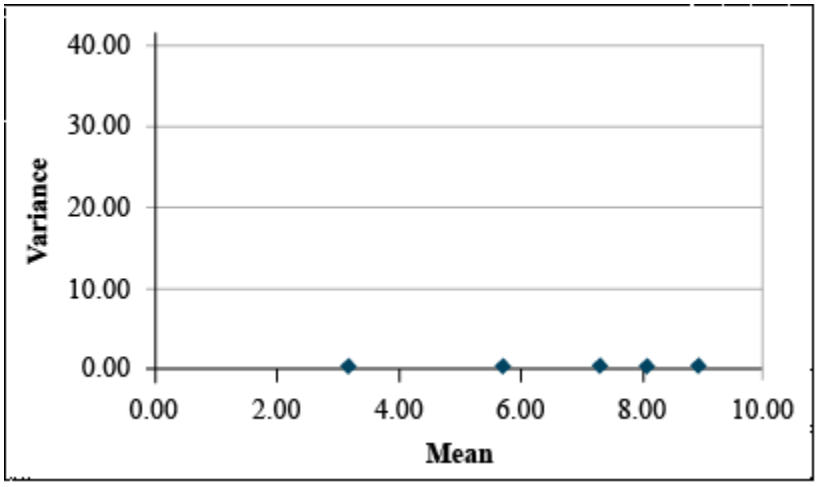



Figure 15 Transformed data: the variance of the treatments now have no apparent relationship to the mean.

## The Arcsine Transformation

The arcsine transformation is most often necessary on data expressed as a percentage or proportion. These data are most likely to conform to a binomial distribution, where the variance tends to be greater near the center of the distribution.

For example, observe the following data set (Table 8)

*Table 8 Untransformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 38.75 | 68.20 | 59.63 | 59.39 | 83.93 |
| 2 | 30.64 | 34.68 | 50.96 | 72.17 | 83.65 |
| 3 | 43.42 | 51.45 | 49.60 | 73.44 | 93.05 |
| 4 | 52.41 | 62.91 | 57.15 | 77.48 | 85.44 |
| 5 | 51.73 | 44.03 | 58.91 | 73.74 | 84.82 |
| Mean | 43.39 | 52.25 | 55.25 | 71.24 | 86.18 |
| Variance | 83.72 | 186.18 | 21.62 | 47.85 | 15.24 |

### Data before Transformation

It is clear from looking at the data set that the standard deviation differs among treatments. A plot of the standard deviations by treatment means reveals that the standard deviation tends to increase with treatment mean. (Figure 16)

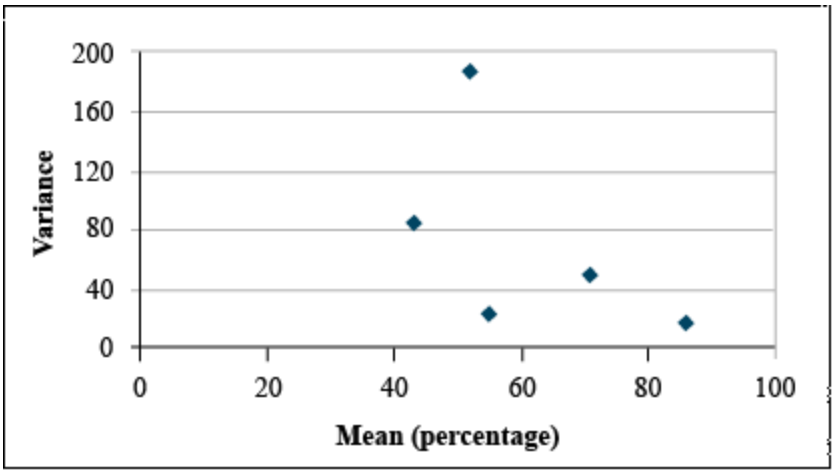



Figure 16 Untransformed data: the variance tends to fall away from the center of the distribution.

## Transform using Arcsine

If the Bartlett's test confirms that these differences in variance are significant, then we should transform these data using the arcsine transformation. This gives us the data set and plot at right (Table 9 and Figure 17).

*Table 9 Transformed data*

| Block | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 38.50 | 55.67 | 50.55 | 50.41 | 66.37 |
| 2 | 33.61 | 36.08 | 45.55 | 58.16 | 66.15 |
| 3 | 41.22 | 45.83 | 44.77 | 58.98 | 74.71 |
| 4 | 46.38 | 52.48 | 49.11 | 61.67 | 67.57 |
| 5 | 45.99 | 41.57 | 50.13 | 59.17 | 67.07 |
| Mean | 41.14 | 46.33 | 48.02 | 57.68 | 68.37 |
| Variance | 28.66 | 63.26 | 7.18 | 18.23 | 12.86 |

It is easy to see that the transformation has reduced the relationship between variance and mean.

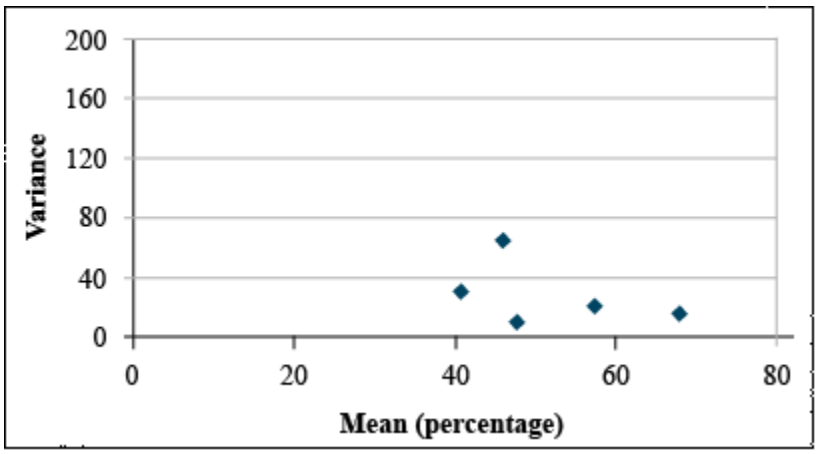



Figure 17 Transformed data: the previous variance trend has been removed, leaving a more uniform distribution.

## Exercise 3: Data Transformation using Angular

Generally, the most appropriate data transformation for percentage data is the angular or arcsine transformation. In this exercise, we will transform the germination data using the angular transformation in Excel.

The angular transformation is more complicated to calculate than appears on the surface. Although we will calculate it using a single equation, the calculation actually involves the following steps:

1. The percentage data are converted to proportions by dividing by 100.
2. The square root of the proportion is taken using the SQRT function.
3. The arcsine of the square root is taken using the ASIN function.
4. Convert from radians to degrees by multiplying by (180/PI). This last step is not required. Either degrees or radians give the same results. However, converting back to degrees makes the transformed data look more like the original data - only compressed.

**Ex. 3: Natural Log Transformation**

Once you have decided how to transform the data, this can be done quickly in Excel. Using the germination percentages found in the Exercise 12.1 data, follow these steps. In each case, you will need to create a new column heading. Since you will likely be importing these data into R, it is easiest to give them a one-word heading. Don't worry about originality — just select something simple that will be easy to use in R code.

**The Natural Log Transformation**

The natural log is sometimes referred to as "Ln", so change the title of column D to "LnGerm." Below that, type in "=ln(C2)". Select that cell, then double click on the tiny box in the right hand corner to copy down the rest of the data. Your table should look like the one below (Figure 18).

| Treatment | Rep | Germination | LnGerm |
|---|---|---|---|
| 1 | 1 | 12.7 | 2.5 |
| 1 | 2 | 11.3 | 2.4 |
| 1 | 3 | 13.4 | 2.6 |
| 1 | 4 | 10.8 | 2.4 |
| 1 | 5 | 12.0 | 2.5 |
| 2 | 1 | 25.9 | 3.3 |
| 2 | 2 | 24.5 | 3.2 |
| 2 | 3 | 26.2 | 3.3 |
| 2 | 4 | 23.9 | 3.2 |
| 2 | 5 | 24.1 | 3.2 |
| 3 | 1 | 40.3 | 3.7 |
| 3 | 2 | 45.0 | 3.8 |
| 3 | 3 | 36.8 | 3.6 |
| 3 | 4 | 32.9 | 3.5 |
| 3 | 5 | 43.7 | 3.8 |
| 4 | 1 | 52.3 | 4.0 |
| 4 | 2 | 51.4 | 3.9 |
| 4 | 3 | 57.4 | 4.1 |
| 4 | 4 | 60.8 | 4.1 |
| 4 | 5 | 60.5 | 4.1 |
| 5 | 1 | 72.3 | 4.3 |
| 5 | 2 | 76.2 | 4.3 |
| 5 | 3 | 77.1 | 4.3 |
| 5 | 4 | 77.9 | 4.4 |
| 5 | 5 | 75.0 | 4.3 |

Figure 18 Natural Log Transformation Table

**Ex. 3: Square Root Transformation**

**The Square Root Transformation**

Change the title of column E to "SqRtGerm" (real original, see?). Below that, type in "=sqrt(C2)". Select that cell, then double click on the tiny box in the right hand corner to copy down the rest of the data. Your table should look like this (Figure 19):

| Treatment | Rep | Germination | LnGerm | SqRtGerm |
|---|---|---|---|---|
| 1 | 1 | 12.7 | 2.5 | 3.563705936 |
| 1 | 2 | 11.3 | 2.4 | 3.361547263 |
| 1 | 3 | 13.4 | 2.6 | 3.660601044 |
| 1 | 4 | 10.8 | 2.4 | 3.286335345 |
| 1 | 5 | 12.0 | 2.5 | 3.464101615 |
| 2 | 1 | 25.9 | 3.3 | 5.08920426 |
| 2 | 2 | 24.5 | 3.2 | 4.949747468 |
| 2 | 3 | 26.2 | 3.3 | 5.118593557 |
| 2 | 4 | 23.9 | 3.2 | 4.888762625 |
| 2 | 5 | 24.1 | 3.2 | 4.909175083 |
| 3 | 1 | 40.3 | 3.7 | 6.348228099 |
| 3 | 2 | 45.0 | 3.8 | 6.708203932 |
| 3 | 3 | 36.8 | 3.6 | 6.066300355 |
| 3 | 4 | 32.9 | 3.5 | 5.73585216 |
| 3 | 5 | 43.7 | 3.8 | 6.610597552 |
| 4 | 1 | 52.3 | 4.0 | 7.231873893 |
| 4 | 2 | 51.4 | 3.9 | 7.169379332 |
| 4 | 3 | 57.4 | 4.1 | 7.57627877 |
| 4 | 4 | 60.8 | 4.1 | 7.797435476 |
| 4 | 5 | 60.5 | 4.1 | 7.778174593 |
| 5 | 1 | 72.3 | 4.3 | 8.502940668 |
| 5 | 2 | 76.2 | 4.3 | 8.729261137 |
| 5 | 3 | 77.1 | 4.3 | 8.780660567 |
| 5 | 4 | 77.9 | 4.4 | 8.826097665 |
| 5 | 5 | 75.0 | 4.3 | 8.660254038 |

Figure 19 Square Root Transformation Table

**Ex. 3: Arc Sin (Angular) Transformation**

**The Arc Sin (Angular) Transformation**

The arc sin transformation is somehow legitimate statistical voodoo. There is a whole lot going on, converting percentages to proportions, taking the square root and arcsin, and converting from radians to degrees.

Change the title of column F to “AsinGerm. Below that, type in “=asin(sqrt(C2/100))*180/PI()”. Select that cell, then double-click on the tiny box in the right-hand corner to copy down the rest of the data. Your table should look like Figure 20, below.

| Treatment | Rep | Germination | LnGerm | SqRtGerm | ArcSinGerm |
|---|---|---|---|---|---|
| 1 | 1 | 12.7 | 2.5 | 3.563705936 | 20.8774683 |
| 1 | 2 | 11.3 | 2.4 | 3.361547263 | 19.64277155 |
| 1 | 3 | 13.4 | 2.6 | 3.660601044 | 21.4728379 |
| 1 | 4 | 10.8 | 2.4 | 3.286335345 | 19.18585757 |
| 1 | 5 | 12.0 | 2.5 | 3.464101615 | 20.26790106 |
| 2 | 1 | 25.9 | 3.3 | 5.08920426 | 30.59194658 |
| 2 | 2 | 24.5 | 3.2 | 4.949747468 | 29.66808513 |
| 2 | 3 | 26.2 | 3.3 | 5.118593557 | 30.78776004 |
| 2 | 4 | 23.9 | 3.2 | 4.888762625 | 29.2667483 |
| 2 | 5 | 24.1 | 3.2 | 4.909175083 | 29.40090459 |
| 3 | 1 | 40.3 | 3.7 | 6.348228099 | 39.40684385 |
| 3 | 2 | 45.0 | 3.8 | 6.708203932 | 42.13041476 |
| 3 | 3 | 36.8 | 3.6 | 6.066300355 | 37.34622945 |
| 3 | 4 | 32.9 | 3.5 | 5.73585216 | 35.0006141 |
| 3 | 5 | 43.7 | 3.8 | 6.610597552 | 41.38074591 |
| 4 | 1 | 52.3 | 4.0 | 7.231873893 | 46.31826812 |
| 4 | 2 | 51.4 | 3.9 | 7.169379332 | 45.80224576 |
| 4 | 3 | 57.4 | 4.1 | 7.57627877 | 49.25552069 |
| 4 | 4 | 60.8 | 4.1 | 7.797435476 | 51.23710086 |
| 4 | 5 | 60.5 | 4.1 | 7.778174593 | 51.06117612 |
| 5 | 1 | 72.3 | 4.3 | 8.502940668 | 58.24366812 |
| 5 | 2 | 76.2 | 4.3 | 8.729261137 | 60.80047406 |
| 5 | 3 | 77.1 | 4.3 | 8.780660567 | 61.40994579 |
| 5 | 4 | 77.9 | 4.4 | 8.826097665 | 61.95879848 |
| 5 | 5 | 75.0 | 4.3 | 8.660254038 | 60 |

Figure 20 Arc Sin Angular Transformation Table

**Ex. 3: Data Transformation using R**

So, what kind of data transformation should we perform? There are many different ways, and each type of transformation is appropriate in different circumstances. We will explore only 3 of them in this activity.

**The Natural Log Transformation**

This transformation is good for when the standard deviation of the treatments is more or less proportional to the means of the treatments and where the effects seem to be multiplicative instead of additive.

R code:

lnGerm<-log(Germination)

**The Square Root Transformation**

This is most often used when the data consists of counting rare events and tends to follow a Poisson distribution, not a normal distribution. In this instance, the variance tends to be proportional to the mean. R code:

sqrtGerm<-sqrt(Germination)

**The Arcsine (Angular) Transformation**

This is typically used on data that is expressed in percentages or proportions because they are more likely to have a binomial distribution where the variance tends to be greater at the center of the distribution. Calculating the arcsine transformation is a bit tricky and involves several steps:

- Percentage data is divided by 100
- Then the square root is taken
- Then the arcsine is taken
- The last step of converting from radians to degrees is optional and is done by multiplying by 180/pi

R code:

asingerm<-asin(sqrt(Germination/100))*180/pi

**Ex. 3: Combine All Transformations in One Table**

We can combine each of these transformations into a single table with the following code:

transform<-cbind(germdata[,1:3], lnGerm, sqrtGerm, asinGerm)

transform

Treatment Rep Germination lnGerm sqrtGerm asinGerm

1 1 1 12.7 2.541602 3.563706 20.87747

2 1 2 11.3 2.424803 3.361547 19.64277

3 1 3 13.4 2.595255 3.660601 21.47284

4 1 4 10.8 2.379546 3.286335 19.18586

5 1 5 12.0 2.484907 3.464102 20.26790

6 1 1 25.9 3.254243 5.089204 30.59195

7 1 2 24.5 3.198673 4.949747 29.66809

8 1 3 26.2 3.265759 5.118594 30.78776

9 1 4 23.9 3.173878 4.888763 29.26675

10 1 5 24.1 3.182212 4.909175 29.40090

In the previous part of the activity, Bartlett's test showed us that our variances were not homogenous and, therefore, we should transform the data. Since we have percentage data and our scatterplot of the means and variances shows that the variance tends to be greatest in the center of the distribution, the arcsine transformation is most appropriate for our data.

## Ex. 3: Bartlett's Test and ANOVA

Once we have transformed our data, we can run Bartlett's test again with the new data:

bartlett.test(asinGerm~Treatment, germdata)

Bartlett test of homogeneity of variances

data: asinGerm by Treatment

Bartlett's K-squared = 9.7311, df = 4, p-value = 0.04521

Since our p-value is now >0.01, it is safe to assume that the variances are homogenous and we can proceed to the ANOVA. Use this formula so we can calculate LSDs next. For some reason, using the anova(lm()) format doesn't work for calculating LSDs in R.

anovasin<-aov(asinGerm~Treatment, data=transform)

summary(anovaasin)

Df Sum Sq Mean Sq F value Pr(>F)

Treatment 4 4930 1232.4 331 <2e-16 ***

Residuals 20 74 3.7

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

According to the ANOVA, the chances of getting this F-value or a higher one are very small (P<0.0001), and we can conclude that there are one or more differences among the different treatments.

## Ex. 3: LSD and Conclusions

To take a closer look at the differences between the treatments, we can calculate the least significant difference and see which of the treatment means differ.

library(agricolae)

LSD.Treatment<- LSD.test(anovaasin, "Treatment")

LSD.Treatment

$Statistics

Mean CV MSerror LSD

39.70057 4.860259 3.723166 2.545616

$parameters

Df ntr t.value

20 5 2.085963

$means

asinGerm std r LCL UCL Min Max

1 20.28937 0.9196122 5 18.48934 22.08939 19.18586 21.47284

2 29.94309 0.7002702 5 28.14307 31.74311 29.26675 30.78776

3 39.05297 2.9304004 5 37.25295 40.85299 35.00061 42.13041

4 48.73486 2.5682770 5 46.93484 50.53488 45.80225 51.23710

5 60.48258 1.4479205 5 58.68255 62.28260 58.24367 61.95880

$comparison

NULL

$groups

trt means M

1 5 60.48258 a

2 4 48.73486 b

3 3 39.05297 c

4 2 29.94309 d

5 1 20.28937 e

Based on this output, we can see that the least significant difference is ~2.55 and each of our treatment means have been placed in their own group which means that each treatment is significantly different from each other. Treatment 5 has the best germination, while treatment 1 (the control) has the worst germination.

The last thing to be done is to transform the means back to their original scale. Since we used the arcsine function, this is the code to transform the data:

inverse<-(sin(asinmeans$asinGermMean*(pi/180)))^2*100

inverse

[1] 12.02436 24.91403 39.69486 56.50011 75.72583

## Review Questions

1. Why is it sometimes necessary to transform data?
2. How do you decide which method of transformation is appropriate?

# Summary

## Analysis of Variance Assumptions

- Errors are independent with normal distribution.
- Error variances are constant and independent of treatment means.
- The additive model is correct.

## Bartlett's Test

- Tests for heterogeneity of variance
- Accept $H_0$ : Variances are similar if the P-value is greater than 0.10 and reject if less than 0.01.

## Normal Quantile Plot

- Visual verification of normal distribution.

## Data Transformations

- Three transformations for constancy of variance.
- Natural log (ln) if standard deviation is proportional to mean.
- Square root if variance is proportional to mean.
- Arc Sine ($\sqrt{y}$) if percentage or binomial proportion data. (y = proportion.)

**How to cite this chapter**: Mowers, R., K. Moore, M.L. Harbur, L. Merrick, and A. A. Mahama. 2023. Data Transformation. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.12: Data Transformation is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Ken Moore; M. L. Harbur; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.13: Multiple Regression

## Observing Variables

In Chapter 7 on Linear Correlation, Regression, and Prediction, we discussed determining the correlation and possible regression relationships between an independent variable **X** and a dependent variable **Y**. Specifically, in regression, the discussion was based on how the change in one variable (**X**) produced an effect on another (**Y**). This is the essence of regression. But from experience, we know that often multiple causes interact to produce a certain result. For example, yield from a crop is based on the amount of water a plant has to use, the soil fertility of the field, the potential of the seed to produce a plant, pest and pathogen pressures, and numerous other factors. In this lesson, we'll explore how we can determine linear relationships between multiple independent variables and a single dependent variable.

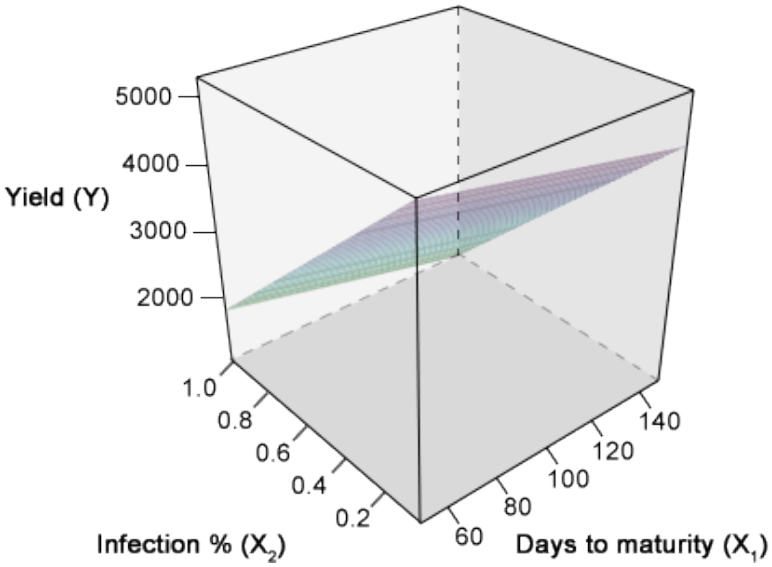



Figure 1 Barley yield as a function of rust infection and days to maturity.

## Exploring Multiple Variables

Multiple regression functionally relates several continuous independent variables to one dependent variable. In the above example, barley yield per plot (Y) is shown as a function of the percentage of plants in the plot affected by rust ($X_1$) and the days to maturity required by the cultivar grown in a particular plot ($X_2$). Yield is modeled as a linear combination of these two X variables in the response surface on the previous slide. Before we can relate the dependent variable Y to the independent X variables, we need to know the interrelationships between all of the variables. Multiple correlation and partial correlation provide measures of the linear relationship among the variables.

Separating the individual factor's effect on the whole result, such as the effect of rust infection or the number of days a particular cultivar requires to reach maturity, can be difficult and, at times, confounding. The objective of this chapter is to explain and illustrate the principles discussed in Chapter 7 on Linear Correlation, Regression and Prediction or correlating two variables or enumerating the effect of one variable on another, but now expanded to multiple variables.

### Learning Objectives

- To define correlation relationships among several variables
- To separate the individual relationships of multiple independent variables with a dependent variable
- To test the significance of multiple independent variables and to determine their usefulness in regression analysis
- To recognize some of the potential problems resulting from improper regression analysis

# Multiple Correlation and Regression

The correlation of multiple variables is similar to the correlation between two variables. The same assumptions apply; the sampled Y's should be independent and of equal variance. Error (variance) is associated with the Y's while X's have no error, or the error is small. But now, since there are multiple factors involved, the correlations are somewhat more complex, and interactions between the $X_i$ variables are expected. A note on notation: We now include a subscript with the "X" to indicate which independent variable to which we are referring. Three levels of correlation are used in determining the multi-faceted relationships; simple correlation, partial correlation, and total correlation.

## Simple Correlation

The simple correlation between one of the $X_i$'s and Y is computed for a simple correlation of X and Y. This calculation assumes a direct relationship between the particular $X_i$ and Y. It is also useful in stating the simple relationship between two $X_i$'s in the multiple correlation. When determining the significance of regression coefficients, the variable with the largest simple correlation with Y is usually the starting point. Some interaction among numerous X and Y variables is likely to occur. Because two $X_i$'s have large simple correlations with a resulting Y does not necessarily indicate that their relationships to Y are independent of each other. They may be measuring the same effect on Y. The number of hours of sunlight (cloud-free skies) and GDDs both have a good (simple) correlation to the rate of crop development. But their effects would not be additive. There would be a significant interaction between these two variables in describing crop development. The two variables measure two different factors, light and temperature. However, the amount of sunlight and temperature are generally highly correlated during the summer. So there would be a significant relationship between the two variables. These individual effects can be separated using partial correlation.

## Partial Correlation

Quantifying which continuous X variables are best correlated with the continuous Y-variable requires an understanding of the interactions between the $X_i$'s. To break down the interaction requires partial correlation coefficients. These use the simple correlation coefficients to explain the correlation of two variables with all other variables held constant. One such example is, "How much yield will result from nitrogen applications assuming the seasonal amount of rainfall will be average?" Here, rainfall is held constant, and the effect of nitrogen on yield would be used for a partial correlation. This relationship is given for the partial coefficient of determination between Y and $X_1$ where two X's are involved, as shown in Equation 1.

$$r_{YX_1X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{1 - r^2_{YX_2}}\sqrt{1 - r^2_{YX_1}}} \tag{1.13.1}$$

Equation 1 (1.13.2)

**where:**
$X_1, X_2$ = continuous variables X,
$Y$ = continuous variable Y,
$r$ = correlation coefficient.

To calculate the partial coefficient of determination between Y and $X_2$, just reverse the equation, i.e. use $r_{YX1}$ and vice versa. Don't panic! You will not be asked to hand-calculate this on the homework or exam. That is why we use R!

### Correlation Matrix

The value $r_{YX1}$ is the simple correlation between Y and $X_1$. The whole equation describes the correlation between Y and $X_1$ with $X_2$ held constant. The relationship between the $X_1$ and Y is displayed within the effect of the interaction. Partial correlations can be calculated for all variables involved. They can also be calculated for more than three variables, but the equation becomes more complex. Often, the total and partial correlations are calculated and displayed in a table with the individual $X_i$s and Y listed across the top and down the left side. The correlations for each variable pair are displayed at the intersection of the variables.

*Table 1 Correlation Matrix*

| n/a | $X_1$ | $X_2$ | Y |
|---|---|---|---|
| **$X_1$** | 1 | 0.462 | 0.693 |
| **$X_2$** | 0.462 | 1 | 0.354 |
| **Y** | 0.693 | 0.354 | 1 |

## Total Correlation

The combination of these partial effects leads to a multiple correlation coefficient, R, which states how related the Y is to the combined effects of the $X_i$'s. For $X_1$, $X_2$, and Y, the total correlation is determined once again using the simple correlations (Equation 2):

$$R^2_{Y \cdot X_1 X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2 r_{YX_1} r_{YX_2} r_{Y_1 X_2}}{1 - r^2_{X_1 X_2}} \tag{1.13.3}$$

$$\text{Equation 2} \tag{1.13.4}$$

In this equation,$r^2_{YX_1}$ and $r^2_{YX_2}$ are just the squares of partial correlation coefficients.

The calculations for 2 $X_i$'s are relatively straightforward, but for three or more variables, the calculations involve a large number of terms with different correlations among individual variables. Consequently, total correlation is calculated with computer programs, such as R.

Similar to the linear correlation coefficient, the total correlation coefficient, when squared, produces the multiple coefficient of determination, $R^2$. This value explains the proportion of the Y variation, which can be accounted for by a multiple regression relationship. The partial correlation coefficients squared produce the partial coefficients of determination, $r^2$, or that proportion of variance which can be described by one variable, while the partial coefficients will be used in testing individual regression coefficients for significance.

## Calculating the Correlation

Graphing data to visualize the correlation relationships in multiple dimensions is difficult. The graphing of data involving 2 $X_i$'s with Y is possible in 3-dimensional space. Using the variables mentioned, the regression equation would be a plane in the $X_1$, $X_2$, Y space (Figure 1). The partial regression coefficients for $X_1$ with Y and $X_2$ with Y in this space could be used to produce lines where the plane intersected a certain X value. For example, the following equation would produce a plane on a graph.

$$\mathrm{Y} = 2.4X_1 + 3.9X_2 - 7.1 \tag{1.13.5}$$

Setting X1 equal to 0 would reduce the above equation of a plane to a linear equation:

$$\mathrm{Y} = 3.9X_2 - 7.1 \tag{1.13.6}$$

Either $X_1$ or $X_2$ could be set to any value producing any number of different linear relationships in the plane. With more than two X values, graphing the relationship in 3 dimensions is not easily done. Instead of graphing, interpreting the data numerically and conceptually is the preferred method.

### Exercise 1: Correlation-Multiple Regression Analysis

**Ex. 1, Step 1**

R CODE FUNCTIONS

- cor
- cor.test

- install.packages
- library
- pcor

Multiple regression functionally relates several **continuous** independent variables (X), to one dependent variable, Y. For example, we could carry out multiple regression with yield as the dependent response variable (Y), $X_1$ as an independent variable indicating the amount of fertilizer applied and $X_2$ as an independent variable indicating the amount of water each plot received. In this example, we model yield as a linear combination of the amount of water and fertilizer applied to each plot in multiple regression. However, before we can relate Y to the other variables, we need to know the interrelationships of all the variables. Multiple correlation provide measures of the linear relationship among variables.

head(data)

cor(data$perc.inf, data$Yield)

cor(data$perc.inf, data$Yield)

R returns the simple correlation matrix.

dtm perc.inf yield
dtm 1.00000000 0.00352555 -0.2268896
perc.inf 0.00352555 1.00000000 -0.9475068
yield -0.22688955 -0.94750681 1.0000000

Great! Now we have the simple correlation matrix showing the correlations between RIL (Line), days to maturity (dtm), infection rate (perc.inf), and yield. The correlation matrix returned by R is constructed with the variables listed as both row and column headings. The top number at the intersection of a row and column is the correlation coefficient for those two variables. For example, the simple correlation between yield and perc.inf is -0.94750681.

**Ex. 1, Step 2**

First, calculate the p-value for the simple correlation of perc.inf and yield.

data<-read.csv("barley.csv", header = T)

R returns

Pearson's product-moment correlation

data: data$perc.inf and data$yield
t = -29.3362, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9644361 -0.9228353
sample estimates:
cor
-0.9475068

The p-value for the correlation between **perc.inf** and **Yield** is $2.2*10^{-16}$, which is extremely low. This low p-value tells us that the correlation between the two variables (yield and perc.inf) is highly significant.

**Ex. 1, Step 3**

Now, let's calculate the p-value for the correlation of **dtm** and **yield**.

cor.test(data$dtm, data$yield)

R returns

Pearson's product-moment correlation

data: data$dtm and data$yield
t = -2.3062, df = 98, p-value = 0.0232
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.40524772 -0.03189274
sample estimates:
cor
-0.2268896

The p-value, though not as low as that which was calculated for the correlation between **perc.inf** and **Yield**, is still significant at =0.05. Thus, the correlation between **dtm** and **yield** is also significant. Now let's see if there is a significant p-value for the correlation of perc.inf and DTM (the two X variables).

**Ex. 1, Step 4**

Calculate the p-value for the correlation between DTM and perc.inf.

cor.test(data$dtm, data$perc.inf)

R returns

Pearson's product-moment correlation

data: data$dtm and data$perc.inf
t = 0.0349, df = 98, p-value = 0.9722
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.41930262 -0.19980053
sample estimates:
cor
0.00352555

*Based on the extremely high p-value, the correlation between perc.inf and DTM is not significant.*

**Ex. 1, Step 5**

Interpret the results:

Which of the variables are the most correlated? Which will contribute the most to the final regression of **yield** on **dtm** and infection rate **perc.inf**? The first question can be answered by looking at the simple correlation matrix that we created in step 5. **perc.inf** and **yield** have a simple correlation of -0.94750681, and dtm and yield have a simple correlation of -0.22688955. The correlation of **dtm** and **yield** has a smaller absolute magnitude. Thus, infection rate (**perc.inf**) will contribute the most to the regression equation when we calculate it.

Before we construct a regression model for yield, we need to analyze how days to maturity (dtm) interact with infection rate (perc.inf) in the multiple regression. Despite the simple correlation between dtm and perc.inf being not statistically significant, calculating the partial correlation between these two variables may help explain a possible relationship between them. Simple correlations are the basis for calculating the additional correlation relationships.

**Ex. 1, Step 6**

Now, let's calculate the partial correlation matrix for the 3 variables. To do this, we'll first need to get the package 'ppcor'.

Install.packages('ppcor')

library (ppcor)

ppcor(data)

R returns

$estimate

dtm perc.inf yield
dtm 1.0000000 -0.6790490 -0.6991729
perc.inf -0.6790490 1.0000000 -0.9720637
yield -0.6991729 -0.9720637 1.0000000

$p.value

dtm perc.inf yield
dtm 0.000000e+00 8.210329e-20 5.887334e-22
perc.inf 8.210329e-20 0.000000e+00 0.000000e+00
yield 5.887334e-22 0.000000e+00 0.000000e+00

$statistic

dtm perc.inf yield
dtm 0.000000 -9.110368 -9.631483
perc.inf -9.110368 0.000000 -40.788323
yield -9.631483 -40.788323 0.000000

Note: Using the pcorr function, we obtain test statistics (t) and p-values without having to use any other function, such as cor.test. The R output **$estimate** gives the partial correlations, where one of the 3 variables is held constant as a partial variable. For example, the partial correlation of **yield** and **perc.inf** is -0.9720637; **dtm** is held constant as a partial variable for this correlation. The p-value for this correlation, or the probability of the correlation equal to zero, is so small that R returns a p-value of 0. From this incredibly small p-value, we would conclude there is a significant correlation between **yield** and organic **perc.inf**. The partial correlation coefficient between **yield** and **dtm** is -0.6991729, and the probability of this correlation being equal to zero is only 5.887334e-22. Thus, we would conclude that dtm is very much correlated with the yield, at least under these disease conditions.

Did the partial correlation follow the simple correlation in magnitude? The partial correlation of **dtm** with **yield** (with **perc.inf** held constant) was -0.6991729, while that of **perc.inf** with **yield** (dtm held constant) is -0.9720637. The squared values of the partial coefficients of determination are used in calculating the contribution of each variable to the regression analysis. These values are calculated as in equation 2 from above for the simplest case of multiple regression, where there are two X's and one Y. More complex equations result from equations with more than two X variables.

1. Set your working directory to the folder containing the data file barley.csv
2. Read the file into the R data frame, calling it data.

   data<-read.csv("barley.csv", header=T)

3. Check the head of the data to make sure it was read in correctly.
4. Calculate the correlation between the fusarium infection rate (perc.inf) and barley yield.
5. Calculate the correlation between DTM and yield.
6. Install the package 'ppcor'.
7. Load the package.
8. Calculate the partial correlation between **yield**, **dtm**, and **perc.inf**.

# Multiple Regression

## Relationships Among Multiple Variables

Multiple regression determines the nature of relationships among multiple variables. The resulting Y is based on the effect of several X's (Equation 3). How much of an effect each has must be quantified to determine the equation (below). The degree of effect each X has on the Y is related through partial regression coefficients. The b-value estimate of each regression coefficient

can be determined by solving simultaneous equations. Usually, computer programs determine these coefficients from the data supplied.

$$\mathrm{Y} = a + b_1 x_1 + b_2 x_2 + ... + b_i x_i + \epsilon \tag{1.13.7}$$

$$\text{Equation 3} \tag{1.13.8}$$

**where:**
$a$ = the $Y$ intercept,
$b$ = estimates of the true partial regression coefficients, β.

The $a$ is the Y-intercept, or Y estimate, when all of the X's are 0. The $< spanstyle =" font - size : 1em "> b's < /span >< spanstyle =" text - align : initial; font - size : 1em ">$ b's" title="b's" class="latex mathjax"> are estimates of the true partial regression coefficients $\beta$, the weighting of each variable's effect on the resulting Y. The b's are interpreted as the effect of a change in that X variable on Y, assuming the other X's are held constant. These can be tested for significance. The weighting of effects now will be based on regression techniques.

The simplest example of multiple linear regression is where two X's are used in the regression. The technique of estimating $b_1$ and $b_2$ minimizes the error sums of squares of the actual from estimated Y's. The variability of the data (Y's) can be partitioned into that caused by different X variables or into error.

## Example of Multiple Correlation and Regression

The simplest example of multiple correlation involves two X's. Calculations with more variables follow a similar method but become more complex. Computer programs have eased the computational problems. Proper analysis of the data and interpretation of analyses are still necessary and follow similar procedures.

The following two-variable research data were gathered relating the yield of inbred maize to the amount of nitrogen applied and the seasonal rainfall data (Table 2).

*Table 2 Yield response of inbred maize to nitrogen fertilizer and rainfall amounts.*

| Yield of Maize<br>bu/Ac | Fertilizer<br>lb N/Ac | Rainfall<br>in. |
|---|---|---|
| 50 | 5 | 5 |
| 57 | 10 | 10 |
| 60 | 12 | 15 |
| 62 | 18 | 20 |
| 63 | 25 | 25 |
| 65 | 30 | 25 |
| 68 | 36 | 30 |
| 70 | 40 | 30 |
| 69 | 45 | 25 |
| 66 | 48 | 30 |

## Review the Data

The first issue is to review how highly correlated the data are. Since visualization of multiple data is more difficult, numerical relationships must be emphasized. The first step is to examine the correlations among the variables. The simple correlations (calculated as in Chapter 7 on Linear Correlation, Regression and Prediction) may be computed for the three variables (see below).

## Simple Correlations

$$r_{YX_1} = 0.895 \quad (1.13.9)$$
$$r_{YX_2} = 0.944$$
$$r_{X_1X_2} = 0.905$$

Study Questions 1: X Variables

Query 1.13.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Partial Coefficients of Determination

All are highly correlated. But these simple correlations include the interactions among variables. To determine individual relationships, calculations of partial coefficients of determination are helpful (Equation 4).

$$r^2_{YX_1 \cdot X_2} = \frac{(r_{YX_1} - r_{YX_2} r_{X_1X_2})}{(1 - r^2_{X_1X_2})(1 - r^2_{X_1X_2})} = 0.081 \quad (1.13.10)$$

$$r^2_{YX_2 \cdot X_1} = 0.499 \quad (1.13.11)$$

$$r^2_{YX_2 \cdot X} = 0.170 \quad (1.13.12)$$

Equation 4 (1.13.13)

These values are the additional variability that can be explained by a variable, such as that by $X_1$, after the variability of $X_2$ alone has been accounted for. These values are used in computing the ANOVA for multiple regression. The partial correlations may be found by taking the square root of these partial coefficients of determination.

### Total Coefficients of Determination

The $R^2$-value is the total coefficient of determination, which combines the X's to describe how well their combined effects are associated with the Y's. This is determined as shown below (Equation 5).

$$R^2_{Y \cdot X_1X_2} = \frac{0.801 + 0.891 - 2(0.895)(0.944)(0.905)}{1 - 0.819} = 0.900 \quad (1.13.14)$$

Equation 5 (1.13.15)

The $R^2$ value is the proportion of variance in Y that is explained by the regression equation. This can be used to partition the variability in the ANOVA. The square root of this value gives the correlation of the X's with Y. It is obvious that the correlations are not additive. The simple correlations are all greater than 0.8, and the correlation between $X_1$ and $X_2$ is 0.905. This is where partial correlation comes into play.

## Partial Regression Coefficients

Before we can create an ANOVA and test the regression, we need a regression equation as determined by R. The estimate of the regression relationship is found to be in the below equation.

$$Y = 49.53 + 0.089x_1 + 0.515x_2 \tag{1.13.16}$$

The partial regression coefficients indicate that for the data gathered here, each additional pound of nitrogen applied per acre would produce an additional 0.089 bushels of maize per acre, and for each additional inch of rainfall, an additional 0.516 bushels per acre. An estimate of the yield is determined by entering the amount of nitrogen applied to the field and the amount of rainfall into the equation. The number produced is the regression equation estimate of the yield based on the data gathered.

The next issue is deciding if this equation is useful and explains the relationship in the gathered data. The sums of squares are partitioned in an ANOVA table, and the significance of the regression equation as a whole and the individual regression coefficient estimates are tested for significance in the next section.

Ex. 2: Multiple Regression and Anova Using R (1)

This exercise contains 26 steps (including this statement)

## Ex. 2: Multiple Regression and Anova Using R (1)

This exercise contains 26 steps (including this statement).

## Ex. 2: Multiple Regression and Anova Using R (2)

**R CODE FUNCTIONS**

- anova
- summary
- lm
- pf
- ppcor

$$Y = a + b_1X_1 + b_2X_2 + ... + b_iX_i + \epsilon \tag{1.13.17}$$

In equation 3 above, the a term is the Y-intercept, or the estimate of Y when all of the X's are 0. The *b* with each X is an estimate of the true partial regression coefficient *β* for that X variable, the weighting of each variable's effect on the resulting Y. The *b*'s are interpreted as the effect of a change in that X variable on Y, assuming the other X's are held constant. These coefficients can also be tested for significance. The weighting of effects will now be based on regression techniques.

The simplest example of multiple linear regression is where two X variables are used in the regression. The technique of estimating $b_1$ and $b_2$ via multiple regression minimizes the error sums of squares of the actual data from the estimated Y's. The variability in the data can be partitioned into that which is caused by different X variables or that which is caused by error.

In the file QM-Mod13-ex2.csv, we have yield data from one inbred maize line under all factorial combinations of 9 different levels of nitrogen treatment and 9 different levels of drought treatment. We'll use these data to investigate correlations between the variables, to do a multiple regression analysis, and to carry out an analysis of variance (ANOVA).

## Ex. 2: Multiple Regression and Anova Using R (3)

Read the dataset into R, and have a look at the structure of the data.

data<-read.csv("ex2_data.csv", header=T)

head(data)

R returns

drought N yield
1 -4 0 1886.792
2 -4 28.025 2590.756

3 -4 56.05 3743.000
4 -4 84.075 4910.937
5 -4 112.1 5656.499
6 -4 140.125 5689.165

The data contain entries for yield (kg/ha), level of nitrogen applied (kg/ha), and a “drought” score to indicate the level of drought stress applied (i.e. a level of -4 is the maximum drought stress applied, and a value of 4 is the minimum level of drought stress).

Note: Even though we have fixed treatments assigned to each test plot, we will run the analyses in this ALM as if they were random treatments (i.e. keeping the values for drought and N as numeric). This will allow us to investigate simple and partial correlations.

## Ex. 2: Multiple Regression and Anova (4)

### Simple Correlation

The correlation between X and Y, ($P_{XY}$), is calculated as the covariance of X and Y divided by the product of the standard deviations of X and Y, i.e., $P_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$
The first step is to review how highly correlated the data are. Since visualization of multiple data is more difficult, numerical relationships must be emphasized. Let’s examine the correlations among the variables. Calculate the simple correlations between the 3 variables by entering into the console window.

cor(data)

R returns the simple correlations matrix

drought N yield
drought 1.0000000 0.0000000 0.7364989
N 0.0000000 1.0000000 0.2147159
yield 0.7364989 0.2147159 1.0000000

## Ex. 2: Multiple Regression and Anova Using R (5)

### Partial Correlation

Taking $X_1$ to be nitrogen, $X_2$ to be drought, and Y to be yield, we can list the simple correlation variables.

Simple Correlations

$r_{YX_1} = 0.2147159$ $r_{YX_1} = 0.7364989$ $r_{Y_1X_2} = 0$
Both nitrogen and irrigation are correlated with yield, but these simple correlations include interactions among the variables. To determine individual relationships, calculations of partial correlation coefficients are helpful. Partial correlation coefficient values are the additional variability in the response variable that can be explained by an independent variable, such as that by $X_1$, after the variability of another independent variable, such as $X_2$, alone has been accounted for. These values are also used in computing the ANOVA for multiple regression.

We will now do a quick investigation of the partial correlations between the variables in the dataset. If you haven’t already, load the ‘ppcor’ package. Then, use the pcor command to obtain the matrix of partial correlations between all variables in the data set.

library(ppcor)

pcor(data)

## Ex. 2: Multiple Regression and Anova Using R (6)

R returns 3 matrices: a matrix with the partial correlation coefficient estimates ($estimate), a matrix with the test statistic for the estimate ($statistic), and a matrix for the p-value of the test statistic ($p.value).

$estimate

drought N yield
dtm 1.0000000 -0.239363 0.7540868
N -0.2393630 1.000000 0.3174210
yield 0.7540868 0.317421 1.0000000

$p.value

drought N yield
drought 0.000000e+00 0.029458897 3.658519e-24
N 2.945890e-02 0.000000000 3.113831e-03
yield 3.658519e-22 0.003113831 0.000000e+00

$statistic

drought N yield

dtm 0.000000 -2.177291 10.140333
perc.inf -2.177291 0.000000 2.956271
yield 10.140333 2.9562

## Ex. 2: Multiple Regression and Anova Using R (7)

Let us calculate the partial correlation coefficient for nitrogen on yield by hand, using Equation 1, to check the calculation returned by R in the *estimate* matrix.

$$r_{YX_1X_2} = \frac{(0.2147159 - 0.7364989 * 0)}{\sqrt{(1 - 0.7364989^2)(1 - 0)}} = \frac{0.2147159}{0.6764387} = 0.317421 \tag{1.13.18}$$

You can see that the value in the estimate matrix for the partial correlation coefficient between nitrogen and yield is identical to the value obtained by our hand calculation. Also, based on the p-value matrix, all of the partial-correlation estimates are statistically significant.

The test-statistic matrix contains values calculated from the standard normal distribution (with a mean of 0 and standard deviation of 1). The test statistic for the partial correlation of nitrogen on yield is 12.12521. We can check that this value is correct by calculating the p-value for this value from the standard normal distribution by entering

(1-pnorm(2.95621, mean=0, sd=1))*2

## Ex. 2: Multiple Regression and Anova Using R (8)

R returns

[1]0.00311834

The p-value for the partial correlation coefficient given in the R output from calculating the partial correlation coefficients is identical to that given in the R output using the pcor function.

The $R^2$ value is the total coefficient of determination, which combines the explanatory variables (X's) to describe how well their combined effects are associated with the response variable (Y). This is determined by the following equation:

$$R^2_{YX_1X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r^2_{X_1X_2}} \tag{1.13.19}$$

The $R^2$ is very useful for interpreting how well a regression model fits. Its value is the proportion of variance in Y that is explained by the regression equation. The closer to 1.0, the better the fit; a value of 1 would mean all of the data points fall on the regression

line. The square root of this value gives the correlation of the X's with Y. It is obvious that the correlations are not additive. This is where partial correlation comes into play.

## Ex. 2: Multiple Regression and Anova Using R (9)

The drawback of relying on the R2 value as a measure of fit for a model is that the value of $R^2$ increases with each additional term added to the regression model, regardless of how important the term is in predicting the value of the dependent variable. The Adjusted $R^2$ value (or $R^2_{Adj}$) is a way to correct for this modeling issue. The formula for Adjusted $R^2$ is:

$$R^2_{Adj} = 1 - \frac{(1 - R^2) - (N - 1)}{n - K - 1} \tag{1.13.20}$$

where R2 is the regression coefficient, n is the sample size, and k is the number of terms in the regression model. The R-squared value increases with each additional term added to the regression model, so taken by itself, can be misleading. The $R^2_{Adj}$ takes this into account and is used to balance the cost of adding more terms; i.e. it penalizes the $R^2$ for each additional term (k) in the model. The $R^2_{Adj}$ value is most important for comparing and selecting from a set of models with different numbers of regression terms. It is not of great concern until you are faced with choosing one model to describe a relationship over another. We'll carry out hand calculations for both $R^2$ and $R^2_{Adj}$ after we run the regression in R.

## Ex. 2: Multiple Regression and Anova Using R (10)

### Regression Model Significance

The initial test is to determine if the total regression equation is significant. As in linear regression, "does the regression relationship explain enough of the variability in the response variable to be significant?"

The testing of the regression equation partitions the total sum of squares using the total coefficient of determination, $R^2$. Note that this is not the same as the square of total correlation.

$$R^2_{Adj} = \frac{(\text{Regression SS})}{\text{Total SS}} \tag{1.13.21}$$

Initially, the null hypothesis being tested is that the whole regression relationship is not significantly different from 0.

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \\ H_a &: \beta_1 = \beta_2 \neq 0 \end{aligned} \tag{1.13.22}$$

## Ex. 2: Multiple Regression and Anova Using R (11)

The F-test for multiple linear regression uses the regression mean square to determine the amount of variability explained by the whole regression equation. If the regression mean square is significant at your specified level, the null hypothesis that all of the regression coefficients are equal to 0 is rejected. This F-test does not differentiate between coefficients; all are significant, or none are according to the test.

Individual regression coefficients ($b_1$, $b_2$, etc.) may be tested for significance. The simple coefficient of determination between each X and Y explains the sum of squares associated with each regression coefficient, including interactions with other X's. The partial coefficient of determination between each X and Y explains the additional variability without interaction. These can be tested with the residual error not explained by the regression model to test the significance of each X.

Each coefficient may also be tested with a t-test; R does this automatically when you run a multiple regression model using the **lm** function.

## Ex. 2: Multiple Regression and Anova Using R (12)

MULTI-LINEAR REGRESSION

Let's run a multiple regression analysis where yield is the response variable and drought and nitrogen are the explanatory variables. We will keep nitrogen and drought as numeric variables for this analysis but later will run the same analysis with these variables as factors.

In the console window, enter

summary(1m(data=data,yield~drought+N))

Let's go through this command from the inside out.

1. **data = data** indicates that we want to run the linear model with dataset 'data'
2. **yield~ drought + N** specifies that the regression equation we are analyzing is yield = 'the amount of nitrogen applied' + 'the amount of drought applied'
3. **lm** indicates to R that we want to run a linear regression model
4. **Summary** indicates that we want R to return all of the useful information from the regression analysis back to us.

## Ex. 2: Multiple Regression and Anova Using R (13)

R returns

Call:

lm(formula = yield ~ drought + N, data = data)

Residuals:

Min IQ Median 3Q Max

-4204.2 -1118.0 1.8 1251.6 3148.9

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7829.436 363.787 21.522 <2e-16 ***

drought 774.828 76.410 10.140 6.78e-16 ***

N 8.060 2.727 2.956 0.00412 ***

—

Signif. codes:

0 '***' 0.001 '**' 001 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1776 on 78 degrees of freedom

Multiple R-squared: 0.5885,

Adjusted R-squared: 0.578

F-statistic: 55.78 on 2 and 78 DF, p-value: 9.1e-16

## Ex. 2: Multiple Regression and Anova Using R (14)

The $R^2$ value is given at the bottom of the R output as 0.5885. This means that the model explains 58.85% of the variation in yield. Let's calculate the $R^2$ value by hand using the simple correlation coefficient matrix from above (Equation 2).

$$R^2_{Y \cdot X_1 X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1} r_{YX_2} r_{X_1 X_2}}{1 - r^2_{X_1 X_2}} \tag{1.13.23}$$

$$R^2_{Y \cdot X_1 X_2} = \frac{0.2147159^2 + 0.7364989^2 - 2(0.2147159)(0.7364989)(0)}{1 - 0} = 0.5885 \tag{1.13.24}$$

The value obtained for $R^2$ obtained by our hand calculation is identical to the value returned by R.

Now, let's calculate the $R^2_{Adj}$for the model by hand. Use the value of $R^2$ from the R output (0.5885).

$$R^2_{Adj} = 1 - \frac{(1 - R^2) - (N - 1)}{n - K - 1} \tag{1.13.25}$$

This is the same value for $R^2_{Adj}$ as given in the R output (under "Adjusted R-squared").

## Ex. 2: Multiple Regression and Anova Using R (15)

VARIABLE INTERACTION

Should we include a term in the linear model indicating the interaction between nitrogen and drought? Let's run the regression again, this time adding a variable accounting for the interaction between the two independent variables into the model. (i.e. the amounts of drought and nitrogen applied). The interaction variable is specified using a multiplication sign (*) with the explanatory variables that you are analyzing for interaction.

summary(1m(data=data,yield~N+drought+N*drought))

## Ex. 2: Multiple Regression and Anova Using R (16)

R returns

Call:

lm(formula = yield ~ N + drought + N * drought, data = data)

Residuals:

Min IQ Median 3Q Max

-3857.9 -1076.4 22.8 1244.8 3148.9

Coefficients:

Estimate Std. Error t value Pr(<|t|)

(Intercept) 7829.436 364.898 21.457 < 2e-16 ***

N 8.060 2.735 2.947 0.00424 **

drought 688.736 141.324 4.873 5.75e-06 ***

N:drought 0.768 1.059 0.725 0.47061

—

signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1781 on 77 degrees of freedom

Multiple R-squared: 0.5913

Adjusted R-squared: 0.5754

F-statistic: 37.14 on 3 and 77 DF, p-value: 5.986e-15

## Ex. 2: Multiple Regression and Anova Using R (17)

Compare the $R^2_{Adj}$ value and the F-statistic of the model, including the interaction, to the model not including the interaction. Which model fits the data better?

With interaction: $R^2_{Adj}$ = 0.5754, F = 37.14

Without interaction: $R^2_{Adj}$ = 0.578, F = 55.78

The model without the interaction between Nitrogen and Irrigation has a slightly better fit for these data than the model including the interaction. Also, the regression coefficient on the interaction term has a very high p-value, indicating that is not statistically significant. Save the model without the interaction as 'm1'.

m1<-1m(data=data,yield~N+drought)

## Ex. 2: Multiple Regression and Anova Using R (18)

Calculate the ANOVA table for the multiple regression models with and without the interaction between N and drought.

First carry out the ANOVA for the model without the interaction.

Enter into the console window

anova(lm(data=data,yield~drought+N))

R returns the ANOVA table

Now run the ANOVA with the linear model excluding the interaction term.

anova(lm(data=data,Yield~drought+N))

## Ex. 2: Multiple Regression and Anova Using R (19)

R returns the ANOVA table

Analysis of Variance Table

Response: yield

Df Sum Sq Mean Sq F value Pr(<F)

drought 1 324193187 324193187 102.8263 6.785e-16 ***

N 1 27554215 27554215 8.7395 0.004118 **

Residuals 78 245920315 3152822

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Interpret the results of these ANOVA tables

## Ex. 2: Multiple Regression and Anova Using R (20)

$$\frac{(102.8263 + 8.7395)}{2} \tag{1.13.26}$$

To get the p-value for this F-statistic, in the R console window, enter

1-pf(55.78,2,80)

## Ex. 2: Multiple Regression and Anova Using R (21)

The value returned is $3.725908*10^{-13}$. The probability of the F-statistic value of 78.4533 occurring by chance is only incredibly small, so we conclude that the model we have developed explains a significant proportion of the variation in the data set.

$$\frac{81959 + 6966}{81959 + 6966 + 62171} = 0.5885 \tag{1.13.27}$$

This is the same value as was reported for $R^2$ in the regression output.

MLR with factors

Let us run the same multiple regression model again, but this time having N and drought as factors instead of numbers. We must tell R that we want entries for these variables to be considered factors and not numbers. As factors, there are 9 specified treatment amounts for each of the 2 independent variables and 81 possible combinations between the 2-factor variables.

Convert the data for N and drought into factor variables.

## Ex. 2: Multiple Regression and Anova Using R (22)

Enter into the R console

data$N<-as.factor(data$N)

data$drought<-as.factor(data$drought)

Test to make sure that R now recognizes the N variable as a factor.

Enter into the R console

is.factor(data$N)

R returns

[1]TRUE

Great, now let's run the multiple regression. Save this model as 'm2'.

m2<-summary(1m(data=data,yield~drought+N))

summary(m2)

## Ex. 2: Multiple Regression and Anova Using R (23)

R returns

Call:

lm(formula = yield ~ drought + N, data = data)

Residuals:

Min IQ Median 3Q Max

-344.37 -86.09 0.00 86.09 344.37

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1542.42 73.95 20.86 < 2e-16 ***

drought-3 1849.62 76.10 24.31 < 2e-16 ***

drought-2 3661.00 76.10 48.11 < 2e-16 ***

drought-1 5180.37 76.10 68.08 < 2e-16 ***

drought0 6225.43 76.10 81.81 < 2e-16 ***

drought1 6730.03 76.10 88.44 < 2e-16 ***

drought2 6760.31 76.10 88.84 < 2e-16 ***

drought3 6198.62 76.10 85.40 < 2e-16 ***

drought4 6198.62 76.10 81.46 < 2e-16 ***

N25 790.06 76.10 10.38 2.35e-15 ***

N50 2028.39 76.10 26.66 < 2e-16 ***

N75 3282.42 76.10 43.14 < 2e-16 ***

N100 4114.07 76.10 54.06 < 2e-16 ***

N125 4232.83 76.10 55.62 < 2e-16 ***

N150 3597.21 76.10 47.27 < 2e-16 ***

N175 2429.25 76.10 31.92 < 2e-16 ***

N200 1136.95 76.10 14.94 < 2e-16 ***

## Ex. 2: Multiple Regression and Anova Using R (24)

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 161.4 on 64 degrees of freedom

Multiple R-squared: 0.9972,

Adjusted R-squared: 0.9965

F-statistic: 1430 on 16 and 64 DF, p-value: < 2.2e-16

Explain how these results differ from the linear regression with our explanatory variables as numbers (how do the $R^2$ values differ?)
.

Under the “Coefficients” heading, in the “Estimate” column, we find the intercept, as well as the X variable coefficients for the

multiple regression equation. You'll notice that the variables for drought = -4 and N = 0 are not listed in the "Coefficient" output. The reason for this is that the "intercept" encapsulates these variables, meaning N=0 and drought = -4 is the baseline in the regression model. All of the other effects of variable combinations on yield are quantified with respect to this baseline.

## Ex. 2: Multiple Regression and Anova Using R (25)

Write the equation from the linear regression output of the model **yield~drought + N** for N=25 and drought=0, with "drought" and "N" as numeric variables. Then write out the equation for the same linear model and parameters but with "drought" and "N" as factors. Compare the predicted yields

Intercept + N + drought ~Yield

#numeric model

124.4880 + 0.1437*(25) + 12.3198*(0) = 128.0805

#factored model

24.525 + 12.562 + 98.984 = 136.071

You can calculate a prediction from a linear model with R too. For the same parameters (N = 25, drought = -1) For the non factored linear model ('m1'), enter

predict(m1,list(N=25,drought=0))

## Ex. 2: Multiple Regression and Anova Using R (26)

R returns

8030.944[kg/hectare]

#factored model (m2)(notice the quotes around the numbers to indicate factors).

predict(m2,list(N="28.025",drought="o"))

R returns

8557.915[kg/hectare]

### Exercise 3: Correlation, Multiple Regression and Anova

This exercise contains 9 steps (including this statement).

**R CODE FUNCTIONS**

- anova
- summary
- lm
- install.packages
- library('ppcor')
- cor
- pcor

You are a maize breeder in charge of developing an inbred line for use as the 'female' parent in a hybrid cross. Yield of the inbred female parent is a major factor affecting hybrid seed production; a high level of seed production from the hybrid cross leads to more hybrid seed that can be sold. Only 2 lines remain in your breeding program, and your boss wants you to determine which of the two lines has the best yield-response to variable Nitrogen fertilizer (N) applications under several different drought levels. The three-variable dataset relating the yield (per plot) of the 2 inbred lines to the amount of N and level of drought applied to each

plot can be found in the file ex3.csv.

Determine the simple and partial correlation between yield and the amount of nitrogen fertilizer applied and drought for each of the lines. Then, develop a regression equation to predict yield from the independent variables. Test to see if an interaction between drought and N should be included in the linear model. Decide on a model to evaluate these data and decide which of the 2 lines should be selected.

Load the file ex3.csv into R.

data<-read.csv("ex3.csv",header=TRUE)

Check the head of the data to make sure the file was read into R correctly.

head(data)

drought N yield rep line
1 -4 0.000 2991.842 1 1
2 -4 28.025 3533.566 1 1
3 -4 56.050 2900.837 1 1
4 -4 84.075 5759.073 1 1
5 -4 112.100 7630.583 1 1
6 -4 140.125 6723.432 1 1

## Ex. 3: Correlation, Multiple Regression and Anova (3)

All data should be of the numeric class (that is, R recognizes all entries for all explanatory variables as numbers). Calculate the simple correlation matrix for the data.

cor(data)

drought N yield rep line

drought 1.0000000 0.0000000 0.71394080 0.00000000 0.00000000

N 0.0000000 1.0000000 0.20700055 0.00000000 0.00000000

yield 0.7139408 0.2070006 1.00000000 -0.05199154 0.09821234

rep 0.0000000 0.0000000 -0.05199154 1.00000000 0.00000000

line 0.0000000 0.0000000 0.09821234 0.00000000 1.00000000

Interpret the results:

Drought has a very high simple correlation with yield, and N has a moderate correlation with yield. Keeping in mind that all variable data are classified as numbers, what does the positive correlation between line and yield imply (think about how line is coded in the data)?

## Ex. 3: Correlation, Multiple Regression and Anova (4)

There are only 2 lines in the data. The positive correlation between yield and line means that the two variables move in the same direction; a higher value for line (i.e. 2) corresponds to higher values of yield, and vice versa. This positive correlation provides evidence for line 2 being the higher-yielding line.

If there were more than 2 lines and more than 2 reps in these data, could we analyze the data in the same way (i.e. could 'line' and 'rep' be classified as numbers in the analysis)? Could we calculate the correlation between yield and line and rep and yield? If we had more than 2 lines and reps in these data, we'd have to reclassify the 'line' and 'rep' variables as factors. We would then not be able to calculate the correlation between yield and line and rep and yield.

You should've already installed the package 'ppcor'. If you have, ignore the 'install.package' command and simply load the package using the 'library' command.

install.packages ('ppcor')

library(ppcor)

Calculate the partial correlations between the 3 variables.

pcor(data)

## Ex. 3: Correlation, Multiple Regression and Anova (5)

$estimate

drought N yield rep line

drought 1.00000000 -0.21992608 0.73450000 0.05771514 -0.10816995

N -0.21992608 1.00000000 0.29942284 0.02352788 -0.04409606

yield 0.73450000 0.29942284 1.00000000 -0.07857745 0.14727018

rep 0.05771514 0.02352788 -0.07857745 1.00000000 0.01157211

line -0.10816995 -0.04409606 0.14727018 0.01157211 1.00000000

$p.value

drought N yield rep line

drought 0.000000e+00 5.659152e-05 2.912615e-83 0.3018162 0.051970262

N 5.659152e-05 0.000000e+00 2.082329e-08 0.6742387 0.430493426

yield 2.912615e-83 2.082329e-08 0.000000e+00 0.159193 0.007829716

rep 3.018162e-01 6.742387e-01 1.591930e-01 0.0000000 0.836245389

line 5.197026e-02 4.304934e-01 7.829716e-03 0.8362454 0.000000000

$statistic

drought N yield rep line

drought 0.000000 -4.0265902 19.331597 1.0325464 -1.9433800

N -4.026590 0.0000000 5.605018 0.4203378 -0.7883476

yield 19.331597 5.6050183 0.000000 -1.4077910 2.6593260

rep 1.032546 0.4203378 -1.407791 0.0000000 0.2066984

line -1.943380 -0.7883476 2.659326 0.2066984 0.0000000

What linear model would you use to analyze these data? Should you include the interaction between N and drought? Should you include rep? Should you include line? Test some possible models, then explain which model you think is best and why.

## Ex. 3: Correlation, Multiple Regression and Anova (6)

The coefficient on the interaction (drought*N) is significant (barely) at =0.1. Coefficient on rep is not significant. Thus, 'rep' should be excluded, and the interaction term should be included.

summary(1m(data,yield~line+drought+N+drought*N))

Call

lm(formula = yield ~ line + drought + N + drought * N, data = data)

Residuals:

Min IQ Median 3Q Max

-4849.1 -1269.8 16.7 1296.9 4790.0

Coefficients:

Estimate Std. Error t value Pr(<|t|)

(Intercept) 8093.5905 367.4571 22.026 < 2e-16 ***

line 555.6503 208.7016 2.662 0.00815 **

drought 678.7811 74.5214 9.109 < 2e-16 ***

N 8.0924 1.4421 5.612 4.36e-08 ***

drought:N 0.9225 0.5585 1.652 0.09959 .

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1878 on 319 degrees of freedom

Multiple R-squared: 0.5659,

Adjusted R-squared: 0.5605

F-statistic: 104 on 4 and 319 DF, p-value: < 2.2e-16

## Ex. 3: Correlation, Multiple Regression and Anova (7)

Run the anova for the linear model you chose in the previous question.

anova(lm(data=data,yield~line+drought+N+N*drought))

Analysis of Variance Table

Response: yield

Df Sum Sq Mean Sq F value Pr(>F)

line 1 25008526 25008526 7.0885 0.008151 **

drought 1 1321540272 1321540272 374.5793 < 2.2e-16 ***

N 1 111096151 111096151 31.4893 4.357e-08 ***

drought:N 1 9624425 9624425 2.7280 0.099589 .

Residuals 319 1125452898 3528066

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Based on the regression equation, what yield value would you predict to obtain for each line under N = 100 and drought = 0?

#Note: line=0 is used for ‘line 1’ and line=1 is used for ‘line 2’.

predict(m1, list(N=140.125,drought=0,line=1))

9783.187

predict(m1,list(N=140.125,drought=0,line=2))

10338.84

## Ex. 3: Correlation, Multiple Regression and Anova (8)

Interpret the results; which line would you choose and why?

*Line 2, higher predicted yield.*

Designate ‘line’ as a factor, and run the linear regression again with the same model.

data$line<-as.factor(data$line)

data$rep<-as.factor(data$rep)

summary(1m(data=data,yield~line+rep+drought+N+N*drought))

Interpret the results of the linear regression output with line as a factor (i.e. why is ‘line2’ listed and ‘line1’ not listed in the output?).

*’line2’ indicates that if we are predicting a yield for line 2 based on linear regression, we need to add 555.6503 kg/ha to the predicted yield. If a yield value for ‘line 1’ is being predicted, we do not add anything to the predicted yield value based on the line. In effect, the ‘Intercept’ includes ‘line1’. ‘line1’ can be considered the baseline, and the ‘line2’ a deviation from the baseline.*

## Ex. 3: Correlation, Multiple Regression and Anova (9)

Call:

lm(formula = yield ~ line + drought + N + drought * N, data= data)

Residuals:

Min IQ Median 3Q Max

-4996.2 -1303.3 11.6 1389.3 4937.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8796.3156 242.1130 39.331 < 2e-16 ***

line2 555.6503 208.3777 2.667 0.00806 **

rep -294.1495 208.3777 -1.412 0.15904

drought 678.7811 74.4057 9.123 < 2e-16 ***

N 8.0924 1.4399 5.620 4.17e-08 ***

drought:N 0.9225 0.5577 1.654 0.09907 .

—

Signif. codes: 0 ‘***’ 0.001 “**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 1875 on 318 degrees of freedom

Multiple R-squared: 0.5686,

Adjusted R-squared: 0.5686

F-statistic: 83.83 on 5 and 318 DF, p-value: < 2.2e-16

# Testing Multiple Regression

## Regression Model Significance

Because multiple effects are involved in multiple regression, the determination of which terms and variables are of importance adds a level of difficulty to the analysis. Not only are there direct effects from certain variables, but combinations of effects among separate variables. These are caused by the interaction between several variables. The effects are estimated by using the associated regression coefficients.

The initial test is to determine if the total regression equation is significant. As in linear regression, “Does the regression relationship explain enough of the variability in Y to be significant?” Partitioning the sums of squares into an ANOVA table can be used to resolve the hypothesis test. The ANOVA table for multiple regression is similar to that in linear regression. Additional regression degrees of freedom are included for each X variable. Two df’s are used for a regression relationship with two variables.

The testing of the regression equation partitions the total sum of squares using the total coefficient of determination, $R^2$, (Equation 6). Note that this is the same as the square of the total correlation, as given in the Total Correction section above.

Formula for calculating total coefficient of determination.

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} \tag{1.13.28}$$

$$\text{Equation 6} \tag{1.13.29}$$

Total Correlation Equation

$$R^2_{YX_1X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1}r_{YX_2}r_{Y_1X_2}}{1 - r^2_{X_1X_2}} \tag{1.13.30}$$

## The Whole Regression Relationship

The hypothesis being tested, initially, is to test the whole regression relationship to see if it is significantly different from 0, Equation 7.

$$H_0 : \beta_1 \text{or} \beta_2 = 0 \tag{1.13.31}$$

$$H_A : \beta_1 \text{or} \beta_2 \neq 0 \tag{1.13.32}$$

$$\text{Equation 7} \tag{1.13.33}$$

The F-test uses the regression mean square, RegMS, to determine the amount of variability explained by the whole regression equation. If the RegMS is significant at your alpha level, the null hypothesis that all of the partial regression coefficients equal 0 is rejected. This F-test does not differentiate any coefficients, all are significant, or none are according to the test.

The total sum of squares in this data set can be calculated as 338 (see Exercise 3). The $R^2$ was calculated in the last section as 0.900. The ANOVA table (Table 3) with two degrees of freedom is calculated.

*Table 3 ANOVA for treatment and block effects.*

| Source | df | SS | MS | F | P |
|---|---|---|---|---|---|
| **Treatment** | 2 | 304.2 | 152.10 | 31.51 | 0.00032 |
| **Block** | 7 | 33.8 | 4.83 | n/a | n/a |
| **Total** | 9 | 338.0 | n/a | n/a | n/a |

The complete regression model is significant at a probability much less than 0.01. The regression equation is significant, explaining sufficient variability in the data.

## Regression Coefficient Significance

Individual regression coefficients ($b_1$, $b_2$, etc.) may be tested for significance. The simple coefficient of determination between each X and Y explains the sum of squares associated with each regression coefficient, including interactions with other X's. The partial coefficient of determination between each X and Y explains the additional variability without interaction. These can be tested with the residual error not explained by the regression to test the significance of each b.

Each coefficient may also be tested with a t-test, and this is done in the Parameter Estimates table of the lm() Output. Tested individually, $X_2$ is significant, while the coefficient for $X_1$ is not significant. The nitrogen term would probably be dropped because it explains little additional variance beyond that from the rainfall. The final equation would be a simpler linear equation obtained by just adding Rainfall to the Fit Model, not including N Fertilizer. This equation is:

$$\text{Y} = 49.53 + 0.515 X_2 \tag{1.13.34}$$

$$\text{Equation 8} \tag{1.13.35}$$

You will note from the original regression equation that the $X_1$ coefficient was small. This does not necessarily mean that small regression coefficients are not significant. They need to be tested to determine their significance. The testing is not to attempt to remove terms but to remove terms which add to the complexity without explaining variability in the regression analysis. Dropping terms from an equation is not always done. All coefficients in the regression equation may be significant and may be kept to explain the variability in the response.

### Exercise 4: Non-Linear Regression and Model Comparison (1)

This exercise contains 16 steps (including this statement).

**R CODE FUNCTIONS**

- anova
- summary
- lm
- install.packages
- library
- cor
- pcor
- ggplot

- reshape2

If the assumptions necessary for multiple regression are not met, a number of problems can arise. These problems can usually be seen when examining the residuals, the difference between the actual Y's from the predicted Y's.

First, if the Y's are not independent, serial correlation (or auto-correlation) problems can result. These can be seen if the residuals are plotted versus the X values, showing a consistently positive or negative trend over portions of the data. When collecting data over a period of time, this can be a problem since temporal data has some relationship to the value at the previous time. For instance, a temperature measurement 5 minutes after a previous one is going to be strongly correlated with the previous measurement because temperatures do not change that rapidly. These problems can be overcome by analyzing the data using different techniques. One of these is to take the difference of the value at the current time step from the value at the last time step as the Y value instead of the measured value.

## Ex. 4: Non-Linear Regression and Model Comparison (2)

Second, violating the equal variances assumption leads to heteroscedasticity. Here the variance changes for changing values of X. A plot of residuals where the spread gradually increases toward lower or higher X's can also occur.

The third problem is multicollinearity. Here two or more independent variables (X's) are strongly correlated (for example, the growing degree days (GDD) and hours of sunlight). The individual effects are hard to separate and lead to greater variability in the regression. Large $R^2$ values with insignificant regression coefficients are seen with this problem. Eliminating the least significant variable after testing will often solve this problem without changing the $R^2$ very much.

## Ex. 4, Non-Linear Regression and Model Comparison (3)

**POLYNOMIAL FUNCTION**

A set of functions that can be useful for describing quantitative responses are the various orders of polynomial functions. Polynomial functions have a general form:

$$Y = a + bx + cx^2 + dx^3 + ... + nx^{n-1} + \epsilon \tag{1.13.36}$$

$$\text{Equation 9} \tag{1.13.37}$$

A horizontal line is a polynomial function of order 0. Linear relationships are polynomial functions of the first order. The highest exponent of X in the function determines the order of the polynomial (0 for a horizontal line, 1 for simple linear regression equation). Each order has a distinctive shape. First-order polynomials produce a straight line, second-order polynomials produce a parabola, third-order polynomials produce a parabola with one inflection point and fourth-order polynomials produce a parabola with 2 inflection points. Graphs of the first 4 orders are shown Figure 2 below.

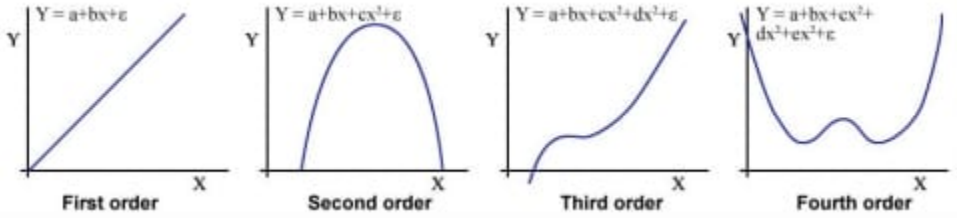



Figure 2. Graphs of first, second, third, and fourth order polynomials.

## Ex. 4, Non-Linear Regression and Model Comparison (4)

As with the other functions, an infinite number of curves may be created by carrying the coefficients. A polynomial function can usually be fit to most sets of data. The value of such relationships can be questioned at very high orders, though. Important in most functional relationships is the physical or biological relationship represented in the data. Higher-order relationships sometimes produce detailed equations which have a relatively limited physical or biological relevance.

**ADDING ORDERS OF FUNCTIONS**

Polynomial relationships are calculated to reduce the variability around the regression line, whatever the order. The usual technique is to begin with a linear equation. If the deviation from this line is significant, add a term to reduce the sum of squares (SS) about the line. Adding another order to the polynomial reduces the sum of squares. When the reduction of the sum of squares by adding another order becomes small, the limit of the equation has been reached. Enough terms can be added to fit any dataset. Generally, a third-order equation is the upper limit of terms in an equation to have any relevance. More terms often simply fit the error scatter of the data into the equation without adding additional relevance.

## Ex. 4, Non-Linear Regression and Model Comparison (5)

Each additional order should be tested for significance using the hypothesis $H_0$: highest order coefficient = 0. This can be tested using the equation:

$$F = \frac{\text{Regression SS for higher degree model - Regression SS for lower degree model}}{\text{Residual MS for higher degree model}} \tag{1.13.38}$$

**where:**
Numerator df = 1,
Denominator df = residual df for the higher order model.

Test additional models using the same approach.

Now we'll look at some very simple data and try to find the best model to fit the data. In the file QM-Mod13-ex4.csv, you'll find a very small data set giving the rate of runoff for various amounts of rainfall. Read the file QM-Mod13-ex4.csv into R and take a look at it (there are only 10 entries, so don't use the "head" command).

data<-read.csv("ex4.csv",header=T)

data

## Ex. 4, Non-Linear Regression and Model Comparison (6)

R returns

Rainfall Runoff

1 3.00 0.00

2 12.00 1.00

3 14.00 2.50

4 14.50 3.25

5 15.00 8.50

6 15.50 9.50

7 16.00 12.50

8 17.50 13.50

9 19.00 16.00

10 19.25 19.00

Let's plot the data quickly to see if we visualize any obvious trends.

```
library(ggplot2)

gplot(data=data,x=Rainfall,y=Runoff)
```

## Ex. 4, Non-Linear Regression and Model Comparison (7)

R returns the plot in Figure 3.

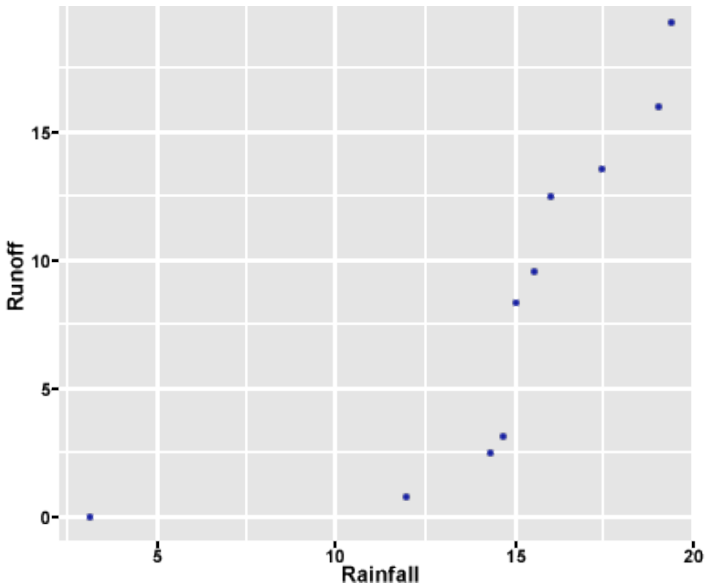



Figure 3 Rainfall vs. runoff data generated by R

Let us run the regression models of the 1st and 2nd order (i.e. Runoff ~ Rainfall and Runoff ~ Rainfall^2) and compare them visually and statistically. We will plot the predictive function given by each regression model output on a scatterplot with these data and compare the models visually.

We use "I" in front of the "x" variable in the "lm" command to indicate to R that we want the higher-order of x included in the model (i.e., for a second-order model, we would indicate $x^2$ by entering: I(x^2).

## Ex. 4, Non-Linear Regression and Model Comparison (8)

Enter the models into the R console. Call the "Rainfall" variable x, and the "Runoff" variable y.

```
x<-data$Rainfall
y<-data$Runoff
m1<-1m(y~x,data-data) #1st order
m2<-1m(y~x+I(x^2),data=data) #2nd order
m3<-1m(y~x+I(x^2)+I(x^3),data=data) #3rd order
```

Here, we create the points on the line or parabola given by each model. Because the distance between these points is so small, they will appear as a line on our figure.

```
1d<-data.frame(x=seq(0,20,by=0.5))

result<-1d

result$m1<-predict(m1,newdata=1d)

result$m2<-predict(m2,newdata=1d)

result$m3<-predict(m3,newdata=1d)
```

## Ex. 4, Non-Linear Regression and Model Comparison (9)

Here, we use the package “reshape2” to change the format of the data to facilitate graphing in the next step.

```
library(reshape2)

library(ggplot2)

result<-melt(result,id,vars=”x”,variable.name=”mode1″,value.name=”fitted”)

names(result)[1:3]<-c(“rainfall”,”order”,”runoff”)

levels(result$order)[1:3]<-c(“1st”,”2nd”,”3rd”)
```

Finally, we are ready to plot the 1st, 2nd, and 3rd order regression models on top of the original data.

```
ggplot(result,aes(x=rainfall,y-runoff))+

theme bw()+ggtitle(“Rainfall/Runoff data with 3 regression models”)+

geom point(data=data,aes(x=x,y=y))+

xlab(“Rainfall (mm)”)+

ylab(“Runoff(m^3/sec)”)+

geom line(aes(colour=order),size=1)
```

## Ex. 4, Non-Linear Regression and Model Comparison (10)

R returns the plots for the different-order polynomials in a single graph (Figure 4).

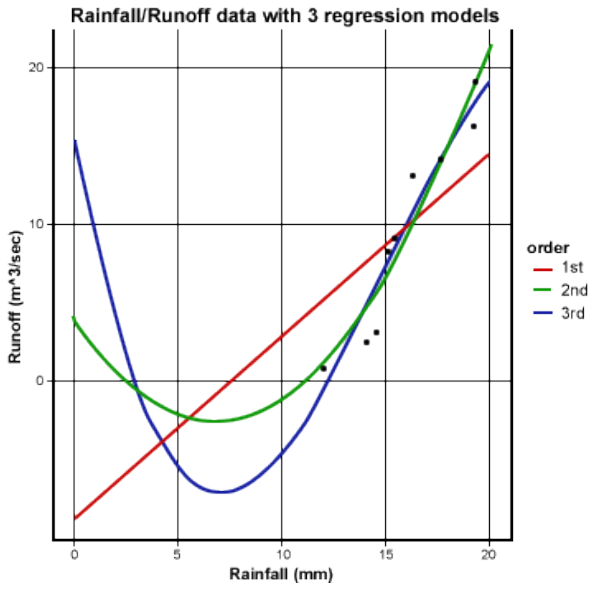



Figure 4 Rainfall vs. runoff data with 3 regression models

1ST ORDER MODEL

Let’s take a look at the output for the 1st-order regression model and ANOVA.

```
summary(m1)
```

anova(m1)

## Ex. 4, Non-Linear Regression and Model Comparison (11)

R outputs,

Call:

lm(formula = y ~ x, data = data)

Residuals:

Min 1Q Median 3Q Max

-5.4079 -3.5828 0.6918 2.2866 5.0015

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.3336 4.5632 -1.826 0.10524

x 1.1601 0.2997 3.871 0.00473

(Intercept)

x **

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 4.174 on 8 degrees of freedom

Multiple R-squared: 0.6519,

Adjusted R-squared: 0.6084

F-statistic: 14.98 on 1 and 8 df, p-value: 0.004735

## Ex. 4, Non-Linear Regression and Model Comparison (12)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 261.11 261.105 14.984 0.004735

Residuals 8 139.40 17.425

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

The equation given by the linear model is $y_{Runoff}$ = -8.3336 + 1.1601$x_{Rainfall}$. The intercept is not statistically significant; the x variable is. The $r^2$ value is 0.6519, and the linear regression is significant, but there is scatter about the regression line. The ANOVA shows a regression SS of 261.11 and a residual SS of 139.4 for the 1st-order model (Figure 4).

## Ex. 4, Non-Linear Regression and Model Comparison (13)

2ND ORDER MODEL

summary(m2)

anova(m2)

R outputs,

Call:

lm(formula = y ~ x + I(x^2), data = data)

Residuals:

Min 1Q Median 3Q Max

-2.512 -1.558 0.205 1.189 3.292

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.20020 3.44216 1.220 0.26189

x -1.87695 0.63627 -2.950 0.02141 *

I(x^2) 0.13687 0.02785 4.915 0.00172 **

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.115 on 7 degrees of freedom

Multiple R-squared: 0.9218,

Adjusted R-squared: 0.8995

F-statistic: 41.26 on 2 and 7 df, p-value: 0.0001337

## Ex. 4, Non-Linear Regression and Model Comparison (14)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 261.105 261.105 58.363 0.0001222 ***

I(x^2) 1 108.085 108.085 24.160 0.0017229 **

Residuals 8 31.316 4.474

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here you can see that the $R^2$ value increased, indicating that more of the variance in the data is explained by the regression equation. Testing the reduction using the F-test produces a very significant decrease in unexplained variability as the residual SS drops from 139.4 to 31.316. The regression line (green parabola) follows the data closely (Figure 4).

## Ex. 4, Non-Linear Regression and Model Comparison (15)

### 3RD ORDER MODEL

Let us take a look at the output for the 1st order regression model and ANOVA.

summary(m3)

anova(m3)

Call:

lm(formula = y ~ x + I(x^2) + I(x^3), data = data)

Residuals:

Min 1Q Median 3Q Max

-2.7284 -1.2079 0.3878 1.1311 2.4531

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.51888 10.04750 1.545 0.173

x -6.94363 4.28634 -1.620 0.156

I(x^2) 0.63545 0.41826 1.519 0.180

I(x^3) -0.01393 0.01166 -1.195 0.277

Residual standard error: 2.053 on 6 degrees of freedom

Multiple R-squared: 0.9368

Adjusted R-squared: 0.9052

F-statistic: 29.66 on 3 and 6 df, p-value: 0.0005382

## Ex. 4, Non-Linear Regression and Model Comparison (16)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

x 1 261.105 261.105 61.9225 0.0002229 ***

I(x^2) 1 108.085 108.085 25.6330 0.0023041 **

I(x^3) 1 6.017 6.017 1.4269 0.2773480

Residuals 8 25.300 4.217

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here, you see that not much more information about the response has been gained. The $R^2$ (and $R^2_{Adj}$) increases little, and very little additional variability is explained in the third-order regression (Figure 4). In the ANOVA table, the F-value for the third-order regression is not significant at even the 0.10 level. The second-order polynomial, therefore, is the best polynomial equation for describing the response. Physically, we are trying to fit a relationship between rainfall to runoff. The negative runoff or infiltration after rain begins makes sense. The $x^2$ relationship may be explainable since we are considering a volume of runoff from a depth of rainfall. The equation does fit the data well. Again, this fits only the data gathered. Use of this relationship beyond the scope of this dataset would be improper.

## Problems in Multiple Regression

### Examining Problems

Recall the assumptions for regression discussed at the beginning of the lesson and in Chapter 10 on Mean Comparisons. If the assumptions necessary for multiple regression are not met, a number of problems can arise. These problems can usually be seen when examining the residuals, the difference between the actual Y’s and the predicted Y’s.

First, if the Y’s are not independent, serial correlation or **auto-correlation** problems can result. These can be seen if the residuals are plotted versus the X values, showing a consistently positive or negative trend over portions of the data. When collecting data over a period of time, this can be a problem since temporal data has some relationship to the value at the previous time. For instance, a temperature measurement 5 minutes after a previous one is going to be strongly correlated with the previous measurement because temperatures do not change that rapidly. These problems can be overcome by analyzing the data using different techniques. One of these is to take the difference of the value at the current time step from the value at the last time step as the Y value instead of the measured value.

Second, violating the equal variances assumption leads to **heteroscedasticity**. Here the variance changes for changing values of X. A plot of residuals where the spread gradually increases toward lower or higher X’s can also occur. The residual plot from the replicated data regression in Chapter 7 on Linear Correlation, Regression and Prediction shows a hint of this. Notice how the residuals start to spread slightly as date of harvest, X, increases (Figure 5).

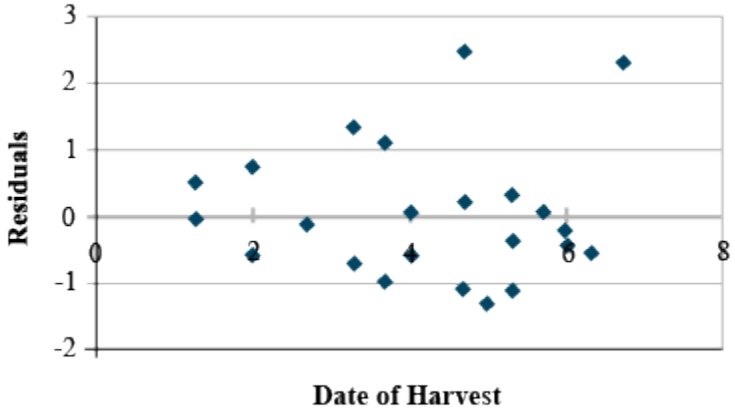



Figure 5 Residuals, or deviation of each data point from the calculated regression equation.

## Multicollinearity

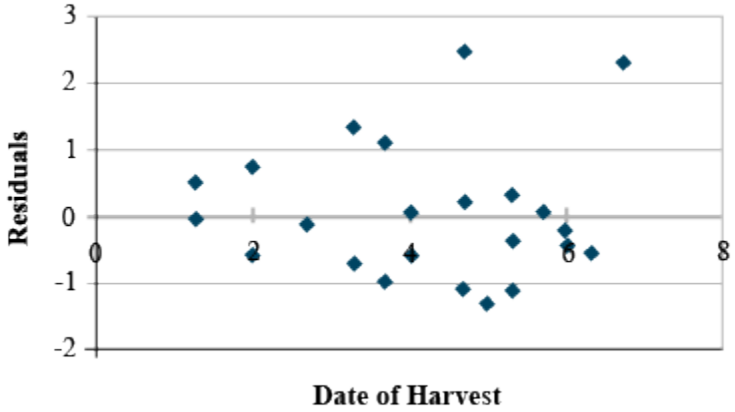



Figure 6 Residuals, or deviation of each data point from the calculated regression equation.

The third problem is **multicollinearity**, as we discussed in the first part of the unit. Here two or more independent variables (x) are strongly correlated (for example, the GDD and hours of sunlight variables). The individual effects are hard to separate and lead to greater variability in the regression. Large $R^2$ values with insignificant regression coefficients are seen with this problem. Eliminating the least significant variable after testing will often solve this problem without changing the $R^2$ very much. The example just discussed showed such a property, where the $X_1$ and $X_2$ values were strongly correlated (r=.9005). The insignificant coefficient can be eliminated, usually solving the problem (Figure 6).

## Polynomial Relationships

The application of the various orders of polynomial functions in describing quantitative responses has been covered in the earlier section – Polynomial Functions above. The equations, which are linear in the parameters (a, b, c, in Equation 9), are used to fit experimental data similar to the methods described earlier in this unit. Also covered earlier is enough terms can be added to fit any data set, with, in general, a third-order equation being the upper limit of terms in an equation to have any relevance. More terms often simply fit the error scatter of the data into the equation without adding additional relevance.

Each additional order should be tested for significance using the hypothesis H0: highest order coefficient = 0 using the F-test equation in section (Ex. 4. Non-Linear Regression and Model Comparison (5)) above.

# Polynomial Regression

## Polynomial Example

Let's use the example from Chapter 7 on Linear Correlation, Regression and Prediction. The data set was approximated using a linear model (Figure 7).

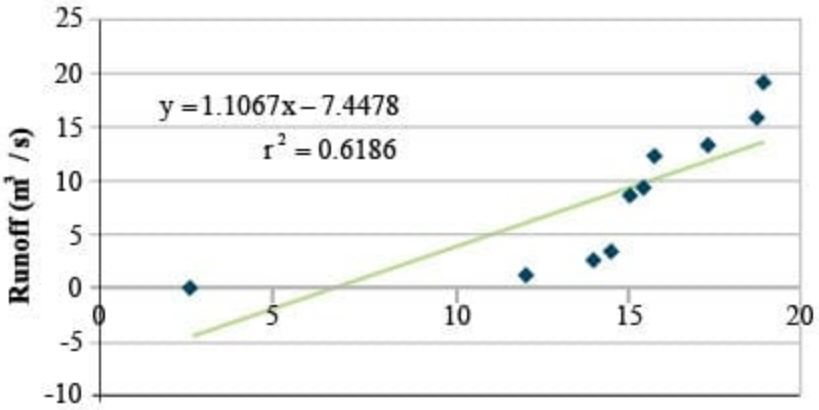



Figure 7 Linear regression applied to runoff from a field based on rainfall data.

The $R^2$ value is 0.62, with a regression SS of 242.1 and a residual SS of 149.2. The linear regression is significant, but there is a scatter about the regression line. Fitting the same data with a second-order polynomial produces:

*Table 4 ANOVA for 2nd Order Polynomial*

| n/a | df | SS | MS |
|---|---|---|---|
| **Regression** | 2 | 362.2 | 181.10 |
| **Residual** | 7 | 29.1 | 4.15 |
| **Total** | 9 | 391.3 | n/a |

Formula does not parse\small y = 4.32 - 2.05x + 0.15x^2" title="\small y = 4.32 - 2.05x + 0.15x^2" class="latex mathjax">.

$$r^2 = 0.93 \tag{1.13.39}$$

$$F = \frac{362.2 - 242.1}{4.15} = 28.9 \tag{1.13.40}$$

$$\text{Critical F} = 12.25;\ \text{p-value} = 0.01 \tag{1.13.41}$$

$$\text{with df } 1,\ 7; p - value <<< 0.01 \tag{1.13.42}$$

## Variance in the Data

Here you can see that the R2 value increased, indicating that more of the variance in the data is explained by the regression equation. Testing the reduction using the F-test produces a very significant decrease in unexplained variability as the residual SS drops from 149.2 to 29.1. The regression line follows the data closely (Figure 8).

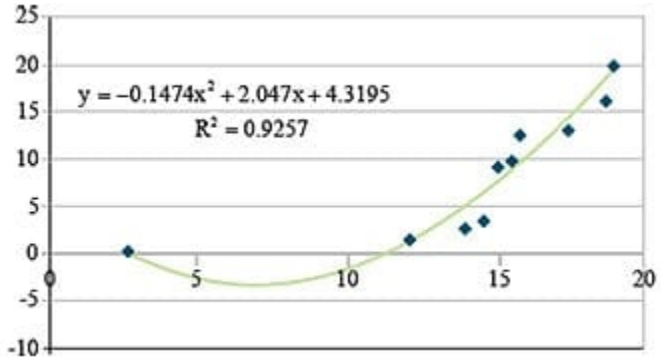



Figure 8 Linear regression applied to runoff from a field based on rainfall data.

Going a step further to assure that most of the variance is explained by the regression equation, we fit a third order polynomial (Table 5).

*Table 5 ANOVA for 3rd Order Polynomial.*

| n/a | df | SS | MS |
|---|---|---|---|
| **Regression** | 3 | 365.6 | 121.90 |
| **Residual** | 6 | 25.7 | 4.28 |
| **Total** | 9 | 391.3 | n/a |

$$y = 12.14 - 5.94x^2 + 0.01x^3 \tag{1.13.43}$$

$$r^2 = 0.93 \tag{1.13.44}$$

$$F = \frac{365.6 - 262.2}{4.28} = 0.79 \tag{1.13.45}$$

$$\text{Critical F} = 3.29;\ \text{p-value} = 0.01 \tag{1.13.46}$$

$$\text{with df } 1,\ 7;\ \text{not significant, even at p-value} = 0.01 \tag{1.13.47}$$

## Calculating Polynomial Equations

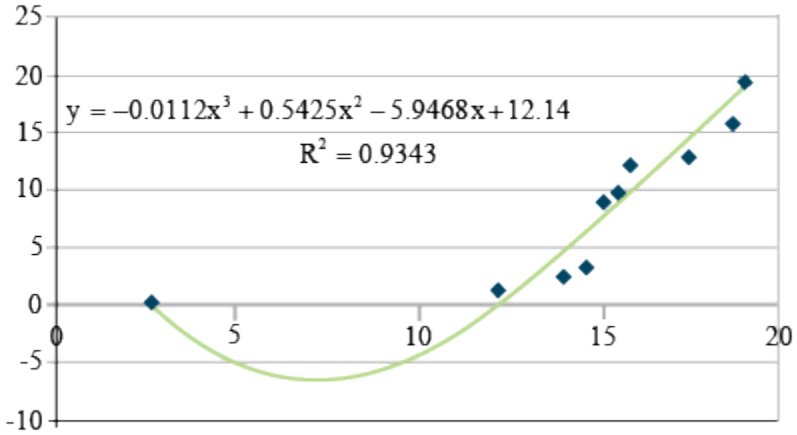



Figure 9 Third-order polynomial regression for the rainfall data.

Here, you see that not much more information about the response has been gained. The $R^2$ increases little, and very little additional variability is explained in the third-order regression (Figure 9). The F-value for third-order regression is not significant at even the 0.10 level. The second-order polynomial, therefore, is the best polynomial equation for describing the response. Physically, we are trying to fit a relationship between rainfall to run-off. The negative run-off or infiltration after rain begins makes sense. The $X^2$ relationship may be explainable since we are considering a volume of run-off from a depth of rainfall. The equation does fit the data well. Again, this fits only the data gathered. Use of this relationship beyond the scope of this data set would be improper.
 

### Exercise 5: Non-Linear Multiple Regression Analysis (1)

This exercise contains 5 steps (including this statement).

R CODE FUNCTIONS

- anova
- summary
- lm
- install.packages

- library
- cor
- pcor

You are a maize breeder in charge of developing an inbred line for use as the 'female' parent in a hybrid cross. Yield of the inbred female parent is a major factor affecting hybrid seed production; a high level of seed production from the hybrid cross leads to more hybrid seed that can be sold. Only 3 lines remain in your breeding program, and your boss wants you to determine 1. Which is the best model to use to analyze the data, and 2. Which of the three lines should be selected for advancement in the breeding program? The three-variable data set relating the yield (per plot) of the 3 inbred lines (evaluated in 3 reps) to the amount of N and level of drought applied to each plot can be found in the file QM-mod13_ex5.csv.

Helpful questions to ask in writing the R code:

1. Should "line" be classified as a numeric or factor variable?
2. Should "rep" be classified as a numeric or factor variable?
3. Should "rep" be included in the model?
4. Should the interaction between N and drought be included?
5. What higher orders of N and drought, if any, should be included in the model?

## Ex. 5: Non-Linear Multiple Regression Analysis (2)

Answers:

1. "Line" should be a factor
2. "Rep" should be a factor
3. "rep" should be included in the model (see ANOVA and Regression analysis below)
4. Yes (see ANOVA and Regression analysis below)
5. $N^2$,drought$^2$ should be included. $N^3$ has a slightly higher $R^2_{Adj}$ value, but it is only ~0.006 better than the model, including both N and drought as second-order variables.

Students should test models on their own to find the best one.

## Ex. 5: Non-Linear Multiple Regression Analysis (3)

The correct model is: yield ~ N + drought + line + rep + N*drought + $N^2$ + drought$^2$ "rep" and "line" should be factors, as the numeric values (1 to 3) are identifiers only and don't indicate a treatment amount.

The model including drought and N as a 2nd-order variable is the best. The model that has drought as a 2$^{nd}$ order polynomial and N as a 3$^{rd}$ order polynomial technically has a better $R^2_{Adj}$, however since the difference between the $R^2_{Adj}$ values of the 2 models is incredibly small AND since the coefficient $N^2$ is not significant in the model with the higher-order polynomial, we choose the simpler of the 2.

```
data$line<-as.factor(data$line)

data$rep<-as.factor(data$rep)

summary(1m(data,yield~N+drought+line+rep+N*drought+I(N^2)+I(drought^2)))
```

## Ex. 5: Non-Linear Multiple Regression Analysis (4)

R outputs,

Call:

lm(formula = yield ~ N + drought + line + rep + N * drought + I(N^2) + I(drought^2), data = data)

Residuals:

Min 1Q Median 3Q Max

-2872.09 -550.53 11.45 517.71 2123.26

Coefficients:

Estimate Std. Error t value Pr(<|t|)

(Intercept) 7.997e+03 9.944e+01 80.422 < 2e-16 ***

N 6.936e+01 1.504e+00 46.120 < 2e-16 ***

drought 7.172e+02 2.118e+01 33.864 < 2e-16 ***

line2 6.026e+02 7.264e+01 8.296 5.31e-16 ***

line3 8.256e+02 7.264e+01 11.366 < 2e-16 ***

rep2 -2.080e+02 7.264e+01 -2.864 0.00431 **

rep3 -8.158e+02 7.264e+01 -11.231 < 2e-16 ***

I(N^2) -2.730e-01 6.454e-03 -42.296 < 2e-16 ***

I(drought^2) -1.920e+02 5.069e+00 -37.869 < 2e-16 ***

N:drought 4.716e-01 1.587e-01 2.971 0.00306 **

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 800.7 on 719 degrees of freedom

Multiple R-squared: 0.9211, Adjusted R-squared: 0.9201

F-statistic: 933.1 on 9 and 719 df, p-value: < 2.2e-16

## Ex. 5: Non-Linear Multiple Regression Analysis (5)

anova(m2)

Analysis of Variance Table

Response: yield

Df Sum Sq Mean Sq F value Pr(>F)

N 1 254135518 254135518 396.4202 < 2.2e-16 ***

drought 1 2881646910 2881646910 4495.0156 < 2.2e-16 ***

line 2 88656727 88656727 69.1468 < 2.2e-16 ***

rep 2 87339382 87339382 68.1194 < 2.2e-16 ***

I(N^2) 1 1146830086 1146830086 1788.9142 < 2.2e-16 ***

I(drought^2) 1 919337101 919337101 1434.0531 < 2.2e-16 ***

N:drought 1 5659813 5659813 8.8286 0.003064 **

Residuals 719 460933695 641076

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

*Line 3 has the best predicted yield and should be advanced.*

## Summary

**Multiple Regression**

- Effects of several continuous independent variables (Xs) on continuous dependent variable Y
- With just $X_1$ and $X_2$, get the plane of best fit.

**Multiple Correlation**

- Start with all pairwise simple correlations of Y and Xs.
- Partial correlation of Y and $X_1$ holds all other Xs at their average value.
- Total multiple correlation of Y on Xs squared ($R^2$) is the coefficient of determination

**Calculating Multiple Regression**

- In R, use the REG procedure
- Get prediction equation $Y = a + b_1X_1+...+b_kX_k$ from Parameter Estimates
- Get Analysis of Variance for Regression

**Testing Multiple Regression**

- $R^2$ gives the proportion of variation accounted for by Regression
- Overall F-test of all coefficients equal to zero
- Each regression coefficient tested in Parameter Estimates or Effect Tests

**Problems in Multiple Regression**

  - Ys not independently distributed
- Unequal variances
- Either of these is seen in residual plots
- Multicollinearity from high pairwise correlations of Xs

**Polynomial Regression**

- Add successive powers of X: X, $X^2$, $X^3$, ...
- Test for significance with F-test.

## Acknowledgments

**How to cite this chapter**: Mowers, R., D. Todey, K. Moore, L. Merrick, and A. A. Mahama. 2023. Multiple Regression. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.13: Multiple Regression is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Dennis Todey; Ken Moore; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.14: Nonlinear Regression

Many of the relationships between variables encountered in agronomy are nonlinear. The growth of plants and most other organisms approaches a physiological limit as they age, so their rate of growth diminishes with time. Many other natural phenomena occur in a nonlinear manner with respect to time and can best be described using nonlinear functions. In Chapter 7 on Linear Correlation, Regression, and Prediction, we discussed how to recognize and define the linear relationship between two variables and how the change in one variable can be used to predict the resulting change of another. In Chapter 13 on Multiple Regression, we learned how to fit polynomial equations to approximate nonlinear relationships between two variables. In this chapter, some of the most common nonlinear relationships and their application are presented and discussed.

**Learning Objectives**

- To identify strong relationships that are not strictly linear
- How to perform a regression with nonlinear terms
- How to fit data with and test the usefulness of various types of nonlinear regression equations

## Approximation of Non-Linear Data

### Relationships Among Variables

**Many relationships are curvilinear rather than linear.** Relationships among variables in agronomic data are often assumed to be linear. Many statistics are based on this assumption of linearity between variables because calculations are simpler. The growing degree day formula for corn and other warm-season crops, for instance, assumes that a plant sustains linear growth between 50° F (10° C) and 86° F (30° C) (Figure 1). Much of the experimental data that is gathered, however, is inherently nonlinear. This is often caused by the nonlinear reaction of many physical and biological processes to time, temperature, and other conditions. Distributions of these data often follow other more complex but definable equations.

Almost any relationship can be fit using higher-order polynomials, as you learned in Chapter 13 on Multiple Regression. However, while the numerical relationship can be modeled with a polynomial, it may be devoid of any practical meaning or significance. We often refer to such relationships as “black box” or empirical because there is no clear or obvious relationship between the model parameters and the biology of the response. The parameters of many nonlinear models are often better defined and correspond to biological processes that can be interpreted with respect to them.

Since computation becomes easier when relationships are linear or can be approximated as linear, efforts are undertaken to create linear relationships. In Chapter 12 on Data Transformation, we discussed transformations such as the log, square root, and arcsine to make data conform to the assumptions of the ANOVA, simplifying the calculations to be performed in the analysis. Another method of approximation is to assume that a linear relationship is valid over a portion of the data. While not appropriate for a whole data set, it may be useful over a small part of it.

**In Detail – Linear Growth**

Assumptions such as this are wrong to a certain extent. A corn plant develops more slowly at colder temperatures (GDDs overestimate growth) but develops more rapidly at higher temperatures (GDDs underestimate growth in the upper 70s and low 80s). Some error is then introduced by this assumption.

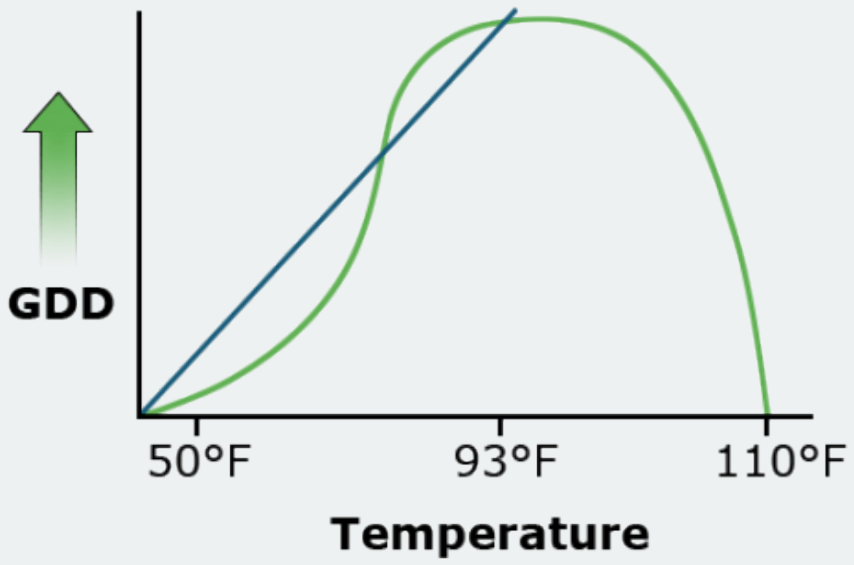



Figure 1 GDD linearly simulated growth and actual corn plant growth influenced by temperature.

## Interpolating Data

When interpolating data in the small area of interest, the linear approximation may be acceptable, especially when small errors are acceptable and their existence is understood. But when such approximations are extrapolated beyond this region, errors can grow quickly.

Statisticians refer to linear versus nonlinear equations with respect to their parameters. For example, the equation Y=∝ + β1 X + β2$X^2$ is linear in the parameters ( ∝, β1, β2) even though it is quadratic in X. It follows a linear model because the multiple linear regression can be done with $X^2$ considered as a variable in the equation.

Some other equations, such as the power curve Y = ∝Xβ are nonlinear in the parameters. However, as we see next, this equation can be linearized by taking logarithms. This is an advantage because we can use the familiar linear regression methods to fit data to this equation.

Still, other functions are not easily linearized by taking logarithms, for example, the S-shaped logistic curve of plant growth, Y =∝ /(1 + β exp(-δXβ)), where exp refers to Euler's constant e raised to the power in parentheses. These nonlinear functions require a complex iterative solution technique rather than the common linear regression methods.

A log transformation works to linearize many functions which involve an exponent. This is performed by taking the log of both sides of an equation. Other transformations may be applied in different situations.

## Difference Comparisons

The power curve (Equation 1.14.1) is a simple example of this application.

$$Y = ax^b \tag{1.14.1}$$

Taking the log or exponent of each side of Equation 1.14.1 produces linear equations (Equations 1.14.2 and 1.14.3)

$$\log \mathrm{Y} = \log \mathrm{a} + \mathrm{b} \log \mathrm{x} \tag{1.14.2}$$

$$\mathrm{Y}^1 = a^1 + bx^1 \tag{1.14.3}$$

where the primed values are used in place of the log values. The two different plots produced by these equations are shown in Figure 2.

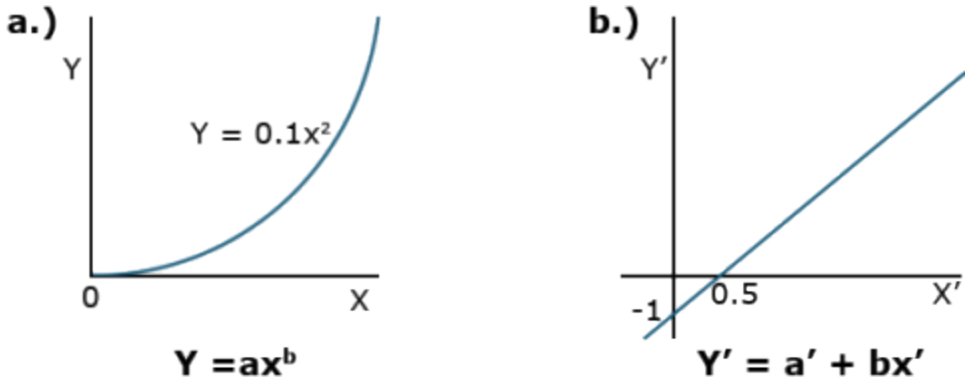



Figure 2 Equivalent graphs for Equations 1 and 2 (or 3) using a linear scale.

Study Questions 1: Difference Comparisons

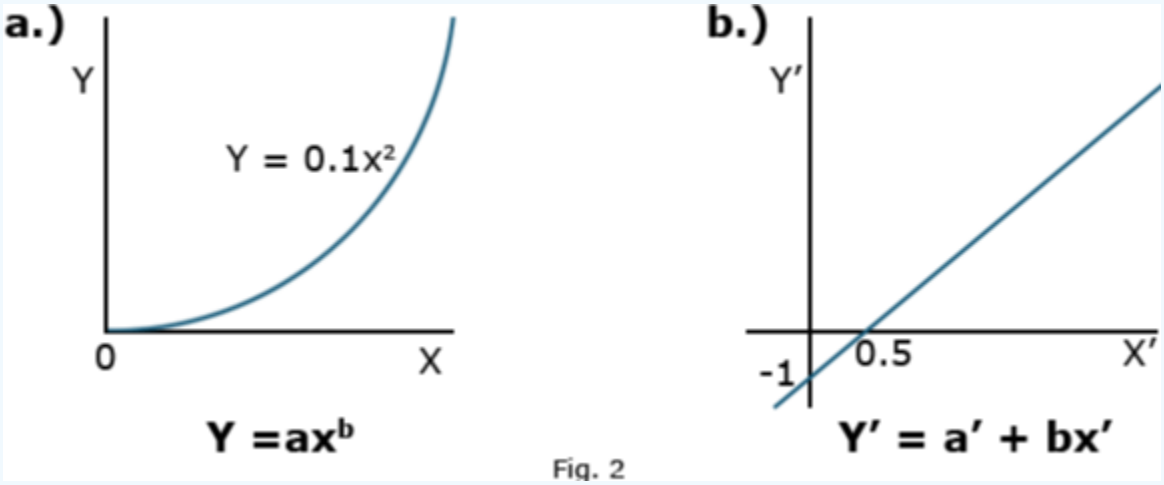



Fig. 2

Query 1.14.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Comparing Equations

Instead of the non-linear relation of the first plot, the linear equation plotted is produced. Correlation and regression equations from Chapter 7 on Linear Correlation, Regression, and Prediction can then be applied to measure the relationship. Taking the anti-log of both sides removes the logs and leaves the original equation form using X and Y.

# Functional Relationships

## Nonlinear Relationships

Several other nonlinear relationships are applicable to agricultural data and analysis. The structure and application of some of

the major ones are described here.

The **exponential curve** describes slow change at small values of X with rapidly increasing values at large X's (exponential growth). A negative exponent changes the distribution to one of exponential decay. Its shape can change into a nearly infinite number of curves depending on the parameters ***a*** and ***b*** of the equation. One form is given as Equation 4.

$$Y = ab^x \tag{1.14.4}$$

The parameters ***a*** and ***b*** can be any value. The more common representation uses the exponential function (Equation 5).

$$Y = ae^{bx} \tag{1.14.5}$$

## Exponential Graph

These equations produce a graph that appears similar to the power curve (Figure 3).

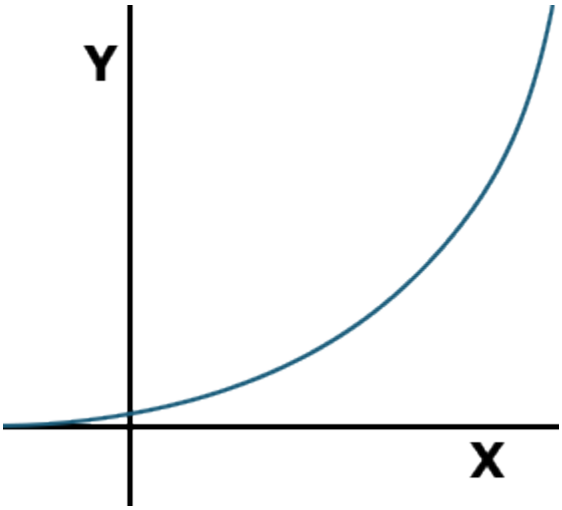



Figure 3 Generalized Exponential Graph

In Equation 5, where the b is contained in the exponent, e is Euler's constant. It is an irrational number, which is approximately equal to 2.7183. The value e is raised to the power in the exponent to calculate the value of the function.

### Study Questions 2: Y-Intercept

#### Query 1.14.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Exponential Relationships

A positive ***b*** models exponential growth, while a negative ***b*** models exponential decay toward a value ***a***. When discussing exponential growth or decay, the X is often replaced by ***t*** for time since growth or decay is often a function of time.

Taking the log of this function produces Equation 6.

$$\log y = \log a + bx \tag{1.14.6}$$

Again, a straight line is produced, which is easier to work with computationally. Biological relationships of the exponential function can be seen in the early growth of plants, where initial growth is slow, followed by a rapid increase. Another exponential relationship is that between air temperature and the saturation vapor pressure, or the amount of water needed to saturate air. As air temperature increases, the amount of water needed to saturate it increases dramatically (Figure 4).

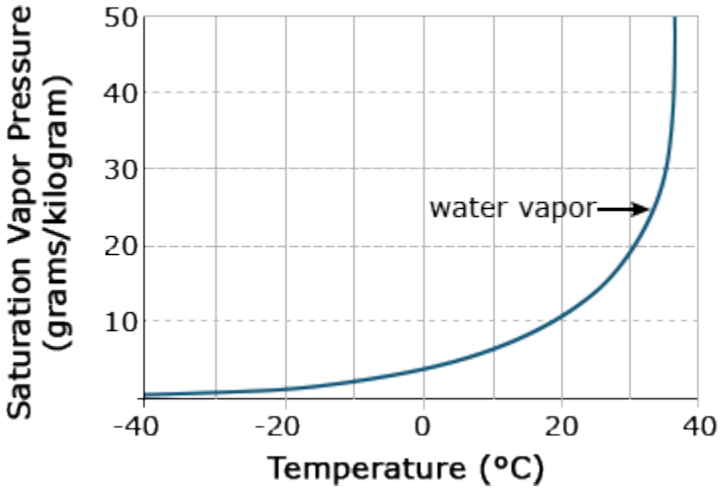



Figure 4 Saturation vapor pressure for water as a function of temperature.

## Exercise 1: Calculating the Regression Equation for an Exponential

An experiment was conducted to study the development of cabbage over an 8-week period following emergence. Height (cm) of the cabbage above the cotyledon was measured at weekly intervals. In this example, we will use simple linear regression to fit a straight line describing height as a function of time (weeks) to evaluate how well it fits the data. We will continue by transforming the Y variable so that we can fit the nonlinear exponential function to the data, also using simple linear regression. We will be using the linear regression program in Excel that you learned in Chapter 7 on Linear Correlation, Regression, and Prediction for this example.

Fit a linear regression equation for data on the growth of cabbage and determine if a nonlinear (transformed) model will fit better.

**Steps:**

1. Open the Excel data file Module 14 Example 1 data [XLS].
2. Select **Data Analysis** from the Data menu at the top of the window and select **Regression** from the list of **Analysis Tools** that appear. Click **OK**.
3. Enter the **Input Y Range**: by clicking on the spreadsheet icon to the right of the input box.
4. Using your mouse, select the data in the **Height** column, including the column heading. Click on the icon to the right of the input box labeled **Regression,** which will input the range and return you to the **Regression** window.
5. Repeat step 4 for the **Input X Range**: this time selecting the **Week** column of data.
6. Check the **Labels** box. This tells Excel that the first row will contain data labels.
7. Under **Output options**, select **New Worksheet Ply**: which will cause the results to be listed in a new worksheet.
8. Under **Residuals**, select **Residual Plots** and **Line Fit Plots,** then click **OK**.

The **SUMMARY OUTPUT** for the analysis should appear in a new worksheet. If not, go back to the steps above and make sure the input data are correct and all the other options have been selected.

We are most interested in the fit statistics that are presented in the ***Regression Statistics*** table (Table 1):

*Table 1 Summary statistics from regression analysis.*

| Regression Statistics | |
|---|---|
| Multiple R | 0.981418 |
| R Square | 0.963181 |
| Adjusted R Square | 0.957044 |
| Standard Error | 0.982233 |
| Observations | 8 |

**Ex. 1: Examining the Fit of Data**

Based on these statistics alone, we would likely conclude that the straight-line fits pretty well. The R Square is very high at 0.963, so the equation does a pretty good job of describing the relationship. However, when we look at the residual plot, we see that the equation actually overpredicts early and late in the time period and underpredicts from weeks 2 – 5. This is cause for concern because we expect residuals to be distributed randomly about the regression line. When that is not the case, as is here, it indicates that the model may not describe the relationship as well as we thought.

data<-read.csv("14_ex1.csv")

head(data)

Week Height

1 0 4.5

2 1 5.5

3 2 6.5

4 3 8.0

5 4 10.0

6 5 12.0

1m_1<-1m(data-data,Height~Week)

summary(1m_1)

Residuals:

Min 1Q Median 3Q Max

-0.9881 -0.8110 -0.1399 0.8408 1.2083

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.2917 0.6340 5.192 0.00203 **

Week 1.8988 0.1516 12.528 1.58e-05 ***

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Residual standard error: 0.9822 on 6 degrees of freedom
>
> Multiple R-squared: 0.9632
>
> Adjusted R-squared: 0.957
>
> F-statistic: 157 on 1 and 6 DF, p-value: 1.582e-05

## Ex. 1: Calculating Residuals

Calculate the residuals of the model. (Figure 5)

1m.res1<-resid(1m_1)

plot(data$Week,lm.res1,xlab="Week",ylab="residuals",main="Residual Plot,pch=20,ylim=c(-2,2))

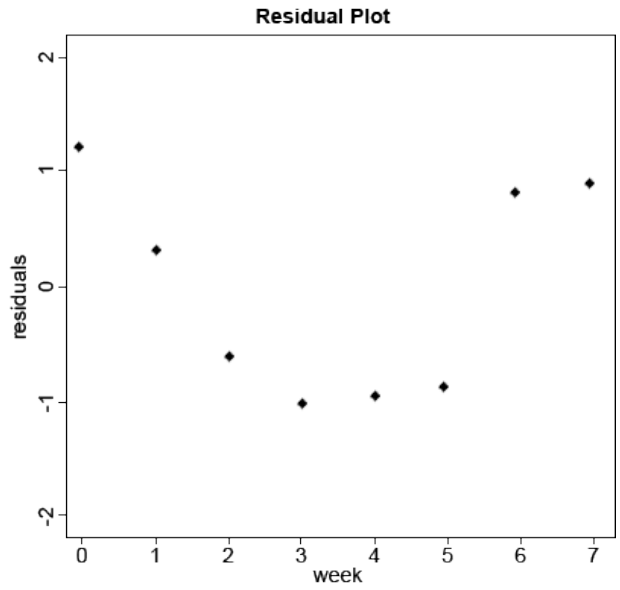



Figure 5 Residual plot from R

## Ex. 1: Calculating ANOVA

Calculate the anova table for the linear model Height~Week.

y<-aov(lm_1)

summary(y)

Df Sum Sq Mean Sq F value Pr(>F)

Week 1 151.43 151.43 157 1.58e-05 ***

Residuals 6 5.79 0.96

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The significance of the test indicates that a linear model does account for enough of the variability to be useful, but the bias

discovered when examining the residuals leads us to believe that perhaps another equation might describe the relationship better. Knowing that plant growth is inherently nonlinear, let's examine a nonlinear relationship.

### Ex. 1: Transforming the Data and Calculating Residuals

Add a column to the data set where each entry is the natural log of the corresponding entry for height.

data$lnHeight<-log(data$Height)

head(data)

Week Height lnHeight

1 0 4.5 1.504077

2 1 5.5 1.704748

3 2 6.5 1.871802

4 3 8.0 2.079442

5 4 10.0 2.302585

6 5 12.0 2.484907

## Plotting Residuals With Log-Transformed Data/Fitting Parameters 'a' and 'b'

Run the linear model with lnHeight as the response variable and Week as the explanatory variable and then look at the residuals of the model.

lm_2<-1m(data=data,1nHeight~Week)

summary(lm_2)

Calculate the residuals of the model and plot them from the log-transformed data (Figure 6).

lm.res2<-resid(lm_2)

plot(data$Week,lm.res2,xlab="Week",ylab="residuals,main="Residual
Plot(ln(Height))",pch=20,ylim=c(-0.5,0.5))

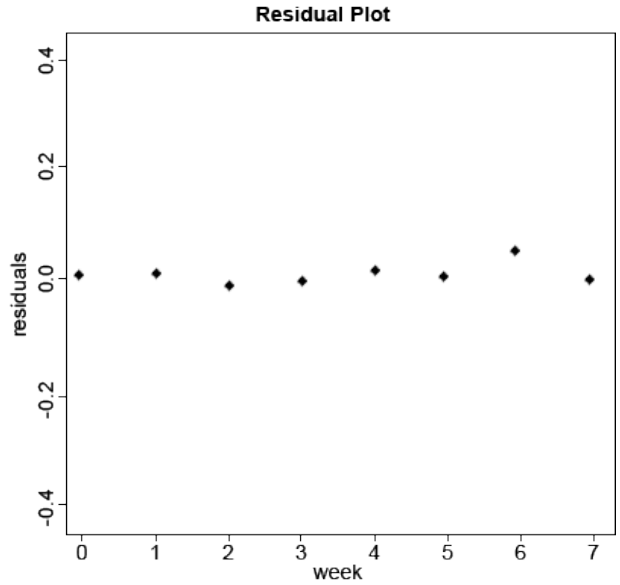



Figure 6 Residual plot of transformed data using R.

## Exercise 2: Estimating Nonlinear Regression

We have seen how the linearized exponential equation can be fit to data using a least-squares approach in the second part of Exercise 1. Being able to use this approach is nice because it allows an algebraic solution for estimating the model parameters. Many nonlinear equations, however, cannot be linearized easily and cannot be solved using the least-squares approach. Other regression methods have been developed to estimate the parameters of nonlinear equations. The process is called nonlinear regression and arrives at a solution for the estimated parameters by fitting them iteratively until the error SS for the complete model is minimized. There are different algorithms for doing this, some more complicated than others, but most work by trying different values of the parameters until no further improvement in the fit is realized by doing so.

Execute the nonlinear least square, 'nls', procedure.

a<-5

b<-0.2

fit1=nls(data=data,Height~a*exp(b*Week),start=list(a=a,b=b))
#Look at the confidence interval

confint(fit1,level=0.95)

## Ex. 2: Summary of the Model

Look at the summary of the model.

summary(lm_2)

lm(formula=lnHeight~Week,data=data)

Residuals:

Min 1Q Median 3Q Max

-0.029532 -0.016742 0.000069 0.009151 0.048509

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.495918 0.017216 86.89 1.57e-10 ***

Week 0.199402 0.004115 48.45 5.18e-09 ***

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 0.02667 on 6 degrees of freedom

Multiple R-squared: 0.9975

Adjusted R-squared: 0.997

F-statistic: 2348 on 1 and 6 DF, p-value: 5.182e-09

summary(fit1)

Formula: Height~a*exp(b*Week)

Parameters:

Estimate Std. Error t value Pr(>|t|)

a 4.504885 0.169353 26.60 1.86e-07 ***

b 0.197576 0.006714 29.43 1.02e-07 ***

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Residual standard error: 0.3788 on 6 degrees of freedom

Number ofiterations to convergance: 3

Achieved convergence tolerance: 1.785e-06

## Ex. 3: Plotting the Exponential Curve

Download and read Module 14 Example [CSV] in R and plot the data with Week on the x-axis and Height on the y-axis. Use nls outputs (***a*** as the intercept and ***b*** as the slope) to overlay the non-linear regression line (Figure 7).

```
data<-read.csv('14_ex1.csv')

plot(data$Week, data$Height,xlab="Week",
ylab="Height,main="Plot with nl regression
line",pch=20,ylim=c(4,18))

x<-seq(0,8,0.1)

y<-4.504885*exp(0.19757*x)

lines(x,y,col="red")
```

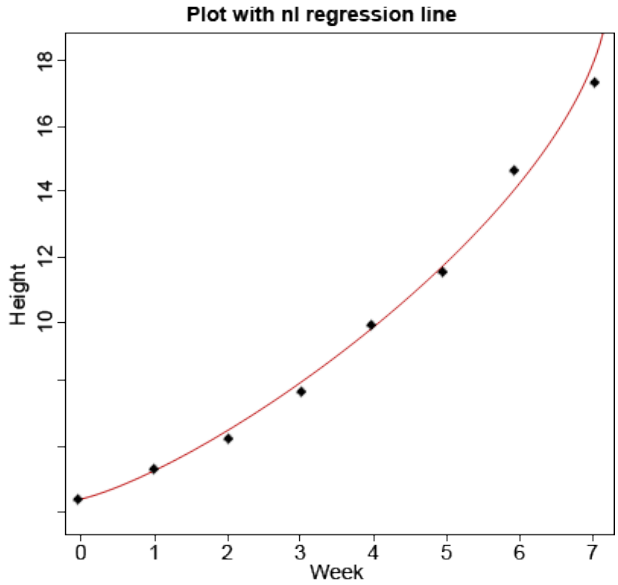



Figure 7 A nonlinear regression plot from R.

## Ex. 3: ANOVA

anova(lm_1)

Analysis of Variance Table

Response: Height

Df Sum Sq Mean Sq F value Pr(>F)

Week 1 151.430 151.430 156.96 1.582e-05 ***

Residuals 6 5.789 0.965

—

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

anova(lm_2)

Analysis of Variance Table

Response: lnHeight

Df Sum Sq Mean Sq F value Pr(>F)

Week 1 1.66997 1.66997 2357.6 5.182e-09 ***

Residuals 6 0.00427 0.00071

—

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1

Study Questions 3: ANOVA F-test

Query 1.14.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Monomolecular Function

Other nonlinear (in parameters) functions are not easily linearized and require nonlinear regression to fit. A **monomolecular function** (Equation 8) is an inverted form of the exponential function. It rises rapidly initially and then approaches an asymptote, or some limiting value. The asymptote, which is parameter estimate **a** in Equation 7, can be thought of as the maximum possible response.

$$Y = a(1 - be^{-cx}) \tag{1.14.7}$$

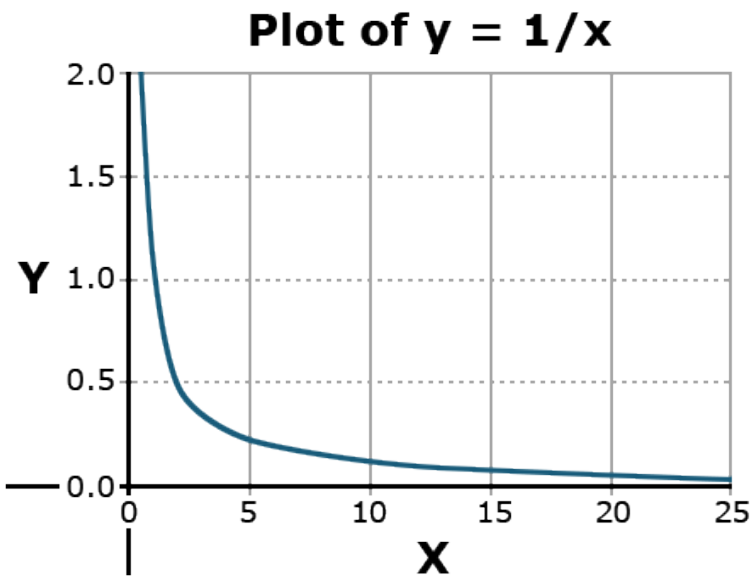



Figure 8 Plot of y = 1/x

The value Y is a(1-b) at X=0 and approaches a maximum at larger values of X (Figure 8). Thus, **a** is called the asymptote, the value which is approached but never reached. A practical application of this model would be the response of crops to fertilizer application. Applying additional fertilizer increases yields up to a point. The rate of yield increase drops off rapidly as that value is approached. This is often referred to as "diminishing returns." In the area of soil fertility, the monomolecular function is often referred to as Mitscherlich's equation.

In Detail – Maximum Possible Response

An asymptote is a value that a function will approach infinitely closely without ever reaching. A simple example is the function y = 1/x.

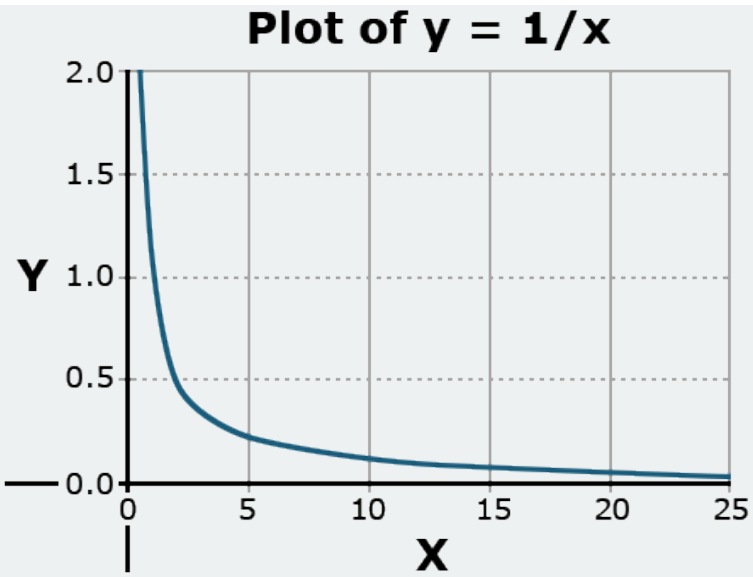



Figure 8 Plot of y = 1/x

Why doesn't this value ever reach the y or x-axis?

## Total Growth Functions: Logistic

Two functions have applications in describing the total growth or complete life cycle of a plant. Both are exponential in form, beginning from the origin. This makes sense because, at time zero, there should be no growth. They have an inflection point, where the concavity of the curve changes, and approach an asymptotic Y value as X increases. Each has different parameters.

The logistic function has the form listed in Equation 8.

$$Y = \frac{a}{1 - be^{-cx}} \tag{1.14.8}$$

Equation 8 (1.14.9)

This function is also asymptotic, approaching a maximum as X becomes very large (Figure 9).

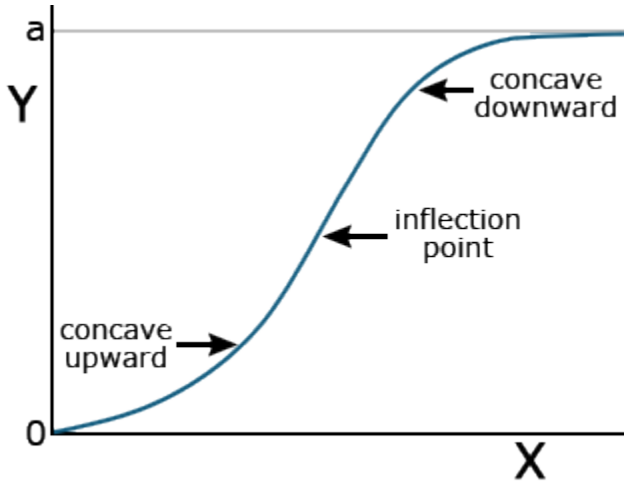



Figure 9 Generalized curve of a logistic function with associated facets of its graph.

## Total Growth Functions: Gompertz

The Gompertz function is another common equation for describing plant growth. It has the form:

$$Y = ae^{-be^{(-cx)}} \tag{1.14.10}$$

While the logistic function is more symmetric about the inflection point (the point where the curve changes from being concave upward to concave downward), the Gompertz function levels off more rapidly than the logistic function (Figure 10).

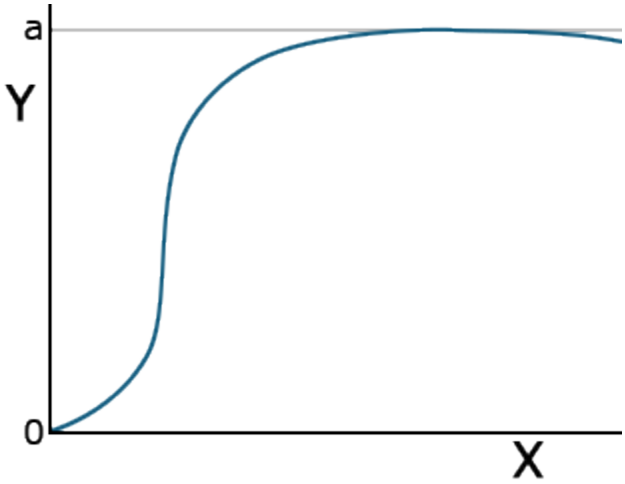



Figure 10 Generalized figure of a Gompertz function

# Nonlinear Model Calculation

## Functions Are Compared Using Error Mean Squares or $R^2$

Some nonlinear models have linear forms using a log or other transformation that improves the ease of computation. Using the linear transformation allows the use of linear regression techniques from the chapters on Linear Correlation, Regression and Prediction, and Multiple Regression. Other nonlinear functions can not be linearized and require nonlinear modeling software. The use of computers has reduced the difficulty of obtaining parameters for equations. The technique for finding the parameters of these equations is qualitatively the same as for a linear equation. The idea is to minimize the deviations of the data around the line. The calculation is much less straightforward, however. Generally, it requires an initial guess of the values of the constants and then iterates closer to a solution by nudging the values closer to the “best fit.”

The choice of functional relationship is somewhat arbitrary. There are accepted functions for certain applications. Often, testing several functions for a “best fit” approach works well.

Exercise 4

**Ex. 4: Estimating Regression Equations**

Start R, set your working directory, and make sure all of the data sets for Nonlinear Regression are in the working directory folder. Verify the file reads in correctly by checking the ‘head’ of the data (first download the .xls or .xlsx file and save it as a .csv format).

data<-read.csv(“14_ex4.csv“)

a<-500

b<-25

```
c<-0.5

Plot the logistic function line over data.

m1 = nls(data=data,Yield ~ a/(1+b*exp(-c*Week)), start=list(a=a,b=b,c=c))
confint(m1, level=0.95)

plot(data$Week,data$Yield,xlab = “Week”, ylab=”Yield”, main = “data +
Logistic”,pch=20)

x<-seq(o,10,0.1)

y<-496,3023/(1+27.7107*exp(-0.6156 *x))

lines(x,y,col=”red”)
```

**Ex. 4: Plot Monomolecular and Gompertz**

```
Plot the monomolecular function line over data.

data<-read.csv(“14_ex4.csv”)

a<-500

b<-10

c<-0.1

m2 = nls(data=data,Yield ~`a*(1-(b*exp(-c*Week))), start=list(a=a,b=b,c=c))

confint(m2, level=0.95)

plot(data$Week,data$Yield,xlab = “Week”, ylab=”Yield”, main = “data + Monomolecular”,pch=20)

x<-seq(0,10,0.1)

y<-2.076e+03*(1-(1.037*exp(-3.075e-02*x)))

lines(x,y,col=”red”)

Plot the Gompertz function over data.

data<-read.csv(“14_ex4.csv”)

a<-500

b<-5

c<-0.25

m3 = nls(data=data,Yield ~ a*exp(-b*exp(-c*Week)), start=list(a=a,b=b,c=c))
```

**Ex. 4: Computation**

```
Compute the confidence interval of the model and create a plot of data with the Gompertz function overlaid.

confint(m3,level=0.95)
```

```
plot(data$Week,data$Yield,xlab = “Week”,
ylab=”Yield”, main = “data +
Goompertz”,pch=20)
#Plot the Gompertz function line

x<-seq(0,10,0.1)

y<-555.1372 *exp(-5.1347*exp(-0.3499*x))

lines(x,y, col = “red”)
```

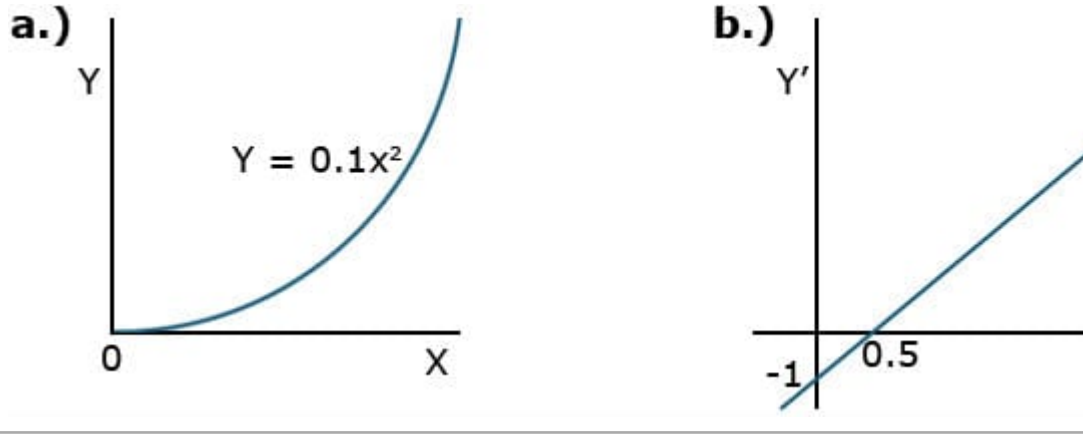
## Selecting the Best Function

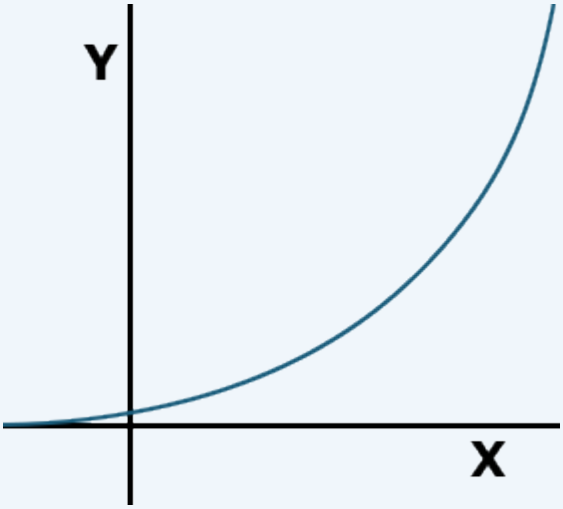
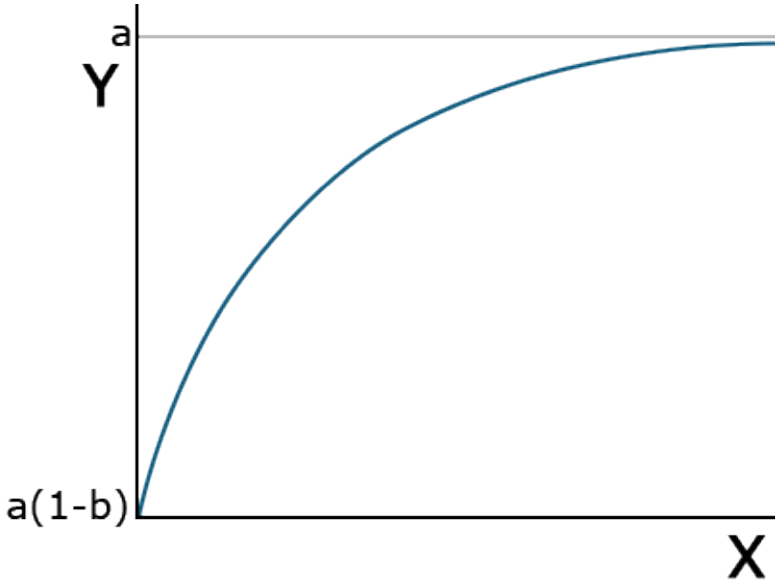
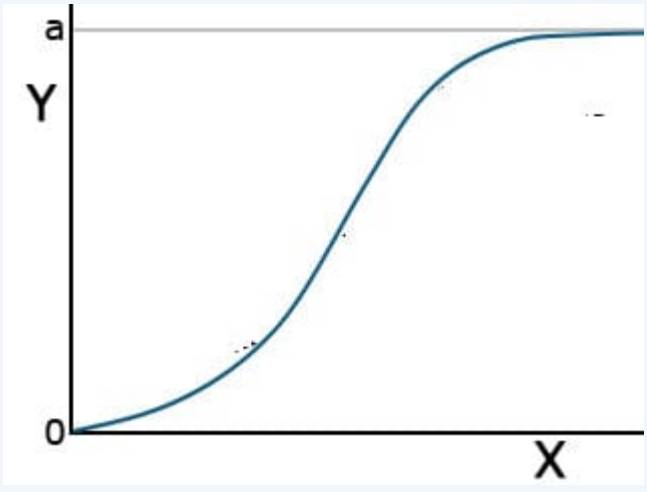
If we have several nonlinear models and want to select the best of these functions, there are several considerations. First, knowledge of theoretical reasons that one of these functions should be superior is probably the most important consideration. For example, the monomolecular model is theoretically a good function to relate crop growth to fertilizer application. The logistic and Gompertz models are theoretically more appropriate for modeling growth as a function of time. If we do not have a strong theoretical model or want to choose among several potential models, we can use statistics from fitting the models to make the comparison. First, we can compare the error mean squares of the models. We want the smallest error mean square possible. Secondly, we can try to compare $R^2$ values. However, this is difficult for two reasons. We can always fit a model perfectly ($R^2 = 1$) if we just include enough parameters or variables in the model. It is also difficult to use $R^2$ because the statistic is computed as the proportion of variation accounted for based on the sums of squares after correcting for the mean. Nonlinear models often do not even have a mean value as one of the parameters, and such a statistic is not generally computed for these models.

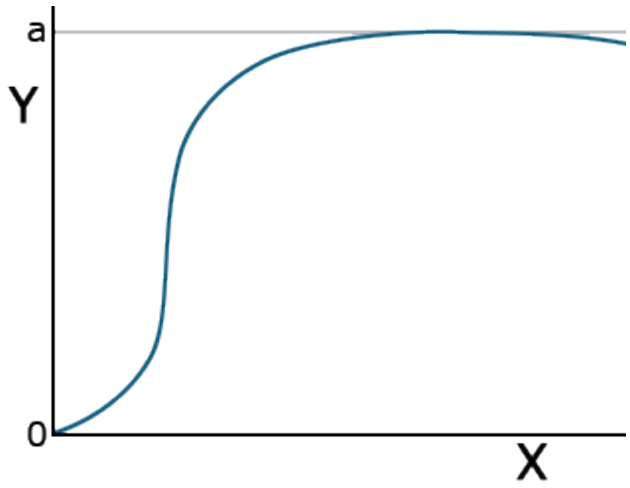
## Summary of Nonlinear Functions

The main functions discussed in this lesson are summarized for easy referral.

*Table 2 Different nonlinear functions, equations, and graphs for evaluating of calculating them and their application.*

| Function | Equation | Graph |
|---|---|---|
| Power Curve | $Y = ax^b$<br><br>$\log Y = \log a + b \log x$ |  |

| Function | Equation | Graph |
|---|---|---|
| Exponential Growth/Decay | $Y = ae^{bx}$ |   |
| Monomolecular | $Y = a(1 - be^{cx})$ |   |
| Logistic | $Y = \frac{a}{1+be^{-cx}}$ |   |

| Function | Equation | Graph |
|---|---|---|
| Gompertz | $Y = ae^{-bc^{-cx}}$ |   |

# Summary

## Curvilinear Relationships

- Nonlinear (in parameters) which can be linearized Examples: Power curve, Exponential Growth
- Nonlinear not easily transformed Examples: Monomolecular, Logistic, Gompertz

## Nonlinear Functions

- Can fit with R NLIN
- Compare models using error SS (or $R^2$ if mean is in model)
- Test for significance with F-test

**How to cite this chapter**: Mowers, R., D. Todey, K. Moore, L. Merrick, and A. A. Mahama. 2023. Nonlinear Regression. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods.* Iowa State University Digital Press.

This page titled 1.14: Nonlinear Regression is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ron Mowers; Dennis Todey; Ken Moore; Laura Merrick; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.15: Multivariate Analysis

Multivariate datasets: for each unit, various variables have been assessed. What information do we want to extract?

- Group the units based on the various variables into groups.
- Determine how the variables interact, and which of them have the main effects on the units.
- Eliminate variables that eventually overshadow the effects of those variables we are really interested in.
- Reduce the complexity of the dataset so we can use graphical tools to help us in interpretation.

**Learning Objectives**

- Analysis, graphical display and interpretation of complex datasets

## Measures that Describe Similarities/Dissimilarities Between Units or Variables



Figure 1 Initial example similarities/dissimilarities.

If we are asked to describe a group of diverse objects, as given in Figure 1, our mind would automatically start to look for attributes these objects have in common and others that allow us to divide them into groups.

- **Common:** these are all letters
- **Different:** Letters: A & B
- **Font size**: large & small
- **Case**: upper & lower
- **Color**: red & black

We are still able to say which of the objects are identical = highest similarity (for example, the 2 black uppercase “A” in large font – positions 1 and 4 in Figure 1, or the 2 red lowercase “a” in small font – positions 2 and 9, but we have to take a closer look to find the objects pairs, that share the least of the variables = lowest similarity.

### Example 1: Data Sheet

Even in a very small dataset, it would be nice to have algorithms at hand to compute similarities/dissimilarities between the objects.

| Object | Letter | Font | Case | Color |
|---|---|---|---|---|
| 1 | A | L | U | B |
| 2 | A | S | L | R |
| 3 | B | S | U | B |
| 4 | A | L | U | B |
| 5 | A | S | U | R |
| 6 | B | L | L | B |
| 7 | B | S | U | R |
| 8 | B | S | L | R |
| 9 | A | S | L | R |



| Object | Letter | Font | Case | Color |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 |
| 5 | 1 | 0 | 1 | 0 |
| 6 | 0 | 1 | 0 | 1 |
| 7 | 0 | 0 | 1 | 0 |
| 8 | 0 | 0 | 0 | 0 |
| 9 | 1 | 0 | 0 | 0 |

Figure 2 Datasheet for the initial example and conversion to binary variables.

Figure 2 (above) shows a datasheet we can create based on the objects in Figure 1. We have 9 objects and 4 variables. These are

categorical/ordinal variables. For this example, we can transform the variables into binary variables since each of them has only two states.

## Example 1: R Output

There are many different approaches how to compute the dissimilarity between objects. Without looking into detail, let's just try out one and use R to calculate the distance matrix for the example.

## load packages

library("cluster")

#load the binarey datasheet (Ex1.csv)

Ex1_data <- read.csv("Ex1.csv", header = T)

Ex1_data

Letter Font Case Color

1 1 1 1 1

2 1 0 0 0

3 0 0 1 1

4 1 1 1 1

5 1 0 1 0

6 0 1 0 1

7 0 0 1 0

8 0 0 0 0

9 1 0 0 0

#calculate the simple matching coefficient with the function daisy

daisy(Ex1_data, metric = "gower")

Dissimilarities:

1 2 3 4 5 6 7 8

2 0.75

3 0.50 0.75

4 0.00 0.75 0.50

5 0.50 0.25 0.50 0.50

6 0.50 0.75 0.50 0.50 1.00

7 0.75 0.50 0.25 0.75 0.25 0.75

8 1.00 0.25 0.50 1.00 0.50 0.50 0.25

9 0.75 0.00 0.75 0.75 0.25 0.75 0.50 0.25

Metric : mixed ; Types = I, I, I, I

Number of objects : 9

I

**R output for the initial example.**

The simple matching coefficient computes the percentage of variables that are different between two objects. If you have a look into the dissimilarities matrix, you can see that, for example, the object pair 1-4 has a dissimilarity value of zero ($d_{14}$ = 0) – these are the 2 identical large, black, uppercase “A”, we identified earlier.

You can also see that for cases where the dissimilarities are expressed as numbers between 0 and 1, there is a simple relation between similarities ($s_{ij}$) and dissimilarities ($d_{ij}$) (Equation 1):

$$S_{ij} = 1 - d_{ij} \tag{1.15.1}$$

$$\text{Equation 1} \tag{1.15.2}$$

**where:**
$S_{ij}$ = the similarity,
$d_{ij}$ = the dissimilarity.

Furthermore, a dissimilarity, $d_{ij}$, is also called a distance or metric if it fulfills certain properties:

**[1]** $d_{ij} >= 0$ and $d_{ij} = 0$ if and only if i = j

**[2]** $d_{ij} = d_{ij}$

**[3]** $d_{jk} =< d_{ij} + d_{jk}$

# Calculating Similarities/Dissimilarities for Different Data Types

## Calculating Similarities and Dissimilarities in Binary Data

In molecular marker data based on allelic non-informative marker systems (for example, AFLP data), only the presence or absence of a specific band can be scored.

Similarity coefficients calculated in a binary marker data set depend on the question if the absence of the marker alleles in both observed objects should be taken into account or not and also if alleles present in both observed objects should be counted once or twice.

The first step is to determine which bands are common or different in two objects:

**$a_{ij}$** – number of bands present in both objects i and j

**$b_{ij}$** – number of bands present in object i but not in j

**$c_{ij}$** – number of bands present in object j but not in i

**$d_{ij}$** – number of bands absent in both objects i and j

# Different Coefficients

## Simple matching coefficient:

$$D(SM) = 1 - \frac{a_{ij} + d_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}} \tag{1.15.3}$$

Equation 2 (1.15.4)

**where:**
$a_{ij}$ = number of bands present in both objects i and j,
$b_{ij}$ = number of bands present in object i but not in j,
$c_{ij}$ = number of bands present in object j but not in i,
$d_{ij}$ = number of bands absent in both objects i and j.

The simple matching coefficient (Equation 2) computes the percentage of variables that are different between two objects related to all variables observed.

## Jaccard's coefficient:

$$d(J) = 1 - \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}} \tag{1.15.5}$$

Equation 3 (1.15.6)

**where:**
$a_{ij}$ = number of bands present in both objects i and j,
$b_{ij}$ = number of bands present in object i but not in j,
$c_{ij}$ = number of bands present in object j but not in i,
$d(J)$ = Jaccard's coefficient.

The Jaccard's coefficient takes only those observations into account where an observation was made – if, for example, a marker band is absent in both objects observed, then the observation is considered as non-informative and not included in the calculation (Equation 3).

## Dice coefficient:

$$D(SOR) = 1 - \frac{2a_{ij}}{2a_{ij} + b_{ij} + c_{ij}} \tag{1.15.7}$$

Equation 4 (1.15.8)

**where:**
$2a_{ij}$ = two times the number of bands present in both objects i and j,
$b_{ij}$ = number of bands present in object i but not in j,
$c_{ij}$ = number of bands present in object j but not in i,
$d(SOR)$ = Dice coefficient.

The Dice coefficient (Equation 4) is similar to Jaccard's coefficient as non-informative observations are not included, but a band present in both objects is counted twice.

# Calculating Dissimilarities for Marker Data

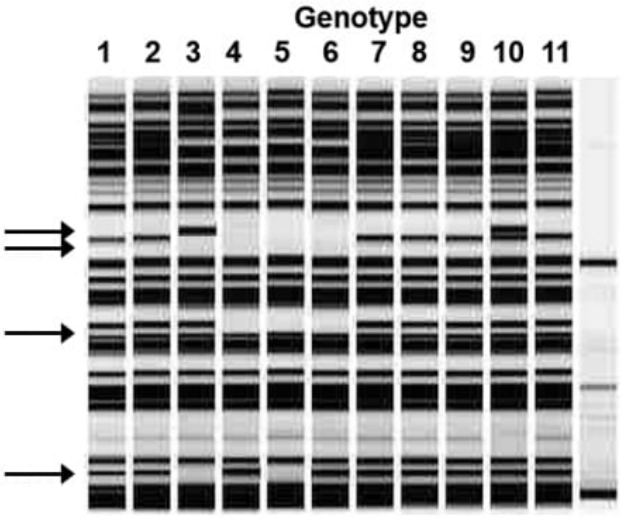



Figure 3 Partial AFLP gel, 4 polymorphic marker bands are found within 11 genotypes analyzed.

## Example 2: Marker Data

First, we score the presence/ absence of the marker bands in each genotype (Figure 3), and Table 1.

*Table 1 Marker data for the second example. The 0-1 matrix is saved as a .csv in Excel Example 2 [CSV].*

| | m1 | m2 | m3 | m4 |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 |
| 3 | 1 | 0 | 1 | 0 |
| 4 | 0 | 0 | 0 | 1 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 1 |
| 7 | 0 | 1 | 1 | 1 |
| 8 | 0 | 1 | 1 | 1 |
| 9 | 0 | 1 | 1 | 1 |
| 10 | 1 | 1 | 1 | 1 |
| 11 | 0 | 1 | 1 | 1 |

Then we can calculate the dissimilarities $d_{SM}$, $d_J$, and $d_{SOR}$ using R (Example 2 [CSV]).

For the simple matching coefficient, we can use the function daisy{cluster}; the other coefficients can be calculated with the function betadiver{vegan}. Encoding for the different diversity measures can be found in Koleff et al. (2003) Journal of Animal Ecology, 73, 367-382.

## Dissimilarity Matrices (Figure 4)

```
> DSM
Dissimilarities :
      1    2    3    4    5    6    7    8    9   10
2  0.00
3  0.75 0.75
4  0.50 0.50 0.75
5  0.75 0.75 0.50 0.25
6  0.50 0.50 0.75 0.00 0.25
7  0.00 0.00 0.75 0.50 0.75 0.50
8  0.00 0.00 0.75 0.50 0.75 0.50 0.00
9  0.00 0.00 0.75 0.50 0.75 0.50 0.00 0.00
10 0.25 0.25 0.50 0.75 1.00 0.75 0.25 0.25 0.25
11 0.00 0.00 0.75 0.50 0.75 0.50 0.00 0.00 0.00 0.25

Metric :  mixed ;  Types = I, I, I, I
Number of objects : 11
> DJ
           1         2         3         4         5         6         7         8         9        10
2  1.0000000
3  0.2500000 0.2500000
4  0.3333333 0.3333333 0.0000000
5  0.0000000 0.0000000 0.0000000 0.0000000
6  0.3333333 0.3333333 0.0000000 1.0000000 0.0000000
7  1.0000000 1.0000000 0.2500000 0.3333333 0.0000000 0.3333333
8  1.0000000 1.0000000 0.2500000 0.3333333 0.0000000 0.3333333 1.0000000
9  1.0000000 1.0000000 0.2500000 0.3333333 0.0000000 0.3333333 1.0000000 1.0000000
10 0.7500000 0.7500000 0.5000000 0.2500000 0.0000000 0.2500000 0.7500000 0.7500000 0.7500000
11 1.0000000 1.0000000 0.2500000 0.3333333 0.0000000 0.3333333 1.0000000 1.0000000 1.0000000 0.7500000
> DSOR
           1         2         3         4         5         6         7         8         9        10
2  1.0000000
3  0.4000000 0.4000000
4  0.5000000 0.5000000 0.0000000
5  0.0000000 0.0000000 0.0000000 0.0000000
6  0.5000000 0.5000000 0.0000000 1.0000000 0.0000000
7  1.0000000 1.0000000 0.4000000 0.5000000 0.0000000 0.5000000
8  1.0000000 1.0000000 0.4000000 0.5000000 0.0000000 0.5000000 1.0000000
9  1.0000000 1.0000000 0.4000000 0.5000000 0.0000000 0.5000000 1.0000000 1.0000000
10 0.8571429 0.8571429 0.6666667 0.4000000 0.0000000 0.4000000 0.8571429 0.8571429 0.8571429
11 1.0000000 1.0000000 0.4000000 0.5000000 0.0000000 0.5000000 1.0000000 1.0000000 1.0000000 0.8571429
```

Figure 4 Dissimilarity matrices for the three coefficients.

# Calculating Similarities and Dissimilarities in Categorical Data

Categorical variables are non-numerical, and it is not possible to apply any order to them.

In order to calculate distances between objects described by categorical variables, the variables can be transformed into binary variables, as we did already in example 1, for the special case when only two states for each variable are possible. If there are more than 2 states, we will have to proceed a bit differently.

*Table 2 Data extracted from FSE Documents [PDF]: Key characteristics for differentiating thistles.*

| Common Name | Species Name | Origin | On noxious weed list | Flower | Growth form |
|---|---|---|---|---|---|
| Musk thistle | Carduus | nonnative | 1 | Purple | biennial |
| Scotch thistle | Onopordum | nonnative | 1 | Purple | biennial |
| Canada thistle | Cirsium | nonnative | 1 | Purple | perennial |
| Bull thistle | Cirsium | nonnative | 0 | Purple | biennial |
| Anderson's thistle | Cirsium | native | 0 | Red | perennial |
| Snowy thistle | Cirsium | native | 0 | Red | biennial |
| Douglas or swamp thistle | Cirsium | native | 1 | White | biennial |
| Elk or Drummond thistle | Cirsium | native | 1 | White | biennial |
| Perennial sow-thistle | Sonchus | nonnative | 1 | Yellow | perennial |

In the example in Table 2 (Example 3 [CSV]), only the variables "origin", "on noxious weed list," and "Growth Form" can readily be transformed into a binary variable. "Species name" and "Flower" have 4 different stages each.

## Creating Placeholder Variables

One way to handle this would be to annotate these variables the following way—we create a set of binary placeholder variables (Table 3), which define the states the original variable can assume:

For example, Flower—to describe the 4 states of flower color, we will need two binary variables Fl1 and Fl2:

| | Purple | Red | White | Yellow |
|---|---|---|---|---|
| Fl1 | 1 | 1 | 0 | 0 |
| Fl2 | 1 | 0 | 1 | 0 |

Same for species names:

| | Carduus | Onopordum | Cirsium | Sonchus |
|---|---|---|---|---|
| Sn1 | 1 | 1 | 0 | 0 |
| Sn2 | 1 | 0 | 1 | 0 |

How to estimate the number of placeholder variables (N) needed to represent the categorical data with X states:

$$\mathrm{N} \tag{1.15.9}$$

## Binary Placeholder Variables

*Table 3 Binary placeholder variables for example 3 (Example 3-tr [CSV])—Flower and Species name.*

| Common Name | Fl1 | Fl2 | Sn1 | Sn2 |
|---|---|---|---|---|
| Musk thistle | 1 | 1 | 1 | 1 |
| Scotch thistle | 1 | 1 | 1 | 0 |
| Canada thistle | 1 | 1 | 0 | 1 |
| Bull thistle | 1 | 1 | 0 | 1 |
| Anderson's thistle | 1 | 0 | 0 | 1 |
| Snowy thistle | 1 | 0 | 0 | 1 |
| Douglas or swamp thistle | 0 | 1 | 0 | 1 |
| Elk or Drummond thistle | 0 | 1 | 0 | 1 |
| Perennial sow-thistle | 0 | 0 | 0 | 0 |

# Calculating Similarities or Dissimilarities in Quantitative Data

Quantitative variables can be discrete values (for example, ear row counts, pod numbers) or continuous values (for example, yield (t/ha) or temperatures (°C)).

We will use the data in Excel file (Example 4 [CSV]) to calculate a few of the most commonly used similarities/dissimilarities in quantitative datasets.

## Euclidean Distance

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{1.15.10}$$

Equation 5 (1.15.11)

**where:**
$d_{ij}$ = Euclidean distance between objects i and j,
$x_i$ = number of bands present in object i but not in j,
$x_j$ = number of bands present in object j but not in i,
$k$ = number of occurrences of objects i and j.

The Euclidean distance (Equation 5) calculates the square root of the sum over all squared differences between two objects.

In our example, the distance between the two hybrids from AGRIGOLD would be calculated as:

$$d_{12} = \sqrt{(24.7 - 35.4)^2 + (12.93 + 11.74)^2 + (10.7 - 1.7)^2 + (100 - 99)^2} = \sqrt{197.906} = 14.07 \tag{1.15.12}$$

You can already see that the unit the respective variable is measured in has a great influence on the results of the Euclidean Distance. Some standardization of the data set is advisable, and we will come to that later.

## Manhattan Distance

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \tag{1.15.13}$$

Equation 6 (1.15.14)

**where:**
$d_{ij}$ = Manhattan distance between objects i and j,
$x_i$ = number of bands present in object i but not in j,
$x_j$ = number of bands present in object j but not in i,
$k$ = number of occurrences of objects i and j.

The Manhattan distance (Equation 6) describes the distance between two points in a grid, allowing only strictly horizontal or vertical paths. The distance is the sum of the horizontal and vertical components of the path between two points.

In our example, comparing the two first hybrids of Example 4 [CSV]:

$d_{12}$ = (24.7-35.4) + (12.93-11.74) + (10.7-1.7) + (100-99) = 21.89

## Euclidean and Manhattan Distance Results

Below (Fig, 5 and Figure 6) are the results for Euclidean and Manhattan distances generated in R:

```
> DEUC<-daisy(Ex4_matrix, metric = "euclidean")
> DEUC
Dissimilarities :
           1         2         3         4         5         6         7         8         9
2  14.0679103
3   7.7757572 13.8249810
4   9.7514922 10.7931877  6.0735821
5  13.0216166  3.6512464 11.9716707  7.8080791
6   4.8127435 11.6405155  5.4443457  5.0105988  9.5932268
7   7.4450050  6.7119297  8.2962702  6.9055702  5.9610066  5.3929213
8   1.8096685 13.1151668  6.9579092  9.0940035 12.0887716  4.3552727  6.5088862
9   6.7803835 12.5730863  1.5223666  5.3198120 10.7709842  4.1538055  6.8317275  5.9160882
10  5.5509008 10.1973330  8.1276626  9.8311800 10.2844543  6.5345237  4.5329461  4.4672587  6.9259007
11 15.5194233 15.8380428  8.3111311 11.0346772 14.5787517 13.0157597 13.2832827 14.3388424  9.1649386
12  8.9751045  9.5804802  6.1023684  1.6947271  6.7446275  4.3335897  5.4504679  8.1262784  5.0742586
13  2.2433011 12.8595062  5.7427868  7.7012726 11.5173304  2.9780698  6.1540962  1.7669465  4.6243270
14  9.0288925 10.4392720  5.8003534  1.3675160  7.5360732  4.3037658  6.1971284  8.2055835  4.9143565
15 14.3984583  2.0300000 13.3970295  9.8346327  2.3259622 11.3925458  7.0309957 13.3623688 12.2289983
          10        11        12        13        14
2
3
4
5
6
7
8
9
10
11 13.4599406
12  8.5440037 11.1915146
13  5.0170609 13.5061060  6.9390850
14  9.0107935 11.0855040  0.8912912  6.9701148
15 10.8911019 15.1131764  8.7060956 12.9771183  9.5113669
```

Figure 5 R output: Euclidean distances for the example 4 data.

```
Metric :  euclidean
Number of objects : 15
> DMAN<-daisy(Ex4_matrix, metric = "manhattan")
> DMAN
Dissimilarities :
       1     2     3     4     5     6     7     8     9    10    11    12    13    14
2  21.89
3  12.92 21.61
4  10.74 12.63 10.38
5  18.35  6.54 20.07  9.69
6   5.15 17.44  7.83  6.21 13.90
7  11.99  9.90 12.51 10.33  9.24  7.54
8   3.57 20.46 12.45 10.83 18.92  6.62 10.56
9  10.36 19.45  2.56  8.22 17.91  5.79  9.95 10.41
10  9.05 16.84 13.57 15.39 15.70 11.80  6.94  7.62 11.01
11 23.05 24.94 13.13 12.31 21.80 17.90 22.64 21.72 13.49 23.70
12 11.75 11.44 10.83  2.81  9.90  6.60  8.74  9.02  9.19 13.80 13.90
13  3.18 19.07  9.74  7.84 17.53  3.97  9.17  2.99  7.42  7.83 19.87  8.57
14 11.43 12.52 10.51  2.29 10.98  6.28 10.02  8.54  8.87 15.08 13.18  1.28  8.25
15 21.46  3.43 21.18 10.80  3.11 17.01 10.33 20.03 19.02 16.79 22.91 11.01 18.64 12.09

Metric :  manhattan
Number of objects : 15
```

Figure 6 R output: Manhattan distances for the example 4 data.

# Correlation

Correlation (linear correlation coefficient, Pearson correlation coefficient):

The correlation coefficient can obtain values between -1 and 1; it measures similarity (Equation 7).

$$S_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x_i})^2 \sum_{k=1}^{n}(x_{jk} - \bar{x_j})^2}} \tag{1.15.15}$$

Equation 7 (1.15.16)

**where:**
$S_{ij}$ = correlation coefficient between objects i and j,
$x_i$ = variable $x_i$,
$x_j$ = variable $x_j$,
$\bar{x}$ = mean of variable x,
$k$ = number of occurrences of variables i and j.

The function cor() in R calculates similarities between columns/variables (which is the more common application for the function)—if you want to compare the rows/objects, you will have to transpose your data matrix first.

## Calculating the Correlation

```
> Ex4_dft <-t(Ex4_df)
> cor(Ex4_dft)
          1         2         3         4         5         6         7         8         9        10
1  1.0000000 0.9826690 0.9981800 0.9948651 0.9863283 0.9986156 0.9950108 0.9998018 0.9982792 0.9976970
2  0.9826690 1.0000000 0.9871757 0.9922141 0.9991313 0.9884615 0.9962344 0.9853336 0.9888368 0.9925406
3  0.9981800 0.9871757 1.0000000 0.9988980 0.9916960 0.9998900 0.9968445 0.9991332 0.9999210 0.9972193
4  0.9948651 0.9922141 0.9988980 1.0000000 0.9961106 0.9987994 0.9980771 0.9966828 0.9990677 0.9965968
5  0.9863283 0.9991313 0.9916960 0.9961106 1.0000000 0.9923954 0.9976532 0.9890464 0.9928390 0.9940142
6  0.9986156 0.9884615 0.9998900 0.9987994 0.9923954 1.0000000 0.9976652 0.9994629 0.9999813 0.9982070
7  0.9950108 0.9962344 0.9968445 0.9980771 0.9976532 0.9976652 1.0000000 0.9964164 0.9977348 0.9991482
8  0.9998018 0.9853336 0.9991332 0.9966828 0.9890464 0.9994629 0.9964164 1.0000000 0.9992452 0.9982855
9  0.9982792 0.9888368 0.9999210 0.9990677 0.9928390 0.9999813 0.9977348 0.9992452 1.0000000 0.9980360
10 0.9976970 0.9925406 0.9972193 0.9965968 0.9940142 0.9982070 0.9991482 0.9982855 0.9980360 1.0000000
11 0.9914272 0.9935427 0.9970930 0.9995541 0.9973337 0.9968943 0.9972666 0.9938303 0.9973335 0.9945567
12 0.9946078 0.9931737 0.9986356 0.9999681 0.9967572 0.9986315 0.9984489 0.9964690 0.9989030 0.9968495
13 0.9997199 0.9857536 0.9992743 0.9969813 0.9894901 0.9995793 0.9966177 0.9999929 0.9993845 0.9983338
14 0.9947282 0.9918507 0.9988967 0.9999929 0.9958968 0.9987422 0.9978366 0.9965717 0.9990228 0.9963163
15 0.9821653 0.9996967 0.9879341 0.9933595 0.9996333 0.9888779 0.9959653 0.9851556 0.9893745 0.9916689
         11        12        13        14        15
1  0.9914272 0.9946078 0.9997199 0.9947282 0.9821653
2  0.9935427 0.9931737 0.9857536 0.9918507 0.9996967
3  0.9970930 0.9986356 0.9992743 0.9988967 0.9879341
4  0.9995541 0.9999681 0.9969813 0.9999929 0.9933595
5  0.9973337 0.9967572 0.9894901 0.9958968 0.9996333
6  0.9968943 0.9986315 0.9995793 0.9987422 0.9888779
7  0.9972666 0.9984489 0.9966177 0.9978366 0.9959653
8  0.9938303 0.9964690 0.9999929 0.9965717 0.9851556
9  0.9973335 0.9989030 0.9993845 0.9990228 0.9893745
10 0.9945567 0.9968495 0.9983338 0.9963163 0.9916689
11 1.0000000 0.9996191 0.9942380 0.9995707 0.9950933
12 0.9996191 1.0000000 0.9967732 0.9999367 0.9942136
13 0.9942380 0.9967732 1.0000000 0.9968763 0.9856432
14 0.9995707 0.9999367 0.9968763 1.0000000 0.9930824
15 0.9950933 0.9942136 0.9856432 0.9930824 1.0000000
```

Figure 7 Calculating the correlation between objects after transposing the data matrix.

# Preparing Data for Statistical Analysis

If we have a closer look at the data from Example 4 [CSV] and the distances we calculated (Figure 7, we realize that the values depend a lot on in what unit, for example, the yield is measured. Changing the data from t/ha into kg/ha would result in completely different results.

The Euclidean distance between the first two hybrids would change from 14.0679 to a value of 1190.08256! As the numerical value of yield is then much larger than the value of the other variables, yield would have a much larger weight than the other variables.

The example shows that raw data cannot just be used for statistical analysis. It has to be prepared before any statistical analysis is applied.

Real data are never perfect; there are missing values, outliers, or other inconsistencies, which have to be dealt with.

As an example of how to prepare a raw dataset for statistical analysis, we will use the data collected in the file “RawDataEarPhenotypes.xlsx”:

Four inbred lines, their respective F1 (6), F2 (6), and the 2 possible BC1 (2 x 6) were grown in a field trial with three replications. From each of the 3 x 28 plots, 10 randomly sampled ears were evaluated for 14 different ear phenotypes: row number (RowNo.), Kernels per row (K/Row), ear length (EarL), cob length (CobL), ear diameter at the base (EarDB), middle (EarDM), and tip (EarDT), cob diameter at the base (CobDB), middle (CobDM), and tip (CobDT), ear weight (EWt), grain weight (GrWt), cob weight (CobWt), and the 300 kernel weight (300KWt). So we expect a datasheet with 10 x 28 x 3 = 840 rows with data for the 14 variables measured...

## Looking for Obvious Inconsistencies

The following inconsistencies can be found in the dataset (Figure 6):

1. There are 832 instead of the expected 840 observations: in two replications of inbred line PHG84, 6 instead of 10 cobs were measured!
2. Using the Excel function Min() and Max(), we observe in the 300KWt variable an unexpectedly high value for the 300 kernel weight of 8706—obviously a typing mistake; if we correct this value to 87.6, another obviously too-high value, 998.2 shows up, we correct it to 99.82 In CobL the value of 183 as a maximum value is too high, we change it to 18.3 In EWt the value of 735.4 is too high, we change it to 73.54

3. The Min() value in 300KWt column also seems to be very low—what happened?—Obviously, not all cobs had the required number of 300 seeds—so the seeds were counted and weighed, and in an additional column #Kernels, the number of seed weighed is given. There are different ways to handle this problem
    1. Eliminate the value generated with less than 300 seeds completely from the data set = missing values.
    2. Replace the value generated with less than 300 seeds completely from the data set with the mean 300KWt calculated based on the available correct data.
    3. Calculate an estimate for the weight of 300 seeds based on the available information—The problem is, if only a few seeds can be weighed, this estimate can be very insecure—we might consider doing this estimate only for cases where a certain number of seed (for example more than 150) are available.

The column "ca300KWt" is the result of applying first solution 3c and then 3a to the 300 kernel weight data.

4. Once you start reading the data into a program like R, you will realize that there is another small typo in the EarL—one value has 2 decimal points: 21..9

## Typical Data Clean-up Example

This is an example of a typical data clean-up (Figure 8).

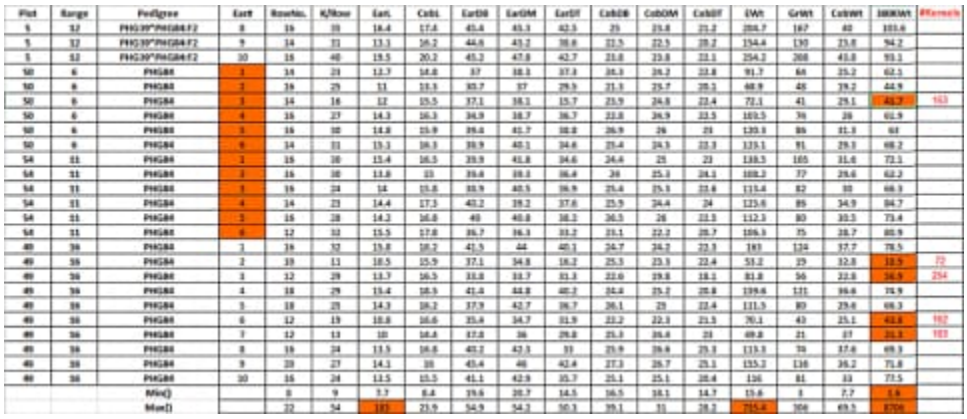


Figure 8 Extract from the raw data file – "RawDataEarPhenotypes.xlsx" – colored cells are inconsistencies in the data set that have to be dealt with before any statistical analysis.

## Generating Consistent Data: Missing Values

Some computer programs tolerate missing values, but eventually, it can become necessary to replace them with estimates for the real value.

One way to go is to replace the missing value with the mean overall values or, better, the mean over the values within the group of your missing value.

In order to make our data set a bit more manageable, we continue with the means over the 10 cobs for all repetitions and variables: "MeanEarPhenotypes.csv"

In this dataset, only the variable ca300KWt still has missing values—4 in total. We replace these with the mean calculated with the two remaining values within each pedigree (Figure 9).

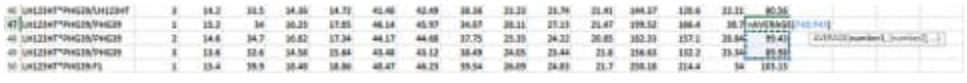


Figure 9 The final cleaned dataset is saved as a .csv file: "MeanEarPhenotypes.csv".

# Cluster Analysis

**Cluster analysis** is an exploratory technique which allows us to subdivide our sample units into groups, such that similarities of

sample units within a group are larger than between groups. Cluster analysis is applied if we have no idea how many groups there are. This is in contrast to another technique called the **Discriminant Analysis**, where the groups are given, and the sample units are distributed to the groups so they fit best.

Besides the grouping of sample units, cluster analysis may also reveal a natural structure in the data and eventually allow to define prototypes for each cluster in order to reduce the complexity of datasets.

There are several algorithms to perform the task of grouping; we will have a closer look at 2 of them:

1. Agglomerative hierarchical clustering
2. K-means clustering

## Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering methods produce a hierarchical classification of the data.

There are two ways to do so: (1) starting from a single “cluster” containing all units of the datasets in a series of partitions, the units are divided into n clusters, containing each individual unit, (2) starting out from n individual “cluster”, a series of fusions are performed until only one cluster containing all units is formed (=agglomerative techniques).

Hierarchical classifications can be represented in 2-dimensional diagrams called **dendrograms**, which will show the stage at which a fusion between units has been made.

## Hierarchical Clustering Example

- Download the dataset: Mean Ear Phenotypes [CSV]

The dataset contains the means over all repetitions done in the 4 inbred lines, the 6 F1, the 6 F2, and the 2 x 6 BC1 families. We will perform the cluster analysis based on the Euclidean distance matrix calculated with the raw (non-standardized) data (Figure 10).

```
Source on Save                                                    Run
1  ## load packages
2  library("cluster")
3
4  # set working directory for module and check
5  setwd("C:/Users/ufrei/Desktop/WD-module")
6  getwd()
7
8  #load the MeanMeanEarPhenotypes dataset and look at the structure
9  data <- read.csv("MeanMeanEarPhenotypes.csv", header = T)
10 str(data)
11 #remove the first column (Pedigree) of the dataset by substting
12 datap <- subset(data, select=-Pedigree)
13 head(datap)
14
15 ###### Generating the distance matrix
16 # calculate an Euclidean distance matrix and round the data to three digits
17 Dist <- dist(datap, method="euclidean")
18 round(Dist, digits=3)
19 Dist
```

Figure 10 R codes for using Euclidean function.

# Different Agglomeration Methods

We will be using the hclust() function of the stats package of R:

The function allows choosing between different agglomeration methods, which basically differ in how to calculate the distance between a group of units. Given the distance matrix, one can calculate the distance between two groups of units (cluster) based on the minimal distance between two individuals of the group (Fig 11a) or the maximal distance between the two individuals of a group (Fig 11b).

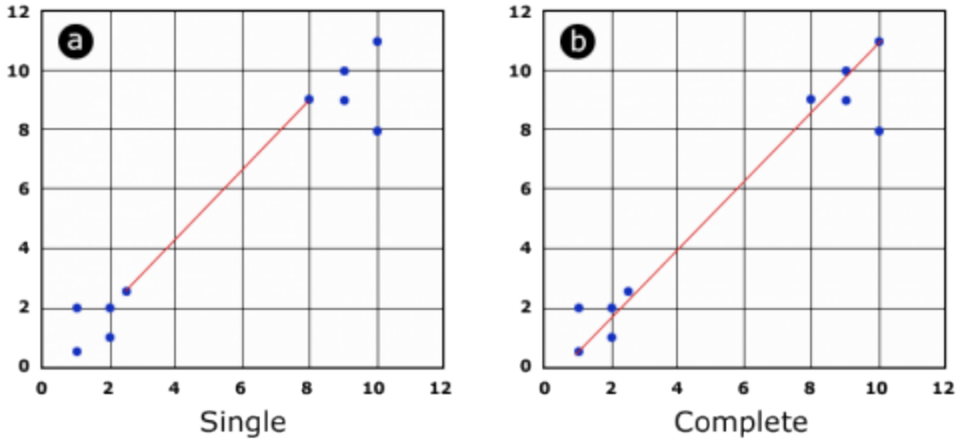



Figure 11 Illustration of the single and complete agglomeration method.

## Cluster Analysis Results

Alternatively, the distance between two clusters can be calculated as the average distance of the units within these clusters; this method is also called Unweighted Pair Group Method with Arithmetic Mean (UPGMA) and is widely applied.

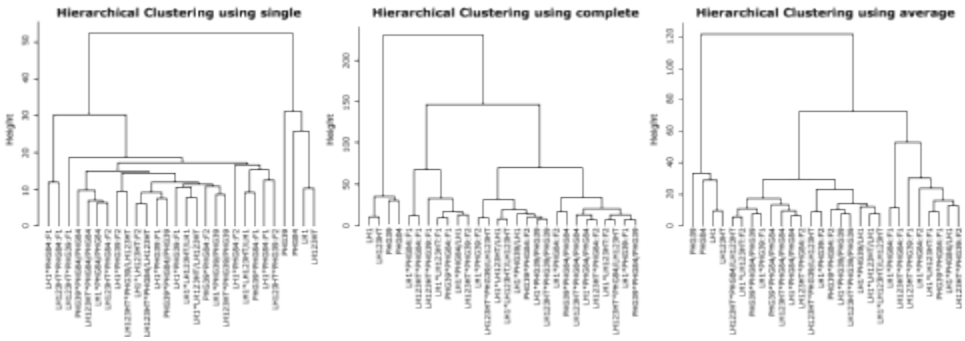



Figure 12 Results of the cluster analysis using different agglomeration methods.

All clustering methods partition the inbred lines clearly apart from the other families (F1, F2, and BC2) (Figure 12). The methods “complete” and “average” return a similar partition of the units, although a clear partitioning in F1 versus F2 versus BC1 is not achieved. By choosing a height at which to cut off, the researcher decides how many groups or clusters he wants to form within the dataset (Figure 13).

```
#### Hierarchical Clustering using different methods: single, complete, average=UPGMA

x<- hclust(Dist, method="single")
plot(x, labels = data$Pedigree, hang = -0.1, main = "Hierarchical Clustering using single")

x<- hclust(Dist, method="complete")
plot(x, labels = data$Pedigree, hang = -0.1, main = "Hierarchical Clustering using complete")

x<- hclust(Dist, method="average")
plot(x, labels = data$Pedigree, hang = -0.1, main = "Hierarchical Clustering using average")
```

Figure 13 R-code for generating the cluster and dendrogram (Example-HierarchicalClusterAnalysis.R).

## Deciding a Cut-off Height

Deciding on a cut-off height, we can divide our dataset into 3, 4, or more groups. In Figure 14, the most obvious cut-off height will be at the level of 3 clusters, dividing inbred lines against mainly F1 and the rest (F2, BC1).

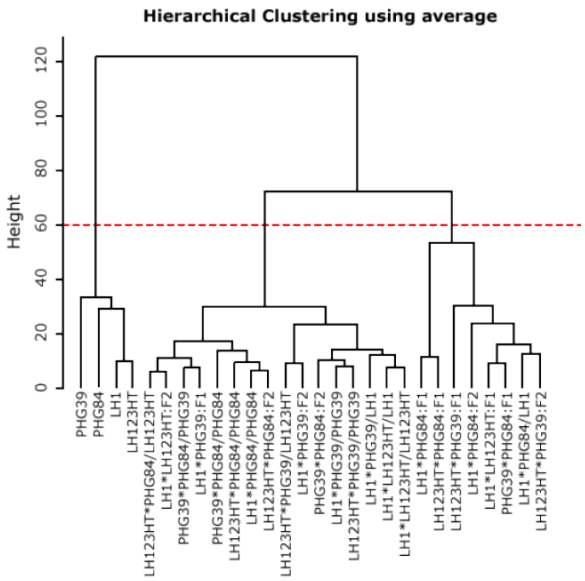



Figure 14 Dendrogram after UPGMA.

## K-means Clustering

The k-means clustering algorithm tries to partition your dataset into a given number of groups with the goal to minimize the "within group sum of squares" over all variables. Checking each possible partition of your n sample units into k groups for the lowest within-group sum of squares is not practical, as the numbers of calculations necessary rise exponentially. In our example dataset, it would be more than 2,375,000 possible partitions to check if we assume *k=3* groups. The k-means clustering algorithm, therefore, starts out making an initial partition in the number of groups requested and rearranges these so that the sum of squares is minimized. The technique is called an unsupervised learning technique and can result from each time it is initialized in slightly different results. For k-means clustering, it is recommended to use standardized datasets.

### K-means Clustering Example

- Download the dataset: MeanEarPhenotypes [CSV]

To prepare the dataset for the k-means cluster analysis, we remove the first two columns (Pedigree and Type) and standardize the data (Figure 15)—see file Example-kmeansClusterAnalysis.txt

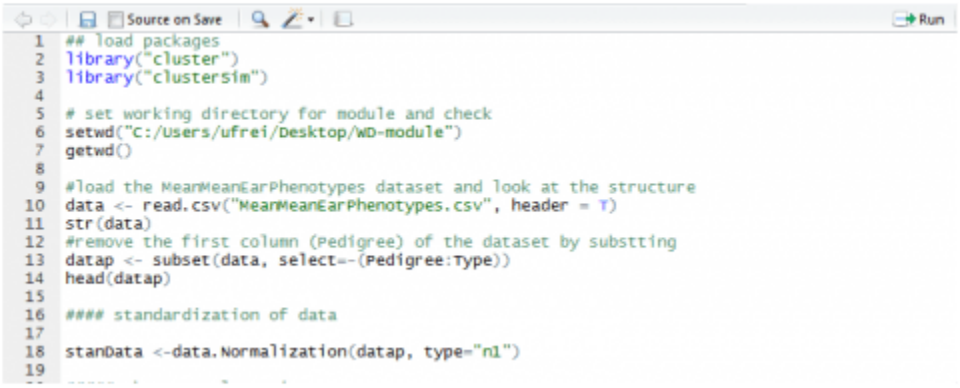



Figure 15 Preparing the data for analysis with k-means – data standardization with the function data.Normalization{clusterSim}.

## K-means Cluster Analysis

```
>
> #####  k-means Clustering
>
> cl3 <-kmeans(stanData, 3)
> table(data$Type,cl3$cluster)

         1 2 3
  BC1    8 4 0
  F1     1 5 0
  F2     2 4 0
  Inbred 0 0 4
>
> cl4 <-kmeans(stanData, 4)
> table(data$Type,cl4$cluster)

         1 2 3 4
  BC1    0 1 4 7
  F1     0 3 2 1
  F2     0 1 3 2
  Inbred 4 0 0 0
>
> cl5 <-kmeans(stanData, 5)
> table(data$Type,cl5$cluster)

         1 2 3 4 5
  BC1    1 3 6 2 0
  F1     3 2 0 1 0
  F2     1 2 1 2 0
  Inbred 0 0 0 0 4
>
```

Figure 16 R-code for k-means cluster analysis with k=3, k=4, and k=5 groups, and assignment of the Types into these groups.

If we compare the clustering results, depending on the number of groups we allow, we see that all k-means clustering analyses clearly put the inbred lines into a single group (Figure 16).

$F_1$, $F_2$, and $BC_1$ are distributed over the remaining groups. It looks like the “k=3 cluster analysis” is able to assign the $F_1$ and $BC_1$ at least to some extent, mainly to their individual group, but overall, k-means gives us here the same result as the hierarchical cluster analysis.

## Distribution of Types

The graphics below (Figure 17) show the data for the variables GrWt and ca300KWt, as an example of how the types are distributed compared to the group assignment through k-means cluster analysis.

```
plot(data[c("GrWt", "ca300KWt")],
     col=cl3$cluster,
     pch=16,
     main="K-means assignment")
legend("topleft", c("1", "2", "3"), pch = 16, col=c("black","green","red"))

plot(data[c("GrWt", "ca300KWt")],
     col=data$Type,
     pch=16,
     main= "Distribution of Types")
legend("bottomright", c("Inbred", "F1", "F2", "BC1"), pch = 16, col=c("blue","black","green","red"))
```

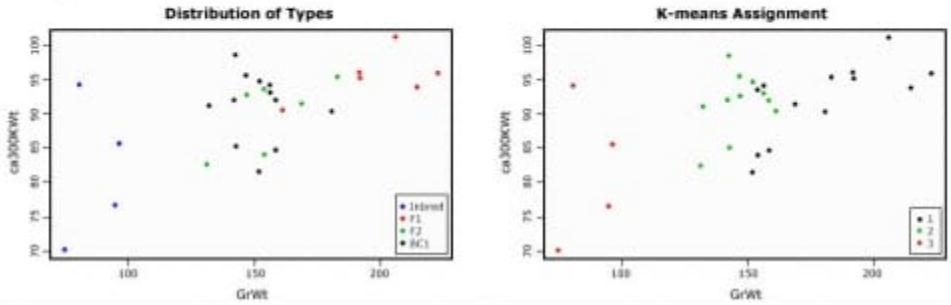



Figure 17 Graphical display based on the Grain Weight and 300 Kernel Weight, how the types are distributed compared to the k-means assigned groups, R code to generate plots is given above.

# Principal Components Analysis

The more variables there are in a multivariate dataset, the more difficult it becomes to describe and extract useful information

from it. Also, displaying the data in a graphic becomes harder when more than 3 variables are included. In this case, a Principal Component Analysis (PCA) helps to reduce the dimensionality of our dataset by changing the set of correlated variables we want to describe into a new set of uncorrelated variables. The new variables are sorted in order of importance, the first one accounting for the largest portion of variation found in the original dataset, the latter for less and less large portions of the variation. The hope is that a few of these new variables will be sufficient to describe most of the variability found in your original dataset, so we can replace our data with only a few variables and end up with less complexity and dimensionality.

In our first example, we will look at a dataset that has only two variables for an easy demonstration of PCA. We will use the following dataset: MeanEarPhenotypes [CSV], and create a subset consisting of the columns GrWt and ca300KWt only.

## PCA Step by Step

- Download the dataset: MeanEarPhenotypes [CSV], R-code: Example-PCA1 [TXT]

First, we will have to standardize our data; if the scales in your dataset are similar, we can use mean centering for standardization and do the subsequent calculation based on covariance. If the scales in a dataset are very different (weights, temperatures...), it is recommended to divide the mean center values by the standard deviation and use the correlation for subsequent calculations.

Head of the standardized data for the variables GrWt and ca300KWt

head(PCAdata)

GrWt ca300KWt

1 -0.3157250 0.23311021

2 -0.2982722 1.18856570

3 -0.2955165 -0.76213572

4 -0.1779398 0.75378250

5 0.7562440 -0.02213083

6 0.1371291 -0.83589358

#display data in a plot

plot(PCAdata[c(“GrWt”, “ca300KWt”)]
+ col-data$Type,
+ pch-16
+ main-“Distribution of data for PCA example”)

legend(“bottomright”, c(“Inbred”, “F1”, “BC1”), pch = 16, col=c(“blue”,”red”,”green”,”black”))

## Distribution of Data

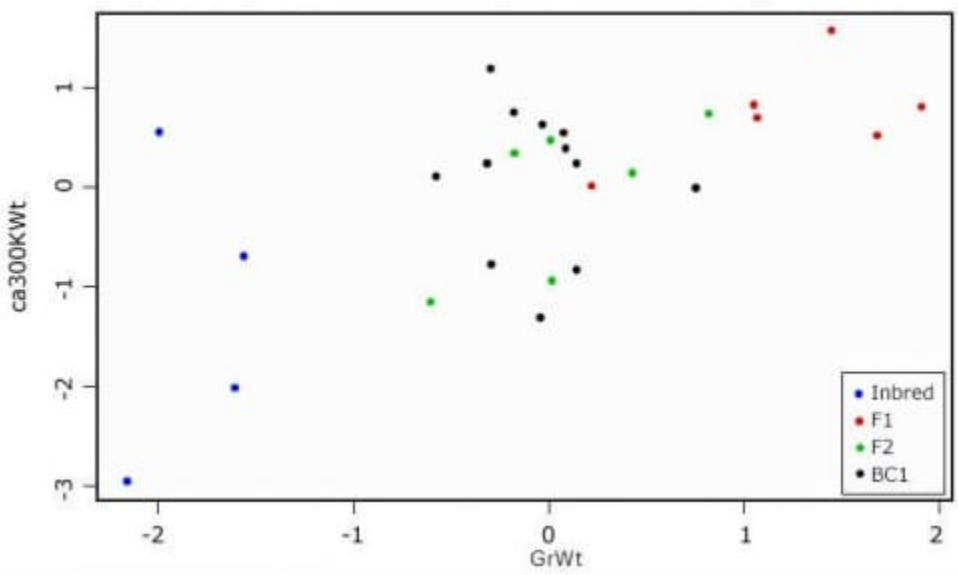



Figure 18 Scatterplot to display the standardized data for the PCA example.

Looking at the scatterplot (Figure 18), we see that there is some correlation between the two measures. If we want to know which of the variables contributes the most to the overall variance of the dataset, we have to calculate the Eigenvalues for the covariance (or correlation) matrix.

## Eigenvectors Output Matrix

my.eigen$values

[1] 1.6208572 0.3791428

rownames(my.eigen$vectors) <- c("GrWt", "ca300KWt")

colnames(my.eigen$vectors) <-c("PC1", "PC2")

my.eigen$vectors

PC1 PC2

GrWt -0.7071068 0.7071068

ca300KWt -0.7071068 -0.7071068

sum(my.eigen$values)

[1] 2

var(PCAdata$GrWt) + var(PCAdata$ca300KWt)

[1] 2

## I

In the output matrix of the Eigenvectors, the first column is also our first Principal Component and the second column is the second Principal Component. The values are a measure of the strength of association with the Principal Component. While values in the first Principal Component trend together, in the second, they trend apart. We will try to visualize this in a graphic.

## Display of Principal Components

PC1.slope <- my.eigen$vectors[1,1]/my.eigen$vectors[2,1]

PC2.slope <- my.eigen$vectors[1,2]/my.eigen$vectors[2,2]

abline(0, PC1.slope, col="green")

abline(0, PC2.slope, col="red")

Graphical display of the Principal Components. PC1 in green, PC2 in red—the two components are orthogonal to each other (Figure 19).

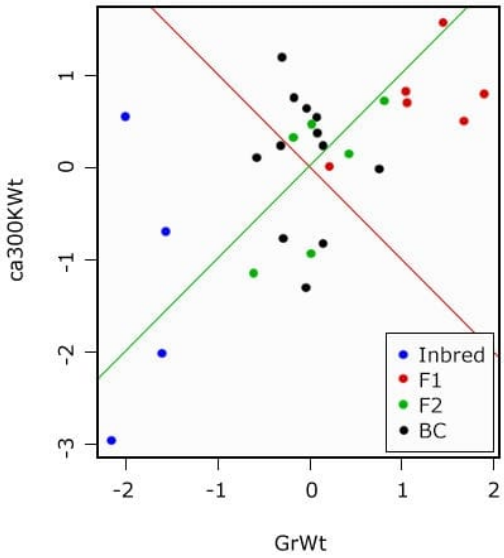



Figure 19 Distribution of data for PCA example

## Percentage of Overall Variance

Finally, we want to ask how much (or what percentage) of the overall variance in our data is represented by the first Principal Component:

PC1.slope <- my.eigen$vectors[1,1]/my.eigen$vectors[2,1]

PC2.slope <- my.eigen$vectors[1,2]/my.eigen$vectors[2,2]

abline(0, PC1.slope, col="green")

abline(0, PC2.slope, col="red")

PC2.var <-100 * (my.eigen$values[1]/sum(my.eigen$values))

PC2.var <-100 * (my.eigen$values[2]/sum(my.eigen$values))

PC1.var

[1] 81.04286

PC2.var

[1] 18.95714

|

PC1 explains ca. 81% of the variance, and PC2 explains about 19%.

## Calculate the PCA Scores

By multiplying the original variable with the respective Eigenvectors, we calculate the PCA scores for each sample unit in our dataset. Plotting the scores shows that the plot is rotated such that the Principal Components form the x and y axis (Figure 20). The relations between the sample units are unchanged, although one should be aware that a Principal Component per se has no biological meaning.

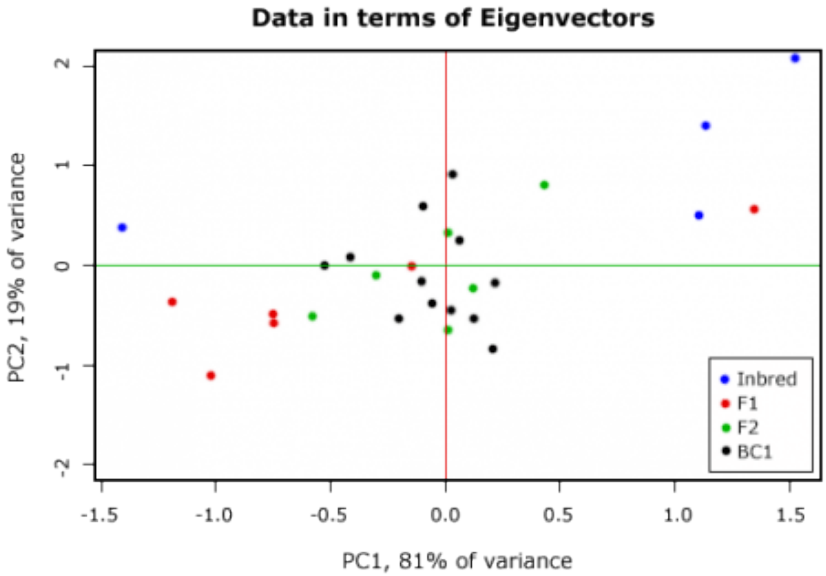



Figure 20 Plotting the PCI scores.

It is possible to plot the variables as vectors into the graphic (Figure 20)—how to do this and how to interpret the arrows will be shown in the next example.

## Perform a Principal Component Analysis

#### PERFORMING A PRINCIPAL COMPONENT ANALYSIS USING INBUILT R FUNCTIONS

- Download the dataset: MeanEarPhenotypes [CSV],
- Create a subset consisting of the columns: RowNo., EarDM, EarL, CobWt, ca300KWt (Figure 21); we will use the standardized data.
- Download the R code: R-code: Example-PCA2.R [TXT]

Example-HierarchicalClusterAnalysis.R | Example-kmeansClusterAnalysis.r | Example-PCA1.R | Example-PCA2.R*

Source on Save

```
1  ## load packages
2  library("gclus")
3  library("stats")
4  library("cluster")
5  library("clustersim")
6
7  # set working directory for module and check
8  setwd("C:/Users/ufrei/Desktop/WD-module")
9  getwd()
10
11 #load the MeanMeanEarPhenotypes dataset and look at the structure
12 data <- read.csv("MeanMeanEarPhenotypes.csv", header = T)
13 str(data)
14
15
16 #subsetting dataset to 5 variables
17 datap <- subset(data, select=c("RowNo.","EarDM","EarL", "CobWt", "ca300KWt"  ))
18 # and data standardization
19 PCAdata <-data.Normalization(datap, type="n1")
20
```

Figure 21 Preparing an example dataset with 5 variables.

## Generate a Scatterplot Matrix

With the cpair() function, we can generate a scatterplot matrix of our dataset (Figure 22).

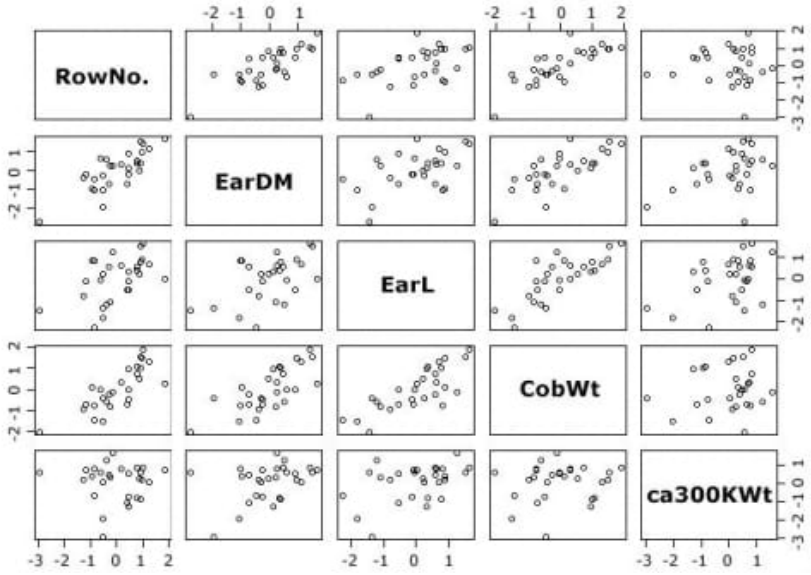



Figure 22 Scatterplot matrix for the 5 variables dataset.

Not unsurprising, for example, there is a strong correlation between the RowNo. and ear diameter.

## Calculate the Principal Components

To calculate the Principal Components, we will use the function prcomp() {stats}.

PCA2 <- prcomp(PCAdata, center=T, scale=T)

class(PCA2)

[1] “prcomp”

1s(PCA2)

[1] “center” “rotation” “scale” “sdev” “x”

summary(PCA2)

Importance of components:

PC1 PC2 PC3 PC4 PC5

Standard deviation 1.7858 1.0466 0.7404 0.33744 0.23121

Proportion of Variance 0.6378 0.2191 0.1096 0.02277 0.01069

Cumulative Proportion 0.6378 0.8569 0.9665 0.98931 1.00000

Looking at the results of prcomp, 5 vectors are listed; note that “x” stands for the Principal component scores. In the summary, each Principal component is assigned its standard deviation, the proportion of the overall variance, which is explained with this component as well as a cumulative proportion. The first three components explain already more than 96% of the overall variance in the dataset. As the main goal of the PCA is to reduce the complexity of the dataset, we could ask ourselves how many Principal components we have to keep.

## Loadings of the Principal Components

In order to see which variable contributes mainly to the Principal Components, you will have to look at the rotation data or loadings:

PCA2$rotation

PC1 PC2 PC3 PC4 PC5

RowNo. 0.4766670 -0.40713953 0.2610523 0.6767104 -0.2845007

EarDM 0.4964644 0.01002647 0.5805497 -0.3360285 0.5508806

EarL 0.4595622 0.22644933 -0.6767696 0.2657527 0.4570357

CobWT 0.5197882 -0.18893398 -0.2894435 -0.5855696 -0.5171605

ca300KWt 0.2119776 0.86438506 0.2302587 0.1250267 -0.3731665

While all variables contribute almost equally to the first component, the ca300KWt has a strong influence on the second component; in the third component, variables that measure the ear dimensions are prevalent. This can also be shown in a simple graphic (Figure 23).

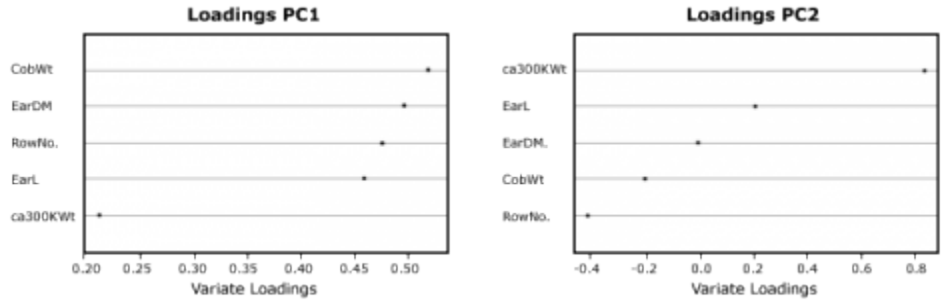



Figure 23 Loadings of the Principal components 1 and 2 in a dot plot.

## Scree Plot

```
>
> #calculate Eigen values for each component: square of standard deviation
> PCA2$sdev^2
[1] 3.18910550 1.09535795 0.54821611 0.11386400 0.05345645
> # Graphical display
> screeplot(PCA2, type="line", main="Scree Plot")
>
```

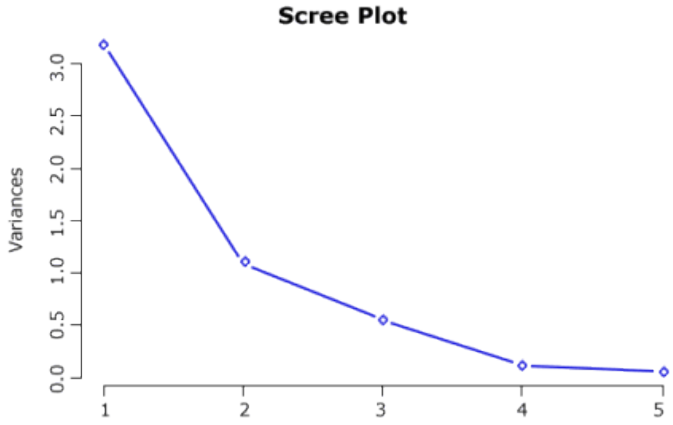



Figure 24 R-output for Eigen Values of Principal Components and Scree plot.

Coming back to the question of how many Principal Components should be considered – there are two ways to answer this question:

1. The Kaiser criterion: as long as the Eigen Value of a component is larger than 1, keep it (Figure 24).
2. Make a decision based on a Scree plot: keep components as long as the rate of change is still larger than 0.

So based on the Kaiser criterion, we would keep the first two components; based on the Scree plot, maybe the first three components would be kept; this is up to the researcher.

## Create a Biplot

Finally, we want to create a biplot to visualize the results of our PCA (Figure 25); the observations are plotted as points, and the variables as vectors:

```
#graphical display:
biplot(PCA2,cex =c(1, 0.7) )
biplot(PCA2,cex =c(1, 0.7),xlabs=data$Type )
```

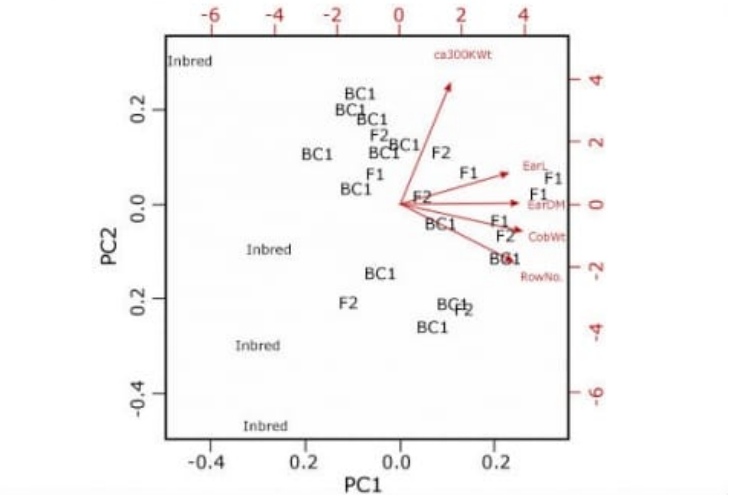



Figure 25 R-output for Eigen Values of Principal Components and Scree plot.

The biplot reveals what we saw looking at the loadings of the principal components: although trending in the same direction as

the other 4 variables, the ca300KWt vector is set apart from the other 4. The cosines of the angles between the vectors actually reflect the correlation between these variables.

**How to cite this chapter**: Frei, U., R. Howard, W. Beavis, and A. A. Mahama. 2023. Multivariate Analysis. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.15: Multivariate Analysis is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Ursula Frei; Reka Howard; William Beavis; and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.16: Algebra Review Guide

## Linear Equations, Formulas, and Inequalities

1. Eliminate denominators (multiply by Least Common Denominator)
2. Remove parentheses (distribute)
3. Get variable terms on one side (add/subtract principles)
4. Combine like terms (for formulas, factor the variable if it appears in more than 1 term)
5. Get the variable alone (multiply/divide principles)
   **Note:** For inequalities, division or multiplication by a negative number will switch the inequality symbol around
6. Check the solution

### Practice Problems

#### Query 1.16.1

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

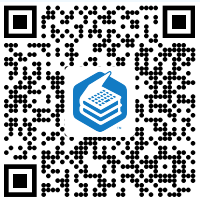
View this Interactive Content online



#### Query 1.16.2

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



#### Query 1.16.3

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



Query 1.16.4

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.16.5

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.16.6

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 1.16.7

*This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

## Terminology

Various terms are used in carrying out algebraic operations. A good understanding of these enables calculations to be carried out correctly. Some common terms are shown in Table 1.

*Table 1 Algebraic vocabulary terms and their definition.*

| Vocabulary | Definition |
|---|---|
| **Variable** | A letter that can be replaced by any number |
| **Constant** | A capital letter that represents a fixed number |
| **Expression** | Consists of variables, numbers and operation symbols (ex: $2x-5+3y^2-\frac{x}{3}$) |

| Vocabulary | Definition |
|---|---|
| **Equation** | When "=" is between two algebraic expressions<br>(equation can be true $x+2=x+2$, false $x+2=x+3$<br>or neither $\text{or } 3\text{x -}9=4\text{x - }14$) |
| **Solution** | A value replacing the variable to make the equation true<br>($x=5$ for $3x-9=4x-14$) |
| **Solved** | Having found the values that make the equation true |
| **Translate** | To convert words to an algebraic expression or equation |
| **Substitute** | Replacing a variable with a given number |
| **Evaluate** | To find a solution |
| **Factor** | Multiple or Product |
| **Factoring** | Reversing distributive law and turning it into the factors (multiples) |
| **Equivalent** | Expressions that, when evaluated, produce the same value |
| **Terms** | A number, var (variable), or a product/quotient of numbers and/or variables separated by + or – signs<br>(Expression $5-3xy-\frac{2x}{5}$ contains three terms) |
| **Fraction notation** | A way of showing the division of two numbers<br>(Numerator: top, Denominator: bottom) |
| **Undefined** | $\frac{a}{0}$: Division by zero is undefined — There is no solution |
| **Zero fraction** | $\frac{0}{a}$: Zero in the numerator makes fraction equal to zero |
| **Fraction notation for 1** | Any nonzero number divided by itself is 1: $\frac{a}{a}=1$ |
| **Reciprocal** | Multiplicative inverse: if $a\neq 0$, $\frac{a\cdot 1}{a}=\frac{a}{b}\cdot\frac{b}{a}=1$<br>Zero has no reciprocals. The reciprocal of a is $a=\frac{1}{a}$. |
| **Opposite** | Additive inverse: opposite of "*a*" is "-*a*" and $a+(-a)=0$. |
| **Prime number** | A natural number that is divisible by only two different factors: itself and 1 |
| **Inequality** | A statement about the relative size of two objects. Indicated by <, >, $\ll$, $\gg$ |
| **Law** | Definition |
| **Commutative** | For any real a and b, $a+b=b+a$, and<br>$a\cdot b=b\cdot a$ |
| **Associative** | For any real a, b, and c, $a\cdot(b+c)=a\cdot b+a\cdot c$ |
| **Distributive** | For any real a, b and c, $\text{a}+(\text{b}+\text{c})=(a+b)+c, a\cdot(b\cdot c)=(a\cdot b)\cdot c$ |

# Operation With Fractions

## Simplifying Fractions

1. Prime factorize each numerator and denominator
2. Remove factors that are the same from the numerator and denominator and replace them with "1".
   These form fractions that are equal to "1".
3. Multiply the remaining factors in the numerator
4. Multiply the remaining factors in the denominator

### Simplifying Fractions Example

$$\frac{60}{126} = 2^2 \cdot 3 \cdot \frac{5}{2} \cdot 3^2 \cdot 7 = \frac{10}{21} \tag{1.16.1}$$

## Multiplying Fractions

1. Prime factorize each numerator and denominator
2. Remove factors that are the same from any numerator and any denominator and replace them with "1". These form fractions that are equal to "1".
3. Multiply the remaining factors in all numerators
4. Multiply the remaining factors in all denominators

### Multiplying Fractions Example

$$\frac{7}{12} \cdot \frac{30}{21} = \frac{7}{2^2} \cdot 3 \cdot 2 \cdot 3 \cdot \frac{5}{3} = \frac{5}{6} \tag{1.16.2}$$

## Dividing Fractions

1. Change division to multiplication by the inverse of the second fraction
2. Multiply as above

### Dividing Fractions Example

$$\frac{4}{18} \div \frac{10}{21} = \frac{4}{18} \cdot \frac{21}{10} = \frac{2^2}{2} \cdot 3^2 \cdot 3 \cdot \frac{7}{2} \cdot 5 = \frac{1}{3} \cdot \frac{7}{5} = \frac{7}{15} \tag{1.16.3}$$

## Add and Subtract

1. If denominators are the same, go to step 4
2. Otherwise, prime factorize each denominator
3. Find the Least Common Denominator (LCD) by multiplying each fraction by a fraction equal to "1", made out of the missing factors from each denominator
4. Once denominators are the same, add/subtract numerators and write over the LCD
5. Simplify as above

### Adding/Subtracting Fractions Example 1: Same denominator

$$\frac{2}{14} + \frac{5}{14} = \frac{7}{14} = \frac{7}{2} \cdot 7 = \frac{1}{2} \tag{1.16.4}$$

---

### Adding/Subtracting Fractions Example 2: One denominator is a multiple of the other

$$\frac{3}{2} - \frac{5}{6} = \frac{3}{2} - \frac{5}{2} \cdot 3 = \frac{3}{2} \cdot \frac{3}{3} - \frac{5}{2} \cdot 3 = 9 - \frac{5}{2} \cdot 3 = \frac{4}{2} \cdot 3 = \frac{2^2}{2} \cdot 3 = \frac{2}{3} \tag{1.16.5}$$

---

### Adding/Subtracting Fractions Example 3: Different denominators

$$\frac{9}{42} - \frac{3}{15} = \frac{9}{2} \cdot 3 \cdot 7 - \frac{3}{3} \cdot 5 = \frac{9}{2} \cdot 3 \cdot 7 \cdot \frac{5}{5} - \frac{3}{3} \cdot 5 \cdot 2 \cdot \frac{7}{2} \cdot 7 = 45 - \frac{42}{2} \cdot 3 \cdot 5 \cdot 7 = 32 \cdot 3 \cdot 5 \cdot 7 = \frac{1}{2} \cdot 5 \cdot 7 = \frac{1}{70} \tag{1.16.6}$$

# Rules of Exponents

## Zero and One

- $a^0 = 1$
- $a^1 = a$

## Multiply and Divide

- $\frac{a}{b} \times \frac{a}{n} = a^{b+n}$
- $\frac{a^b}{a^n} = a^{b-n}$

## Distribute

- $(a^b c^n)^p = a^{bp} c^{np}$
- $(\frac{a^b}{c^n})^p = a^{bp} c^{np}$

## Negative

- $a^{-n} = \frac{1}{a^n}$
- $\frac{1}{a^{-n}} = a^n$
- $\frac{a^{-n}}{b^{-c}} = \frac{b^c}{a^n}$

# Rules of Radicals

## Add and Subtract Radicals

Simplify each radical by removing perfect square roots out of the radical to get like radicals, then combine the number of like radicals.

## Multiply and Divide Radicals

$$\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab} \tag{1.16.7}$$

$$\frac{\sqrt[n]{a}}{\sqrt[n]{b}} = \sqrt[n]{\frac{a}{b}} \tag{1.16.8}$$

If no *n* is given, assume that *n* is 2, and a square root is required.

# Plus/Minus Sign

Used to indicate that a value can be of either sign. Often used to construct confidence intervals.

$$p \pm x + y = p + x + y, \text{or p - x + y} \tag{1.16.9}$$

# Data Transformation

## Trigonometry and Natural Logarithms

*Table 2 Trigonometric transformation terms and their reciprocals.*

| Transformation | Reciprocal of Transformation |
|---|---|
| Sine (sin) | Cosecant (csc or cosec) or Arcsine (arcsin) |
| Cosine (cos) | Secant (sec) or Arccosine (arccos) |
| Tangent (tan) | Cotangent (cot) or Arctangent (arctan) |
| Arcsine (arcsin) | Sine (sin) |
| Natural logarithm (in or log or $\log_e$) | Exponential (exp or $e^n$) |

# Summation

Summation (S) signifies that a series of terms should be added together.

Given a series of numbers $x_1$, $x_2$, $x_3$ ... $x_n$

Sum all $x_i$ starting at $i = 1$ through $i = n$.

$$\sum_{i=1}^{n} x^i = x_1 + x_2 + x_3 + x_4 + ... + x_n \tag{1.16.10}$$

$$\text{Equation 1} \tag{1.16.11}$$

**How to cite this chapter:** Kendra Meade and A. A. Mahama. 2023. Algebra Review Guide. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Methods*. Iowa State University Digital Press.

This page titled 1.16: Algebra Review Guide is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Kendra Meade and Anthony Assibi Mahama (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# 1.17: Applied Learning Activities

2

The following downloadable Applied Learning Activities (ALAs) and recommended readings are associated with the chapters linked below. These files cover the use of R applications, as discussed in chapters 8 onward.

*Additional exercises in R format will be added later. Stay tuned!*

## Chapter 8

### Files for Associated Learning Activities

- CRD 1 Data [CSV]
- CRD 2 Data_[CSV]

## Chapter 9

### Files for Associated Learning Activities

- ANOVA 2-factor CRD [CSV]
- Exercise 9.1 larger set [CSV]
- Randomization 2-factor CRD [CSV]

## Chapter 10

### Files for Associated Learning Activities

- Exercise 10.2 Data [CSV]

## Chapter 11

### Files for Associated Learning Activities

- Exercise 11.2 Data [CSV]
- Exercise 11.3 Data [CSV]

## Chapter 12

### Files for Associated Learning Activities

- Exercise 12.1 Data [CSV]
- QM Chapter 12 ALA 12.3 Set 1 [CSV]

# Chapter 13

## Files for Associated Learning Activities

- QM Chapter 13 Barley Data [CSV]
- Exercise 13.2 Data [CSV]
- Exercise 13.3 [CSV]
- Exercise 13.4 [CSV]
- Exercise 13.5 [CSV]

# Chapter 14

## Files for Associated Learning Activities

- QM Chapter 14 Exercise 1 [CSV]
- QM Chapter 14 Exercise 4 [CSV]

# Chapter 15

## Files for Associated Learning Activities

- QM Chapter 15 Exercise 1 [CSV]
- QM Chapter 15 Exercise 2 [CSV]
- QM Chapter 15 Exercise 3 [CSV]
- QM Chapter 15 Exercise 4 [CSV]
- QM Chapter 15 15.4 RawDataEarPhenotypes
- QM Chapter 15 15.5 MeanMeanEarPhenotypes

This page titled 1.17: Applied Learning Activities is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# INDEX


## A

agricultural analysis
*1.13: Multiple Regression*
agricultural research
*1.10: Mean Comparisons*
*1.9: Two Factor ANOVAs*
agricultural science
*1.8: The Analysis of Variance (ANOVA)*
agriculture
*1.1: Basic Principles*
agronomy
*1.14: Nonlinear Regression*
ALAs
*1.17: Applied Learning Activities*
algebra
*1.16: Algebra Review Guide*
algorithms
*1.15: Multivariate Analysis*
analysis
*1.14: Nonlinear Regression*
*1.15: Multivariate Analysis*
analysis of variance
*1.12: Data Transformation*
AnalysisOfVariance
*1: Chapters*
ANOVA
*1.10: Mean Comparisons*
*1.11: Randomized Complete Block Design*
*1.12: Data Transformation*
*1.8: The Analysis of Variance (ANOVA)*
*1.9: Two Factor ANOVAs*
appliedlearning
*1.17: Applied Learning Activities*

## B

BasicPrinciples
*1: Chapters*
Bernoulli trial
*1.4: Categorical Data - Binary*
binomial
*1.4: Categorical Data - Binary*
binomial probability
*1.4: Categorical Data - Binary*
blocking
*1.11: Randomized Complete Block Design*

## C

CategoricalData
*1: Chapters*
central limit theorem
*1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests*
CentralLimitTheorem
*1: Chapters*
clustering
*1.15: Multivariate Analysis*
confidence intervals
*1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests*
*1.6: Continuous Data*
ConfidenceIntervals
*1: Chapters*
constant
*1.16: Algebra Review Guide*
contingency tables
*1.5: Categorical Data Multivariate*
corn
*1.9: Two Factor ANOVAs*
correlation
*1.13: Multiple Regression*
*1.7: Linear Correlation, Regression and Prediction*
correlation coefficient
*1.7: Linear Correlation, Regression and Prediction*
crop
*1.14: Nonlinear Regression*
crop varieties
*1.1: Basic Principles*
CSV
*1.17: Applied Learning Activities*
cultivars
*1.11: Randomized Complete Block Design*

## D

data
*1.14: Nonlinear Regression*
data analysis
*1.15: Multivariate Analysis*
*1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests*
*1.5: Categorical Data Multivariate*
*1.7: Linear Correlation, Regression and Prediction*
*1.8: The Analysis of Variance (ANOVA)*
*1.9: Two Factor ANOVAs*
data collection
*1.1: Basic Principles*
data transformation
*1.12: Data Transformation*
data visualization
*1.13: Multiple Regression*
dataanalysis
*1.17: Applied Learning Activities*
datasets
*1.15: Multivariate Analysis*
*1.17: Applied Learning Activities*
dependent variable
*1.13: Multiple Regression*
descriptive statistics
*1.2: Distributions and Probability*
discrete data
*1.4: Categorical Data - Binary*
distance matrix
*1.15: Multivariate Analysis*
distribution
*1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests*
*1.4: Categorical Data - Binary*

## E

Education
*1.17: Applied Learning Activities*
environmental science
*1.7: Linear Correlation, Regression and Prediction*
equation
*1.16: Algebra Review Guide*
exercises
*1.17: Applied Learning Activities*
experimental design
*1.11: Randomized Complete Block Design*
*1.12: Data Transformation*
*1.2: Distributions and Probability*
*1.9: Two Factor ANOVAs*
experimental research
*1.1: Basic Principles*
experimental units
*1.6: Continuous Data*
exponents
*1.16: Algebra Review Guide*
expression
*1.16: Algebra Review Guide*

## F

factor
*1.16: Algebra Review Guide*
failure
*1.4: Categorical Data - Binary*
field trial
*1.8: The Analysis of Variance (ANOVA)*

## G

growth
*1.14: Nonlinear Regression*

## H

homogeneity of variances
*1.12: Data Transformation*
hybrids
*1.9: Two Factor ANOVAs*
hypothesis testing
*1.1: Basic Principles*
*1.10: Mean Comparisons*
*1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests*
*1.4: Categorical Data - Binary*
*1.5: Categorical Data Multivariate*
*1.6: Continuous Data*
*1.8: The Analysis of Variance (ANOVA)*
HypothesisTests
*1: Chapters*

## I

independence
*1.12: Data Transformation*
independent variables
*1.13: Multiple Regression*
inequalities
*1.16: Algebra Review Guide*

interaction
1.9: Two Factor ANOVAs

L

least significant difference
1.10: Mean Comparisons
linear regression
1.7: Linear Correlation, Regression and Prediction
linear_equations
1.16: Algebra Review Guide
LinearCorrelation
1: Chapters
LSD
1.10: Mean Comparisons

M

mean
1.2: Distributions and Probability
mean comparison
1.10: Mean Comparisons
multiple comparison procedures
1.10: Mean Comparisons
multivariate
1.15: Multivariate Analysis
MultivariateAnalysis
1: Chapters

N

nonlinear
1.14: Nonlinear Regression
normal distribution
1.2: Distributions and Probability
1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests
normality
1.12: Data Transformation
null hypothesis
1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests
1.5: Categorical Data Multivariate

P

plant breeding
1.1: Basic Principles
plants
1.14: Nonlinear Regression
populations
1.2: Distributions and Probability
prediction
1.13: Multiple Regression
predictive analysis
1.7: Linear Correlation, Regression and Prediction
probability
1: Chapters
1.2: Distributions and Probability
1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests
proportion
1.4: Categorical Data - Binary
proportions
1.5: Categorical Data Multivariate

R

random distribution
1.12: Data Transformation
randomization
1.11: Randomized Complete Block Design
randomized complete block design
1.11: Randomized Complete Block Design
Rapplications
1.17: Applied Learning Activities
regression
1: Chapters
1.13: Multiple Regression
1.14: Nonlinear Regression
1.7: Linear Correlation, Regression and Prediction
relationships
1.14: Nonlinear Regression
research design
1.8: The Analysis of Variance (ANOVA)

S

sample size
1.6: Continuous Data
sampling
1.2: Distributions and Probability
1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests
scatter plot
1.7: Linear Correlation, Regression and Prediction
scientific method
1.1: Basic Principles
1.2: Distributions and Probability
significance
1.5: Categorical Data Multivariate
significance testing
1.13: Multiple Regression
solution
1.16: Algebra Review Guide
standard deviation
1.2: Distributions and Probability
1.6: Continuous Data
statistical analysis
1.1: Basic Principles
1.11: Randomized Complete Block Design
1.12: Data Transformation
1.15: Multivariate Analysis
1.2: Distributions and Probability
1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests
1.7: Linear Correlation, Regression and Prediction
1.8: The Analysis of Variance (ANOVA)
1.9: Two Factor ANOVAs
statistical methods
1.13: Multiple Regression
statistical modeling
1.14: Nonlinear Regression
statistical science
1.1: Basic Principles
statistical significance
1.11: Randomized Complete Block Design
statistical tests
1.10: Mean Comparisons
1.5: Categorical Data Multivariate
statistics
1.17: Applied Learning Activities
1.4: Categorical Data - Binary
1.6: Continuous Data
strength of relationship
1.7: Linear Correlation, Regression and Prediction
success
1.4: Categorical Data - Binary

T

treatment
1.11: Randomized Complete Block Design
treatment effects
1.12: Data Transformation
treatment means
1.6: Continuous Data
treatment response
1.10: Mean Comparisons
treatments
1.9: Two Factor ANOVAs

V

variable
1.16: Algebra Review Guide
variables
1.15: Multivariate Analysis
variance
1.6: Continuous Data
1.8: The Analysis of Variance (ANOVA)
visualization
1.15: Multivariate Analysis

Y

yield
1.9: Two Factor ANOVAs
yield comparison
1.8: The Analysis of Variance (ANOVA)
yield measurement
1.1: Basic Principles
yield trial
1.11: Randomized Complete Block Design

This page titled Index is shared under a license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# GLOSSARY

## S

**Sample Word 1** | Sample Definition 1

This page titled Glossary is shared under a license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey (Iowa State University Digital Press) via source content that was edited to the style and standards of the LibreTexts platform.

# DETAILED LICENSING

## Overview

**Title:** Quantitative Methods for Plant Breeding (Suza and Lamkey)

**Webpages:** 30

**Applicable Restrictions:** Noncommercial

**All licenses found:**

- CC BY-NC 4.0: 66.7% (20 pages)
- Undeclared: 33.3% (10 pages)

## By Page

- Quantitative Methods for Plant Breeding (Suza and Lamkey) - *CC BY-NC 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
    - About the PBEA Series - *CC BY-NC 4.0*
    - Contributors - *CC BY-NC 4.0*
  - 1: Chapters - *Undeclared*
    - 1.1: Basic Principles - *CC BY-NC 4.0*
    - 1.2: Distributions and Probability - *CC BY-NC 4.0*
    - 1.3: Central Limit Theorem, Confidence Intervals, and Hypothesis Tests - *CC BY-NC 4.0*
    - 1.4: Categorical Data - Binary - *CC BY-NC 4.0*
    - 1.5: Categorical Data Multivariate - *CC BY-NC 4.0*
    - 1.6: Continuous Data - *CC BY-NC 4.0*
    - 1.7: Linear Correlation, Regression and Prediction - *CC BY-NC 4.0*
    - 1.8: The Analysis of Variance (ANOVA) - *CC BY-NC 4.0*
    - 1.9: Two Factor ANOVAs - *CC BY-NC 4.0*
    - 1.10: Mean Comparisons - *CC BY-NC 4.0*
    - 1.11: Randomized Complete Block Design - *CC BY-NC 4.0*
    - 1.12: Data Transformation - *CC BY-NC 4.0*
    - 1.13: Multiple Regression - *CC BY-NC 4.0*
    - 1.14: Nonlinear Regression - *CC BY-NC 4.0*
    - 1.15: Multivariate Analysis - *CC BY-NC 4.0*
    - 1.16: Algebra Review Guide - *CC BY-NC 4.0*
    - 1.17: Applied Learning Activities - *CC BY-NC 4.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*

This page titled Quantitative Methods for Plant Breeding (Suza and Lamkey) is shared under a CC BY-NC 4.0 license and was authored, remixed, and/or curated by Walter Suza and Kendall Lamkey via source content that was edited to the style and standards of the LibreTexts platform.