Introduction

Statistics is the theory and practice of analyzing data. **Data** are numbers with context.

A **data set** is a collection of **observations** and **variables** that are presented on a **spreadsheet** or **table**.

Example Table

Observations	**Variable 1**	**Variable 2**
Observation A	Value	Value
Observation B	Value	Value
Observation C	Value	Value
…	…	…

Observations are anything that you observe. Usually, you can count the number of observations. The number of observations is usually abbreviated with *n*.

Observations can also be measured. The feature or characteristic that is being measured is called a **variable**. Variables vary. A big part of statistics is about understanding what causes variables to vary.

The **value** is the actual number that is "attached" to an observation. For a given variable, each observation will have its own value. A variable is measured in **units**.

Bringing is all together in an example: We may be measuring the height of a group of 5^{th} graders. Each student is the observation, their height is the variable, the value is each child's height and the units would be feet or meters.

If you couldn't tell already, the hardest part about statistics is getting the language right.

Table of Contents

Statistics
introduction

The term **statistics** actually has two definitions:

1. Statistics as practice
2. Statistics as numbers

Statistics as practice: the act of collecting, organizing, summarizing, analyzing, and interpreting data.

Referring to definition 2, statistics are numbers that summarize a data set.

Statistics have different interpretations. There is an intuitive, verbal explanation of how the statistic relates to the data set. There is also a more formal, mathematical definition. This guide covers both for each statistic.

This guide will introduce you to many, many statistics. But to paint a broad stroke, there are two different statistics: the average and the variance.

The average is the single number that best describes the data set. The **variance** gives a measurement of the distribution of the data.

Central Tendency
introduction

The big idea behind the **central tendency** or the **average** is that you want to boil your data down to a single number that is most representative of the data set.

Averaging is achieved by a sort of mathematical compression- smoothing out the individual differences between the values of the observations.

We are going to discuss three measures of central tendency, the **mean**, the **median** and the **mode**, using verbal, numerical and graphical methods.

Central Tendency
the arithmetic mean

The **arithmetic mean**, $\bar{x}$, "x-bar" is the most common measure of central tendency. The arithmetic mean is so ubiquitous that it is colloquially referred to as "the mean."

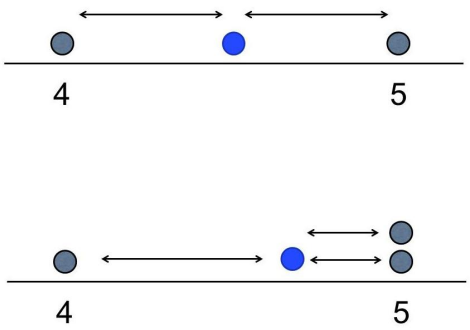
To find the arithmetic mean of a set of observations, add all the values of the individuals in the sample and divide by the number of observations.

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

The formula represents each of the observations in the data set as x_i. x is the value of each observation. The subscript *i* assigns order to the observations: the first observation, the second observation, through the final observation, *n.*

The sigma symbol $\sum_{i=1}^{n}$ represents the summation operation, from the first observation ($i=1$) to the final observation, n.

Central Tendency
the arithmetic mean, visual



In the image above, the mean is the blue dot halfway between 4 and 5- $\bar{x} = 4.5$. 50% of the distance lays on each side of the mean.

In the second image, we've added a third observation, $x_3 = 5$.

In the second image, the single line of the left of the mean has a length of 2/3 and the two lines to the right each have a length of 1/3.

From this example, we can derive a more general definition of the arithmetic mean: it is the value such that the cumulative distance between the observations to the left of the mean and the mean *equals* the cumulative distance from the mean to each observation to the right of the mean.

Central Tendency
the median

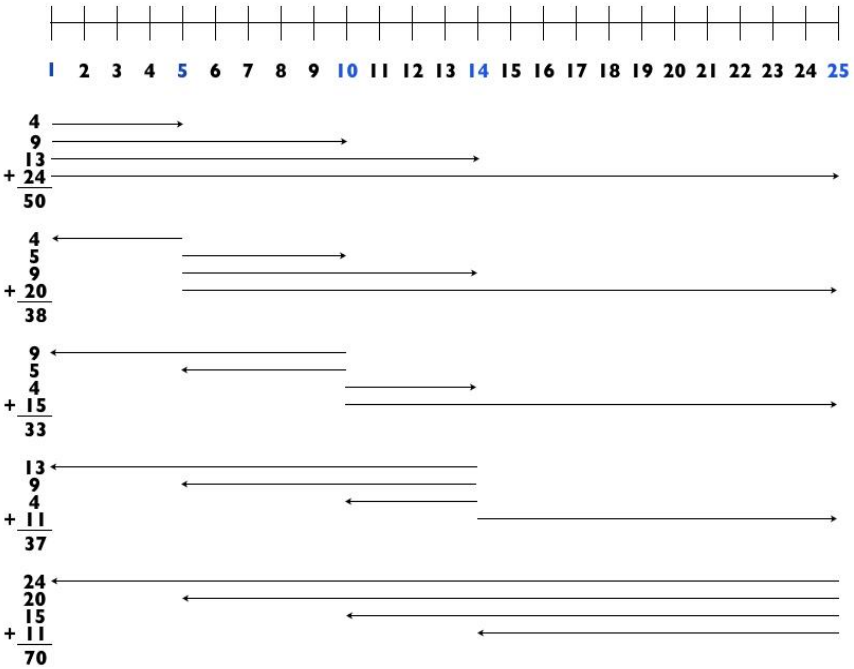
The **median**, M, is easily understood as the observed value in the middle of a series of sequentially ordered observations. That is, the median is the value with an equal count of observations above and below it.

This is only true when the data set contains an odd number of observations. When the data set contains an even number of observations, there is no middle observation. In such a case, the median is the arithmetic mean of the two numbers in the middle.

For example, take the data set [1, 5, 10, 14, 25] The median is 10. 10 is the median because it is the observation in the middle.

Take the data set [1, 5, 10, 14, 25].

Below the number line are arrows that measure the absolute distances from each particular observation in the data set to all other observations. The distances are summed on the left side of the picture below.



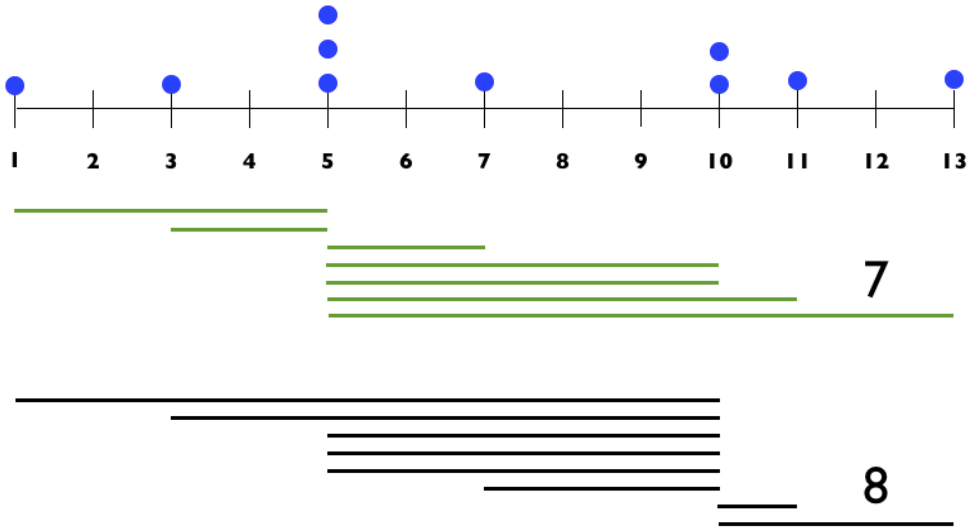
From this example, we can deduce a more general definition of the median: The median value is the value in the data set that minimizes the sum of the absolute values of the distances from each value to all other values.

Central Tendency
the mode

The mode is the value that occurs most frequently in the data set. For example, take the data set [1, 3, 5, 5, 5, 7, 10, 10, 11, 13]. The mode is 5.

Again, let's invoke the number line.

Let's measure all distances as equal to 1 and sum the distances between all the observations and the value 5 versus all of the observations and the value 10.



When the mode is 5, the distance is minimized at 7.

The mode can also be defined as the number in the data set that minimizes the distances between the mode and all other values when each distance equals 1.

Variance
introduction

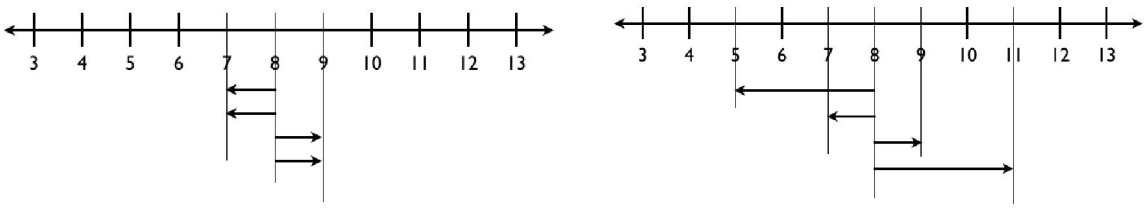
Variance[1] is used to gauge the accuracy or precision of the average. A low amount of dispersion shows that the average accurately represents the data set, and the converse is true as well.

Two very different data sets can have two different measures of dispersion but the same central tendency. Take two example data sets: [7, 7, 9, 9] and [5, 7, 9, 11]. First, let's calculate their means.

1. $\frac{7+7+9+9}{4}=8$ 2. $\frac{5+7+9+11}{4}=8$

Although the data sets have the same mean, there's something different about the observations. The observations in data set 2 are further from the mean.

We can measure the deviation scores**.**



We can sum the deviation scores to get the **absolute deviation**- the total of the distances that each observation is from the mean.

$$\sum_{n=1}^{i}\left|x_i-\bar{x}\right|$$

However, comparing these values to the mean doesn't make any intuitive sense because the mean is an average but the absolute deviation is a total of deviation scores.

Rather than using the absolute deviation, we'll be using the **average deviation**, which divides the absolute deviation by the number of observations.

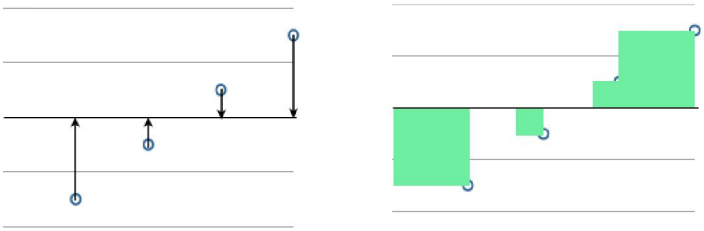
$$\frac{\sum\left|x_j-\mu_X\right|}{N}$$

The average deviation is not commonly used in statistics either.

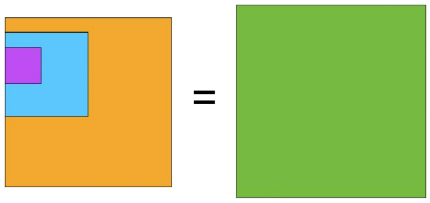
[1] deviation, dispersion, spread, wildness, scatter

Variance
useful statistics

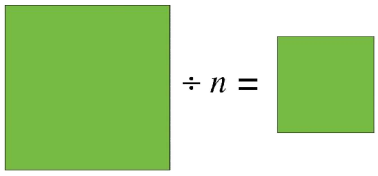
The **root mean square** (RMS) outlines the most common statistical process for measuring dispersion.
First, we measure the deviation scores. Next, we're going to square the distances from the observations to the mean, which eliminates any negative values.



After that, we combine the squares into a single area- the **sum of squares**. Understand that variability can always be thought of in terms of geometrical areas.

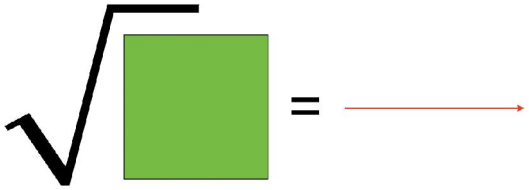


When we divide this large square by the sample size, n, we get the **mean square or variance.**



$$\text{var} = s^2 = \frac{\sum_{j=1}^{N}(x_j - \mu)^2}{N}$$

The variance cannot be compared to the mean because the variance is in squared units, so we take the square root of the mean square. This is the root mean square. (Now we know where the name comes from.) This is also called the **standard deviation**. The standard deviation is the root mean square of the deviations from the mean. It tells us how far the "typical observation" will be from the mean.



$$sd = s = \sqrt{\frac{\sum_{j=1}^{N}(x_j - \mu)^2}{N}}$$

Variables
Quantitative Variables

Variables are equated with X.

A quantitative variable is a variable that can be described with a number.

If you can add or subtract the values among different individuals and produce a result that makes logical sense, then you are working with a quantitative variable.

Discrete quantitative variables are quantitative variables that take on whole number values, 1, 2, 3... If you can count it on your fingers, it's a discrete quantitative variable.

Continuous quantitative variables are variables that can assume any value, even irrational numbers. If you can measure it with a ruler or a scale, then it's a continuous quantitative variable.

Variables
Categorical Variables

Qualitative variables describe variables that have an implicit ordering- they are **ordinal**.

Categorical or **nominal variables** describe variables whose values have no distinct relation to one another.

Qualitative	Categorical
population density: low/medium/high	Rural/Urban Area
Height: short/medium/tall	hair color: blond/brunette/redhead
Weight: slender/average/overweight	Gender: m/f
Age: young/middle age/old	Ethnicity: Asian/black/latino/white

The values for each variable are also referred to as **levels** in further analysis. **Levels** can apply to class ranges as well when dealing with quantitative variables.

The fact that a variable is labeled with a number does not make the variable quantitative. For example, zip codes, phone numbers and social security numbers are numbers, but it wouldn't make sense to add these variables.

Variables
Categorical Variables (cont)

The **count** or **frequency** is the number of observations that take a particular value for each category.

A **frequency table** lists the frequency of each category on a table. It is used to represent the distribution of a categorical variable.

A **proportion** is equal to the count of each category divided by the total number of observations. A proportion can also be thought of as the **relative frequency**, or the percentage, of individuals in each category.

This is called a **relative frequency table**, as it lists the percentage of observations in each category.

Category	Count	Proportion (Percentage)
Group 1	3	3/10 (30%)
Group 2	5	5/10 (50%)
Group 3	2	2/10 (20%)
Total	10	10/10 (100%)

A **bar chart** or **bar graph** is used to visualize the distribution of counts within a categorical variable.

First, we create a standard x/y-axis. The x-axis isn't set up numerically, since the categorical variables are not quantitative.

The x-axis simply serves as the foundation for rectangular, vertical **bars** or **blocks**. Each bar represents a different category. The category is labeled under each bar. The y-axis represents the count of observations in each category. The height of the bar is equal to the count of observations.

There are two caveats to keep in mind with a bar chart. First, the height of the bars is important, but the width isn't. Second, leave a space between bars. If the chart were ordered by size, with the largest chart (bananas) to the smallest chart (oranges), one would be building a **Pareto chart**.

Distributions
Introduction

Analysis happens at the variable level. We can rise above the individual observations and analyze the landscape of values by looking at the variable's **distribution**.

A distribution has three components. The first two are the center and spread. We describe these using the average and variance.

The third component is the distribution's **shape**.

We display distributions by first constructing a table, then graphing the distribution.

This section will focus on analysis for two types of variables: categorical/qualitative and quantitative.

Finally, we will discuss more general distributions on a theoretical level.

Distributions
Histogram tables

We summarize quantitative variables on a **histogram table**.

Class	Frequency	Relative Frequency	Cumulative Frequency	Relative Cumulative Frequency
60-62	n_1	$n_1 / \sum n$	n_1	$n_1 / \sum n$
62-64	n_2	$n_2 / \sum n$	$n_1 + n_2$	$(n_1 + n_2) / \sum n$
64-68	n_3	$n_3 / \sum n$	$n_1 + n_2 + n_3 = \sum n$	$(n_1 + n_2 + n_3) / \sum n$

A quantitative variable can have an infinite number of values, so instead using pre-determined categories to construct a distribution, we group observations into **classes**.

> Classes are just small ranges of data. For example, fighters are put into "weight classes," light-weight through heavyweight. While weight classes are described qualitatively, they depend on the quantitative weight of the athlete. For example, lightweight may be 60 to 62 kg, welter weight, etc.
>
> The **class frequency** is the number of individuals in each class. The **class interval** is the sub-range of data.
>
> The **class width** is just the distance from the lower limit to the upper limit. In the example above, the class width is 3. The **mark** is at the center of the class width, ex. 61.
>
> The **class limits** are the numbers at the beginning and the end of the class. For the class interval 60-62, the **lower class limit** is 60 and the **upper class limit** is 62.

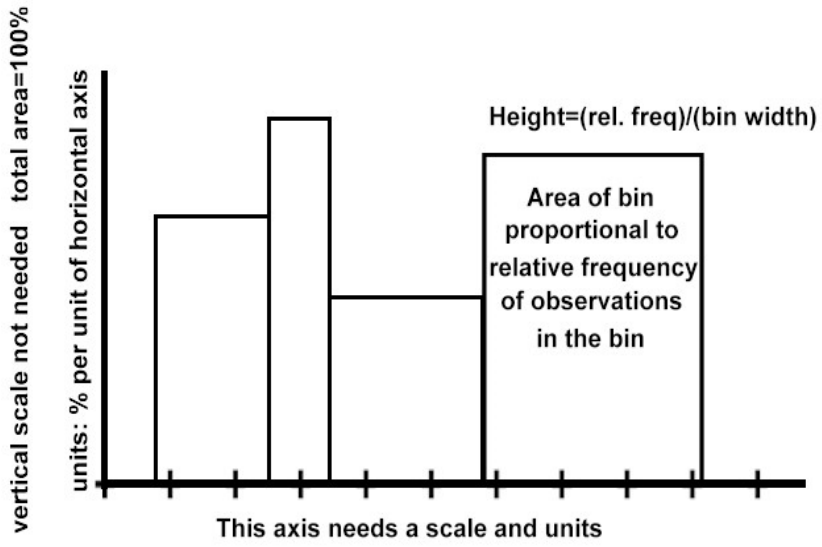
The first and second columns are the frequency and relative frequency.

The second is the **relative cumulative frequency**, which is the cumulative frequency for each class (and previous classes) divided by the total number of observations.

The third is the **cumulative frequency**, which for each class is the sum of all previous cases and current case frequencies.

Distributions
histograms

A **histogram** measures the relative frequency distribution using block or **bins** to represent the class intervals**.** The area of all of the blocks adds to 100%.



The x-axis is usually is the unit of the variable, for example, weight in lbs.

The vertical axis on a histogram is called a **density scale**. The density scale measures the relative frequency per unit on horizontal axis.

For example, if a block had a class width of 3 inches and a relative frequency of 30%, it would have a **common unit** of 1 inch and a height of 10%.

This makes calculating the relative frequency for a class easy: it is simply the product of the density scale and the class width.

The scale of the vertical axis is automatically imposed by the fact that the class widths are pre-determined and the total area of all the blocks is 100%.

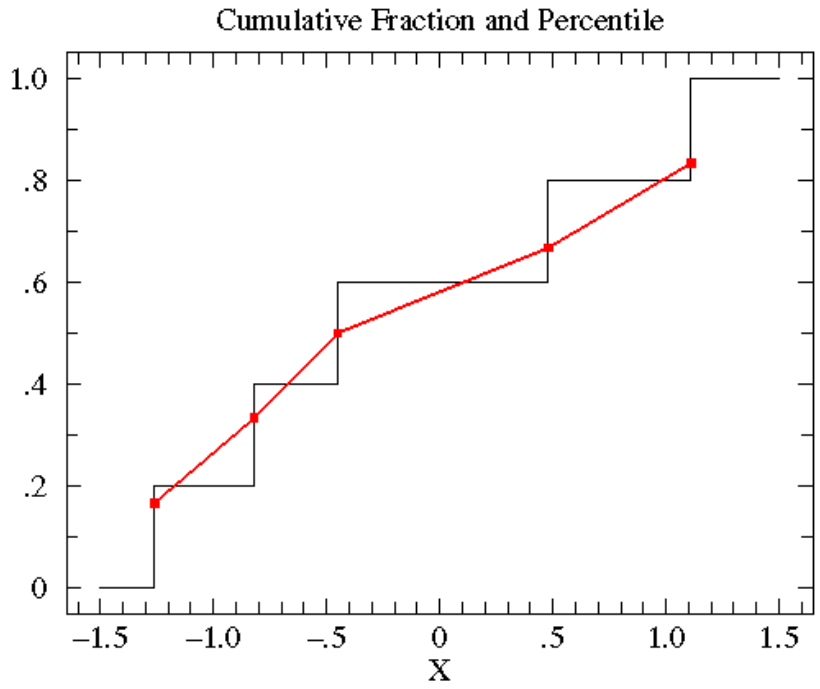
Distributions
ogives (OH-Jives)

Ogives graph the relative cumulative frequency. They are used to find the relative position of an individual within a group of observations

The x-axis on an ogive is the same as for a histogram, but the y-axis is the relative cumulative frequency.

After setting up your axes, put a dot at the intersection of the class mark and relative cumulative frequency for each class. Connect the dots to build an ogive curve.

To find the cumulative percentile within a bin, locate the class mark, draw a vertical line to the ogive curve, and then a horizontal line to the y-axis.



For example, if you're interested in the percentage of presidents older than 50 at inauguration, you would find 50 on the x-axis, move up to the ogive curve, then to the y-axis for the cumulative percent. If you have a % and want the value, work backwards.

You can use a "break-in-scale" symbol on the x-axis (//) so it doesn't start at the origin. If you're constructing a histogram of the age at which someone becomes president, no one is president at the age of 12, so you can break in the x-axis.

Distributions
Endpoint convention

If the class limits coincide, then we need to establish an **endpoint convention**. Each range must be disjoint, with no overlapping values.

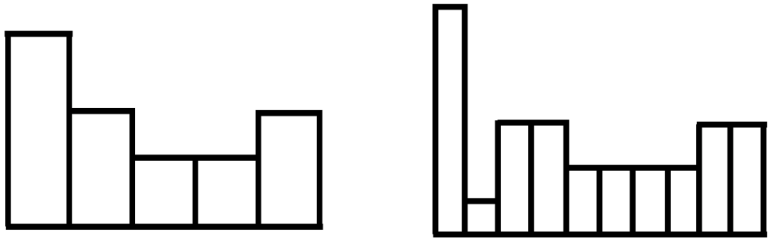
To establish an endpoint convention, one limit (either upper or lower) will be **inclusive**, that is, it includes observations with values equal to the limit, and the other limit will be **exclusive**, that is, it includes observations with values equal to the exclusive limit will not be included in the interval. The key is to be consistent.

An alternative to "inclusive/exclusive" is to choose class limits that do not coincide. This only works with discrete quantitative variables.

Classes are determined at the discretion of the statistician. There is no right or wrong way.

It's important to avoid putting to many or too few observations in a single class.

When we bin the data, we lose information about the distribution of values *within* each bin. For example: A call center puts in a policy mandating that operators minimize the duration of calls. Predictably, the average duration of calls declines.

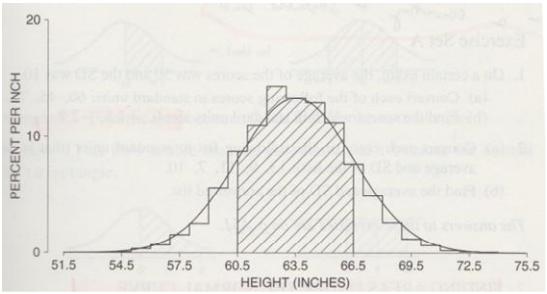


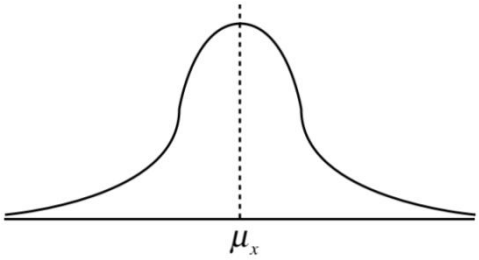
Originally, the class interval was 60 seconds. But when the class intervals were narrowed down to 10-second intervals, we find that 7.6% of calls were less than 10 seconds! As an unintended consequence, operators simply hung up on customers with difficult questions.

Distributions
The normal distribution

A normal or bell shaped distribution[1] is a symmetrical, bell-shaped curve.

A density curve or frequency curve is a curve that fits the irregular bars of a histogram. A density curve describes the overall, idealized shape of a distribution and ignores minor irregularities as well as any outliers.





A density estimator is software that looks at the data and draws a density curve that describes the overall shape of the data. It does not start with any specific shape. Software draws density curves without the user having to specify the shape of the curve.

The equation of a line that traces out the shape of a density curve is called a density function.

In this section, we'll be looking at some general shapes with an emphasis on the normal distribution.

The distribution for the normally distributed variable X is specified by the mean and the standard deviation, abbreviated as $X \sim N(\mu, \sigma)$, where μ is the population mean and σ is the population standard deviation.

As you can see, the normal distribution has most of the area near the average. It's symmetrical and the probability decreases as we move away from the center.

Normally distributed variables are either independent or highly dependent.

For example, the roll of a die is independent. Height is highly dependent. Both are normally distributed.

Furthermore, while a population distribution for a variable may not be normal, a sub-sample of the data may be. For example, the fuel consumption for all vehicles isn't normal, but the fuel consumption for light trucks is.

The points on the normal curve at which the curvature changes from convex to concave are called inflection points. They are always one standard deviation from the mean in both directions.

What is unique about the bell curve is that its area is tabulated according to the empirical rule, such that:

---

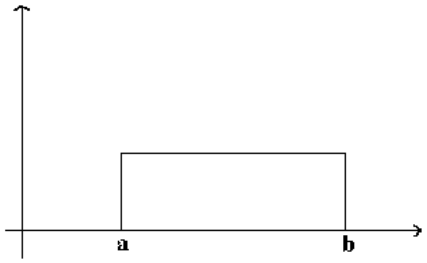
[1] It is also called the Gaussian curve for the German mathematician Carl Friedrich Gauss. In 1908, and English statistician by the name of Karl Pearson uncovered historical papers proving that the normal distribution was actually discovered a century before by the English mathematician Abraham De Moivre.

68% of all observations fall within $\pm\sigma$
95% of all observations fall within $\pm 2\sigma$
99.7% of all observations fall within $\pm 3\sigma$

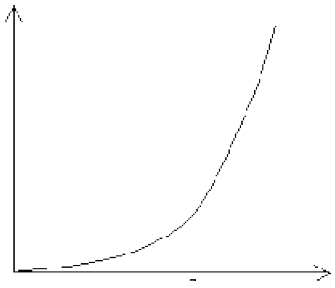
Only about 6 out of 100,000 observations are outside of the interval from -4 standard deviations to 4 standard deviations. The area outside of the interval of -5 to 5 is 0.00000057.

Visualizing Distributions
shapes, cont.

Uniform distributions have an equal number of observations across values. ex. cola temperature in fridge, measured at random times.

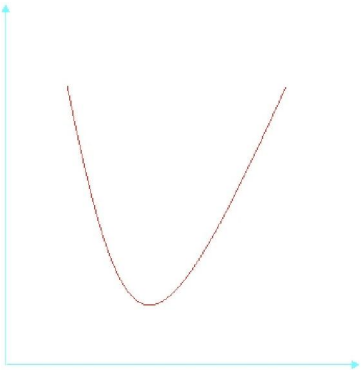


J-shaped or reverse J-shaped-maximum occurs at one end or another. J-shaped curves are often cumulative, like ogives, and are measured over time, for example, the number of calories eaten in a day, measured by the hour.

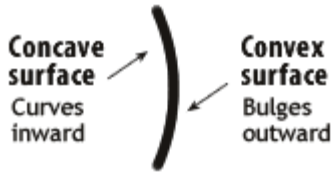


Visualizing Distributions
shapes, cont.

U-shaped- maxima @ both ends and a minimum in between, for example, the daily temperature measured from August 1st to July 31st. There is also **reverse u-shaped** curves.

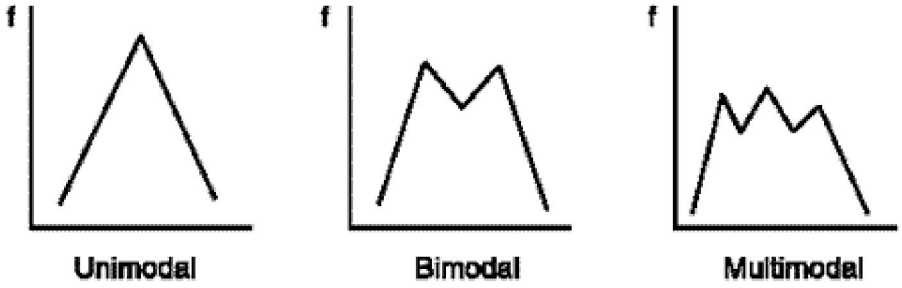


Remember that each curve has two surfaces: a concave surface and a convex surface.



Uni-modal, bi-modal and multi-modal curves have 1, 2 or 3 maxima respectively. Note the term "mode." We're counting the number of modes.

Multiple modes usually implies multiple groups. For example, a bimodal distribution can be the heights of men and women together on the same distribution. Bimodal distributions are equal to the sum of two different distributions.



An example of the tri-modal distribution is the number of colleges and their tuition and fees. You would have a lot of colleges that chart a little (community colleges), another modal that charges the median amount (state schools), and the final modal charging a large amount (private colleges).

Probability
introduction

Probability or **chance theory** is humanity's attempt to understand and measure uncertainty. Probability is about calculating how likely it is that an event going to occur.

Chance and randomness are philosophical minefields. Several theories of probability underlie statistical argument in particular.

In the **theory of equally likely outcomes**, probability assignments depend on the assertion that no particular outcome is preferred over any other by nature. The chance of heads or tails is 50/50 because a head or tail is equally likely.

Or is it? Enter logic and mathematics. Enter science and scientific assumptions. Enter experimentation.

We can flip the coin infinity times. We can take a small sample and use probabilities to measure our uncertainty. In the **frequency theory of probability**, we use some calculus to compare our sample to the infinite limit of our model.

Subjective theory is just what we predict. So it's open to bias of all kinds.

It's something of a trip to think of different realities when only one reality exists. The scope of all possible events and outcomes, given preceding causes and effects, raises questions about the nature of reality.

An **event** describes a singular reality. We abbreviate events with capital letters A, B, C, etc.

A reality can have many potential **outcomes**.

A **trial** is an event in practice, for example, the rolling of a die or the flipping of a coin.

For example, event A can be the flipping of a coin and the outcomes would be either heads or tails.

We desire successful outcomes, and events are to be taken as given and measured. In the event of drawing a card, drawing a spade is an outcome. But in the event of drawing a spade, the ace or the king is an outcome.

This section provides the basics of probability and logic needed for understanding statistical inference.

Probability
Set notation

When we are setting up a probability problem, we want to write it in mathematical or **set notation**.

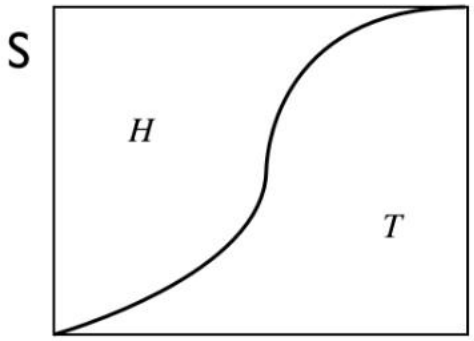
Rather than asking, "What is the probability of the **successful** outcome *k* in event A?" we would simply write:

$$X = P(k = A)$$

Event A will have many potential outcomes:

$$S = \{x_1, x_2 ... x_n\}$$

The **sample space**, S, also called the **space of possibilities**, is nothing more than a graphical representation of the set of all outcomes.



$\cap$ is called **intersect** and means "and."

$$A \cap B$$

$\cup$ is called **union** and means "or."

$$A \cup B$$

An event that is not another event is its **complement**.

$$\sim A$$

The probability that event B occurs after event A occurs is **conditional probability**. Conditional probabilities are read right to left; first, A happens, second, B happens.

$$P(B \mid A)$$

Probability
complement rule

Both an event and its nonevent can not occur. This is the **complement rule**.

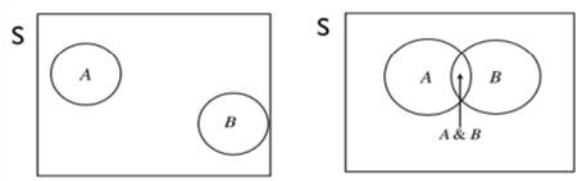
$$P(A^c) = 1 - P(A)$$

This is only true if the events are mutually exclusive, $A \cap A^c = 0$.

If there is more than one outcome classified as a failure, then their summed probability is the probability of the complement. For example, if we're rolling a die and a success is a two, then a failure is a one, three, four, five or six.

Probability
Joint and disjoint events

Joint events share some number of outcomes. That is, at least a single outcome belongs to both events. **Disjoint** or **mutually exclusive events** have no outcomes belonging to both events.



If $P(A \cap B)$ is greater than zero, then events A and B are joint.

$P(4 = A \cup \Diamond = B)$?

Spades	2	3	4	5	6	7	8	9	10	J	Q	K	A
Diamonds	2	3	4	5	6	7	8	9	10	J	Q	K	A
Hearts	2	3	4	5	6	7	8	9	10	J	Q	K	A
Clubs	2	3	4	5	6	7	8	9	10	J	Q	K	A

The **addition rule for joint events** states that the probability event A union event B is the sum of the events individual probabilities minus the probability of all outcomes for event A intersect event B.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(K = A \cup A = B)$?

Spades	2	3	4	5	6	7	8	9	10	J	Q	K	A
Diamonds	2	3	4	5	6	7	8	9	10	J	Q	K	A
Hearts	2	3	4	5	6	7	8	9	10	J	Q	K	A
Clubs	2	3	4	5	6	7	8	9	10	J	Q	K	A

The **addition rule for disjoint events** tells that, as long as the events are disjoint, the probability of event A or event B is the sum of their probabilities.

$$P(A \cup B) = P(A) + P(B)$$

Probability
independence

Two or more events are **independent** if the occurrence of one outcome or event does not change the probability that the other outcome or event occurs. The probabilities are referred to as **unconditional**.

$P(4 = A \cap \Diamond = B)$?

Spades	2	3	4	5	6	7	8	9	10	J	Q	K	A
Diamonds	2	3	4	5	6	7	8	9	10	J	Q	K	A
Hearts	2	3	4	5	6	7	8	9	10	J	Q	K	A
Clubs	2	3	4	5	6	7	8	9	10	J	Q	K	A

The **multiplication rule for independent events** says that, as long as the events are independent, the probability of event A intersect event B is equal to the product of the probabilities of the separate events.

$$P(A \cap B) = P(A)P(B)$$

Probability
conditional probabilities

Event A is unconditional (or independent) and event B is conditional.

If $P(A \cap B)$ is a **conditional probability**, then it relies on the unconditional event occurring as well as the conditional event occurring.

$$P(A \cap B) = P(A)P(B \mid A)$$

A more useful form of the equation:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

The probability of drawing an ace given that we've drawn a spade is:

$$P(B \mid A) = \frac{1/52}{1/4} = 4/52$$

The probability of drawing a spade given that we've drawn an ace is:

$$P(A \mid B) = \frac{1/52}{1/13} = 13/52$$

Expand the number of events to three and torture your mind.

$$P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B)$$

Probability
law of unions of independent events

The **law of unions of independent events** gives us the probability that at least one successful event in a disjoint sample set happens.

$$P(A \cup B) = 1 - P(\sim A \cap \sim B)$$

Say you're looking for a job, and since the market is awful, you have a 10% chance of getting a call back.

If you apply to 2 jobs, then the probability of getting at least one call back $P(A \cup B)$ is the product of the complements (getting rejected by both companies) subtracted from the total sample space.

$$P(A \cup B) = 1 - P(\sim A) * P(\sim B) = 1 - 0.9 * 0.9 = 19\%$$

By applying to 10 jobs, the probability of getting at least one call back has increased to 65%!

Even with a small probability of success, the probability of at least a single positive outcome increases as the number of trials increase.

Probability

combinatorial analysis

Combinatorial analysis or **combinatorics** is nothing more than a sophisticated way of counting. Flipping a coin has two potential outcomes. What about flipping two coins? Counting the total number of outcomes can be difficult.

There are two basic formulas for combinatorial analysis: **combinations** and **permutations**. All you really need to remember is that with combinations, you are counting the number of unique groups, and with permutations, you are counting the number of unique groups *as well as* the number of unique arrangements of individuals within each group.

For example, let's say we want to know how many ways we can seat 4 people with only 3 available chairs.

The number of combinations, ${}_nC_r$, is the number of unique 3-person groups. n is the number of individuals and r is the size of the group. Of course, n is always greater than r. What we say is, "Out of n, choose r."

$$ {}_nC_r = \frac{n!}{r!(n-r)!} = \frac{4!}{3!(4-3)!} = 4 $$

So when we have 4 people and 3 seats, there are 4 unique 3-person groups that we can form with 4 individuals.

A **permutation**, ${}_nP_r$, tells us the number of unique groups and the number of unique arrangements.

If the three people sitting in chairs are Joe, Jeff and Jack, this is a single permutation. If the order changes to Jack, Joe and Jeff, this is a different permutation. The formula permutation is:

$$ {}_nP_r = \frac{n!}{(n-r)!} \qquad {}_nP_r = \frac{n!}{(n-r)!} = \frac{4!}{(4-3)!} = 24 $$

Think of the additional *r!* as the number of ways to arrange the individuals in each unique group. This point is illustrated in the diagram below. Combinations are blue and the permutations are **both** the red and the blue.

$$ \binom{4}{3} = \begin{array}{cccccccccccccccccccccccc} A & B & C & A & C & B & B & A & C & B & C & A & C & A & B & C & B & A \\ A & B & D & A & D & B & B & A & D & B & D & A & D & A & B & D & B & A \\ A & C & D & A & D & C & C & A & D & C & D & A & D & A & C & D & C & A \\ B & C & D & B & D & C & C & B & D & C & D & B & D & B & C & D & C & B \end{array} $$

We can break down the formulae for permutations and combinations a bit more.

$n!$ in the numerator is just the number of people who can fill each seat. 4 people can fill the first seat, multiplied by 3 the second seat, multiplied by 2 in the third...

$(n - r)!$ eliminates the number of objects we're not considering. When we have more people than seats, we have to account for the surplus number of people by mathematically truncating the factorial in the numerator.

Again, $r!$, eliminates different arrangements of the same group of objects. Out of a group of 3, there are 3 places for the first man, multiplied by 2 for the second…

Inference
Introduction

This section covers the theory of statistical inference. Let's start by defining some terms.

The **population** is everyone and everything- all possible observations in the universe. For example, the population of orangutans comprises every orangutan on Earth.

A **finite population** is limited, like the number of people in a town. An **infinite population** is unlimited, like the number of stars in the universe.

Sometimes, defining the population can be tricky. For example, the population of voters isn't everyone in a given country- it is everyone over the age of 18 in the country.

When a statistic describes a population, we call the statistic a **parameter**. Parameters measure characteristics of populations. We use Greek symbols to represent parameters. We symbolize the population mean with the Greek letter μ. The standard deviation of the population is σ.

Α α	alpha	Ν ν	nu
Β β	beta	Ξ ξ	ksi
Γ γ	gamma	Ο ο	omicron
Δ δ	delta	Π π	pi
Ε ε	epsilon	Ρ ρ	rho
Ζ ζ	zeta	Σ σς	sigma
Η η	eta	Τ τ	tau
Θ θ	theta	Υ υ	upsilon
Ι ι	iota	Φ φ	phi
Κ κ	kappa	Χ χ	chi
Λ λ	lambda	Ψ ψ	psi
Μ μ	mu	Ω ω	omega

Greek alphabet chart © by deTraci Regula; licensed to About.com

Learn your Greek, intimidate your colleagues.

Parameters are numerical facts. While the parameter is out there, the problem is, we don't know its value. No one knows for certain, even if they happen to guess it correctly.

A **sample** is a set of observations drawn from the population. We can measure **sample statistics** or **estimators** from our sample, in contrast to measuring parameters.

We use the English alphabet to represent estimators. We symbolize the **sample mean** with $\bar{x}$, or "x-bar." We use the lower case s or s^2 when we're describing the **sample standard deviation** and the **sample variance** respectively.

Finally, when describing the number of observations in the population we use a capital N, but for the **sample size**, we use a lower-case n.

Inference
chance variability

Statistical inference seeks to answer the question: “How different is my sample data from the population?” The goal is to use imperfect information (our data) to infer facts, make predictions, and make decisions.

The model for **chance variability** is:

population parameter = estimator + chance error

Chance variability if uncertainty formalized in an equation.

The quantifiable amount of chance variability depends on the population distribution- if there were a 100% chance of getting the population parameter when sampling, then there would be no chance variability.

A more sophisticated equation disaggregates **chance error** further:

population parameter = estimator + sampling error + bias

The first cause of chance error is **sampling error**. It is a measure of randomness and uncertainty, which is expected in any random sample.

The second component of chance error is **systemic bias**. Like a weighed die, systemic bias influences all trials and samples. The key distinguishing feature of bias is that, unlike randomness, bias influences all observations in the same direction.

We’ll explore sampling error in this section and review bias later.

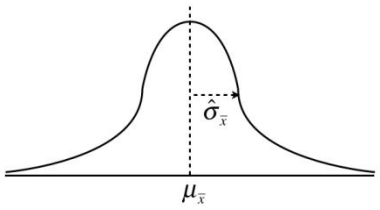
Inference
Sampling Distributions

Imagine you're sampling 30 women to measure their height. You record each of their individual heights, take the average, and plot it on the x-axis.

Then, you take a new sample of 30 different women, average their height, and plot it on the x-axis. You continue to do this for each new group of 30 women, again and again, infinity times.

The **sampling distribution** describes the distribution of infinitely many sample statistics of size *n* from the population. Sampling distributions model chance variability.

The sampling distribution shows the probability of estimating a sample statistic with a given value.



The **sampling distribution of sample means** is normally distributed with the parameters $N(\mu_{\bar{x}}, \sigma_{\bar{x}})$.

The mean of the sampling distribution of sample means, $\mu_{\bar{x}}$, is equal to ("is an unbiased estimator of") the population mean, μ, regardless of the shape of the population distribution.

The parameter that measures the dispersion of the sampling distribution of sample means is the standard error of the sample mean, $\sigma_{\bar{x}}$.

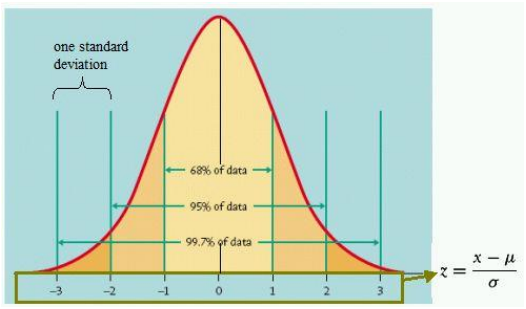
A **standardized sampling distribution** will be normally distributed with a mean of 0 and a standard deviation of 1.

Shapes

standard normal distribution

The standard normal distribution provides a common scale for comparing all variables that are normally distributed, regardless of the values of its observations.

The standard normal distribution is a normal distribution that has a mean of 0 and a standard deviation of 1, $N(0,1)$.



Any normal distribution can be standardized by converting its values into standard units (aka standardized scores, sigma scores, z-scores or z-statistics, z .)

Standardizing an observation amounts to measuring the number of standard deviations that the observation or sample is above or below the mean.

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\hat{\sigma}_{\bar{x}}} \, or \, \frac{\mu_{\bar{x}} - \bar{x}}{\hat{\sigma}_{\bar{x}}}$$

Standardizing the standard deviations implies that you divide the standard deviation by itself. So one standard deviation will be equal to a z-score of one, two standard deviations a z-score of two, and so on.

If you are given an area and want to find the test statistic, divide the area in half, then look it up on a z-table.

A p-value, P , is the probability of 'drawing' a given value (outcome, sample) by chance. We add the caveat "by chance" to cover our asses- we assume that there is no bias.

Based on the standard normal distribution above, we can infer that the probability of falling to the right of the mean by plus one standard deviation is P = 16% (or P = 32% on both).

The probability of falling to the right of the mean by 2 standard deviations is P = 2.5%, or P = 5% for both.

The probability of falling to the right of the mean by 3 standard deviations is P = 0.15%, or P = 0.3% for both.

A random variable X that has a normal distribution is abbreviated as $X \sim N(E(X), SE(X))$.

Shapes

probabilities on standard normal distributions

The standard normal distribution serves two practical purposes: given an observation, we can find the p-value of that observation occurring, or, given a probability, we can find the value of the observation.

These are two sides of the same coin: given x, find P(x), or given P(x), find x.

We begin with three values: our sample statistic, our (hypothesized) population parameter, and our standard error.

First, we normalize our sample statistic by subtracting our hypothesized population parameter from the sample mean and dividing by the standard error.

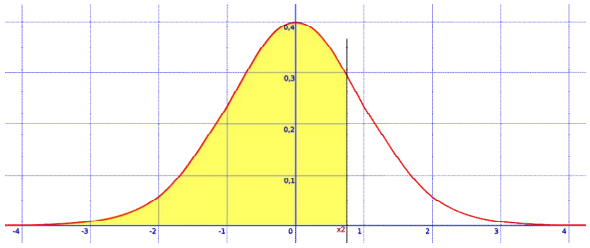
We call this value our **test statistic**. Depending on our sample size, it might be a z-statistic or a t-statistic. Let's assume that it's a z-stat.

We look up the test statistic's probability on a **z-table**, which is the intimidating table at the front of every stats book.

More generally, the z-table lists the percentage area under the standard normal distribution to the left of the test-statistic. The percentage can also be interpreted as a probability.

This is our p-value- the probability of getting a test statistic equal to or less than our test statistic. (There are variations on this theme- the probability of a test statistic equal to or greater than, or between, etc.)

Our test statistic, z, and its negative, -z, will fall somewhere on the standardized sampling distribution of sample means.



Conversely, if you're wondering about the probability of getting a z-statistic, or any value of a normally distributed random variable, simply find the probability on the z-table, then locate your z-statistic on the z-table.

Next, un-standardize your z-statistic by multiplying it by the standard deviation and adding your mean. (This is done less often.)

$$z = \frac{x_j - \mu}{\sigma} \rightarrow z * \sigma + \mu = x_j$$

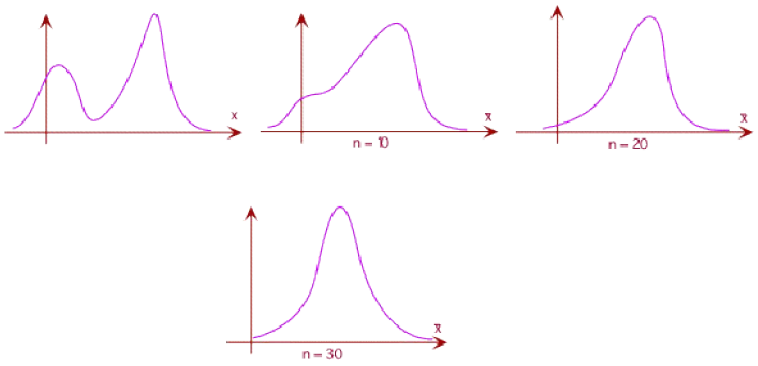
Notice that the equation of the curve is completely determined by the mean mu and the standard deviation sigma.

Inference
Central Limit Theorem

Of course, we don't need to, nor can we, build a sampling distribution for infinitely many sample statistics. But a verifiable theory tells us what would happen if we did.

The central limit theorem states that, if you plot an infinite amount of samples from the population, the sampling distribution will approach a normal distribution, *even if the population is not normally distributed*. This allows us to perform inference on non-normally distributed populations.

How large must n be in order to have a normally-distributed sampling distribution of sample means? The shape of the sampling distribution is normal if either n > 30 or population distribution is normal. (This may be higher depending on the skew of the population.)



Inference
sampling error and standard error

Recall the model for chance variability:

population parameter = estimator + sampling error + bias

According to the equation above, the **sampling error** is the difference between the population parameter and the estimator, ignoring sampling bias. This difference is due to chance. Even when spinning a color wheel, you won't get a perfectly uniform distribution.

The sampling error mainly depends on the size of the sample relative to the population. Recall that according to the weak definition of the law of large numbers, sampling error decreases as sample size increases.

The sampling error is the standard deviation of the sampling distribution, $\sigma_{\bar{x}}$. It depends on the variance in the population- as the variance increases, the sampling error increases.

The **standard error** is the statistical estimate of the sampling error.

The **standard error for the mean** is the population sample standard deviation divided by the square root of the sample size.

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The **standard error of a proportion** follows the same formula, alternatively, $\sigma_{\hat{p}} = \sqrt{pq} / \sqrt{n}$.

As you can see, the standard error is always less than the standard deviation. This is because the domain of the sample is always less than the domain of the population.

Because we don't know the population standard deviation, we generally calculate the **estimated standard error**, $\hat{\sigma}_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \hat{\sigma}_{\bar{x}}$$

It helps to think of the standard error as being analogous to the standard deviation of a sample. While the standard deviation describes how far a typical observation will be from the average, the standard error measures how far a typical sample will be from the population parameter.

Inference
random variables and probability distributions

A **random** or **stochastic variable** is a variable whose value is a numerical outcome of a random phenomenon. We denote a random variable with upper-case capital letters at the upper end of the alphabet, X, Y, Z.

A **probability distribution** is a description of the sample space in terms of a random variable. It assumes that we know all of the outcomes and their probability.

Value of X	x_1	x_2	x_3	x_k
Probability	p_1	p_2	p_3	p_k

A **probability histogram** gives the probability of each possible value of the random quantity in question. The Y-axis is a decimal. The area of a probability histogram or probability density function is 100%, 1 or **unity**.

We call the center of the distribution the **expected value**, $E(X)$ or the **expectation of X**. The expected value of a random variable is essentially the mean of the random variable. The formula for the expected value is

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

Note that the expected value may not be included in the set of outcomes.

The **standard deviation** on a probability distribution is similar to our prior formula, except now it substitutes the probability for the sample size.

$$\sigma = \sum_{i=1}^{k} (x_i - E(X)) * p_i$$

Probability density functions describe the distribution of continuous random variables. Probability density functions also displayed on probability density curves, which are smooth measurements of a probability histogram.

A probability density function describes the outcomes within a range- below or above a single value, or between two values. A spinning top never falls on the two same points. The probability of each outcome is 0.

A probability density function is a smooth approximation to the irregular bars of a probability histogram. It is an *idealized* distribution because the probabilities are based on the theoretical long-run, ignoring short-run irregularities and outliers.

Inference

degrees of freedom, d.f.

Say we're trying to calculate the average height of a population. Because we don't know the population parameters, we'll carry out statistical inference by calculating two statistics: the sample mean, which represents our best guess, and the standard error, which represents our uncertainty.

The degrees of freedom, d.f., is equal to the number of independent statistics used to estimate another sample statistic in a sample. We usually subtract this value from the sample size.

Let's use the sample variance as an example. The formula for the population variance is

$$\sigma^2 = \frac{\sum_{j=1}^{N}(x_j - \mu)^2}{N}$$

For the sample variance, the degrees of freedom is $n-1$ because we include another sample statistic, the sample mean, in the formula for the sample variance:

$$s^2 = \frac{\sum_{j=1}^{N}(x_j - \bar{x})^2}{n-1}$$

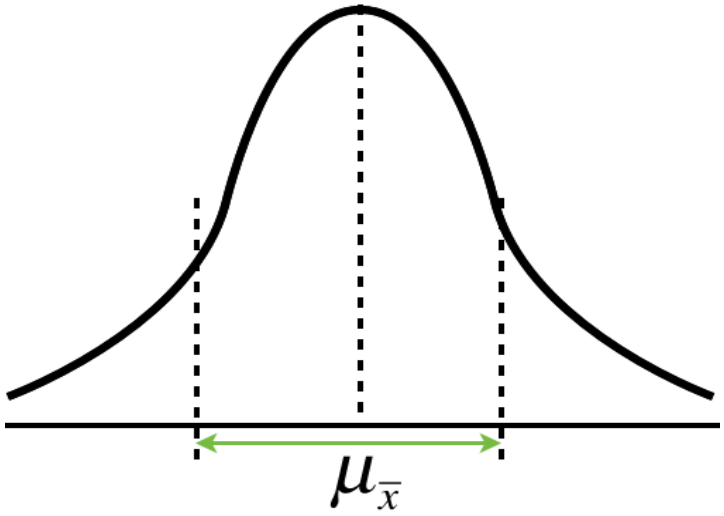
You'll see the degrees of freedom come up for various formulas. Don't worry about why. The short version is that it eliminates an inherent bias in sampling.

Des stats- div by n
Infer- div by n-1

Confidence Intervals
introduction

The fundamental question of inference is: is the sample representative of the population? An alternative to "point estimates" interval estimates. Neither is more accurate.

Rephrased, statistically: Within what range on the x-axis of the sampling distribution will we consider a sample mean to be representative mean of the sampling distribution of sample means (which is an unbiased estimator of the population parameter)?



an interval with the mean of the sampling distribution of sample means at its center

This range is called a **confidence interval**. It extends out in both directions and equal distances from the mean of the sampling distribution of sample means.

Any sample statistic that falls within the range of the confidence interval is considered representative of the population. If the sample statistic is outside of the confidence interval, it is out of bounds, so to speak, and is not considered representative of the population.

We will begin with a theoretical model of the confidence interval, then address constructing confidence intervals using sample data.

We know we have the sampling error, so let's be honest. Sampling error implies each point estimate will be off.

Confidence Intervals
Re-defined

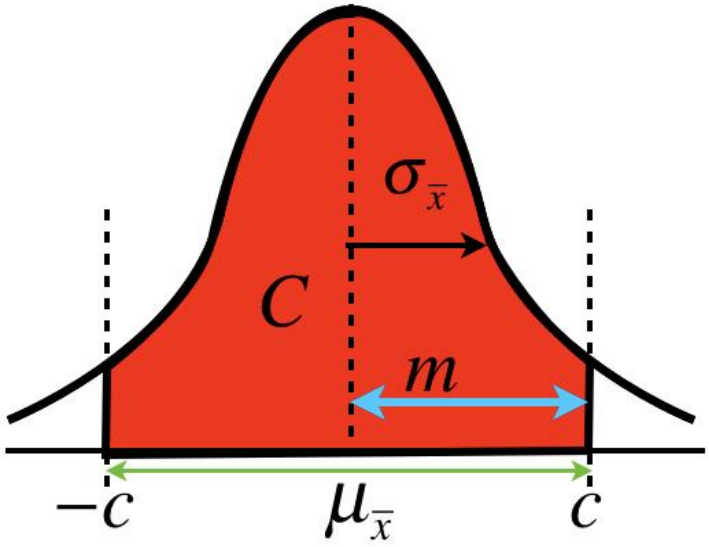
Let's redefine confidence intervals as a theoretical model. There are three parts.

$$\mu_{\bar{x}} \pm \sigma_{\bar{x}} * c \quad t \pm c * SE(t)$$

$\mu_{\bar{x}}$ is simply the mean of the sampling distribution of sample means. (Although we don't know the value of $\mu_{\bar{x}}$, pretend we know it for now.)

The second part of the equation is the **standard error**, $\sigma_{\bar{x}}$. We don't know this value either.

The third part of the equation is our **confidence level**, C. C is our level of confidence (higher is better) and it is also the area under the sampling distribution and within the confidence interval.



The confidence level doesn't factor directly into the formula. We're interested in the range on the x-axis of the sampling distribution that contains C percentage of the observations.

In practice, we choose our confidence level C, which will determine the **critical values** or **confidence coefficient**, *c*. These are the upper and lower bounds on the x-axis of the standardized normal distribution containing area C.

The **margin of error**, *m*, is the interval equal to the product of the critical values and the standard error. Think of the margin of error as the standard error "stretched" by the critical values so that the standard error covers ½ * C area of the sampling distribution.

$$m = c * \sigma_{\bar{x}}$$

Confidence Intervals
confidence intervals with sample data

We usually don't know the mean of the sampling distribution of sample means, $\mu_{\bar{x}}$, nor do we know the standard error, so we can't build a confidence interval for the population mean.

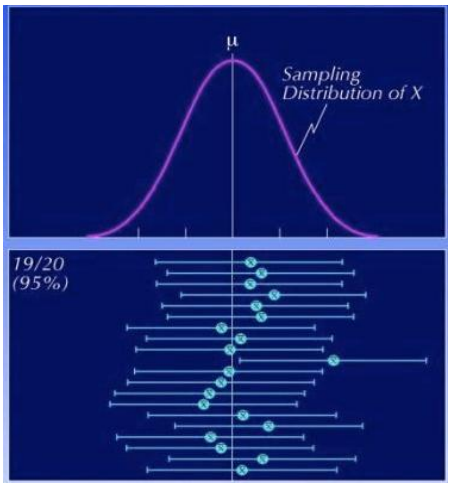
However, using statistical inference, we can compare the theoretical model to the confidence interval for the sample mean:

$$\bar{x} \pm \hat{\sigma}_{\bar{x}} * c$$

where $\bar{x}$ is the sample mean and $\hat{\sigma}_{\bar{x}}$ is the estimated standard error.

Upon first impression, this seems strange. However, it is just a subtle shift in perspective that invokes the frequency theory of probability.

Say we collect 20 sample means and place each sample statistic at the center of our margin of error with a confidence level of 95%.



Just as we can expect 19/20 of our sample means to fall within the 95% confidence interval for the population mean, we can also expect the population mean to fall within 19/20 of the confidence intervals for the sample mean. [1]

Of course, our probabilities are not exact because they are based on sample data. This is why we invoke the term "confidence."

Remember: You can't say that there is a 95% chance that the population parameter will fall within the confidence interval for the sample- it either falls in the interval or it doesn't! Instead, you must say that you are 95% confident that the population parameter is within a given sample mean interval.

---

[1] More generally, if there is an X% probability that A is Y-distance from B, then there is also an X% probability that B is Y-distance from A.

Confidence Intervals
trade-offs

Ideally, we would like a high confidence level and a small margin of error. Higher confidence levels imply greater certainty. A smaller margin of error implies more accuracy.

But we can't have it both ways. There is an inherent trade-off between confidence levels and confidence intervals: the larger the confidence level, the larger the margin of error.

$$m = \pm c * \frac{s}{\sqrt{n}}$$

Use can rearrange this formula to optimize confidence and sample size. To calculate the To obtain a desired margin of error m :

$$n = \left(\frac{c * s}{m} \right)^2$$

Tests of Significance
intro

The purpose of a **significance test** is to determine the probability of "drawing" a sample's value in a random, independent draw from the population.

Significant in the statistical sense does not mean "important." It means, "not likely to happen by chance."

When we carry out a significance test, we measure the distance between the sample statistic (called a **test statistic** in this context) and the mean of the sampling distribution of sample means on the x-axis of the sampling distribution.

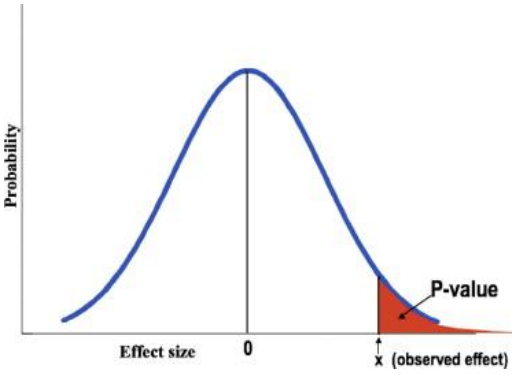
Using various statistical models, we can measure a test statistic's **p-value**[1]. The p-value is the probability that any given test statistic will occur by chance.

Graphically, the p-value is the area under the sampling distribution and to the right of the test statistic.

The p-value is equal to the probability of a test statistic's occurrence given that the null is true…

$$p = P(data \mid H_0 = true)$$

…*as well as* the probability of drawing a sample statistic whose value exceeds that of the test statistic.



The closer the test statistic is to the mean of the sampling distribution of sample means, the greater the p-value. The further the test statistic is from mean of the sampling distribution of sample means, the smaller the p-value.

If a test statistic is far from the mean of the sampling distribution of sample means, this implies that something other than chance variability is at play.

While p-values can have any value, the standard for a significant result is the **significance level**, α. Alpha-values are also p-values, but they are also strictly defined thresholds. There is no sharp border between "significant" and "insignificant," meaning significance tests can be arbitrary. Significance levels are usually $\alpha = 1\%$, $\alpha = 5\%$ or $\alpha = 10\%$.

Any p-value less than 5% is considered a **statistically significant** result. Any p-value less than 1% is considered **highly significant**.

[1] aka **observed significance level**

Tests of Significance
one vs. two-tailed tests

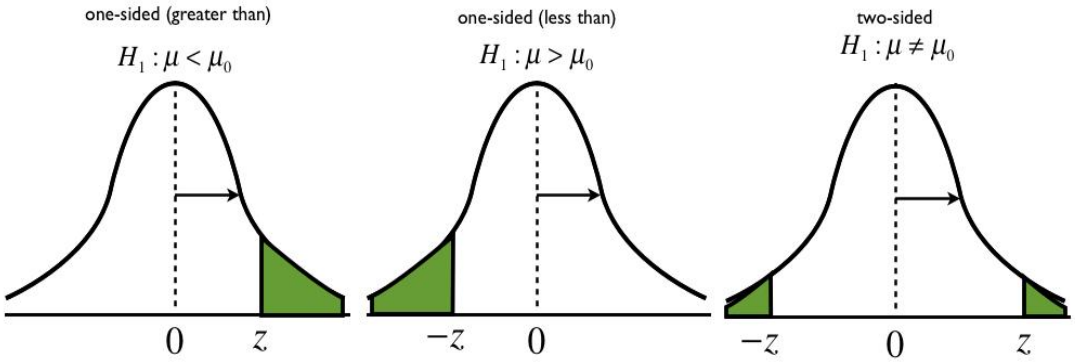
When carrying out a significance or hypothesis test, we can choose between a one-tailed test and a two-tailed test. It depends entirely on context.

A directional test is a one-tailed test. A non-directional test is two-tailed.

In the introduction to this section, we described a **one-tailed test**, where we measure the p-value on one side of the distribution. One tailed tests define the mean of the sampling distribution as either greater than or less than the population mean.

For example, if we're only interested in whether a drug *increases* serotonin levels in the brain then we would conduct a one-tailed hypothesis test.

A **two-tailed test** makes the assumption the mean of the sampling distribution can be either greater than or less than the population parameter. For example, if we're looking at a drug's effect on serotonin levels, the drug could potentially increase or decrease the patient's serotonin levels.



This implies that the critical value for a one-sided test is less (in absolute terms) than for a two-sided test. Therefore, two sided tests require more extreme critical test statistics for a given level of significance.

Some regression packages only compute p-values for two sided alternatives. But it is simple to obtain the one-sided p-value: just divide the two-sided p-value by 2.

Tests of Significance
hypothesis tests

A **hypothesis test** is a particular kind of significance test that is inferential in nature. While we may know the value of the population parameter with a significance test, we do not know it with a hypothesis test.

A hypothesis test asks, "How likely is it that my sample is from the population, if the population parameter is equal to X?"

For a hypothesis test, we compare our sample statistic to a hypothesized population parameter in order to test the hypothesis that our population parameter takes such-and-such a value.

We begin our hypothesis test by declaring our **null hypothesis** H_0 that a hypothesized population parameter is equal to the mean of the sampling distribution of sample means.

$$H_0 : \mu = \mu_0$$

The null hypothesis ("H-nought") is a statement about the population parameter, not the sample statistic.

Next, we calculate the test statistic with reference to the null hypothesis, thus assuming that the null hypothesis is true.

Finally, we calculate the p-value on the appropriate sampling distribution. The p-value is the probability of observing a t-statistic as extreme as we did if the null hypothesis is true. On the basis of the p-value, we then either reject the null hypothesis or fail to reject it.

A hypothesis test is a logical proof that works by eliminating one or more alternatives. It is a proof by contradiction. If the population parameter isn't A, then it may be B.

In this case, B is our **alternative hypothesis**, H_1 or H_a, which simply states the opposite of the null hypothesis.

$$H_1 : \mu \neq \mu_0$$

If we reject the null hypothesis, this does imply that we accept the alternative hypothesis. We can never accept a hypothesis as true because we don't know the population parameter.[2] We can only study other samples and continue to reject or fail to reject the null hypothesis.

We should not draw any conclusion based on a single significance test. We call our results **robust** when our test statistics remain significant while testing multiple samples under varying assumptions.

---

[2] We're assuming the truth of our null hypothesis and our test statistic in order to make an inference about a population. This is why we use the term "p-value" rather than "probability." Population parameters do not occur with probability. We're pretending they do, so we say "p-value."

Tests of Significance
type I and type 2 errors

Truth / Decision	Retain H0	Reject H0
H0 is true	Correct decision	**Type 1 error (false alarm)**
H0 is false	**Type II Error (miss)**	Correct decision

To commit a **type 1 error** is to reject the null hypothesis when it is, in fact, correct.

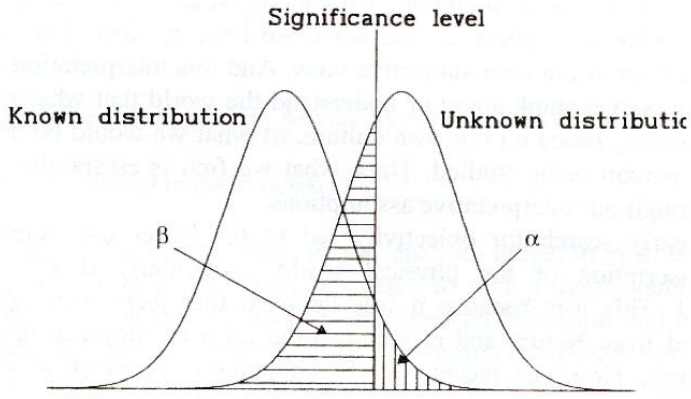
The probability of a type 1 error is equal to α. If α=5%, we should expect 5% of the sample statistics to fall outside of the 95% confidence interval for the hypothesized population parameter.

We minimize the probability of making a type 1 error by choosing an extreme significance level.

The probably of a type 1 error increases when researches conduct multiple NHST's, especially with a dependent date. (Replicate significant effects. ROBUST.)

To commit a **type 2 error,** β, is to mistakenly accept the null hypothesis when the alternative hypothesis is true. We can avoid this problem altogether by never accepting the alternative hypothesis, which is standard practice.

Measuring β involves drawing two sampling distributions on the same x-axis- one sampling distribution for the null hypothesis and another for the alternative hypothesis.



Confusingly enough, we attempt to *minimize* the probability of committing a type 2 error by *maximizing* the power. The **power**, $1-\beta$, is the probability of correctly rejecting the null hypothesis.

There are several ways to increase the power. (80% power is standard.) First, you can increase alpha, but this increases the chance of making a type 1 error. Second, you can choose an alternative hypothesis further from the null, but decreases the strength of the underlying hypothesis test. Third, you can increase the sample size, which may be costly.

Tests of Significance
Problems and remedies

All NHST are biased by sample size. The explanation is the following: A large N leads to a small standard error, since the standard error has *n* in the denominator. This leads to a high t-value and a low p-value.

So along with reporting the results of the NHST, also report the effect size and p-value. Also, reporting standard errors as well.

Yokel Local Test- It's all people know → weak hypothesis testing. Learn other forms of hypothesis testing. Consider multiple alternative hypothesis and a model comparison.

Shady logic- Modus tollens: if p then q. Not q. Therefore P. This is a logical fallacy.

If the null is true, the data can not occur. The data occurred. Therefore, the null is false.

For example, if the pop is healthy, then no one sample should have a high body temp. I obtained a high sample mean value. Therefore the population is not healthy.

Chance Models
introduction

So far in our discussion of inference, we have been using the **Gaussian** model. This model applies to a single quantitative variable where the population has a normal distribution. This variable's test statistic is a z-statistic.

There are actually many statistical models, called **chance models**. They have the same features we've discussed with the Gaussian model, such as having a distribution and a test statistic.

We will begin by exploring the t-distribution, which is very similar to the Gaussian model.

After that, we will explore models where the variables may be categorical. The table below outlines the chance models we will be exploring in this section:

X/Y	Quantitative	Categorical
Quantitative	Gaussian	logit
Categorical	ANOVA	Chi-Square

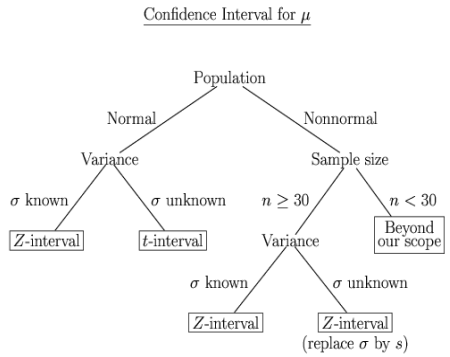
Finally, we will conclude by examining several other distributions: the binomial, the two-t and the the poisson.

t
introduction

We generally use the z-statistic under two circumstances:

1. The population is normally distributed and the population standard deviation is known
2. The population has a non-normal distribution, the variance is unknown and n>30 (CLT)

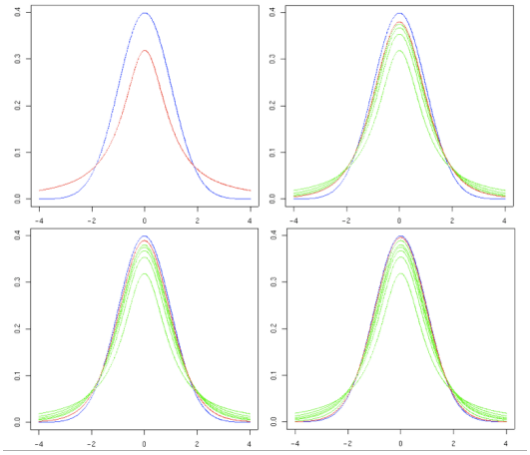
If the population is normal and we do not know the population standard deviation, then we use the t-statistic. (The T-statistic is also used on small sample comparison tests- more later.)



The key assumption in using the t-statistic is that the population is normally distributed. This assumption can be tested by looking at the sample data.

You can do this if the distribution of that sample is roughly normal. Stem and leaf plots are particularly useful for confirming these assumptions among smaller samples.

The t-distribution is similar to the z-distribution in that it is symmetrical and bell-shaped. Unlike the normal distribution, however, there are actually many t-distributions.



In the image above, the blue curve is the normal distribution. The orange curve is the t-distribution for 1, 5, 10 and 30 degrees of freedom.

Although the t-distribution never exactly equals the normal distribution, they're approximately equal beyond a df of 30 or 40.

The degrees of freedom is the single parameter that specifies the particular t-distribution, *k=n-1*. We abbreviate a particular t-distribution with k degrees of freedom as $t(k)$.

As the pictures above show, t-distribution contains slightly more area in the tails than the normal distribution. These "fat tails" imply that we're slightly more likely to get an extreme sample because our sample size is small. There is more room for extreme variability.

t
inference with a single t-statistic

The t-statistic is the test-statistic of the t-distribution. We use the t-statistic to create confidence intervals and carry out hypothesis tests.

To construct a **confidence interval for a sample**:

$$\overline{x} \pm t_{\alpha/2} * \hat{\sigma}_{\overline{x}}$$

Where $\hat{\sigma}_{\overline{x}} = s / \sqrt{n}$.

Note that the **t-statistic has n-1 degrees of freedom**.

We use the t-statistic for confidence intervals under the same assumptions as those for a hypothesis test. Namely, if we don't know the population standard deviation but we can assume that the population is normally distributed, we use a t-test. If we know sigma, or have a sufficiently large sample size on a non-normal population (n>30), then we can use z.

To perform a **hypothesis test**,

$H_0 : \mu = \mu_o$	$H_1 : \mu \neq \mu_o$	$\lvert t \rvert \geq t_{\alpha/2}$	$p = 2P(t \geq \lvert computed\ t \rvert)$
$H_0 : \mu \leq \mu_o$	$H_1 : \mu > \mu_o$	$t \geq t_{\alpha}$	$p = P(t \geq computed\ t)$
$H_0 : \mu \geq \mu_o$	$H_1 : \mu < \mu_o$	$t \leq -t_{\alpha}$	$p = P(t \leq computed\ t)$

To calculate the **t-statistic**, divide the sample mean by the estimated standard error:

$$t = \frac{(\overline{x} - \mu_0)}{\hat{\sigma}_{\overline{x}}}$$

Where $\hat{\sigma}_{\overline{x}} = s / \sqrt{n}$ and μ_0 is the hypothesized population parameter.

On the statistical table, the t-statistic will be given within bounds of alpha. So after you calculate the value of your t-statistic, you'll have to see where it falls between $t_{\alpha_1} < t < t_{\alpha_2}$ for a given degree of freedom.

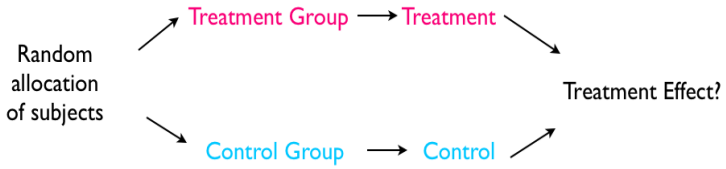
For testing 2 variances, see the next chapter, "comparing means."

Two-Sample Statistics

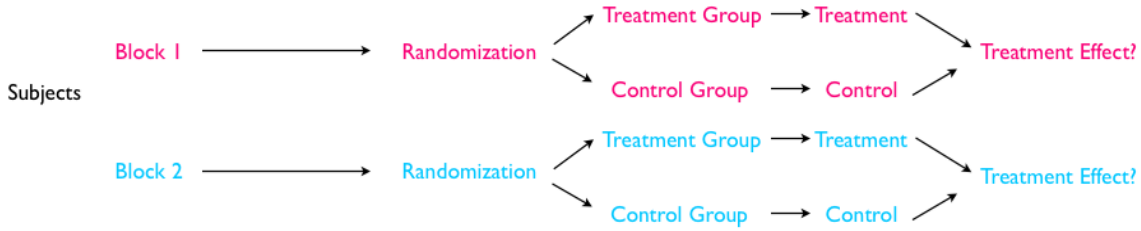
methods of comparison

A comparative experiment has three methods of comparison: completely randomized design, randomized block design and matched pairs design.

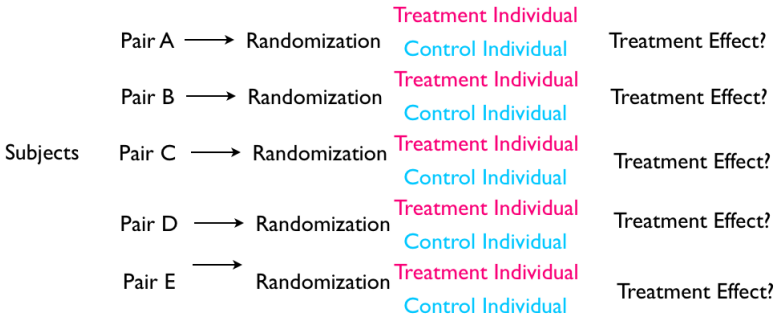
Completely randomized comparative experiments sort experimental units randomly among all treatments, just like the case with the doctor and his or her patients.



In a **randomized block design**, the experimenter first matches individuals into classes based on a third variable. Then, the subjects are randomly assigned treatments within each class. This allows a class-by-class comparison of the treatment.



Matched pairs design matches the subjects in pairs on the basis of having the same characteristic in common. A treatment is then applied to an individual in each pair.



Introduction

When we are comparing two population parameters, we compare either the mean or the standard deviation.

For two sample means, we can take two different approaches. The first method is called **independent sample comparison**. The second method is called a **paired comparison.**

The key criterion of deciding which approach to use depends on the independence of the data. If each the subgroups are independent, then you start with the independent sample comparison. If the subgroups are matched (such as before and after, placebo vs. treatment applied to twins, etc.), then you do a paired comparison. In this case, the subgroups are not dependent, but the difference between the subgroups is.

There are three key assumptions that must be met in performing inference. The first is independence. This isn't as cut and dry as it sounds. The **cluster effect** states that any population, even if divided into subgroups, will still have some commonalities. For example, a classroom of 50 students that are treated with 5 different tutoring techniques will still have a common teacher. **Serial correlation** and **spatial correlation** imply that observations are less independent when they're gathered in similar time or space intervals. All-in-all, these effects may not ruin the assumption of independence but they can lead to a higher type I error. As these effects mount, consider doing a paired test instead of an independent sample comparison.

The second is normality for each subgroup. If the independent sample data is not normally distributed, you can check the difference and perform a paired comparison. If the data still is not normally distributed, then you would perform a rank-test. In estimating normality, note that Shapiro-Wilk is just a numerical representation of the distribution- look also at the graph and test the power.

The third assumption is homogeneity of variance. Regardless of approach, each subgroup must have approximately equal variance. If the variances are approximately equal, then you can use a pooled standard deviation. If the variances are less than approximately equal, you can perform a Satterwhaite test and include both standard errors in the analysis.

To pool or not to pool the variance? In SAS, just run the F-test for homogeneity of variance. Note that the F-test is highly sensitive to non-normal distributions. If they're not equal, use Satterwhaite. Otherwise, pool the variance.

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$ and df=n1 +n2 -2

The pooled variance is a weighted average of the sample variances.

If the variances are really not equal, you can perform a log transformation of the data and try again. If the variances are still not equal, even after a transformation, then the assumption fails.

Mean Comparison
Independent Samples, pooled variance

In the case of independent samples, you can either pool the variance of not.

If the variance is pooled, the **confidence interval** is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} * s_p * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The **hypothesis test for independent samples** is:

H0	HA	RR
$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$t \geq t_{\alpha, n_1 - n_2 - 2}$
$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$t \leq -t_{\alpha, n_1 - n_2 - 2}$
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$\lvert t \rvert \geq t_{\alpha/2, n_1 - n_2 - 2}$

The test statistic in this case is

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Note that D_0 is often 0 but it doesn't have to be.

Note also that the SAS computes the difference in means automatically (alphabetically by variable), meaning you need to be consistent in how you set up the problem. If you get a negative confidence interval for the difference, watch out. If $D_0 \neq 0$, then you may have to make it positive or negative.

Note also that SAS reports the p-value for a two-tailed test, so you'll have to divide by two for the one-sided p-value.

Mean Comparison
Independent Samples, unequal variance

If you don't pool the variance, you can use this formula for the test statistic.

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In this case, the degrees of freedom are below. If the degrees of freedom are not an integer, round down.

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1-c)^2(n_1 - 1) + c^2(n_2 - 1)}, \quad c = \frac{s_1^2 / n_1}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Mean Comparison
Paired data

The hypothesis test:

H0	HA	RR
$\mu_d \leq 0$	$\mu_d > 0$	$t \geq t_{\alpha,n-1}$
$\mu_d \geq 0$	$\mu_d < 0$	$t \leq -t_{\alpha,n-1}$
$\mu_d = 0$	$\mu_d \neq 0$	$\lvert t \rvert \geq t_{\alpha/2,n-1}$

With test statistic:

$$t^* = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$$

And confidence interval:

$$\bar{d} \pm t_{\alpha/2} * \frac{s_d}{\sqrt{n}}$$

Where n is the number of differences.

Two-Sample Statistics

Introduction

When we're comparing two samples to determine whether they are significantly different, then we model the distribution of two-sample statistics, x-bar and y-bar.

A comparative experiment uses the two-sample statistical model. We'll use a clinical trial as an example throughout this section.

In clinical trial, we attempt to measure the difference in outcomes between two groups, the **treatment group** and the **placebo group**.

A **treatment** is a potential cause, such as a pill, whose effect we attempt to measure in an experiment.

The **effect** of any treatment is twofold: the **treatment effect** and the **placebo effect**.

The treatment effect is the actual effect of the treatment. It is the effect we wish to measure, such as the chemical effect of a pill on a subject's body.

The placebo effect is the observed effect that is not due to the treatment effect. We can **control** for the placebo effect by giving the other group a sugar pill. While the placebo may be effective, it is not what we are measuring.

This experiment is called an **independent t-test**, since the samples are independent of one another. If we were comparing two related samples (such as a "pre/post" scenario), we would be carrying out a dependent t-test.

Two-Sample Statistics

The model: the difference between two-sample statistics

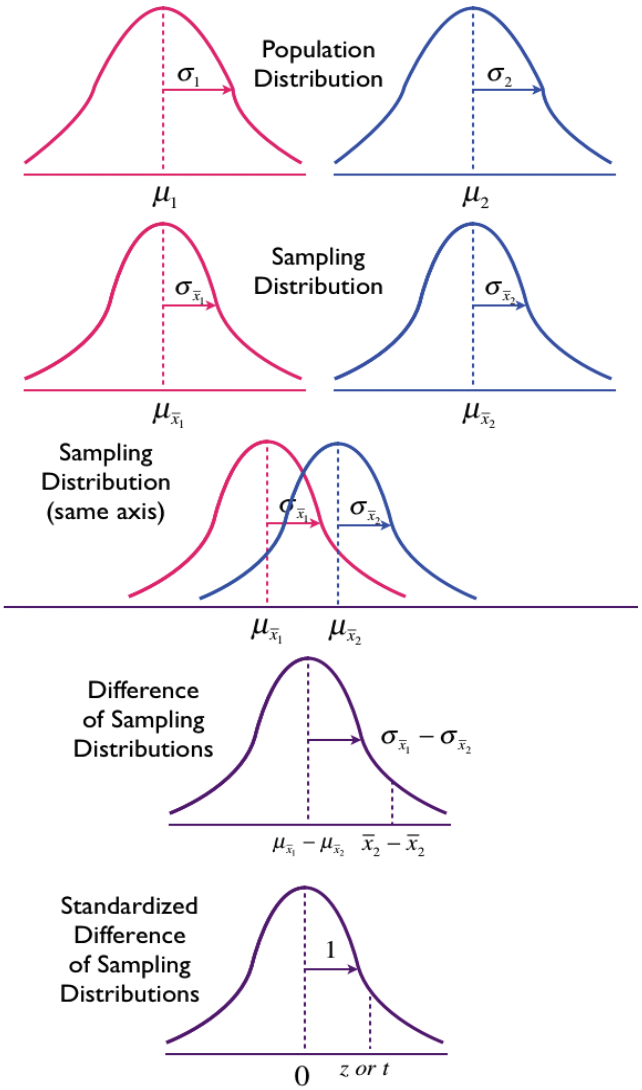
The trick to constructing an independent two-sample model is to make some simplifying assumptions and then combine the two parameters into a single parameter. Once we have a single parameter, then we can carry out a standard t-test.

When we compare two population parameters, we combine them into a single value by taking their difference.

$$\mu_{x-y} = \mu_x - \mu_y$$

The mean of the sampling distribution of the difference of two-sample statistics is simply the difference in sampling means.

$$\mu_{\bar{x}-\bar{y}} = \mu_{\bar{x}} - \mu_{\bar{y}}$$



Two-Sample Statistics

the standard error, unpooled and pooled

While we can take the difference of means, taking the difference between two standard errors is slightly more complex.

By the **rule for variances**, the **difference of two estimated standard errors** is

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Alternatively, we can use a **pooled estimate**, specifically, the **pooled estimate for the sample standard deviation**, s_p

.

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

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

This is called a pooled estimate because we combine 2 estimators, s_1 and s_2 into a single estimate.

If you're not sure about whether or not the pooled estimator can be used, test the assumption that the population standard deviations for both groups are equal using an F-test.

Two-Sample Statistics

standardized sampling distributions

Putting it altogether, the test statistic for the difference of two-sample statistics is

$$\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{(\sigma_{\bar{X}}-\sigma_{\bar{Y}})}=z\ or\ t$$

Keep the sign of the test statistic in mind!

The **two-sample z-statistic** is normally distributed, with an $N(0,1)$ sampling distribution.

If you are using a **two-sample t-statistic**, *t(k)*, there are two methods for measuring the degrees of freedom.

The conservative method:

$$k=(n_1+1)+(n_2+1)$$

The more accurate estimate is to use a **pooled estimate of the degrees of freedom**.

$$k=(\sigma_{\bar{X}}+\sigma_{\bar{Y}})*\left(\frac{n_1-1}{\sigma_{\bar{X}}^2}+\frac{n_2-1}{\sigma_{\bar{Y}}^2}\right)$$

The downside of the more accurate estimate is that the degrees of freedom is probably not reported as an integer, so you can't use a t-table.

Two-Sample Statistics

hypothesis testing

Our null hypothesis is that there is no difference between the groups: $(\mu_X - \mu_Y) = 0$.

Therefore, the equation of the test statistic becomes

$$\frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{(\sigma_{\bar{X}} - \sigma_{\bar{Y}})} \rightarrow \frac{(\bar{x} - \bar{y})}{(\sigma_{\bar{X}} - \sigma_{\bar{Y}})}$$

Our null hypothesis the null hypothesis is

$$H_0 = \mu_x - \mu_y = 0 \rightarrow \mu_x = \mu_y$$

Confidence Intervals?

Testing a Single Population Variance (chapter 7)

Recall the formula for the sample variance:

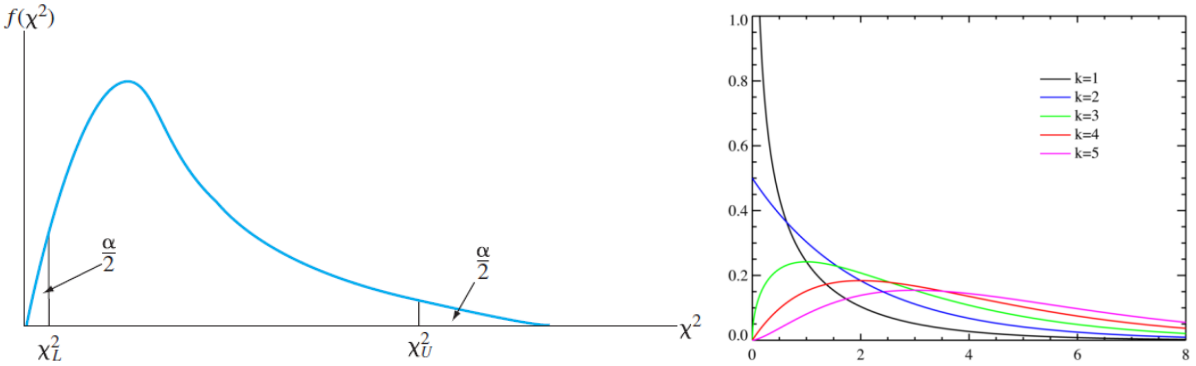
$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

The sample variance is an unbiased estimator of σ^2 (if the sample is random).

The sampling distribution of s^2 follows the chi-square distribution with df=n-1.

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

When performing inference on the variance, **the key assumption is that the population is normally distributed**, even if the sample size is large**.** If a box plot or normal plot of the sample data shows substantial skewness or a large number of outliers, you can't use the chi-square based inference procedure.



The **confidence intervals for the population variance** with df = n-1 and confidence level $1-\alpha$ is

$$\frac{(n-1)s^2}{\chi^2_{df,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{df,1-\alpha/2}}$$

Note that we have to take the square root to get the confidence interval for the population standard deviation.

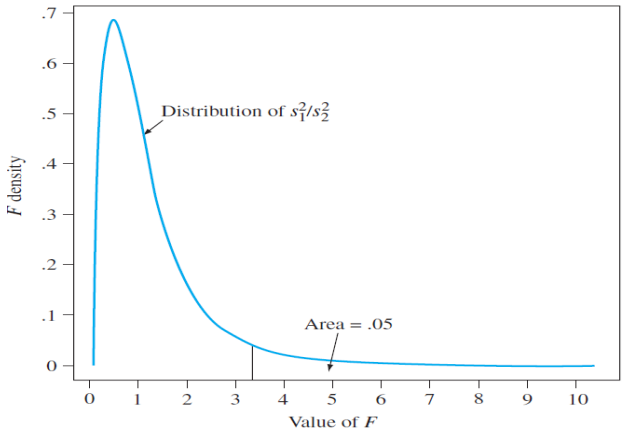
A **hypothesis test for the population variance** with test statistic $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ is

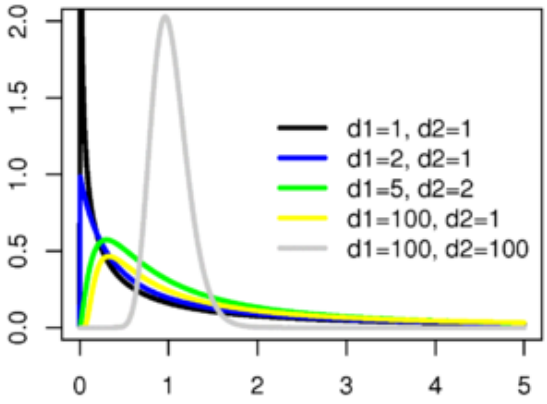
H0	Ha	RR
$\sigma^2 \le \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2_{df,\alpha}$
$\sigma^2 \ge \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2_{df,1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	$\chi^2 > \chi^2_{df,\alpha}$ or $\chi^2 < \chi^2_{df,1-\alpha}$

Testing Two Population Variances

Again, we have to assume that the two populations being tested are normally distributed.

A statistical test comparing σ_1^2 and σ_2^2 utilizes the test statistic s_1^2 / s_2^2 which fits the F-distribution. There are many **f sampling distributions**, one for each combination of the two parameters with DF1=n1-1 and DF2= n2-1.





Note that when $\sigma_1^2 = \sigma_2^2$, σ_1^2 / σ_2^2=1.

The **hypothesis test for equal variances** uses the test statistic is $F = s_1^2 / s_2^2$:

H0	Ha	RR
$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F \leq F_{1-\alpha/2,df_1,df_2}$
$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F > F_{\alpha,df_1,df_2}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F \geq F_{\alpha/2,df_1,df_2}$ or $F \leq F_{1-\alpha/2,df_1,df_2}$

It is important to note that $F_{1-\alpha/2,df_1,df_2} = \dfrac{1}{F_{\alpha/2,df_2,df_1}}$. Note that it's df2, df1 in the denominator.

Note also that the DF in the F-statistic subscript are always in the other df1, df2 referencing the table. (Ex. 7.7)

DF1 represents the degrees of freedom for the "between samples" or "treatment sum of squares" (t-1). DF1 represents the degrees of freedom for "within samples" or the "error sum of squares" n-t.

The **confidence interval for comparing population variances** is:

$$\left(\frac{s_1^2}{s_2^2} F_{1-\alpha/2,df_2,df_1} \leq \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} F_{\alpha/2,df_2,df_1}\right)$$

Note that this is the confidence interval for the ratio of the variances. Generally, confidence intervals are built for each variance independently using the chi-square statistic.

When comparing two variances, one typically looks at CI's for each individual population variance and performs a hypothesis test for equal variances. Note that the CI for the individual population variances will tell you whether they're significantly different depending on whether the upper and lower bound of each variance overlap.

Chi-Square
Chi square distribution

More generally, we can derive the chi-square statistic from the standardized test statistic, *z*. We can square a z-statistic and "make it Greek."

$$z = \frac{(X-\mu)}{\sigma} \rightarrow z^2 = \frac{(X-\mu)^2}{\sigma^2} = \chi^2$$

The chi-square statistic is the variance, normalized. The numerator is the total sum of squares, which is normalized by the population variance.

As the number of degrees of freedom increase, the **chi-square distribution** becomes more and more normal. The sampling distribution of the chi-square statistic with one degree of freedom turns out to be the distribution of the variance for the standardized normal distribution.

At $\chi_c^2 = 0$, the variance is 0. For most of the curves, most of the area is between $\chi_c^2 = 0$ and $\chi_c^2 = 1$.

The chi-square distribution table provides critical chi-square statistics at varying significance levels.

Chi Square Distribution Table

d.f.	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$	$\chi^2_{.001}$
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5

When comparing 2 or more distributions:

$$H_0 : \sigma_1 = \sigma_2 = \sigma_3$$

$$H_A : notallpop.stdde\,\mathrm{v.ar}\,e.equal$$

BFL: Both "BF" and "L" use F and have similar results. The Barlett uses the chi-square.

Test against L $\geq F_{\alpha,df1,df2}$ where $df1 = t - 1$ (t= the number of groups) and $df2 = N - t$

Chi-Square
Caveats

The weakness of the chi-square test is that it is broad. If the test allows us to reject the null hypothesis that all the proportions are equal, we then want to do a follow-up analysis that examines the difference in detail.

The chi-square statistic itself is additive for multiple chi-square statistics with varying degrees of freedom:

$$\chi^2_{(\nu_1+\nu_2)} = \chi^2_{(\nu_1)} + \chi^2_{(\nu_2)} = z_1^2 + z_2^2$$

For multiple samples, we add z-squared statistics, which have already been normalized. Each squared z-statistic represents the variance for a single observation. Summing multiple squared z-statistics gives us the total sum of squares.

The number of degrees of freedom is an additive function of the number of z-statistics.

The expected value of the chi-square distribution is the number of degrees of freedom.

The variance of the chi-square distribution is 2df, and the standard deviation is the square root of 2df.

ANOVA
introduction

ANOVA (ANalysis Of VAriance) is fundamentally about comparing different subgroups. Typically, a treatment is applied to each **experimental unit.** The level of the treatment determines which subgroup the unit falls into. The treatment/subgroup (two terms which are analytically identical) is the independent variable which is by definition categorical- there are as many subgroups as there are **levels** of the treatment. We measure the effect of the treatment on the **measurement unit.** This is the dependent variable, which always quantitative.

A factor is selected by the researcher for comparison. It must be categorical.

A response variable is a measure of treatment that is not controlled.

Below is what a typical data set would look like for the purposes of statistical analysis. Often we'll have to clean the data in order for it to take this form.

Obs. #	Treatment Group (X)	Outcome (Y)
1	1	36
2	2	29
3	1	34

There are several kinds of ANOVA. If there is only one factor, we do **one-way ANOVA.** If there are two factors, we do **two-way ANOVA.**

If there are more than two factors we do a **factorial design**. A **treatment** is a specific combination of factors aka a factor combination. If there are 3 factors which each have four levels, then the number of treatments is 4^3 or 64. The number of **replications** is the number of **experimental units** on which each treatment is applied. So if there are 256 experimental units and 64 treatments, then there are 4 replications. N = t * r

Another example with a single factor: if there are 32 experimental units and four different treatments, then there are 8 replications. Think of a replication as an execution of the experiment. A replication is like a flip of a coin in that sense. A replication is a plot. It is a row on a table.

If you hear the phrase, "each individual was assigned to one of three groups…" then you know you're dealing with ANOVA. While a t-test compares two groups, ANOVA is used to compare three or more groups.

The **assumptions for ANOVA** are normality, equal variances and SRS.

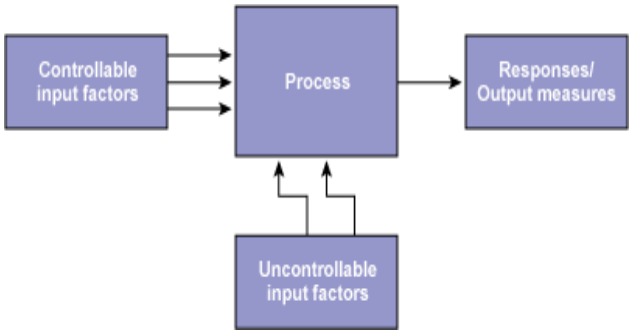
The normality assumption is that all subpopulations are normally distributed. It is tested by looking at a normal probability plot of the sample data and also test of residuals.

The equal variance assumption is tested by a residual plot of the fitted values vs. the residuals. The points must be in a horizontal band to claim constant variances. This ANOVA test is not sensitive to the constant variances assumption, so slightly different variances will not change our conclusions much. If the equal variances assumption fails, you can transform the response variable. If the problem cannot be fixed, use a non-parametric procedure.

The randomness/independence assumption implies that there is little dependence between data points. To test this, run order plot: order vs. residuals.

ANOVA
Experimental Design

The model looks like this:



If you apply the same treatment to a group of individuals, it is unlikely to yield the exact same quantitative outcome, even under controlled conditions. We can attribute this to four causes of **experimental error:**

1. Natural variation in experimental units. No two people are alike
2. Measurement error. This is why they take your blood pressure twice and then average at the hospital.
3. Variation in treatment. No two pills are identical.
4. Exogenous factors. Everything else. Context.

Although there is no perfect answer, we can control for experimental error with good design. *Think* about the experiment and how bias might creep in.

1. Have a precise experimental procedure.
2. Careful selection of experimental and measurement units.
3. Reduce variance by using a blocked design (subgroups) to control for experimental error i.e. confounding of the blocked variable with the response.
4. Reduce variance by …Use **covariates-** continuous variables that are also confounding, used to reduce variance. This is covered in ANCOVA.

As for the uncontrollable input factors, we want to randomize them in their application. We call these factors **blocks**. A block is a subgroup. Randomization eliminates the block effect to isolate the treatment effect. This can be done using a CRBD- a **completely randomized block design**.

Note that if the uncontrolled variable is categorical, it is a block. If it is continuous, it is a covariate.

One way to think of the difference between a factor and a block is that a factor can be randomized while a block can't. Another way to think about it is that a block must contain all treatment combos within it. Otherwise it is not orthogonal. If it is expected that the questionable factor/block would have no impact on y, then it would be more likely to be a block.

Blocking lowers the number of replications. If we have a single factor with four levels, a sample size of 32, and four blocks (two individuals per block), then 32/4/4 = 2 replications.

ANOVA
a graphical interpretation

To perform ANOVA, we begin by taking the mean of each subgroup.

Next, we measure the variance. There are three types of variance in ANOVA.

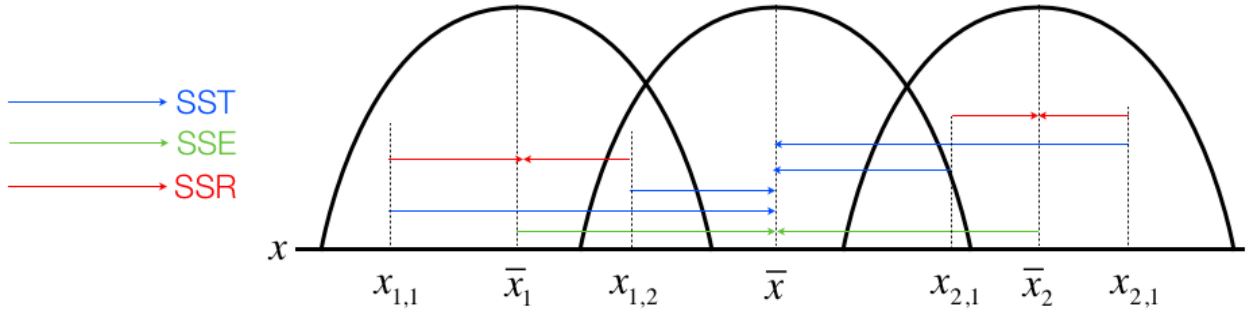
1. The variance within the subgroup itself (SSW (within) or SSR (residual))
2. The variance between the subgroups (SSB (between) or SSE (explained))
3. The total variance (SST)

The SSW represents the experimental error and the SSB represents the difference that can be attributed to the factor.

By measuring (1) and (2) we can sum them for (3). The formula looks something like this:

Total variance = variance within the groups + variance between the groups

The number line below gives us a visual interpretation of how the different variances are calculated. x_{ij} is the outcome for subgroup i, observation j. $\overline{x}_i$ is the mean of subgroup *i*, and $\overline{x}$ is the **grand mean**- the mean of all of the individuals. (note that we should be talking about y, not x)



Unlike the number line above, in actuality we are measuring *squared* distances. Mathematically:

$$\left(X_{ij}-\overline{X}\right)^2=\left(\overline{X}_j-\overline{X}\right)^2+\left(X_{ij}-\overline{X}_j\right)^2$$

ANOVA
a statistical model for one-way ANOVA

The beginning model for one-way ANOVA gives way to the more intuitive model:

$$y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad \mu_i = \mu + \tau_i$$

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}$$

y_{ij} is the observed outcome for the ith treatment and jth experimental unit. *i x j* is the sample size.
μ is the **grand mean** or **overall mean**, which is the average of the observed outcomes to all experimental units.
τ_i is the differential effect of the treatment for subgroup i. $\sum \tau_i = 0$
$\varepsilon_{i,j}$ is the noise or error due to other factors associated with the (i,j)th value. We assume that the errors are ~iid, meaning independent, identical distributions with common variance $N(0, \sigma_\varepsilon^2)$.

μ, τ_i, and $\varepsilon_{i,j}$ are all unknown population parameters.

The equation for the estimators of the above variables in respective order is as follows:

$$y_{ij} = \bar{y}_{..} + (\bar{y}_i - \bar{y}_{..}) + (y_{ij} - \bar{y}_i)$$

Where $\bar{y}_{..}$ is the grand mean for the sample and $\bar{y}_i$ is the mean for subgroup i. The dots in the subscript represent the average for the value.

To calculate the variance, we move the grand mean to the left-hand side of the equation and then sum the squares:

$$\sum (y_{ij} - \bar{y}_{..})^2 = \sum ((\bar{y}_i - \bar{y}_{..}) + (y_{ij} - \bar{y}_i))^2$$

$$\sum (y_{ij} - \bar{y}_{..})^2 = \sum (\bar{y}_i - \bar{y}_{..})^2 + \sum (y_{ij} - \bar{y}_i)^2 + 2\sum ((\bar{y}_i - \bar{y}_{..})(y_{ij} - \bar{y}_i))$$

Where the last term equals zero. We're left with:

$$\sum (y_{ij} - \bar{y}_{..})^2 = \sum (\bar{y}_i - \bar{y}_{..})^2 + \sum (y_{ij} - \bar{y}_i)^2$$

The left hand term is the **total sum of squares (TSS),** followed by the **treatment sum of squares (SST)** and the **error sum of squares (SSE)**.

ANOVA
Summary table

The **ANOVA summary table** sums up the three measures of variance:

SS_	sum of squares	DF	mean square	F
SST	$\sum(\bar{y}_i - \bar{y}_{..})^2$	i-1	SST/DFT	MST/MSE
SSE	$\sum(y_{ij} - \bar{y}_i)^2$	i (j-1)	SSE/DFE	
TSS	$\sum(y_{ij} - \bar{y}_{..})^2$	N(=ij)-1	TSS/TDF	

There are some nice relationships on this table: the SST and DFT are the sum of the first two rows in each respective column.

In the ANOVA context, the F-statistic is

$$F_{\alpha,df_1,df_2} = \frac{MSB}{MSW} = \frac{s_B^2}{s_W^2}$$

This implies that a large amount of between group variance and a small amount of within group variance will lead to a larger F-statistic. This would imply that the means, subgroups and treatments are statistically different. So we reject with a large F.

The total "within group" variation is

$$s_w^2 = \frac{(n_1 - 1)s_1^2 + ... + (n_i - 1)s_i^2}{n_1 + n_2 + ... + n_i - i}$$

The total "between group" variation is

$$s_B^2 = \frac{\sum(\bar{x}_i - \bar{x})^2}{i - 1}$$

The hypothesis test for ANOVA is:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_i \text{ or } H_0 : \tau_1 = \tau_2 = ... = \tau_i$$
$$H_a : \text{At least one of the treatments is different.}$$

When the p-value of the F-statistic is less than α, reject the null hypothesis for the alternative which states that at least one level is different from the others. (see ex. 8.1 and 8.2, match with SAS output.)

SS_	sum of squares	DF	mean square	F
SST	$\sum(\bar{y}_i - \bar{y}_{..})^2$	t-1	SST/DFT	MST/MSE
SSE	$\sum(y_{ij} - \bar{y}_i)^2$	t(r-1)=N-t	SSE/DFE	
TSS	$\sum(y_{ij} - \bar{y}_{..})^2$	tr-1=N-1	TSS/TDF	

F= MST/MSE.

The way to think of the degrees of freedom is as follows. For SST, the DF is the number of treatments minus 1. Recall that $\sum \tau_i = 0$, therefore we can calculate the last tau without knowing the value. That is the DF…? The total sum of squares df is the sample size minus one. DF for the SSE is can be figured out from the other two, but it is basically the sample size minute the number of treatments.

In the ANOVA context, the F-statistic is

$$F_{\alpha,df_1,df_2} = \frac{MSB}{MSW} = \frac{s_B^2}{s_W^2}$$

It is important to note that $F_{1-\alpha/2,df_1,df_2} = \frac{1}{F_{\alpha/2,df_2,df_1}}$. Note that it's df2, df1 in the denominator.

Note also that the DF in the F-statistic subscript are always in the other df1, df2 referencing the table. (Ex. 7.7)

DF1 represents the degrees of freedom for the "between samples" or "treatment sum of squares" (t-1). DF1 represents the degrees of freedom for "within samples" or the "error sum of squares" n-t.

Multiple Comparisons

If we perform one-way ANOVA and reject H0 and are interested in which treatment mean is different, we can perform a **multiple comparisons** test. Is treatment 1 different from the rest? Or is some combination of them different?

The **linear contrast statement L** is an equation that contains all of the means being tested with a coefficient for each mean.

$$L = a_1\mu_1 + a_2\mu_2 + a_t\mu_t \text{ subject to } \sum a_t = 0$$

Using sample data,

$$\hat{L} = a_1\bar{y}_1 + a_2\bar{y}_2 + a_t\bar{y}_t = \sum a_i\bar{y}_i \text{ subject to } \sum a_t = 0$$

Using data with a single factor and four treatments, we apply a value to each treatment:

Treatment 1 (a1)	Treatment 2 (a2)	Treatment 3 (a3)	Treatment 4 (a4)	Comparison	Sample
0	0	1	-1	$H_0 : \mu_3 = \mu_4$	$\hat{L} = \bar{y}_1 - \bar{y}_2$

A 0 means that the treatment is not part of the comparison. Note that the row must always add to zero, $\sum a_t = 0$.

Comparisons can be for more than one group, such as:

Treatment 1 (a1)	Treatment 2 (a2)	Treatment 3 (a3)	Treatment 4 (a4)	Comparison	Sample
-1	-1	3	-1	$H_0 : \mu_2 = \frac{\mu_1 + \mu_3 + \mu_4}{3}$	$\hat{L} = \bar{y}_1 - \bar{y}_2$

We will assume that each treatment the sample size is the same.

Two linear contrast statements are said to be **mutually orthogonal contrast statements (MOCS)** if the sum of the products of each columns equals zero. So if $L_1 = \sum a_i\mu_i$ and $L_2 = \sum b_i\mu_i$, $0 = \sum a_i b_i$.

Treatment 1 (a1)	Treatment 2 (a2)	Treatment 3 (a3)	Treatment 4 (a4)	Comparison	Sample
-1	-1	3	-1	$H_0 : \mu_2 = \frac{\mu_1 + \mu_3 + \mu_4}{3}$	$\hat{L} = \bar{y}_1 - \bar{y}_2$

The maximum number of MOCS is the number of treatments minus 1.

The importance of MOCS is that it splits up the data in such a way that it allows an ANOVA to be performed. In a one-way ANOVA, the SST and the SSE are orthogonal. In a factorial design, the SST is broken up into MOCS L1, L2, etc. It is able to partition the effect by contrasts.

In such a case, the F-test would give a MSC/MSE, where the df for the MSC = 1. The DF for each contrast statement is 1 and the DF for all contrast statements is t-1.

In a sense, each factor is a linear contrast statement.

Multiple Comparisons-Subsetting Data

L.C.S. are cool because you can compare two groups to two other groups with a control.

When looking at a L.C.S., comparing two means is a **pairwise comparison.** In a pairwise comparison, if you have 5 means, you have 5-choose-2 or 10 pairwise comparisons that you can carry out.

Why not just do multiple t-tests between all of the two groups? Because with each additional chance, the probability of a type I error increases.

How do we arrive at this conclusion? Each test has a probability of a type 1 error (rejecting a null when it is true) of alpha of say 5%. This is called the individual type one error.

The experimental type 1 error or the overall alpha is the probability of getting at least 1 type 1 error when carrying out multiple pairwise comparisons.

You can calculate the overall alpha using the binomial distribution. If we have n=10 pairwise comparisons with alpha = 0.05, the probability of a type 1 error is

$$P(Y \geq 1) = 1 - \binom{10}{0} .05^0 (1 - .05)^{10} = 40.1\%$$

In order to control the experimental type 1 error you can decrease your alpha. The Bonferroni procedure says to divide alpha equally by the number of tests. Alternatively, doing an interim analysis means testing the hypothesis in multiple time periods as the sample data is collected in order to get an idea of whether it is worthwhile to continue the test. One may not have to divide alpha equally in this situation

ANOVA
Kruskal-Wallis

The non-parametric alternative to ANOVA is the Kruskal-Wallis test. It is used when the dependent variable is not normally distributed and/or the constant variances assumption fails (even after transformation). It is used when the number of levels is greater than or equal to 3, otherwise Mann-Whitney or Wilcoxon is used.

H_0 : Probability distributions are identical for each level of the factor

H_a : Not all distributions are the same

Note that this is NOT a comparison of means. Like other rank tests, each subgroup is sorted, combined, and ranked. Ties are averaged- and this is the only averaging that happens! The sum of the ranks is then plugged into the test statistic:

$$H = \frac{12}{N(N+1)} * \sum_{i=1}^{i} \frac{T_i^2}{n_i} - 3(N+1)$$

The H-statistic can be approximated by a χ^2_{i-1} distribution.

Categorical Data
introduction

Categorical data can be analyzed in a method similar to the analysis of quantitative variables. This section will cover inference for a single proportion and for two proportions. In addition, it will cover one-way ANOVA for proportions.

In this section, we use the z-statistic only. This is because, as long as the sample size is sufficiently large, proportions have a binomial distribution that approaches the normal distribution.

The key assumption of $np>5$ and $nq<5$ is there to judge the adequacy of the normal approximation to the binomial distribution. If $np<5$, we use the exact binomial distribution, as we will see with Fisher's exact test. On a two-way categorical table, all that matters is that each cell value exceeds 5.

Categorical Data
Inference for a single proportion

The sample proportion is $\hat{\pi} = Y/n$ where Y is the number of successes in n trials for group xi. We perform inference on the sampling distribution of proportions, which is a binomial distribution with B~N($\mu_{\hat{\pi}}, \sigma_{\hat{\pi}}$).

$\mu_{\hat{\pi}}$ is an unbiased estimator of the population mean, π .

$\sigma_{\hat{\pi}}$ is an unbiased estimator of the population standard deviation, $\sqrt{\frac{\pi(1-\pi)}{n}}$.

The approximate confidence interval (aka **Wald**) for π is:

$$\hat{\pi} \pm Z_{\alpha/2} * \hat{\sigma}_{\hat{\pi}}$$

Where **the standard error of the proportion** is:

$$\hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Note: Another kind of standard error is the **asymptotic standard error (ASE)**, which uses the null π_0 instead of $\hat{\pi}$. The standard error above is listed as the "score" in SAS.

For a desired sample size, use

$$n = \frac{Z_{\alpha/2}^2 * \pi * (1-\pi)}{E^2}$$

where 2E is the width of the entire confidence interval at the desired confidence level. Note also that you can assume $\pi = 0.5$ for the largest sample size.

When $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$, the **hypothesis test for the value of the population proportion** with test statistic:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Note that the hypothesis test uses the null hypothesized proportion in the standard error.

H0	Ha	RR
$\pi \leq \pi_0$	$\pi > \pi_0$	$z > z_\alpha$
$\pi \geq \pi_0$	$\pi < \pi_0$	$z < -z_\alpha$
$\pi = \pi_0$	$\pi \neq \pi_0$	$\lvert z \rvert > -z_{\alpha/2}$

Categorical Data

Inference for a comparison of proportions, $\hat{\pi}_1 - \hat{\pi}_2$

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2$$

$$\sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

The **confidence interval for the difference of two proportions** is

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} * \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}$$

Where the **standard error of the difference of proportions** is the sum of the standard deviations for each sample:

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

The **hypothesis test for the difference of two proportions** with test statistic $z = \dfrac{(\hat{\pi}_1 - \hat{\pi}_2)}{\sqrt{\dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$ is

H0	Ha	RR
$\pi_1 - \pi_2 \leq 0$	$\pi_1 - \pi_2 > 0$	$z > z_\alpha$
$\pi_1 - \pi_2 \geq 0$	$\pi_1 - \pi_2 < 0$	$z < -z_\alpha$
$\pi_1 - \pi_2 = 0$	$\pi_1 - \pi_2 \neq 0$	$\lvert z \rvert > -z_{\alpha/2}$

Categorical Data
Fisher's Exact Test

When the sample size is too small, we can perform Fisher's exact test. The example here is the "Lady's Tea Tasting" problem.

Some wealthy socialite from the 18th century said that she could taste whether tea was added to milk or milk was added to tea. In order to test this, Fisher produced 8 cups of tea, 4 with milk first and 4 with tea first. Note the small sample size.

All of the possible outcomes are in the tables below. Note that the margins have to be the same.

		Worst		
		Lady		Total
		Milk 1st	Tea 1st	
Truth	Milk 1st	0	4	4
	Tea 1st	4	0	4
	Total	4	4	

2nd Worst			
	Lady		Total
	Milk 1st	Tea 1st	
Milk 1st	1	3	4
Tea 1st	3	1	4
Total	4	4	

Random			
	Lady		Total
	Milk 1st	Tea 1st	
Milk 1st	2	2	4
Tea 1st	2	2	4
Total	4	4	

		2nd Best		
		Milk 1st	Tea 1st	
		Lady		Total
Truth	Milk 1st	3	1	4
	Tea 1st	1	3	4
	Total	4	4	

Best			
	Milk 1st	Tea 1st	
	Lady		Total
Milk 1st	4	0	4
Tea 1st	0	4	4
Total	4	4	

Let's say the lady got ¾ correct- the 2nd best outcome. Is this enough to show that she can taste the difference?

The null hypothesis is that she can't taste the difference. The alternative is that she can.

All of these possibilities fit the geometric distribution.

What is the probability of four out of four?

$$\frac{\binom{4}{4}*\binom{4}{0}}{\binom{8}{4}} = 1/70$$

3 out of four?

$$\frac{\binom{4}{3}*\binom{4}{1}}{\binom{8}{4}} = 16/70$$

So the probability of getting 4/4 or ¾ is 17/70, which is 0.24. This is a large enough probability to fail to reject the null.

Categorical Data
Chi-Square Test of Independence

The chi-square test is essentially ANOVA for categorical variables when there are more than two classes. When the outcomes have more than two possible classes (ex. "success/partial success/failure"), the data no longer have a binomial distribution. This is a multinomial experiment that essentially compares the classes of each of the two categorical variables.

The chi-square test of independence tests whether two categorical variables are associated.

$$H_0: \text{ The two variables are independent}$$
$$H_A: \text{The two variables are associated}$$

The test of independence compares the distribution of outcomes with what is expected. We want to compare the **observed frequencies** to what we expect in the population. To compute the **expected frequencies** for each outcome, multiply the relative frequency of each outcome by the column total.

Starting with a two-way **contingency table of observed frequencies**:

	Y1	Y2	Y3	Total
X1	40	10	10	60
X2	90	40	10	140
Total	130	50	20	200

To compute expected frequencies, use

$$\hat{E}_{ij} = \frac{n_i * n_j}{n}$$

which yields the **two-way contingency table of expected frequencies**:

	Y1	Y2	Y3	Total
X1	39	15	6	60
X2	91	35	14	140
Total	130	50	20	200

In a stacked table, calculate the **squared normal deviates:**

	Observed	Expected	(O-E)	(O-E)^2	((O-E)^2)/E
X1/Y1	40	39	1	1	0.03
X1/Y2	10	15	-5	25	1.67
X1/Y3	10	6	4	16	2.67
X2/Y1	90	91	1	1	0.01
X2/Y2	40	35	5	25	0.71
X2/Y3	10	14	-4	16	1.14
Total	200	200	0	84	**6.23**

The **chi-square statistic** is the sum of the square normal deviates. It can mathematically be formulated as:

$$\chi^2 = \sum_i \frac{(\mathrm{O}_i - \mathrm{E}_i)^2}{\mathrm{E}_i}$$

We reject the null hypothesis if $\chi^2 > \chi^2_{\alpha,df}$, where df=(r-1)(c-1)

Remember, the test of independence is used when 80% or more cells must have expected values greater than or equal to 5: $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

Categorical Data
Chi Square Goodness of Fit

In the chi-square test for goodness of fit, the population proportions are already hypothesized. Our hypothesis test is that the observed and expected proportions are relatively equal.

$$H_0 : \pi_i = \pi_0 \text{ for all } \pi_i$$
$$H_a : \pi_i \neq \pi_0 \text{ for one or more } \pi_i$$

Technically, the value of the last population proportion is implied.

A goodness of fit table will look like this.

	Sub1	Sub2	Sub3	Sub4	Total
Observed Frequency	45	28	11	16	100
Hypothesized Proportion	50	25	12	13	100

The key assumption here is of adequate cell counts- 80% or more cells must have expected values greater than or equal to 5 ($n\pi \geq 5$ and $n(1-\pi) \geq 5$).

The degrees of freedom are the number of subgroups minus 1.

Completely Randomized Design
Factorial Treatment Structure

When we have two or more explanatory, categorical variables, we move from one-way ANOVA to a **factorial treatment structure**. If we're working with exactly two explanatory variables, then it is technically **two-way ANOVA**. These types of experiments are involved with examining the effect of two or more explanatory variables on a response variable *y*.

Then, assuming that the experimenter has chosen the levels of each independent variable, he or she must decide which factor–level combinations are of greatest interest and are viable. In some situations, certain of the factor–level combinations will not produce an experimental setting that can elicit a reasonable response from the experimental unit. Certain combinations may not be feasible due to toxicity or practicality issues. **one-at-a-time approach.** To examine the effect of a single variable, an experimenter changes the levels of this variable while holding the levels of the other independent variables fixed. If this happens, the two factors, nitrogen and phosphorus, are said to **interact.** That is, the effect of one factor on the response does not remain the same for different levels of the second factor, and the information obtained from the one-at-a-time approach would lead to a faulty prediction.

We include an interaction if the difference in the mean response of the first factor is not constant across all levels of the second factor. Draw a line chart or a "treatment mean profile plot" or "mean response profile plot" to see whether there is an interaction. **(ex. 14.10)**

For a given level of factor A, are the differences between all levels of factor B the same? If there is not a constant difference between levels, then there may be an interaction. The magnitude of the interaction is more evident when the actual direction of the effects change. An **ordinal** interaction means that the directions are always the same, but the magnitudes are different.

We have seen that the one-at-a-time approach to investigating the effect of two factors on a response is suitable only for situations in which the two factors do not interact.

Factorial treatment structures are useful for examining the effects of two or more factors on a response *y*, whether or not interaction exists.

Classically, factorial treatment structures have not been referred to as designs because they deal with the choice of levels and the selection of factor–level combinations (treatments) rather than with how the treatments are assigned to
Experimental units. Unless otherwise specified, we will assume that treatments are assigned to experimental units at random. The factor–level combinations will then correspond to the "treatments" of a completely randomized design.

A **factorial treatment structure** is an experiment in which the response *y* is observed at all factor–level combinations of the independent variables.

Using our previous example, if we are interested in examining the effect of two levels of nitrogen, $x1$, at 40 and 60 pounds per plot and two levels of phosphorus, $x2$, at 10 and 20 pounds per plot on the yield of a crop, we could use a completely randomized design where the four factor–level combinations (treatments) of Table 14.9 are assigned at random to the experimental units.

(Basically, each combination sis a treatment.)

One final comparison should be made between the one-at-a-time approach and a factorial treatment structure. Not only do we get information concerning factor interactions using a factorial treatment structure, but also, when there are no interactions, we get at least the same amount of information about the effects of each individual factor using fewer observations.

The Model

The beginning model for one-way ANOVA gives way to the more intuitive model:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \mu_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij}$$

$$y_{i,j} = \mu + \tau_i + \varepsilon_{i,j}$$

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$$

y_{ijk} is the observed outcome (aka the response) from the kth experimental units receiving the ith level of factor A and the jth level of factor b. (Note that k can be interpreted at the number of replications.)

μ_{ij} is the treatment mean for the ith and jth level of factors A and B, respectively.

μ is the **grand mean** or **overall mean**, which is the average of the observed outcomes to all experimental units.

τ_i is the differential effect to the ith level of factor A. $\sum \tau_i = 0$, an unknown constant

β_j is the effect due to the jth level of factor B, an unknown constant. $\sum \beta_j = 0$

$\tau\beta_{ij}$ is the interaction effect of the ith level of factor A and the jth level of factor B, an unknown constant. $\sum \tau\beta_{ij} = 0$. Note that tao and beta are not multiplied in any sense.

ε_{ijk} is a random error associated with the response from the kth experimental unit receiving the ith level of factor A combined with the th level of factor B. We assume that the errors are ~iid, meaning independent, identical distributions with common variance $N(0, \sigma_\varepsilon^2)$.

If the model is **fixed effects**, then the main effects are constants, i.e. they do not have multiple levels. If the model is **random effects**, then the main effects have their own distributions with various possibilities. A **mixed effects** model incorporates fixed and random effects into a single model.

TABLE 14.13
Expected values for a 2×2 factorial treatment structure without interactions

	Factor B	
Factor A	**Level 1**	**Level 2**
Level 1	$\mu + \tau_1 + \beta_1$	$\mu + \tau_1 + \beta_2$
Level 2	$\mu + \tau_2 + \beta_1$	$\mu + \tau_2 + \beta_2$

The Table

Below is a table for 3 factors.

TABLE 14.19
AOV table for a completely randomized design with an $a \times b \times c$ factorial treatment structure

Source	SS	df	MS	F
Main effects				
A	SSA	$a-1$	$\text{MSA} = \text{SSA}/(a-1)$	MSA/MSE
B	SSB	$b-1$	$\text{MSB} = \text{SSB}/(b-1)$	MSB/MSE
C	SSC	$c-1$	$\text{MSC} = \text{SSC}/(c-1)$	MSC/MSE
Interactions				
AB	SSAB	$(a-1)(b-1)$	$\text{MSAB} = \text{SSAB}/(a-1)(b-1)$	MSAB/MSE
AC	SSAC	$(a-1)(c-1)$	$\text{MSAC} = \text{SSAC}/(a-1)(c-1)$	MSAC/MSE
BC	SSBC	$(b-1)(c-1)$	$\text{MSBC} = \text{SSBC}/(b-1)(c-1)$	MSBC/MSE
ABC	SSABC	$(a-1)(b-1)(c-1)$	$\text{MSABC} = \text{SSABC}/(a-1)(b-1)(c-1)$	MSABC/MSE
Error	SSE	$abc(n-1)$	$\text{MSE} = \text{SSE}/abc(n-1)$	
Total	TSS	$abcn-1$		

If the study is a **balanced design**, then the sum of squares for the **main effects,** the **interactions**, and the **error** is equal to the total sum of squares. They are also mutually orthogonal.

The interactions are AB or A*B. If there are x factors in the study, then there are $\binom{x}{x-1}$ two way interaction effects and if there are y two-way interaction effects, there are $\binom{x}{y}$ three-way interaction effects.

The degrees of freedom for the SST is the sum of the df for the main and interaction effects.

R^2 if the difference in Y explained by the model.

Adjusted R^2 factors out bullshit explanatory variables. We want to explain as much as possible but keep it simple.

CV is the coefficient of variation, the mean over the standard devation.

The root MSE (root-mean-squared error) is an estimate of σ_ε, $\hat{\sigma}_\varepsilon$.

If your treatments = 12 then you need N>12 to get the error degrees of freedom.

Hypothesis Tests

The first hypothesis test is to make sure at least one of the means is significantly different. This is the one-way ANOVA.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_i \text{ or } H_0 : \tau_1 = \tau_2 = ... = \tau_i$$
$$H_a : \text{At least one of the treatments is different.}$$

The second hypothesis test includes the fixed effects- it is the CRD. First, we look to see if the interactions are significant. (If the interactions are significant, then the main effects are too.)

$$H_0 : \tau\beta_{ij} = 0$$
$$H_a : \tau\beta_{ij} \neq 0$$

If there are more than two factors, we do not have to reject all of the interaction terms. If we do reject all of the interaction terms, then the model is said to be an **additive model.**

Finally, we can test the individual main effects.

Calculate a confidence interval using percentage points of studentized range distribution (ex 14.11)

$$(\bar{y}_i - \bar{y}_i^{'}) \pm q_\alpha(t,v)\sqrt{\frac{s_\varepsilon^2}{n_t}}$$

$q_\alpha(t,v)$, where alpha is the degrees of freedom, t is the number of treatment means and v is the degrees of freedom for the MSE.

$s_\varepsilon = \sqrt{MSE}$

n_t is the number of observations in each treatment group.

So this is about comparing the means of one factor without controlling for the other factor. We're comparing the means on the margin, using t=the total number of treatments for the factor and the MSE gives us v and S-sub-E.

This is used to calculate the difference in mean reaction times between two levels of a factor.

So first, we perform an F-test to make sure the main effect is significant. Then we would place confidence intervals on the difference between any pair of factor-level means, $\mu_{i.} - \mu_{i.}^{'}$ or $\mu_{.\mathrm{j}} - \mu_{.\mathrm{j}}^{'}$ for factor A or B respectively.

Next, we would want to determine which pairs of factor level means are significantly different using multiple comparison procedures (chapter 9) controlling for the error rate.

Non-parametric Procedures
Fisher and Tukey

Fisher's Least Significant Difference (LSD) test is sued to compare different groups in order to determine whether they are significantly different. We would run this test after rejecting the null hypothesis that all the means are equal.

In order to calculate the LSD, we use

$$LSD = t_{\alpha/2} * \sqrt{MSE} * \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

The value of the LSD is at p=.05.

μ_1	μ_2	μ_3	μ_4	μ_5
7	12	13	15	18

Given the above data and a LSD of 5, how can we divide the means into significantly different groups? (Note that the LSD must be "equal to or greater than" 5.)

1. μ_1 is significantly different from μ_{2-5}. Note that a single mean can count as its own group.
2. Both μ_2 & μ_3 are significantly different from μ_5. But μ_4 will fall into both groups. (Note that the groups do not have to be mutually exclusive.) Because of this fact, μ_4 & μ_5 are <u>not</u> significantly different from μ_{2-4}.

We can use Tukey's W procedure (ex. 14.12), where $W = q_\alpha(t,v)\sqrt{\frac{s_W^2}{n}}$, where n is the number of observations per mean and S-squared-W is the MSE from the AOV table.

Again we're looking at the means for a factor on the margin. We list them from lowest to highest. If the difference is more than W, we declare them to be statistically different from each other,

Completely Randomized Design
Confidence Intervals and Hypothesis Tests

In statistics, someone will give you the null and alternative hypothesis. It's up to you to come up with the design and the sample size. The sample size is driven by the design.

You can choose the replication count based on either accuracy or power.

If you select **accuracy**, you're minimizing the confidence interval, or specifically the margin of error.

This is the same formula for calculating n for a confidence interval (in fact it's derived from the formula), except instead of n it's r.

$$E = z_{\alpha/2} * \frac{\hat{\sigma}}{n}$$

$$n = r = \frac{z^2_{\alpha/2} * \hat{\sigma}^2}{E^2}$$

The population variance is unknown so it is either drawn from other research of estimated through a pilot study. The most basic estimator is

$$\hat{\sigma}^2 = \frac{\max - \min}{4}$$

If you select **power,** you're maximizing the power (against cost). Recall that power is the probability of rejecting the null hypothesis when the alternative is true (as apposed to failing to reject the null hypothesis when the alternative is true, a type II error). In this case, we need to know 5 things: alpha, the difference that is considered significant which is related to the alternative hypothesis (ex. cost effective if difference is…), the power, the variance, and the number of treatments.

Given an alpha of 5%, D=15, beta of 10% a variance of 7.5 and a t=4 treatments…

$$\lambda = \frac{rD^2}{2\sigma^2} = \ldots = 2r \qquad \phi = \sqrt{\frac{\lambda}{t}} = \sqrt{\frac{r}{2}}$$

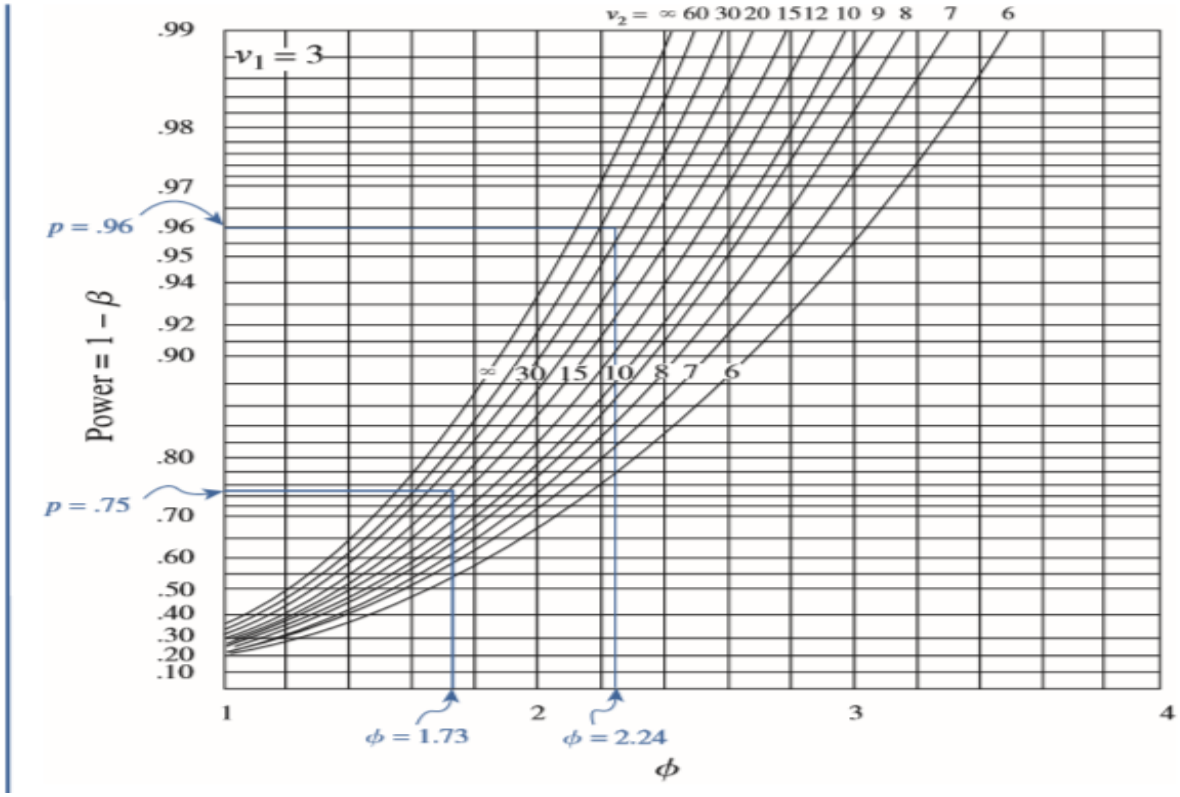
Where r is the number of replications. Note that this is an "open form" equation. There is no closed form. You need to know df1 and try out different df2s.

r	DF2 (t(r-1))	ϕ	Power

3	8	2.597535115	0.88
4	12	3.463380153	0.925

You can use a table to solve this problem.

Remember that n = r * t.



ANOVA for Blocked Designs

CRBD Model

Now it's time to add blocks! A block is treated just like a factor.

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

τ_i is the treatment

β_j is the block

Interactions, either between blocks or between blocks and treatments, are not considered. Furthermore, we assume that the block effects are subject equally to each observation.

Within each block, treatments are randomly assigned. So to apply 16 observations to a block with four levels and a factor with four treatments:

1. Randomize all the observations 1-16. Put the first four observations in block one, the second set of four observations in block two, etc.
2. Within each block, randomize the four treatments, then apply to subjects. This is essentially a CRD within each block.

It's OK if each treatment appears only once in each block (i.e. no replications). For example, if you have a block with four levels and a treatment with four levels, you'll have a minimum of 16 observations. With 3 df for both the treatment and block and 15 df total, you still have 10 df to estimate your errors.

If you don't have any replications, you can present the data on a two-way table.

	Block				
Treatment	**1**	**2**	**⋯**	***b***	**Mean**
1	y_{11}	y_{12}	⋯	y_{1b}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	⋯	y_{2b}	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
t	y_{t1}	y_{t2}	⋯	y_{tb}	$\bar{y}_{t.}$
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	⋯	$\bar{y}_{.b}$	$\bar{y}_{..}$

ANOVA for Blocked Designs

AOV Table

Source	SS (sum of square)	Df	MS (mean of square)	F
Treatments	SST	t-1	MST=SST/(t-1)	MST/MSE
Block	SSB	b-1	MSB=SSB/(b-1)	MSB/MSE
Error	SSE	(b-1)*(t-1)	MSE=SSE/(b-1)(t-1)	
Total	TSS	btk-1		

An AOV output may also display the F-statistic for the block and treatment individually.

K is the number of replications.

ANOVA for Blocked Designs

Latin Square Design (LSD)

A Latin Square Design has in an ANOVA model with one factor and exactly two blocks.

$$y_{ij(k)} = \mu + \tau_{(k)} + \beta_i + \gamma_j + \varepsilon_{ij(k)}$$

$\tau_{(k)}$ is the effect of the treatment. The subscript k represents the replication. If a LSD has only one replication, then it is determined by the level of both blocks- k is a function of i and j. (If there is more than one replication, each replication gets its own table.)

Randomization applied to row, column and treatment.

On an LSD table, each treatment appears exactly once in each row and column. There are an equal number of treatments and levels within both blocks.

Positions	Applications			
	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Why must it be this way? It is the only way for the extraneous variation among blocks to balance out and to accurately measure the treatment effect.

Positions	Applications			
	1	2	3	4
1	A	A	C	A
2	B	D	A	D
3	C	B	D	B
4	D	C	B	C

$$\tau_1 = \bar{y}_{..1} = (y_{211} + y_{321} + y_{431} + y_{431}) * 1/4$$

$$y_{211} = \mu + \beta_2 + \gamma_1 + \tau_1 + \varepsilon_{211} \;/\; y_{321} = \mu + \beta_3 + \gamma_2 + \tau_1 + \varepsilon_{321}$$

$$\bar{y}_{ij1} = 1/4(\beta_2 + \beta_3 + \beta_4 + \beta_3) + 1/4(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4) + \tau_1 + \varepsilon$$

$$\bar{y}_{ij3} = 1/4(\beta_3 + \beta_4 + \beta_1 + \beta_4) + 1/4(\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4) + \tau_3 + \varepsilon$$

$$\bar{y}_{ij1} - \bar{y}_{ij3} = (\tau_1 - \tau_3) + \frac{(\beta_3 + \beta_2) - (\beta_4 + \beta_1)}{4}$$

The LSD AOV table is below.

Source	SS	Df	MS	F
Treatments	SST	t-1		
Rows	SSR	r-1		
Columns	SSC	c-1		
Error	SSE	(t-1)(t-2)		
Total	TSS	T^2-1		

You can see that t has to be equal to 3 or greater as a minimum

When null hypothesis testing:

$$H_0 : \tau_1 = \tau_2 = ... = \tau_5 = 0$$

When hypothesis testing the blocks, what we want to know is if there is a random effect. We do not want to know if the 5 houses or whatever are equal. We want to know whether the variances for each block are equal.

$$\gamma_1 \sim N(0, \sigma_\gamma^2)$$

$$H_0 : \sigma_\gamma^2 = 0$$

$$H_a : \sigma_\gamma^2 > 0$$

ANCOVA
ANCOVA Model

The model takes several forms. The **full model** is:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 x_{ij} + \varepsilon_{ij}$$

Where τ_i is the treatment, β_1 is the effect of the covariate, β_0 is the intercept. It should be noted that in this case β_1 is constant for all values of x_{ij} and all treatments.

The **general linear model** treats each level of the factor as a separate parameter:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_t x_t + \varepsilon$$

β_1 is the effect of the covariate, β_2 through β_t represent the effects of different levels of the treatment where x_2 through x_t are **dummy variables** that are equal to 1 only if the treatment specifically applies to that replication and is equal to zero otherwise. (Note we can exclude one level of the treatment, generally the control, from the model, for a total of t-1 estimators.) Note also that we don't write the "ij" subscript because we treat it as a vector.

We start with the full model because it makes fewer assumptions about the treatment effects.

We begin by testing the **equal slope hypothesis.** $H_0 : \beta_i = \beta_1$. We can test this by including an interaction term in the model.

$$y_{ij} = \beta_0 + \tau_i + \beta_i x_{ij} + \varepsilon_{ij}$$

If we reject the null, we conclude the slope is the same for all treatments. We follow-up by testing whether the covariate is significant at all.

$$H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$$

You also include $\beta_2 x_{ij}^2$ to test for non-linear effects.

For the treatment, we want to test whether there is a difference in the treatments.

$$H_0 : \tau_1 = \tau_2 = ... = \tau_t$$

We can verify this with an F-test. If we reject the null, then at least one treatment is different.

If we reject the either the factor or covariate parameter altogether, we are left with a **reduced model**, where only one explanatory variable remains.

ANCOVA
AOV Table

For the ANCOVA table, we actually reference two separate tables.

The first table is the type III SS table.

Source	Df	SS	MS	F
Factor	t-1	SSF	MSF=SSF/(t-1)	MSF/MSE
Covariate	1	SSC	MSC=SSC	MSC/MSE

A few observations on this table. The degrees of freedom for a single covariate is only 1! This is a good argument for including a variable as a covariate rather than a factor if possible.

Second, the SSF or SSC can be calculated using the AoV tables for the reduced model. For each explanatory variable, you want to take the difference between the SSE from the reduced model for the OTHER explanatory variable and the SSE for the full model. (For the factor, take the difference of the SSE between the regression and the ANCOVA models. For the covariate, take the difference of the SSE between the ANOVA and the ANCOVA model.)

Note that the error for the ANCOVA model isn't on the above table. The SSE and MSE for the ANCOVA model is on the main AoV table (below). Furthermore, the statistical significance of each of the explanatory variable is calculated by dividing the MS for the explanatory variable (table above) by the MSE for the model (table below).

Source	SS	Df	MS	F
Model	SST	T	MST=SST/(t-1)	MST/MSE
Error	SSE	n-t-1	MSE=SSE/(b-1)(t-1)	
Total	TSS	n-1		

Random Effects

In a random-effects model for an experiment, the levels of factors used in the experiment are randomly selected from a population of many possible levels. It also applies to blocks.

A one-way random effect model looks like this:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

While the model is the same, the assumptions are different. For a fixed effects model, we assume that the expected value of y_ij for a given level of the treatment is:

$$E(y_{ij}) = \mu + \tau_i$$

In a random effects model, we the null hypothesis is that the expected value of the response for any given level is equal to the grand mean only.

$$E(y_{ij}) = \mu$$

The reason this is the case is because we are assuming that the (random) treatment effect is normally distributed with a mean of 0 and a variance σ_τ^2.

The hypothesis for the random treatment effect is the same as for a fixed effect- the null hypothesis is that the treatment effect is zero. However, with random effects, we test the variance of the treatment. Under the null hypothesis, the variance is equal to zero. If the variance for at least one treatment is significantly greater than zero, then at least one treatment effect is not equal to zero.

$$H_0 : \sigma_{ij}^2 = 0, H_a : \sigma_{ij}^2 > 0$$
$$T.S.: F = MSIJ / MSE, df_1 = 12, df_2 = 20$$
$$F = 3.89; P = .0037$$

Most importantly, the null hypothesis for the fixed effect model is that we are only interested in the specific treatments in the model. In the random effects model, we are interested in the samples treatments, as well as all of the other treatments in the population.

A **one-factor experiment with random treatment effects** uses one factor (with replications) only. In a **mixed-effects model** for an experiment, the levels of some of the factors used in the experiment are randomly selected from a population of possible levels, whereas the levels of the other factors in the experiment are predetermined.

AOV Table

The **expected mean square** table gives us different formula for the value of the mean square depending on your assumption about whether a variable is random or not.

You can also use an interaction effect. If one of the variables in the interaction are random, then both are random. You'll definitely need more than one replication for this.

Source	**SS**	**df**	**EMS**		
			MS	Fixed Effect	Random Effect
Factor A	SSA	a-1	MSA	$\sigma_\varepsilon^2 + bn\theta_A$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2 + bn\sigma_A^2$
Factor B	SSB	b-1	MSB	$\sigma_\varepsilon^2 + an\theta_A$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2 + an\sigma_B^2$
A*B	SSAB	(a-1)(b-1)	MSAB	$\sigma_\varepsilon^2 + n\theta_{AB}$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2$
Error	SSE	ab(n-1)	MSE	σ_ε^2	σ_ε^2
Total	TSS	abn-1			

If the variable is random, then in order to calculate the F-statistic from the mean square for a factor, divide the mean square for the factor by the correct error term. There are multiple error terms if there is an interaction effect; check the EMS output. If the variable is not random, use the MSE.

$$E(MS_A) = \sigma^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2 \Longrightarrow F_0 = \frac{MS_A}{MS_{AB}}$$

$$E(MS_B) = \sigma^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2 \Longrightarrow F_0 = \frac{MS_B}{MS_{AB}}$$

$$E(MS_{AB}) = \sigma^2 + n\sigma_{\tau\beta}^2 \Longrightarrow F_0 = \frac{MS_{AB}}{MS_E}$$

$$E(MS_E) = \sigma^2$$

Variance Components

The formula equals the value of the mean square, and can be used to determine the variance component if the variable is random.

In order to algebraically derive the random variance component for the mean square for a factor, subtract the correct error term (the MSAB for non-interaction effects), then divide by the coefficients for the variance component.

Source	**SS**	**df**	**EMS**		
			MS	Fixed Effect	Random Effect
Factor A	SSA	a-1	MSA	$\sigma_\varepsilon^2 + bn\theta_A$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2 + bn\sigma_A^2$
Factor B	SSB	b-1	MSB	$\sigma_\varepsilon^2 + an\theta_A$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2 + an\sigma_B^2$
A*B	SSAB	(a-1)(b-1)	MSAB	$\sigma_\varepsilon^2 + n\theta_{AB}$	$\sigma_\varepsilon^2 + n\sigma_{AB}^2$
Error	SSE	ab(n-1)	MSE	σ_ε^2	σ_ε^2
Total	TSS	abn-1			

$$\hat{\sigma}_\tau^2 = \frac{MS_A - MS_{AB}}{bn}$$

$$\hat{\sigma}_\beta^2 = \frac{MS_B - MS_{AB}}{an}$$

$$\hat{\sigma}_{\tau\beta}^2 = \frac{MS_{AB} - MS_E}{n}$$

$$\hat{\sigma}^2 = MS_E$$

The sum of all of the variance components is equal to the total variance in the model.

$$\sigma_y^2 = \sigma_\tau^2 + \sigma_\beta^2 + \sigma_{\tau\beta}^2 + \sigma^2$$

It follows that $Var(y_{ij}) = \sigma_\tau^2 + \sigma_\varepsilon^2$, where the error is also random.

Confidence Intervals

$\bar{y}_{..} \pm t_{\alpha/2,t-1} * SE(\hat{\mu})$

$SE(\hat{\mu}) = \sqrt{MST / tn}$, where MST is the treatment.

If there are two-factors, $SE(\hat{\mu}) = \sqrt{MSA + MSB - MSE / \text{ab}}$

SAS reports the "[y-var-name] mean" as the estimate of mu.

Nested Designs

Consider a company that purchases medical pills from three suppliers and the material comes in batches. We can randomly select four batches from each supplier and can then sample pills from each batch.

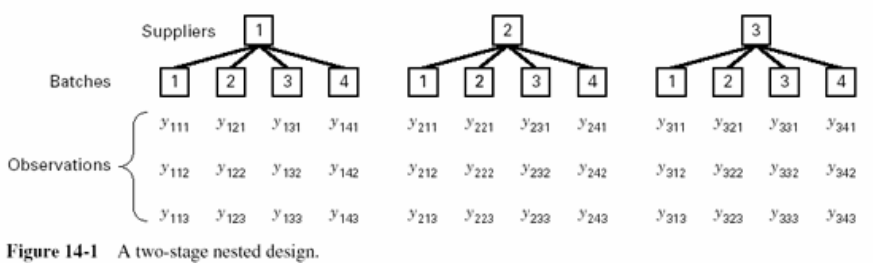


Figure 14-1 A two-stage nested design.

(typo- batches should be 1-12)

At first glance, we may want to treat each supplier as a block and measure the variability in the batches along with an interaction term. But this would be assuming that the batches are i.i.d. from each factory.

What if they're not? What if we don't know? Essentially the question that we want to answer is, "Is the purity of the material the same across suppliers?"

In a nested design, each level of factor (B) does not appear with each level of factor (A). It is not batch 1-4, repeated 3 times. It is batch 1-12. Because not every level of B appears with every level of A, there is no interaction between A and B.

It is not always easy to tell whether a variable is nested or not. There are two things to look for. The first thing to do is ask whether the data is crossed or hierarchical (nested). If the variable is intended to explain the variation in the higher-order variable (while also depending on it), it is likely to be hierarchical. However, in order to be nested, it must also be the case though that the treatments are not i.i.d. If the treatments are assumed to be coming from the same population (under the null), then it's a factorial design.

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{k(ij)}$$

The subscript j(i) indicates that jth level of factor B is nested under the ith level of factor A. Furthermore, it is useful to think of replicates as being nested under the treatment combinations; thus, k(ij) is used for the error term. (In most of our designs, the error is nested in the treatments, but we only use this notation for error when there are other nested factors in the design).

A design that includes some nested variables is a **partially nested design.**

AOV Table

Here is the design question: How many batches should you take and how many measurements should you make on each batch? How many batches versus how many samples per batch? It is the average of the batches and the variability across the batches that are most important.

At a minimum you need at least two measurements per batch (n = 2) so that you can estimate the variability within the batches, σ2, and at least two batches per supplier (b = 2) so you can estimate the variability among batches, σ2β.

				Expected Mean Squares		
Source	SS	df	MS	A&B Fixed	A Fixed, B Random	A&B Random
A	SSA	a-1	MSA	$\sigma_\varepsilon^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2 + bn\theta_\tau$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2 + bn\sigma_\tau^2$
B(A)	SSB	a(b-1)	MSB(A)	$\sigma_\varepsilon^2 + n\theta_{\beta(\tau)}$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2$	$\sigma_\varepsilon^2 + n\sigma_{\beta(\tau)}^2$
Error	SSE	ab(n-1)	MSE	σ_ε^2	σ_ε^2	σ_ε^2
Total	SST	abn-1				

In this example the model assumes that the batches are random samples from each supplier, i.e. suppliers are fixed, the batches are random, and the observations are random. It follows that:

$$\hat{\sigma}_{\beta(\tau)}^2 = \frac{MS_{B(A)} - MSE}{n} \qquad \hat{\sigma}_\tau^2 = \frac{MSA - MS_{B(A)}}{bn}$$

The F-statistics follow as well.

If there are no replications in the study, the MS for the nested variable can be used as the error for the higher-order variable.

Repeated Measures

Each level of a single treatment is administered to every individual *n*. By definition there is more than one observation per subject.

$$y_{ij} = \mu + \tau_i + \delta_j + \varepsilon_{ij}$$

Mu is the mean response, tau is the effect of the compound, gamma is the patient.

One assumption is that the errors are correlated across compounds but are independent across subjects. Another assumption is that of **compound symmetry**, where the covariance for any two observations from patient *j,* y_{ij} and $y_{i'j}$ is constant. This is displayed on a variance-covariance matrix.

Source	**SS**	**df**	**EMS**	
			MS	Fixed Effect, patients random
Patient	SSP	n-1	MSA	$\sigma_\varepsilon^2 + a\sigma_\delta^2$
Factor A	SSA	a-1	MSB	$\sigma_\varepsilon^2 + n\theta_\tau$
Error	SSE	(a-1)(n-1)	MSE	σ_ε^2
Total	TSS	an-1		

The first hypothesis measures the difference in factor A. You can also check the means for each class of the factor as well as calculate the variance component.

Two-Factor Experiments with Repeated Measures on One Factor

Each level of a treatment *m* is administered to each experimental unit *n* over time period *p*.

The **period effect** controls for exogenous effects that would occur over time, such as changes in patient disposition or the season.

The interaction controls for longer mean responses for different treatments. If the effect lasts equally long for all treatments, the interaction will not be significant.

Note that if different individuals are assigned to each treatment, then the design is completely randomized. This will reduce variation in the estimation of the treatment mean.

$$y_{ij} = period + treatment + EU(treatment) + treatment * period + \varepsilon_{ij}$$

Source	**SS**	**df**	**EMS**	
			MS	…
Treatment	SST	m-1	MST	**SEE BOOK PAGE 1020**
EU(Treatment)	SSE(T)	m(n-1)	MSE(T)	$\sigma_\varepsilon^2 + n\theta_\tau$
Period	SSP	p-1	MSP	
T*P	SSTP	(m-1)(p-1)	MSTP	
Error	SSE	M(p-1)(n-1)	MSE	σ_ε^2
Total	TSS	mpn-1		

The experimental unit is nested in the treatment. To measure the main effect of the treatment, we use EU(Treatment) as the error term to control for between subjects effects. The within-subjects effects include the time period, the interaction, and the error term MSE.

$$F_T = MST / MSEU(T)$$

Where the patient effect is random, the time period is fixed, the treatment effect is fixed, and the experimental error is random.

Crossover Designs

In a **crossover design**, each experimental unit is assigned to a **sequence** with *t=p* periods and receives a level of the treatment for each period. The experimental unit is nested in the sequence.

If all experimental unit received the same treatments in the same sequence of periods, the treatment effect would be confounded by the period effect. **(See Ex 18.5)** To avoid this problem, multiple sequences must be applied to each experimental unit. There are p! possible periods. *r* Patients could be randomly assigned to each of the sequences, or a subset of *n* periods selected.

Because the treatments are compared on the same units, between unit variation is greatly reduced. Each experimental unit serves as a block in order to reduce the SSE.

Source	**SS**	**df**	**EMS**	
			MS	…
Sequence	SSS	n-1	MST	**SEE BOOK PAGE 1028**
EU(Seq)	SSE(S)	n(r-1)	MSE(T)	$\sigma_\varepsilon^2 + n\theta_\tau$
Period	SSP	p-1	MSP	
Treatment	SST	t-1		
T*P	SSTP	n-1	MSTP	
Error	SSE	nt(r-1)	MSE	σ_ε^2
Total	TSS	tpr-1		

$$y_{ijk} = seq_i + period_k + treatment_{d(i,k)} + EU(sequence)_{j(i)} + treatment * period + \varepsilon_{ij}$$

Where the sequence effect is fixed, the patient effect is random, the time period is fixed, the treatment effect is fixed, and the experimental error is random.

If you know the period and the sequence, you'll know the treatment.

We always test the **carryover effect** first, which essentially means we include an interaction term (see above) to measure the lingering effect of the previous treatment. This is also called a **washout period**. If the carryover effect (or the period effect) is significant, then the results are invalidated except for the first period. This becomes a CRD parallel design where Y = treatment.

After rejecting the carryover effect, you can re-fit the model without the interaction term, and test the treatment and also the sequence (by dividing the sequence by the EU(sequence).)

With sufficient washout, there will not be a period effect. With homogeneous subjects, there will not be an EU effect. With randomization of EU to sequence, there shouldn't be a sequence effect. That leaves the treatment.

There are two experimental units. The experimental unit for the sequence is the subject. The experimental unit for the treatment is the time period.

Bivariate Data Sets

introduction

A **bivariate data set** has a pair of values for each observation. As such, no single value can describe an observation for a bivariate data set.

What we are focused on is the various aspects of the relationship between the two variables.

Data sets are basically tables. The variable(s) are listed in the first row on the top of each column. The individuals are listed in each row.

Datasets come in different forms, depending on the variables they describe.

Datasets can be thought of in two dimensions, based on (1) the individuals that they describe and (2) the time period over which the individuals are described.

	instant	over time
same n	cross-sectional data	panel data
different n		pooled cross-section or longitudinal

Bivariate Data Sets

cross-sectional data sets

A **cross-sectional data set** consists of sample individuals taken at a given point in time. Think of it like a photograph or a snapshot.

Cross-Sectional Data Set					
individual	wage	education	experience	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
...	3.00	11	2	0	0

Bivariate Data Sets

pooled cross-sections

A **pooled cross-section** data set describes the same variables over time, but the individuals are different.

For example, the census is taken in 1985 and in 1990. The same individuals aren't sampled every 5 years.

Pooled Cross Section						
Observation	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2
2	1993	67300	36	1440	3	2.5
...	...	...	...	...	...	...
251	1995	243600	16	1250	2	1
252	1995	65000	20	2200	4	2

The year the data was collected is just another variable on the spreadsheet.

The **data frequency** is the number of times the data is collected. The most common frequencies are daily, weekly, monthly, quarterly, and annually.

Bivariate Data Sets

panel data aka longitudinal data

A **panel data** or **longitudinal data set** measures variables for the same individuals over time. Panel data comes from the same group of people like on a game show.

This is the holy grail of data because taking multiple observations on the same individuals over time allows us to control for certain unobserved characteristics of individuals.

Panel Data Set						
observation	individual	year	murders	pop.	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	651000	5.5	75

In a spreadsheet, you need to distinguish observations from individuals. There is a unique row for each observation, while there are two observations for each individual.

As you have seen, depending on the data set, the individuals that are being followed and the time periods are both variables as well. More generally, each individual will have t rows for t time periods.

Bivariate Data Sets

time series data

A special case is **time series data set** consists of observations on a single individual over time. Although the data set follows a single individual, it can contain many variables relating the individual.

For example, the time series data set below follows a single country from 1950 onwards but records many variables: the gnp, the unemployment rate, etc.

Time Series Data Set						
observation	year	avgmin	avgcov	unemp	gnp	t-1
1	1950	0.2	20.1	15.4	878.7	1949
2	1951	0.21	20.7	16	925.0	1950
...	...	...	...	...	...	...

Because past events can influence future events (history is not independent), previous years are often included as variables in a data set. For example, the data set above includes the variable "t-1" which lists the previous year. These variables are called **lags**.

A **time plot** of a variable plots each observation against the time at which it was measured, a day, week month, quarter, year.

Trends can be emphasized by connecting the points by a jagged line. You can also add a **trend line**- a straight line that heads in the direction of the trend.

Seasonal variations are patterns that repeat themselves over regular time intervals.

Data can be **seasonally adjusted** to control for the effect of such seasonal variations. The data is "decomposed" into separate components- trends, seasonal variation and residuals (explained later).

You can plot times series for different groups by using different lines.

Scatterplots

introduction

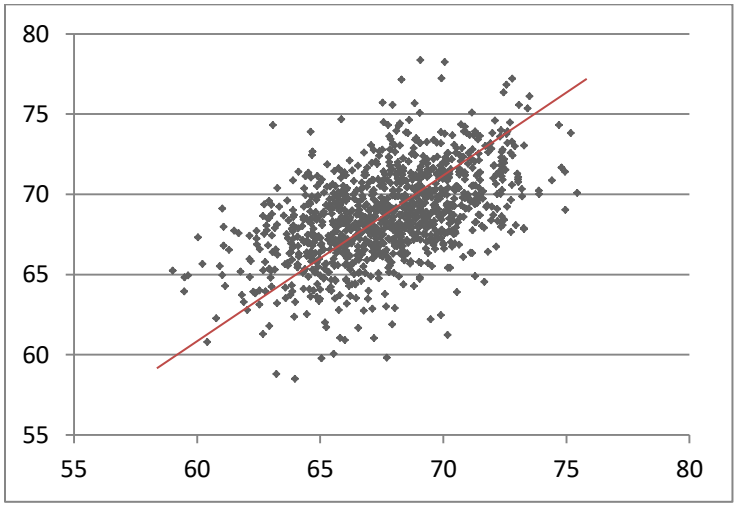
Statistics is a relatively new subject. It wasn't until the late 1800's that Sir Francis Galton created the first **scatterplot** in order to measure the relationship between two variables, i.e. **bivariate data**.

Galton and his assistant, Pearson, were fascinated by the idea of quantifying hereditary influences. They measured the height of 1078 fathers and their 1078 sons.

Galton and Pearson were interested in understanding how heights varied across generations. Are fathers taller than their sons, or are they shorter, or neither? How do we answer this question statistically?

Scatterplots allow us to leverage our analytical skills from the level of the individual to the level of the group. We can study the relationship between the variables as a whole instead of looking at each father/son pair.

We create a scatterplot by placing the observations on a set of Cartesian coordinates, where each axis represents a single variable. Each pair of values are represented by a single point on the scatterplot that is fixed by its corresponding values for both variables.



Imagine a line going through the data set. This is the **regression line**. It is the line "of best fit." We will explore what this means later. First, we will explain the methods behind interpreting bivariate data on scatterplot.

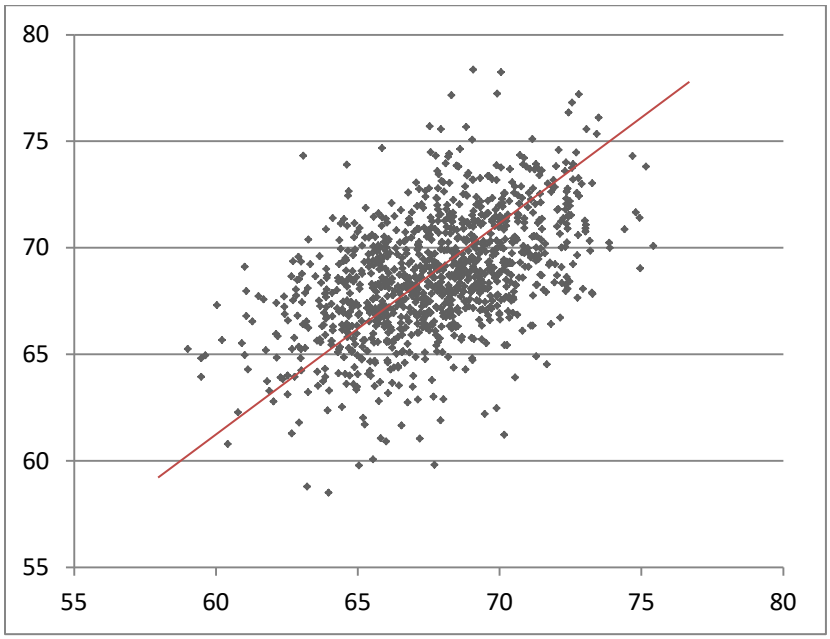
y	X
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control Variable
Predicted variable	Predictor variable
Regressand	Regressor

Scatterplots

direction

The **direction** of a scatterplot describes the slope of the regression line, that is, whether the regression line is positively sloped or negatively sloped.

As a whole, if the x-values are increasing as the y-values are increasing, then direction of the regression line is positive. We can say that there is a **positive association** or a **positive relationship** between the two variables.



We can see from Galton's data that taller fathers tend to have taller sons.

When above-average values of one variable tend to accompany below-average values of the other, the slope of the regression line is negative. Ergo, there is a **negative association** or an **inverse relationship** between the two variables.

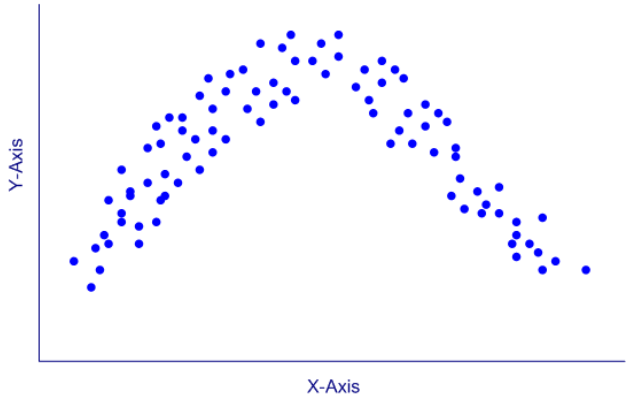
Scatterplots

form

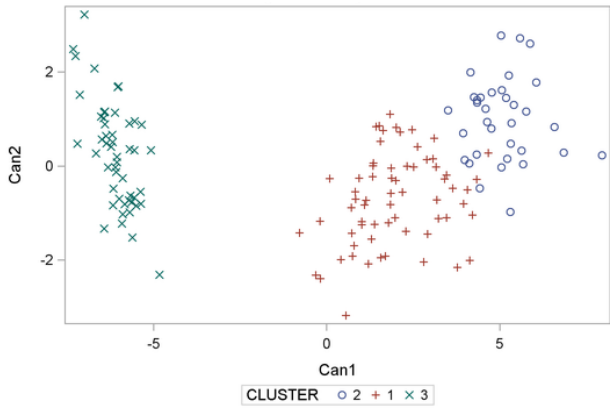
The **form** or **shape** of the bivariate data set, as projected on a scatterplot, describes the general shape of the data.

The most common shape of a scatterplot is **linear**- the points are more-or-less evenly scattered around a straight line. Galton's data is fairly linear.

Non-linear association or **curved**- The data are scattered more-or-less evenly around a curve, which may change direction more than once.



A **clustered** data set essentially has multiple shapes. This is usually the case when a variable is broken down categorically.



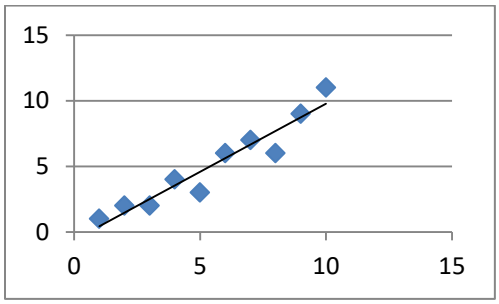
Scatterplots

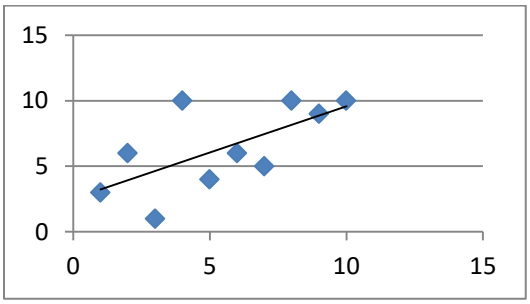
spread and association

Finally, we come to **spread**. Spread is analogous to variance for a single variable.

For single variable descriptions of spread, we measured the distance from each observation to the mean. For bivariate descriptions of spread, we measured the distance from each observation to the regression line.

This distance is usually measured vertically. If there is a lot of spread, then the observations are far from the regression line. If the points fall almost perfectly on a line, then there is minimum spread.



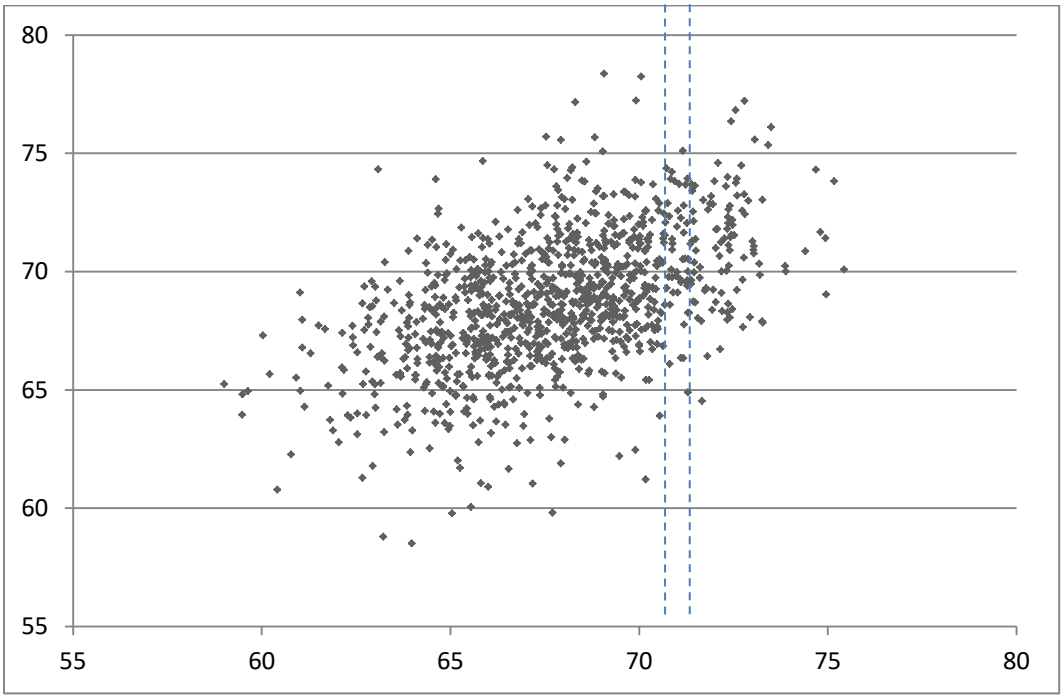


Scatterplots

Subpopulations and chimneys

A **subpopulation**, $x = x^*$, consists of a class of individuals having the same or approximately the same value of x. You can study the spread of the observations whose x-value is equal to $x = x^*$.

Chimneys isolate scatterplots into vertical strips. They are pictured as dashed lines on both sides of an x-value. All the values within the chimney all have the same value of x.



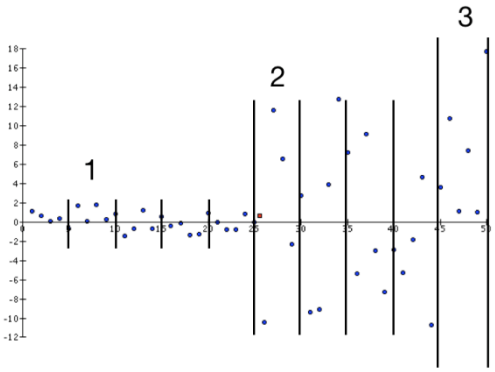
We can measure the average and standard deviation (i.e. spread) for each subpopulation. We can go so far as to construct frequency histograms of y in a given sub-population.

Scatterplots

scedasticity

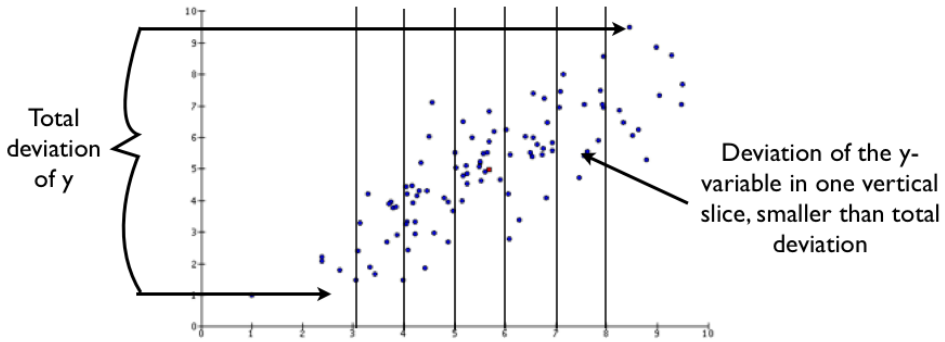
"Scedastic" means scatter. The amount of scatter is simply the amount of spread. One corollary question is, is the amount of spread consistent across all subpopulations of x?

Heteroscedasticity describes varying degrees of spread for each subpopulation. The scatter in vertical slices depends on where you take the slice.



Subpopulation 1 has a smaller standard deviation than subpopulation 2 or subpopulation 3. This obviously warrants further investigation.

Homoscedastic data is generally "football shaped."



Although the total spread in y across all of x is much larger than the deviation within any single chimney, the spread in each chimney is approximately the same size.

Scatterplots

regression to the mean

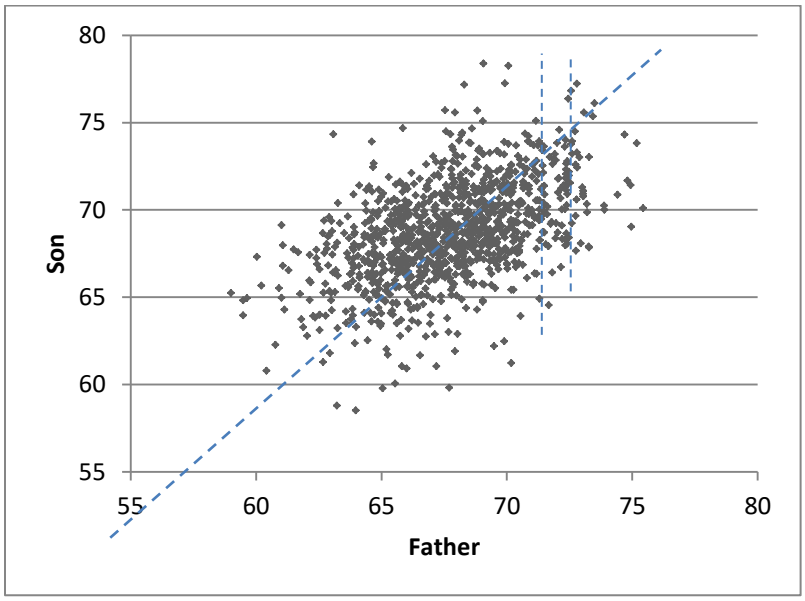
Regression toward the mean (RTM) is when a sample statistic is extreme due to an extreme value in the chance error, and we mistakenly assume that the sample statistic is representative of the population parameter in that particular case.

For example, the "sports illustrated" cover jinx. Or, when we punish groups that do poorly, they will probably do better next time.

The underlying mechanism is that observations with extreme values revert to the mean value during repeated testing of the same individuals.

When we look at a subpopulation of fathers whose height is larger than average, we expect the percentage of sons that are shorter to be greater than the percentage of sons that are taller. The father is the outlier, the son the regression (hence the term).

Say we're looking at the subpopulation of fathers who are one standard deviation taller than the average. They're 72" tall.



The number of observations within the chimney and above the SD line is less than the number of observations within the chimney and below the SD line.

As it turns out, a larger proportion of sons are shorter than their fathers than taller than their fathers. This is regression to the mean in action.

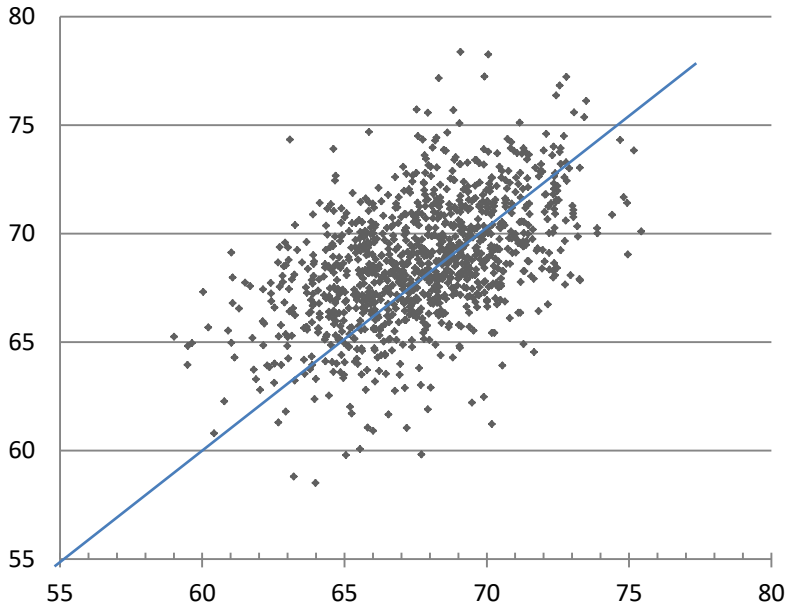
Lines

45 degree lines

If there is a linear association between the two variables, then we can use various lines to help us interpret more subtle features of this relationship.

The first line is the **45 degree line**, y = x. Apply this line to your data set when you want to examine how closely the two variables have a 1-to-1 relationship.

In Galton's data set, fathers and sons of equal height will lie on the 45 degree line. In the chart below, the 45 degree line highlights the fact that there are more observations above the line than below. This implies that, in this sample, there are more sons who are taller than their fathers than vice versa.



A 45 degree line is also useful in any case where there is a "before-and-after" treatment applied to a single individual. For instance, we might measure the weights of a particular set of people, before and after they begin an exercise regimen. In this case, the 45 degree line serves as a metric of the status quo.

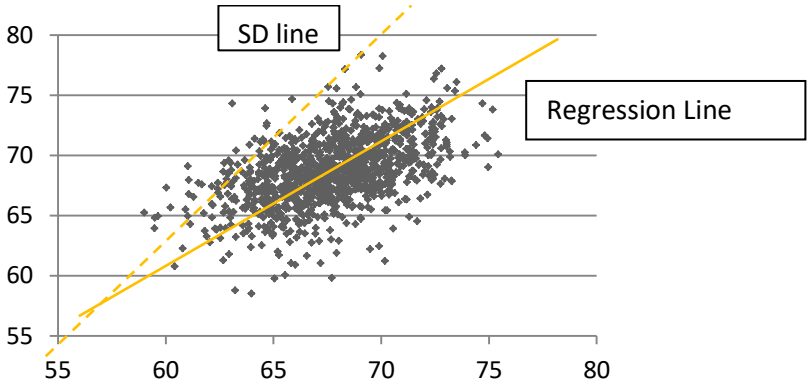
Lines

the standard deviation line

The **standard deviation line**, y=sdy/sdx, is a dashed line that measures the relationship between the variances of X and Y.

The standard deviation line goes through the points that are the same number of standard deviations from the mean, for example, the points $(\bar{x}+\sigma_x),(\bar{y}+\sigma_y)$ or $(\bar{x}-2\sigma_x),(\bar{y}-2\sigma_y)$.

The standard deviation line allows a graphical interpretation of the covariance. If a one standard deviation increase in x suggests a predicted increase in y by less than on standard deviation, than the observation will lay below the SD line.



Lines

the regression line

Generally, the regression line is the line that best fits a linear relationship to the data. More specifically, the regression line minimizes the root mean square of the residuals. We'll cover this derivation later.

$$\hat{y}_i = b_o + b_1 * x_i$$

We can actually derive the line quickly if we use the 5 key statistics: $\bar{x}, \bar{y}, s_x, s_y, r$.

We can measure **the slope of the regression line**, b1, by multiplying the slope of the SD line by the correlation coefficient.

$$b_1 = \frac{s_y}{s_x} * r$$

If the correlation coefficient is equal to 1, the standard deviation line and the regression line will have the same slope.

The point of averages falls on the regression line, so we can plug in these values, along with the slope, to find **the y-intercept of the ordinary least squares regression line**, b_0 .

$$\bar{y} = b_o + b_1 * \bar{x} \rightarrow b_o = \bar{y} - b_1 * \bar{x}$$

Covariance

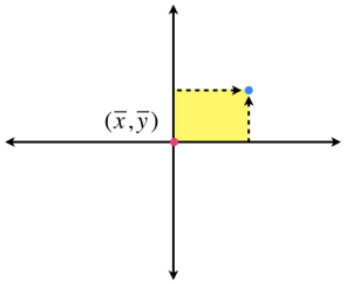
intro

When we measure variance for two-variables, we call it **covariance**.

Measuring covariance means that we measure how much the variable vary *together*. When one variable increases, does the other variable increase or decrease?

Let's define covariance using a single set of observations, (xn and yn).

First, we measure the deviation score from the **point of averages**, $(\bar{x}, \bar{y})$.[1] We've graphed some Cartesian coordinates with the point of averages at the center. We can measure the deviation scores on both the x and y axis.



Next, we multiply those distances. We're forming our sum of squares here, except it no longer is necessarily a square per se:

$$(x_i - \bar{x}) * (y_i - \bar{y})$$

This is called the **cross product** because it is the product of two individual deviation scores:

Observation	Deviation	Deviation	Cross Products X,Y
1	$(x_1 - \bar{x})$	$(y_1 - \bar{y})$	$(x_1 - \bar{x})\ (y_1 - \bar{y})$
2	$(x_2 - \bar{x})$	$(y_2 - \bar{y})$	$(x_2 - \bar{x})\ (y_2 - \bar{y})$
3	$(x_3 - \bar{x})$	$(y_3 - \bar{y})$	$(x_3 - \bar{x})\ (y_3 - \bar{y})$
	SST	**SST**	**SSCP**

The **sum of squares of cross products** (**SSCP** or **Sum of X-P**) is

$$\sum_{i=1}^{N}[(x_i - \bar{x}) * (y_i - \bar{y})]$$

The SSCP contains the total amount variability for all observations. That's a large area, and it contains all of the variability in the data set.

---

1 In the previous section, we defined spread as the distance from the regression line, which has a complicated derivation in and of itself. Since this section involves formal math, we're going to start with a single point rather than a line.

Covariance

defined

Covariance is the SSCP divided by the sample size. Think of the covariance as the multiplied distances that the "typical" observation is from the point of averages.

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{n}$$

The sign of the covariance allows you to determine whether two variables are positively or negatively associated.

Positive covariance indicates a positive association, **negative covariance** indicates a negative association, and **zero covariance** suggests that the two random variables are independent.

How can covariance be 0? Recall that when taking the square root of the distance for a single variable, the square is always becomes positive. However, a distance between an observation and its mean can be negative in one direction but positive in another.

An alternative formula for the covariance shows it as the average of the products of x and y minus the product of their means.

$$\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$$

Covariance

Correlation coefficient

The **correlation coefficient**[2], *r*, is the most commonly used measure of **linear association**. The correlation coefficient measures how tightly the data cluster around the regression line.

Mathematically, the correlation coefficient is simply the covariance, standardized. The distances between the observations and their mean are expressed in standard units.

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)}{n}$$

A correlation of 1 or -1 is called a **perfect correlation**- each observation falls on the regression line. A correlation coefficient of +1 means the data has a perfectly linear, positive association. A correlation coefficient of -1 means the data has a perfectly linear, negative association.

A correlation of 0 means there is no linear relationship between the two variables. This is equivalent to the statement that x and y are independent.

Correlation values of 0.5 or higher up to 0.8 denote a weak association. Correlation values less than 0.5 are considered weak.

We can quickly move from the covariance to the correlation coefficient by dividing the covariance by the product of the standard deviations of each variable.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Also,

$$r = \frac{sscp}{\sqrt{SS_x * SS_y}}$$

Or, r can be the average of the cross-products.

$$r = \frac{\sum z_x z_y}{N}$$

[2] aka **Person Product Moment Correlation Coefficient** or **Person's r**. The population correlation is symbolized by ρ, pronounced "rho."

Covariance

the regression method

We can use the correlation coefficient to figure out the **fitted value** or **predicted value** of y, $\hat{y}_i$, given any value of x.

The term "predicted" is something of a misnomer because nothing is being predicted. Prediction implies forecasting or looking into the future. While $\hat{y}_i$ can be used for these purposes, it does not need to be.

More simply, the predicted value is just our "best guess" of y within a subpopulation x*.

Let's calculate $\hat{y}_i$ with a sample.

Say a test is given to all students at the end of the first year of law school, Y. Furthermore, we have the data on these students LSAT score, X. $\bar{x} = 162, s_x = 6, \bar{y} = 68, s_y = 10, r = 0.6$

Here is the question: of the students who scored 165 on the LSAT, what is their predicted end-of-year score?

We answer this question using the **regression method**.

$$\hat{y}_i = \bar{x} + \left(\frac{x^* - \bar{x}}{\sigma_x} \right) * \sigma_y * r$$

1. First, we measure how many standard deviations above or below the point of averages is x*.
2. We multiply this value by the standard deviation of y because we want to increase y-bar by the same number of standard deviations. This is a value on the SD line.
3. Then we weigh it by the correlation coefficient, to get a value on the regression line.
4. Finally, we add this value to x-bar. Think of this step as just orienting ourselves to the point of averages.

$$\hat{y}_i = \bar{x} + \left(\frac{x^* - \bar{x}}{\sigma_x} \right) * \sigma_y * r \rightarrow 68 + 0.5 * 10 * 0.6 = 71$$

Covariance

prediction error and the r.m.s. error

For any observation, the distance between the actual y-values and the predicted value of y is called the **prediction error** or **residual**, e_i **.**

$$e_i = \hat{y}_i - y_i$$

The residual represents all the other variables that can effect y besides x including randomness.

The **root mean square error** (**r.m.s. error**) is the length of the typical residual.

$$s_{e_i} = \sqrt{\frac{\sum e_i^2}{n}}$$

The goal of constructing a regression line is to minimize the root mean square of errors.

The **residual coefficient** is a shortcut to figuring out the r.m.s error.

$$\sqrt{1-r^2} * \sigma_y = r.m.s.e$$

All three variables measuring bivariate variability in one lovely equation.

The Regression Model
introduction

The multivariate regression model uses an **additive function** to estimate the partial effect of each variable.

The primary goal of multivariate regression is to estimate the **partial effect** of one variable on another. The partial effect of an explanatory variable is its effect on a dependent variable in isolation from other explanatory variables ("ceteris paribus").

The functional form of the multivariate regression model:

$$E(y \mid x_k) = \beta_0 + \sum \beta_k X_k$$

Multivariate analysis the most popular tool for empirical analysis because it can be used to simulate experiments using observational data as long as certain assumptions hold. All else is held equal. Other variables are "controlled for" or "held constant".

The partial effect can be interpreted in the context of an experiment or, in economics, an elasticity:

Experiment: if we compared 2 observations that were identical in all other ways except for X, how do the observations differ in Y when X varies?

Elasticity: if a randomly chosen observation had a change in x, what is the change in y?

The Regression Model
the population model

The **population model for multiple regression** is:

$$E(y \mid x_1, x_2, ..., x_p) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + ... + \beta_p x_p + u_i$$

Like ANOVA, we can break this equation down into systematic variance and unsystematic. $E(y \mid x_1, x_2, ..., x_p)$ is the **systematic** part of the regression and the error term is **unsystematic**.

$E(y \mid x_1, x_2, ..., x_p)$ is the expected value of y for the given values of x_p, where *p* is the number of explanatory variables. $E(y \mid x_1, x_2, ..., x_p)$ is called the **mean response** and can be abbreviated as μ_y.

β_p measures the **partial effect** of x_p on the mean response of y. It is more commonly called the **regression coefficient**.

u_i represents the **error term**, **disturbance**, or **unobservable**. The error term "contains" all of the variance due to the variables that we are not measuring, including variables that are impossible to measure. This term can be thought of as randomness. The error term should not be confused with the residuals, which are computed from the data.

The Regression Model
the sample regression function

The sample regression function or the ordinary least squares (OLS) regression line is an estimate of the population model using actual data.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}$$

Note that there are two indices. The first index measures the observation number and the second represents the order of the *p* explanatory variables.

$\hat{y}_i$ is the **predicted y.**

$\hat{\beta}_p$ is the **estimated partial effect**

$\hat{u}_i$ or ε_i or $\hat{e}_i$ are the **residuals** or the **prediction error**

The hats indicates that it is an estimated equation.

The parameters of a multivariable equation is presented on a **table of coefficients**, which lists all of the variables in the first column (along with the y-intercept), and the regression coefficients, standard errors, t-values, p-values and possibly more.

You can plug in any value for X to get the estimated Y:

$$\Delta\hat{y} = \beta_0 + \beta_1 \Delta x$$

If utilizing the formula above, keep in mind that all of the other variables are assumed to equal to their average value.

The Regression Model

The coefficient of determination, R^2

The population multivariate regression function also includes a measurement of model's accuracy, ρ^2 or "rho".

Its equivalent sample statistic is the **coefficient of determination**, R^2.

> R^2 is used to compare competing models. It measures the percentage of variation in y that is explained by the regression model's explanatory variables.
>
> By contrast, R is equal to the square of the sample correlation coefficient between the predicted scores and the observed scores, $R_{\hat{Y}Y}$.

A low R^2 does not imply that our estimate partial effect is biased. You can often accurately estimate a partial effect with a low R^2. A low R^2 means that we have not accounted for several factors that affect y.

The Regression Model
residuals explained

The residuals are the difference between the predicted y (from the regression) and the actual value of y (from the data).

$$\varepsilon_i = \hat{y}_i - y_i$$

In most cases, residuals are not equal to 0. In other words, none of the predicted y-values lie on the OLS line.

Obs #	x_{i1}	x_{i2}	y_i	$\hat{y}_i$	$\hat{e}_i$
1	14.1	12.7	1095	1224	-129
2	10.9	8.5	1001	1164	-163
3	23.5	25.5	1122	1397	149

β_0 and β_1 are chosen in such a way to make the residuals add up to 0. The sum and therefore the sample average of the OLS residuals is 0.

$$\sum_{i=1}^{e_i} \hat{e}_i = 0$$

.

It is also true that,

$$\overline{\hat{y}} = \overline{y}$$

And

$$\text{cov}(\hat{y}_i, \hat{e}_i) = 0$$

Economists enjoy **residual analysis** because it can be used to find unmeasured value.

“The house with the most negative residual is… the most underpriced one relative to its observed characteristics.”

Assumptions of the Regression Model

introduction

There are various assumptions behind doing regression:

1. X and Y are normally distributed
2. Reliability & Validity

It is always good to examine these assumptions by plotting histograms and scatterplots.

For the regression function to be an unbiased estimate of the population model, the **Gauss-Markov assumptions** must be satisfied. A regression model must be:

1. Linear Parameters
2. Random Sampling
3. No Perfect Collinearity
4. Zero Conditional Mean
5. Homoscedasticity

If a Gauss-Markov assumption is violated, then $\hat{\beta}_j$ and R^2 are biased[1].

$$E(\hat{\beta}_j) \neq \beta_j \quad E(R^2) \neq \rho^2$$

This section will address the Gauss-Markov assumptions as well as the normality assumption which is used for statistical inference.

[1] Bias is not a feature of a sample regression function. It is possible that our regression may be unbiased. The partial effect and coefficient of determination might be spot on. The problem is, we don't know!

Assumptions of the Regression Model

GM1: linear in parameters

The mean response is a linear function of the explanatory variables. That is, each explanatory variable has a linear relationship with y. Linearity implies the same partial effect β_j regardless of the value of x.

We can test for linearity by creating scatterplots for each independent variable and Y.

If the partial effect is non-linear, then we can transform the functional form of either x or y to allow for nonlinear relationships, but we are still restricted to linear parameters, β.

Some non-linear effects may not be captured by transformations, for example, “diploma effects” for schooling on wage.

Assumptions of the Regression Model

GM2: Random Sampling

Is the sample randomly selected? This is a meta-question that is to be asked of the data set.

If the sample wasn't randomly selected, it doesn't really matter what variables are held constant, you're open to the selection bias.

Sampling or **selection bia**s is a systematic bias caused by a non-random sample of a population for an observational study, causing some members of the population to be more likely selected.

During WWII, statistician Abraham Wald was asked to help the British decide where to add armor to their bombers. After analyzing the returned planes, he recommended adding more armor to the one spot on the plane where none of the returning planes were damaged. Wald surmised all the planes with damage in that particular spot had been shot down!

Not all cross-sectional samples can be viewed as outcomes of random samples, but many can be.

Assumptions of the Regression Model

GM3: No Perfect Collinearity

In the sample, no two explanatory variables are perfectly correlated.

The simplest way that two independent variables can be perfectly correlated is when one variable is a constant multiple of another.

A **correlation matrix** measures the correlation between all pairs of multiple variables, including residuals. The value in each cell represents the correlation between variables occupying the row and column headers.

	x_1	y	$\hat{y}_i$	$\hat{y}_i - y_i$
x_1	1			
y	.72	1		
$\hat{y}_i$	1	.72	1	
$\hat{y}_i - y_i$	0	.69	0	1

The correlation matrix will allow you to check for perfect collinearity.

Assumptions of the Regression Model

GM4: zero conditional mean assumption

The **zero conditional mean assumption** is that there are no **lurking** or **omitted variables**. There is no variable that is statistically related to both Y and x_k.

This is usually defined in terms of x_k and *u* rather than x_k and *y*.

Formally, all unobserved or unmeasured variables have an expected value of 0 for all values of all independent variables.

$$E(u \mid x_1, x_2, ..., x_p) = 0$$

An **exogenous** variable is included in the model and is uncorrelated with the unsystematic variance. An **endogenous** variable is included in the model and is correlated with the unsystematic variance. You want your variables to be **exogenous** and you want to avoid **endogeneity**.

Violation of this assumption is almost always a concern in regression analysis with observational data, since it is impossible to account for all unmeasured variables.

The defense should go something like this:

> Challenge: "What about variable Z that you did not include in your model? Doesn't excluding this variable bias the model?"
>
> Retort: "While we can not explain the variance in Y due to Z, our model is not biased under GLM4 because x_k is not correlated with Z."

Assumptions of the Regression Model

GM5: homoscedasticity

The homoscedasticity assumption states that the variance of the unobservables is constant, conditional on the value of each of the explanatory variables.

$$Var(u \mid x_p) = \sigma^2$$

Both $Var(u \mid x)$ and $Var(y \mid x)$ are equal to the error variance. When the homoscedasticity assumption is satisfied,

$$E(\hat{\sigma}^2) = \sigma^2$$

If $E(\hat{\sigma}^2) = \sigma^2$, then we have unbiased estimators of $Var(\hat{\beta}_k)$.

Heteroskedasticity or **non-constant variance** is present whenever σ^2 is dependent on x. This implies that the estimators of the variances, $Var(\hat{\beta}_j)$ are not equal to the variance in the population.

This does ***not*** imply that the expected values of either the partial effect $E(\hat{\beta}_j)$ or the coefficient of determination $E(R^2)$ are biased and are therefore not equal to their respective means on the sampling distribution of sample means.

Assumptions of the Regression Model

Breusch-Pagan

(This is jumping ahead here, but whatever.)

We may wish to test for heteroskedasticity. The null hypothesis would look something like this:

$$H_0 : Var(u \mid x_1, x_2, ..., x_k) = \sigma^2$$

We use the **Breusch-Pagan** regression model:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + v$$

where δ_k is the slope, u^2 is the squared unobservable for each value of x_k, and v is the unobservable.

The **Breusch-Pagan** sample regression function is:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k$$

We want to test whether u^2 is related to one or more of the IV's.

$$H_0 : \delta_1 = \delta_2 = ... = \delta_0 = 0$$

If we cannot reject H_0, we conclude that heteroskedasticity isn't a problem.

We can compute the F-statistic for joint significance in order to test the null hypothesis:

$$F = \frac{R^2_{\hat{u}^2} / k}{(1 - R^2_{\hat{u}^2}) / (n - k - 1)}$$

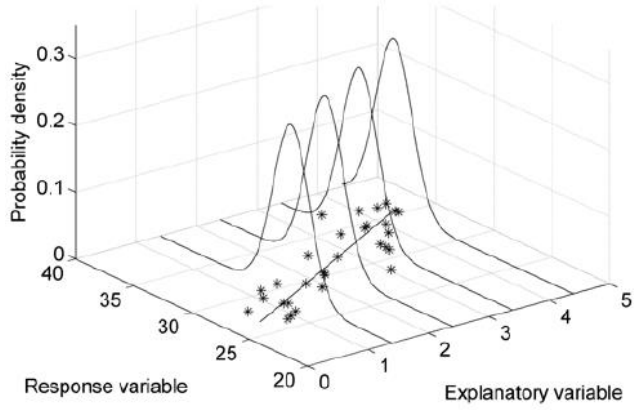
where k is the number of regressors. The F-stat has approximately an $F_{k,n-k-1}$ distribution under the null.

We can use **weighted least squares** to explicitly account for different variances in the errors.

Assumptions of the Regression Model
normality

We add one final assumption when performing inference with the regression model- this is the **normality assumption**.

Within any subpopulation, the distribution of $\hat{y}$ is centered about $E(y \mid x)$.



The normality assumption is more formally stated in terms of the error term. The unobserved error is normally distributed in the population.

$$u \sim N(0, \sigma^2)$$

u is the same across all slices of x.

For small samples, we need to argue for a normal distribution for u when performing inference. With larger samples, we can drop the normality assumption, because we're covered by the CLT.

If we make this assumption, we are also necessarily assuming MLR4 and MLR5.

Sometimes this is clearly false, such as when y is a binary variable.

Inference with Regression
introduction

Inference can be performed on various pieces of the regression model, most notably the regression coefficient and the correlation coefficient.

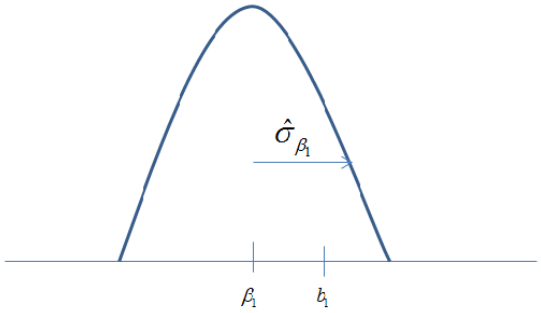
This section focuses on inference for:

1. The regression coefficient
2. The mean response
3. The predicted response
4. The correlation coefficient

Inference with Regression

regression coefficient as a random variable

We can imagine the regression coefficient as a random variable that is some distance from its parameter on a sampling distribution.



Each regression coefficient beta has a standard error as reported in the table of coefficients.

The formula for the standard error:

$$SE_{\hat{\beta}_j} = \frac{\hat{\sigma}}{\sqrt{SST_j}} = \frac{\hat{\sigma}}{\sqrt{\left(\sum x_j - \bar{x}\right)^2}}$$

For multiple explanatory variables, standard errors have slightly different formulas:

$$SE_{\hat{\beta}_j} = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R^2)}}$$

Given Gauss-Markov, we assume that the potential sample regression coefficients fall on a normal distribution.

$$\hat{\beta}_j = N[\beta_j, Var(\hat{\beta}_j)]$$

Therefore,

$$\frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)} \sim N(0,1)$$

Inference with Regression

confidence intervals and hypothesis tests for the regression coefficient

A **level C confidence interval for** β_j is

$$b_j \pm t^* SE_{b_j}$$

where SE_{b_j} is the standard error of b_j and t* is the critical t-value.

The hypothesis test for β_j tests whether the regression coefficient is equal to some constant. The most common constant is 0, suggesting that the regression coefficient has no partial effect.

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

In this case, the t-statistic is:

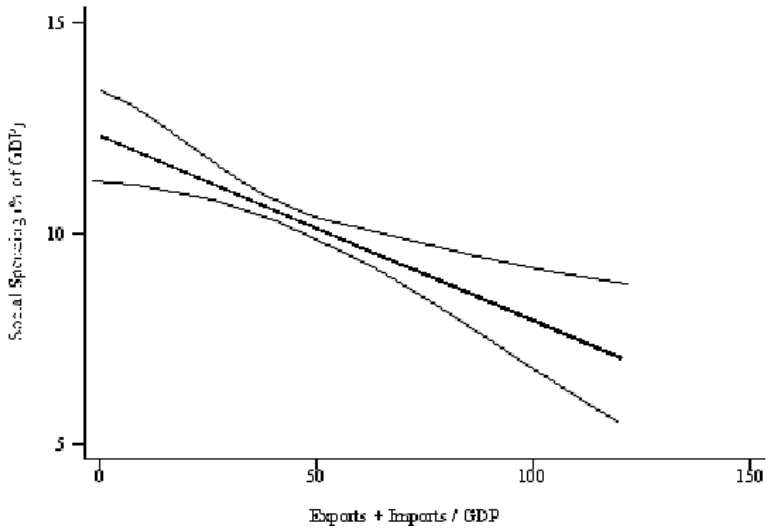
$$t = \frac{(\hat{\beta}_j - 0)}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

A larger beta leads to higher values for the t-stat.

Inference with Regression

confidence intervals for the mean response

We can imagine a confidence interval for the entire regression line. The confidence interval for the mean response shows the range of expected regression lines, with the population regression line falling somewhere in between. We can get an idea of how steep or how shallow the population regression line is.



Confidence intervals for the mean response will change depending on the distance from the point of averages.

This is usually computed automatically.

The standard error of the mean response is

$$SE_{\hat{\mu}_y} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

A level C confidence interval for the mean response is

$$\hat{\mu}_y = t^* SE_{\hat{\mu}}$$

Inference with Regression

predicted response

We estimate the predicted response, $\hat{y}_i$, given certain constant values (c_k) for the independent variables.

$$\theta_0 = E(y \mid x_1 = c_1, x_2 = c_2, ..., x_k = c_k)$$

We estimate the predicted response by creating new variables that subtract the constant from each variable:

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1(x_1 - c_1) + \hat{\beta}_2(x_2 - c_2) + ... + \hat{\beta}_k(x_k - c_k)$$

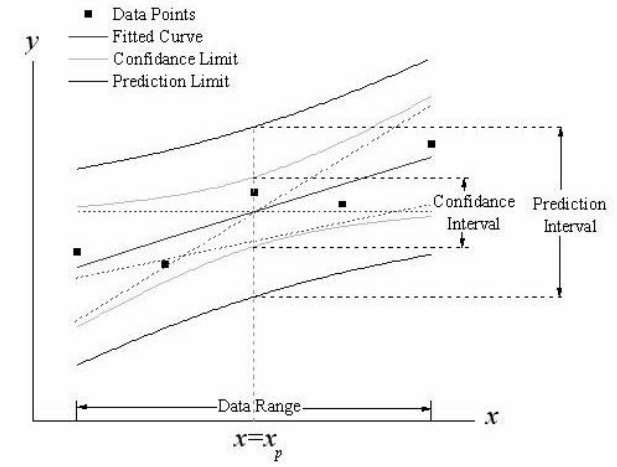
The predicted value and its standard error are obtained from the intercept.

A confidence interval for the predicted value is

$$\hat{\beta}_0 \pm t * SE_{\hat{\beta}_0}$$

Note that this doesn't change the slope, only the intercept.

A **prediction interval** is the confidence interval for a predicted response, $\hat{y}_i$.

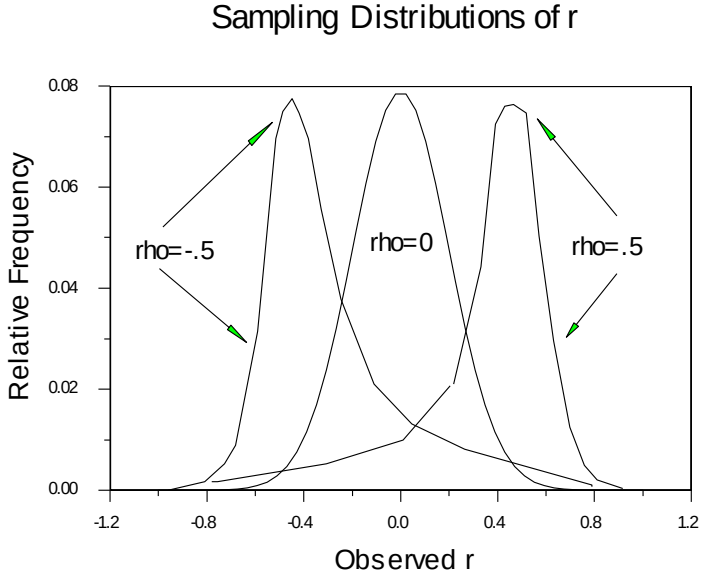


The confidence interval for a prediction interval is wider than for a confidence interval for the mean response because there is more variation around predicted responses for individual outcomes.

Inference with Regression

inference with ρ

One can imagine r existing on the **sampling distribution of the correlation coefficient**. At the center of this standardized sampling distribution is the parameter ρ.



ρ=0 implies that two variables are independent. If we calculate an r that is far from 0, then we may use inference to show that the two or more variables in our population are not independent.

For estimating regression coefficients from a sample, R^2 is biased because it includes chance fluctuations in the explained sum of squares.

To reduce this bias, use the **adjusted r-squared,** $\bar{R}^2$ **,** or the **shrunken estimate**:

$$\bar{R}^2 = 1-(1-R^2)*\frac{N-1}{N-k-1} = 1-\frac{[SSR/(n-k-1)]}{[SST/(n-1)]}$$

This estimate accounts for the number of variables. It can go up or down when a new independent variable is added to a regression. In fact, the adjusted r-squared can actually be negative.

We'll be looking at confidence intervals for ρ and three hypothesis tests using $\bar{R}^2$**.**

Inference with Regression

confidence intervals and hypothesis test for independence

A confidence interval for estimating ρ:

$$r \pm z^* m$$

For this test, we use **Fisher's *r* to *z* Transformation** to calculate the z-critical value.

$$z = .5 \ln\left(\frac{(1+r)}{(1-r)}\right)$$

This transformation pulls out short tail to make better (normal) distribution.

A hypothesis test that two variables are independent states that:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

To create the t-statistic in order to carry out this hypothesis test:

$$t_{n-2} = \frac{r-\rho}{\hat{\sigma}_r} = \frac{r}{\hat{\sigma}_r} = \frac{r}{\sqrt{\frac{(1-r^2)}{n-2}}} = r^* \frac{\sqrt{n-2}}{\sqrt{(1-r^2)}} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r}}$$

For example, say r=.25, N=100

$$t = \sqrt{98}\,\frac{.25}{\sqrt{1-.25^2}} = 9.899\,\frac{.25}{.986} = 2.56$$

Inference with Regression

hypothesis test for degree of association

For the **hypothesis test** that ρ is equal to a hypothesized population value:

$$H_0 : \rho = value$$
$$H_1 : \rho \neq value$$

We standardize our statistic using Fisher:

$$z = \frac{.5\log_e \frac{1+r}{1-r} - .5\log_e \frac{1+\rho}{1-\rho}}{1/\sqrt{N-3}}$$

For example, say N=200, r = .54, and ρ is hypothesized as ρ=0.3.

$$H_0 : \rho = 0.3$$
$$H_1 : \rho \neq 0.3$$

$$z = \frac{.5\log_e \frac{1+.54}{1-.54} - .5\log_e \frac{1+.30}{1-.30}}{1/\sqrt{200-3}}$$

$$z = \frac{.60-.31}{.07} = 4.13$$

Compare to unit normal, e.g., 4.13 > 1.96 so it is significant. Thus, our sample was likely not drawn from a population in which rho is .30.

Inference with Regression

hypothesis test for equality between two groups

When we are testing the equality of correlations from 2 independent samples, rejecting the null suggests that the two samples are not independently selected or are from different populations.

$$H_0 : \rho_1 = \rho_2$$
$$H_1 : \rho_1 \neq \rho_2$$

Again, using Fisher's transformation:

$$z = \frac{.5\log_e \frac{1+r_1}{1-r_1} - .5\log_e \frac{1+r_2}{1-r_2}}{\sqrt{1/(N_1-3) + 1/(N_2-3)}}$$

For example, say N_1=150, r_1=.63, N_2=175, r_2=70.

$$z = \frac{.5\log_e \frac{1+.63}{1-.63} - .5\log_e \frac{1+.70}{1-.70}}{\sqrt{1/(150-3) + 1/(175-3)}}$$

$$z = \frac{.74-.87}{.11}$$

$$= -1.18, \text{ n.s.}$$

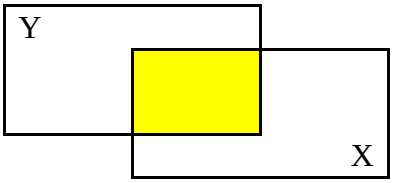
Evidence suggests that our samples were independently selected.

Partial Correlations
introduction

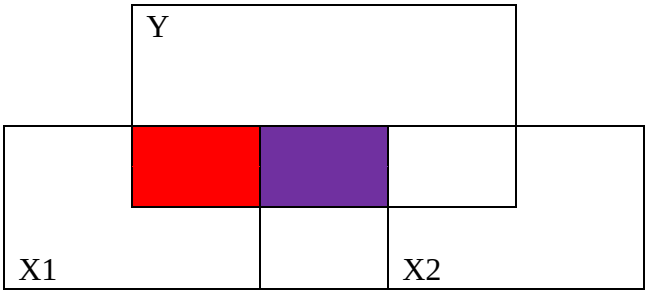
When handling multiple IV's, the degree of association between a given IV and the DV may depend on the other IV's.

This presents two problems: if we fail to include the IV's that are associated with other IV's, then our regression coefficient and our R^2 will be biased according to the zero conditional mean assumption.

We can think of the correlation between X and Y as an area where the variability in y and the variability in X overlap- X and Y vary together.



However, there are other variables associated with both x and y. This presents a particular problem when doing a multivariable analysis.



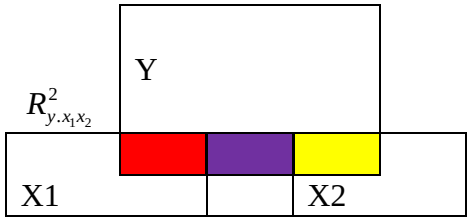
The purple area on the diagram above is double counted when only a single IV is included

The Venn diagram illustrates a key idea: when we hold a third variable constant, we are actually holding two relationships constant: the association between X2 and Y as well as the association between X2 and X1.

We'll be looking at how to handle this problem from the direction of looking at the regression coefficient and the correlation coefficient. For the correlation coefficient, we can calculate the partial correlation. For the regression coefficient, we can moderate our IV.

Partial Correlations
partial correlations

A **partial correlation** $R^2_{y.x_1x_2}$ is the proportion of the unexplained variation in y caused by x1 alone, i.e. the red area.



The correlation between X1 and X2 on Y…minus the correlation between X2 and Y… yields the red area- the isolated association of X1 on Y.

$$r^2_{yx_1.x_2} = \frac{R^2_{y.x_1x_2} - R^2_{y.x_2}}{1 - R^2_{y.x_2}}$$

However, this is measured as the sum of squares, so we need to standardize it by dividing it by the remaining unexplained variation in y.

A low r shows that there is little partial correlation between X1 and Y when X2 is held constant. Conversely, a high r shows that the two variables are correlated after holding X2 constant.

A **semi-partial correlation** is the correlation between a single variable (holding others constant) and a dependent variable, expressed as a relative percentage of the total variation in Y, rather than the remaining unexplained variation.

$$r^2_{y(x_1.x_2)} = R^2_{y.x_1x_2} - R^2_{y.x_2}$$

To measure the partial correlation between x_1 and y mathematically, first measure the association between the variable we wish to isolate and both the dependent variable and the second independent variable that we are holding constant.

$$\hat{y} = b_0 + b_1x_1 \quad \hat{e}_{\hat{y}_i} = (\hat{y}_i - y_i)$$
$$\hat{x}_2 = b_0 + b_1x_1 \quad \hat{e}_{\hat{x}_2} = (\hat{x}_2 - x_2)$$

Next, create two new variables that measure the difference between the predicted value and the actual value for both variables. Finally, measure the correlation between these residuals.

$$r_{(\hat{y}-y)(\hat{x}_2-x_2)} = r_{yx_1.x_2}$$

Partial Correlations
moderation

If $cor(y, x_1) > 0$ and $cor(x_1, x_2) > 0$, then the partial effect x_1 on *y* will vary depending on the value of its **correlate**, x_2. The relationship between x_1 and *y* varies depending on x_2.

This important effect is called a **moderation effect.** x_2 "moderates" the effect of x_1 on *y*. A moderator variable implies that the effect of x_1 on *y* is not consistent across the distribution of x_2.

Measuring moderation effects along with partial effects allows us to more accurately predict *y* for a given value of the correlate.

The main take-away from this section is that if we want an accurate estimate of the regression coefficient, then we need to include all the other variables that are also associated with the IV and DV.

The purple area below represents the area that is double counted when only a single IV is included, otherwise, both our regression coefficients and our correlation coefficient will be biased in the upward direction.

To measure an moderation effect, create an **moderation term**, $x_1 x_2$.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

With the sample regression model, we can take the derivative of the regression function to determine the partial effect and the moderation effects for x_1.

$$\partial x_1 / \partial y = \beta_1 + \beta_3 \Delta x_2$$

When $x_2 = 0$, β_1 is the partial effect of x_1. We can enter any value for x_2, such that the combined partial and moderation effects are $\beta_1 + \beta_3 \Delta x_2$.

If you suspect a moderation effect, it is best to first look at means across groups and correlations between variables.

We'll be looking at several scenarios with different moderation variables

1. A binary variable moderates two groups
2. A binary variable moderates multiple groups
3. A continuous variable moderates two or more groups

Partial Correlations
moderation- a binary variable moderates two groups

In an experimental context, we may want to compare an outcome for two groups, for example, male and female. We use **dummy coding** to code binary or categorical variables in a regression analysis.

In this example, gender is the moderates the effect of x_2 on *y*.

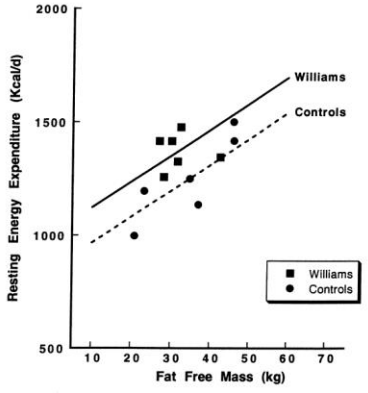
$$y = \beta_0 + \delta_0 x_1 + \beta_1 x_2 + u$$

x_1 is a binary moderating variable. One group is defined as the **base group** or **referent**, such that $x_1 = 0$

x_2 is a continuous variable

δ_0 is the partial effect of x_1.

We can imagine two regression lines for y on x_2. These lines have the same slope, β_1, but the two separate intercepts depending on the moderator, β_0 and $\beta_0 + \delta_0$.



Partial Correlations
moderation- a binary variable moderates multiple groups

We also use moderation when comparing multiple groups where each individual is in a single class in each group. For example, gender and hemisphere (North or South).

x_1 is a binary moderating variable.

x_2 is a binary moderating variable.

Note that both groups have base groups.

To model the partial effects, create a moderating variable for each combination of subgroups, save the base group:

$$y = \beta_0 + \beta_1(x_1 = 1)(x_2 = 1) + \beta_2(x_1 = 0)(x_2 = 1) + \beta_3(x_1 = 1)(x_2 = 0)$$

In this case, the base group is $(x_1 = 0)(x_2 = 0)$.

When you want to test whether a moderating variable is significant, make that given moderating variable the base group and compute the t-statistic using $SE(\hat{\beta}_0)$.

Partial Correlations
moderation- a continuous variable moderates two or more groups

x_1 is a categorical variable with three levels

x_2 is a quantitative variable whose effect we wish to measure

The effect of x_2 on y will be moderated by x_1. This allows for difference slopes across the regression of y on x_2. This is commonly called the **dose response curve**.

$$y = \beta_0 + \beta_1(x_1 = 1) + \beta_2(x_1 = 2) + \beta_3 x_2 + \beta_4(x_1 = 1)x_2 + \beta_5(x_1 = 2)x_2 + u$$

The regression coefficients on the moderation terms represent the change in the slope between groups, compared to the slope for X, which is the slope for the control condition.

We model this by dummy coding x_1 into two separate variables, x_1 and x_2. The quantitative variable is now x_3.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + u$$

To calculate the regression coefficient, take the derivative of the group in question.

$$\frac{\partial x_1}{\partial y} = \beta_1 x_1 + \beta_4 x_3$$

You don't have to calculate the regression coefficient at the mean of x_3. For constant c_3:

$$\frac{\partial x_1}{\partial y} = \beta_1 x_1 + \beta_4 (x_3 - c_3)$$

To calculate the predicted response, ensure x_1 is the referent and subtract a given constant from x_3 .

$$\hat{y} = \beta_0 + \beta_3 (x_3 - c_3)$$

A model comparison between restricted and unrestricted models can be done using ANOVA and F statistics.

This is effectively a test for a difference in slopes. This hypothesis puts no restrictions on the difference in intercepts.

Graphically, you can plot the reg. line for all three groups on the same scatterplot. Parallel lines imply no moderation effect (the slope is the same).

Partial Correlations

mediation- standard approach & path analysis

A **mediation analysis** is conducted to better understand the observed effect of X on Y.

A **mediator variable**, M, is introduced to test (in effect) the robustness of the correlation between x and y.

There are three regression models in a mediation analysis:

1. (y~X)- regression coefficient for X should be significant
2. (M~X)- regression coefficient for x should be significant
3. (Y~X + M)- regression coefficient for M should be significant.

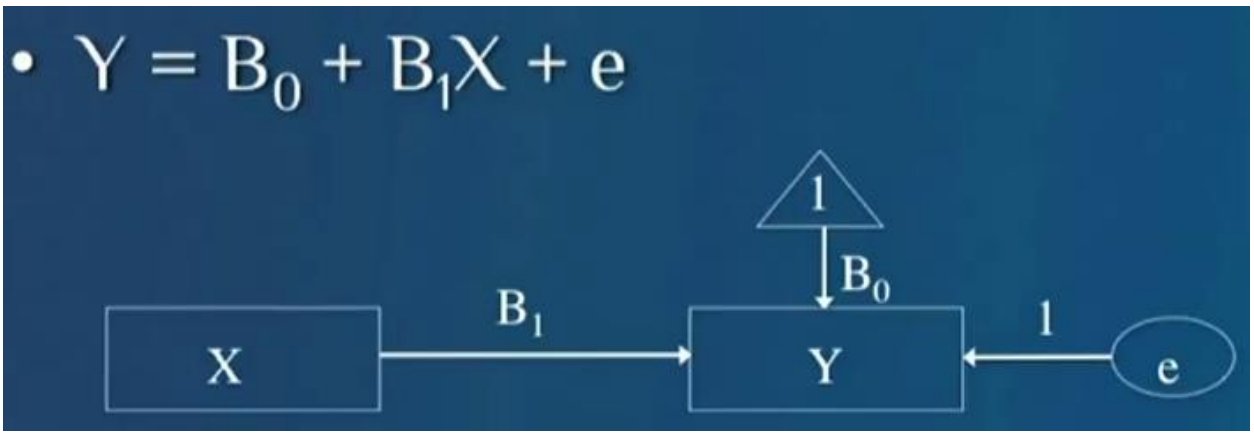
If the above cases are true, then we can ask the key question: is the regression coefficient for X still significant?

If yes, then there is **partial mediation**. If no, then **full mediation**.

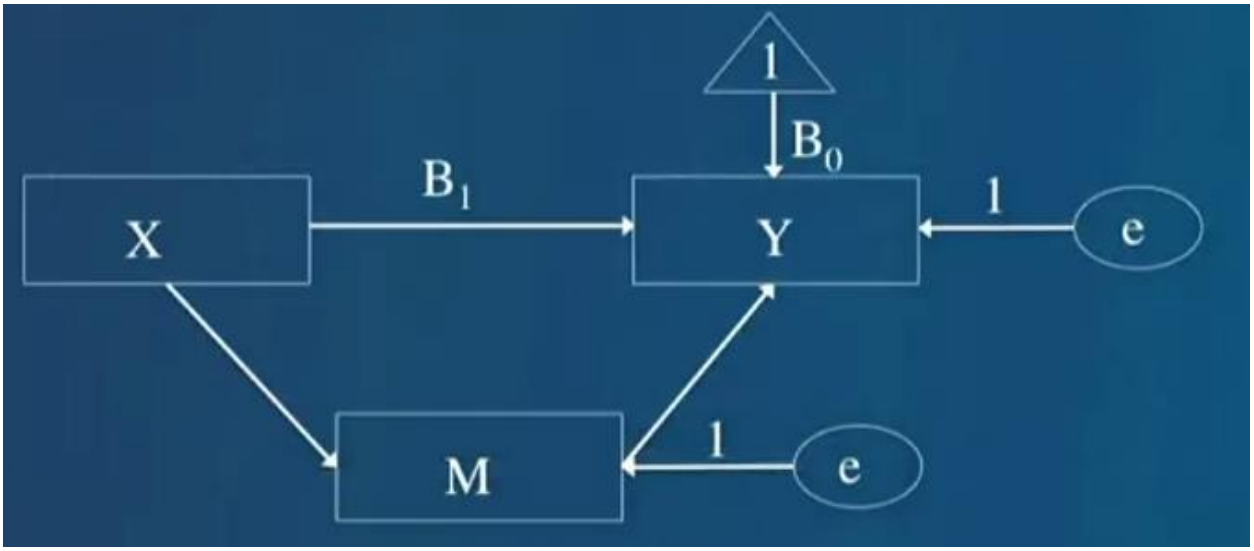
Partial Correlation

Mediation analyses are illustrated using path models.

Rectangles: observed variables (X,Y,M); circles (unobserved variables e), triangles for constants and arrows for associations (not causations). It can be presented graphically below.



Path model with a mediator



The sobel test tests the indirect path (from X to M to Y)

The test is whether the indirect effect is zero. There is a formula here, but it can be done in r.

Multiple Regression

introduction

Thus far, we've only focused on a single variable when approaching inference with regression.

Say we have two explanatory variables, x_1 and x_2, and a single dependent variable, y. We want to know the partial effect of x_1 on y. Now comes the kicker- we also suspect that x_1 and x_2 are correlated.

In order to estimate an accurate partial effect, we need to include all of the correlates of x_j. A low R^2 does not imply that a partial effect is inaccurate.

By including x_2 in our model, we "factor out" the correlation between the two explanatory variables. Adding x2 pulls x2 out of the error term and puts it explicitly in the equation, measuring the correlations between x2 and y as well as x2 and x1.

What is left is the correlation (or lack thereof) between x_1 and y. Knowing the correlation, we can measure the slope of the regression line, β_1, and we have estimated the partial effect of x_1 on y!

This section will explore statistical tests when there are **multiple restrictions**, that is, two or more regression coefficients are tested simultaneously.

When a group of independent variables are highly correlated, it may be useful to test for the significance of the variables as a group.

There are two models that this section covers. The first model is the difference of two statistics and the second is the F-test for multiple regression coefficients.

Multiple Regression

difference of two statistics

To measure and compare the partial effect of two alternative treatments on a dependent variable, we use the **difference of two statistics** test.

For example, in terms of financial outcomes, is a year of junior college less valuable than a year at a four-year university?

$$E(y \mid x_1, x_2, ..., x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + u$$

Where x_1 and x_2 are years of junior college and years at a four-year college.

The null hypothesis is that they have the same partial effect. The alternative hypothesis is that junior college has a smaller partial effect.

$$H_0 : \beta_1 = \beta_2$$
$$H_1 : \beta_1 < \beta_2$$

The trick here is that we're going to define a new random variable that is equal to the difference between the two variables. This allows us to treat two variables as one.

$$\theta = \beta_1 - \beta_2$$

By combining the two variables, inference is a cinch.

$$H_0 : \beta_1 - \beta_2 = 0; \qquad t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$
$$H_1 : \beta_1 - \beta_2 = 0$$

Since it's a pain to figure out the standard error for the difference of two parameters, we can use an alternate regression formula (derived below) where we can measure the significance of the difference directly from the computation.

$$\theta = \beta_1 - \beta_2 \rightarrow \beta_1 = \theta + \beta_2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$y = \beta_0 + (\theta_1 + \beta_2) x_1 + \beta_2 x_2 + u$$

$$y = \beta_0 + \theta_1 x_1 + \beta_2 x_1 + \beta_2 x_2 + u$$

$$y = \beta_0 + \theta_1 x_1 + \beta_2 (x_1 + x_2) + u$$

Whether x_1 is significant or not confirms the hypothesis.

Multiple Regression

joint significance

When there are multiple restrictions, we compare two regression models, the **restricted** and the **unrestricted** model.

This is called a joint hypothesis test.

Say we have a regression equation where we want to test whether parameters 3-5 are jointly significant.

$$H_0 : \beta_{3-5} = 0$$

If the null hypothesis is rejected, then the excluded variables are **jointly statistically significant**. If the null hypothesis is not rejected, then the excluded variables are **jointly insignificant**.

It is possible that some (or all) of the jointly insignificant variables are still significant in isolation.

The F-statistic is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated.

$$F = \frac{(SSR_r - SSR_{ur}) / q}{SSR_{ur} / (n - k - 1)}$$

Think of F as measuring the relative increase in SSR when moving from the unrestricted to the restricted model.

We can also use the correlation coefficient of the two sample regression functions to determine the F-statistic.

$$F = \frac{(R_{ur}^2 - R_r^2) / q}{(1 - R_{ur}^2) / n - k - 1}$$

where q is the number of exclusion restrictions.

In most regression packages, an F-statistic is automatically reported for the **overall significance of a regression**, which tests all explanatory variables.

Transformations

introduction

A **transformation** is the replacement of a variable by a monotonic function of that variable.

A **monotonic function** f(t) moves in one direction as its argument t increases. On the graph of a monotonic function, *x* doesn't double back: there are never two y-values for a single x-value.

We'll start with simpler transformations and make our way to more complicated ones.

A **linear transformation** is when we add, subtract, multiply and divide our data. It does not effect the value of our R^2.

A special kind of linear transformation is the **affine transformation,** which is performed on the statistics themselves. Unit conversions are affine transformations.

Say we want to measure the effect of cigarettes on birthweight and the units on our variable for birthweight is ounces, but we want to measure weight in lbs.

When we're changing the units of the dependent variable, we can divide the entire equation by the constant.

$$\hat{y}/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)income$$

Meanwhile, the units for cigarettes are the number of cigarettes smoked, but we want the number of packs.

When changing the units of the independent variable, we only change the independent variable and its corresponding partial effect.

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 * 20)\frac{cigs}{20} + (\hat{\beta}_2)income \rightarrow \hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 * 20)packs + (\hat{\beta}_2)income$$

Transforming statistics that are observations, such as the mean, will have the same effect on the statistic as on any given observation. For example, if you add 5 to all of the observations, the mean will increase by 5. If you multiply all the observations by 3, then the mean will triple.

Transforming statistics that are distances, such as the standard deviation, will have a different effect. Increasing the mean does nothing to the standard deviation. Multiplying the standard deviation by *x*, will stretch the distance between the observations by *x*.

For example, if you add 1 to x (or 1 to y, or 1 to both) the "cloud" of data shifts to the right by one unit, including the standard deviation and the averages.

Transformations

standardized coefficients

The partial effect is presented as an **unstandardized coefficient**. In order to compare "apples to apples" which variable has the largest partial effect, we need to compare the **standardized coefficient**.

While $\hat{\beta}_j$ may be statistically significant, it may not be practically significant. The magnitude of a regression coefficient may be too small to warrant including the variable in the model.

Starting from the sample regression function:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

$$y_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 (x_{i1} - \bar{x}_1) + \hat{\beta}_2 (x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k (x_{ik} - \bar{x}_k) + \hat{u}_i$$

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \hat{\beta}_0 + \frac{\hat{\sigma}_1}{\hat{\sigma}_y} \hat{\beta}_1 (\frac{x_{i1} - \bar{x}_1}{\hat{\sigma}_1}) + \frac{\hat{\sigma}_2}{\hat{\sigma}_y} \hat{\beta}_2 (\frac{x_{i2} - \bar{x}_2}{\hat{\sigma}_2}) + \ldots + \frac{\hat{\sigma}_k}{\hat{\sigma}_y} \hat{\beta}_k (\frac{x_{ik} - \bar{x}_k}{\hat{\sigma}_k}) + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

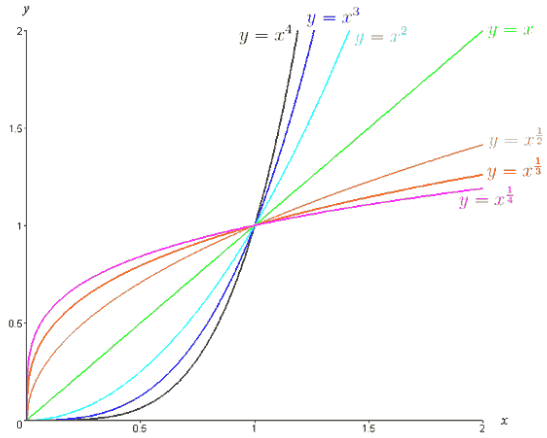
$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + error$$

Where $\hat{b}_j = \frac{\hat{\sigma}_j}{\hat{\sigma}_y} * \hat{\beta}_j$ for j=1,…,k. Note that the intercept drops out altogether.

Transformations
power functions

We can make a non-linear regression line with non-constant partial linear effects by applying a transformation to a variable.

A **power** or **exponential function follows the form C^x.** A power function takes off quickly.



The **square** or **quadratic transformation** follows x^2.

The **cube** or **cubic transformation** follows x^3 and can have two peaks.

We can start a regression model with a linear interpretation of a function, and then square it. If it accounts for additional variance, we can cube it, etc. With real data, it is rare to have a quadratic term add significant variance, and even more rare for a cubic term to add variance. In applied work, you will probably never need anything beyond the cubic.

The **reciprocal transformation**, x^{-1} , flips a power function across the x-axis. It cannot be applied to zero values. The reciprocal flips a ratio upside down. Mi/gallon becomes gallon/mi. This reverses the magnitudes of y. The largest y becomes the smallest and the smallest, largest.

The **negative reciprocal transformation,** $-x^{-1}$, is a monotonic decreasing function in that it reverses the order of the data. That is, if a>b, then f(a)<f(b). The negative reciprocal preserves order among values of the same sign after the reciprocal is made.

A **root transformation,** $x^{1/2}$ is applied to decreasing monotonic functions where the power is less than 1.

We model a **quadratic function** by including x_k as both a linear and a squared term (hence, quadratic).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2$$

The ceteris paribus effect of x_k is the derivative of the sample regression function. It is moderated by itself:

$$\Delta\hat{y} / \Delta x_1 \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x_1)$$

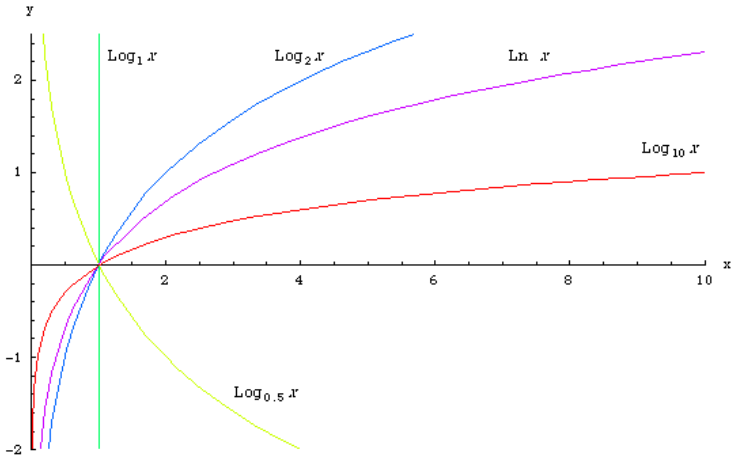
We can set $y = (\hat{\beta}_1 + 2\hat{\beta}_2 x_1) = 0$ in order to find the maximum/minimum of the function. If we do, we get:

$$x_1^* = | \hat{\beta}_1 / (2\hat{\beta}_2) |$$

Transformations

Log functions

Log functions of x are asymptotic to y. The **ladder of transformations:**



$$y = b_0 + b_1 \log(x)$$

A percentage increase in x leads to an increase in y units.

The **constant-elasticity model** or **log-log** borrows the standard $\%\Delta x / \%\Delta y$ interpretation from economics:

$$\log(y) = \beta_0 + \beta_1 \log(x)$$

The constant elasticity model follows a power distribution.

$$y = b_0 * x^{b_1} \rightarrow \log(y) = b_0 + b_1 \log(x)$$

For powers greater than 1, the curve is concave up, if less than 1, concave down. In other words, if beta-1 (the slope of the log-log regression) is greater than 1, we're increasing, otherwise, decreasing.

The **semi-elasticity model** or **log-level**:

$$\log(y) = \beta_0 + \beta_1 x_1 + u$$

Taking the log of y will capture a constant percentage partial effect. 100* β_1 is the percentage change in y when x_1 changes by one unit. We can say that, "the semi-elasticity of y with respect to x_1 is 100* β_1."

For the exact $\hat{\beta}_1$:

$$\hat{y} = 100 * e^{(\hat{\beta}_1 \Delta x_1)} - 1$$

Transformations
log rules

To calculate value of $\hat{y}$ from $\widehat{\log(y)}$, one can not simply exponentiate $\widehat{\log(y)}$.

$$\hat{y} = e^{\log(\hat{y})}$$

The method of choice is the unfortunately named **smearing estimate,** $\hat{\alpha}_0$.

1. From the regression of $\log(y)$ on the independent variables, obtain the fitted values, $\hat{y}_i$, and the residuals, $\hat{u}_i = \hat{y}_i - y$
2. Calculate $\hat{\alpha}_0 = \sum_{i=1}^{n} e^{\hat{u}_i} \Big/ n$
3. Plug in values for the independent variables to estimate the change in $\log(\hat{y})$
4. Obtain $\hat{y}$ using $\hat{y} = \hat{\alpha}_0 * e^{\log(\hat{y})}$

There are some standard rules of thumb for taking logs. None are written in stone.

- If the variable is measured in dollars (wages, income, etc.), we take the log. If the variable is a count of individuals, we take the log. If the variable is time (years, experience, age, tenure), we do not take the log.
- Be careful if presenting percentages or proportions as logs- it is a percentage change in percentage points.
- It is not statistically valid to compare the coefficient of determination for log(y) vs. y.
- $\log(y)$ also mitigates heteroskedasticity and outliers.

Also, changing the unit of measurement for the log variables won't affect any of the slope estimates.

$$\log(c, y_i) = \log(c_1) + \log(y_i), c_1 > 0$$

If the dependent variable frequently takes a value of 0, then try using $\log(1+y)$. $\log(1+y)$ cannot be normally distributed.

Generalized Linear Model

introduction

An extension of the general linear model (which is what we've been looking at so far), the **generalized linear model** allows for non-normal distributions in the outcome variable and therefore also allows testing of non-linear relationships between a set of predictors and the outcome variable.

GLM is the mathematical framework used in many common statistical analyses, including multiple regression and ANOVA.

Characteristics of GLM:

- Linear- relationships between X and Y are linear
- Additive- effects (i.e. beta-coefficients) are additive

When assumptions failed, we have some "quick fixes"- transformations or product terms.

If these quick fixes are not enough, go GLM.

GLM* is an extension of GLM

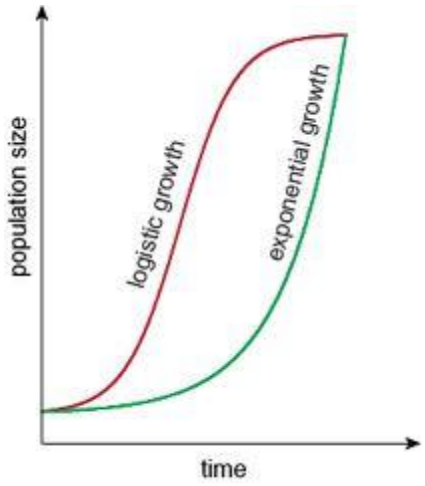
The key to the generalized linear model is that it is allowed to generalize to other forms by adding a **link functions**. A link function is applied to the outcome variable.

We'll look at two different examples of GLM functions: the logit and the poisson.

Generalized Linear Model

Binary logistic regression

A good first example of the GLM* is the **binary logistic regression**. The link function is the log function of the odds. Why the odds? Because we want the outcome to be between zero and one.



This is the **logistic curve**. The shape of the graph would be an **s-shaped** or **sigmoid function**, where the response variable increases slowly, then exponentially, then increases slowly again, eventually settling at 100%.

An example- given an individual's height, we want to know the probability that the individual is male. To the graph above, if we plotted P(X) against height, P(X) should be near 0 at lower heights and approach 100% at taller heights.

Recall the **odds**, *p:q*. An event with a 2/3 probability of success has an odds of 2/3:1/3 or 2:1.

With respect to the binary logistic regression, the dependent variable is the odds, $y = p / q$.

We convert the odds into a probability by taking the **logit**- the natural log of the odds.

$$\log it(y) = \ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = \beta_0 + \sum(\beta_k x_k)$$

Generalized Linear Model

Logit regression analysis

To evaluate predictive power of the overall model, compare the chi-square for the model to the chi-square of a model with no predictors (**the null model**).

To test each regression coefficient, we use the **Wald Test**, which tests how much the chi-square changes with the model vs. with the model without the predictor.

The regression coefficient on a logit regression is more easily interpreted if it is translated from a logit into an odds ratio. This is done with exponentiation:

$$\log it(y) = \ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = \beta_0 + \sum(\beta_k x_k) \rightarrow \beta_x = \frac{\hat{y}}{1-\hat{y}} = e^{\beta \Delta X}$$

The regression coefficients can also be presented in terms of probability, which makes the most sense.

$$P(y=1) = \frac{e^{a+bX}}{1+e^{e+bX}}$$

It also helps to view a **classification table**, a two-way table which tells us how well the model predicts the actual outcomes, as a frequency, percentage and total percentage.

You can graph a predictor against the logit for a straight-line scatterplot.

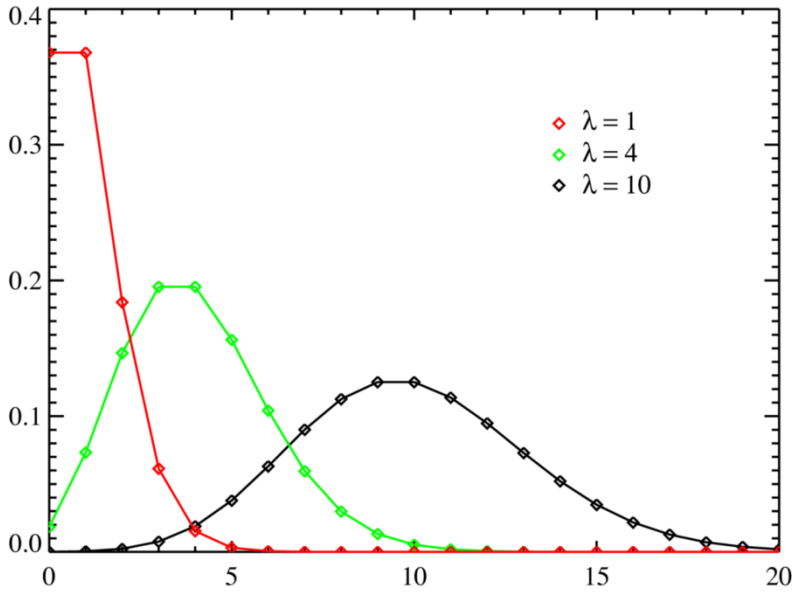
Adding more than 2 categories on the outcome (no longer a binary logistic regression) is fine- do a **multinomial logistic regression**, which still uses the logit along with dummy codes.

Poisson
Introduction

The **Poisson** statistic, λ ("lamda") is a random variable that tells you how many times you expect to see streaks arise from a random process.

It assumes that events are rare, have a fixed average rate, and are independent.

Using the **Poisson distribution**, you can then compare the number of predicted streaks to the real number of streaks in your data, and mathematically test whether a set of events is random or not.



The Poisson distribution is unique in that it has a single variable that defines the distribution, λ. λ represents the average value within a given number of trials or time periods.

The **link function is the log function.**

A Poisson distribution has a cumulative dependent variable. We're working with count data over time. For example, the number of traffic accidents as a function of weather conditions. (clear weather, rain, snow…)