Module 1: Introduction to Statistics

Section 1.1: Big idea

What is Statistics? Dictionary definitions

"The science of using information discovered from collecting, analyzing, and organizing numbers"

- Cambridge Academic Dictionary

"The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling."

- American Heritage Dictionary

"The science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

- Random House Dictionary

Less dry "definition"

- The world is incredibly complex. We want to learn about it anyway. So we collect observable information that we call "data".
- Statistics deals with the means by which we use that data in the pursuit of knowledge. It is imperfect and our answers will be imprecise. We do our best anyway.
- More precisely, **we use statistical methods to summarize and simplify the information contained in data.** We also use statistical methods to attempt to learn about the underlying mechanisms that created the data in the first place.

Uncertainty

- We use statistics to answer questions about things that are unknown, for example:
  - the effectiveness of a marketing strategy
  - the difference in % of men and women who voted for Trump

- Quantifying uncertainty is critical. We can use our data to estimate the value of an unknown quantity. We also use our data to estimate how wrong we think our estimate might be.

- Uncertainty will be a theme in this class. "We can't know for sure" will to some extent qualify the answers to all of our questions.

Some statistical claims, all containing uncertainty

- "I sleep seven hours per night, on average"
- "56% of likely voters intend to vote for the incumbent, with a margin of error of +/- 3%"
- "Students who attend class regularly tend to get higher grades than those who do not"
- "There is a 20% chance of snow tomorrow"

JMP

To explore these examples and more, we need some kind of software that will do computations for us. In this class, we will use the statistical computing software JMP (*pronounced "jump"*) to analyze data:

- Easy to use (no writing code by hand!)
- Reliable
- Excellent visualization

Example: Planets

On August 24, 2006, the International Astronomical Union voted that Pluto is not a planet. Some members of the public were reluctant to accept that decision. The data show a variety of facts about the 8 planets, Pluto, and 4 dwarf planets, including mean distance from the sun, number of moons, atmospheric composition, etc.

Using statistics we can explore the question, does Pluto behave like a planet?

Example: Brain waves

Researchers were interested in whether sensory deprivation over an extended period of time has any effect on the alpha-wave patterns produced by the brain. (Alpha-waves were thought to represent brain activity in an idle state.) To determine this, 20 inmates were randomly split into two groups of 10: members of one group were placed in solitary confinement and those in the other group remained in their cells. A week later, alpha-wave frequencies were measured for all subjects.

Using statistics, we can explore the researchers' question: does sensory deprivation have an effect on alpha-wave patterns?

Section 1.2: Variables

Variables and observations

- **Variables** are items (usually represented by letters or symbols) of interest which can take on different values.
- An **observation** is the person or thing the variables are measured on.

Example: Planets data in JMP

Distance from the sun, number of moons, etc. are variables ⇩

Each planet is an observation ⇨

	Planet	Diameter	Distance	Moons	Order	Atmosphere	Rings	Mass
1	Mercury	4878	0.39	0	1	none	no	0.055
2	Venus	12104	0.72	0	2	CO2	no	0.815
3	Earth	12756	1	1	3	N2pO2	no	1
4	Mars	6787	1.52	2	4	CO2	no	0.107
5	Jupiter	142800	5.2	79	5	H2pHe	yes	318
6	Saturn	120000	9.54	62	6	H2pHe	yes	95
7	Uranus	51118	19.18	27	7	H2pHe	yes	15
8	Neptune	49528	30.06	14	8	H2pHe	yes	17
9	Ceres	974.6	2.77	0	9	none	no	0.00016
10	Pluto	2300	39.44	5	10	CH4	no	0.002
11	Haumea	1518	43.34	2	11	none	no	0.0007
12	Makemake	1600	45.79	0	12	CH4	no	0.00067
13	Eris	2700	67.67	1	13	CH4	no	0.0028

This is how data are typically formatted:

- rows are observations
- columns are variables

Variable types

Variables can be broken into two types:

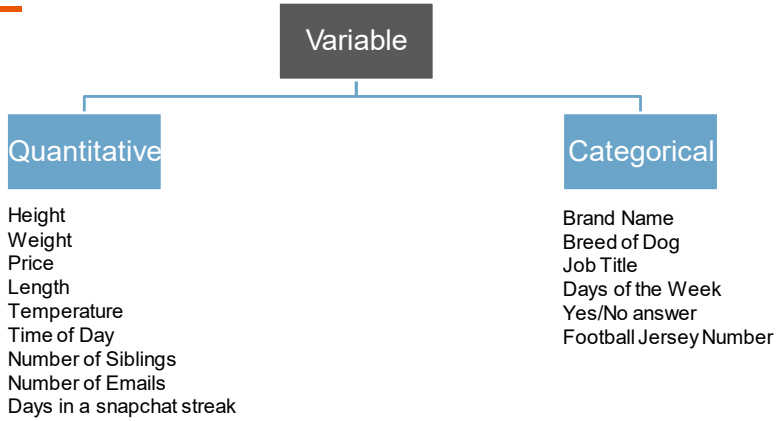
Quantitative:
Quantitative variables have values that we can do sensible math with.

Numbers which do not represent quantities are not quantitative.

Categorical:
Names or categories

Sometimes categorical variables are also referred to as qualitative.

Variable Type Examples



Example: Planets

Classify the variables from the planets data:

- Diameter (km)
- Distance (AU)
- Number of moons
- Order (from the Sun)
- Atmosphere type
- Rings (yes or no)
- Mass (Earth = 1)

iClicker: Classifying Variables

Total points a team scores in a basketball game is:

A. Quantitative
B. Categorical

Coffee sizes (Small/Medium/Large):

A. Quantitative
B. Categorical

iClicker: Classifying Variables

ZIP code is :

A. Quantitative
B. Categorical

Color is:

A. Quantitative
B. Categorical

Classifying a Variable

- As we've just seen, some variables can be measured in different ways. How a variable is measured will affect whether we consider it to be quantitative or categorical.
- Later we will see that what kinds of variables we have (quantitative/categorical) will change what graphs we can make and what statistical approaches to use.
- You should always think about how you are going to measure a variable BEFORE you go out and collect your data.

Classifying variables in JMP

- JMP lists variable names to the left of the data display.
- JMP distinguishes between quantitative and two different types of categorical data: ordinal (the categories are ordered in some way) and nominal (the categories aren't ordered).

Classifying variables in JMP

- We won't worry about distinguishing between ordinal and nominal in 201, but just so you are aware:

- *The blue triangle icon indicates quantitative data*
- *The green bars icon indicates ordinal data*
- *The red bars icon indicates nominal data*
- JMP will attempt to determine variable types / levels automatically. If you don't agree with JMP's choices, they can be changed manually.

Predictor and response variables

- Sometimes we want to use one or more variables (called predictors) to predict or explain another variable (called a response).
- Remember dependent and independent variables from math classes? Predictors are like independent variables and the response is like the dependent variable.

Predictor and response variables

- **Predictor variable:**
  Sometimes called the "independent variable", or the "explanatory variable".

  These are variables which we think will be useful in predicting or in explaining the response variable. There may be more than one predictor variable in a study.
- **Response variable:**
  Sometimes called the "dependent variable".

  This is the variable whose behavior we want to explain, or whose value we want to predict.
- *If* we are investigating a "cause and effect" relationship, then the predictor is the "cause" and the response is the "effect".

Example: Brain waves variables

- For the alpha-waves study, we have only two variables: one that denotes confinement (cell or confined) and alpha waves (measured for all 20 prisoners).
- Notice that each row is an observation, a prisoner who was either in confinement or a cell and whose alpha wave is shown in the column to the right.

	Confinement	Waves
1	cell	10.7
2	cell	10.7
3	cell	10.4
4	cell	10.9
5	cell	10.5
6	cell	10.3
7	cell	9.6
8	cell	11.1
9	cell	11.2
10	cell	10.4
11	confined	9.6
12	confined	10.4
13	confined	9.7
14	confined	10.3
15	confined	9.2
16	confined	9.3
17	confined	9.9
18	confined	9.5
19	confined	9
20	confined	10.9

Example: Brain waves variables

Recall that researchers were interested in whether sensory deprivation has any effect on the alpha-wave patterns produced by the brain.

Response variable and type:

Predictor variable and type:

Section 1.3: Populations and samples

Populations

- A population is the entire overall group we are interested in.
- If we are trying to get an idea of the average height of US adult females, then the population of interest is all US adult females.
- Populations can be large or small. Usually they are large.

Small population examples:

- Every student in this statistics class
- Business Students

Large population examples:

- Everyone between the ages of 18-25 in the US
- All adults

Samples

- A sample is a subset of the entire population that we collect data on. The variable(s) of interest is/are measured on each member of the sample.
- We take samples because the entire population of interest is usually not available to us (though when it is, we call this a census).
- For example, if we are studying the height of US adult females, we will measure the heights of a sample of women.

  We are not going to measure the height of every woman in the United States!

Example: Planets

What is the population? What is the sample?

Example: Brain waves

What is the population? What is the sample?

Parameters and Statistics

- A parameter is a numeric characteristic pertaining to a population. We never get to know the true value of a parameter. We can only try to estimate it.
- A statistic is any number you calculate using data. We often use statistics to estimate parameters.

For example, if we use the average height of the women in our sample to estimate the average height of women in the country:

- the parameter of interest is: the average height of all US adult females
- the statistic we use is: the average height of our sample

Example: Brain waves

What is a parameter and a statistic we might be interested in?

Parameter:

Statistic:

Descriptive Statistics

- Descriptive statistics involve describing a dataset: we could make a graph of it, or tell you its average and how spread out it is. We could tell you any interesting features about it.
- However, in descriptive statistics, we limit ourselves to describing the data itself. **We do not generalize facts about the dataset to a larger group.**

Inferential Statistics

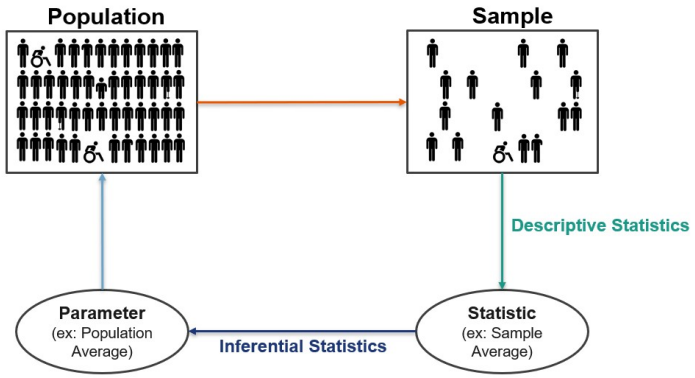
- If we generalize from a sample to a population, we are performing statistical inference.
- For example, we can take a sample of 100 women and use their average height to "draw inference" on the average height of women in the entire country. The average of our sample will be our best guess (or "estimate") for the average of the entire population.
- Statistical inferences always contain uncertainty. Our estimate for average height will be wrong! And we will have to decide how to deal with this uncertainty.

Putting it all together

So, in our example, we can take a **sample** of US women from the **population** of all US women and measure their height, which is the **variable** of interest. Each individual woman in our data set is considered an **observation**.

We can then calculate the average height of the women in our sample, which is our **statistic**. We use this statistic to draw **inference** on the average height of all women in the country, which is our **parameter**.

Putting it all Together



Example: social media activity

Let us say we want to judge the social media activity levels of college students. Students in this class take a two question survey asking their age and a rating of their self-assessed activity level on a 1 to 10 scale, defined as:

1 = "What is social media?" … 10 = "I actually speak in 140 characters at a time."

Identify the:

- Population
- Sample
- Variables
- Observations

Example: social media activity

Let us say we want to judge the social media activity levels of college students. Students in this class take a two question survey asking their age and a rating of their self-assessed activity level on a 1 to 10 scale, defined as:

1 = "What is social media?" … 10 = "I actually speak in 140 characters at a time."

- What parameter might we want to estimate?

- What resulting statistic might we find?

Drawing inference from sample to population

- After we complete a study, we often intend to use our results to make conclusions for a larger population. About what population is it valid to make an inference?

- General rule: we draw inference to the population from which we sample. Note: as we saw with the planets data sometimes the population isn't clearly defined or obvious.

Statistical inference: what to watch for

Statistical inference refers to the process of taking information from a sample and applying it to a population

Two major considerations in inference:

1. Is the sample representative of the population of interest? If there is bias in the study, it might not be.

2. Is the sample large enough for us to be confident in our inferences?

1. Sample Sizes

- If our sample is small, we can't confidently draw inference to the population.

- For example, if we measure the height of three women, their average height might not be a good estimate for the true average height of all women. Maybe we happened to sample three tall women.

- Smaller samples mean greater uncertainty when drawing inference to a population.

2. Representation in Our Sample

- After we complete a study, we often intend to use our results to make conclusions for a larger population.
- To what population is it valid to make an inference about?
- What population does our sample accurately represent?
- General rule: we draw inference to the population from which we sample. Note: sometimes this population isn't clearly defined or obvious.

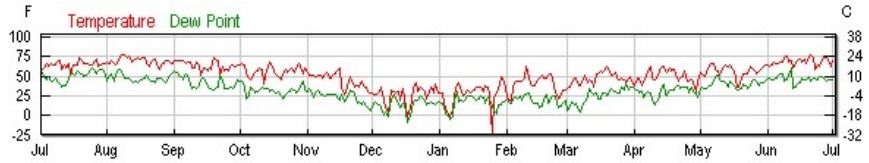
Section 1.4: Variation and Bias

The vital role of variation

- Data contains variation: our measurements will not all be identical.
- We would like this variation to be "explainable", e.g. knowing the value of the predictor variable tells us the value of the response variable.
- There will (nearly) always be "unexplained" variation in data: we won't be able to fully explain the reasons for the differences in measurements.
- Our goal is typically to explain as much variation as we can. But we also want to quantify the extent to which variation is not explained.

Example: temperature data

- Here is a plot of data (taken from wunderground.com) showing temperatures in Fort Collins between July 1st, 2016 and July 1st, 2017:



- What kind of variability is easy to explain here? What kind is difficult?

Example: temperature data

- Here is a similar plot, zoomed in to just temperatures between May 1st, 2017 and June 1st, 2017:

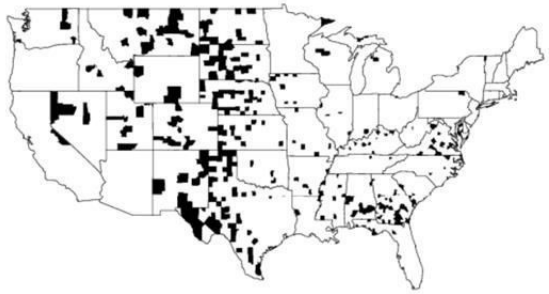
- Now what kind of variability is easy to explain? What kind is difficult?

Example: kidney cancer deaths by county

Here is a map showing the US counties in the top 10% for age-adjusted rates of kidney cancer deaths. Do you notice any pattern? Can you think of an explanation for what you see?

Example: kidney cancer deaths by county

Here is a map showing the US counties in the *bottom* 10% for age-adjusted rates of kidney cancer deaths. Do you notice any pattern? Can you think of an explanation for what you see?



Example: kidney cancer deaths by county

The maps seem very similar. Why would the patterns of counties with the top 10% and bottom 10% of kidney cancer death rates look so similar?

Bias

Bias occurs when a study is set up in such a way that its results will tend to be systematically wrong (as opposed to just wrong because of random chance and inherent uncertainty).

Sources of bias we will discuss:

- How data is collected from a population
- The amount of data collected
- The placebo effect
- Researcher expectations
- Confounding variables

Sampling bias

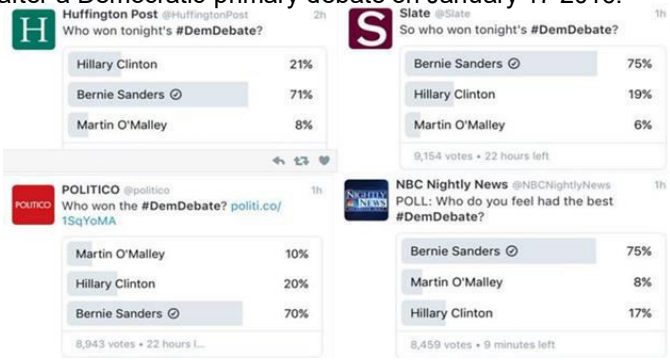
- Sampling bias can occur when a sample is taken in such a way that we would expect it to differ systematically from the population of interest.
- For example, if we are estimating the average height of all CSU students but only sample the basketball team our estimate might be biased upwards.

Self selection bias

- Self-selection bias can occur when choose if they want to be included in a sample. If the reason for their choosing to be in the sample is related to what is being measured, bias can result.
- Example: online reviews on websites like amazon.com or yelp.com
  - Often you will see that most reviews are either 5 stars or 1 star.
  - Who is most motivated to write a review? Those who feel lukewarm, or those who feel passionate?

Online polls after a political debate

- We also see self-selection bias in online political polls. These are from right after a Democratic primary debate on January 17 2016:



Scientific polls from the same week...

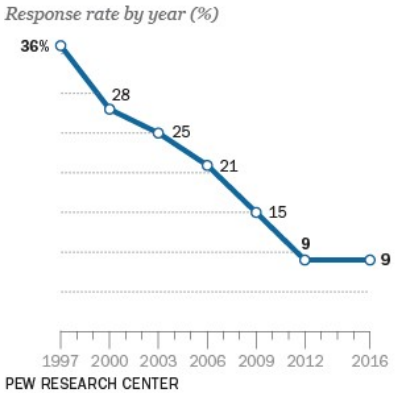
POLLSTER	DATES	POP.	CLINTON	SANDERS	O'MALLEY
Monmouth University NEW	1/15 - 1/18	352 RV	52	37	2
Morning Consult NEW	1/14 - 1/17	1,805 RV	54	30	2
NBC/SurveyMonkey NEW	1/11 - 1/17	3,259 RV	52	36	1
Ipsos/Reuters	1/9 - 1/13	696 A	54	35	4
NBC/WSJ	1/9 - 1/13	400 LV	59	34	2
YouGov/Economist NEW	1/9 - 1/11	620 LV	58	33	3
Gravis Marketing/One America News	1/10 - 1/10	890 LV	65	26	9
Morning Consult	1/8 - 1/10	909 RV	49	32	3
CBS/Times	1/7 - 1/10	389 LV	48	41	2
NBC/SurveyMonkey	1/4 - 1/10	2,619 RV	52	37	2

Non-response bias

- Non-response bias can occur when certain types of respondents are less likely to answer a survey, and the reason they don't respond is related to the variable being studied.
- Example:consider a survey designed to estimate the percentage of a city's population who are eligible to receive income-based social services.
  - The people most in need of social services are also the least likely to have permanent addresses or phone numbers, and are thus more difficult to reach via survey than people not in need of social services.

Discussion question: survey non-response

- This image shows the decline in response rates to telephone surveys conducted by the Pew Research Center
- Question: can you think of a good way to take a survey, in which non-response bias shouldn't be a problem?



Simple random sample

- To mitigate bias, samples should be collected randomly. A simple random sample is a sample of the population where every unit has an equal opportunity to be selected, as in drawing names from a hat.
- Random samples are more likely to be representative of the population of interest than non-random samples. This does not mean that a random sample is ALWAYS representative of the population. (maybe, just by chance, we sampled lots of basketball players!) It just means that we are not introducing bias via the way we collect the sample.
- Note that taking a SRS might be difficult. How does a polling firm make sure that every likely voter has an equal chance of being sampled?

Why sample randomly?

- Self-selection bias can be overcome by making sure to select the members of your sample randomly. If your sample is taken randomly, then people do not get to choose to be a part of it.
- Sampling bias can also be overcome via random sampling from the whole population. For instance, if we take a SRS of CSU students' heights, chances are all of them won't be basketball players.
- Non-response bias cannot be eliminated, though there are tools for detecting it and attempting to correct for it.

Blinding

- A blinded study is one in which participants do not know which group they are in, or which treatment is which.
- In the context of a taste test, the study is blinded if the cups are not labeled and the participants do not know which cup contains which product.

https://commons.wikimedia.org/wiki/File:Blind_taste_test.jpg

Double blinding

- Even if the study is blinded, there could still be a problem. If the experimenter knows which product is in which cup, then they could subtly (and possibly unintentionally) influence the outcome of the experiment.
- A **double blinded** study is one in which neither the participant nor the experimenter know which group is which.
- Why not always blind or double blind? (Think about the alpha-waves study.)

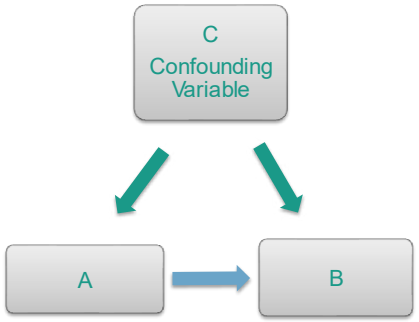
https://commons.wikimedia.org/wiki/File:Sleep_mask.jpg

The placebo effect

- Blinding is used to mitigate the "placebo effect": the phenomenon in which subjects' expectations of what will happen influence what does happen.
- Example: a drug for treating pain might "work" only because a patient expects it to work.
- Subjects in an experiment can be split into two groups: one gets the drug, the other gets the placebo. They are blind to which they are getting.
- If the drug group experiences different outcomes than the placebo group, this is evident that the drug has a real effect.

Confounding variables

- A **confounding variable** is a variable that influences both the predictor and response variables (but is usually not accounted for in the study).
- Formally, if variable C affects both variables A and B, then we might observe an association between A and B even though A and B have not causal relationship with each other.
- Sometimes called a moderator variable.



Discuss: Confounding Variables

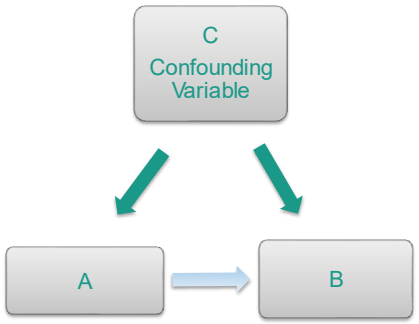
What is the confounding variable in each of the following examples:

- The number of murders is higher on days when more ice cream is sold
- A child's vocabulary is correlated with the number of cavities they have
- Retail stores report higher sales when their bathrooms are dirty

Confounding variables

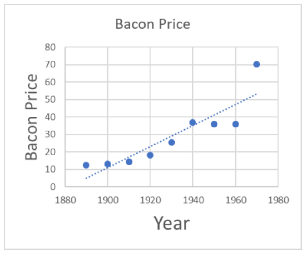
To explain a confounding variable completely make sure to explain:

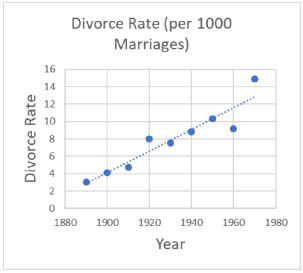
- How C affects A
- How C affects B
- That the relationship between A and B could really be explained by C (and we can not conclude that A causes B)

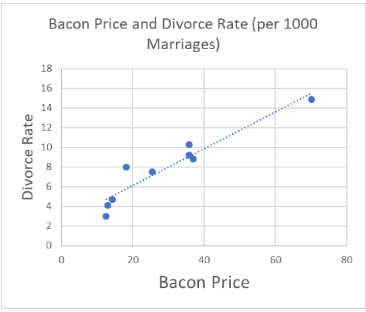


Confounding examples

Time is often a confounding variable that we forget about.

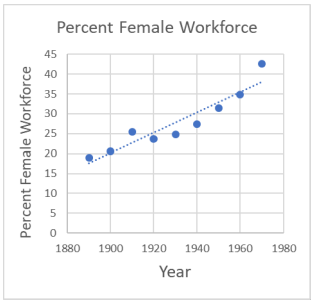


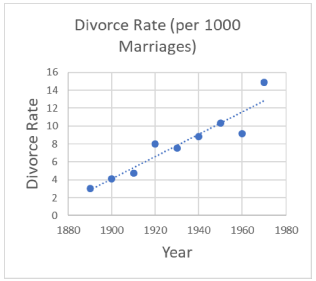


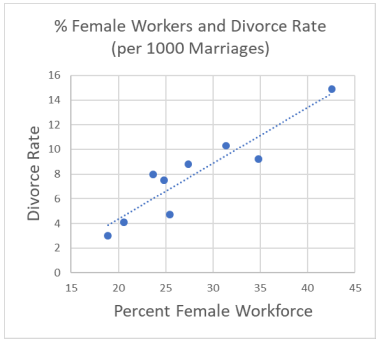


Confounding examples

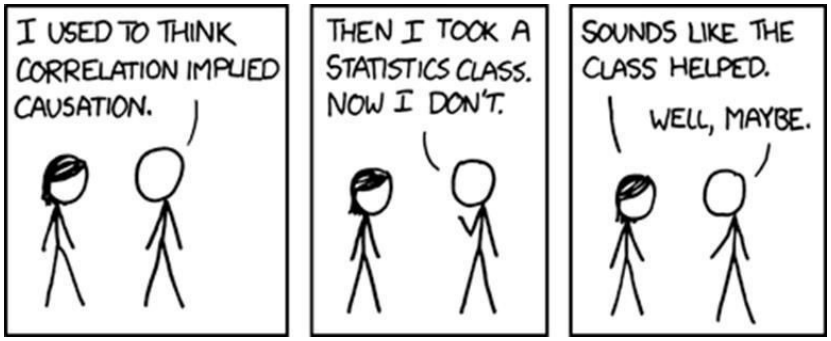
Time is often a confounding variable that we forget about.







Via XKCD.com



Section 1.5: Observational vs. experimental studies

Observational vs Experimental

In **observational** studies, predictor and response variable values are "observed", but the researcher does not manipulate anything.

- Observational studies can be retrospective (looking at past data) or prospective (set up to look at future data).

In an **experimental** study, the predictor variable(s) are assigned/manipulated by the researcher, and the response variable is observed.

Example: Planets vs. Brain waves

- In the Planets example scientists recorded information on all of the planets without changing anything, so the planets study is _____.

- In the brain waves study, scientists assign subjects (prisoners) to treatments (cell or confined) and observed their brain waves. Since the researcher assigned the predictor variable (cell/confined) this study is _____.

Advantages and disadvantages

	Observational	Experimental
Advantages		
Disadvantages		

Establishing causation example

- In general, association does not imply causation, especially in observational studies, because there may be a *confounding variable* in the background.

- Causation can only be inferred from a randomized experiment (in which the confounding variable can be controlled for in some way).

Example: Sunscreen

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen causes skin cancer?

How could we repeat this study in a way that might establish causation?

Wrap Up

In the discipline of statistics, we investigate scientific questions about a **population**. In order to do this we collect or analyze data from some **sample** of that population that seems to be representative.

We will try to explain as much **variation** as we can, but there will be some that goes unexplained. Also, studies should be conducted in such a way that we reduce **bias** whenever possible.

Module 1 Part 2: Summarizing Data

Wrap Up

Our **sample** will contain many **observations**, and each observation (person, planet, etc) will have one or more **variables** (height, brain waves, color, etc) measured and recorded.

We will find **statistics** about that sample and use it to perform **inference** and make a claim about the corresponding **parameter** in the population.

Example: STAT 204 class grades

Which of these is easier to understand?

B B B B B B A B D F A C B D
B C A B B B B C B C C D B
F C C A C A C C F D C D B
C A D A B B A B A C B C A
D A B C A C B B B C B B B
F A B A B A A B A A C

-OR-

Grade	Percent of Students
A	23%
B	39%
C	23%
D	9%
F	5%

Why do we need summary statistics?

Descriptive/Summary statistics help us glean important information from data without being overwhelmed.

The table on the right made the grades way easier to understand, and also tells us useful things that are hard to gather from the raw grades. However it doesn't tell us how many students are in the class.

When we use *summary* statistics, we lose some information but gain better understanding (like reading the *summary* of a book).

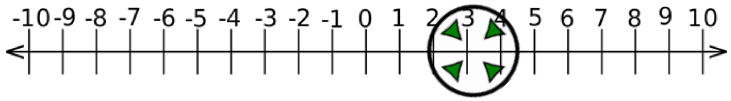
Section 1.6: Measures of Location

What is "location"?

"Location" - where the data are located on a number line

Measures of Location:

- Mean
- Median
- Quartiles
- Minimum, maximum
- Five number summary



Mean (Average)

The "mean", or "sample mean", or "sample average" is one statistic that identifies the data's **center**.

It is the sum of all of the sample data, divided by the number of data points, or sample size.

This is the sum of all of the data:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

This is the sample average, pronounced "x bar"

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

This is the number of data points, or sample size

Example: Planets

Below are the values collected for mass (Earth = 1) for the eight planets and Pluto:

Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
0.055	0.815	1	0.107	318	95	15	17	0.00016

The mean/average mass is given by:

$$\bar{x} = \frac{0.055 + 0.815 + 1 + 0.107 + 318 + 95 + 15 + 17 + 0.00016}{9} =$$

Median

- The median is another measure of the data's **center**.
- It is the middle data point when the data are arranged from smallest to largest.
- The median splits data in half: 50% below and 50% above.

Example: Planets Median

Let's use our mass example from before:

Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
0.055	0.815	1	0.107	318	95	15	17	0.00016

First, we put the data in order:

0.00016 0.055 0.107 0.815 1 15 17 95 318

Example: Planets Median

Imagine we left Pluto out. Then our ordered data would be:

0.055 0.107 0.815 1 15 17 95 318

Now there is no middle number. To get the median we average the two middle numbers instead:

Quartiles

- The median breaks the data set into two halves
- **Quartiles** break the data set into 4 quarters

- Therefore quartiles don't measure the center of the data, but they are still measures of location.

There are three quartiles:

- The **lower quartile**, **Q1** is the median of all the data *below* the overall median.
- The overall **median, Q2.**
- The **upper quartile**, **Q3** is the median of all the data *above* the overall median.

Example: Planets quartiles

Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune	Pluto
0.055	0.815	1	0.107	318	95	15	17	0.00016

We order the data to get:

0.00016 0.055 0.107 0.815 1 15 17 95 318

Example: Planets quartiles

Let's leave Pluto out again. Then our ordered data would be:

0.055 0.107 0.815 1 15 17 95 318

Extremes

- The **minimum** and **maximum** (or the extremes) of a data set are also measures of location.
- These are very far away from the center of the data, but they still tell us something about where the data is *located*.
- All of the data is below the maximum, and all of the data is above the minimum.

The Five Number Summary

- The **five number summary** is a convenient way to summarize a set of data:

  Minimum, Q1, Median, Q3, Maximum

- Roughly 25% of the data lies in each **quartile** defined by these statistics.

Ex: The five number summary for the planets (including pluto) would be:

0.00016, 0.081, 1, 56, 318

iClicker: Five number summary

Find the five number summary for the following data:

3 7 8 10 5 4 7 12 14 9 16 3 12 17

A. 3, 10, 9.5, 16, 17
B. 3, 9, 9.5, 9.5, 17
C. 3, 5, 8.5, 12, 17
D. 3, 4.5, 8.5, 13, 17
E. None of the above

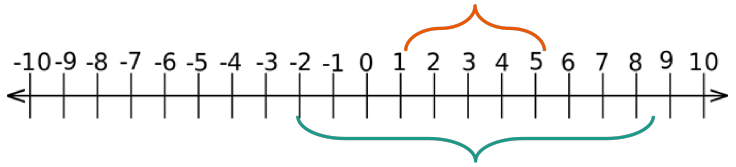
Section 1.7: Measures of Dispersion

What is “dispersion”?

Dispersion is how spread out data are on the number line.

Measures of dispersion:

- Range
- IQR
- Variance
- Standard Deviation



Example: Dispersion Concepts

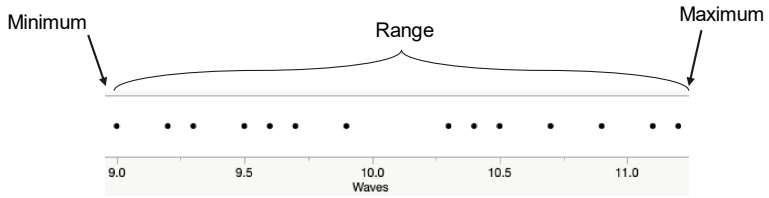
Weights of 20 randomly selected iPhone X's	Weights of 20 randomly selected loaves of bread	Weights of 20 randomly selected CSU students

Less dispersed ⟷ More dispersed

Range

The range of a data set is the largest value minus the smallest value.

It tells you the farthest distance between any two points in the data.



IQR

The IQR of a data set is the distance between its quartiles.

It tells you how much space the *middle 50%* of the data takes up.

iClicker: Range and IQR

Based on the five number summary for the data:

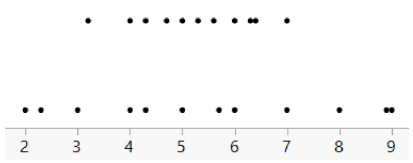
3 7 8 10 5 4 7 12 14 9 16 3 12 17

What is the range and interquartile range of the data?

A. Range = 14 , IQR = 6
B. Range = 14 , IQR = 7
C. Range = 17 , IQR = 0.5
D. Range = 17, IQR = 8.5
E. Range = 14, IQR = 8.5

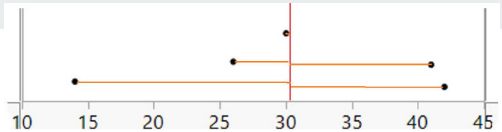
Sample Variance

- The sample variance, or just "variance", is a very common measure of dispersion.
- The variance is denoted as s^2.
- The idea of variance is that it measures how spread out the data are by adding together how far points are from the middle.



In this picture the top row has a smaller variance than the bottom row

Calculating Variance



To compute the variance, we first need the sum of squared deviations (called SS, for "sum of squares")

- Deviation = difference between one observation and the mean.

  $x_i - \bar{x}$, for some value of i
- Squared deviation = the deviation of an observation, squared.

  $(x_i - \bar{x})^2$, for some value of i
- SS = the sum of the squared deviations for all observations.

$$\text{SS} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Example: Calculating SS

Let's calculate the SS for mass of the inner planets (Mercury, Venus, Earth, and Mars):

i	1	2	3	4
x_i	0.055	0.815	1	0.107
$\bar{x}$				
$x_i - \bar{x}$				
$(x_i - \bar{x})^2$				

$$\text{SS} = \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

Variance

Once we have SS, we can calculate the variance as

$$s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

In a sense, variance is the "average" squared amount any observation deviates from the mean.

iClicker: Variance

Here is the formula for variance again: $s^2 = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

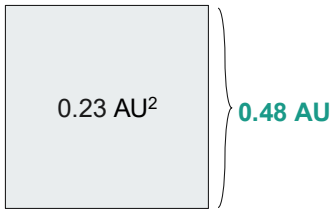
Which of the following statements is ***not*** true?

A. A larger sample variance means the data are more spread out
B. The units of variance is the original units squared.
C. Two different samples can have the same variance
D. The sample variance is always larger than the sample mean
E. Sample variance is always positive

Standard Deviation

Variance is mathematically useful, but not easy to think about. (*What does (astronomical units)^2 even mean?*)

A more scientifically useful and intuitive statistic is the **standard deviation**, which we can calculate from the variance.



Sample Standard Deviation

The **sample standard deviation**, or just "standard deviation", is the square root of the sample variance.

We use the letter s to denote it.

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Interpret the Standard Deviation

- The **standard deviation** is what it sounds like: a standardized amount by which observations deviate from the mean.
  - It's kind of: *on average how far are the points from the mean*.
- Large standard deviation implies data are highly dispersed, or spread out.
- Note that "large" or "small" depends on the data itself.

iClicker: Mean and Standard Deviation

You have a data set with the numbers:

4 5 6 7.2 7.5 8 10

The mean is 6.8 and the standard deviation is 2.

If we took a new observation and added it to the data set, what value would increase the mean and increase the standard deviation?

A. 3
B. 6
C. 7.5
D. 25
E. Not possible

iClicker: Mean and Standard Deviation

You have two variables with the following values:

Variable A: 1 3 4 4 4 5 6 7 7

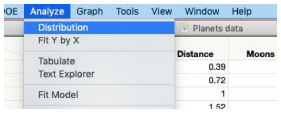
Variable B: 4 4 5 6 8 9 15 20 25

Without calculating, which of the following is true:

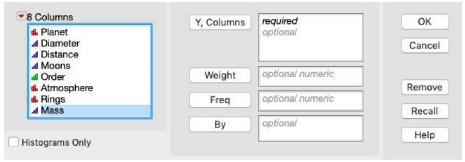
A. Variable A has a larger mean and larger standard deviation
B. Variable A has a larger mean and smaller standard deviation
C. Variable A has a smaller mean and a smaller standard deviation
D. Variable A has a smaller mean and a smaller standard deviation
E. Variable A and Variable B have the same mean and standard deviation

Summary Statistics in JMP

JMP will give you basic summary statistics via "Analyze > Distribution"



In this box you select variables of interest and place them in the "Y" field. The "by" box is useful if you want to look at statistics for one variable, separated by categories of another.



Example: planets SD vs. brain waves SD

For example, when we compare the alpha waves and planet diameter with JMP summary statistics output, a SD of 100 would be (very!) small for the planet diameter data but large for the alpha waves data.

Alpha waves:

Summary Statistics

Mean	10.18
Std Dev	0.6614179
Std Err Mean	0.1478975
Upper 95% Mean	10.489553
Lower 95% Mean	9.8704469
N	20

Planet diameter (km):

Summary Statistics

Mean	31466.431
Std Dev	47769.949
Std Err Mean	13249
Upper 95% Mean	60333.522
Lower 95% Mean	2599.3397
N	13

Section 1.8: Summarizing Categorical Data

Categorical Data

- Categorical data is data that is not numeric - instead, it represents several categories.
  - E.g. the letter grades from the start of the module

- Categorical data needs to be summarized differently from numerical data.
- Measures of location and dispersion depend on the data existing on a number line, so we can't use them on categorical data.

Summarizing Categorical Data in Tables

- The most obvious way to quantify categorical data is to count how many observations are in each category.
- We can put these numbers in a *frequency table* as a simple summary.
  - When one variable is being summarized, we use a *one-way table*.
  - When two variables are being jointly summarized, we use a *two-way table*.

Example: One-way table

Let's revisit the atmospheric composition variable in the planets data:

none	H2+He
CO2	none
N2+O2	CH4
CO2	none
H2+He	CH4
H2+He	CH4
H2+He	

We count the number of occurrences of each composition and put them in the table to the right:

Composition	Frequency
none	3
CO2	2
N2+O2	1
H2+He	4
CH4	3
Total	**13**

Proportions

Another way to summarize categorical data is to use *proportions*.

$$p = \frac{\text{\# observations of interest}}{\text{Total \# of observations under consideration}}$$

Example continued: Relative frequency table

We can take a frequency table and divide each value by the total to create a relative frequency table.

Composition	Frequency
none	3
CO2	2
N2+O2	1
H2+He	4
CH4	3
Total	**13**

Composition	Proportion of planets
none	$3/13 = 0.23$
CO2	$2/13 = 0.15$
N2+O2	$1/13 = 0.08$
H2+He	$4/13 = 0.31$
CH4	$3/13 = 0.23$
Total	**1.000**

Proportions and Percentages

Proportions and percentages can be easily converted by multiplying or dividing by 100. For example,

$$.302 = 30.2\%$$

(Be sure to only include a "%" with percentages, not with proportions!)

Example continued: Proportions

Composition	Frequency
none	3
CO2	2
N2+O2	1
H2+He	4
CH4	3
Total	**13**

What proportion of planets have an atmosphere?

What percent of planets have an atmosphere made of CO2?

Example: Planets two-way table

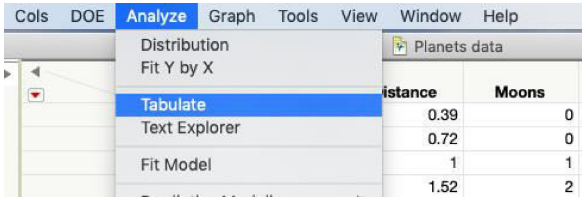
In the planets data set, in addition to their composition, we recorded whether or not they have rings:

	none	CO2	N2+O2	H2+He	CH4
Yes	0	0	0	4	0
No	3	2	1	0	3

Do you see any relationship between presence/absence of rings and atmospheric composition?

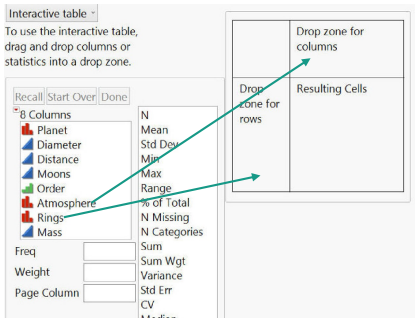
Tables in JMP

Let's explore making table in JMP. Load the planets dataset, then go to Analyze > Tabulate.



Tables in JMP

Drag the "Atmosphere" variable into the columns space and the "Rings" variable into the rows space to create the table below:



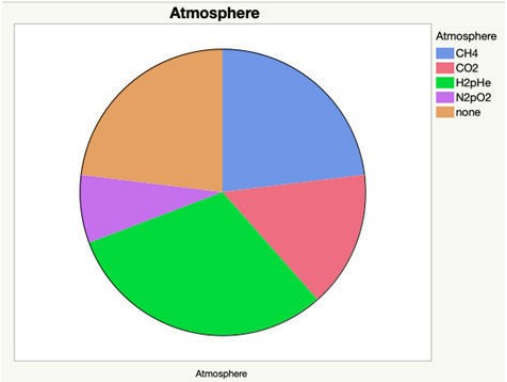
	Atmosphere				
Rings	CH4	CO2	H2pHe	N2pO2	none
no	3	2	0	1	3
yes	0	0	4	0	0

Section 1.9: Visualizing Data

Visualizing data

- All of the graphs we will look at show frequency distributions of data. Often this is shortened to just distribution.
- A **distribution** tells you the values a variable takes on, and the frequency with which those values are taken on.
- The tables we saw for categorical variables is an example of a distribtion.

Pie Charts



Pie charts can be used to summarize one categorical variable.

Pie slices represent the proportion of observations in a category.

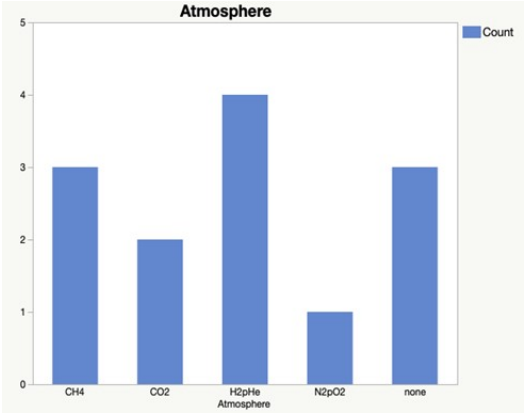
Here is a pie chart from JMP, for the *atmospheric composition* variable in the planets data.

Bar charts

Bar charts can be used wherever pie charts are used (i.e. the distribution of a categorical variable)

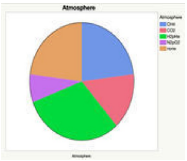
Each bar shows the frequency of its category.

Example: Here is the *atmospheric* data again as a bar chart:



Pie VS Bar charts

PIE CHART



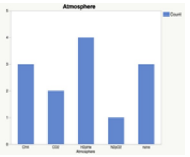
People aren't good at comparing relative areas

Are the slices the same size?

No axes

How big is the slice? 15%? 30%?

BAR CHART



Easy to compare groups

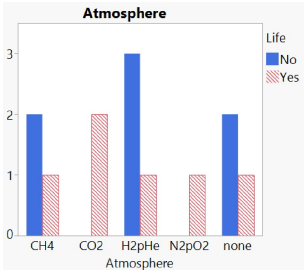
Comparing horizontally instead of around a circle.

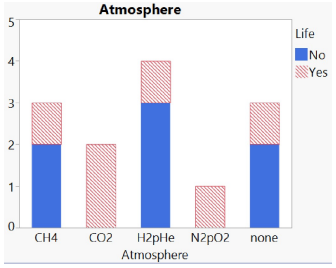
Labeled axis

Easy to read exact percentages.

Split/Stacked bar plot

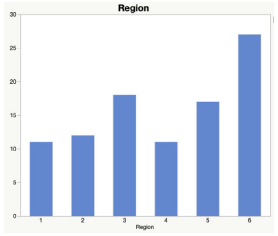
Bar plots can be used to show the joint distribution of two categorical variables at once. The split bar plot on the left and the stacked bar plot on the right show the same data: the atmosphere and if there is life (made up because the rings graph is not very interesting).



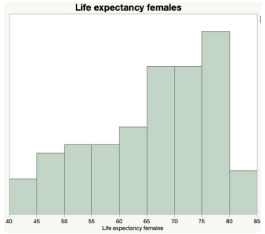


Histograms

A histogram displays the distribution of a quantitative variable. It looks a lot like a bar chart, but there are some important differences that we will see on the next slide.



Bar Graph



Histogram

Histograms vs. Bar Charts

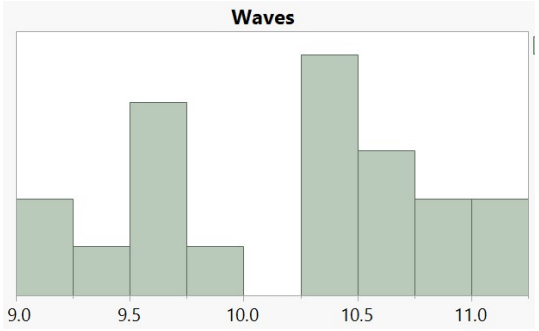
- The difference between a histogram and a bar chart is that bar charts are for categorical data and histograms are for quantitative data.
- With bar charts, each bar represents a different distinct group or category. With histograms, each bar represents the number of observations which fall into an **interval**, also known as a **bin**.
- **Like bar charts**, the height of each bar corresponds to either frequency or relative frequency.

Example: histogram of brain waves

In the Brain Waves data set, we can make a histogram of the "Waves" variable.

Notice how:

- The variable, "Waves," is quantitative not categorical
- The bars touch each other



Histograms

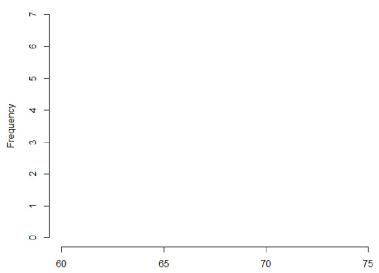
The number of bins on a histogram is arbitrary (i.e. you get to choose it). If you choose to use more bins, the bin size (the length of the interval) will be smaller.

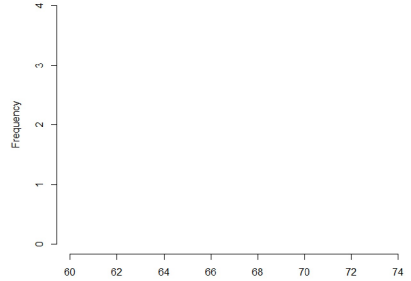
Changing the number of bins can produce different looking histograms, even if the underlying data is exactly the same.

Example: making a histogram

Let's make two different histograms for the dataset below

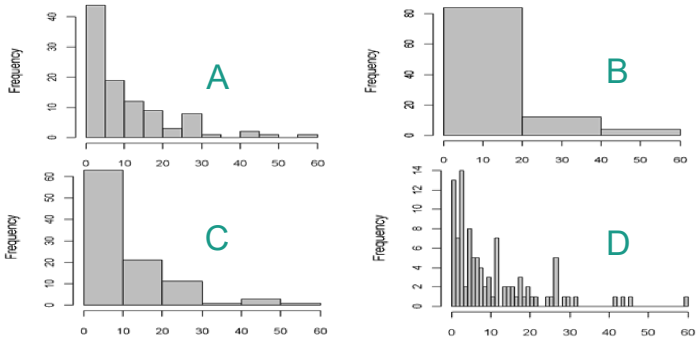
Heights of 10 randomly selected statistics students									
65	67	66	69	69	66	64	64	63	72





iClicker: Histogram bin size?

These four histograms represent the exact same data. Which of these histograms is "best"? How about most useful?



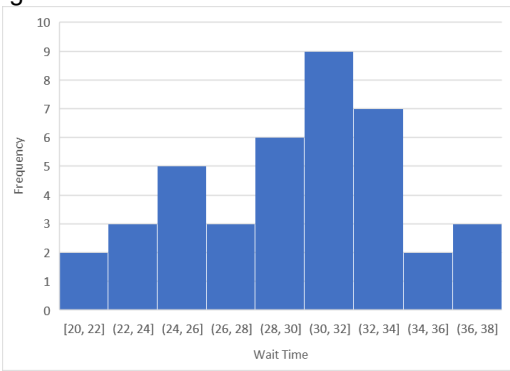
Discuss

You collect a sample of how long customers are waiting for their food at your restaurant during a busy night. The histogram shows the results below. Answer the questions below:

What is the approximate range of the data?

What wait times occur most frequently?

How many people waited longer than 34 minutes?



Boxplots

Boxplots are used to display the distribution of a quantitative variable (like histograms)

Like a histogram, boxplots help us determine the shape of a distribution and identify possible outliers.

Unlike histograms, boxplots let us compare two categories side by side.

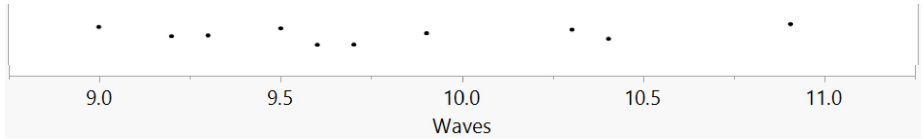
Boxplots are the visual representation of the five number summary.

Example: Making a Boxplot

Consider the alpha waves for the prisoners in solitary confinement:

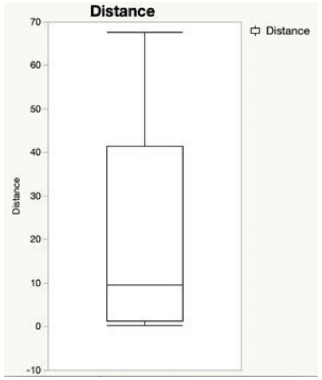
9 9.2 9.3 9.5 9.6 9.7 9.9 10.3 10.4 10.9

Give the five number summary:



Example continued: boxplot of distance

- To the right is a boxplot of the distance variable from the planets data.
- The line in the center is the median.
- The bottom and top edges of the box are Q1 and Q3, respectively.
- The “whiskers” extend to the minimum and maximum



Example: Making a Boxplot

What if we put one more prisoner in solitary confinement, and the new value is 14?

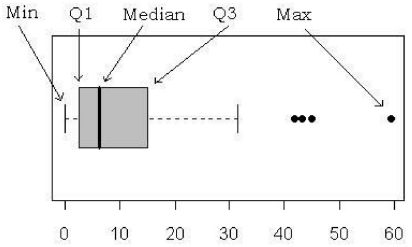
9 9.2 9.3 9.5 9.6 9.7 9.9 10.3 10.4 10.9 14

Give the five number summary:

Outliers

- **Outliers** are data points that are located far away from the majority of the data is.
- There isn't universal agreement on what the exact standard should be. Data analysts and software will all use different methods to identify outliers.
- An outlier is usually a data point that you should look closer at. Outliers could be:
  - Improperly entered data
  - Measurement error
  - Accurate observations that are just unusual

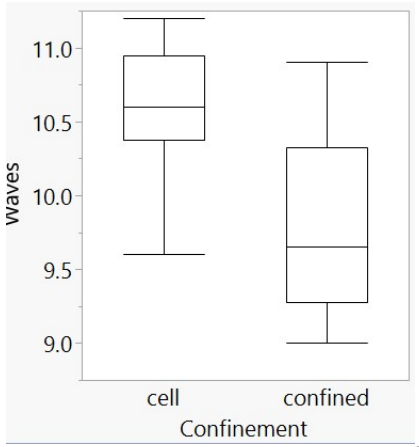
Anatomy of a Boxplot



- Boxplots can be displayed horizontally or vertically.
- If there are outliers, they are drawn as dots beyond the whiskers.
- The whiskers extend to either the min/max or the furthest non-outler. (In this plot, the max is an outlier.)
- Remember that the five number summary (and therefore the boxplot) divides our data into quartiles. So 50% of the data is inside the box, 25% is below the box, and 25% is above the box.

Boxplots

Boxplots are great for comparing the values of a quantitative variable split up by a categorical variable!



Discuss

Answer the following.

Which group has the largest mean?

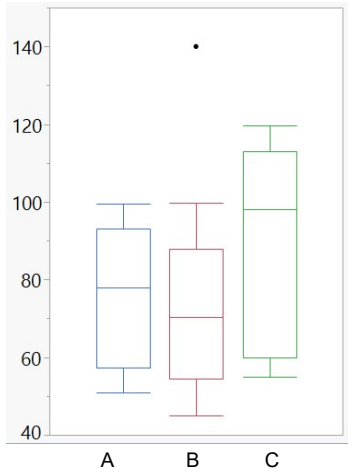
Which group has the smallest minimum?

Which group has the largest maximum?

Which group has the lowest quartile 1?

Which group has the largest range?

Which group has the largest interquartile range?



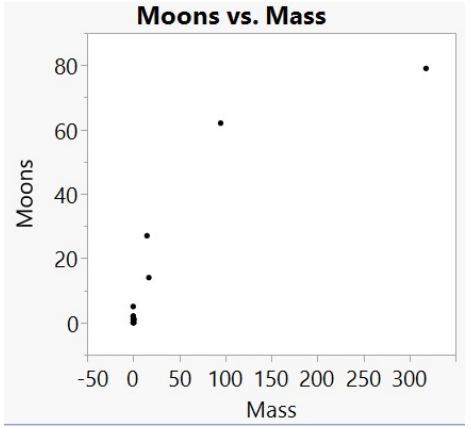
Scatterplots

- Scatterplots are used to plot the values of two quantitative (and usually continuous) variable against one another. Generically we call these X and Y.
- On a scatterplot, each dot show the X and Y coordinates for a single data point.
- It is conventional to plot a response variable on the Y axis and a predictor variable on the X axis. But this is not a hard and fast rule.

Example: Scatterplot

This scatterplot shows Number of Moons(Y) and Mass (X) of all the planets in the data set.

The scatterplot allows us to see the association between these two variables:
As mass increases the number of moons tends to increase.



iClicker: Scatterplot

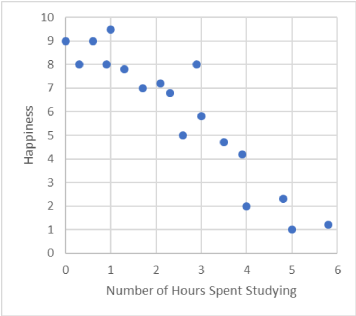
You poll your friends and ask how many hours they spent studying last night and asked them to rate their happiness on a scale from 1-10 with 10 being the most happy.

What relationship do you see?

A. As studying increases happiness tends to increase.

B. As studying increases happiness tends to decrease.

C. Happiness does not depend on studying.



Distribution Shape

Plotting data can also tell us about the "shape" of the data.

Shape can mean lots of things but we'll focus on whether the data is:

-
-
-

Distribution shape: symmetrical

If the two halves of the data look almost like mirror images, then we say a distribution is symmetrical or has no skew

This is a histogram of the percent daily change in oil price.

Notice how the data is mirrored roughly around 0%.

Distribution Shape: Skewed right

If there are a lot of low values and only a few high values, then we say a distribution is skewed to the right or positively skewed

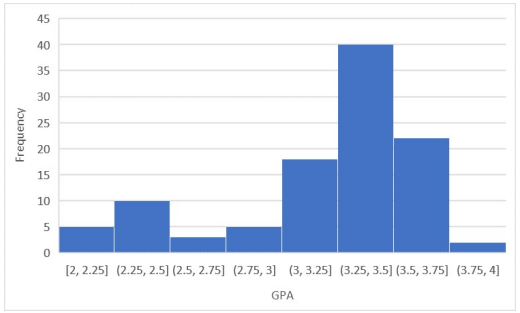
We can see from this histogram of diamond prices, that low priced diamonds are relatively common and high priced diamonds are relatively rare. So diamond prices are positively skewed

(Think about it as the direction the "tail" goes.)

Distribution Shape: Skewed left

If there are a lot of high values and only a few low values, then we say a distribution is skewed to the left or negatively skewed

This histogram of GPAs shows that GPAs are most commonly between 3-3.75 but there is a tail of lower scores.



iClicker: Distribution Shape

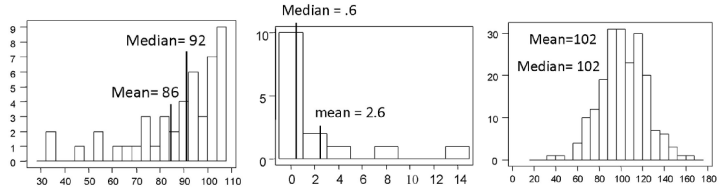
Label the shapes of the distributions below in order.

A. Left skewed, Right skewed, Symmetrical

B. Right skewed, Left skewed, Symmetrical

C. Symmetrical, Left skewed, Right skewed

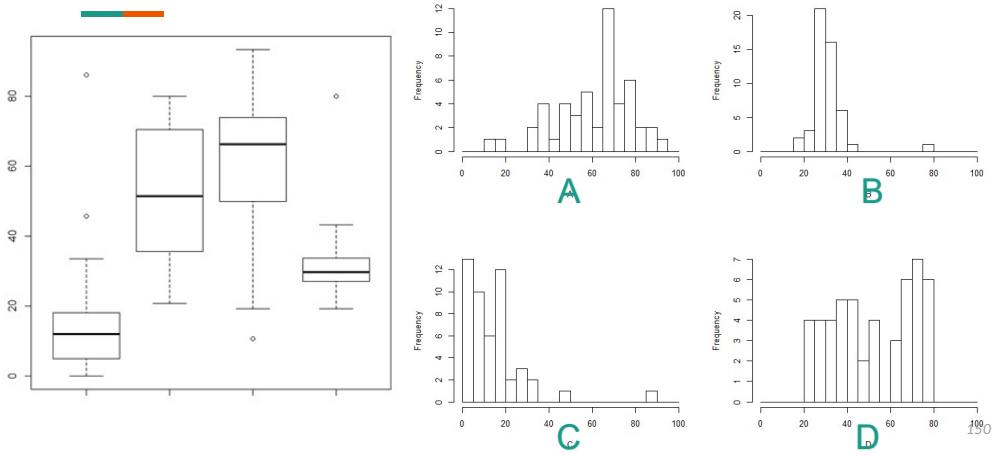
D. Right skewed, Symmetrical, Left skewed



Discuss: Mean, median, and shape

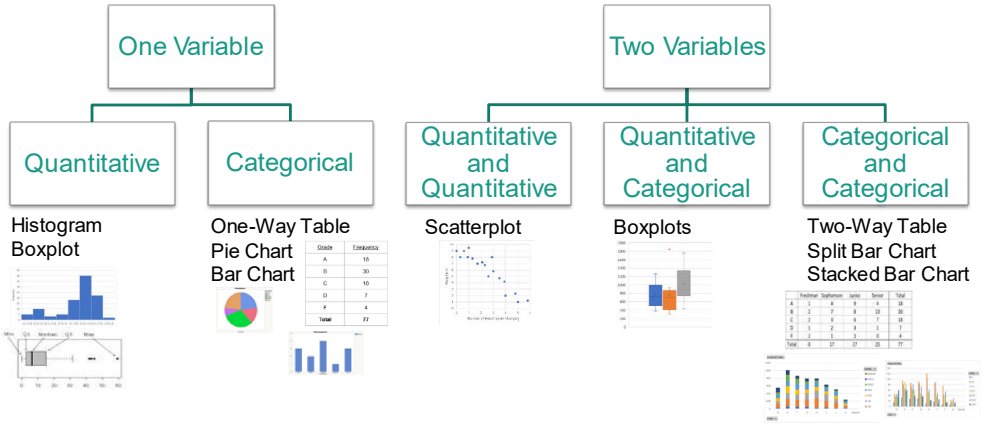
- If the mean is greater than the median then the distribution is skewed to the _____.
- If the mean is less than the median then the distribution is skewed to the _____.
- If the mean and median are (approximately) equal then the distribution is (approximately) _____.

Comparing Histograms and Boxplots



The type of variables you have will affect how you can graph the data!

Summary of Visual Options



Module 2: Probability

Review

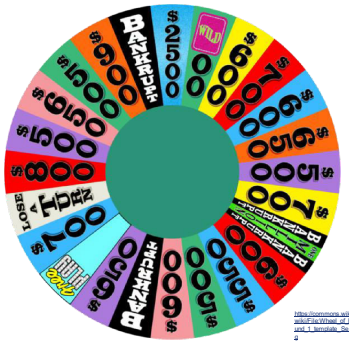
Module 1:

- Summarizing Categorical Data
  - Proportions
  - Percents
  - One-way and Two-way tables
- Confounding Varaibles

iClicker: Probability warm-up

On the show "Wheel of Fortune" contestants spin a wheel with 24 options. What is the probability that a contestant will lose a turn?

A. .0417
B. .0833
C. .125
D. 1
E. None of the above



https://commons.wikimedia.org/wiki/File:Wheel_of_Fortune_Round_1_template_Season_31.png

2.1 Overview of Probability

Random Events

- "Random" is a difficult word to define. In statistics, we use it to refer to events that are unpredictable.
- A **random event** is something that may or may not occur, and that which we can assign a **probability** to.
- We can think of a random event as a possible value that a **random variable** takes on.
- For instance, a random variable "x" may be the outcome of a roll of a die, and a random event might be x = 6.

Probability

Probability is a way of quantifying the chance that some random event occurs.

Probabilities are often related as percentages, but formally they should be given as proportions.

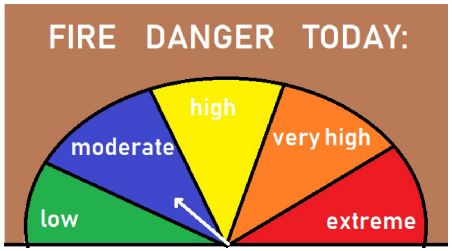
- For example, if there is a 50% chance of something happening, then its probability is 0.5.

A probability MUST be a number between 0 and 1. Think of 0 as "impossible" and 1 as "absolutely certain".

iClicker: Probability warm-up

What is the probability that the fire danger will be extreme?

A. .1
B. .2
C. .5
D. .8
E. None of the above



Interpretations of Probability

There are different ways to think about what probabilities refer to.

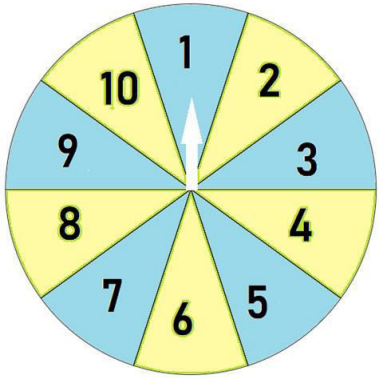
Plausibility: This interpretation of probability reflects a state of knowledge / information / uncertainty

Relative frequency: This interpretation thinks of if this 'event' was repeated over and over and over again, how often would this outcome occur? Refers to a property of "the real world", independent of human knowledge.

Example: Comparing Interpretations

You are asked to spin the wheel. What is the probability you will get a 10?

Now suppose you spin the wheel and see that it is 1. What is the probability that you spun a 10?



Example: Comparing interpretations

Now suppose you are blindfolded when you spin the wheel. Your friend sees that you spun a 1. What is the probability that you spun a 10?

Frequentist Interpretation of Probability

- Classical statistical inference is based on relative frequency, and is sometimes called “frequentist” statistics.
- STAT 201 will treat probabilities as relative frequencies.
- Be aware that this might sometimes feel counter-intuitive. For example, what is meant by the claim “There is a 60% chance that the incumbent representative wins re-election”?

Expressing Relative Frequency

- According to the department of transportation, 1 out of every 324 airline passengers lost their luggage (at least temporarily) in 2012.
- Here are two ways of expressing this probability / relative frequency:
  - *The proportion of passengers who lost their luggage is 1/324 or about .00309.*
  - *The probability that a randomly selected passenger lost his/her luggage is about .00309.*
- Notice that we are using the concepts of probability and relative frequency interchangeably.

2.2 Formalizing Probability

Probability notation

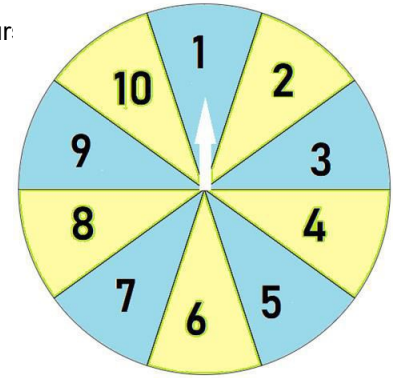
$P(X)$ is the probability that event X occur

For example on the wheel:

Probability of spinning a 5
P(Spin 5) = 1/10 = .1

Probabilitiy of spinning a 1 or 2
P(Spin 1 or 2) = 2/10 = .2

Probability of spinning a blue
P(Spin Blue) = 5/10 = .5



Example: Titanic

"RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in 1912 after the ship struck an iceberg during her maiden voyage from Southampton to New York City." – Wikipedia

We have data on 891 passengers in a data set called Titanic. In that data set:
Survival: 0 = No, 1 = Yes
Pclass: Ticket class: 1 = 1st, 2 = 2nd, 3 = 3rd
Sex: male or female
Age: age in years

Image: https://en.wikipedia.org/wiki/RMS_Titanic#/media/File:St%C3%B6wer_Titanic.jpg

Example: Titanic

Let's start by looking at only the survival. Fill in the relative frequency.

Survived	Frequency	Relative Frequency
No	549	
Yes	342	
Total	891	

What is the probability that a randomly selected passenger will survive?

Example: Titanic

Now let's split this up by the class of the passenger as well.

Survived	1st class	2nd class	3rd class	**Total**
No	80	97	372	549
Yes	136	87	119	342
Total	216	184	491	891

What is the probability that a randomly selected passenger is 1st class?

What is the probability that a randomly selected passenger is not 1st class?

Example: Titanic

Now let's split this up by the gender of the passenger.

Survived	Female	Male	**Total**
No	81	468	549
Yes	233	109	342
Total	314	577	891

What is the probability that a randomly selected passenger is female?

Conditional probability

- Conditional probabilities are probabilities of certain events occurring given that some other event occurs or has occurred.
- We write the conditional probability of "A given B" as: **P(A|B)**
- The bar "|" means "given" or "conditional upon".
- So, $P(A|B)$ means, "What is the probability of event A given that event B has occurred?"

Probabilities

When we see P(Thing 1) the whole is always the grand total, and the part is how many are in Thing 1.

$$\mathrm{P}(\text{Thing 1}) = \frac{how\ many\ Thing\ 1}{Grand\ Total}$$

When have conditional probability P(Things 1|Cateogry 1) we change the whole to what is after the conditional bar.

$$\mathrm{P}(\text{Thing 1}|\text{Category 1}) = \frac{how\ many\ Thing\ 1\ in\ Category\ 1}{how\ many\ category\ 1}$$

	Category 1	Category 2	Total
Thing 1	How many thing 1 in category 1		How many thing 1
Thing 2			
Total	How many category 1		GRAND TOTAL

Example: Titanic

Survived	Female	Male	**Total**
No	81	468	549
Yes	233	109	342
Total	314	577	891

Notice that in conditional probability the order is important:

P(male | survive) is not equal to P(survive | male)

Example: Titanic

Survived	Female	Male	**Total**
No	81	468	549
Yes	233	109	342
Total	314	577	891

Consider:
P(survive | female)

P(survive | male)

Is survival related to whether the passenger is male or female?

Independence and Dependence

We often want to know whether or not two categorical variables are related. By "are they related?", we mean "If an observation falls into a certain category for one variable, does this make it more likely to fall into a certain category of the other?"

In statistics we refer to this as independent and dependent.

If they are 'not related' we call them independent.
If they are 'related' we call them dependent.

Independence and Dependence

Two variables are **dependent** when knowing what category one falls into changes the probability of the other occurring.

Here the variable Letter can take on A or B.
The variable Number can take on 1 or 2.

We would say Letter is dependent on Number if
$P(A|1) \neq P(A|2)$

We would say Letter is not dependent, or independent of Number, if $P(A|1) = P(A|2)$

		Number		
		1	2	Total
Letter	A			
	B			
	Total			GRAND TOTAL

Example: Titanic

Split up by the class of the passenger.

Survived	1st class	2nd class	3rd class	**Total**
No	80	97	372	549
Yes	136	87	119	342
Total	216	184	491	891

What is the probability that a passenger survives given that they are in 1st class?

What is the probability that a passenger survives given that they are in 2nd class?

Example: Titanic

Survived	1st class	2nd class	3rd class	**Total**
No	80	97	372	549
Yes	136	87	119	342
Total	216	184	491	891

What is the probability that a passenger survives given that they are in 3rd class?

Is survival dependent on the class of the passenger?

iClicker: Interpret a Table

Some people take Vitamin C in hopes of not getting sick during flu season. A survey of 210 people is done on December 10th to ask if they regularly take a vitamin containing vitamin C and if they have gotten sick during the past month. The results are shown in the table below:

	Sick	Not Sick
Vitamin C	9	69
No Vitamin C	17	115

What probabilities are we interested in?

a. P(Sick|Vitamin C), P(Sick|No Vitamin C)
b. P(Sick|Vitamin C), P(Not Sick|No Vitamin C)
c. P(Vitamin C|Sick), P(No Vitamin C|Not Sick)
d. P(Vitamin C|Sick), P(No Vitamin C|Sick)
e. P(Sick) and P(Vitamin C)

iClicker: Interpret a Table

	Sick	Not Sick
Vitamin C	9	69
No Vitamin C	17	115

What should we conclude from these results?

a. Vitamin C prevents you from getting sick
b. Vitamin C actually makes you get sick more often
c. Vitamin C does not affect if you will get sick
d. We can't really conclude anything about vitamin C from these results

Discuss: Interpret a 2x2 Table

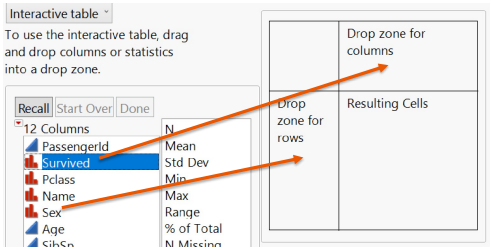
Some people take Vitamin C in hopes of not getting sick during flu season. A survey of 210 people is done on December 10th to ask if they regularly take a vitamin containing vitamin C and if they have gotten sick during the past month. The results are shown in the table below:

	Sick	Not Sick
Vitamin C	104	41
No Vitamin C	63	20

What concerns might we have about theses results or the set up of this study? (think bias, observational/experimental, sample size, etc)

Tables In JMP

Remember to make a table in JMP you can go to Analyze > Tabular and drag the variable names into the spots of row and column. Usually we also want to check the box for 'Add Aggregate Statistics' in order to get the totals.



☑ Add Aggregate Statistics

2.3 Accounting for Confounding Variables

Simpson's Paradox

Sometimes when we break data apart by a confounding variable we see that a particular relationship might change.

This is called **Simpson's Paradox**.

A **paradox** is a seemingly absurd or self-contradictory statement or proposition that when investigated or explained may prove to be well founded or true.

Example: Ambulance Rides

A hospital wants to determine the effectiveness of using helicopters to bring in patients to the hospital. They take the records of 178 patients who were brought in to the hospital by ambulance or helicopter and record whether the patient had died in the two-week after admittance.

	Death	No Death	TOTAL
Ambulance	23	130	153
Helicopter	5	20	25
TOTAL	28	150	178

Example: Ambulance Rides

	Death	No Death	TOTAL
Ambulance	23	130	153
Helicopter	5	20	25
TOTAL	28	150	178

Based on the information in the table does dying depend on the mode of transportation?

Example: Ambulance Rides

There is more information we can consider when looking at these cases. Let's split the table up by the severity of the patients condition.

Not Severe Condition

	Death	No Death	TOTAL
Ambulance	8	92	100
Helicopter	0	2	2
TOTAL	8	94	102

Severe Condition

	Death	No Death	TOTAL
Ambulance	15	38	53
Helicopter	5	18	23
TOTAL	20	56	76

Example: Ambulance Rides

Now let's rethink: does dying depend on the mode of transportation?

Clicker: Interpret

Which of the following headlines would be true?

a. Tell your 911 operator that you want a helicopter!
b. Ambulances are safer than helicopters.
c. We should ban helicopters.
d. Sign this petition to put helicopter wings on all ambulances.
e. Helicopters shows slightly less deaths overall for patients over a 2 week period. However, helicopters and ambulances are used in different situations and most people only resort to helicopters when they are in a severe condition, which leads to the interesting result of helicopters having lower survival rates. Choose your mode of transportation based on your immediate needs.

Example: Death Penalty

In 1991 a study was done using data from Florida's criminal justice system. There were 674 Florida homicides and death sentences with complete records from the years 1976-1987. Let's look at the information:

	Death Penalty	No Death Penalty	TOTAL
White Defendant	53	430	483
Black Defendant	15	176	191
TOTAL	68	606	674

Example: Death Penalty

	Death Penalty	No Death Penalty	TOTAL
White Defendant	53	430	483
Black Defendant	15	176	191
TOTAL	68	606	674

Based on the information in the table is receiving the death penalty dependent on the race of the defendant?

Example: Death Penalty

There is more information we can consider when looking at these cases. Let's split the table up by the race of the victim.

White Victim

	Death Penalty	No Death Penalty	TOTAL
White Defendant	53	414	467
Black Defendant	11	37	48
TOTAL	64	451	515

Black Victim

	Death Penalty	No Death Penalty	TOTAL
White Defendant	0	16	16
Black Defendant	4	139	143
TOTAL	4	155	159

Example: Death Penalty

Now let's rethink: Are white defendants more likely to get the death penalty?

iClicker: Death Penalty

Which of the following headlines would be true?

a. White people are more likely to get the death penalty.
b. Black people are more likely to get the death penalty.
c. Black and white people are equally likely to get the death penalty.
d. Death penalty verdict may be biased based on the race of the victim.
e. Death penalty verdict is totally based on the race of the victim.

Example: Electoral College vs Popular Vote

The president of the united states is elected based on the electoral college. 48 states and the district of Columbia have a winner take all system for their elector college votes. For example California has 55 Electoral votes, and whichever candidate wins in the state gets all 55 of those votes in the electoral college.

There have been 5 times in America's history where a president won the electoral college votes and became president even though they did not win the popular vote.

This idea is similar to Simpson's paradox, when we split votes up by state we get a different result than when we have them all together as a country.

What to do about Simpson's Paradox

Always be aware of possible confounding variables.

- Sometimes the confounding variables will be in your data and you'll be able to take a deeper look right away.
- Sometimes the confounding variable is impossible to measure.
- No matter what, remind yourself that one or more may exist.

Then think critically:

- Is it better to split the data up into smaller groups based on this other variable or do I want a generalization?
- This will depend on your situation

2.4 Percent Change

Percent Change

Percent change is one way to compare two values (and old and a new) or compare two groups.

The formula for percent change is:

$$percent\ change = \frac{new - old}{old} * 100$$

Unlike probability percent change can be negative (if the value decrease) and can be greater than 100%.

Recognizing Percent Change

Some possible ways you might see percent change worded:

- Bike accidents have decreased 10 percent on campus this semester.
- Sales have increased 5% this year.
- You are 200% more likely to be single on valentines day if you eat anchoives.
- Study shows milk leads to 3% increase in bone health
- Bee populations are down 20% since last year.
- Netflix has 400% more subscribers than Hulu.

Note that when we talk about Thing 1 is 30% higher than Thing 2 we would consider Thing 1 to be 'new' and Thing 2 to be 'old' in our formula.

Example: Pedestrian Deaths

Why are US drivers killing so many pedestrians?

By Joe Cortright | 27.6.2019

US drivers are killing 50 percent more pedestrians, European drivers are killing a third fewer

If anything else-a disease, terrorists, gun-wielding crazies-killed as many Americans as cars do, we'd regard it as a national emergency. Especially if the death rat had grown by 50 percent in less than a decade. But as new data from the Governor's Highway Safety Association (via Streetsblog) show, that's exactly what's happened with the pedestrian death toll in the US. In the nine years from 2009 and 2018, pedestrian deaths increased 51 percent from 4,109 to 6,227:

http://cityobservatory.org/why-are-us-drivers-killing-so-many-pedestrians/

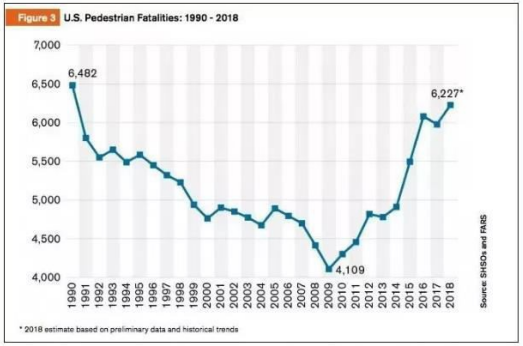
Example: Pedestrian Deaths

In 2009 there were 4109 pedestrian deaths and in 2018 there were 6,227 deaths.

$$percent\ change = \frac{new - old}{old} * 100$$

Thus pedestrian deaths have increased about 51% in the past year.

Example: Pedestrian Deaths



This is what a pedestrian safety crisis looks like. Graph: Governor's Highway Safety Association

http://cityobservatory.org/why-are-us-drivers-killing-so-many-pedestrians/

Example: Pedestrian Deaths

In 1990 there were 6482 pedestrian deaths and in 2018 there were 6,227 deaths.

$$percent\ change = \frac{new - old}{old} * 100$$

We could also say that pedestrian deaths have _____since 1990.

Example: Lottery

In Colorado the chance of winning the jackpot prize on the powerball is 1 in 292,201,338 (0.000000003422298). Your chance of winning a million dollars is 1 in 11,688,054 (.0000000855574418). Let's call the jackpot 'old' and the million dollars 'new'

$$percent\ change = \frac{new - old}{old} * 100$$
$$= \frac{0.0000000855574418 - 0.000000003422298}{0.000000003422298} * 100$$
$$= 23.99 * 100$$
$$= 2399\%$$

You are 2399% more likely to win a million dollars than you are to win the jackpot.
Does that mean you are likely to win the million dollars?

Percent Change

Whenever you see information given as a percent change you should always ask:

What was the original value?

Think about shopping: You have a 10% off coupon. This means that the 'new' price at the register will be 10% lower than the 'old' price on the tag.
If you buy something that costs $200 you'll get $20 off.
If you buy something that is $2 you get $0.20 off.
The starting value makes a big difference.

Example: Interest Rates

The Fed Just Raised Interest Rates. Here's What That Means for Your Wallet.

https://www.nytimes.com/2018/12/19/your-money/interest-rates-consumers.html

In December 2018 the New York Times wrote about interest rates:
"The Fed raised short-term rates by a quarter of a percentage point to a range of 2.25 to 2.5 percent"

$$percent\ change = \frac{new - old}{old} * 100$$

Old rate is 2.25%
New rate is 2.5%

You could say that the interest rates have increased by 11%. But is this confusing?

Percent Change vs Change in Percentage Points

If you are measuring something that is already in percentages it is usually easier to just talk about 'percentage points', where a 'percentage point' is 1%.

For example if your grade in this class increases from a 80% to an 85% you should say it increased by 5 **percentage points.** The percent change is 6.25% but this would be confusing to tell people and you may mislead them.

Watch out for language about percent change and percentage point change when talking about things that are already measured in percentages.

iClicker: Ambulance Rides

Let's revisit the ambulance rides.
P(Death|Ambulace)=.15
P(Death|Helicopter)=.20

	Death	No Death	TOTAL
Ambulance	23	130	153
Helicopter	5	20	25
TOTAL	28	150	178

Which of the following is false?

A. Your chance of dying after taking an ambulance is 15%
B. You are 33% more likely to die after a helicopter than an ambulance.
C. You are 25% less likely to die after an ambulance than a helicopter.
D. Your chance of dying after a helicopter is 5% higher than after an ambulance.
E. Your chance of dying after an ambulance is 5 percentage points lower than dying after a helicopter.

Main Ideas

We've looked at calculating probabilities from a two-way/contingency table.

- We can check if one variable might depend on another by comparing probabilities.
- We can split a table apart by a confounding variable, and we may run into Simpson's paradox and see a relationship change.

We can compare probabilities using percent change or a change in percentage points.

When reading statistics or presenting statistics to others it is very important to be careful with our language. We must think critically about what information is important and what numerical summaries will convey information the best. There isn't always one 'right' way to present information.

Module 3: Standardization and Percentiles

3.1 Normal Distribution and Z-scores

Example: Bad Grades

You are your friend are both complaining about your horrible grades. They got a 65 on their English paper and you got a 55 on your math exam. Both of you seem to think you did worse.

Whose score was really worse?

Example: Bad Grades

They got a 65 on their English paper and you got a 55 on your math exam.

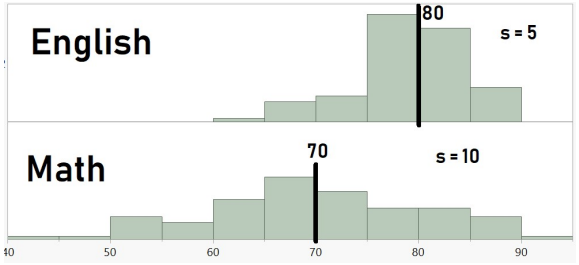
You take a deeper look. Both classes have 100 people. You find the average test score from each class. The average grade on the English papers is 80, and the average grade on the math exam is 70.

Whose score was really worse?

Example: Bad Grades

You look even deeper. You are able to see the grade distribution. English average was 80 with a standard deviation of 5, and math average was 70 with a standard deviation of 10.

Whose score was really worse?



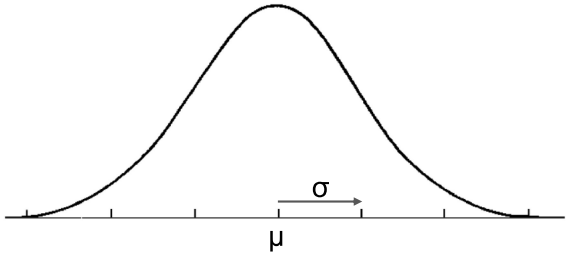
Normal Distribution

The most common distribution we encounter in statistics is the **normal** distribution. (This is also known as the **Gaussian** distribution.)

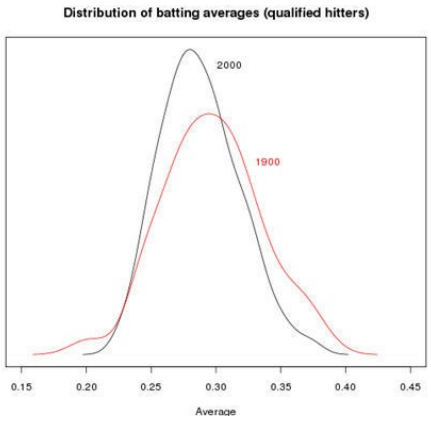
- A variable that follows a normal distribution is said to be **normally distributed**.

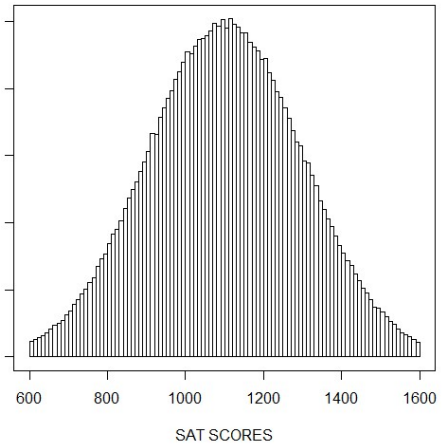
The normal distribution is bell shaped, and it is defined by its:

- μ mean (tells us the center)
- σ standard deviation (tells us how spread out) (or could use variance σ^2)



Examples of Normal Distributions

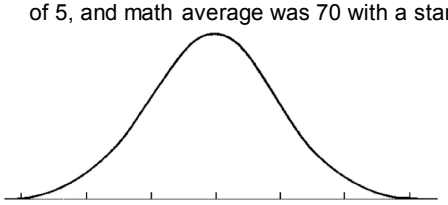


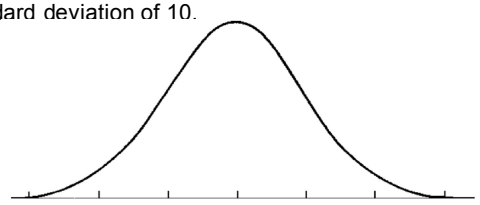


Example: Bad Grades

When labeling a normal distribution put the mean at the center and use the tick marks to go up and down one standard deviation at a time.

The English and math grades were both pretty close to a normal distribution, but spread out and centered differently. English average was 80 with a standard deviation of 5, and math average was 70 with a standard deviation of 10.





Standardization and z-scores

When values are **standardized**, they are put into some kind of common unit. In statistics, we commonly standardize data by converting its units into **number of standard deviations from the mean**.

- We call a value that has been standardized in this manner a **z-score**.
- The units for a z-score are always "number of standard deviations from the mean". Once you convert a value to z, its original units go away.

Standardization and z-scores

We use this formula to standardize a value into a z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{how\ far\ away\ is\ the\ observation\ from\ the\ mean}{standard\ deviation}$$

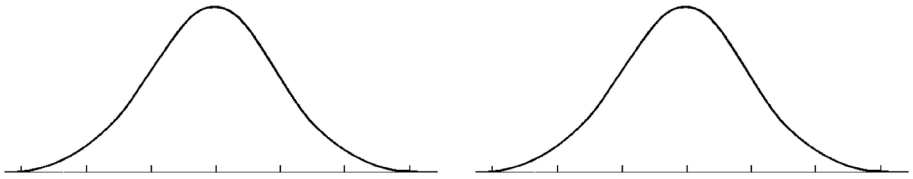
- *x* is the value to be standardized
- *μ ("mu")* is the population mean
- *σ ("sigma")* is the population standard deviation

Z-scores: Interpretation

- **Z-score** for an observations tells the **number of standard deviations away from the mean.**
- The sign of the z-score (+ or -) indicates whether the data point lies above or below the mean.
- A positive z-score means that the observation is above the mean.
- A negative z-score means that the observation is below the mean.
- The closer the z score is to 0, the more likely it is that it will occur.
- The farther the z score is from 0, the less likely it is that it will occur.

Example: Snowfall

You are debating if you should get a ski pass to Breckenridge or Aspen. Yearly snowfall is distributed normally. Data from 2009-2018 shows that Breckenridge gets an average of 770 inches of snow a year with a standard deviation of 200 inches. Aspen gets an average of 720 inches of snow a year with a standard deviation of 150 inches.



Example: Snowfall

Which resort is more likely to get below 500 inches of snow?

Which resort is more likely to get above 800 inches of snow?

Example: Mountain Lions

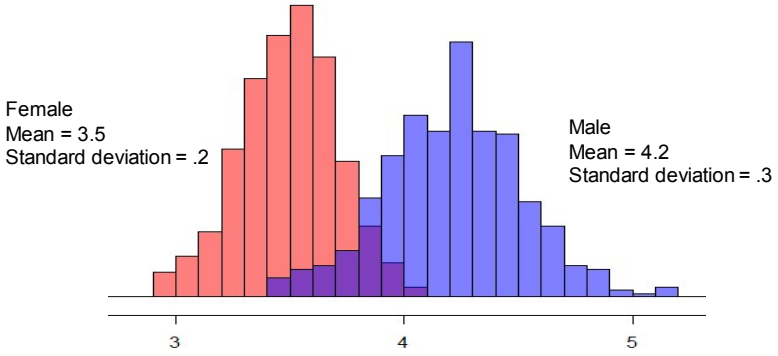
You are studying mountain lions in Colorado, which are hard to find. You come across some tracks on a trail. Male and female mountain lions have different track sizes, and you want to determine if the animal is a male or female.

Male mountain lions have an average track width of 4.2” and a standard deviation of .3”.

Female mountain lions have an average track width of 3.5” and a standard deviation of .2”.

Example: Mountain Lions

Here’s what a histogram would look like if we could sample many mountain lions and plot their track width.



iClicker: Mountain Lions

The track you come across has a width of 3.8”. Is it more likely to be male or female?

A. Male
B. Female
C. Impossible to know

iClicker: Mountain Lions

You come across another set of tracks on a different trail with a width of 4". Is it more likely to be male or female?

A. Male
B. Female
C. Impossible to know
D. I don't care I'm running away.

Example: Mountain Lions

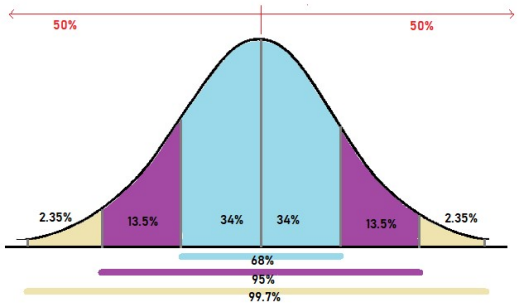
A small female would be a female with a track size of 3.2" or less. What would be the equivalent track size for a male?

Normal Distribution: A Handy Factoid

The graphic below illustrates a handy fact, called **the empirical rule** or the **68/95/99.7 rule**.

All normal distributions have the same shape, and when you mark off where the standard deviations are they will have the same percentage of observations falling in the same regions

- 68% of the distribution lies within 1 standard deviation of the mean
- 95% lies within 2 standard deviations
- 99.7% lies within 3 standard deviations



Example: Mountain Lions

About what percent of female mountain lions have a track width between 3.3" and 3.7"?

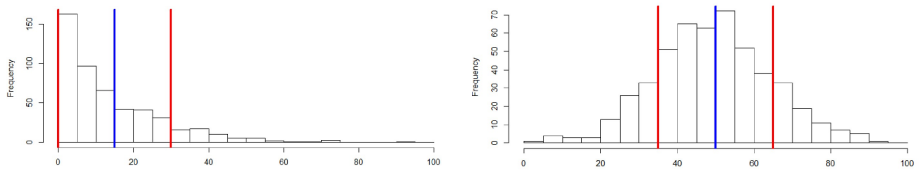
iClicker: Mountain Lions

About 95% of male mountain lions have a track width between what two values?

A. 3.9” and 4.5”
B. 3.3” and 3.7”
C. 3.6” and 4.8”
D. 3.1” and 3.9”
E. None of the above

Warning

Be careful when using z-scores for two data sets that do not have the same distribution shape. If some data looks normal and some data is skewed then likelihoods will not match up perfectly.



For example, in the distributions above if we go one standard deviation below the mean we have very different percentages above and below those values.

Z-scores and beyond

Using z-scores is great when we want to quickly compare two similar shaped distributions. We can compare the likelihoods of an observation from each of those distributions.

If the distributions are approximately normal we can quickly estimate the percentages using the 68/95/99.7 rule.

However, if we want to know the exact probability or compare distributions that look different, we’ll need more tools.

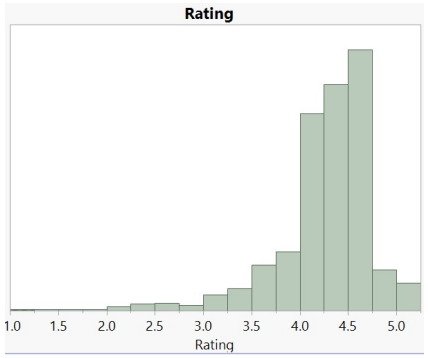
3.2 Percentiles

Example: Probabilities in JMP

We have a data set called GooglePlayStore with information about 2151 apps, including their rating, price, category, etc. Below is a histogram of the ratings of those apps from 1 to 5 stars.

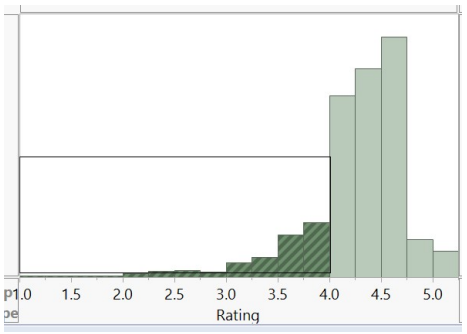
Notice that this is not quite a normal distribution. It is skewed left.

We won't be able to use the 68/95/99.7 rule to estimate probabilities. Instead we'll use a different strategy.



Example: Probabilities in JMP

What proportion of apps in the data set have a rating less than 4 stars?



We can do this in JMP:

- **Graph > Graph Builder**
- **Drag 'Rating' to the x or y axis**
- **Choose the histogram option**
- **Select the bars you want, and in the bottom left corner JMP will tell you how many observations have been selected**

Example: Probabilities in JMP

JMP reports that 328 rows have been selected. The sample size is n = 2151. With this information, we can compute the probability/relative frequency:

$P(Ratings < 4) =$

Percentiles

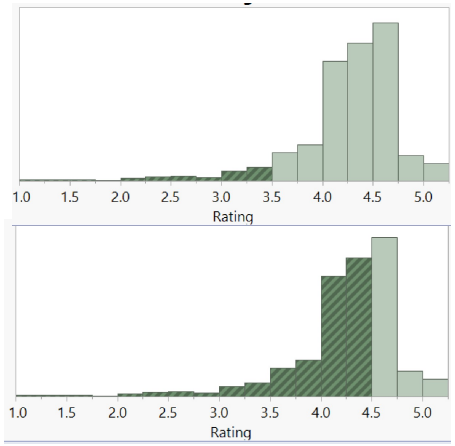
A **percentile** is the value of a variable for which the given percentage of values fall below it.

- In the Google Play Store data set, 4 stars is the _____ percentile for price.

Example: Ratings

$P(Ratings < 3.5) = 122/1251$

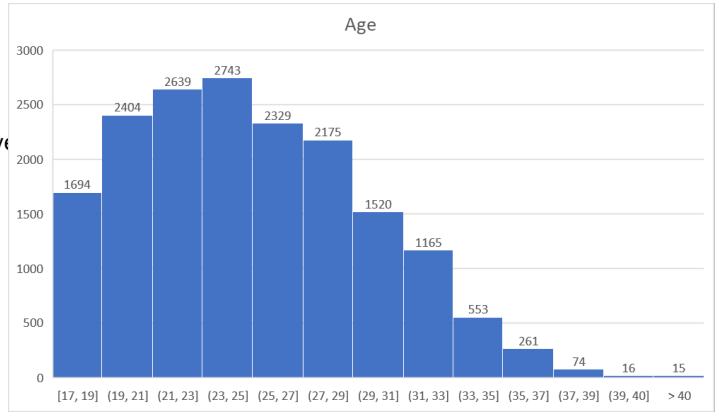
$P(Ratings < 4.5) = 1164/1251$



iClicker: Probability

The following histogram shows the ages of 17588 FIFA soccer players. What is the probability that a randomly selected player is less than 21 years old?

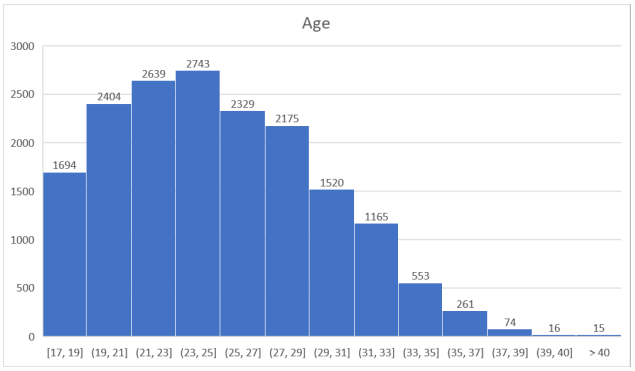
A. .001
B. .10
C. .13
D. .23
E. None of the above



iClicker: Percentiles

This means that 21 years old is the _____ percentile of age for FIFA soccer players.

A. 11th
B. 21st
C. 23rd
D. 77th
E. None of the above

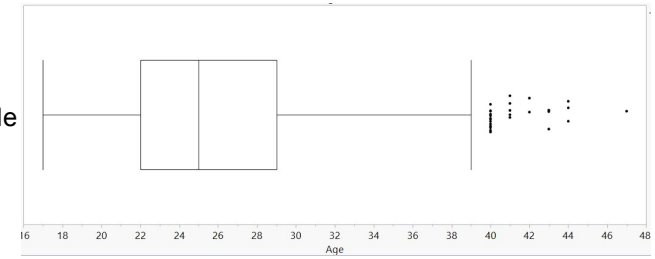


Boxplots and Quartiles

- Recall, the median is the value in a data set such that 50% of values fall below it. Therefore the median is the 50th percentile.

On a boxplot:

- Min = 0th percentile
- Q1 = 25th percentile
- Median = 50th percentile
- Q3 = 75th percentile
- Max = 100th percentile



Percentiles In JMP

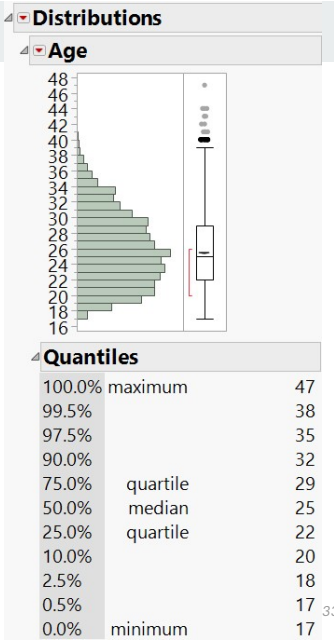
Another way to see percentiles in JMP is to go to:
Analyze > Distribution
Drag ‘Age’ into the ‘Y’

We can now pick out the five number summary:

- Min = 0th percentile =
- Q1 = 25th percentile =
- Median = 50th percentile =
- Q3 = 75th percentile =
- Max = 100th percentile =

And other percentiles like:

- 90th percentile =
- 2.5th percentile =



Wrap Up

A percentile is the value of a variable for which the given percentage of values fall below it.

Percentiles are a great way to asses probabilities and likelihoods on data that is not normally distributed or when comparing two variables which have different distributions.

Module 4:
Sampling Distributions

Review Concepts

Module 1:

- Descriptive vs. Inferential Statistics
- Histograms
- Sample Mean
- Standard Deviation

Module 3:

- Normal Distribution and Z-Score
- Percentiles

iClicker: Warm-up Question

Recall: What is statistical inference?

A. The process by which we tap our messy data with a magical statistics wand and turn it into solid, gold TRUTH

B. The process of when statistics try to infer things by themselves

C. The process by which we use information obtained from data and apply it to a larger population.

D. None of the above

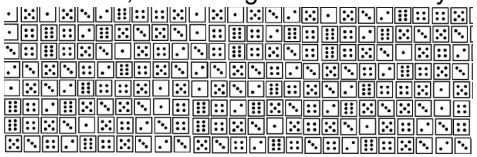
Section 4.1: The Law of Large Numbers

Example: Rolling 6-Sided Dice

A single 6-sided die has the values 1,2,3,4,5, and 6. The average of all these values is 3.5. Consider:

- If you only roll two dice, the average roll could easily be extreme as 1 or 6.

- If you roll 1000 dice, the average roll will be very close to 3.5.



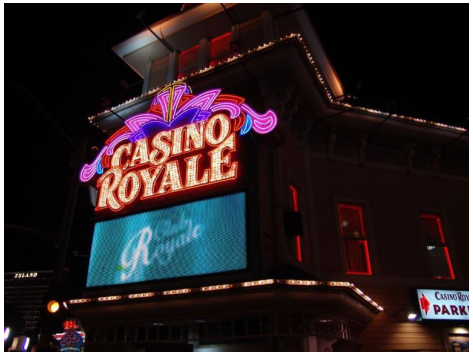
The Law of Large Numbers

- The Law of Large Numbers (LLN) is a mathematical result that tells us what tends to happen to a sample mean (i.e. average dice roll) as the sample size (i.e. # of dice rolled) gets bigger.

The Law of Large Numbers states: *as our sample size increases, the average of our sample will tend to get closer and closer to the true average of the population from which we are sampling.*

Example: Gambling

The Law of Large Numbers can be applied to gambling. Casinos do not go out of business because of too many gamblers getting lucky! The games are designed, on average, for gamblers to lose money in the long run.



https://en.wikipedia.org/wiki/Casino#/media/File:Casino_Royale_&_hotel.jpg

iClicker: Gambling

Which scenario would you rather bet $1000 on? Why?

A. That a single six-sided die roll is less than 5. (roll 1, 2, 3, or 4)

B. That the average of 10 dice rolls is less than 5.

Can we compute the actual chances?

What if we wanted to know exactly how great the chances are of 10 die rolls averaging less than 5? This is a probability that can actually be computed! We will need some new tools first.

Classical statistical inference is built on the behavior a collection of statistics (in our case the sample average) taken from lots of samples. The frequency interpretation of probability defines the probability of an outcome as its long run relative frequency.

If we want to know the probability that a statistic takes on some value, then we can take lots of samples and computing the same statistic repeatedly.

Section 4.2: Activity on Central Limit Theorem

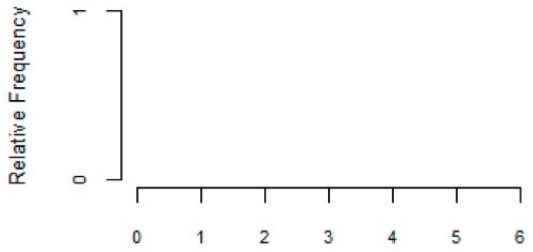
These slides match your CLT example handout. Use that to follow along.

Demonstrating CLT with dice

The next topic in our class will be the central limit theorem (CLT). Before we describe it mathematically in the lecture notes, we will first observe it using real data. The central limit theorem describes the behavior of a collection of sample means, taken under "repeated sampling".

We will simulate repeated sampling by rolling dice and recording the values of the rolls. Our variable of interest will be the result of a single die roll. The sample size will be the number of dice rolled.

First, let's consider the distribution of our variable at the "population" level. Record the probability of obtaining each dice value in the histogram below.



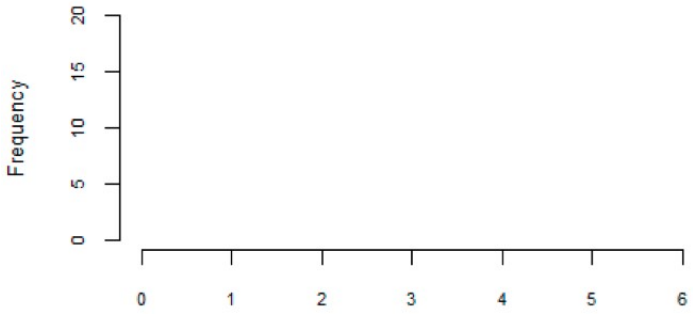
This is clearly not a normal distribution. In fact, it has its own name:

We will now take random samples of varying sizes from this distribution, calculate their means, and observe the behavior of these sample means. We will then plot histograms showing the overall results for the entire class.

Let's start by taking a "sample" of size n=1.

X_1

And we can make a histogram of the class's results:

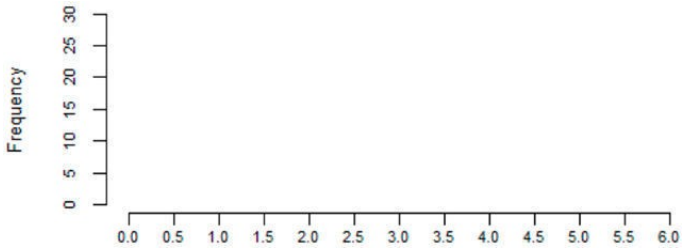


Next, we'll take samples of size n = 2. Roll a die twice, record the values, and record the sample mean:

X_1	X_2

$\bar{X} =$

Now, we can make a histogram showing the distribution of these means for the whole class:



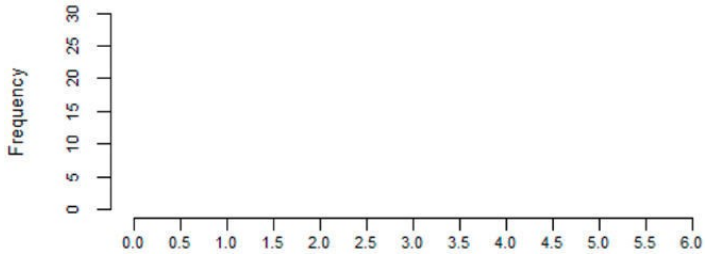
Compare this histogram to the histogram of the original data. How is it different?

Now let's do the same thing again, but using samples of size n=5:

X_1	X_2	X_3	X_4	X_5

$\bar{X} =$

Again, we can make a class histogram out of these sample means:



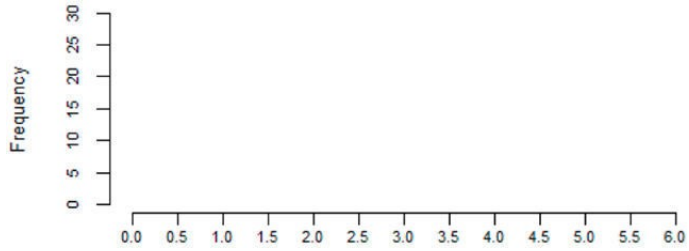
Compare this histogram to the previous two. How is it different?

Finally, let's do the same thing again, using samples of size n = 20 and recording their means:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}

$\bar{X} =$

Again, we can make a histogram out of these sample means:



Does this histogram fit with the pattern we've already seen?

Conclusions and Discussion:

- Think - Pair - Share with each other: What's a reason you can think of that led us to seeing the change in the histogram shapes as we increased the number of rolls?
- Even if we had a skewed distribution to start (as opposed to uniform), the CLT tells us that the "average of the averages" will be normally distributed as the sample size increases. Discuss why that is.
- Keep this example in mind as we move on to define CLT and do calculations based on this idea.

Example: Bringing the dice back!

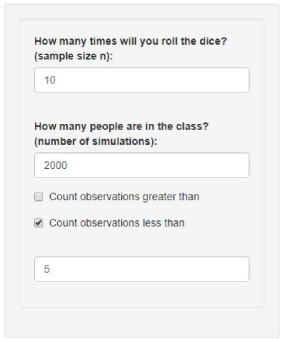
We said that we can compute the chances are of 10 die rolls averaging less than 5. Now let's actually do it.

The six numbers on a die have a mean of 3.5 and a standard deviation of 1.87. What is the probability that the average of 10 die rolls will be less than 5?

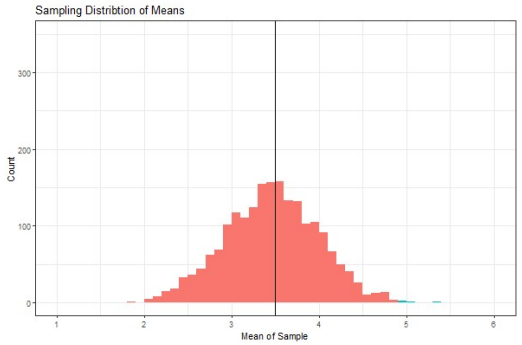
Let's use a computer simulation to pretend that we have a really big class of 2000 students and see how many times we get an average less than 5.

Example: Dice

The following program will simulate the dice example but with more people in the class.



COUNT: 1996 of the 2000 people (99.8 %) got a sample mean less than 5



Section 4.3: The Central Limit Theorem

The Central Limit Theorem

The **Central Limit Theorem** tells us that for any population distribution (no matter how skewed or strange the population is), if we repeatedly take new random samples from this distribution and calculate the mean each time, then:

- As sample size increases, the sample average should get close to the population average. (This is the Law of Large Numbers)
- As sample size increases, the sample averages will be less spread out.
- As sample size increases, the distribution of the sample averages will look more like a normal distribution.

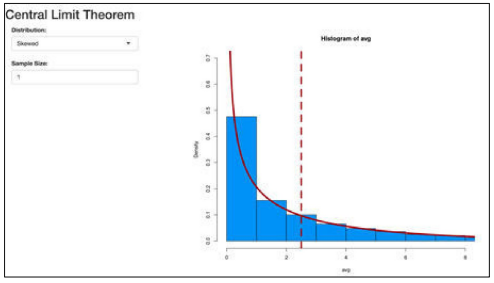
The Central Limit Theorem

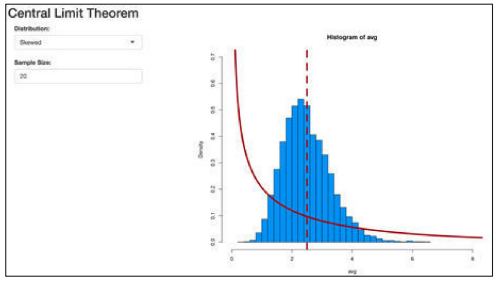
Why do we need this?

- We often work with non-normally distributed data, and usually the statistic we care most about is the mean.
- CLT is of great importance in classical statistical inference. It allows us to use the properties of a normal distribution (and tools like a z-score) when drawing inference on unknown a population mean, even when the variable of interest is not normally distributed.

The Central Limit Theorem

There is a simulation on Canvas demonstrating the central limit theorem. Here is a direct link:
https://benprytherch.shinyapps.io/CentralLimitTheorem/





How big of a sample size do we need?

- How large must our sample size be before we can be confident that CLT applies (the distribution of sample means will be normal)? This depends upon how far from (or close to) normal the underlying distribution is.
- Extremely skewed distributions require larger sample sizes. Distributions that are already normal will always have normally distributed sample means.
- No exact sample size cutoff, but many textbooks use ***n = 30*** as a safe rule-of-thumb

Section 4.4: Connection to Sampling Distribution

Defining Sampling Distribution

- **Distribution**: tells us the values that a ***variable*** takes on, and how often it takes them on.
- **Sampling Distribution**: tell us the values that a ***statistic*** takes on, and how often it takes them on. *It is the distribution of a statistic "under repeated sampling".*
- In our dice example, we plotted the sampling distributions of average die roll out of 2, 5, and 20 die rolls at a time. We kept taking new samples of these various sizes from the population of die rolls over and over again, and each time we recorded the sample mean.
  Each histogram we created displayed a sampling distribution of means.

Sampling Distribution

We use the term **"sampling variability"** to refer to the fact that new random samples will produces different values for the same statistic.

- The mean of one sample won't be the same as the mean of the next. Hence there is sampling variability in the value of a sample mean.

Note: In classical (frequentist) inference, all statistics (not just means) have associated sampling distributions.

- We could make a sampling distribution for a median, for a maximum, for the interquartile range, etc. However we would not be guaranteeed that those sampling distribtuions would look normal.

Example: Polling Data

Here is a tangible example. This table shows the results of opinion polls listed on RealClearPolitics.com

- Notice that "Economist/YouGov" reports lots of different numbers. These are different samples of people. When we take new samples repeatedly, we see sampling variability in our statistics.

Polling Data

Poll	Date	Sample	Approve	Disapprove	Spread
RCP Average	7/5 - 9/12	--	15.6	72.0	-56.4
Economist/YouGov	9/10 - 9/12	1313 RV	10	65	-55
Reuters/Ipsos	9/8 - 9/12	1669 A	24	62	-38
Economist/YouGov	9/3 - 9/5	1309 RV	10	63	-53
Reuters/Ipsos	9/1 - 9/5	1672 A	25	64	-39
FOX News	8/27 - 8/29	1006 RV	15	74	-59
Economist/YouGov	8/27 - 8/29	1278 RV	11	69	-58
Reuters/Ipsos	8/25 - 8/29	1767 A	22	66	-44
Economist/YouGov	8/20 - 8/22	1327 RV	11	67	-56
Quinnipiac	8/17 - 8/23	1514 RV	10	83	-73
Reuters/Ipsos	8/18 - 8/22	2744 A	21	67	-46
PPP (D)	8/18 - 8/21	887 RV	9	73	-64
Economist/YouGov	8/13 - 8/15	1291 RV	8	68	-60

iClicker: Notation review

Select the sequence of symbols that correctly corresponds with the order of the following terms:

population standard deviation, sample mean, population mean, sample size

A. $s, \mu, \bar{x}, n$

B. $\sigma, \mu, \bar{x}, n$

C. $\sigma, \bar{x}, \mu, n$

D. $\mu, \bar{x}, n, s$

E. $s, n, \bar{x}, \mu$

Formalizing the Central Limit Theorem

We need to get a fancier definition of the Central Limit Theorem:

For any variable X with mean μ and variance σ², the distribution of $\bar{x}$ converges to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, as n goes to infinity.

- Here, "as *n* goes to infinity" can just be thought of as the sample size getting larger and larger.
- Notice that the mean of the sampling distribution stays at the same as the original population mean.

What about that change in standard deviation?

- We see in the previous slide that if the variance of *X* is σ^2, then the variance of $\bar{x}$ is $\frac{\sigma^2}{n}$.
- This means that the standard deviation of $\bar{x}$ *is* $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$
- **Discuss:** Why do we see this change? What does this mean about the sampling variability of $\bar{x}$, relative to the variance of *X*? (for sample size $n > 1$)

Standard Error

- Recall that "**standard deviation**" refers to the standard amount by which the value of a *variable* will deviate from *its mean* (Module 2).
- The **standard error** is the standard deviation of the statistic under repeated sampling. It refers to the standard amount by which the value of a *sample statistic* will deviate from *the value of the unknown population parameter it is estimating*.
  - This corresponds to how spread out the sampling distribution of that statistic is.
  - When we use a statistic to estimate the value of a parameter, we know our estimate will be wrong a.k.a "in error", so, "standard error" is the standard amount by which a statistic will be in error.

Example: Drawing the Sampling Distribution

Lets say you have a variable X that is normally distributed with population mean μ = 40 and standard deviation σ = 10. You repeatedly take samples of size n = 25. Draw the population distribution and the sampling distribution.

Almost there: Introduction to Calculations

- Since we know that the sampling distribution of a sample mean will converge to a normal distribution with mean μ and standard error of $\frac{\sigma}{\sqrt{n}}$, we can convert any sample mean to a z-score and find probabilities associated with it, using a slightly modified z-score formula:

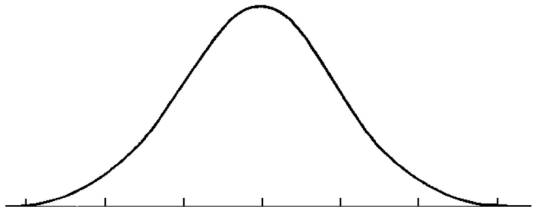
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- In order to use this formula, μ and σ must either be known or assumed.

iClicker: Practice the formula

Donations to your charity are known to have mean μ=$50 and standard deviation σ=$18. You take a sample of 36 donations and your sample mean is $46. What is the resulting z score?

A) -1.33
B) -.22
C) -10.45
D) -.05
E) None of the above

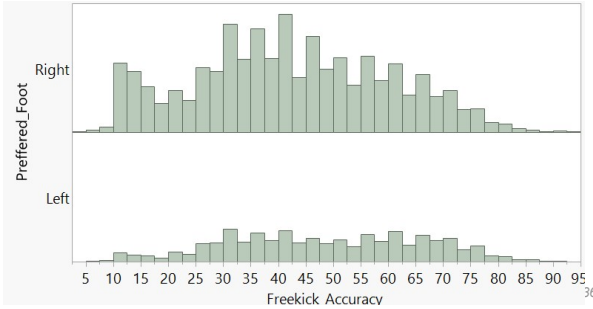


Example: Soccer Players

You are interested in comparing the freekick accuracy of right and left footed soccer players. The histogram below shows the preferred foot and freekick accuracy of 17588 FIFA soccer players (let's consider this the population). Then at the population level:

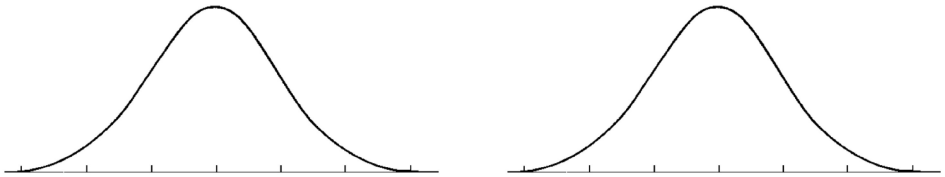
Right foot
Average = 42%
Standard deviation = 18%

Left foot
Average = 48%
Standard deviation = 17%



Example: Soccer Players

You make a teams of 11 random left footed soccer players and 11 random right footed soccer players. Let's sketch the sampling distributions of the teams.



Example: Soccer Players

You make a team of 11 random left footed soccer players and 11 random right footed soccer players. In which of the two samples are you more likely to get an average accuracy of 45%.

Right foot
Average = 42%
Standard deviation = 18%

Left foot
Average = 48%
Standard deviation = 17%

Example: Soccer Players

If your left footed team had an average accuracy of 60% you might consider that to be unusual. What would be the equivalent average for the right footed team?

Right foot
Average = 42%
Standard deviation = 18%

Left foot
Average = 48%
Standard deviation = 17%

iClicker: Soccer Players

If you made many left footed teams, then about 95% of the time you would get a team with an accuracy between what two values?
(Hint: Use the 68/95/99.7 rule)

A. 14 to 82
B. 37.8 to 58.2
C. 42.6 to 53.4
D. 31.8 to 64.2
E. None of the above

Left foot
Average = 48%
Standard deviation = 17%

Wrap-up: Remember the sampling distribution!

In the next two modules, we will introduce two formal statistical inferential procedures: constructing confidence intervals and testing hypotheses.

In both procedures, we will be using the sample data to make claims about population parameters. These population parameters will represent quantities of scientific interest.

We will be justifying our inferential claims by appealing to the properties of sampling distributions.

Module 5: Confidence Intervals

Review Topics

- Concept of statistical inference (Module 1)
- Sample standard deviation vs. population standard deviation meaning and notation (Module 1)
- Z-distribution: conversion and characteristics (Module 3)
- Concept of sampling distribution and variability; standard error (Module 4)

iClicker: Warm-Up Question

What is a population parameter and how does it differ from a sample statistic?

A. Population parameter is a number. Sample statistic is a letter.

B. A population parameter is a quantity/number pertaining to the whole population of interest. A sample statistic is a quantity/number pertaining to a subset of the population.

C. A population parameter smells better than a sample statistic.

D. A population parameter is a quantity/number pertaining to a subset of a population. A sample statistic is a quantity/number pertaining to a population and a sample.

Section 5.1: Introduction to Confidence Intervals

Introduction to Confidence Intervals

- These notes cover a very popular inferential method: constructing confidence intervals.
- The basic idea: we use sample statistics to estimate the value of unknown population parameters. This is an "inferential" process; we are generalizing from a sample to a population.
- The only problem: We know our estimates are almost always wrong.

Example: Estimating mean U.S. male height

- We randomly select $n = 20$ U.S. adult males and measure their heights. We compute a sample mean of $\bar{x} = 70.2$ inches
- This can be treated as an *estimate* for the population mean height, which we denote as μ.
- Do we therefore really believe that $\mu = 70.2$? Of course not! This is merely our best guess from our sample.
- So what can we do? We can build an interval around the sample mean. If the interval is wide enough, it should contain the population mean.

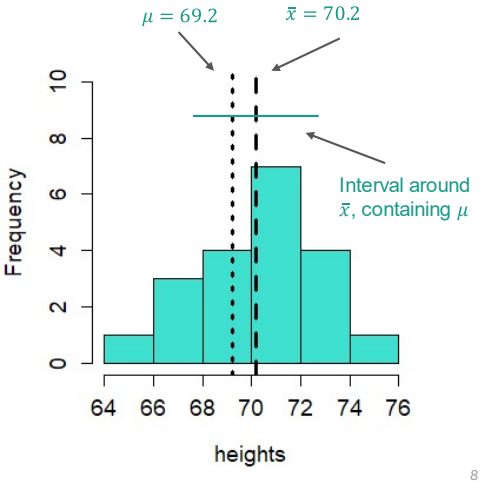
Confidence Interval - definition

- An interval is a range of values. For example "68 inches to 72.4 inches" might refer to all heights between 68 and 72.4 inches. This would be written as $(68, 72.4)$.
- A confidence interval is a range of values created for the purpose of capturing the value of an unknown population parameter value.
- A confidence interval has an associated "confidence level", which is the success rate. For example, a 95% confidence interval is an interval made in such a way that, 95% of the time, this kind of interval will capture the unknown population parameter value.

Example: Estimating mean U.S. male height

- We have $\bar{x} = 70.2$ and $\mu = 69.2$
- We want to build an interval around $\bar{x}$ that will contain μ. Let's use $(68, 72.4)$.
- The Center for Disease Control (CDC) reports* average adult male height to be 69.2 inches. We'll assume this is the population mean, μ. In practice, we usually do not μ.

* https://www.cdc.gov/nchs/data/series/sr_03/sr03_039.pdf



Confidence Interval - formula

- A confidence interval is created by taking a "point estimate" and then adding and subtracting a "margin of error".
- Point estimate: the value of a statistic that is being used to estimate an unknown population parameter value. For example, $\bar{x} = 70.2$ was our point estimate for population mean U.S. adult male height (μ).
- Margin of error: the maximum amount by which our point estimate should be wrong. For example, a margin of error of 2.2 inches means that we think $\bar{x} = 70.2$ could differ from μ by as much as 2.2 inches.

Confidence Interval - formula

$$Confidence\ Interval = point\ estimate \pm margin\ of\ error$$

- In our example, we have $\bar{x} = 70.2$ and $margin\ of\ error\ = 2.2$, so our confidence interval is :

$$Confidence\ Interval = 70.2 \pm 2.2 = (68, 72.4)$$

- So, our data suggest that μ, the population mean U.S. male height, lies somewhere between 68 inches and 72.4 inches.
- We will see how to calculate margin of error soon.

Confidence Intervals in Real Life

Survey shows strong demand for brick and mortar banking

Younger people between the ages of 21 and 34 are using brick and mortar banks more frequently than their older counterparts, according to a recent Consumer Insights Survey.

Of those surveyed between the ages of 21 to 34, 54 percent reported they walk into a local banking branch at least once a month.

The survey of 600 Texans was released by SWACHA, a non-profit electronic payments association based in Dallas. Dennis Simmons, CEO of SWACHA, says the survey results show that there is still a strong demand today for brick and mortar banking.

"Most people assume that the internet is eliminating the need for brick and mortar banks, and the data simply does not support that," Simmons says. "We have found that those who use online banking at higher rates also use the services at their local bank branch more often, and younger people are leading the way back into the banks."

The survey was conducted in April 2013 by Decision Analyst with a confidence interval of 95 percent and a corresponding margin of error of +/- 4 percent.

https://www.bizjournals.com/sanantonio/blog/morning-edition/2014/01/survey-shows-strong-demand-for-brick.html

iClicker: Confidence Interval

Fitness trackers need to know the average number of steps taken per mile. In an experiment done a team of researchers reports a 95% confidence interval for number of steps per mile while casually walking to be (2000,2500).

What is the average from their sample and margin of error used?

A. Average: 2000, Margin of Error: 500
B. Average: 2000, Margin of Error: 250
C. Average: 2250, Margin of Error: 500
D. Average: 2250, Margin of Error: 250
E. I don't know

Section 5.2: Quantifying Uncertainty with Confidence Intervals

General Formula of CI

- "Confidence interval" will be abbreviated "CI"

CI = point estimate ± margin of error

- Margin of error is just some number of standard errors. We call the number of standard errors the "**critical value**":

margin of error = critical value * standard error

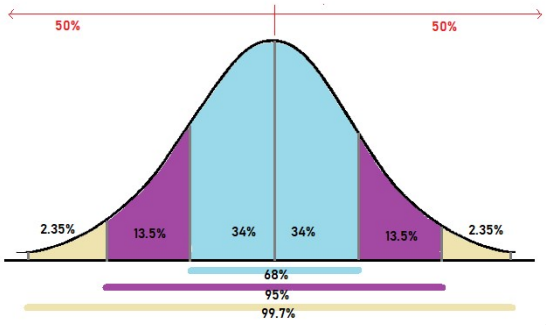
- So the general formula of a confidence interval is:

CI = point estimate ± critical value * standard error

General Formula and Big Idea of CI

The Empirical Rule (68-95-99.7 rule) says that the middle 95% of the normal distribution lies within 2 standard deviations of the mean (technically it is 1.96 but we round to 2).

Thus since 95% of values for a sample mean will fall withing two standard errors of the population mean, in order to make a 95% confidence interval we will take our sample mean and extend out to two times standard error.



Margin of Error

The margin of error is calculated by multiplying the standard error of our estimate by a **critical value.** The critical value is based on the desired Confidence Level.

Critical Value	Confidence Level

Confidence level: the rate at which confidence intervals successfully capture an unknown population parameter, under repeated sampling of data from a single population.

Confidence interval simulation

- Here is a link to a simulation demonstrating this idea of making confidence intervals using repeated sampling. This link is also on the Canvas page vis Additional Class Stuff / Simulations: http://students.brown.edu/seeing-theory/frequentist-inference/index.html#section2
- To make 95% confidence intervals, move the "$1 - \alpha$" slider so that its value is 0.95 (we will cover what "$1 - \alpha$" refers to in the next module)
- You can experiment with this success rate. See what happens when you reduce it to 0.5
- You can also experiment with the sample size. What happens when it gets larger? What about when sample size gets smaller?

Approximate 95% confidence interval

- A 95% confidence interval requires a margin of error of approximately 2 standard errors:

95% critical value ≈ 2

95% CI ≈ point estimate ± 2 * standard error

- This formula is approximate. The critical value will be larger than 2 if the sample size is small, and smaller than 2 if the sample size is large.
- JMP will compute exact 95% CIs. To approximate 95% CIs, you will always use a critical value of 2.

(side note: the exact critical value depends on "degrees of freedom", a concept we will not address in this class. Ask your instructor if you're curious what it refers to)

Practice: CI for mean lightblub lifespan

- Suppose we are interested in estimating the population average lifespan of a certain brand of incandescent light bulb. We denote this average μ.
- We collect a sample of $n = 50$ bulbs and run each bulb until it fails. We record how long each bulb lasts.
- Sample average lifespan is $\bar{x} = 1{,}958$ hours, and sample standard deviation of lifespan is $s = 65$ hours.
- We want to create a 95% CI for μ

Practice: CI for mean lightblub lifespan

Create a 95% confidence interval for μ

iClicker: CI for mean lightblub lifespan

Suppose we take a new sample of $n = 100$ light bulbs. We calculate $\bar{x} = 1979$ and $s = 42$. Create a 95% confidence interval for μ.

A. (1974.8,1983.2)
B. (1970.6,1987.4)
C. (1979.84,1978.16)
D. (1979,1978.16)
E. I don't know

Practice: CI for mean lightblub lifespan

- What population are we drawing inference to in this example?
- Can you think of any sources of bias?

Confidence interval interpretation

- Informal interpretation: we don't know what the true population mean lifespan of these bulbs is, but we think it's in the confidence interval.
- Formal interpretation: this is a confidence interval computed using a method that has a 95% success rate for capturing μ.
- So, we have "confidence" that μ is in the interval, because the interval was created using a method that successful captures μ 95% of the time.

What affects the width of a CI?

- This is a very important question you will need to be able to answer.
- Wider CIs represent greater uncertainty in our estimates. The wider the interval, the greater the range of plausible values for the unknown population parameter
- Conversely, narrower CIs represent less uncertainty (more certainty) in our estimates

iClicker: What affects the width of a CI?

The formula for a CI for a mean is

$$CI\ for\ \mu = \bar{X} \pm \underbrace{critical\ value \cdot \frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

If you collect a sample with a larger sample standard deviation the confidence interval will be:

A. The same
B. Wider
C. Smaller
D. I don't know

iClicker: What affects the width of a CI?

The formula for a CI for a mean is

$$CI\ for\ \mu = \bar{X} \pm \underbrace{critical\ value \cdot \frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

If you collect a sample with a larger sample size the confidence interval will be:

A. The same
B. Wider
C. Smaller
D. I don't know

iClicker: What affects the width of a CI?

The formula for a CI for a mean is

$$CI\ for\ \mu = \bar{X} \pm \underbrace{critical\ value \cdot \frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

If you make the confidence interval at a higher confidence level (99.7% instead of 95%) the confidence interval will be:

A. The same
B. Wider
C. Smaller
D. I don't know

Section 5.3: Confidence Intervals in JMP

How to get confidence intervals in JMP

- You've already seen 95% CIs in JMP, you just weren't familiar with them yet.
- Use Analyze / Distribution to obtain descriptive statistics for a variable. By default, JMP gives the lower and upper limits for the 95% CI for μ.
- Let's see how this works in practice using an example.

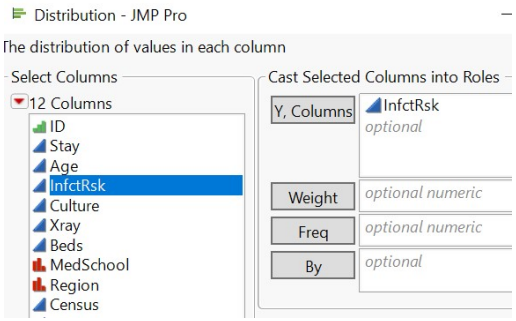
Example: Hospital infections

The data set called Hospital Infections. In 1980, an effort to reduce the risk of infection which can result from hospital stays, researchers collected data on several variables for 113 patients at various hospitals in the United States. Variables included infection risk, age, and length of stay, amongst others. Shown below is a snapshot of some of the data.

ID	Stay	Age	InfctRsk	Culture	Xray	Beds	MedSchool	Region	Census	Nurses	Facilities
1	7.13	55.7	4.1	9	39.6	279	2	4	207	241	60
2	8.82	58.2	1.6	3.8	51.7	80	2	2	51	52	40
3	8.34	56.9	2.7	8.1	74	107	2	3	82	54	20
4	8.95	53.7	5.6	18.9	12...	147	2	4	53	148	40
5	11.2	56.5	5.7	34.5	88.9	180	2	1	134	151	40

Example: Hospital infections

Using the hospital data, let's get summary statistics for infection risk:



Summary Statistics	
Mean	4.3548673
Std Dev	1.340908
Std Err Mean	0.126142
Upper 95% Mean	4.6048015
Lower 95% Mean	4.104933
N	113

Example: Hospital infections

- JMP says that the 95% CI for population mean infection risk is $(4.10, 4.60)$
- Verify that you can calculate this interval using the mean, standard deviation, and sample size:

Summary Statistics	
Mean	4.3548673
Std Dev	1.340908
Std Err Mean	0.126142
Upper 95% Mean	4.6048015
Lower 95% Mean	4.104933
N	113

Example: Hospital infections

We can interpret this 95% CI as giving a range of plausible values for the population mean infection risk.

What is the population here?

Mean	4.3548673
Std Dev	1.340908
Std Err Mean	0.126142
Upper 95% Mean	4.6048015
Lower 95% Mean	4.104933
N	113

- Notice how narrow the CI is. The fairly small standard deviation and fairly large N result in a small standard error.
- The small standard error results in a narrow confidence interval.

iClicker: Interpret

This data is a sample of hospitals taken in 1980 in the USA.

Suppose someone said that they believe that the population mean infection rate for hospitals in the USA in 1980 was 5.0? Based on our confidence interval, does this seem like a reasonable claim? Why?

A. No, because 5.0 is in the interval we got.
B. No, because 5.0 is not in the interval we got.
C. Yes, because 5.0 is in the interval we got.
D. Yes, because 5.0 is not in the interval we got.
E. I have no idea

Section 5.4: Estimating a Difference in Means – Part 1

A. Inference for a difference
B. Standard error of $\bar{x}_1 - \bar{x}_2$
C. JMP Example: Brain waves
D. Interpreting the 95% CI for $\mu_1 - \mu_2$

Inference for a Difference

- We have looked at how confidence intervals can be used to draw inference on a single population mean.
- Many times, we are interested in comparing two means to each other.
- **Examples:**
  - In a controlled experiment, we might be interested in the difference in means between a control group and a treatment group.
  - Opinion pollsters: They don't just report estimates for how much support *one* candidate has. They report estimates for how big the difference is between support for one candidate and support for another.

Confidence Interval for a Difference

- We can construct confidence intervals for differences in population means or for differences in population proportions
- Here is the generic formula for a confidence interval, again. It will still apply for differences!

$$CI\ for\ any\ parameter = point\ estimate \pm margin\ of\ error$$

where

$$margin\ of\ error = critical\ value\ * standard\ error\ of\ the\ point\ estimate$$

Confidence Interval for a Difference

Applying this general formula to a difference in means, we have:

Parameter of interest:

Point estimate of parameter:

The standard error of the point estimate:
It is **not** the difference in standard errors for $\bar{X}_1$ and $\bar{X}_2$ separately. It has its own formula:

$$\text{Standard error of } \bar{X}_1 - \bar{X}_2 = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval for a Difference

- Putting it all together:

$$CI\ for\ parameter = point\ estimate \pm (critical\ value\ * \frac{standard\ error\ of}{the\ point\ estimate})$$

95% CI for $\mu_1 - \mu_2$ example: Brain waves

Journal of Abnormal Psychology
1972, Vol. 79, No. 1, 54–59

CHANGES IN EEG ALPHA FREQUENCY AND EVOKED RESPONSE LATENCY DURING SOLITARY CONFINEMENT [1]

- Recall the brain waves data from module 2.
- These data were collected on Canadian prisoners who were randomly assigned to spend a week in solitary confinement, or to stay in their regular cell. Alpha waves were measured after one week.
- We are interested in estimating the difference in mean alpha wave level between prisoners in solitary confinement and those in their regular cells.

Brain waves example: "long form"

- First though, take a look at the data and notice the formatting:
- Each row is an observation (in this case, a prisoner).
- Each column is a variable ("confinement – cell or confined") and each row is an observation.
- This format is called "long form".

	Confinement	Waves
1	cell	10.7
2	cell	10.7
3	cell	10.4
4	cell	10.9
5	cell	10.5
6	cell	10.3
7	cell	9.6
8	cell	11.1
9	cell	11.2
10	cell	10.4
11	confined	9.6
12	confined	10.4
13	confined	9.7
14	confined	10.3
15	confined	9.2
16	confined	9.3
17	confined	9.9
18	confined	9.5
19	confined	9
20	confined	10.9

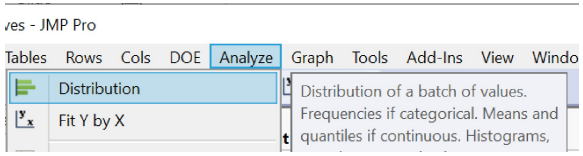
Brain waves example: "wide form"

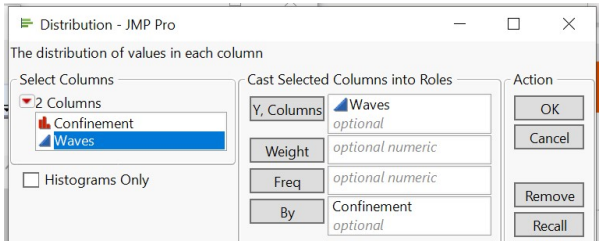
- Here is the WRONG way to format these data:
- This is known as "wide form", where there is a separate column for each category of a variable.
- This form implies that there are 10 prisoners, each of whom were measured twice: once in their regular cell and once in solitary confinement.
- But this is not true. There were 20 prisoners measured in one or the other. So, there should be 20 rows.

	cell	confined
1	10.7	9.6
2	10.7	10.4
3	10.4	9.7
4	10.9	10.3
5	10.5	9.2
6	10.3	9.3
7	9.6	9.9
8	11.1	9.5
9	11.2	9
10	10.4	10.9

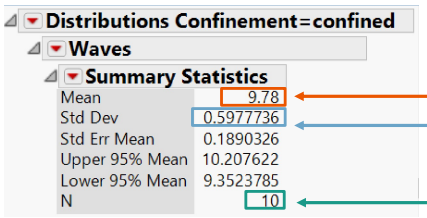
Brain waves example in JMP

- We can use Analyze / Distribution in JMP to obtain means and standard deviations:
- Recall that the "By" field is used to split summary statistics for the variable in "Y,Columns"





Brain waves example in JMP

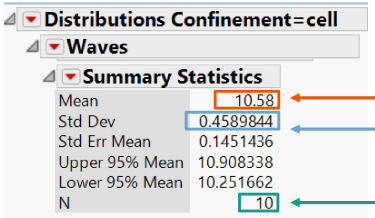


Let's call group 1 "Confinement" and group 2 "Cell"

$\bar{x}_1 = 9.78$

$s_1 = 0.5978$

$n_1 = 10$



$\bar{x}_2 = 10.58$

$s_2 = 0.4590$

$n_2 = 10$

Brain waves example

- Create an approximate 95% confidence interval for the difference in population mean alpha wave levels between prisoners who spend 7 days in their regular cell vs. those who spend 7 days in solitary confinement:

$$(\bar{X}_1 - \bar{X}_2) \pm 2 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

Interpreting the 95% CI for $\mu_1 - \mu_2$

- Recall that group 1 were the prisoners who were in solitary confinement, and group 2 were the prisoners who were in their regular cell.
- So the 95% confidence interval for $\mu_1 - \mu_2$ is a confidence interval for population mean alpha waves level for prisoners in solitary confinement, minus population means alpha waves level for prisoners in their regular cell.

Interpreting the 95% CI for $\mu_1 - \mu_2$

- The 95% CI is -1.276 to -0.324. Note that the negative values imply that the mean of the 2nd group is larger than the mean of the 1st group.
- So, mean alpha wave levels for prisoners in their regular cell was larger than mean alpha wave levels for prisoners in solitary confinement.
- Specifically, our data suggest that population mean alpha wave level for prisoners in solitary confinement after 7 days (μ_1) is less than the population mean alpha wave level for prisoners in their normal cells (μ_2) by anywhere between 1.276 units to 0.324 units.

iClicker: One more practice example

Suppose we are comparing tomato yield in pounds (lbs) for two kinds of fertilizers, A and B.

- 25 tomato plants got fertilizer A; they had a mean yield of 14.2 lbs with a standard deviation of 3.1 lbs.
- 28 tomato plants got fertilizer B; they had a mean yield of 12.7 lbs with a standard deviation of 4.2 lbs.

Calculate an approximate 95% CI for the different in population mean yields between plants receiving fertilizers A vs. B

A. (.5,2.5)
B. (-3.5,0.5)
C. (-0.5,3.5)
D. (1,2)
E. None of the above

Section 5.5: Estimating a Difference in Means – Part 2

Inference on "paired" data differences

- In the previous examples, we had two groups of data whose means we were comparing
    - Prisoners in solitary confiement vs. those in their regular cells
    - Tomato plants receiving fertilizer A vs. fertilizer B
- Sometimes, instead two *separate* groups of subjects, we have data for which each subject has been measured twice, and we want to draw inference on the average difference between the two observations on the same subject

Example: Paired data type studies

Examples of "paired" data:

- "Before" and "after" type studies (e.g. compare blood pressure before being put on a drug to after being on the drug)
- Studies where each subject is measured under two different treatments, or a treatment and a control (e.g. conduct a vision test with your right eye, then with your left eye).

Statistical Inference for "paired" data

- When we measure each subject twice, we call the two observations "paired" observations.
- If we want to draw inference using differences between paired observations, we can just subtract the paired observations from each other and make a confidence interval for the mean of the differences

Example: taking the ACT twice

- The ACT is a national standardized exam taken by students who are applying to college.
- Sometimes students will take the ACT more than once. Common wisdom is that scores should improve on repeated attempts, as students become more skilled at taking the exam.
- The dataset "ACT_SCORES.csv" contains sets of ACT scores for students who took the exam twice.

Data formatting: wide vs. long form

- Here are the first 10 rows of the data set.
- Remember that each row is an observation and each column is a variable.
- Because each student took the exam twice, "wide form" is the appropriate format for these data. "Long form" would suggest that each student took it only once.

"Wide" form

	ACT_1	ACT_2
1	21	24
2	16	19
3	16	19
4	29	33
5	13	14
6	29	27
7	25	26
8	30	34
9	19	19
10	23	27

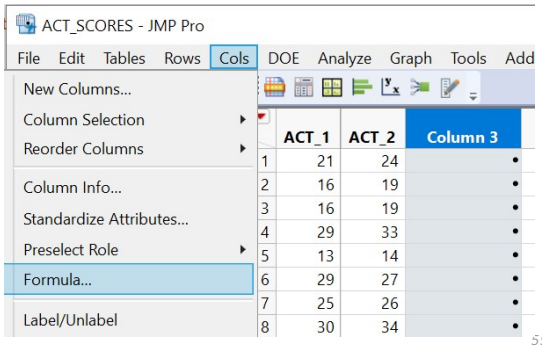
"Long" form

	Attempt	Score
1	ACT_1	21
2	ACT_2	24
3	ACT_1	16
4	ACT_2	19
5	ACT_1	16
6	ACT_2	19
7	ACT_1	29
8	ACT_2	33
9	ACT_1	13
10	ACT_2	14
11	ACT_1	29
12	ACT_2	27
13	ACT_1	25

(this continues to row 20)

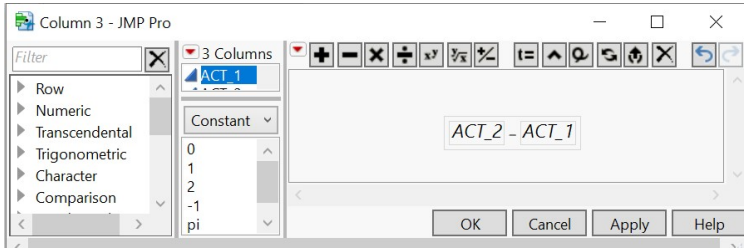
Getting the differences

- We want to make a confidence interval for the population mean *difference* between 1st and 2nd attempts at the ACT.
- So, we will make a new variable that is defined as the difference, and then make the confidence interval for this variable.
- Double click on a new column, then go to Cols / Formula.



Getting the differences

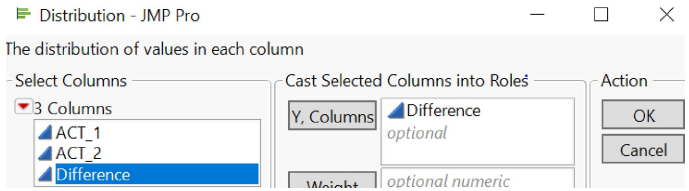
- The formula box lets you use the existing variables (under "columns"), basic mathematical operations, and more complex operations. See the JMP reference guide on Canvas for more details.
- Here, we just want to subtract the two columns. We are subtracting ACT_1 from ACT_2 so differences will be positive when score increases.



Getting the differences

- Click “OK” to finish. This screenshot shows that the differences were calculated. We also gave the column a more meaningful title than “Column 3”.
- We can also get summary statistics using Analyze / Distribution:

	ACT_1	ACT_2	Difference
1	21	24	3
2	16	19	3
3	16	19	3
4	29	33	4
5	13	14	1
6	29	27	-2
7	25	26	1
8	30	34	4
9	19	19	0



Making a 95% CI for the mean difference

Summary Statistics

Mean	1.8636364
Std Dev	2.0539267
Std Err Mean	0.4378986
Upper 95% Mean	
Lower 95% Mean	
N	22

The confidence interval is:

A. (-0.02,3.62)
B. (0.95,2.77)
C. (0.99,2.73)
D. (0.924,2.676)
E. None of the above

Now we can make a 95% confidence interval for the mean of the differences. This is just one population mean, so we can use the 95% CI formula for a single mean:

$$95\%\ CI\ for\ \mu = \bar{x} \pm 2 * \frac{s}{\sqrt{n}}$$

Paired difference vs. two sample difference

- Give an interpretation for the 95% CI on the previous slide:
- Notice that this refers to the mean amount by which ACT scores change when the same student takes it twice.

Paired difference vs. two sample difference

- What if we had analyzed this data as though it came from two different samples of students, each of whom had taken the ACT once?
- Here are the results when we analyze this as a 95% CI for $\mu_1 - \mu_2$:

Distributions

ACT_1 — **Summary Statistics**

Mean	21.863636
Std Dev	5.2217079
Std Err Mean	1.1132719
Upper 95% Mean	24.178812
Lower 95% Mean	19.548461
N	22

ACT_2 — **Summary Statistics**

Mean	23.727273
Std Dev	5.6415281
Std Err Mean	1.2027778
Upper 95% Mean	26.228586
Lower 95% Mean	21.225959
N	22

$$(\bar{X}_2 - \bar{X}_1) = 1.86$$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.64$$

Paired difference vs. two sample difference

- The 95% CI for the mean difference μ:

$$1.86 \pm 2 * 0.44 \approx (0.95, 2.77)$$

- The 95% CI for the difference in means $\mu_1 - \mu_2$:

$$1.86 \pm 2 * 1.64 \approx (-1.44, 5.17)$$

- What's going on here?
- First, notice that the point estimate is 1.86 in both cases. But the standard error is much smaller for the paired case.

Paired difference vs. two sample difference

- Here are descriptive statistics for ACT_1 (before), ACT_2 (after), and the difference between before and after:

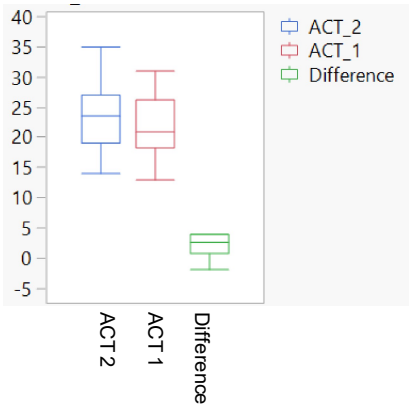
Distributions

ACT_1		ACT_2		Difference	
Summary Statistics		**Summary Statistics**		**Summary Statistics**	
Mean	21.863636	Mean	23.727273	Mean	1.8636364
Std Dev	5.2217079	Std Dev	5.6415281	Std Dev	2.0539267
Std Err Mean	1.1132719	Std Err Mean	1.2027778	Std Err Mean	0.4378986
Upper 95% Mean	24.178812	Upper 95% Mean	26.228586	Upper 95% Mean	2.7742964
Lower 95% Mean	19.548461	Lower 95% Mean	21.225959	Lower 95% Mean	0.9529763
N	22	N	22	N	22

- Notice the standard deviations!

Paired difference vs. two sample difference

- We could look at this using boxplots too:
- The standard deviation of the differences is much smaller than the standard deviations of the raw scores.
- This should make sense! There is a lot of variability in how students perform on the ACT, but this doesn't mean there should be a lot of variability in how much students' scores change from the first attempt to the second attempt.



Paired difference vs. two sample difference

- Look again at the raw data. Notice that there are big differences in students' scores. Also notice that there are not so big differences in how much students' scores change.
- This is the advantage of paired designs. We can get more precise estimates of what we're interested in when we measure the same people repeatedly.

	ACT_1	ACT_2	Difference
1	21	24	3
2	16	19	3
3	16	19	3
4	29	33	4
5	13	14	1
6	29	27	-2
7	25	26	1
8	30	34	4
9	19	19	0
10	23	27	4
11	21	23	2
12	19	20	1
13	27	27	0
14	21	24	3
15	23	21	-2
16	16	18	2

Paired or Not paired (Two Sample)

- Randomly sample 500 households for their tap water quality to measure level of contaminants in milligrams. 300 houses are in Fort Collins and 200 houses are in Greeley.
- Determine whether a new engine type will improve miles per gallon for a car by testing 20 cars' mpg with the old engine and 25 cars' mpg with the new engine.
- You measure the temperature and humidity level on 40 random days throughout the year.
- Select 20 bikes, 10 with tires of brand 1 and 10 with tires of brand 2, and track how many miles the bike was ridden before the tires need to be replaced.
- Survey 30 people entering a debate to find their approval on a mayoral candidate from 1 to 10 and follow up after the debate with the same people to get their new approval rating.

Section 5.6: Explaining Confidence and Error

How 95% confidence implies a 5% error rate

- We have seen that confidence intervals can be used to "exclude" values which fall outside the interval.
- In particular, if a confidence interval for a difference excludes zero, this may be treated as evidence that there is a real difference between the groups, at the population level.
- BUT – we shouldn't forget that confidence intervals sometimes fail to capture the population parameters they are trying to capture.
- In particular, a 95% CI will fail to capture the parameter of interest 5% of the time. In other words, a 95% CI has a 5% error rate. This is something to always keep in mind when we draw conclusions

Model assumptions

- When we construct and interpret confidence intervals, we make some assumptions.
- Notably for now, we assume that our data came from a normally distributed population. We write this:

$$x_i \sim Normal(\mu, \sigma), \qquad i = 1, \dots, n$$

- This is a very simple statistical "model". The model says that we collect a sample of size n from a normal distribution with population mean μ and population standard deviation σ.

Model assumptions

- The assumption that our data were sampled from a normally distributed population is probably false.
- Thankfully, the Central Limit Theorem says that sample means should be approximately normally distributed, as long as the data came from a population that isn't *too* skewed and the sample size isn't *too* small.
- Our statistical inferences are only as good as the assumptions that they depend on!

Sampling Error vs. "Plain Old" Error

- The 5% error rate for 95% confidence intervals refers only to sampling error (sometimes called "statistical error").
- This is the error that is just due to the randomness of the data. If we collect a new data set, we'll get different values for our statistics.
- This error rate does NOT account for error that could arise from violating the assumption of normality.
- And there are many other sources of error that are not accounted for in the "5%" error rate…

Other Potential Sources of Error

Confidence intervals attempt to account for sampling error. Confidence intervals do not account for:

- Biased samples
- Faulty measuring devices
- Poor experimental design (e.g. failing to double-blind)
- The violation of underlying assumptions (e.g. normally distributed data)
- Using parameters that don't reflect reality (e.g. treating population mean lifespan as a fixed value over the last 100 years)
- Failure to account for relevant differences between sub-groups (e.g. you estimate mean effect of a drug for all people, but the drug affects old and young people differently)
- Any other kind of error that isn't due to sampling variability

Module 5 summary

- We make confidence intervals because population parameters are unknown. We hope that the unknown parameter value falls somewhere inside the confidence interval.
- The general formula for a CI is $point\ estimate \pm margin\ of\ error$
- The formula for a 95% CI for μ is $\bar{x} \pm 2 * \frac{s}{\sqrt{n}}$
- The formula for a 95% CI for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 \pm 2 * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Under repeated sampling, 95% of 95% CIs should contain the true population parameter value.
- Confidence intervals do not correct for bias or low quality data. Garbage in, garbage out.

Module 6: Hypothesis Testing

Previous topics we will apply in this module

- Confidence Intervals (Module 5)
  - "Plausible" values
- Population parameters vs Sample statistics (Module 4 and 5)
- Sampling Distributions (Module 4)
- Central Limit Theorem (Module 4)
- Probability (Module 2 and 3)
- Calculating Mean/Standard Deviation (Module 1)

iClicker: Confidence Interval Review

Suppose you just made a 95% CI for a mean and got (30,56). Which best describes what the "95%" in a 95% confidence interval refers to?

A. *95% of intervals created this way will contain the population parameter*

B. *95% of intervals created this way will contain the point estimate*

C. *There's a 95% chance that the point estimate will equal the population parameter*

D. *There's a 95% chance that this interval contains the point estimate*

E. *I have no idea*

Section 6.1: Review

Review: Brain waves example

- Recall the brain waves example from module 5. Prisoners in a Canadian prison were randomly assigned to stay in their regular cells or spend a week in solitary confinement.
- We created a 95% confidence interval for the difference in population mean alpha wave levels between the two groups:

$$(\bar{X}_1 - \bar{X}_2) \pm 2 \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 9.78 - 10.58 \pm 2 \cdot \sqrt{\frac{0.5978^2}{10} + \frac{0.4590^2}{10}}$$

$$-0.8 \pm 0.48 = (-1.28, -0.32)$$

Review: Brain waves example

$(-1.28, -0.32)$

Population mean alpha wave level for prisoners in solitary confinement

Population mean alpha wave level for prisoners in regular cells

- We interpreted this as a range of plausible values for $\mu_1 - \mu_2$
- Note that both endpoints are negative. This means that all of our plausible values for the difference are negative.
- So, this confidence interval seems to give good evidence that μ_1 is smaller than μ_2 by some amount.

Review: Brain waves example

What if the confidence interval was $(-1.28, .32)$

- Since there are positive numbers in the CI it would be plausible that μ_1 was larger than μ_2.
- If the left endpoint is negative and the right endpoint is positive, then the confidence interval does not give strong evidence regarding which population mean was larger than the other.
- If zero is in the interval, then the interval does not provide strong evidence regarding which population mean is larger. If zero is not in the interval, then the interval gives a strong suggestion that one population mean is larger than the other.

Section 6.2: Hypothesis testing overview

Testing against a null hypothesis

- Prior to conducting the brain waves study, the researchers were interested in knowing if there would be evidence for lower alpha wave levels among prisoners in solitary confinement.
- Another way of thinking about this: they wanted to see if the data gave good evidence that the population mean alpha wave levels were not equal.
- This can be stated formally as a test against a **null hypothesis**, denoted H_0:

$$H_0: \mu_1 = \mu_2$$

Testing against a null hypothesis

$$H_0: \mu_1 = \mu_2$$

- This null hypothesis says that the two population means are equal.
- This can be re-written as:

$$H_0: \mu_1 - \mu_2 = 0$$

- This is the preferred form of the null hypothesis, because it makes it clear that we will be testing against a null value of zero difference.

Testing against a null hypothesis

$$H_0: \mu_1 - \mu_2 = 0$$

- Our confidence interval of $(-1.28, -0.32)$ does not contain zero.
- So, our data provides evidence against this null hypothesis, and we will reject it.
- In rejecting this null hypothesis, we are claiming to have statistical evidence that the difference in population means is **not** zero. Specifically, we are claiming to have evidence that μ_1 is smaller than μ_2.

Hypothesis testing: outline

A hypothesis test can be broken down into three broad steps:

1. State the null hypothesis
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision to either **reject** or **fail to reject (FTR)** the null hypothesis

Hypothesis testing: state the null hypothesis

- The null hypothesis (H_0) is the proposition we are testing **against**.
- The results of the hypothesis test will be to either reject H_0 or fail to reject H_0
- The example we just saw used $H_0: \mu_1 - \mu_2 = 0$. Other examples:
  - $H_0: \mu = 30$ ("The population mean equals 30")
  - $H_0: \mu_1 - \mu_2 = 5$ ("The difference between population means equals 5")
  - $H_0: \sigma_1 - \sigma_2 = 0$ ("There is no difference between population standard deviations")

Hypothesis testing: state the null hypothesis

- If you've taken a Statistics class before, or if you've seen hypothesis testing in another class, you may have heard of the "alternative hypothesis", typically denoted H_A or H_1
- The alternative hypothesis is typically just the negation of the null hypothesis.
- Example: if $H_0: \mu_1 - \mu_2 = 0$, then $H_A: \mu_1 - \mu_2 \neq 0$
- **We will not explicitly state alternative hypotheses in this class**. Which means you don't need to know the stuff on this slide!

Hypothesis testing: compute a statistic that tests the null hypothesis

- The null hypothesis will be tested using some statistic that we calculate from our data.
- In the brain waves example, a confidence interval was used to test the null hypothesis.
- There are other kinds of statistics called "test statistics" and "p-values" that can also be used to test a null hypothesis. We will consider these later.

Hypothesis testing: make the statistical decision

- There are two possible conclusions from a hypothesis test:
  - $Reject\ H_0$ ("reject the null hypothesis")
  - $FTR\ H_0$ ("fail to reject the null hypothesis")
- There will always be a rule for how you use a statistic to decide between rejecting and failing to reject H_0.
- When using confidence intervals, we $reject\ H_0$ if the hypothesized value (such as 0) is outside the interval. We $FTR\ H_0$ if this value is inside the interval.

Hypothesis testing: the statistical decision

If we **reject H_0**, we are saying that our data provide evidence against the null hypothesis.

- More specifically, we are saying that if the null hypothesis was true then the statistical results we obtained would be unlikely to happen.

If we **fail to reject (FTR) H_0**, then we say that our data do not provide strong evidence against H_0.

- Failing to reject H_0 is **not** the same thing as saying that our data suggest that H_0 is true. We do not "accept" H_0.

Hypothesis testing: interpreting "reject H_0"

- Consider the example:

$$H_0: \mu_1 - \mu_2 = 0; \qquad 95\%\ CI = (1.5, 2.6)$$

- Here, the 95% CI does not contain the null value, zero.
- That means that a difference of zero is not plausible.
- That means that there is a difference between the groups.

Hypothesis testing: interpreting "fail to reject H_0"

- Consider the example:

$$H_0: \mu_1 - \mu_2 = 0; \qquad 95\%\ CI = (-7.1, 2.6)$$

- Here, the 95% CI contains the null value, zero.
- That means a difference of zero is plausible.
- But the interval also contains -5, so a difference of -5 is plausible.
- And the interval contains 2, so a difference of 2 is plausible.
- Since there are many plausible values for the difference we can't say for sure that the difference is zero, but it might be
- This is why we say fail to reject H_0

iClicker: Quick Dry Cement

- A new brand of quick dry cement claims to set up (dry) in 15 minutes. A quality control team makes and pours 30 randomly selected bags. They compute a sample mean of 16.7 min, with a standard deviation of 5.3 min.
- Conduct a hypothesis test to see if there is strong evidence that this brand's population mean set up time exceeds 15min.

1. State the null hypothesis:
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision
   a. Reject H_0
   b. Fail to Reject H_0

iClicker: Interpret Cement

- A new brand of quick dry cement claims to set up (dry) in 15 minutes. A quality control team makes and pours 30 randomly selected bags. They compute a sample mean of 16.7 min, with a standard deviation of 5.3 min.

- Conduct a hypothesis test to see if there is strong evidence that this brand's population mean set up time exceeds 15min.

How do we interpret our decision?

A. We have enough evidence to claim that the average set up time is greater than 15 min
B. We have enough evidence to claim that average set up time is equal to 15min
C. We do not have enough evidence to claim that the average set up time is greater than 15 min
D. We have no evidence of anything.
E. I have no idea.

iClicker: Tomato Yeilds

- Tomato yield (in lbs) is compared for plants receiving fertilizers A and B.
  - 42 plants received fertilizer A and had sample mean yield of 15.7 lbs with a standard deviation of 5.5 lbs.
  - 41 plants received fertilizer B and had a sample mean yield of 14.1 lbs with a standard deviation of 6.2 lbs.

- Conduct a hypothesis test against the null of no difference in population means.

a. Reject H_0
b. Fail to Reject H_0

iClicker: Interpret Tomato Yeilds

Tomato yield (in lbs) is compared for plants receiving fertilizers A and B.

- 42 plants received fertilizer A and had sample mean yield of 15.7 lbs with a standard deviation of 5.5 lbs.
- 41 plants received fertilizer B and had a sample mean yield of 14.1 lbs with a standard deviation of 6.2 lbs.

Conduct a hypothesis test against the null of no difference in population means.

How do we interpret our decision?

A. We have enough evidence to claim that fertilizer A works better than fertilizer B
B. We have enough evidence to claim that the fertilizers produce the same yeild
C. We do not have enough evidence to claim that the fertilizers produce different yields
D. We have no evidence of anything.
E. I have no idea.

Hypothesis testing: assumptions

- First, we should consider standard assumptions made when performing hypothesis tests.

- We assume that population parameters exist, and that our data are sampled in an unbiased manner from some well defined population or populations.

- Typically we assume that our data come from a normally distributed population. For example, in the fertilizer example, we might assume:

$$X_{iA} \sim Normal(\mu_A, \sigma_A), \qquad iA = 1 \ldots n_A$$
$$X_{iB} \sim Normal(\mu_B, \sigma_B), \qquad iB = 1 \ldots n_B$$

Note: that the Central Limit Theorem is helpful here when we have large sample sizes.

Errors in Hypothesis Testing

- Recall that 95% confidence intervals capture the true population parameter value 95% of the time.
- This implies that confidence intervals (with 95% confidence level) fail to capture the true population parameter 5% of the time.
- Example: suppose that we test $H_0: \mu_1 - \mu_2 = 0$, and in reality $\mu_1 = \mu_2$ (H_0 is true)
- There is a 5% chance that a 95% CI will fail to capture the true population mean difference of zero. So, there is a 5% chance of rejecting H_0, even though H_0 is true. This scenario is called a Type I error (we will see Type II errors in a few slides.)

Type I and Type II errors

- A Type I Error is rejecting H_0, even though H_0 is true.
  - These will happen more often if our standard of evidence is too weak.
- A Type II Error is failing to reject H_0 , even though H_0 is false.
  - These will happen more often if our standard of evidence is too strong.

"The truth"

"The statistical decision"	H_0 is true	H_0 is false
Reject H_0		
Fail to Reject H_0		

Type I and Type II errors

- We (almost) never get to know if we are committing an error, because we (almost) never get to know whether H_0 is true or false.
- We can try to control the probabilities of committing a Type I or Type II error.
- When using a 95% CI, the probability of a Type I error is 0.05.
- The probability of a Type II error depends on statistical "power", which we will not cover in this class.

A courtroom analogy for hypothesis testing

- Some people find the following analogy useful when thinking about hypothesis testing:
- Suppose a defendant is on trial, accused of a crime. The defendant is considered "innocent until proven guilty".
- But guilt is rarely "proved". Instead there is some standard of evidence that must be met in order to overturn the presumption of innocence.
- Also, if the defendant is found "not guilty", this does not necessarily imply that the evidence shows they are innocent. It only means that the evidence was not strong enough to show guilt.

A courtroom analogy for hypothesis testing

- In hypothesis testing, the null hypothesis is analogous to the presumption of innocence. We only reject H_0 if we have strong enough evidence against it.
- And if we fail to reject the null, this does not necessarily mean we have evidence that the null is true. Just like a finding of "not guilty" does not necessarily mean there is evidence of innocence.
- We can also make errors. Finding an innocent defendant "guilty" is analogous to a Type I error. Finding a guilty defendant "not guilty" is analogous to a Type II error.

iClicker: Statistical decisions and errors

Suppose we are testing $H_0: \mu_1 - \mu_2 = 0$. We calculate the values below. Which option gives the correct statistical decision and possible error?

$$\bar{x}_1 - \bar{x}_2 = 24.2 \quad \text{and} \quad \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 18.8$$

A. $Reject\ H_0$, we may have committed a Type I error
B. $Reject\ H_0$, we may have committed a Type II error
C. $FTR\ H_0$, we may have committed a Type I error
D. $FTR\ H_0$, we may have committed a Type II error
E. I have no idea

iClicker: Statistical decisions and errors

Suppose we are testing $H_0: \mu = 9.0$. But in truth, $\mu = 9.8$. We calculate the values below. Which option gives the correct statistical decision and statement of truth?

$$\bar{x} = 10.7 \quad \text{and} \quad \frac{s}{\sqrt{n}} = 0.4$$

A. $Reject\ H_0$, we have committed a Type I error
B. $Reject\ H_0$, we have made the correct decision
C. $FTR\ H_0$, we have committed a Type II error
D. $FTR\ H_0$, we have made the correct decision
E. I have no idea

iClicker: Errors with Cement

- A new brand of quick dry cement claims to set up (dry) in 15 minutes. A quality control team makes and pours 25 randomly selected bags. They compute a sample mean of 16.7 min, with a standard deviation of 5.6 min.
- Conduct a hypothesis test to see if there is strong evidence that this brand's population mean set up time exceeds 15min.

What would a Type I error be in this situation?

A. The cement really dries in 15 minutes but we think it takes longer.
B. The cement really dries in 15 minutes and we agree with their claim.
C. The cement really takes longer than 15 minutes to dry and we correctly tell them they're wrong.
D. The cement really takes longer than 15 minutes to dry but we mistakenly believe it might take only 15 minutes.
E. I have no idea.

iClicker: Errors with Cement

- A new brand of quick dry cement claims to set up (dry) in 15 minutes. A quality control team makes and pours 25 randomly selected bags. They compute a sample mean of 16.7 min, with a standard deviation of 5.6 min.
- Conduct a hypothesis test to see if there is strong evidence that this brand's population mean set up time exceeds 15min.

What would a Type II error be in this situation?

A. The cement really dries in 15 minutes but we think it takes longer.
B. The cement really dries in 15 minutes and we agree with their claim.
C. The cement really takes longer than 15 minutes to dry and we correctly tell them their wrong.
D. The cement really takes longer than 15 minutes to dry but we mistakenly believe it might take only 15 minutes.
E. I have no idea.

Section 6.3: From the 95% CI to the t-test statistic

From the 95% CI to the t-test statistic

- We have seen that we reject H_0 when a 95% CI does not contain the null value.
- You may have noticed that a 95% CI will not contain the null value is more than 2 standard errors away from the point estimate.
- This is another way of conducting a hypothesis test: find the difference between the point estimate and the null value, then divide this difference by the standard error. If this value is greater than 2, reject H_0:

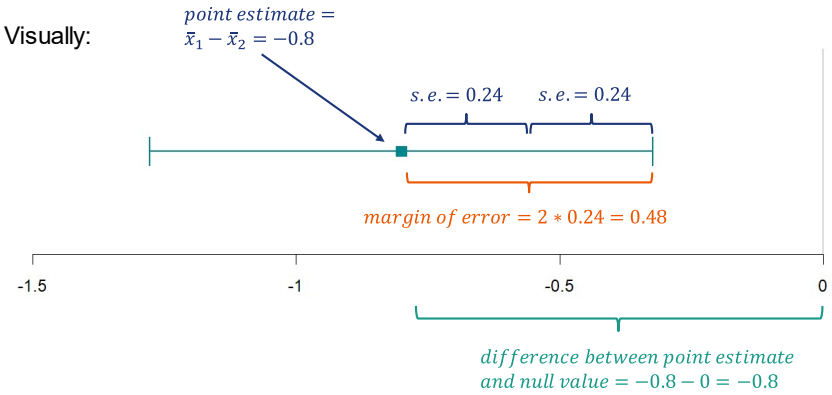
$$if \quad \frac{point\ estimate - hypothesized\ value}{standard\ error} > 2 : \ reject\ H_0$$

From the 95% CI to the t-test statistic

- In the brain waves example, we computed 95% CI = $(-1.28, -0.32)$, which did not contain the null value of zero, so we rejected $H_0: \mu_1 - \mu_2 = 0$.
- Here the unknown parameter of interest is $\mu_1 - \mu_2$. The point estimate for this unknown parameter is $\bar{x}_1 - \bar{x}_2 = -0.8$
- The standard error of the point estimate is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.24$
- The margin of error is $2 * standard\ error = 2 * 0.24 = 0.48$

From the 95% CI to the t-test statistic

- Visually:



The "t" test statistic

- We can calculate the number of standard errors away the point estimate and the null value are with this fraction, called the "t" test statistic:

$$t = \frac{point\ estimate - null\ value}{standard\ error}$$

Example: Brain Waves

In the brain waves example, we computed 95% CI = $(-1.28, -0.32)$, which did not contain the null value of zero, so we rejected $H_0: \mu_1 - \mu_2 = 0$.

The point estimate for this unknown parameter is $\bar{x}_1 - \bar{x}_2 = -0.8$

The standard error of the point estimate is $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0.24$

So t =

iClicker: t-test statistic

Let's back the previous example with cement drying. Consider:

$$H_0: \mu = 15; \quad \bar{x} = 16.7; \quad s = 5.6; \quad n = 25$$

Standard error =

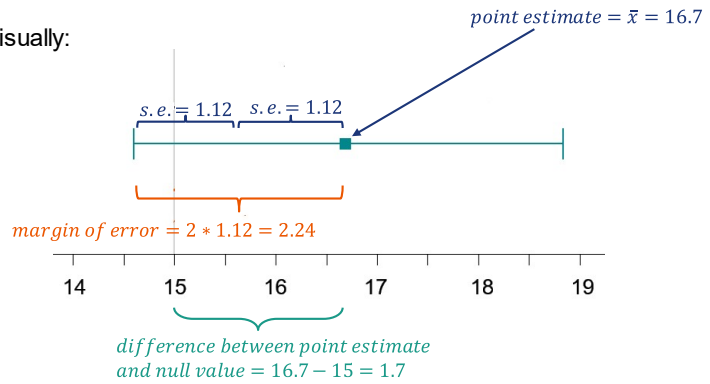
How many standard errors are between the null value and the point estimate?
AKA calculate the t-statistic

$$t = \frac{point\ estimate - null\ value}{standard\ error}$$

A. -1.5
B. 0.3
C. 1.5
D. 7.5
E. I don't know

From the 95% CI to the t-test statistic

- Visually:



From the 95% CI to the t-test statistic

If the point estimate is more than 2 standard errors from the null value, $reject\ H_0$.

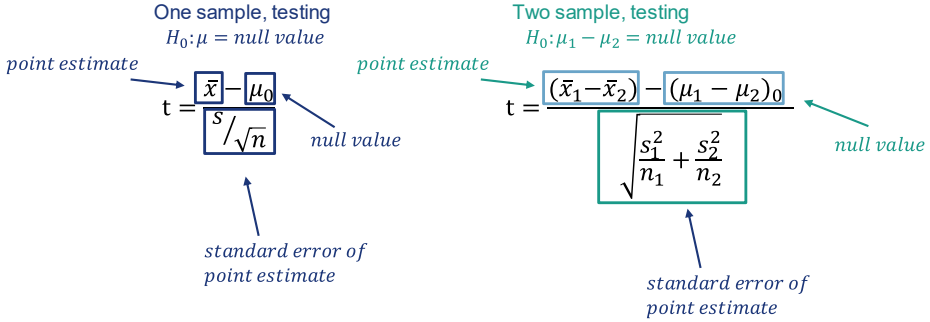
If $|t| > 2$: $Reject\ H_0$

If the point estimate is fewer than 2 standard errors from the null value, $FTR\ H_0$.

If $|t| < 2$: $Fail\ to\ Reject\ H_0$

The "t" test statistic

- The specific formulas for one sample and two sample problems:



Brain waves example

- Back to the brain waves example, use the JMP output obtained via "Analyze / Distribution" to calculate the test statistic:

Distributions Confinement=confined

Waves

Summary Statistics

Mean	9.78
Std Dev	0.5977736
Std Err Mean	0.1890326
Upper 95% Mean	10.207622
Lower 95% Mean	9.3523785
N	10

Distributions Confinement=cell

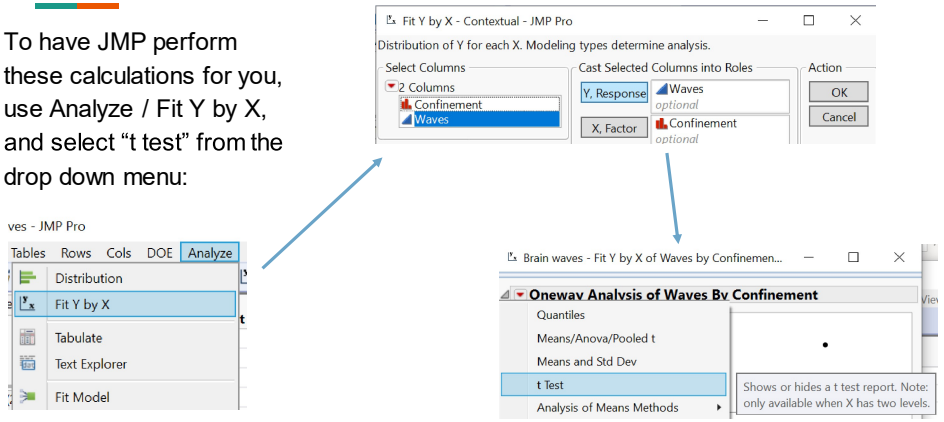
Waves

Summary Statistics

Mean	10.58
Std Dev	0.4589844
Std Err Mean	0.1451436
Upper 95% Mean	10.908338
Lower 95% Mean	10.251662
N	10

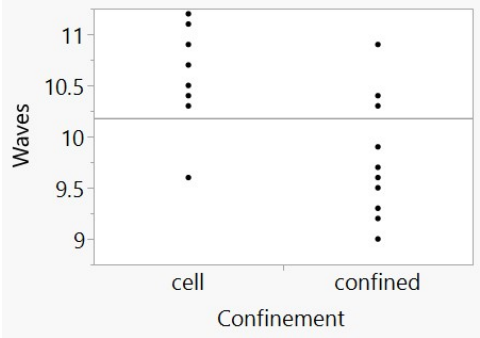
t-test in JMP: Brain Waves

- To have JMP perform these calculations for you, use Analyze / Fit Y by X, and select “t test” from the drop down menu:



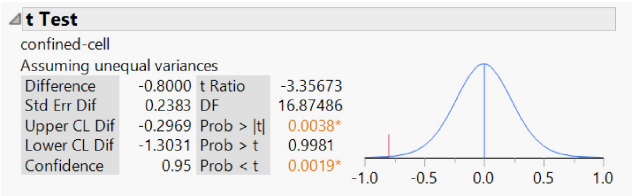
Visualizing

- JMP gives a nice visualization of the data when performing a t-test.
- Here we notice that brain waves seem higher for those in their regular cell and lower for those in confinement.



Brain waves example

- JMP calls the t-test statistic the “t-ratio”
- Verify that you got the same test statistic as JMP



- Using the t-test statistic, what is the statistical decision?

Brain waves example - summarized

We can summarize this in the 3-step hypothesis testing framework.

Steps 1. and 3. are the same as before. But this time, the statistic we are using is the t-test statistic

1. State the null hypothesis:
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision

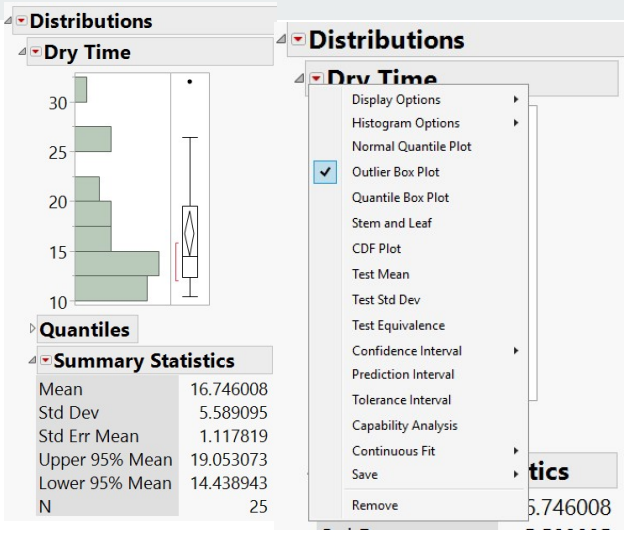
iClicker: Interpret

How do we interpret these results?

A. There is a difference between brain activity for prisoners in confinement vs cell.
B. There is not a difference between brain activity for prisoners in confinement vs cell.
C. There is not enough evidence to show that there is a difference between brain activity for prisoners in confinement vs cell.
D. I don't know.

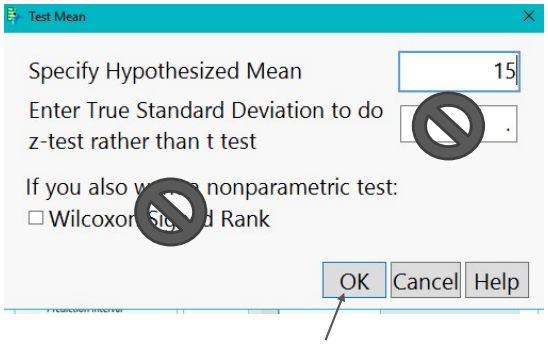
t-test in JMP: Brain

We can also use JMP to perform a t-test for the mean of one sample. Use Analyze / Distribution to find the summary statistics as normal. Then use the red drop down arrow next to the variable name and choose Test Mean.



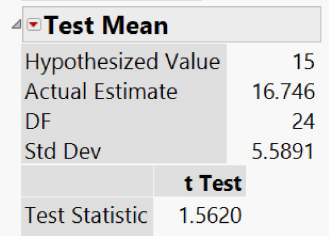
t-test in JMP: Brain Waves

Enter the value from the null hypothesis and then ignore everything else and say ok.



Example: Cement Dry Time

Outline the hypothesis test using the JMP output:



Section 6.4: The p-value

The p-value

- We've seen how to conduct a hypothesis test using a 95% CI, or by using a t-test statistic. Both methods will lead to the same result!
- Another equivalent way of conducting a hypothesis is by using a p-value.
- The p-value is defined as:

> *The probability that, from a random set of data, you would calculate a test statistic that is at least as large as the one you calculated, if the null hypothesis were true.*

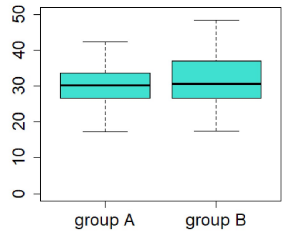
The p-value

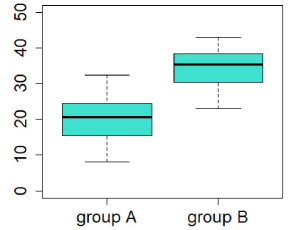
- Smaller p-values indicate that data like yours would be less likely to occur by chance if the null hypothesis were true.
- So if the p-value is small, you will reject the null hypothesis.

- Larger p-values indicate that data like yours would be more likely to occur by chance if the null hypothesis were true.
- If the p-value is large, you will fail to reject the null hypothesis.

The p-value logic, visually

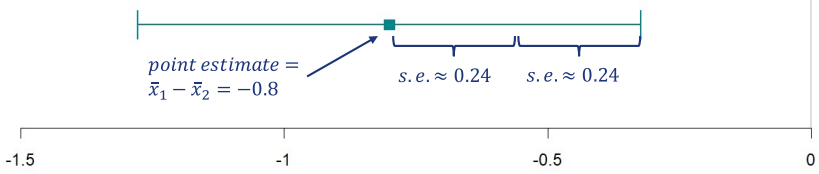
- The plot on top shows data that looks like it would be likely to occur by chance, if the population means for groups A and B were equal *(n = 50 per group)*.
- The plot on the right shows data that looks like it would not be likely to occur by chance, if the population means for groups A and B were equal.
- The p-value for the top plot is 0.304.
- The p-value for the bottom plot is 0.000000000002





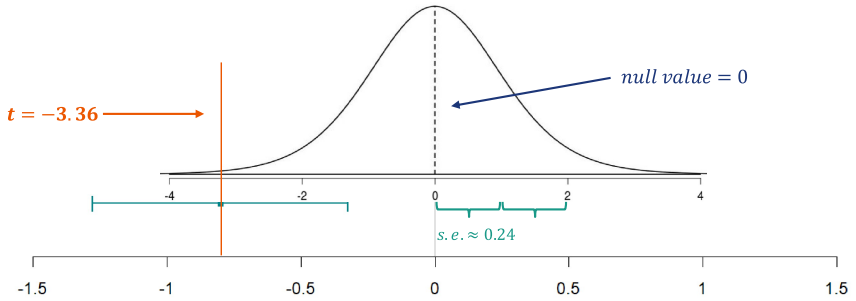
The p-value, visually

- Recall the diagram from the brain waves example of a confidence interval not containing zero, and thus rejecting $H_0: \mu_1 - \mu_2 = 0$.
- Also recall the test statistic: $t = \frac{point\ estimate - null\ value}{standard\ error} = \frac{-0.8-0}{0.2383} = -3.6$



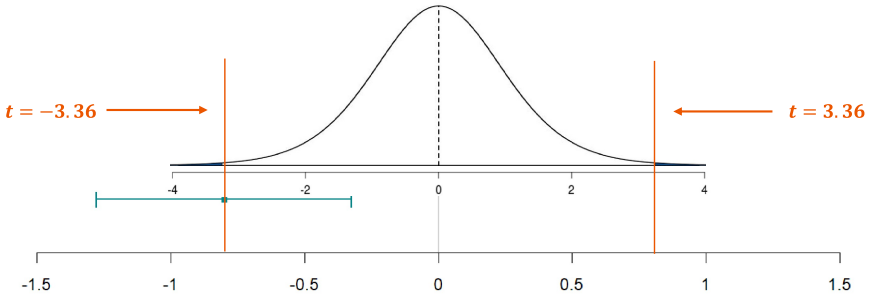
The p-value, visually

The p-value is computed under the assumption that H_0 is true. If the null is true and you repeat the experiment over and over again, most of the time you should get a t statistic close to zero. Every once in a while you would get a weird sample that has a large or small t statistic. Here is a t-distribution, centered on the null value of zero.



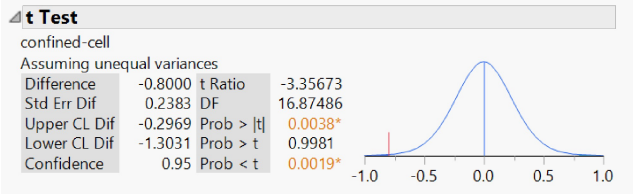
The p-value, visually

- Here, the p-value is the combined small blue shaded area to the left of $t = -3.36$ and to the right of $t = 3.36$. This represents the probability of getting a test statistic at least as far from zero as $t = -3.36$, if H_0 were true.



The p-value, in JMP output

- Here again is the JMP output for the brain waves example:



- Notice how similar the diagram looks to the one on the previous slides.
- JMP reports the p-value as "Prob > $|t|$". The p-value here is 0.0038.

Interpreting the p-value

- The p-value of 0.0038 is the probability we would get a test statistic at least as large as the one we got ($t = -3.36$), if the null hypothesis of no difference in population means were true.
- This is very small. It tells us that the data we got is the kind of data that would be very unlikely to happen just by chance, if the null hypothesis was true.
- So, we reject H_0.
- This is the same conclusion we came to when we saw that the 95% CI did not contain zero, and that the absolute value of the test statistic was greater than 2.

Making the statistical decision using the p-value

- The rule for deciding whether to $reject\ H_0$ or $FTR\ H_0$ using the p-value is:
  - If p-value < 0.05, $reject\ H_0$
  - If p-value > 0.05, $FTR\ H_0$
- 0.05 is referred to as the "level of significance". It doesn't have to be 0.05, but this is by far the most popular value to use, and it is the only one we will use in STAT 201.
- The level of significance of 0.05 corresponds to a 95% confidence level.
- Example: if we made a 99% CI, the level of significance would be 0.01.

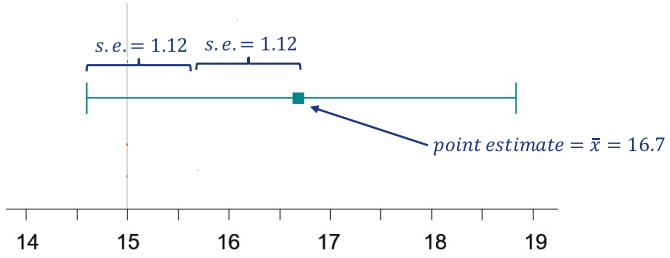
Example: Cement

- Here is the other confidence interval diagram for cement drying we saw earlier.

$H_0: \mu = 15; \quad \bar{x} = 16.7; \quad s = 5.6; \quad n = 25$

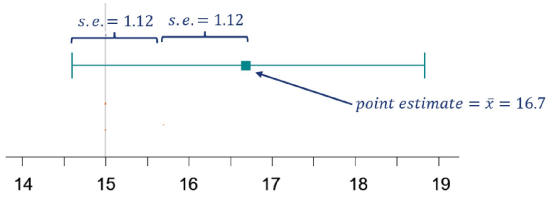
CI: (14.46,18.94)

- The 95% CI contains the null value of 15, so we $FTR\ H_0$.



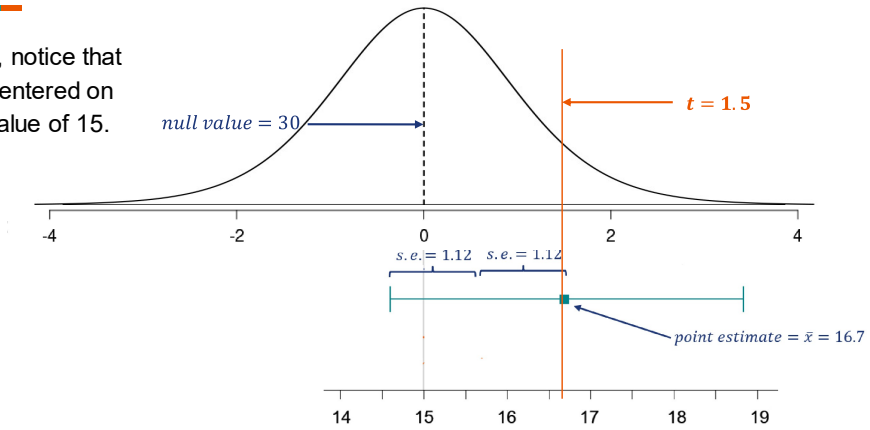
Example: Cement

- Also, the test statistic is $t = \frac{point\ estimate - null\ value}{standard\ error} = \frac{16.7-15}{2.24} = 1.5$
- This is smaller than 2, which also tells us that we $FTR\ H_0$
- Because we $FTR\ H_0$, we should expect the p-value to be larger than 0.05.



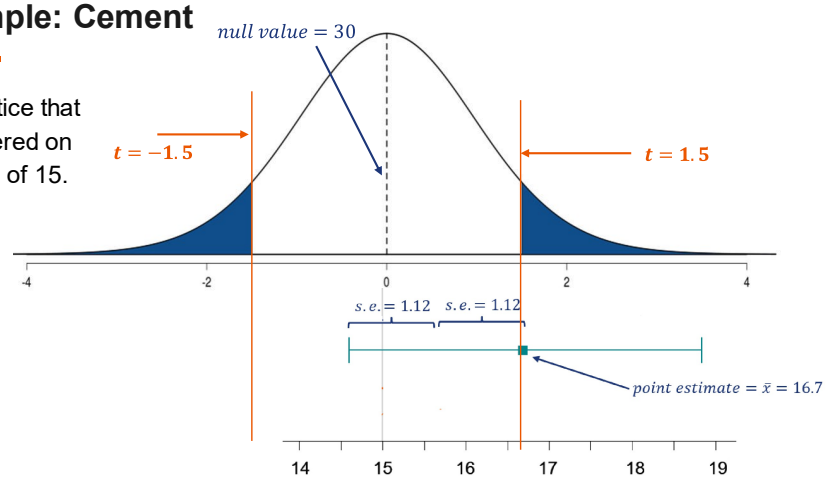
Example: Cement

This time, notice that $t = 0$ is centered on the null value of 15.



Example: Cement

This time, notice that $t = 0$ is centered on the null value of 15.



Example: Cement Dry Time

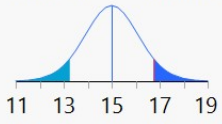
Using the same JMP output from the test we did earlier using the t-statistic we can see the JMP output gives us the same picture for the p-value.

Outline the hypothesis test using the JMP output:

Test Mean

Hypothesized Value	15
Actual Estimate	16.746
DF	24
Std Dev	5.5891

	t Test
Test Statistic	1.5620
Prob > |t|	0.1314
Prob > t	0.0657
Prob < t	0.9343



iClicker: Practice example

Suppose we are testing $H_0: \mu_1 - \mu_2 = 0$. Given these JMP results, state whether you should $reject\ H_0$ or $FTR\ H_0$. Justify your decision using the 95% CI, the test statistic, and the p-value.

t Test

B-A

Assuming unequal variances

Difference	7.82762	t Ratio	9.198089
Std Err Dif	0.85100	DF	117.6476
Upper CL Dif	9.51289	Prob > |t|	<.0001*
Lower CL Dif	6.14234	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

Statistical decision:	Justification using 95% CI:	Justification using test statistic:	Justification using p-value:
a. Reject H_0 b. Fail to Reject H_0			

Section 6.5: The t-distribution

What is the "t distribution"?

- We've now seen "t" statistics and pictures of the "t distribution". But we haven't seen where this distribution comes from.
- The t-distribution is very similar to the z-distribution, but used when the population standard deviation is unknown and the sample standard deviation is used instead.

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

σ is the population standard deviation

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

s is the sample standard deviation

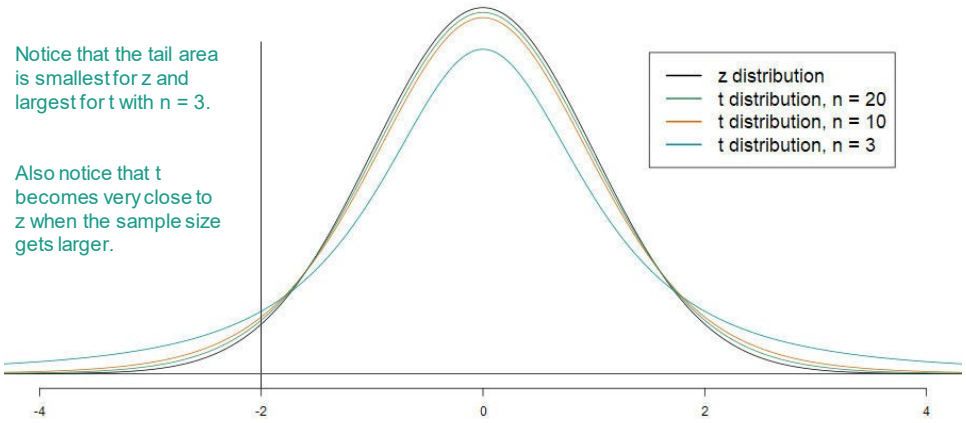
What is the "t distribution"?

- So, the t-statistic has extra sampling variability built in because it uses sample standard deviation, which changes every time you take a new sample. Whereas the z-statistic uses the population standard deviation which does not change.
- This means that the t-distribution will be more spread out than the z-distribution.
- How much more spread out depends on the sample size. When n is smaller, there is more sampling variability in s, and so the t-distribution is more spread out.
- As n gets larger, the t-distribution becomes more and more similar to the z-distribution. Technically, we say that t "converges" to z as n increases.

What is the "t distribution"?

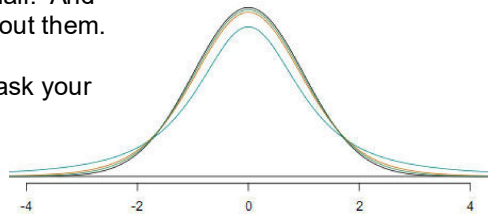
Notice that the tail area is smallest for z and largest for t with n = 3.

Also notice that t becomes very close to z when the sample size gets larger.



What is the "t distribution"?

- Because the t-distribution slightly changes shape when sample size change, our p-values and margins of error technically also change when sample size changes.
- However, for any reasonably large sample size (say, $n > 20$), these changes are small. And so, in STAT 201, we will not worry about them.
- If you want to know more about this, ask your instructor or a tutor in the SSC!



Origin of the t-distribution

- As a side note, the t-distribution is sometimes called *"Student's t-distribution"*.
- The statistician who first discovered (or created, depending on your perspective) this distribution was William Gosset, who worked for Guinness Brewing. His work involved drawing inference from small samples – Guinness wasn't going to let him brew 100 batches of beer to run an experiment.
- At the time, the consensus was that inferential statistics should never be done using small samples. So, Gosset mathematically derived the t-distribution.

(Photo: public domain)

Origin of the t-distribution

- Guinness had a policy of not allowing its employees to publish work they had done for the company.
- Gosset knew that his discovery would have broad scientific applicability, as it would allow researchers to do statistical inference using small sample sizes.
- He felt his work should be published, and Guinness allowed him to publish it under a pseudonym; he chose "Student".

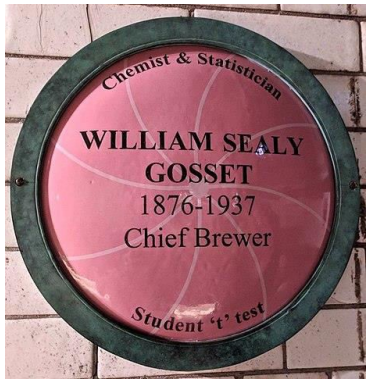


Image: Tim Bates, Creative Commons license

Section 6.6: More examples

Example: the logic of hypothesis testing

- There is a classic example of hypothesis testing, sometimes known as “the lady tasting tea”. The example is meant to elucidate the overall reasoning behind hypothesis testing.
- The characters in this example are the statistician Ronald Fisher and the phycologist (a scientist who studies algae) Muriel Bristol

Muriel Bristol

Photo source

Ronald Fisher

Photo: public domain

The lady tasting tea

- In 1935, Ronald Fisher was at a social gathering and offered Muriel Bristol a cup of tea. She refused, saying that she preferred the taste of tea when the milk had been poured in first.
- Fisher was skeptical that Bristol could distinguish whether the milk or tea had been poured in first, just from tasting it.
- So, a taste test was set up. 8 cups of tea were made, 4 of which had the milk poured in first. Bristol’s task was to identify which 4 these were.

The lady tasting tea

- For this test, the null hypothesis is:
- What would constitute a Type I error?
- What would constitute a Type II error?

The lady tasting tea

- Dr. Bristol correctly identified all 4 cups in which the milk had been poured in first.
- Fisher calculated the probabilities of guessing any number of cups correctly.
- So, the p-value for this experiment was 0.014: the probability of guessing all four correctly if the null hypothesis of random guessing were true.

# of cups correct	Probability of guessing randomly
4	0.014
3	0.229
2	0.514
1	0.229
0	0.014

The lady tasting tea

Outlining the 3 steps of a hypothesis test:

1. State the null hypothesis:
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision

iClicker: ACT before vs. after

- Recall the ACT example from module 5. Here are the summary statistics for the differences in ACT scores (after minus before):

Summary Statistics	
Mean	1.8636364
Std Dev	2.0539267
Std Err Mean	0.4378986
Upper 95% Mean	2.7742964
Lower 95% Mean	0.9529763

- Conduct a hypothesis against the null that population mean change in ACT scores is zero.

1. State the null hypothesis:
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision
   a. Reject the null
   b. Fail to reject the null

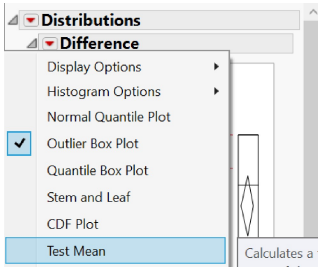
iClicker: Interpret

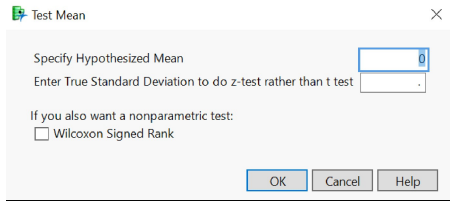
How do we interpret these results?

A. ACT scores do not change when taken a second time.
B. There is not enough evidence to show that ACT scores change when taken a second time.
C. ACT scores change when taken a second time.
D. I don't know.

Example: ACT before vs. after

- Here is what happens when we tell JMP to conduct a test against the hypothesized mean of zero:





Test Mean	
Hypothesized Value	0
Actual Estimate	1.86364
DF	21
Std Dev	2.05393

	t Test
Test Statistic	4.2559
Prob > |t|	0.0004*
Prob > t	0.0002*
Prob < t	0.9998

Example: ACT before vs. after

- Note that all three methods for making the statistical decision here agree:
  - The 95% CI does not contain zero
  - The t-test statistic is larger than 2
  - The p-value is smaller than 0.05

iClicker: Your Opinion

Which method do you prefer for hypothesis testing?
A. Confidence Interval
B. T-statistic
C. P-value
D. I have no opinion

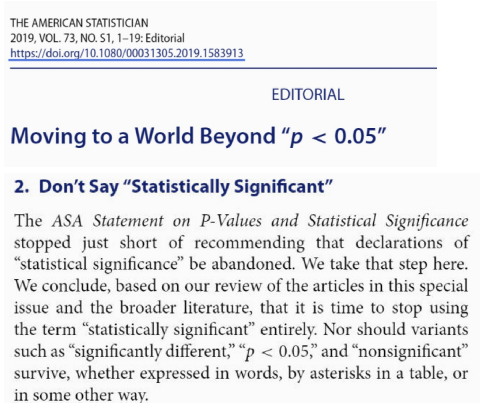
Section 6.7: Criticisms of hypothesis testing and "statistical significance"

"Statistical significance"

- When a hypothesis test is performed and H_0 is rejected, the phrase "statistically significant" is sometimes used to describe the result.
- Example: suppose we are testing to see if a pain relief drug is more effective than a placebo.
  - $H_0: \mu_{treatment} - \mu_{placebo} = 0$
  - 95% CI is $4.6 \pm 2.4 = (2.2, 7.0)$
  - Statistical decision is $reject\ H_0$
- It might be said that the sample means "differ significantly" or "are statistically significantly different".

American Statistical Association editorial

- But – we do not recommend this language.
- Here is a recommendation co-authored by the president of the American Statistical Association, in its flagship journal "The American Statistician".
- It is from a special issue, encouraging researchers to move "to a world beyond $p < 0.05$"



THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

https://www.tandfonline.com/doi/full/10.1080/00031305.2019.1583913

Criticisms of hypothesis testing and p-values

- Hypothesis testing with p-values is a very popular inferential method. It is also very controversial, with many statisticians recommending against it.

  The criticisms are:

  - p-values are commonly misinterpreted
  - p-values are commonly calculated using invalid methods
  - hypothesis testing encourages dichotomous thinking
- We have spent these notes learning hypothesis testing. We will now look at the dangers of hypothesis testing.

First criticism: p-values are commonly misinterpreted

- Here is the definition of the p-value that we saw earlier:

  The probability that you would calculate a test statistic from a random set of data that is at least as large as the one you calculated, if the null hypothesis were true.

- It is tempting, but wrong, to shorten this to:

  The probability that the null hypothesis is true.

This can be seen more clearly by writing the p-value as a conditional probability.

First criticism: p-values are commonly misinterpreted

- $P(data \mid H_0 \; is \; true)$ is read as "the probability of the data, given the null hypothesis is true".
- If we mistakenly interpret the p-value as the "the probability that the null hypothesis is true, given our data", then would write $P(H_0 \; is \; true \mid data)$
- But, $P(data \mid H_0 \; is \; true)$ and $P(H_0 \; is \; true \mid data)$ are not the same thing!
- We cannot simply "flip" the conditional probability statement.

Flipping the conditional probability: don't do it!

- Here is an illustrative example. Suppose you go to the doctor for a checkup and they call you up the next day with bad news.

- You've tested positive for a rare disease.
- How rare? Only 1 in 200 people get the disease.
- Also, the test has a false positive rate of **1 in 100**. Meaning, if 100 people who don't have the disease are tested, only 1 will test positive.

What's the probability that you don't actually have the disease given that you tested positive?

Flipping the conditional probability: don't do it!

- Putting these numbers into a contingency table:

	Disease	No Disease	**Total**
Test Positive	1	2	**3**
Test Negative	0	197	**197**
Total	**1**	**199**	**200**

- Based on this table, we have:

$P($ no disease | test positive $)$ =

$P($ test positive | no disease $)$ =

Flipping the conditional probability: don't do it!

$$P(\,no\ disease \mid test\ positive\,) \neq P(\,test\ positive \mid no\ disease\,)$$

- These two values are not equal. And the different interpretation between these two quantities is important. Using the wrong one could be a huge mistake!

- This same problem can occur if we mistake $P(data \mid H_0\ is\ true)$ and $P(H_0\ is\ true \mid data)$.

"The probability of H_0"

- It is actually not meaningful to calculate $P(H_0\ is\ true)$ when using classical (frequentist) statistical methods.

- In classical statistics, the null hypothesis refers to a fixed truth. And so the null is either true or it is false, but there is no such thing as the probability of the null being true.

Misinterpreting the p-value

The p-value is routinely misinterpreted in one of the previous ways.

- Most popular articles get it wrong.
- Most *scientists* get it wrong.
- Most *textbooks* get it wrong.

"What's wrong with null hypothesis significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!"

- *Jacob Cohen,* The Earth Is Round ($p < 0.05$)

Second criticism: p-values are commonly calculated using invalid methods

- The definition of a p-value, once again:

The probability that you would calculate a test statistic from a random set of data that is at least as large as the one you calculated, if the null hypothesis were true.

- This definition implies that you would have conducted the same hypothesis test if the data set had been different.
- But what if this isn't true? What if the hypothesis test was chosen based on the data?

Second criticism: p-values are commonly calculated using invalid methods

- Suppose you plan to perform a t-test, but you have options as to which response variable to use.
- Example: you suspect that a certain studying style will result in improved academic performance among students. You conduct a randomized experiment in which some students use this new style of studying, and some use any other method that they prefer.
- You aren't quite sure how best to quantify academic performance. So, you collect data on exam grades, overall course grade, and a specialized concept inventory quiz.

Second criticism: p-values are commonly calculated using invalid methods

- Now, you have at least three options for your analysis. You will compare the means of the two groups of students. You could compare their mean exam grades, mean overall class grades, or mean scores on the concept inventory. (You could also combine some of these)
- Why is this a problem for the p-value? The p-value is the probability of obtaining a test statistic at least as large as the one you obtained, if the null hypothesis were true.
- This interpretation assumes that you would have performed the exact same data analysis if you had a different set of data. Would you?

Second criticism: p-values are commonly calculated using invalid methods

- In general, the correct interpretation of a p-value is only valid if the researcher would have performed the same analysis even if the data had been different.
- In our thought experiment, maybe the researcher looks at the data, notices that the two groups differ the most on the concept inventory quiz, and decides to compute a p-value for this variable.
- If the data had been different, a different variable might have been chosen. And so this p-value is not valid.

Third criticism: hypothesis testing encourages dichotomous thinking

- There are only two possible outcomes of a hypothesis test: $reject\ H_0$ or $fail\ to\ reject\ H_0$. This gives the impression that statistical tests determine whether a hypothesis is true or false.
- But things are very rarely this simple. Complicating factors:
  - H_0 may be trivially false.
  - Model assumptions might be unrealistic.
  - Uncertainty might still large.

H_0 may be trivially false

- An excerpt from a research paper titled "How many cents on the dollar? Women and men in product markets":

  "Women had a slightly higher percentage of transactions for which positive feedback had been given in the year preceding the current transaction (99.60% for women and 99.58% for men, P < 0.05)"

  http://advances.sciencemag.org/content/2/2/e1500599.full

- Here, the null hypothesis was that population percentage of transactions with positive feedback for women and men are equal. The sample percentages differed by 0.02 percentage points, and the null hypothesis was rejected.

iClicker – How should this result be interpreted?

"Women had a slightly higher percentage of transactions for which positive feedback had been given in the year preceding the current transaction (99.60% for women and 99.58% for men, P < 0.05)"

A. We've discovered evidence for a meaningful population level difference between positive feedback rates for women and men.

B. We've discovered evidence for some population level difference between positives feedback rates for women and men, but it's too small to care about.

C. We've discovered a difference that's statistically significant, but it's likely that there are also issues with sampling bias or model assumptions being false.

D. We've discovered basically nothing.

E. I don't like any of these choices.

H_0 may be trivially false

- How can such a tiny difference result in rejecting H_0?
- In this case, the sample size was $n = 1{,}106{,}741$.
- When sample sizes are very large, tiny differences will be statistically significant.
- So, the null hypothesis can be false in a way that is trivial. Should we really care about a difference as small as 0.02 percentage points?

H_0 may be trivially false

- We should also ask if H_0 is believable in the first place.
- We recently saw an example where the null hypothesis was that the mean pain reliving effect of a medication was no different than placebo:

$$H_0: \mu_{treatment} - \mu_{placebo} = 0$$

- Does it seem possible that the treatment does absolutely *nothing* compared to placebo? That their difference is *exactly* zero?
- If the null hypothesis doesn't seem plausible in the first place, then we don't get a lot of value in rejecting it.

Model assumptions might be unrealistic

- Consider again the null: $H_0: \mu_1 - \mu_2 = 0$
- This null says that there are two population means that are equal.
- It also assumes that our data came from two populations that are both normally distributed. More formally:

$$X_{i1} \sim Normal(\mu_1, \sigma_1), \qquad i = 1 \ldots n_1$$

$$X_{i2} \sim Normal(\mu_2, \sigma_2), \qquad i = 1 \ldots n_2$$

Model assumptions might be unrealistic

- Most vitally, we are assuming that our data set constitutes a random, unbiased sample from a population.
- This is very hard to do! Examples:
  - *We used a sample of people who were willing to volunteer for a study. Do those who are willing to volunteer differ in an important way from those who don't?*
  - *We collected a sample from a small geographic area. Does this area differ in an important way from other areas?*
  - *Our sample is of data that were easy and convenient to collect. Do easy to collect data differ in an important way from hard to collect data?*

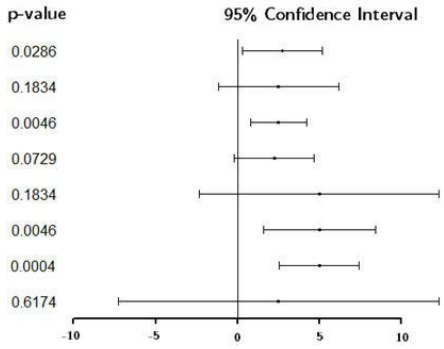
Uncertainty might still be large

- Statistical inference inherently involves uncertainty.
- The fact that we make one of two decisions (reject or fail to reject H_0) does not remove this uncertainty. It is important not to treat these decisions as definitive answers to questions, or as "proof" of anything.
  - Even if we reject H_0, H_0 could still be true.
  - Even if we fail to reject H_0, H_0 could still be false.
- To have an appreciation of statistical uncertainty, **it is best to use confidence intervals rather than p-values.**

Example: CI's vs p-values

Here we have a plot of several hypothesis test results with both a p-value and confidence interval.

- Each test is for $\boldsymbol{H_0: \mu_1 - \mu_2 = 0}$, and each CI is for $\boldsymbol{\mu_1 - \mu_2}$.
- CI's and p-values on the same row come from the same data.
- The vertical line at zero makes it easier to see which CI's contain zero.



The ASA statement on p-values

What do professional statisticians think of p-values?

In 2016, the American Statistical Association put out a statement on the use and misuse of p-values, in which they state 6 principles

ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

The ASA's 6 p-value principles:

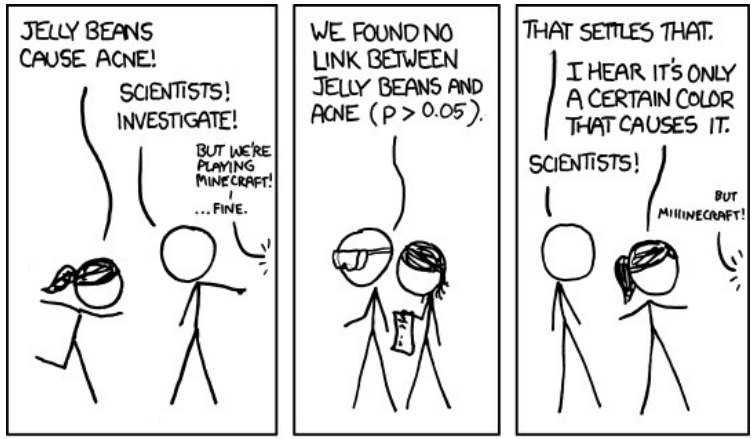
1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Hypothesis testing, p-values, and STAT 201

The ASA statement on p-values begins with this set of questions:

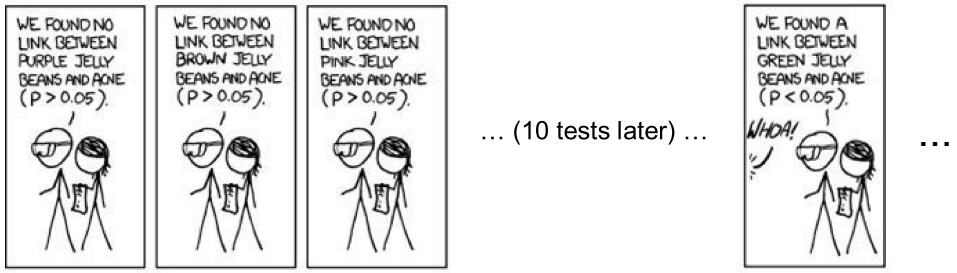
Q: Why do so many college and grad schools teach p = 0.05?
A: Because that's still what the scientific community and journal editors use

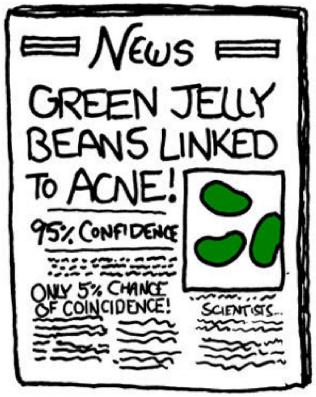
Q: Why do so many people still use p = 0.05?
A: Because that's what they were taught in college or grad school



Hypothesis testing, p-values, and STAT 201

- We are teaching hypothesis testing and p-values because these methods are very popular, and to be an educated consumer of statistics you need to understand them.
- You should know that these methods are controversial, commonly misused, and commonly misunderstood. Proceed with caution.





https://xkcd.com/882/

Module 6 summary

- A hypothesis test is conducted against a null hypothesis (H_0). The two possible outcomes are to $reject\ H_0$ or $fail\ to\ reject\ (FTR)\ H_0$.
- The three ways to decide the outcome of a hypothesis test at the 0.05 level of significance are:
  - If the 95% CI for the unknown parameter does not contain the null value, $reject\ H_0$.
  - If the t-test statistic is greater than 2, $reject\ H_0$
  - If the p-value is less than 0.05, $reject\ H_0$
  - Otherwise, $FTR\ H_0$
- The p-value is a very popular statistic, whose correct interpretation is "the probability of getting a test statistic at least as large as the one we got, assuming the null hypothesis is true"
- When rejecting H_0, some people call the result "statistically significant". This is a controversial phrase; some prominent statisticians have recommended against using it. Hypothesis testing and p-values are controversial overall. They are popular but also frequently misused and misunderstood. In STAT 201 we recommend using confidence intervals rather than hypothesis tests, when possible.

Module 7: Correlation & Simple Linear Regression

7.1 Correlation

Discussion: Causation

Here's a quote by Nate Silver:

"You don't light a patch of the Montana brush on fire when you buy a pint of Haagan-Dazs."

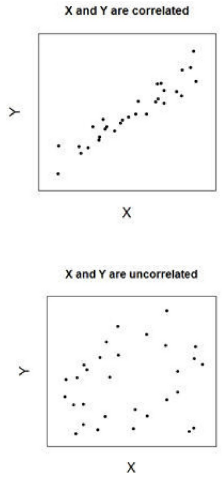
What do you think he means by this?

What is correlation?

- In this module we will cover a very popular statistical procedure called linear regression. In order to motivate this procedure, we will first consider **correlation**.
- If two variables are **correlated**, then they "move together", meaning that as one increases, the other one either tends to increase or tends to decrease.
- If two variables are **uncorrelated**, then there is no such tendency.

What is correlation?

- In the top **scatterplot**, knowing the location of X gives you information about the probable location of Y. On the bottom scatterplot, you could be almost anywhere on the Y axis regardless of where you are on the X axis.
- We could also say that, in the first case, Y gets bigger as X gets bigger. In the second case, Y and X do not follow any noticeable pattern.



What is correlation?

- Note that this is the same idea behind **dependent** and **independent** variables that we saw back when studying categorical variables.
- Typically, we use "correlated" and "uncorrelated" when talking about quantitative variables. "Correlated" means roughly the same thing as "dependent"; "uncorrelated" means roughly the same thing as "independent".
- If Y gets bigger as X gets bigger, then X and Y are **positively correlated**. If Y gets smaller as X gets bigger, then X and Y are **negatively correlated**. If there is no pattern in how Y changes as X changes, then X and Y are **uncorrelated**.

The linear correlation coefficient, "r"

- We quantify the strength of the correlation between X and Y by giving it a value between -1 and 1.
- The lowercase letter "*r*" is used to denote the **linear correlation coefficient**. It is used to show how much linear correlation there is between X and Y.
- Note the word "linear" in "linear correlation coefficient". We use *r* to quantify "straight line" type relationships, but not "curved" relationships. Non-linear correlations will not be well identified by *r*.

Note: There is a formula for this, which we will not cover.

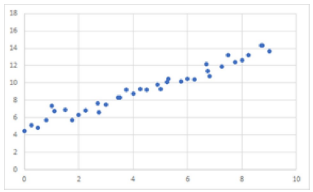
The linear correlation coefficient, "r"

r = 1 means perfect positive correlation
r = -1 means perfect negative correlation
r = 0 means no correlation

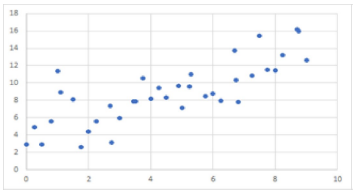
iClicker: Correlation

Which has a stronger linear correlation (r closer to 1 or -1) ?

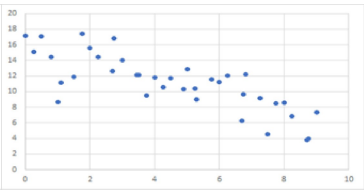
A



B



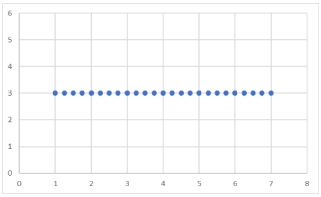
C



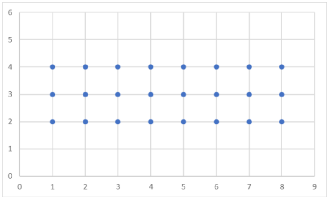
iClicker: Correlation

Which has a stronger linear correlation (r closer to 1 or -1) ?

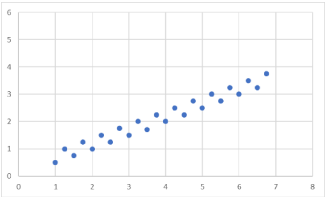
A



B

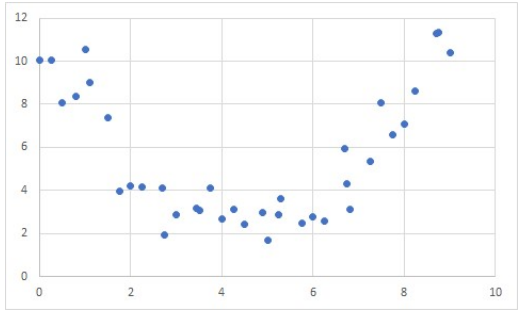


C



Discussion: Non-linear trend

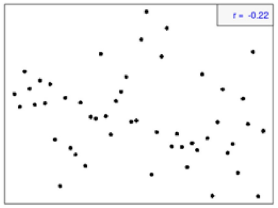
Note that, for this plot, X and Y very clearly move together. Why is the value of *r* near zero?

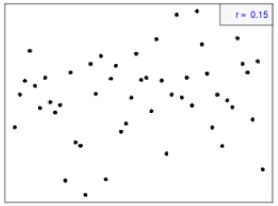


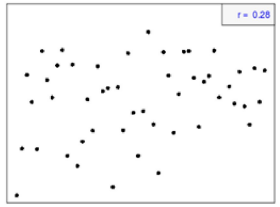
r = .03

Inference for correlation

- Note that when r is near zero, it can be difficult to tell whether the data are correlated by just looking.
- This leads to the question, is the strength of the correlation in our data strong enough that we would not expect to see it by chance alone?







Inference for correlation

We can answer this using a hypothesis test. Let ρ (a Greek "rho") denote the population linear correlation. We use our data to test:

$$H_0 : \rho = 0$$

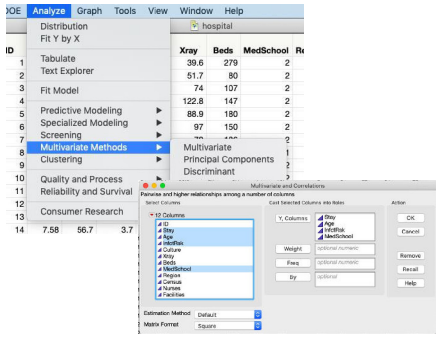
By creating a 95% CI for ρ.

Note: We will not go over the mathematics behind these. JMP will do that for us.

Inference for correlation

In JMP, use Analyze / Multivariate Methods / Multivariate:

- From here, you can choose two or more continuous variables and JMP will calculate correlations, p-values, and 95% CIs.



Note: if a numeric variable is coded as ordinal or nominal, you must change it to continuous to be able to calculate the correlation.

Example: Hospital data

Recall the hospital data set from module 5. In an effort to reduce the risk of infection which can result from hospital stays, researchers collected data on several variables for $n = 113$ hospitals in the United States. Variables included infection risk, age, and length of stay, amongst others. Shown below is a snapshot of some of the data.

Is this an observational or experimental study?

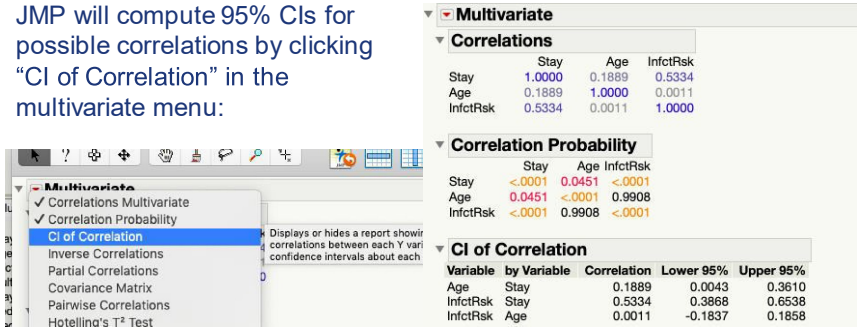
What is the population of interest?

ID	Stay	Age	InfctRsk
91	8.86	51.3	2.9
92	8.93	56	2
93	8.92	53.9	1.3
94	8.15	54.9	5.3
95	9.77	50.2	5.3
96	8.54	56.1	2.5
97	8.66	52.8	3.8
98	12.01	52.8	4.8
99	7.95	51.8	2.3
100	10.15	51.9	6.2
101	9.76	53.2	2.6
102	9.89	45.2	4.3

Example: Hospital data and correlation

In the Hospital data set, we can look for correlations between length of stay, infection risk, and age.

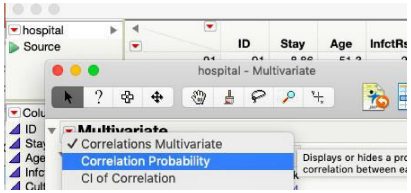
JMP will compute 95% CIs for possible correlations by clicking "CI of Correlation" in the multivariate menu:

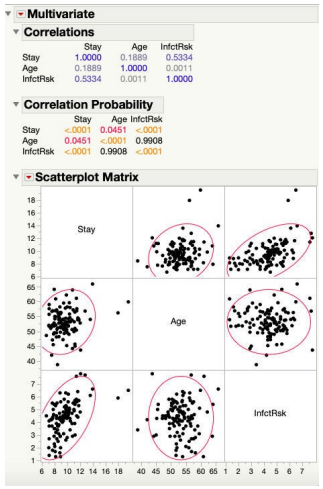


Example: Hospital data and correlation

JMP will also give p-values:

In the multivariate menu, click "Correlation Probability," and JMP will give the output shown to the right. The p-values that test the null hypothesis that $\rho = 0$ are shown for each pair under "Correlation Probability."

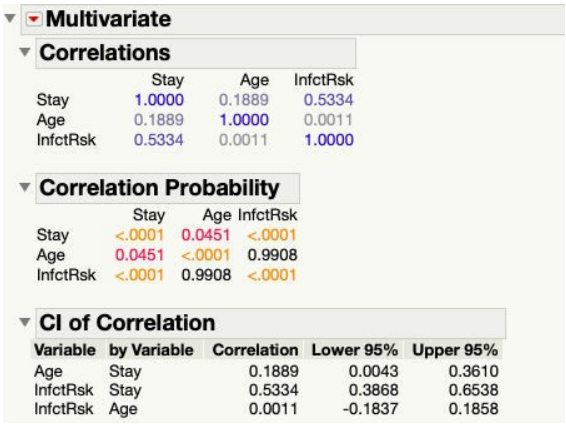




iClicker: Significant correlations

For which pairs would we say that we have evidence that the correlation is not equal to zero?
(*Hint: you should be able to answer this using either a p-value or a CI.*)

A. Infection Risk/Age
B. None of these
C. Infection Risk/Stay
D. Age/Stay



Warning about significant correlations

- If we reject the null hypothesis that $\rho = 0$, we call our correlation "significant".
- Remember that statistically significant results are large enough to be unlikely to occur by chance, assuming that the null hypothesis is true.
- So, a significant sample correlation (r) is one that is larger than we would expect if the population correlation (ρ) was zero.
- Remember that larger sample sizes (n) make it "easier" to reject the null as standard error goes down as n goes up.

Example: Bitter personality

A published scientific paper found a significant correlation between preferences for bitter foods and drinks (e.g. IPA, black coffee) and scores for anti-social personality traits (e.g. psychopathy, narcissism, Machiavellianism, "everyday sadism").

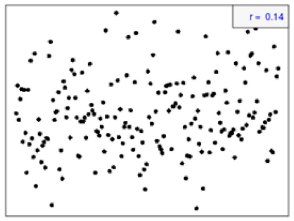
The paper is titled "Individual differences in bitter taste preferences are associated with antisocial personality traits". It can be found at:

doi.org/10.1016/j.appet.2015.09.031

Note: the sample size for this study was n = 500.

Example: Bitter personality

The correlations between mean bitter preference and "psychopathy" and "everyday sadism" are r = 0.14, with p-values less than 0.01. Here is an example of what r = 0.14 looks like:



Remember – "statistically significant" effects need not be large! When it comes to significant correlations, they need not even be detectable to the naked eye.

Example: Bitter personality

Some news coverage…

IF YOU ARE A FAN OF IPA, SCIENCE SAYS YOU'RE MORE LIKELY TO BE A PSYCHO

1 Minute Read
Produced by VinePair Staff / @VinePair
Updated on 2016-03-04

(VinePair: https://vinepair.com/booze-news/if-you-are-a-fan-of-ipa-science-says-youre-more-likely-to-by-psychotic/)

SCIENCE

Love Beer And Coffee? You Might Be A Psychopath

A new study associates taste preferences with personality traits

(Popular Science: https://www.popsci.com/psychopaths-may-prefer-bitter-foods)

News › Science

How you drink your coffee 'could point to psychopathic tendencies'

The study found 'bitter taste preferences were a robust predictor for Machiavellianism, psychopathy, narcissism and everyday sadism'

(Indepdendent: http://www.independent.co.uk/news/science/psychopathic-people-are-more-likely-to-prefer-bitter-foods-according-to-new-study-a6688971.html)

The linear correlation coefficient, "r"

- Here is a very general rule of thumb:

  if $|r| > 0.9$, the correlation is "strong"
  if $0.9 > |r| > 0.6$, the correlation is "moderate"
  if $0.6 > |r|$, the correlation is "weak"

- Often it is of interest to see not just how strong X and Y are correlated, but also how they "move" together. In other words, as X increases, how does Y tend to change?

- To answer this question, we use a technique called **linear regression**, which will be introduced in the next section.

7.2 Simple Linear Regression

Overview: Linear regression

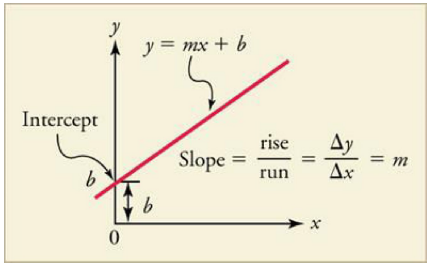
- Linear regression is a popular statistical technique used to quantify the nature and magnitude of relationships between multiple quantitative variables.
- The most basic kind of linear regression involves making a two dimensional scatterplot of data showing X and Y values for each data point, and then drawing a straight line through this scatterplot in order to show the overall trend in the data.
- The line we draw through the data is called a **"regression line"** or **"line of best fit" (LOBF).**

Equation of a line

The first thing we need to do in order to develop this procedure is to review the equation of a line:

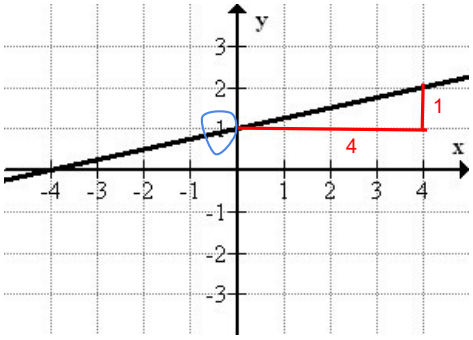
$$Y = mx + b$$

- *Y* is the location on the Y axis
- *X* is the location on the X axis
- *m* is the slope
- *b* is the intercept



Note: if you have numbers for *m* and *b* you can use the formula to find every point on a line: plug in a value for *X*, and the equation tells you the value of Y that will place you on the line.

Equation of a line



The equation for this line is:

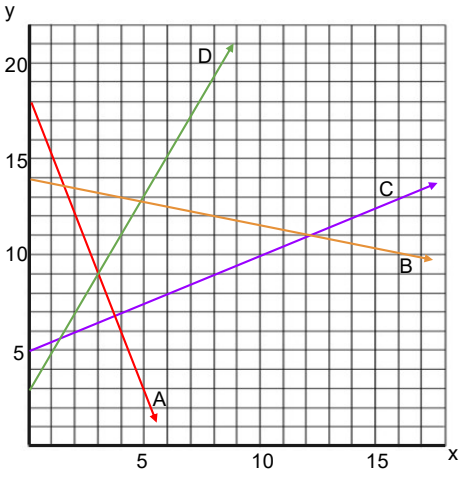
m = slope = $\frac{change\ in\ y}{change\ in\ x}$ =

b = intercept =

If you want to know the value of y when x is a certain number, you can use the equation and plug in your value for x.

When x=20:

Equation of a line



A $y = -3x + 18$

B $y = -\frac{1}{4}x + 14$

C $y = \frac{1}{2}x + 5$

D $y = 2x + 3$

iClicker: equation of a line

What is the equation for this line?

a) y = 9x – 2
b) y = -9x + 2
c) y = 2x + 9
d) y = -2x + 9
e) y = 2x - 9

Linear regression as a model

- As we have seen, our data won't be in a perfectly straight line.
- We need a way of describing the relationship (AKA "model" the relationship) between two variables, X and Y, that is a straight line plus some random variation:

$$y = mx + b + variation$$

- In statistics we usually put the intercept first:

$$y = b + mx + variation$$

Linear regression as a model

Statistics also uses greek letters for parameters (population values), so we'll replace b and m with greek letters, β_0 "beta 0" and β_1 "beta 1"

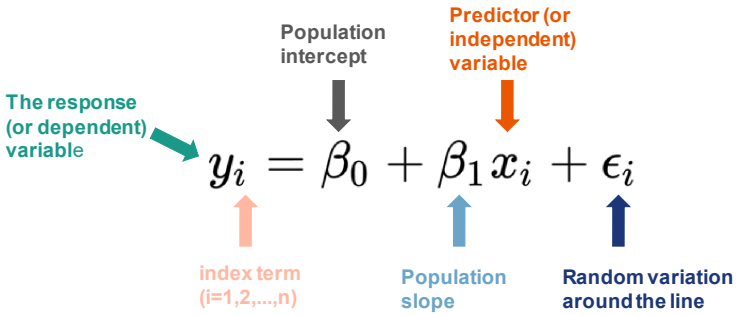
$$y = b + mx + variation$$
$$y = \beta_0 + \beta_1 x + variation$$

- Also, it's tedious to keep writing "variation" over and over again. The common notation is to use ε, the greek letter 'epsilon'

$$y = \beta_0 + \beta_1 x + \varepsilon$$

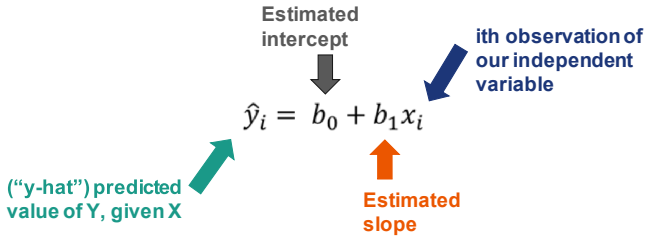
Linear regression as a model

One last thing: Remember that we have multiple observations, which are rows in our data sets, so we need some letter to index which observation we're talking about. Let's call it "i" for "index." That gives us the **theoretical regression model**:



Linear regression as a model

- We use our data to estimate the values of the parameters from our theoretical model. We won't worry about these formulas. We will instead get b_0 and b_1 from JMP.
- Below is the **estimated model**, i.e. the equation of our **line of best fit (LOBF)**. This is the version we will plug numbers into.

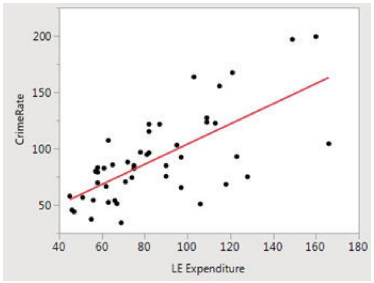


Example: Crime rate and LE expenditure

Here is a plot of crime rate vs. expenditure level on law enforcement.

The red line going through the data is our line of best fit (LOBF). It has a slope and an intercept, which are calculated from the data.

In creating this line, we are modeling the response variable (Crime Rate) and a linear function of the predictor variable (LE Expenditure).



Example: Crime Rate and LE Expenditure

Now, let's apply this to the crime rate vs. LE expenditure example and write:

This assumes:

1. The relationship between crime rate and LE expenditure can be expressed as a straight line, with some population slope and intercept.
2. Observed values of crime rate are determined by going to this line and then adding or subtracting a random number that we represent with ε

Note: We don't really believe that the observed values of crime rate are "created" by observing a value of law enforcement expenditure, putting it into the equation for a line, and then assign a random number from a normal distribution.

iClicker: Regression Model

Let's think back to the brain waves example. Let's say we perform the experiment differently and put prisoners into solitary confinement for different lengths of time. If we wanted to write a regression model to predict alpha waves based on time spent in solitary confinement it would look like:

A. $y = Time + Alpha\ Waves + \varepsilon$

B. $Time = \beta_0 + \beta_1 Alpha\ Waves + \varepsilon$

C. $Alpha\ Waves = \beta_0 + \beta_1 Time + \varepsilon$

D. $y = Time + Alpha\ Waves\ x + \varepsilon$

E. I don't know

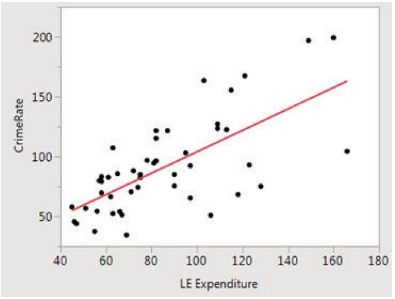
Linear regression as a model: Assumptions

What we saw is only a model. Regression models may be useful for:

- Predicting values of a response variable, using values of a predictor (and assessing the uncertainty in these predictions).
- Quantifying how much a response variable tends to change when a predictor changes (the slope).
- Quantifying how much of the variability in a response variable can be attributed to variability in a predictor variable (this is R^2, which we will consider shortly).

Linear regression as a model: Assumptions

- One assumption we are NOT making: we are not assuming that the predictor variable *causes* changes in the response variable.
- As usual, *correlation does not imply causation*.



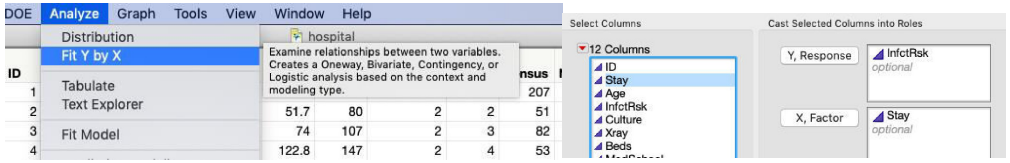
So, if we make crime rate the response variable and LE enforcement the predictor, we are not saying changing the level of expenditure on law enforcement *causes* changes in the crime rate.

It's possible the causality here goes both ways, and possibly many other variables affect both.

Example: Infection risk by length of stay

Let's return to the hospital data. We'd like to use linear regression to determine the nature of the relationship between a patient's length of stay and their risk of infection. The model is:

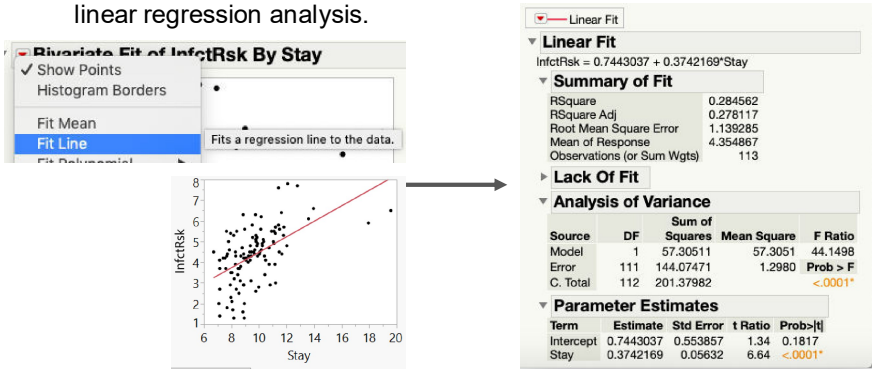
To run a regression model in JMP, use Analyze / Fit Y by X, then select the appropriate predictor and response variables:



Example: Infection risk by length of stay

JMP will produce a scatter plot of the two variables.

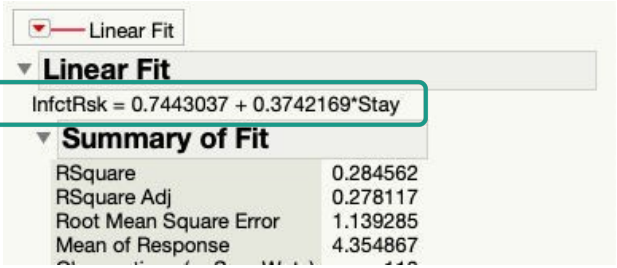
- From the drop down menu (click the red arrow), you can select "fit line" to add a LOBF to the plot, and to obtain detailed results of the linear regression analysis.



Example: Infection risk by length of stay

The very top line under "linear fit" gives the estimated regression equation:

$$InfctRsk = 0.744 + 0.374 * Stay$$



The rest of the output refers to topics we will cover later in these notes and the next set of notes.

Example: Infection risk by length of stay

We can use the equation for the estimated LOBF to predict infection risk from length of stay.

For example, to predict infection risk for a patient with a length of stay of 10 days, use the formula and plug in 10 for Stay:

Note:

- The LOBF contains all of the predicted values we could make for Y (in this case infection risk) using X (in this case length of stay).

iClicker: Prediction

$$InfctRsk = 0.744 + 0.374 * Stay$$

What is the predicted infection risk for a person with a length of stay of 0 days?

A. 0.370
B. 0.744
C. 0.374
D. 1.118
E. I don't know

iClicker

$$InfctRsk = 0.744 + 0.374 * Stay$$

What is the predicted infection risk for a person with a length of stay of 0 days?

A. 0.370
B. 0.744
C. 0.374
D. 1.118
E. I don't know

What do we call this value?

A. b_0
B. b_1
C. y_i
D. x_i
E. I don't know

Interpreting b_0

- The intercept, b_0, is the predicted value of Y when X=0.
- This may make mathematical sense, and it is certainly necessary for computing predicted values, but by itself it does not make real world sense.
- The intercept is rarely of scientific interest. Unless we have a special reason for being interested in the value of Y when X=0, we should not treat b_0 as a statistic worth interpreting.

Interpreting b_1

- b_1, the slope, is the change in Y for a one unit increase in X.
- This is often of great interest. So, unlike b_0, we should ensure that we can correctly interpret b_1.
- In the previous example, we had $b_1 = 0.374$. This means that as length of stay increases by one day, our prediction for infection risk increases by 0.374.

Notes:

- This does not mean a one unit increase in high school GPA "results in" a one unit increase in college GPA.
- This also does not mean changing X causes a change in Y. Correlation does not imply causation!

Example: Vocabulary and height

Suppose we have a regression model where the response variable is a child's vocabulary (in number of words) and the predictor is height (in inches). Our estimated regression model is given by:

$$vocabulary\ size = -14788 + 500 * height$$

How should we interpret the slope?

Example: Vocabulary and height

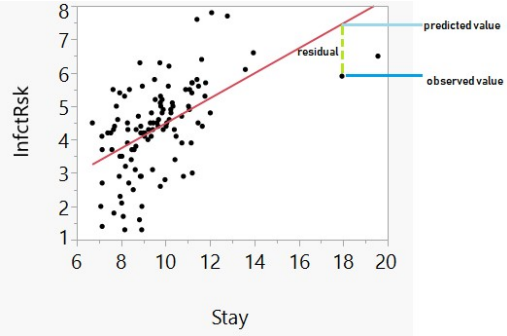
Here are some interpretations to avoid:

- Increasing height by one inch leads to a 500 word increase in a child's vocabulary.
- The impact of height on vocabulary is 500 words per inch.
- If you want children to know more words, just make them taller!

What makes a regression line best? Residuals

A residual (or "error") is the difference between an actual observed value of Y and the predicted value of Y at an observed value of X:

$$residual_i = y_i - \hat{y}_i = observed\ value - predicted\ value$$



Visually, this is the vertical distance between a data point and the LOBF.

Example: Infection risk by length of stay

The person representing observation 93 in our hospital study had a length of stay of 8.92 and an observed infection risk of 1.3.

ID	Stay	Age	InfctRsk
91	8.86	51.3	2.9
92	8.93	56	2
93	8.92	53.9	1.3
94	8.15	54.9	5.3
95	9.77	50.2	5.3

What is the residual for this patient?

1. Locate and write down the observed value:
2. Calculate the predicted value using the LOBF:
3. Subtract the predicted value from the observed value:

iClicker: Prediction

Now, you try! The person representing observation 94 in our hospital study had a length of stay of 8.15 and an observed infection risk of 5.3.

ID	Stay	Age	InfctRsk
91	8.86	51.3	2.9
92	8.93	56	2
93	8.92	53.9	1.3
94	8.15	54.9	5.3
95	9.77	50.2	5.3

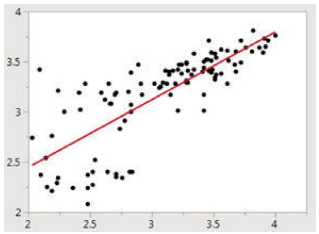
What is the residual for this patient?

A. 1.5079
B. 2.5738
C. 2.7262
D. 3.7921
E. I don't know

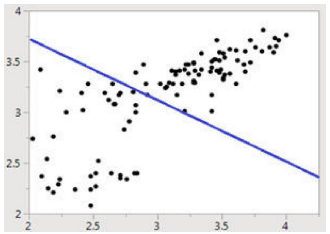
What makes a regression line best? SSE

- The LOBF has the very important property that it minimizes the sum of the squared residuals (SSE). This is what makes it "best".
- If we took any line other than the LOBF and calculated its SSE, this would be bigger than the SSE we get from the LOBF.

Best fitting line:



Bad fitting line:



R Squared

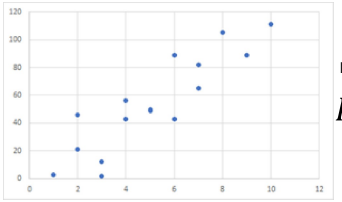
R-squared is the proportion of variation in our response variable that is explained by variation in our predictor variable.

In simple linear regression (where we only have one predictor), R-squared is just taking "r" (the linear correlation coefficient) and squaring it (r^2).

Note that since r is a number between -1 and 1, R-squared must be between 0 and 1.
Also, note that: $|r| \geq R^2$.

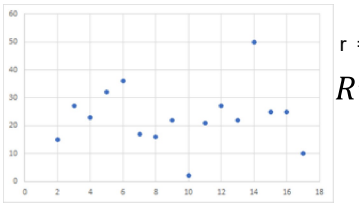
R Squared

Visually, here is a comparison of large vs. small R^2:



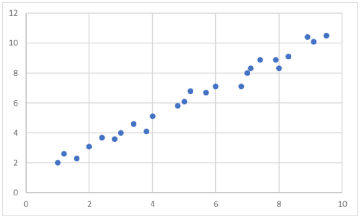
r = .887

$R^2 = .787$



r = -.363

$R^2 = .132$

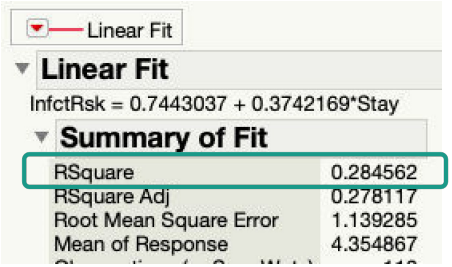


r = .996

$R^2 = .992$

R Squared in JMP

- In JMP we can find R^2 by looking at the regression output at Rsquared.



Here it is .284562.

If we wanted to know the linear correlation coefficient, r, we could take the square root

$r =$

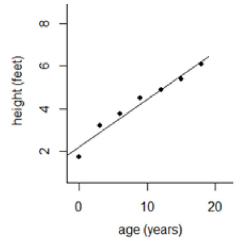
iClicker: Prediction

Suppose we collect data on the age (in years) and height (in feet) of different people and find an estimated line of best to be:

$$height = 2.21 + 0.2232 * age$$

Predict the height when age is 40 years.

A. -8.6 ft
B. 3.1028 ft
C. 9.0632 ft
D. 11.138 ft
E. This is ridiculous



Extrapolation

- In linear regression, we define **extrapolation** as making predictions for Y at values of X that are beyond the range of our data. This is dangerous! By definition, we have no data regarding the relationship between X and Y for values outside the range of our data.

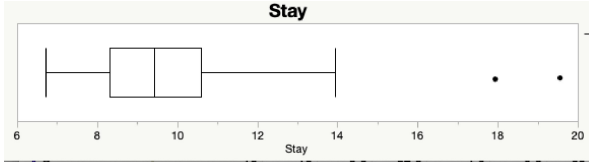
- It could be the case that X values beyond the range of our data don't make sense, or it could be the case that the relationship that we observe between X and Y doesn't hold outside of the range of our data

Extrapolation

- One common example, which we have seen briefly, comes from interpreting the estimated intercept b_0.

- Recall that b_0 is defined as the predicted value of Y when X=0. But what if X=0 is outside the range of our data?

- For example, in our hospital data, no one has a length of stay of zero. Looking at a boxplot for *length of stay* shows that in fact no data fall below 6, so we can't even predict for any value below that:



7.3 Inference in Regression

Regression and Inference

Linear regression models are used for two broad purposes:

1. To estimate how much the response variable tends to change when the predictor variables changes.This is the slope of the LOBF.

2. To make predictions for what the value of the response variable will be, given some value of the predictor variable.

In both cases, if we wish to generalize our estimates or predictions to the overall population, we will have to use an inferential procedure.

Regression and Inference

- Most of the content in these notes will be analogous to what we've already seen regarding inference for comparing means.
- The foundation of our inferential methods will still be the sampling distribution.
    - This time, rather than referring to distributions of sample means, we will refer to distributions of sample regression lines.

The Sampling Distribution of the LOBF

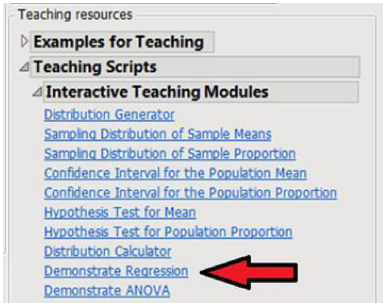
Recall…

- When we first constructed sampling distributions of means, by rolling dice repeatedly and recording the means of each roll.
- The idea was that our statistic (the sample mean) will take on different values when we take new random samples. We are interested in the distribution of these values.
- Of particular interest was the fact that, as the size of each individual sample increased, the sampling distribution of the mean became less spread out. Larger samples led to sample means that tended to be closer to the population mean.

The Sampling Distribution of the LOBF

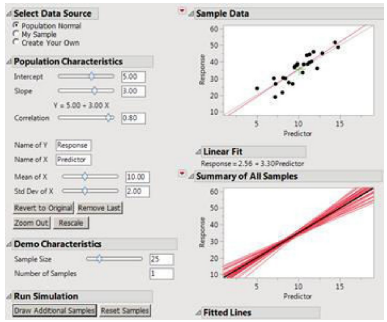
- This exact reasoning can be applied to the line of best fit (LOBF). Imagine there is a population LOBF, and that we sample repeatedly from this population.
- Each new sample will produce a new sample LOBF. This means we will have a sampling distribution of lines of best fit.
- When sample sizes are small, the sampling distribution of the LOBF will be very "spread out" around the population LOBF.
- When sample sizes are large, each sample LOBF will tend to be closer to the population LOBF, and so the sampling distribution of the LOBF will be less spread out.

The Sampling Distribution of the LOBF



- JMP has a built in simulation that shows this.
- Select Help / Sample Data.
- Under "Teaching resources", select Teaching Scripts / Interactive Teaching Modules / Demonstrate Regression.

The Sampling Distribution of the LOBF



- This simulator displays individual samples in the top window, and the simulated sampling distribution of the LOBF in the bottom window.
- You can manipulate the population LOBF with the controls on the left. By far the most important ones are the "correlation" and "sample size" sliders.

I recommend selecting "Fit LS Line" and "Show True Line" from the drop down menu next to the sample data window. This way you can see that the "true" (i.e. population) line stays constant, while the fit line changes for each new sample.

iClicker: Sampling Distribution of LOBF

What should happen as the true correlation gets smaller?

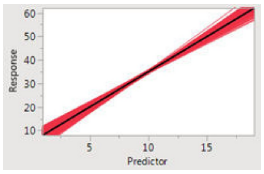
A. The lines of best fit should get more spread out
B. The lines of best fit should get less spread out
C. The lines of best fit will not change

What should happen as n gets smaller?

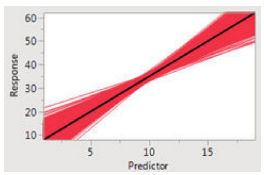
A. The lines of best fit should get more spread out
B. The lines of best fit should get less spread out
C. The lines of best fit will not change

The Sampling Distribution of the LOBF

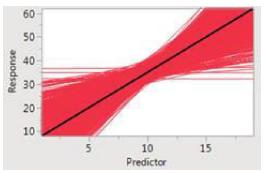
$r = 0.9$ $n = 100$



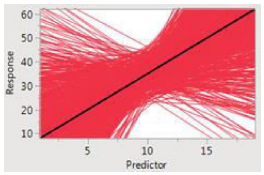
$r = 0.9$ $n = 20$



$r = 0.3$ $n = 100$

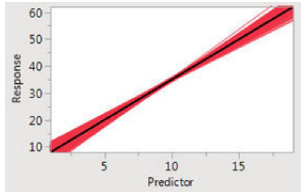


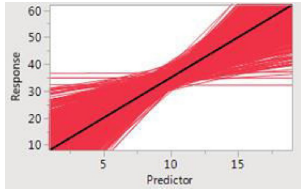
$r = 0.3$ $n = 20$



The Sampling Distribution of the LOBF

- Every time you collect data and calculate a LOBF, it is as though you are sampling one line from a distribution like these.
- We hope to get a sample line that is close to the population line! Our chances of getting a sample line close to the population line get better when the sample size is large, and / or the correlation is strong.
- This is analogous to using a sample mean to estimate a population mean: our estimates are better when the sample size is larger and/ or the data are less spread out.

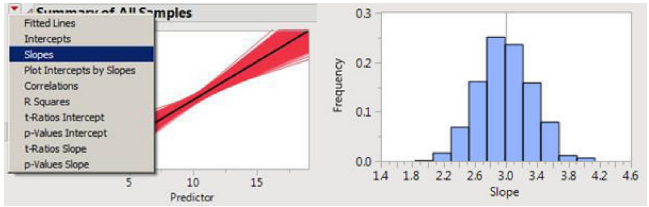




The Sampling Distribution of the LOBF

The simulator will also display the sampling distribution of the slope of the LOBF (which we have denoted b_1).

From the menu by "Summary of All Samples", select "Slopes".



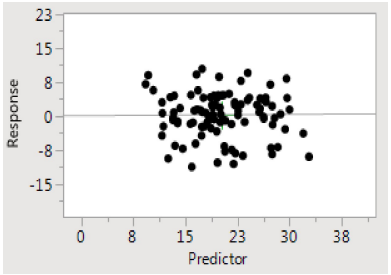
It turns out that the sampling distribution of the slope is normal! This means that we can use all the same inferential methods for slopes that we did for means.

Inference on b_1: Standard error

- The sampling distribution of b_1 will be normal as long as the residuals are normal, or if *n* is large enough (*think Central Limit Theorem*).
- So, we can conduct t-based inferential procedures:
  confidence intervals for β_1
  hypothesis tests on β_1
- Both techniques use the standard error (s.e.) of b_1. JMP will compute this standard error for us.

Inference on b_1: Hypothesis Test

- We can perform a hypothesis test to see whether we have sufficient evidence to conclude that the slope is not zero. This is of interest because, if the slope is zero, then there is no tendency for Y to change as X changes.



- Here, X and Y are uncorrelated: for all X values, the predicted Y value is the same.
- This is represented by a horizontal LOBF (i.e. b_1=0)

Inference on b_1: Hypothesis Test

- Formally, the null hypothesis is:
- And the test statistic is:
- Although test statistics and p-values are often used to test the null that the slope is zero, as we've seen we will primarily test the null using an approximate 95% CI.

Inference on b_1: Confidence Interval

To make a confidence interval, we will use the usual format:

$$CI = estimate \pm (critical\ value)(s.e.\ of\ estimate)$$

Using the approximate 95% t-critical value of 2, we have:

Note: As usual, we treat this interval as one created in such a way that, under repeated sampling, we will successfully capture β_1 95% of the time.

Example: Infection risk and inference with JMP

To produce regression output using JMP, load the hospital data, select Analyze/Fit Y by X, select Infection risk as the Y variable and Stay as the X variable, and then select "Fit Line" from the drop down menu.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7443037	0.553857	1.34	0.1817
Stay	0.3742169	0.05632	6.64	<.0001*

The first term is the intercept. The second term is the slope for length of stay. For each, JMP gives the estimate, standard error, test statistic, and p-value. So, we have:

$b_0 = 0.744$ $\quad se_{(b_0)} = 0.554$ $\quad t = \frac{b_0}{se_{(b_0)}} = 1.34$

$b_1 = 0.374$ $\quad se_{(b_1)} = 0.056$ $\quad t = \frac{b_1}{se_{(b_1)}} = 6.64$

iClicker: Infection risk and inference with JMP

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7443037	0.553857	1.34	0.1817
Stay	0.3742169	0.05632	6.64	<.0001*

Make an approximate 95% CI for β_1:

A. (-.364,1.852)
B. (.262,.486)
C. (.318,.43)
D. (.19,1.298)
E. I don't know

iClicker: Infection risk and inference with JMP

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7443037	0.553857	1.34	0.1817
Stay	0.3742169	0.05632	6.64	<.0001*

If we want to test $H_0: \beta_1 = 0$ using the CI, t-statistic, or p-value we should conclude:

A. Reject H_0, There is enough evidence to say the slope is not equal to zero
B. Reject H_0, There is enough evidence to say the slope is not equal to .375
C. Reject H_0, There is not enough evidence to say the slope is not equal to zero
D. Fail to Reject H_0, There is not enough evidence to say the slope is not equal to zero
E. I don't know

Inference on b_1: Hypothesis Test

If we reject the null, then we say that the observed slope is large enough that it would be unlikely to occur by chance, if the population slope were zero.

Note: Often, if the null is rejected, the slope is said to differ "significantly" from zero. The usual caveat regarding significance applies: we should not confuse a "significant" slope with a "large" slope. If we want to know how large the population slope could reasonably be, we should look at the confidence interval for the slope.

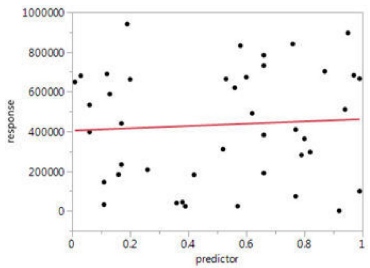
Extreme ranges for X and Y

- Whether an estimated slope is "big enough" to be considered "significant" isn't based on its value alone. It is based on how big its value is, relative to its standard error.
- This is because the significance of the slope is determined by the p-value, which is determined by:

$$t = \frac{b_1}{se_{(b_1)}}$$

- The next slide shows a simulated case where the predictor variable takes on very small values, the response variable takes on very large values, and the data was simulated using a population slope of 0.

Example: Extreme ranges for x and y



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	403920.18	83865.37	4.82	<.0001*
predictor	57627.913	140819.3	0.41	0.6846

Visually, there is clearly no correlation between the predictor and response variables.

- The slope is not "significantly different" from zero:

$$t = 0.41,\ p = 0.6846$$

- The slope also looks huge:

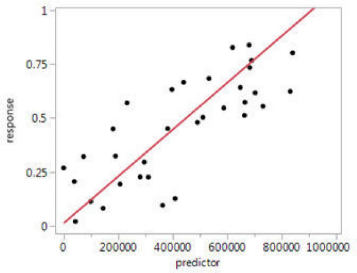
$$b_1 = 57,627$$

- But standard error is even larger:

$$se_{(b_1)} = 140,819$$

Note the scales of the axes!

Example: Extreme ranges for x and y, reversed



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0126702	0.054892	0.23	0.8186
predictor	1.0834e-6	1.065e-7	10.17	<.0001*

Visually, there is clearly a strong correlation between the predictor and response variables.

- The slope is highly "significant":

$$t = 10.17,\ p < 0.0001$$

- The slope also looks tiny:

$$b_1 = 0.00000108$$

- But standard error is even smaller:

$$se_{()b_1} = 0.000000107$$

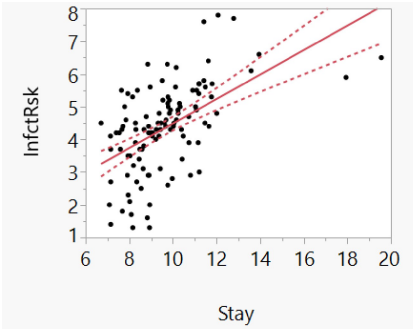
Again, note the scales of the axes!

Confidence Bands

- Just as we can make a confidence interval for a mean, difference in means, slope, etc.. we can also make a confidence interval for the line of best fit.
- We call this confidence interval for the line of best fit a **'confidence band'**

- In JMP, after fitting the LOBF, select the drop down menus by "Linear Fit" and select "Confid Curves Fit"
- This puts 95% confidence bands around the LOBF.

Confidence Bands in JMP



- Note that these bands form a shape that is similar to the sampling distribution of the LOBF we saw earlier.

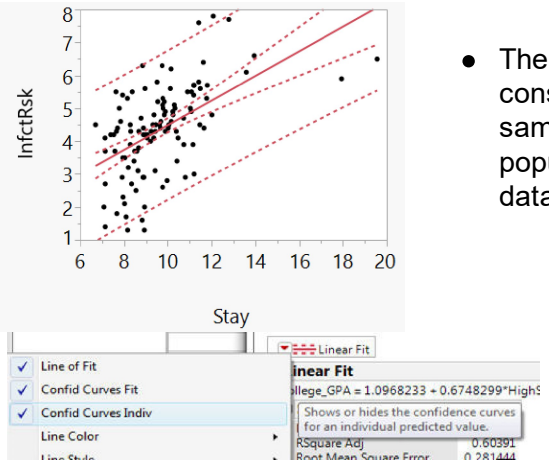
- The 95% confidence bands are constructed in such a way that, under repeated sampling, 95% of such bands will capture the population line.

Prediction Bands

- We can also make a confidence interval for individual points.
- We call this confidence interval for each individual point a **'prediction band'**
- The interval for will need to be much wider, because there is far more uncertainty involved in predicting the value of a single observation than there is in estimating the entire line.

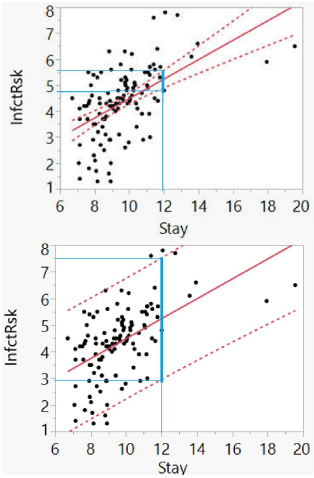
- In JMP, go back to the "Linear Fit" menu and select "Confid Curves Indiv"
- This puts 95% prediction bands around the individual points.

Prediction bands in JMP



- The 95% prediction bands are constructed in such a way that, if we sample new data from the population, 95% of individual new data points will fall within the bands.

Prediction bands vs. CI in JMP



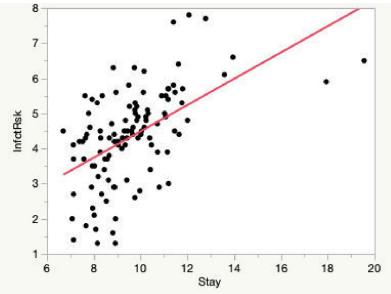
- Confidence Band: For any value of length of stay, these bands show the endpoints for a 95% CI for population mean infection risk.
- Prediction Band: For any value of length of stay, these bands show the endpoints for a 95% CI for an individual observed infection risk.

Outliers in Regression

- The last aspect of simple regression we will consider is the effect of outliers. Outlying observations can seriously influence the results of a study.
- Just as an outlier in a sample can have a large influence on the value of the sample mean, and outlier can also have a large influence on the estimated LOBF.

Example: Infection risk and outliers

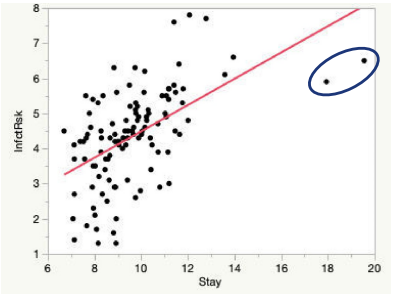
Let's take a closer look at the linear fit of Infection risk by Length of stay:



Does this look like a good linear fit?

Example: Infection risk and outliers

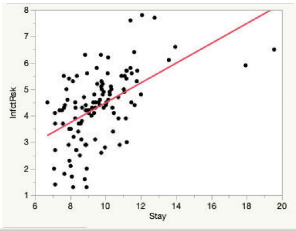
Let's take a closer look at the linear fit of Infection risk by Length of stay:



What if we remove the two values circled in blue? Do we expect the LOBF to differ? What about the R-squared statistics?

Example: Infection risk and outliers

Linear Fit with outliers:



RSquare	0.284562

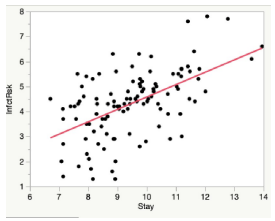
Linear Fit

InfctRsk = 0.7443037 + 0.3742169*Stay

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7443037	0.553857	1.34	0.1817
Stay	0.3742169	0.05632	6.64	<.0001*

Linear Fit without outliers:



RSquare	0.301914

Linear Fit

InfctRsk = -0.374503 + 0.495146*Stay

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.374503	0.692118	-0.54	0.5895
Stay	0.495146	0.072116	6.87	<.0001*

Example: Infection risk and outliers

Linear Fit with outliers:

RSquare	0.284562

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.7443037	0.553857	1.34	0.1817
Stay	0.3742169	0.05632	6.64	<.0001*

Linear Fit without outliers:

RSquare	0.301914

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.374503	0.692118	-0.54	0.5895
Stay	0.495146	0.072116	6.87	<.0001*

How do the two fits compare?

Outliers in Regression

- If we are trying to quantify the nature of the relationship between X and Y, we don't want our numbers depending heavily on a single data point. Ideally we would like all data points to contribute a roughly equal amount of information when determining the slope and intercept for the LOBF.

- The previous example illustrated the potential impact that an outlier(s) can have.

iClicker: Outliers

What should we do about the outliers?

A. Keep them in the model
B. Get rid of them and use the new model
C. See if we can call up the hospital to see what was going on
D. Give up and never do math again

7.4 Extra example

Example: Diamonds Data

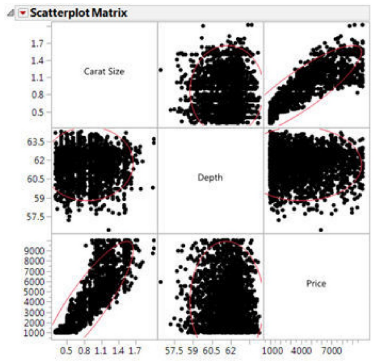
Gem-grade diamonds (the kinds put into rings) vary greatly in price. This data set contains information on 2,690 diamonds for sale on a diamond retail website.

Diamond prices are thought to be determined primarily by the "four C's" - Clarity, Color, Carat, and Cut. The dimensions of a diamond may also be relevant, such as Depth.

Which is a better predictor of price, Carat Size or Depth?

Example: Diamonds Data

We can look for correlations between Carat Size, Depth, and Price:



Multivariate

Correlations

	Carat Size	Depth	Price
Depth	0.0178	1.0000	-0.0174
Price	0.8617	-0.0174	1.0000

Correlation Probability

	Carat Size	Depth	Price
Carat Size	<.0001	0.3562	<.0001
Depth	0.3562	<.0001	0.3660
Price	<.0001	0.3660	<.0001

CI of Correlation

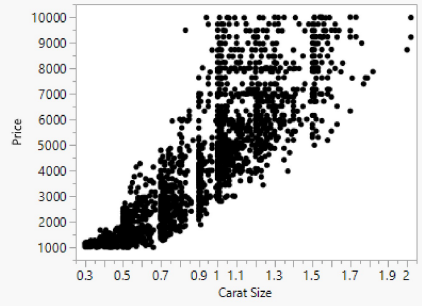
Variable	by Variable	Correlation	Lower 95%	Upper 95%
Depth	Carat Size	0.0178	-0.0200	0.0556
Price	Carat Size	0.8617	0.8517	0.8712
Price	Depth	-0.0174	-0.0552	0.0204

Example: Diamonds Data

Start by using Carat Size to predict Price.

You are modeling:

With the estimate:



Example: Diamonds Data

$$\widehat{Price} = b_0 + b_1 * Carat\ Size$$
$$= -1661 + 6473 * Carat\ Size$$

Test the hypothesis: $H_0 : \beta_1 = 0$

Linear Fit

Price = -1660.618 + 6472.8914*Carat Size

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1660.618	68.19971	-24.35	<.0001*
Carat Size	6472.8914	73.50489	88.06	<.0001*

Example: Diamonds Data

You run regression and get the estimated equation of the line as:

$$\widehat{Price} = b_0 + b_1 * Carat\ Size$$
$$= -1661 + 6473 * Carat\ Size$$

Linear Fit

Price = -1660.618 + 6472.8914*Carat Size

How do we interpret the intercept?

How do we interpret the slope?

iClicker: Prediction

$$\widehat{Price} = b_0 + b_1 * Carat\ Size$$
$$= -1661 + 6473 * Carat\ Size$$

What do we expect the price of a diamond to be if it has a carat size of 1.6?

A. $4812
B. $8134
C. $8695.8
D. $10356.8
E. I don't know

iClicker: Residuals

$$\widehat{Price} = b_0 + b_1 * Carat\ Size$$
$$= -1661 + 6473 * Carat\ Size$$

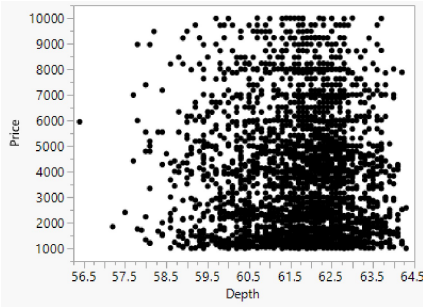
If you find a diamond at the online retailer that has a carat size of 1.6 but the real price is $8999, what is the residual for that diamond?

A. $1.03
B. -$303.2
C. $303.2
D. $17694.8
E. I don't know

Example: Diamonds Data

Now try using Depth to predict Price.

You are modeling:



With the estimate:

Example: Diamonds Data

$$\widehat{Price} = b_0 + b_1 * Depth$$
$$= 6124 - 35 * Depth$$

Test the hypothesis: $H_0 : \beta_1 = 0$

Linear Fit

Price = 6123.5753 - 34.873636*Depth

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6123.5753	2380.732	2.57	0.0102*
Depth	-34.87364	38.57101	-0.90	0.3660

Module Summaries

Module 1: Introduction to Statistics

Variables: Variables are items of interest which can take on different values (usually represented by words, letters, or symbols).

Observations: Each individual person/thing/unit we have measured in our data is an observation. Usually in a data set each row represents a unique observation.

Population: A population is the entire overall group we are interested in.

Sample: A sample is a subset of the entire population that we collect data on. The variable(s) of interest is/are measured on each member of the sample.

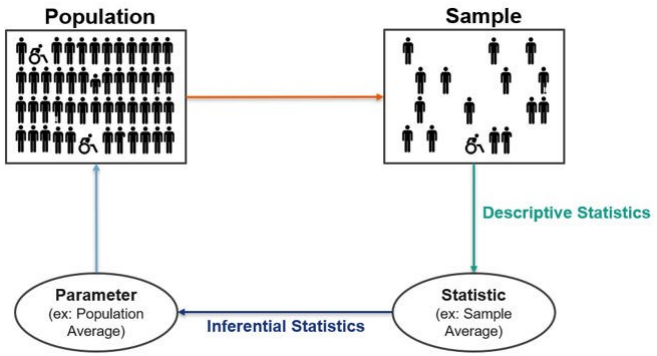
Sample Size (n): The number of observations in the sample.

Descriptive Statistics: The process of describing a dataset (ex: making a graph, finding an average). Strictly limited to the data itself. We do not generalize facts about the dataset to a larger group.

Inferential Statistics: The process of generalizing information from a sample to make claims about a population.

Parameter: A parameter is a numeric characteristic pertaining to a population. We usually never get to know the true value of a parameter, but we try to estimate it.

Statistic: A statistic is any number you calculate using data. We often use statistics to estimate parameters.



Response variable: This is the variable whose behavior we want to explain, or whose value we want to predict. Sometimes called the "dependent variable".

Predictor variable: These are variables which we think will be useful in predicting or in explaining the response variable. There may be more than one predictor variable in a study. Sometimes called the "independent variable", or the "explanatory variable".

Quantitative: Quantitative variables are numerical values that we can do sensible math with.

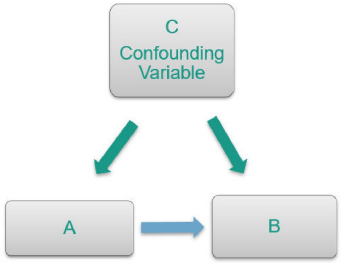
Categorical: Categorical variables take on values that are names/categories. They are usually words but can sometimes be numbers that represent categories (sometimes called qualitative variables).

Bias: Bias occurs when a study is set up in such a way that its results will tend to be systematically wrong (as opposed to just wrong because of random chance and inherent uncertainty). Some forms of bias are:

- **Sampling bias:** Can occur when a sample is taken in such a way that we would expect it to differ systematically from the population of interest.
- **Self-selection bias:** Can occur when choose if they want to be included in a sample. If the reason for their choosing to be in the sample is related to what is being measured, bias can result.
- **Non-response bias:** Can occur when certain types of respondents are less likely to answer a survey, and the reason they don't respond is related to the variable being studied.

Simple Random Sample: To mitigate bias, samples should be collected randomly. A simple random sample is a sample of the population where every unit has an equal opportunity to be selected, as in drawing names from a hat.

Confounding Variable: A confounding variable is a variable that influences both the predictor and response variables (but is usually not accounted for in the study).Formally, if variable C affects both variables A and B, then we might observe an association between A and B even though A and B have not causal relationship with each other.



Observational Study: In observational studies, variable values are "observed", but the researcher does not manipulate anything. Observational Studies do not result in cause and effect conclusions.

Experimental Study: In an experimental study, the researcher assigns members of the study to different experimental conditions (or "treatments"). Experimental Studies have the advantage of being useful for making cause and effect claims.

Module 1 Part 2: Summarizing Data

Measures of Location: They tell us where the data are located on a number line.

- **Mean:** It is the sum of all of the sample data, divided by the number of data points, or sample size. Identifies the data's center.

$$\text{sample mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Median:** It is the middle data point when the data are arranged from smallest to largest. Identifies the data's center.
- **Quartiles:** Quartiles break the data set into 4 quarters. The lower quartile, Q1 is the median of all the data below the overall median. The upper quartile, Q3 is the median of all the data above the overall median.
- **Minimum:** The smallest value in the sample.
- **Maximum:** The largest value in the sample.
- **Five Number Summary:** A convenient way to summarize a set of data. List the : Minimum , Q1 , Median , Q3 , Maximum

Measures of Dispersion: They tell us how spread out data are on the number line.

- **Range:** The range of a data set is the largest value minus the smallest value. It tells you the farthest distance between any two points in the data.

$$\text{Range} = \text{Maximum - Minimum}$$

- **Interquartile Range (IQR):** The IQR of a data set is the distance between its quartiles. It tells you how much space the middle 50% of the data takes up.

$$\text{IQR} = \text{Q3 - Q1}$$

- **Variance:** Measures how spread out the data are. In a sense, variance is the "average" squared amount any observation deviates from the mean.

$$\text{sample variance} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Standard Deviation:** Measures how spread out the data are. The standard deviation is a standardized amount by which observations deviate from the mean. It's kind of on average how far are the points from the mean.

$$\text{sample standard deviation} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Outlier: Outliers are data points that are located far away from the majority of the data is. Data analysts and software will all use different methods to identify outliers. An outlier is usually a data point that you should look closer at. They are most commonly: improperly entered data, measurement error, or accurate observations that are just unusual.

Frequency Table: A table that summarizes how often categorical variable takes on a given category.

- **One-Way Table:** Summarizes one categorical variable.
- **Two-Way/Contingency Table:** Summarizes two categorical variables at the same time.

Proportions/ Relative Frequency: Proportions are given as numbers between 0 and 1.

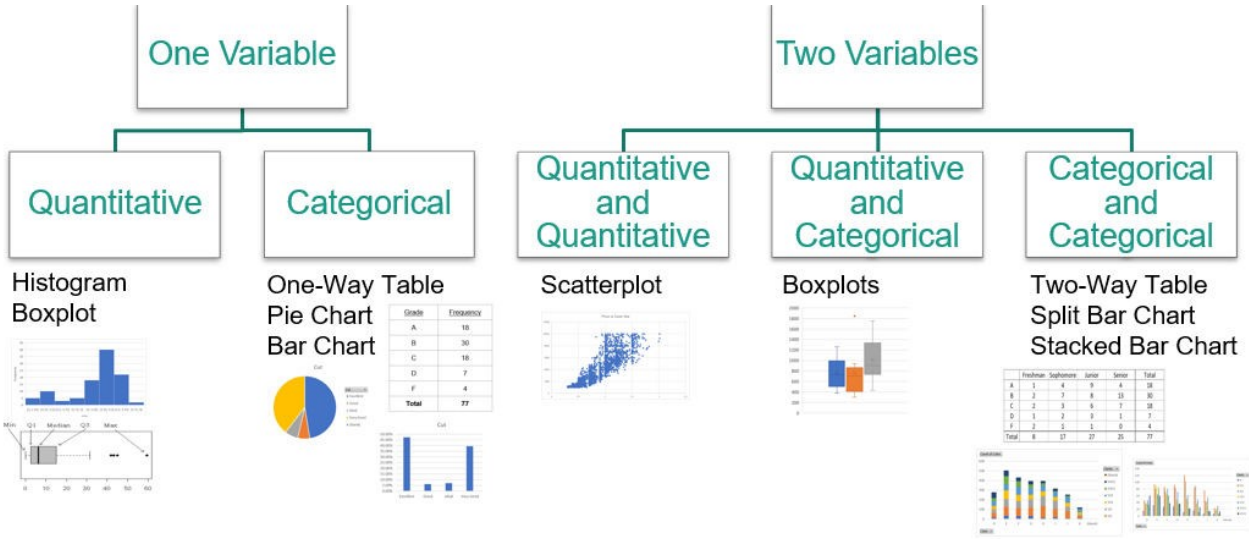
$$\text{proportion} = \frac{\text{number of observations of interest}}{\text{total number of observations under consideration}} = \frac{\text{part}}{\text{whole}}$$

Percents: Percents give a proportion in a slightly different format. To convert a percent into a proportion divide by 100 (move the decimal place to the left 2). To convert a proportion into a percent multiply by 100 (move the decimal place to the right 2). Make sure to use a % sign on any values that are percents.

Distribution: A distribution tells you the values a variable takes on, and the frequency with which those values are taken on. Ex: Tables, Histograms, etc

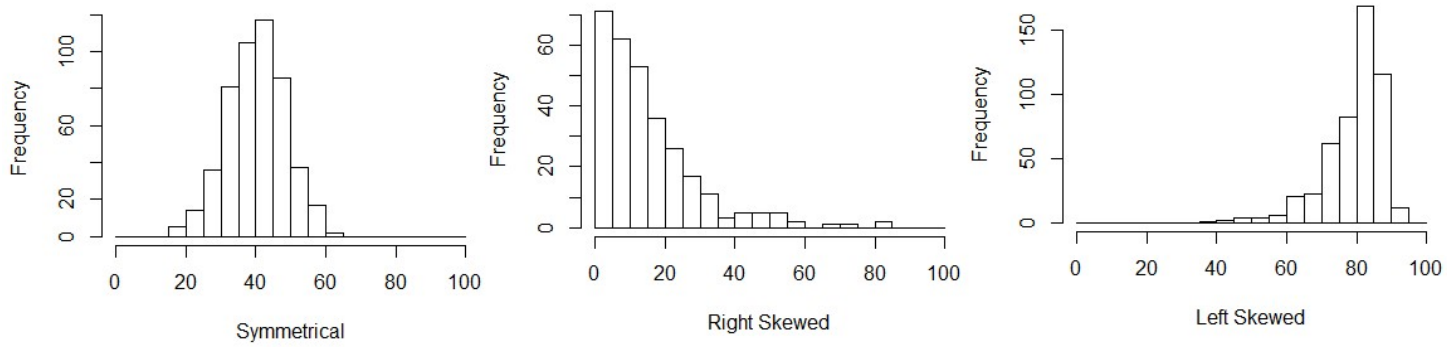
Graphs:

- **Pie Chart:** Pie charts can be used to summarize one categorical variable. Pie slices represent the proportion of observations in a category.
- **Bar Chart:** Bar charts can be used to summarize one categorical variable. The height of each bar represents either the frequency or relative frequency of observations in a category.
- **Split/Stacked Bar Chart:** Split bar charts can be used to summarize two categorical variable. The height of each bar represents either the frequency or relative frequency of observations in a category.
- **Histogram:** A histogram displays the distribution of a quantitative variable. Each bar represents the number of observations which fall into an interval (bin). The height of each bar corresponds to either frequency or relative frequency.
- **Boxplots:** Boxplots are used to display the distribution of one quantitative variable, or of one quantitative variable split up by the categories of one categorical variable. Boxplots are the visual representation of the five number summary. They can be displayed horizontally or vertically. If there are outliers, they are drawn as dots beyond the whiskers and the whiskers extend to either the min/max or the furthest non-outlier.
- **Scatterplot:** Scatterplots are used to plot the values of two quantitative variable against one another. Generically we call these X and Y. On a scatterplot, each dot show the X and Y coordinates for a single data point.



Shape (of a distribution): The shape of a distribution is classified by how often values are taken on. There are three options we cover:

- **Symmetrical:** If the two halves of the data look almost like mirror images, then we say a distribution is symmetrical or has no skew.
- **Right Skewed:** If there are a lot of low values and only a few high values, then we say a distribution is skewed to the right or positively skewed.
- **Left Skewed:** If there are a lot of high values and only a few low values, then we say a distribution is skewed to the left or negatively skewed.



Module 2: Probability

Random Event: something that may or may not occur, and that which we can assign a probability to.

Probability: is a way of quantifying the chance that some random event occurs. Formally a probability should be given as a proportion, between 0 and 1.

$P(X)$ is the probability that event X occurs

Conditional Probability: are probabilities of certain events occurring given that some other event occurs or has occurred.

$P(A|B)$ is the probability of event A occurring, given that event B has occurred.

Independent/Dependent: Two events are dependent when knowing what category one falls into changes the probability of the other occurring. They are independent when knowing what category one falls into does not change the probability of the other occurring.

Here the variable Letter can take on A or B. The variable Number can take on 1 or 2. We would say Letter is dependent on Number if $P(A|1) \neq P(A|2)$. We would say Letter is not dependent, or independent of Number, if $P(A|1) = P(A|2)$.

		Number		
		1	2	Total
Letter	A			
	B			
	Total			GRAND TOTAL

Simpson's Paradox: When we break data apart by a confounding variable we see that a particular relationship changes.

Percent Change: Percent change is one way to compare two values (and old and a new) or compare two groups.

$$\text{percent change} = \frac{\text{new} - \text{old}}{\text{old}} * 100$$

Percent change is given as a percent and can be negative (if the value decrease) or can be greater than 100%.

Percentage Points: This language is used when comparing two values that are measured in percents. One percentage point is equal to 1%.

Module 3: Standardization and Percentiles

Standardization: Putting measurements into a common unit.

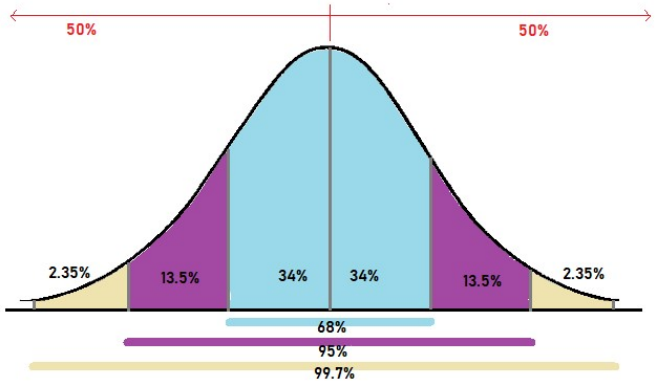
Z-score: The z-score of an observation tells the number of standard deviations that observation is away from the population mean.

$$z = \frac{x - \mu}{\sigma} = \frac{\text{how far away is the observation from the mean}}{\text{standard deviation}}$$

Where x is the value to be standardized, μ is the population mean, σ is the population standard deviation.

- A positive z-score means that the observation is above the mean. A negative z-score means that the observation is below the mean.
- The closer the z score is to 0, the more likely it is that it will occur. The farther the z score is from 0, the less likely it is that it will occur.

Empirical Rule: (also known as the **68/95/99.7 rule**). All normal distributions have the same shape, and when you mark off where the standard deviations are they will have the same percentage of observations falling in the same regions. 68% of the distribution lies within 1 standard deviation of the mean, 95% lies within 2 standard deviations, and 99.7% lies within 3 standard deviations.



Percentile: A percentile is the value of a variable for which the given percentage of values fall below it.

Module 4: Sampling Distributions

The Law of Large Numbers: As a sample size increases, the average of the sample will tend to get closer and closer to the true average of the population from which it is being sampled.

The Central Limit Theorem: For any population distribution (no matter how skewed or strange the population is), if we repeatedly take new random samples from this distribution and calculate the mean each time, then:

- As sample size increases, the sample average should get close to the population average. (This is the Law of Large Numbers)
- As sample size increases, the sample averages will be less spread out.
- As sample size increases, the distribution of the sample averages will look more like a normal distribution. That normal distribution will have the same mean as the population, μ, but the standard deviation, now called standard error, will be $\frac{\sigma}{\sqrt{n}}$

Sampling Distribution: Tells us the values that a statistic takes on, and how often it takes them on. It is the distribution of a statistic under repeated sampling.

Sampling Variability: A term that refers to the fact that new random samples will produces different values for the same statistic.

Standard Error: The standard deviation of the statistic under repeated sampling. It refers to the standard amount by which the value of a sample statistic will deviate from the value of the unknown population parameter it is estimating.

Z-score (modified): The z-score of a *sample mean* tells the number of standard *errors* that *sample mean* is away from the population mean.

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Module 5: Confidence Intervals

Confidence Interval: A confidence interval is a range of plausible values created for the purpose of capturing the value of an unknown population parameter value, at some level of confidence.

Confidence Level: The rate at which confidence intervals successfully capture an unknown population parameter, under repeated sampling of data from a single population.

Point Estimate: The value of a statistic that is being used to estimate an unknown population parameter value.

Margin of Error: The amount added and subtracted from a point estimate. It is $\frac{1}{2}$ the width of the confidence interval.

Critical Value: The number of standard errors that the confidence will cover above and below the point estimate.

Critical Value	Confidence Level
1	68%
2	95%
3	99.7%

Confidence Interval Formula:

$$\begin{aligned} \text{Confidence Interval} &= \text{Point Estimate} \pm \text{Margin of Error} \\ &= \text{Point Estimate} \pm (\text{Critical Value} * \text{Standard Error of Point Estimate}) \end{aligned}$$

95% CI for a mean μ:

$$95\% \text{ CI for} \mu = \bar{x} \pm 2 * \frac{s}{\sqrt{n}}$$

95% CI for a difference in means $\mu_1 - \mu_2$:

$$95\% \text{ CI for } \mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm 2 * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Paired Data: Studies where each subject is measured twice. Commonly used with a before and after study.

Module 6: Hypothesis Testing

Hypothesis Test: A formal way to test an idea in statistics. A hypothesis test can be broken down into three broad steps:

1. State the null hypothesis
2. Compute a statistic that tests the null hypothesis
3. Make the statistical decision to either reject or fail to reject (FTR) the null hypothesis

Null Hypothesis H_0: The hypothesis or proposition that we are testing against.

Statistical Decision: The result of the hypothesis test. There are only two options:

- **Reject** H_0 The data provide enough evidence against H_0. More specifically, we are saying that the statistical results we obtained would be unlikely to happen, if the null hypothesis was true.
- **Fail to Reject** H_0 The data do not provide enough evidence against H_0. This is NOT saying that the data prove that H_0 is true.

Errors: There are two types of errors that may result during a hypothesis test. We (almost) never get to know if we are committing an error, because we (almost) never get to know whether H_0 is true or false.

- **Type I Error:** Rejecting H_0 when H_0 is actually true. (When using a 95% CI, Critical Value of 2, or P-value of .05, the probability of a Type I error is 0.05.)
- **Type II Error:** Failing to reject H_0 when H_0 is actually false.

t - Test Statistic: The t-statistic measures how many standard errors away the point estimate is away from the value in the null hypothesis.

$$t = \frac{\text{Point Estimate} - \text{Null Value}}{\text{Standard Error}}$$

For one sample testing with $H_0 : \mu =$ *(null value)* then $t = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$

For two sample testing $H_0 : \mu_1 - \mu_2 =$ *(null value)* then $t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$

P-value: The probability that, from a random set of data, you would calculate a test statistic that is at least as large as the one you calculated, if the null hypothesis were true.

	Method Used		
Decision	Confidence Interval	T-statistic	P-value
Reject H_0	Null Value outside CI	$\|t\|>2$	$p < .05$
Fail to Reject H_0	Null Value inside CI	$\|t\|<2$	$p > .05$

Criticisms of the p-value/hypothesis testing:

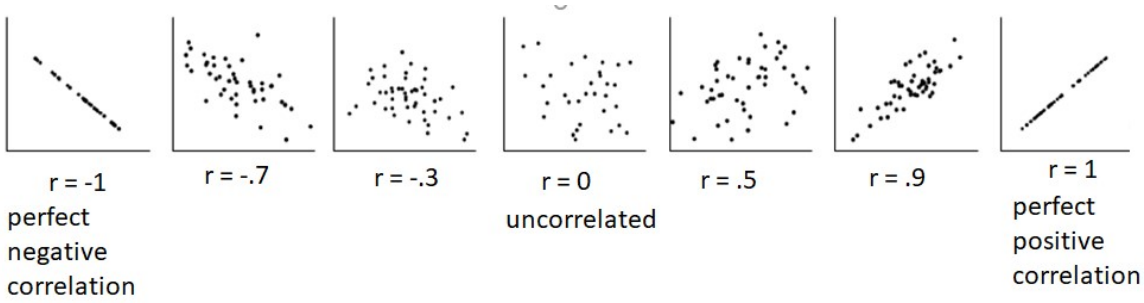
- **p-values are commonly misinterpreted** The p-value is NOT the probability that the null hypothesis is true.
- **p-values are commonly calculated using invalid methods** This definition of p-value assumes that you would have performed the exact same data analysis if you had a different set of data, which is not always the case.
- **hypothesis tests encourage dichotomous thinking** Using only 'Reject' or 'Fail to Reject' give people the impression that they have proven whether a hypothesis is true or false. There are many other things to consider, such as if the null hypothesis is worth considering, if model assumptions are realistic, or if there are outside sources of bias/uncertainty.

Module 7: Correlation & Simple Linear Regression

Correlated: If two variables are correlated, then they "move together" meaning that as one increases, the other one either tends to increase or tends to decrease. Variables x and y are correlated if knowing the value of x gives you insight on the value of y. Sometimes this is called 'dependent'.

Uncorrelated: Two variables are uncorrelated if they are not correlated. Sometimes this is called 'independent'.

Linear Correlation Coefficient r: A number between -1 and 1 that quantifies the correlation between two variables. This value is a *statistic*.



A general rule of thumb is:

|r| > .9	the correlation is 'strong'
.9 > |r| > .6	the correlation is 'moderate'
.6 > |r|	the correlation is 'weak'

Linear Correlation Coefficient ρ: The population correlation between two variables. This value is a *parameter* that we estimate with the statistic r.

Inference on correlation: A hypothesis test can be done to test if population correlation is equal to 0. If the population correlation is equal to zero this would imply that there is not a relationship between the predictor and the response. The hypothesis test is usually set up as: $H_0 : \rho = 0$ and JMP will compute the confidence interval, t-statistic, and p-value needed to complete the test.

Regression Line / Line of Best Fit (LOBF): If x and y are variables plotted on a scatterplot the regression line is the best line that passes through those points. It is considered best because it minimizes the sum of squared residuals. The LOBF contains all of the predicted values we could make for Y using X. We see the regression line in two forms:

- **Theoretical Model:** The population regression line. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- **Estimate:** The sample regression line calculated from the collected data. $\hat{y}_i = b_0 + b_1 x_i$

Intercept b_0: The intercept is the predicted value of y when x equals 0. Sometimes it does not make sense to interpret the intercept for your data, but it still needs to be included for mathematical reasons. The parameter β_0 is used in the theoretical model to represent the population intercept, which is estimated using the statistic b_0 calculated from the sample.

Slope b_1: The slope is the predicted change in y for a 1 unit increase in x. This does not imply a cause and effect relationship. The parameter β_1 is used in the theoretical model to represent the population slope, which is estimated using the statistic b_1 calculated from the sample.

Inference on a Slope: Generally when a model is being used a hypothesis test will be done on the slope to test if population slope is equal to 0. If the slope is equal to zero this would imply that there is not a strong relationship between the predictor and the response. The hypothesis test is usually set up as: $H_0 : \beta_1 = 0$ with

$$t = \frac{b_1 - 0}{\text{standard error}} \qquad \text{and} \qquad 95\%\text{ CI for } \beta_1 = b_1 \pm 2 * \text{standard error}$$

where standard error is found by a computer.

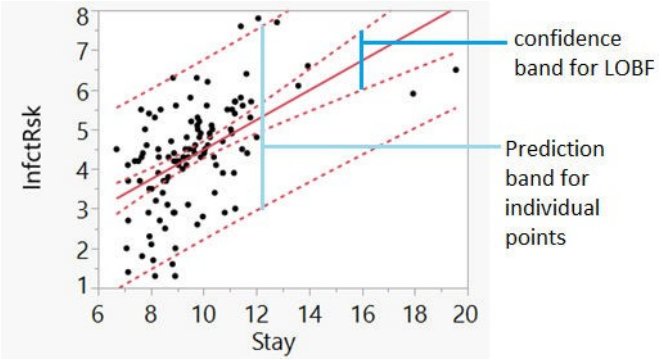
Residual: A residual, or *residual error*, is the difference between an actual observed value of Y and the predicted value of Y at an observed value of X. The observed value of y can be found in the data, and the predicted value is found by using x and the equation of the line of best fit to find the value of $\hat{y}$ on the line of best fit.

$$residual_i = y_i - \hat{y}_i = observed\ value - predicted\ value$$

R Squared R^2: The proportion of variation in the response variable that is explained by variation in the predictor variable. In simple linear regression, where there is only one predicto, R-squared is just taking "r" (the linear correlation coefficient) and squaring it. R^2 will always be between 0 and 1.

Confidence Bands: A confidence band is essentially a confidence interval for the line of best fit. The 95% confidence bands are constructed in such a way that, under repeated sampling, 95% of such bands will capture the population line (mean value of y for each x value).

Prediction Bands: A prediction band is essentially a confidence interval for individual points. The 95% prediction bands are constructed in such a way that, if we sample new data from the population, 95% of individual new data points will fall within the bands.



Extrapolation: Making predictions for Y at values of X that are beyond the range of our data. This is dangerous! It could be the case that X values beyond the range of our data don't make sense, or it could be the case that the relationship that we observe between X and Y doesn't hold outside of the range of our data.

Simple Linear Regression: The term simple implies that there is one 1 predictor variable and 1 response variable. The term linear implies that we are only making a model that follows a straight line relationship. The term regression is a fancy way of saying determining the relationship between two variables.