

Data Science for Agricultural Professionals

Marin L. Harbur

2023-01-14

Table of Contents

Preface	7
Welcome	7
R-language	8
1 Population Statistics	11
1.1 Populations	11
1.2 Case Study: Yield Map	12
1.3 Distributions	14
2 Distributions and Probability	29
2.1 Case Study	29
2.2 The Normal Distribution Model	31
2.3 The Z-Distribution	36
3 Sample Statistics	41
3.1 Samples	42
3.2 Case Study	42
3.3 Distribution of Sample Means	45
3.4 Central Limit Theorem	48
3.5 Standard Error	48
3.6 Degrees of Freedom	49
3.7 The t-Distribution	50
3.8 Confidence Interval	57
3.9 Confidence Interval and Probability	60

4 Two-Treatment Comparisons	63
4.1 Side-by-Side Trials	63
4.2 Blocked Design	65
4.3 Case Study	67
4.4 Confidence Interval	71
4.5 T-Test	73
4.6 Conclusion	74
5 Understanding Statistical Tests	75
5.1 Research Question	75
5.2 The Model	76
5.3 Hypotheses	82
5.4 P-Value	83
5.5 The P-Value and Errors	84
5.6 One-Sided vs Two-Sided Hypotheses	86
6 Multiple Treatment Trials	91
6.1 Case Study	92
6.2 The Linear Additive Model	93
6.3 Analysis of Variance	97
6.4 The F statistic	98
6.5 The ANOVA Table	98
6.6 Visualizing How the Anova Table Relates to Variance	102
7 Multiple Treatment Designs	103
7.1 Randomized Complete Block Design	104
7.2 Factorial Design	108
7.3 Split-Plot Design	122
7.4 Linear Additive Model	124
7.5 Conclusion	126

<i>TABLE OF CONTENTS</i>	5
8 Means Separation and Data Presentation	127
8.1 Case Study	127
8.2 Least Significant Difference	128
8.3 LSD Output in R	129
8.4 Comparisonwise versus Experimentwise Error	132
8.5 Tukey's Honest Significant Difference	133
8.6 Linear Contrast	135
8.7 Means Presentation	142
9 Messy and Missing Data	149
9.1 Inspecting data for Normal Distributions	150
9.2 Inspecting Data for Equal Variances	156
9.3 Dealing with Messy Data	161
9.4 Dealing with Missing Data	165
9.5 Summary	171
10 Correlation and Simple Regression	173
10.1 Correlation versus Regression	174
10.2 Correlation	176
10.3 Regression	183
10.4 Extrapolation	195
11 Nonlinear Relationships and Multiple Linear Regression	199
11.1 Multiplie Linear Regression	200
11.2 Nonlinear Relationships	211
11.3 Summary	222
12 Spatial Statistics	225
12.1 Projection (General)	226
12.2 Shape Files	235
12.3 Rasters	244

13 Machine Learning	255
13.1 Machine Learning	255
13.2 Cluster Analyses	256
13.3 k-Nearest-Neighbors	268
13.4 Classification Trees	286
13.5 Summary	300
14 Putting it all Together	301
14.1 Scenario 1: Yield Map (Population Summary and Z-Distribution)	301
14.2 Scenario 2: Yield Estimate (Sampling t-Distribution)	304
14.3 Scenario 3: Side-By-Side (t-Test)	305
14.4 Scenario 4: Fungicide Trial (ANOVA CRD or RCB)	307
14.5 Scenario 5: Hybrid Response to Fungicide Trial (ANOVA Factorial or Split Plot)	310
14.6 Scenario 6: Foliar Rate-Response Trial (Linear or Non-Linear Regression)	313
14.7 Scenario 7: Application Map (Shapefiles and Rasters)	317
14.8 Scenario 8: Yield Prediction (Multiple Linear Regression and other Predictive Models)	320
14.9 Summary	327

Preface

Welcome

Welcome to Data Science for Agricultural Professionals. I have written these course materials for a few reasons. First, it is my job. The more powerful motivation, however, was to write a guide that satisfied the following criteria:

- covers basic statistics used in reporting results from hybrid trials and other controlled experiments
- also addresses data science tools used to group environments and make predictions
- introduced students to R, an open-source statistical language that you can use after your studies at Iowa State University and can use without installing on your laptop, or using a VPN connection, which your work laptop may not allow.

I also wanted to develop a text that presented statistics around the situations in which you are most likely to encounter data:

- yield maps used by farmers and agronomists
- side-by-side or split-field trials used often at the retail level
- smaller-plot controlled experiments used in seed breeding and other product development
- rate recommendation trials for fertilizers and crop protection products
- fertilizer prediction maps
- decision support tools

I began my career as an a university researcher and professor, but in 2010 entered the private sector, working first in retail as a technical agronomist for a regional cooperative and then as a data scientist for a major distributor, developing product data and insights for a team of researchers and agronomists. In seeing how data were used at the retail and industry levels, I gained an appreciation for what areas of statistics were more often used than others.

What is important to the agricultural professional, in my experience, is data literacy and a basic ability to run analyses. It is easy, after years in the field, to lose skills gained as an undergraduate. My hope is that all of you upon completing this course will look at statistics you receive (from any source, but especially manufacturers) a little more critically. If you are involved in field research, I hope you will understand how to better layout and more creatively analyze field trials

I wanted to develop a reference that appreciated not all of us are mathematical

progedies, nor do we have flawless memories when it comes to formula. At the risk of redudancy, I will re-explain formulas and concepts – or at least provide links – so you aren't forever flipping back and forth trying to remember sums of squares, standard errors, etc. I am committed to making this as painless as possible.

R-language

In this course, we will use the R language to summarise data, calculate statistics, and view trends and relationships in data. R is open-source (free) and incredibly versatile. I have used multiple languages, including SAS and SPSS, in my career; in my opinion, R is not only the cheapest, but the best, and I now use it exclusively and almost daily in my work.

R also has personal connections to Iowa State: Hadley Wickam, Carson Seivert, two authors of the R language, are Iowa State graduates.

We will use an application called R-Studio to work with R language. R Studio allows you to write and save scripts (code) to generate analyses. It also allows you to intersperse Markdown (html code) in between your statistical output so that you can explain and discuss your results. The beauty of Markdown is your report and your statistics are together in a single file; no cutting and pasting is needed between documents.

R is all about having options, and so with R-Studio you have the option of installing it on your laptop or, in case you are using a work laptop without administrative priviledges, you can also use a cloud site, R-Studio Cloud, to work with R and save your work.

I know that for most of you R (and coding itself) will be a new experience. I am sure the idea of coding will intimidate many of you. To head off your anxiety as much as possible, I offer this: I understand that coding is challenging. I spend my days making mistakes, searching for bugs, and looking up how to do something for the umpteenth time. If something confuses you, that is normal.

But I can also assure you that you will likely have an easy time remembering what functions to use. In other words, which code to use will not be the problem. Most of your bugs will be due to misspellings, forgetting to close a parentheses, or referring to a dataset by the wrong name, e.g. “soy_data” instead of “soybean_data”. And you will get better at avoiding these mistakes over time – there is not more to learn, just practice. Our exercises in R will be designed to give you that practice, and introduce new functions as slowly as possible.

At the end of this course, you will have not only your completed work, but the course materials themselves as a resource from which you can borrow code for future projects. There is no shame in copying lines of codes (or whole chunks

of code) into your own original analyses. All of us data scientists do that, and it is one of the best ways to continue learning.

R is supported by many great books which may be accessed for free online, or purchased online for very reasonable prices. These may be used as references during the course, but also to continue learning for years to come. These include many references from bookdown.org:

- Introduction to Data Science: although there are many “Introduction to R” books, this one closely matches how we approach it in Agronomy 513. (<https://rafalab.github.io/dsbook/>)
- R Graphics Cookbook: a comprehensive explanation of how to create just about any plot you could ever imaging in R, mainly using the ggplot package. (<https://r-graphics.org/>)
- Geocomputation with R: The best book I have found for learning how to work with spatial data. (<https://geocompr.robinlovelace.net/>)
- Hands-On Machine Learning with R: I have not read this book, but it appears to be a robust and beautifully-illustrated guid to machine learning concepts in R. (<https://bradleyboehmke.github.io/HOML/>)

Chapter 1

Population Statistics

1.1 Populations

Almost every statistics text begins with the concept of a **population**. A population is the complete set of individuals to which you want to predict values. Let's dwell on this concept, as it is something that did not hit home for me right away in my career. Again, the population is all of the individuals for which you are interested in making a prediction. What do we mean by individuals? Not just people – individuals can plants, insects, disease, livestock or, indeed, farmers.

Just as important as what individuals are in a population is its extent. What do you want the individuals to represent? If you are a farmer, do you want to apply the data from these individuals directly to themselves, or will you use them to make management decisions for the entire field, or all the fields in your farm? Will you use the results this season or to make predictions for future seasons? If you are an animal nutritionist, will you use rations data to support dairy Herefords, or beef Angus?

If you are a sales agronomist, will you use the data to support sales on one farm, a group of farms in one area, or across your entire sales territory? If you are in Extension, will the individuals you measure be applicable to your entire county, group of counties, or state? If you are in industry like me, will your results be applicable to several states?

This is a very, very critical question, as you design experiments – or as you read research conducted by others. To what do or will the results apply? Obviously, an Iowa farmer should not follow the optimum planting date determined for a grower in North Carolina, nor should an Ohio farmer assume our pale clays will be as forgiving as the dark, mellow loam of southern Minnesota.

Drilling down, you might further consider whether research was conducted in

areas that have similar temperature or rainfall, how different the soil texture might be to the areas to which you want to apply results. At the farm level, you might ask how similar the crop rotation, tillage, or planting methods were to that operation. At the field level, you might wonder about planting date or the hybrid that was sown.

When you use the results from set of individuals to make predictions about other individuals, you are making inferences – you are using those data to make predictions, whether it be for that same field next month or next year, or for other locations (areas in a field, fields in a farm, counties, or states). When we speak of an inference space, then, that is the total group of individuals to which you will apply your results. Or, in another word, your population.

In summary, one of the most critical skills you can apply with data science has no formula and, indeed, little to do with math (at least in our application). It is to ask yourself, will my experiment represent the entire population in which I am interested? Indeed, one field trial likely will not address the entire population in which you are interested – it is up to you to determine the population to which you are comfortable applying those results.

In fact, statistics or data science done without this “domain” knowledge whether a dataset is useful or experimental results are reasonable can be disastrous. Just because a model fits one dataset very well or treatments are significantly different does not mean they should be used to make decisions. Your competency as an agronomic data scientist depends on everything you learn in your program of study. Soil science, crop physiology, and integrated pest management, to name just a few subjects, are as much a prerequisite as any math course you have taken.

In some cases all of the individuals in a population can be measured – in such case, we will use the basic statistics described in this unit. The yield map we will analyze in this unit is a loose example of a case where we can measure

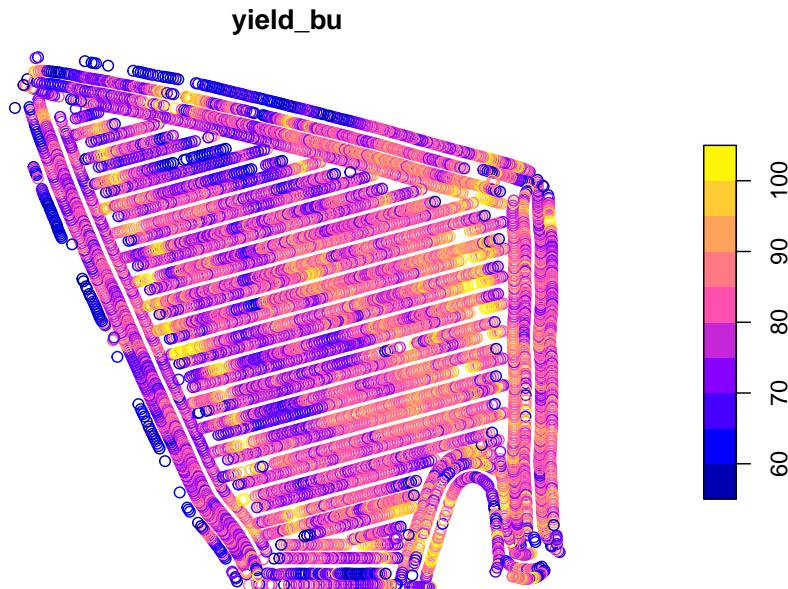
In most cases, however, it is not physically or financially feasible to measure all individuals in a population. In that case, subsets of the population, called samples, are used to estimate the range of individuals in a population.

1.2 Case Study: Yield Map

For our first case study, we will use a situation where every individual in our population can be measured: a single soybean field in central Iowa. In this case, yield data were gathered using a combine monitor. In case you don’t live and breathe field crops like me, combines (the machines that harvest grain) are usually equipped with a scale that repeatedly weighs grain as the combine moves across the field. The moisture of the grain is also recorded. These data are combined with measures of the combine speed and knowledge of the number

of rows harvested at once to calculate the yield per area of grain, adjusted to the market standard for grain moisture.

Almost all of you have seen a yield map somewhat like the one below. In this map, blue circles represent lower yields, while yellow and orange circles represent higher yields.



We will learn in the Exercises portion of this lesson how to create a map like this using just a few lines of code.

Each dataset has a structure – the way data are organized in columns and rows. To get a sense of the structure of our soybean dataset, we can examine the first 6 rows of the dataset using R.

```
##  
## Attaching package: 'kableExtra'  
  
## The following object is masked from 'package:dplyr':  
##  
##     group_rows
```

DISTANCE	SWATHWIDTH	VRYIELDVOL	Crop	WetMass	Moisture	Time
0.9202733	5	57.38461	174	3443.652	0.00	9/19/2016 4:45:46
2.6919269	5	55.88097	174	3353.411	0.00	9/19/2016 4:45:48
2.6263101	5	80.83788	174	4851.075	0.00	9/19/2016 4:45:49
2.7575437	5	71.76773	174	4306.777	6.22	9/19/2016 4:45:51
2.3966513	5	91.03274	174	5462.851	12.22	9/19/2016 4:45:54
3.1840529	5	65.59037	174	3951.056	13.33	9/19/2016 4:45:55

Each row in the above dataset is a **case**, sometimes called an **instance**. It is an single observation taken within a soybean field. That case may contain one or more **variables**: specific measurements or observations recorded for each case. In the dataset above, variables include *DISTANCE*, *SWATHWIDTH*, *VRYIELDVOL*, *Crop*, *WetMass*, and many others.

The two most important to us in this lesson are *yield_bu* and *geometry*. That this dataset has a column named *geometry* indicates it is a special kind of dataset called a **shape file** – a dataset in which the measures are geo-referenced. That is, we know where on Earth these measurements were taken. The *geometry* column in this case identifies a point with each observation.

1.3 Distributions

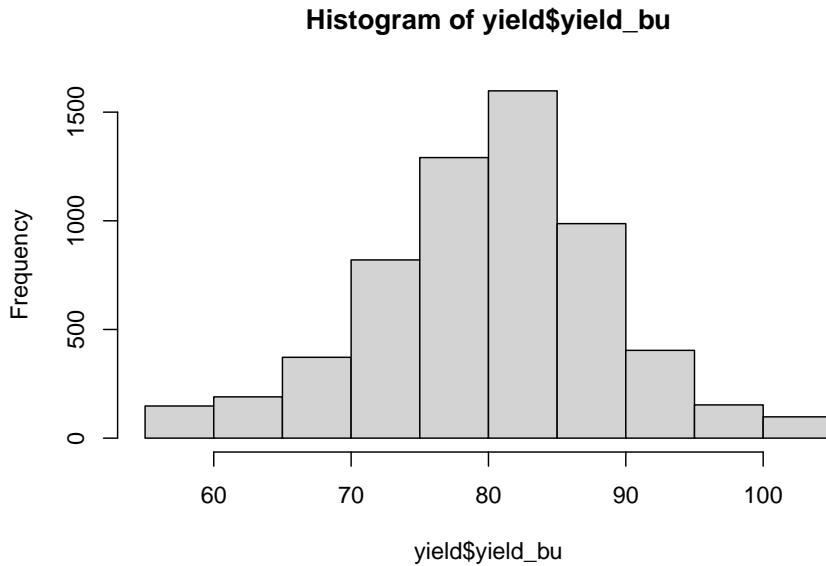
At this point, we have two options any time we want to know about soybean yield in this field. We can pull out this map or the complete dataset (which has over 6,500 observations) and look at try to intuitively understand the data. Or we can use statistics which, in a sense, provide us a formula for approximating the values in our dataset with just a few numbers.

A **distribution** describes the range of values that occur within a given variable. What is the range of values in our measured values? In this example, what are the highest and lowest yields we observed? What ranges of values occur more frequently? Many times, we want to see whether the distribution of one

1.3.1 Histograms

Before we get into the math required to generate these statistics, however, we should look at the shape of our data. What is the range of values in our measured values? In this example, what are the highest and lowest yields we observed? What ranges of values occur more frequently? Do the observed values make general sense?

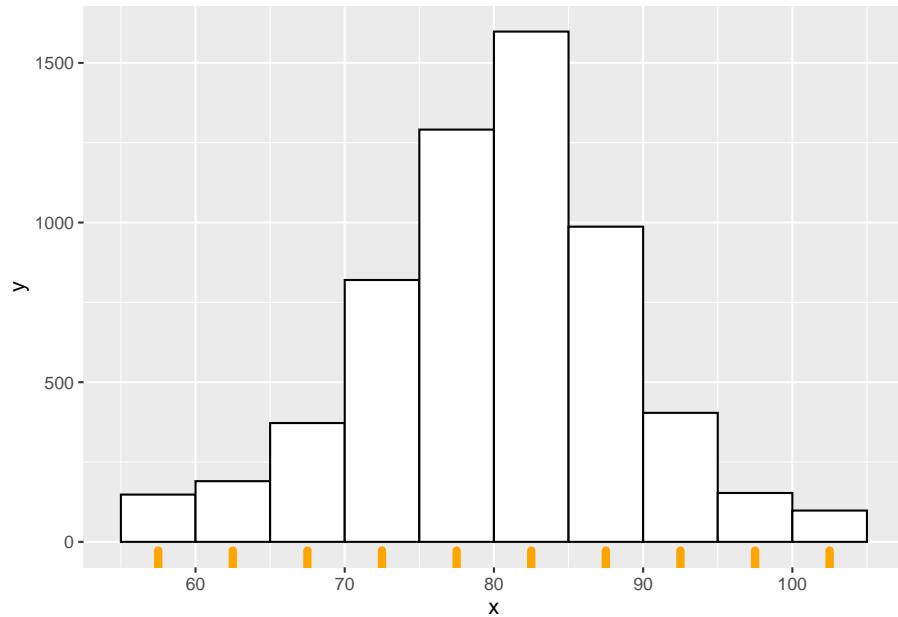
One of the easiest and most informative things for us to do is to create a particular bar chart known as a **histogram**.



In the histogram above, each bar represents range of values. This range is often referred to as a *bin*. The lowest bin includes values from 50 to 59.0000. The next bin includes values from 60 to 69.9999. And so on. The height of each bar represents the **frequency** within each range: the number of individuals in that population that have values within that range.

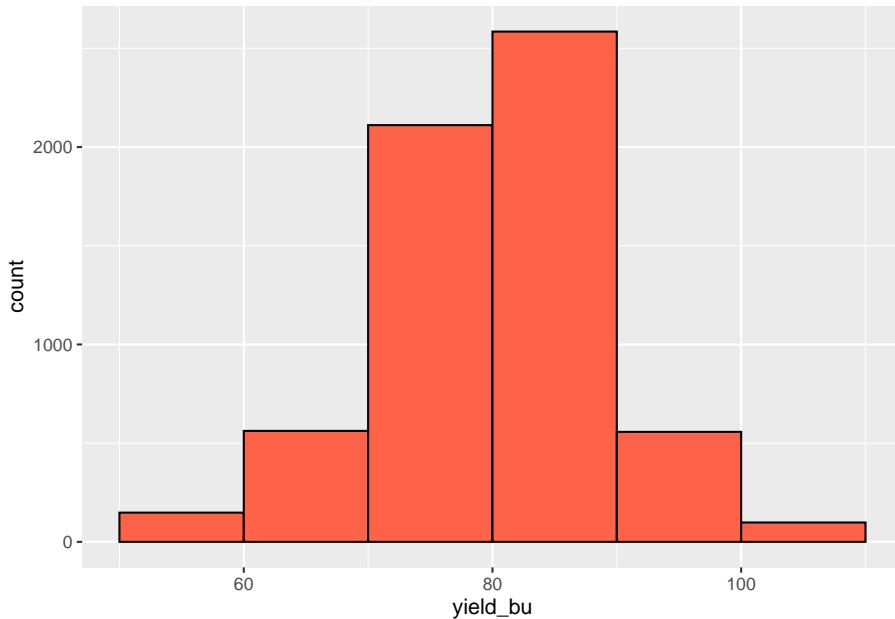
Each bin can also be defined by its **midpoint**. The midpoint is the middle value in each range. For the bin that includes values from 50 to 59.9999, the midpoint is 55. For the bin that includes values from 60 to 69.9999, the midpoint is 65.

In the plot above, the midpoint for each bar is indicated by the orange bar beneath it.



There are many ways in which we can draw a histogram – and other visualizations – in R. We will learn more in this course about an R package called `ggplot2`, which can create just about any plot you might imagine. Here is a simple taste:

```
ggplot(data=yield, aes(x=yield_bu)) +  
  geom_histogram(breaks=seq(50, 110, 10), fill="tomato", color="black")
```



Varying the *bin width* provides us with different perspectives on our distribution. Wide bins, which each include a greater range of values, will provide more gross representations of the data, while narrower bins will provide greater detail. When bins are too narrow, however, there may be gaps in our histogram where no values occur within particular bins.

Throughout this course, I have created interactive exercises to help you better visualize statistical concepts. Often, they will allow you to observe how changing the variables or the number of observations can affect a statistical test or visualization.

These exercises are located outside of this textbook. To access them, please follow links like that below. The exercises may take several seconds to launch and run in your browser. I apologize for their slowness – this is the best platform I have found to date.

Please click on the link below to open an application where you can vary the bin width and see how it changes your perspective:

<https://marin-harbur.shinyapps.io/01-app-histogram/>

1.3.2 Percentiles

We can also use **percentiles** to describe the values of a variable more numerically. Percentiles describe how values the proportional spread of values, from lowest to highest, within a distribution. To identify percentiles, the data are

numerically ordered (ranked) from lowest to highest. Each percentile is associated with a number; the percentile is the percentage of all data equal to or less than that number. We can quickly generate the 0th, 25th, 50th, and 75th, and 100th percentile in R:

```
summary(yield$yield_bu)
```

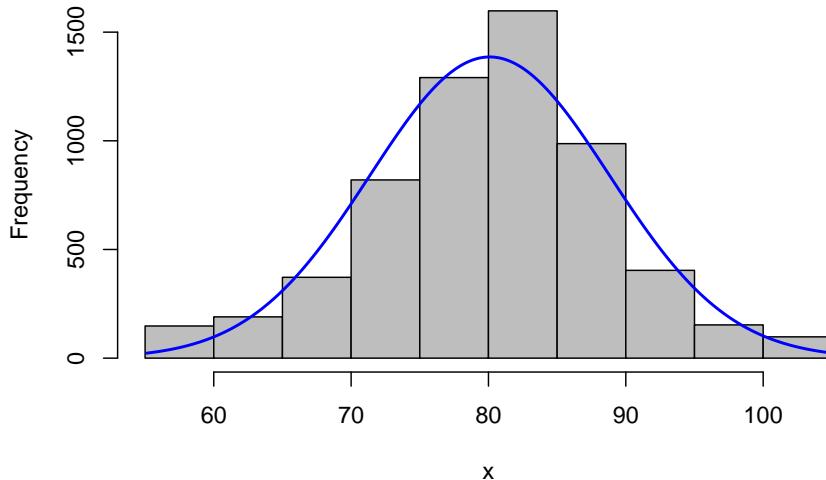
```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    55.12    74.96   80.62    80.09   85.44  104.95
```

This returns six numbers. The 0th percentile (alternatively referred to as the minimum) is 55.12 – this is the lowest yield measured in the field. The 25th percentile (also called the 1st Quartile) is 74.96. This means that 25% of all observations were equal to 74.96 bushels or less. The 50th percentile (also known as the median) was 80.62, meaning half of all observations were equal to 80.62 bushels or less. 75% of observations were less than 85.44, the 75th percentile (or 3rd quartile). Finally, the 100th percentile (or maximum yield) recorded for this field was 104.95.

We are now gaining a better sense of the range of observations that were most common. But we can describe this distribution with even fewer numbers.

1.3.3 Normal Distribution Model

Let's overlay a curve, representing the **normal distribution**, on our histogram. You have probably seen or heard of this curve before. Often it is called a *bell curve*; in school, it is the *Curve* that many students count on to bring up their grades. We will learn more about this distribution in *Lesson 2*.



In a perfect scenario, our curve would pass through the midpoint of each bar. This rarely happens with real-world data, and especially in agriculture. The data may be slightly **skewed**, meaning there are more individuals that measure above the mean than below, or vice versa.

In this example, our data do not appear skewed. Our curve seems a little too short and wide to exactly fit the data. This is a condition called **kurtosis**. No, kurtosis doesn't mean that our data stink; they are just more spread out or compressed than in a "perfect" situation.

No problem. We can – and should – conclude it is appropriate to fit these data with a normal distribution. If we had even more data, the curve would likely fit them even better.

Many populations can be handily summarized with the normal distribution curve, but we need to know a couple of statistics about the data. First, we need to know where the center of the curve should be. Second, we need to know the width or dispersion of the curve.

1.3.4 Measures of Center

To mathematically describe our distribution, we first need a **measure of center**. The two most common measures of center are the arithmetic mean and median. The **mean** is the sum of all observations divided by the number of observations.

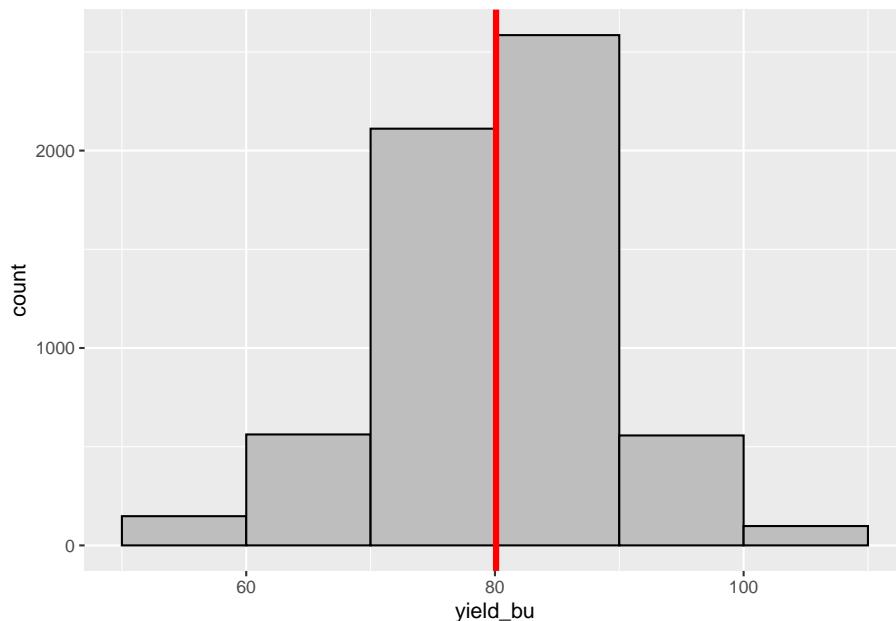
$$\mu = \frac{\sum x_i}{n}$$

The μ symbol (a u with a tail) signifies the true mean of a population. The \sum symbol (the character next to x_i which looks like the angry insect alien from *A Quiet Place*) means “sum”. Thus, anytime you see the \sum symbol, we are summing the variable(s) to its right. x_i is the value x of the i th individual in the population. Finally, n is the number of individuals in the population.

For example, if we have a set of numbers from 1:5, their mean can be calculated as:

$$\frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

The mean yield for our field is about 80.09 bushels per acre. This is represented by the red line in the histogram below.



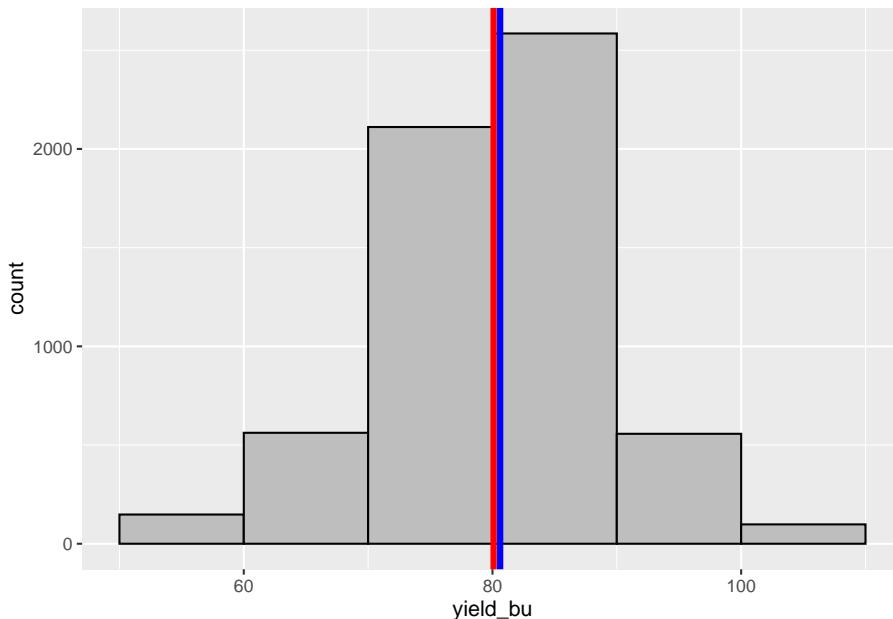
Earlier, you were introduced to the median. As discussed, the **median** is a number such that half of the individuals in the population are greater and half are less. If we have an odd number of individuals, the median is the “middle” number as the individuals are ranked from greatest to least.

$$\{1, 2, 3, 4, 5\} \text{median} = 3$$

If we have an even number of measures, the median is the average of the middle two individuals in the ranking:

$$\{1, 2, 3, 4, 5, 6\} \text{median} = 3.5$$

Let's add a blue line to our histogram to represent the median.



As you can see, they are practically identical. When the mean and median are similar, the number of individuals measuring greater and less than the mean are roughly equivalent. In this case, our data can be represented using the normal distribution.

We also need a statistic that tells us how wide to draw the curve. That statistic is called a measure of dispersion, and we will learn about it next.

1.3.5 Measures of Dispersion

To describe the spread of a population, we use one of three related **measures of dispersion**: sum of squares, variance, and standard deviation. Although there is a little math involved in these three statistics, please make yourself comfortable with their concepts because they are *very* important in this course. Almost every statistical test we will learn during this course is rooted in these measures of population width.

1.3.5.1 Sum of Squares

The first measure of population width is the **sum of squares**. This is the sum of the squared differences between each observation and the mean. The sum of squares of a measurement x is:

$$S_{xx} = (x_i - \mu)^2$$

Where again x_i is the value x of the i th individual in the population and μ is the true mean of a population.

Why do we square the differences between the observations and means? Simply, if we were to add the unsquared differences they would add to exactly zero. Let's prove this to ourselves. Let's again use the set of numbers (1, 2, 3, 4, 5). We can measure the distance of each individual from the mean by subtracting the mean from it. This difference is called the residual.

```
sample_data = data.frame(individuals = c(1,2,3,4,5))

library(janitor)

## 
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
## 
##     chisq.test, fisher.test

first_resid_table = sample_data %>%
  mutate(mean = 3) %>%
  mutate(residual = individuals - mean) %>%
  mutate(mean = as.character(mean))

first_resid_totals = data.frame(individuals = "Total",
                                 mean = "",
                                 residual = sum(first_resid_table$residual))

first_resid_table %>%
  rbind(first_resid_totals) %>%
  kbl()
```

individuals	mean	residual
1	3	-2
2	3	-1
3	3	0
4	3	1
5	3	2
Total		0

The first column of the above dataset contains the individual observations. The second column contains the population mean, repeated for each observation. The third column is the residuals, which are calculated by subtracting each observed value from the population mean.

And if we sum these residuals we get zero.

$$(-2) + (-1) + (0) + (+1) + (+2) = 0$$

Let's now do this with our field data. The number of residuals (almost 6800) is too many to visualize at once, so we will pick 20 at random.

```
set.seed(080921)
yield_sample = data.frame(yield = sample(yield$yield_bu, 10))

second_resid_table = yield_sample %>%
  mutate(yield = round(yield,2),
        mean = round(mean(yield),2),
        residual = yield-mean)

second_resid_totals = data.frame(yield = "Total",
                                  mean = "",
                                  residual = sum(second_resid_table$residual))

second_resid_table %>%
  rbind(second_resid_totals) %>%
  kbl()
```

yield	mean	residual
83.61	76.52	7.09
86.82	76.52	10.30
68.39	76.52	-8.13
81.91	76.52	5.39
80.75	76.52	4.23
57.06	76.52	-19.46
62.58	76.52	-13.94
86.6	76.52	10.08
80.05	76.52	3.53
77.42	76.52	0.90
Total		-0.01

If we sum up all the yield residuals, we get -0.04. Not exactly zero, but close. The difference from zero is the result of rounding errors during the calculation.

The sum of squares is calculated by squaring each residual and then summing the residuals. For our example using the set (1, 2, 3, 4, 5):

```
first_squares_table = first_resid_table %>%
  mutate(square = residual^2)

first_squares_totals = data.frame(individuals = "Total",
                                   mean = "",
                                   residual = "",
                                   square = sum(first_squares_table$square))

first_squares_table %>%
  rbind(first_squares_totals) %>%
  kbl()
```

individuals	mean	residual	square
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
Total			10

And for our yield data:

```
second_squares_table = second_resid_table %>%
  mutate(square = round(residual^2, 2)) %>%
  mutate(residual = round(residual, 2))

second_squares_totals = data.frame(yield = "Total",
```

```

mean = "",
residual = "",
square = sum(second_squares_table$square))
second_squares_table %>%
  rbind(second_squares_totals) %>%
  kbl()

```

yield	mean	residual	square
83.61	76.52	7.09	50.27
86.82	76.52	10.3	106.09
68.39	76.52	-8.13	66.10
81.91	76.52	5.39	29.05
80.75	76.52	4.23	17.89
57.06	76.52	-19.46	378.69
62.58	76.52	-13.94	194.32
86.6	76.52	10.08	101.61
80.05	76.52	3.53	12.46
77.42	76.52	0.9	0.81
Total			957.29

1.3.5.2 Variance

The sum of squares helps quantify spread: the larger the sum of squares, the greater the spread of observations around the population mean. There is one issue with the sum of squares, though: since the sum of square is derived from the differences between each observation and the mean, it is also related to the number of individuals overall in our population. In our example above, the sum of squares was 10.

Now, let's generate a dataset with two 1s, two 2s, two 3s, two 4s, and two 5s:

```

first_squares_table = first_resid_table %>%
  mutate(square = residual^2)

double_squares_table = first_squares_table %>%
  rbind(first_squares_table)

double_squares_totals = data.frame(individuals = "Total",
                                    mean = "",
                                    residual = "",
                                    square = sum(double_squares_table$square))

double_squares_table %>%

```

```
rbind(double_squares_totals) %>%
  kbl()
```

individuals	mean	residual	square
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
Total			20

You will notice the sum of squares increases to 20. The spread of the data did not change: we recorded the same five values. The only difference is that we observed each value twice.

The moral of this story is this: given any distribution, the sum of squares will always increase with the number of observations. Thus, if we want to compare the spread of two different populations with different numbers of individuals, we need to adjust our interpretation to allow for the number of observations.

We do this by dividing the sum of squares, S_{xx} by the number of observations, n . In essence, we are calculating an “average” of the sum of squares. This value is the variance, σ^2 .

$$\sigma^2 = \frac{S_{xx}}{n}$$

We can calculate the variance as follows.

In our first example, the set $\{1,2,3,4,5\}$ had a sum of squares of 10. in that case, the variance would be:

$$\frac{10}{5} = 2$$

In our second example, the set $\{1,2,3,4,5,1,2,3,4,5\}$ had a sum of squares of 20. In that example, the variance would be

$$\frac{20}{10} = 2$$

As you can see, the variance is not affected by the size of the dataset, only by the distribution of its values.

Later on in this course, we will calculate the variance a little differently:

$$\sigma^2 = \frac{S_{xx}}{n-1}$$

$n-1$ is referred to as the **degrees of freedom**. We use degrees of freedom when we work with samples (subsets) of a population. In this unit, however, we are working with populations, so we will not worry about those.

1.3.5.3 Standard Deviation

Our challenge in using the variance to describe population spread is it's units are not intuitive. When we square the measure we also square the units of measure. For example the variance of our yield is measured in units of bushels². Wrap your head around that one. Our solution is to report our final estimate of population spread in the original units of measure. To do this, we calculate the square root of the variance. This statistic is called the standard deviation, σ .

$$\sigma = \sqrt{(\sigma^2)}$$

For the dataset {1,2,3,4,5}, the sum of squares is 10, the variance 2, and the standard deviation is:

$$\sigma = \sqrt{2} = 1.4$$

For our yield dataset, the sum of squares is 957.29, and based on 10 observations. Our variance is therefore:

$$\sigma^2 = \frac{957.29}{10} = 9.57 \text{ bushels}^2 / \text{acre}^2$$

Our sum of squares is :

$$\sqrt{9.57 \text{ bushels}^2 / \text{acre}^2} = 3.09 \text{ bushels} / \text{acre}$$

That is enough theory for this first week of class. The remainder of this lesson will focus on introducing you to **RStudioCloud**.

Chapter 2

Distributions and Probability

In this unit we will continue with the **normal distribution model** introduced in the previous unit. As you will recall, the normal distribution is a symmetrical curve that represents the frequency with which individuals with different particular measured values occur in a population.

The peak of the curve is located at the population mean, μ . The width of the curve reflects how spread out other individuals are from the mean. We learned three ways to measure this spread: the sum of squares, the variance, and the standard deviation. These statistics can roughly be thought of as representing the sums of the squared differences, the average of the squared distances, and the distances presented in the original units of measure. These four statistics – mean, sum of squares, variance, and standard deviation – are among the most important you will learn in this course.

2.1 Case Study

This week, we will continue to work with the Iowa soybean yield dataset introduced to us in Unit 1.

Here again is the structure of this dataset:

```
head(yield)
```

```
## Simple feature collection with 6 features and 12 fields
## Geometry type: POINT
## Dimension: XY
```

```

## Bounding box: xmin: -93.15033 ymin: 41.66641 xmax: -93.15026 ymax: 41.66644
## Geodetic CRS: WGS 84
##   DISTANCE SWATHWIDTH VRYIELDVOL Crop  WetMass Moisture           Time
## 1 0.9202733          5 57.38461 174 3443.652    0.00 9/19/2016 4:45:46 PM
## 2 2.6919269          5 55.88097 174 3353.411    0.00 9/19/2016 4:45:48 PM
## 3 2.6263101          5 80.83788 174 4851.075    0.00 9/19/2016 4:45:49 PM
## 4 2.7575437          5 71.76773 174 4306.777    6.22 9/19/2016 4:45:51 PM
## 5 2.3966513          5 91.03274 174 5462.851   12.22 9/19/2016 4:45:54 PM
## 6 3.1840529          5 65.59037 174 3951.056   13.33 9/19/2016 4:45:55 PM
##   Heading VARIETY Elevation           IsoTime yield_bu
## 1 300.1584 23A42 786.8470 2016-09-19T16:45:46.001Z 65.97034
## 2 303.6084 23A42 786.6140 2016-09-19T16:45:48.004Z 64.24158
## 3 304.3084 23A42 786.1416 2016-09-19T16:45:49.007Z 92.93246
## 4 306.2084 23A42 785.7381 2016-09-19T16:45:51.002Z 77.37348
## 5 309.2284 23A42 785.5937 2016-09-19T16:45:54.002Z 91.86380
## 6 309.7584 23A42 785.7512 2016-09-19T16:45:55.005Z 65.60115
##   geometry
## 1 POINT (-93.15026 41.66641)
## 2 POINT (-93.15028 41.66641)
## 3 POINT (-93.15028 41.66642)
## 4 POINT (-93.1503 41.66642)
## 5 POINT (-93.15032 41.66644)
## 6 POINT (-93.15033 41.66644)

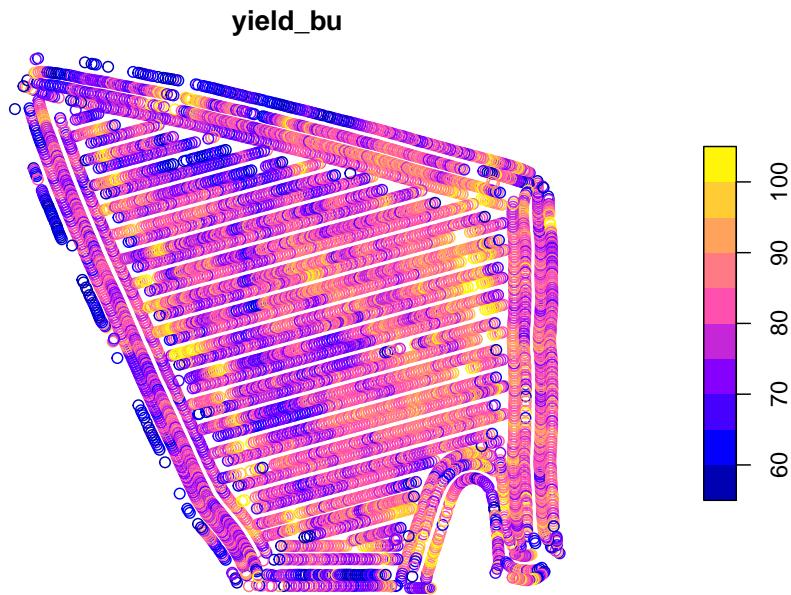
```

And here is the map of the field:

```

library(sf)
plot(yield["yield_bu"])

```



2.2 The Normal Distribution Model

Mean, sum of squares, variance, and standard deviation are so important because they allow us to reconstruct the normal distribution model. Before we go further, what is a **model**? Is it frightening we are using that term already in the *second chapter of this text???*

Models can be very complex, but in there essence they all have the following in common: they are simplified representations of reality. No, that doesn't mean that models are "fake" (SMH). It means that they summarize aspects of data, both measured and predicted. The normal distribution model describes the relationship between the values of individuals and how frequently they appear in the population. The model is useful because we can use it to approximately reconstruct the dataset at any time by knowing just two things about the original dataset – its mean and its standard deviation.

2.2.1 The Bell Curve

The normal distribution curve is often referred to as the *bell curve*, since it is taller in the middle and flared on either side. This shape reflects the tendency of measures within many populations to occur more frequently near the population mean than far from it. Why does this occur and how do we know this?

As agronomists, we can reflect on what it takes to produce a very good – or very bad crop. For a very good crop, many factors need to coincide: temperature, precipitation, soil texture, rate of nitrogen mineralization, proper seed singulation (spacing during planting), pest control, and hybrid or variety selection, to name just a few. In a typical season or within a field, we might optimize a few of these factors, but the possibility of optimizing every one is exceedingly rare. Thus, if we are measuring yield, measures near the mean yield will occur more frequently. Extremely high yields will occur less frequently.

Conversely, very low yields require we manage a crop very badly or that catastrophic weather conditions occur: a hailstorm, flood, or tornado. A frost at exactly the wrong time during seed germination or, in corn, excessive heat or a drought during pollination or grain fill. A planter box running out of seed or a fertilizer nozzle jamming. These things do occur, but less frequently.

The distribution of individuals around the mean is also a the result of measurement inaccuracies. Carl Friedrich Gauss, who introduced the normal distribution model, showed that it explained the variation among his repeated measurements of the position of stars in the sky. All measurements of continuous data (those that can be measured with a ruler, a scale, a graduated cylinder, or machine) have variation – we use the term **accuracy** to explain their variation around the population mean.

2.2.2 Distribution and Probability

Some areas of mathematics like geometry and algebra identify theorems: consistent, proven relationships between variables. In statistics, however, we typically solve for the **probability** that one or more variables are have a given value or range of values. To be more specific, most of the statistical tests we will learn can be reduced to the probability that a particular value is observed in a population. These probabilities include:

- that a given value or mean is observed for a population; we calculate this using the *normal distribution*.
- that the difference between two treatments is not zero; we calculate this using a *t-Test* or *Least Significant Difference* test.
- that a value will be observed in one variable, given a specific value of another variable; we calculate this using *Linear Regression*.
- that the spread of individual measures in a population is better predicted by treatment differences than random variation; we calculate this using an *F-test* and *Analysis of Variance*)

Each of these probabilities is calculated from a distribution – the frequency with which individuals appear in a population. Another way of stating this is that

probability is the proportion of individuals in a population that are expected to have values within a given range. Examples could include:

- the proportion of individual soybean yield measurements, within one field, that are less than 65 bushels per acre
- the proportion of individual corn fields that have an average less than 160 bushels per acre
- the proportion of trials in which the difference between two treatments was greater than zero
- the proportion of observations in which the actual crop yield is greater than that predicted from a regression model

2.2.3 Probability and the Normal Distribution Curve

Probability can be calculated as the proportion of the area underneath the normal distribution that corresponds to a particular range of values. We can visualize this, but first we need to construct the normal distribution curve for our soybean field.

We need just two statistics to construct our distribution curve: the mean and standard deviation of our yield. In the last lesson, we learned how easily these both can be calculated in R:

```
library(muStat)
yield_mean = mean(yield$yield_bu)
yield_sd = stdev(yield$yield_bu, unbiased = FALSE)

# to see the value of yield_mean and yield_sd, we just run their names in our code
yield_mean

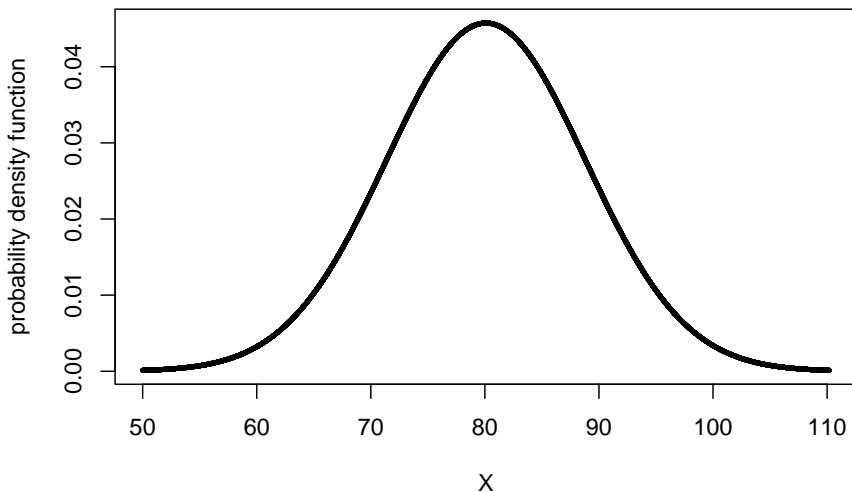
## [1] 80.09084

yield_sd
```

```
## [1] 8.72252
```

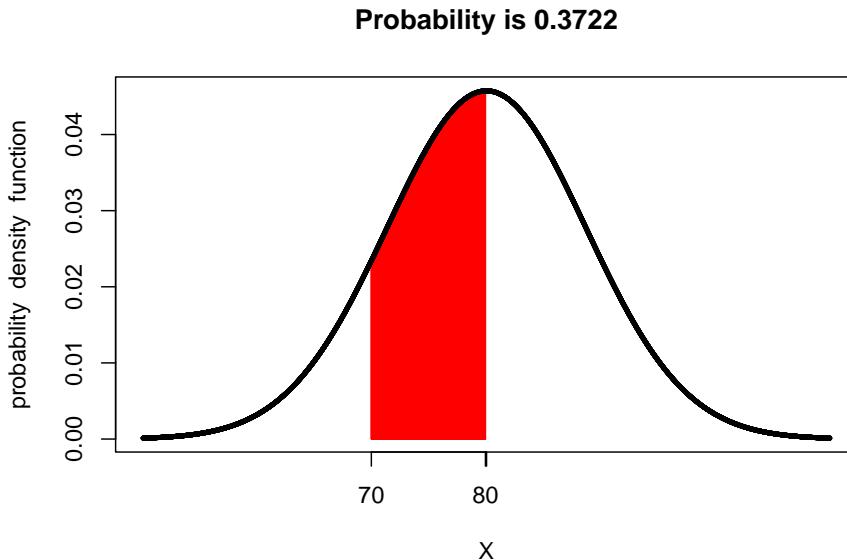
In this lesson's exercise section, we will learn to use R to construct the distribution curve for any population, given the population mean and population standard deviation.

```
library(fastGraph)
plotDist("dnorm", yield_mean, yield_sd)
```



Let's now shade the area underneath the normal curve corresponding to X values from 70 - 80. This area will represent the proportion of the population where individuals were measured to have values between 70 and 80 bushels. We will use a function in R called `shadeDist` to do this. You will learn more about this function this lesson in the exercise section.

```
shadeDist(xshade=c(70,80), ddist = "dnorm", yield_mean, yield_sd, lower.tail = FALSE)
```



Pretty cool, huh? The red area is the proportion of the soybean yield population that was between 70 and 80 bushels/acre. At the top of the output, `shadeDist` has also reported the proportion of the curve represented by that area, which it has labelled *Probability*. The probability in this case is 0.3722. What does that number mean?

The total area of the curve is 100%, or 1.0000. The proportion of the area under the curve that corresponds with yields from 70 to 80 bushels, then, is 37.22 percent of the area. This means that 37.22 percent of the individuals in our yield population had values from 70 and 80 bushels

But wait a second – why is R using the term *Probability*? Think of it this way. Imagine you sampled 1000 individuals from our population. If 37.22 percent of our individuals have values from 70 to 80 bushels, then about 37% of the individuals in your sample should have values from 70 to 80 bushels. In other words, there is a 37% probability that any individual you select, at random, from the population will have a value from 70 to 80 bushels.

Let's test this. Let's randomly sample 1000 individuals from our population. Then lets count the number of individuals that have yields between 70 and 80 bushels. For the curious, this is how we do this in R. We will run three lines of R code to do this:

- First, use the `set.seed()` function to specify a certain point in the population where R will begin sampling. R uses an algorithm to choose samples. This is not quite the same thing as random sampling – by setting a seed, we ensure that we can generate the same “random” set in the future should

we need to. Our seed number can be any numeric value; in this case, the date this section was revised.

- Second, use the `sample` function to randomly sample our population and return a vector of those numbers.
- Third, use the `subset` function to subset our data into those that meet logical conditions and return a dataframe or vector.
- Fourth, use the `length` function to count the number of observations.

For everyone else, just understand this is how we came up with the sample of 1000 individuals.

```
# 1) set seed
set.seed(081521)

# 2) take sample
yield_sample = sample(yield$yield_bu, 1000)

# "yield_sample >=70 & yield_sample <=80 tells it to only include measures from 70 to 80

# 3) Subset data into values between 70 and 80
yield_subset = subset(yield_sample, yield_sample >=70 & yield_sample <=80)

# 4) Count number of samples in subset
length(yield_subset)

## [1] 353
```

Our sample had 1000 individuals. 353, or 35.3%, were had yields between 70 and 80 bushels/acre.

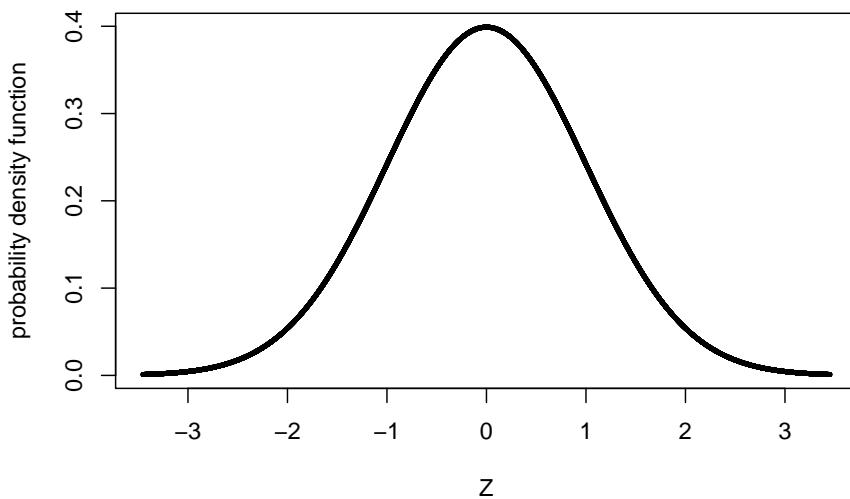
Is the proportion predicted by the normal distribution curve exactly that of the actual population? No. The normal distribution curve is, after all, a model – it is an approximation of the actual population. In addition, our sample is a subset of the population, not a complete accounting.

We will talk more about sampling in the next unit.

2.3 The Z-Distribution

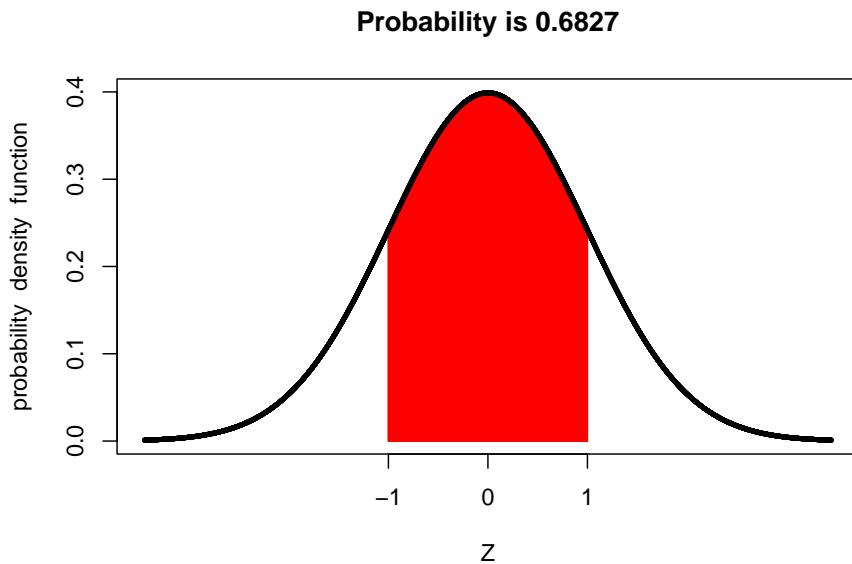
The relationship between probability and the normal distribution curve is based on the concept of the Z-distribution. In essence, the **Z-distribution** describes a

normal distribution curve with a population mean of 0 and a standard deviation of 1.

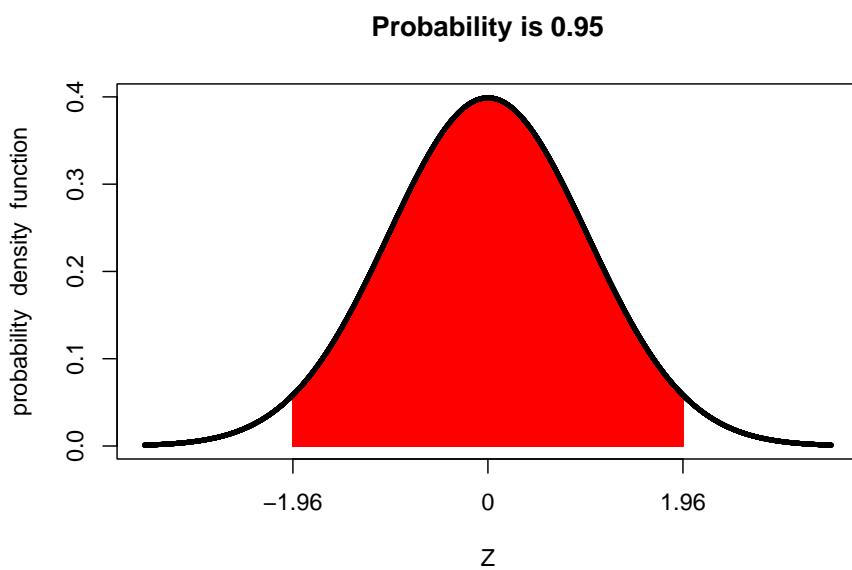


The Z-distribution helps us understand how probability relates to standard deviation in a normal distribution, regardless of the nature of a study or its measurement units.

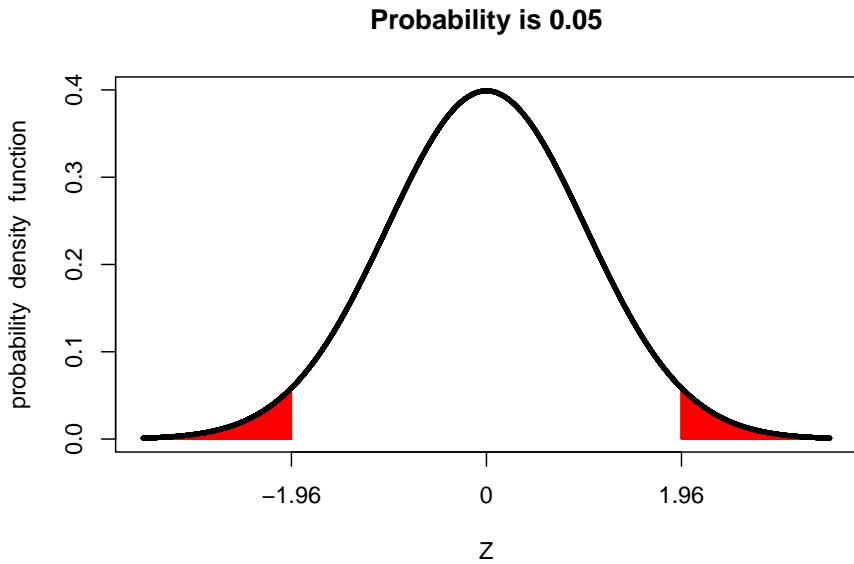
For example, the proportion of a population within one standard deviation of the mean is about 68 percent:



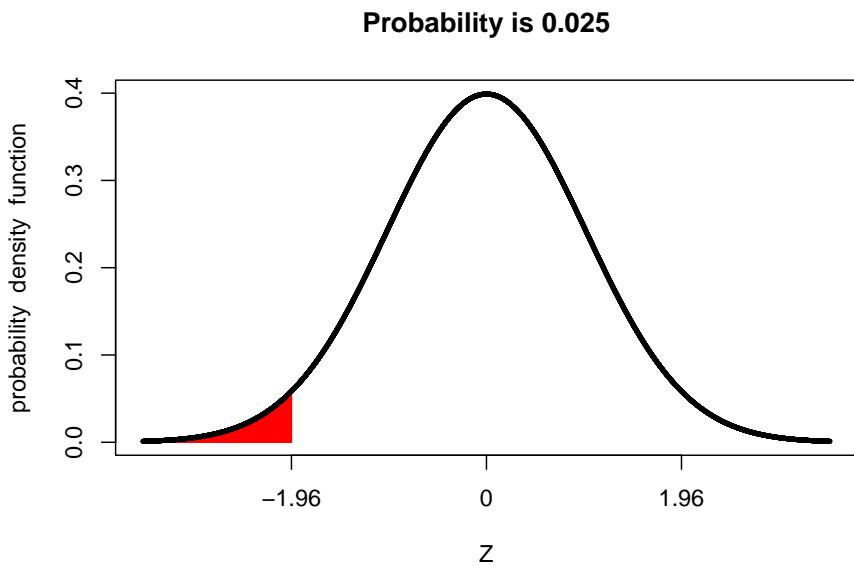
Similarly, the proportion of a population within 1.96 standard deviations of the mean is about 95 percent:



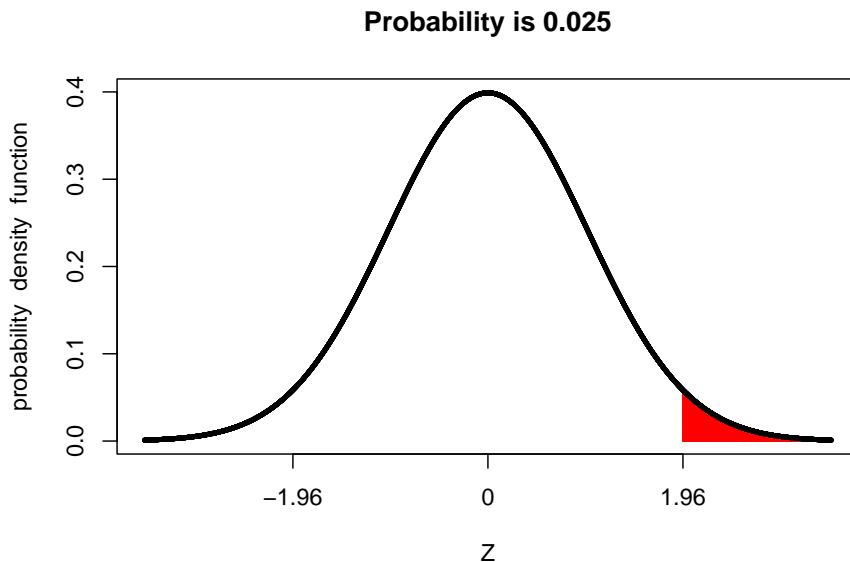
Conversely, the proportion of a population beyond 1.96 standard deviations from the mean is about 5 percent.



We refer to the upper and lower ends of the distribution as **tails**. In a normal distribution we would expect about 2.5% of observations to less than -1.96 standard deviations of the mean.



And 2.5% of the population to be more than +1.96 above the mean:



2.3.1 Important Numbers: 95% and 5%

Above we learned that 95% of a normal distribution is between 1.96 standard deviations of the mean, and that 5% of a normal distribution is outside this range. Perhaps these numbers sound familiar to you. Have you ever seen results presented with a 95% confidence interval? Have you ever read that two treatments were significantly different at the $P=0.05$ level?

For population statistics, the normal distribution is the origin of those numbers. As we get further into this course, we will learn about additional distributions – t and F – and the unique statistical tests they allow. But the concept will stay the same: identifying whether observed statistical values are more likely to occur (i.e., within the central 95% of values expected in a distribution), or whether the values are unusual (occurring in the remaining 5%).

Chapter 3

Sample Statistics

In the previous two units, we studied populations and how to summarize them with statistics when the *entire* population was measured. In other words, the measure of center (mean) and measure of spread (standard deviation) were the summary of all observations.

In the case of a yield monitor map, these are appropriate statistics. In most every other agricultural reality, however, we cannot measure every individual in a population. Instead, we only have enough resources to collect a **sample** the population, that is, measure a subset of individuals from the population. In this case, we cannot measure the exact population mean or population standard deviation of the population. Instead, we can only estimate them using our **sample mean** or **sample standard deviation**.

This, of course, raises questions. Was the sample representative of the population? Would another random sample result in a similar estimate of the population mean or population standard deviation? And, perhaps, how much could our sample mean deviate from the population mean?

In other words, there is always uncertainty that statistics calculated from samples represent the true values of a population. You might even say we lack complete *confidence* that a sample mean will closely estimate the population mean.

Enter statistics. By taking multiple sets of samples, and calculating their means, we can use the differences among those sample means to estimate the distribution of sample means around the true population mean. Indeed, this is a fundamental concept of research and statistics – using the measured variance of samples to determine how accurate they are in predicting population statistics.

3.1 Samples

To measure the variation of sample means, we need at least two samples to compare. Ideally we can gather even more. As we will see, the more samples included in our estimates of the population mean, the more accurate we are likely to be.

A second comment, which may seem intuitive, but at the retail level may be overlooked, is randomization. Samples – for example individual plants or areas where yield will be measured – are ideally selected at random. In reality, the plants or areas selected for measures may be less than random. When I used to count weed populations, we used square quadrats (frames) to consistently define the area that was measured. We would throw them into different areas of the plot and count weeds where ever they landed.

The most important thing about selecting samples, however, is that the researcher work to minimize bias. Bias is when the samples selected consistently overestimate or underestimate the population mean. The most egregious example of this would be a researcher who consistently and purposely sampled the highest- or lowest-measuring parts of a field.

But bias can enter in other ways. For example, if our weed populations were very uneven, our thrown quadrat might be more likely to skid to a stop in weedy areas. A researcher might unconsciously choose taller plants to sample. In August, we might be tempted to sample a corn field from the edge than walk deep into that sweltering, allergenic hell.

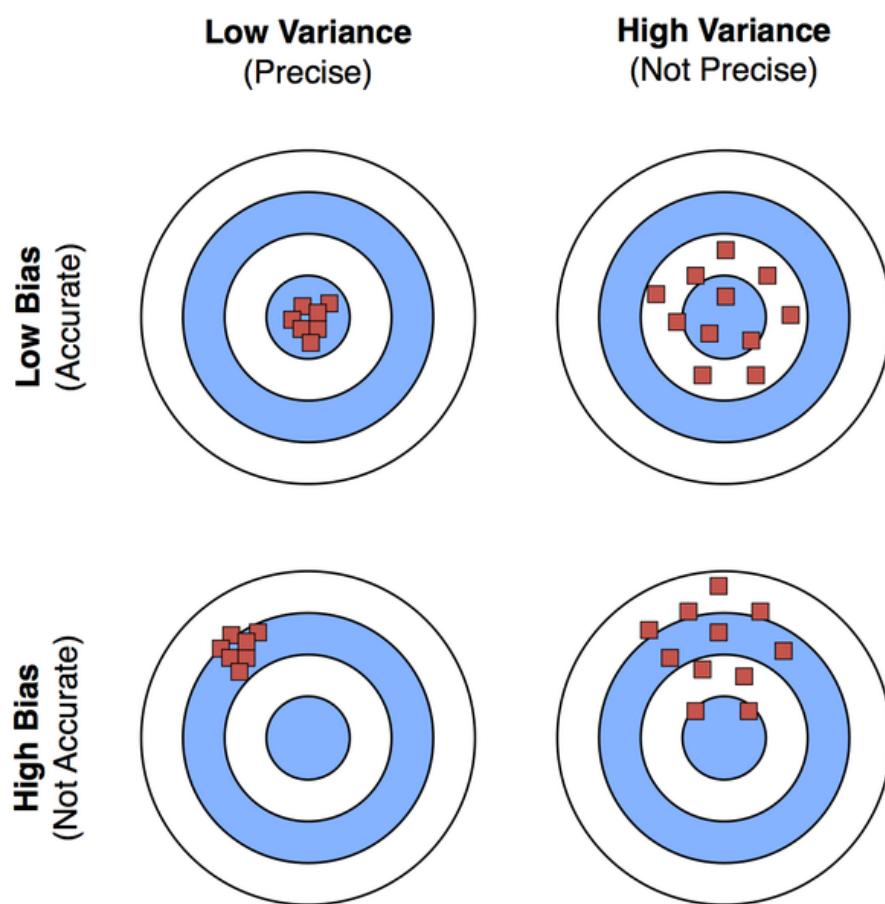
Remember, our goal is to generate estimates of population values that are as precise and accurate as our resources allow. **Precise** means our sample means have a low variance around the population mean. **Accurate** means our sample means are equivalently scattered above and below the population mean.

3.2 Case Study

Once more, we will work with the Iowa soybean yield dataset from Units 1 and 2.

Let's review the structure of this dataset:

```
## Simple feature collection with 6 features and 12 fields
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: -93.15033 ymin: 41.66641 xmax: -93.15026 ymax: 41.66644
## Geodetic CRS: WGS 84
##   DISTANCE SWATHWIDTH VRYIELDVOL Crop  WetMass Moisture           Time
## 1  0.9202733          5  57.38461   174 3443.652      0.00 9/19/2016 4:45:46 PM
```



This work by Sebastian Raschka is licensed under a
Creative Commons Attribution 4.0 International License.

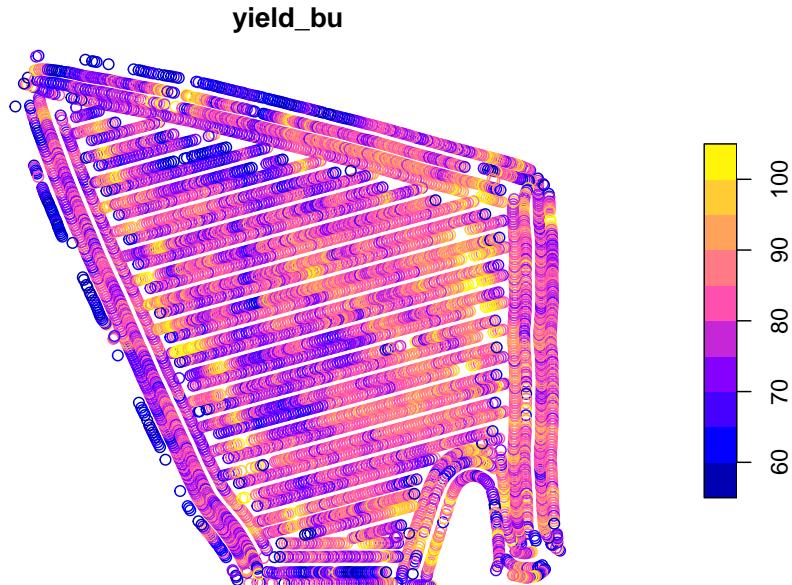
Figure 3.1: Accuracy versus Bias

```

## 2 2.6919269      5  55.88097 174 3353.411    0.00 9/19/2016 4:45:48 PM
## 3 2.6263101      5  80.83788 174 4851.075    0.00 9/19/2016 4:45:49 PM
## 4 2.7575437      5  71.76773 174 4306.777    6.22 9/19/2016 4:45:51 PM
## 5 2.3966513      5  91.03274 174 5462.851   12.22 9/19/2016 4:45:54 PM
## 6 3.1840529      5  65.59037 174 3951.056   13.33 9/19/2016 4:45:55 PM
##   Heading VARIETY Elevation           IsoTime yield_bu
## 1 300.1584 23A42 786.8470 2016-09-19T16:45:46.001Z 65.97034
## 2 303.6084 23A42 786.6140 2016-09-19T16:45:48.004Z 64.24158
## 3 304.3084 23A42 786.1416 2016-09-19T16:45:49.007Z 92.93246
## 4 306.2084 23A42 785.7381 2016-09-19T16:45:51.002Z 77.37348
## 5 309.2284 23A42 785.5937 2016-09-19T16:45:54.002Z 91.86380
## 6 309.7584 23A42 785.7512 2016-09-19T16:45:55.005Z 65.60115
##   geometry
## 1 POINT (-93.15026 41.66641)
## 2 POINT (-93.15028 41.66641)
## 3 POINT (-93.15028 41.66642)
## 4 POINT (-93.1503 41.66642)
## 5 POINT (-93.15032 41.66644)
## 6 POINT (-93.15033 41.66644)

```

And map the field:



In Unit 2, we learned to describe these data using the normal distribution model. We learned the area under the normal distribution curve corresponds to the proportion of individuals within a certain range of values. We also discussed

how this proportion allows inferences about probability. For example, the area under the curve that corresponded with yields from 70.0 to 79.9 represented the proportion of individuals in the yield population that fell within that yield range. But it also represented the probability that, were you to measure individual points from the map at random, you would measure a yield between 70.0 and 79.9.

3.3 Distribution of Sample Means

In the last unit, we sampled the yield from 1000 locations in the field and counted the number of observations that were equal to or greater than 70 and equal to or less than 80.

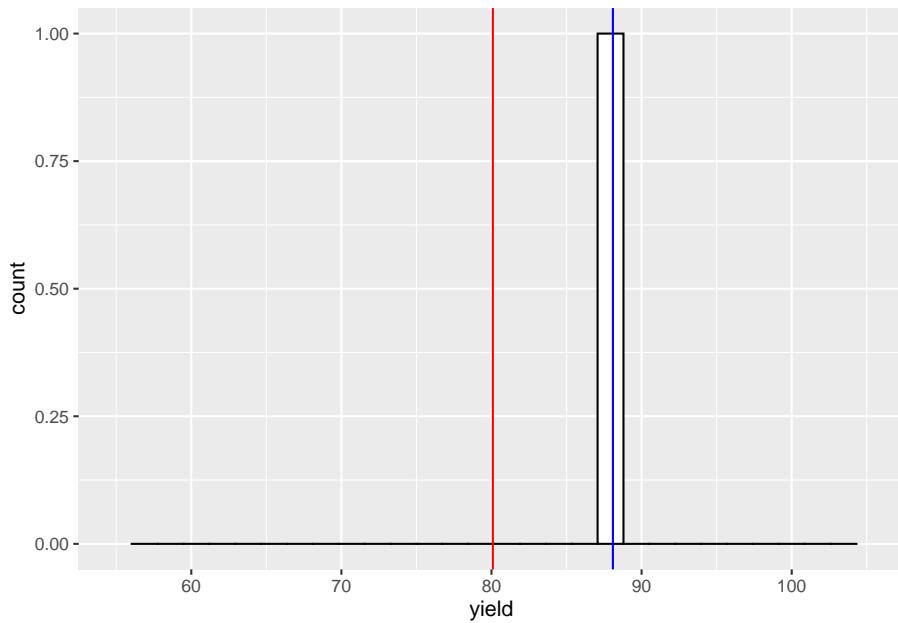
What would happen if we only sampled from one location. What would be our sample mean and how close would it be to the population mean?

In the histograms below, the red vertical line marks the population mean. The blue line marks the sample mean.

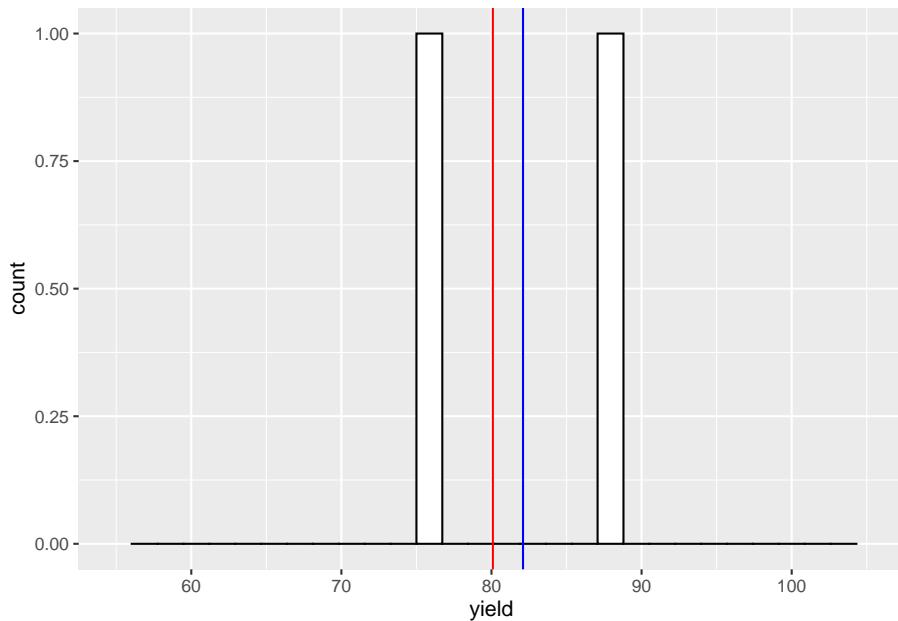
```
set.seed(1771)
yield_sample = sample(yield$yield_bu, 1) %>%
  as.data.frame()
names(yield_sample) = c("yield")
ggplot(yield_sample, aes(x=yield)) +
  geom_histogram(fill="white", color="black") +
  geom_vline(xintercept = mean(yield$yield_bu), color = "red") +
  geom_vline(xintercept = mean(yield_sample$yield), color = "blue") +
  lims(x=c(55,105))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

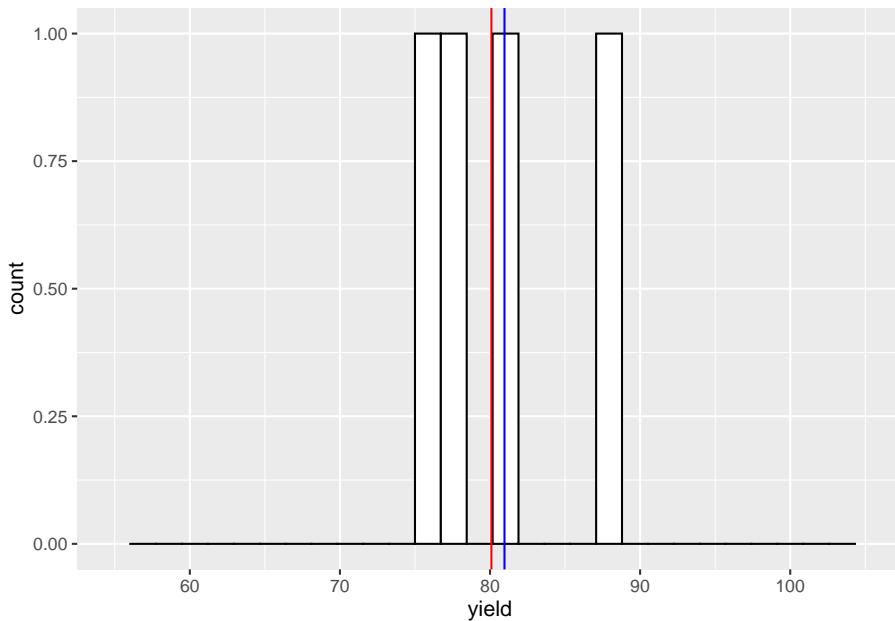
## Warning: Removed 2 rows containing missing values (geom_bar).
```



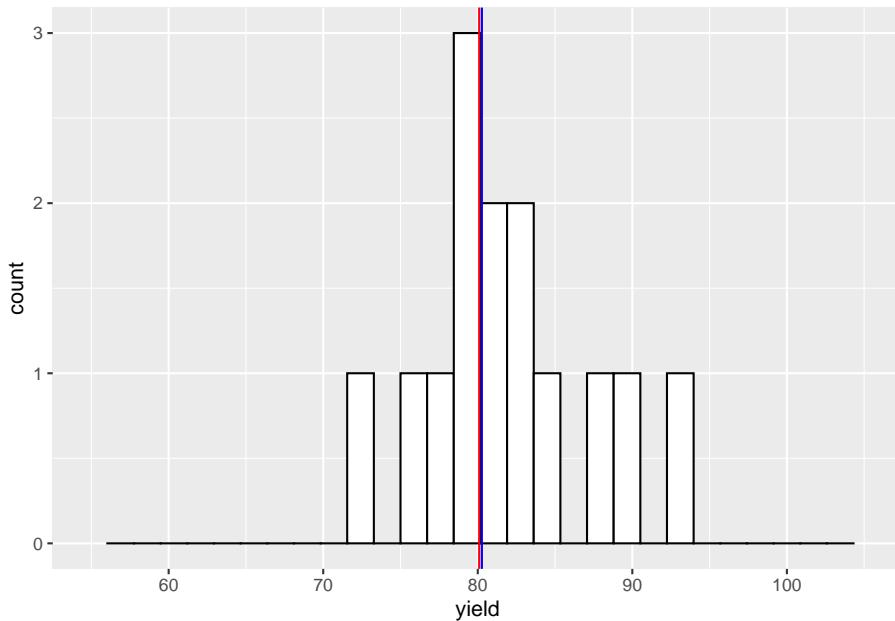
With one sample, our sample mean is about 8 bushels above the population mean. What would happen if we sampled twice?



Our sample mean is now about 2 bushels greater than the population mean. What would happen if we sampled four times?



Our sample mean is now about 1 bushel greater than the population mean. What would happen if we sampled 15 times?



The sample mean and population mean are almost equal.

Click on this link to access an app to help you further understand this concept:

https://marin-harbur.shinyapps.io/03-sampling_from_normal_distn/

3.4 Central Limit Theorem

The **Central Limit Theorem** states that sample means are normally distributed around the population mean. This concept is powerful because it allows us to calculate the probability that that a sample mean is a given distance away from the population mean. In our yield data, for example, the Central Limit Theorem allows us to assign a probability that we would observe a sample mean of 75 bushels/acre, if the population mean is 80 bushels/acre. More on how we calculate this in a little bit.

In our yield dataset, the population data are approximately normally distributed. It makes sense that the sample means would be normally distributed, too. But the Central Limit Theorem shows us that our sample means are likely to be normally distributed even if the population *does not* follow a perfect normal distribution.

Let's take this concept to the extreme. Suppose we had a population where every value occurred with the same frequency. This is known as a uniform distribution. Click on the following link to visit an app where we can explore how the sample distribution changes in response to sampling an uniform distribution:

<https://marin-harbur.shinyapps.io/03-sampling-from-uniform-distn/>

You will discover that the sample means are normally distributed around the population mean even when the population itself is not normally distributed.

3.5 Standard Error

When we describe the spread of a normally-distributed population – that is, all of the individuals about which we want to make inferences – we use the population mean and standard deviation.

When we sample (measure subsets) of a population, we again use two statistics. The **sample mean** describes the center of the samples.. The spread of the sample means is described by the **standard error of the mean** (often abbreviated to **standard error**). The standard error is related to the standard deviation as follows:

$$SE = \frac{\sigma}{\sqrt{n}}$$

The standard error, SE, is equal to the standard deviation (σ), divided by the square root of the number of samples (n). This denominator is very important

– it means that our standard error shrinks as the number of samples increases.
Why is this important?

A sample mean is an estimate of the true population mean. The distribution of sample means the range of possible values for the population mean. I realize this is a fuzzy concept. This is the key point: by measuring the distribution of our sample sample means, we are able to describe the probability that the population mean is a given value.

To better understand this, please visit this link:

<https://marin-harbur.shinyapps.io/03-sample-distn/>

If you take away nothing else from this lesson, understand whether you collect 2 or 3 samples has tremendous implications for your estimate of the population mean. 4 samples is much better than 3. Do everything you can to fight for those first few samples. Collect as many as you can afford, especially if you are below 10 samples.

3.6 Degrees of Freedom

In Unit 1, the section on variance briefly introduced **degrees of freedom**, the number of observations in a population or sample, minus 1. Degrees of Freedom are again used below in calculating the **t-distribution**. So what are they and why do we use them? Turns out there are two explanations.

In the first explanation, *degrees of freedom* refers to the number of individuals or samples that can vary independently given a fixed mean. So for an individual data point to be free, it must be able to assume any value within a given distribution. Since the population mean is a fixed number, only $n - 1$ of the data have the freedom to vary. The last data point is determined by the value of all the other data points and the population mean.

Confusing, huh? Who starts measuring samples thinking that the data point is fixed, in any case? But if you think about it, the purpose of the sample is approximate a real population mean out there – which is indeed fixed. It's just waiting for us to figure it out. So if our sample mean is equal to the population mean (which we generally assume), then the sample mean is also fixed. But it is a very weird way of thinking.

Yet this is the answer beloved by all the textbooks, so there you go.

The second answer I like better: samples normally *underestimate* the true population variance. This is because the sample variance is calculated from the distribution of the data around the sample mean. Sample data will always be closer to the sample mean – which is by definition based on the data themselves – than the population mean.

Think about this a minute. Your sample data could be crazy high or low compared to the overall population. But that dataset will define a mean, and the variance of the population will be estimated from that mean. In many cases, it turns out that using $n - 1$ degrees of freedom will increase the value of the sample variance so it is closer to the population variance.

3.7 The t-Distribution

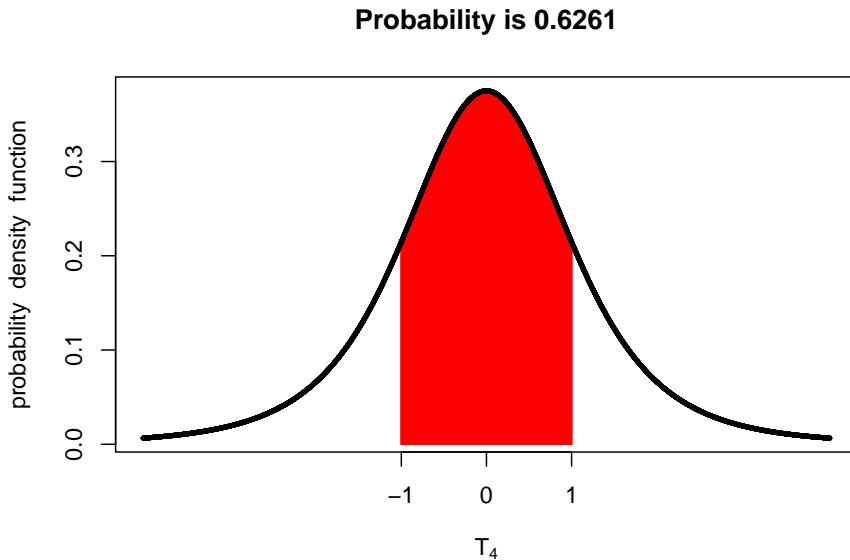
In the last unit, we used the *Z-distribution* to calculate the probability of observing an individual of a given value in a population, given its population mean and standard deviation. Recall that about 68% of individuals were expected to have values within one standard deviation, or Z , of the population mean. Approximately 95% of individuals were expected to have values within 1.96 standard deviations of the population mean. Alternatively, we can ask what the probability is of observing individuals of a particular or greater value in the population, given its mean and standard deviation.

We can ask a similar question of our sample data: what is the probability the population mean is a given value or greater, given the sample mean? As with the Z-distribution, the distance between the sample mean and hypothesized population mean will determine this probability.

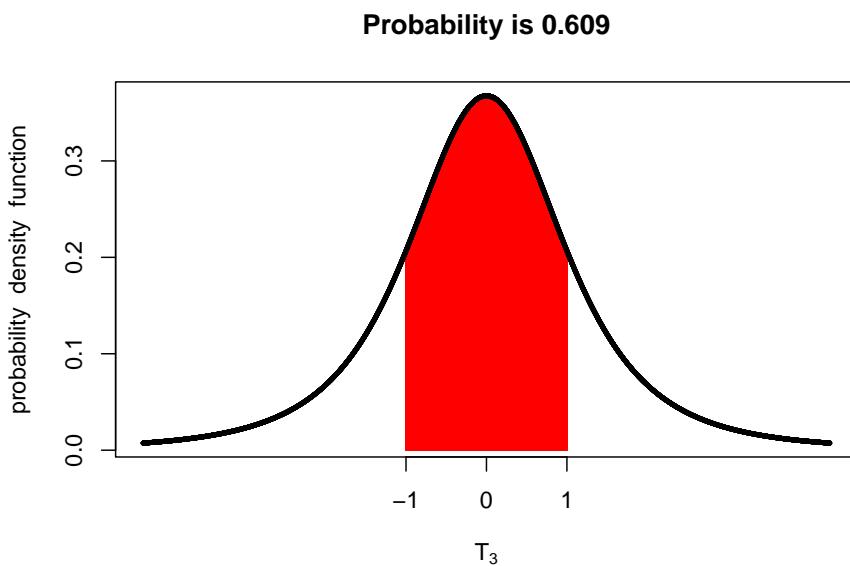
There is one problem, however, with using the Z-distribution: it is only applicable when the population standard deviation is *known*. When we *sample* from a population, we do not know its true standard deviation. Instead, we are estimating it from our samples. This requires we use a different distribution: the t-distribution.

Unlike the Z-distribution, the **t-distribution** changes in shape as the number of samples increases. Notice in the animation above that, when the number of samples is low, the distribution is wider and has a shorter peak. As the number of samples increases, the curve becomes narrower and taller. This has implications for the relationship between the distance of a hypothetical population mean from the sample mean, and the probability of it being that distant.

We can prove this to ourselves with the help of the `shadeDist()` function we used in Unit 2. You will learn how to plot the t-distribution in an exercise this week.

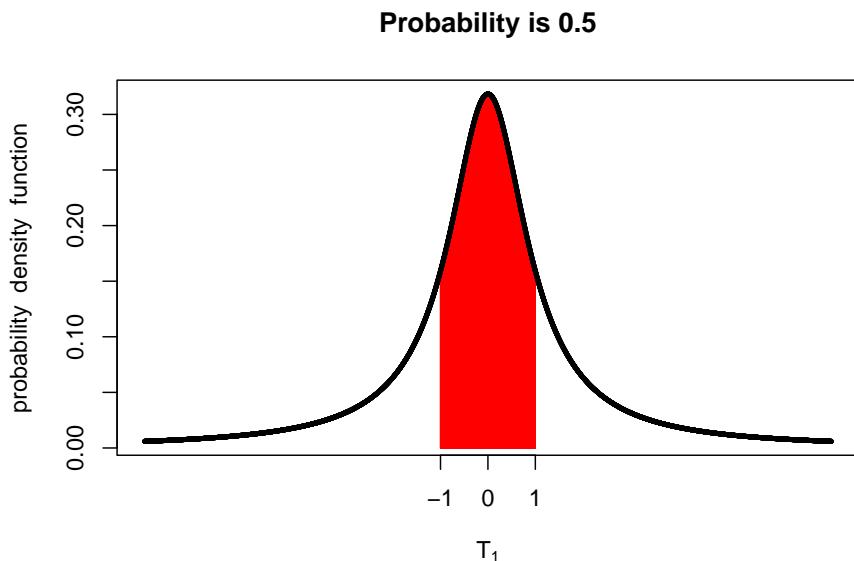


With 4 degrees of freedom, there is about a 63% probability the population mean is within 1 standard error of the mean. Let's decrease the sample mean to 3 degrees of freedom



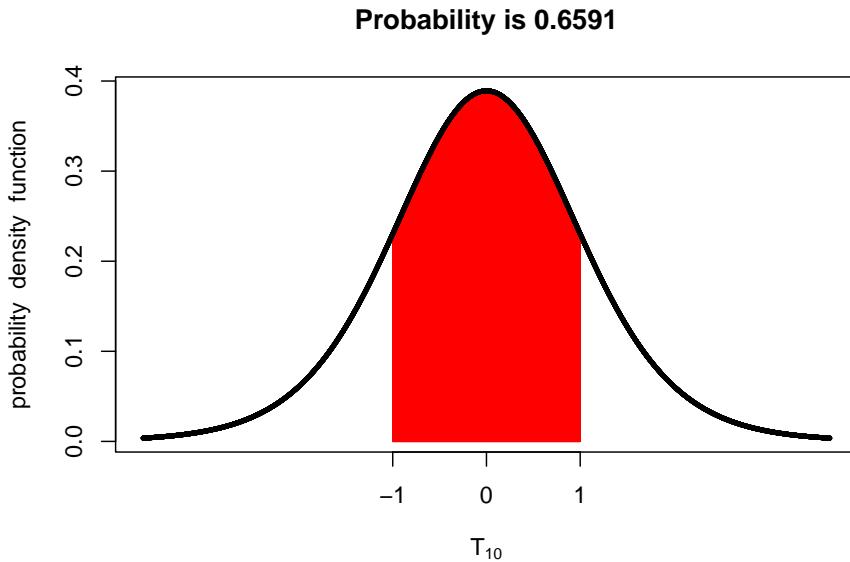
With only 3 degrees of freedom (4 samples), there is only a 61% probability the

population mean is within one standard error of the mean. What if we only had one degree of freedom (two samples)?

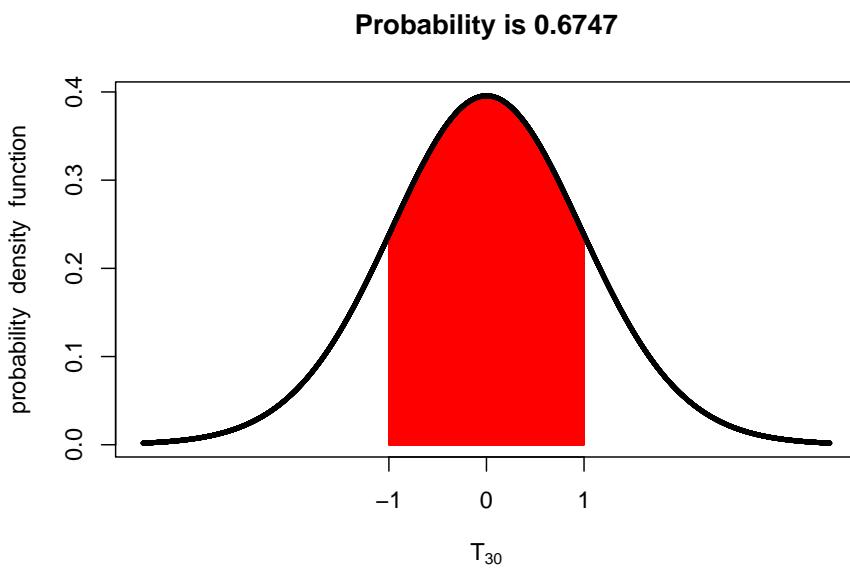


You should see the probability that the population mean is within 1 standard error of the sample mean falls to 50%.

If we have 10 degrees of freedom (11 samples), the probability increases to about 66%.

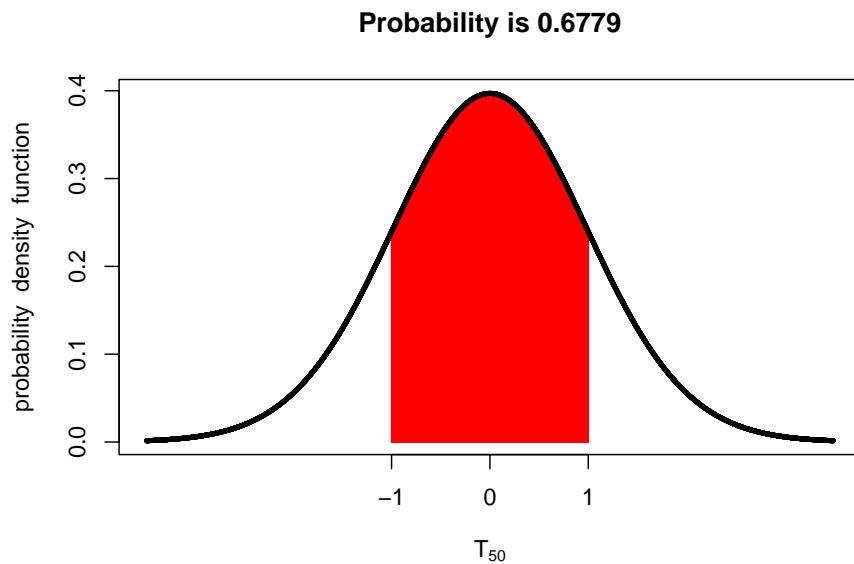


With 30 degrees of freedom the probability the population mean is within 1 standard error of the sample mean increases to 67%.

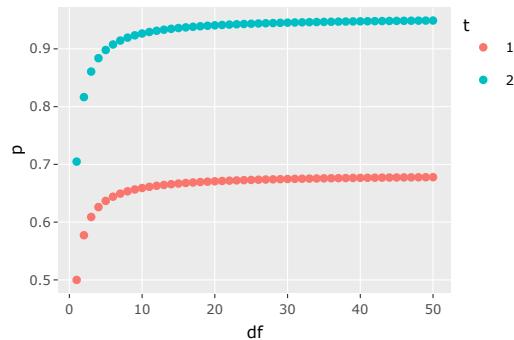


With 50 degrees of freedom (51 samples) the probability is about 68%. At this point, the t-distribution curve approximates the shape of the z-distribution

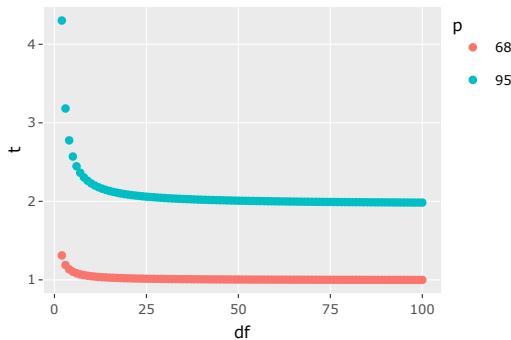
curve.



We can sum up the relationship between the t-value and probability with this plot. The probability of the population mean being within one standard error of the population mean is represented by the red line. The probability of the population mean being within 2 standard errors of the mean is represented by the blue line. As you can see, the probability of the population mean being within 1 or 2 standard errors of the sample mean *increases* with the degrees of freedom (df). Exact values can be examined by tracing the curves with your mouse.



Conversely, the t-value associated with a given proportion or probability will also decrease as the degrees of freedom increase. The red line represents the t-values that define the area with a 68% chance of including the population mean. The blue line represents the t-values that define the area with a 95% chance of including the population mean. Exact values can be examined by tracing the curves with your mouse. Notice the t-value associated with a 68% chance of including the population mean approaches 1, while the t-value associated with a 95% chance approaches about 1.98.



Takeaway: the number of samples affects not only the standard error, but the t-distribution curve we use to solve for the probability that a value will occur, given our sample mean.

3.8 Confidence Interval

The importance of the number of samples the standard error, and the t-distribution becomes even more apparent with the use of confidence interval. A **confidence interval** is a range of values around the sample mean that are selected to have a given probability of including the true population mean. Suppose we want to define, based on a sample size of 4 from the soybean field above, a range of values around our sample mean that has a 95% probability of including the true sample mean.

The 95% confidence interval is equal to the sample mean, plus and minus the product of the standard error and t-value associated with 0.975 in each tail:

$$CI = \bar{x} + t \times se$$

Where CI is the confidence interval, \bar{x} is the sample mean, t is determined by desired level of confidence (95%) and degrees of freedom, and se is the standard error of the mean

Since the t-value associated with a given probability in each tail decreases as the degrees of freedom increase, the confidence interval narrows as the degrees of freedom increase – even when the standard error is unaffected.

Lets sample our yield population 4 times, using the same code we did earlier. Although I generally try to confine the coding in this course to the exercises, I want you to see how we calculate the confidence interval:

1. Let's set the seed using `set.seed(1776)`. When R generates what we *call* a random sample, it actually uses a very complex algorithm that is generally unknown to the user. The `seed` determines where that algorithm starts. By setting the seed, we can duplicate our random sample the next time we run our script. That way, any interpretations of our sample will not change each time it is recreated.

```
# setting the seed the same as before means the same 4 samples will be pulled
set.seed(1776)
```

2. Now, let's take 4 samples from our population using the `sample()` function. That function requires has arguments. First, `yield$yield_bu` tells R to sample the `yield_bu` column of the `yield` data.frame. The second argument, `size=4`, tells R to take four samples.

```
# collect 4 samples
yield_sample = sample(yield$yield_bu, size=4)
#print results
yield_sample
```

```
## [1] 82.40863 71.68231 73.43349 81.27435
```

3. Next, we can calculate our sample mean and sample standard deviation. We will assign these values to the objects `sample_mean` and `sample_sd`.

```
sample_mean = mean(yield_sample)
sample_sd = sd(yield_sample)

sample_mean
```

```
## [1] 77.1997
```

```
sample_sd
```

```
## [1] 5.427144
```

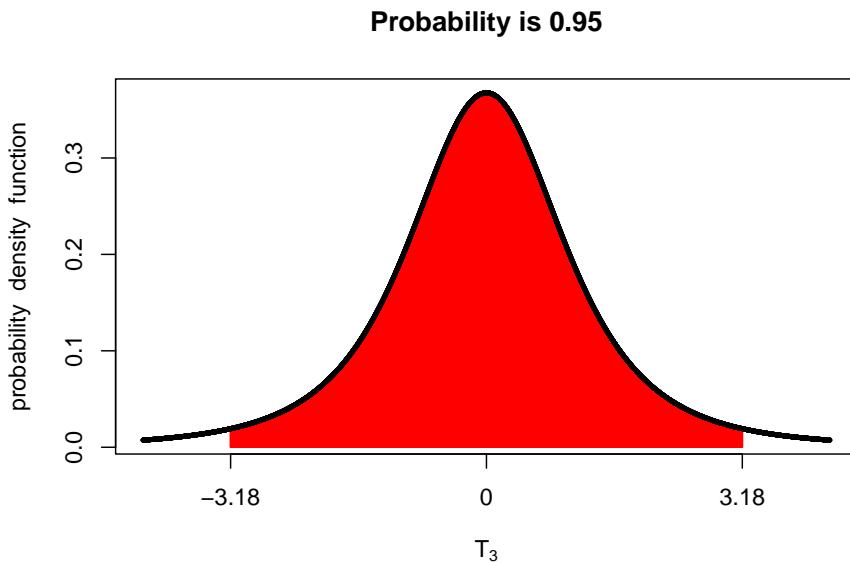
4. Finally, we can calculate the standard error, by dividing `sample_sd` by the square root of the number of observations in the sample, 4 . We assign this value to the object `sample_se`.

```
sample_se = sd(yield_sample)/sqrt(4)

sample_se
```

```
## [1] 2.713572
```

5. At this point, we have everything we need to calculate the confidence interval except the t-value. The middle 95% of the of the population will look like this:



There will be about 2.5% of the distribution above this range and 2.5% below it. We can calculate

```
# t-value associated with 3 df
upper_t = qt(p=0.975, df=3)
upper_t
```

```
## [1] 3.182446
```

6. Our lower limit is the t-value *below* which only 2.5% of the t-distribution occurs.

```
## [1] -3.182446
```

You will notice that "lower_t", the t-value that measures from the sample mean to the lower limit

7. We can then calculate our upper confidence limit. The upper limit of the confidence interval is equal to:

$$\text{Upper CL} = \bar{x} + t \cdot SE$$

Where \bar{x} is the sample mean, t is the t-value associated with the upper limit, and SE is the standard error of the mean.

```
upper_limit = sample_mean + upper_t*sample_se
upper_limit
```

```
## [1] 85.83549
```

8. We can repeat the process to determine the lower limit.

```
lower_limit = sample_mean + lower_t
lower_limit
```

```
## [1] 74.01725
```

9. Finally, we can put this all together and express it as follows. The confidence interval for the population mean, based on the sample mean is:

$$CI = 81.6 \pm 3.2$$

We can also express the interval by its lower and upper confidence limits.

$$(78.5, 85.4)$$

We can confirm this interval includes the true population mean, which is 80.1.

3.9 Confidence Interval and Probability

Lets return to the concept of **95% confidence**. This means if we were to collect 100 sets of 4 samples each, 95% of them would estimate confidence intervals that include the true population mean. The remaining 5% would not.

Click on this link to better explore this concept:

<https://marin-harbur.shinyapps.io/03-confidence-interval/>

Again, both the standard error and the t-value we use for calculating the confidence interval decrease as the number of samples increase, so the confidence interval itself will decrease as well.

Click on the link below to explore how the size (or range) of a confidence interval changes with the number of samples from which it is calculated:

<https://marin-harbur.shinyapps.io/03-sample-size-conf-interval/>

As the number of samples increases, the confidence interval shrinks. 95 out of 100 times, however, the confidence interval will still include the true population

mean. In other words, as our sample size increases, our sample mean becomes less biased (far to either side of the population mean), and it's accuracy (the proximity of the sample mean to the population mean) increases. In conclusion, the greater the number of samples, the better our estimate of the population mean.

In the next unit, we will use these concepts to analyze our first experimental data: a side by side trial where we will us the confidence interval for the difference between two treatments to test whether they are different.

Chapter 4

Two-Treatment Comparisons

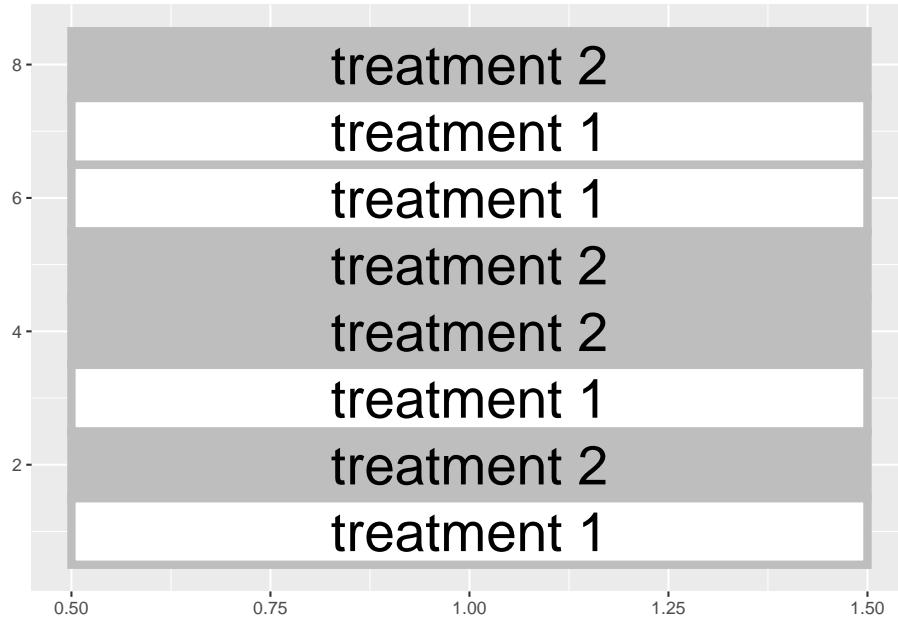
Until now, we have worked with a single population. To recap our progress: -

- In Unit 1, we learned a **population** is a complete group of individuals for which we wanted to develop a summary or prediction. We learned to describe the center of this population with the population mean and its spread using the sum of squares, variance, and standard deviation.
- In Unit 2, we used the **normal distribution** model to describe the pattern with which individuals in many populations are spread around the mean. Individuals closer in value to the population mean were predicted to occur more frequently than those further in value. Based on the frequency with which values were predicted to occur, we calculated the probability an individual from that population would fall within a given range of values.
- In Unit 3, we worked with **samples**, subsets drawn from a population. We saw how sample means are normally distributed, regardless of the population distribution from which they come. The standard error describes the spread of individual samples around the sample mean. This distribution was modeled with the t-distribution, which is wider when the number of samples is low and taller when the number of samples was greater.

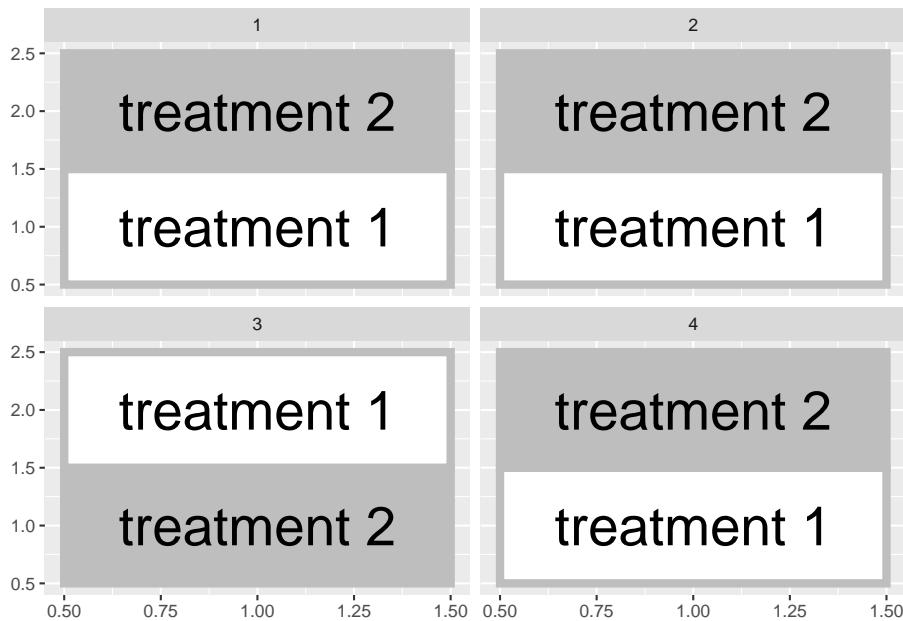
4.1 Side-by-Side Trials

In this unit, we will finally put our statistical knowledge to work to test treatment differences. We will work with a simple but important experimental design

– the two-treatment comparison. Here in Ohio, this is referred to as a side-by-side trial, but you may have a different term for it where you work. If you work in retail agronomy, you have probably conducted these trials. Typically, you would split one or more fields into treated and untreated fields. For example, you might “stripe” an individual field with treated and untreated areas:



Or, you might divide multiple fields in half like this:



In either case, a side-by-side trial deals with two treatments and can be analyzed using a t-test. In these next two units we will compare different designs for side-by-side (or paired) trials, use R to randomly assign treatments, understand how the t-distribution allows us to test for significance, and run these tests in R.

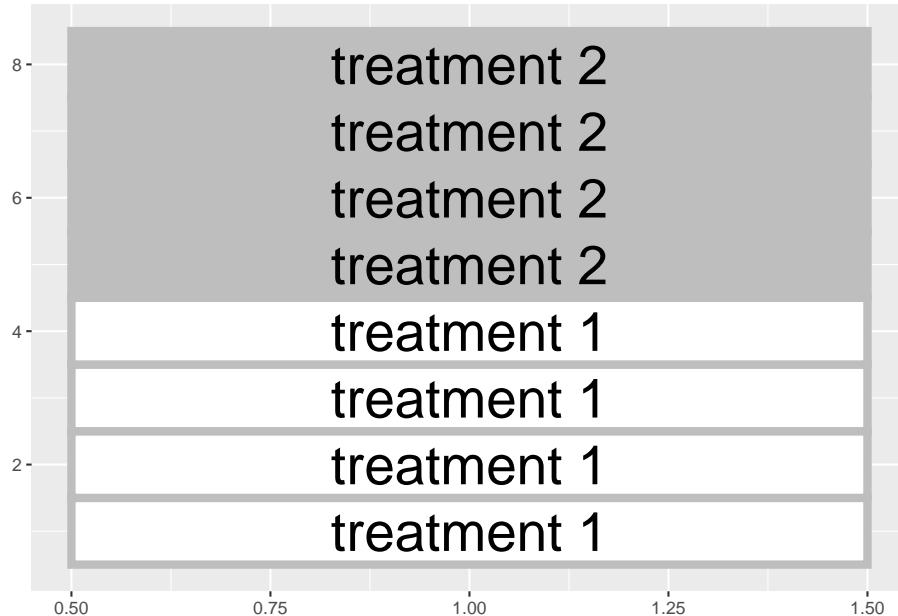
4.2 Blocked Design

In Unit 1 we learned the hallmarks of a designed trial are the **randomization** and **replication** of treatments. Each treatment should be observed several times in different experimental units. In our work, often **experimental unit** is a fancy term for a plot, half a field, or a pot.

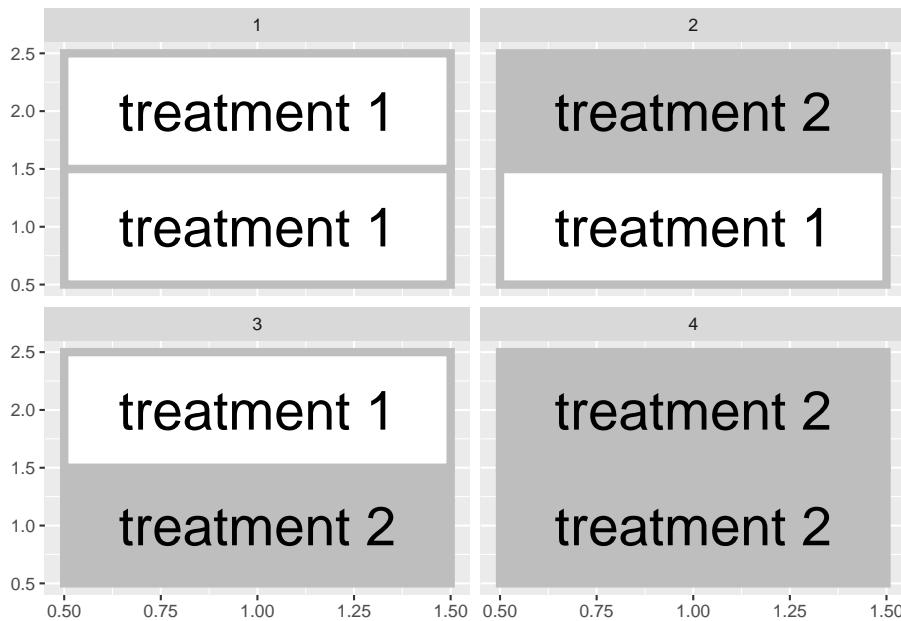
Observing the treatment several times has two benefits. First, the average of those observations – the sample mean – will likely be closer to the true population average for that treatment than the individual observations. Second, variation among the observations can be used to give us a sense how much environmental factors – in contrast with our treatment – cause our observations to vary.

It is also important to randomize treatments in order to reduce intentional or unintentional biases that might skew our interpretation of results. For example, one treatment might be biased by always putting it to the north of another treatment or, more insidiously, in the better half of a field. Deliberate randomization reduces the probability of either scenario.

That said, reality sometimes intervenes. Soil types change across a field, as do microclimates around field edges and management histories (e.g. an old feedlot). Though randomization reduces the likelihood that treatments are concentrated in one areas, it may not produce as even a distribution of treatments across the experimental area as we would like. We could conceivably end up with all replicates of a treatment level concentrated in one half of the field:



Similarly, if we are using multiple fields, both halves of a field could receive the same treatment in a randomized design



Blocked experimental designs place a restriction on the random assignment of treatments to plots. Instead of assigning treatments randomly within a field or across a state. We instead force both treatments to occur within a field section or in the same field, so that our treatment maps look more those we showed in the first two figures of this unit.

Here is another way to think about this:

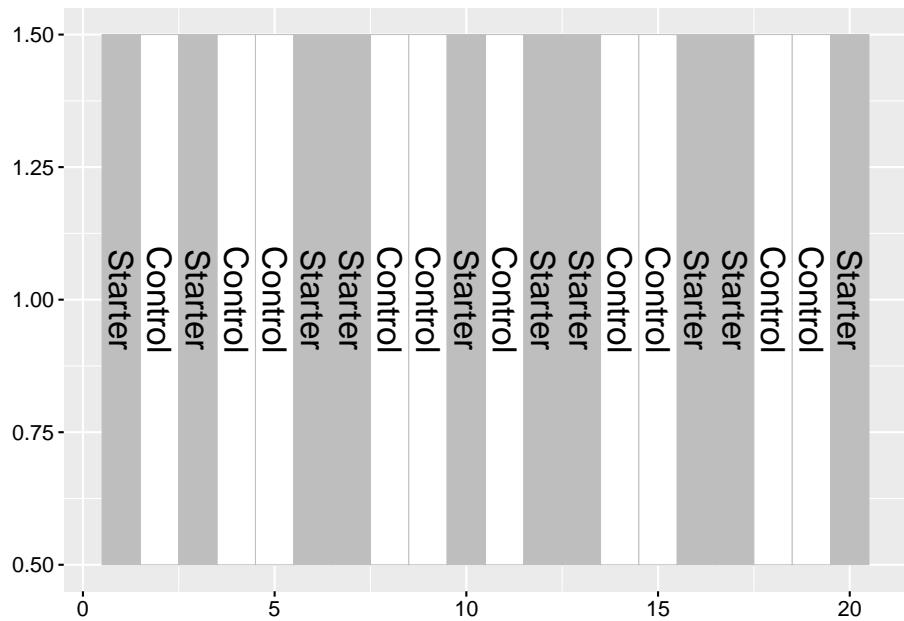
- Our statistical test this week is based on comparing samples from two populations, the control population and the population receiving starter.
- We want to design our experiment so that aside from the effect of the treatment, the two populations are as identical as possible.

The blocked approach, in general, helps create two populations that are similar.

In the exercise `exercise_randomizing_plots` in R, you will learn how to design your own blocked two-treatment trial.

4.3 Case Study

An in-furrow corn starter (*6-24-6 plus micronutrients*) was tested against a control (*no starter*) in a trial in western Ohio. Treatments were blocked (paired) so that each treatment occurred once per block. Plot numbers are given on the x-axis of the map below. There were 10 blocks.

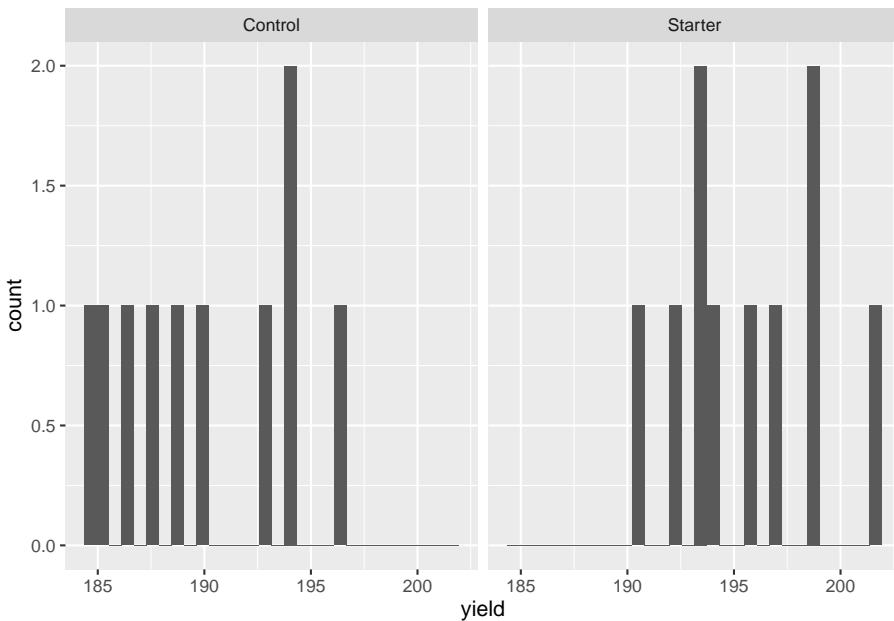


Here are the first six rows of our dataset.

block	plot	treatment	yield
1	11	Starter	193.4
1	12	Control	194.2
2	21	Starter	192.2
2	22	Control	189.0
3	31	Control	193.8
3	32	Starter	194.2

Note the column headings: *treatment* refers to the treatment level (Control or Starter) and *yield* refers to the measured yield.

Let's make a quick histogram plot of the data using ggplot. For the first time, in this unit, we are working with samples from two populations: the control and treatment populations. We will have separate histograms for each set of samples.



The mean yield of the sample from the population that received starter is greater than the mean yield of the sample of the population that received the control (no starter).

```
## # A tibble: 2 x 2
##   treatment yield
##   <chr>     <dbl>
## 1 Control    190.
## 2 Starter    196.
```

Instead of studying the control and treatment sample sets independently, however, we can restructure our dataset to analyze it another way. Data set structure describes how our columns and rows are organized. Do our treatment levels, for example, differ down rows or across columns? In the dataset above, for example, our two treatment levels (*control* and *starter*) vary among rows in the *treatment* column.

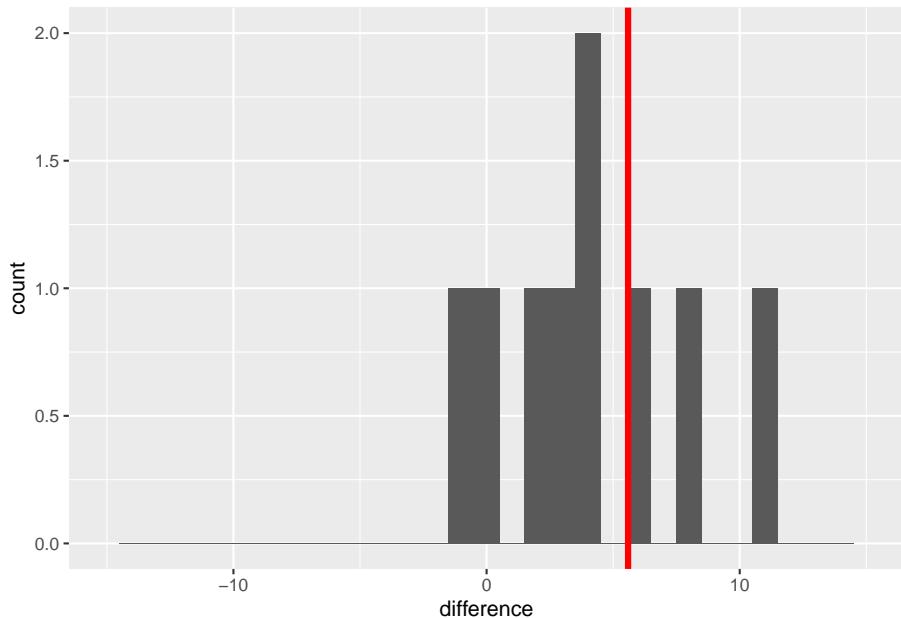
When we **restructure** a dataset, we rearrange it so that values that formally varied among rows vary across columns, or vice versa. If you have every used a *PivotTable* in *Excel*, you have restructured data. In **exercise_restructuring_data** in RStudio, you will learn how to similarly restructure data in R.

In our case study, we will restructure the treatment column by separating its values among two new columns, *control* and *starter*. We will create a new *difference* column by subtracting the yield of the control plot from the yield of

the starter plot in each block. By doing this, each block is now summarized by one value, the difference between treatments.

```
##    block Control Starter difference
## 1      1    194.2   193.4     -0.8
## 2      2    189.0   192.2      3.2
## 3      3    193.8   194.2      0.4
## 4      4    186.4   190.5      4.1
## 5      5    196.4   198.7      2.3
## 6      6    189.8   193.4      3.6
## 7      7    187.6   196.0      8.4
## 8      8    185.3   196.7     11.4
## 9      9    184.5   201.5     17.0
## 10    10    192.6   198.9      6.3
```

In this new population, individuals with a positive value indicate a block where the starter out-yielded the control. Negative differences indicate the control out-yielded the starter. We can plot these differences in a histogram, also.



Most of the individuals in this new population are greater than 0. The population mean, indicated by the vertical red line, is 5.6 bushels/acre. So far, our starter shows promise: a 5.6 bushel/acre mean *increase* in yield. Of course, this immediately raises the next questions:

- How consistently is there a positive response to this starter product?

- If we repeated this experiment, would we likely see a difference equal to zero or – even worse – one less than zero, in which case, yield actually *decreased* with the use of starter fertilizer?

We can use the t-distribution to answer this question, either by calculating the confidence interval of the mean difference, 5.6 bushels/acre, or testing the difference between the control and treatment means using a t-test.

4.4 Confidence Interval

In Unit 3, we worked with the sample mean, which is based on individual samples from an individual population. We defined the confidence interval around a sample mean, based on the variation in sample values around that sample mean.

We can use the same approach now on our sample differences: we will calculate the confidence interval around the mean difference between our control and treatment. If that interval contains zero or negative values, that means it is possible that control and treatment yields are the same, or even that the control yields more than the treatment.

Recall that to calculate the confidence interval we need two values:

- the minimal t-value that: is associated with our degrees of freedom and is expected to encompasses 95% of the distribution of sample means
- the standard error of the difference.

We calculate the minimal t-value using the `qt()` function in R.

```
min_t = qt(0.975, df=9)
min_t
## [1] 2.262157
```

Remember, degrees of freedom is equal to $n-1$. Our new population has 10 degrees of freedom, therefore, `df` equals 9 above. The minimal t-value is about 2.26 .

We also need to know the standard error of the population. We will calculate that the same as we did for a single variable in Unit 3. Given our sample set is composed of treatment differences, though, we will call this statistic the *standard error of the difference*.

First, we calculate the standard deviation of the treatment differences using the `sd()` function:

```
sd = sd(corn_differences$difference)
sd
```

```
## [1] 5.404001
```

Next, we divide the standard deviation by the square root of the number of differences. This is equal to the number of blocks.

Then we divide the standard deviation by the square root of the number of observations (10) to get the standard error of the difference:

```
sed = sd/sqrt(10)
sed
```

```
## [1] 1.708895
```

Finally, since the confidence interval is constructed with the sample mean as its center, we calculate the sample mean:

```
sample_mean = mean(corn_differences$difference)
```

Our sample mean is 5.59 and the standard error of the difference is about 1.71. We additionally calculated pop_mean, the mean of our population.

We can now calculate the confidence interval, using `sample_mean`, `min_t`, and `sed`. To calculate the lower limit, we subtract the product of `min_t` and `sed` from the sample mean:

```
lower_limit = sample_mean - (min_t * sed)
lower_limit
```

```
## [1] 1.724211
```

Similarly, the upper limit is calculated by adding the product of `min_t` and `sed` to the sample mean:

```
upper_limit = sample_mean + (min_t * sed)
upper_limit
```

```
## [1] 9.455789
```

It is common practice to present this confidence interval in parentheses, with the lower and upper limits separated by commas.

```
## [1] "(1.72421086829983,9.45578913170016)"
```

It is also good practice to round your final results so that your answers have either the same number of decimal places as the original measures. By doing this, we more fairly the precision of the original measurements.

```
## [1] "(1.7,9.5)"
```

The 95% confidence interval ranges from 1.7 to 9.5 bushels/acre and, notably, does not include zero. Since zero is outside the 95% confidence interval, there is greater than a 95% probability the population mean is not equal to zero. Another way of saying this is there is less than a 5% probability that the population mean is equal to zero. Or, finally, the population mean is significantly different from zero at the 5% (or 0.05) level.

Going our population was composed of the difference of starter and control, we can say the starter effect is significantly different from the control at the 5% level.

Learn how to construct your own confidence interval in `exercise_confidence_interval` in RStudio.

4.5 T-Test

An alternative to the confidence interval is the t-test. The first step of the t-test is to calculate our observed t-value:

$$t = \frac{\bar{x} - \mu}{SED}$$

Where \bar{x} is the observed population mean, μ is the hypothesized population mean (usually 0, meaning no treatment difference), and SED is the standard error of the difference.

In our example above:

$$t = \frac{5.59 - 0}{1.71} = 3.27$$

Our observed t-value is about 3.27. If there were no difference between the starter and the control, of course, t would be equal to zero. So there is a difference between our observed mean and zero of $t = 3.27$. Using R, we can quickly calculate the probability of observing a value $t = 3.27$ above – or below our population mean of 5.59.

```
library(fastGraph)
shadeDist(xshade=c(-3.27, 3.27), "dt", parm2 = 9, lower.tail = TRUE)
```

Note that this time with this function we used the argument “lower.tail=TRUE”. This tells R to calculate the probability of values further than $t = 3.27$ both above and below the mean.

Again there were 10 observed differences in our population, so there were 9 degrees of freedom. The value that is returned is the probability the population mean is actually zero, given the population mean. The probability of this t-value (sometimes abbreviated as $Pr \geq |t|$) is very low, less than 0.01, or 1%.

Learn how to conduct your own t-test in **exercise_t_test** in RStudio.

4.6 Conclusion

In this unit, we learned to create the experimental design for a two-treatment test and compare two treatments using the confidence interval of the treatment differences and the t-test.

Before the first plot is planned, however, we must consider what is the objective of the experiment. Specifically, what question is it designed to address? What are the potential answers to that question, and how should they be tested to determine which answer is most probably correct?

In Unit 5, we will learn about the nuances of developing research questions and expressing them in hypotheses we can then test with experimental data.

Chapter 5

Understanding Statistical Tests

In the last unit, we introduced the concept of statistical testing and, although I endeavor to make this course as practical and painless as possible, I believe it worthwhile to spend a unit on some of the theory of statistical testing. This will help reinforce what we have learned so far in this course, and prepare us for the other statistical tests that lie ahead. Otherwise, it is easy to become ambiguous about what we are really testing, and unclear in reporting our results.

In the last unit, we discussed experimental design and quickly jumped into data analysis. This unit, we will walk through the thought processes that surround our trial, including: - identifying our research question - creating a model from our question - identifying the hypotheses our data will be used to test - recognizing that we can mistakenly accept or reject these hypotheses - understanding how the confidence interval and p-value describe our measured difference - incorporating pre-existing knowledge into our hypotheses and tests

This list seems intimidating, but we will take our time and break these down into as much plain language as statistics will allow.

5.1 Research Question

As I have likely said before in this course, the first think you must have to design an experiment is a clear, testable research question. The question should be answerable using quantitative data and specific about what those data will measure. Here are some examples of bad and good questions:

Bad: Is this fertilizer better than another? Good: Does corn treated with 6-24-6 fertilizer at planting differ in yield from corn that is not treated with an

in-furrow starter.

Bad: Is the winter wheat variety MH666 (“the Beast”) different than variety MH007 (“the Bond”)? Good: Does variety MH666 differ in test weight from variety MH007 ?

Bad: Does herbicide deposition agent “Stick-It!” perform differently than agent “Get-Down!” ? Good: Do “Stick-It!” and “Get-Down!” differ in the number of live weeds two weeks after their application with glyphosate?

Remember to be clear about what we are measuring. Otherwise, it is unclear whether we are testing fertilizer affects on corn yield or moisture at harvest. We don’t know whether we are comparing winter wheat yield or head scab. We don’t know whether we are measuring the effect of our deposition agent on weed survival or crop injury.

5.2 The Model

The word “model” probably makes you shudder and think of a crowded blackboard filled with mathematical equations.

Yes, models can be quite complex. All of you have worked with models, however, and most of you should recall this one:

$$y = b + mx$$

Where y is the vertical coordinate of a point on a graph, x is its horizontal coordinate, and b is the Y-intercept (where the line crosses the y-axis). The most interesting variable is often m , the slope. The slope is the unit increase in y with each unit increase in x .

Suppose we took a field of corn and conducted a side-by-side trial where half of the plots were sidedressed with 150 lbs treated with an N stabilizer. The other half were sidedressed with 150 lbs actual N plus 1 unit of nitrogen stabilizer. The mean yield of plots treated with N plus nitrogen stabilizer was 195 bu and the mean yield of plots treated with N alone was 175 bu. How could we express this with a slope equation?

First, let’s state this as a table. We will express the N stabilizer quantitatively. The “No Stabilizer” treatment included zero units of N stabilizer. The “Stabilizer” treatment received 1 unit of stabilizer.

	Nitrogen	Yield
No Stabilizer	0	177.7
Stabilizer	1	198.0

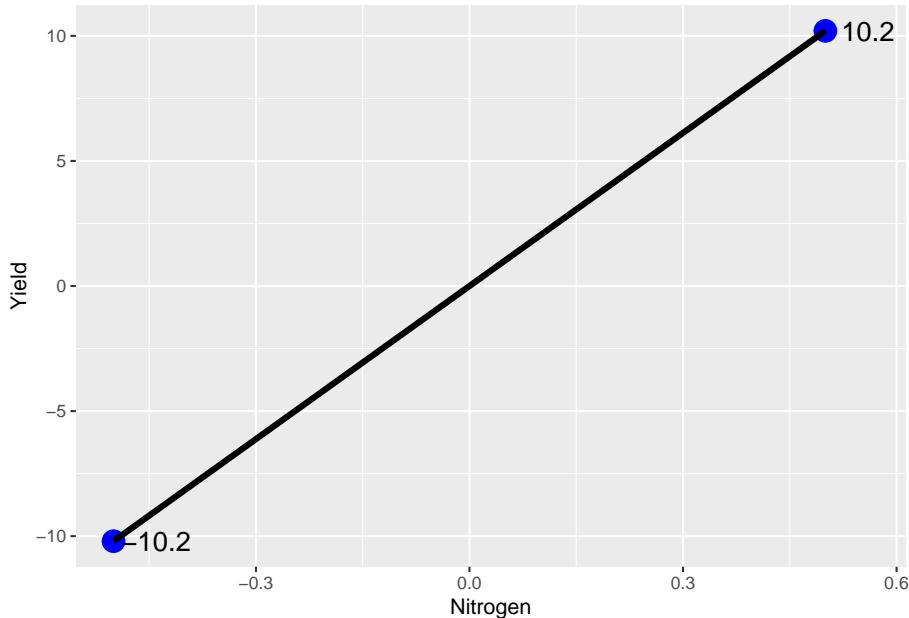
In creating this table, we also calculated the mean stabilizer rate and corn yield across all plots. These are are population means for the field.

Now, let's express the stabilizer rate and yield little differently, this time by their differences from their population mean. In half of the plots, the N stabilizer rate was 0.5 less than the population mean of 187.9; in the other half, the rate was 0.5 greater. Similarly, the yield in half of the plots was about 10 bushels less than the population mean of 188.9; in the other half of the plots, it was 10 bushels greater. Our table now looks like this:

```
##          Nitrogen Yield
## No Stabilizer    -0.5 -10.2
## Stabilizer       0.5  10.2
```

What we have just done is a statistical process called *center scaling*. Centering expresses measures by their distance from the population mean, instead of as absolute values.

Now let's plot this out using our line equation. y equals yield. x equals nitrogen rate. b equals the mean yield, the y-intercept, which is zero in our centered data.



Ta da: our line plot. If we were to write this as a linear equation, it would be:

$$Yield = 0 + 20 * Stabilizer$$

Thus, as the N stabilizer rate increase from 0 (no stabilizer) to 1 (stabilizer), yield increases 20 bushels.

5.2.1 Treatment Effect

Another way of expressing the effect of the treatment levels is in terms of their distance from the mean yield across all plots. Where sidedressed with nitrogen alone, our mean yield is equal to the population mean minus 10. Conversely, where we sidedressed with nitrogen plus stabilizer, our yield is the population mean *plus* 10.2. We can express these treatment effects as:

$$\text{Unstabilized : } T_0 = -10.2$$

$$\text{Stabilized : } T_1 = +10.2$$

Our mean yield when corn is sidedressed with N without stabilizer is then equal to the mean yield across all plots plus the treatment effect:

$$\text{Unstabilized : } Y_0 = \mu + T_0$$

$$\text{Stabilized : } Y_1 = \mu + T_1$$

We can prove this to ourselves by plugging in the actual yields for Y and μ and the actual treatment effects for T_0 and T_1 :

$$\text{Unstabilized : } 175 = 185 + (-10.2)$$

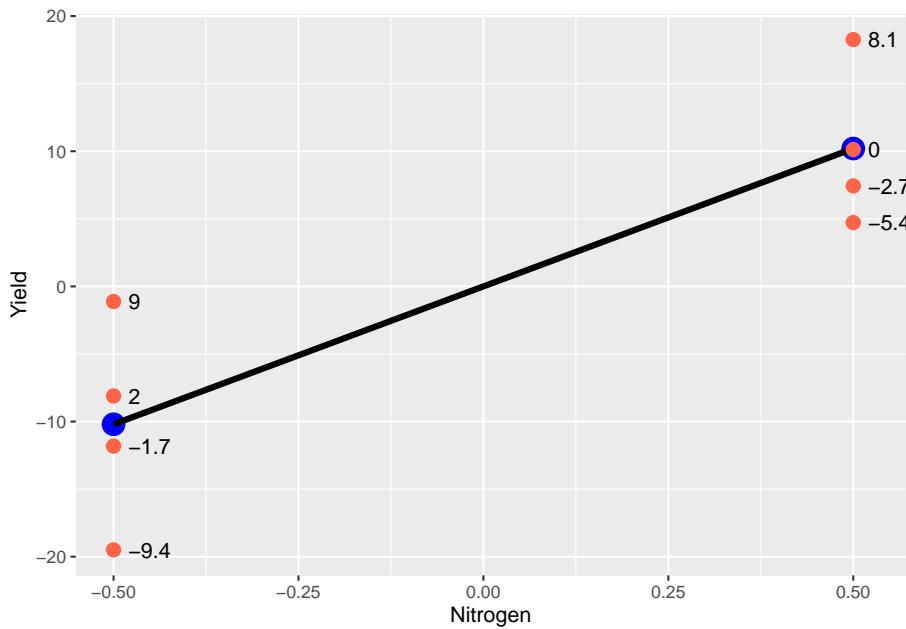
$$\text{Stabilized : } Y_1 = 185 + (+10.2)$$

5.2.2 Error Effect

The treatment effect is known as a *fixed* effect: we assume it will be consistent across all plots within our trial. That said, will every plot that receives nitrogen plus stabilizer will yield 195 bushels? Will every field sidedressed with nitrogen without stabilizer yield 175?

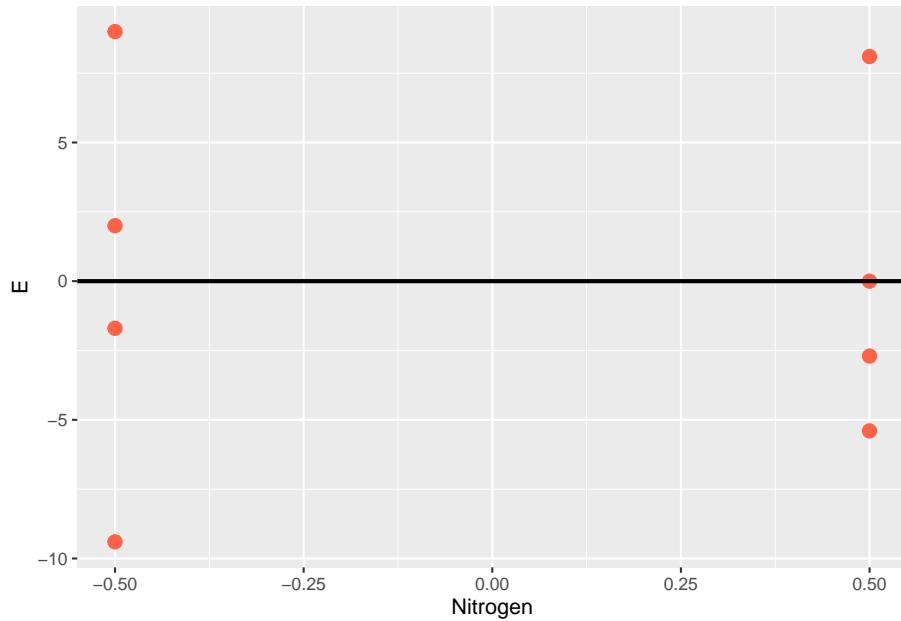
Of course not. Any yield map will show variation in yield within a field, even when the entire field has been managed uniformly. Differences in soil texture and microclimates, inconsistencies in field operations, and inaccuracies in measuring equipment contribute to variations in the values recorded. These variations will also add to or subtract from the mean yield across all plots.

We can visualize this in the plot below.



The blue points still represent the treatment mean, and the black line represents the difference between treatments. The red points are the original data – note how they are distributed around each treatment mean. Any individual observation is going to add to or subtract from its treatment mean. The value which each point adds to the treatment mean is shown to the right of the point. This is the error effect for that observation.

Sometimes it is easier to view the experimental unit effect another way, by excluding the treatment effect so that just the effects are plotted around their mean of zero:



This kind of plot is often called a *residual plot*, because the error can be thought of as the unexplained, leftover (ie “residue”) effect after the population mean and treatment effects are accounted for. When a correct model is fit to the data, about half the observations for each treatment should be greater than zero, and half below zero. The residual plot is a valuable tool to inspect and verify this assumption.

The yield observed in each plot, then, will be the sum of three values: - the mean yield across all plots - the effect of the treatment applied to that plot - the combined effect of environment, field operations, and measurements

This model can be expressed as:

$$Y_{ij} = \mu + T_i + \epsilon_{ij}$$

Where: - Y_{ij} is the yield of the i^{th} treatment level in the j^{th} block - μ is the yield mean across all plots - T_i is the effect of the i^{th} level of stabilizer - ϵ_{ij} is the *random* effect associated with the plot in the j^{th} block that received the i^{th} level of stabilizer

For example, a plot in the 3rd block that received nitrogen treated with stabilizer (T_1) would be indicated by the equation:

$$Y_{13} = \mu + T_1 + \epsilon_{13}$$

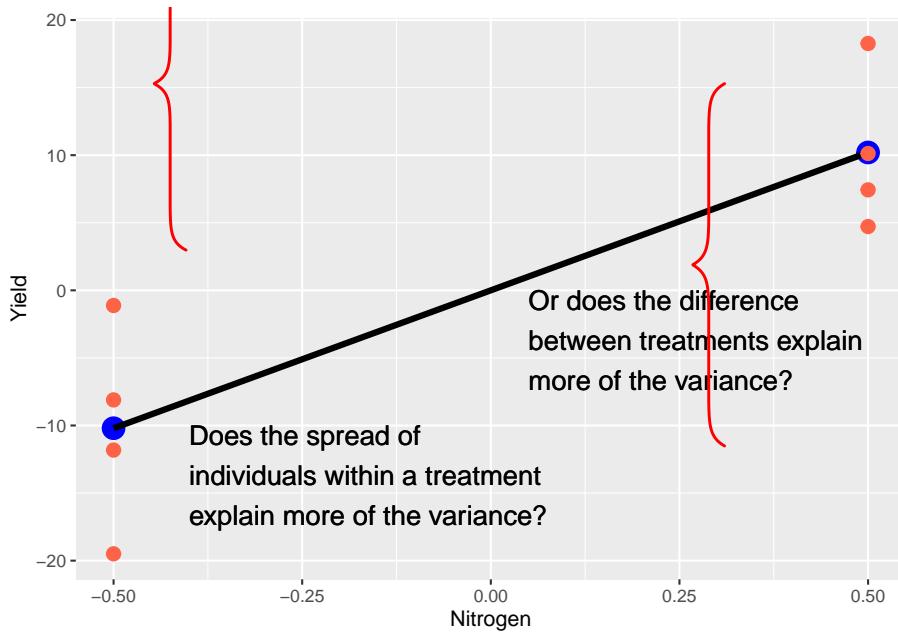
If the error for this plot, ϵ_{13} , was -2, the observed yield would be:

$$Y_{13} = 185 + 10 - 2 = 193$$

Why bother with the linear model when we simply want to know if one treatment yields more than the other? There are two reasons. First, although in agriculture we often think of field trials as testing differences, what we are really doing is using the data from those trials to *predict* future differences. In my opinion, this is one of the key differences between classical statistics and data science. Classical statistics describes what has happened in the past. Data science predicts what will happen in the future.

The linear model above is exactly how we would use data from this trial to predict yields if the product is used under similar conditions. Adding the stabilizer to nitrogen during sidedress will increase the mean yield for a field by 10 bushels. But any given point in that field will have a yield that is also determined by the random effects that our model cannot predict: soil and microclimate, equipment, and measurement errors.

Second, the linear model illustrates what statistics will test for us. Ultimately, every statistical test is a comparison between fixed and random effects: what explains the differences in our observations more: the fixed effects (our treatment) or random effects (error)? In our current example, we can visualize this as follows:



The purpose of a trial is to measure both types of effects and render a verdict. Which hypotheses are important, as we will now see.

5.3 Hypotheses

Before we design any experiment, however, we have to define our research question. In the case of a side-by-side trial, the question is generally: “Is one treatments better than the other? This question then needs to be translated into hypotheses.

Outside of statistics, a hypothesis is often described as “an educated guess.” Experiments are designed, however, to test two or more hypotheses. We may casually describe a side-by-side trial as comparing two treatments, but the data are formally used as evidence to test two, opposing hypotheses:

- H_0 : The difference between the two treatments is zero.
- H_a : The difference between the two treatments is not zero.

The first hypothesis, H_0 , is called the *null hypothesis*. The second hypothesis, H_a , is the *alternative hypothesis*. Typically, we tend to focus our effort on gathering enough evidence to support the alternative hypothesis: after all this work, we typically want to see a treatment difference! But we need to remember the null hypothesis may also be supported.

This process, like the linear model ahead, probably seems overly-formal at first. But like the linear model, it helps us to understand what statistics really tell us. We cannot prove either of these hypotheses. The world is full of one-off exceptions that will prevent either hypothesis from being universal truths. Our science is about comparing the evidence for each hypothesis, and selecting the hypothesis that is probable enough to meet our standards.

To illustrate this, look no further than the t-test we used in the last unit:

$$t = \frac{\bar{x} - \mu}{SED}$$

Recall that \bar{x} was our treatment difference, μ was the hypothesized treatment difference (zero), and SED was the standard error of the difference. The numerator is our treatment effect on plot yield. The denominator quantifies the random effects on plot yield. As this ratio increases, so does t . As t increases, the probability that the true population difference is zero decreases.

Another nuance of hypotheses is this, especially when it comes to the alternative hypothesis. If the evidence fails to support the alternative hypothesis, that does not mean it is wrong. The fixed (treatment) effect we observed was real. But the random effect was so great we could not rule out the differences we observed were the result of chance.

Simply put, our confidence interval was too big to rule out the true difference between treatments was actually zero. There was too much variation among plots. In a trial with a lower standard error of the difference, our t -value would have

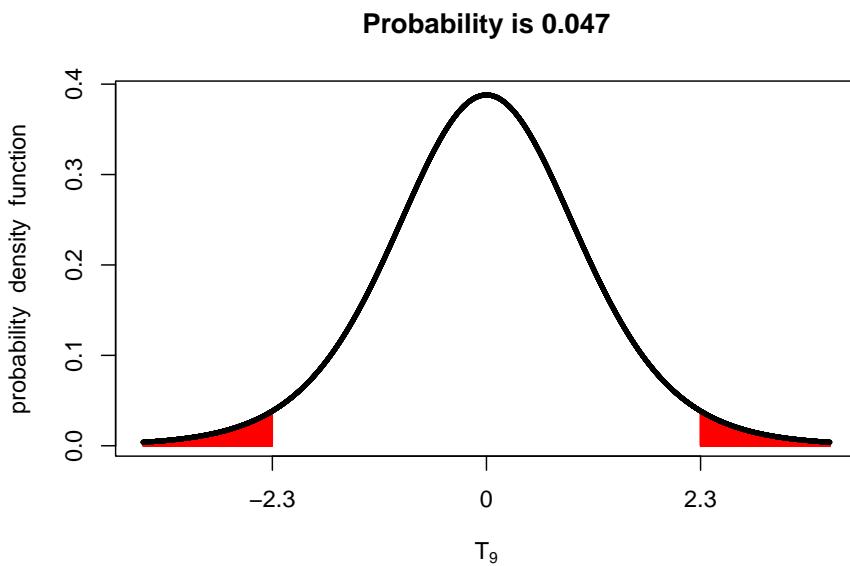
been greater, and the probability that the true difference between treatments was zero would be lesser.

Statistics is not about finding the truth. It is about quantifying the probability an observed difference is the result of chance. Lower probabilities suggest less chance in our observations, and the greater likelihood this difference will be observed if the trial is repeated by someone else, in a laboratory, or in a production field.

5.4 P-Value

The P-Value is always a source of confusion in statistics. What does it mean? What is so magical about the 0.05, or 5%, cutoff for declaring populations different? Even if you think you've mastered the P_value concept already, let's review it one more time.

The P-value, as applied to a distribution, is the probability of observing a value with a given (or greater) difference from the population mean. For example, if we have a t-distribution with a mean of 0 and a standard error of 1, the probability we will, the probability we will observe a value 2.3 standard errors away than the mean, given a population size of 4, is 0.047, or 4.7%.



What does a P-value of 0.047 mean? It means the probability of us measuring this value, by dumb luck, when the true population mean is 0, is about 4.7%. Put another way, given the standard error we observed, if the true population

mean was zero, we would observe a value equal to or more than 2.3 units away from the mean in less than 5 out of 100 trials.

If this concept sounds the same as that described for a confidence interval, that's because it is the same principle. If we constructed a 95% confidence interval around the sample mean of 2.3, we would see it excludes zero.

Knowing this, we have three options. We can conclude, for now, that the population mean really was zero and we were just very lucky (or unlucky) in our sampling.

We could also repeat the trial multiple times, to see what other values occur. If this is a field trial, that will incur additional research expenses. Even worse, it means delaying a recommendation for several seasons.

Our final option would be to conclude that our population mean is probably *not* zero. If the probability of observing a sample mean 2.3 or more units away from the mean, when the true population mean is zero, is 4.7% or less, then we can also say that the probability that the population mean has a value of zero or less, given our sample mean of 2.3, is 4.7% or less. Given this knowledge, we may conclude the true population mean is different from zero.

This is the power – and beauty! – of statistics. By properly designing an experiment (with sufficient replication and randomization), we can estimate the variability of individuals within a population. We can then use these estimates to test the probability of a hypothetical population mean, given the variability in our sample. And from that, we decide whether one population (which, for example, may have received a new product) is different from the other (which was untreated).

5.5 The P-Value and Errors

There are two kinds of P-value: the P-Value we measure, and the maximum P-Value we will accept before determining an observed difference between populations is insignificant. This maximum P-Value is referred to as *alpha* (α). You have probably seen statistical summaries that included whether treatments were “different at the $P \leq 0.05$ level. In this case, the α is 0.05, or 5%.

Before we discuss why an alpha of 0.05 or 5% is so often used for statistical tests, we need to understand how it relates to the likelihood we will reach the correct inference when comparing populations. You see, once we have gathered our data, calculated the variance in those populations (or, in the case of the paired t-test, the variance in their differences), and run our test(s), we will conclude either that the two populations are the same, or that they are different.

Either conclusion may be right. Or it may be wrong. And since we rarely measure entire populations, we never know their exact population means. We work with probabilities. In the above example, there was a 4.7% chance we

could observe a sample mean 2.3 units from a true population of zero. That means there is a 95.3 % (100 - 4.7) chance we would not see that value by chance. But there is still a chance. In other words, there is still a chance we could conclude the population mean is not zero, when in fact it is.

When we infer the mean of one population is significantly different from another (whether the second mean be measured or hypothesized), when in fact the two population means are equal, we commit a *Type I Error*. One example would be concluding the yield of one population, having received additional fertilizer, yielded more than an unfertilized population, when in fact their yields were equal. Another example would be concluding there is a difference in the percent of corn rejected from two populations, each treated with a different insecticide.

The P-value is the probability of making that Type I Error: of observing a sample mean so improbable enough that it leads us to conclude two populations are different, when for all purposes they are the same. If we are worried that recommending a product to a grower that does not increase yield will cost us their business, then we are worried about making a Type I Error. Outside of agriculture, if we are worried about releasing a treatment for COVID-19 that does not work and will leave users unprotected, we are worried about a Type I Error.

If we are really, really worried about wrongly inferring a difference between populations, we might use an even lower P-value. We might use $P=0.01$, which means we will not infer two treatments are different unless the mean difference we observe has less than a 1% probability of being observed by chance. This might be the case if the product we recommend to a grower is not \$10 per acre, but \$30. If our COVID treatment is very expensive or has serious side effects, we might insist on an even lower alpha, say $P=0.001$, or 0.1%.

So why not always use an alpha of 0.01 or 0.001 to infer whether two populations are different?

There is a second error we can make in the inferences from our research: we can conclude two populations are not different, when in fact they are. In this case, we observed, by chance, a sample mean from one population that was very close to the mean (hypothesized or measured) of another population, when in fact the two population means are different.

For example, a plant breeder might conclude a there performance of a new hybrid is similar to an existing hybrid, and fail to advance it for further testing. An agronomist might erroneously conclude there is no difference in plants treated with one of two micronutrient fertilizers, and fail to recommend the superior one to a grower.

Even more dangerously, a health researcher might conclude there is no difference in the incidence of strokes between a population that receives a new medication and an untreated population, and erroneously conclude a that mdeciation is safe.

Thus, there is a tradeoff between Type I and Type II errors. By reducing the alpha used as criterial to judge whether an one value is significantly different from another, we reduce the likelihood of a Type I error, but increase the likelihood of a Type II error.

In agronomic research, we conventionally use an alpha of 0.05. I cannot explain why we use that particular alpha, other than to suggest it provides an acceptable balance between the risks of Type I and Type II errors for our purposes. It is a value that assures most of the time we will only adopt new practices that very likely to increase biomass or quality, while preventing us wrongly rejecting other practices. In my research, I might use an alpha of 0.05 in comparing hybrids for commercial use. But I might use an alpha of 0.10 or 0.15 if I was doing more basic work in weed ecology where I was testing a very general hypothesis to explain their behavior.

To make things simple, we will use an alpha of 0.05 for tests in this course, unless states otherwise.

5.6 One-Sided vs Two-Sided Hypotheses

So far, we have treated our hypotheses as:

H₀: there is no difference between two populations, each treated with a different input or practice
H_a: there is a difference between two populations, each treated with a different input or practice

We could casually refer to these as “no difference” and “any difference”. But often in side-by-side trials, we have an intuitive sense (or hope) that one population will be “better” than another. If we are the yield of the two populations, one planted with an older hybrid and the other with a newer hybrid, we may be trying to determine whether the new hybrid is likely to yield more. If we comparing the number of infected plants in populations treated with different fungicides, we may hope the population that receives the new technology will have fewer diseased plants than the population that receives the older technology..

Is this bias? Yes. Is it wrong? I would argue no. The whole purpose of product development, in which I work, is to identify better products. Better hybrids, better micronutrients, better plant growth regulators. If we were equally satisfied whether a new product performed significantly better or significantly worse than an older product – well, in that case, I’d better look into teaching another section of this course.

It is okay to have this kind of bias, so long as we keep our research honest. Proper experimental design, including replication and randomization of plots in the field, or pots in the greenhouse, will go a long way towards that honest. So will selection of a P-value that acceptably minimizes Type I errors, so that

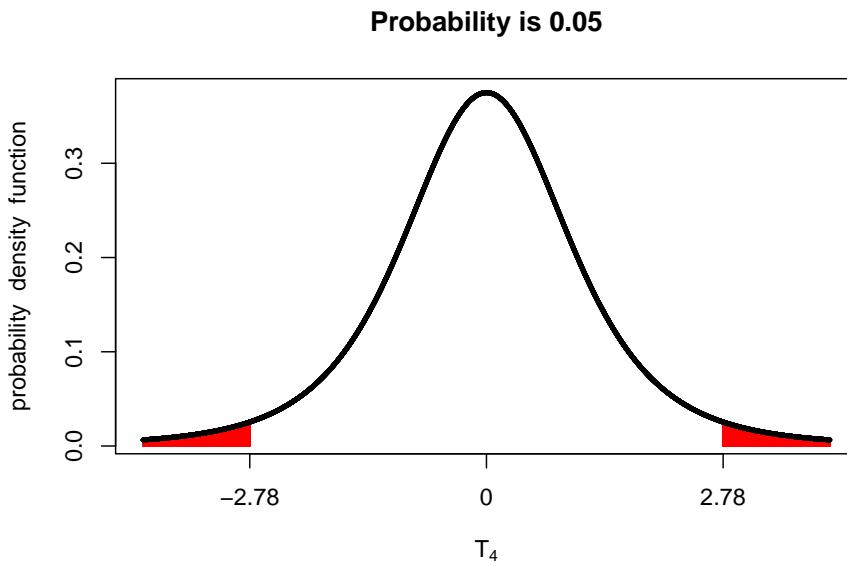
we don't advance a new product which isn't highly likely to perform better in future trials, or in the grower's field.

When we have this kind of bias, or interest, however, it also changes our hypotheses. Our null hypothesis is that the population treated with the new product will not perform better than the population treated with the old product. Our alternative hypothesis is the population treated with the new product will perform better.

If we are comparing yield in corn populations treated with a new fungicide, our hypotheses will be:

- H_0 : The yield of the population that receives the new fungicide will be equal too – or *lesser* – than the yield of the population that receives the old fungicide.
- H_a : The yield of the population that receives the new fungicide will be *greater* than the yield of the population that receives the old fungicide.

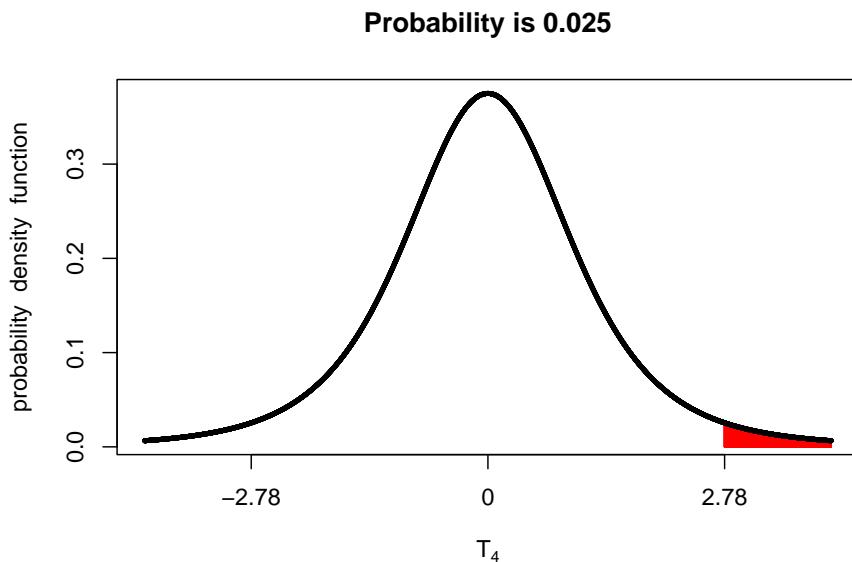
The reason these are called one-sided hypotheses is because we are only interested in one-side of the normal distribution. In the two-sided hypotheses we have worked with, we would only care if the yield of the two populations (one receiving the old fungicide, the other receiving the new fungicide) were different. To visualize this:



If either $t \leq -2.78$ or $t \geq 2.78$ (either of the red areas above), we declare the two populations different. In other words, the observed t-value can occur

in either of the two tails and be significant. Accordingly, we refer to this as a two-tailed test.

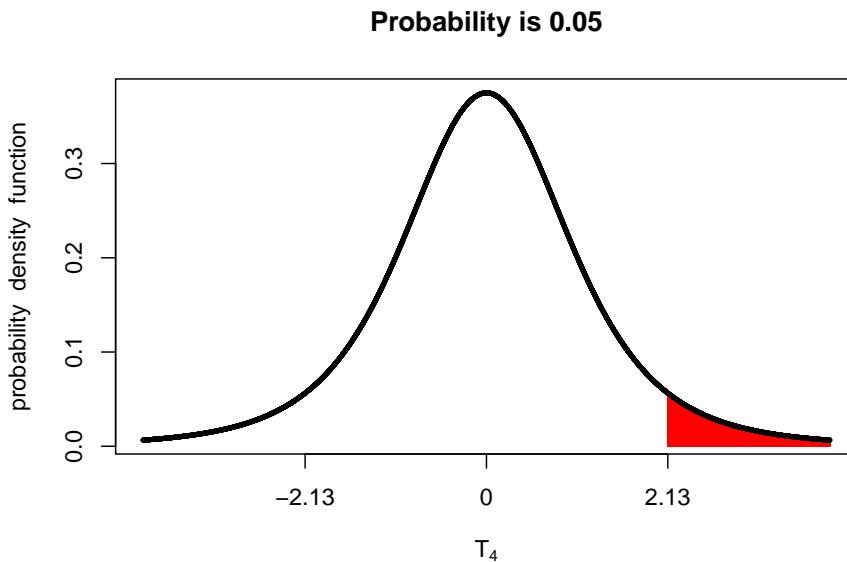
In testing a one-sided hypothesis, we only care if the difference between the population that received the new fungicide and the population that received the old fungicide (ie new fungicide - old fungicide) has a t-value of 1.98 or greater. We would visualize this as:



Only if the measured t-value falls in the upper tail will the population that receives the new fungicide be considered significantly better than the population that received the old fungicide. We therefore – you guessed it! – refer to this test as a one-tailed test.

In using a one-tailed test, however, we need to use a different t-value to achieve our alpha (maximum p-value for significance). If you look at the plot, only 2.5% of the distribution is in the upper tail. If we leave this as it is, we will only conclude the populations are different if their P-value is equal to or less than 2.5%. Reducing our P-value from 5% to 2.5% will, indeed, reduce our probability of Type I errors. But it will increase our likelihood of Type II errors.

If we are going to conduct a one-tailed test with an alpha of 0.05, we need to adjust the percentage of the distribution in the upper tail to 5% of the distribution:



The implication is that the minimum difference between populations to be significant at an alpha=0.05 is lesser than for a two-tailed test.

A common first response to the one-tailed test is: “Isn’t that cheating? Aren’t we just switching to a one-tailed test so we can nudge our difference passed the goal line for significance”? And, indeed, if you switch to a one-tailed test for that reason alone, it is cheating. That is why it is important we declare our hypotheses before we begin our trial. If we are going to run a one-tailed test, it needs to be based on a one-sided hypothesis that declares, from the beginning, we are only testing the difference in one direction, either because we have intuition that the difference will be one-sided, or we have a practical reason for only being interested in a positive or negative difference between populations.

Chapter 6

Multiple Treatment Trials

Here is our course, so far, in a nutshell:

- statistics is all about populations
- when we compare treatments, we are actually comparing populations that have been exposed to different products or management
- when we can measure every individual in that population, we can use the Z-distribution to describe their true population means and variance
- most of the time in agricultural research, however, we must use samples (subsets from those populations) to *estimate* their population means and variance
- when we use the sample mean to estimate the population mean, we use the t-distribution to describe the distribution of sample means around the mean of their values
- the distribution of sample means can be used to calculate the probability (the p-value) the true population mean is a hypothetical value
- in a paired t-test of two populations, we create a new population of differences, and calculate the probability its mean is zero
- proper hypotheses and alphas (maximum p-values for significance) can reduce the likelihood we conclude populations are different when they are likely the same, or the same when they are likely different

I include this brutal distillation so you can see how the course has evolved from working with complete populations to samples, from “true” or “certain” estimates of population means to estimates, from working with single populations to comparing differences between two populations.

In the last two units, we learned how to design and evaluate the results of side-by-side trials: trials in fields or parts of fields were divided into two populations that were managed with different treatments or practices. This was a practical, powerful, jumping-off point for thinking about experiments and their analyses.

Let's face it, however: if we only compared two treatments per experiment in product testing or management trials, our knowledge would advance at a much slower pace. So in the next three units, we will learn how to design and evaluate trials to test multiple *categorical* treatments. By *categorical*, we mean treatments we identify by name, not quantity. Treatments that, in general, cannot be ranked. Hybrids are a perfect example of categorical treatments. Herbicides, fungicides, or fertilizers that differ in brand name or chemical composition are also categorical treatments. Comparisons of cultural practices, such as tillage systems or crop rotations are categorical treatments as well.

6.1 Case Study

For our case study, we will look at a comparison of four hybrids from the Marin Harbur seed company. We will compare hybrids MH-052672, MH-050877, MH-091678, and MH-032990, which are not coded by relative maturity or parent lines, but represent some of the owner's favorite Grateful Dead concerts:

- MH-052672 has a great early disease package and is ideal for environments that have heavy morning dews - MH-050877 is very resilient in poor fertility soils and among the last to have its leaves fire under low nitrogen conditions. It can also grow well on everything from flatlands to mountains.
- MH-091678 has a very strong root base that will not be shaken down by late-season wind storms - MH-032990 offers a very rich performance and responds well to additional inputs. (This one catches growers' eyes at every field day – not to blow our own horn.)

plot_number	Hybrid	Yield	col	row
1	MH-050877	189.5	1	1
2	MH-052672	186.4	1	2
3	MH-091678	196.9	1	3
4	MH-050877	191.2	1	4
5	MH-091678	191.8	2	1
6	MH-032990	198.9	2	2

The hybrids were grown in a field trial with four replications as shown below.

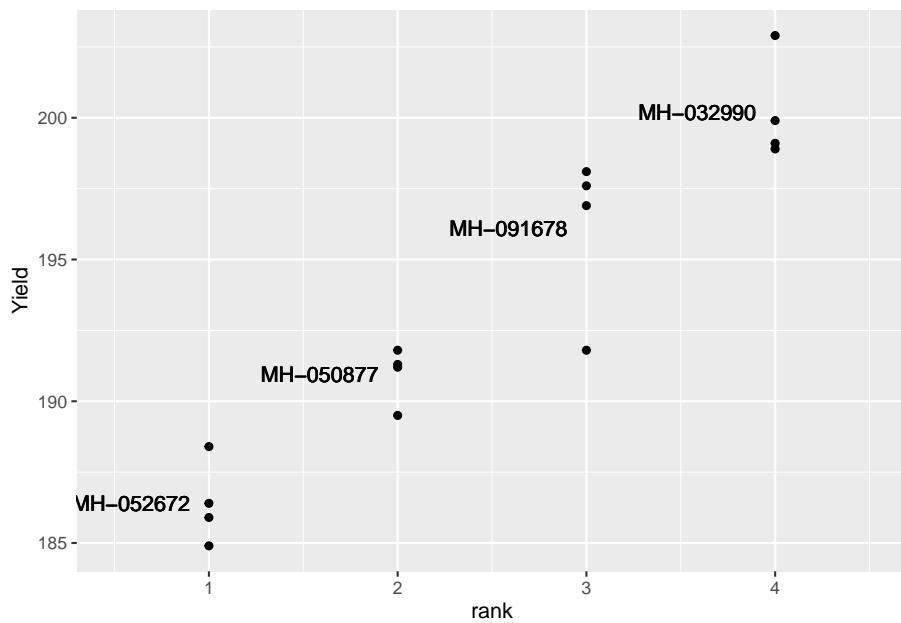
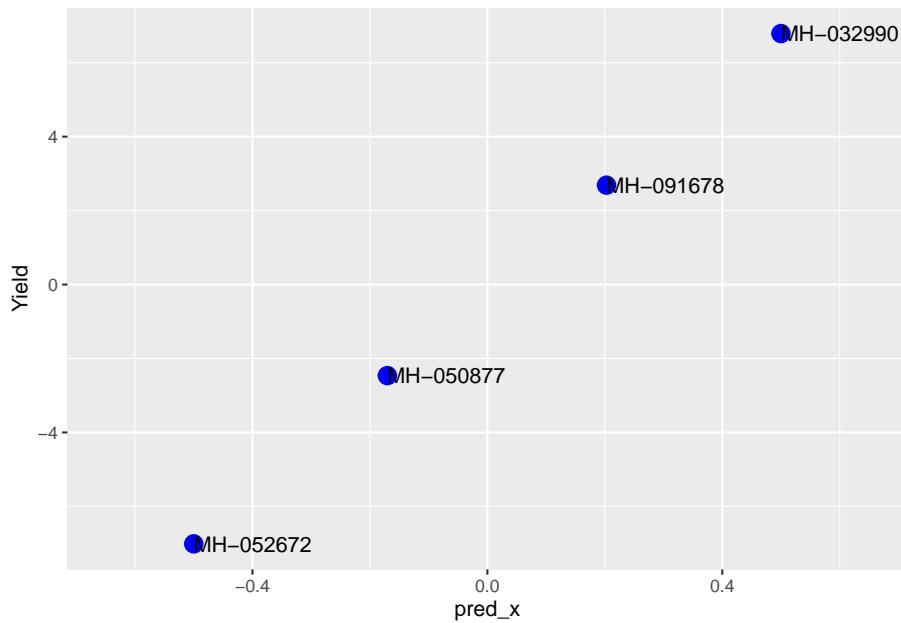
	1	2	3	4
row	MH-050877	MH-052672	MH-032990	MH-050877
1	MH-050877	MH-091678	MH-032990	MH-052672
2	MH-052672	MH-032990	MH-050877	MH-091678
3	MH-091678	MH-032990	MH-032990	MH-052672
4	MH-050877			

The arrangement of treatments above is a *completely randomized design*. This means the hybrids were assigned at random among all plots – there was no grouping or pairing of treatments as in our side-by-side trials earlier.

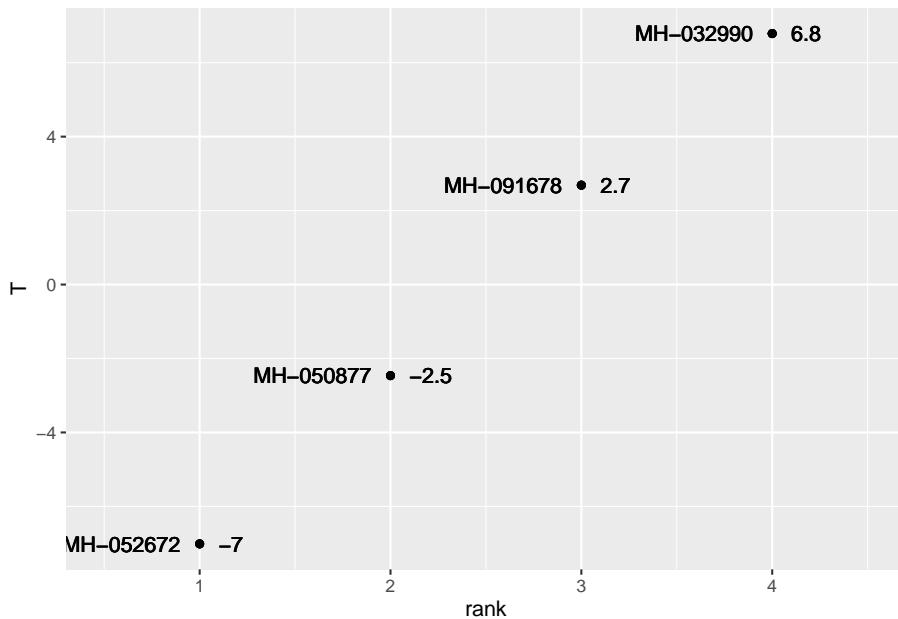
This design – to be honest – is used more often in greenhouse or growth chamber research where the soil or growing media are more uniform. Still, it is the most appropriate design to start with in discussing multiple treatment trials.

6.2 The Linear Additive Model

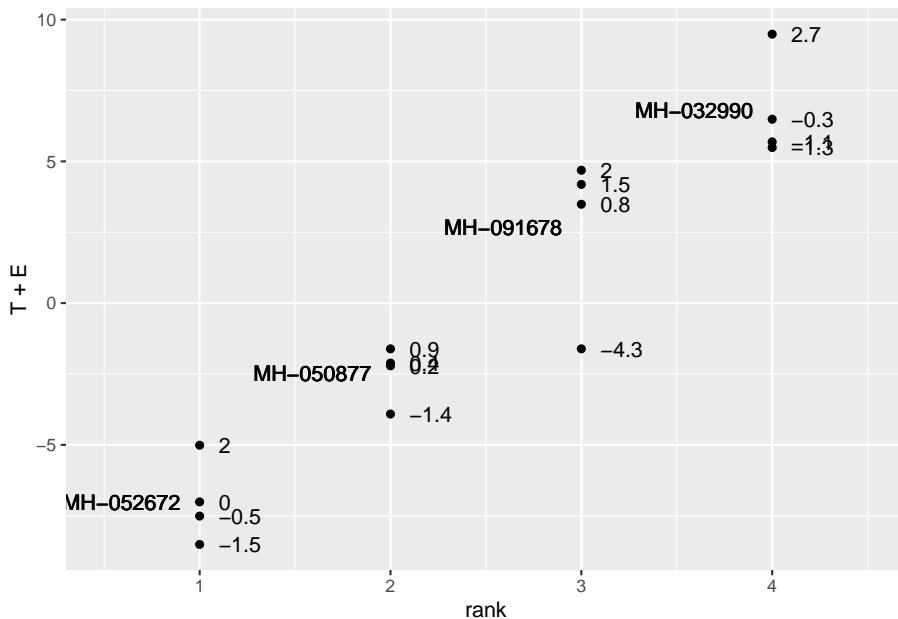
In order to understand how we will analyze this trial, let's plot out our data.



Our treatment means are shown below:



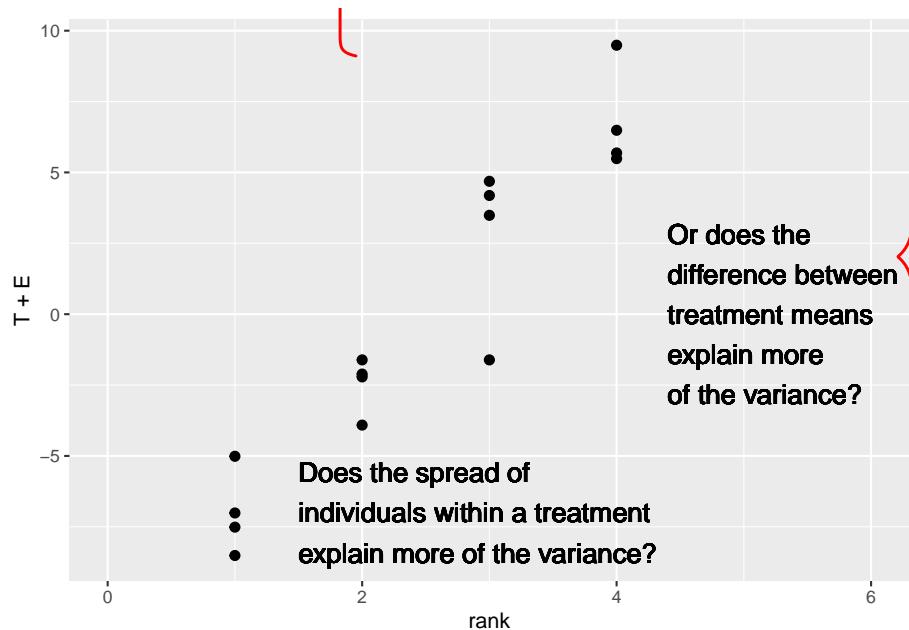
We can see the plot effects as well.



We can observe the linear model using this table:

Hybrid	mu	T	E	Yield (mu + T + E)
MH-032990	193.4	6.8	-1.3	198.9
MH-032990	193.4	6.8	-1.1	199.1
MH-032990	193.4	6.8	2.7	202.9
MH-032990	193.4	6.8	-0.3	199.9
MH-050877	193.4	-2.5	-1.4	189.5
MH-050877	193.4	-2.5	0.2	191.2
MH-050877	193.4	-2.5	0.9	191.8
MH-050877	193.4	-2.5	0.4	191.3
MH-052672	193.4	-7.0	0.0	186.4
MH-052672	193.4	-7.0	-1.5	184.9
MH-052672	193.4	-7.0	2.0	188.4
MH-052672	193.4	-7.0	-0.5	185.9
MH-091678	193.4	2.7	0.8	196.9
MH-091678	193.4	2.7	-4.3	191.8
MH-091678	193.4	2.7	2.0	198.1
MH-091678	193.4	2.7	1.5	197.6

Just as in the side-by-side trials, our statistical test is based on the ratio of the variation among treatment means to the variation among observations within each treatment.



6.3 Analysis of Variance

Chances are if you have spent time around agronomic research you have probably heard of *ANOVA*.



Figure 6.1: A Nova

No, not that fine example of Detroit muscle, but a statistical test, the *Analysis of Variance (ANOVA)*. The ANOVA test performs the comparison described above when there are more than two more populations that differ categorically in their management. I'll admit the nature of this test was a mystery to me for years, if not decades. As the name suggests, however, it is simply an analysis (a comparison, in fact) of the different sources of variation as outlined in our linear additive model. An Analysis of Variance tests two hypotheses:

- Ho: There is no difference among population means.
- Ha: There is a difference among population means.

In our hybrid trial, we can be more specific:

- Ho: There is no difference in yield among populations planted with four different hybrids.
- Ha: There is a difference in yield among populations planted with four different hybrids.

But the nature of the hypotheses stays the same.

6.4 The F statistic

The Analysis of Variance compares the variance from the treatment effect to the variance from the error effect. It does this by dividing the variance from treatments by the variance from Error:

$$F = \frac{\sigma^2_{\text{treatment}}}{\sigma^2_{\text{error}}}$$

The *F-value* quantifies the ratio of treatment variance to error variance. As the ratio of the variance from the treatment effect to the variance from the error effect increases, so does the F-statistic. In other words, a greater F-statistic suggests a greater treatment effect – or – a smaller error effect.

6.5 The ANOVA Table

At this point, it is easier to explain the Analysis of Variance by working with the output table of statistics.

term	df	sumsq	meansq	statistic	p.value
Hybrid	3	434.1275	144.709167	38.44388	2e-06
Residuals	12	45.1700	3.764167	NA	NA

As we did with the t-test a couple of units ago, lets go through the ANOVA output above column by column, row by row.

6.5.1 Source of Variation

The furthest column to the left, *term* specifies the two effects in our linear additive model: the Hybrid and Residual effects. As mentioned in the last chapter, the term Residual is often used to describe the “leftover” variation among observations that a model cannot explain. In this case, it refers to the variation remaining when the Hybrid effect is accounted for. This column is also often referred to as the *Source of Variation* column.

6.5.2 Sum of Squares

Let’s skip the *df* column for a moment to expain the column titled *sumsq* in this output. This column lists the sums of squares associated with the Hybrid and Residual Effect. Remember, the sum of squares is the sum of the squared

differences between the individuals and the population mean. Also, we need to calculate the sum of squares before calculating variance

The Hybrid sum of squares based on the difference between the treatment mean and the population mean for each observation in the experiment. In the table below we have mu, the population mean, and T, the effect or difference between the treatment mean and mu. We square T for each of the 16 observations to create a new column, T-square:

Hybrid	mu	T	T-square
MH-032990	193.4125	6.7875	46.070156
MH-032990	193.4125	6.7875	46.070156
MH-032990	193.4125	6.7875	46.070156
MH-032990	193.4125	6.7875	46.070156
MH-050877	193.4125	-2.4625	6.063906
MH-050877	193.4125	-2.4625	6.063906
MH-050877	193.4125	-2.4625	6.063906
MH-050877	193.4125	-2.4625	6.063906
MH-052672	193.4125	-7.0125	49.175156
MH-052672	193.4125	-7.0125	49.175156
MH-052672	193.4125	-7.0125	49.175156
MH-052672	193.4125	-7.0125	49.175156
MH-091678	193.4125	2.6875	7.222656
MH-091678	193.4125	2.6875	7.222656
MH-091678	193.4125	2.6875	7.222656
MH-091678	193.4125	2.6875	7.222656

We then sum the squares of T to get the Hybrid sum of squares.

```
## [1] 434.1275
```

We use a similar approach to calculate the Error (or Residual) sum of squares. This time we square the error effect (the difference between the observed value and the treatment mean) for each observation in the trial.

Hybrid	mu	E	E-square
MH-032990	193.4125	-1.30	1.6900
MH-032990	193.4125	-1.10	1.2100
MH-032990	193.4125	2.70	7.2900
MH-032990	193.4125	-0.30	0.0900
MH-050877	193.4125	-1.45	2.1025
MH-050877	193.4125	0.25	0.0625
MH-050877	193.4125	0.85	0.7225
MH-050877	193.4125	0.35	0.1225
MH-052672	193.4125	0.00	0.0000
MH-052672	193.4125	-1.50	2.2500
MH-052672	193.4125	2.00	4.0000
MH-052672	193.4125	-0.50	0.2500
MH-091678	193.4125	0.80	0.6400
MH-091678	193.4125	-4.30	18.4900
MH-091678	193.4125	2.00	4.0000
MH-091678	193.4125	1.50	2.2500

We again sum the squared error effects to get the Error or Residual sum of squares.

```
## [1] 45.17
```

6.5.3 Degrees of Freedom

The *df* column above refers to the degrees of freedom. Remember, the variance is equal to the sum of squares, divided by its degrees of freedom. The hybrid sum of squares is simply the number of treatments minus 1. In this example, there were 4 hybrids, so there were three degrees of freedom for the Hybrid effect. The concept behind the hybrid degrees of freedom is that if we know the means for three hybrids, as well as the population mean, then we can calculate the fourth hybrid mean, as it is determined by the first three hybrids and the population mean. Degrees of freedom are a weird concept, so try not to overanalyze them.

The degrees of freedom for the error or residual effect are a little more confusing. The degrees of freedom are equal to the Hybrid degrees of freedom, times the number of replications. In this case, the error degrees of freedom are 12. The idea behind this is: if for a hybrid you know the values of three observations, plus the hybrid mean, you can calculate the value of the fourth observation.

6.5.4 Mean Square

In the Analysis of Variance, the Sum of Squares, divided by the degrees of freedom, is referred to as the “Mean Square”. As we now know, the mean

squares are also the variances attributable to the Hybrid and Error terms of our linear model. Our hybrid mean square is about 144.7; the error mean square is about 3.8.

6.5.5 F-Value

The F-Value, as introduced earlier, is equal to the hybrid variance, divided by the error variance. In the ANOVA table, F is calculated as the hybrid mean square divided by the error mean square. When the F-value is 1, it means the treatment effect and error effect have equal variances, and equally describe the variance among observed values. In other words, knowing the treatment each plot received adds nothing to our understanding of observed differences.

6.5.6 P-value

The F-value is the summary calculation for the relative sizes of our Hybrid and Error variances. Its value is 38.4, which means the Hybrid variance is over 38 times the size of the Error variance. In other words, the Hybrid variance accounts for much, much more of the variation in our observations than the Error variance. But, given this measured F-value, what is the probability the true F-value is 1, meaning the Hybrid and Error variances are the same?

To calculate the probability our F-value could be the product of chance, we use the F-distribution. The shape of the F-distribution, like the t-distribution, changes with the number of replications (which changes the Error degrees of freedom). It also changes with the treatment degrees of freedom.

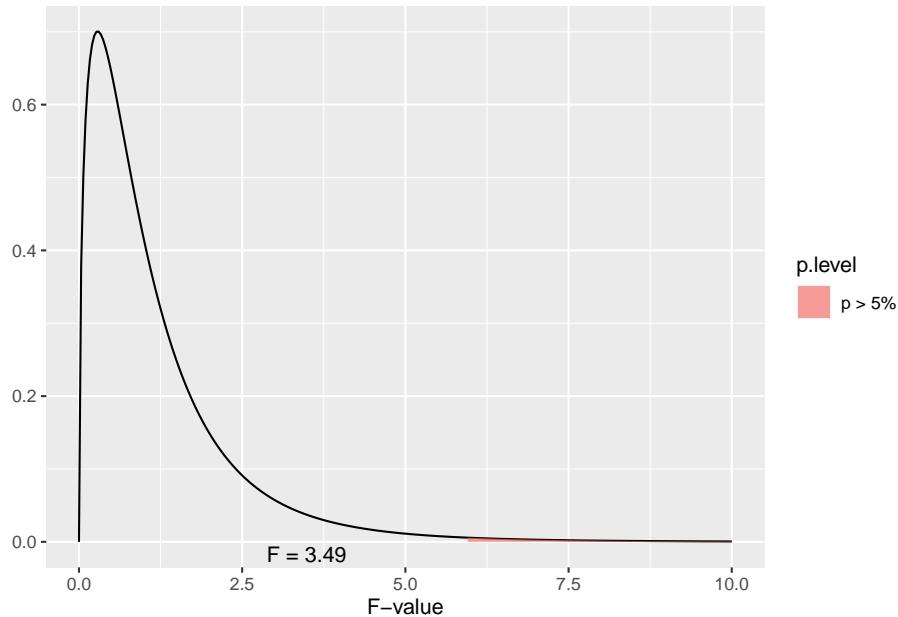
Please click the link below to access an app where you will be able to adjust the number of treatments and number of replications:

<https://marin-harbur.shinyapps.io/06-f-distribution/>

Adjust those two inputs and observe the change in the response curve. In addition, adjust the desired level of significance and observe how the shaded area changes. Please ignore the different color ranges under the curve when you see them: any shaded area under the curve indicates significance.

The F-distribution is one-tailed – we are only interested in the proportion remaining in the upper tail. If we were to visualize the boundary for the areas representing $P \geq 0.05$ for our example above, we would test whether F was in the following portion of the tail.

As we can see, our observed F of 38.4 is much greater than what we would need for significance at $P \geq 0.05$. What about $P \geq 0.01$?



Our value is also way beyond the F-value we would need for $P \geq 0.05$.

6.6 Visualizing How the Anova Table Relates to Variance

Please follow the following link to an app that will allow you to simulate a corn trial with three treatments:

<https://marin-harbur.shinyapps.io/06-anova-variances/>

Use your observations to address the following four questions in Discussion 6.1:

- 1) What do you observe when distance between the trt means increases?
- 2) what do you observe when the pooled standard deviation decreases?
- 3) Set the treatment means to treatment 1 = 180, treatment 2 = 188, and treatment 3 = 192. What do you observe about the shapes of the distribution curve for the treatment means (black curve) and treatment 2 (green curve)?
- 4) What does an F-value of 1 mean?

Chapter 7

Multiple Treatment Designs

In the last unit, we were introduced to multiple-treatment experiments, using an example with which many of you are familiar: a hybrid research or demonstration trial. Recall how the analysis of variance worked: we compared two sources of variation to see how much of that variation each of them explained. There were two effects in our original trial: treatment and error

$$Y_{ij} = \mu + T_i + \epsilon_{i(j)}$$

The Analysis of Variance (ANOVA) was used to calculate and compare these variances. First, the sums of squares from the treatment means and the error (the summed distributions of observations around each treatment mean) were calculated. By dividing the sum of squares by their degrees of freedom, the we obtained the treatment and error mean squares, also known as their variances. The F-value was derived from the ratio of the treatment variance to the error variance. Finally, the probability that the difference among treatments was zero, given the F-value we observed, was calculated using the F-distribution.

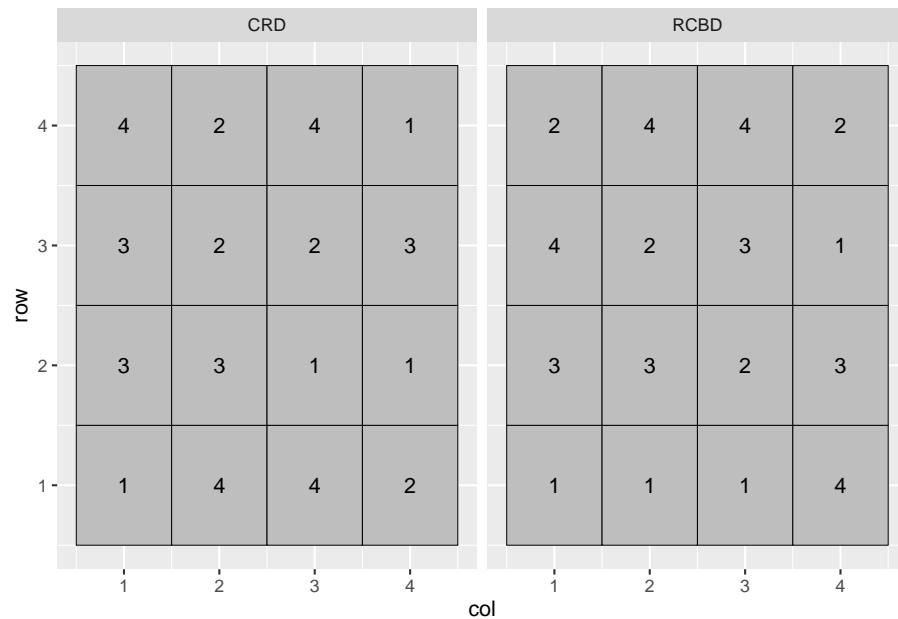
This experimental design is known as a *Completely Randomized Design*. It is the simplest multiple-treatment design there is. In this unit, we will learn two other designs commonly used in trials:

- Randomized Complete Block Design
- Two-Way Factorial Design

These designs, we will see, use additional sources of variation to either expand the number of treatments we can evaluate, or reduce the error (unexplained variation) in our trials so that we can better identify treatment effects.

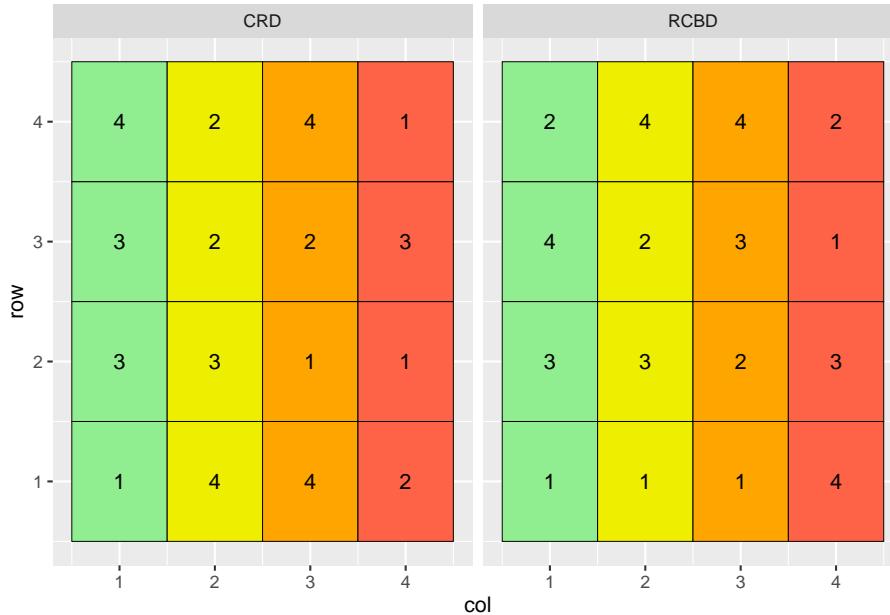
7.1 Randomized Complete Block Design

If you have participated in agronomic research, you have likely heard references to a Randomized Complete Block Design (RCBD). We first discussed blocking when we learned about side-by-side trials. When we block treatments, we force treatments to occur in closer proximity to each other than they likely would were they assigned at random. The best way to understand this is to look at a plot map.



In the plot above, the Completely Randomized Design (CRD) is shown on the left, and the Randomized Complete Block Design (RCBD) on the right. In the Completely Randomized Design, any treatment can occur anywhere in the the plot. Note that in the left plot, treatment 3 occurs twice in the first column of plots while treatment 2 does not occur at all. Treatment 2 occurs twice in the second column, but there is no treatment 1. In the Randomized Complete Block Design, each treatment must occur once, and only once, per column. In this case, the treatments are blocked on column.

Why block? Let's suppose each column in the plot map above has a different soil type, with soils transitioning from more productive to less productive as columns increase from 1 to 4:



Note that treatment 3 occurs three times in the more productive soils of columns 1 and 2, and only once in the less productive soils of columns 3 and 4. Conversely, treatment 1 occurs three times in the less productive soils of columns 3 and 4, but only once in the more productive soil of columns 1 or 2.

If the mean effect of treatment 3 is greater than the mean effect of treatment 1, how will we distinguish the effects of treatment and error? It is a moot question: we can't. Our linear model is *additive*:

$$Y_{ij} = \mu + T_i + \epsilon_{i(j)}$$

The term additive is important: it means we assume that treatment and error effects do not interact – they independently add or subtract from the population mean. If the measured effect of treatment is dependent on plot error, the model fails.

The plot on the right has blocked treatments according to column. Doing that allows us to remove the effect of soil type, which consistently varies from column to column, from the effect of error, which is random. Our linear model changes as well:

$$Y_{ij} = \mu + B_i + T_j + BT_{ij}$$

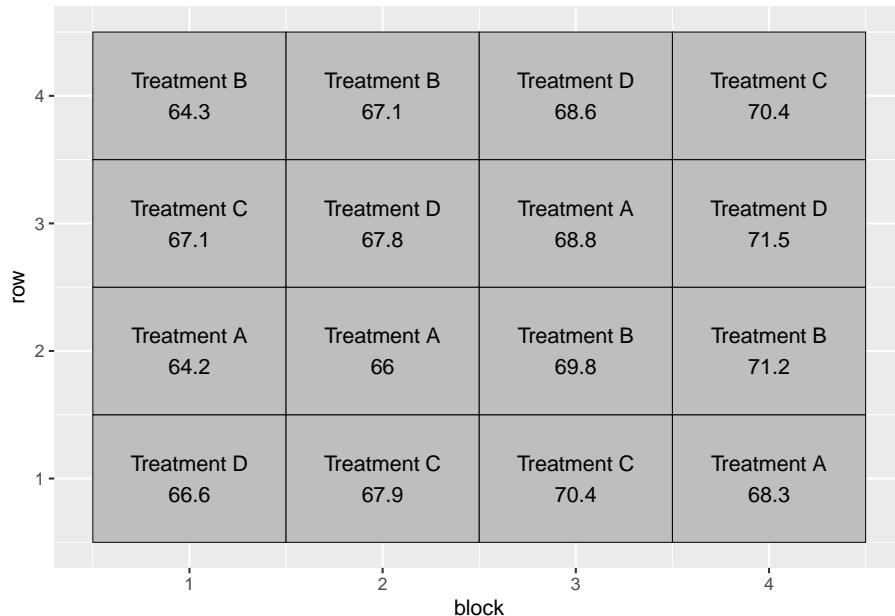
Where Y_{ij} is the individual value, μ is the population mean, B_i is the block effect, T_j is the treatment effect, and BT_{ij} is the interaction of block and treatment, also known as the error effect.

7.1.1 Case Study: Randomized Complete Block Design

A field trial outside Goshen, Indiana, evaluated the effect of seed treatments on soybean yield. Treatments were as follows:

- A: untreated
- B: metalaxyl only
- C: metalaxyl + insecticide
- D: metalaxyl + insecticide + nematicide

Treatments were arranged in a Randomized Complete Block Design.



7.1.1.1 Linear Additive Model

In this example, the linear additive model is:

$$Y_{ij} = \mu + B_i + T_j + BT_{ij}$$

Or, with regard to our particular trial:

$$\text{Yield} = \text{Population Mean} + \text{Block Effect} + \text{Treatment Effect} + \text{Block} \times \text{Treatment Interaction}$$

We can see how the additive model works in the following table:

Block	row	Treatment	mu	B	T	BT	Y
1	1	D	68.3	-2.2	1.6	-1.1	66.6
1	2	A	68.3	-2.2	-1.1	-0.8	64.2
1	3	C	68.3	-2.2	0.8	0.2	67.1
1	4	B	68.3	-2.2	-0.3	-1.5	64.3
2	1	C	68.3	-0.8	0.8	-0.4	67.9
2	2	A	68.3	-0.8	-1.1	-0.4	66.0
2	3	D	68.3	-0.8	1.6	-1.3	67.8
2	4	B	68.3	-0.8	-0.3	-0.1	67.1
3	1	C	68.3	0.3	0.8	1.0	70.4
3	2	B	68.3	0.3	-0.3	1.5	69.8
3	3	A	68.3	0.3	-1.1	1.3	68.8
3	4	D	68.3	0.3	1.6	-1.6	68.6
4	1	A	68.3	1.2	-1.1	-0.1	68.3
4	2	B	68.3	1.2	-0.3	2.0	71.2
4	3	D	68.3	1.2	1.6	0.4	71.5
4	4	C	68.3	1.2	0.8	0.1	70.4

In the first row of the table, we see that the observed yield, Y, is:

$$Y = 68.3 + (-2.2) + (1.3) + (-1.1) = 66.3$$

Similarly, in the fifth row:

$$Y = 68.3 + (-0.8) + (0.5) + (-0.4) = 67.6$$

7.1.1.2 Analysis of Variance

We can use the linear additive model above with R to create the proper linear model:

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Block      3 56.250 18.7500 23.7175 0.0001313 ***
## Treatment  3 10.485  3.4950  4.4209 0.0359018 *
## Residuals  9  7.115  0.7906
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that in the above model, we only specify the Block and Treatment sources of variation. Any source of variation not included in the model and, if you will “leftover” from the model, is pooled into the Residuals, or error. In the model

above, the interaction of Block and Treatment (BT) is not specified, for it is the source of any variation we observe from plot to plot.

Two sources of variation are tested above: Block and Treatment. The F-value for both is calculated by dividing their Mean Square by the Residual (or Error) Mean Square. The probability that the difference among blocks or treatment is zero, given their observed F-value, is reported in the $Pr(> F)$ column.

The Block effect is usually of less interest in analyzing table results. If it is insignificant, that may indicate we don't need to block in this location in the future, but in the vast majority of trials there will be at least some benefit to blocking.

The most important effect of blocking is seen upon examining the Sum of Squares ("Sum Sq") column. Here we can see just how much the Residual Sum of Squares was reduced by including blocks in the model. Had we not included the Block term in our model, our Residual Sum of Squares would have been about 63.4. Even given the greater residual degrees of freedom (which would have included the three degrees of freedom that were assigned to Block, the residual mean square would have been about $64 \div 12 = 5.3$. Without even calculating F, we can see the error mean square would have been larger than the treatment mean square, meaning there was more variance within treatments than between them. The Treatment effect would not have been significant.

7.2 Factorial Design

Agronomy is all about interactions. How do hybrids differ in their response to nitrogen? Response to fungicide? Does a fungicide increase yield more when it is sprayed on corn at V5 or VT? Does the effect of a starter fertilizer depend whether it is applied in-row or in a 2x2 band? How does a crop respond to population in 30-inch rows vs 15-inch rows?

To grow a successful crop requires not a single input, but dozens, some of which we can manage and others we can't. So the issue of how different treatments interact is critical. It is also often more efficient and informative for us to study these interactions together in a single trial, than to study them separately in two or more trials.

Before we go further, some nomenclature. In factorial design, treatments that differ in a single variable are called levels, and these levels together compose a factor. Here are some examples of levels and factors:

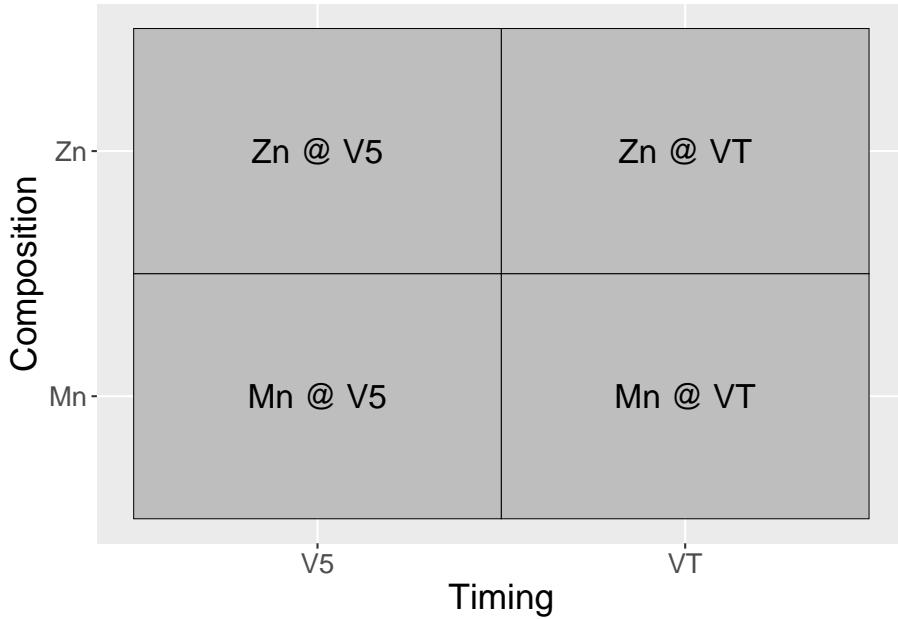
- multiple micronutrient products all applied at the V5 stage in corn – the product composition is the level, the timing the factor
- a single fungicide product, applied at V5, V10, or VT in corn – the timing is the level, the fungicide composition the factor

- a single adjuvant, tested with different nozzles – the nozzle is the level, the adjuvant composition the factor
- multiple hybrids, grown in rotation following soybean – the hybrid is the level, the crop rotation is the factor

As you may have anticipated, a factorial design combines two or more factors, each containing two or more levels. For example:

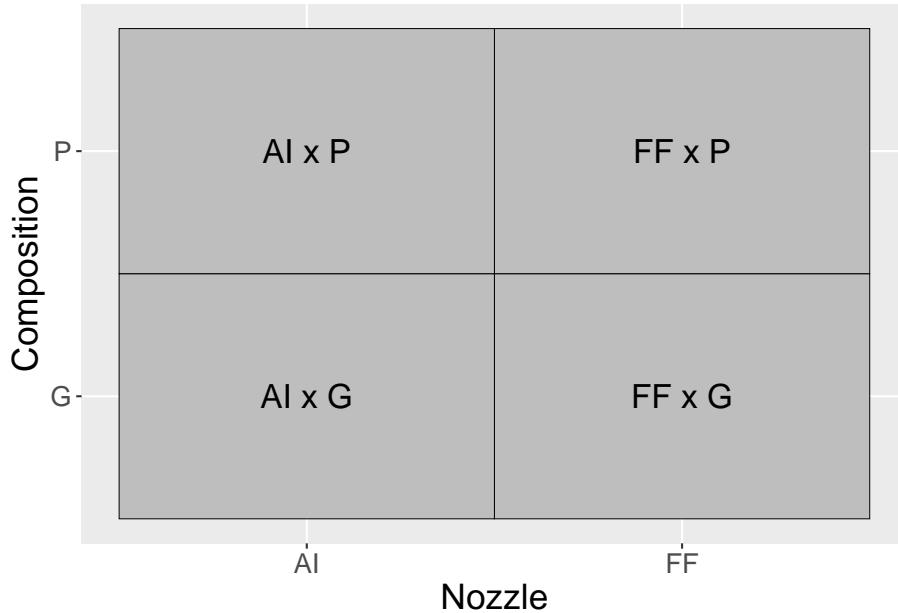
- Factor “Composition” includes two levels of foliar micronutrient product: Mn and Zn
- Factor “Timing” includes two levels of timing: V5 and VT

In a factorial design, every level of factor Composition will occur with every level of factor Timing. We can visualize these treatment combinations the same way we might visualize a Punnet square in Mendelian genetics. The main effects are given on the axes and the particular treatment combinations are in the cells.



In another example: - Factor “Composition” consists of two adjuvant ingredients: guar (G) or an polyacrylamide (P) - Nozzles are Flat Fan (F) or AI nozzle (A)

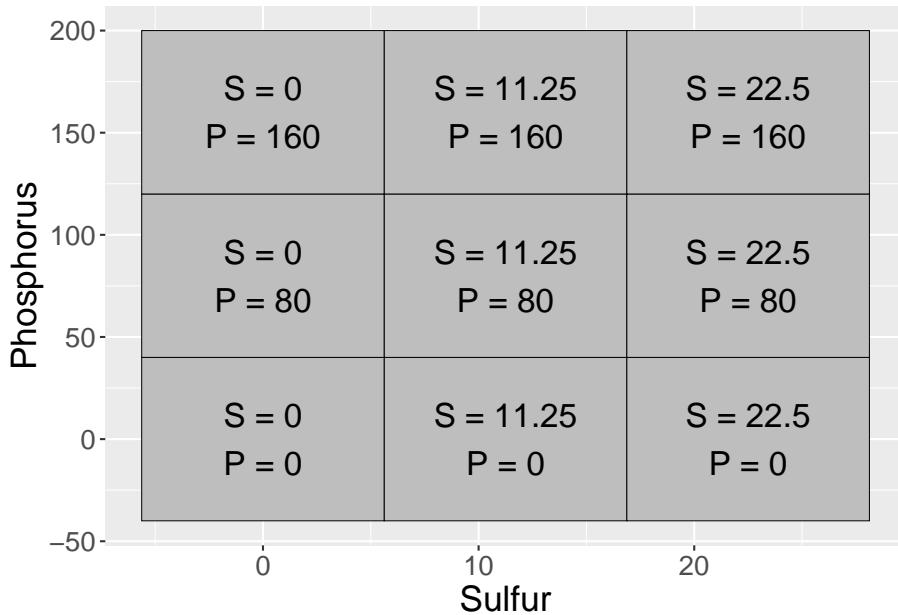
Our treatments in the factorial design, then, are:



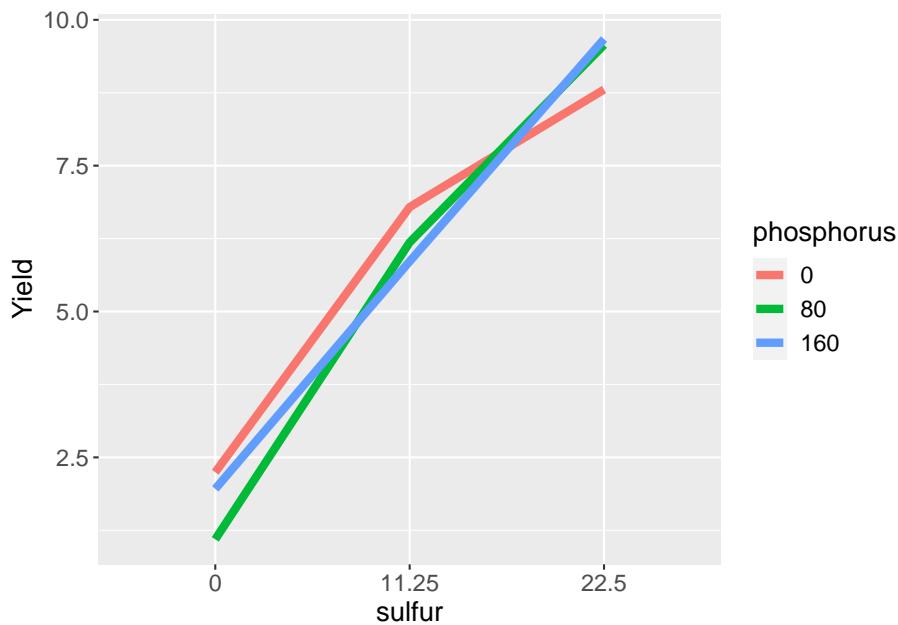
7.2.1 Case Study 1

Our case study is a clover forage trial conducted in New Zealand from 1992 to 1994. This dataset is publically available as part of the *agridat* package in R. For this first case study, we will focus on a subset of the data. The two factors were sulfur (S) and phosphorus (P) fertilizer. Sulfur was applied at 0, 11.25, or 22.5 kg/ha, while phosphorus was applied at 0, 40, and 80 kg/ha. Yield is reported in tons/hectare.

Factorial trials are often nicknamed by the number of levels of each factor. In this case, we have a *three – by – three*, or 3×3 trial. We can visualize the factorial combinations as producing the following nine treatments:



When we deal with factorial designs, it is important to visualize the data. We can observe the data patterns using a line plot. We see clover yield increased with sulfur and phosphorus. We notice, however, the difference in yield between P=80 and P=160 is greater when S=0 than when S=11.25 or S=22.5. We also notice the difference between sulfur=0 and sulfur=22.5 is greater for P=160 than P=0.



These differences are called *interactions* and are the most interesting part of a factorial design. In this case, it is no surprise that crop yield increases with sulfur fertilization in a sulfur soil. It is interesting, however, that the response to sulfur appears to be dependent on phosphorus fertilization, and lends support to Liebig's *Law of the Minimum*, which states that addition of a particular nutrient will not increase yield if other nutrients are more deficient.

7.2.1.1 Linear Additive Model

For our trial with two factors, the linear model is:

$$Y_{ijk} = \mu + S_i + P_j + SP_{ij} + \epsilon_{ij(k)}$$

Each observed value, then, is the sum (additive value) of the population mean (μ), sulfur effect (S), phosphorus effect (P), the interaction of sulfur and phosphorus (SP), and the error (ϵ). As in previous models, error is the random source of variation due to environmental and instrumental inconsistency. S_i and P_j are described as *main effects*. SP_{ij} is the interaction.

In the table below, the effects from this linear model are broken out:

block	sulfur	phosphorus	mu	S	P	SP	Error	Yield
1	0	0	5.72	-2.06	-2.06	0.37	-0.1	1.87
1	0	80	5.72	-2.06	-2.06	-0.38	0.2	1.42
1	0	160	5.72	-2.06	-2.06	0.01	0.1	1.71
1	11.25	0	5.72	0.24	0.24	0.29	0.2	6.69
1	11.25	80	5.72	0.24	0.24	0.16	-0.4	5.96
1	11.25	160	5.72	0.24	0.24	-0.44	0.0	5.76
1	22.5	0	5.72	1.81	1.81	-0.66	0.6	9.28
1	22.5	80	5.72	1.81	1.81	0.23	0.2	9.77
1	22.5	160	5.72	1.81	1.81	0.43	-0.3	9.47
2	0	0	5.72	-2.06	-2.06	0.37	0.4	2.37
2	0	80	5.72	-2.06	-2.06	-0.38	0.0	1.22
2	0	160	5.72	-2.06	-2.06	0.01	0.2	1.81
2	11.25	0	5.72	0.24	0.24	0.29	1.0	7.49
2	11.25	80	5.72	0.24	0.24	0.16	0.1	6.46
2	11.25	160	5.72	0.24	0.24	-0.44	-0.4	5.36
2	22.5	0	5.72	1.81	1.81	-0.66	-0.2	8.48
2	22.5	80	5.72	1.81	1.81	0.23	-0.1	9.47
2	22.5	160	5.72	1.81	1.81	0.43	0.2	9.97
3	0	0	5.72	-2.06	-2.06	0.37	0.4	2.37
3	0	80	5.72	-2.06	-2.06	-0.38	-0.1	1.12
3	0	160	5.72	-2.06	-2.06	0.01	0.7	2.31
3	11.25	0	5.72	0.24	0.24	0.29	0.2	6.69
3	11.25	80	5.72	0.24	0.24	0.16	-0.6	5.76
3	11.25	160	5.72	0.24	0.24	-0.44	0.9	6.66
3	22.5	0	5.72	1.81	1.81	-0.66	0.4	9.08
3	22.5	80	5.72	1.81	1.81	0.23	-0.3	9.27
3	22.5	160	5.72	1.81	1.81	0.43	0.0	9.77
4	0	0	5.72	-2.06	-2.06	0.37	0.4	2.37
4	0	80	5.72	-2.06	-2.06	-0.38	-0.6	0.62
4	0	160	5.72	-2.06	-2.06	0.01	0.4	2.01
4	11.25	0	5.72	0.24	0.24	0.29	-0.2	6.29
4	11.25	80	5.72	0.24	0.24	0.16	0.2	6.56
4	11.25	160	5.72	0.24	0.24	-0.44	-0.1	5.66
4	22.5	0	5.72	1.81	1.81	-0.66	-0.3	8.38
4	22.5	80	5.72	1.81	1.81	0.23	0.2	9.77
4	22.5	160	5.72	1.81	1.81	0.43	-0.3	9.47

7.2.1.2 Analysis of Variance

In R, we use the same approach as previous weeks. All terms from the linear model, except the population mean and error, are included in the model statement.

term	df	sumsq	meansq	statistic	p.value
sulfur	2	349.0468222	174.5234111	1227.117734	0.0000000
phosphorus	2	0.6720889	0.3360444	2.362812	0.1133374
sulfur:phosphorus	4	5.7705111	1.4426278	10.143477	0.0000381
Residuals	27	3.8400000	0.1422222	NA	NA

In the table, we see that the main effect of sulfur is significant at the $P \leq 0.05$ level. The phosphorus effect is not significant. The interaction (sulfur:phosphorus) effect is also significant at the $P \leq 0.05$ level.

When an interaction is significant, we should examine the effect of one factor independently at each level of the other. We can group analyses of one factor by levels of another factor using the *group_by* command in R. In this case we will tell R to run the analysis of variance for the sulfur effect separately for each level of P.

phosphorus	term	df	sumsq	meansq	statistic	p.value
0	sulfur	2	90.33447	45.1672333	264.8242	0
0	Residuals	9	1.53500	0.1705556	NA	NA
80	sulfur	2	145.58927	72.7946333	671.9505	0
80	Residuals	9	0.97500	0.1083333	NA	NA
160	sulfur	2	118.89360	59.4468000	402.2716	0
160	Residuals	9	1.33000	0.1477778	NA	NA

We see that the effect of sulfur is significant at each level of phosphorus. We can group analyses of phosphorus by each level of sulfur.

sulfur	term	df	sumsq	meansq	statistic	p.value
0	phosphorus	2	2.869267	1.4346333	17.331141	0.0008196
0	Residuals	9	0.745000	0.0827778	NA	NA
11.25	phosphorus	2	1.782067	0.8910333	3.734249	0.0659399
11.25	Residuals	9	2.147500	0.2386111	NA	NA
22.5	phosphorus	2	1.791267	0.8956333	8.507335	0.0084258
22.5	Residuals	9	0.947500	0.1052778	NA	NA

Whoa, what is going on here! We can now see the phosphorus effect is significant at S=0 and S=22.5, and almost significant at S=11.25. If we look at the line plot above, we see that phosphorus increases yield when S=0 and S=11.25, but decreases yield when S=22.5. The positive and negative effects cancelled each out when we looked at the overall analysis of variance – the interaction *masked* the sulfur effect so that its significance was not reflected in the results.

This demonstrates why it is so important to investigate interactions before making inferences about treatments. If we concluded from the first analysis of variance that sulfur affected yield, we would have been accurate. But if we had not analyzed the phosphorus effect separately by sulfur level, we would have erroneously concluded it did not affect yield.

7.2.2 Case Study 2

For the second case study, we are going to look at an experiment where turnips were planted at different densities in different row spacings. There were 5 densities and 4 row widths. This trial also blocked treatments, so it combines the randomized complete block and factorial designs.

7.2.2.1 Linear Additive Model

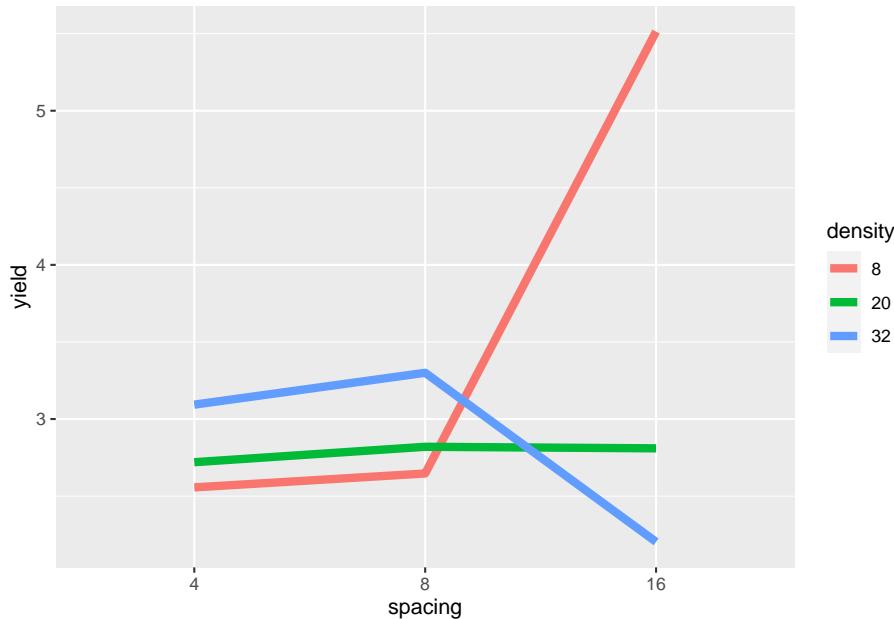
The linear additive model for this trial is:

$$Y_{ijk} = B_i + D_j + S_k + DS_{jk} + BDS_{ijk}$$

In this model, Y_{ijk} is yield, B_i is block, D_j is density, S_k is row spacing, DS_{jk} is the interaction of planting density and row spacing, and BDS_{ijk} is the interaction of block, density, and row width, which is used as the residual or error term in this model. We can see the additive effects of the factor levels and their interactions in the table below.

yield	block	spacing	density	mu	B	S	D	SD	Error
2.40	B1	4	8	2.67	-0.26	0.02	-0.17	0.29	-0.16
2.56	B1	4	20	2.67	-0.26	0.02	0.11	0.18	-0.16
2.78	B1	4	32	2.67	-0.26	0.02	0.06	0.30	-0.31
2.33	B1	8	8	2.67	-0.26	0.08	-0.17	0.32	-0.32
2.56	B1	8	20	2.67	-0.26	0.08	0.11	0.21	-0.26
3.03	B1	8	32	2.67	-0.26	0.08	0.06	0.25	-0.27
5.21	B1	16	8	2.67	-0.26	-0.10	-0.17	0.17	-0.30
2.56	B1	16	20	2.67	-0.26	-0.10	0.11	0.38	-0.25
1.90	B1	16	32	2.67	-0.26	-0.10	0.06	0.23	-0.30
2.60	B2	4	8	2.67	0.07	0.02	-0.17	-0.03	0.04
2.89	B2	4	20	2.67	0.07	0.02	0.11	-0.15	0.17
3.06	B2	4	32	2.67	0.07	0.02	0.06	-0.02	-0.03
2.63	B2	8	8	2.67	0.07	0.08	-0.17	-0.01	-0.02
2.69	B2	8	20	2.67	0.07	0.08	0.11	-0.11	-0.13
3.48	B2	8	32	2.67	0.07	0.08	0.06	-0.08	0.18
5.57	B2	16	8	2.67	0.07	-0.10	-0.17	-0.16	0.06
3.04	B2	16	20	2.67	0.07	-0.10	0.11	0.06	0.23
2.31	B2	16	32	2.67	0.07	-0.10	0.06	-0.10	0.11
2.67	B3	4	8	2.67	0.19	0.02	-0.17	-0.16	0.11
2.71	B3	4	20	2.67	0.19	0.02	0.11	-0.27	-0.01
3.44	B3	4	32	2.67	0.19	0.02	0.06	-0.15	0.35
2.98	B3	8	8	2.67	0.19	0.08	-0.17	-0.13	0.33
3.21	B3	8	20	2.67	0.19	0.08	0.11	-0.24	0.39
3.39	B3	8	32	2.67	0.19	0.08	0.06	-0.21	0.09
5.76	B3	16	8	2.67	0.19	-0.10	-0.17	-0.29	0.25
2.83	B3	16	20	2.67	0.19	-0.10	0.11	-0.07	0.02
2.40	B3	16	32	2.67	0.19	-0.10	0.06	-0.22	0.20

As we did before, we visually inspect the data for insights into how they may interact.



The plot gives critical insight into these data. Increasing spacing from 4 to 8 to 16 seems to cause a slight increase in yield where density=20. But where density=8, yield seems to increase rapidly between row spacing 8 and row spacing 16. Where density = 32, yield increases slightly with row spacing from 4 to 8, and then decreases markedly from 8 to 16.

The mean yield, averaged across row spacings, changes little across planting densities – even though the yields change dramatically within each of the individual row spacings. If we did not visually examine the data above, and instead relied on the ANOVA to alert us to an affect, we could miss this very important insight.

7.2.2.2 Analysis of Variance

Our analysis of variance is similar to that we ran for the first case study, except for it now includes the block term.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
block	2	0.9770963	0.4885481	25.06684	1.17e-05
density	2	3.3854519	1.6927259	86.85182	0.00e+00
spacing	2	2.6353852	1.3176926	67.60929	0.00e+00
density:spacing	4	16.3880593	4.0970148	210.21312	0.00e+00
Residuals	16	0.3118370	0.0194898	NA	NA

The planting density and plant spacing main effects were significant, as was their interaction.

Was the spacing effect significant at each level of density? We can slice the data to find this out.

density	term	df	sumsq	meansq	statistic	p.value
8	spacing	2	16.9677556	8.4838778	125.0694513	0.0000129
8	Residuals	6	0.4070000	0.0678333	NA	NA
20	spacing	2	0.0182000	0.0091000	0.1341523	0.8770078
20	Residuals	6	0.4070000	0.0678333	NA	NA
32	spacing	2	2.0374889	1.0187444	12.8701572	0.0067549
32	Residuals	6	0.4749333	0.0791556	NA	NA

We can see above that the effect of spacing on yield is only significant at $P \leq 0.05$ density=8 and density=32. If we examining the effect of density separately for each level of spacing:

spacing	term	df	sumsq	meansq	statistic	p.value
4	density	2	0.4540667	0.2270333	4.347447	0.0680698
4	Residuals	6	0.3133333	0.0522222	NA	NA
8	density	2	0.6872889	0.3436444	3.670979	0.0909484
8	Residuals	6	0.5616667	0.0936111	NA	NA
16	density	2	18.6321556	9.3160778	135.037365	0.0000103
16	Residuals	6	0.4139333	0.0689889	NA	NA

We similarly see that density is only significant at $P \leq 0.05$ where spacing=16.

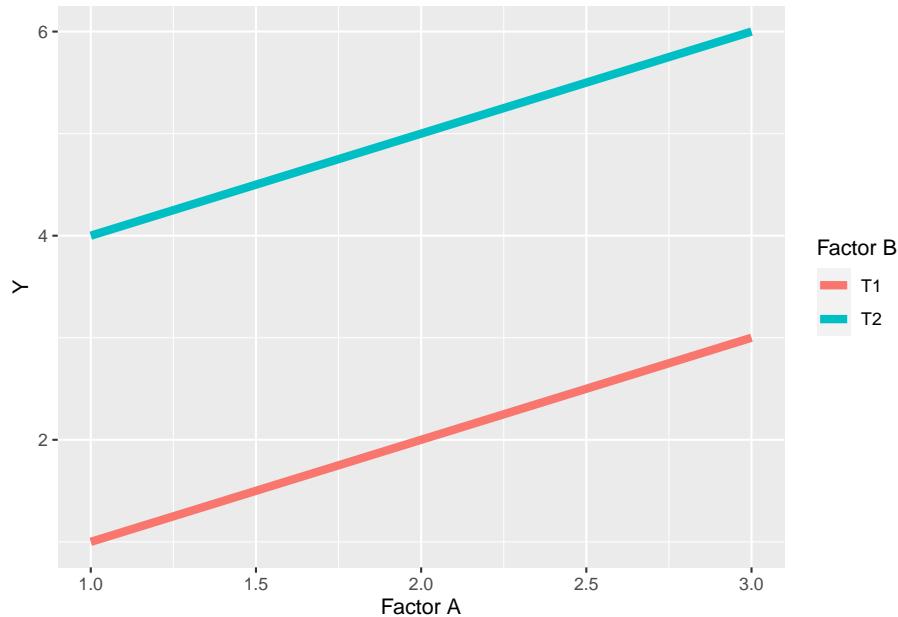
In this trial, both main (density and spacing) effects are significant. But if don't explore and explain our analysis further, we might miss how the both the magnitude of and the rank rank of row spacing effects on yield changes with plant density.

7.2.3 Discussing Interactions

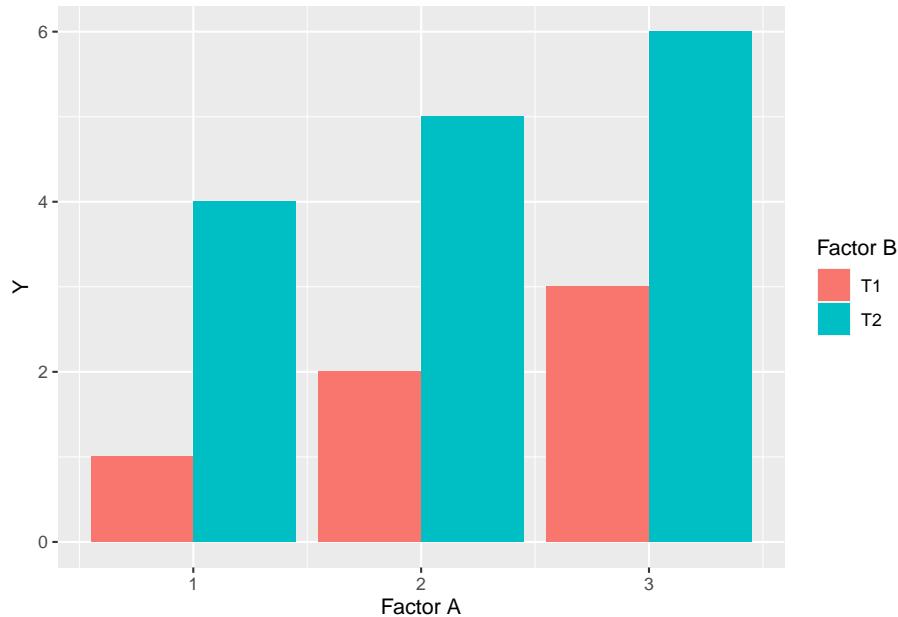
In factorial experiments, there are three kinds of interaction that may occur. We can have no interaction, a spreading interaction, or a crossover interaction.

7.2.3.1 No Interaction

Where there is no interaction, the treatments effects are simply additive – the observed value of each observation in the plot above is the sum of the effect of the level of Factor A plus the effect of the level of Factor B. The difference between levels of Factor A are consistent across levels of Factor B, and vice versa. If we plot the data using a line plot, it will look like this:



In a bar plot, it should look like this:

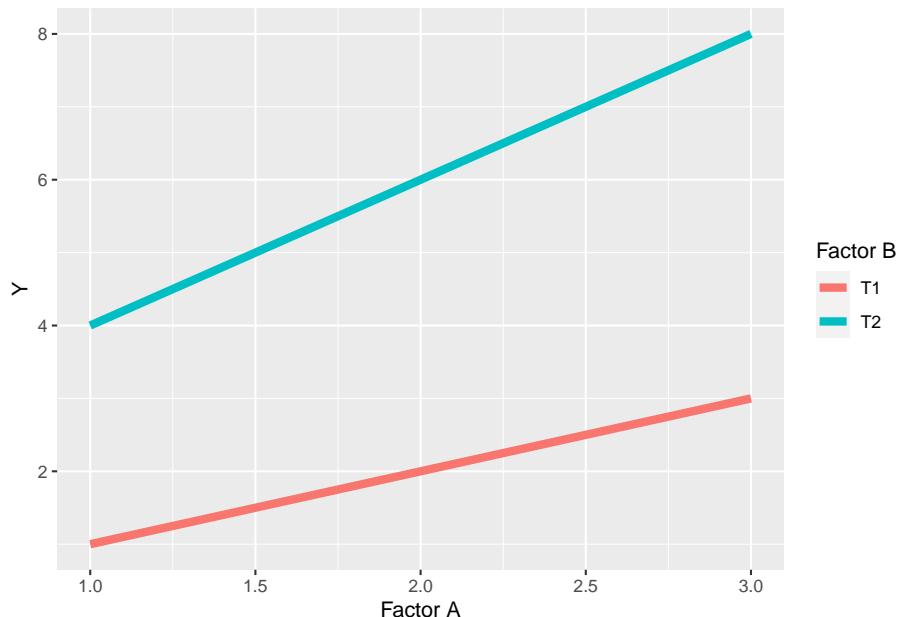


Finally, there will be no change in the ranking of levels within factors. Rank is the order of levels according to their observed effects, from least to greatest. For factor A, the ranking of levels within Factor A is $1 > 2 > 3$, while within Factor

B, level T2 always ranks higher than level T1.

7.2.3.2 Spreading Interaction

In a spreading interaction, the ranking of levels within factors does not change, but the difference between them does.

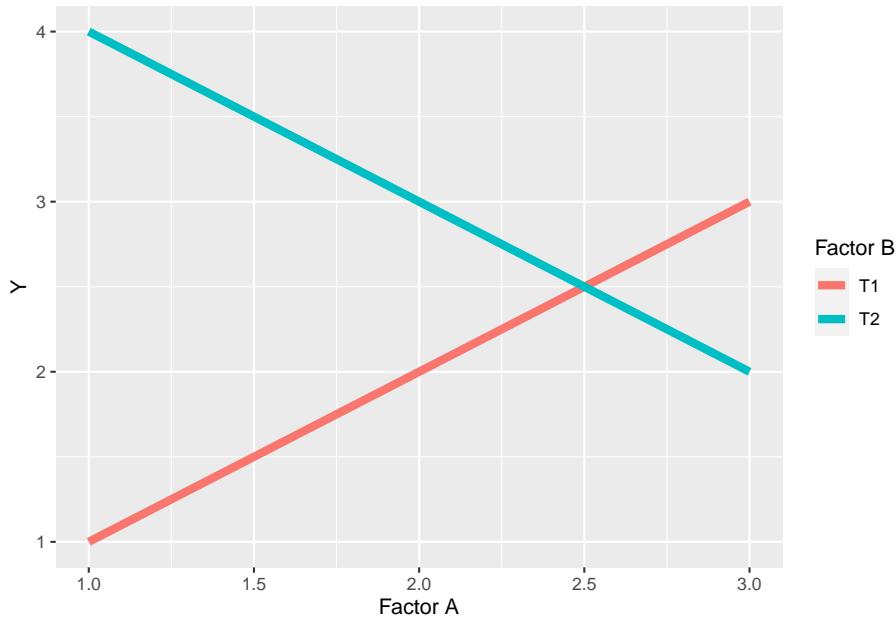


In the above plot, we can see the levels of Factor A still rank $1 < 2 < 3$ in their effect on the observed value Y, for both level T1 and level T2 of Factor B. We also note that the levels of Factor B rank $T1 < T2$ at each level of Factor A. In this spreading interaction, however, the difference between T1 and T2 of factor B increases as the levels of Factor A increase. Similarly, the differences among levels of Factor A are greater for level T2 than level T1 of Factor B.

We saw a spreading interaction before in Case Study 1. The effect of sulfur increased with the level of phosphorus, and vice versa.

7.2.3.3 Crossover Interaction

A crossover interaction is similar to a spreading interaction in that the differences between levels within one factor change with the levels of a second factor, but different in that the ranking of levels changes as well. In addition, as we saw above, crossover reactions can mask the effects of factor levels.



In the plot above, the ranking of levels within Factor B is $T2 > T1$ for levels 1 and 2 of Factor A, but $T1 > T2$ for level 3 of Factor A. In other words, whether T2 is greater than T1 depends on the level of Factor A. In addition, the levels of Factor B behave differently in response to the levels of Factor A. Level T1 of Factor B increases with the level of Factor B, while level T2 decreases.

We observed a crossover reaction in Case Study 2 above, where the effect of the widest row spacing on yield was greater than the narrow row spacings at the lowest planting density, but was less than the narrow spacings at the greatest planting density.

7.2.3.4 Discussing Interactions

Interactions are exciting. Spreading interactions show us how the proper combination of management practices or inputs can have a combined effect that is greater than the individual effect of inputs. Conversely, crossover interactions show us how the advantage of one practice can be lost with the mismanagement of a second practice. This is the essence of agronomy.

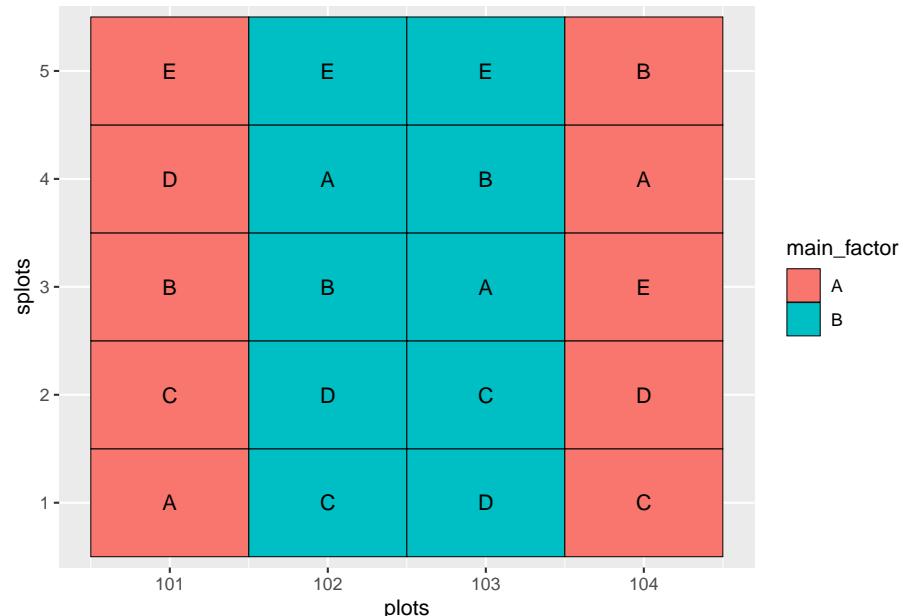
When we identify spreading interactions, we learn how to build more productive *cropping systems*, as opposed to one-off solutions. Don't get me wrong – there are some wonderful one-off solutions. But almost every input or practice can be made more effective by understanding how it interacts with other practices. In addition, trials can be designed to test how the effect of that input interacts with different environmental conditions, such as temperature and precipitation.

Interactions should always be highlighted when discussing the findings of an Analysis of Variance. The essence of an interaction is that the effect of one factor is *dependent* on the level of a second factor. If you don't discuss the interaction between factors, you don't completely describe the effect of either factor.

As we saw above, crossover interactions can mask the effects of factors. In the second case study, had we simply concluded from the main effect that yield did not differ with plant density, we would have failed to report what the data actually show: yield differs profoundly with plant density – but that row spacing affects the direction of that difference.

7.3 Split-Plot Design

The *Split-Plot Design* is a special kind of factorial experiment, in which one factor is nested within another. Let's jump to an illustration.



In the experiment shown above, there are two factors: the main factor and the sub-factor. The main factor has two levels, while the sub-factor has 5 levels. The design is a randomized complete block design. As with any randomized complete block design, all treatments must occur once within each block.

Main factor levels are indicated by color. Sub-factor levels are indicated by letter.

In the above plot map, the main factors occur once in Block 1 (plots 101 and 102) and Block 2 (plots 103 and 104). Within each occurrence of the main factor, all five levels of the sub-factor are nested. The levels of the main factor are randomized within each block, and the levels of the subfactor are randomized within each main plot.

So now that we know what a split-plot design looks like, we will address the question: why would we want to do this? The first answer has to do with practicality. Say the first factor is composed of fungicide treatments that we are going to apply with a Hagge applicator with a 100 ft boom. The second factor is composed of treatments (hybrid, in-furrow fertilizer, etc) that can be applied in 20-foot widths. We can apply our treatments much more easily if we use a split-plot design and nest the second factor (the sub-factor) within the first factor (the main factor).

The second answer is that a thoughtfully-designed split-plot experiment will be more sensitive to differences among levels of the sub-factor than a randomized complete block trial. This is because, no matter how hard we try, plots that are closer together are more likely to be similar than plots that are further apart. By putting the sub-factor treatments close together, we can better estimate and test treatment effects.

The greater sensitivity to subfactor treatments, however, comes at a cost: we sacrifice some of our ability to detect differences among levels of our main factor. Sometimes, however, the levels of the main plot factor are so markedly different that we know we will detect differences among them, even if they are placed further apart. Plus, we may be more interested in the interaction between the main factor and sub-factor, and our ability to estimate and test this interaction, too, is enhanced by the split plot design.

####Case Study: Habenero Peppers If you are a chili pepper fan like me, you enjoy habanero chilies, also called “Scotch bonnets” in strict moderation. They are hotter than the jalapeno, but not devastating like the Carolina Reaper. In this study, habanero peppers were grown in pots of soils characterized by their color. They were harvested at two stages of ripening, green and orange, and their polyphenol concentrations measured. Soil was the main factor and harvest time the subfactor.

block	soil	harvest_stage	total_polyphenols
R1	red	green	109.84958
R1	red	orange	186.71777
R1	brown	green	130.53991
R1	brown	orange	207.00041
R1	black	green	97.86705
R1	black	orange	212.19618

7.4 Linear Additive Model

$$Y_{ijk} = \mu + M_j + \text{Error}(B_i + BM_{ij}) + S_k + MS_{ik} + \text{Error}(BS_{ik} + BMS_{ijk})$$

Bear with me. We can break this model down in to two sections. Let's start with the first three terms. These focus on explaining the observed variance among the main plots.

B_i is the effect of block i

M_j is the effect of level j of the main factor

$\text{Error}(BM_{ij})$ is the error (unexplained variance) associated with the block i and level j of the main factor. When we calculate the F-value for the main factor, we use this error term.

The second three terms focus on explaining the observed variance among subplots.

S_k is the effect of level k of the sub-factor

MS_{ik} is the interaction between level i of the main factor and level k of the subfactor

$\text{Error}(BS_{ik} + BMS_{ijk})$ is the unexplained variance associated with the given levels of the sub factor and the main factor - sub-factor interaction. It is used in testing the significance of both those effects

Ugh. What a dumpster fire. Let's look at the ANOVA table for our habenero trial to make more sense of this.

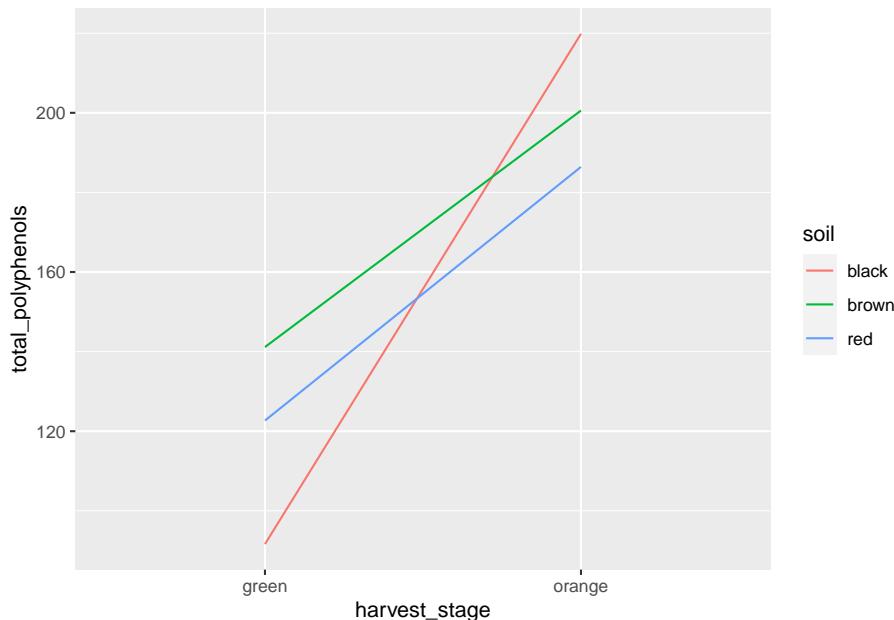
```
## 
## Error: block:soil
##           Df Sum Sq Mean Sq F value Pr(>F)
## soil          2 1322.7   661.3   8.691 0.00791 ***
## Residuals    9   684.9     76.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## harvest_stage     1  42120   42120  176.72 3.2e-07 ***
## soil:harvest_stage 2   5938    2969   12.46 0.00256 **
## Residuals         9   2145     238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have two ANOVA tables. The top table evaluates the effect of soil in our main plots. Looking at the degrees of freedom, we can see we have 3 soils - 1

$= 2$ degrees of freedom for soil and $(4 \text{ blocks} - 1) \times 3 = 9$ degrees of freedom for the error. The F-value is the ratio of the mean square for soil and the mean square for the main factor error (residuals).

The second table evaluates the effect of the harvest stage and the interaction between soil and harvest stage in the subplots. There are $2 \text{ stages} - 1 = 1$ degree of freedom for harvest stage and $*(3 \text{ soils} - 1) \times (2 \text{ stages} - 1) = 2$ degrees of freedom for the interaction. The F-values for the harvest stage and interaction effects are the ratios of their mean squares to the mean square for the subplot error (residuals).

We analyze the ANOVA results the same as we would for any factorial trial. Looking at our table, we see significant results for soil, harvest stage, and their interaction. So lets go straight to the interaction plot.



We can see the total polyphenol concentration was affected by harvest stage in each soil. Orange habaneros had greater total polyphenols than green ones. Furthermore, we can see that black soils produced the fewest total polyphenols when habaneros were harvested green, but the most when they were harvested orange.

Looking at our interaction tables, we see soil has a significant effect on total polyphenols at both harvest stages.

harvest_stage	term	df	sumsq	meansq	statistic	p.value
green	block	3	312.1119	104.03731	0.4324894	0.7375426
green	soil	2	5006.9642	2503.48208	10.4071271	0.0112036
green	Residuals	6	1443.3275	240.55458	NA	NA
orange	block	3	79.2278	26.40927	0.1592069	0.9199513
orange	soil	2	2253.3551	1126.67757	6.7921174	0.0287564
orange	Residuals	6	995.2810	165.88017	NA	NA

And that harvest stage had a significant effect at each level of soil.

soil	term	df	sumsq	meansq	statistic	p.value
black	block	3	79.03859	26.34620	0.0555253	0.9798439
black	harvest_stage	1	32872.72483	32872.72483	69.2801410	0.0036344
black	Residuals	3	1423.46960	474.48987	NA	NA
brown	block	3	214.87947	71.62649	0.5937286	0.6605111
brown	harvest_stage	1	7062.17490	7062.17490	58.5400106	0.0046367
brown	Residuals	3	361.91529	120.63843	NA	NA
red	block	3	390.95481	130.31827	1.0869202	0.4734925
red	harvest_stage	1	8122.82365	8122.82365	67.7484520	0.0037542
red	Residuals	3	359.69045	119.89682	NA	NA

7.5 Conclusion

Experimental design of multiple treatment experiments plays a critical role in the quality of our data and the inferences that can be made. In most cases, you will probably run randomized complete block design experiments with a single factor, where blocking will reduce the error among plots within each block.

The factorial design allows greater efficiency if you are interested in studying two factors. It is also the only way to study the interaction between two factors.

Finally, where one factor is known to be more subtle (and therefore, harder to test significance for) than a second factor, or where differences in equipment size make it more practical to nest one factor inside the other, the split-plot design can be very useful.

Chapter 8

Means Separation and Data Presentation

The previous two units focused on the design and analysis of effects in multiple treatment files. Our focus was to determine whether treatment effects explained more of the variation among individuals in a population than error (or residual) effects, which are based on unexplained differences among individuals.

In the first half of this unit, we will learn three common tools used for testing the differences *between* treatments. This is often the key purpose of a research trial. We know going in that some treatments will be different. But we don't know how they will rank, and whether the difference between them will be great enough to infer one is better than another.

In the second half, we will learn how to present treatment means in tables and plots. Proper data allows the reader to not only grasp results, but even incorporate some of your findings into their own work.

8.1 Case Study

Our sample dataset is inspired by Salas, M.C.; Montero, J.L.; Diaz, J.G.; Berti, F.; Quintero, M.F.; Guzmán, M.; Orsini, F. Defining Optimal Strength of the Nutrient Solution for Soilless Cultivation of Saffron in the Mediterranean. *Agronomy* 2020, 10, 1311.

Saffron is a spice made from the anthers of the saffron flower. It has a nuanced, sweet, complex flavor, and is used in dishes from Cornish saffron rolls to Spanish paella to Iranian tahdig. It comes from the anthers of the field-grown saffron flower and must be hand picked, making it very expensive.

In this study, saffron was grown hydroponically in 15-L pots filled with perlite, with four nutrient concentrations as defined by electroconductivity (EC): low (EC 2.0), medium (EC 2.5), high (EC 3.0), and very high (EC 4.0). The effect of the solutions on corm production (needed to propagate the crop) was measured.

plot	block	rate	corm_number
1	1	1X	246
2	1	4X	240
3	1	3X	254
4	1	2X	285
5	2	4X	233
6	2	3X	251

8.2 Least Significant Difference

Perhaps the most straightforward method of means separation is the infamous Least Significant Difference test. Not to be confused with the psychadelic parties in Marin County, California, the LSD test is as simple as this:

- calculate the least significant difference
- any treatment differences that are equal to or greater than the least significant difference are – you guessed it – significant

The least significant difference is calculated as follows:

$$LSD_{df,\alpha} = t_{df,\alpha} \cdot SED$$

Where $LSD_{df,\alpha}$ is the least significant difference, given the degrees of freedom associated with the error effect and the desired signifance.

Does this formula look vaguely familiar? Does it remind you of what you were doing back in Unit 4? Great, because this is the same formula we used to calculate the distance between confidence limits and the sample mean back when we learned t-tests. Back then, we saw how the confidence interval was used to test the probability our observed difference between treatments was different from zero. Recall if zero fell outside our confidence interval, we inferred the two treatments were different. Similarly, if the difference between two treatments is greater than the least significant difference, we infer the treatments are significantly different.

In R, we use a function, *LSD.test()*, which is part of the *agricolae* package, to calculate the LSD. First, however, lets run an analysis of variance on our t data. The experiment was a randomized complete block design, so our linear additive model is:

$$Y_{ij} = \mu + B_i + T_j + BT_{ij}$$

Where Y_{ij} is the number of corms, μ is the overall population mean for the trial, B_i is the block effect, T_j is the treatment effect, and BT_{ij} is the block effect.

Our analysis of variance result is below. The effect of fertilization rate is highly significant. And this brings us to an important rule for using the LSD test. We only use the LSD test to separate means *if* the treatment effect is significant in our analysis of variance. Doing otherwise can lead to errors, as we will discuss below.

term	df	sumsq	meansq	statistic	p.value
block	3	85.25	28.41667	2.46506	0.1288285
rate	3	4472.75	1490.91667	129.33253	0.0000001
Residuals	9	103.75	11.52778	NA	NA

8.3 LSD Output in R

Now that we know the effect of fertilization rate is highly significant, we want to know how the individual treatments rank, and whether they are significantly different from one another. The results of our LSD test are below.

```
## $statistics
##      MSerror Df     Mean      CV t.value      LSD
##      11.52778  9 257.125 1.32047 2.262157 5.43101
##
## $parameters
##          test p.adjusted name.t ntr alpha
## Fisher-LSD    none     rate   4  0.05
##
## $means
##      corm_number      std r      LCL      UCL Min Max      Q25      Q50      Q75
## 1X      249.25 4.272002 4 245.4097 253.0903 246 255 246.00 248.0 251.25
## 2X      284.75 2.629956 4 280.9097 288.5903 281 287 284.00 285.5 286.25
## 3X      254.25 2.872281 4 250.4097 258.0903 251 258 253.25 254.0 255.00
## 4X      240.25 5.439056 4 236.4097 244.0903 233 246 238.25 241.0 243.00
##
## $comparison
## NULL
##
## $groups
##      corm_number groups
## 2X      284.75      a
```

```

## 3X      254.25    b
## 1X      249.25    b
## 4X      240.25    c
##
## attr(,"class")
## [1] "group"

```

Lets unpack this piece by peace. The output from the LSD test is in a list of tables.

8.3.1 Statistics Table

	MSerror	Df	Mean	CV	t.value	LSD
	11.52778	9	257.125	1.32047	2.262157	5.43101

Let's start with the `$statistics` table. This explains how our LSD was calculated:

- MSerror is the error or residual mean square. It should match that value from the ANOVA table above. Recall that MSerror is an estimate the variance within treatments – that is, the variation among plots unexplained by our model. Therefore, its square root is the standard deviation of the observations within each treatment.
- DF is the degrees of freedom, which is used to calculate our t.value
- Mean is just that – the overall mean of the trial.
- CV is the coefficient of variance. By dividing the standard deviation by the mean, and multiplying by 100, we arrive at this value. Recall from Unit 6 that the CV is a measure of the quality control of the trial: how consistent were our experimental units?
- The t-value is based on the degrees of freedom and α , the desired p-value (often 0.05) to be used to to test significance.
- LSD is the product of the t-value and the standard error of the difference, which can be derived from MSerror and the number of replicates.

This is a lot to absorb, I realize. The most important two statistics for you to understand from this table are the CV and LSD. The other numbers are intermediate values, although if you list the LSD in a report you should also report the degrees of freedom used to calculate the t-value.

8.3.2 Means Table

	corm_number	std	r	LCL	UCL	Min	Max	Q25	Q50	Q75
1X	249.25	4.272002	4	245.4097	253.0903	246	255	246.00	248.0	251.25
2X	284.75	2.629956	4	280.9097	288.5903	281	287	284.00	285.5	286.25
3X	254.25	2.872281	4	250.4097	258.0903	251	258	253.25	254.0	255.00
4X	240.25	5.439056	4	236.4097	244.0903	233	246	238.25	241.0	243.00

The `$means` table explains the distribution of observations within a treatment level around their sample mean. Most of these concepts we discussed in Unit 4. The statistics are:

- `corm_number`: These are our sample means for each level of treatment
- `std`: the standard error of the sample mean. This is unique to the sample mean for each treatment.
- `r`: the number of replicates per treatment level
- `LCL` and `UCL`: the lower confidence limit and upper confidence limit for each mean. These are calculated just as we did in Unit 4.

The remainder of the statistics show the minimum and maximum values, and the quartiles.

8.3.3 Groups Table

	corm_number	groups
2X	284.75	a
3X	254.25	b
1X	249.25	b
4X	240.25	c

Often, the `$groups` table is the most interesting, for it tests the differences among levels of a treatment. The treatment means among levels are ranked from greatest to least.

- `corm_number`: again, the sample mean for each level of treatment
- `group`: this groups treatments that are statistically equivalent. Any means followed by the same letter are considered equal. In the table above, the mean corm numbers associated with the 1X and 3X rates of fertilizer are considered equal. Any means not followed by the same letter are considered statistically different at the p-level chosen to test those differences. The 2X rate produced significantly greater corm numbers than the other fertilizer rates. The 4X rate produced statistically lesser corm numbers than the other fertilizer rates.

8.4 Comparisonwise versus Experimentwise Error

It is important LSD tests not be used indiscriminately to separate means. The problem is that each time we test the difference between two means, we have a 5% chance of committing a type I error. This is the probability of committing a Type I error in *that* comparison. This is known as the *comparisonwise error rate*.

The probability of committing a type I error across all the comparisons is known as the *experimentwise error rate*. If we only conduct one comparison, our experimentwise error rate is equal to the comparisonwise error rate. But if we conduct two comparisons, our experimentwise error rate is equal to the probability that comparisonwise errors will not occur in both comparison.

We can demonstrate this with some simple calculations. Before we start our comparisons, there is a 100% chance we have not committed an experimentwise Type I error. We cold express this as:

$$\text{Experimentwise Error} = 1$$

If we once comparison, there is 5% probability of a comparisonwise Type I error – and a 95% chance the error will not occur. We can express this as

$$\text{Experimentwise Error} = 1 - 0.95 = 0.05$$

If we have two comparisons, there is a 95% probability a comparisonwise Type I error won't occur in the first comparison – and a 95% probability it won't occur in the second comparison. But the probability it doesn't occur in both comparisons is $0.95 * 0.95$:

$$\text{Experimentwise Error} = 1 - 0.95 \times 0.95 = 1 - 0.9025 = 0.0975$$

Now the Experimentwise error rate is 0.0975, or 9.75%.

What about three comparisons?

$$\text{Experimentwise Error} = 1 - 0.95 \times 0.95 \times 0.95 = 1 - 0.8573 = 0.1427$$

The Experimentwise error rate is now 0.1427, or 14.27%.

Finally, what if we had 10 comparisons?

$$\text{Experimentwise Error} = 1 - 0.95^{10} = 1 - 0.5987 = 0.4013$$

Our experimentwise error rate is now about 0.40, or 40%.

As the number of our comparisons increases, so does the probability of an experimentwise error. How can we avoid this? The first method, mentioned above, is to not use the LSD test unless the ANOVA shows a significant treatment effect. We call this approach the *F-protected LSD test*.

The second approach is to use a multiple range test that increases its minimum significant difference for comparing treatments as the number of treatments increases.

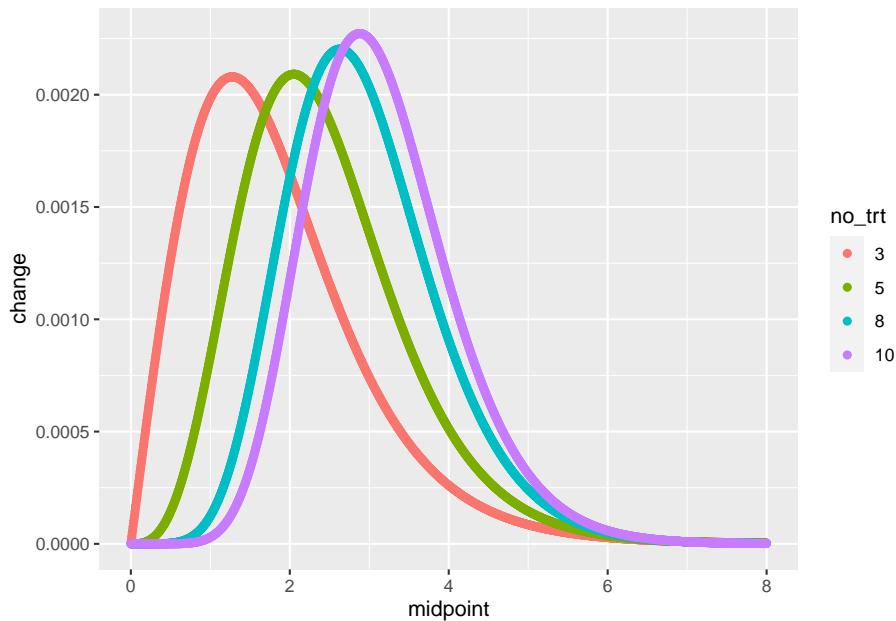
8.5 Tukey's Honest Significant Difference

If we are going to be comparing many treatments, it is better to use a minimum significant difference, like Tukey's Honest Significant Different (HSD) Test. Tukey's HSD is calculated very similarly to the least significant difference:

$$\text{Tukey's HSD} = Q_{\alpha, df, k} \cdot SED$$

The difference in Tukey's HSD is that we use Q , the “studentized range distribution” (you can just call it Q in this course) in place of the t-distribution. Q differs from t in that its value is determined not only by α , the desired probability of a Type I error, and df , the degrees of freedom associated with the error mean square, but also k , the number of treatments.

The plot below includes four “ Q ” distribution curves, associated with 3, 5, 6, and 10 treatments. Observe how the distributions shift to the right as the number of treatments increases. This means that the minimum difference for significance also increases with the number of treatments.



The Tukey test output below is very similar to the LSD test output. The “\$statistics” section is identical to that of the LSD output, except it now reports the minimum significant difference instead of the least significant difference. In the “\$parameters” section, the “Studentized Range” value (Q) is given in place of the t-value. The “\$groups” section can be interpreted the same for the minimum significant difference as for the least significant difference.

```
## $statistics
##   MSerror Df      Mean       CV      MSD
##   11.52778  9 257.125 1.32047 7.494846
##
## $parameters
##   test name.t ntr StudentizedRange alpha
##   Tukey    rate  4        4.41489  0.05
##
## $means
##   corm_number      std r Min Max     Q25     Q50     Q75
##   1X          249.25 4.272002 4 246 255 246.00 248.0 251.25
##   2X          284.75 2.629956 4 281 287 284.00 285.5 286.25
##   3X          254.25 2.872281 4 251 258 253.25 254.0 255.00
##   4X          240.25 5.439056 4 233 246 238.25 241.0 243.00
##
## $comparison
## NULL
##
```

```

## $groups
##   corm_number groups
## 2X      284.75    a
## 3X      254.25    b
## 1X      249.25    b
## 4X      240.25    c
##
## attr(,"class")
## [1] "group"

```

Unlike the LSD test, the Tukey test does not need to be “protected” by first examining whether the Analysis of Variance treatment effect is significant. That said, the Tukey test is unlikely to indicate a significant difference between treatments without the ANOVA treatment effect being significant as well.

8.6 Linear Contrast

The last type of means comparison we will learn in this lesson is the linear contrast. Unlike the LSD and Tukey tests, linear contrasts may be used to separate two groups of treatments. While a lot of math may be introduced to the curious, in a linear contrast the statistician defines two groups of treatments through the use of *coefficients*; R then calculates their means and standard errors, and compares them using a t-test. There are multiple ways we can use a linear contrast, and our saffron dataset is a great way to introduce them.

8.6.1 Coefficients

Recall how a t-test between two treatments works. We calculate t as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SED}$$

Where $\bar{x}_1 - \bar{x}_2$ is the difference between sample means, $(\mu_1 - \mu_2)$ is the hypothesized difference between treatments (usually zero), and SED is the standard error of the difference.

For most situations, we could simplify the above equation to:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SED}$$

What about if we are comparing more than two treatments? For our saffron example, what if we wanted to calculate the difference between the mean of the

two lower rates (1X and 2X) and the mean of the two higher rates (3X and 4X)? Lets call this difference L . We would calculate this difference as:

$$L = \frac{\bar{x}_{1X} + \bar{x}_{2X}}{2} - \frac{\bar{x}_{3X} + \bar{x}_{4X}}{2}$$

All we are doing above is 1) calculating the means for the two groups and 2) subtracting the mean of the 3X and 4X rates from the mean of the 1X and 2X rates. Now let's express this same formula a little differently. What we are doing in the above equation is multiplying each number by 1/2:

$$L = \frac{1}{2}(\bar{x}_{1X} + \bar{x}_{2X}) - \frac{1}{2}(\bar{x}_{3X} + \bar{x}_{4X})$$

In addition (bear with me!), when we subtract the mean of treatments 3X and 4X, it is the equivalent of adding the negative value of their mean to the mean of treatments 1X and 2X:

$$L = \frac{1}{2}(\bar{x}_{1X} + \bar{x}_{2X}) + (-\frac{1}{2}(\bar{x}_{3X} + \bar{x}_{4X}))$$

Finally, we can arrange the equation above as:

$$L = \frac{1}{2}\bar{x}_{1X} + \frac{1}{2}\bar{x}_{2X} - \frac{1}{2}\bar{x}_{3X} - \frac{1}{2}\bar{x}_{4X} +$$

The reason for this tortuous algebra flashback is to show you where the contrast coefficients come from. Each of the $\frac{1}{2}$ s in the equation above is a contrast coefficient.

Let's demonstrate this with our saffron data. Our saffron treatment means are:

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25

Now let's add in a column with our coefficients:

rate	corm_number	coefficient
1X	249.25	0.5
2X	284.75	0.5
3X	254.25	-0.5
4X	240.25	-0.5

We see that R has converted these to decimals. We then create a new column, “mean_x_coefficient”, that is the product of the original mean and the coefficient. We see these products are approximately half the value of the original sample mean (some may be less than half because of rounding errors).

rate	corm_number	coefficient	mean_x_coefficient
1X	249.25	0.5	124.625
2X	284.75	0.5	142.375
3X	254.25	-0.5	-127.125
4X	240.25	-0.5	-120.125

Finally, we can sum the mean_x_coefficient column to get the total difference among treatments.

```
## [1] "total difference = 19.75"
```

One critical rule about requirements is their sum must always equal zero. Otherwise you are not completely identifying two groups to compare. Using coefficients that do not sum to zero can also suggest you are weighting one group unfairly compared to another.

###Contrast Calculations

To calculate t , of course, we must divide L by the standard error of the difference:

$$t = \frac{L}{SED}$$

So how is the standard error of the difference calculated?

Well, we know that the error mean square from our ANOVA is equal to the mean variance within treatments. And we know that if we take the square root of a variance, we get the standard deviation. And if we divide the standard deviation by the number of observations in each sample mean, we get the standard error.

In a paired t-test, where we are simply calculating a standard error for one set of numbers, the differences between each pair, the standard error of the difference is calculated the same as the standard error:

$$SED = SE = \frac{s}{\sqrt{n}}$$

Where s is the standard deviation and n is the number of observations (reps) per treatment. This formula is equal to that below, where we use the variance, s^2 , in place of the standard deviation.

$$SED = \sqrt{\frac{s^2}{n}}$$

When we work with two-treatment trials where the treatments are not paired or blocked, we account for the variance and number replicates individually. For treatment levels “1” and “2”, the standard error of the difference becomes

$$SED = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In most trials, however, we assume that the variances and number of replications are equal among treatments. So we can also express the above trial as:

$$SED = \sqrt{\frac{2 \cdot s^2}{n}}$$

Recall that in the analysis of variance for a multiple treatment trial, the error mean squares *is* the mean variance within treatments. So the equation above is equivalent to:

$$SED = \sqrt{\frac{2 \cdot EMS}{n}}$$

In a linear contrast, however, our standard error of the difference must be scaled according to the coefficients we use, since we are no longer comparing two treatments, but multiple. So our equation becomes:

$$SED = c \cdot \sqrt{\frac{EMS}{n}}$$

Where c is the square root of the sum of the squared constants is the For our example above, this sum, c , would be:

$$c = s \sqrt{\sum \left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 + \left(-\frac{1}{2} \right)^2 + \left(-\frac{1}{2} \right)^2}$$

Which is equal to:

$$c = \sqrt{\frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = \sqrt{1} = 1$$

We can now calculate the standard error of the difference for our contrast. First, lets go back to our analysis of variance.

term	df	sumsq	meansq	statistic	p.value
rate	3	4472.75	1490.917	94.66138	0
Residuals	12	189.00	15.750	NA	NA

We can see from the ANOVA output that our error mean square is 15.75. We know from the design itself that we had 4 replicates

So our final standard error of the difference for our contrast is:

$$SED = 1 \cdot \sqrt{\frac{15.75}{4}} = 1 \cdot \sqrt{3.9375} = 1.984$$

The t-value for our test, would then be:

$$t = \frac{L}{SED} = \frac{19.75}{1.984} = 9.954$$

The probability of a t-value of 10, given the 12 degrees of freedom associated with the error sum of squares, would be:

```
## [1] 1.882084e-07
```

8.6.2 Linear Contrasts with R

We can automate linear contrast testing, of course, using R and the *glht()* function of the *multcomp* package. First, we need to define a matrix (table) of contrast coefficients for R. To get the coefficients in the correct order, lets double check the order in which the rate levels are listed in R. It is important to make sure our treatment is classified as a factor. We can do this using the *as.factor()* function

We can then list the order of the levels in treatment rate:

```
levels(saffron$rate)
## [1] "1X" "2X" "3X" "4X"
```

We can see they follow the order 1X, 2X, 3X, 4X. We will therefore form a matrix, K, with the appropriate coefficients.

```
K = matrix(c(1/2, 1/2, -1/2, -1/2), 1)
```

```
K
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  0.5  0.5 -0.5 -0.5
```

We are now ready to run our contrast. First, we need to slightly alter our ANOVA model by adding a zero (0) between the tilde (~) and the treatment name. This is one of of those one-off ideoyncrasies of R.

```
library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##      select

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##      geyser

saffron_model_for_contrast = aov(corm_number ~ 0 + rate, saffron)

low_vs_high = glht(saffron_model_for_contrast, linfct=K)
summary(low_vs_high)

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = corm_number ~ 0 + rate, data = saffron)
##
## Linear Hypotheses:
##             Estimate Std. Error t value Pr(>|t|)    
## 1 == 0     19.750     1.984    9.953 3.77e-07 ***
## ---                                                 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

And there we have it, our contrast in seconds. Now let's interpret this. Estimate is the difference between groups. Since we subtracted the mean of the 3X and 4X rates from the mean of the 1X and 2X rates, the positive estimate value indicates the lower rates produced greater corm numbers than the higher rates.

"Std. Error" is the standard error of the difference, as we calculated above. The t-value is equal to the estimate divided by the standard error of the difference. Finally, "Pr(>|t|)" is the probability of observing the t-value by chance. In this case, it is 3.77×10^{-7} , very close to zero. We conclude the lower two rates, as a group, produce greater corms than the upper two rates as a group.

We can quickly ask other questions of our data. For example, do the middle two rates (2X and 3X) produce a greater corm number than the lowest (1X and highest (4X) rates? Again, let's examine the order of our rate levels.

```
levels(saffron$rate)

## [1] "1X" "2X" "3X" "4X"
```

In order to subtract the mean corm number of rates 1X and 4X from the mean corm number of rates 2X and 3X, we will need to calculate the difference as:

$$L = \left(-\frac{1}{2}\right)\bar{x}_{1X} + \left(\frac{1}{2}\right)\bar{x}_{2X} + \left(\frac{1}{2}\right)\bar{x}_{3X} + \left(-\frac{1}{2}\right)\bar{x}_{4X} +$$

So our contrasts coefficients would be $-\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}$. Our contrast matrix is then:

```
K = matrix(c(-1/2, 1/2, 1/2, -1/2), 1)

K

##      [,1] [,2] [,3] [,4]
## [1,] -0.5  0.5  0.5 -0.5
```

We are now ready to run our contrast. First, we need to slightly alter our ANOVA model by adding a zero (0) between the tilde (~) and the treatment name. This is one of those one-off idiosyncrasies of R.

```
library(multcomp)

saffron_model_for_contrast = aov(corm_number ~ 0 + rate, saffron)

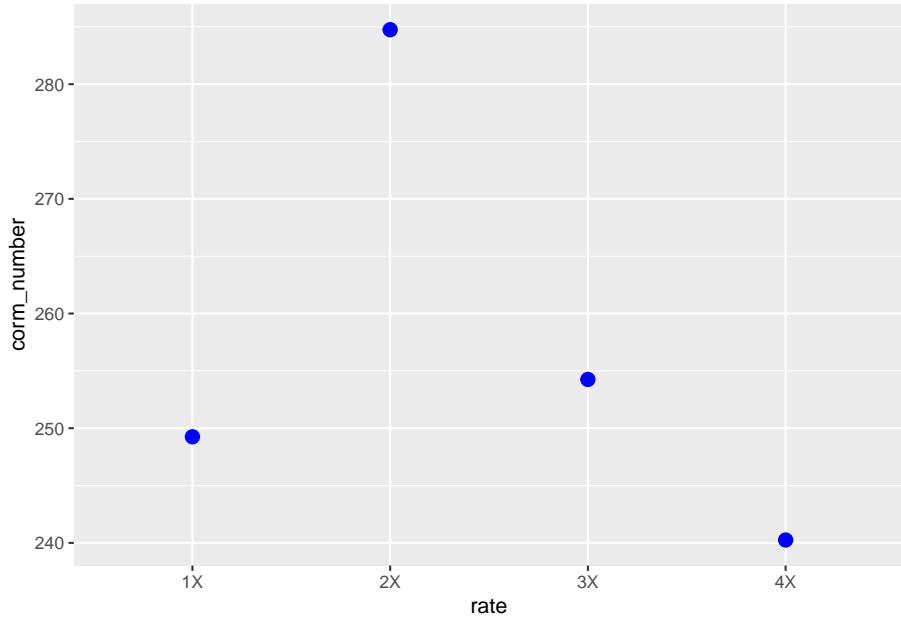
low_high_vs_middle = glht(saffron_model_for_contrast, linfct=K)
summary(low_high_vs_middle)
```

```

## Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = corm_number ~ 0 + rate, data = saffron)
##
## Linear Hypotheses:
##          Estimate Std. Error t value Pr(>|t|)
## 1 == 0    24.750     1.984   12.47 3.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)

```

We see this estimated difference is even greater than the previous contrast. The significant t-value suggests there may be a parabolic (curved) response of corm number to nutrient solution rate. Indeed, if we plot the mean, we can see a parabolic response



8.7 Means Presentation

Something I have always overlooked in teaching this course is also something that is most basic: once we have separated our means, how do we present them to others? There are two ways to present means: in a table, or in a plot. Both have advantages and disadvantages. In a table, you provide raw statistics

(treatment means, standard errors, and perhaps, means groupings from an LSD or Tukey test). A reader can use these values to recalculate your statistics, say, if they wanted to separate the means at $p=0.10$, $p=0.1$, or $p=0.001$.

A figure, one the otherhand, allows the reader to quickly grasp treatment differences, or patterns among treatments. In addition, they are colorful and – lets face it – more inviting than a dense table of numbers.

Whichever format we use, however, we need to present our treatment means and some reference for the reader to gauge whether those means are statistically equal or different.

8.7.1 Means Tables

Tables can be tricky, especially when many types of measures are included. If we are reporting results from a factorial experiment, we may be tempted to list the levels of one factor down rows, and the other factor across columns. I would generally discourage this, however, unless required to fit the table neatly into a publication space. Generally, the long form of means presentation is best.

In the long form of data, cases (individual observations or treatment means) are listed down the rows. Measurements from each cases are listed across rows.

For our saffron data, our table would start like this:

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25

To this table, we may wish to add the standard error of the difference. Remember, the standard error of the difference for an LSD or Tukey test is equal to:

$$SED = \sqrt{\frac{(2 \cdot EMS)}{n}}$$

So for our saffron trial, where the error mean square is 15.75 (we can get this from either the ANOVA table or the LSD output)and the number of replications is 4, the standard error of the difference is:

$$SED = \sqrt{\frac{(2 \cdot 15.75)}{4}} = \sqrt{7.875} = 2.80$$

We can add this below the means

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25
SED	2.8

It is also important to indicate the number of replications of the treatments. We can add another row to the table with N.

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25
SED	2.8
N	4

We should add the LSD, to make it easy for readers to compare treatments. We will want also, to include the α which was used to calculate the LSD. That way, the reader will know whether the risk of a Type I error – that the LSD will separate treatments that are not truly different – is 5% or some other probability.

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25
SED	2.8
N	4
LSD (0.05)	6.11

We might want to include the p-value from the ANOVA table, so the reader knows that the LSD is protected. The p-value for the saffron trial is 1.28×10^{-8} . This is an extremely small number. It is acceptable for us to simplify this in the table, indicating that the probability of F was less than 0.001.

rate	corm_number
1X	249.25
2X	284.75
3X	254.25
4X	240.25
SED	2.8
N	4
LSD (0.05)	6.11
Pr<F	<0.001

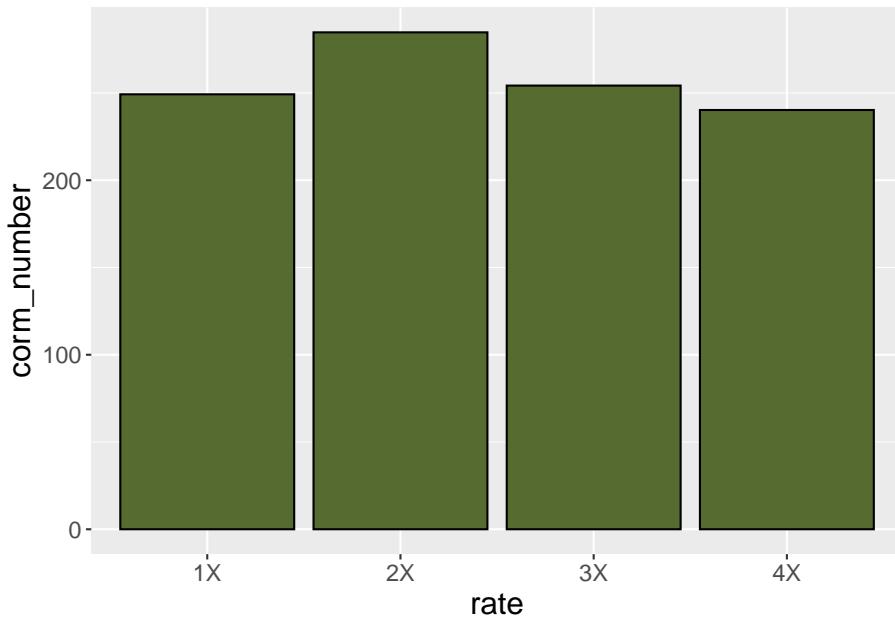
Finally, if imperial (pounds, acres, etc) or metric units were used to measure

the response variable, it is important to indicate that in the table. Indicating those in parentheses after the variable name is appropriate.

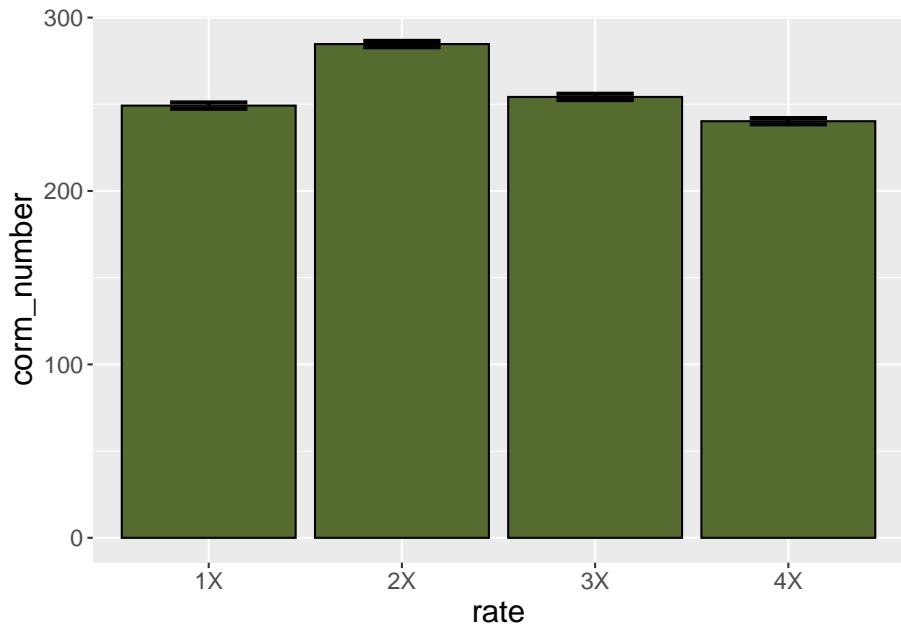
8.7.2 Plotting Means

When we work with categorical treatments (that is, treatments levels that are defined by words), or even numerical variables treated as categorical variables (as in the saffron example) we should use a *bar plot*, not a line plot, to visualize the data. An easy way to determine whether a bar plot should be used is this: if you are using an LSD or Tukey's test to separate your means, you should use a bar plot. A line plot, which suggests treatment levels are numerically related to (higher or lower than) each other should be used to fit regression models, where the analysis defines a continuous relationship between Y and X.

A basic bar plot will have a bar representing each treatment mean. The treatment level is indicated along the x (horizontal) axis. The sample mean is indicated along the y (vertical) axis. The bar height indicates the sample mean.

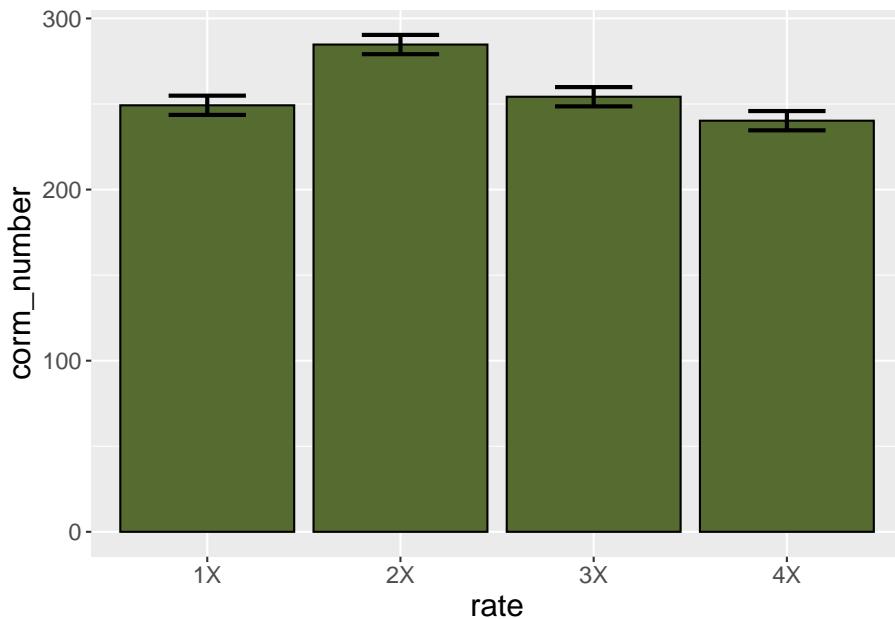


This plot, however, does not provide the viewer any sense of whether corm number is significantly different among treatments. For that purpose, we can add error bars.



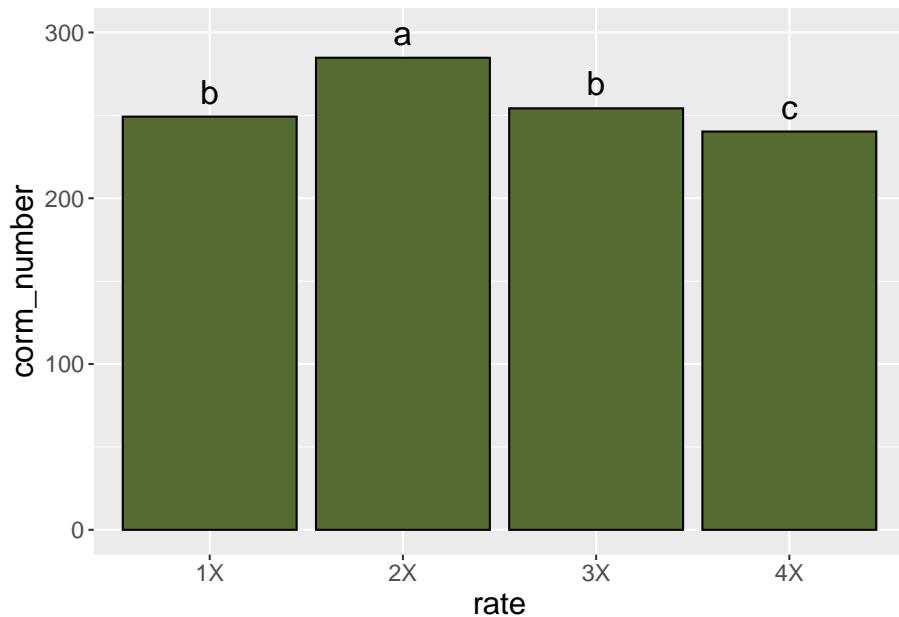
These error bars stretch from one standard error of the mean below the sample mean to one standard error of the mean above the sample mean. As a general rule of thumb, the least significant difference is approximately 2 standard errors of the mean. If the error bars from two treatment levels don't overlap, then the two treatments are likely different. We can see in the plot that the corm number for the 1X and 3X fertilizer rates are very similar.

Alternatively, we could set the error bar height equal to the least significant difference itself.



This allows the viewer to use the actual LSD to separate treatment means. If the range of the error bar from one treatment does not include the mean of another treatment, then the two treatments are not equal.

Significant differences among treatments can also be indicated by including letter groupings from an LSD or Tukey's test.



We will learn more about graphics in the exercises this unit. A very comprehensive resource for creating plots in R is *R Graphics Cookbook*, by Winston Chang. An online copy is available at <https://r-graphics.org/>. Print copies can also be purchased from common booksellers. This book explains not only how to create plots, but how to adjust labels, legends, axis labels, and so on.

Chapter 9

Messy and Missing Data

You have to love the nice, complete datasets I have served up to you in this and other statistics texts. Indeed, some trials will work out this way – in general, the simpler the treatment (and the technology used to apply it) the greater your chances of a complete dataset. Planters will jam, nozzles will plug, and if a trial has 12 treatments a couple of them may end up in the wrong plot. Best to avoid those if possible.

As well, the smaller the experiment, or the more controlled the environment, the greater your odds of a complete dataset. For that reason, decades of university and industry research has been performed in 10- x 40- plots stacked closely together on manicured, table-top-flat ground. If you were a seed-breeder, you had dibs on these fields. If you were an ecologist like me, you might have to head to the back 40 (j/k).

If you can farm in your head or your basement, drop me a note and I will exempt you, with envy from this unit. For those of us who farm by the acre or section, however, the issue of data quality is an agonizing and sometimes subjective topic – but critical. Remember, most of our statistics are models, and the saying goes: “junk in, junk out.” Your models are only as good as the data you use to build them.

Part of your job as a researcher or end-user of any data, before you conduct or read the results from any test, is to ask yourself – are the data reliable enough to support the inferences? A trial with wildly-different experimental units may bias results – if all the replicates of one treatment end up in the better units and others are concentrated in the poorer ones. You may falsely recommend a product if you don’t catch this.

At the other extreme, the failure conduct research on similar experimental units will introduce background variance (or noise) that prevents a statistical test from concluding a difference is not the result of chance, even though the treatments

are, in fact, different. In that case, you may fail to introduce – or adopt – a product or technology with real promise.

In this unit, we will first learn ways to inspect datasets for extreme values which, even given the inherent variability of data, may be suspect. Boxplots, histograms, and probability plots will be our tools for these exercises.

We will then address the uncomfortable question of what to do when we have missing data, either because a plot was compromised during the trial, or because we rejected that plot because its extreme value.

9.1 Inspecting data for Normal Distributions

I know, I know, it is so exciting to have data! The hard physical work of the research is done, the data is painstakingly entered into Excel. Let's run the ANOVAs and regressions now – if we hurry we can still make it to happy hour!

It is so tempting to jump right into deep analyses as the data roll in. But it is important to realize these analyses are based on assumptions:

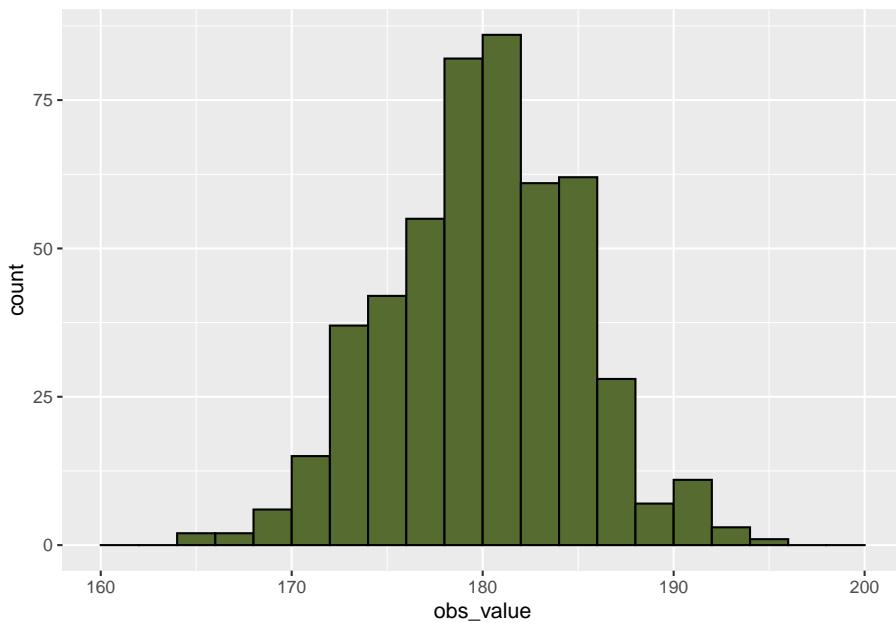
- That the observations for each treatment level are normally distributed around the treatment mean
- That the variance or spread of observations around each level of treatment is roughly equal.
- That experimental units are comparable among treatment levels, so that the treatment and error effects are appropriately separated.

It is very possible, based on the nature of trials, that one or more of these assumptions may be violated. If you have ever counted weeds in a herbicide trial, you have noted that well-treated plots have weed counts that are consistently near zero – but that weedy checks have wildly variable counts (100, 200, 300 weeds). Growth analysis (where plants of different sizes are measured) is also prone to messy data issues, because variance in measurements increases numerically as plants grow. Experimental units (plots) in both commercial and research farm fields can vary because of prior management unknown to the researcher.

9.1.1 Histograms

In the very first unit of this course Unit 2, you were introduced to the histogram. Recall the histogram is a vertical bar chart in which the width of bars defines the different ranges into which observations are grouped, and the height represents the count or proportion of observations falling into each range.

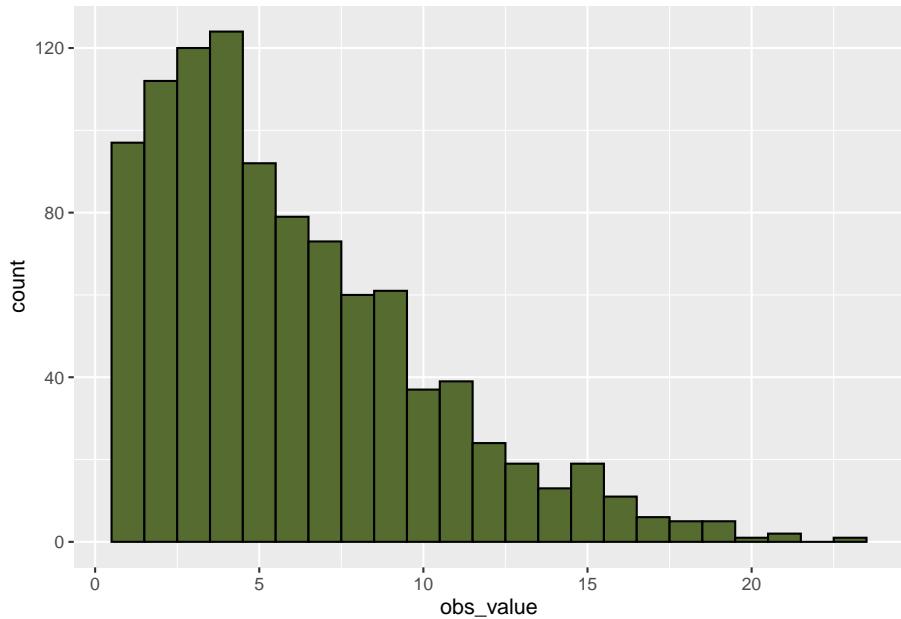
In the data below, we have a dataset with 500 observations of corn yield. The mean is approximately 180. We can see the data distribution is approximately normal.



The summary data are shown below. We can see that the median and mean are both approximately 180 bushels per acre. We can also see the 1st and 3rd quantiles (equal to the 25th and 75th percentiles) are a little over three bushels from the median. The minimum and maximum observations are also similarly spaced from the median.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
164.8697	176.8281	180.1694	180.0362	183.3801	194.2372

When we are dealing with data such as pest counts, our data may be non-normal. Rather than being symmetrical, the data may be skewed to one side or another. For example, in the dataset below, total velvetleaf dry weight in grams per square meter was measured. If you have worked like me with weed populations, you realize weed competitiveness is all about outracing the crop to the sun. If the weed loses, which it will in most cases, it will be small. But the proud few weeds who beat the crop will be huge. That is reflected in the data below.



When we look at the histogram, the data are skewed to the right. The histogram is not symmetrical.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.7386857	2.822542	5.053556	6.010417	8.333231	23.2785

When we look at the summary data, we first notice the mean and median are different. For a dataset this size (1000 observations, we would expect them to be more similar.) We notice that the first quantile is closer to the median than the third quantile. The greatest indication the data is skewed, however, is that the minimum is about 4 plants less than the median, while the maximum is about 18 plants greater.

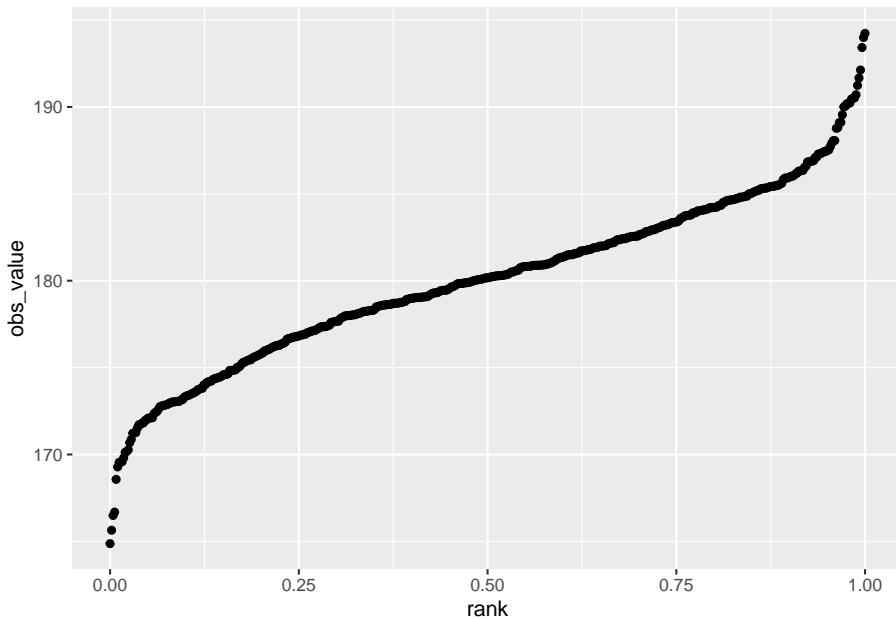
Data like this may be transformed (mathematically re-scaled) so that it is more normal for analyses. We will cover this below.

9.1.2 Rank Percentile Plots

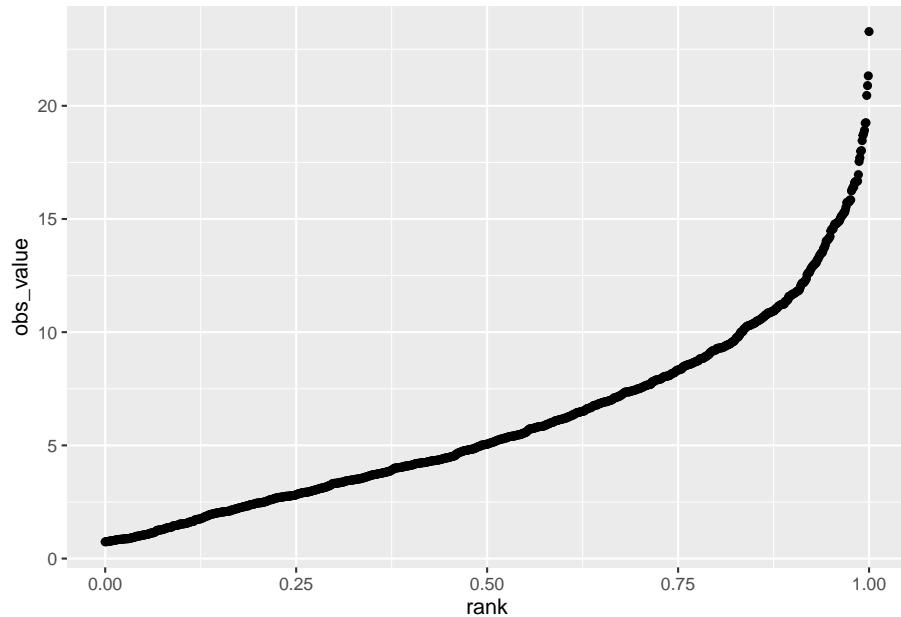
Another way to inspect the normality of datasets is to use a rank percentile plot. This plot uses the percentile rank of each observation, from lowest to highest, as the x-value of each point. The y-value of the point is its observed value.

The data for our normally-distributed corn yield dataset are plotted in the rank percentile plot below. Normally-distributed data tend to be strongly linear in the middle of the plot. If we draw a regression line through the plot, you can see most of the data are close to that line. The lowest percentile points fall below

the line. That means they are a little lower in value than the normal distribution function might predict. The opposite is true of the highest percentile points. This indicates our distribution is a little bit wider than normal, but not enough that we cannot use it for analysis.



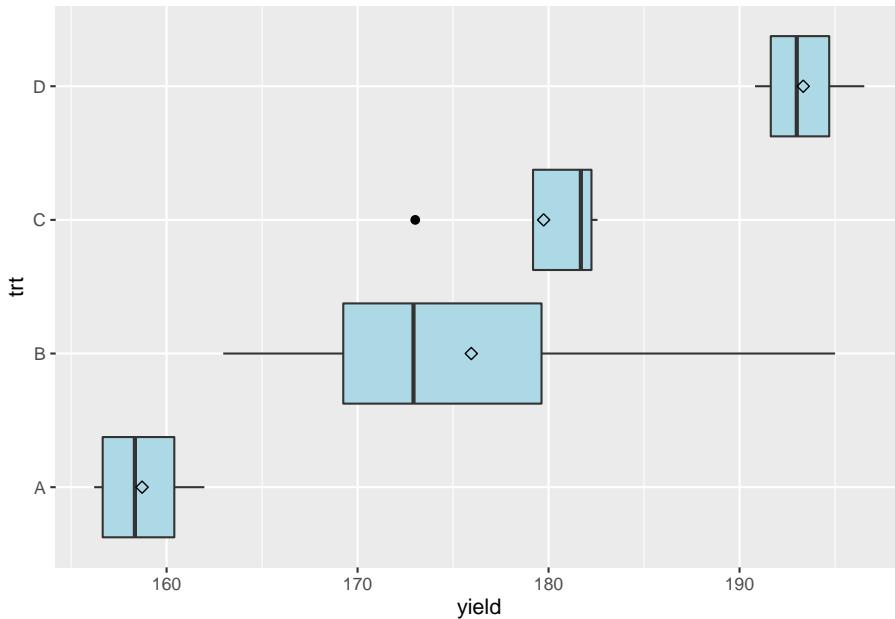
Our skewed data, however, shows up quite differently in the rank percentile plot. We can see that most of the data closely fit a line. But starting around the 75th percentile, the observed values are much greater than the predicted values – almost twice as much. This means the distribution is much wider to the right of the distribution curve than to the left, and that the data are non-normal



9.1.3 Box Plots

The first two methods I have shown you, the histogram and rank percentile plots, are useful if you have a few treatments with a large number of replicates. They are taught in every statistics course and you should know about them. But, in my experience, they are not useful if you have a trial with fewer replications. A normal distribution is not going to appear in a histogram if you only have four replicates – instead you will just see the four individual measurements.

Box plots, on the other hand, are very useful for inspecting multiple treatments. In the plot below, boxplots for four corn treatments are shown. The treatments are labeled A, B, C, and D. The data are plotted so their treatments are listed along the vertical axis, and their values are listed along the y-axis.



The boxplots help us understand the distribution of the data. Lets start with the box, which tells about the spread of the data. The left side of the box is the 25th percentile, the line in the middle is the 50th percentile (or median), and the right side of the box is the 75th percentile. So the box shows us the spread of the middle half of the observations for each treatment.

The diamond shown within the box is the mean. In a normal distribution, the median and mean should be close.

The lines extending from the left and right side of the box are called whiskers. The whiskers extend to the lowest and highest observations for a treatment. The whiskers extend no more than 1.5 times the *inter-quartile range*, which for the lower whisker is the difference between the 25th and 50th percentiles, and for the upper whisker is the difference between the 50th and 75th percentiles.

In treatment B, we can see the upper whisker is missing, and instead there is a point to the right of the bar. If an observation is beyond 1.5 times the interquartile range, the whisker is not shown and the observation is instead represented by a point. This observation is called an *outlier*, meaning that it is outside the range of values expected in a normal distribution. We will talk more about outliers in the next section.

The boxplot tells us something beyond the distribution of the individual treatments. If the boxes are markedly different in their width, the data may have substantially different variances. We should investigate these further using a *mean-variance plot*, and perhaps a statistical *test of heterogeneity*.

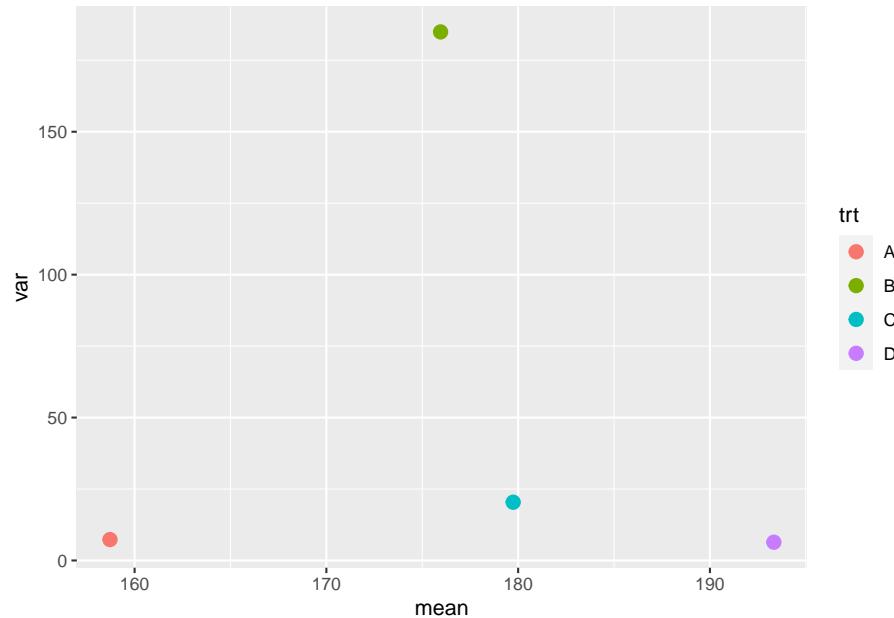
9.2 Inspecting Data for Equal Variances

So far, we have learned to use the t-test and analysis of variance to test named treatments (that is, hybrids, management practices, and other products that can be described by name). These tests generally assume not only that observed values are normally distributed, but that the variances are approximately equal among the different treatments in our experiment. If the variances are unequal, we may calculate least significant differences (LSDs) or honest significant differences (HSDs) that are inappropriate. Among treatments that have smaller variances, our LSD or HSD may be overestimated; among treatments that have larger variances, the LSD or HSD may be underestimated.

The visual inspection of individual treatment distributions in the box plot above, followed by a scatter plot of the treatment variances versus their means, can give us a visual sense of unequal variances. These suspicions can then be tested using a Test for Homogeneity that calculates the probability of differences in variances as greater as those observed.

9.2.1 Mean-Variance Plot

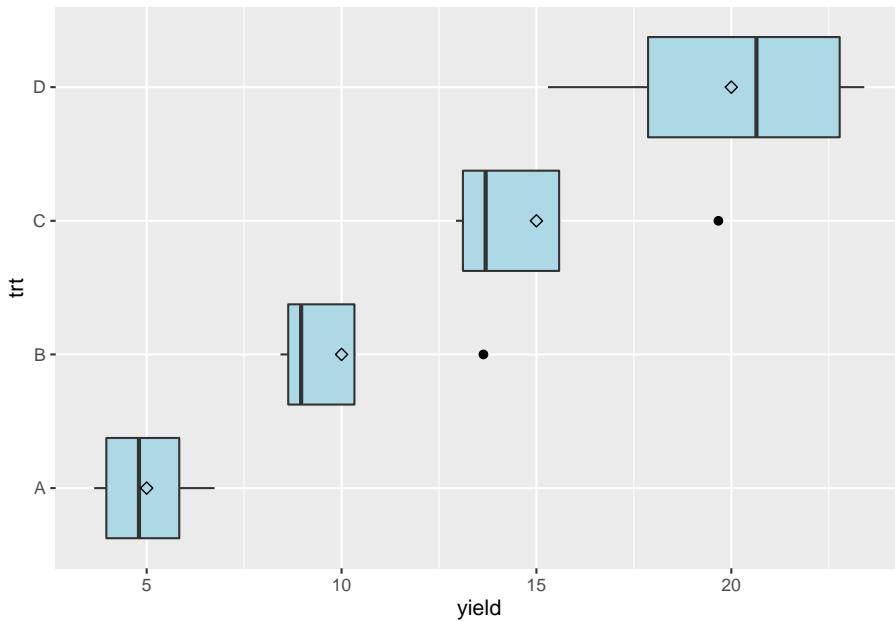
In a mean-variance plot, the treatment means are plotted along the horizontal axis and the variances are plotted along the vertical axis. The plot for the corn yield dataset we have used so far is shown below.



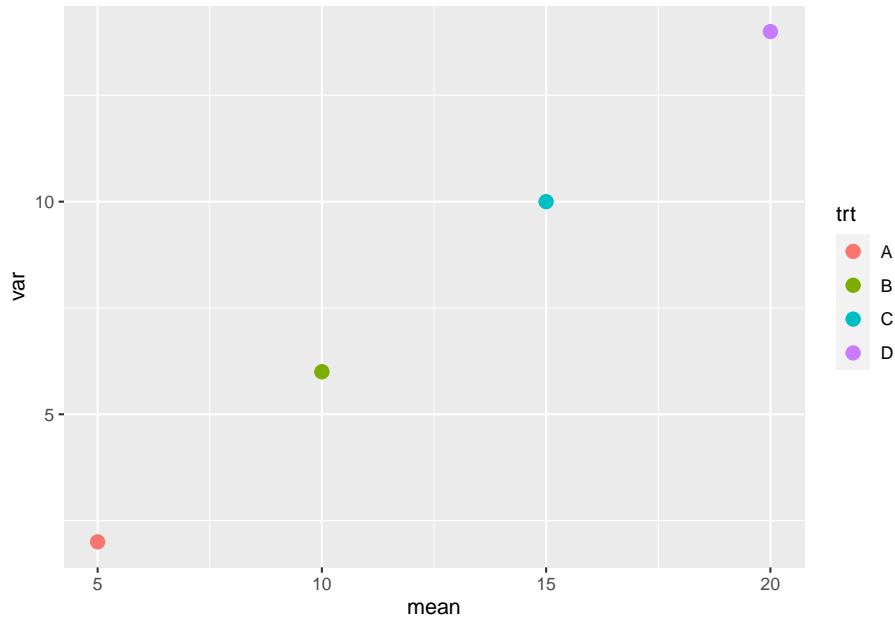
We can see the variance of treatment B is many times greater than that of the

other treatments. In general, we like to see the variances differ by no more than a factor of 2.

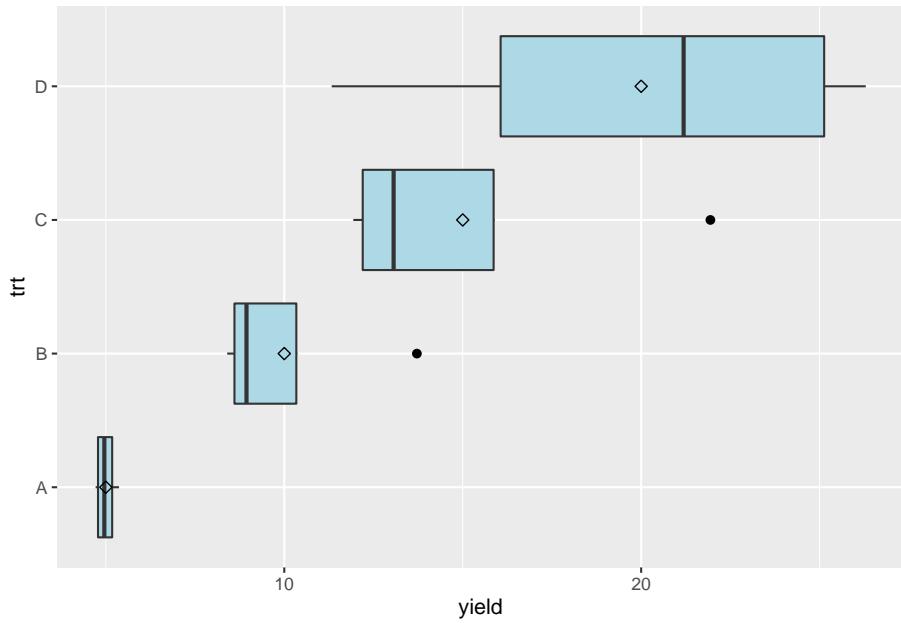
In cases where we are dealing with populations that either “thrive or die” based on environment – particularly pest populations – we may see relationships between the mean and variance. Pest count data is often like this. In our velvetleaf counts, for example, we might find that our greater treatment means are also associated with greater variations in counts between plots.



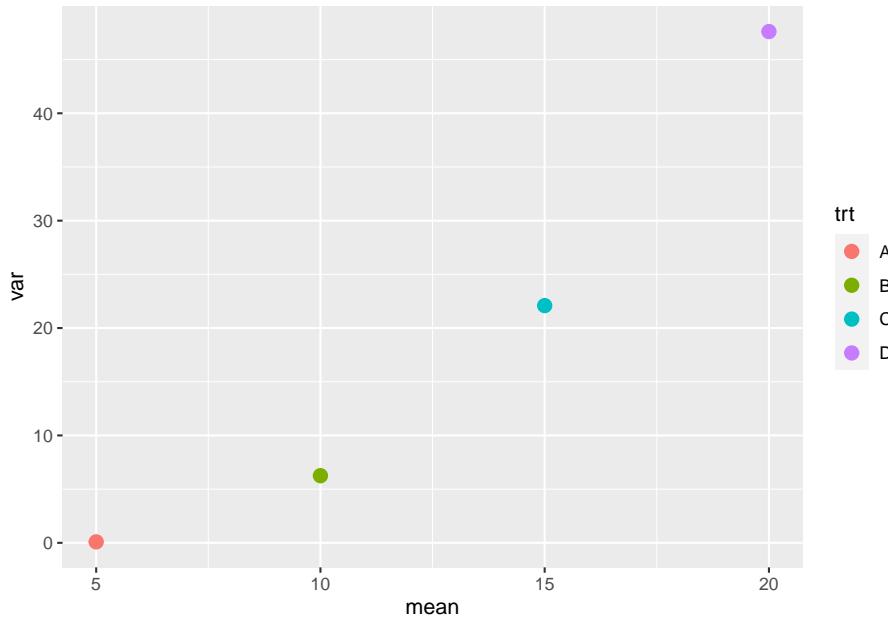
In this case, the mean-variance plot may show a linear relationship between variance and mean.



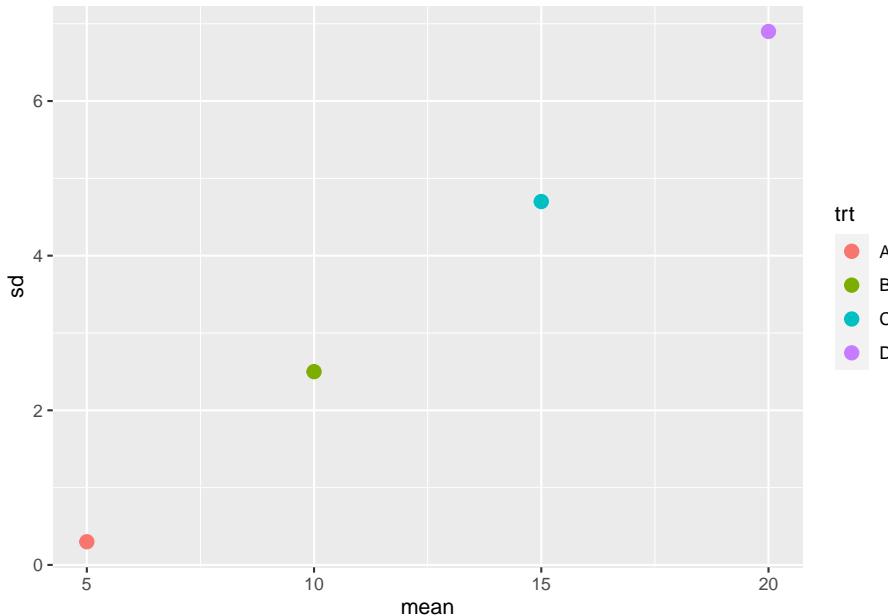
Finally, we may observe a dataset in which the distributions not only increase with means, but seem to do so exponentially.



In this case, the mean-variance plot may show a curved relationship between variance and mean.



We may want to check whether the standard deviation, which is the square root of the variance, has a more linear relationship to mean.



The largest mean has a significant difference of 6.9, while the smallest mean has a significant difference of 0.3. In other words, the largest significant difference

is 23 times the smallest significant difference.

9.2.2 Homogeneity of Variance Tests

In previous units, we learned how to compare two variances – using the F-test. In the Analysis of Variance, we tested whether the variance among treatment means was greater than the variance within treatments. If the treatment variance was sufficiently greater than the error variance, we concluded the treatment effect explained a significant amount of the variation in observed values.

In this unit, we want to do something similar – we want to compare the variances associated with multiple treatments to see if they are significantly different. When data are normally-distributed, the method for comparing multiple variances is *Bartlett's Test*. (If you continue your statistical travels, you may come across Levene's Test, but that is for non-normal data.)

Bartlett's test, as best as I can tell (the formula is awful and no one seems willing to explain it), acts like a sum of squares for variances, comparing the variances of the individual treatments to their mean when pooled together. This is an incomplete explanation, but I hope it will satisfy the curious. Fortunately for us, the test is easy to implement in R and the output simple.

If we run Bartlett's test on our corn data above, we get the following results.

```
##  
##  Bartlett test of homogeneity of variances  
##  
## data: yield by trt  
## Bartlett's K-squared = 10.355, df = 3, p-value = 0.01578
```

Let's go through the output. "Bartlett's K-squared" is the statistic produced by the nasty formula I referenced above. Don't worry about that. The degrees of freedom refers to the four treatments whose variances we are comparing. Most important, of course, is our p-value. There are many opinions on when to transform data – but I would recommend against transforming data unless the p-value is less than 0.01. I would also recommend running your data on both the transformed and untransformed data and comparing results. If transformation does not change your inferences, then skip it.

Here is the Bartlett's test on our velevleaf data where the mean and standard deviation were linearly related:

```
##  
##  Bartlett test of homogeneity of variances  
##  
## data: yield by trt  
## Bartlett's K-squared = 14.217, df = 3, p-value = 0.002625
```

In this case, we will want to analyze both the transformed and untransformed data before deciding which to us for our final inferences.

9.3 Dealing with Messy Data

Dealing with messy data is one of the more uncomfortable aspects of statistics, but also one of the most important. Our tests and summary statistics are based on assumptions. For tests, we assume the data are from a populations that have approximately normal distributions. We also assume they have variances that are equal – otherwise our mean separation tests will fail.

And finally, and this is a concept that I learned just while writing this: the validity of our inferences is based on the assumption that our samples represent the population that we are studying. Which brings us back to outliers.

9.3.1 Outliers

Outliers can have a powerful effect in skewing data. Particularly in smaller datasets (i.e. fewer than 10 replicates per treatment), an outlier can have noticeable effects on a distribution's normality and its variance. In regression analyses, one outlier can significantly change the slope of the model.

Does this mean that outliers should be omitted from the dataset? Not necessarily – first we should inspect the data more closely. The outlier might be a mistake in recording a measurement. It could be an inconsistent scale in a combine. These would be experimental errors that mean the outlier is an *artifact* of our methods, rather than a representative sample from our population. In that case, we may want to remove that observation from our dataset.

But investigating the outlier may also include examining the location where it was taken. This can be difficult if you are not the primary researcher and on-site where the data were gathered. But if you can overlay your plot map with a soils map, or work with aerial imagery, or examine as-applied maps, you may be able to identify differences in environment or management that caused a dramatic difference in the observed value.

In such a case, you may decide that plot did not represent the environment about which you were trying to draw inferences, and choose to omit it from the dataset. At the same time, however, knowing that the outlier's environment or management had a dramatic effect on its performance, you may generate new hypotheses about that product. In fact, you may learn more from your outlier, through the new research it inspires, than you do from the original experiment.

Also, before removing an outlier, it is a good idea to run your tests with and without it to see whether it changes your conclusions. When you run your model, look at the standardized residuals. How many standard errors is the

outlier from the predicted value? As a rule, if an observed value is more than two standard deviations from the predicted value, I scrutinize it before allowing it into the final analysis.

If you are comparing different treatments, does it change the significance of tests or differences among treatments? If you are generating a regression model, does the change in slope have a dramatic effect on the values you will predict? Will the change in slope have a dramatic effect on grower inputs and cost, or will the effect be more modest?

These are important questions, because as uncomfortable as it is to identify and/or remove outliers, working with incomplete datasets can be even nastier. If the statistical significance of tests or means separations are not affected by the outlier, it is best to leave it in the dataset if possible, especially if treatment replications are limited.

9.3.2 Non-normal Data and Unequal Variances

Above, we discussed two other problems with data: data that were skewed (or non-normal) and therefore could not be modelled based on the normal distribution, and data where treatment variances were not equal – they were, in statistical terminology, they suffered from *heterogeneity of variances*.

Both issues can arise out of trials where observation values vary widely: counts that include rare events or where one treatment (like a check plot) can “blow up” and have a huge value. Growth studies, where size increases occur at exponential rates, are another.

These two issues may be similarly addressed by transforming the data. When we transform data, we use a different measuring system to rescale it. The easiest example of rescaling data is pH. Recall pH is the concentration of hydrogen atoms in a solution. This concentration can range from 0 to 10^{-14}

So when is the last time you read that your soil pH was 6.5×10^{-6} ? You wouldn't. We commonly speak of pH as ranging from 1 (highly acidic) through 7 (neutral) to 14 (highly basic). You are used to using the logarithmic scale (10, 1, 0.10, 0.010), rather than the arithmetic scale (1,2,3,4,5). The decibel scale for sound and the Richter scale for earthquakes also use the logarithmic scale.

9.3.2.1 Natural Logarithm

There are several ways to transform data, but the one I have most-often used is the natural logarithm. The natural logarithm transformation is often used when we are working with data that have a wide range of values. What constitutes a wide range of values? Think growth analysis, or counts of common events (for example, weed counts in a herbicide trial that includes treatments that vary

widely in effectiveness). In these trials, it is not uncommon for observed values to vary by two or more orders of magnitude (powers of 10).

Our process for working with transformed data is as follows:

- Transform the original observations to their natural logs.
- Calculate the ANOVA
- Calculate treatment means using transformed data
- Back-transform the treatment means to the original measurement scale so they are more intuitive to users

Lets work with our velvetleaf data above. Below is the analysis of variance of our initial, untransformed data:

term	df	sumsq	meansq	statistic	p.value
trt	3	500.00	166.6667	8.767315	0.0023712
Residuals	12	228.12	19.0100	NA	NA

Similarly, here is the means separation using the least significant difference test:

	yield	groups
D	20	a
C	15	ab
B	10	bc
A	5	c

The treatment effect is significant, and some of the treatments are significantly different from each other, but there are also noticeable overlaps.

Now let's transform the data using the natural log. The original and transformed data are show below.

trt	yield	log_yield
A	4.802644	1.569167
A	5.113508	1.631886
A	5.369530	1.680740
A	4.714318	1.550604
B	9.211467	2.220449
B	8.401127	2.128366
B	8.671603	2.160054
B	13.715802	2.618549
C	11.933264	2.479330
C	21.939493	3.088288
C	13.841261	2.627654
C	12.285982	2.508459
D	11.328175	2.427293
D	26.292831	3.269296
D	24.739139	3.208387
D	17.639854	2.870161

Now when we run our Bartlett's test, we see the p-value is 0.09 – there is no longer a significant difference among the treatment variances.

```
##  
##  Bartlett test of homogeneity of variances  
##  
## data: log_yield by trt  
## Bartlett's K-squared = 6.4701, df = 3, p-value = 0.09085
```

The largest standard deviation is still 6.5 times the smallest standard deviation – but the difference has decreased dramatically. When we run our analysis of variance, we see that our p-value has decreased by several orders of magnitude.

term	df	sumsq	meansq	statistic	p.value
trt	3	4.043462	1.347821	18.95189	7.58e-05
Residuals	12	0.853416	0.071118	NA	NA

When we run our LSD test, we notice more significant differences among the means, especially treatments A and B, which were associated with lower treatment means.

	log_yield	groups
D	2.943784	a
C	2.675933	ab
B	2.281854	b
A	1.608099	c

Our last step is to back-transform the means from our LSD test. Each back-transformed mean is equal to Euler's constant, e^x , where x is each treatment mean in the transformed data.

treatment	yield	groups
D	18.987563	a
C	14.525893	ab
B	9.794826	b
A	4.993311	c

In the above table, we have back-transformed the means in our LSD table to their original scale.

9.4 Dealing with Missing Data

Missing data can be very problematic. Whether missing data are the result of deleting outliers, or plots lost to weather or human damage, there are three options:

- Drop the treatment entirely from the dataset
- Drop one observation from each of the other treatments; if a Randomized Complete Block Design was used, delete an entire block
- Predict the value for the plot that was omitted or lost

As you see, these are ugly choices to make. If you drop the treatment entirely from the dataset, you lose all ability to test it against the other treatments. If you drop other replicates or the remainder of a block, you retain all treatments but reduce the degrees of freedom for statistical tests, rendering them less sensitive.

The third option, predicting the value that is missing, comes with its own challenges. The missing value for a plot is generally calculated using the linear additive model. For a completely randomized design, the linear model is:

$$Y_{ij} = \mu + T_i + \epsilon_{(i)j}$$

So the missing value would be equal to $\mu + T_i$, where i would be whatever treatment level the missing plot received.

In a randomized complete block design, the linear additive model is:

$$Y_{ij} = \mu + B_i + T_j + BT_{ij}$$

The missing value would be equal to $\mu + B_i + T_j$, where i is the block to which the missing plot occurred, and j is the treatment the missing plot received.

Note that we did not include $\epsilon_{(i)j}$ or BT_{ij} in estimating the missing values. Although this approach is widely used, this is a shortcome. When we predict a missing value from the effects of treatment or treatment within block, we are using *mean* effects. So the predicted value will be exactly the mean for a given treatment or treatment within block. Because the predicted value is closer to the treatment and block means than it would be otherwise, it will contribute less to the treatment variance than it would normally.

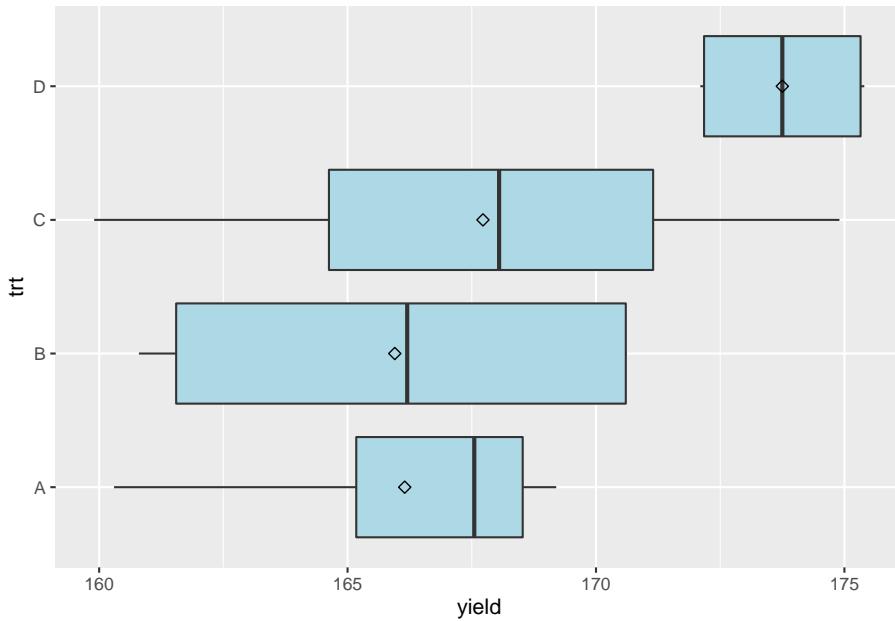
We can demonstrate this with a normally-distributed dataset.

plot	trt	yield
1	A	166.8
2	C	169.9
3	B	170.6
4	C	174.9
5	B	161.8
6	A	168.3
7	A	169.2
8	C	159.9
9	B	160.8
10	C	166.2
11	A	160.3
12	D	172.1
13	D	175.4
14	D	175.3
15	B	170.6
16	D	172.2

Here is the anova for the original data:

term	estimate	std.error	statistic	p.value
(Intercept)	166.150	2.352204	70.6358743	0.0000000
trtB	-0.200	3.326519	-0.0601229	0.9530474
trtC	1.575	3.326519	0.4734679	0.6443764
trtD	7.600	3.326519	2.2846705	0.0413279

And here is the boxplot:



Here is the table with the treatment and error effects broken out. Unlike previous effects tables, I have also added the within treatment variance.

plot	trt	yield	mu	trt_var	trt_effect	error_effect
1	A	166.8	168.3938	16.190000	-2.24375	0.650
2	C	169.9	168.3938	39.922500	-0.66875	2.175
3	B	170.6	168.3938	28.996667	-2.44375	4.650
4	C	174.9	168.3938	39.922500	-0.66875	7.175
5	B	161.8	168.3938	28.996667	-2.44375	-4.150
6	A	168.3	168.3938	16.190000	-2.24375	2.150
7	A	169.2	168.3938	16.190000	-2.24375	3.050
8	C	159.9	168.3938	39.922500	-0.66875	-7.825
9	B	160.8	168.3938	28.996667	-2.44375	-5.150
10	C	166.2	168.3938	39.922500	-0.66875	-1.525
11	A	160.3	168.3938	16.190000	-2.24375	-5.850
12	D	172.1	168.3938	3.416667	5.35625	-1.650
13	D	175.4	168.3938	3.416667	5.35625	1.650
14	D	175.3	168.3938	3.416667	5.35625	1.550
15	B	170.6	168.3938	28.996667	-2.44375	4.650
16	D	172.2	168.3938	3.416667	5.35625	-1.550

Plot 8 has a greater error effect than most other plots. Let's treat it as an outlier, delete it, and recalculate the treatment means. Let's delete it and see how that changes the treatment effect for treatment C.

plot	trt	yield	mu	trt_var	trt_effect	error_effect
1	A	166.8	168.96	16.190000	-2.810000	0.6500000
2	C	169.9	168.96	19.063333	1.373333	-0.4333333
3	B	170.6	168.96	28.996667	-3.010000	4.6500000
4	C	174.9	168.96	19.063333	1.373333	4.5666667
5	B	161.8	168.96	28.996667	-3.010000	-4.1500000
6	A	168.3	168.96	16.190000	-2.810000	2.1500000
7	A	169.2	168.96	16.190000	-2.810000	3.0500000
8	C	NA	168.96	19.063333	1.373333	NA
9	B	160.8	168.96	28.996667	-3.010000	-5.1500000
10	C	166.2	168.96	19.063333	1.373333	-4.1333333
11	A	160.3	168.96	16.190000	-2.810000	-5.8500000
12	D	172.1	168.96	3.416667	4.790000	-1.6500000
13	D	175.4	168.96	3.416667	4.790000	1.6500000
14	D	175.3	168.96	3.416667	4.790000	1.5500000
15	B	170.6	168.96	28.996667	-3.010000	4.6500000
16	D	172.2	168.96	3.416667	4.790000	-1.5500000

We can see that removing the yield data from plot 4 causes the treatment effect of treatment C to change – in fact, it has gone from negative to positive. The other treatment effects have also changed. The within-treatment variance for treatment C has also decreased by about one-third. When we re-run our analysis of variance, we see the treatment effect is 0.062 – almost significant at the P=0.05 level.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## trt          3 165.3   55.09   3.294 0.0617 .
## Residuals    11 183.9   16.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

What would happen if we estimated the yield for plot 8 using the population mean, mu, and the treatment effect?

$$Y_{3,4} = \mu + T_i = 168.96 + 1.37 = 170.33$$

We see that mu, treatment variance, treatment effect, and error have again changed. The variance within Treatment 3 has again decreased by about one-third.

plot	trt	yield	mu	trt_var	trt_effect	error_effect
1	A	166.80	169.0456	16.190000	-2.895625	0.6500
2	C	169.90	169.0456	12.708892	1.286875	-0.4325
3	B	170.60	169.0456	28.996667	-3.095625	4.6500
4	C	174.90	169.0456	12.708892	1.286875	4.5675
5	B	161.80	169.0456	28.996667	-3.095625	-4.1500
6	A	168.30	169.0456	16.190000	-2.895625	2.1500
7	A	169.20	169.0456	16.190000	-2.895625	3.0500
8	C	170.33	169.0456	12.708892	1.286875	-0.0025
9	B	160.80	169.0456	28.996667	-3.095625	-5.1500
10	C	166.20	169.0456	12.708892	1.286875	-4.1325
11	A	160.30	169.0456	16.190000	-2.895625	-5.8500
12	D	172.10	169.0456	3.416667	4.704375	-1.6500
13	D	175.40	169.0456	3.416667	4.704375	1.6500
14	D	175.30	169.0456	3.416667	4.704375	1.5500
15	B	170.60	169.0456	28.996667	-3.095625	4.6500
16	D	172.20	169.0456	3.416667	4.704375	-1.5500

And here is the kicker, wait for it...

```
## # A tibble: 2 x 6
##   term      df sumsq meansq statistic p.value
##   <chr>     <dbl> <dbl>   <dbl>    <dbl>    <dbl>
## 1 trt       3    167.    55.7     3.63    0.0450
## 2 Residuals 12   184.    15.3     NA      NA
```

Our treatment differences are now significant. Why? Because when we estimate a missing value using only the population mean and treatment effect, we decrease the overall variance. And why does that happen? Because we have now created that is almost exactly equal to the treatment mean. Was there a change in the originally observed values associated with this change in significance? No. And this is problem. But there is a way to reduce it.

The problem with the model we have used so far is we did not include the error effect in our yield estimate. If we added it in, our yield estimate for plot 8 would be more appropriate. Of course, we cannot calculate the error effect because it is random and changes among plots. But, knowing that error effects are normally distributed around the treatment mean, we can model that distribution and draw an individual from it at random, to use as the error effect in our estimate.

The error distribution has its own mean, which should be close to zero:

```
## [1] 1.894781e-15
```

And its own standard deviation:

```
## [1] 3.624684
```

Knowing these two parameters, we can select a value for our error effect from that distribution.

```
## [1] -5.36659
```

Let's plug that into our yield estimate and see how our statistics change.

$$Y_{3,4} = \mu + T_i = 168.96 + 1.37 - 5.37 = 164.96$$

We see that mu, treatment variance, treatment effect, and error have again changed. The variance within Treatment 3 has again decreased by about one-third.

plot	trt	yield	mu	trt_var	trt_effect	error_effect
1	A	166.80	168.71	16.190000	-2.56	0.65
2	C	169.90	168.71	19.927067	0.28	0.91
3	B	170.60	168.71	28.996667	-2.76	4.65
4	C	174.90	168.71	19.927067	0.28	5.91
5	B	161.80	168.71	28.996667	-2.76	-4.15
6	A	168.30	168.71	16.190000	-2.56	2.15
7	A	169.20	168.71	16.190000	-2.56	3.05
8	C	164.96	168.71	19.927067	0.28	-4.03
9	B	160.80	168.71	28.996667	-2.76	-5.15
10	C	166.20	168.71	19.927067	0.28	-2.79
11	A	160.30	168.71	16.190000	-2.56	-5.85
12	D	172.10	168.71	3.416667	5.04	-1.65
13	D	175.40	168.71	3.416667	5.04	1.65
14	D	175.30	168.71	3.416667	5.04	1.55
15	B	170.60	168.71	28.996667	-2.76	4.65
16	D	172.20	168.71	3.416667	5.04	-1.55

When we look at the ANOVA, we see that the mean square and p-value are approximately the same as they were before the missing value was interpolated.

```
##          Df Sum Sq Mean Sq F value Pr(>F)
## trt         3 158.6   52.87   3.086  0.068 .
## Residuals   12 205.6   17.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In R, there is a nice package called *mice* that does this for us. We will learn about it in an exercise this week.

One final note, however: we should not interpolate missing values if 5% or more of the data are missing. Why? Because, as we have seen above, that interpolated value can markedly change our interpretation of the data. Restricting interpolation to datasets where a small percentage of data are missing reduces the leverage one observation has on conclusions from the data. It also increases the accuracy of the interpolated values.

In smaller datasets, then, we may to use the approaches above. How important is it to have that treatment in the trial, versus losing a replicaton of the other treatments? For the latter option, you may want to test whether including or omitting that replicate changes your conclusion from the data. If not, it may be easiest to drop the replication.

9.5 Summary

Agronomic data are rarely nice and neat. The scale in which we work (acres and hectares), plus the variations in soil types, equipment performance, weather, weeds and other pests, make it virtually impossible to ensure our experimental units are near-equivalent. Above all, our subjects are alive and integrate every aspect of their environment into their growth. Unlike human subjects, corn plants cannot answer a history questionnaire. Months or years of a trial can be erased by one plugged nozzle, one broken singulator, one strong wind, one “white combine”. It’s a nasty business.

It is important that we inspect our data and consider its shortcomings. We have ways to address these shortcomings. Outliers may be trimmed, or we may use other techniques to overcome them. We have also learned how to re-scale data (using logarithms or square roots) so that variances are closer to equal, and how to “fill in” missing values using imputation so that our datasets can be balanced.

If stuck, consult with other data scientists. There are “robust” statistics that are based on using the median, rather than the mean, for summaries and tests. There are also non-parametric tests, some of which we will be introduced to towards the end of this course. Non-parametric methods don’t use linear models – therefore, they are more insulated from the problems discussed above. We will learn about these in a future unit.

Chapter 10

Correlation and Simple Regression

Let's review our progress so far.

In Units 1 to 3, we learned about populations, distributions, and samples. A population was a group of individuals in which we were interested. We use statistics to describe the spread or distribution of a population. When it is not possible to measure every individual in a population, a sample or subset can be used to estimate the frequency with which different values would be observed were the whole population measured.

In Units 4 and 5, we learned how to test whether two populations were different. The t-distribution allowed us to calculate the probability the two populations were the same, in spite of a measured difference. When the two populations were managed differently, the t-test could be used to test whether they, and therefore their management practices, caused different outcomes.

In Units 6 and 7, we learned how to test differences among multiple qualitative treatments. By qualitative, we mean there was no natural ranking of treatments: they were identified by name, and not a numeric value that occurred along a scale. Different hybrids, herbicides, and cropping systems are all examples of qualitative treatments. Different experimental designs allowed us to reduce error, or unexplained variation among experimental units in a trial. Factorial trials allowed us to test multiple factors and once, as well as test for any interactions among factors.

In Unit 8, we learned about to test for differences among multiple qualitative treatments. The LSD and Tukey tests can be used to test the difference among treatments. Contrasts can be used to compare intuitive groupings of treatments. We learned how to report results in tables and plots.

In Units 9 and 10, we will learn to work with quantitative treatments. Quantitative treatments can be ranked. The most obvious example would be rate trials for fertilizers, crop protection products, or crop seeds. What makes this situation different is that we are trying to describe a relationship between x and y along within a range of x values, rather at only at discrete values of x .

10.1 Correlation versus Regression

There are two ways of analyzing the relationship between two quantitative variables. If our hypothesis is that an change in x , the independent variable (for example, pre-planting nitrogen fertilization rate), *causes* a change in y , the dependent variable (for example, yield), then we use a *regression model* to analyze the data. Not only can the relationship between tested for significance – the model itself can be used to predict y for any value of x within the range of those in the dataset.

Sometimes, however, we suspect that x and y are associated, but we don't know whether x causes y , or y causes x . This is the chicken-and-egg scenario. Outside animal science, however, we can run into this situation in crop development when we look at the allocation of biomass to different plant parts or different chemical components of tissue. In this case, we analyze the *correlation* between x and y .

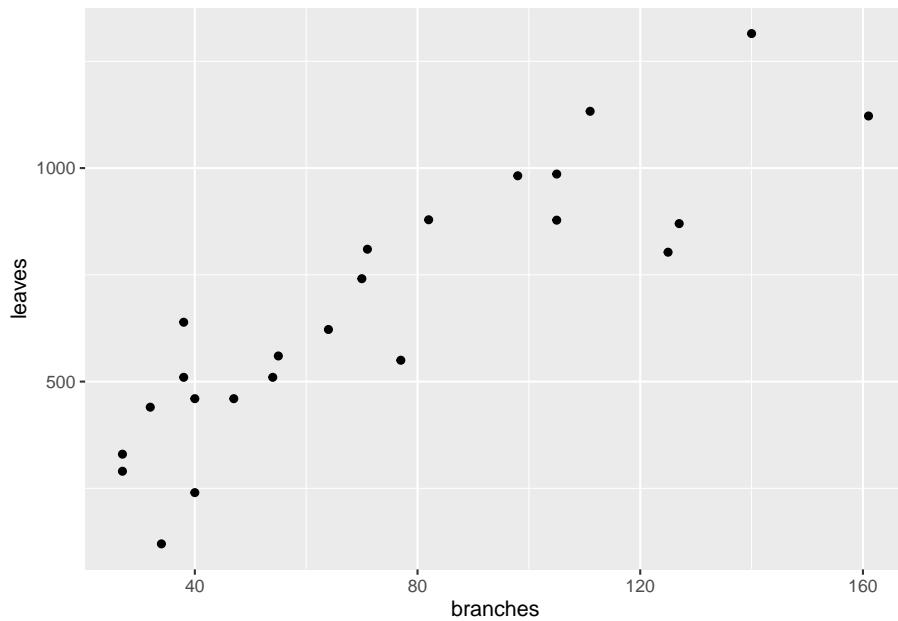
10.1.1 Case Study: Cucumber

In this study, cucumbers were grown the their number of leaves, branches, early fruits, and total fruits were measured.

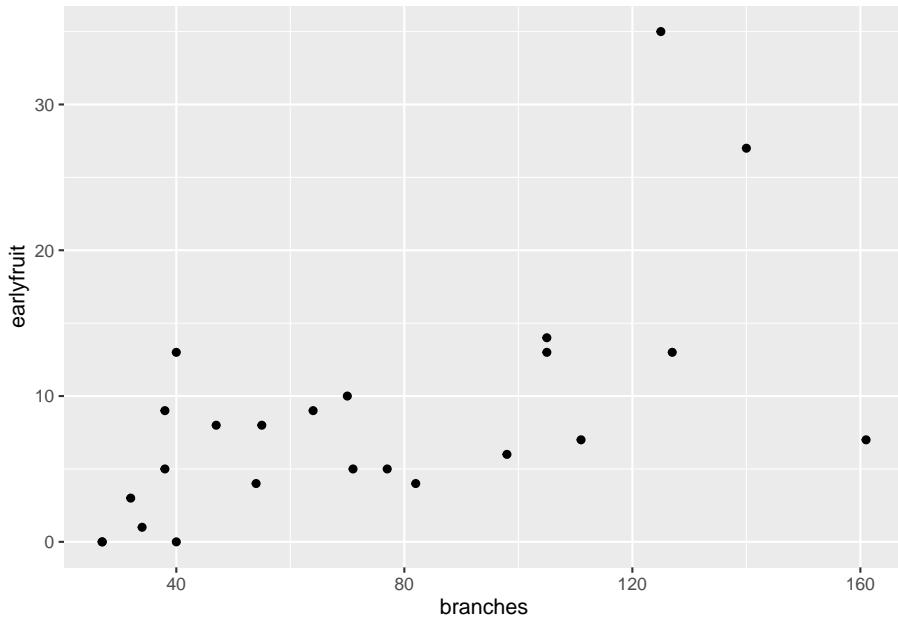
First, let's load the data and look at its structure.

cycle	rep	plants	flowers	branches	leaves	totalfruit	culledfruit	earlyfruit
1	1	29	22	40	240	1	0	0
1	2	21	17	34	120	17	2	1
1	3	31	52	32	440	29	15	3
1	4	30	49	77	550	25	9	5
1	5	28	88	140	1315	58	13	27
1	6	28	162	105	986	39	9	14

What is the relationship between branches and leaves? Let's plot the data.



Similarly, what is the relationship between earlyfruit and the number of branches?



In both cases, we can see that as one variable increases, so does the other. But we don't know whether the increase in one causes the increase in the other, or whether there is another variable (measured or unmeasured, that causes both to increase).⁴

10.2 Correlation

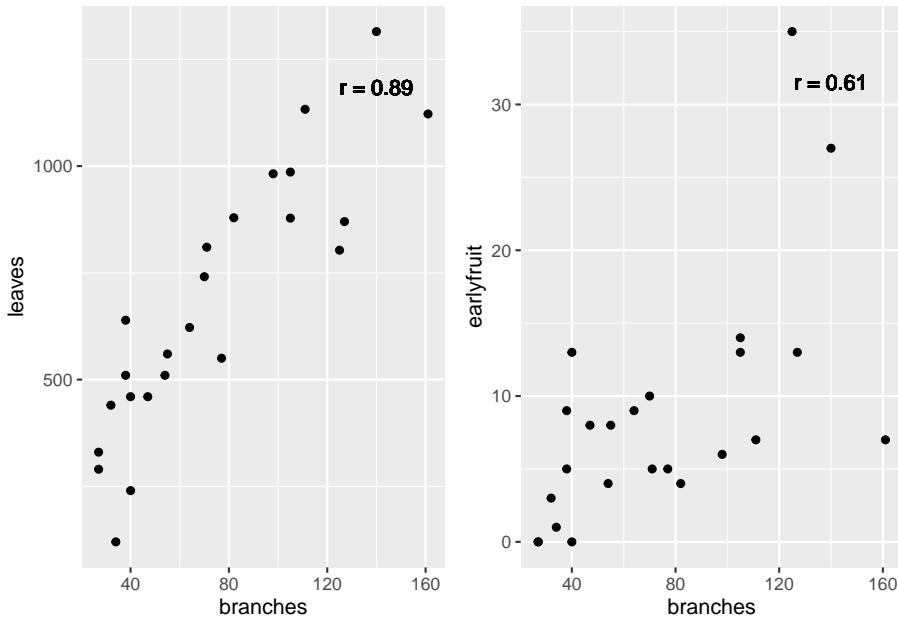
Correlation doesn't measure *causation*. Instead, it measures *association*. The first way to identify correlations is to plot variables as we have just done. But, of course, it is hard for us to measure the strength of the association just by eye. In addition, it is good to have a way of directly measuring the strength of the correlation. Our measure in this case is the *correlation coefficient*, r . r varies between -1 and 1 . Values near 0 indicate little or no association between Y and X . Values close to 1 indicate a strong, positive relationship between Y and X . A positive relationship means that as X increases, so does Y . Conversely, values close to -1 indicate a strong, negative relationship between Y and X . A negative relationship means that as X increases, Y decreases.

Experiment with the application found at the following link:

<https://marin-harbur.shinyapps.io/10-correlation/>

What happens as you adjust the value of r using the slider control?

For the cucumber datasets above, the correlations are shown below:



10.2.1 How Is r Calculated (optional reading)

Something that bothered me for years was understanding what r represented – how did we get from a cloud of data to that number?. The formula is readily available, but how does it work? To find the explanation in plain English is really hard to find, so I hope you enjoy this!

To understand this, let's consider you are in Marshalltown, Iowa, waiting for the next Derecho. You want to go visit your friends, however, who are similarly waiting in Des Moines for whatever doom 2020 will bring there. How will you get there?

First, you could go “as the crow flies” on Routes 330 and 65. This is the shortest distance between Marshalltown and Des Moines. In mathematical terms this is known as the “Euclidian Distance”. Now you probably know the Euclidean Distance by a different name, the one you learned in eighth grade geometry. Yes, it is the hypotenuse, the diagonal on a right triangle!

Second, you might want to stop in Ames to take some barbecue or pizza to your friends in Des Moines. In that case, you would travel a right angle, “horizontally” along US 30-and then “vertically” down I-35. You might call this “going out of your way”. The mathematical term is “Manhattan Distance”. No, not Kansas. The Manhattan distance is named for the grid system of streets in the upper two thirds of Manhattan. The avenues for the vertical axes and the streets form the horizontal axes.

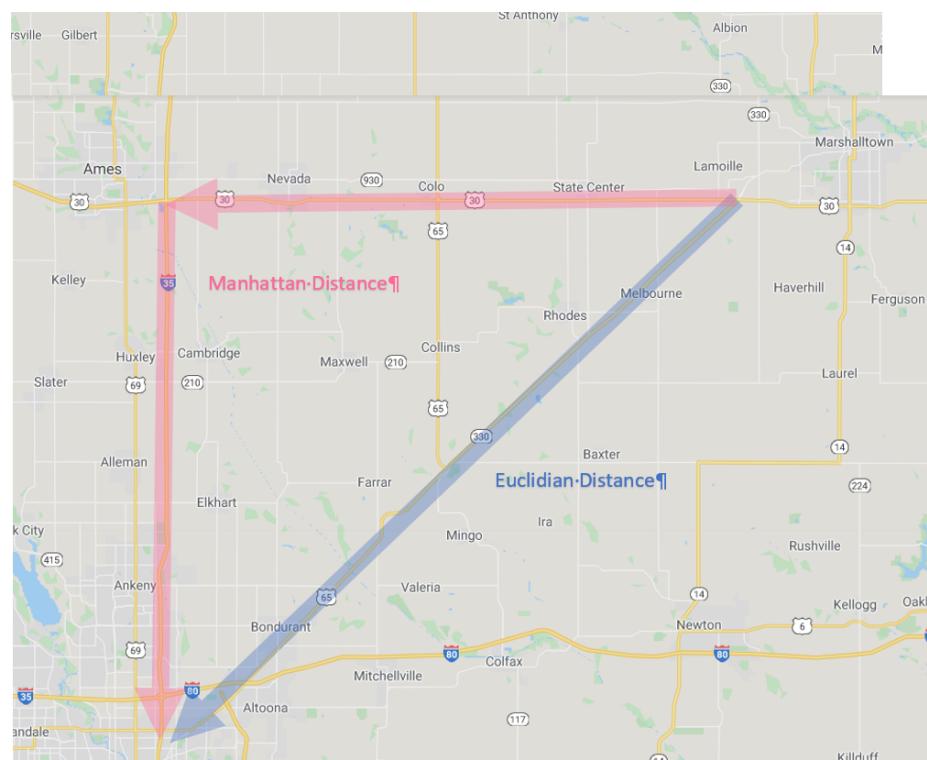


Figure 10.1: Iowa Map

As you might remember, the length of the hypotenuse of a right triangle is calculated as:

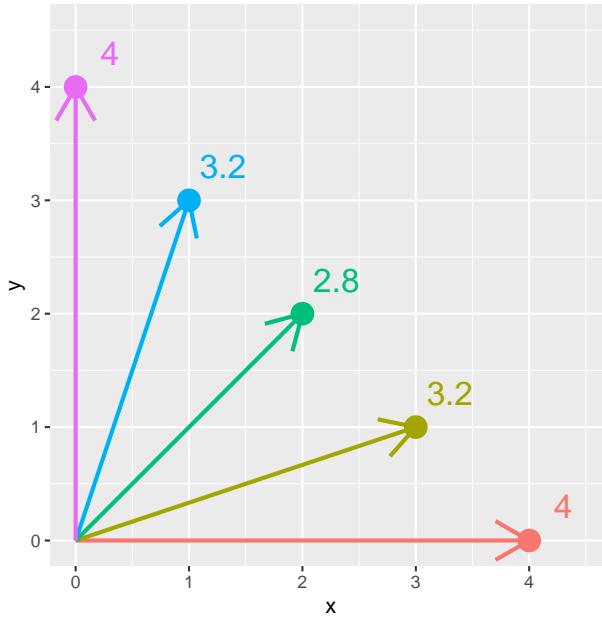
$$z^2 = x^2 + y^2$$

Where z is the distance as the crow flies, x is the horizontal distance, and y is the vertical distance. This is the Euclidian Distance. The Manhattan distance, by contrast, is simply $x + y$.

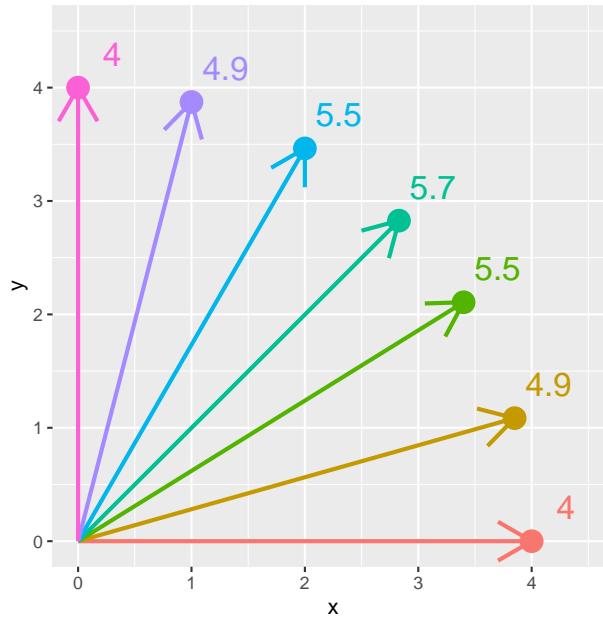
Now what if we were simply driving from Marshalltown to Ames? Would the Euclidian distance and the Manhattan distance be so different? No, because both Marshalltown and Ames are roughly on the x-axis? Similarly, what if we drove from Ames to Des Moines? The Euclidian distance and Manhattan distance would again be similar, because we were travelling across the x-axis.

The difference between the Euclidian distance and the Manhattan distance is greatest when we must travel at a 45 degree angle from the X axis. We can demonstrate this with the following plot. Every point below is four units from the origin ($x = 0, y = 0$). Notice that when the point is on the x or y axis, the Euclidian distance and Manhattan distance are equal. But as the angle increases to zero, the Euclidian distance decreases, reaching its lowest value when $x=y$ and the angle from the axis is 45 degrees.

In the plot below, each point has a Manhattan distance ($x + y$) of 4. The Euclidian distance is shown beside each point. We can see the Euclidian distance is least when $y = x = 2$.



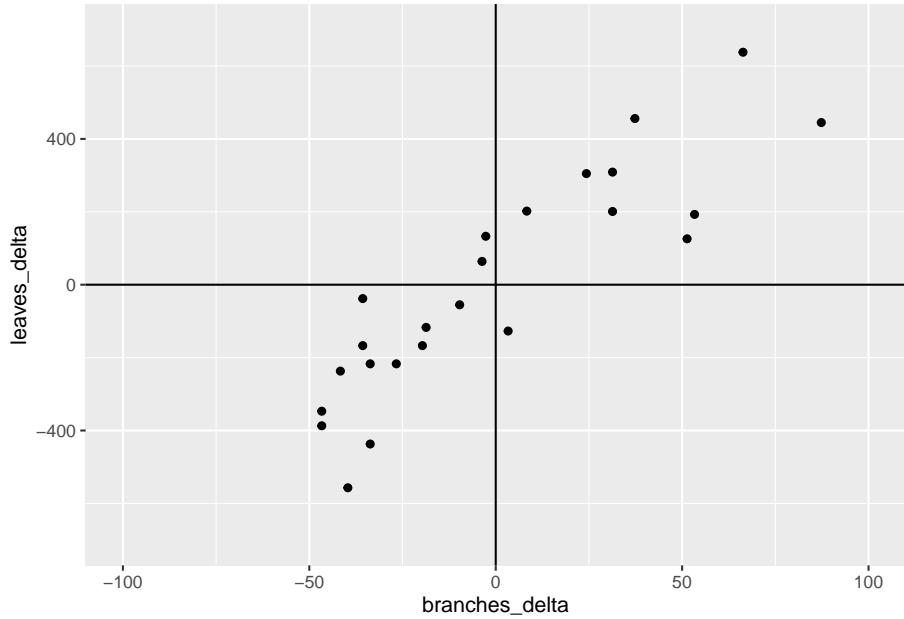
Conversely, in the plot below, each point is the same Euclidean distance (4 units) from the origin ($x = 0, y = 0$). The Manhattan distance is shown beside each point. We can see the Manhattan distance is greatest when the point is at a 45 degree angle from the origin.



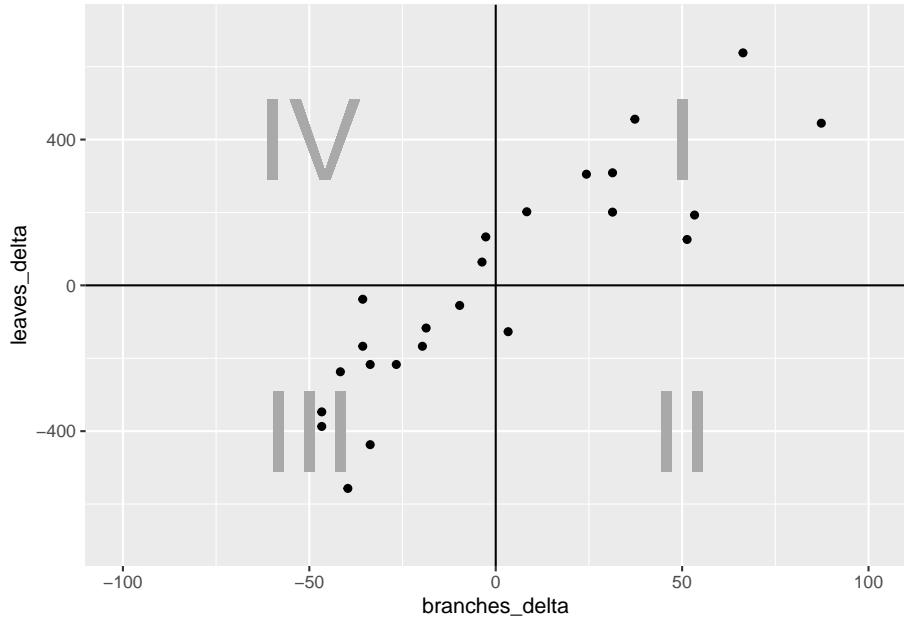
The calculation of the correlation coefficient, r , depends on this concept of covariance between x and y . The covariance is calculated as:

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

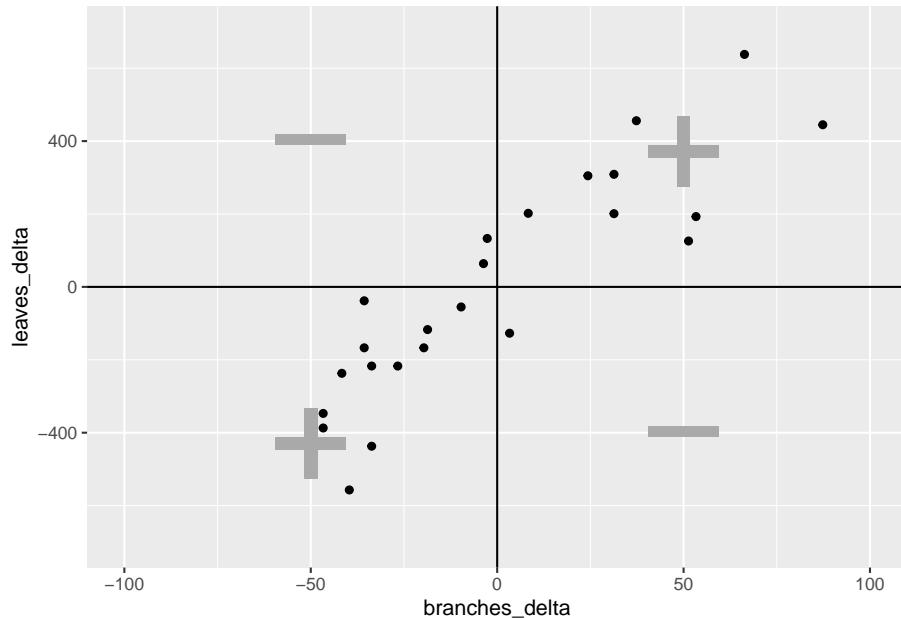
Let's go back to our cucumber data. We will calculate the difference of each point from the \bar{x} and \bar{y} . The points are plotted below.



What we have now are four quadrants. The differences between them are important because they affect the *sign* of the covariance. Remember, the covariance of each point is the product of the x-distance and the y-distance of each point.



In quadrant I, both the x-distance and y-distance are positive, so their product will be positive. In quadrant II, the x-distance is still positive but the y-distance is negative, so their product will be negative. In quadrant III, both x-distance and y-distance are negative, so their negatives will cancel each other and the product will be positive. Finally, quadrant IV, will have a negative x-distance and positive y-distance and have a negative sign.



Enough already! How does all this determine r ? It's simple – the stronger the association between x and y , the more linear the arrangement of the observations in the plot above. The more linear the arrangement, the more the points will be in diagonal quadrants. In the plot above, any observations that fall in quadrant I or III will contribute to the positive value of r . Any points that fall in quadrants II or IV will subtract from the value of r .

In that way, a loose distribution of points around all four quadrants, which would indicate x and y are weakly associated, would be penalized with an r score close to zero. A distribution concentrated in quadrants I and III would have a positive r value closer to 1, indicating a *positive* association between x and y . Conversely, a distribution concentrated in quadrants II and IV would have a negative r value closer to -1, and a *negative* association between x and y .

One last detail. Why is r always between -1 and 1? To understand that, we look at the complete calculation of r .

$$r = \frac{S_{xy}}{\sqrt{S_{xx}} \cdot \sqrt{S_{yy}}}$$

You don't need to memorize this equation. Here is what it is doing, in plain English. S_{xy} is the covariance. It tells us, for each point, how its x and y value vary together. S_{xx} is the sum of squares of x. It sums the distances of each point from the origin ($x = 0, y = 0$) along the x axis. S_{yy} is the sum of squares of y. It is the total y distance of points from the origin. By multiplying the square root of S_{xx} and S_{yy} , we calculate the maximum theoretical covariance that the points in our measure could have.

r is, after all this, a proportion. It is the measured covariance of the points, divided by the covariance they would have if they fell in a straight line.

I hope you enjoy this explanation. I don't like to go into great detail about calculations, but one of my greatest hangups with statistics is how little effort is often made to explain where the statistics actually come from. If you look to Wikipedia for an explanation, you usually get a formula that assumes you have an advanced degree in calculus. Why does this have to be so hard?

Statistics is the end, about describing the *shape* of the data. How widely are the observations distributed? How far do they have to be from the mean to be too improbable to be the result of chance? Do the points fall in a line or not? There is beauty in these shapes, as well as awe that persons, decades and even millenia before digital computers, discovered how to describe them with formulas.

Then again, it's not like they had *Tiger King* to watch.

10.3 Regression

Regression describes a relationship between an dependent variable (usually represented by the letter y) and one or more independent variables (x_1, x_2 , etc). Regression differs from correlation in that the model assumes that the value of x is substantially determined by the value of x . Instead of describing the *association* between y and x , we now refer to causation – how X determines the value of Y .

Regression analysis may be conducted simply to test the hypothesis that a change in one variable drives a change in another, for example, that an increase in herbicide rate causes an increase in weed control. Regression, however, can also be used to predict the value of y based on intermediate values of x , that is, values of x between those used to fit or “train” the regression model.

The prediction of values of y for intermediate values of x is called *interpolation*. In the word interpolation we see “inter”, meaning between, and “pole”, meaning end. So interpolation is the prediction of values between actually sampled values. If we try to make predictions for y outside of the range of values of x in which the model was trained , this is known as *extrapolation*, and should be approached very cautiously.

If you hear a data scientist discuss a predictive model, it is this very concept to which they refer. To be fair, there are many tools besides regression that are used in predictive models. We will discuss those toward the end of this course. But the regression is commonly used and one of the more intuitive predictive tools in data science.

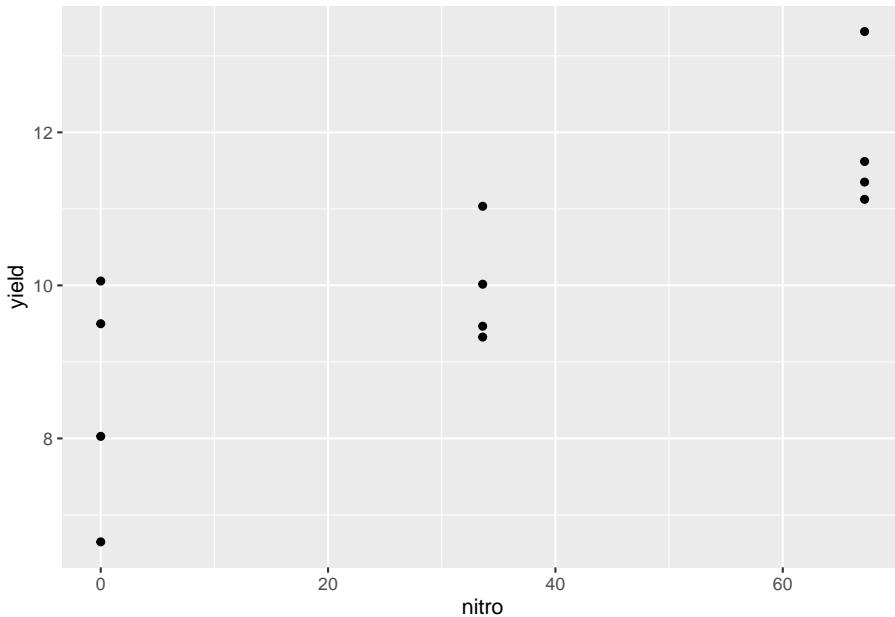
In this lesson, we will learn about one common kind of regression, simple linear regression. This means we will learn how to fit a cloud of data with a straight line that summarizes the relationship between Y and X . This assumes, of course, that it is appropriate to use a straight line to model that relationship, and there are ways for us to test that we will learn at the end of this lesson.

10.3.1 Case Study

A trial was conducted in Waseca, Minnesota, to model corn response to nitrogen. Yields are per plot, not per acre.

site	loc	rep	nitro	yield
S3	Waseca1	R1	0.0	9.49895
S3	Waseca1	R2	0.0	10.05715
S3	Waseca1	R3	0.0	8.02693
S3	Waseca1	R4	0.0	6.64823
S3	Waseca1	R1	33.6	10.01547
S3	Waseca1	R2	33.6	11.03366

The first thing we should do with any data, but especially data we plan to fit with a linear regression model, is to visually examine the data. Here, we will create a scatterplot with yield on the Y-axis and nitro (the nitrogen rate) on the X-axis.



We can see that nitrogen was applied at three rates. It is easy to check these five rates using the unique command in R.

```
unique(nitrogen$nitro)
## [1] 0.0 33.6 67.2
```

We can see that the centers of the distribution appear to fall in a straight line, so we are confident we can model the data with simple linear regression.

10.3.2 Linear Equation

Simple linear regression, unlike correlation, fits the data could with an equation. In Unit 5, we revisited middle school, where you learned that a line can be defined by the following equation:

$$y = mx + b$$

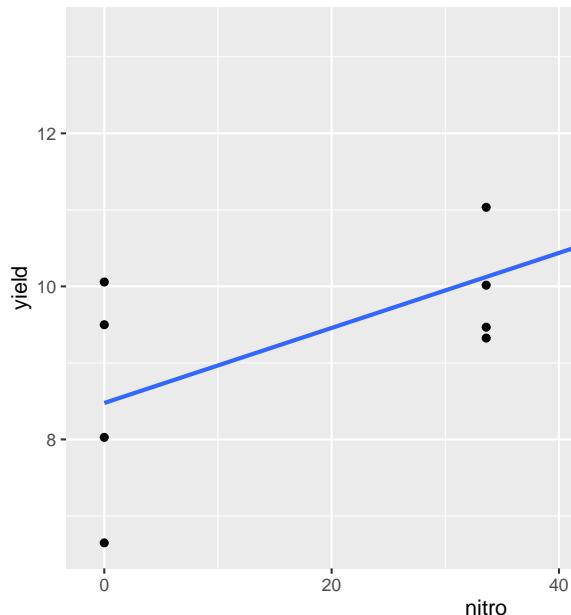
Where y and x define the coordinate of a point along that line, m is equal to the slope or “rise” (the change in the value of y with each unit change in x , and b is the y -intercept (where the line crosses the y -axis. The y -intercept can be seen as the “anchor” of the line; the slope describes how the line is pivoted on that anchor.

In statistics we use a slightly different equation – same concept, different annotation

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

The y intercept is represented by the greek letter α , the slope is represented by the letter β . y , α , and β all have hat-like symbols called carats ($\hat{\cdot}$) above them to signify they are estimates, not known population values. This is because the regression line for the population is being estimated from a sample. \hat{y} is also an estimate, since it is calculated using the estimated values $\hat{\alpha}$ and $\hat{\beta}$. Only x , which in experiments is manipulated, is a known value.

10.3.3 Calculating the Regression Equation



We can easily add a regression line to our scatter plot.

The blue line represents the regression model for the relationship between yield and nitro. Of course, it would be nice to see the linear equation as well, which we can estimate using the `lm()` function of R.

```
##  
## Call:  
## lm(formula = yield ~ nitro, data = nitrogen)  
##  
## Residuals:
```

```

##      Min    1Q Median    3Q    Max
## -1.8278 -0.6489 -0.2865  0.9384  1.5811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.47604   0.50016 16.947 1.08e-08 ***
## nitro       0.04903   0.01153  4.252  0.00168 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 10 degrees of freedom
## Multiple R-squared:  0.6439, Adjusted R-squared:  0.6083 
## F-statistic: 18.08 on 1 and 10 DF,  p-value: 0.001683

```

The estimates above define our regression model. The number to the right of “(Intercept)” is the coefficient for the y-intercept, or $\hat{\alpha}$. The number to the right of nitro is the coefficient for the slope, or $\hat{\beta}$. Knowing that, we can construct our regression equation:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$$\hat{y} = 8.476 + 0.04903x$$

This tells us that the yield with 0 units of n is about 8.5, and for each unit of nitrogen yield increases about 0.05 units. If we had 50 units of nitrogen, our yield would be:

$$\hat{y} = 8.476 + 0.04903(50) = 10.9275$$

Since nitrogen was only measured to three significant digits, we will round the predicted value to 10.9.

If we had 15 units of nitrogen, our yield would be:

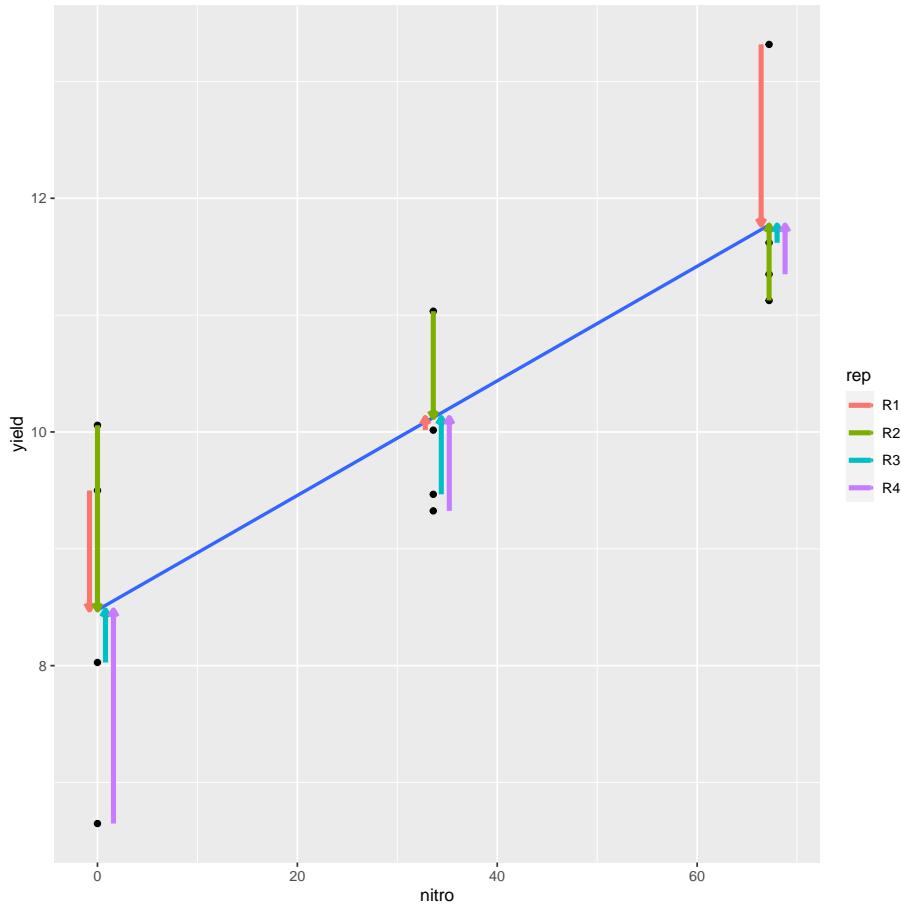
$$\hat{y} = 8.476 + 0.04903(15) = 9.21145$$

Which rounds to 9.21. So how is the regression line calculated?

10.3.4 Least-Squares Estimate

The regression line is fit to the data using a method known as the least-squares estimate. To understand this concept we must recognize the goal of a regression equation is to make the most precise estimate for Y as possible. We are not estimating X, which is already known. Thus, the regression line crosses the data

cloud in a way that minimizes the vertical distance (Y-distance) of observations from the regression line. The horizontal distance (X-distance) is not fit by the line.



In the plot above, the distances of each the four points to the regression line are highlighted by arrows. The arrows are staggered (“jittered”, in plot lingo) so you can see them more easily. Note how the line falls closely to the middle of the points at each level of R.

Follow the link below to an appl that will allow you to adjust the slope of a regression line and observe the change in the error sum of squares, which measures the sums of the differences between observed values and the value predicted by the regression line.

<https://marin-harbur.shinyapps.io/10-least-squares/>

You should have observed the position of the regression line that minimizes the sum of squares is identical to that fit using the least squares regression technique.

The line is fit using two steps. First the slope is determined, in an approach that is laughably simple – after you understand it.

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

What is so simple about this? Let's remember what the covariance and sum of squares represents. The covariance is the sum of the manhattan distances of each individual from the “center” of the sample, which is the point located at (\bar{x}, \bar{y}) . For each point, the Manhattan distance is the product of the horizontal distance of an individual from \bar{x} , multiplied by the sum of the vertical distance of an individual from \bar{y} .

The sum of squares for x , (S_{xx}) is the sum of the squared distances of each individual from the \bar{x} .

We can re-write the fraction above as:

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})}$$

In the equation above, we can cancel out $(x_i - \bar{x})$ from the numerator and denominator so that we are left with:

$$\hat{\beta} = \frac{\sum (y_i - \bar{y})}{\sum (x_i - \bar{x})}$$

) In other words, the change in y over the change in y .

Once we solve for slope ($\hat{\beta}$) we can solve for the y-intercept ($\hat{\alpha}$). Alpha-hat is equal to the

$$\hat{\alpha} = \hat{y} - \hat{\beta}\bar{x}$$

This equation tells us how much the line descends (or ascends) from the point (\bar{x}, \bar{y}) to where $x=0$ (in other words, the Y axis).

10.3.5 Significance of Coefficients

What else can we learn from our linear model?

```
##  
## Call:  
## lm(formula = yield ~ nitro, data = nitrogen)  
##
```

```

## Residuals:
##      Min     1Q Median     3Q    Max
## -1.8278 -0.6489 -0.2865  0.9384  1.5811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.47604   0.50016 16.947 1.08e-08 ***
## nitro       0.04903   0.01153  4.252  0.00168 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 10 degrees of freedom
## Multiple R-squared:  0.6439, Adjusted R-squared:  0.6083
## F-statistic: 18.08 on 1 and 10 DF,  p-value: 0.001683

```

The table above shows the coefficient estimates, as we saw before. But it also provides information on the standard error of these estimates and tests whether they are significantly different.

Again, both $\hat{\beta}$ and the $\hat{\alpha}$ are estimates. The slope of the least-squares line in the actual population may tilt a more downward or upward than this estimate. Similarly, the line may be higher or lower in the actual population, depending on the actual Y axis. We cannot know the actual slope and y-intercept of the population.

From the sample we can define confidence intervals for both values, by multiplying the standard error by the statistic (“statistic” in this table is short for “t-statistic), and calculate the probability that they differ from hypothetical values by chance.

We forego the discussion how these values are calculated – most computer programs readily provide these – and instead focus on what they represent. The confidence interval for the Y-intercept represents a range of values that is likely (at the level we specify, often 95%) to contain the true Y-intercept for the true regression line through the population.

In some cases, we are interested if the estimated intercept differs from some hypothetical value – in that case we can simply check whether how the true population Y-intercept compares to a hypothetical value. If the value falls outside the confidence interval, we conclude the values are significantly different. In other words, there is low probability the true Y-intercept is equal to the hypothetical value.

More often, we are interested in whether the slope is significantly different than zero. This question can be represented by a pair of hypotheses:

$$H_o : \beta = 0$$

$$H_a : \beta \neq 0$$

The null hypothesis, H_0 , is the slope of the true regression line is equal to zero. In other words, y does not change in a consistent manner with changes in x . Put more bluntly: there is no significant relationship between y and x . The alternative hypothesis, H_a , is the slope of the true regression line is *not* equal to zero. Y *does* vary with X in a consistent way, so there is a significant relationship between Y and X in the population.

The significance of the difference of the slope from zero may be tested two ways. First, a t-test will directly test the probability that 0 . Second, the significance of the linear model may be tested using an Analysis of Variance, as we learned in Units 5 and 6.

10.3.6 Analysis of Variance

Similarly, we can use the *summary.aov()* command in R to generate an analysis of variance of our results.

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## nitro       1  21.71  21.713   18.08 0.00168 **
## Residuals  10  12.01   1.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis of variance works the same as those you learned previously. The variance described by the relationship between Y and X (in this example identified by the “nitro” term) is compared to the random variance among data points. If the model describes substantially more variance than explained by the random location of the data points, the model can be judged to be significant.

For a linear regression model, the degree of freedom associated with the effect of x on y is always 1. The concept behind this is that if you know the mean and one of the two endpoints, you can predict the other endpoint. The residuals will have $n - 1$ degrees of freedom, where n is the total number of observations.

Notice that the F-value is the square of the calculated t-value for slope in the coefficients table. This is always the case.

10.3.7 Measuring Model Fit with R-Square

Let’s return to the summary of the regression model.

```
## 
## Call:
## lm(formula = yield ~ nitro, data = nitrogen)
```

```

## 
## Residuals:
##   Min     1Q Median     3Q    Max
## -1.8278 -0.6489 -0.2865  0.9384  1.5811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.47604   0.50016 16.947 1.08e-08 ***
## nitro       0.04903   0.01153  4.252  0.00168 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 10 degrees of freedom
## Multiple R-squared:  0.6439, Adjusted R-squared:  0.6083
## F-statistic: 18.08 on 1 and 10 DF,  p-value: 0.001683

```

At the bottom is another important statistic, *Multiple R-squared*, or R^2 . How well the model fits the data can be measured with the statistic R^2 . Yes, this is the square of the correlation coefficient we learned earlier. R^2 describes the proportion of the total sum of squares described by the regression model: it is calculated by dividing the model sum of squares.

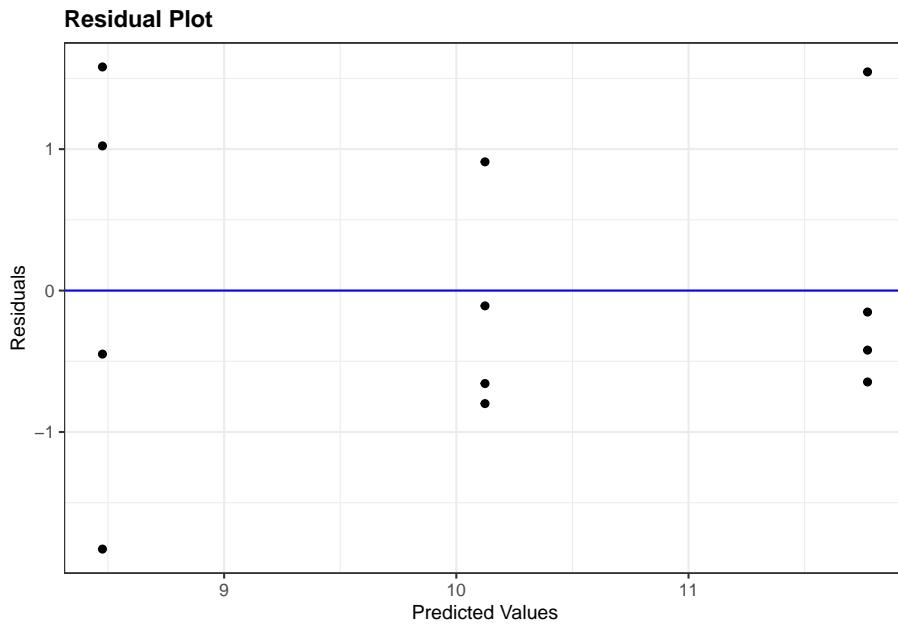
The total sum of squares in the model above is $21.71 + 12.01 = 33.72$. We can confirm the R^2 from the model summary by dividing the model sum of squares, 21.71, by this total, 33.72. $21.71 \div 33.72 = 0.6439$. This means that 64% of the variation between observed values can be explained by the relationship between yield and nitrogen rate.

R^2 has a minimum possible value of 0 (no relationship at all between y and x) and 1 (perfect linear relationship between y and x). Along with the model coefficients and the analysis of variance, it is the most important measure of model fit.

10.3.8 Checking whether the Linear Model is Appropriate

As stated earlier, the simple linear regression model is a predictive model – that is, it is not only useful for establishing a linear relationship between Y and X – it can also be used under the correct circumstances to predict Y given a known value of X. But while a model can be generated in seconds, there are a couple of cautions we must observe.

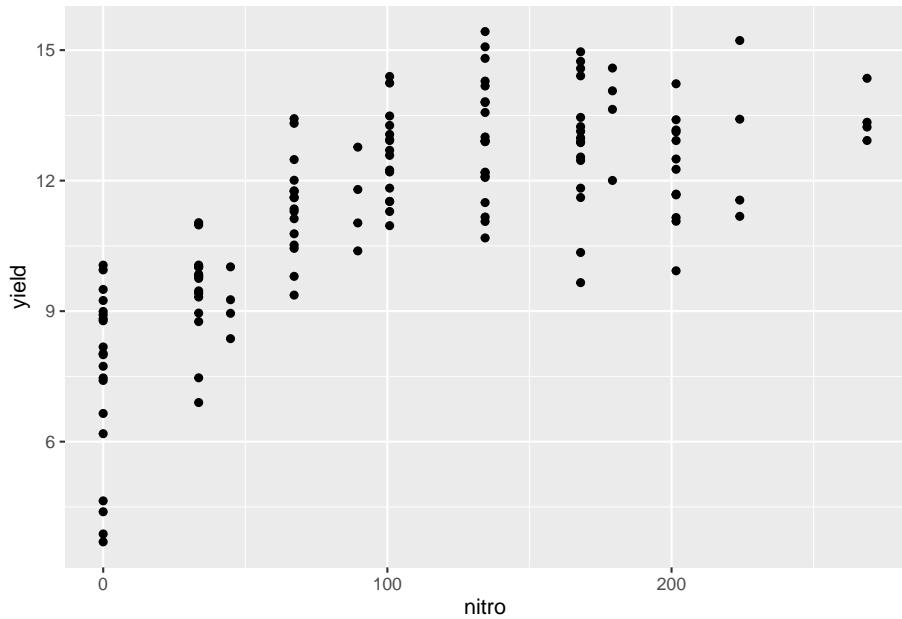
First, we must be sure it was appropriate to fit our data with a linear model. We can do this by plotting the residuals around the regression line. The *ggResid_panel* package in R allows us to quickly inspect residuals. All we do use run *resid_panel()* function with two arguments: the name of our model (“regression_model”) and the plot we want (plots = “resid”).



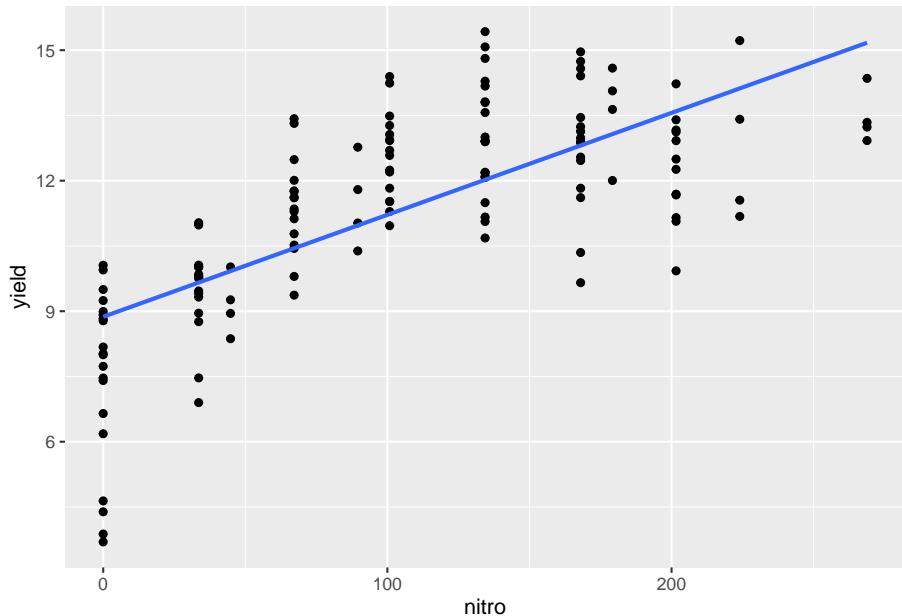
In my mind, the residual plot is roughly equivalent to taking the regression plot above and shifting it so the regression line is horizontal. There are a few more differences, however. The horizontal axis is the y-value predicted by the model for each value of x. The vertical axis is the standardized difference (the actual difference divided by the mean standard deviation across all observations) of each observed value from that predicted for it. The better the regression model fits the observations the closer the points will fall to the blue line.

The key thing we are checking is whether there is any pattern to how the regression line fits the data. Does it tend to overpredict or underpredict the observed values of x? Are the points randomly scattered about the line, or do they seem to form a curve?

In this example, we only modelled a subset of the nitrogen study data. I intentionally left out the higher rates. Why? Let's plot the complete dataset.



We can see the complete dataset does not follow a linear pattern. What would a regression line, fit to this data, look like?



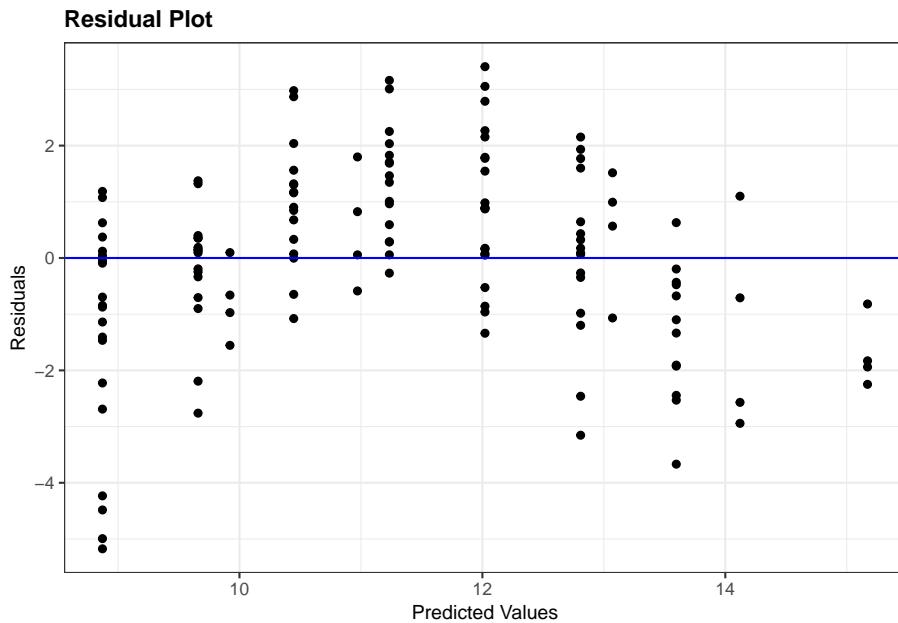
We can see how the regression line appears seems to overpredict the observed values for at low and high values of nitrogen and underpredict the intermediate values.

values.

What does our regression model look like?

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## nitro          1  395.1   395.1  140.9 <2e-16 ***
## Residuals     134  375.7      2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is still highly significant, even though it is obvious it doesn't fit the data! Why? Because the simple linear regression model only tests whether the slope is different from zero. Let's look at the residuals:



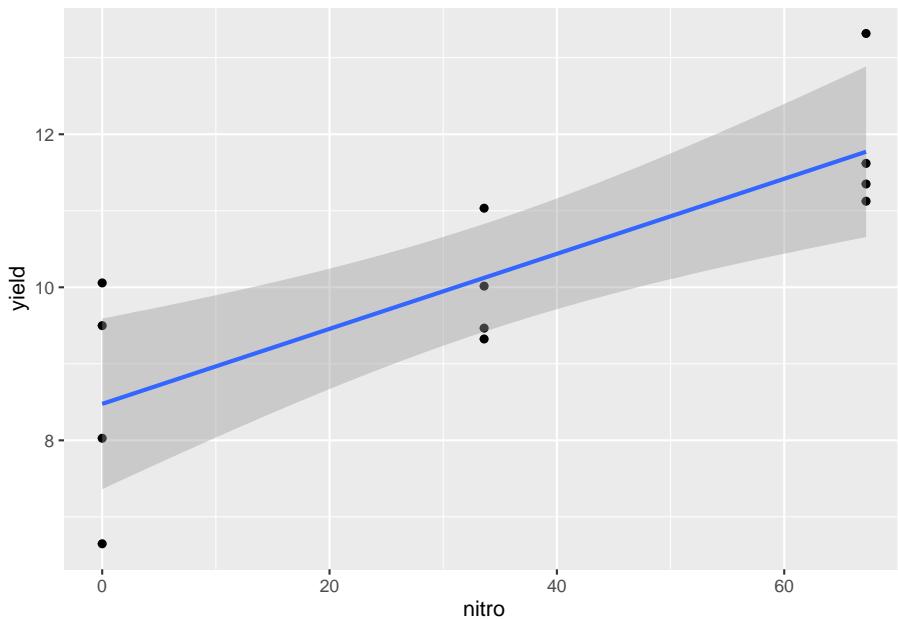
As we expect, there is a clear pattern to the data. It curves over the regression line and back down again. If we want to model the complete nitrogen response curve, we will need to use a nonlinear model, which we will learn in the next unit.

10.4 Extrapolation

The above example also illustrates why we should not extrapolate: because we do not know how the relationship between x and y may change. In addition, the

accuracy of the regression model decreases as one moves away from the middle of the regression line.

Given the uncertainty of the estimated intercept, the entire true regression line may be higher or lower – i.e. every point on the line might be one unit higher or lower than estimated by our estimated regression model. There is also uncertainty in our estimate of slope – the true regression line may have greater or less slope than our estimate. When we combine the two sources of uncertainty, we end up with a plot like this:



The dark grey area around the line represents the standard error of the prediction. The least error in our estimated regression line – and the error in any prediction made from it occurs closer at \bar{x} . As the distance from \bar{x} increases, so does the uncertainty of the Y-value predicted from it. At first, that increase in uncertainty is small, but it increases rapidly as we approach the outer data points fit with the model.

We have greater certainty in our predictions when we predict y for values of x between the least and greatest x values used in fitting the model. This method of prediction is called interpolation – we are estimating Y for X values within the range of values used to estimate the model.

Estimating Y for X values outside the range of values used to estimate the model is called extrapolation, and should be avoided. Not only is our current model less reliable outside the data range used to fit the model – we should not even assume that the relationship between Y and X is linear outside the of the range of data we have analyzed. For example, the middle a typical growth curve (often

called “sigmoidal”, from sigma, the Greek word for “S”) is linear, but each end curves sharply.

When we make predictions outside of the range of x values used to fit our model, this is extrapolation. We can now see why it should be avoided.

Chapter 11

Nonlinear Relationships and Multiple Linear Regression

In the last unit, we learned how to describe linear relationships between two variables, X and Y . *Correlation* was used when we wanted to measure the *association* between the two variables. This was appropriate when when, based on our “domain knowledge” (agronomy or our other specialty), we did not have insight into whether one variable affected the other, or whether a third, unknown variable affected the value of both. *Simple Linear Regression* was used when we had insight into the *causation* that linked two variables: we knew or hypothesized that a single response variable, Y , was affected (and could be predicted) by a single predictor variable, X .

Simple linear regression, however, has its limitations. First, simple linear regression is often too inadequate for modelling more complex systems. For example, a simple linear regression model might fit the effect of rainfall on corn yield. But we all realize that a prediction of corn yield based on rainfall alone will not be particularly accurate.

If we included additional predictor variables, such as a measure of heat (cumulative growing degree days) or soil texture (water holding capacity or clay content), we would likely predict yield more accurately. *Multiple Linear Regression* allows us to build model the relationship between a response variable, Y , and multiple predictor variables, X_1 , X_2 , X_3 , etc.

Second, a linear model assumes the relationship between Y and X can be fit with a straight line. If you have taken a soil science course, however, you learned about *Leibig’s Law of the Minimum* and the *Mitscherlich Equation* that describe the relationship between nutrient availability and plant biomass. These

relationships are curved and need to be fit with *Nonlinear Regression* models.

In this unit, we will learn how to use multiple linear regression to model yield responses to environment. We will also learn how to model nonlinear relationships, such as fertilizer response and plant growth

11.1 Multiple Linear Regression

In multiple linear regression, response variable Y is modeled as a function of multiple X variables:

$$Y = \mu + X_1 + X_2 + X_3 \dots X_n$$

11.1.1 Case Study: Modelling Yield by County

We are going to work with a county-level dataset from Purdue university that includes soil characteristics, precipitation, corn, and soybean yield. I've become best-friends with this dataset during the past few years and have combined it with county-level weather data to build complex yield models. The authors used land use maps to exclude any acres not in crop production. Click the following link to access a little app with which to appreciate this dataset.

Here is the top of the data.frame:

county	state	stco	ppt	whc	sand	silt	clay	om	k
Abbeville	SC	45001	562.2991	22.94693	39.26054	21.38839	39.351063	36.92471	0.1
Acadia	LA	22001	751.1478	37.83115	12.82896	54.19231	32.978734	105.79133	0.1
Accomack	VA	51001	584.6807	18.56290	74.82042	16.82072	8.358863	103.11425	0.1
Ada	ID	16001	124.9502	15.54825	44.49350	40.48000	15.026510	77.61499	0.1
Adair	IA	19001	598.4478	28.91851	17.82011	48.44056	33.739326	259.53255	0.1
Adair	KY	21001	698.6234	20.87349	15.66781	48.16733	36.164866	73.05356	0.1

How does corn yield respond to soil properties and precipitation in Minnesota and Wisconsin? Let's say we want to model corn yield as a function of precipitation (ppt), percent sand (sand), percent clay (clay), percent organic matter (om) and soil pH (spH)? Our linear additive model would be:

$$Y = \alpha + \beta_1 \text{ppt} + \beta_2 \text{sand} + \beta_3 \text{clay} + \beta_4 \text{om} + \beta_5 \text{spH} + \epsilon$$

First, we need to filter the dataset to MN and WI

```
counties_mn_wi = counties %>%
  filter(state %in% c("MN", "WI"))

head(counties_mn_wi) %>%
  kableExtra::kbl()
```

county	state	stco	ppt	whc	sand	silt	clay	om	kwfactor	ki
Adams	WI	55001	548.2443	15.67212	78.38254	11.71815	9.899312	438.3402	0.1170651	0.11
Aitkin	MN	27001	485.5737	28.30823	47.94610	38.02355	14.030353	1760.8355	0.3114794	0.31
Anoka	MN	27003	551.9312	24.75840	75.16051	18.69882	6.140666	1787.7790	0.1505643	0.15
Ashland	WI	55003	495.9158	21.11778	43.10370	32.20112	24.695183	274.1879	0.2977456	0.31
Barron	WI	55005	528.5029	22.30522	58.10407	31.03742	10.858506	231.5082	0.2886149	0.29
Bayfield	WI	55007	490.7059	21.75064	34.08119	29.80771	36.111094	137.1594	0.2927169	0.29

To fit our model, we use the linear model *lm()* function of R. For multiple regression, we don't list any interactions between the terms.

```
model_mn_wi = lm(corn ~ ppt + sand + clay + om + sph, data=counties_mn_wi)
summary(model_mn_wi)
```

```
##
## Call:
## lm(formula = corn ~ ppt + sand + clay + om + sph, data = counties_mn_wi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -42.432  -4.342   1.108   7.035  23.278 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -189.02125  22.76727 -8.302 5.52e-14 ***
## ppt          0.37140   0.02165 17.159 < 2e-16 ***
## sand         -0.33075  0.10746 -3.078  0.00248 **  
## clay         -1.16506  0.24579 -4.740 4.92e-06 ***
## om           -0.01679  0.00307 -5.469 1.85e-07 ***
## sph          24.21303  2.02991 11.928 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.61 on 150 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.7651, Adjusted R-squared:  0.7573 
## F-statistic: 97.71 on 5 and 150 DF,  p-value: < 2.2e-16
```

Let's go through the results. The first output item, the Call, is the linear model we gave to R.

The second item, residuals, describes the distribution of residuals around the regression model. We can see the minimum residual is 42 bu/acre below the predicted value. The maximum residual is 23 bu/acre above the predicted value. The middle 50% (between 1Q and 3Q) were between 4.3 below and 7.0 above the predicted values.

The Coefficients table is the highlight of our output. The table shows the estimated slopes $\beta_1, \beta_2, \beta_3 \dots \beta_n$ associated with each environmental factor, plus the individual t-tests of their significance. We can see that each factor of our model is significant.

In the bottom, we see that three observations were deleted due to missingness (they didn't have a value for each factor). Two R^2 's are presented. The multiple R^2 is the proportion of the total variance in the model explained by our regression model. This is the same concept as for simple linear regression. Our value is 0.77, which is pretty good for an environmental model like this, especially because we did not include any temperature data.

11.1.2 Beware of Bloated Models

Multiple Linear Regression is a powerful tool, but it generally pays to be conservative in how many factors you include in a model. The more terms you include, the more you are likely to encounter problems with overfitting, multicollinearity, and heteroscedasticity.

11.1.2.1 Overfitting

Every model contains fixed terms (which the model is meant to measure and predict) and random terms, such as error, which are beyond the scope of the model to predict. Not only that: since the error effect is random, it would be wrong for a model to try to predict it. If we add enough factors to a model, however, it will do exactly that. It will *overfit* the data. The problem with overfitting is that the model contorts itself to fit every nook and cranny of one dataset – but fails to accurately predict values for additional datasets that, randomly, have different error structures.

Here is an analogy. You let your spouse or best friend borrow your car. When they return it, the seating settings are completely messed up. The lumbar support is either poking you in your but or about to break your neck. You either can't reach the pedals and the steering wheel, or the seat is up so far you cannot even get in the car. In addition, the climate control is too hot or too cold. And, seeing what they left on the radio, you start to rethink the whole relationship.

This is exactly the problem with overfitting. The more we perfect the fit of a model one dataset, the more unlikely it is to make accurate predictions for another dataset.

Adding factors to a model will always increase the R^2 value, even if the new factor has nothing to do with what the model is predicting. For fun, lets create a column that randomly assigns an average number of Grateful Dead concerts attended per person, from 1 to 40, to each county.

county	state	stco	ppt	whc	sand	silt	clay	om	kwfactor	kd
Adams	WI	55001	548.2443	15.67212	78.38254	11.71815	9.899312	438.3402	0.1170651	0.11
Aitkin	MN	27001	485.5737	28.30823	47.94610	38.02355	14.030353	1760.8355	0.3114794	0.31
Anoka	MN	27003	551.9312	24.75840	75.16051	18.69882	6.140666	1787.7790	0.1505643	0.15
Ashland	WI	55003	495.9158	21.11778	43.10370	32.20112	24.695183	274.1879	0.2977456	0.31
Barron	WI	55005	528.5029	22.30522	58.10407	31.03742	10.858506	231.5082	0.2886149	0.29
Bayfield	WI	55007	490.7059	21.75064	34.08119	29.80771	36.111094	137.1594	0.2927169	0.29

And then let's check out our model:

```
##
## Call:
## lm(formula = corn ~ ppt + sand + clay + om + sph + gd_concerts,
##      data = counties_mn_wi_gd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -41.881  -4.347   1.033   6.632  23.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.897e+02  2.284e+01 -8.304 5.66e-14 ***
## ppt          3.707e-01  2.173e-02 17.059 < 2e-16 ***
## sand         -3.334e-01  1.078e-01 -3.093  0.00236 **
## clay         -1.169e+00  2.464e-01 -4.745 4.84e-06 ***
## om           -1.690e-02  3.082e-03 -5.483 1.75e-07 ***
## sph          2.426e+01  2.036e+00 11.917 < 2e-16 ***
## gd_concerts  4.393e-02  7.361e-02   0.597  0.55155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.64 on 149 degrees of freedom
## (3 observations deleted due to missingness)
## Multiple R-squared:  0.7656, Adjusted R-squared:  0.7562
## F-statistic: 81.13 on 6 and 149 DF,  p-value: < 2.2e-16
```

See? Our Multiple R^2 increased slightly from 0.765 to 0.766. While this example is absurd, it points to a temptation for statisticians and data scientists: keep

adding terms until the *MultipleR*² is an accepted value. Less nefariously, the researcher may just think the more variables the researcher can add, the better. We can now see this is wrong.

11.1.2.2 Multicollinearity

Another problem that can occur as we add factors to our model is *multicollinearity*. Our linear model assumes that each factor has an *independent* effect on the response variable. We know, however, that is not always the case. For example, all sorts of weather and soil factors in our model can be related. Cloudy days associated with precipitation may decrease temperature. Warm temperatures may reduce soil moisture. Areas with greater precipitation will tend to have more woody vegetation, leading to lower soil pH and soil organic matter. Coarser soils, which are better drained, will tend to have lower soil organic matter.

Here is a hypothetical scenario: suppose we look at the regression of yield on precipitation and pH and conclude yield is significantly affected by both of them. Do we know that soil pH caused the change in yield? No, it's possible that precipitation affected both pH and yield so that they appeared to change together. That's not to say we can't include all of these factors in our model. But we need to make certain that they are directly causing changes *to* the response variable, rather than responding *with* the response variable to a third variable in the model.

11.1.2.3 Heteroscedasticity

The third problem we can have with a multiple linear regression model is that it is *heteroscedastic*. This intimidating term refers to unequal variances in our model. That is, the variance of observed yields varies markedly with the value being predicted. Multiple linear regression, like other linear models, assumes that the variance will be the same along all levels of the factors in the model. When heteroscedasticity – unequal variances – occurs, it jeopardizes our ability to fit the that factor. Our least squared estimate for that factor will be more influenced by factor levels with greater variances and less influences by levels with lesser variances.

One of the causes of heteroscedasticity is having many predictors – each with their own scales of measure and magnitudes – in a model. Since linear regression relies on least squares – that is, minimizing the differences between observed and predicted values – it will be more influenced by factors with greater variances than factors with small variances, regardless of how strongly each factor is correlated with yield.

The end result is that the regression model will fit the data more poorly. Its error will be greater and, thus, its F-value and p-value will be reduced. This may lead us to conclude a factor or the overall model does not explain a significant

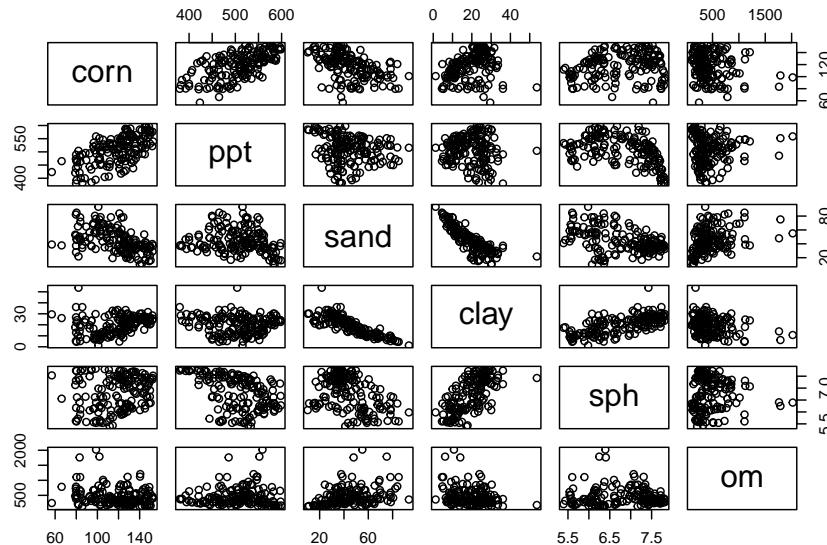
proportion of the variation in data, when in fact it does. In other words, we may commit at Type II error.

11.1.3 Methods for Avoiding Bloated Models

There are multiple approaches, and surely books written, about how to avoid overfitting and multicollinearity in models. After all that, also, model tuning (selecting the factors to include) seems to be art as much as science. This section provides an overview of methods used, which you might consider if presented with someone elses results, or if you are trying to construct a model with a few factors on your own.

11.1.3.1 Create a Covariance Matrix

A matrix (not “the matrix”) is a mathematical term that describes what you and I would call a table. So a covariance matrix is a table composed of correlation plots that we can use to inspect the covariance (or relationship) between each possible pair of variables in our dataset.



To use the matrix, find the intersection between a column and a row containing two variables whose relationship you want to inspect. For example, in the fourth column, clay is plotted in relationship to the four other predictor variables, plus the response corn, in our model. In each plot, clay is on the X-axis, and the intersecting variable is on the Y-axis. Looking at the matrix, we notice the

relationship of clay with sand and soil pH (sph) is visible. We may wonder, then, if the addition of clay to our model is improving our prediction.

11.1.3.2 Partial Correlation

Another thing we can do to look for multicollinearity is to calculate the partial correlations. Partial correlation shows the individual correlations between variables, with all other variables being held constant. What this does is allow us to quantify the correlation between two variables without worrying that both may be affected by a third variable. For example, we can look at the correlation between soil pH and soil organic matter without worrying that precipitation might be driving changes in both we could mistake for a relationship.

	corn	ppt	sand	clay	sph	om
corn	1.0000000	0.8139276	-0.2437413	-0.3609330	0.6977065	-0.4077095
ppt	0.8139276	1.0000000	-0.0390449	0.2401821	-0.7798037	0.4618512
sand	-0.2437413	-0.0390449	1.0000000	-0.7743363	0.1767851	0.1169185
clay	-0.3609330	0.2401821	-0.7743363	1.0000000	0.5772645	-0.1893945
sph	0.6977065	-0.7798037	0.1767851	0.5772645	1.0000000	0.4490715
om	-0.4077095	0.4618512	0.1169185	-0.1893945	0.4490715	1.0000000

The output of partial correlation is a matrix, which cross-tabulates the correlations among every predictor variables and reports their values in a table. The output above tells us that sand and clay are both negatively correlated with corn. That's odd – we would expect that as sand decreases, clay would increase, or vice versa – unless, both are being impacted by a third value, silt, which is not in our model.

11.1.3.3 Cross Validation

A third way to evaluate the performance of a model and to avoid mistaking over-prediction for true model performance is to use cross-validation. I'll confess to graduating from Iowa State and making it another 15 years in academia and industry without having a clue about how important this is. If you plan to use your model not only to test hypotheses about variable significance, but to make predictions, the cross-validation is critical.

In cross-validation, the initial data are divided into *training* and *testing* groups. The model parameters (coefficients for slopes) are solved for using the training dataset. In general, all models will better fit the data used to train them. Therefore, the predictive performance of the model is measured using the testing dataset. In this way, the true performance of the model can be measured. In addition, the effect of adding or removing factors may be measured.

We will use a technique that sounds vaguely like an adult-alternative music group – Repeated 10-Fold Cross Validation. While the name sounds scary as all get-out, how it works is (pretty) straight forward:

1. Divide the data into 10 groups.
2. Randomly select 9 groups and fit the regression model to them. These groups are called the “training” dataset
3. Predict the responses for each observation in the tenth dataset, using the model fit to the other 9 datasets. This 10th dataset is called the “testing” dataset.
4. Use linear regression to determine the strength of the relationship between the predicted value and the actual values. Review summary statistics in comparing models.

Here is the cross validation of our first yield model for Illinois and Ohio:

```
## Linear Regression
##
## 156 samples
##   5 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 140, 140, 141, 140, 140, ...
## Resampling results:
##
##   RMSE     Rsquared    MAE
##   10.73952  0.7632216  8.203451
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We can see in the top of the output confirmation we are working with the original 5 predictors. To measure model performance, let’s look particularly at the MRSE and Rsquared statistics at the bottom. RMSE is the Root Mean Square Error, which you earlier learned is the square root of the Mean Square Error, and equivalent to the standard deviation of our data. This is expressed in our regression output above as residual standard error. An RMSE of 10.7 means the distribution of residuals (observed values) around our model predictions has a standard deviation of about 10.7. Thus 95% of our observed values would be expected to be within 21.4 bushels ($2 * 10.7$) of the predicted value.

The Rsquared statistic is the same as we have seen previously, and describes the amount of variation in the observed values explained by the predicted values.

11.1.4 Tuning and Comparing Models

Here is our original model for reference. Note that sand has a negative effect (-0.33075) on yield and that it is highly significant ($p=0.00248$).

```
## 
## Call:
## lm(formula = corn ~ ppt + sand + clay + om + sph, data = counties_mn_wi)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -42.432  -4.342   1.108   7.035  23.278 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -189.02125   22.76727 -8.302 5.52e-14 ***
## ppt          0.37140    0.02165  17.159 < 2e-16 ***
## sand         -0.33075   0.10746 -3.078  0.00248 **  
## clay         -1.16506   0.24579 -4.740 4.92e-06 ***
## om           -0.01679   0.00307 -5.469 1.85e-07 *** 
## sph          24.21303   2.02991 11.928 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.61 on 150 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.7651, Adjusted R-squared:  0.7573 
## F-statistic: 97.71 on 5 and 150 DF,  p-value: < 2.2e-16
```

We saw above that clay was strongly correlated with both sand and soil pH. Let's drop clay from our model and see what happens:

```
## 
## Call:
## lm(formula = corn ~ ppt + sand + om + sph, data = counties_mn_wi)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -50.342 -5.438   1.410   7.135  22.276 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.031e+02  2.412e+01 -8.421 2.69e-14 ***
## ppt          3.815e-01  2.302e-02 16.575 < 2e-16 ***
## sand         5.577e-02  7.479e-02   0.746   0.457
```

```

##   om      -1.607e-02  3.277e-03 -4.903 2.41e-06 ***
##   sph      1.953e+01  1.895e+00 10.306 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.34 on 151 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.7299, Adjusted R-squared:  0.7227
## F-statistic:  102 on 4 and 151 DF,  p-value: < 2.2e-16

```

First, let's look at the model fit. The Multiple R-squared decreased from 0.7651 to 0.7299. The residual standard error increased from 10.61 to 11.34. These would suggest the fit of the model has decreased, although we also notice the F-statistic has increased from 97.7 to 102, which suggests the model itself is more strongly detecting the combined effects of the factors on yield,

An additional statistic that we want to watch is the Adjusted R-Squared. This statistic not only takes into effect the percentage of the variation explained by the model, but how many factors were used to explain that variance. The model is penalized for according to the number of factors used: of two models that explained the same amount of variation, the one that used more factors would have a lower Adjusted R-square. We see the adjusted R-square decreased from our first model to the second.

Now let's go back to sand. In the first model, it had a negative effect of -0.33075 and was highly significant. Now it has a positive effect of 5.577e-02 and an insignificant effect on corn yield. Remember, each factor in a linear regression model should be independent of the other factors. If we compare the other four factors, we will see their coefficients have changed slightly, but they have remained highly significant. This suggests that clay was affecting both sand content (after all, if you have more clay, you are likely to have less sand) and yield.

Let's cross-validate the new model.

```

## Linear Regression
##
## 156 samples
##   4 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 140, 140, 141, 140, 140, ...
## Resampling results:
##
##   RMSE      Rsquared     MAE
##   11.18499  0.7413634  8.347023

```

210CHAPTER 11. NONLINEAR RELATIONSHIPS AND MULTIPLE LINEAR REGRESSION

```
##  
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We see the Rsquared and RMSE (root mean square error) are statistics are slightly lower than the original model in the cross-validation, too.

Since sand was insignificant in the second model, let's remove it and rerun our model.

```
##  
## Call:  
## lm(formula = corn ~ ppt + om + sph, data = counties_mn_wi)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -50.786  -5.803   1.263   6.637  22.088  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.904e+02  1.702e+01 -11.187 < 2e-16 ***  
## ppt          3.724e-01  1.943e-02  19.167 < 2e-16 ***  
## om           -1.501e-02  2.949e-03 -5.089 1.05e-06 ***  
## sph          1.863e+01  1.465e+00 12.720 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 11.32 on 152 degrees of freedom  
##   (3 observations deleted due to missingness)  
## Multiple R-squared:  0.7289, Adjusted R-squared:  0.7236  
## F-statistic: 136.2 on 3 and 152 DF,  p-value: < 2.2e-16
```

We see little change in the Multiple R-squared or Adjusted R-squared, but the F-statistic has again increased. Let's cross-validate the model.

```
## Linear Regression  
##  
## 156 samples  
##   3 predictor  
##  
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)  
## Summary of sample sizes: 140, 140, 141, 140, 140, 140, ...  
## Resampling results:  
##  
##   RMSE      Rsquared     MAE
```

```
##   11.15776  0.7424514  8.284529
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

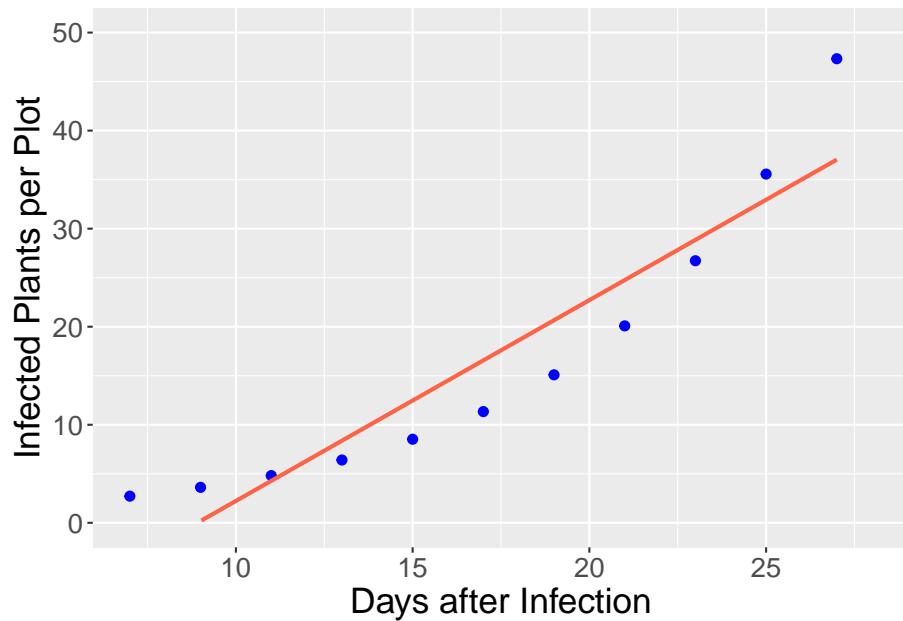
We see the new model fits slightly better.

We could go on and on with the process. We might try adding silt into the model to replace the sand and clay we removed. Water holding capacity was a factor in the original dataset – we might try using that as a proxy, too. But the iterations we have gone through show us that bigger models are not necessarily better (not by much, in any case). While we can build complex models with multiple linear regression, it is better not to when possible.

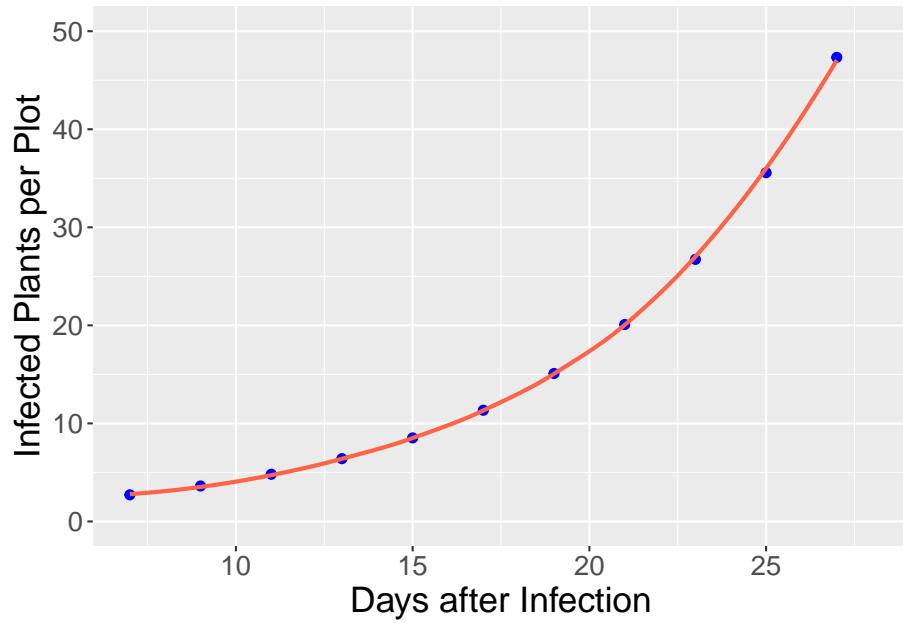
11.2 Nonlinear Relationships

As mentioned in the introduction, there are many relationships between variables that are nonlinear – that is, cannot be modelled with a straight line. In reality, few relationships in agronomy are perfectly linear, so by nonlinear we mean relationships where a linear model would systematically over-predict or underpredict the response variable. In the plot below, a disease population (infected plants per plot) is modeled as a function of days since infection.

As with many pest populations, the number of infections increases exponentially with time. We can see how a linear model would underpredict the number of plants from 7 to 9 days after infection and again from 25 to 27 days after infection, while overpredicting the number of infected plants from 15 to 21 days after infection. Systematic overpredictions or underpredictions are called *bias*.



Instead, we can fit the data with an exponential model that reduces the bias and increases the precision of our model:



11.2.1 Fitting Nonlinear Responses with Linear Regression

Fitting a relationship with a simple linear regression model is simpler than fitting it with a nonlinear model, which we will soon see. Exponential relationships, like the one above can be fit by *transforming* the data to a new scale. This is the same concept as we used in an earlier unit to work with messy data.

11.2.1.1 Exponential Model

For example, the data above (and many exponential relationships in biology and other disciplines can be transformed using the natural log:

$$y = \log(x)$$

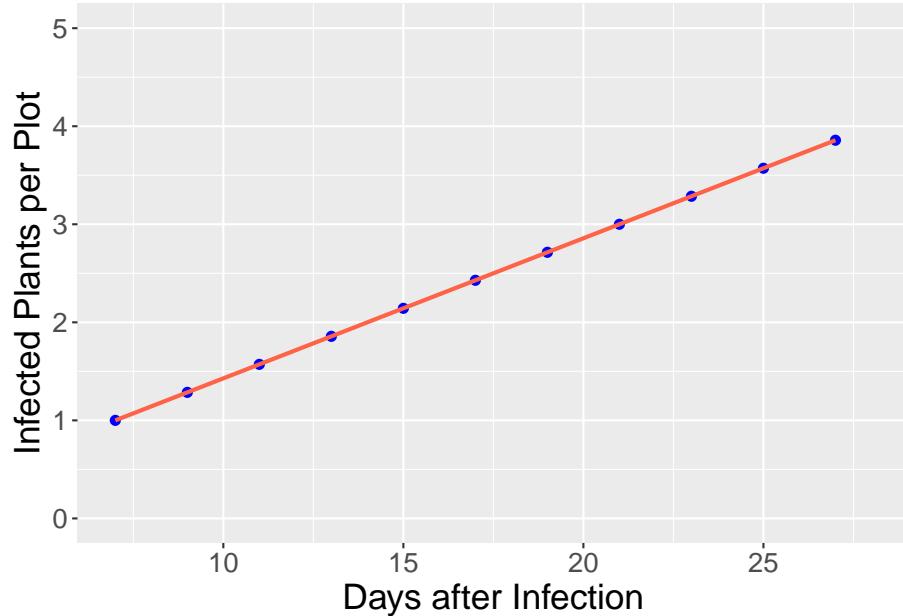
If our original dataset looks like:

x	y
7	2.718282
9	3.617251
11	4.813520
13	6.405409
15	8.523756
17	11.342667
19	15.093825
21	20.085537
23	26.728069
25	35.567367
27	47.329930

Where x is the days since infection and y is the mean number of infections per plot. We can create a new column, `log_y` with the natural log of infected plants.

x	y	log_y
7	2.718282	1.000000
9	3.617251	1.285714
11	4.813520	1.571429
13	6.405409	1.857143
15	8.523756	2.142857
17	11.342667	2.428571
19	15.093825	2.714286
21	20.085537	3.000000
23	26.728069	3.285714
25	35.567367	3.571429
27	47.329930	3.857143

When we fit the log of the infected plants, we see the relationship between the number of infected plants and days since infection is now linear.



We can now fit a linear model to the relationship between infected plants and days after infection.

```
## 
## Call:
## lm(formula = log_y ~ x, data = exponential_final)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -4.375e-16 -3.117e-16 -3.760e-18  2.469e-16  5.395e-16 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.142e-15  3.108e-16 6.893e+00 7.12e-05 ***
## x           1.429e-01  1.713e-17 8.337e+15 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 3.594e-16 on 9 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:      1 
## F-statistic: 6.951e+31 on 1 and 9 DF,  p-value: < 2.2e-16
```

Our linear model, from the equation above, is:

$$\log_y = 0 + 0.1429 \times x$$

Just to test the model, let's predict the number of infected plants when $x = 15$

```
## [1] "log_y = 2.1435"
```

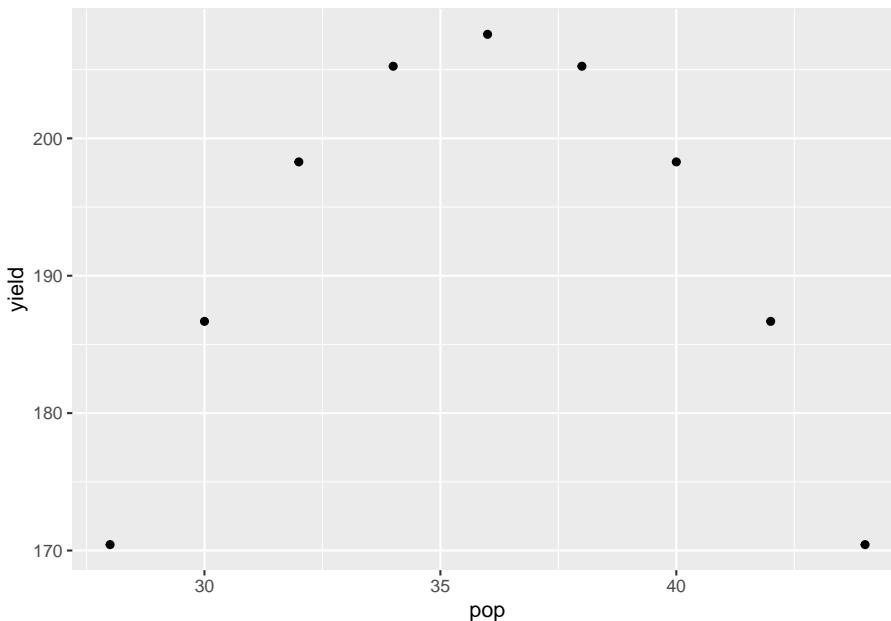
We can see this value is approximately the number of infected plants 15 days after infection in the table above. Any predicted value can be transformed from the logarithm back to the original scale using the `exp()` function. Compare this with the original count, y , in the table above.

```
## [1] "y = 8.52923778043145"
```

11.2.1.2 Parabolic

In nutrient availability and plant density models, we sometimes encounter data that are parabolic – the relationship between Y and X resembles a \cap or \cup shape. These data can be fit with a *quadratic model*. Remember that beast from eighth grade algebra. Don't worry – we don't have decompose it!

Let's say we have data from a plant density study in corn:



The quadratic model is:

$$Y = \alpha + \beta X + \gamma X^2$$

Where α is the Y-intercept, and β and γ are the coefficients associated with X and X^2 . We can run this model the same as we would a simple regression.

```
## 
## Call:
## lm(formula = yield ~ pop + pop2, data = plant_density_data)
## 
## Residuals:
##       Min      1Q  Median      3Q     Max 
## -8.565e-14 -3.115e-14 -7.869e-15  4.033e-14  7.242e-14 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -5.446e+02  1.043e-12 -5.223e+14 <2e-16 ***
## pop          4.179e+01  5.877e-14  7.111e+14 <2e-16 ***  
## pop2         -5.804e-01  8.146e-16 -7.125e+14 <2e-16 ***  
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 5.718e-14 on 6 degrees of freedom
## Multiple R-squared:      1, Adjusted R-squared:      1 
## F-statistic: 2.538e+29 on 2 and 6 DF,  p-value: < 2.2e-16
```

This output should look very similar to the simple linear regression output. The only difference is that there are now three coefficients returned: the intercept (α above), the coefficient for pop (β above), and the coefficient for pop2 (γ above).

11.2.2 Fitting Nonlinear Responses with Nonlinear Regression

Other nonlinear relationships must be fit with nonlinear regression. Nonlinear regression differs from linear regression in a couple of ways. First, a nonlinear regression model may include multiple coefficients, but only X as a predictor variable. Second, models are not fit to nonlinear data using the same approach (using least square means to solve for slope) as with linear data. Models are often fit to nonlinear data often do so using a “trial and error” approach, comparing multiple models before converging on the model that fits best. To help this process along, data scientists must often “guesstimate” the initial values of model parameters and include that in the code.

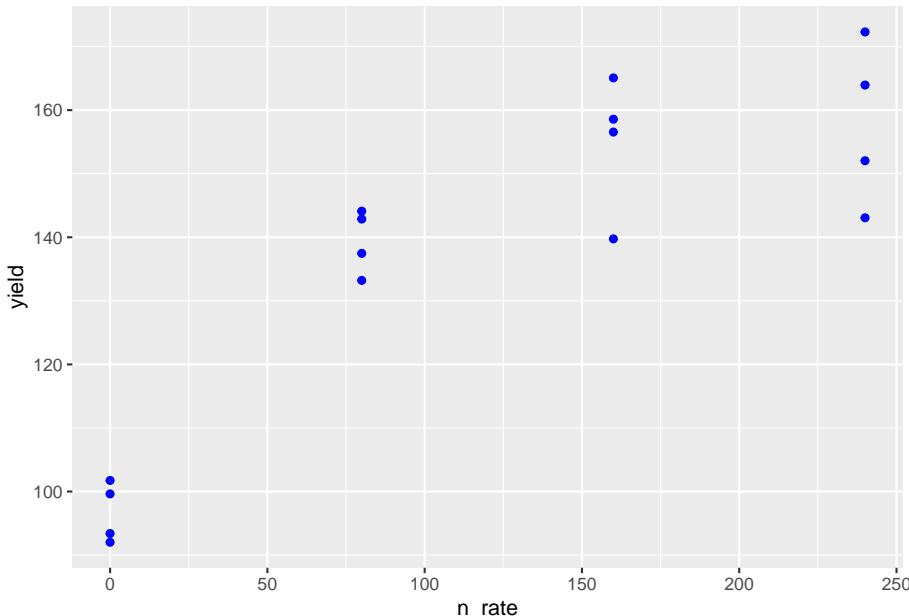
11.2.2.1 Monomolecular

In the Monomolecular growth model, the response variable Y initially increases rapidly with increases in X. Then the rate of increase slows, until Y plateaus and does not increase further with X. In the example below, the response of corn yield to nitrogen fertilization rate is modelled with the monomolecular (asymptotic) function.

First, we load and plot the data.

```
corn_n_mono = read.csv("data-unit-11/corn_nrate_mono.csv")
p = corn_n_mono %>%
  ggplot(aes(x=n_rate, y=yield)) +
  geom_point(color="blue")

p
```



Then we fit our nonlinear model.

```
corn_n_mono_asym = stats::nls(yield ~ SSasymp(n_rate, init, m, plateau), data=corn_n_mono)
summary(corn_n_mono_asym)

##
## Formula: yield ~ SSasymp(n_rate, init, m, plateau)
##
```

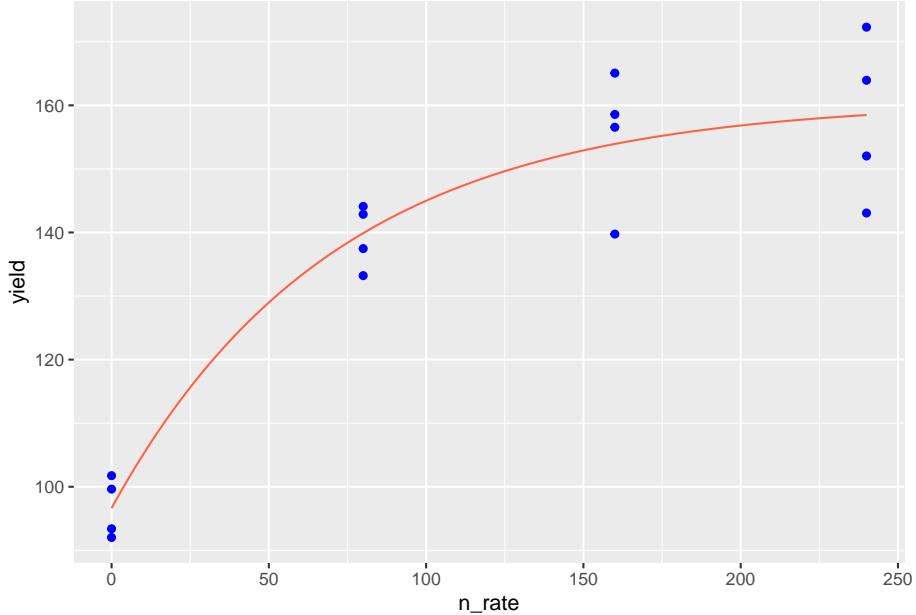
```

## Parameters:
##             Estimate Std. Error t value Pr(>|t|)
## init      160.6633    5.7363 28.01 5.24e-13 ***
## m         96.6297    4.3741 22.09 1.08e-11 ***
## plateau   -4.2634    0.3019 -14.12 2.90e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.76 on 13 degrees of freedom
##
## Number of iterations to convergence: 0
## Achieved convergence tolerance: 6.133e-08

```

The most important part of this output is the bottom line, “Achieved convergence tolerance”. That means our model successfully fit the data.

Finally, we can add our modelled curve to our initial plot:

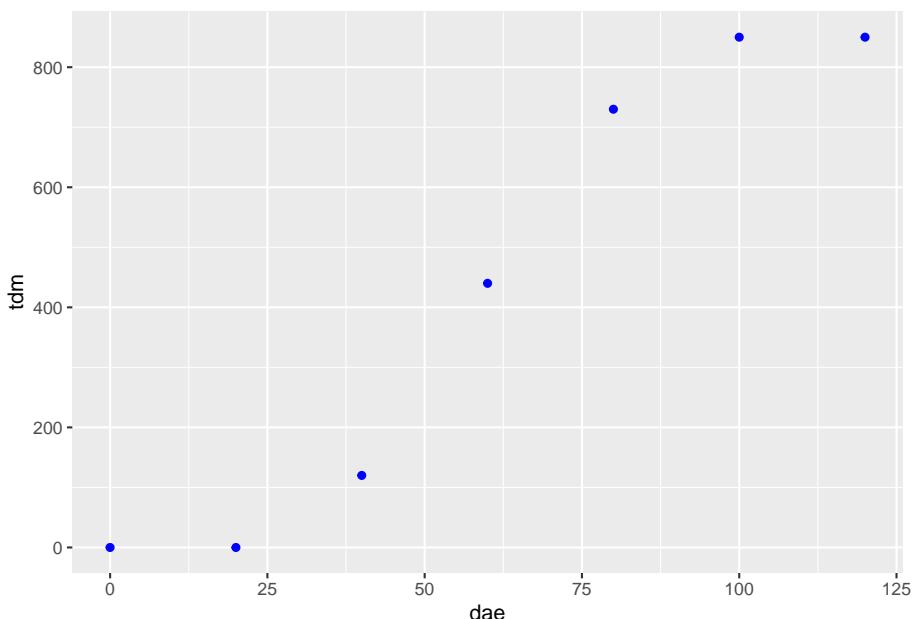


In the plot above, the points represent observed values, while the red line represents the values predicted by the monomolecular model. The monomolecular function is often used to represent fertilizer rate or soil nutrient content, since the response to many fertilizers plateaus or “maxes out”. Fertilizer rate recommendations are developed to increase fertilization up to the point where the cost of adding another unit of fertilizer exceeds the benefit of the increase in yield. The monomolecular function can also measure other “diminishing returns” responses, such as the response of photosynthesis to increasing leaf area.

11.2.2.2 Logistic Model

The *Logistic* model is often used in “growth analysis”, studies that highlight patterns of plant growth, particularly cumulative biomass accumulation over time. Data generally follow a “sigmoidal”, or S-shaped, pattern. In early growth the rate of biomass accumulation slowly increases. In the intermediate growth phase, biomass accumulation is rapid and linear. In the final growth phase, the rate of growth decreases and, if the trial is long enough, may plateau.

In the plot below, total dry matter accumulation (tdm) is shown in relation to days after emergence (dae).



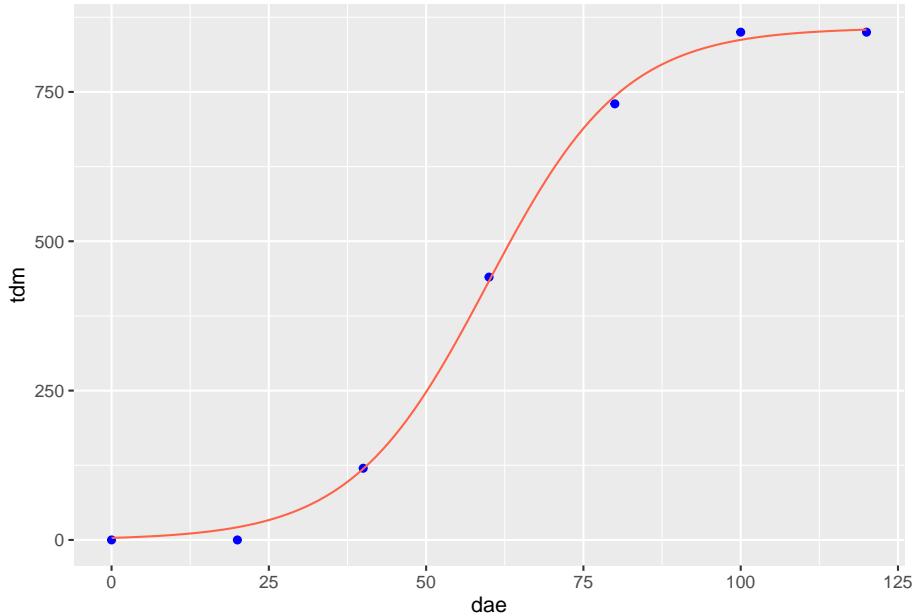
The “S” shape of the data is very pronounced. Next, we fit a logistic growth curve to the data:

```
##  
## Formula: tdm ~ SSlogis(dae, Asym, xmld, scal)  
##  
## Parameters:  
##      Estimate Std. Error t value Pr(>|t|)  
## Asym 857.5043   11.7558  72.94 2.12e-07 ***  
## xmld  59.7606    0.7525  79.42 1.51e-07 ***  
## scal  10.8261    0.6574  16.47 7.96e-05 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##
```

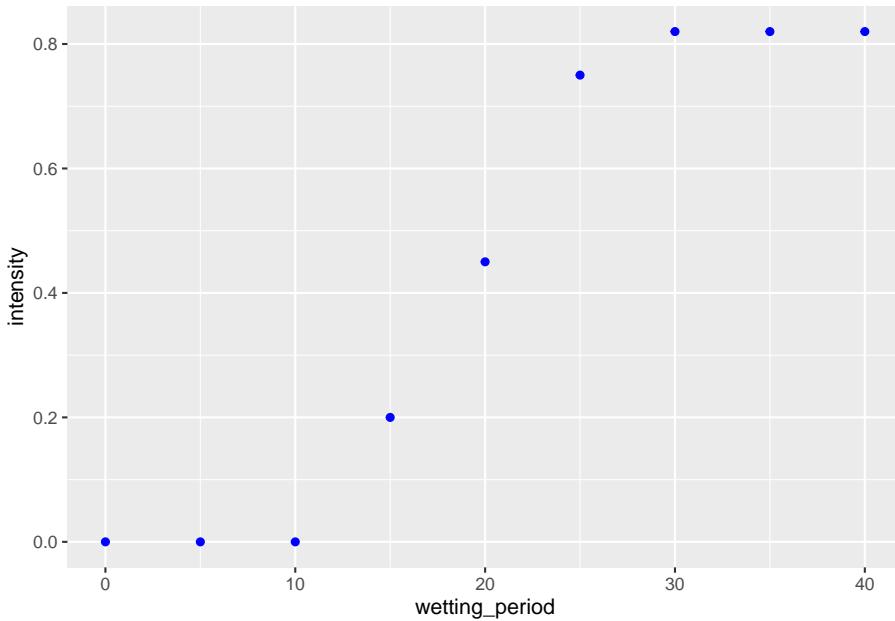
```
## Residual standard error: 14.64 on 4 degrees of freedom
##
## Number of iterations to convergence: 0
## Achieved convergence tolerance: 3.029e-06
```

We see again in the output that the model achieved convergence tolerance. Another thing to note is the “Number of iterations to convergence”. It took this model 9 steps to fit the data. The algorithm will typically quit after 50 unsuccessful attempts. When that occurs, it may be an indication the data should be fit with a different model.

Here is our data with the fitted prediction model:



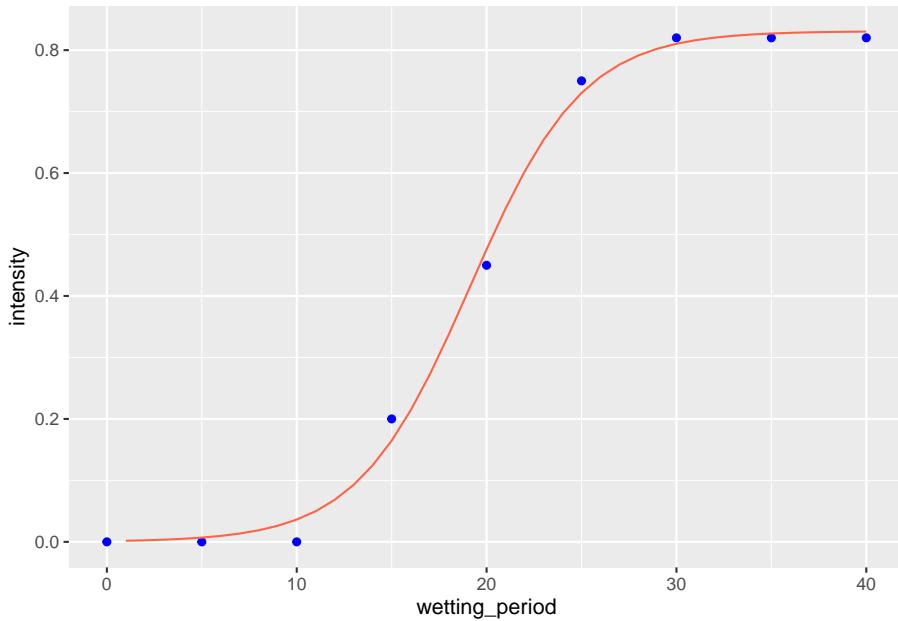
We can see this again illustrated in the next plot, which illustrates the effect of wetting period on the intensity of wheat blast.



We fit the data with the logistic model.

```
## 
## Formula: intensity ~ SSlogis(wetting_period, Asym, xmid, scal)
## 
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
##   Asym  0.83092   0.01606  51.74 3.50e-09 ***
##   xmid 19.13425   0.33505  57.11 1.94e-09 ***
##   scal  2.95761   0.28825  10.26 5.00e-05 ***
##   ---
##   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02552 on 6 degrees of freedom
## 
## Number of iterations to convergence: 1
## Achieved convergence tolerance: 4.024e-06
```

Convergence criteria was met. Here is the data plotted with the logistic model.



11.2.2.3 Starting Variable Values

As mentioned earlier, nonlinear regression algorithms can be difficult to use in that they don't fit a regression model using the least squares approach. Instead, nonlinear regression "nudges" the different coefficients in its equation until it zeros in on coefficient values that best fit the data. Often, traditionally, algorithms have not been able to solve for these coefficients from scratch. The user would have to provide initial coefficient values and hope they were close enough to true values that the nonlinear algorithm could then fit the curve.

What a pain in the butt. While some coefficients could be easily guessed, others required more complex calculations. With R, some packages have introduced "self-starting" non-linear models that do not require the user to enter `initial` values for coefficients. I have used those to create the examples above and we will cover them in the exercises. Just be aware, were you to get into more intensive nonlinear modelling, that you may need to specify those variables.

11.3 Summary

Multivariate models (those that model yield or another response to multiple variables) are very powerful tools for unlocking the mysteries of how changes in environment drive what we observe in our research. However, they require skill to use. Remember that more variables do not automatically make a better

model. Too many variables can cause our model to overfit our original data, but be less accurate or unbiased in fitting future datasets. Ask yourself whether it makes sense to include each variable in a model. Covariance matrixes and partial correlations can help us identify predictor variables that may be correlated with each other instead of the response variable.

Cross validation is also an important tool in assessing how a model will work with future data. In this unit, we learned a common practice, 10-fold cross validation, in which the data were divided into 10 groups. Each group took turns being part of the datasets used to train the model, and the dataset used to test it.

Finally, nonlinear regression is used to fit variables that have a nonlinear relationship. Unlike multiple linear regression, there is usually only one predictor variable in non-linear regression. In addition, nonlinear regression is often based on a theoretical relationship between the predictor and response variable. In this unit, we focused on two models: the monomolecular model for relationships between yield and soil fertility, and the logistic model for predicting growth of plants and other organisms.

Chapter 12

Spatial Statistics

One of the most powerful ways I use and present data is to explain spatial patterns in our data. How does a product perform in Ohio versus Iowa? What might be the underlying weather or soil causes of these data? How do they vary with geography?

Quantitative data and hard numbers are fantastic. But as we have already seen with plots, visualizations can be much more engaging. Our minds are evolved to recognize patterns – in fact, we are so focused on looking for patterns that we need tools like statistics to keep us honest. So a statistics-based plot or map is a very powerful way to convey information to your audience or customers.

This is one of two brand-new units in Agronomy 513. You might not think a statistics software like R might be equipped to work with spatial data, especially after spending the first 11 weeks working with some ugly code. But R can readily work with shape files and rasters (think of a fertilizer application map), both creating and analyzing them. We will learn how to overlay polygons to relate soil to yield, and how to create a application map based on gridded soil tests.

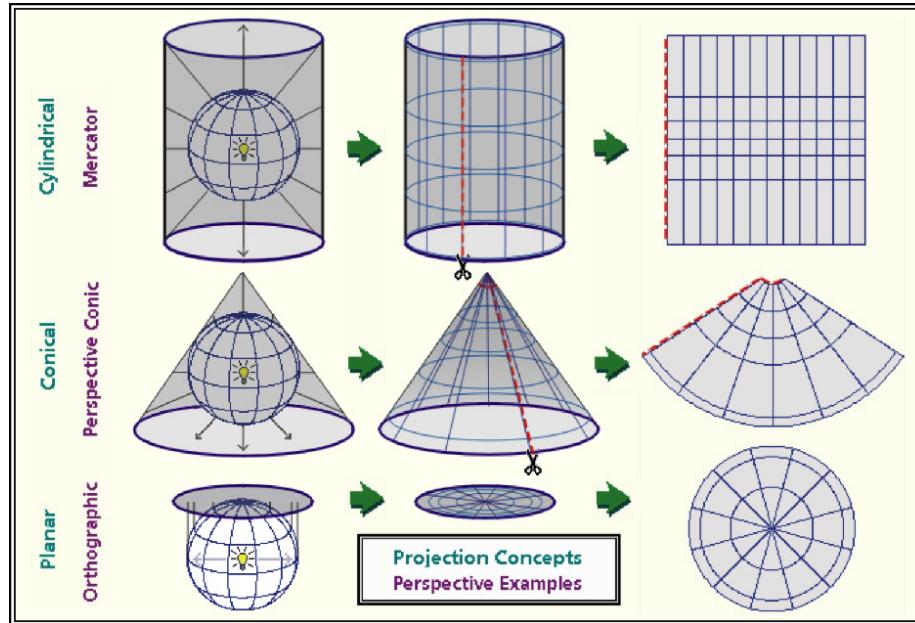
This unit will be light on calculations (yea!) and focus more on three key areas. First, what is projection and why do we need to worry about it if all we want to do is draw a map or average some data?

Second, what is a shapefile? How do we make sure it's projection is correct for our analyses and for joining with other shapefiles? How can layer the shapefile data with soil survey data readily accessible through R? How can we create an attractive map with our shapefile data?

Finally, we will study rasters. Rasters organize data in grids of cells that are of equal dimensions. Using point data from a shapefile, we can use tools like kriging to interpolate (predict) the value of each cell in the raster, creating another kind of map we can use to understand spatial data trends.

12.1 Projection (General)

One of the most challenging concepts for me when I began working with spatial data was *projection*. To be honest, it is still a challenging concept for me! Projection describes how we represent points on the surface of the earth, which is spheroidal, using maps, which are flat.



As we can see in the figure above, projection differ in how they are positioned relative to the earth's surface. Some are positioned relative to the equator, others might be centered between the equator and the poles, while yet others may be positioned at the poles. Each map will represent the center of it's geography better than the edges.

Each map is a compromise between the representation of boundaries (positions on the earth's surface) and the areas within those boundaries. Maps that pursue the accurate representation of boundaries on the earth's surface are going to end up distorting the area of geographies outside the focal point of the map. Maps that accurately represent areas are going to distort the position of geographic boundaries on the earth's surface. Thus, there are hundreds of different projection systems, focused on different areas of the earth, and using different units to describe the position of borders.

“Whoa, Marin”, you may be thinking. “I’m not trying to represent the world, the United States, Minnesota, or even my county! It’s just a freaking yield map!” And you would be absolutely correct: none of these projection systems are going to vary much in how they represent the location or area of a single section of land.



Figure 12.1: from <https://datacarpentry.org/r-raster-vector-geospatial/09-vector-when-data-dont-line-up-crs/>

But, in working with spatial data from that field, you will encounter differences among systems in how they locate your field on the face of the earth. Therefore, it is important we look at a few examples so you understand how to process those data.

We will start in the lower corner with WGS 84. This is the geographic system with which most of you are probably familiar. It is also how I roll with most of my analyses. It's simplistic, but it works just fine for point geographies – that is, single points on the earth's surface.

12.1.1 WGS 84 (EPSG: 4236)

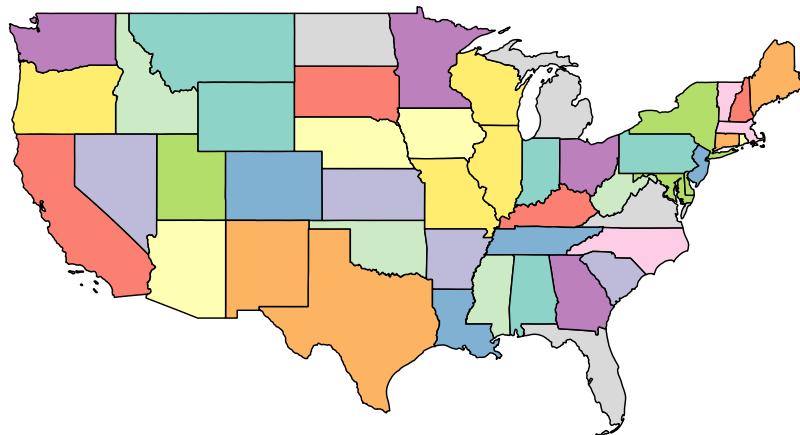
WGS 84 refers to “World Geodetic System”; 84 refers to 1984, the latest (!) revision of this system. WGS 84 uses the earth’s center as its *origin*. An origin is the reference point for any map – each location is then geo-referenced according to its position relative to the origin. In WGS 84, the position of each location is described by its angle, relative to the origin. We usually refer to these angles as degrees latitude and longitude.

EPSG (EPSG Geodetic Parameter Dataset) is a set of many, many systems used to describe the coordinates of points on the Earth’s surface. and how they are projected onto flat maps. The EPSG stands for “European Petroleum Survey

Group” – presumably, for the purpose of locating oil fields. 4326 is the code that EPSG uses to represent the WGS 84 system.

We can map the continental United states using

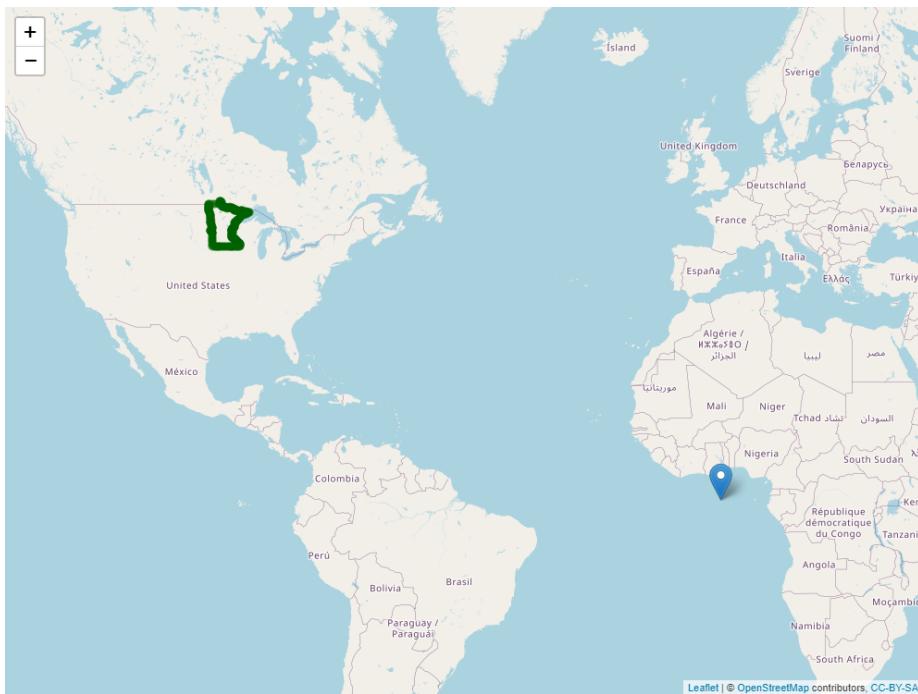
state_name



This is the flat map with which most of us are familiar. Latitude and longitude are drawn as parallel lines on this map. The map data are in a shapefile, a format we encountered at the beginning of this course. Let’s look at the top few rows of this shapefile.

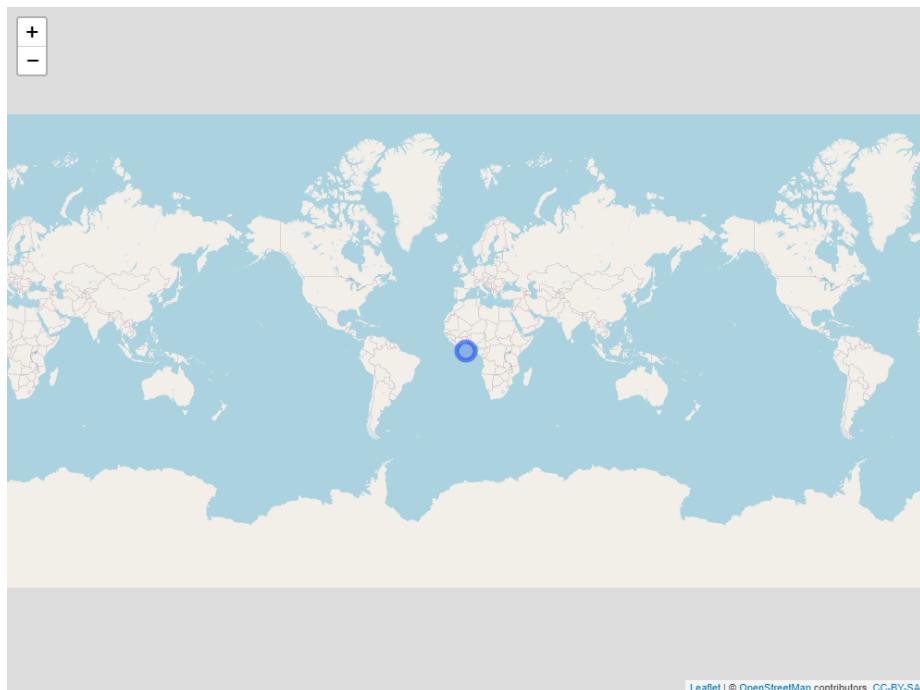
```
## Simple feature collection with 6 features and 1 field
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -124.4096 ymin: 32.53416 xmax: -86.80587 ymax: 49.38436
## Geodetic CRS: WGS 84
##   state_name           geometry
## 1 California MULTIPOLYGON (((-118.594 33...
## 2 Wisconsin MULTIPOLYGON (((-86.93428 4...
## 3 Idaho MULTIPOLYGON (((-117.243 44...
## 4 Minnesota MULTIPOLYGON (((-97.22904 4...
## 5 Iowa MULTIPOLYGON (((-96.62187 4...
## 6 Missouri MULTIPOLYGON (((-95.76564 4...
```

This is a complex dataset, so we will use the Minnesota state boundary as an example. In the map below, there are two objects. The pin in the map represents the map origin. The green dots indicate the Minnesota border.



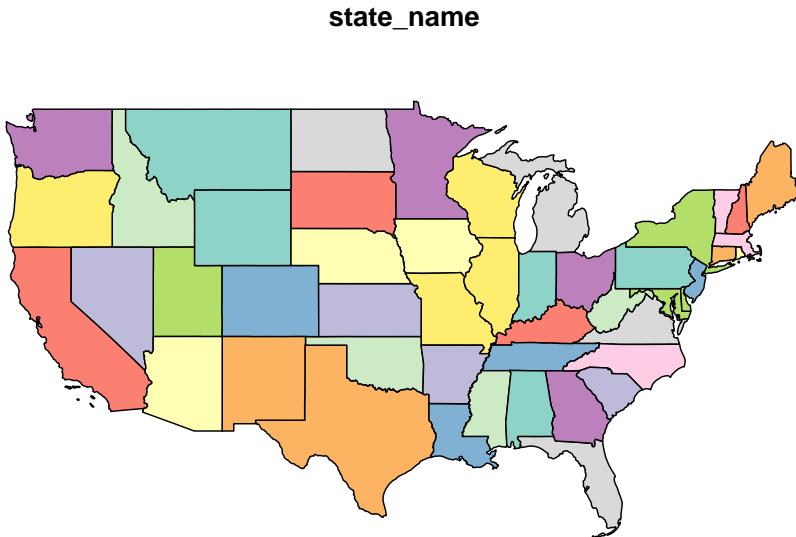
Zoom in on Minnesota and click around its borders. You will notice two things. First, each point is specified in latitude and longitude. Second, longitude (the first number) is always negative while latitude (the second number) is always positive.

The sign and size of geocoordinates in a projection system is defined two things: 1) where it places its origin (its reference point for locating objects on the map) and 2) what measurement units it uses. In the case of WGS 84, the origin is the intersection of the Prime Meridian and the Equator. Since all of the continental United States is in the western hemisphere, every state will have a negative longitude and a positive latitude. Since WGS 84 uses angles, the measurement units will be in degrees, which never exceed the range of (-180, 180) for longitude and (-90,90) for latitude.



12.1.2 Mercator (EPSG: 3857)

The Mercator System is commonly used to project data onto a sphere. If you look at the map below, it is very similar (actually related) to the WGS 84 map above but you may be able to see a slight “dome illusion” to the way the map is displayed. This projection is regularly used by online mapping services.



Looking at the top few rows of the Minnesota data points, we can see the units are not latitude and longitude. In this projection, they are easting and northing: measures of the distance east and north of the origin. Easting and northing are usually measured in meters

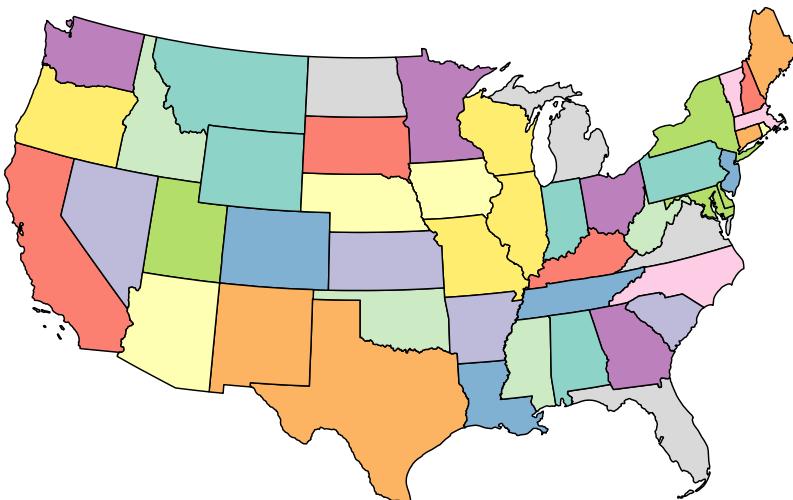
```
## Simple feature collection with 6 features and 1 field
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: -10823490 ymin: 6274648 xmax: -10610960 ymax: 6274978
## Projected CRS: WGS 84 / Pseudo-Mercator
##   state_name           geometry
## 1  Minnesota POINT (-10823487 6274978)
## 1.1 Minnesota POINT (-10790305 6274859)
## 1.2 Minnesota POINT (-10731801 6274859)
## 1.3 Minnesota POINT (-10683932 6274859)
## 1.4 Minnesota POINT (-10613307 6274648)
## 1.5 Minnesota POINT (-10610961 6274651)
```

The origin for the Mercator projection is again the intersection of Prime Meridian and Equator, so each Minnesota border point will have a negative value for easting and a positive value for northing.

12.1.3 US National Atlas Equal Area (EPSG: 2163)

As the name suggests, coordinate systems like the US National Atlas Equal Area project data so that the areas of geographic objects are accurate in the map.

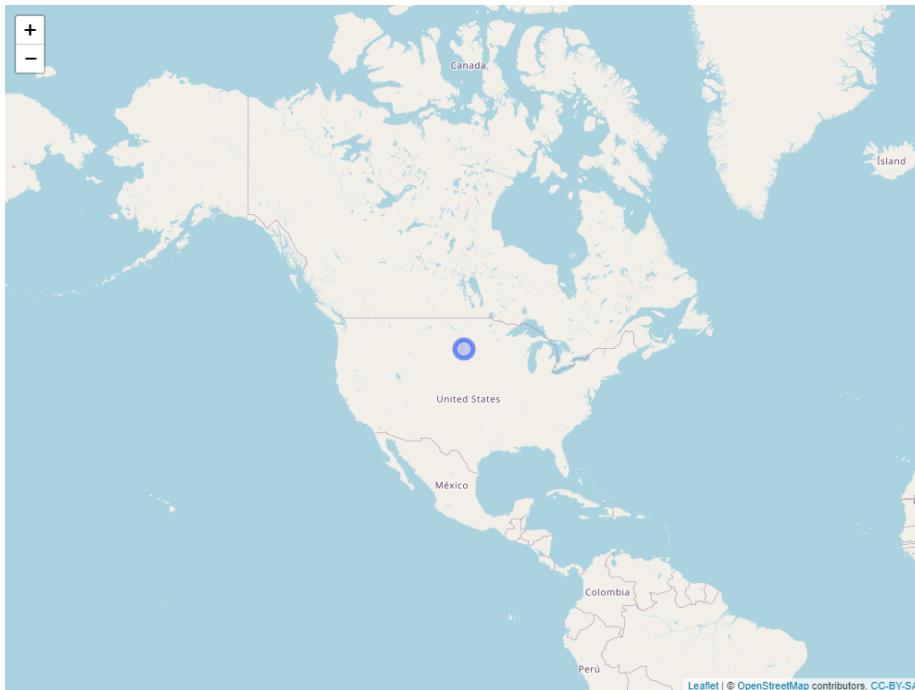
state_name



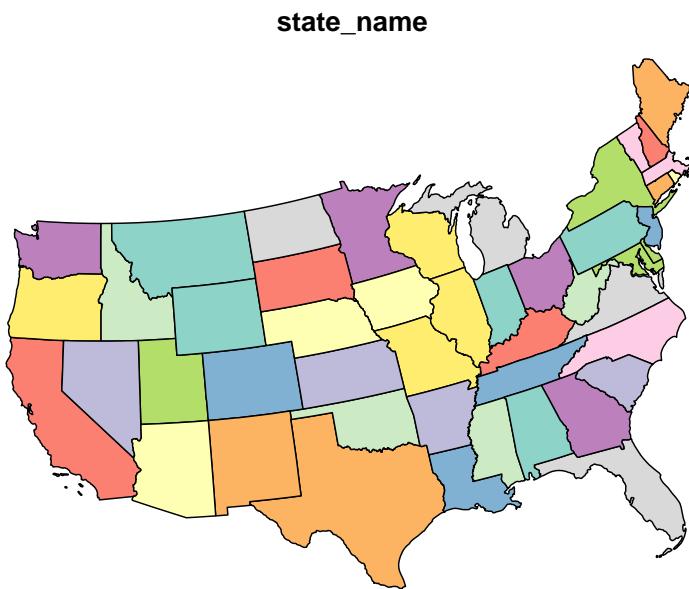
This system, like the Mercator above, uses northing and easting units. But when we look at our Minnesota border coordinates, we now notice our easting values are positive! What happened?

```
## Simple feature collection with 6 features and 1 field
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: 202876 ymin: 448171.5 xmax: 342482.3 ymax: 454473.8
## Projected CRS: NAD27 / US National Atlas Equal Area
##   state_name           geometry
## 1   Minnesota POINT (202876 448171.5)
## 1.1 Minnesota POINT (224686.3 448890.9)
## 1.2 Minnesota POINT (263125.5 450495.2)
## 1.3 Minnesota POINT (294567.2 451996.1)
## 1.4 Minnesota POINT (340942.6 454381.6)
## 1.5 Minnesota POINT (342482.3 454473.8)
```

As you have likely guessed, our origin has changed. For this projection, our origin is in Central South Dakota.



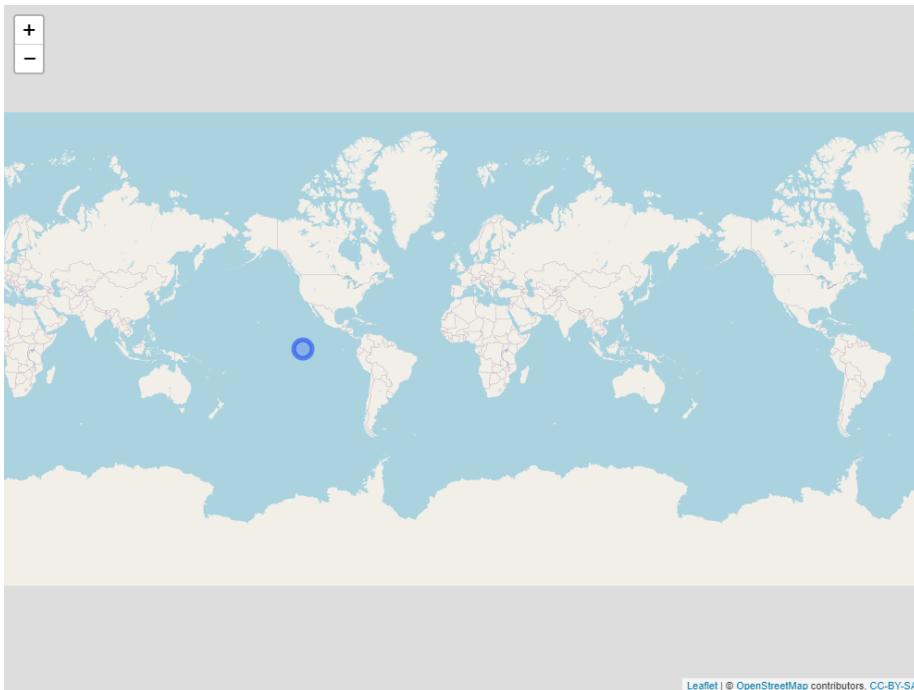
12.1.4 UTM Zone 11N (EPSG: 2955)



```
## Simple feature collection with 6 features and 1 field
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -128697 ymin: 3599652 xmax: 3013312 ymax: 5706611
## Projected CRS: NAD83(CSRS) / UTM zone 11N
##   state_name           geometry
## 1 California MULTIPOLYGON (((351881.3 37...
## 2 Wisconsin MULTIPOLYGON (((2845962 548...
## 3 Idaho MULTIPOLYGON (((480645 4915...
## 4 Minnesota MULTIPOLYGON (((1941564 561...
## 5 Iowa MULTIPOLYGON (((2169056 494...
## 6 Missouri MULTIPOLYGON (((2302630 471...
```

Here are the coordinates for the Minnesota border again.

```
## Simple feature collection with 6 features and 1 field
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: 1941564 ymin: 5618787 xmax: 2079676 ymax: 5658008
## Projected CRS: NAD83(CSRS) / UTM zone 11N
##   state_name           geometry
## 1   Minnesota POINT (1941564 5618787)
## 1.1  Minnesota POINT (1963162 5624612)
## 1.2  Minnesota POINT (2001185 5635247)
## 1.3  Minnesota POINT (2032277 5644167)
## 1.4  Minnesota POINT (2078155 5657549)
## 1.5  Minnesota POINT (2079676 5658008)
```



12.1.5 Projection Summary

It is good to know some of these basic projections, but by far the most important concept of this unit is that it is important you are aware of the projection system that accompanies your spatial data. If you are assembling data from multiple shapefiles, as we will do below with soil and yield maps, you will need to account for the projections of each shapefile, to make sure they all have the same projection system.

In addition, different spatial data operations may prefer one projection system over the other. Operations that summarize areas will require projections that are based on area, not geometry. Similarly, spatial tools like rasters (which divide an area into rectangles or squares), will prefer a system that is square.

12.2 Shape Files

12.2.1 Case Study: Soybean Yield in Iowa

This is not our first encounter with shapefiles – we plotted our first shapefile map in the beginning of our course. Let's return to that dataset!

Let's examine the first few rows of this shapefile

```

## Simple feature collection with 6 features and 12 fields
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: -93.15033 ymin: 41.66641 xmax: -93.15026 ymax: 41.66644
## Geodetic CRS: WGS 84
##   DISTANCE SWATHWIDTH VRYIELDVOL Crop  WetMass Moisture           Time
## 1 0.9202733          5  57.38461  174 3443.652    0.00 9/19/2016 4:45:46 PM
## 2 2.6919269          5  55.88097  174 3353.411    0.00 9/19/2016 4:45:48 PM
## 3 2.6263101          5  80.83788  174 4851.075    0.00 9/19/2016 4:45:49 PM
## 4 2.7575437          5  71.76773  174 4306.777    6.22 9/19/2016 4:45:51 PM
## 5 2.3966513          5  91.03274  174 5462.851   12.22 9/19/2016 4:45:54 PM
## 6 3.1840529          5  65.59037  174 3951.056   13.33 9/19/2016 4:45:55 PM
##   Heading VARIETY Elevation           IsoTime yield_bu
## 1 300.1584 23A42 786.8470 2016-09-19T16:45:46.001Z 65.97034
## 2 303.6084 23A42 786.6140 2016-09-19T16:45:48.004Z 64.24158
## 3 304.3084 23A42 786.1416 2016-09-19T16:45:49.007Z 92.93246
## 4 306.2084 23A42 785.7381 2016-09-19T16:45:51.002Z 77.37348
## 5 309.2284 23A42 785.5937 2016-09-19T16:45:54.002Z 91.86380
## 6 309.7584 23A42 785.7512 2016-09-19T16:45:55.005Z 65.60115
##   geometry
## 1 POINT (-93.15026 41.66641)
## 2 POINT (-93.15028 41.66641)
## 3 POINT (-93.15028 41.66642)
## 4 POINT (-93.1503 41.66642)
## 5 POINT (-93.15032 41.66644)
## 6 POINT (-93.15033 41.66644)

```

The most useful shapefiles, in my experience, are presented in the “spatial feature” format above. It is, essentially, a data frame, but with a single, special geometry column that contains multiple measures per row. The geometry column is, if you will, composed of columns within a column.

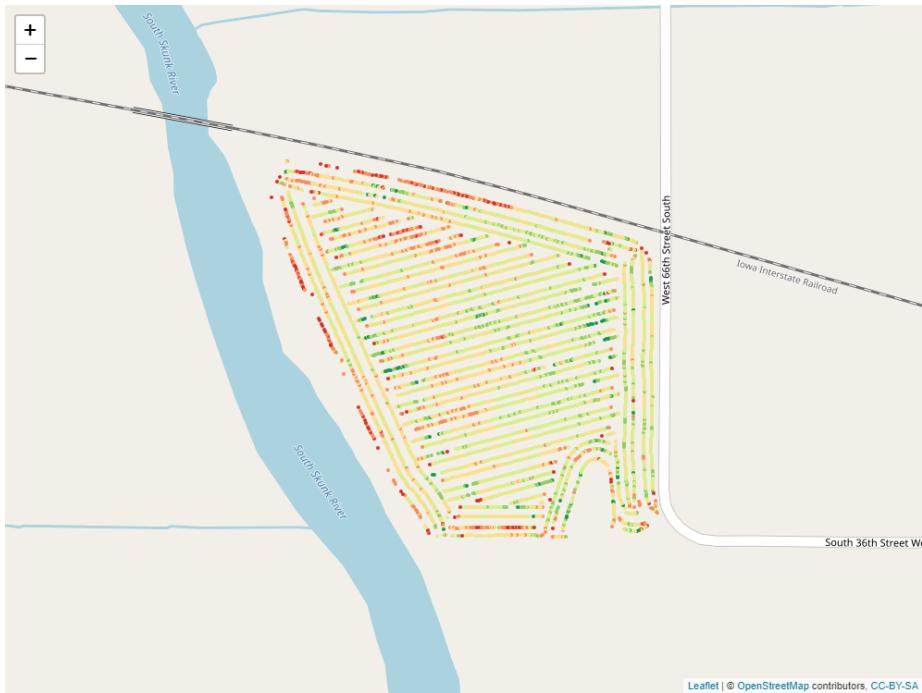
Let’s ignore the data for now and look at the information (the metadata) at the top of the output. First, let’s note the geometry type is POINT. Each row of this datafile defines one point. Shapefiles can be composed of all sorts of objects: points, lines, polygons, sets of multiple polygons, and so forth – and shapefiles can be converted between formats

Creating a successful map includes telling R what kind of object we intend to draw. So knowing the formate of a shapefile is critical!a helpful starting point.

Second, look at the geographic CRS. CRS stands for Coordinate Reference System. In this case, we are already in the standard WGS 84 format we discussed earlier, so our units are latitude and longitude.

One of the things we will learn this lesson is to use *Leaflet* to create maps. Leaflet is an awesome applet whose true appreciation would require using four-

letter conjunctions inappropriate for the classroom. It creates interactive maps that can zoom in, zoom out, and identify the values of individual points.



12.2.2 SSURGO

The Soil Survey Geographic Database (SSURGO) is maintained by the United States Department of Agriculture. It contains extensive soil surveys: soil evaluations for properties, susceptibility to weather extremes, suitability for agriculture, recreation, and buildings. The soil survey used to only be available in county books, which only special libraries had. Now, you can access all these data through R in seconds and match them precisely to a given map location.

The SSURGO data is in a database. A database is a series of tables, all describing different aspects of a data information. Each table contains 1-3 columns that are keys to match the tables with each other. Descriptions of the tables and their data can be obtained for SSURGO at:

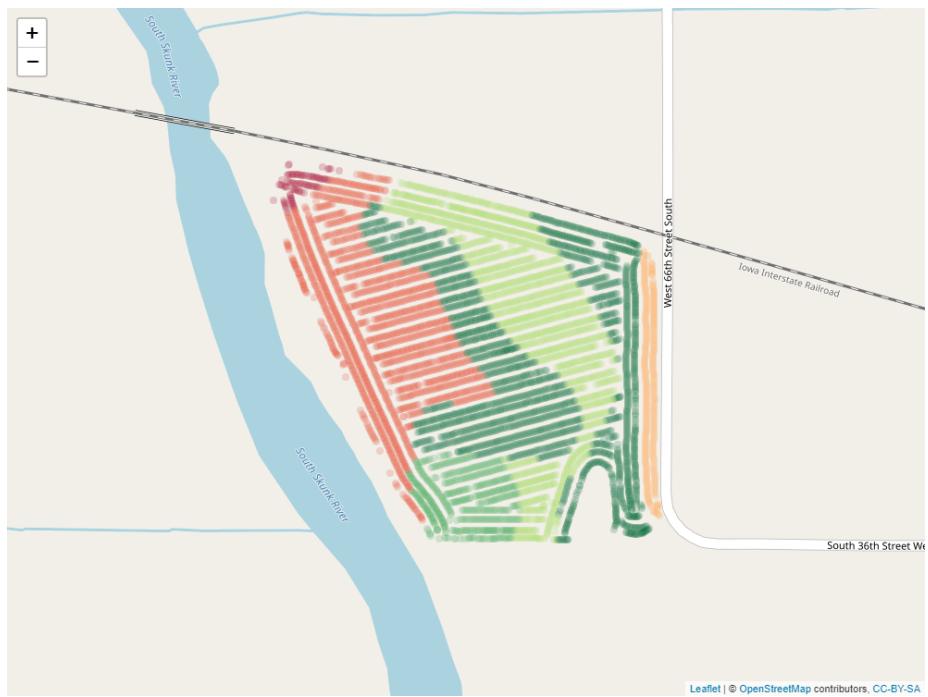
https://data.nal.usda.gov/system/files/SSURGO_Metadata_-_Table_Column_Descriptions.pdf

Putting all these tables together can be messy – fortunately, you only need to do it once, after which you can just change the shapefile you feed to the code. I will give you that code in the exercises this week.

Here is a SSURGO map of soil organic matter.



Here is another map, this time with the percent clay.



Now that we have our SSURGO data, we can join it with our yield data and ask questions how yields were grouped by quantitative descriptors, such as soil map unit name (“muname”), texture (“texdesc”), drainage class (“drainagecl”), or parent material (“pmkind”).

```
##   yield_bu                               muname hzname sandtotal.r
## 1 65.97034 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
## 2 64.24158 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
## 3 92.93246 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
## 4 77.37348 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
## 5 91.86380 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
## 6 65.60115 Wiota silt loam, 0 to 2 percent slopes    H1      9.4
##   silttotal.r claytotal.r om.r awc.r ksat.r cec7.r    chkey   texdesc
## 1       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
## 2       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
## 3       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
## 4       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
## 5       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
## 6       67.1        23.5     4 0.22      9  22.5 59965160 Silt loam
##   drainagecl slope.r pmkind           geometry
## 1 Well drained      1 Alluvium POINT (-93.15026 41.66641)
## 2 Well drained      1 Alluvium POINT (-93.15028 41.66641)
## 3 Well drained      1 Alluvium POINT (-93.15028 41.66642)
## 4 Well drained      1 Alluvium POINT (-93.1503 41.66642)
## 5 Well drained      1 Alluvium POINT (-93.15032 41.66644)
## 6 Well drained      1 Alluvium POINT (-93.15033 41.66644)
```

For example, here are soybean yields by soil texture, which would suggest a trend where soil yield increased with clay content in this field.

```
## # A tibble: 4 x 2
##   texdesc      yield_bu
##   <chr>        <dbl>
## 1 Clay loam     81.6
## 2 Silty clay loam 80.7
## 3 Silt loam     79.2
## 4 Loam          78.0
```

And this table would suggest that soybean preferred poorly drained soil to better-drained soils.

```
##   drainagecl yield_bu
## 1 Poorly drained 81.32400
## 2 Well drained   78.47489
## 3 Somewhat poorly drained 78.09084
```

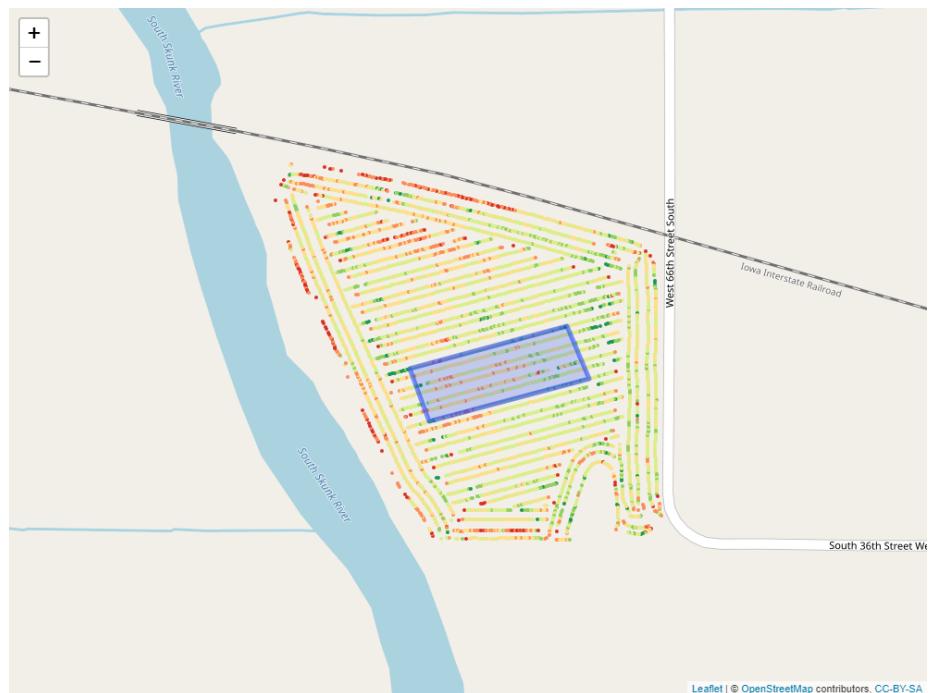
12.2.3 Operations with Shapes

Above, we subsetted our yield data according to different soil properties. In some cases, however, we may want to subset or group data by location.

12.2.3.1 Intersection

Say, for example, we applied a foliar fertilizer treatment to part of the field, as shown in the map below.

```
## [1] "temp\\file4e0c50d64591"
```



How might we find out statistics for yield measures within that applied area?

```
## Warning: attribute variables are assumed to be spatially constant throughout all
## geometries
```

```
## [1] 79.74649
```



12.2.3.2 Difference

What about the yields outside that area?

```
## Warning: attribute variables are assumed to be spatially constant throughout all  
## geometries
```



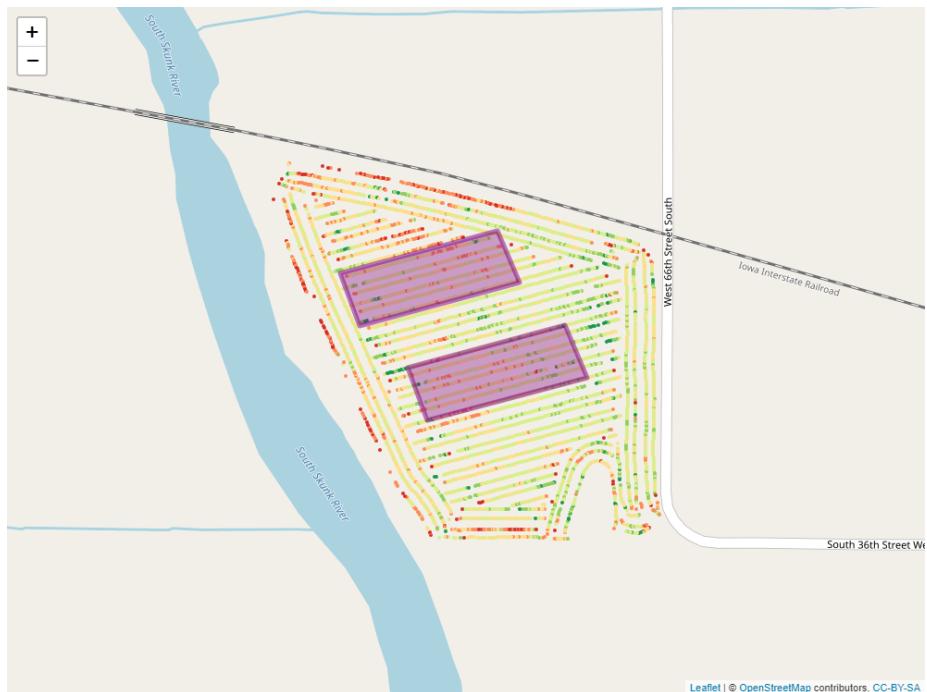
```
mean(field_yield_outside$yield_bu)
```

```
## [1] 80.12065
```

12.2.3.3 Union

What if we had two field plot areas?

If we wanted to analyze two areas together, we could use `st_union()` to combine them:



```
mean(corn_yield$yield_bu)
```

```
## [1] 80.09084
```

12.3 Rasters

```
library(stars)

selected_data = point_data %>%
  filter(attribute=="P_bray")

### make grid
grd = st_bbox(boundary) %>%
  st_as_stars() %>%
  st_crop(boundary)
# %>%
#   st_set_crs(6505)

# ordinary kriging -----
```

```
v = variogram(measure~1, selected_data)
m = fit.variogram(v, vgm("Sph"))
krige_plot = plot(v, model = m)

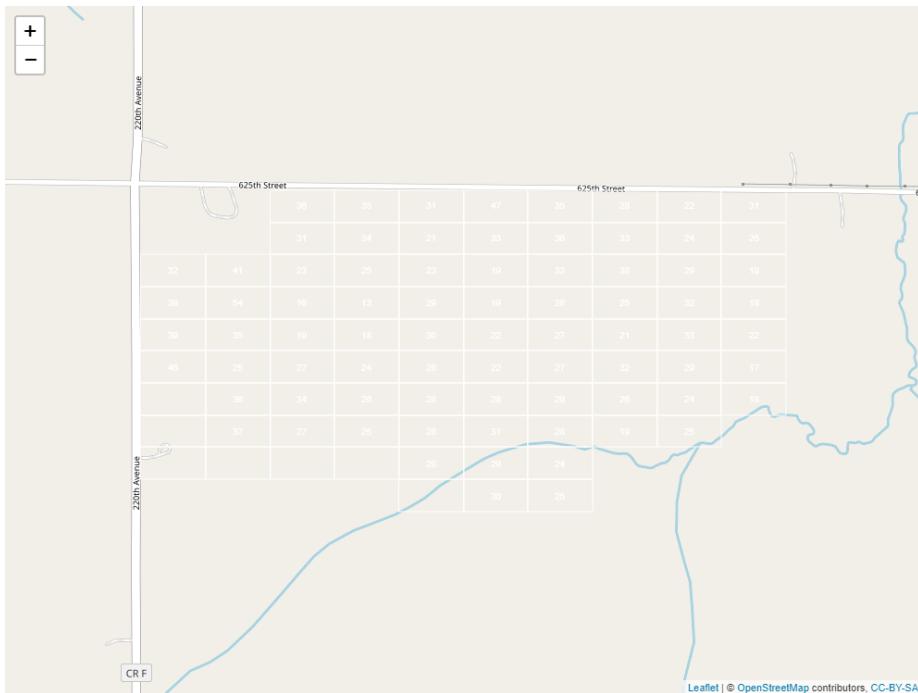
lzn.kr1 = gstat::krige(formula = measure~1, selected_data, grd, model=m)

## [using ordinary kriging]

# plot(lzn.kr1[1])

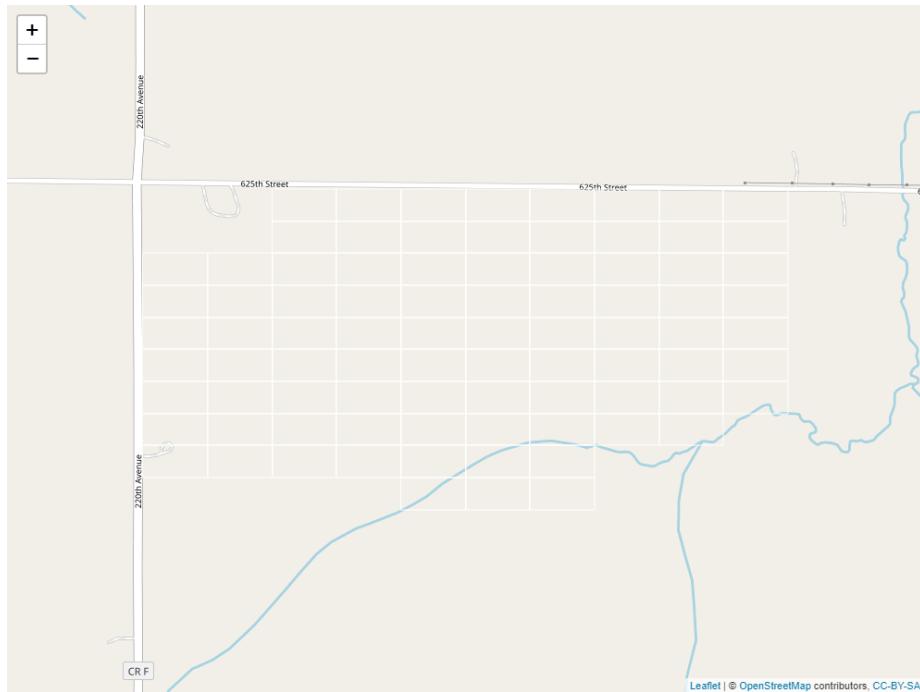
library(leafem)

soil_p_map = leaflet(lzn.kr1[1]) %>%
  addTiles() %>%
  addStarsImage(opacity = 0.5)
```

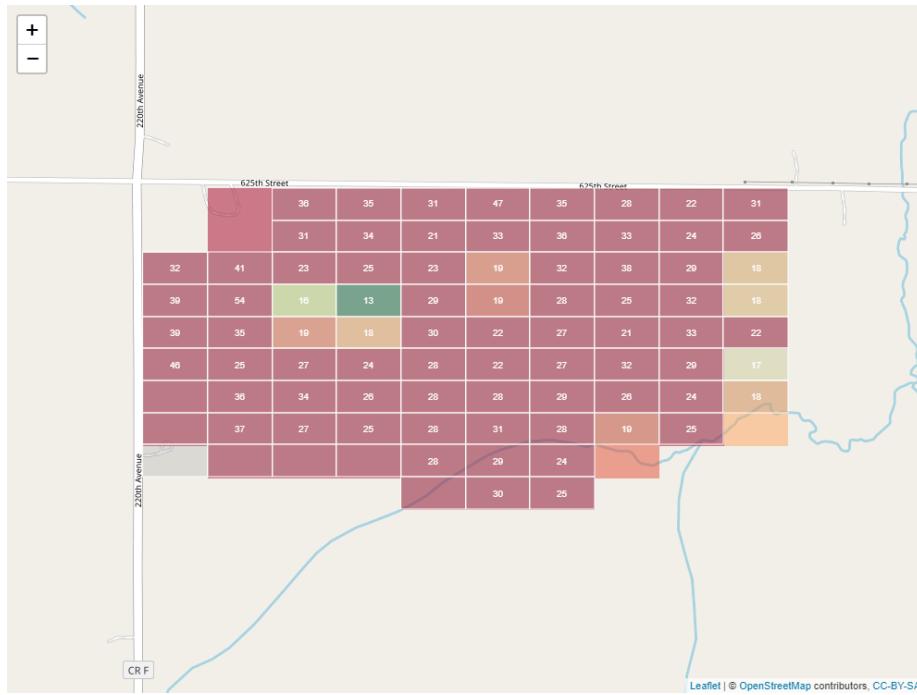


In the previous section, we worked with what are often described as vectors or shapes. That is, points which may or may not have been connected to form lines or polygons.

A raster is a grid system that we use to describe spatial variation. In essence, it is a grid system. Here is the same field we worked with in the previous section, now overlaid with a grid:



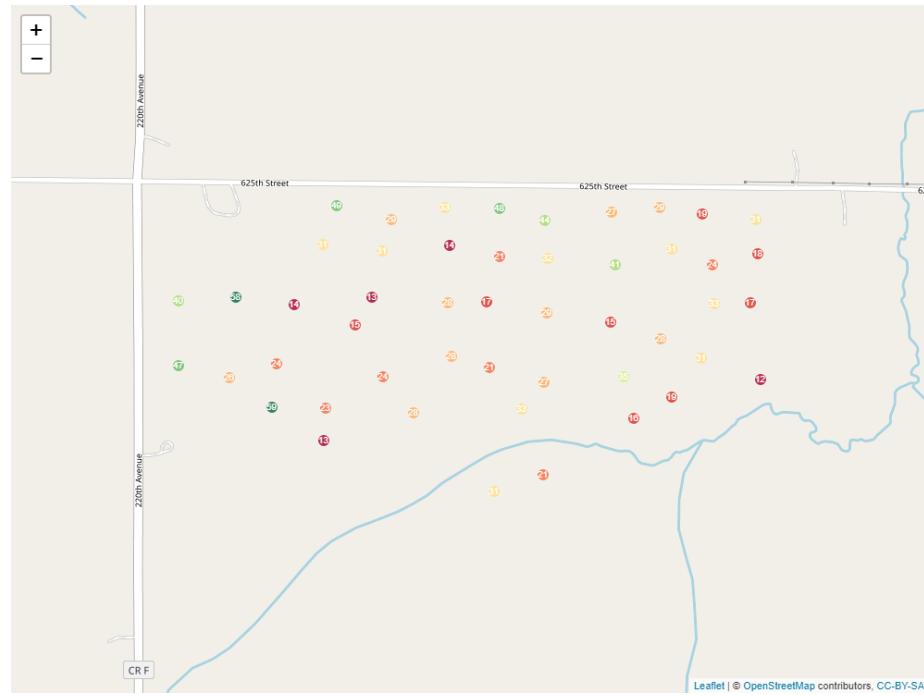
If you think it looks a little like we took a spreadsheet and trimmed it to fit our field, you are exactly right. Taking this analogy further, just as a spreadsheet is composed of cells, each containing a different value, so is a raster. Here it is, filled in with values representing predicted soil P test values:



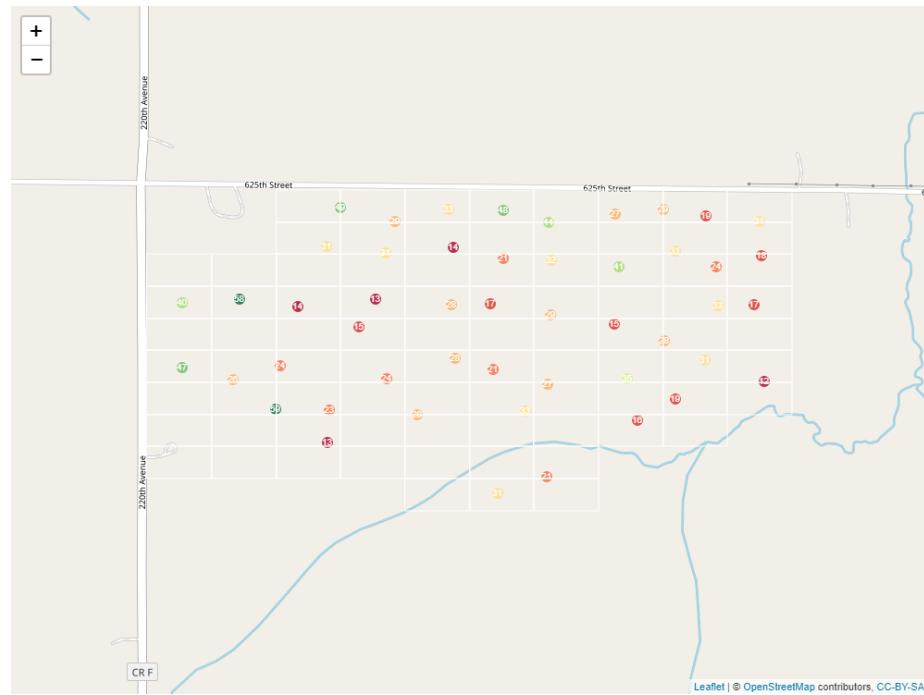
Often, the cells would be colored along a given gradient (say red-yellow-green) according to their values. This helps us to see spatial trends.

12.3.1 Interpolation

To create a raster that represents continuous soil trends across a field, however, we need to do a little modelling. You see, we start out with a set of soil cores like this:

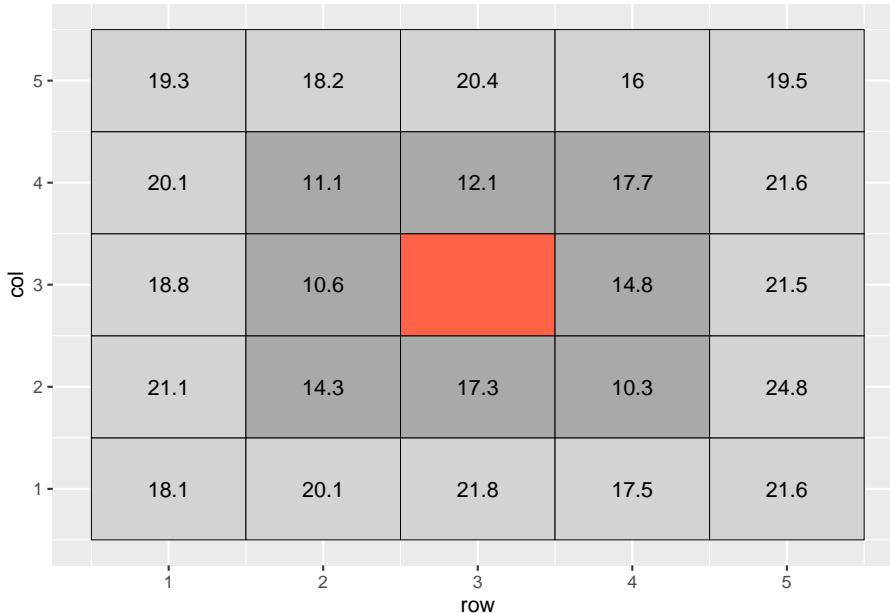


But we are trying to make predictions for each cell of our raster:



Some cells above one measure, others split a measure with a neighboring cell, and yet others contain no measure at all. In addition, even cells that contain a measure may have it in different locations relative to the cell center.

When we interpolate a raster, we make an educated guess about the values of cells in which no measure was taken, based on the values of the other cells. In the following example, the middle cell is missing:



The most basic way to interpolate this missing value would be to estimate its value as the mean value of all cells adjacent to it (dark grey in plot above). So the value of the missing cell would be equal to:

$$\text{cell value} = \text{mean}(16.6, 17.9, 13.1, 19.1, 11.6, 11.1, 16.2, 16.2) = 16.6$$

If we wanted to be a bit more accurate, we might extend out to the next ring of cells around the cell we are trying to estimate. But we probably would not want them to factor into the mean calculation as much as the first ring of cells – the difference between two measurements tends to increase with distance. So if they are two units away from the missing cell of, we might weight them so they contribute 1/4th as much to the estimate as the immediately adjacent cells.

If we were to fill out a table, it would look like this:

```
## # A tibble: 25 x 6
##       row   col cell value weight weighted_value
##   <int> <int> <int> <dbl>  <dbl>
## 1     1     1    1 18.1  1.00      18.1
## 2     1     2    2 20.1  1.00      20.1
## 3     1     3    3 18.8  1.00      18.8
## 4     1     4    4 10.6  1.00      10.6
## 5     1     5    5 21.1  1.00      21.1
## 6     2     1    1 21.1  1.00      21.1
## 7     2     2    2 14.3  1.00      14.3
## 8     2     3    3 17.3  1.00      17.3
## 9     2     4    4 10.3  1.00      10.3
## 10    2     5    5 24.8  1.00      24.8
## 11    3     1    1 18.1  1.00      18.1
## 12    3     2    2 20.1  1.00      20.1
## 13    3     3    3 21.8  1.00      21.8
## 14    3     4    4 17.5  1.00      17.5
## 15    3     5    5 21.6  1.00      21.6
## 16    4     1    1 20.1  1.00      20.1
## 17    4     2    2 11.1  1.00      11.1
## 18    4     3    3 12.1  1.00      12.1
## 19    4     4    4 17.7  1.00      17.7
## 20    4     5    5 21.6  1.00      21.6
## 21    5     1    1 19.3  1.00      19.3
## 22    5     2    2 18.2  1.00      18.2
## 23    5     3    3 20.4  1.00      20.4
## 24    5     4    4 16.0  1.00      16.0
## 25    5     5    5 19.5  1.00      19.5
```

```

##  1    1    1    1 18.1  0.25      4.53
##  2    2    1    2 20.1  0.25      5.03
##  3    3    1    3 21.8  0.25      5.45
##  4    4    1    4 17.5  0.25      4.38
##  5    5    1    5 21.6  0.25      5.4
##  6    1    2    6 21.1  0.25      5.28
##  7    2    2    7 14.3   1        14.3
##  8    3    2    8 17.3   1        17.3
##  9    4    2    9 10.3   1        10.3
## 10   5    2   10 24.8  0.25      6.2
## # ... with 15 more rows

```

The weighted value for each cell is the product of its observed value times its weight. To calculate the weighted value, we sum the weights and the weighted values. The weighted mean is then:

$$\text{weighted mean} = \frac{\sum \text{weighted value}}{\sum \text{weight}}$$

In this example, the calculation would look like:

$$\text{weighted mean} = \frac{220.45}{13} = 17.0$$

What we have just calculated is called the *inverse distance-weighted (IDW) mean* of the surrounding points. It is a simple, elegant way to estimate the missing value.

12.3.2 Kriging

The inverse distance-weighted mean, however, is not as accurate as which we are capable. We assume that the influence of points away from the empty cell decreases exponentially with distance. We don't consider how we would optimally weight the values of the surrounding cells.

We can develop a more complex, but likely accurate, estimate of the cell value using a different interpolation practice called *kriging*. (For some reason, I always want to insert a “d” into this term so it rhymes with “bridging”. But it is pronounced KREE-ging.) The name comes from Danie Krige, a South African geostatistician who was interested in locating gold deposits.

I will take you through the basics of kriging. A more elegant explanation of kriging can be found here: <https://pro.arcgis.com/en/pro-app/tool-reference/3d-analyst/how-kriging-works.htm>

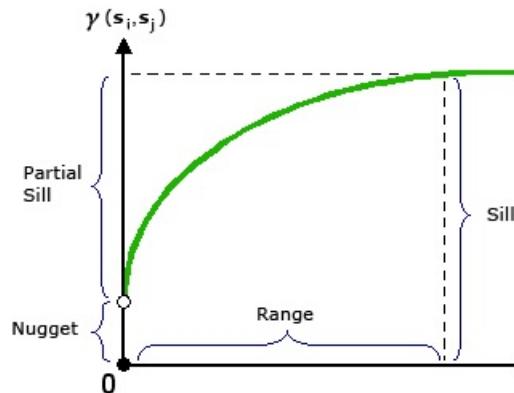


Figure 12.2: from “How Kriging Works” (<https://pro.arcgis.com/en/pro-app/tool-reference/3d-analyst/how-kriging-works.htm>)

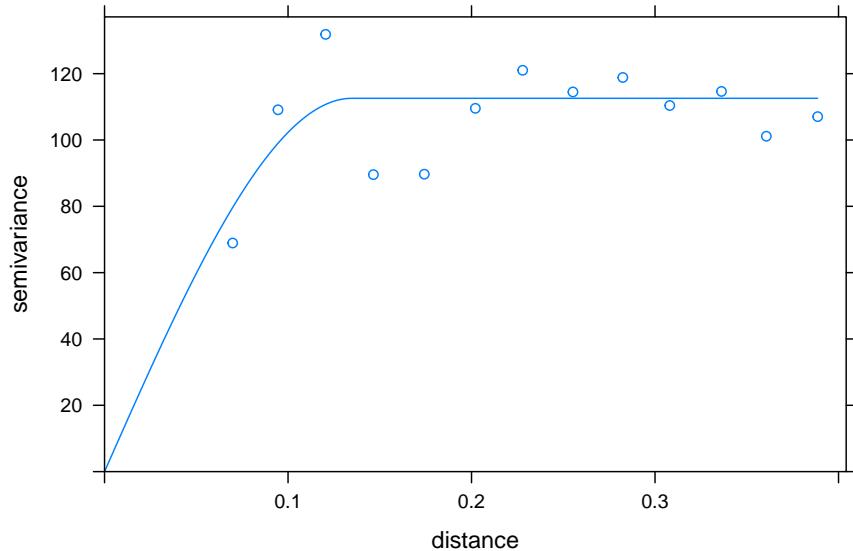
Kriging evaluates how points of varying differences from each other are correlated. These correlations are plotted in a *semivariogram*.

X is the distance between point pairs. Y is the squared difference between each pair of points selected by the software. Sometimes, pairs will be binned (similar to the binning in histograms) into according to distance (called “lag”) in this analysis. The squared differences of all pairs within a lag bin are averaged.

Does this nonlinear function look familiar by any chance? That’s right, it is a monomolecular function! The curve is described by different terms to which we are used to (and, to be honest, they don’t always make much sense.) In the kriging curve, the *Sill* is the maximum value the curve approaches. The *Nugget* is the y-intercept.

Otherwise, this curve is fit with nonlinear regression, just like the others we have seen. The calculated semivariances are then used to weight observations in calculating the weighted me. In this way, observations are weighted according the the strength of surrounding measurements, according to their measured correlation with distance.

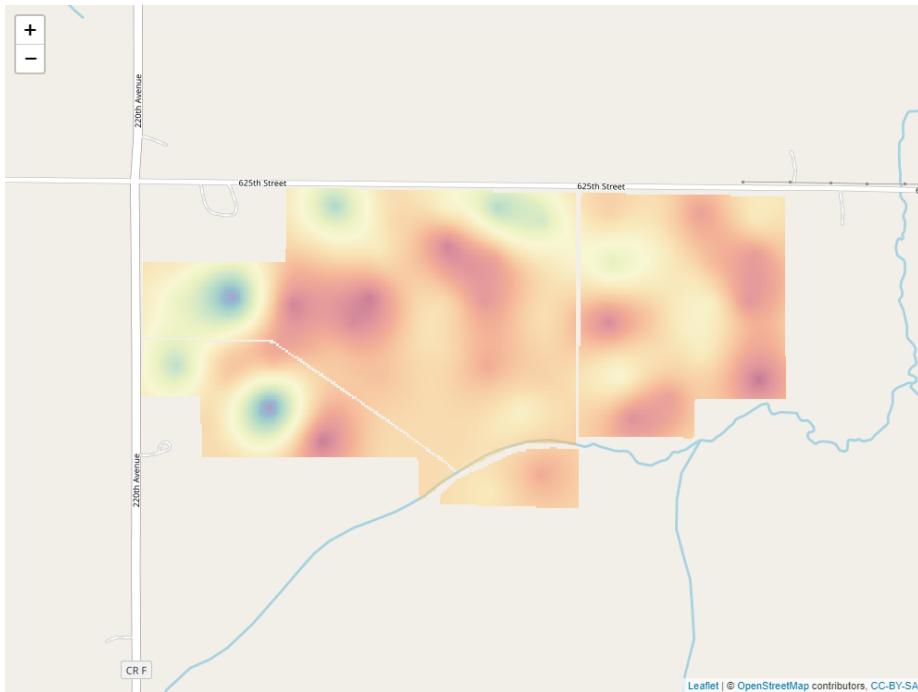
Here is the kridge plot for our soil phosphorus test data.



Using this model of the relationship between covariance and distance, then, R can determine how much to weight each observation, based on distance, to estimate values between measurement points.

When we produced our initial raster, the cell size was especially large for simplification of the raster, and to allow the cell values to be shown. When we build a raster with kriging, however, the cell size can be very small. This has the effect of getting rid of blockiness and allowing us to better predict and visualize trends in values across the landscape.

Here is our soil phosphorus test map, interpolated by R, using kriging. The red areas are areas of lower P test values. The green and blue areas have the greatest test values, and the yellow areas are intermediate.



12.3.3 Operations on Kriged Data

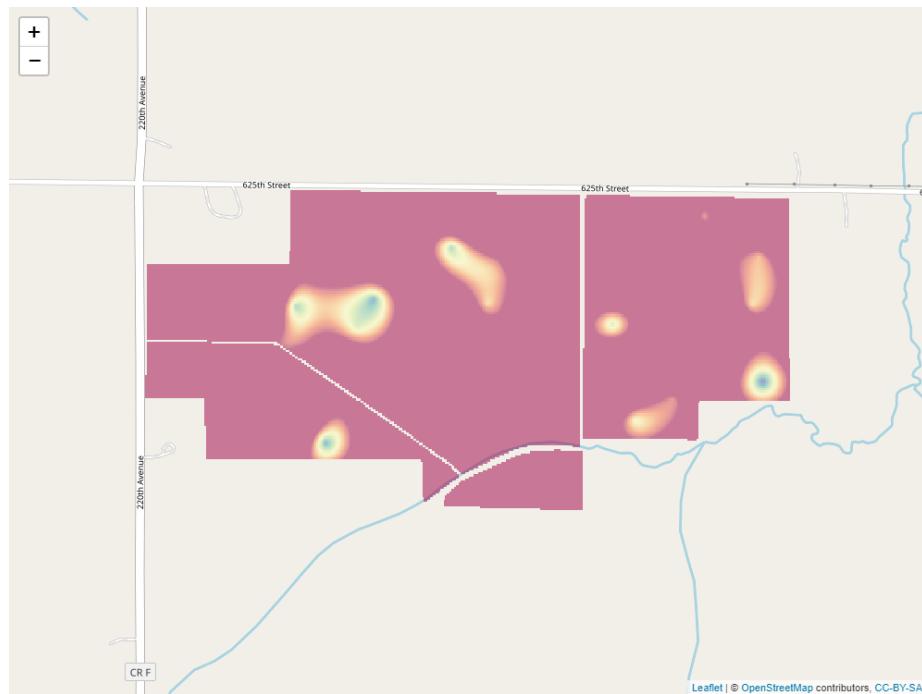
Of course, to take this example to completion, we would like to know the P-application rate for each of the cells within our raster. The University of Minnesota recommended application rate, based on soil P value and yield goal, is:

$$P_2O_5 \text{ recommended rate} = [0.700 - .035(\text{Bray P ppm})](\text{yield goal})$$

We can plug each cell's Bray P test value into this equation. For example, a point with a test value of 17 would, given a yield goal of 200, have a P_2O_5 rate recommendation of:

$$P_2O_5 \text{ recommended rate} = [0.700 - .035(17)](200) = 21$$

Here is the rate recommendation map.



The blue areas now have the highest recommended rates. Yellow are intermediate. Red areas should receive no P fertilizer.

Chapter 13

Machine Learning

Machine Learning is an advanced data science topic that I have struggled over whether to include in this text. On the one hand, it is not generally considered part of classical statistics. In addition, unless your plans include becoming a data analyst, you are unlikely to run these analyses on your own. On the other hand, machine learning is playing an increasing role in agricultural (and all areas of) data. It is a topic in which I have been immersed during the past couple of years.

Most importantly, if you are using tools to select hybrids, adjust nitrogen rates, or predict yield, those tools likely incorporate aspects of machine learning. In your daily lives, machine learning determines what advertisements you see, your social media feed, even the results from your search engines. This lesson emphasizes literacy about machine learning – there are no coding exercises included.

13.1 Machine Learning

To start with, why is it called “machine learning”? It sounds like statistics-meets-steampunk. To answer this, think about a machine that learns. In other words: a robot. In machine learning, we use complex algorithms that, in theory, could be used by artificial intelligence to learn about their environment. There is a strong predictive component to this: the machine could learn to predict the outcomes of future events by building models based on data it has already obtained. Hello, Skynet.

This sword, however, is also a plowshare. Machine learning can help us understand how many variables can simultaneously predict outcomes, in agriculture, medicine, and climate. Furthermore, machine learning takes a different approach to models than the methods we have learned earlier. In previous units, we have learned testing methods that were *parametric*. We defined a linear

model, then used our tests (which were all variations of regression) to parameterize it. Parameteric => parameter.

We've also learned that parametric methods have their challenges. One challenge is non-normal or skewed data. Another is heterogeneity of variances. These challenges required us to transform the data, a step which can generate confusion in analyzing and summarizing data. We were also warned when working with multiple linear regression models to beware of multicollinearity (correlations between independent variables) and heteroscedasticity (unequal variances).

Machine learning uses *nonparametric* tests. As the name suggests, these do not use parameters. Instead of regression models, machine learning tools use logical tests (for example, does an observed value fall in a range) to "decide" what value to predict. They measure the "similarity" between two locations to decide whether an observation in one location is likely to occur in the other. They figure out how to group similar observations into groups or "clusters".

In this unit, we will study three examples of machine learning. First, we will see how cluster analysis might be used to divide a sales territory into multiple regions, based on environment. Second, we will learn how to use *nearest-neighbor* analysis to predict yield for a given county, given its similarity to other counties. Finally, we will learn how *classification trees* can be used to explain yield response to multiple environmental factors.

This unit will be more focused on what I call "data literacy" than execution. I want you to be aware of these three tools, which are being used around you in the agronomy world. I want you to be able to speak of them, to ask questions, if engaged. I will resist going deep into the weeds with them, however, because they are often best suited to hundreds (ok) or many thousands (even better) of data points. It is unlikely you will use them in your Creative Component, or even your day-to-day research.

That said, just a few years ago, neither did I. Awareness and curiosity in data science are always good – there is always a more powerful way to address our questions.

13.2 Cluster Analyses

Cluster analysis takes a set of individuals (sometimes called examples in this analyses) and divides it into a set number of groups, based on the similarities between the individuals. These groups are defined such that the individuals within a group are very similar to each other, but very different from individuals in other groups.

By itself, clustering doesn't answer any questions about the values of the individuals in our data set. Unlike a classification tree or a k-Nearest Neighbor

analysis, it cannot be used to interpolate values between data points, or predict the value of a future observation.

To further illustrate cluster analysis, it is described in the data science community as “unsupervised learning.” This means that we do not begin with a variable of interest, nor a hypothesis about how changes in one variable relate to changes in another.

Why would we want to cluster data? The reason is that sometimes, when we are dealing with many variables, our analyses may become more intuitive if there are ways to break data down into groups. Clustering is a way of reducing or categorizing our data so it becomes less-overwhelming for us to deal with.

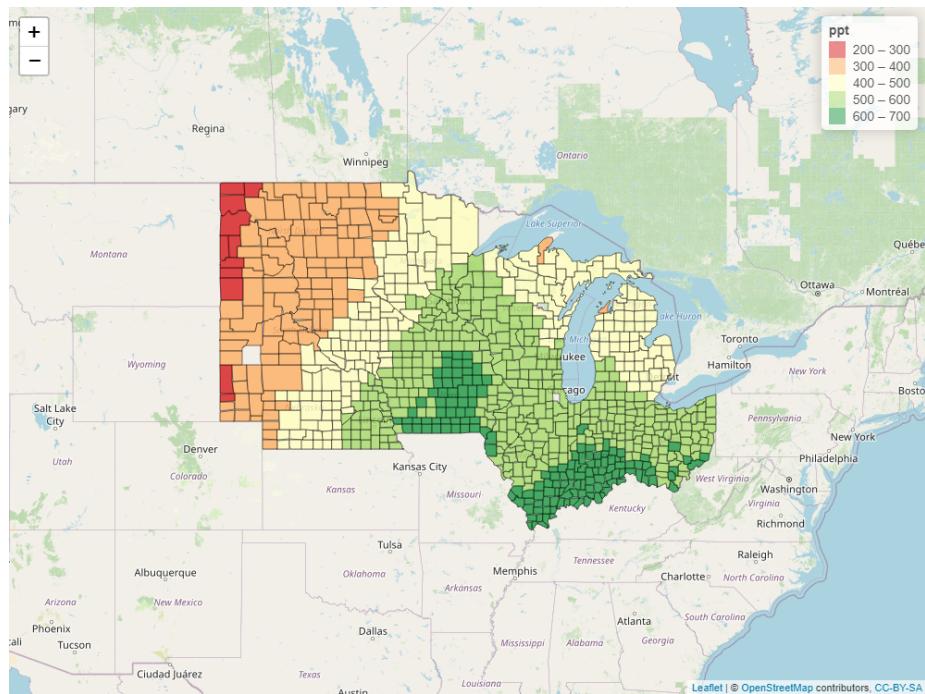
13.2.1 Case Study: Grouping Midwestern Environments

For example, anytime we are working with environmental data, we have potentially enormous datasets. Rather than try juggle the meanings of all these quantitative variables in our heads, we may want to describe the data as a series of groups. For example, we might say that a county in South Dakota is cold and dry, a county in northern Illinois is moderate temperature and moderate wetness, and a farm in western Ohio is warm and rainy. Great, we now have three groups. But how would other counties in the midwest sort into these three categories?

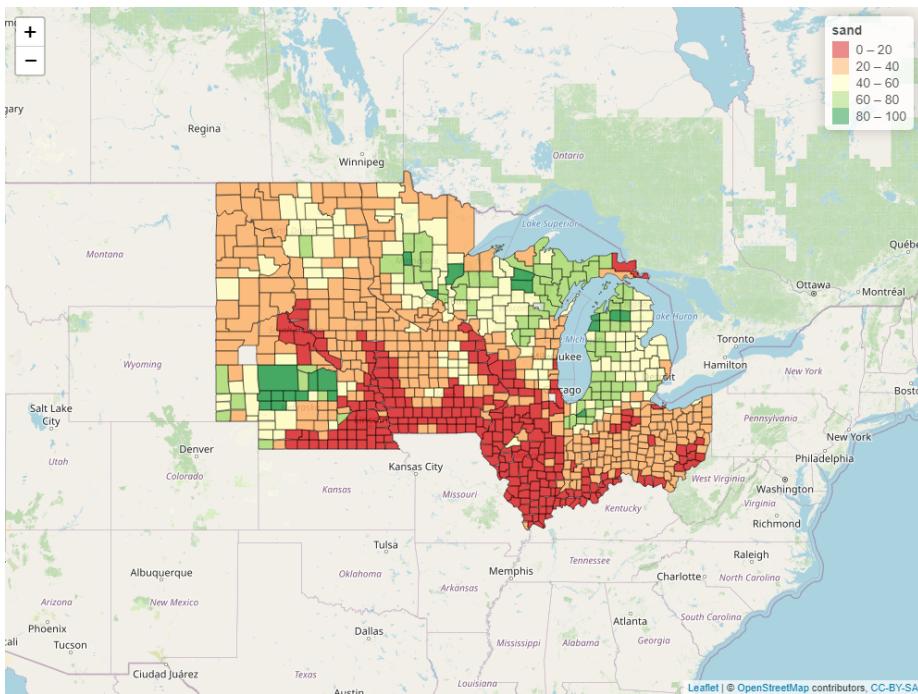
For an agronomic testing program, this is a very important question. If the intent is to address the individual needs of three different environments, we need to define those regions so that a sufficient, but not excessive, number of research locations can be designated within each environment.

Lets start out with dataset of county environments. The county environments are mostly soil data, which are not expected to change (outside of human influence) in the short term. Two data are climatic: growing degree days (gdd) and precipitation (ppt). Their values in this dataset reflect a 20-year mean.

Below I have plotted the mean annual precipitation (in millimeters). We see that precipitation is lowest in the Dakotas and generally increases as we move south and east.



Next up, here is a plot of percentage sand. This is a fascinating map. We can see the path of the glaciers down through Iowa and the Des Moines lobe. Areas with low sand (red below) are largely below the southern reach of the glacier. These areas tend to be high in silt. Conversely, we can see the outwash areas (green below) of the Nebraska Sand Hills, central Minnesota, northern Wisconsin, and Western Michigan, where rapid glacial melting sorted parent material, leaving them higher in sand and rock fragments.



We can continue looking at different variables, and through that process might arrive at a general conclusions. Counties that are further south or east tend to receive greater precipitation and more growing degree days (GDD). Counties that are further north and east tend to have more sand and less silt or clay in their soils.

But where do these boundaries end? Say you have a product, like pyraclostrobin, that may reduce heat, and you want to test it across a range of environments that differ in temperature, precipitation and soil coarseness. How would you sort your counties into three groups? This is where cluster analysis becomes very handy.

13.2.2 Scaling

When we conduct a cluster analysis, we must remember to scale the data first. Clustering works by evaluating the distance between points. The distance between points p and q is calculated as:

$$dist(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$$

Remember, this distance is Euclidean – it is the same concept by which we calculate the hypotenuse of a triangle, only in this case we are working with

a many more dimensions. Euclidean distance is commonly described when we refer to “as the crow flies” – it is the shortest distance between two points.

Cluster analysis will work to minimise the overall distance. If one variable has a much wider range of values than another, it will be weighted more heavily by the clustering algorithm. For example, in our dataset above, growing degree days range from 1351 to 4441, a difference of almost 3100. Precipitation, on the other hand, only ranges from 253 to 694, a different of 341. Yet we would agree the that growing degree days and precipitation are equally important to growing corn!

The solution is to *scale* the data so that each variable has a comparable range of values. Below are the cumulative growing degree days, precipitation, and percent clay values for the first 10 counties in our dataset. You can see how they differ in the range of their values.

cum_gdd	ppt	clay	geometry
3140.368	598.4478	33.73933	MULTIPOLYGON (((-94.70063 4...
3145.911	609.1156	34.32862	MULTIPOLYGON (((-94.92759 4...
2633.780	599.4642	28.56537	MULTIPOLYGON (((-91.61083 4...
3241.873	659.3385	37.16523	MULTIPOLYGON (((-93.09759 4...
2989.952	598.6462	30.46776	MULTIPOLYGON (((-95.09286 4...
2869.339	606.5850	26.67489	MULTIPOLYGON (((-92.29879 4...
2796.425	612.5765	21.99091	MULTIPOLYGON (((-92.55449 4...
2971.723	611.0333	21.81959	MULTIPOLYGON (((-94.164742...
2727.186	618.2398	21.67287	MULTIPOLYGON (((-92.55421 4...
2683.756	620.8813	21.66446	MULTIPOLYGON (((-92.08166 4...

Here are those same 10 counties, with their values for cum_gdd, ppt, and clay scaled. One of the easiest ways to scale the data are to convert their original values to Z-scores, based on their normal distributions. We do this exactly the way we calculated Z-scores in Unit 2.

We can see, for each variable, values fall mainly between -1 and 1.

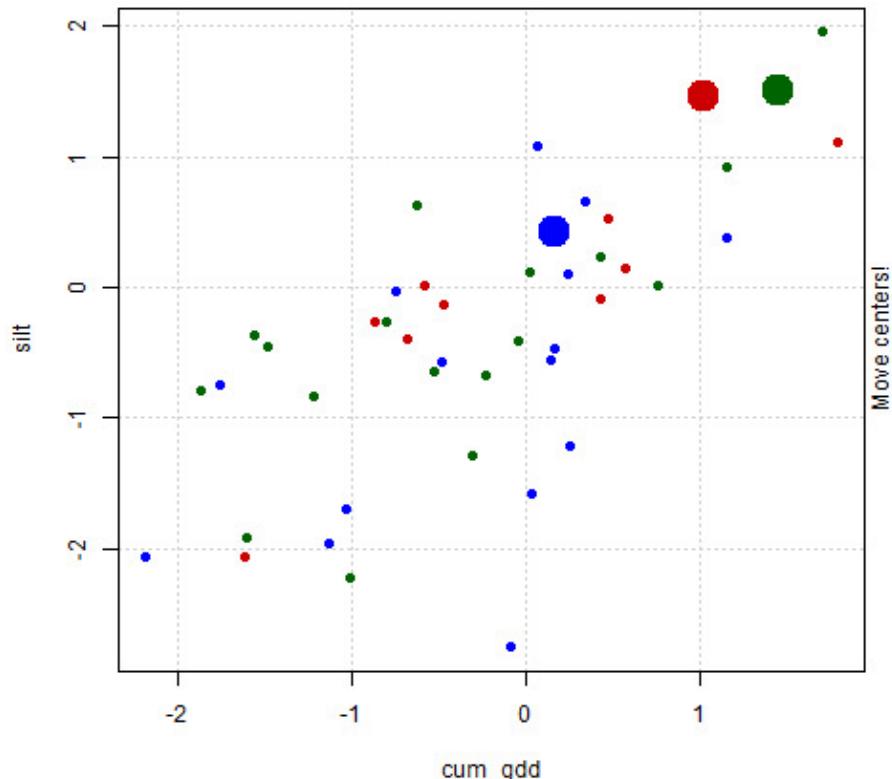
```
##      cum_gdd      ppt      clay
## 1   0.55154310  0.8572820  1.1905194
## 2   0.56106037  0.9708866  1.2651893
## 3  -0.31828780  0.8681063  0.5349214
## 4   0.72583066  1.5057241  1.6246202
## 5   0.29327347  0.8593956  0.7759755
## 6   0.08617543  0.9439380  0.2953761
## 7  -0.03902028  1.0077425 -0.2981361
## 8   0.26197313  0.9913093 -0.3198444
## 9  -0.15790581  1.0680532 -0.3384352
## 10 -0.23247627  1.0961831 -0.3395009
```

13.2.3 Clustering Animation

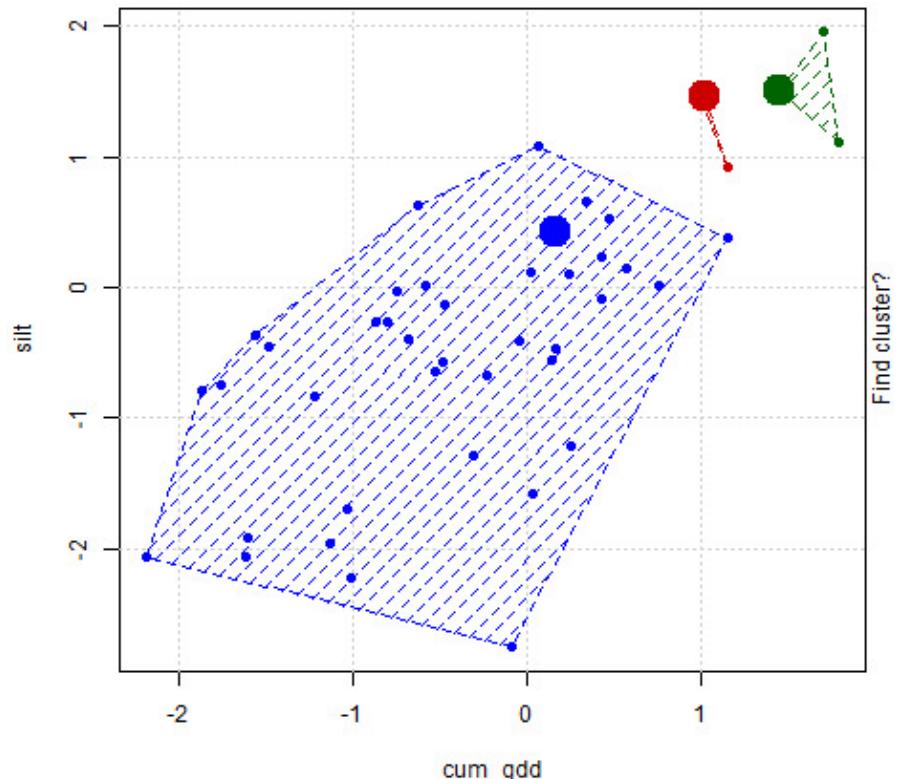
Using two variable, cum_gdd and silt and the animation below, lets walk through how clustering works.

In the plot above, cum_gdd is on the X-axis and silt is on the Y-axis. We want to cluster (or divide) our points into three groups. Colors have been assigned randomly to the points (smaller circles). They are scattered instead of being grouped by proximity to each other. After clustering, the points will be colored according to group.

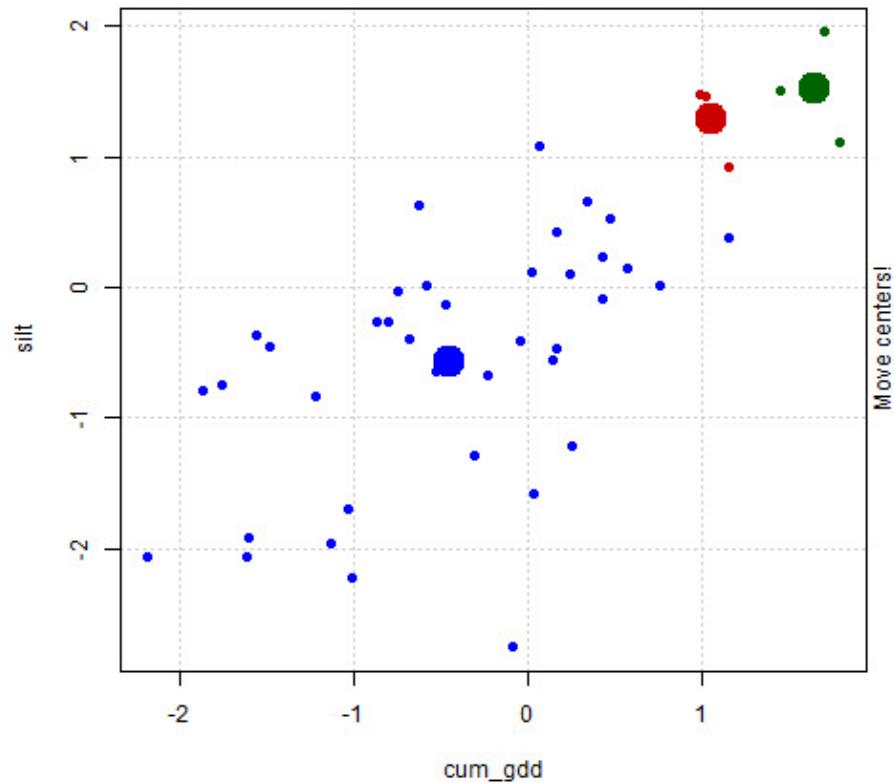
The first step is to randomly place three *centroids* in our plot. These are indicated by the large circles.



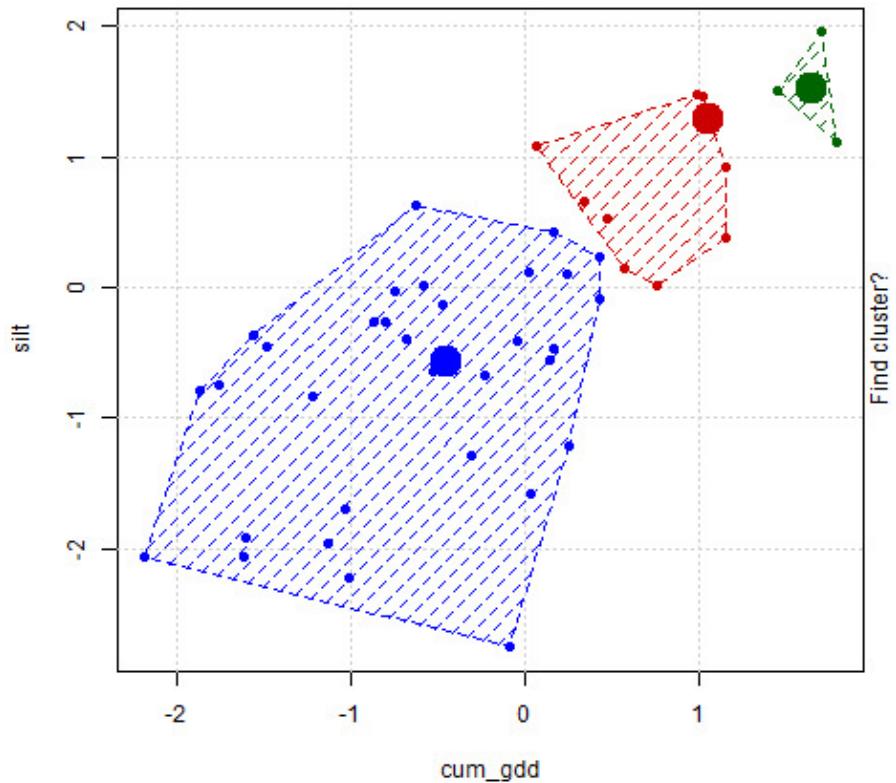
- 2) Next, for each point, we classify each point by its nearest centroid. In the plot below, this is represented by each point taking the same color as its nearest centroid.



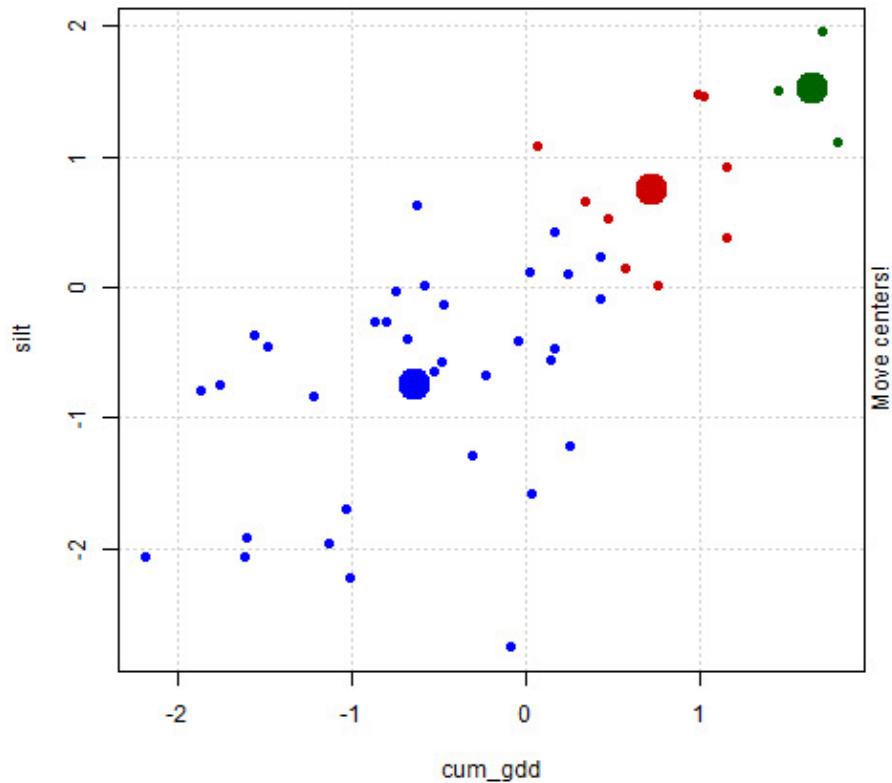
- 3) But now our centroid is not in the center of the group. So we will move it to the center of the group, as seen in the plot below.



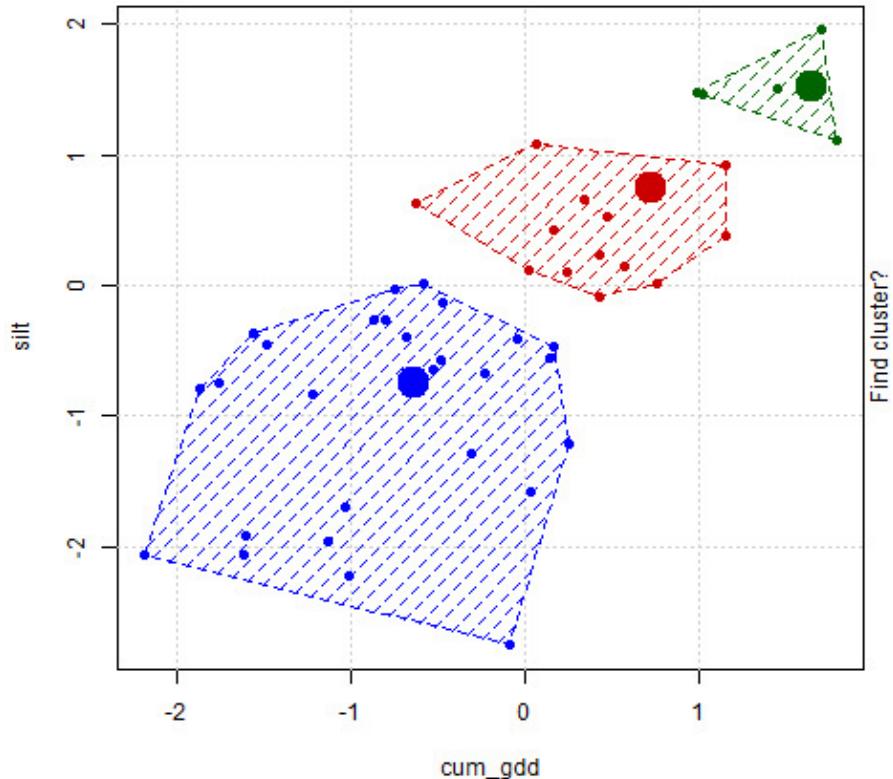
As soon as we move the centroids to the middle of its group, however, a problem occurs. Remember, each point is identified by the nearest centroid. So the groupings of the points again changes. In particular, the number of points in the red grouping increases.



We move our centroids again. You will notice both the red and blue centroids have moved down and to the left.



We reclassify the points. The number of red points has again increased.



This process repeats over and over again until, eventually, there is no more movement in the centroids. At this point, the cluster analysis is said to have converged on its final groupings. The positions of the centroids are typically used to define the new groupings.

cluster	cum_gdd	silt
1	2.2230855	1.0576394
2	-0.7074899	-0.7897348
3	0.6916494	0.9048737

The table above contains the final center estimates from the cluster algorithm. Cluster 1 is warmer and has siltier soil. Cluster 2 has medium numbers of growing degree days and silt. Cluster 3 is cooler and has soils lower in clay.

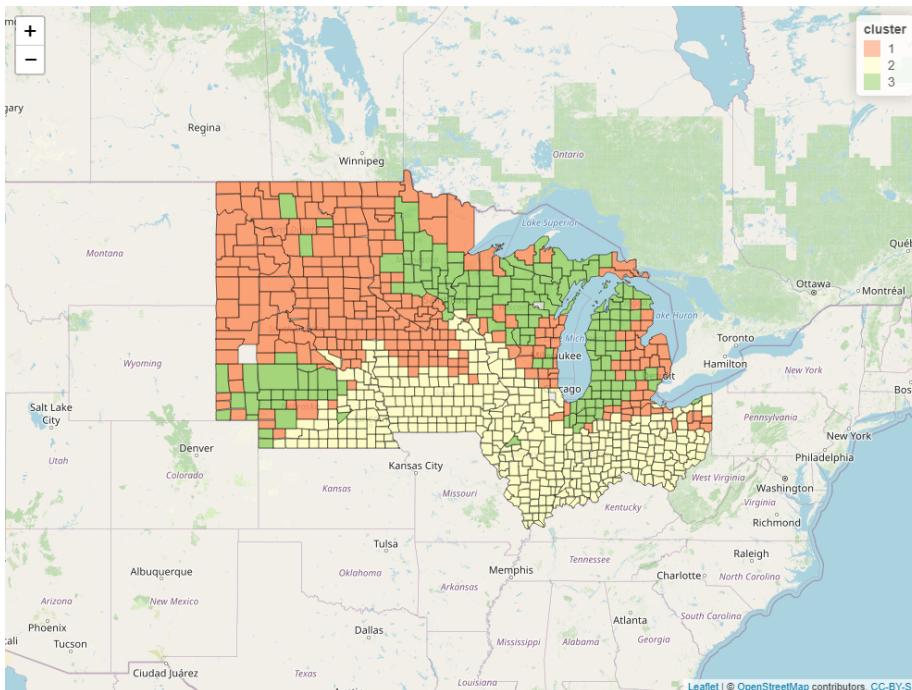
13.2.4 County Cluster Analysis

Now, let's run the cluster analysis with all of our variables. The cluster centers within each variable are shown below.

whc	sand	silt	clay	om	ppt	cum_gdd
-0.0152348	0.1147455	-0.3219194	0.2808926	0.1693996	-0.7947552	-0.6478638
0.4188589	-0.7234396	0.7811347	0.4051389	-0.4065930	0.7064892	0.7585691
-1.0435272	1.6497023	-1.4394576	-1.5196343	0.7460924	-0.4326339	-0.8192046

cluster	clay_class	gdd_class	om_class	ppt_class	sand_class	silt_class	whc_class
1	ave	low	ave	low	ave	ave	ave
2	ave	high	ave	high	low	high	ave
3	low	low	high	ave	high	low	low

We can now plot our clusters on a map. Cluster 1, which has a greater growing degree days, precipitation and silt, accounts for much of the lower third of our map. Cluster 2, which has lower growing degree days and precipitation, accounts for most of northwestern quarter of counties. Cluster 3, which is most remarkable for its higher sand content, accounts for parts of Nebraska, central Wisconsin and counties to the south and east of Lake Michigan.



13.3 k-Nearest-Neighbors

In *k-Nearest-Neighbors* (kNN) analyses, we again use the Euclidian distance between points. This time, however, we are not trying to cluster points, but to guess the value of a variable one observation, based on how similar it is to other observations.

13.3.1 Case Study: Guessing County Yields based on Environmental Similarity

For this example, we will continue to work with our county environmental dataset. This dataset includes yields for corn, soybean, and, where applicable, wheat and cotton. What if, however, our corn yield data were incomplete: that is, not every county had a recorded yield for corn. How might we go about guessing/estimating/predicting what that corn yield might be?

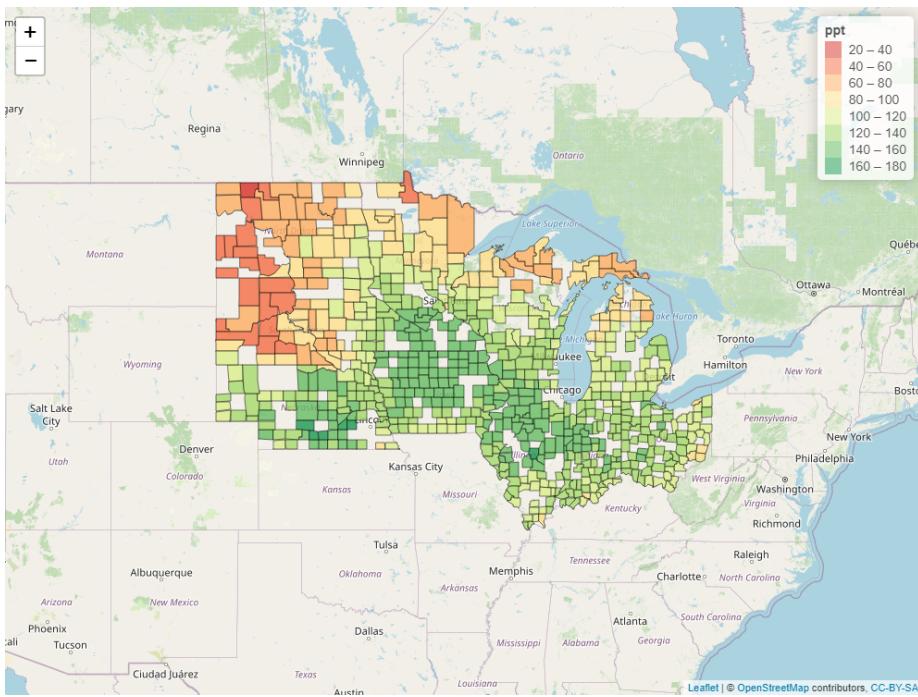
Well, if you are a farmer and you are interested in a new product – what might you do? You might ask around (at least the farmers who aren’t bidding against you for rented land) and see whether any of them have used the product and what response they observed.

In considering their results, however, you might incorporate additional information, namely: how similar are their farming practices to yours? Are they using the same amount of nitrogen fertilizer? Do they apply it using the same practices (preplant vs split) as you? Are they planting the same hybrid? What is their crop rotation? What is their tillage system?

In other words, as you ask around about others experiences with that product, you are also, at least intuitively, going to weigh their experiences based on how similar their practices are to yours. (Come to think of it, this would be great dataset to mock up for a future semester!)

In other words, you would predict your success based on your nearest neighbors, either by physical distance to your farm, or by the similarity (closeness) of your production practices. If that makes sense, you now understand the concept of nearest neighbors.

For our county data scenario, we will simulate a situation where 20% of our counties are, at random, missing a measure for corn yield.



13.3.2 Scaling

Remember, kNN analysis is like cluster analysis in that it is based on the *distance* between observations. kNN seeks to identify k neighbors that are most similar to the individual for whom we are trying to predict the missing value.

Variables with a greater range of values will be more heavily weighted in distance calculations, giving them excessive influence in how we measure distance (similarity).

The solution to this is to scale the data the same as we did for cluster analysis.

13.3.3 k-Nearest-Neighbor Animation

In kNN analysis, for each individual for which we are trying to make a prediction, the distance to each other individual is calculated. As a reminder, the Euclidean distance between points p and q is calculated as:

$$dist(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 \dots (p_n - q_n)^2}$$

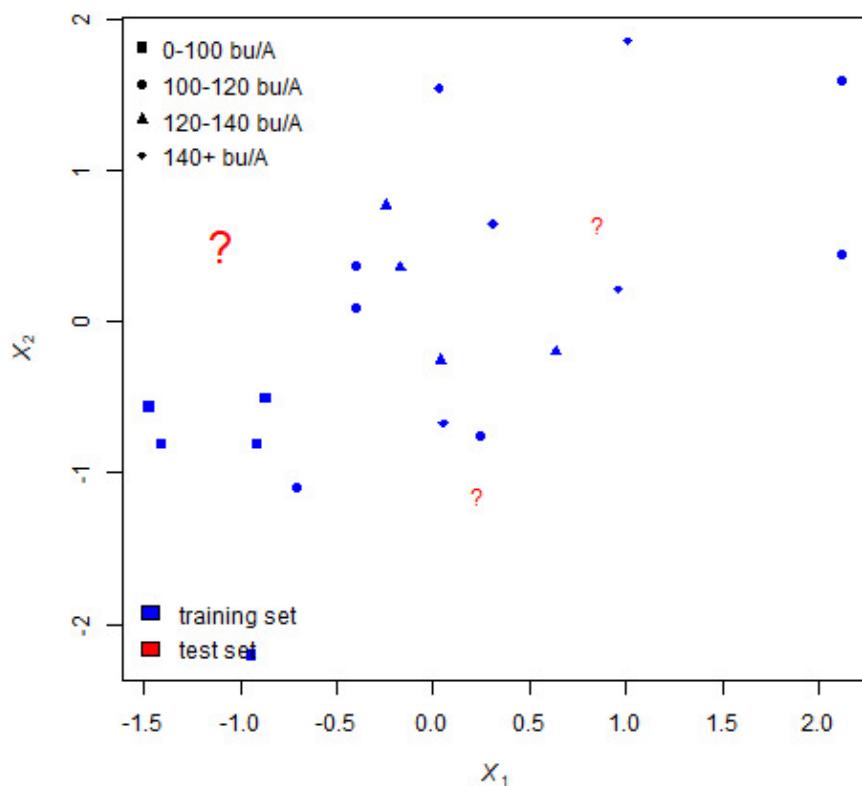
The k nearest (most similar) observations are then identified. What happens next depends on whether the measure we are trying to predict is qualitative

or quantitative. Are we trying to predict a drainage category? Whether the climate is “hot” vs “cold”? In these cases, the algorithm will identify the value that occurs most frequently among those neighbors. This statistic is called the *mode*. The mode will then become the predicted value for the original individual. Let’s demonstrate this below.

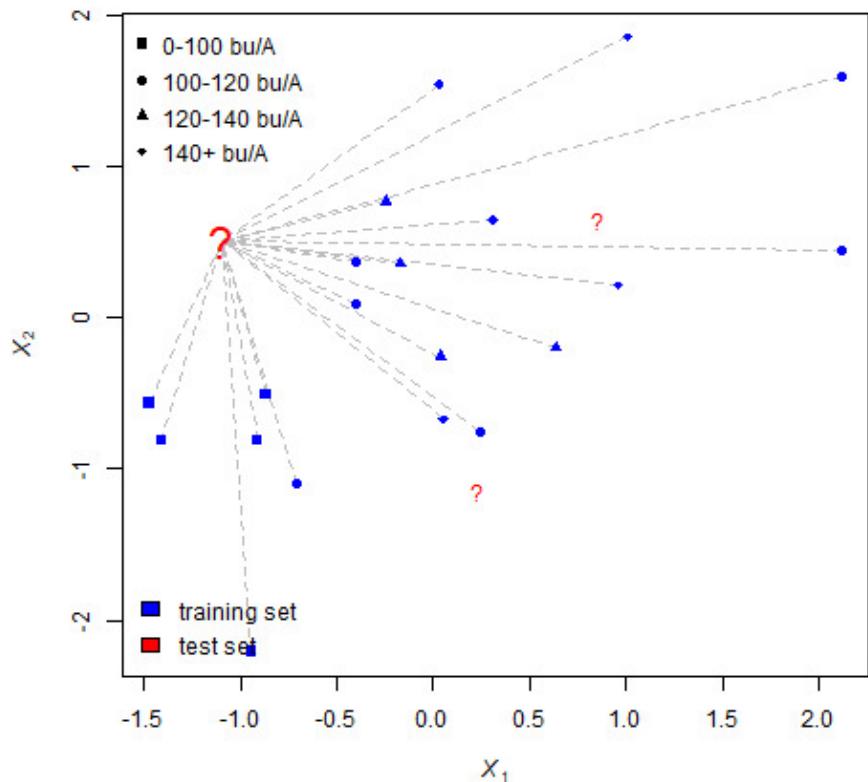
In this example, corn yield is predicted based on cumulative growing degree days and percent silt. Because it is easier to illustrate qualitative predictions, corn yield has been categorized within four ranges: 0-100 (circle), 100-120 (triangle), 120-140 (cross), and 140+ (“x”).

The known yields, referred to as the training set, are colored blue. The individuals for which we are trying to predict yield are represented by red question marks. Although the axes names are ambiguous, each known point is plotted according to its cumulative growing degree days (cum_gdd) and silt content.

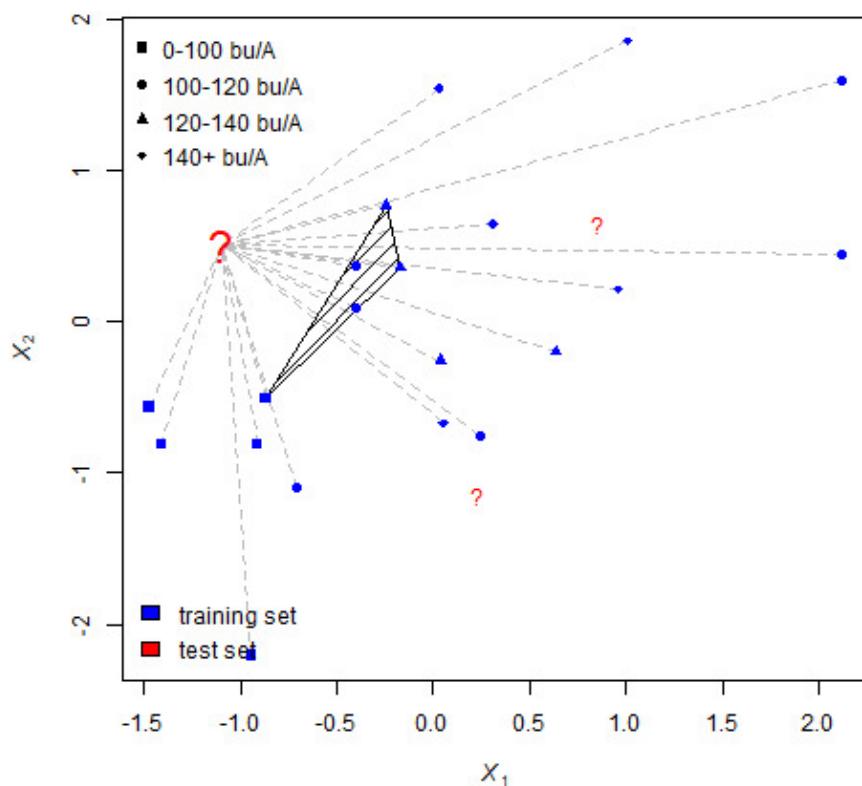
This example uses $k=5$. That is, the missing values of individuals will be calculated based on the most frequent yield classification of their 5 nearest neighbors. There are three points, indicated by red question marks for which we are going to try to pick the value. We start with the leftmost question mark.



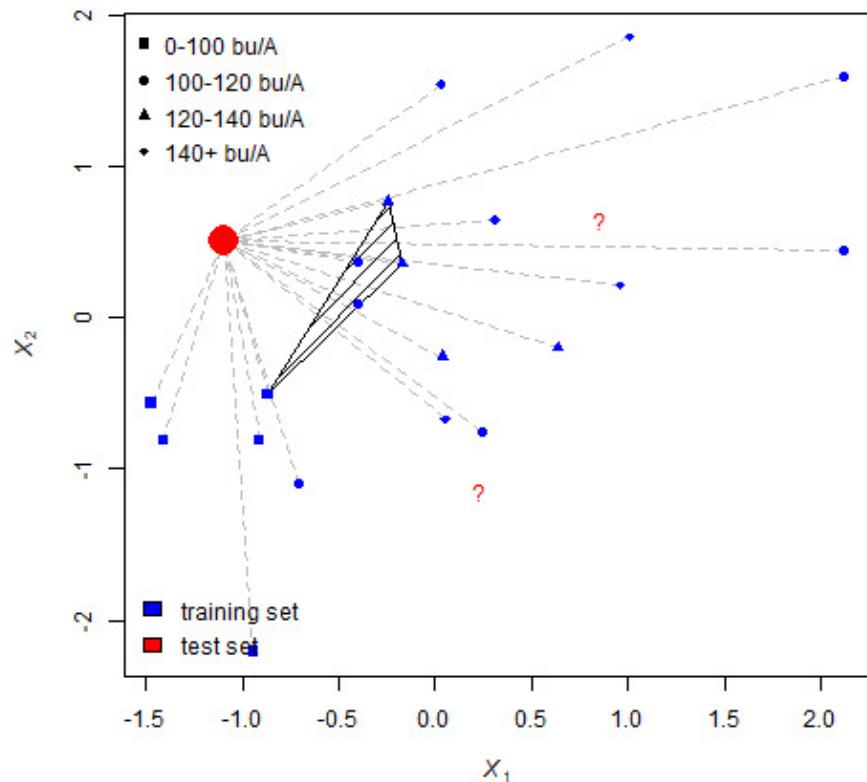
For that point, we measure the distance to every other point:



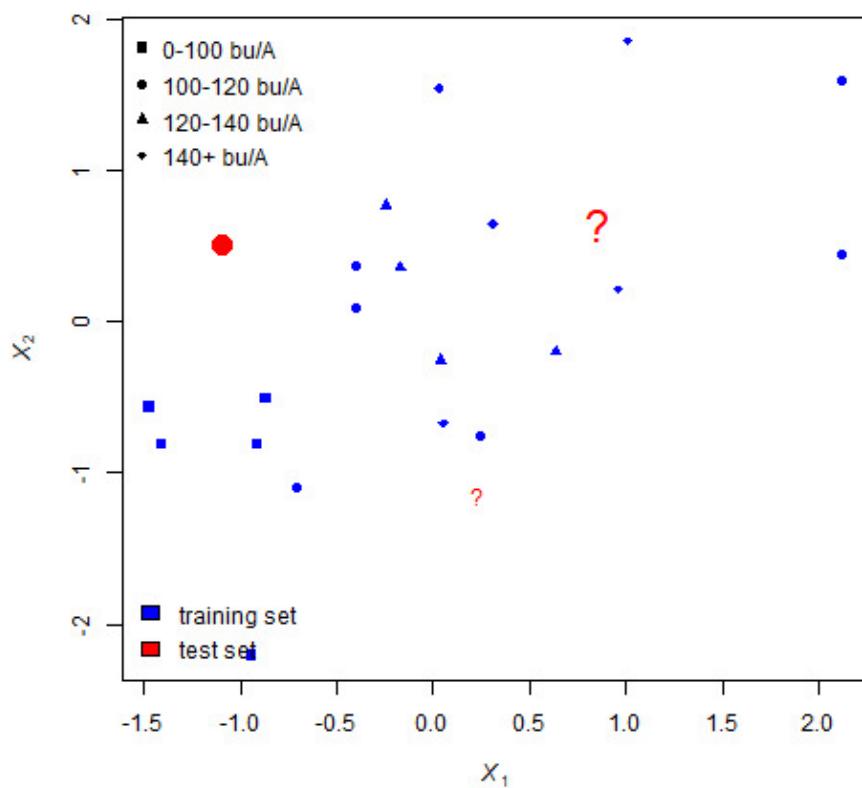
And identify the nearest 5 points (neighbors). Of the five nearest neighbors, there is one square (yield up to 100 bushels), two circles (indicating yields between 100 and 120 bushels) and two triangles (yields between 120 and 140 bushels).



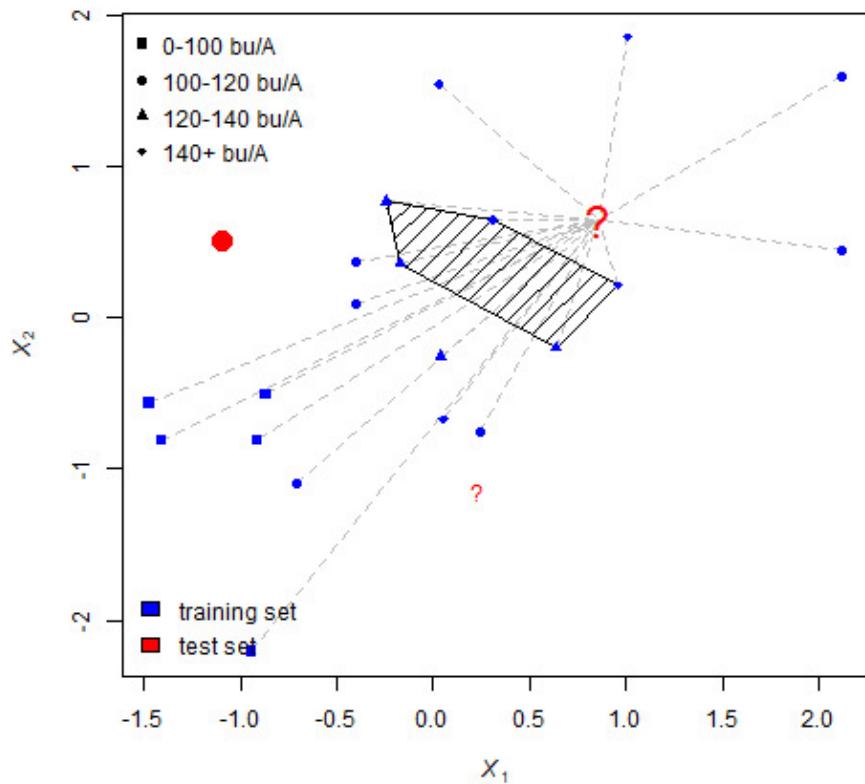
There are two each circles and triangles, so R randomly selects from those two values. The missing value is classified as a circle.



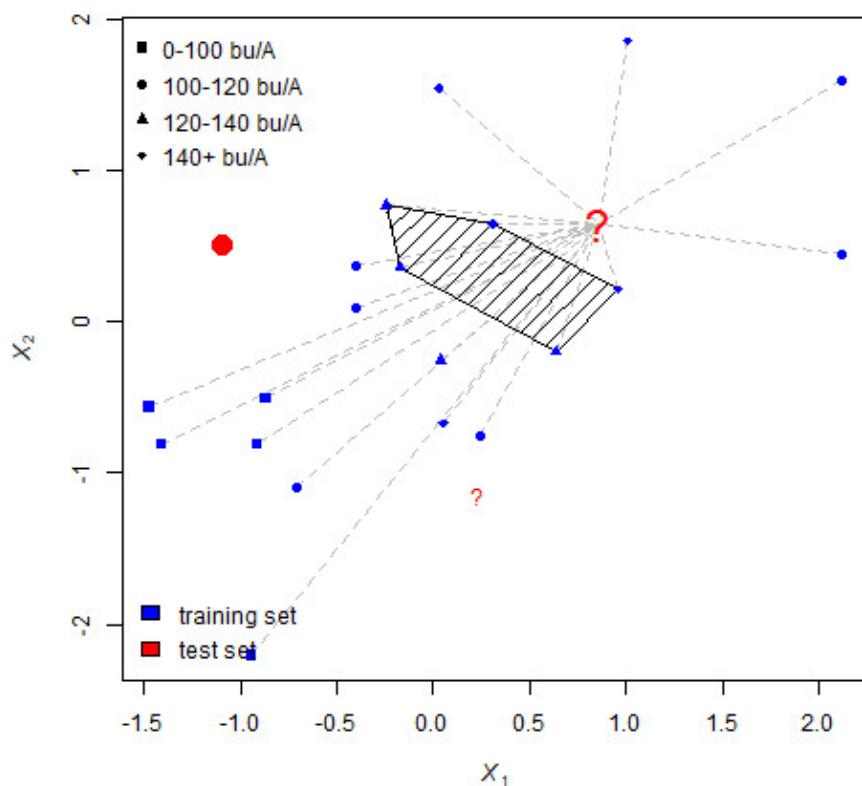
Let's repeat this process for the next missing value:



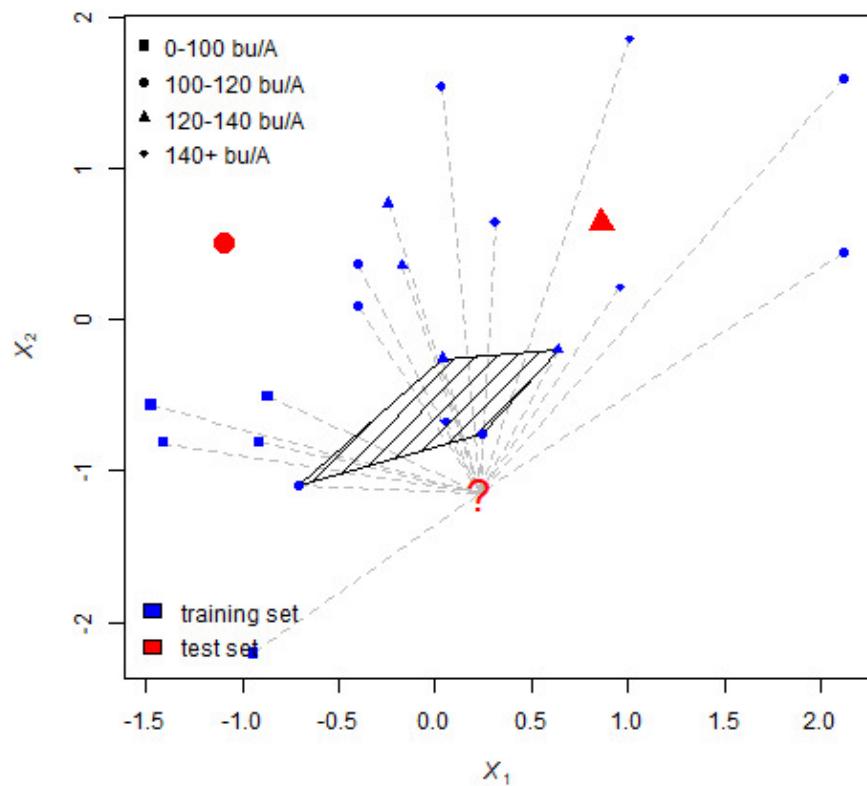
For this point, the five nearest neighbors are again identified.



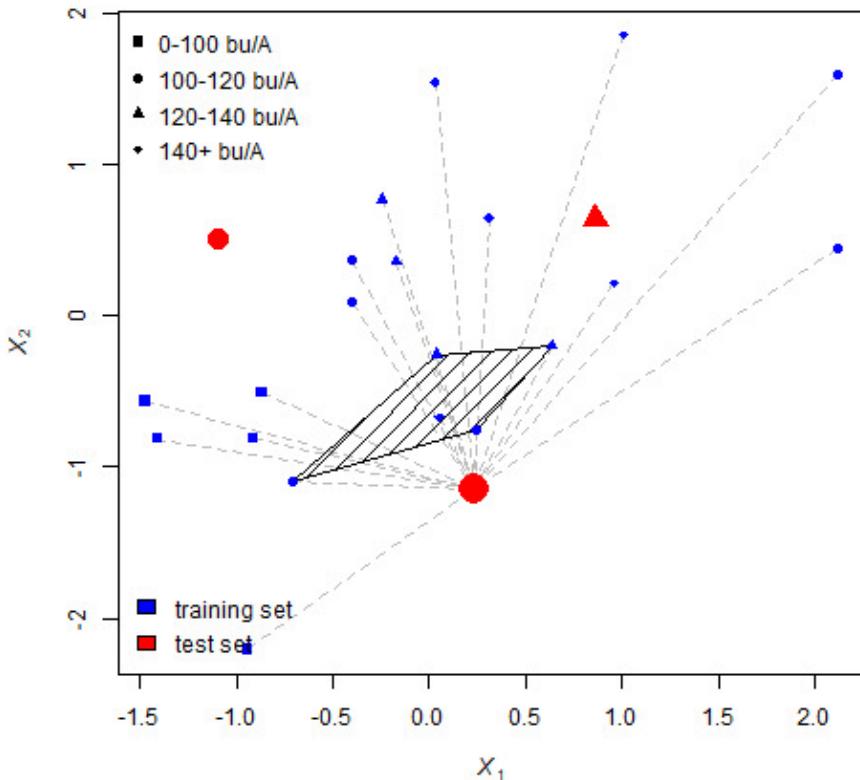
In this case there are three triangles and two circles, so the missing value is estimated to be a triangle.



For the last point, the nearest 5 neighbors include two circles, two triangles, and one diamond.



R randomly chooses between the triangle and circle – again, the circle wins.



13.3.4 Choosing k

In the example above, a value of 5 chosen arbitrarily to keep the illustration from becoming too muddled. How should we choose k in actual analyses?

Your first instinct might be to select a large k . Why? Because we have learned throughout this course that larger sample sizes reduce the influences of outliers on our predictions. By choosing a larger k , we reduce the likelihood an extreme value (a bad neighbor?) will unduly skew our prediction. We reduce the potential for bias.

The flip side of this is: the more neighbors we include, the less “near” they will be to the individual for which we are trying to make a prediction. We reduce the influence of outlier, but at the same time predict a value that is closer to the population mean, rather than the true missing value. In other words, a larger k also reduces our precision.

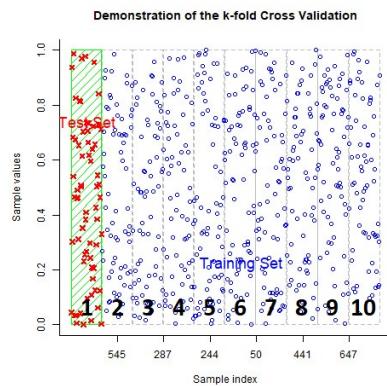
A general rule of thumb is to choose a k equal to the square root of the total number of individuals in the population. We had 50 individuals in the population above, so we could have set our k equal to 7.

13.3.5 Model Cross-Validation

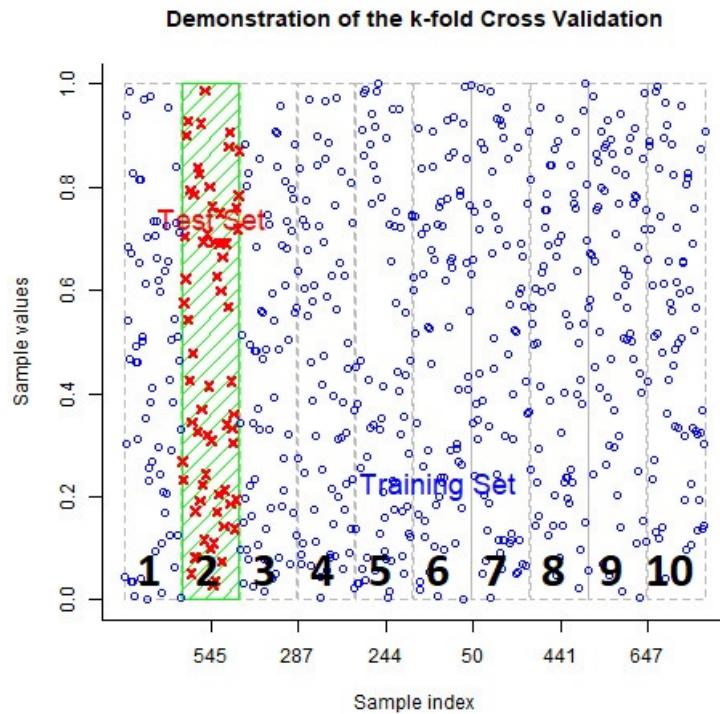
When we fit a kNN model, we should also cross-validate it. In cross-validation, the data are portioned into two groups: a training group and a testing group. The model is fit using the training group. It is then used to make predictions for the testing group.

The figures below illustrate a common cross-validation method, k-fold cross validation. In this method, the data are divided into k number of groups. In this example, $k=10$, meaning the data are divided into 10 groups. Nine of the groups (in blue) are combined to train the model. The tenth group (in red) is the testing group, and used to validate the fit. The validation process goes through 10 rounds.

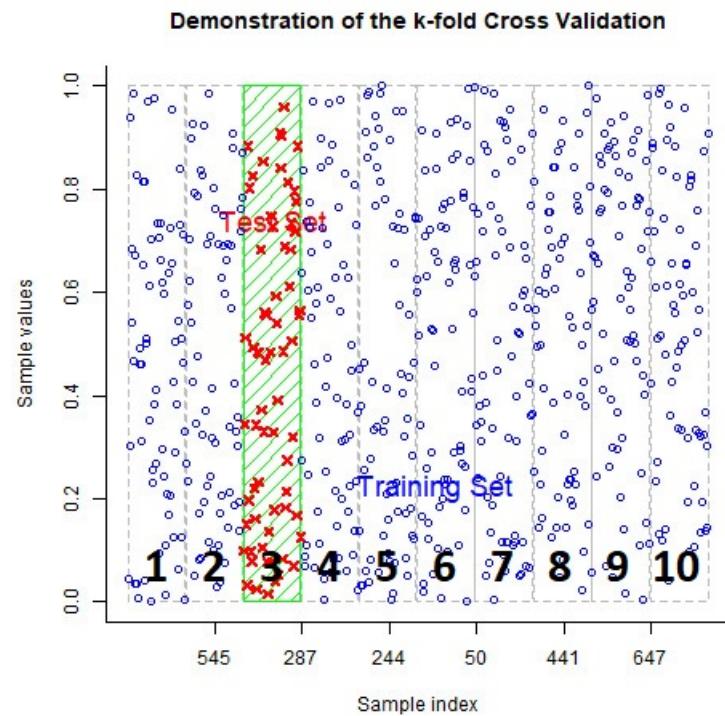
In the first round, Groups 2 through 9 are used to train the model. The fit of the model is then tested by how closely it can predict the observed values in Group 1.



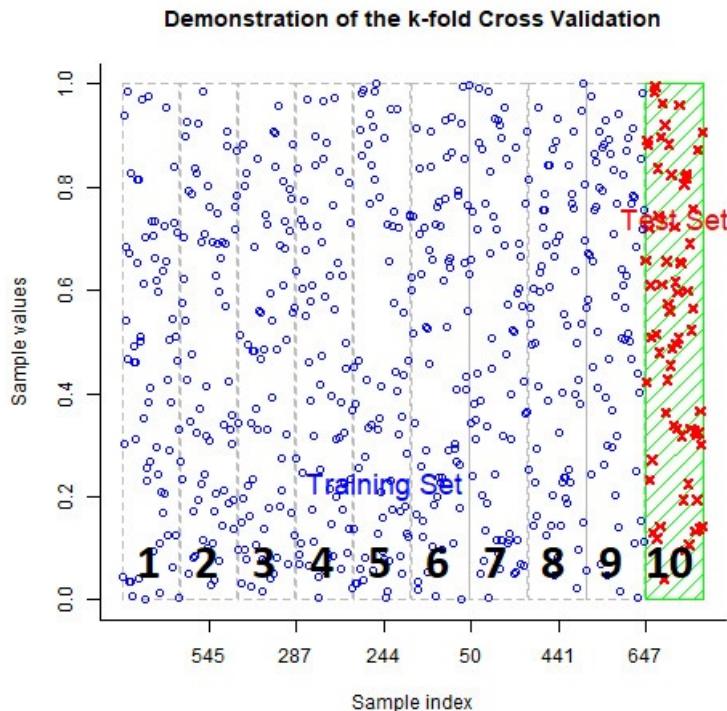
In the first round, Groups 1, 3, 4, 5, 6, 7, 8, 9, and 10 are used to train the model. The fit of the model is then tested by how closely it can predict the observed values in Group 2.



In the third round, Groups 1, 2, 4, 5, 6, 7, 8, 9, and 10 are used to train the model. The fit of the model is then tested by how closely it can predict the observed values in Group 3.



This continues until each group has been used to train and test the model.



A linear regression of the predicted test values on the actual test values is then conducted and the regression coefficient, R^2 , is calculated. The Root Mean Square Error is also calculated during the regression. You will recall the Root MSE is a measure of the standard deviation of the difference between the predicted and actual values.

13.3.6 Yield Prediction with Nearest Neighbor Analysis

As mentioned above, nearest neighbor analysis also works with qualitative data. In this case, the values of the nearest neighbors are averaged for the variable of interest. In R, there are various functions for running a nearest-neighbor analysis. The output below was generated using the *caret* package, a very powerful tool which fits and cross-validates models at once.

```
## k-Nearest Neighbors
##
## 665 samples
##    7 predictor
##
## Pre-processing: centered (7), scaled (7)
```

```

## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 599, 599, 598, 599, 597, 598, ...
## Resampling results across tuning parameters:
##
##   k    RMSE      Rsquared     MAE
##   5   11.04766  0.8129430  7.996958
##   7   10.99521  0.8177150  8.001320
##   9   11.07555  0.8157225  8.147071
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 7.

```

In the above output, the following are important:

- Resampling: this line tells us how the model was cross-validated (10-fold cross validation)
- Identification of the final value used for the model: Remember when we discussed how to select the appropriate k (number of neighbors)? R compares multiple values of k . In this case, it identifies $k=7$ as optimum for model performance.
- Model statistics table. Statistics for three levels of k are given: 5, 7, and 9. RMSE is the root means square error, Rsquared is R^2 , the regression coefficient, and MAE is Mean Absolute Error (an alternative calculation of error that does not involve squaring the error).

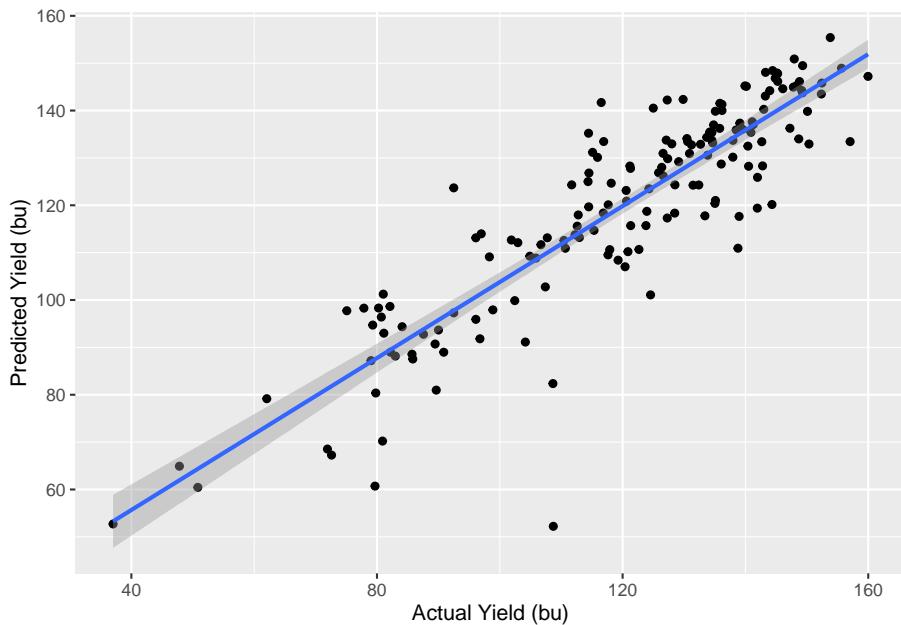
Looking at the table results for $k=7$, we see the model had an R^2 of almost 0.81. This is pretty impressive, given that were predicting mean county corn yield using relatively few measures of soil and environment. The RMSE is about 11.1. Using this with our Z-distribution, we can predict there is a 95% chance the actual corn yield will be within $1.96 \times 11.1 = 21.8$ bushels of the predicted yield.

In this example, as you might have suspected, we did know the corn yield in the counties where it was “missing”; we just selected those counties at random and used them to show how k-nearest-neighbors worked.

So let’s use the our nearest-neighbor model to predict the yields in those “missing” counties and compare them to the actual observed yields. We can use a simple linear regression model for this, where:

$$\text{Predicted Yield} = \alpha + \beta \cdot \text{Actual Yield}$$

Below, the predicted yields (y-axis) are regressed against the actual yields (x-axis).

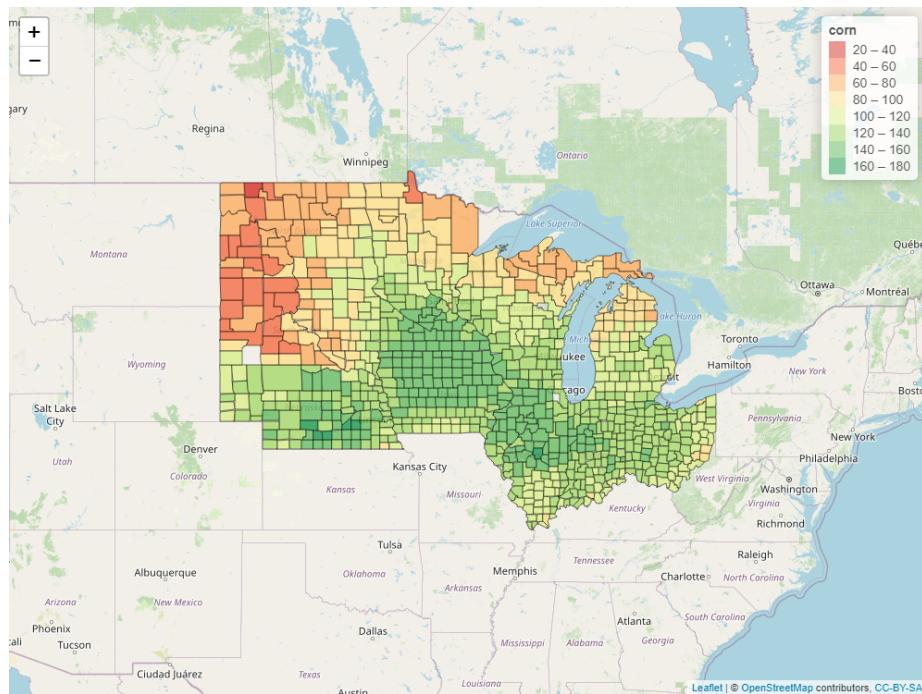


We can see the predicted and actual yields are strongly correlated, as evidenced by their linear distribution and proximity to the regression line. The regression model statistics are shown below.

r.squared	sigma	p.value
0.785811	10.35285	0

Above are select statistics for our model. We see the r.squared, about 0.83 is even better greater than in our cross-validation. Sigma, the standard deviation, is smaller than in the validation, barely 9 bushels. Finally, we see the significance of our model is very, very small, meaning the relationship between predicted and actual yield was very strong.

Finally, we can fill in the missing county yields in the map with which we started this section. We see the predicted yields for the missing county seem to be in line with their surrounding counties; there are no islands of extremely high or extremely low yields. You can see the yield of each county by clicking on it.



13.4 Classification Trees

The last data science tool we will learn in this unit the classification tree. The classification tree uses a “divide and conquer” or sorting approach to predict the value of a response variable, based on the levels of its predictor variables.

The best way to understand how a classification tree works is to play the game Plinko. In Plinko, a player drops disc at the top of a table studded with pins.



The disc falls down the table, bouncing to one side or the other of each successive pin.







The bin determines what fabulous prize the player may have won.



Each level of the decision tree can be thought of as having pins. Each pin is a

division point along variable of interest. If an observation has a value for that variable which is greater than the division point, it is classified one way and moves to the next pin. If less, it moves the other way where it encounters a different pin.

13.4.1 Features

Classification trees are different from cluster and nearest neighbor analyses in that they identify the relative importance of *features*. Features include every predictor variable in the dataset. The higher a feature occurs in a classification tree, the more important they are in predicting the final value.

Classification trees therefore shed a unique and valuable insight into the importance of variables, distinct from all other analyses we have learned. They provide insight into *feature importance*, not just their significance as a predictor.

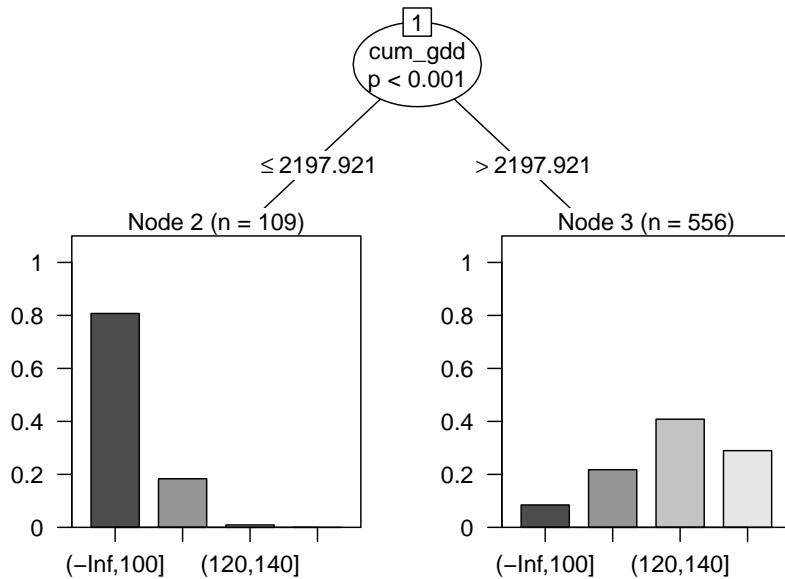
13.4.2 Quantitative (Categorical) Data

A classification tree predicts the category of the response variable. We will start with yield categories, just as we did with the nearest neighbors section, because the results are intuitively easier to visualize and understand.

The classification process begins with the selection of a root. This is the feature that, when split, is the best single predictor of the response variable. The algorithm not only selects the root; it also determines where within the range (the threshold) of that feature to divide the populations.

What is the criterion for this threshold? It is the value that divides the population into two groups in which the individuals within the group are as similar as possible.

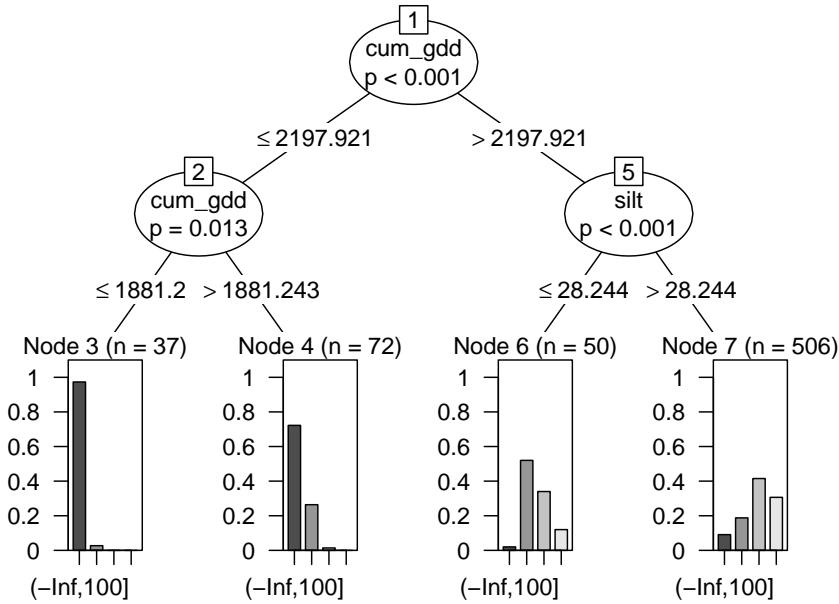
These results are displayed as a tree; thus, the “classification tree”. The results of our first split are shown below:



Cumulative growing degree days (cum_gdd) was selected as the root, and the threshold or split value was approximately 2198 GDD. Counties where the cumulative GDD was less than this threshold tended to have yields in the lower two classes; above this threshold, counties tended to have yields in the greater two classes.

This particular algorithm also calculates a p-value that the distribution of the counts among the yield classes is different for the two groups. We see the difference between the two groups (greater vs less than 2198 GDD) is highly significant, with a p-value < 0.001 .

Next, let's add another level to our tree. The two subgroups will each be split into two new subgroups.



Our tree now has three parts. The root was discussed before. At the bottom of the tree, we have four nodes: Node 3, Node 4, Node 6, and Node 7. These represent the final groupings of our data.

In the middle, we have two branches, labelled as branches 2 and 5 of our tree. Each branch is indicated by an oval. The branch number is in a box at the top of the oval.

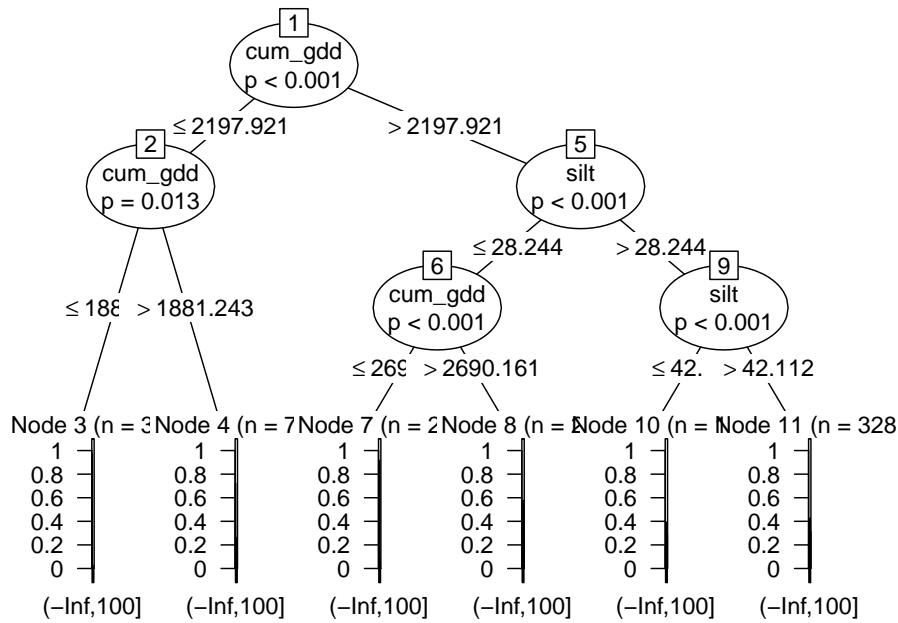
Counties that had *cum_gdd* < 2198 were divided at branch 2 by *cum_gdd*, this time into counties with less than 1881 *cum_gdd* and counties with more than 1881 *cum_gdd*. Almost all counties with less than 1881 *cum_gdd* had yields in the lowest class. Above 1881 *cum_gdd*, most yields were still in the lowest class, but some were in the second lowest class.

Counties with greater than 2198 *cum_gdd* were divided at branch 5 by soil percent *silt*. Counties with soils that with less than 28.2 percent slit had yields that were mostly in the middle two yield classes. Above 28.2 percent silt, most yields were in the upper two classes.

The beauty of the regression tree is we are left with a distinct set of rules for predicting the outcome of our response variable. This lends itself nicely to explaining our prediction. Why do we predict a county yield will be in one of the two highest classes? Because the county has cumulative GDDs greater than 2198 and a soil that has a percent silt greater than 28.2

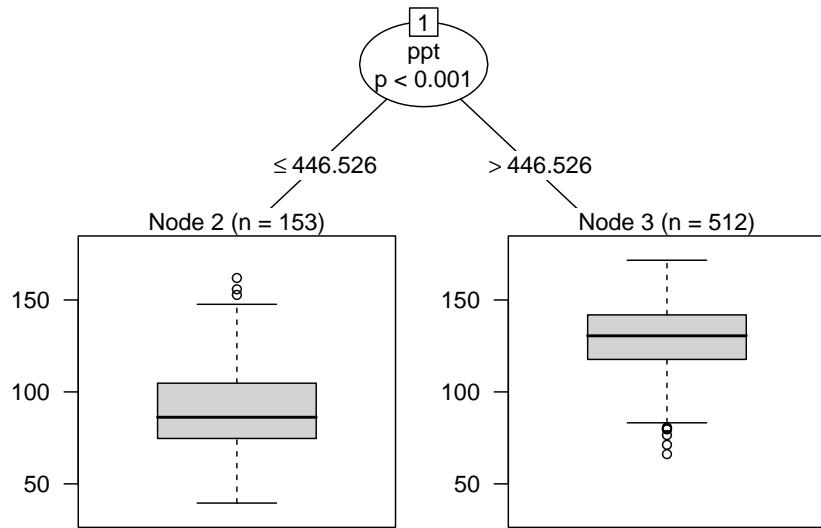
The algorithm will continue to add depths to the tree until the nodes are as homogeneous as possible. The tree, however, will then become too complex to

visualize. Below we have three levels to our tree – we can see how difficult it is to read.



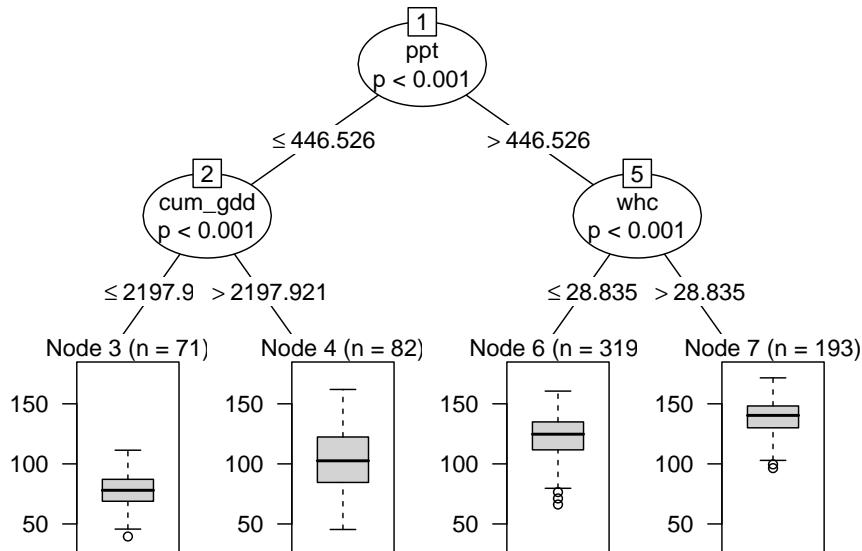
13.4.3 Quantitative (Continuous) Data

The same process above can be used with quantitative data – in this example, the actual yields. When we use yield, the root of our tree is not cumulative growing degree days, but total precipitation. Instead of a bar plot showing us the number of individuals in each class, we get a boxplot showing the distribution of the observations in each node. When we work with quantitative data, R identifies the feature and split that minimizes the variance in each node.



When we add a second level to our tree, we see that observations where precipitation was less than or equal to about 446.5 mm/season, the data was next split by cum_gdd. Counties with a cum_gdd less than about 2198 had lower yields than counties with a cum_gdd $>$ 2198.

Where ppt was greater than 446 mm per season, individuals were next split by water holding capacity (whc). Counties where whc was greater than about 28.8 cm/cm soil had greater yield than counties with less water holding capacity.



Like before, we can continue adding branches, although our virtual interpretation of the data will become more challenging.

13.4.4 Overfitting

In nearest-neighbors analysis, we learned the size of k (the number of neighbors) was a tradeoff: too few neighbors, and the risk of an outlier skewing the prediction increased. Too many neighbors, and the model prediction approached the mean of the entire population. The secret was to solve for the optimum value of k .

There is a similar tradeoff with classification trees. We can continue adding layers to our decision tree until only one observation is left in each node. At that point, we will perfectly predict each individual in the population. Sounds great, huh?

Yeah, but what about when you use the classification model (with its exact set of rules for the training population) on a new population? It is likely the new population will follow the more general rules, defined by the population splits closer to the root of the tree. It is less likely the population will follow the specific rules that define the nodes and the branches immediately above.

Just as a bush around your house can become messy and overgrown, so can a classification tree become so branched it loses its value. The solution in each case is to prune, so that the bush regains its shape, and the classification tree approaches a useful combination of general application and accurate predictions.

13.4.5 Cross Validation

Just as with nearest-neighbors, the answer to overfitting classification trees is cross-validation. Using the same cross-validation technique (10-fold cross-validation) we used above, let's fit our classification tree.

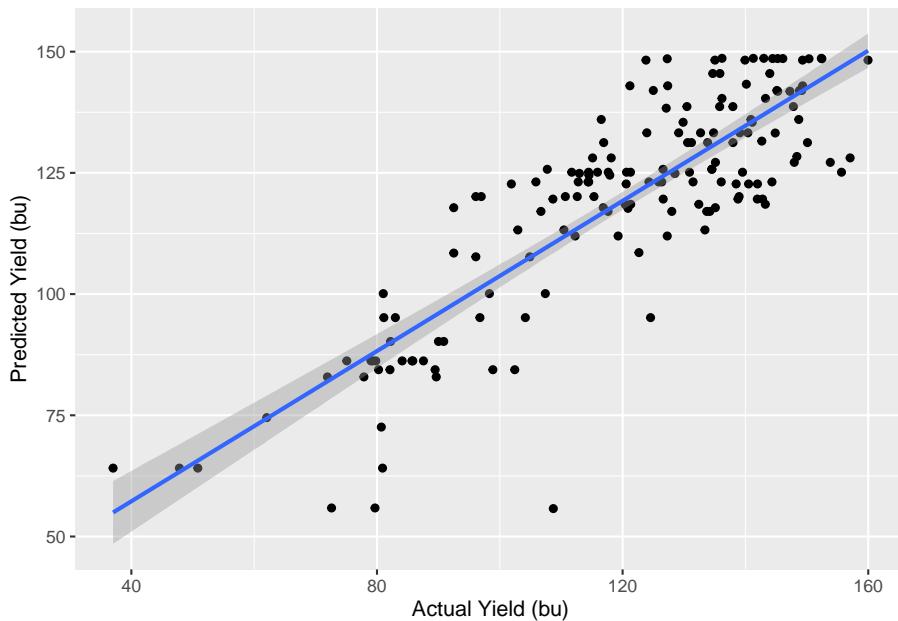
```
## Conditional Inference Tree
##
## 665 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 598, 599, 599, 599, 598, 598, ...
## Resampling results across tuning parameters:
##
##   mincriterion  RMSE      Rsquared     MAE
##   0.01          13.05396  0.7412480  9.689862
##   0.50          13.20743  0.7342159  9.974611
##   0.99          15.13028  0.6567629  11.933852
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mincriterion = 0.01.
```

The output looks very similar to that for nearest neighbor, except instead of k we see a statistic called `mincriterion`. The `mincriterion` specifies the maximum p-value for any branch to be included in the final model.

As we discussed above, the ability of the classification tree to fit new data decreases as the number of branches and nodes increases. At the same time, we want to include enough branches in the model so that we are accurate in our classifications.

In the output above, the `mincriterion` was 0.01, meaning the p-value must be no greater than 0.01. This gave us an `Rsquared` of about 0.745 and and `Root MSE` of about 13.0.

As with the k -Nearest-Neighbors analysis, we can compare the predicted yields for the counties dropped from our model to their actual values.



```
## # A tibble: 1 x 3
##   r.squared sigma  p.value
##       <dbl> <dbl>    <dbl>
## 1     0.720  12.0 1.39e-46
```

We see the model strongly predicts yield, but not as strongly as k -Nearest Neighbors.

13.4.6 Random Forest

There is one last problem with our classification tree. What if the feature selected as the root of our model is the best for the dataset to which we will apply the trained model? In that case, our model may not be accurate. In addition, once a feature is chosen as the root of a classification model, the features and splits in the remainder of the model also become constrained.

This not only limits model performance, but also gives us an incomplete sense of how each feature truly effected the response variable. What if we started with a different root feature? What kind of model would we get?

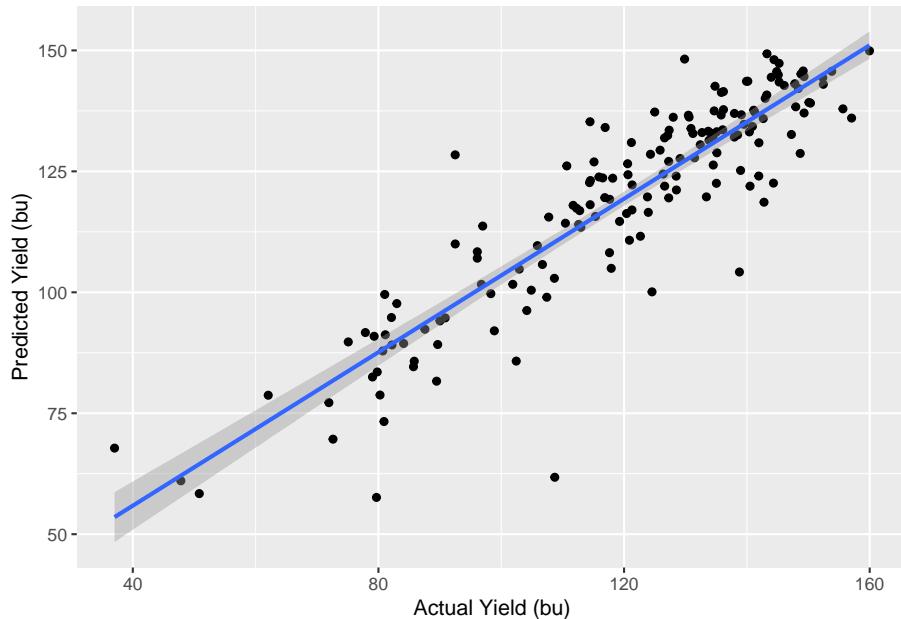
This question can be answered with an *ensemble* of classification trees. This ensemble is known as a *random forest* (i.e. a collection of many, many classification trees). The random forest generates hundreds of models, each starting with a randomly selected feature as the root. These models are then combined as an ensemble to make predictions for the new population.

```

## Random Forest
##
## 665 samples
##    7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 598, 600, 600, 601, 599, 597, ...
## Resampling results across tuning parameters:
##
##   mtry   RMSE      Rsquared     MAE
##   2      10.559385  0.8360096  7.868132
##   4      9.983132   0.8482014  7.327396
##   7      9.788704   0.8519592  7.159581
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 7.

```

If we compare our model results to the individual classification tree above, and to the performance of our nearest-neighbor model, we see the random forest provides a better fit of the data.



```

## # A tibble: 1 x 3
##   r.squared sigma  p.value
##       <dbl> <dbl>    <dbl>

```

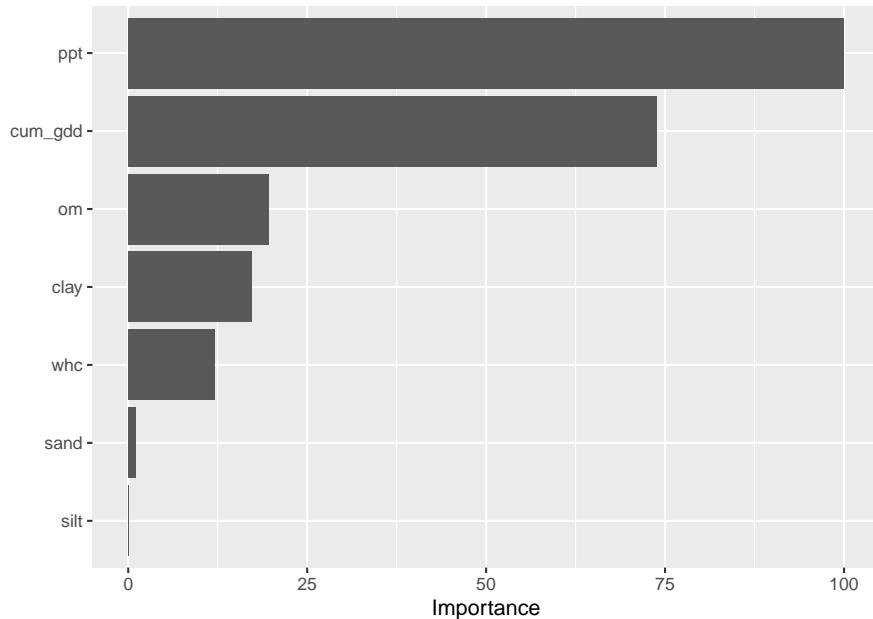
```
## 1      0.810  9.51 3.09e-60
```

13.4.7 Feature Importance

When we use random forest analysis, there is no one tree to review – the forest, after all, is a collection of many hundreds or thousands of trees. There is also no longer one set of rules to explain the predicted values, since all trees were used to predict corn yield.

Random forest analysis, however, allows us to rank the features by importance. Feature importance is calculated from the number of times a feature occurs in the top branches of tree and the most frequent level at which a feature first occurs in the model.

We see in the barplot below the feature importance in our random forest model. Precipitation was by far the single most important feature in predicting yield. The next most important feature, as we might expect, was cumulative growing degree days.



After that, we might wonder whether soil biology or texture might be the next most important feature. Our feature importance ranking suggests that organic matter is slightly more important than clay, followed by water holding capacity. We might infer that soil fertility which increases with organic matter and clay, was more important than drainage, which would be predicted by water holding capacity and our last variable, sand.

13.5 Summary

With that, you have been introduced to three (four, if we include random forest) powerful machine learning tools. I don't expect that you will personally use these tools on a regular basis; thus, this lesson is not hands-on like other lessons. It is most important that you are aware there is a whole world of statistics beyond what we have learned this semester, beyond designed experiments and the testing of treatments one-plot-at-a-time.

This is not to disparage or minimize the statistics we learned earlier in this course. Controlled, designed-experiments will always be key in testing new practices or products. They will be how products are screened in the greenhouse, in the test plot, in the side-by-side trial on-farm. They are critical.

But for big, messy problems, like predicting yield (or product performance) county-by-county, machine learning tools are very cool. They are behind the hybrid recommended for your farm, the nitrogen rate recommended for your sidedress. They are way cool and offer wicked insights and I believe it is worth your while – as agricultural scientists – to be aware of them.

Chapter 14

Putting it all Together

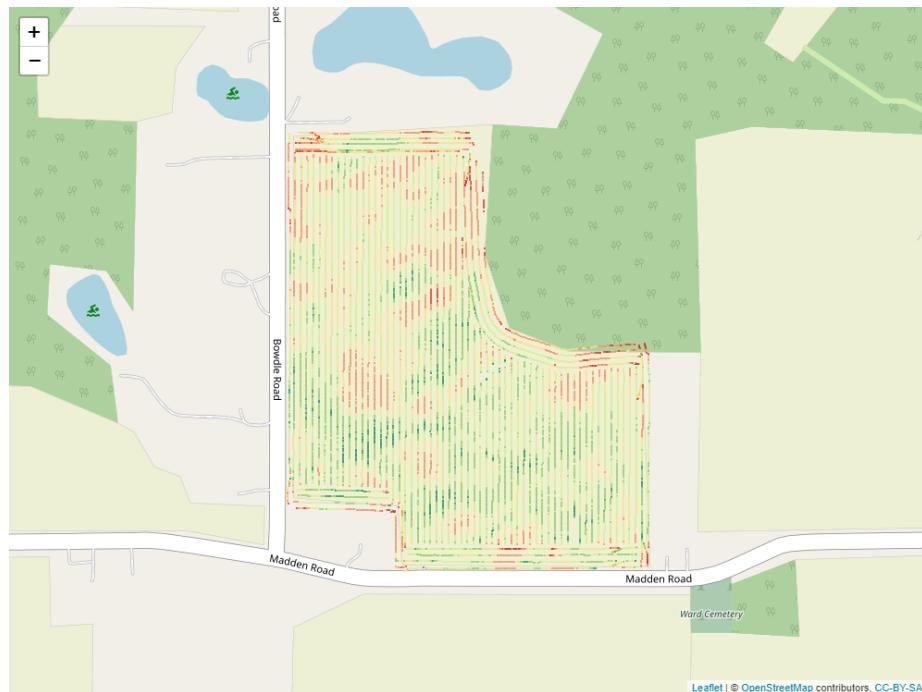
If I have done my job properly, your head is swimming with the different possibilities for collecting and analyzing data. Now it is time for us to back up and revisit the tests we have learned and understand how to choose a design or test for a given situation. Furthermore, it is important we round out this semester by understanding how to report results to others. This includes what to report and how to present it.

In contrast with more general statistics courses, I have tried to build this course around a series of trials which you are likely to conduct or from which you will use data. These range from a simple yield map through more complex factorial trials through nonlinear data and application maps. In this unit, we will review those scenarios and the tools we used to hack them.

14.1 Scenario 1: Yield Map (Population Summary and Z-Distribution)

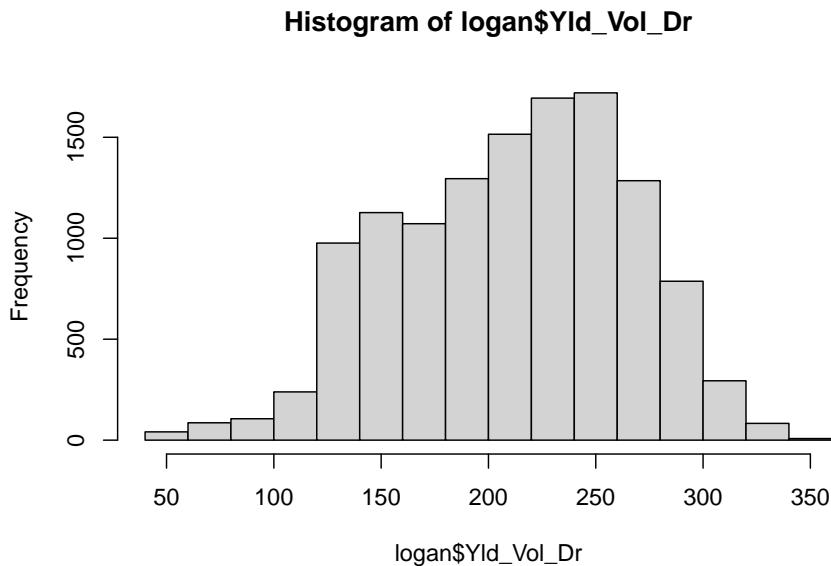
You are presented with a yield map and wish to summarize its yields so you can compare it with other fields. Since you have measured the entire field with your yield monitor, you are summarizing a population and therefore will calculate the population mean and standard deviation.

We started out the semester with a yield map, like this one:



In the map above, the higher-yielding areas are colored green, while the lower-yielding areas are orange to red. How would we summarise the yields above for others?

To start with, let's describe the center of the population. Should we use the mean or median? Either is appropriate if the data are normally distributed; if the data are skewed, the median will be a better measure of center. We can check the distribution using a histogram:



We inspect the data to see if it fits a normal distribution (“bell”) curve. This histogram is admittedly ugly, but is roughly symmetrical. We can also quickly inspect our distribution using the summary command on our yield variable.

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.
##    50.1   169.2  216.6  211.1  252.3  348.2
```

The median and mean are similar, so it would be acceptable to use either to describe the population center. Note the median is a little larger than the mean, so our data is skewed slightly to the right.

How would we describe the spread? That’s right: we would use the standard deviation:

```
## [1] 54.10512
```

Our standard deviation is about 54 bushels. What is the significance of this? We would expect 95% of the individuals (the yield measures) in this population to be within $1.96 \times 54.1 = 106.0$ \$, of the mean. Rounding our mean to 211, we would expect any value less than $211 - 106 = 105$ or greater than $211 + 105 = 316$ to occur rarely in our population.

14.2 Scenario 2: Yield Estimate (Sampling t-Distribution)

We are on a “Pro-Growers” tour of the Corn Belt or we are estimating yield on our own farm in anticipation of harvest or marketing grain.

Let’s take our field above and sample 20 ears from it. After counting rows around the ear and kernels per row, we arrive at the following 20 estimates.

```
## [1] 256 136 244 154 191 241 292 225 183 277 249 207 290 203 213 226 266 275 151
## [20] 181
```

The mean of our samples is 223 bushels per acre. We know this is probably not the actual mean yield, but how can we define a range of values that is likely to include the true mean for this field?

We can calculate a 95% confidence interval for this field. That confidence interval defines a fixed distance above and below our sample mean. Were we to repeat our sampling 100 times, in about 95 of our samples the population mean would be within our confidence interval.

Remember, our confidence interval is calculated as:

$$CI = \bar{x} + t_{\alpha, df} \times SE$$

Where \bar{x} is the sample mean, t is the t-value, based on our desired level of confidence and the degrees of freedom, and SE is the standard error of the mean. In this case, we desire 95% confidence and have 19 degrees of freedom (since we have 20 samples). The t-value to use in our calculation is therefore:

```
## [1] 2.085963
```

Our standard error is equal to the standard deviation of our samples.

```
## [1] 46.97144
```

Our standard error of the mean is our sample standard deviation, divided by the square root of the number of observations (20 ears) in the sample:

```
## [1] 10.50313
```

Our confidence interval has a lower limit of $223 - 2.09 * 10.5 = 201.1$ and an upper limit of $223 + 2.09 * 10.5 = 244.9$. We would present this confidence interval as:

(201.1, 244.9)

We know the combine map above the true population mean was 211.1, which is included in our confidence interval.

14.3 Scenario 3: Side-By-Side (t-Test)

You are a sales agronomist and want to demonstrate a new product to your customer. You arrange to conduct a side-by-side trial on their farm.

Knowing every field has soil variations, you divide a field into 8 pairs of strips. Each strip in a pair is treated either with the farmer's current system (the control) or the farmer's system *plus* the new product. You create the following paired plot layout.

plots	
802	treatment
801	control
702	treatment
701	control
602	control
601	treatment
502	treatment
501	control
402	treatment
401	control
302	control
301	treatment
202	control
201	treatment
102	treatment
101	control

We observe the following values in our trial:

block	treatment	yield
1	control	172.2
1	treatment	170.9
2	treatment	168.7
2	control	169.1
3	treatment	178.3
3	control	144.8
4	control	178.5
4	treatment	183.5
5	control	177.6
5	treatment	193.6
6	treatment	186.8
6	control	176.5
7	control	172.7
7	treatment	196.0
8	control	188.9
8	treatment	183.7

We run a t-test, as we learned in Units 4 and 5, which calculates the probability the difference between the control and treatment is equal to zero. Because we, as sales persons, are only interested in whether our treatment produces greater yield, we run a one-sided test. Our null hypothesis is therefore that the treatment produces yield equal to or less than the control. Our alternative hypothesis (the one we hope to confirm) is that the treatment yields more than the control.

```
##  
## Paired t-test  
##  
## data: yield by treatment  
## t = -2.1424, df = 7, p-value = 0.03469  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
##      -Inf -1.174144  
## sample estimates:  
## mean of the differences  
##                         -10.15
```

In the t-test above, we tested the difference when we subtracted the control yield from the treatment yield. We hoped this difference would be less than zero, which it would be if the treatment yield exceeded the control yield. We see the difference, -10.15, was indeed less than zero. Was it significant? Our p-value was 0.03, indicating a small probability that the true difference was actually zero or greater than zero.

We also see our confidence interval does not include zero or any positive values. We can therefore report to the grower that our treatment yielded more than the control.

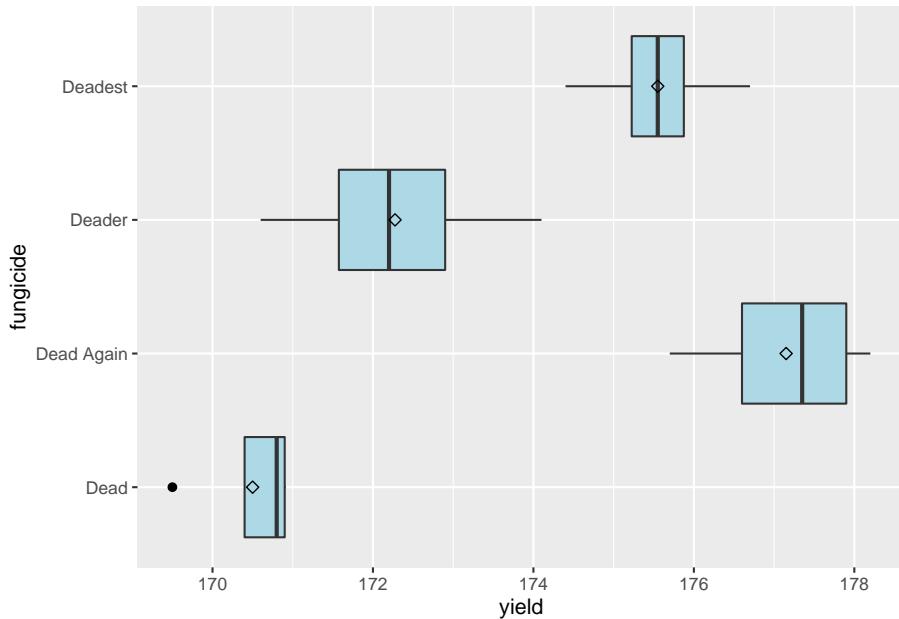
14.4 Scenario 4: Fungicide Trial (ANOVA CRD or RCBD)

We want to compare three or more fungicides that differ in their qualities, as opposed to their quantity.

We design and conduct a randomized complete block design experiment in which four fungicides are compared:

plot	block	fungicide	yield
101	1	Deadest	174.4
102	1	Dead	169.5
103	1	Dead Again	176.9
104	1	Deader	170.6
201	2	Dead Again	175.7
202	2	Deadest	175.5
203	2	Deader	171.9
204	2	Dead	170.9
301	3	Deader	172.5
302	3	Dead Again	178.2
303	3	Dead	170.9
304	3	Deadest	175.6
401	4	Deadest	176.7
402	4	Deader	174.1
403	4	Dead Again	177.8
404	4	Dead	170.7

We can begin our analysis of results by inspecting the distribution of observations within our treatment. We can take a quick look at our data with the boxplot we learned in Unit 9.



Hmmmm, fungicide “Dead” has an outlier. Shoots. Let’s look more closely at our situation. First, back to our field notes. We check our plot notes – nothing appeared out of the ordinary about that plot. Second, we notice the “Dead” treatment has a tighter distribution than the other three treatments: our outlier would not be an outlier had it occurred in the distribution of Deader. Finally, we note the outlier differs from from the mean and median of “Dead” by only a few bushels – a deviation we consider to be reasonable given our knowledge of corn production. We conclude the outlier can be included in the dataset.

We will go ahead and run an Analysis of Variance on these data, as we learned in Unit 8. Our linear model is:

$$Y_{ij} = \mu + B_i + T_j + BT_{ij}$$

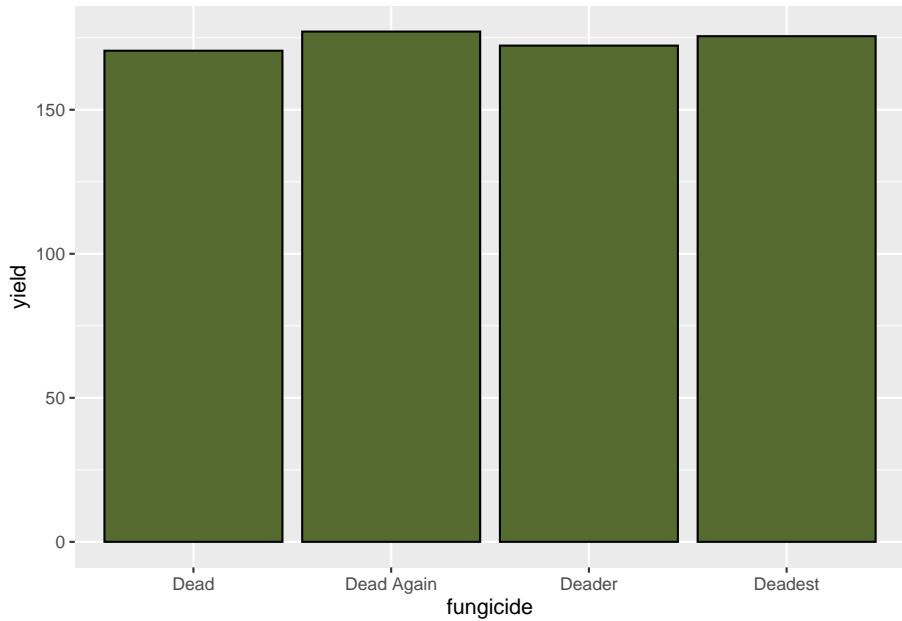
Where μ is the population mean, Y is the yield of the j_{th} treatment in the i_{th} block, B is the effect of the i_{th} block, T is the effect of the j_{th} treatment, and BT_{ij} is the interaction between block and treatment. This model forms the basis of our model statement to R:

```
##          Df Sum Sq Mean Sq F value    Pr(>F)
## block      3   9.10   3.03   5.535  0.0197 *
## fungicide  3 109.93  36.64  66.884 1.79e-06 ***
## Residuals  9   4.93   0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see the effect of the fungicide treatment is highly significant, so we will separate the means using the LSD test we learned in Unit 8.

```
## $statistics
##      MSerror Df     Mean       CV t.value      LSD
## 0.5478472  9 173.8688 0.4257045 2.262157 1.183961
##
## $parameters
##          test p.adjusted   name.t ntr alpha
## Fisher-LSD      none fungicide 4 0.05
##
## $means
##           yield      std.r      LCL      UCL    Min    Max    Q25    Q50
## Dead        170.500 0.6733003 4 169.6628 171.3372 169.5 170.9 170.400 170.80
## Dead Again 177.150 1.1090537 4 176.3128 177.9872 175.7 178.2 176.600 177.35
## Deader     172.275 1.4522970 4 171.4378 173.1122 170.6 174.1 171.575 172.20
## Deadest    175.550 0.9398581 4 174.7128 176.3872 174.4 176.7 175.225 175.55
##           Q75
## Dead        170.900
## Dead Again 177.900
## Deader     172.900
## Deadest    175.875
##
## $comparison
## NULL
##
## $groups
##           yield groups
## Dead Again 177.150     a
## Deadest    175.550     b
## Deader     172.275     c
## Dead        170.500     d
##
## attr(,"class")
## [1] "group"
```

Each of the fungicides produced significantly different yields, with “Dead Again” the highest-yielding. We can plot these results in a bar-plot.



14.5 Scenario 5: Hybrid Response to Fungicide Trial (ANOVA Factorial or Split Plot)

We want to test two factors within the same trial: three levels of hybrid (Hybrids “A”, “B”, and “C”) and two levels of fungicide (“treated” and “untreated”).

The treatments are arranged in a factorial randomized complete block design, like we learned in Unit 7.

plot	block	fungicide	hybrid	yield
101	R1	treated	B	199.1
102	R1	untreated	A	172.1
103	R1	untreated	B	187.4
104	R1	treated	C	204.1
105	R1	untreated	C	195.5
106	R1	treated	A	187.2
201	R2	treated	A	195.7
202	R2	untreated	C	200.3
203	R2	treated	B	200.3
204	R2	treated	C	216.5
205	R2	untreated	A	185.0
206	R2	untreated	B	188.4
301	R3	treated	B	209.4
302	R3	treated	A	198.2
303	R3	untreated	A	185.6
304	R3	untreated	C	203.6
305	R3	untreated	B	194.2
306	R3	treated	C	220.4
401	R4	untreated	B	198.7
402	R4	treated	B	214.4
403	R4	treated	C	219.2
404	R4	treated	A	194.8
405	R4	untreated	A	191.0
406	R4	untreated	C	199.6

Our linear additive model is:

$$Y_{ijk} = \mu + B_i + F_j + H_k + FH_{jk} + BFH_{ijk}$$

where Y_{ijk} is the yield in the i th block with the j th level of fungicide and the k th level of hybrid, μ is the population mean, B_i is the effect of the i th block, F_j is the effect of the j th level of fungicide, H_k is the effect of the k th level of hybrid, FH_{jk} is the interaction of the j th level of fungicide and k th level of hybrid, and BFH_{ijk} is the interaction of block, fungicide, and hybrid.

This translates to the following model statement and analysis of variance:

```

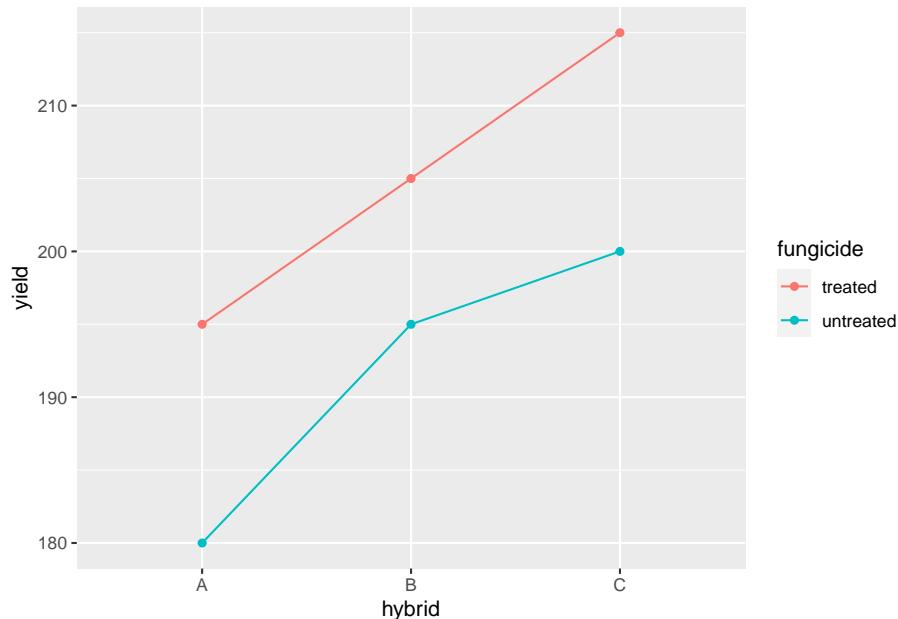
##                               Df Sum Sq Mean Sq F value    Pr(>F)
## block                         3  538.1   179.4   15.92 6.27e-05 ***
## fungicide                      1 1038.9  1038.9   92.22 8.49e-08 ***
## hybrid                          2 1403.4   701.7   62.29 5.43e-08 ***
## fungicide:hybrid               2   23.2     11.6    1.03    0.381
## Residuals                      15  169.0    11.3
## ---

```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Analysis of Variance results above, show that the main effects – fungicide and hybrid – are both highly-significant, but the interaction between fungicide and hybrid is insignificant. The line plot below allows us to further examine that interaction.

```
fung_hyb %>%
  group_by(fungicide, hybrid) %>%
  summarise(yield = mean(yield)) %>%
  ungroup %>%
  ggplot(aes(x=hybrid, y=yield, group=fungicide)) +
  geom_point(aes(color=fungicide)) +
  geom_line(aes(color=fungicide))
```



We can perform means separation on the data the same as we did for our analysis of variance in the previous example. Since fungicide only has two levels, its significance in the analysis of variance means the two levels (“treated” and “untreated”) are significant. To separate the hybrid levels, we can use the least significant difference test.

```
## $statistics
##   MSerror Df      Mean       CV t.value      LSD
##   11.26542 15 198.3625 1.692053 2.13145 3.576998
```

```

## 
## $parameters
##      test p.adjusted name.t ntr alpha
## Fisher-LSD      none hybrid   3  0.05
##
## $means
##      yield     std.r      LCL      UCL    Min    Max    Q25    Q50    Q75
## A 188.7000 8.305420 8 186.1707 191.2293 172.1 198.2 185.450 189.10 195.025
## B 198.9875 9.388966 8 196.4582 201.5168 187.4 214.4 192.750 198.90 202.575
## C 207.4000 9.777817 8 204.8707 209.9293 195.5 220.4 200.125 203.85 217.175
##
## $comparison
## NULL
##
## $groups
##      yield groups
## C 207.4000      a
## B 198.9875      b
## A 188.7000      c
##
## attr(,"class")
## [1] "group"

```

Our means separation results suggest the three hybrids differ in yield.

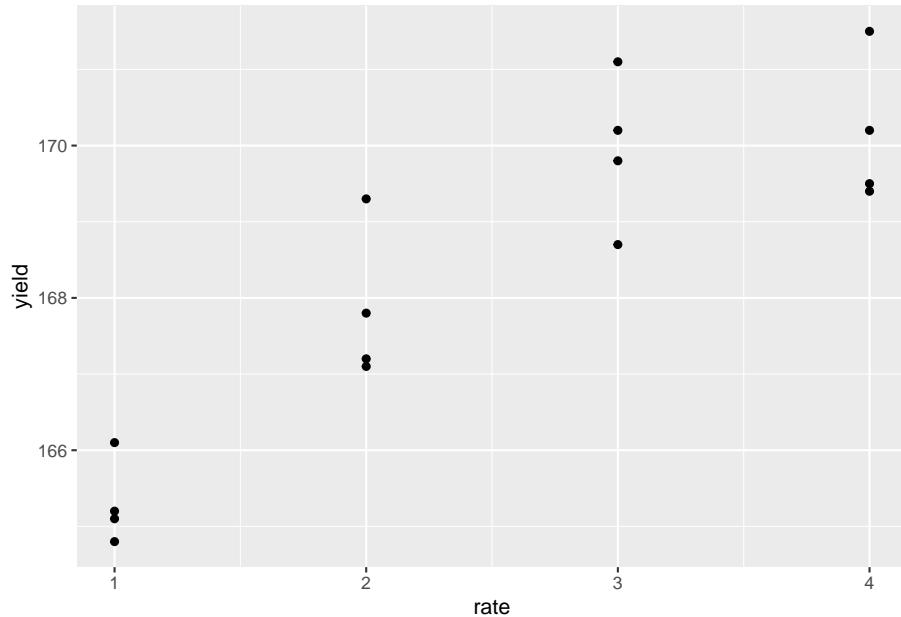
14.6 Scenario 6: Foliar Rate-Response Trial (Linear or Non-Linear Regression)

We want to model how the effect of a foliar product on yield increases with rate, from 1X to 4X.

The data are below:

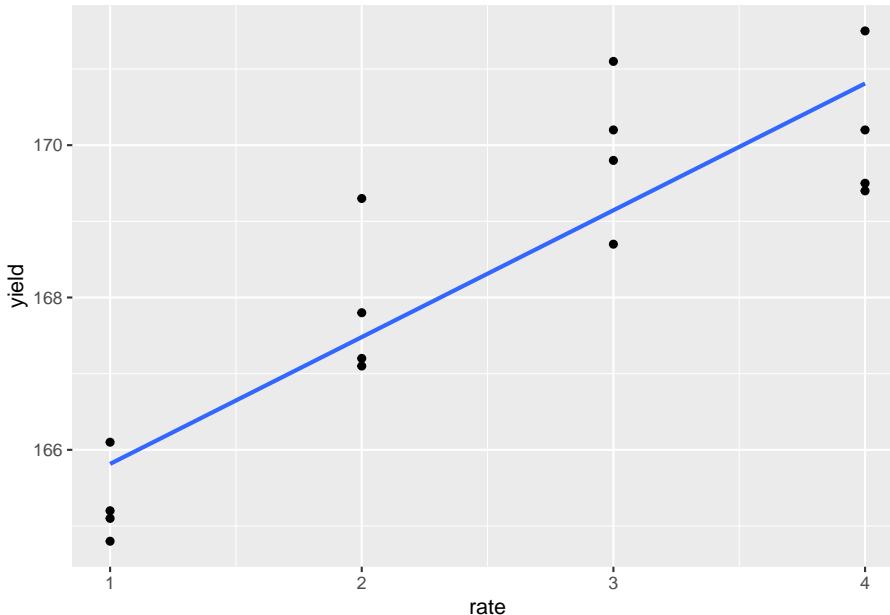
plot	block	rate	yield
101	1	2	167.1
102	1	3	168.7
103	1	4	169.5
104	1	1	165.2
201	2	2	167.2
202	2	1	164.8
203	2	4	169.4
204	2	3	170.2
301	3	2	167.8
302	3	4	170.2
303	3	1	166.1
304	3	3	169.8
401	4	2	169.3
402	4	3	171.1
403	4	4	171.5
404	4	1	165.1

We should start by plotting our data with a simple scatter plot so we can observe the nature of the relationship between Y and X. Do their values appear to be associated? Is their relationship linear or nonlinear?



The response appears to be nonlinear, but we first try to fit the relationship with simple linear regression, as we learned in Unit 10. Our regression line is plotted with the data below:

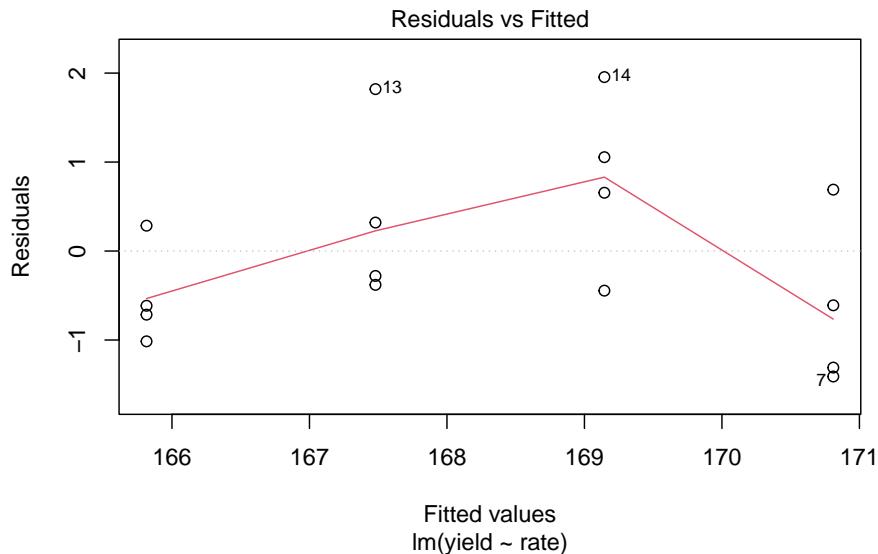
```
p + geom_smooth(method = "lm", se=FALSE)
```



We also run an analysis of variance on the regression, modelling yield as a function of rate, which produces the following results:

```
## 
## Call:
## lm(formula = yield ~ rate, data = foliar_final)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.4100 -0.6400 -0.3300  0.6637  1.9550 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 164.1500    0.6496 252.70 < 2e-16 ***
## rate        1.6650     0.2372   7.02 6.06e-06 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.061 on 14 degrees of freedom
## Multiple R-squared:  0.7787, Adjusted R-squared:  0.7629 
## F-statistic: 49.27 on 1 and 14 DF,  p-value: 6.057e-06
```

Not bad. The slope (rate effect) is highly significant and the $R^2 = 0.75$. To see whether the linear model was appropriate, however, we should plot the residuals.



We see, as we might expect, the residuals are not distributed randomly around the regression line. The middle two yields are distributed mostly above the regression line, while the highest and lowest yields are distributed mostly below the regression line.

A linear model is probably not the best way to model our data. Lets try, instead, to fit the data with a asymptotic model as we did in Unit 11. This model, in which the value of Y increases rapidly at lower levels of X, but then plateaus at higher levels of X, is often also referred to as a monomolecular function.

```
##
## Formula: yield ~ SSasymp(rate, init, m, plateau)
##
## Parameters:
##           Estimate Std. Error t value Pr(>|t|)
## init      171.1636    1.2699 134.789   <2e-16 ***
## m        159.7711    3.0120  53.044   <2e-16 ***
## plateau -0.4221    0.4880  -0.865     0.403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9104 on 13 degrees of freedom
##
```

14.7. SCENARIO 7: APPLICATION MAP (SHAPEFILES AND RASTERS)317

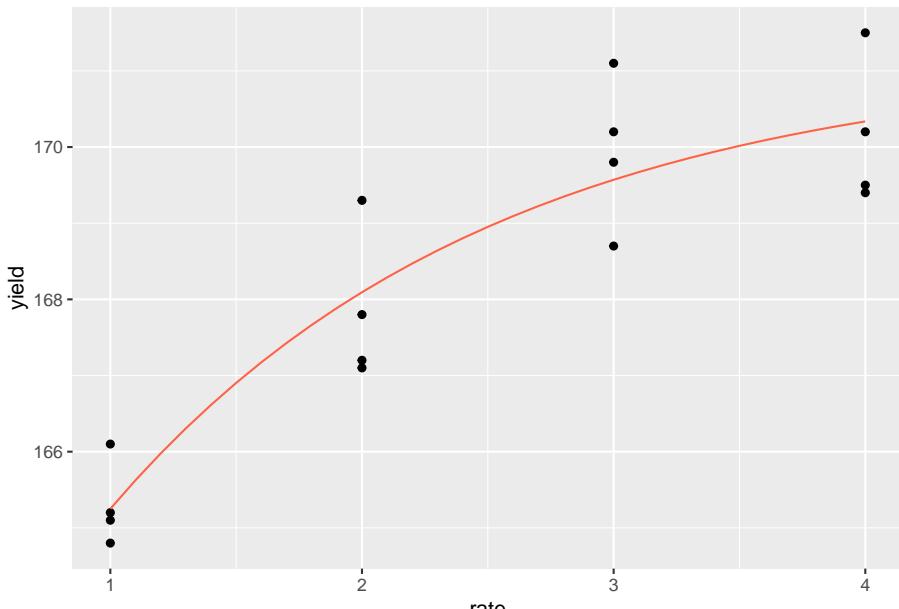
```
## Number of iterations to convergence: 0  
## Achieved convergence tolerance: 4.011e-06
```

We are successful in fitting our nonlinear model. To plot it with our data, however, we have to build a new dataset that models yield as a function of rate, using our new model.

To do this, we first create a dataset with values of rate from 1X and 4X, in increments of tenths.

We then can predict yield as a function of rate, using the new model.

Finally, we can plot our curve and visually confirm it fits the data better than our



linear model.

14.7 Scenario 7: Application Map (Shapefiles and Rasters)

We grid sample our field and want to visualize the results. We are particularly interested in our soil potassium results. We want to first visualize the point values, then create a raster map to predict potassium values throughout the field.

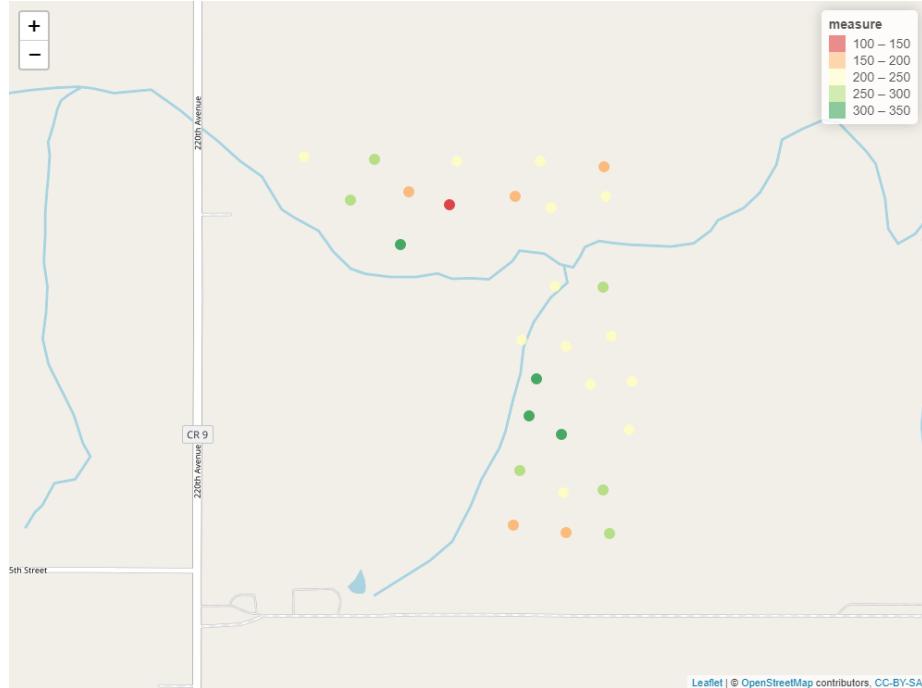
We start by reading in the shapefile with our results.

obs	Sample_id	SampleDate	ReportDate	P2	Grower	Field	attribute	measure
29	21	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3
30	22	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3
31	27	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3
32	5	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3
33	6	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3
34	8	10/26/2018	10/30/2018	0	Tom Besch	Folie N & SE	Om	3

Our next step is to filter our results to potassium ("K") only.

We want to color-code our results with green for the greatest values, yellow for intermediate values, and red for the lowest values. We can do this using the colorBin function.

We then create our map using the leaflet() function, just as we learned in Unit 12.



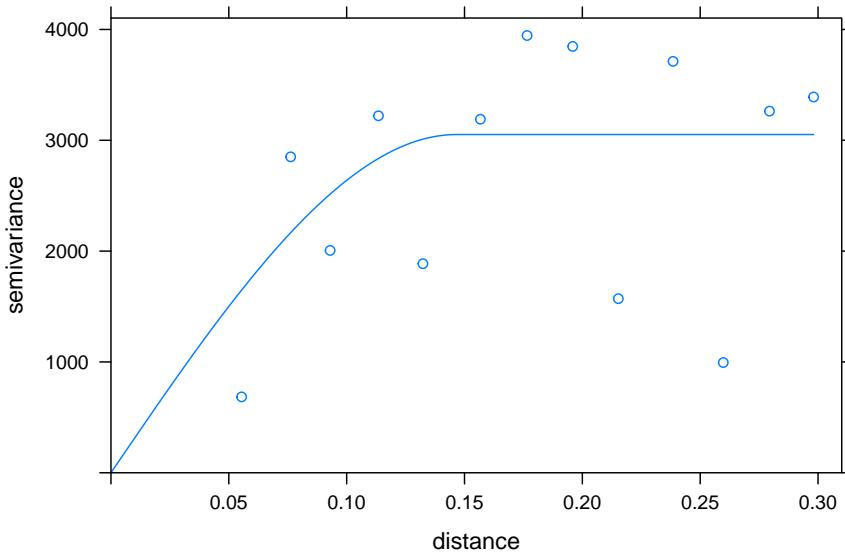
Next, we want to create a raster map that predicts soil potassium levels throughout our field. We first need to define the field boundary, which we do by loading a shapefile that defines a single polygon that outlines our field.

We then use that boundary polygon to create a grid. Each cell of this grid will be filled in when we create our raster.

Each cell in our raster that does not coincide with a test value will be predicted as the mean of the values of other soil test points. Since soils that are closer together are more alike than those that are farther apart, soil test points that

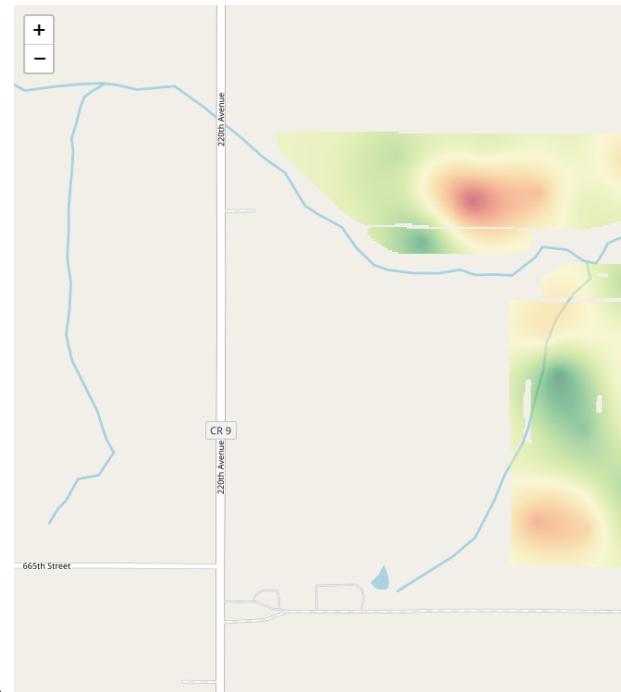
are closer to the estimated cell will be weighted more heavily in calculating the mean than those more distant. How should they be weighted?

To answer this, we fit a variogram to the data. The variogram describes how the correlation between points changes with distance.



We are now able to interpolate our data with kriging, which incorporates the results of our variogram in weighting the effect of sample points in the estimate of a cell value.

```
## [using ordinary kriging]
```



We finish by plotting our kriged data using leaflet().

14.8 Scenario 8: Yield Prediction (Multiple Linear Regression and other Predictive Models)

We want to predict the yield for a particular hybrid across 676 midwestern counties, based on over 300 observations of that hybrid.

We start by reading in the shapefile with our hybrid data.

```
## Reading layer `hybrid_data' from data source
##   `C:\ds_ag_professionals\data-unit-14\hybrid_data.shp' using driver `ESRI Shapefile'
## Simple feature collection with 4578 features and 4 fields
## Geometry type: POINT
## Dimension:      XY
## Bounding box:  xmin: -104.8113 ymin: 37.56348 xmax: -81.08706 ymax: 46.24
## Geodetic CRS:  WGS 84

## Simple feature collection with 6 features and 4 fields
## Geometry type: POINT
## Dimension:      XY
```

14.8. SCENARIO 8: YIELD PREDICTION (MULTIPLE LINEAR REGRESSION AND OTHER PREDICTIVE MODELS)

```
## Bounding box: xmin: -99.80072 ymin: 42.54919 xmax: -89.77396 ymax: 43.77493
## Geodetic CRS: WGS 84
##   book_name year attribute      value           geometry
## 1 ADAMS-MN 2016 bu_acre 215.4100 POINT (-92.26931 43.57388)
## 2 Adams-WI 2016 bu_acre 257.0930 POINT (-89.77396 43.77493)
## 3 Ainsworth-NE 2016 bu_acre 236.5837 POINT (-99.80072 42.55031)
## 4 Ainsworth-NE 2017 bu_acre 244.3435 POINT (-99.80072 42.55031)
## 5 Ainsworth-NE 2019 bu_acre 220.0456 POINT (-99.79738 42.54919)
## 6 Algona-IA 2016 bu_acre 216.0600 POINT (-94.28736 42.97575)
```

The attributes are organized in the long form so we will want to “spread” or pivot them to the wide form first.

```
## Simple feature collection with 6 features and 16 fields
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: -99.80072 ymin: 42.54919 xmax: -89.77396 ymax: 43.77493
## Geodetic CRS: WGS 84
##   book_name year bu_acre     clay      om prcp_0_500 prcp_1001_1500
## 1 ADAMS-MN 2016 215.4100 23.939049 149.7867      152        167
## 2 Adams-WI 2016 257.0930 9.899312 438.3402      140        112
## 3 Ainsworth-NE 2016 236.5837 11.701055 143.7150       97        63
## 4 Ainsworth-NE 2017 244.3435 11.701055 143.7150      148        16
## 5 Ainsworth-NE 2019 220.0456 11.701055 143.7150      234        22
## 6 Algona-IA 2016 216.0600 25.283112 480.4145       96        144
##   prcp_1501_2000 prcp_501_1000     sand      silt tmean_0_500 tmean_1001_1500
## 1             84            122 16.96170 59.09925    14.32143    21.91071
## 2            152            39 78.38254 11.71815    16.74286    22.30357
## 3             79            50 69.87984 18.41910    15.72857    22.38095
## 4            212            12 69.87984 18.41910    15.96428    26.11905
## 5            142            116 69.87984 18.41910   16.32143    23.82143
## 6             55            72 37.38857 37.32832   16.19286    21.26191
##   tmean_1501_2000 tmean_501_1000      whc           geometry
## 1      22.05952      21.45238 27.98555 POINT (-92.26931 43.57388)
## 2      21.83333      20.13095 15.67212 POINT (-89.77396 43.77493)
## 3      23.77381      24.00000 18.41579 POINT (-99.80072 42.55031)
## 4      21.34821      20.64286 18.41579 POINT (-99.80072 42.55031)
## 5      22.46429      21.69048 18.41579 POINT (-99.79738 42.54919)
## 6      22.89286      23.32143 33.21381 POINT (-94.28736 42.97575)
```

Similarly, lets load in the county data. Like the hybrid dataset, it is in long form, so lets again spread or pivot it to the long form so that the attributes each have their own column.

```
## Reading layer `county_climates' from data source
```

```

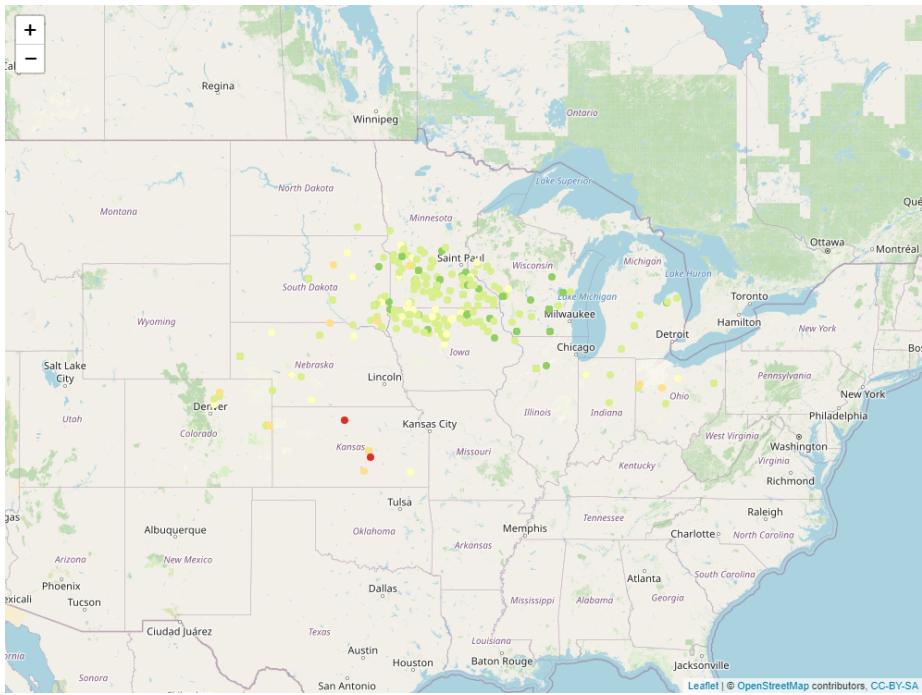
##  `C:\ds_ag_professionals\data-unit-14\county_climates.shp' using driver `ESRI Shapefile'
## Simple feature collection with 8788 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -106.1954 ymin: 36.99782 xmax: -80.51869 ymax: 47.24051
## Geodetic CRS: WGS 84

## Simple feature collection with 6 features and 19 fields
## Geometry type: MULTIPOLYGON
## Dimension: XY
## Bounding box: xmin: -105.6932 ymin: 38.61245 xmax: -102.0451 ymax: 40.26278
## Geodetic CRS: WGS 84
##   stco stt_bbv stt_fps           cnty_nm fps_cls stat_nm    clay      om
## 1 8001     CO     08      Adams County      H1 Colorado 21.88855 102.0773
## 2 8005     CO     08    Arapahoe County      H1 Colorado 23.87947 102.4535
## 3 8013     CO     08    Boulder County      H1 Colorado 25.83687 106.2404
## 4 8017     CO     08    Cheyenne County      H1 Colorado 23.67021 105.8340
## 5 8039     CO     08    Elbert County      H1 Colorado 22.00638 117.1232
## 6 8063     CO     08 Kit Carson County      H1 Colorado 24.78930 159.3644
##   prcp_0_500 prcp_1001_1500 prcp_1501_2000 prcp_501_1000      sand      silt
## 1    72.05263        44.15789        43.84211       31.68421 47.83218 30.27927
## 2    58.20000        43.20000        33.80000       17.00000 41.83093 34.28960
## 3   100.38462        47.84615        67.46154       58.07692 42.52472 31.63841
## 4    57.42105        54.52632        61.73684       37.26316 25.24085 51.08894
## 5    82.35714        57.64286        47.71429       49.28571 45.78082 32.21280
## 6    68.05263        54.26316        76.52632       43.36842 26.70159 48.50911
##   tmean_0_500 tmean_1001_1500 tmean_1501_2000 tmean_501_1000      whc
## 1    16.06967        24.17904        22.77083       21.65316 22.03449
## 2    15.64571        24.02857        22.46845       21.42024 21.34191
## 3    12.94062        18.60742        13.54932       19.00073 22.04795
## 4    17.02484        24.85072        24.18186       22.41526 30.14545
## 5    14.86390        22.27912        20.54928       20.41850 19.80896
## 6    16.74041        24.50439        23.56375       22.10329 26.92315
##   geometry
## 1 MULTIPOLYGON (((-105.0529 3...
## 2 MULTIPOLYGON (((-105.0534 3...
## 3 MULTIPOLYGON (((-105.0826 3...
## 4 MULTIPOLYGON (((-103.1725 3...
## 5 MULTIPOLYGON (((-104.6606 3...
## 6 MULTIPOLYGON (((-103.1631 3...

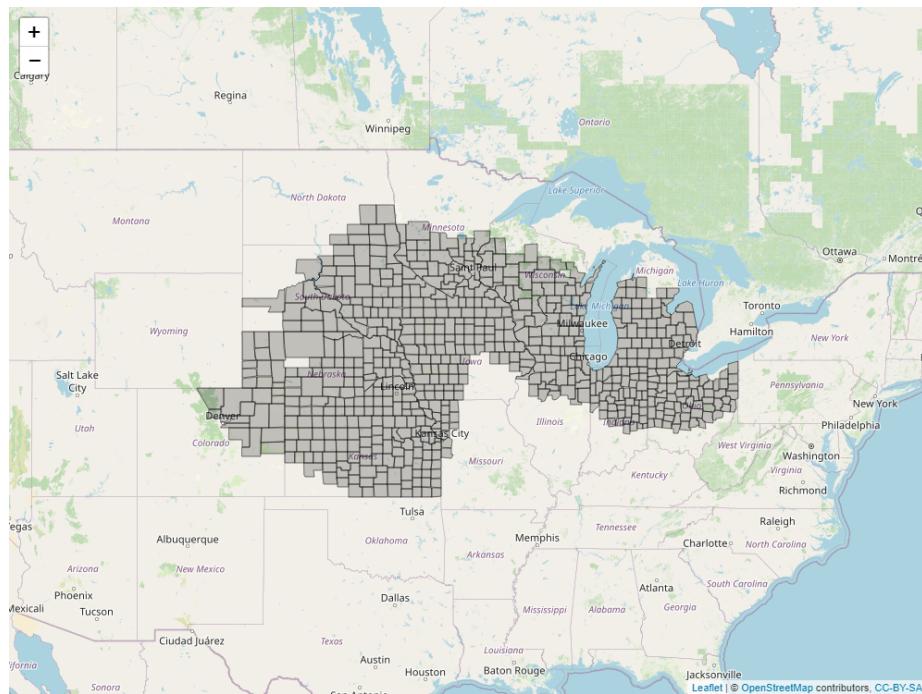
```

First, lets plot the locations of our hybrid trials:

14.8. SCENARIO 8: YIELD PREDICTION (MULTIPLE LINEAR REGRESSION AND OTHER PREDICTIVE MODELS)



We see our trials were conducted mostly in northern Iowa and southern Minnesota, but also in several other states. We will want to constrain our predictions to counties in this general area. That requires a few steps that we didn't cover this semester (the very curious can look up "convex hulls"). Our county dataset has been limited to the appropriate counties for our predictions:



Our next step is to develop our random forest model to predict yield. Recall that the predictor variables in a random forest model are called features. Our data has the following features:

attribute	description
clay	% clay
om	% organic matter
prcp_0_500	precip from 0 to 500 GDD
prcp_1001_1500	precip from 1001 to 1500 GDD
prcp_1501_2000	precip from 1501 to 2000 GDD
prcp_501_1000	precip from 501 to 1000 GDD
sand	% sand
silt	% silt
tmean_0_500	mean temp from 0 to 500 GDD
tmean_1001_1500	mean temp from 1001 to 1500 GDD
tmean_1501_2000	mean temp from 1501 to 2000 GDD
tmean_501_1000	mean temp from 501 to 1000 GDD
whc	water holding capacity

GDD are growing degree days accumulated from the historical average date at which 50% of the corn crop has been planted in each county. For example, prcp_0_500 is the cumulative precipitation from 0 to 500 GDD after planting. This would correspond with germination and seedling emergence.

14.8. SCENARIO 8: YIELD PREDICTION (MULTIPLE LINEAR REGRESSION AND OTHER PREDICTIVE MODELS)

We run our random forest the same as we learned in Unit 13. We have a couple of extraneous columns (`book_name` and `year`) in our `hybrid_wide` dataset. It is also a shapefile; we need to drop the geometry column and convert it into a dataframe before using it. We will do that first.

We will use 10-fold cross validation (indicated by the “`repeatedcv`” option in our `trainControl()` function below.)

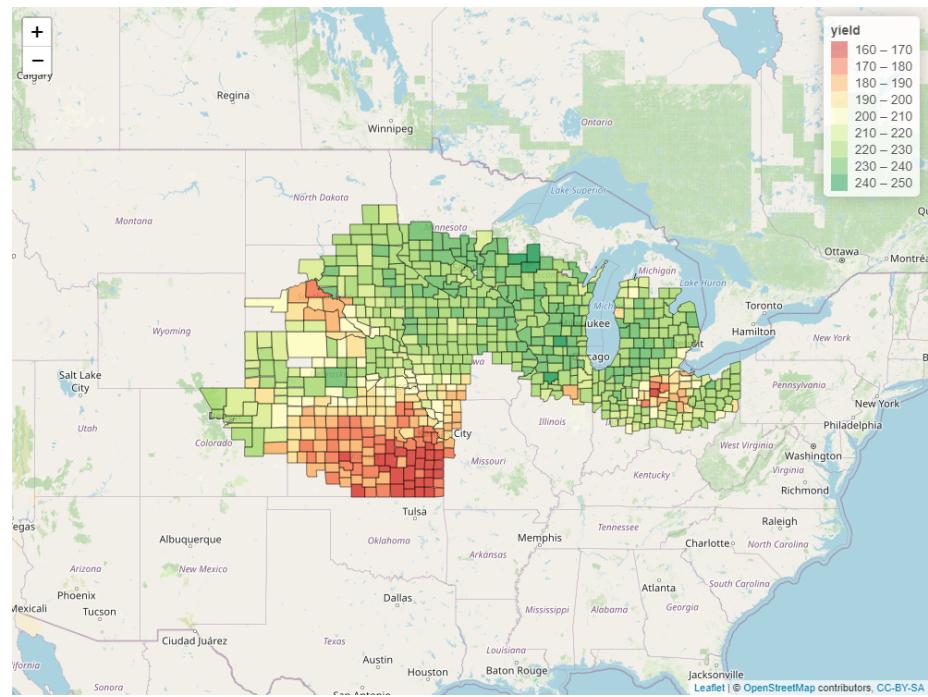
We can now fit our random forest model to the data.

```
## Random Forest
##
## 327 samples
## 13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 295, 294, 295, 294, 295, 295, ...
## Resampling results across tuning parameters:
##
##   mtry   RMSE     Rsquared    MAE
##   2      34.52569 0.1436353  25.86010
##   7      34.95068 0.1426655  26.24718
##  13     35.42862 0.1309118  26.53165
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.
```

These values aren’t great. The R^2 of the optimal model (top line, `mtry = 2`) is only about 0.16 and the root mean square error is above 33 bushels. For simplicity in this example, we left out several additional environmental features. We might consider adding those back in and re-running the model.

Nonetheless, let’s use our fit random forest model to predict yield across each county in our prediction space.

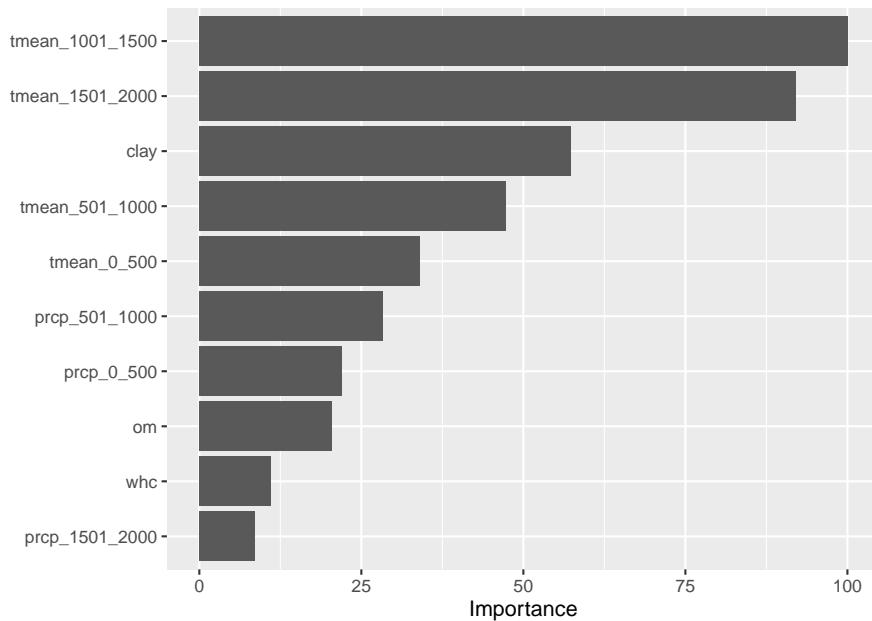
Finally, we can plot the results. We will again use a red-yellow-green scheme to



code counties by yield.

Our hybrid is predicted to yield best in northern Iowa, southern Minnesota, Wisconsin, and Northern Illinois. It is predicted to perform less well in Western Ohio and Kansas.

One last question: of our features, which had the greatest effect on the yield of this hybrid? We can answer that by running the `vip()` function with our model.



We see that mean temperature during later-vegetative (1001-1500 GDD) and reproductive (1501-2000 GDD) phases had the most effect in our model, followed by clay content.

14.9 Summary

And that is it for our whirlwind review of *Data Science for Agricultural Professionals*. While each scenario is discussed briefly, it is my hope that seeing the major tools we have learned side-by-side will give you a better sense where to start with your analyses.

For the sake of brevity, we didn't cover every possible combination of these tools (for example, you should also inspect data distributions and perform means separation when working with factorial trials as well as simpler, single-factor trials). Once you have identified the general analysis to use, I encourage you to go back to the individual units for a more complete "recipe" how to conduct your analysis.

In fact, feel free to as a "cookbook" (in fact, several R texts label themselves as such), returning to it as needed to whip up a quick serving of t-tests, LSDs, or yield maps. Few of us ever memorize more than a small amount of material. Better to remember where to look it up in a hurry.

I hope you have enjoyed this text or, at least, found several pieces of it that can support you in your future efforts. I welcome your feedback and suggestions

how it might be improved. Thank you.