UDACITY
STATISTICS

Lesson 1: Descriptive Statistics

Data Types

CATEGORICAL ORDINAL VS. CATEGORICAL NOMINAL

We can divide categorical data further into two types: Ordinal and Nominal.

Categorical Ordinal data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs).

Categorical Nominal data do not have an order or ranking (like the breeds of the dog).
Have questions? Head to the forums for discussion with the Udacity Community.

CONTINUOUS VS. DISCRETE

We can think of quantitative data as being either continuous or discrete.

Continuous data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

Discrete data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

Summary Statistics

ANALYZING QUANTITATIVE DATA

Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

1. Measures of Center
2. Measures of Spread
3. The Shape of the data.
4. Outliers

Analyzing Categorical Data

Though not discussed in the video, analyzing categorical data has fewer parts to consider. **Categorical** data is analyzed usually be looking at the counts or proportion of individuals that fall into each group. For example if we were looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportion of dogs are of each breed type.

Measures of Center

1. Mean
2. Median
3. Mode

THE MEAN

In this video, we focused on the calculation of the mean. The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

THE MEDIAN

The **median** splits our data so that 50% of our values are lower and 50% are higher. We found in this video that how we calculate the median depends on if we have an even number of observations or an odd number of observations.

Median for Odd Values
If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**. For example, if we have 7 observations, the median is the fourth value when our numbers are ordered from smallest to largest. If we have 9 observations, the median is the fifth value.

Median for Even Values
If we have an **even** number of observations, the **median** is the **average of the two values in the middle**. For example, if we have 8 observations, we average the fourth and fifth values together when our numbers are ordered from smallest to largest.

In order to compute the median we MUST sort our values first.

Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**.

THE MODE
The **mode** is the most frequently observed value in our dataset.
There might be multiple modes for a particular dataset, or no mode at all.

No Mode
If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:
1, 1, 2, 2, 3, 3, 4, 4
There is no mode, because all observations occur the same number of times.

Many Modes
If two (or more) numbers share the maximum value, then there is more than one mode. If we have the dataset:
1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9
There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

NOTATION: AGGREGATIONS
Σ = summation
Π = product
∫ = for continuous values

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

Lesson 2: Descriptive Statistics - Part II

Measures of Spread

Measures of Spread are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

1. Range
2. Interquartile Range (IQR)
3. Standard Deviation
4. Variance

Histograms help you understand the four aspects we outlined earlier regarding a quantitative variable:

- center
- spread
- shape
- outliers

CALCULATING THE 5 NUMBER SUMMARY

The five number summary consist of 5 values:

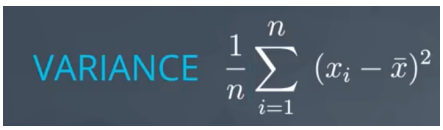
1. Minimum: The smallest number in the dataset.
2. Q1: The value such that 25% of the data fall below.
3. Q2: The value such that 50% of the data fall below.
4. Q3: The value such that 75% of the data fall below.
5. Maximum**:** The largest value in the dataset.

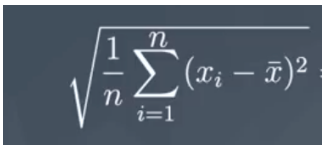
In the above video we saw that calculating each of these values was essentially just finding the median of a bunch of different dataset. Because we are essentially calculating a bunch of medians, the calculation depends on whether we have an odd or even number of values.

Range is the difference between the maximum and the minimum.
IQR the interquartile range is calculated as the difference between

STANDARD DEVIATION AND VARIANCE

The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.





SHAPE

Shape	Mean vs. Median	Real World Applications
Symmetric (Normal)	Mean equals Median	Height, Weight, Errors, Precipitation
Right-skewed	Mean greater than Median	Amount of drug remaining in a blood stream, Time between phone calls at a call center, Time until light bulb dies
Left-skewed	Mean less than Median	Grades as a percentage in many universities, Age of death, Asset price changes

The mode of a distribution is essentially the tallest bar in a histogram. There may be multiple modes depending on the number of peaks in our histogram.

Bell-shaped (symmetric) examples:
Heights, precipitation, weights, standardized test scores, errors in manufacturing processes

Left-skewed examples:
GPA scores, age of death, asset price changes

Right-skewed examples:
Amount of blood in stream over time, distribution of wealth, athletic abilities

Outliers

Outliers Advice
Below are guidelines for working with any column (random variable) in your dataset.

1. Plot your data to identify if you have outliers.
2. Handle outliers accordingly via the methods above.
    At least note they exist and the impact on summary statistics.
    If typo - remove or fix
    Understand why they exist, and the impact on questions we are trying to answer about our data.
3. If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.
    Side note: If you aren't sure if your data are normally distributed, there are plots called normal quantile plots and statistical methods like the Kolmogorov-Smirnov test that are aimed to help you understand whether or not your data are normally distributed. Implementing this test is beyond the scope of this class, but can be used as a fun fact.
4. If you have skewed data or outliers, use the five number summary to summarize your data and report the outliers.

Inferential Statistics

Inferential Statistics is about using our collected data to draw conclusions to a larger population. We looked at specific examples that allowed us to identify the

1. **Population** - our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample** - subset of the population
4. **Statistic** numeric summary about a sample

Simpson's Paradox

In this example lesson, you learned about **Simpson's Paradox**, and you had the opportunity to apply it to a small example with Sebastian, as well as work through similar example in Python. In the lessons ahead, you will be learning a lot by following along with Sebastian, but it is really important to put these ideas to practice using data and computing, because that is how you will apply these skills in a day to day environment as a Data Analyst or Data Scientist.

It is so easy to get caught up in looking at full aggregates of your data. Hopefully, the examples here serve as a reminder to look at your data multiple ways.

Covariance

Just had to add this youtube link by Luis Serrano on PCA in which he first covers the very basics with crystal clear, non-math explanations:

https://www.youtube.com/watch?v=g-Hb26agBFg&t=364s

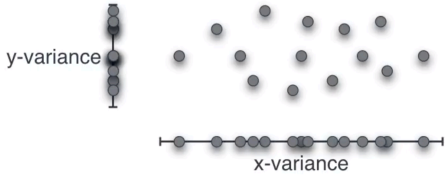
The gist is directly related to Simpson's paradox: sometimes we can distinguish between data by their aggregates (eg: their means are different), but sometimes the means can be the same even though the data are different:

Both these data sets have the same mean

So clearly a measure of spread such as variance shows the difference in these data sets. But what about these cases?

Both these data sets have the same mean and variance

Recall: intuition for variance calculation on 2 dimensions



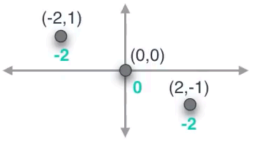
So those two sets above have the same x and y variance:

$$\text{x-variance} = \frac{2^2+0^2+2^2}{3} = 8/3$$

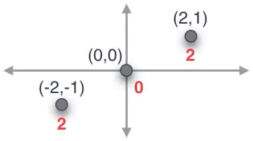
$$\text{y-variance} = \frac{1^2+0^2+1^2}{3} = 2/3$$

So what other aggregate measure can we use to distinguish between these two data sets? Enter covariance. By taking the sum of product of each point's coordinates, we get different values for the two examples:

The covariance values also closely relate to correlation, as the left points are negatively correlated to each other (the greater the x, the smaller the y), and the right points are positively correlated to each other.



$$\text{covariance} = \frac{(-2) + 0 + (-2)}{3} = -4/3$$



$$\text{covariance} = \frac{2 + 0 + 2}{3} = 4/3$$

Lesson 4: Probability

COIN TOSS
H = heads

Given the probability P(H) = 0.5:

What is the probability P(H, H) (ie: 2 heads in a row?

Truth Table: map out all possible scenarios…

FLIP 1	FLIP 2
H	H
H	T
T	H
T	T

So (H, H) only occurs in 1 of the 4 possible scenarios, therefore P(H, H) = 0.25

Which is the same as P(H) * P(H) = 0.5 * 0.5 = 0.25

What if the coin is loaded, so P(H) = 0.6?

FLIP 1	FLIP 2	Probability .
H	H	0.6*0.6 = 0.36
H	T	0.6*0.4 = 0.24
T	H	0.4*0.6 = 0.24
T	T	0.4*0.4 = 0.16

What if for 2 coin flips, the probability of only one heads? (back to fair coin examples)
= 0.5 which is the sum of the probabilities in the truth table with one head one tail

Same but with 3 flips? (to make sure you have all possible cases, test that probability column sums to 1)

FLIP 1	FLIP 2	FLIP 3	Probability .
H	H	H	0.5^3 = 0.125
H	H	T	0.5^3 = 0.125
H	T	H	0.5^3 = 0.125
H	T	T	0.5^3 = 0.125
T	H	H	0.5^3 = 0.125
T	T	**H**	0.5^3 = 0.125
T	**H**	T	0.5^3 = 0.125
T	T	T	0.5^3 = 0.125

— so only 3 occurrences in truth table, so 0.375.
— can do the same with a loaded coin, just adjust the probabilities in the truth table

Can use truth table to calculate probability of rolling even number of a six-sided fair die
— what would be the prob roll die twice, P(double)? ie: same number twice?

PROBABILITY RULES FOR INDEPENDENT EVENTS

1. **P**(**H**)=0.5
2. 1 – P(H) = P(not H) = 0.5 where 'not H' is the event of anything other than heads. Since, there are only two possible outcomes, we have that P(not H) = P(T) = 0.5. In later concepts, you will see this with the following notation: $\neg H$
3. Across multiple coin flips, we have the probability of seeing **n** heads as $P(H)^n$. This is because these events are independent.

We can get two generic rules from this:

1. The probability of any event must be between 0 and 1, inclusive.
2. The probability of the compliment event is 1 minus the probability of an event. That is the probability of all other possible events is 1 minus the probability an event itself. Therefore, the sum of all possible events is equal to 1.
3. If our events are independent, then the probability of the string of possible events is the product of those events. That is the probability of one event **AND** the next **AND** the next event, is the product of those events.

Lesson 5: Binomial Distribution

The **Binomial Distribution** helps us determine the probability of a string of independent 'coin flip like events'.

The probability mass function associated with the binomial distribution is of the following form:

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

where **n** is the number of events, **x** is the number of "successes", and **p** is the probability of "success".

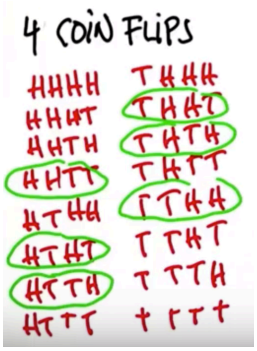
We can now use this distribution to determine the probability of things like:

- The probability of 3 heads occurring in 10 flips.
- The probability of observing 8 or more heads occurring in 10 flips.
- The probability of not observing any heads in 20 flips.

In practice, you will commonly be working with data, which might follow a binomial distribution. So it is less important to calculate these probabilities (though this can be useful in some cases), and it is more important that you understand what the Binomial Distribution is used for, as it shows up in a lot of modeling techniques in machine learning, and it can sneak up in our datasets with tracking any outcome with two possible events. A common place for a Binomial distribution is in logistic regression.

Outcomes

EXAMPLE / WORKINGS:



1. 4 coin flips, how many outcomes could have equal heads and tails?

   = 6 out of 16 distinct outcomes

** trick question: same question for 5 coin flips? = zero (odd number of flips can never have equal heads and tails!)

2. 5 coin flips, how many outcomes could have exactly 1 heads (4 tails)?

   eg: **H** T T T T

   = 5; 1st, 2nd, 3rd 4th OR 5th flips

3. 5 coin flips, how many outcomes could have exactly 2 heads (3 tails)?

   eg: **H1** T **H2** T T

** You have to be careful not to double count because the above outcome of H1 then H2 would be the same if the positions were swapped.

H1 there are 5 ways to place it
H2 there are 4 remaining ways to place it

= 5 * 4 = 20

but since the placements of H1 and H2 could be swapped, need to divide by 2
so the answer = 5 * 4 / 2 = 10

4. 5 coin flips, how many outcomes could have exactly 3 heads (2 tails)?

eg: **H1** T **H2** **H3** T

_____ _____ _____ _____ _____

Well, 3 heads is the same as saying 2 tails, which is the same answer as #3 previously.

But working through the calculation….

H1 there are 5 ways to place it
H2 there are 4 remaining ways to place it
H3 there are 3 remaining ways to place it
— then we have to factor double counting for H2 and triple counting for H3

So the answer is: 5 *4 * 3 / 3 * 2 * 1

The denominator notation simplifies to 3 factorial (3!)
- he expressed is as dividing by the possible ways to arrange 3 coins, which is 6:

eg: 1: a b c 4: a c b
2: b a c 5: c a b
3: b c a 6: c b a

In the numerator, the '5' denotes the number of coin flips. The 3 numbers in the numerator denote the specified number of outcomes we are looking for. It is a truncated factorial, which can also be abbreviated as shown next.

5. 10 coin flips, how many outcomes could have exactly 4 heads (6 tails)?

$$\frac{10 * 9 * 8 * 7}{4 * 3 * 2 * 1} = 5040 / 24 = 210$$

The numerator is abbreviated as **10! / 6!** because it cancels out:

$$\frac{10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1}{6 * 5 * 4 * 3 * 2 * 1} = 10 * 9 * 8 * 7$$

Generalizing, where:

n = number of coin flips
k = number of outcomes

<u>The formula becomes</u>:

$$\frac{n! / (n - k)!}{k!}$$ which is better written as: $$\frac{\mathbf{n!}}{\mathbf{k! (n - k)!}}$$

6. Using the formula and <u>elimination</u> (lh formula), calculate 125 coin flips and 3 heads?

Numerator: 125! / 122! = 125*124*123 = 1,906,500
Denominator: 3! = 6
Answer: 317,750

Probabilities

Apply the previous now to probabilities.

EXAMPLE / WORKINGS:

1. For 5 coins flips, what is the probability P(#heads = 1)?

   5! / 1! (5-1)! = 120 / 24 = 5 possible ways
   2^5 = 32 possible outcomes (size of the truth table)

   So: 5 / 32 = 0.15625

2. For 5 coins flips, what is the probability P(#heads = 3)?

   5! / 3! (5-3)! = 120 / 12 = 10 possible ways
   2^5 = 32 possible outcomes (size of the truth table)

   So: 5 / 16 = 0.3125

The prior 2 examples have been with a fair coin where P(heads) = 0.5. The next example uses a loaded coin where P(heads) = 0.8.

3. For 3 coins flips, what is the probability P(#heads = 1), where P(heads) = 0.8?

   Need to think about the truth table for a loaded coin:
   — 3 flips is 2^3 = 8 possible outcomes (rows) in the truth table

FLIP 1	FLIP 2	FLIP3	Probability
H	H	H	$0.8*0.8*0.8 = 0.8^3 = 0.512$
H	T	H	$0.8*0.2*0.8 = 0.8^2*0.2 = 0.128$
H	H	T	$0.8*0.8*0.2 = 0.8^2*0.2 = 0.128$
H	T	T	$0.8*0.2*0.2 = 0.2^2*0.8 =$ **0.032**
T	H	H	$0.2*0.8*0.8 = 0.8^2*0.2 = 0.128$
T	T	**H**	$0.2*0.2*0.8 = 0.2^2*0.8 =$ **0.032**
T	**H**	T	$0.2*0.8*0.2 = 0.2^2*0.8 =$ **0.032**
T	T	T	$0.2*0.2*0.2 = 0.2^3 = 0.008$
			1

From the truth table, the P(#heads = 1) = 0.032*3 = 0.096

** a faster way is you only need to draw the truth table for the outcomes you are interested in

FLIP 1	FLIP 2	FLIP3	Probability
H	T	T	$0.8*0.2*0.2 = 0.2^2*0.8 =$ **0.032**
T	T	**H**	$0.2*0.2*0.8 = 0.2^2*0.8 =$ **0.032**
T	**H**	T	$0.2*0.8*0.2 = 0.2^2*0.8 =$ **0.032**
			0.096

4. For 5 coins flips, what is the probability P(#heads = 4), where P(heads) = 0.8?

1st attempt:
Break it down: what number of outcomes are there for P(#heads = 4)? Its the same as asking P(#tails = 1), which is 5 possible outcomes.

Instead of taking P(tails)=0.2 and raising to 5, use P(heads)=0.8, so 0.8^5 = 0.32768

***** not quite right…***

Answer
5 possible outcomes is correct, and formal calc is:
5! / 1!(5-1)! = 5

Don't simply take 0.8^5 because we are actually calculating #heads=4 (and therefore #tails=1), so need to calc 0.8^4 * 0.2^1

Then multiply that by 5, so answer is 0.4096
— and actually, 0.8^4 is equal to 0.4096
** so really I was just off in my exponent by 1 (I calculated P(#heads=5)), but because the coin is loaded, that doesn't always hold… **

5. For 5 coins flips, what is the probability P(#heads = 3), where P(heads) = 0.8?

possible outcomes = 5! / 3!(5-3)! = 10
10 * 0.8^3 * 0.2^2 = 0.2048

6. For 12 coins flips, what is the probability P(#heads = 9), where P(heads) = 0.8?

possible outcomes = 12! / 9! / (12-9)! = 12*11*10 / 3! = 220
220 * 0.8^9 * 0.2^3 = 0.2362

Generalized then, the formula is: (where x is k from the video notation)

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Lesson 6: Conditional Probability

Often events are not independent like with coin flips and dice rolling. Instead, the outcome of one event depends on an earlier event.

For example, the probability of obtaining a positive test result is dependent on whether or not you have a particular condition. If you have a condition, it is more likely that a test result is positive. We can formulate conditional probabilities for any two events in the following way:

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

In this case, we could have this as:

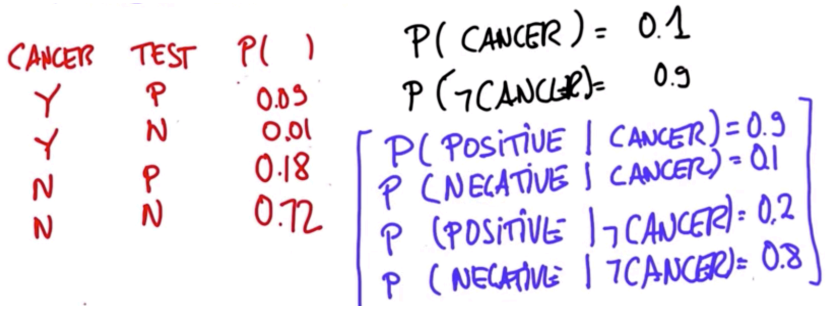
$$P(positive|disease) = \frac{P(\text{positive} \cap \text{disease})}{P(disease)}$$

where | represents "given" and ⋂ represents “and".

Medical example

What are the conditional probabilities of having[not having] cancer if a test is postive[negative]? Well, the probability is normally expressed backwards and we need to do some calcs to reverse it into the question we actually want to answer (the vids don’t explain this well, best to refer to Coursera stuff).

The slide assumes we know the probability of cancer[not cancer], and with that figures out the probabilities that the test results are accurate or not:



The purple section shows the probabilities of a test being positive if the patient has cancer, negative if not (ie: the true postives[negatives]) and the probabilities of the test being positive when the patient actually does not have cancer, and negative when the patient actually has cancer (ie: the false postives[negatives]).

We are asked to calculate the red section: the probability the patient has cancer (‘Y’) given the test is positive etc..

1. P (Y | +ve) = P(cancer) * P(+ve | cancer) = 0.1 * 0.9 = 0.09
2. P (Y | -ve) = P(cancer) * P(-ve | cancer) = 0.1 * 0.1 = 0.01
3. P (N | +ve) = P(not cancer) * P(+ve | not cancer) = 0.9 * 0.2 = 0.18
4. P (N | -ve) = P(not cancer) * P(-ve | not cancer) = 0.9 * 0.8 = 0.72

1

1. Says 9% probability patient has cancer if test is positive
2. Says 1% probability patient has cancer if test is negative
3. Says 18% probability patient DOES NOT have cancer if test is positive
4. Says 72% probability patient DOES NOT have cancer if test is negative

Total Probability

What is the probability we get a positive test result?

- just look at the probabilities calculated where the test results are positive and sum them, so #1 and #3 = 0.27
    - Which is in fact the weighted sum of the positive test probabilities given cancer or not (in purple) and the probabilities of cancer itself

ie: P(+ve) = P(+ve | cancer) * P(cancer) + P(+ve | not cancer) * P(not cancer)

2 COINS EXAMPLE:
Two coins in a bag, coin one is fair, coin 2 is loaded where P_2(heads) = 0.9.

1. What is the probability of picking a coin out of the bag and flipping heads?

P(heads) = P(heads | coin 1) * P(coin 1) + P(heads | coin 2) * P(coin 2)
= 0.5 * 0.5 + 0.9 * 0.5
= 0.7

2. Same as above, but flip the coin twice (you are picking only one coin from the bag, then flipping that one coin twice and observing heads then tails). What is the probability of getting heads then tails

P(heads then tails) = P(coin 1) * P(heads | coin 1) * P(tails | coin 1)
+ P(coin 2) * P(heads | coin 2) * P(tails | coin 2)
= 0.5*0.5*0.5 + 0.5*0.9*0.1
= 0.17

Two coins in a bag, coin 1 is loaded where P_1(heads) = 1, coin 2 is loaded where P_2(heads) = 0.6.

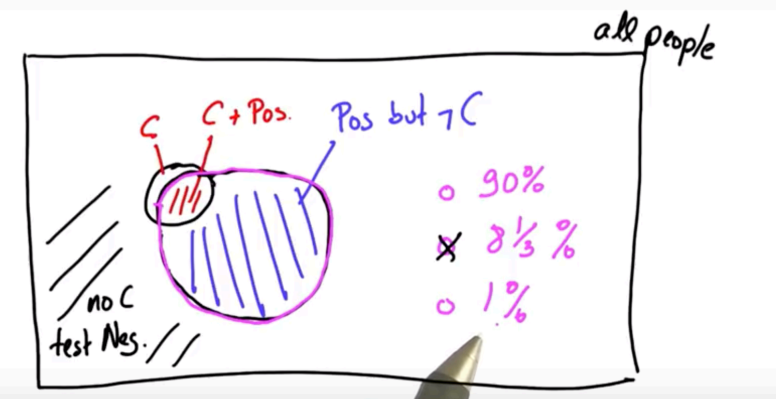
1. What is the probability of picking a coin out of the bag and flipping tails twice?

P(tails then tails) = P(coin 1) * P(tails | coin 1) * P(tails | coin 1)
+ P(coin 2) * P(tails | coin 2) * P(tails | coin 2)
= 0.5*0*0 + 0.5*0.4*0.4
= 0.08

Lesson 7: Bayes Rule

CANCER TEST
1% of the population has cancer. Given that there is a 90% chance that you will test positive if you have cancer (known as 'sensitivity') and that there is a 90% chance you will test negative if you don't have cancer (known as 'specitivity'), what is the probability that you have cancer if you test positive?

To answer, find the ratio of true positive to total positive:



Prior, Posterior and Normalizing

Prior = probability before test
Then add some evidence
Posterior = probability after taking into account new evidence

As in the previous lesson, the probability of cancer is given, this time as 1%. This is the prior.

Prior:	P(cancer)	= 0.01 therefore: P(not c) = 0.99
	P(+ve test | cancer)	= 0.9 therefore: P(-ve test | cancer) = 0.1
	P(-ve test | not c)	= 0.9 therefore: P(+ve test | not c) = 0.1

Conditionals: (joint prob) 'and'	P(cancer, +ve test)	= P(cancer) * P(+ve test, cancer) = prior * sensitivity = 0.01 * 0.9 **= 0.009**
	P(not c, +ve test)	= P(not c) * P(+ve test, not c) = prior * sensitivity = 0.99 * 0.1 **= 0.099**

But the joint posterior probabilities as they stand above do not sum to 1, so they need to be normalized by scaling up each probability so they sum to 1 while maintaining their ratio to the **total probability** of a positive test:

total probability
P(+ve tests) = P(cancer, +ve test) + P(not c, +ve test)

= 0.009 + 0.099
= 0.108

** this *total probability* sum is the normalizer in the denominator

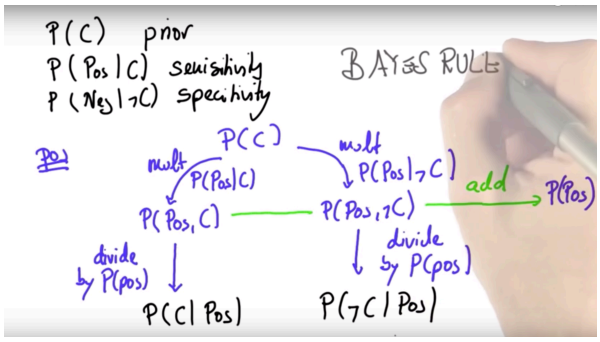
Posterior: P(cancer | +ve tests) = P(cancer, +ve test) / P(+ve tests)
= 0.009 / 0.108
= 0.0833

P(not c | +ve tests) = P(not c, +ve test) / P(+ve tests)
= 0.099 / 0.108
= 0.9167

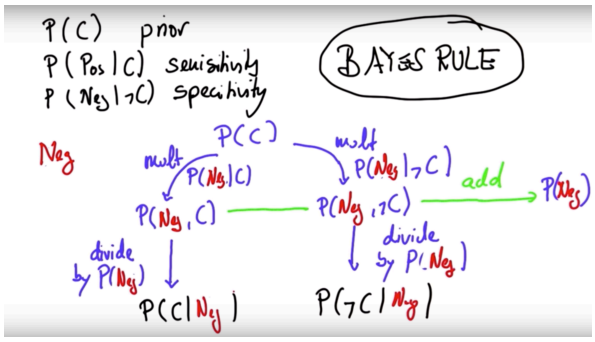
** now the probabilities sum to 1

Bayes Rule Diagram

Diagram for prev +ve example:



and as -ve example:



So, this should be stated as:

P(Pos, C) = P(Pos | C) P(C)
P(Neg, C) = P(Neg | C) P(C)
P(Pos, ~C) = P(Pos | ~C) P(~C)
P(Neg, ~C) = P(Neg | ~C) P(~C)

NEGATIVE TEST EXAMPLE:

Prior: P(cancer) = 0.01 therefore: P(not c) = 0.99
P(+ve test | cancer) = 0.9 therefore: P(-ve test | cancer) = 0.1
P(-ve test | not c) = 0.9 therefore: P(+ve test | not c) = 0.1

Conditionals: (joint prob) 'and'
P(cancer, **-ve test**) = P(cancer) * P(-ve test | cancer)
= prior * sensitivity
= 0.01 * 0.1
= 0.001

P(not c, **-ve test**) = P(not c) * P(-ve test | not c)
= prior * sensitivity
= 0.99 * 0.9
= 0.891

Now normalize by the total probability of a negative test:
P(-ve tests) = P(cancer, -ve test) + P(not c, -ve test)
= 0.001 + 0.891

= 0.892

Posterior: P(cancer | -ve tests) = P(cancer, -ve test) / P(-ve tests)
= 0.001 / 0.892
= 0.001122

P(not c | -ve tests) = P(not c, -ve test) / P(-ve tests)
= 0.891 / 0.892
= 0.9988

So this is Bayes rule in action. Prior to any further information (the tests), we could only say that we might be like the population and have a 1% chance of having cancer. After being tested (and accounting for test accuracy), our posterior belief after receiving a negative test result is that we only have a 0.01% chance of having cancer.

ROBOT EXAMPLE:
Let's start with what we know:

Prior Probabilities
The robot is perfectly ignorant about where it is, so prior probabilities are as follows:

P(at red) = 0.5
P(at green) = 0.5

Conditional Probabilities
The robot's sensors are not perfect. Just because the robot *sees* red does **not** mean the robot is *at* red:

P(see red | at red) = 0.8
P(see green | at green) = 0.8

Posterior Probabilities
From these prior and posterior probabilities we are asked to calculate the following posterior probabilities after the robot sees red:

1. P(at red | see red)
2. P(at green | see green)

and as a reminder, Bayes' rule looks like this: $$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

or, if we want to use our "versions" of A and B (for posterior #1).. $$P(\text{at red}|\text{see red}) = \frac{P(\text{see red}|\text{at red}) \cdot P(\text{at red})}{P(\text{see red})}$$

- We know the numerator from our priors: 0.8 * 0.5

But we still have one unknown! What was the probability that we would see red? The answer is 0.5 and there are two ways I can convince myself of that. The first is intuitive and the second is mathematical.

Why is P(see red) 0.5? ie: the denominator P(B)
Argument 1: Intuitive
Of course it's 0.5! What else could it be? The robot had a 50% belief that it was in red and a 50% belief that it was in green. Sure, its sensors are unreliable but that unreliability is symmetric and **not** biased towards mistakenly seeing either color.

So whatever the probability of seeing red is, that will also be the probability of seeing green. Since these two colors are the only possible colors the probability MUST be 50% for each!

Argument 2: Mathematical (Law of Total Probability)
There are exactly two situations where the robot would see red:

1. When the robot is in a red square and its sensors work correctly — **TRUE POS**
2. When the robot is in a green square and its sensors make a mistake — **FALSE POS**

Just add up these two probabilities to get the total probability of seeing red:

$$P(\text{see red}) = P(\text{at red}) \cdot P(\text{see red}|\text{at red}) + P(\text{at green}) \cdot P(\text{see red}|\text{at green})$$

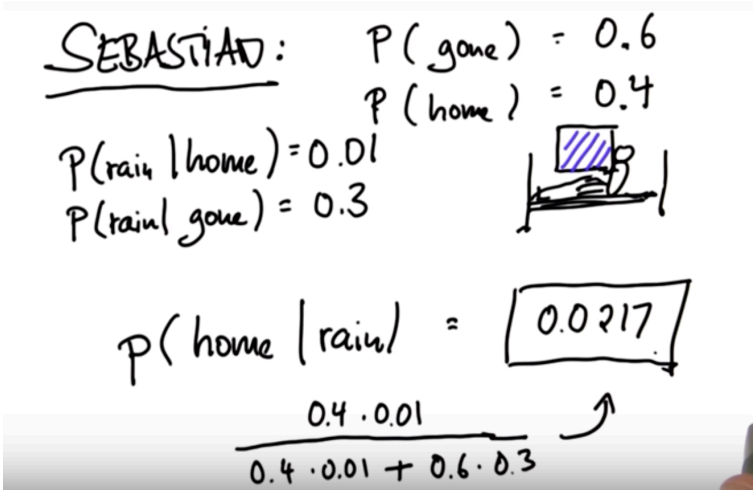
We have these quantities from above:

$$P(\text{see red}) = 0.5 \cdot 0.8 + 0.5 \cdot 0.2$$

$$P(\text{see red}) = 0.4 + 0.1$$

$$P(\text{see red}) = 0.5$$

SEBASTIAN AT HOME OR GONE?:



P(home|rain) = P(rain|home) * P(home) / P(total prob rain?)

where: P(total prob rain) = P(rain|home) * P(home) + P(rain|gone) * P(gone)

= 0.01*0.4 / (0.01*0.4 + 0.3*0.6)
= 0.0217

Learning Objectives - Bayes' Rule

The following questions will help you review what you learned in the Bayes' Rule lesson.

PRIOR KNOWLEDGE
For questions 1-3, assume you already have the following knowledge:

You're interested in finding out the probability of a car stopping if it sees a *yellow* traffic light.

- Past data tells you that the probability of a car stopping at a traffic light intersection is:

  P(S) = 0.40

- You also know that the past probability of a traffic light being yellow (as opposed to red or green) is:

  P(Y) = 0.10

Questions:

1. When a car is stopped at an intersection, data shows that **12%** of the time the light is yellow. So if we know a car is stopped, there's a 12% chance the light is yellow. This is called a *conditional probability*.

   - Given **P(S)** and **P(Y)** above, how would you represent this conditional probability in notation?

     P(Y|S) = 0.12

2. Using what you know from question 1, answer the following: if the traffic light is yellow, what is the chance that the car will stop?

   P(S|Y) = P(Y|S) * P(S) / **(** P(Y|S) * P(S) + P(not Y | S) * P(not S) **)**
   = 0.12 * 0.4 / (0.12*0.4 + 0.9*0.6)
   = 0.048 / (0.048 + 0.54)
   = **wrong!!**

   Using Bayes' rule, we know that:

   P(S|Y) = P(Y|S)*P(S) / P(Y)
   P(S|Y) = 0.12*0.4 / 0.1 = 0.48

   And intuitively this value seems about right; a car should stop about half the time when faced with a yellow light.

3.

PRIOR KNOWLEDGE:
On a four-lane highway, cars are either going fast or not fast. Faster cars should go in the leftmost lanes.

- At any given time, 20% of cars are in the left-most lane: **P(L) = 0.2**
- Overall, 40% of cars on the highway are classified as going fast: **P(F) = 0.4**
- Out of all the cars in the leftmost lane, 90% are going fast: **P(F|L) = 0.9**

4. Given the above information, if a car is going fast, what is the probability that it will be in the leftmost lane?

   P(L|F) = P(F|L) * P(L) / P(F)
   = 0.9 * 0.2 / 0.4
   = 0.45

CONDITIONAL PROBABILITY AND BASES RULE QUIZ:

Use the probabilities given above and Bayes rule to compute the following probabilities.

Probability	Meaning
P(cancer) = 0.105	Probability a patient has cancer
P(~cancer) = 0.895	Probability a patient does not have cancer
P(positive|cancer) = 0.905	Probability a patient with cancer tests positive
P(negative|cancer) = 0.095	Probability a patient with cancer tests negative
P(positive|~cancer) = 0.204	Probability a patient without cancer tests positive
P(negative|~cancer) = 0.796	Probability a patient without cancer tests negative

1. Probability a patient who tested positive has cancer, or **P(cancer|positive)**
2. Probability a patient who tested positive doesn't have cancer, or **P(~cancer|positive)**
3. Probability a patient who tested negative has cancer, or **P(cancer|negative)**
4. Probability a patient who tested negative doesn't have cancer, **P(~cancer|negative)**

1. Using the probabilities above and Bayes rule, compute the following probabilities:
   - **P(cancer|positive)**
     = P(positive|cancer) * P(cancer) /
     [P(positive|cancer)*P(cancer) + P(positive|~cancer)*P(~cancer)
     = 0.905 * 0.105 / [0.905 * 0.105 + 0.204 * 0.895]
     = <u>0.342</u>

   - **P(~cancer|positive)**
     = P(positive|~cancer) * P(~cancer) /
     [P(positive|~cancer)*P(~cancer) + P(positive|cancer)*P(cancer)
     = 0.204 * 0.895 / [0.204 * 0.895 + 0.905 * 0.105]
     = 0.18258 / 0.277605
     = <u>0.65769</u>

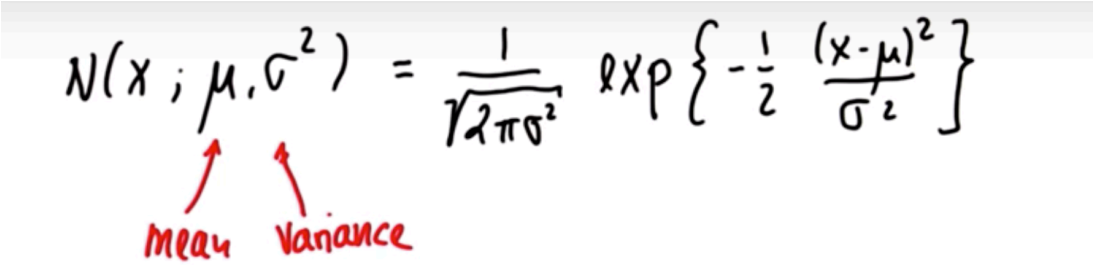
   - **P(cancer|negative)**
     = P(negative|cancer) * P(cancer) /
     [P(negative|cancer)*P(cancer) + P(negative|~cancer)*P(~cancer)
     = 0.095 * 0.105 / [0.095 * 0.105 + 0.796 * 0.895]
     = 0.009975 / 0.722395
     = <u>0.0138</u>

   - **P(~cancer|negative)**

Lesson 9: Normal Distribution Theory

FORMULA FOR NORMAL DISTRIBUTION: GAUSSIAN EXPONENTIAL

- exp = e (euler's number)
- 1 / sqrt(2 pi sigma^2) is an adjustment factor to make the area under the bell curve sum to 1
- The 'quadratics' videos do a short swift job of leading up to this formula

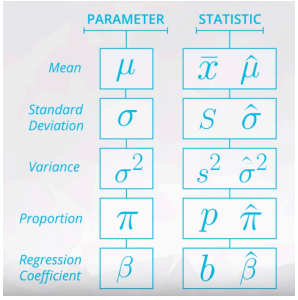


CENTRAL LIMIT THEOREM

Just extension of single probability, to binomial distribution for several probabilities, and normal distributions for many (and possibly infinite) distributions

Lesson 10: Sampling Distributions and the CLT

SAMPLING DISTRIBUTION
The distribution of a statistic. In the coffee drinking example, sample sizes were 5 students. Collecting all the sample sizes of 5 and plotting a histogram to see the differences is what is meant by sampling distribution.



Parameters pertain to a population, while all **statistics** pertain to a sample. Notation uses lower-case greek letter for parameters, and the same but with a hat ^ on top for a statistic.

We simulated the creation of sampling distributions in the previous ipython notebook for samples of size 5 and size 20, which is something you will do more than once in the upcoming concepts and lessons.

Second, we found out some interesting ideas about sampling distributions that will be iterated later in this lesson as well. We found that for proportions (and also means, as proportions are just the mean of 1 and 0 values), the following characteristics hold.

1. The sampling distribution is centered on the original parameter value.

2. The sampling distribution decreases its variance depending on the sample size used. Specifically, the variance of the sampling distribution is equal to the variance of the original data divided by the sample size used. This is always true for the variance of a sample mean!

In notation, we say if we have a random variable, X, with variance of σ^2, then the distribution of X-bar (the sampling distribution of the sample mean) has a variance of σ^2 / n.

Theorems

LAW OF LARGE NUMBERS

The **Law of Large Numbers** says that **as our sample size increases, the sample mean gets closer to the population mean**, but how did we determine that the sample mean would estimate a population mean in the first place? How would we identify another relationship between parameter and statistic like this in the future?

Three of the most common ways are with the following estimation techniques:

- Maximum Likelihood Estimation
- Method of Moments Estimation)
- Bayesian Estimation

Though these are beyond the scope of what is covered in this course, these are techniques that should be well understood for Data Scientist's that may need to understand how to estimate some value that isn't as common as a mean or variance. Using one of these methods to determine a "best estimate", would be a necessity.

CENTRAL LIMIT THEOREM

The Central Limit Theorem states that with a large enough **sample size** the sampling distribution of the mean will be normally distributed.

The **Central Limit Theorem** actually applies for these well known statistics:

1. Sample means ($\bar{x}$)
2. Sample proportions (p)
3. Difference in sample means ($\bar{x}_1 - \bar{x}_2$)
4. Difference in sample proportions ($p_1 - p_2$)

And it applies for additional statistics, **but it doesn't apply for all statistics!**

It turns out no matter how large your sample size, ***variance*** will never be normally distributed. This distribution will actually approach a distribution known as a chi-squared distribution.

Bootstrapping

Simulating the sampling distribution. **Bootstrapping** is sampling with replacement. Using **random.choice** in python actually samples in this way. Where the probability of any number in our set stays the same regardless of how many times it has been chosen. Flipping a coin and rolling a die are kind of like bootstrap sampling as well, as rolling a 6 in one scenario doesn't mean that 6 is less likely later.

Confidence Intervals

Confidence interval: estimate a range of a statistic with a certain level of confidence, rather than a discrete estimate. Like casting a net to get the stat instead of a fishing hook.

We can use bootstrapping and sampling distributions to build confidence intervals for our parameters of interest.

By finding the statistic that best estimates our parameter(s) of interest (say the sample mean to estimate the population mean or the difference in sample means to estimate the difference in population means), we can easily build confidence intervals for the parameter of interest.

DIFFERENCE OF MEANS

Useful for comparing two different drugs for example, or drug vs placebo, etc. This is A/B testing.

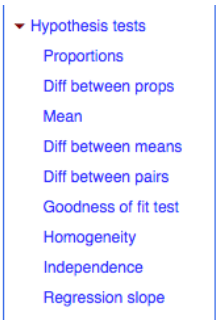
Traditional confidence intervals

These are the tests for statistical significance (hypothesis test) of confidence intervals, such as t-test, two sample t-test, paired t-test, f-test, chi-squared test, z-test and many many more, depending on the way in which the confidence interval was created.

One educated, but potentially biased opinion on the traditional methods is that these methods (bootstrapping?) are no longer necessary with what is possible with statistics with modern computing, and these methods will become even less important with the future of computing. Therefore, memorizing these formulas to throw at particular situation will be a glazed over component of this class. However, there are resources below should you want to dive into a few of the 100s if not 1000s of hypothesis tests that are possible with traditional techniques.

To learn more about the traditional methods, see the documentation **here** on the corresponding hypothesis tests.

In the left margin, you will see a drop down of the hypothesis tests available, as shown in the image below.



Each of these hypothesis tests is linked to a corresponding confidence interval, but again the bootstrapping approach can be used in place of any of these! Simply by understanding what you would like to estimate, and simulating the sampling distribution for the statistic that best estimates that value.

With large sample sizes, these end up looking very similar. With smaller sample sizes, using a traditional methods likely has assumptions that are not true of your interval. Small sample sizes are not ideal for bootstrapping methods though either, as they can lead to misleading results simply due to not accurately representing your entire population well.

Understanding sampling distributions and bootstrapping means that you can simulate the results of any confidence interval you want to build.

CONFIDENCE INTERVALS (& HYPOTHESIS TESTING) VS. MACHINE LEARNING

Confidence intervals take an aggregate approach towards the conclusions made based on data, as these tests are aimed at understanding population parameters (which are aggregate population values).

Alternatively, machine learning techniques take an individual approach towards making conclusions, as they attempt to predict an outcome for each specific data point.

Hypothesis Testing

In this video, you learned a few rules for setting up null and alternative hypotheses:

1. The H_0 (the null) is true before you collect any data.
2. The H_0 usually states there is no effect or that two groups are equal.
3. The H_0 and H_1 *(the alternative)* are competing, non-overlapping hypotheses.
4. H_1 is what we would like to prove to be true.
5. H_0 contains an equal sign of some kind, either =, <=, or =>. *(generally, the null must have an '=' sign)*
6. H_1 contains the opposition of the null, either *not* =, *not* <=, or *not* =>.
7. You saw that the statement, "Innocent until proven guilty" is one that suggests the following hypotheses are true:

   H_0 *: Innocent*
   H_1 *: Guilty*

We can relate this to the idea that "innocent" is true before we collect any data. Then the alternative must be a competing, non-overlapping hypothesis. Hence, the alternative hypothesis is that an individual is guilty.

Because we wanted to test if a new page was better than an existing page, we set that up in the alternative. Two indicators are that the null should hold the equality, and the statement we would like to be true should be in the alternative. Therefore, it would look like this:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

Here mu_1 represents the population mean return from the new page. Similarly, mu_2 represents the population mean return from the old page. Depending on your question of interest, you would change your null and alternative hypotheses to match.

Types of Errors

There are two types of errors that are possible in hypothesis testing: Type I Errors and Type II Errors. Type I Errors are considered the worst type of error.

TYPE I ERRORS
A Type I Error is when the alternative is chosen, but the null is actually true. The type I error is associated with innocence being true, when the jury decides an individual is guilty. This is the worst type of error according to the U.S. judicial system.

Type I errors have the following features:

1. You should set up your null and alternative hypotheses, so that the worse of your errors is the type I error.
2. They are denoted by the symbol α
3. The definition of a type I error is: **Deciding the alternative H_1** is true, when actually H_0 is true.
4. Type I errors are often called **false positives**.
   eg: wrongly convicting someone

TYPE II ERRORS

1. They are denoted by the symbol β
2. The definition of a type II error is: **Deciding the null H_0** is true, when actually H_1 is true.
3. Type II errors are often called **false negatives**.
   eg: setting a guilty person free

In the most extreme case, we can always choose one hypothesis (say always choosing the null) to ensure that a particular error never occurs (never a type I error assuming we always choose the null). However, more generally, there is a relationship where with a single set of data decreasing your chance of one type of error, increases the chance of the other error occurring.

Beta is frequently used to identify a type II error rate. Frequently people will talk about the "power" of a statistical test as 1 - beta (or 1 minus the type two error rate). This is the ability of an individual to correctly choose the alternative hypothesis.

Parachute Example
This example let you see one of the most extreme cases of errors that might be committed in hypothesis testing. In a type I error an individual died. In a type II error, you lost 30 dollars. In the hypothesis tests you build in the upcoming lessons, you will be able to choose a type I error threshold, and your hypothesis tests will be created to minimize the type II errors after ensuring the type I error rate is met.

Types of Hypothesis Tests

You are always performing hypothesis tests on **population parameters**, never on statistics. Statistics are values that you already have from the data, so it does not make sense to perform hypothesis tests on these values.

Common hypothesis tests include:

1. Testing a population mean **(One sample t-test)**.
2. Testing the difference in means **(Two sample t-test)**
3. Testing the difference before and after some treatment on the same individual **(Paired t-test)**
4. Testing a population proportion **(One sample z-test)**
5. Testing the difference between population proportions **(Two sample z-test)**

You can use one of these sites to provide a t-table or z-table to support one of the above approaches: **t-table**, **t-table or z-table**

There are literally 100s of different hypothesis tests! However, instead of memorizing how to perform all of these tests, you can find the statistic(s) that best estimates the parameter(s) you want to estimate, you can bootstrap to simulate the sampling distribution. Then you can use your sampling distribution to assist in choosing the appropriate hypothesis.

How to Choose Between Hypotheses

Videos 16+17. One method is the bootstrapping method to simulate the sample distribution and then choose a confidence interval and compare the result to the hypotheses.

A second method is to “simulate from the closest value under the null to the alternative that is still in the null space” (whatever the hell that means??!). Use the standard deviation of the sampling

distribution to determine what the sampling distribution would look like if it came from the null hypothesis. With a sample size of 150, we know from the central limit theorem that it will follow a normal distribution. So we can use np.random.normal() to draw from the normal distribution:

```
np.random.normal(70 <- null value, np.std(sampling distribution), 10000)
```

Then compare the actual sample mean to the distribution generated above. If it falls within the distribution (to a set confidence interval), then accept the null, otherwise reject the null.

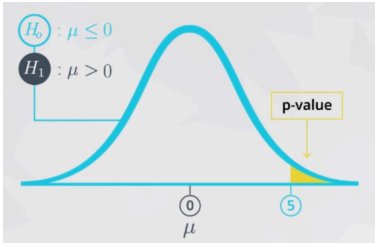
“in hypothesis testing, we could simulate a sampling distribution from the null hypothesis using characteristics that would be true if our data were to have come from the null.”

P-VALUES
The definition of a p-value is the probability of observing your statistic (or one more extreme in favor of the alternative) if the null hypothesis is true.

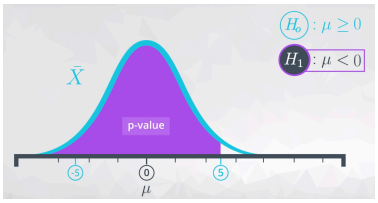
In this video, you learned exactly how to calculate this value. The **more extreme in favor of the alternative** portion of this statement determines the shading associated with your p-value.

Therefore, you have the following cases:

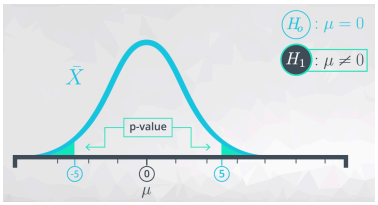
If your parameter is greater than some value in the alternative hypothesis, your shading would look like this to obtain your p-value:



If your parameter is less than some value in the alternative hypothesis, your shading would look like this to obtain your p-value:



If your parameter is not equal to some value in the alternative hypothesis, your shading would look like this to obtain your p-value:



You could integrate the sampling distribution to obtain the area for each of these p-values. Alternatively, you will be simulating to obtain these proportions in the next concepts.

There are a lot of moving parts in these videos. Let's highlight the process:

1. Simulate the values of your statistic that are possible from the null.
2. Calculate the value of the statistic you actually obtained in your data.
3. Compare your statistic to the values from the null.
4. Calculate the proportion of null values that are considered **extreme** based on your alternative.

The **p-value** is the probability of getting our statistic or a more extreme value **if the null is true**. Therefore, **small p-values** suggest our null is not true. Rather, our statistic is likely to have come from a different distribution than the null.

When the **p-value is large**, we have evidence that our statistic was likely to come from the null hypothesis. Therefore, we do not have evidence to reject the null.

By comparing our p-value to our type I error threshold (α), we can make our decision about which hypothesis we will choose.

$$pval \leq \alpha \Rightarrow \text{Reject } H_0$$

$$pval > \alpha \Rightarrow \text{Fail to Reject } H_0$$

The word **accept** is one that is avoided when making statements regarding the null and alternative. You are not stating that one of the hypotheses is true. Rather, you are making a decision based on the likelihood of your data coming from the null hypothesis with regard to your type I error threshold.

Therefore, the wording used in conclusions of hypothesis testing includes: **We reject the null hypothesis** or **We fail to reject the null hypothesis**. This lends itself to the idea that you start with the null hypothesis true by default, and "choosing" the null at the end of the test would have been the choice even if no data were collected.

**** QUIZ NOTEBOOK 26 IS REALLY CONFUSING….**

WHAT IF THE SAMPLE SIZE IS LARGE? DO HYPOTHESIS TESTS MATTER?
One of the most important aspects of interpreting any statistical results (and one that is frequently overlooked) is assuring that your sample is truly representative of your population of interest.

Particularly in the way that data is collected today in the age of computers, **response bias** is so important to keep in mind. In the 2016 U.S election, polls conducted by many news media suggested a staggering difference from the reality of poll results. You can read about how response bias played a role **here**.

Hypothesis Testing vs. Machine Learning
With large sample sizes, hypothesis testing leads to even the smallest of findings as **statistically significant**. However, these findings might not be practically significant at all.

For example, Imagine you find that **statistically** more people prefer beverage 1 to beverage 2 on a study of more than one million people. Based on this you decide to open a shop to sell beverage 1. You then find out that beverage 1 is only more popular than beverage 2 by 0.0002% (but a statistically significant amount with your large sample size). Practically, maybe you should have opened a store that sold both.

Hypothesis testing takes an aggregate approach towards the conclusions made based on data, as these tests are aimed at understanding population parameters (which are aggregate population values).

Alternatively, machine learning techniques take an individual approach towards making conclusions, as they attempt to predict an outcome for each specific data point.

WHAT IF TEST MORE THAN ONCE?
When performing more than one hypothesis test, your type I error compounds. In order to correct for this, a common technique is called the **Bonferroni** correction. This correction is **very conservative**, but says that your new type I error rate should be the error rate you actually want divided by the number of tests you are performing.

Therefore, if you would like to hold a type I error rate of 1% for each of 20 hypothesis tests, the **Bonferroni** corrected rate would be 0.01/20 = 0.0005. This would be the new rate you should use as your comparison to the p-value for each of the 20 tests to make your decision.

Other Techniques
Additional techniques to protect against compounding type I errors include:

1. Tukey correction

2. Q-values

A/B Testing

A/B tests are used to test changes on a web page by running an experiment where a **control group** sees the old version, while the **experiment group** sees the new version. A **metric** is then chosen to measure the level of engagement from users in each group. These results are then used to judge whether one version is more effective than the other. A/B testing is very much like hypothesis testing with the following hypotheses:

- **Null Hypothesis:** The new version is no better, or even worse, than the old version
- **Alternative Hypothesis:** The new version is better than the old version

If we fail to reject the null hypothesis, the results would suggest keeping the old version. If we reject the null hypothesis, the results would suggest launching the change. These tests can be used for a wide variety of changes, from large feature additions to small adjustments in color, to see what change maximizes your metric the most.

A/B testing also has its drawbacks. It can help you compare two options, but it can't tell you about an option you haven't considered. It can also produce bias results when tested on existing users, due to factors like change aversion and novelty effect.

- **Change Aversion:** Existing users may give an unfair advantage to the old version, simply because they are unhappy with change, even if it's ultimately for the better.
- **Novelty Effect:** Existing users may give an unfair advantage to the new version, because they're excited or drawn to the change, even if it isn't any better in the long run.

You'll learn more about factors like these later.

BUSINESS EXAMPLE
In this case study, you'll analyze A/B test results for Audacity. Here's the customer funnel for typical new users on their site:

View home page > Explore courses > View course overview page > Enroll in course > Complete course

Audacity loses users as they go down the stages of this funnel, with only a few making it to the end. To increase student engagement, Audacity is performing A/B tests to try out changes that will hopefully increase conversion rates from one stage to the next.

We'll analyze test results for two changes they have in mind, and then make a recommendation on whether they should launch each change.

The first change Audacity wants to try is on their homepage. They hope that this new, more engaging design will increase the number of users that explore their courses, that is, move on to the second stage of the funnel.

The metric we will use is the click through rate for the Explore Courses button on the home page. **Click through rate (CTR)** is often defined as the the number of clicks divided by the number of views. Since Audacity uses cookies, we can identify unique users and make sure we don't count the same one multiple times. For this experiment, we'll define our click through rate as:

CTR: # clicks by unique users / # views by unique users

Now that we have our metric, let's set up our null and alternative hypotheses:

$$H_0 : CTR_{new} \leq CTR_{old}$$

$$H_1 : CTR_{new} > CTR_{old}$$

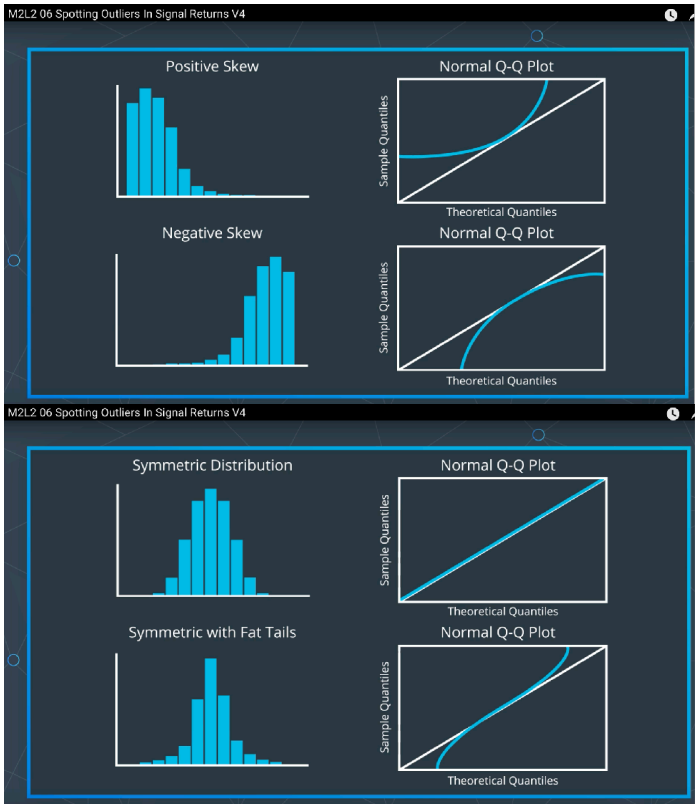
Our alternative hypothesis is what we want to prove to be true, in this case, that the new homepage design has a higher click through rate than the old homepage design. And the null hypothesis is what we assume to be true before analyzing data, which is that the new homepage design has a click through rate that is less than or equal to that of the old homepage design. As you've seen before, we can rearrange our hypotheses to look like this:

$$H_0 : CTR_{new} - CTR_{old} \leq 0$$
$$H_1 : CTR_{new} - CTR_{old} > 0$$

More left to complete….

Lesson 9: Spotting Outliers in Signal Returns

QQ Plots



Lesson 10: Regression

Distributions

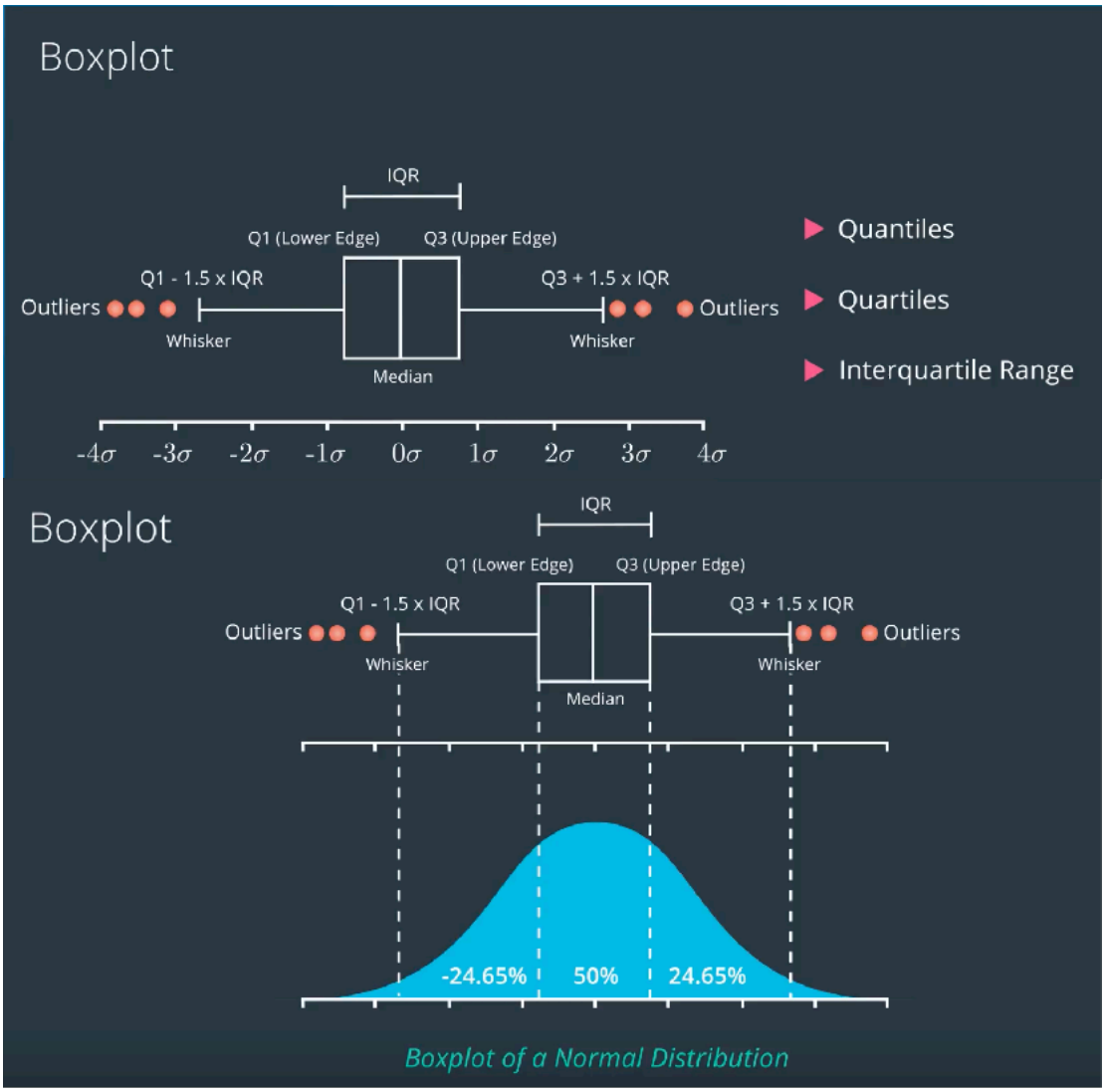
- If not normal, test for validity could erroneously endorse of model that is not valid
  - Types: normal, log-normal, exponential, uniform
- Random variable —> normally distributed within some constraints

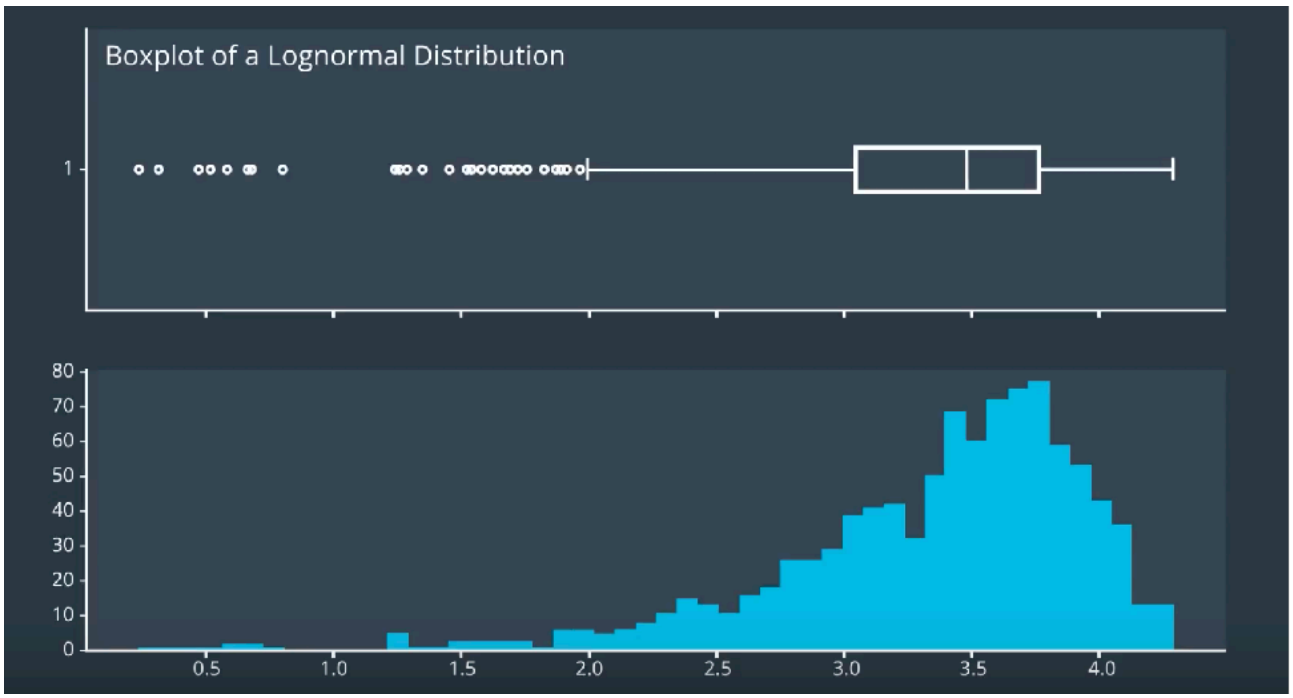
PARAMETERS

- Formula for prob density function (PDF) has parameters to change its shape
- Notation: X ~ D means a random variable ‘X’ follows a probability distribution ‘D’
  - P(x|D) = p(x) reads: probability of x given D
  - So: P(2|N) = p(2) = a number between 0 and 1 <— the probability
  - ’N’ is for normal distribution
- The parameters are:
  - μ = mean
  - σ = standard deviation

Testing for normality

A histogram lets us check if a distribution is symmetric/skewed, and if it has fat tails. QQ plots help us compare any two distributions, so they can be used to compare distributions other than the normal distribution. A box whisker plot lets us check for symmetry around the mean. A QQ-plot lets us compare our data distribution with a normal distribution (or any other theoretical "ideal" distribution). “goodness of fit”

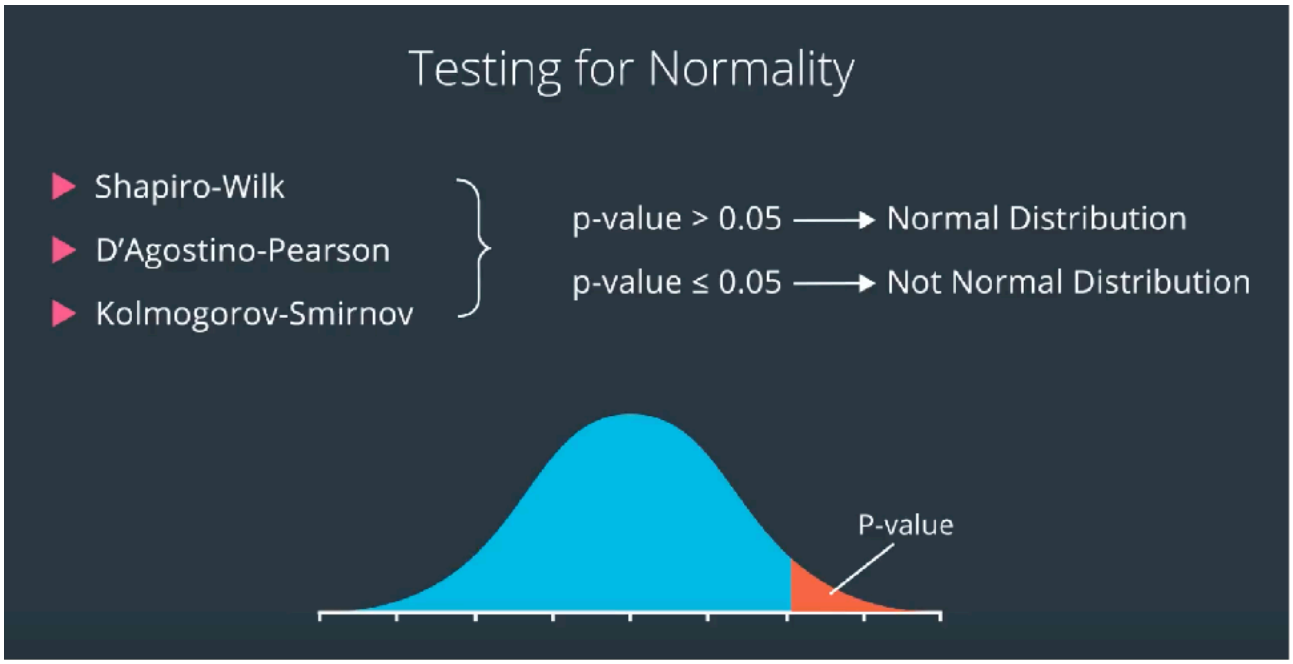




HYPOTHESIS TESTS

There are three hypothesis tests that can be used to decide if a data distribution is normal. These are the Shapiro-Wilk test, the D'Agostino-Pearson test, and the Kolmogorov-Smirnov test. Each of these produce p-value, and if the p-value is small enough, say 0.05 or less, we can say with a 95% confidence that the data is not normally distributed. Shapiro-Wilk tends to perform better in a broader set of cases compared to the D'Agostino-Pearson test. In part, this is because the D'Agostino-Pearson test is used to look for skewness and kurtosis that do not match a normal distribution, so there are some odd non-normal distributions for which it doesn't detect non-normality, where the Sharpiro-Wilk would give the correct answer.

The Kolmogorov Smirnov test can be used to compare distributions other than the normal distribution, so it's similar to the QQ plot in its generality. To do a normality test, we would first rescale the data distribution (subtract the mean and divide by its standard deviation), then compare the rescaled data distribution with the standard normal distribution (which has a mean of zero and standard deviation of 1). In general, the Shapiro-WIlk test tends to be a better test than the Kolmogorov Smirnov test, but not in all cases.



HETEROSKEDASTICITY
One of the assumptions of linear regression is that its input data are homoscedastic. A visual way to check if the our data is homoscedastic is a scatter plot (like the one we saw in the video).
If our data is heteroscedastic, a linear regression estimate of the coefficients may be less accurate (further from the actual value), and we may get a smaller p-value than should be expected, which means we may assume (incorrectly) that we have an accurate estimate of the regression coefficient, and assume that it's statistically significant when it's not.

Breusch-Pagan Test
Note, we'll cover the Breusch-Pagan test for heteroscedasticity in more detail after we learn about regression.
- P-value < 0.05 ==> hetroscedastic
- P-value > 0.05 ==> homoscedastic

Stationary data: mean, variance and covariance stay consistent (stationary) over time

Transforming Data

Make normal by apply log function to it
Make data homoscedastic, can apply the time difference ==> just convert to log returns!!

BOX-COX TRANSFORMATION
A monotonic transformation (evens out the spacing, but maintains order)