Data Science

Yogesh Kulkarni

Inferential Statistics

What is Inferential Statistics?

Inferring from whats not fully available.

- Discover properties of larger group by studying smaller group, with a quantified confidence that results generalize well with the larger group.
- There is a chance (here comes Probability) that sample's behavior can be similar to Population's.
- Used to compare variables/groups

Inferential Statistics Objectives

- To estimate, predict a population 'parameter' using sample value (called 'statistic')
- Eg Predicting election results based of Exit Polls.
- Test Hypotheses.
- E.g. Whether the new method realy works or not.

Estimate/Predict Population Parameter

- Population: Big/Whole dataset
- Sample: Small/test dataset
- Point Estimate: using a single value of sample (called 'statistic') to predict corresponding value in the population (called 'parameter')
- E.g. Mean of sample can very well be used to estimate Population mean.

Test Hypothesis

- Types: Null Hypothesis (H_0) and Alternate Hypothesis (H_1)
- Null Hypothesis: two groups under study are same
- Alternate Hypothesis: two groups are different
- Goal: to prove Null Hypothesis wrong

Inferential Statistics Types

Based on type of distribution:

- Parametric: Variable is assumed normally distributed
- NonParametric: Distribution of scores is severely skewed

Inferential Statistics Types

Based on type of underlying variable:

- For nominal variables: compare distributions, central tendencies, measure spreads
- For ordinal variables: non-parametric tests, rank order differences
- Continuous: Ratio or Interval variables: regression analysis

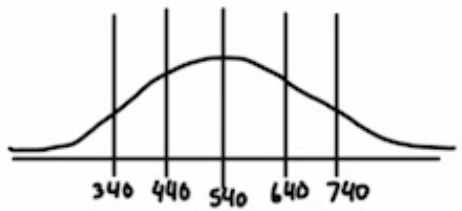
Some Statistical Terms

Z-score

- Meaning: number of std deviations a value is from the mean
- Example:
  - $\mu = 540$
  - $\sigma = 100$
  - $z =$?

GMAT Scores

Average: 540
Standard deviation: 100



Value	Z-score
640	1
340	-2
540	0
690	1.5
420	-1.2

$$\frac{420 - 540}{100} \Rightarrow \frac{-120}{100} = -1.2$$

Z-score

- 1 std deviation would be 540 ± 100 ie 640 or 440
- 2 std deviations would be 540 ± 200 ie 740 or 340
- For 690? Simple intrapolation $\frac{690-540}{100} = 1.5$
- $z = \frac{x-\mu}{\sigma}$
- z of a particular value tells us percentile

Z-score

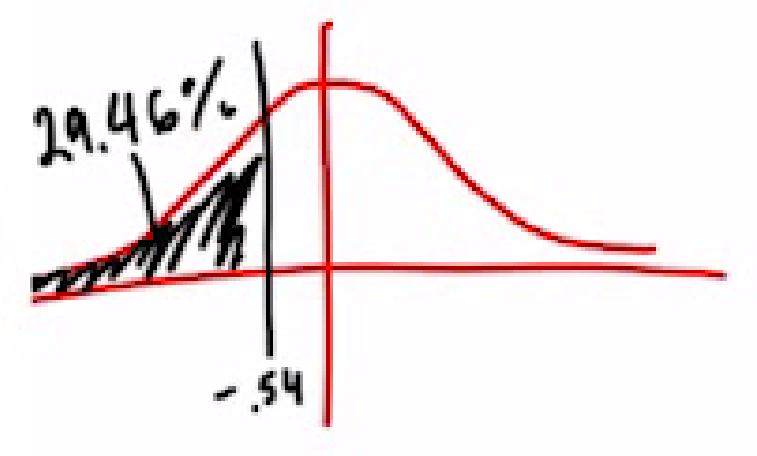
z-table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877

- To look for z score of -0.54 go to row -0.5 and then find column for 0.04. Its 0.2946
- This value tells us proportion of values to the left of that z
- So, for $z = -0.54$ there are 29.46% values to the left.

Z-score

z of a particular value tells us percentile



- You can go backwards as well
- If you want score which is at 95% percentile, then
- Find z value from table (almost 1.645)
- Find x

Inferences for Normal Distribution

GMAT example:

- Finding percentile for score 705
- Find z score first: $z = \frac{x-\mu}{\sigma} = \frac{705-540}{100} = 1.65$
- Go to z table, find the value, make it percentile: 95.05%

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803

Quiz

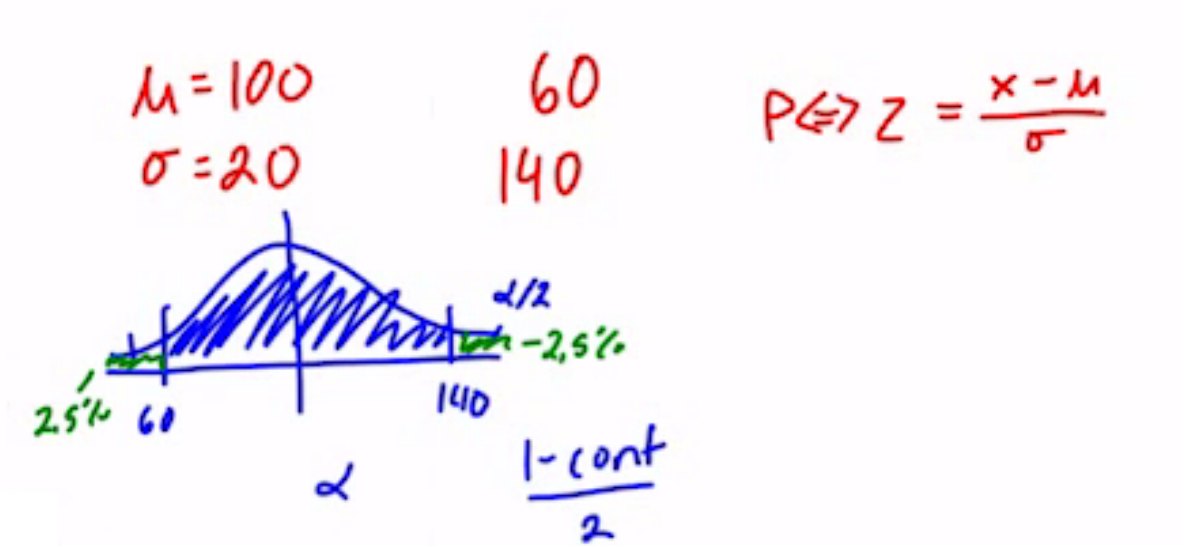
In a telephone call center, the typical agent averages 86 phone calls per day with a standard deviation of 11 calls. Management would prefer that agents make between 80 and 100 phone calls each day. What percentage of agents are meeting management's expectations?

Solution

- z for 100 calls: 100 - 86/11 = 1.27
- z for 80 calls: 80 - 86/11 = -.545
- value from table for z=1.27: 0.8980
- value from table for z=-.545: 0.2929
- Difff between them are the people doing correct amounts of call: about 60%

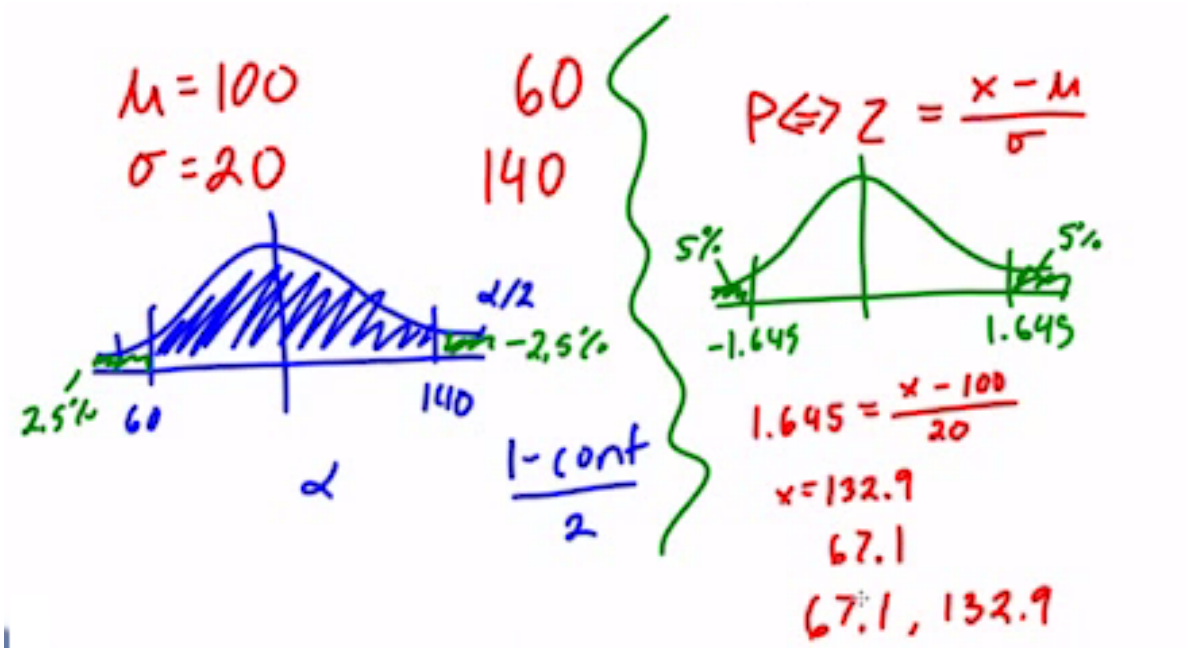
Confidence Interval

- In a restaurant, average 100 customers come, std dev is 20.
- 95% always falls between 2 std dev, so between 60 and 140.
- So, theer is 95% confidence that the number of customers will be between 60 to 140.
- Remaning non-confident portion is α so $\alpha/2$ on both sides.
- What if we want to be only 90% confident?



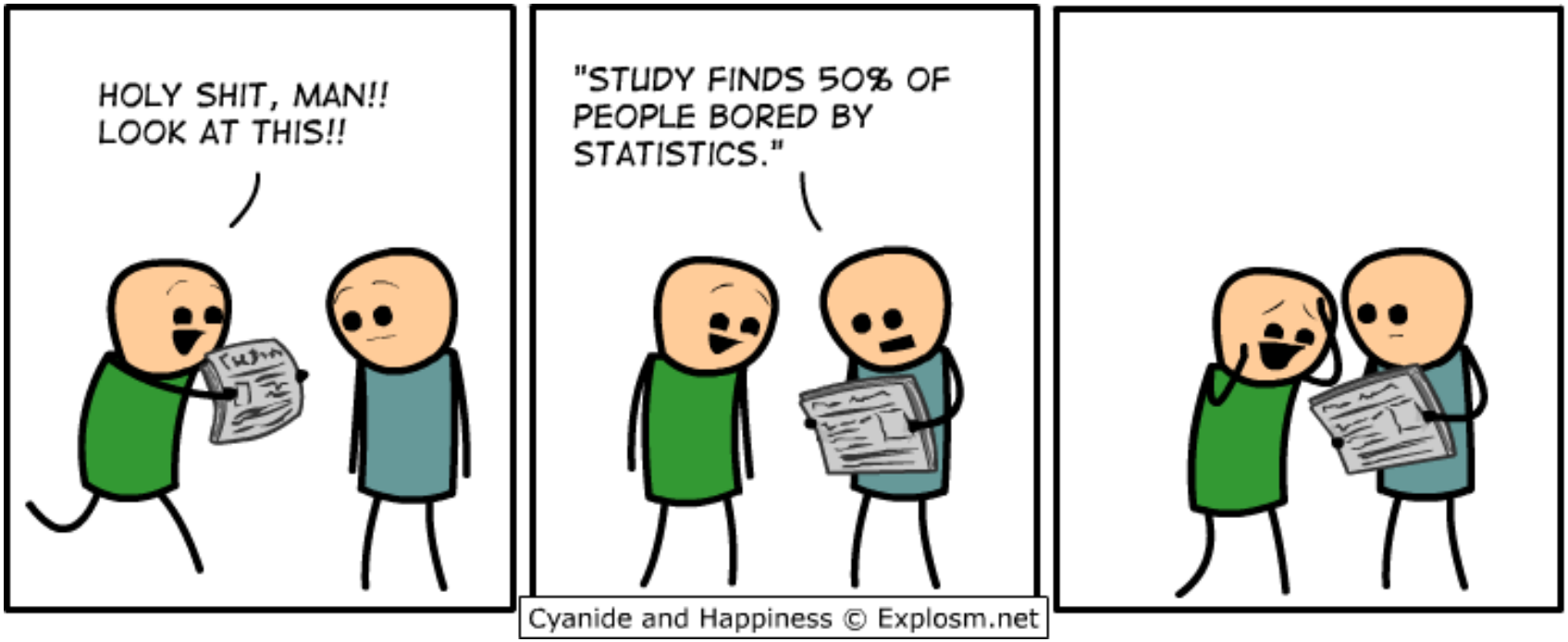
Confidence Interval

- What if we want to be only 90% confident?
- so $\alpha/2 = 0.05$
- Find z for that: Its about - 1.645
- Higher band is 1.645
- Caluclate x for both upper and lower boundaries: 132.9 and 67.1
- For 90% confidence, guest would be in range 67 to 133 guests.

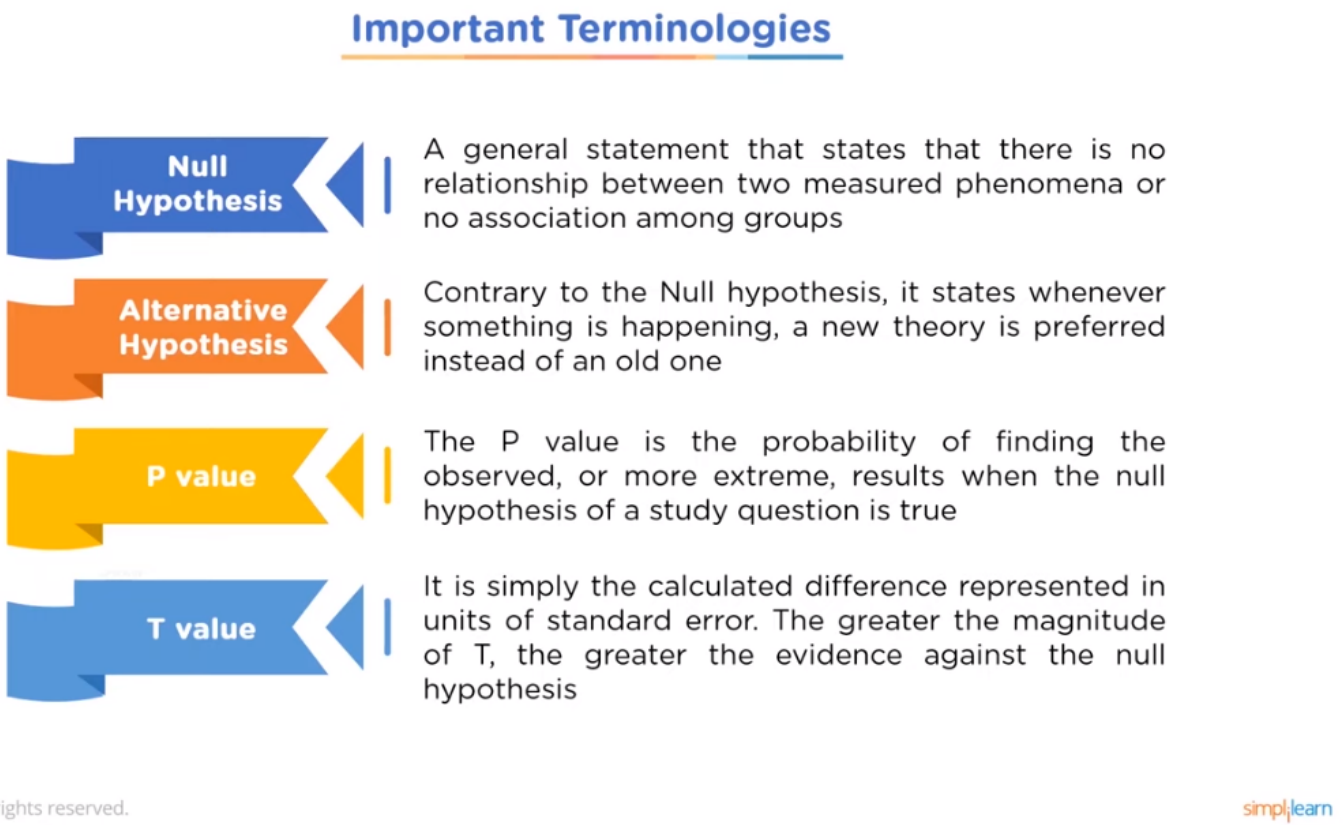


Statistical Testing

Tests

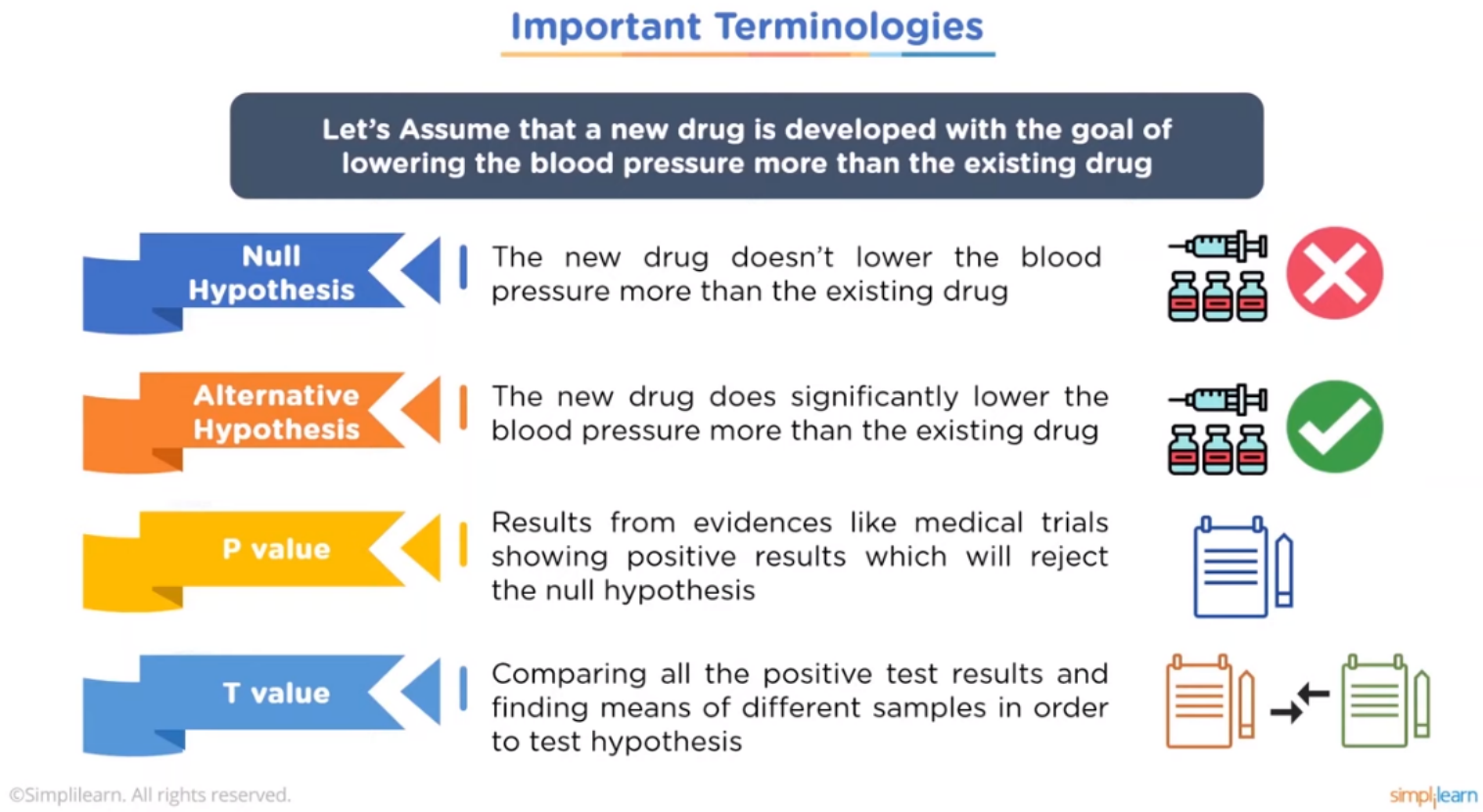

HOLY SHIT, MAN!!
LOOK AT THIS!!
"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."
Cyanide and Happiness © Explosm.net

Test Hypothesis



(Ref: Mathematics For Machine Learning — Essential Mathematics - Machine Learning Tutorial — Simplilearn)

Test Hypothesis: Example



(Ref: Mathematics For Machine Learning — Essential Mathematics - Machine Learning Tutorial — Simplilearn)

Hypothesis

- Formulate effect/relation which needs to be tested
- Null Hypothesis (H_0): Assume Guess/relation does not exist.
- Alternate Hypothesis (H_1): Assume Guess/relation does exist.
- Eg. you are trying to state that your method has positive effect and there is some change.

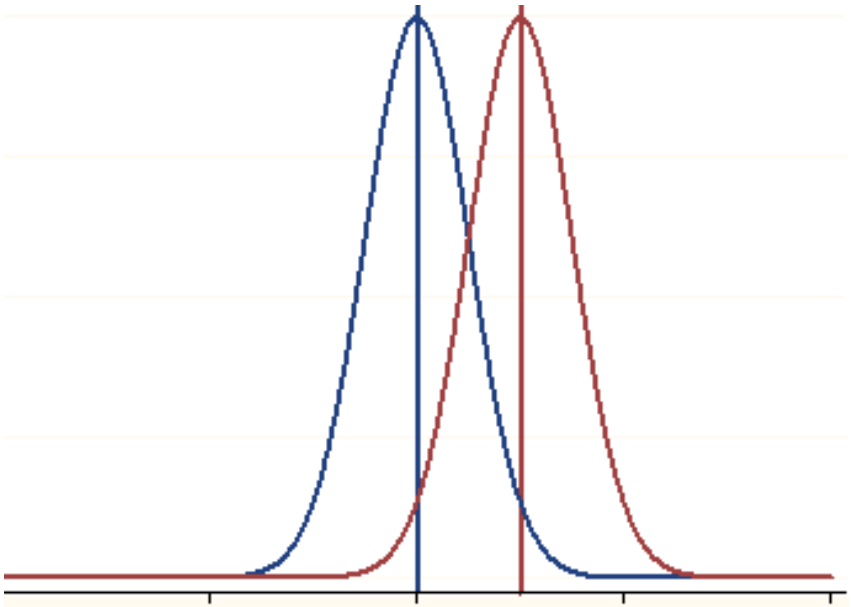
Hypothesis

- Objective: Prove Null Hypothesis wrong.
- Get enough evidence to disprove the Null Hypothesis.
- **Your never prove that H_1 is true, you can only reject H_0**

Hypothesis

Example:

- There is one distribution for H_0.
- There is another distribution for H_1
- Your model will give some output distribution which one needs to check if its in H_0s acceptance region or rejection region, and how much?



Confidence Intervals

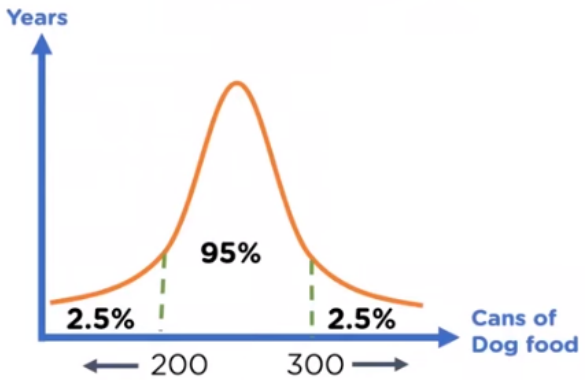
Confidence Intervals

A Confidence Interval is a range of values we are sure our true values of our observations lies in

Let's say you asked dog owners around you and asked them how many cans of food do they buy per year

Through calculations you got to know that the on an average around 95% of the people bought around 200-300 cans of food.

Hence, we can say that we have a confidence interval of (200, 300) where 95% of our values lie



simplilearn

(Ref: Mathematics For Machine Learning — Essential Mathematics - Machine Learning Tutorial — Simplilearn)

Errors

- Whenever you reject a hypothesis, there is a chance that you make some errors there.
- They care classified into Type I and Type II errors

Errors

- Type I: Incorrectly reject H_0.
- 'Reject H_0', meaning H_1 can be accepted, which in turn talks about Test being positive.
- But then 'incorrectly' means Falsely , so the whole thing is False Positive.
- Test is positive but it could actually be wrong. False Positives.
- Probability of making Type I errors is α

Errors

- Type II: Fail to reject H_0 when you should have done it.
- 'reject H_0' meaning H_1 can be accepted, which in turn talks about Test being Positive.
- 'Fails to' do that means the Test came out negative.
- This should have been done, but did not happen, so 'False'. So the whole thing is False Negative.
- Test is negative but but it could actually be wrong, ie the disease is there. False Negative.
- Probability of making Type II errors is β

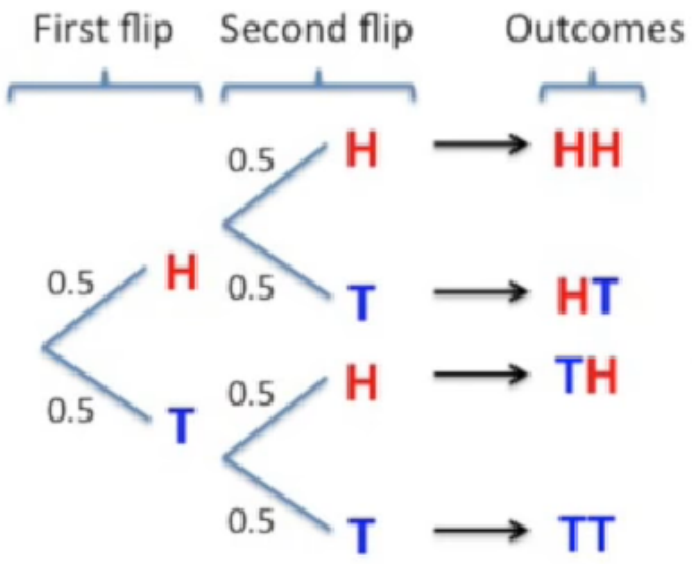
Some Required Definitions

- Normality: Most of the test assume that, the data or test statistics or some function in testing procedure under consideration follows Normal distribution.
- Decision criteria: z critical values: boundaries for rejection or non rejection region

P-Value

P-Value

- Do not confuse this p with probability. They are related but not the same.
- Experiment: What is the probability of getting 2 heads in a row? Whats the p-value of getting 2 heads in a row?
- Probability of getting 1 heads and 1 tails is $0.25 + 0.25 = 0.5$. Here order did not matter, ie HT and TH is same.

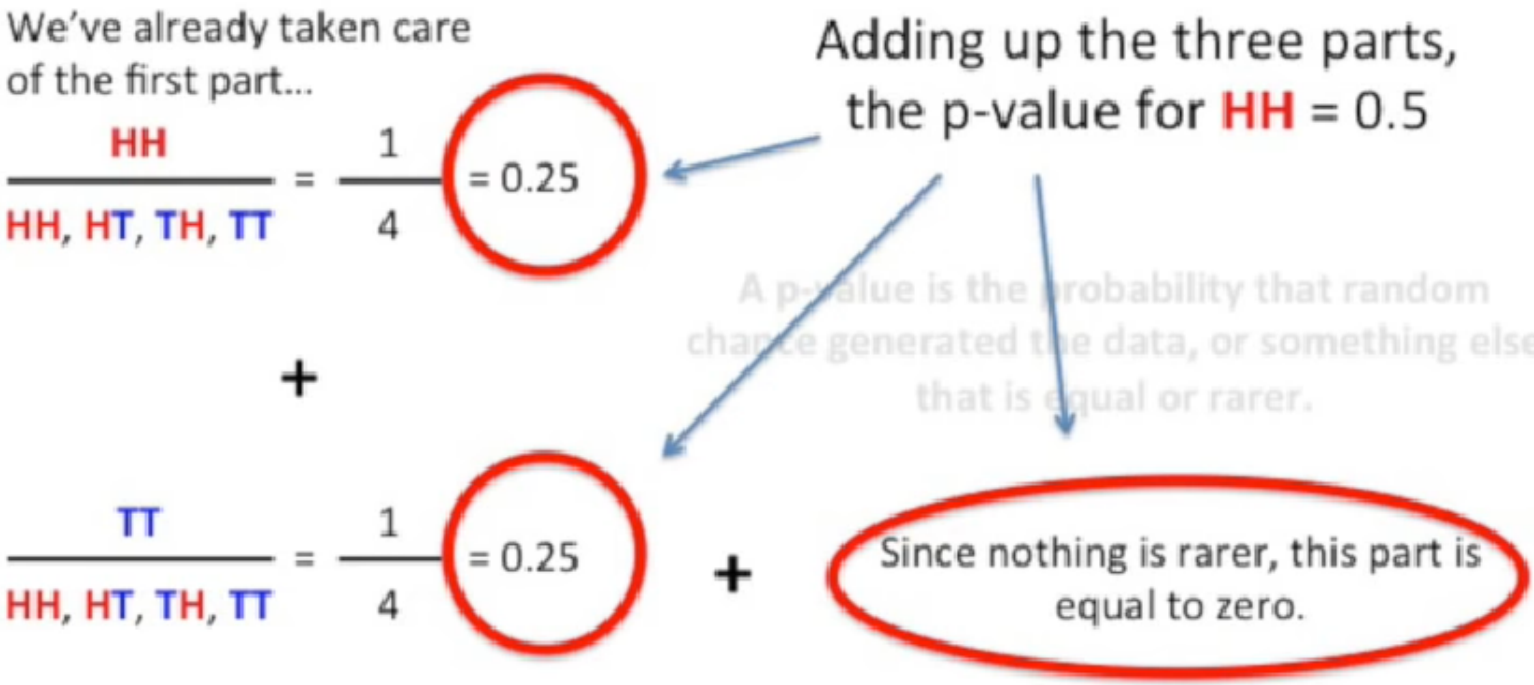


$$\frac{\text{The number of times HH occurred}}{\text{The total \# of outcomes}} = \frac{1}{4} = 0.25$$

(Ref: StatQuest: P Values, clearly explained - Josh Starmer)

P-Value

- P-Value of getting HH?
- P-value is the probability that random chance generated data (which is 0.25) is equal (ie of TT, which has same probability, ie 0.25) or rarer (there are none here)
- So P-value for HH is 0.5



(Ref: StatQuest: P Values, clearly explained - Josh Starmer)

Level of Significance: p-value

- Level of Significance: Probability of rejecting null hypothesis when it is true. Represented by Greek letter 'alpha'.
- if $p < \alpha$: there is statistically significant difference between groups
- if $p > \alpha$: there is NOT MUCH statistically significant difference between groups
- α is generally 0.05
- So, only incorrectly rejecting H_0 is ok upto 5%.
- Type I error only upto 5%

Level of Significance: p-value: Example

Why p-value is the deciding factor for accepting or rejecting a hypothesis we develop before any experiment?

- You have launched a product (e.g. a phone) in the market.
- And you get customer feedback that the phone has over heating problem.
- As the phone is already launched in the market you can't recall all of them to test if the majority of the phones have overheating problem due to some manufacturing problem.

(Ref:
https://www.datasciencecentral.com/profiles/blogs/significance-of-p-value)

Level of Significance: p-value: Example

- Hence to address the issue you have decided to take surveys
- Lets do a statistical test to overrule your apprehension regarding the manufacturing issue
- Now you have a random sample of 500 feedbacks against the total number of 250000 phones you have sold.
- Population size $= 25000$, Sample size $= 500$

Level of Significance: p-value: Example

- The test conducted within the factory says that at maximum 3% of the phone may have the overheating problem which is due to some random event (nothing to do with manufacturing as such), may be due to overcharging or overusing.
- This is acceptable to your company.
- Otherwise you have to recall all the phones from market to do a re-evaluation.

Overheating	Percentage	Number of phones
Yes	3%	7500
No	97%	242500

Now you have to take a decision whether you will recall the phone from the market or not.

Level of Significance: p-value: Example

Hypothesis: Set up null hypothesis and alternate hypothesis first

- Null hypothesis (H0):= Overheating of phones are as expected and due the some random events which was observed during the production process.
- Alternate hypothesis (H1):= Overheating of the phones are not due to some random events. There must be some strong reason behind the overheating.
- If p value is large you accept null hypothesis.
- If p value is small you fail to accept null hypothesis. You believe that the alternate hypothesis is somewhat acceptable. Your test is statistically significant.

Level of Significance: p-value: Example

Data: Scenario 1:

Overheating	Percentage	Number of phones
Yes	4%	20
No	96%	480

Data: Scenario 2:

Overheating	Percentage	Number of phones
Yes	10%	50
No	90%	450

The sample size n = 500.
m = 2. (Number of categorical values (Here they are Overheating & Non-overheating OR Yes & No))

Level of Significance: p-value: Example

Experiments and Results: let's set the confidence interval first and know about the types of error.

- H1 Error: - We reject null hypothesis even though it is true. (In our example, even after observing that the overheating of phones happen due to random events we still reject null hypothesis and assume that the overheating happens due to some manufacturing issue.)
- H2 Error: - We retain null hypothesis even though it is false. (In our example, even after observing that the overheating of phones happen due to some manufacturing issue we still accept null hypothesis and assume that the overheating happens due to random events.)

Level of Significance: p-value: Example

Experiments and Results: let's set the confidence interval first and know about the types of error.

- CI: Confidence interval for our test will be 95%. This means we are 95% confident that the test results of our sample will fall 95% close to the population.
- α (Significance level) is the probability of H1 error. Here $\alpha = 0.05$.

Level of Significance: p-value: Example

Scenario 1:

Observed Number	Observed Percentage (Obs)	Expected (Exp)	Expected percentage (Exp)	Residual=(Obs-Exp)	$(Obs-Exp)^2$	$X^2=(Obs-Exp)^2/Exp$
20	4	15	3	1	1	0.333333333
480	96	485	97	-1	1	0.010309278
					SUMMATION (X^2)	0.343642612
					α	0.05
					p – value (calculated)	0.56
					df = (m -1)	1

- From the experiment we saw the Chi Square (X2) value is 0.3436 and p-value is 0.56 (calculated).
- p-value is greater than the α (=0.05).

Level of Significance: p-value: Example

Scenario 1: Search the critical value of X2 from the table.

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

- The next critical value after X2=0.3436 is 2.706 and its corresponding p-value is 0.1.
- And our X2 value lies between X20.10 and X20.90.
- This means our p-value (though we have already calculated) calculated above is between 0.1 to 0.9 and it is not smaller than 0.05.

Level of Significance: p-value: Example

Scenario 1:

- We fail to prove any evidence against null hypothesis. We can't reject null hypothesis.
- This means the number of overheating phones we found from the survey is not significantly different than what we observed during our production process.

Level of Significance: p-value: Example

Scenario 2:

Observed Number	Observed Percentage (Obs)	Expected (Exp)	Expected percentage (Exp)	Residual=(Obs-Exp)	$(Obs-Exp)^2$	$X^2=(Obs-Exp)^2/Exp$
50	10	15	3	7	49	16.33333333
450	90	485	97	-7	49	0.505154639
					SUMMATION (X^2)	16.83848797
					α	0.05
					p – value (calculated)	4.10E-05
					df = (m -1)	1

- From the experiment we saw the Chi Square (X2) value is 16.8384 and p –value is 4.10E-05 (calculated).
- p-value is smaller than the α (=0.05).

Level of Significance: p-value: Example

Scenario 2: We can also search the critical value of X2 from the table.

	$P(X \leq x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

- The next critical value after X2=X2=16.83 is 2.706 and its corresponding p-value is 0.01.
- This means our p-value (though we have already calculated) should be less than 0.01 and it is obviously less than 0.05 s well.

Level of Significance: p-value: Example

Scenario 2:

- We have to reject null hypothesis.
- This means the number of overheating phones we found from the survey is significantly different than what we observed during our production process. And it is not due to some random events.

Level of Significance: p-value: Example

Outcome of the Problem Statement:

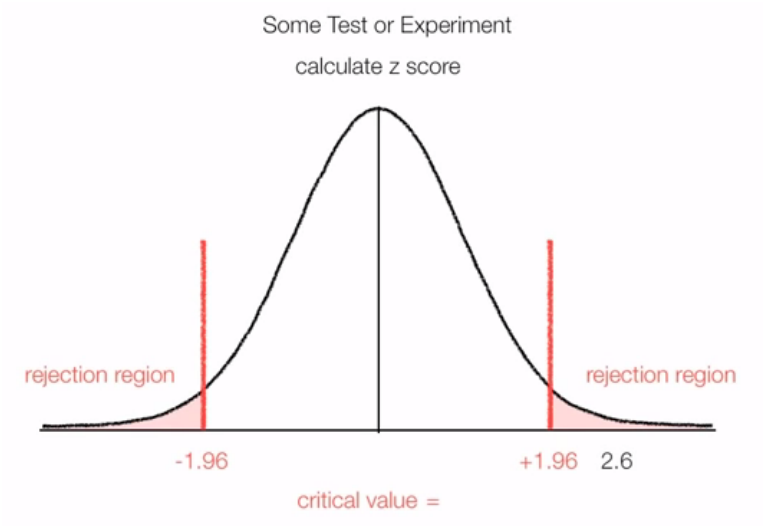
- For the 1st scenario we will accept the null hypothesis and our phones don't have any manufacturing problem.
- For the 2nd scenario we have to check the phones for their manufacturing problem as we have strong evidence against the null hypothesis.

Level of Significance: p-value

- P-value: The probability of getting test statistics as extreme as observed, under the null hypothesis is P-value. It is the observed level of significance.
- E.g. study placebo group vs new medication group to lower BP.
- Mean BP in the treatment group was less by 20mm. Assuming that H_0 is true (ie medication has no effect).
- If we repeated the study, then the difference in means can go AT MAX to 20mm.
- P value is how much data disagrees with the Null Hypothesis.
- High similarity High P
- Low p= reject H_0, as there is much diff
- High p= fail to reject H_0. no real diff exists.
- Whether P is low or high is determined by Level of Significations

Graphical: Critical values

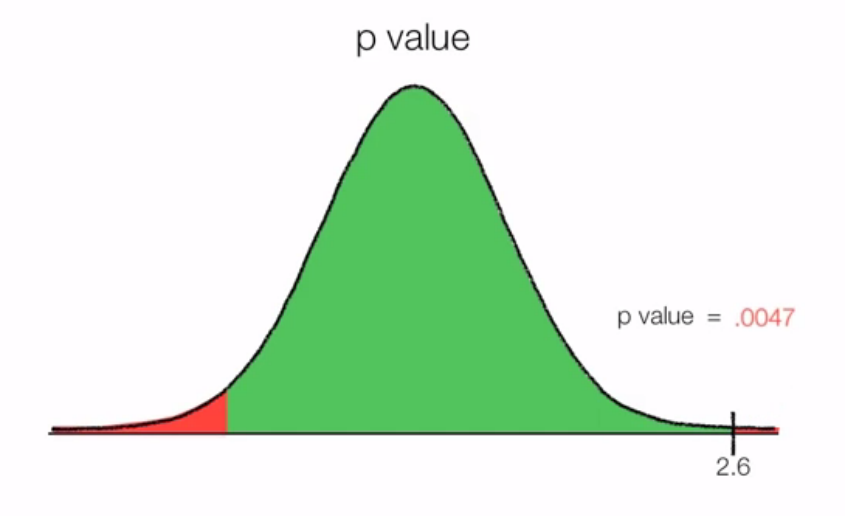
At 95% α



- Compare if calculate z score of 2.6 is it in rejection region?
- Yes, so reject H_0
- Basically, if we are in non-reject region, H_0 holds, so no real diff.
- Only when you cross the band, you are making some diff.
- That band is controlled by α.

Graphical:P Value

How significant my result is.



- In one tail, we have 0.025 area
- Given z score of 2.6, is above critical value of +1.96.
- P is area above 2.6
- It can be calculated from area under normal curve' table to be 0.0047
- $p < 0.025$. So, reject H_0

P-Value : Another Example on How to Calculate P Value

Introduction (recap)

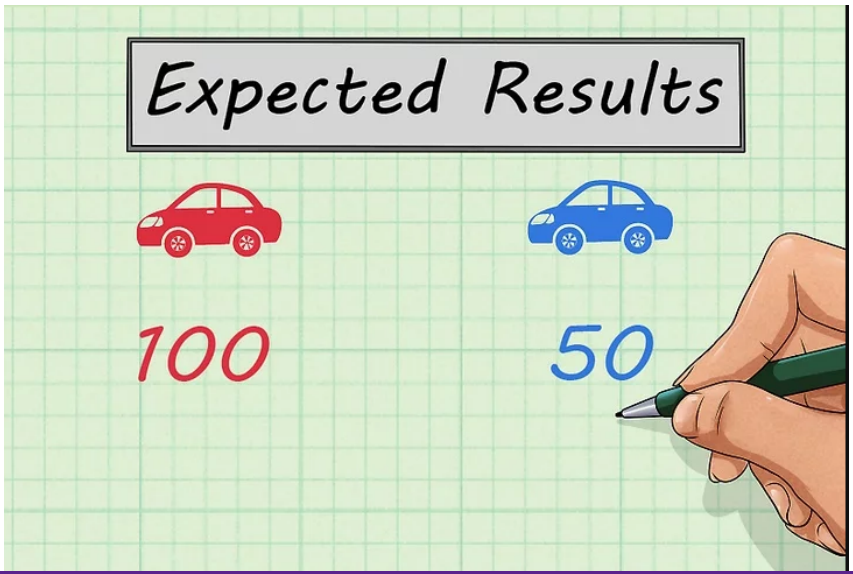
- P value is a statistical measure that helps scientists determine whether their hypotheses are correct.
- Usually, if the P value of a data set is below a certain pre-determined amount (like, for instance, 0.05), scientists will reject the "null hypothesis" of their experiment - in other words, they'll rule out the hypothesis that the variables of their experiment had no meaningful effect on the results.
- Today, p values are usually found on a reference table by first calculating a chi square value.

Expected Results (recap)

- Usually, when scientists conduct an experiment and observe the results, they have an idea of what "normal" or "typical" results will look like beforehand. This can be based on past experimental results, trusted sets of observational data, etc.
- For your experiment, determine your expected results and express them as a number.

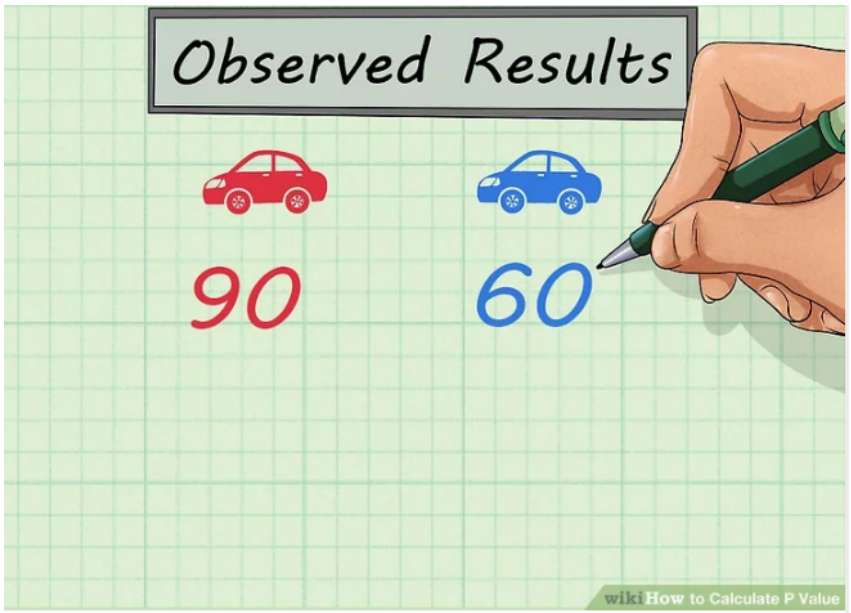
Toy Experiment

- Let's say prior studies have shown that, nationally, speeding tickets are given more often to red cars than they are to blue cars
- Let's say the average results nationally show a 2:1 preference for red cars. We want to find out whether or not the police in our town also demonstrate this bias by analyzing speeding tickets given by our town's police.
- If we take a random pool of 150 speeding tickets given to either red or blue cars in our town, we would expect 100 to be for red cars and 50 to be for blue cars if our town's police force gives tickets according to the national bias.



Determine your experiment's observed results

You can conduct your experiment and find your actual (or "observed") values. Again, express these results as numbers. Following were numbers from LOCAL (not national) observations. See if chaning this data source (National to local) had any significant effect.

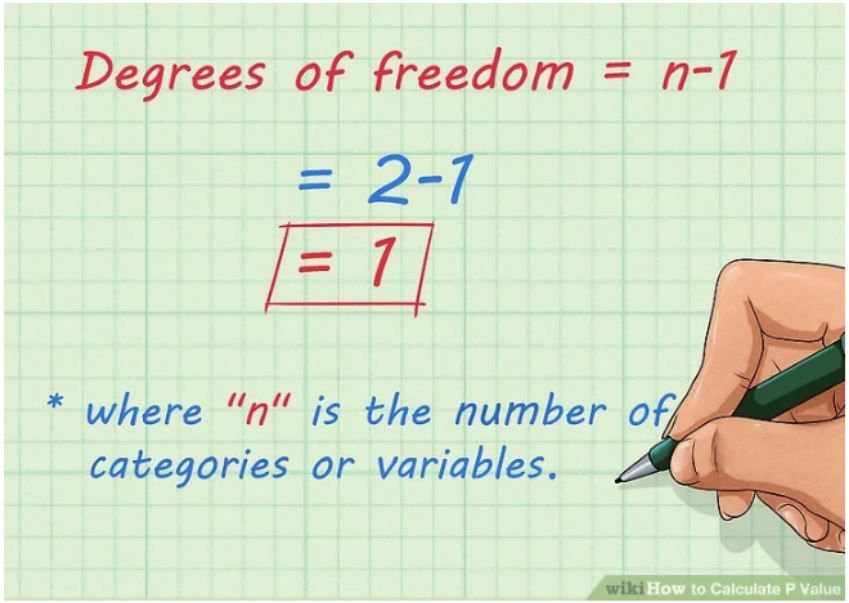


Experiment's observed results

- The observed results differ from this expected results, two possibilities are possible: either this happened by chance, or our experimental variables caused the difference.
- The purpose of finding a p-value is to determine whether the observed results differ from the expected results to such a degree that the "null hypothesis" - the hypothesis that there is no relationship between the experimental variable(s) and the observed results - is unlikely enough to reject. It did not happen by chance.
- Are our local police as biased as the national average suggests, and we're just observing a chance variation? A p value will help us determine this.

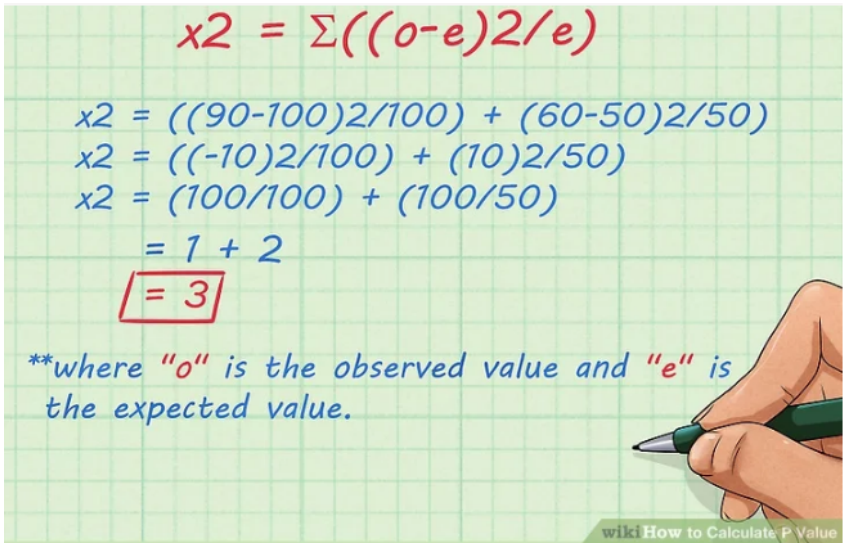
Determine your experiment's degrees of freedom

- Degrees of freedom are a measure the amount of variability involved in the research, which is determined by the number of categories you are examining. The equation for degrees of freedom is Degrees of freedom = n-1, where "n" is the number of categories or variables being analyzed in your experiment.
- Example: Our experiment has two categories of results: one for red cars and one for blue cars. Thus, in our experiment, we have 2-1 = 1 degree of freedom. If we had compared red, blue, and green cars, we would have 2 degrees of freedom, and so on.



Compare expected results to observed results with chi square

- Chi square(written "x2") is a numerical value that measures the difference between an experiment's expected and observed values.
- The equation for chi square is: $x2 = \sum((o - e)2/e)$, where "o" is the observed value and "e" is the expected value
  - $x2 = ((90 - 100)2/100) + (60 - 50)2/50)$
  - $x2 = ((-10)2/100) + (10)2/50)$
  - $x2 = (100/100) + (100/50) = 1 + 2 = 3$

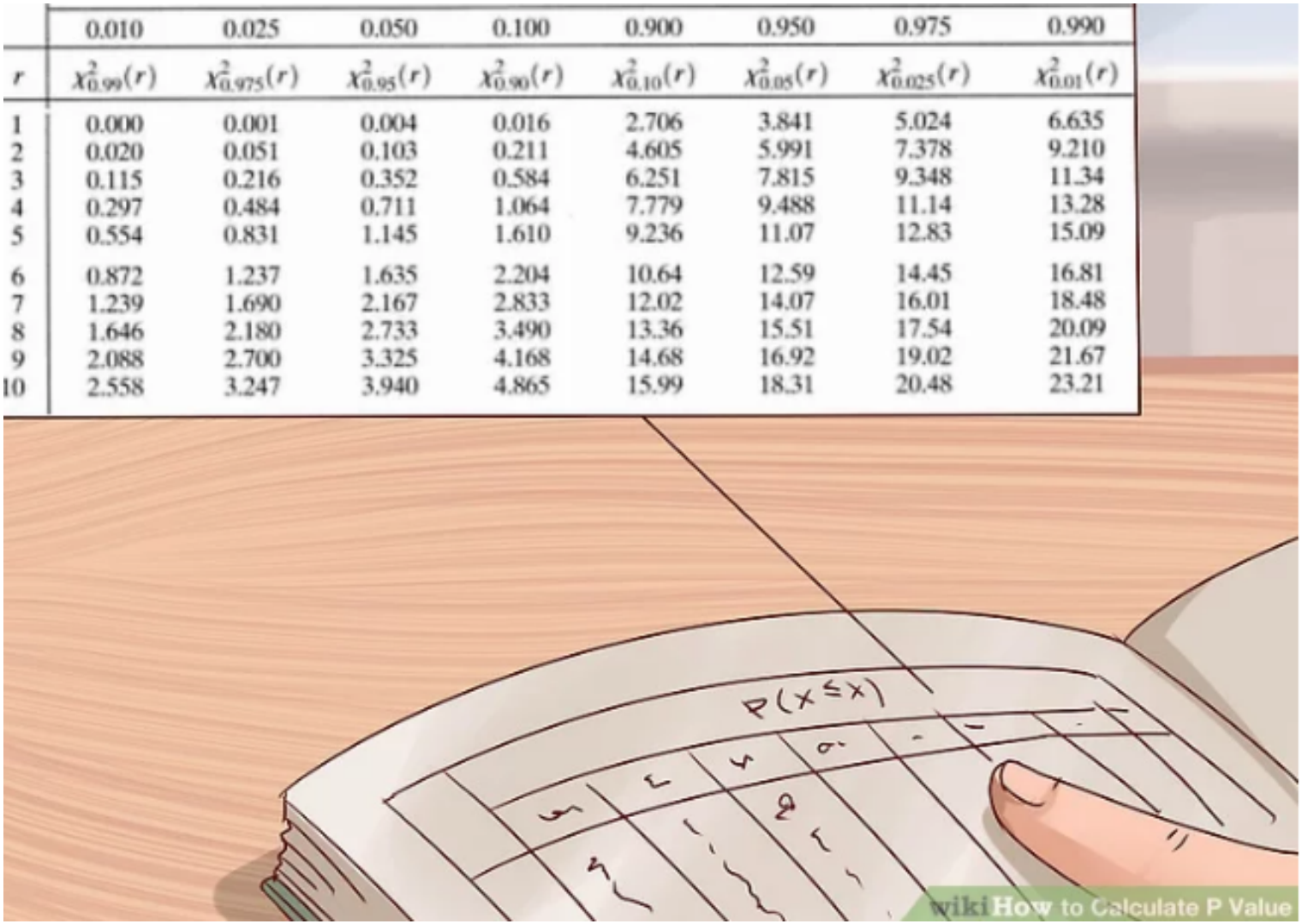


Choose a significance level

- Basically, the significance level is a measure of how certain we want to be about our results - low significance values correspond to a low probability that the experimental results happened by chance, and vice versa.
- By convention, scientists usually set the significance value for their experiments at 0.05, or 5 percent.
- This means that experimental results that meet this significance level have, at most, a 5% chance of being reproduced in a random sampling process

Chi square table

	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21



Use a chi square distribution table to approximate your p-value

- Use these tables by first finding your degrees of freedom, then reading that row across from the left to the right until you find the first value bigger than your chi square value.
- Look at the corresponding p value at the top of the column - your p value is between this value and the next-largest value (the one immediately to the left of it.)
- We'll go from left to right along this row until we find a value higher than 3 - our chi square value. The first one we encounter is 3.84. Looking to the top of this column, we see that the corresponding p value is 0.05.
- This means that our p value is between 0.05 and 0.1 (the next-biggest p value on the table).

Decide whether to reject or keep your null hypothesis

- Our p value is between 0.05 and 0.1 . It is not smaller than 0.05, so, unfortunately, we can't reject our null hypothesis.
- This means that we didn't reach the criterion we decided upon to be able to say that our town's police give tickets to red and blue cars at a rate that's significantly different than the national average.
- In other words, random sampling from the national data would produce a result 10 tickets off from the national average 5-10% of the time.
- Since we were looking for this percentage to be less than 5%, we can't say that we're sure our town's police are less biased towards red cars.

Hypothesis Testing

Statistical Hypothesis Testing

- t-test = compares 1/2 numerical groups
- ANOVA = compares $>$ 2 numerical groups
- Chi-square test = compares categorical variables

t-Test

t-Test

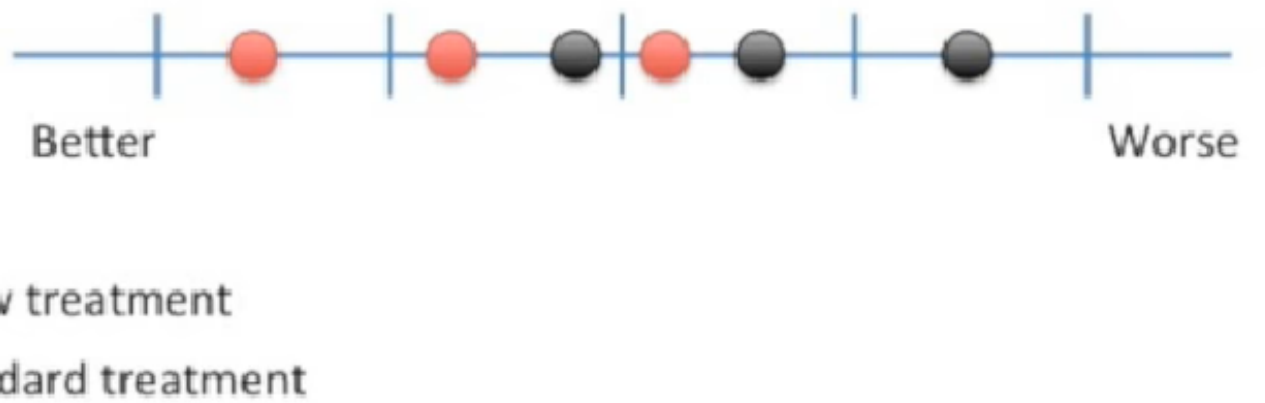
- There are two types of t-tests: paired and unpaired.
- Paired t-tests are used when you have 'before' and 'after' measurements from the same subjects. E.g body temperature before and after medicine.
- Unpaired: you are comparing height data for two groups.
- Subcategories of Unpaired t-tests:
  - Assumes that variations with both groups are same.
  - Does NOT Assume that variations with both groups are same.
- Another classification: 1-tail or 2-tails t-tests
- In case of comparing heights of two groups, the 2-tail t-test will evaluate of group A is Higher than B as well as Shorter than B, ie at both ends.
- It is used when you already know which direction has to be compared, higher or lower.

(Ref: StatQuest: Which t test to use? - Josh Starmer)

1 vs 2 tailed t-Tests

Example: a clinical trial is done on 6 patients to find effect of a new drug.

- Data shows that with new drug, results are much better (more red dots on left)
- The red dot in the middle is a bit problematic.
- After doing stats, you get p-value of 0.03 for 1 tail t test.
- 0.03 is smaller than threshold of 0.05 (CI). Meaning this new drug is significantly different. Its mean result is far away from other means found during existing treatments.
- 2-tailed gives p-value of 0.06. Not good.
- Which p-value to use?

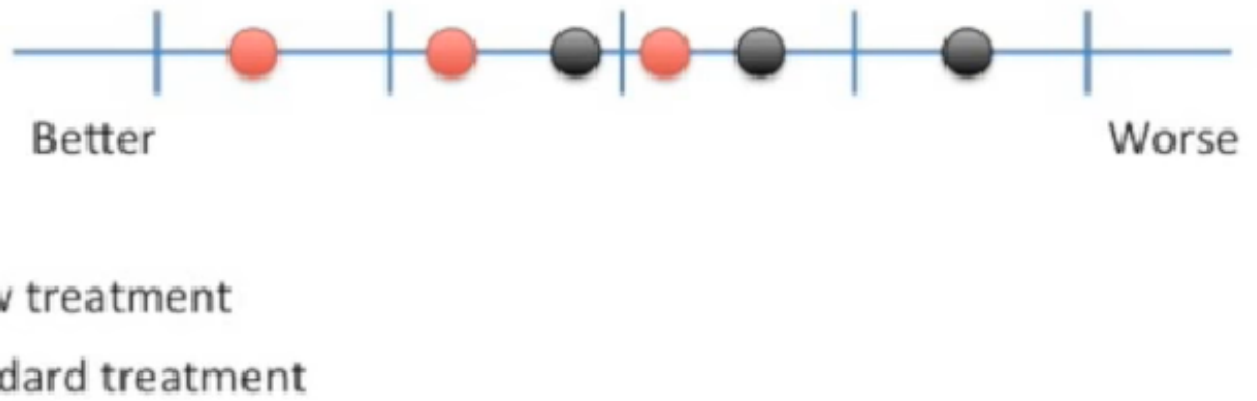


(Ref: StatQuest: One or Two Tailed P-Values - Josh Starmer)

1 vs 2 tailed t-Tests

Example: a clinical trial is done on 6 patients to find effect of a new drug.

- 1 tailed p value tests the hypothesis that your new drug is better than the existing drug. Whereas,
- 2 tailed p-value tests whether the new drug is better , worse or not significantly different.
- 1 tailed p value is smaller because it does not distinguish between "worse" or "not significantly different".
- AS we wish to know if the new drug was worse than the existing treatment, we should use 2 tailed p value test.



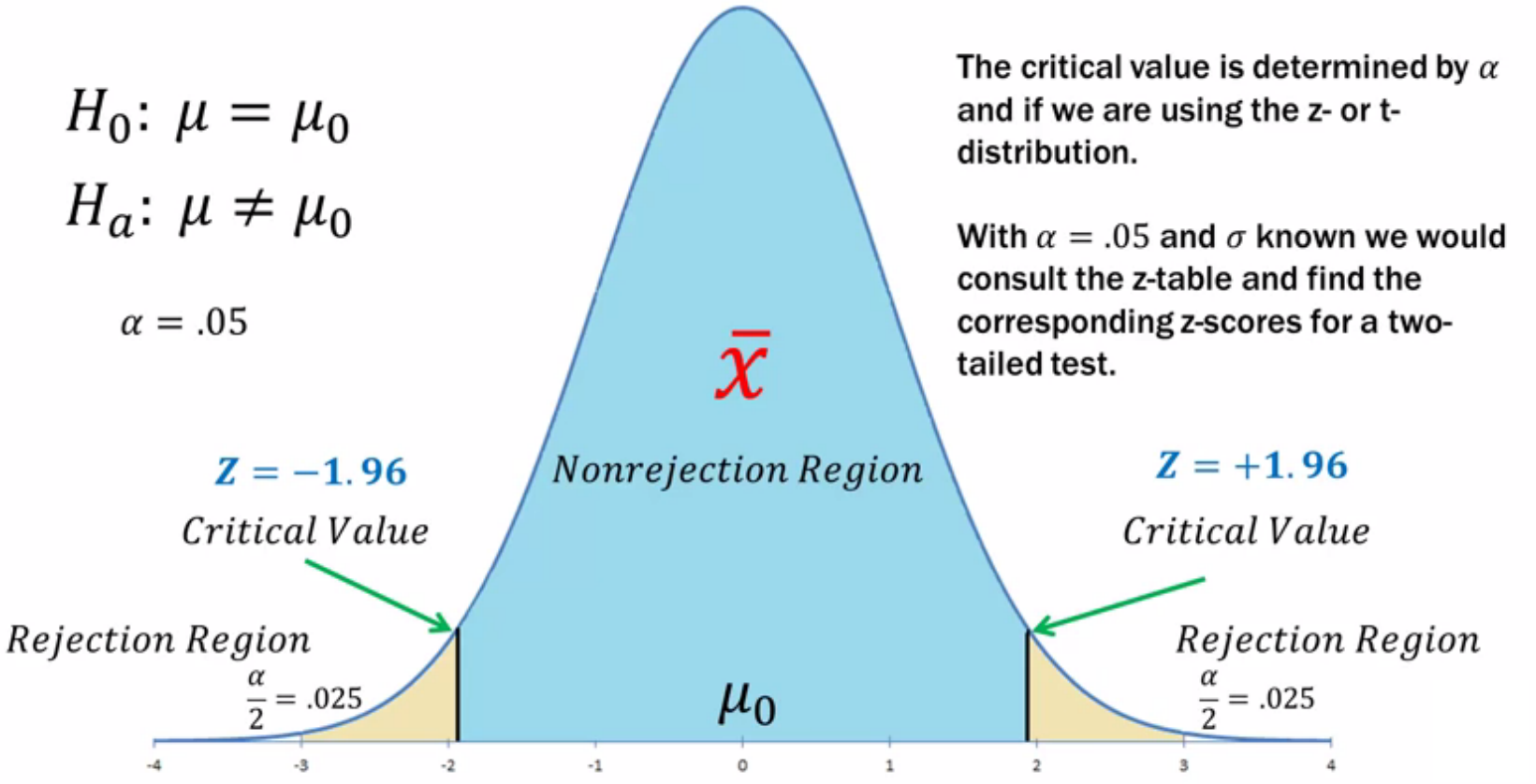
(Ref: StatQuest: One or Two Tailed P-Values - Josh Starmer)

t-tests

- One sample hypothesis testing
  - We are given population σ: use z distribution, z-test
  - We are Not given population σ, need to estimate it: t-distribution, t-test
- Two sample t-test

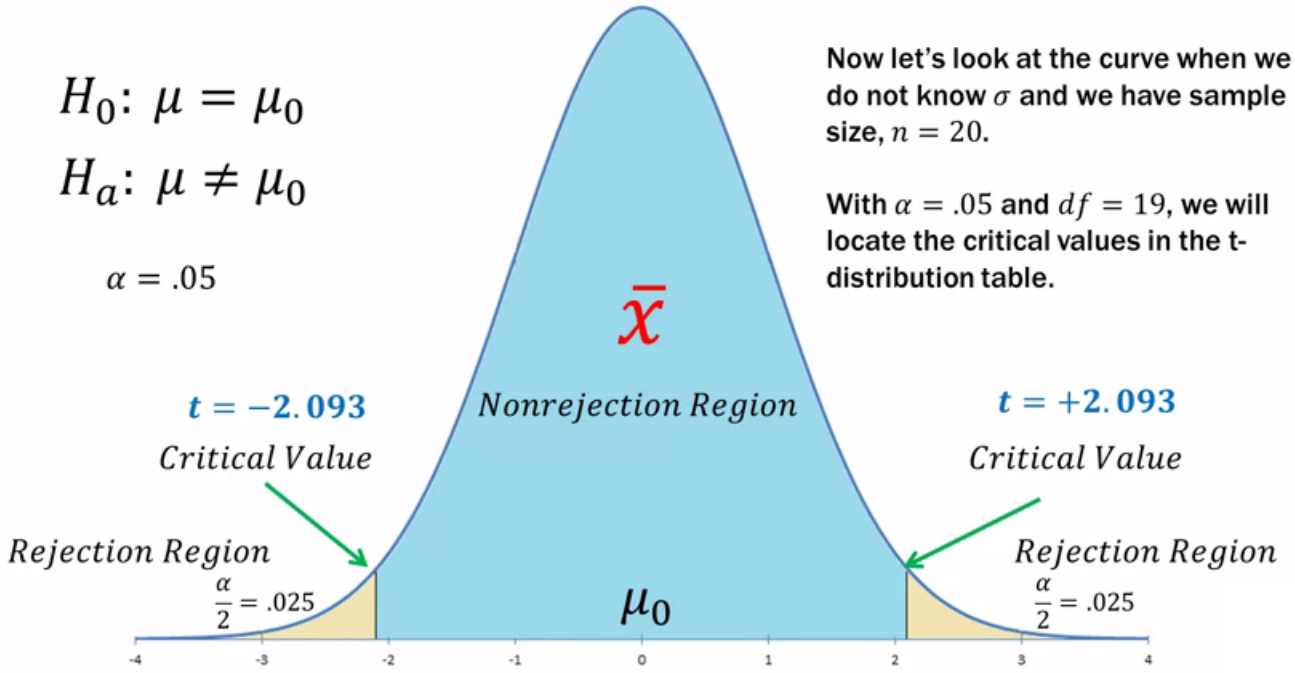
Z distribution

Same as Sampling distribution



t distribution

Every t distribution is unique and is specific to sample size. For n=20, look at t table with degrees of freedom 20-1=19



Range is a bit more than Z dist. Like pushing it down from top!!

One Sample t-test

- Objective: Test if hypothesized population mean matches with the actual population mean.
- μ_0 is hypothesized population mean
- μ is the true/actual population mean
- Method: Test using sample means and confidence interval

One Sample t-test

- When?: To test whether a given sample is coming from a population whose mean is specified value

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- $\bar{x}$: sample mean of n observations
- s: sample std deviation of n observations
- μ_0: hypothesized population mean
- $H_0 : \bar{x} = \mu_0$
- $H_1 : \bar{x} \neq \mu_0$

One Sample t-test

- t value computed is similar to z score.
- Shows how many std deviations you are away from mean.
- With the given t score, in the t table find the location on graph
- Looking at cut off boundaries, are you landing in the non rejection region or the rejection region?

Example: One Sample t-test

- New tyres launched are claim to have life of 40000 km.
- A retailer wants to test this claim.
- He has taken random sample of 8 tyres.
- He is testing life of the tyres under normal condition.

Example: One Sample t-test

- We don't know population std dev. So t test.
- $H_0; \mu_0 = 40000$
- $H_1; \mu_0 \neq 40000$
- Lets take $\alpha = 0.05$
- $df = n - 1 = 8 - 1 = 7$

Example: One Sample t-test

t Table

df	Critical Values				
1-tailed	0.1	0.05	0.025	0.01	0.005
2-tailed	0.2	0.1	0.05	0.02	0.01
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453

- Critical boundary-value = 2.3646
- If t false below -2.3646 or above 2.3646, then reject H_0.

Example: One Sample t-test

- His findings below:

Sr	Km
1	35000
2	38000
3	42000
4	41000
5	39000
6	41500
7	43000
8	38500

- Given goal mean $\mu_0 = 40000$
- $n = 8$
- Calculate sample mean $\bar{x}$, it comes to 39750
- Calculate s, it comes to 2618.615
- $df = n - 1 = 7$

Example: One Sample t-test

- $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{39750-40000}{2618.615/\sqrt{8}} = \frac{-250}{925.82} = -0.27$
- It is right of negative boundary and left of positive boundary,
- So, in valid region.
- So, cannot reject H_0.
- So, no difference.
- Sample mean is same as given mean.

Two Samples t-test

- When?: To test whether given two samples are coming from a population

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- $\bar{x_1}$: sample mean of n_1 observations
- $\bar{x_2}$: sample mean of n_2 observations
- s_p: Pooled std deviation of two samples
- μ_0: specified population mean
- $H_0 : \bar{x_1} = \bar{x_2}$
- $H_1 : \bar{x_1} \neq \bar{x_2}$

Example: Two Sample t-test

- Is there any appreciable difference of IQs of Males and Females?
- $H_0 : \bar{x_{male}} = \bar{x_{female}}$
- $H_1 : \bar{x_{male}} \neq \bar{x_{female}}$
- Lets take $\alpha = 0.05$
- $n_m = n_f = 18; n = 36$
- $\bar{x_{male}} = 98.9$
- $\bar{x_{female}} = 102.4$
- $s = 12$; same for both

Example: Two Sample t-test

- Finding critical values value for t at $1 - \alpha/2$ for two tail test and $df = n - 1 = 35 = 2.030$
- $t = \frac{\bar{x_{female}} - \bar{x_{male}}}{s_p / \sqrt{\frac{1}{n_f} + \frac{1}{n_m}}}$
- $= \frac{102.4 - 98.9}{12 / \sqrt{1/18 + 1/18}} = 0.097$
- Lies in acceptable region.
- Can not reject H_0
- No diff in IQs.

Two Samples t-test

What is t-test?

- Compares two averages (means) and tells you if they are different from each other.
- Tells you how significant the differences are
- It lets you know if those differences could have happened by chance.

Two Samples t-test

What is t-test?

- The t score is a ratio between the difference between two groups and the difference within the groups.
- A large t-score tells you that the groups are different.
- A small t-score tells you that the groups are similar.
- The bigger the t-value, the more likely it is that the results are repeatable.

Two Samples t-test

- Every t-value has a p-value to go with it. A p-value is the probability that the results from your sample data occurred by chance.
- a p-value of .01 means there is only a 1% probability that the results from an experiment happened by chance.

Two Samples t-test

Types of t-tests?

- An Independent Samples t-test compares the means for two groups.
- A Paired sample t-test compares means from the same group at different times (say, one year apart).
- A One sample t-test tests the mean of a single group against a known mean.

Two Samples t-test

Example: to test whether the height of men in the population is different from height of women in general.
Steps:

- Determine a null and alternate hypothesis.
  - Null-Hypothesis: no statistically significant difference, height of men and women are the same
  - Alternate-Hypothesis: statistically significant difference, height of men and women are different
- Collect sample data. Twos sets: men and women. Generally sizes should be same, but can be diferent, n_x and n_y

Two Samples t-test

Steps (cont.):

- Determine a confidence interval and degrees of freedom
- α of 0.005 is 95% confidence that the conclusion of this test will be valid
- The degree of freedom $df = n_x + n_y - 2$
- Calculate the t-statistic

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

M = mean
n = number of scores per group

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

x = individual scores
M = mean
n= number of scores in group

Two Samples t-test

Steps (cont.):

- Calculate the critical t-value from the t distribution

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

- or call ready Python function

Two Samples t-test

Generate data

```
#Sample Size
N = 10
#Gaussian distributed data with mean = 2 and var = 1
a = np.random.randn(N) + 2
#Gaussian distributed data with with mean = 0 and var = 1
b = np.random.randn(N)
```

Two Samples t-test

```
#Calculate the variance to get the standard deviation
#For unbiased max likelihood estimate we have to divide the var by N-1, and
     therefore the parameter ddof = 1

var_a = a.var(ddof=1)
var_b = b.var(ddof=1)

#std deviation
s = np.sqrt((var_a + var_b)/2)
```

Two Samples t-test

```
## Calculate the t-statistics
t = (a.mean() - b.mean())/(s*np.sqrt(2/N))
## Compare with the critical t-value
#Degrees of freedom
df = 2*N - 2
#p-value after comparison with the t
p = 1 - stats.t.cdf(t,df=df)
#Note that we multiply the p value by 2 because its a twp tail t-test

## Cross Checking with the internal scipy function
t2, p2 = stats.ttest_ind(a,b)
```

Paired t-test

- When?: To check, effectiveness of a new treatment, a new method employed
- observations on every unit are made before and after applying the treatment or method.
- Hence if treatment or method is effective, there will be significant difference in observations before applying it and after applying it.

$$t = \frac{\bar{x_D} - \mu_0}{\sigma_D/\sqrt{n}}$$

- $\bar{x_D}$: mean difference of n paired obsrevations
- σ_D: std deviation of difference of n paired obsrevations obsrevations
- μ_0: mean difference of n paired observations under H_0
- H_0 : Mean of samples before and after treatment is same.
- H_1 : Mean of samples before and after treatment is not same.

Chi-Square test

- When?: To test whether any two categorical variables are associated with each other or they are independent of each other,

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- O_i : Observed frequency of ith variable
- E_i : Expected frequency of ith variable
- H_0 : Two variable are independent of each other.
- H_1 : Two variable are not independent of each other.

F test

- When?: to know whether these two sample have same population variance

$$F = \frac{s_1^2}{s_2^2}$$

- S_1^2 Sample variance of first sample of size n_1
- S_2^2 Sample variance of second sample of size n_2
- H_0 : Variance ratio is one
- H_1 : Variance ratio not equal to one

Anova

Analysis of Variance (ANOVA)

The difference between ANOVA and the t tests is that ANOVA can be used in situations where there are two or more means being compared, whereas the t tests are limited to situations where only two means are involved.

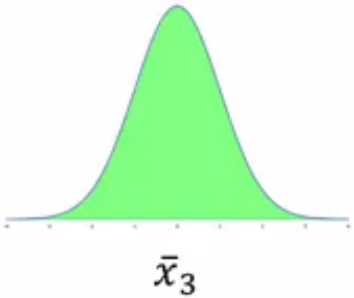
Analysis of Variance (ANOVA)

Suppose we want to compare THREE sample means to see if a difference exists somewhere among them.

Is one mean so far away from the other two that it is likely not from the same population?

What we are asking is:

Do all three of these means come from a common population?



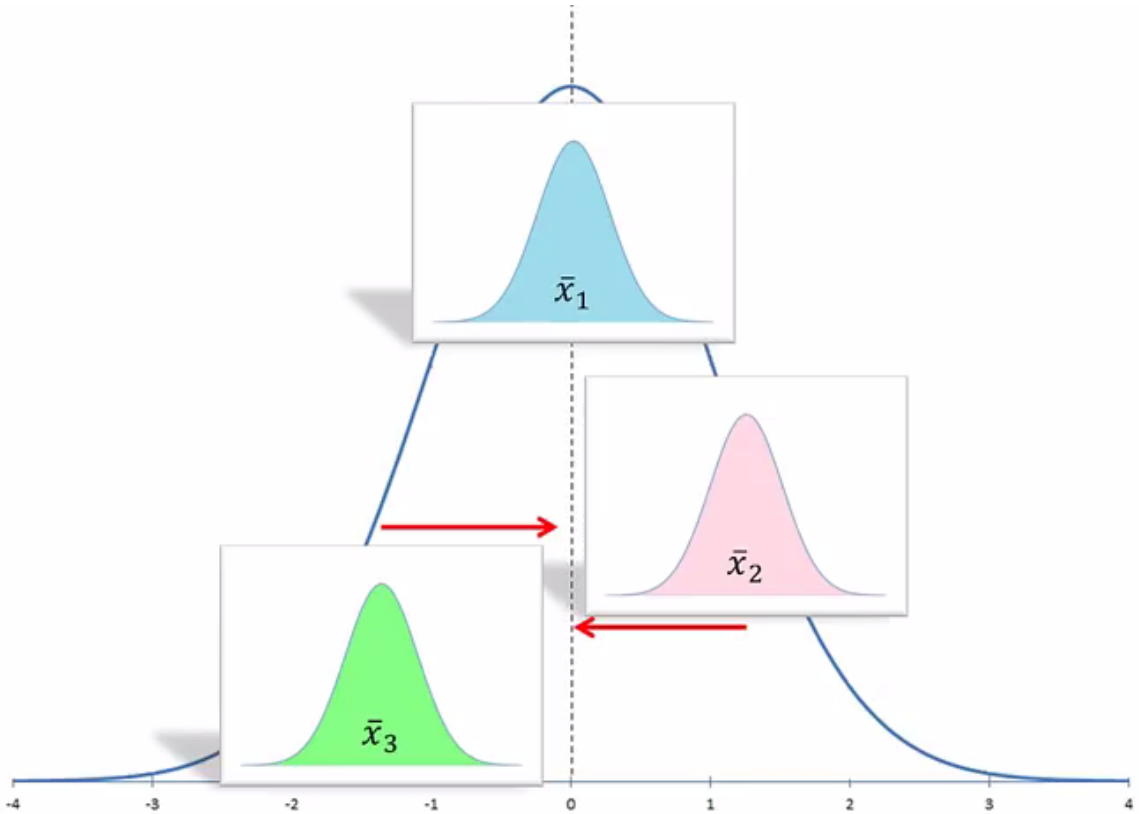
Or are all three so far apart that they ALL likely come from unique populations?

Finding DISTANCE between means

(Reference: Statistics 101 - Brandon Foltz)

Analysis of Variance (ANOVA)

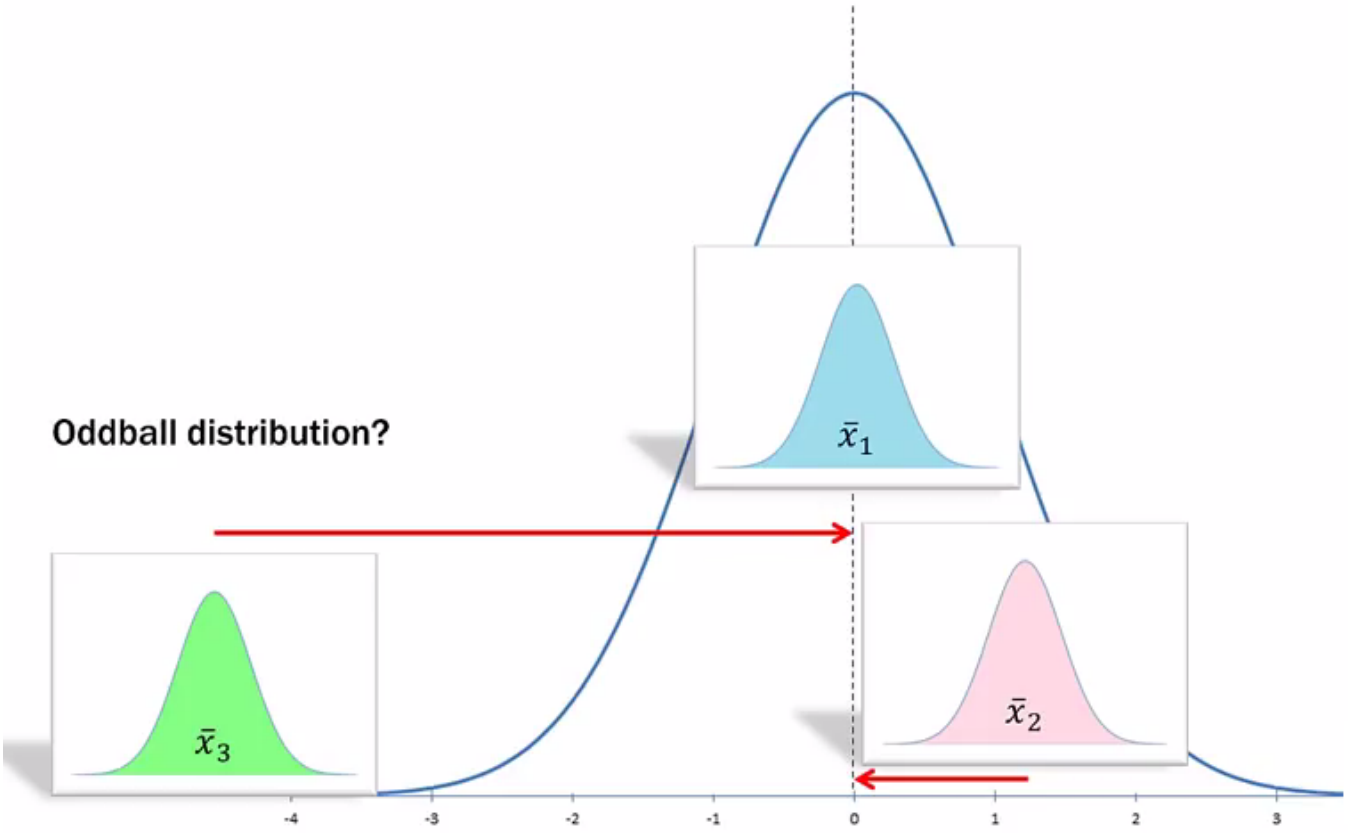
Combine all data points from all 3 samples. Draw its distribution. Then compare of each with the combined one.



$\bar{x}_1$ bang on combined mean, rest are a bit away.

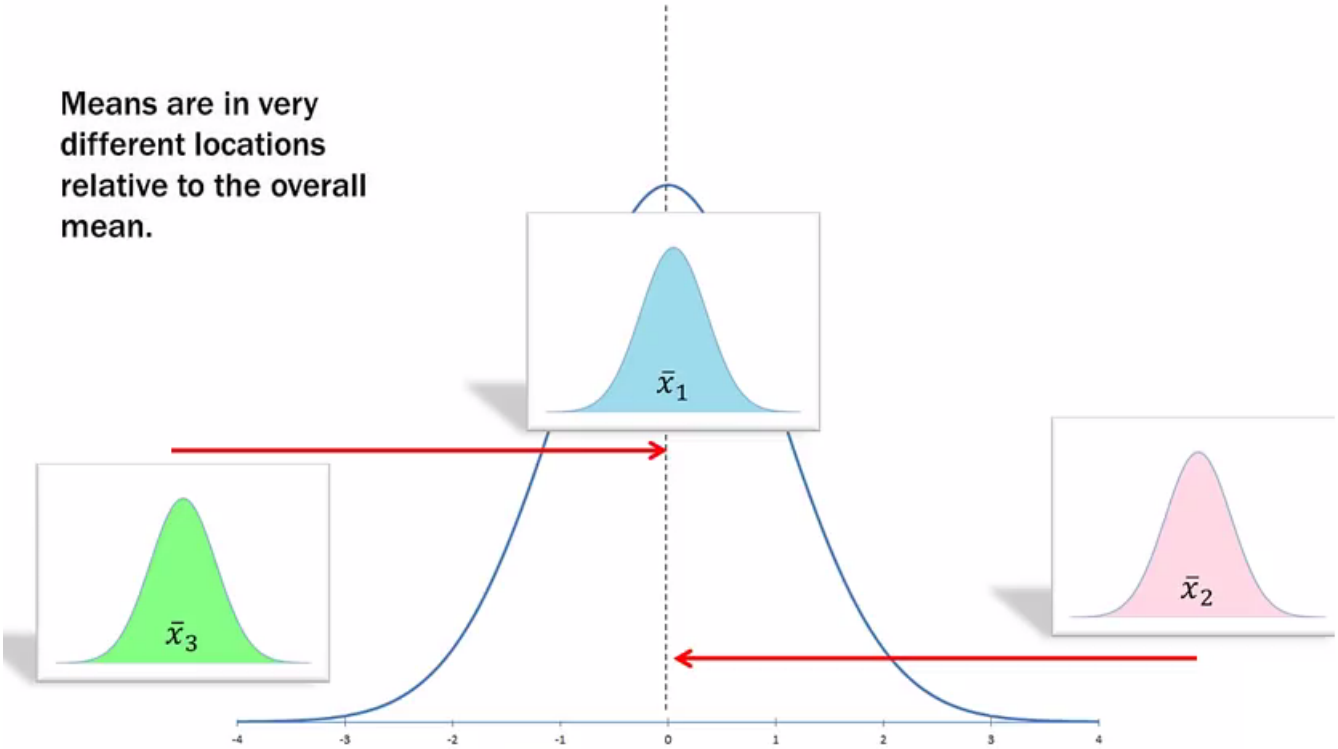
Analysis of Variance (ANOVA)

New example: where $\bar{x}_3$ is far away. So not part of combined population. Outlier.

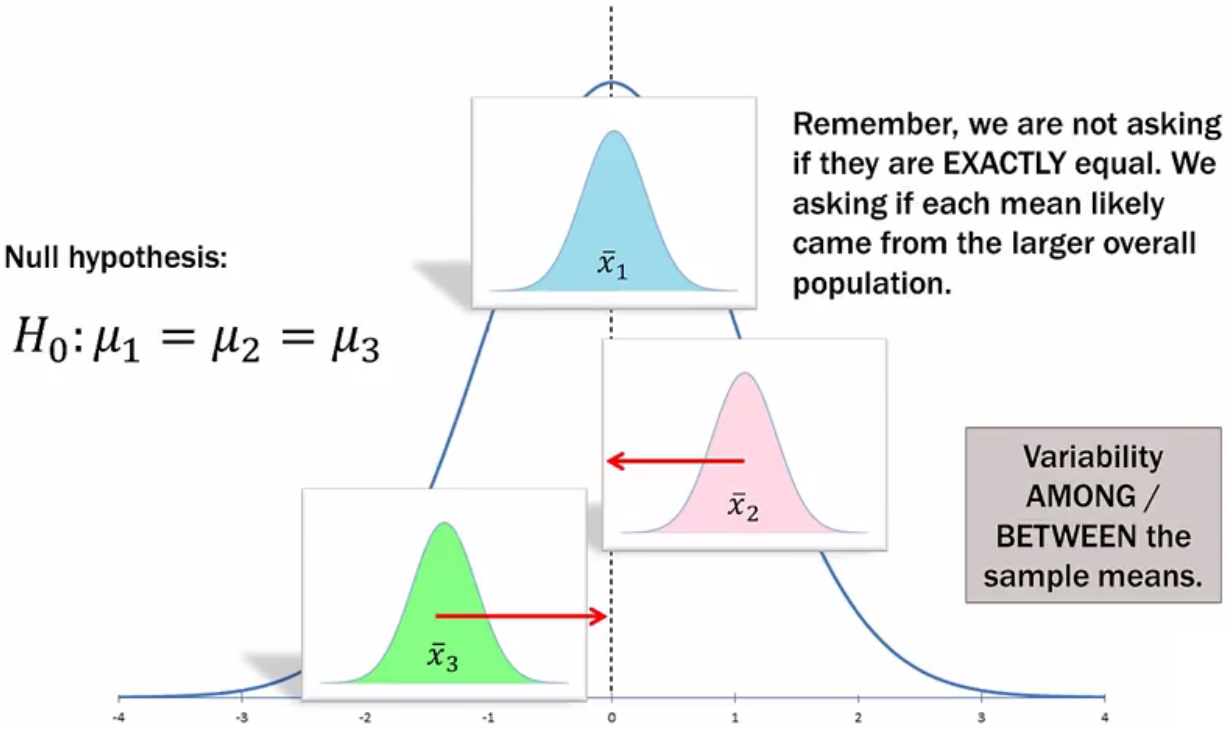


Analysis of Variance (ANOVA)

New example: where both $\bar{x}_2$ and $\bar{x}_3$ are far away. All three are their OWN populations.



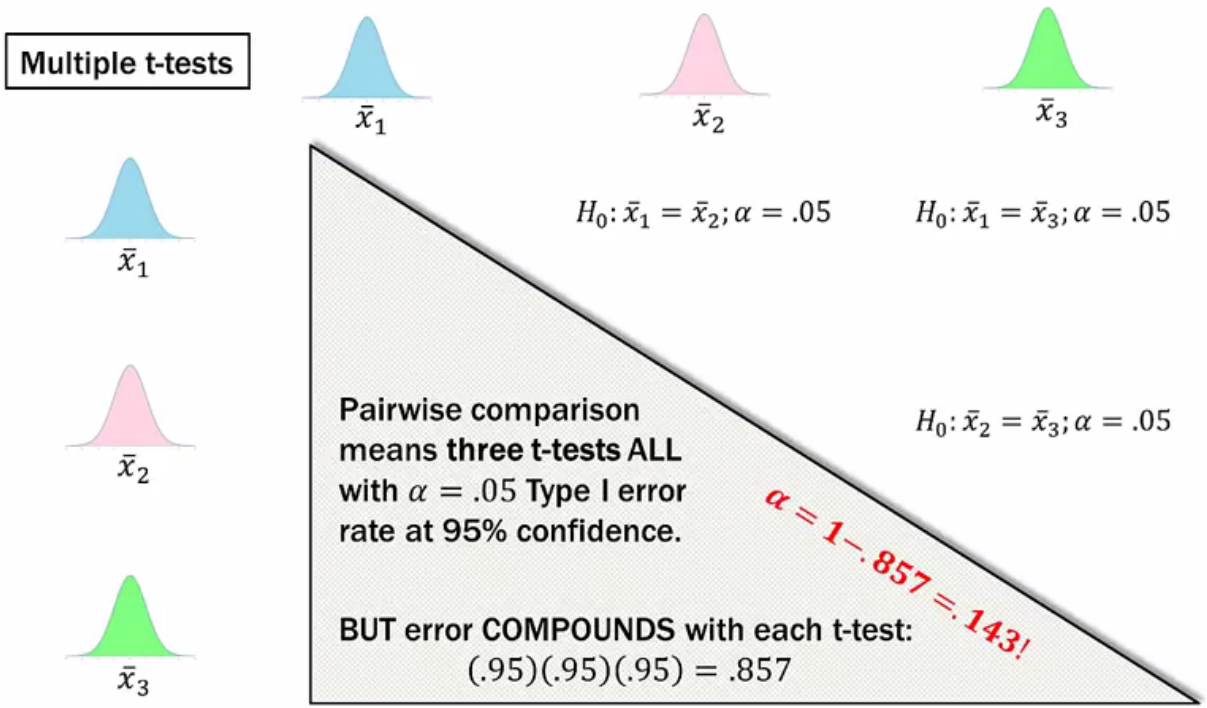
Analysis of Variance (ANOVA)



Variability is between means.

Analysis of Variance (ANOVA)

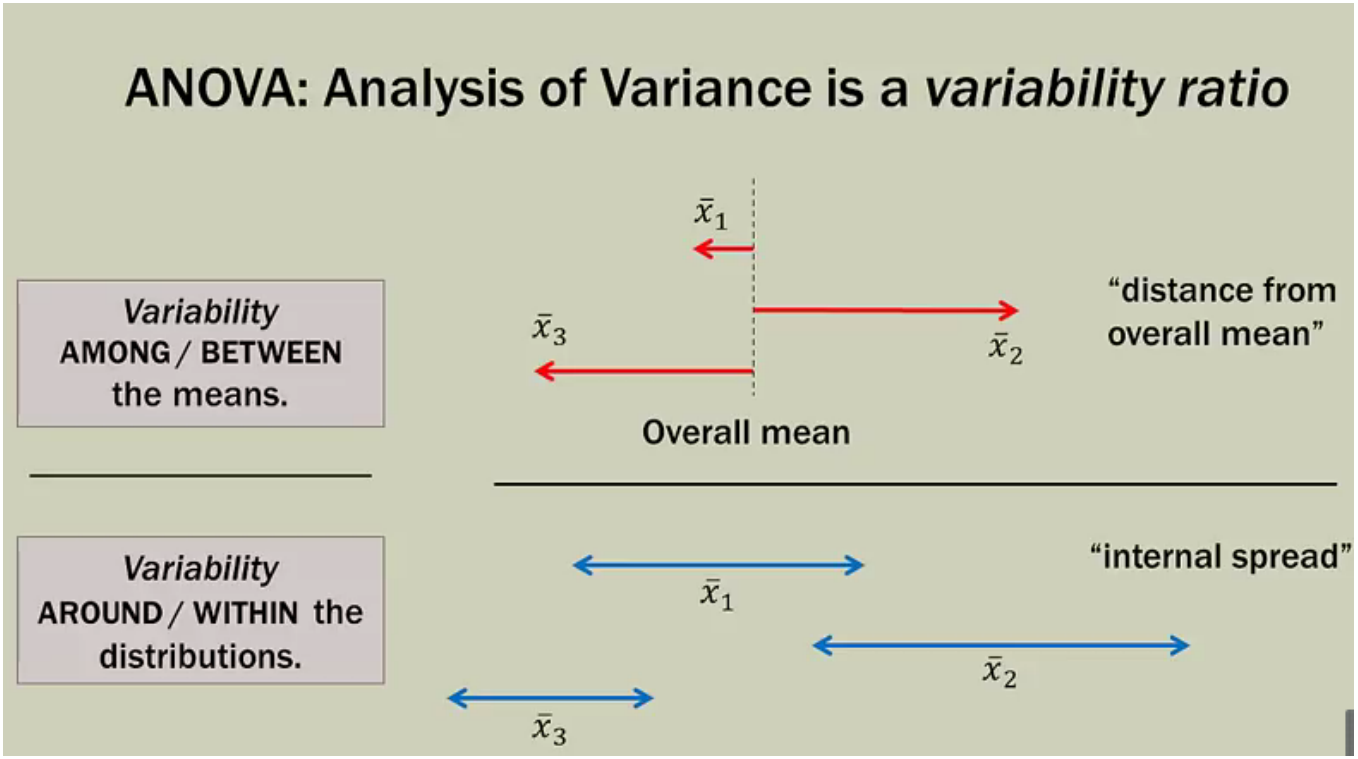
Pair wise t tests



Combined error rate went from 5% to 14.3%. So we do not conduct multiple t tests.

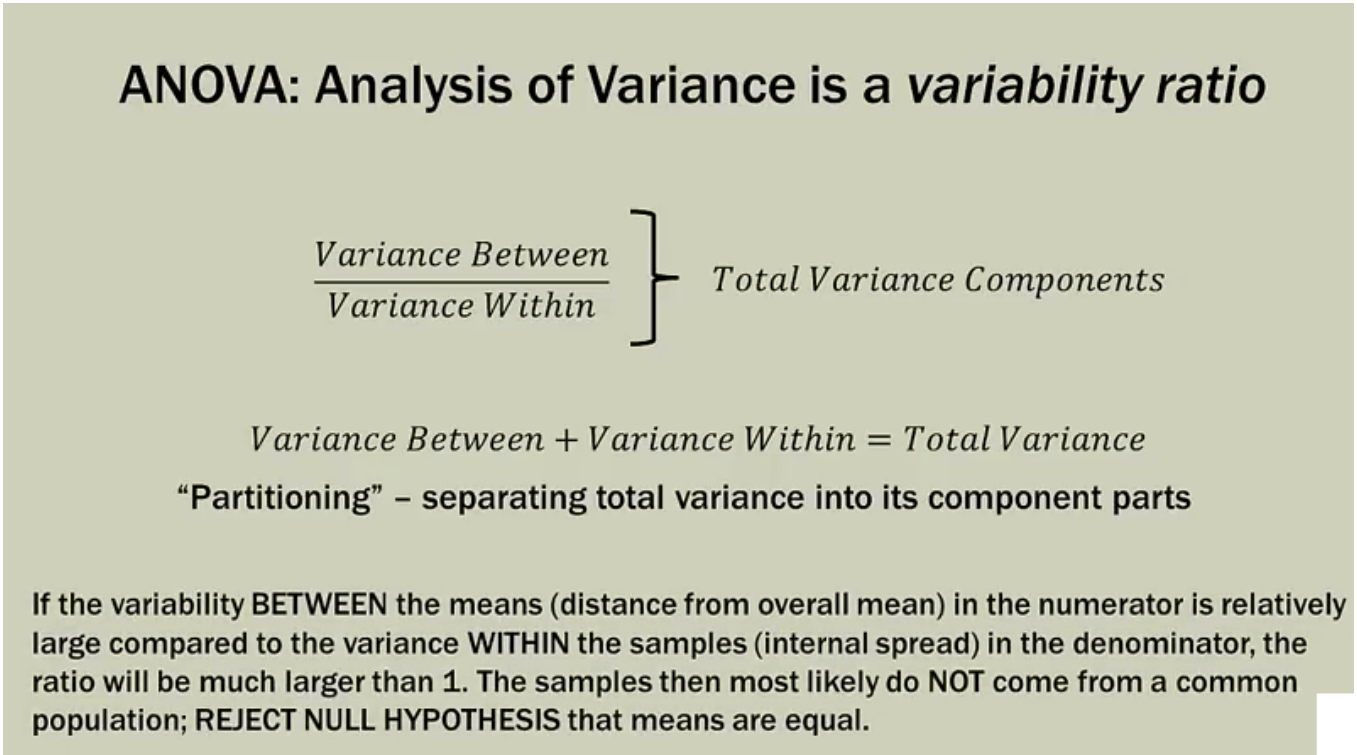
Analysis of Variance (ANOVA)

Each distribution has its own internal variance / variability. Anova is a variability ratio.



Analysis of Variance (ANOVA)

Total variance has two components.



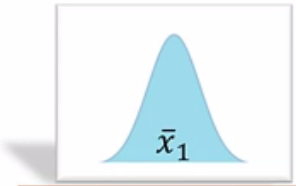
Analysis of Variance (ANOVA)

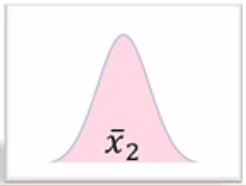
ANOVA: Analysis of Variance is a *variability ratio*

$$\frac{Variance\ Between}{Variance\ Within}$$

$\frac{LARGE}{small} = Reject\ H_0$	At least one mean is an outlier and each distribution is narrow; distinct from each other.
$\frac{similar}{similar} = Fail\ to\ Reject\ H_0$	Means are fairly close to overall mean and/or distributions overlap a bit; hard to distinguish.
$\frac{small}{LARGE} = Fail\ to\ Reject\ H_0$	The means are very close to overall mean and/or distributions "melt" together.

Analysis of Variance (ANOVA)





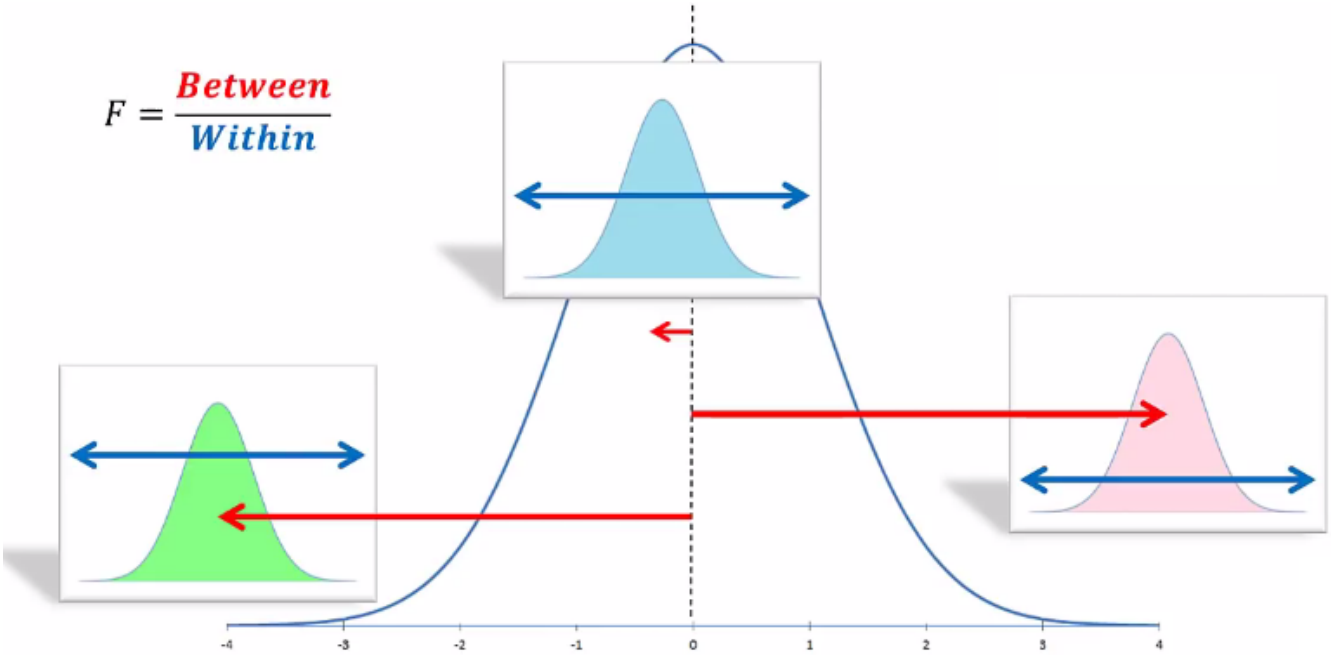
Year 1 Scores	Year 2 Scores	Year 3 Scores
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

Overall Mean:

The mean of all 21 scores taken together.

$$\bar{\bar{x}} = 74.52$$

Analysis of Variance (ANOVA)



Thanks . . .

- Feel free to follow me at:
  - Github (github.com/yogeshhk) for open-sourced Data Science training material, etc.
  - Kaggle (www.kaggle.com/yogeshkulkarni) for Data Science datasets and notebooks.
  - Medium (yogeshharibhaukulkarni.medium.com) and also my Publications:
    - Desi Stack https://medium.com/desi-stack
    - TL;DR,W,L https://medium.com/tl-dr-w-l
- Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- Email: yogeshkulkarni at yahoo dot com