Statistics Basics for Data Scientist

Saturday, August 25, 2018

2:07 PM

**(1) Statistical Analysis**

*What :* Statistics is a branch of mathematics. Which deals with data collection, analysis, interpretation and presentation.

*Statistical Analysis* involves collecting data, analyze the nature of the data, explore the relation of the data, create model

to summarize the the understanding, prove the validity of model and help with predictive analysis.

*Why :* Statistics is essential to make a right decision or statement(after analyzing) .

Example 1: If you want to buy a plot, you visit multiple places, compare and make a decision.

Example 2: Is this candidate good or not? => collect the candidate data, analyses the skillset and competitiveness, and

make the statement.

*Data Scientist is a problem solver. So he needs to collect the right data and apply statistical analysis.*

**(2) Traditional vs Modern Statistics**

*Statistics in Data Science:* Used for the purpose of Analysis and purpose of Prediction.

- Only after learning/analyzing the data, we can make the statement.

Example : You have been to a hotel. Look at the good and bad (analyze) then take decision.

*Purpose of Analysis :* is the process of applying statistical practices to organize, represent, describe, evaluate

and interpret the data.

Example : Go to a veg market. Look at the available vegetables, compare/ analyze and select the right vegetables.

Analysis E.g.: measures of central tendency, MO of dispersion, random variables, mean, mode [ all math concepts]

*Purpose of Prediction:* Extract the information from the data and use it for predicting the trend and behavior.

E.g. entropy, info game, gradient decent, supervised, unsupervised learning, classification, regression, etc.

**(3) Different Data Types**

*(i) Discrete Data* : Data **can't** be plotted on a continuous scale. It usually in any format.

Example 1: Name, Location name (these are fixed and can't be plotted.)

Example 2 : Date of birth

*(ii) Continuous Data* : Data **can** be plotted on a continuous scale. Data only in numeric format.

Example : Salary, Price, Credit score, Age, millage ( continuous and can be plotted)

**(4) Two different sources of Data**

*(i) Observational Data*

- Data Generation process **can't be controlled** by a Data Analyst.

- When the data is generated without any control (huge size), its tough and requires too much time to understand the data.

So we can't understand and predict the result. i.e. we can't apply any ML techniques.

- So change the way to analyze the data and use computer.

- we may need to check the variance of the data.

(e.g.) Facebook, amazon purchase.

*(ii) Experimental Data*

- Data Generation process **can be controlled** by a Data Analyst. Example : Use excel sheet to generate data on your own.

- No need to work on any ML or other statistics techniques.

1. **Different Measurement Scales**

When we share or assign the project to someone, we need to define the entire dataset in different measurement scale.

*(i) Nominal Datatype* - Name or Label of a product.

Example 1: (a) Why there is a hype for Data Scientist. (b) color of the iPhone.

Always speak about the label of the product without any order.

Example 2: Customer name, State name, Product name

*(ii) Ordinal Data* - All about the reviews of the product. In statistics, Rank/Rate of the product in the particular order.

Example 1: 0 - Not satisfied, 1 - Satisfied , 2 - Good, 3 - Excellent

Example 2: Customer Complaint Severity

*(iii) Interval Scale* - Helps to understand the difference between one product and another.

Example 1: I need to buy a car for SGD 50 K.

Example 2: Based on the pricing slab, people change their purchase option.

Example 3: Profit difference in specific intervals.

*(iv) Ratio Scale* : Discount plays a role in product sale and ratio makes difference.

Example: ratio of salary increase, ratio of interest increase

Using ratio scale, we can measure the ratio of profile/loss, customer satisfaction, etc.

**(6-7) Divisions in Statistics**

*(1) Descriptive Statistics*

Process of collecting the data in order to take right decision is called Descriptive Statistics.

- e.g. To Put our kid in a school, get data of different school and compare the features and eventually take the right decision.

*(2) Inferential Statistics*

- deal with extract valuable information (insights) from the data to make the proper statement/decision.

-Ex 1: After the collection of data, understand the details or get information out of it.

Ex 2: Average study hours, average assignments, measure of central tendency, dispersions, sample statistics, random sampling

**(8) Exploratory Data Analysis (EDA)**

- Without understanding the collected data properties and features, we can't make the statement.

- After collecting the data, we need to understand, test and ensure the collected data suitable for creating a model.

Ex 1: checking for features and properties, looking outliers, sufficient or not, typo errors, null values.

- EDA helps to identify the nature, dispersive, variance, typo errors, etc. in the dataset.

-EDA helps to collect the right data and give right solution (for the right problem statement)

**(9) Branches of EDA**

- Aim of EDA is to collect the right amount of data, test it and ensure that its suitable for our analysis.

It helps for the successful delivery of right dataset.

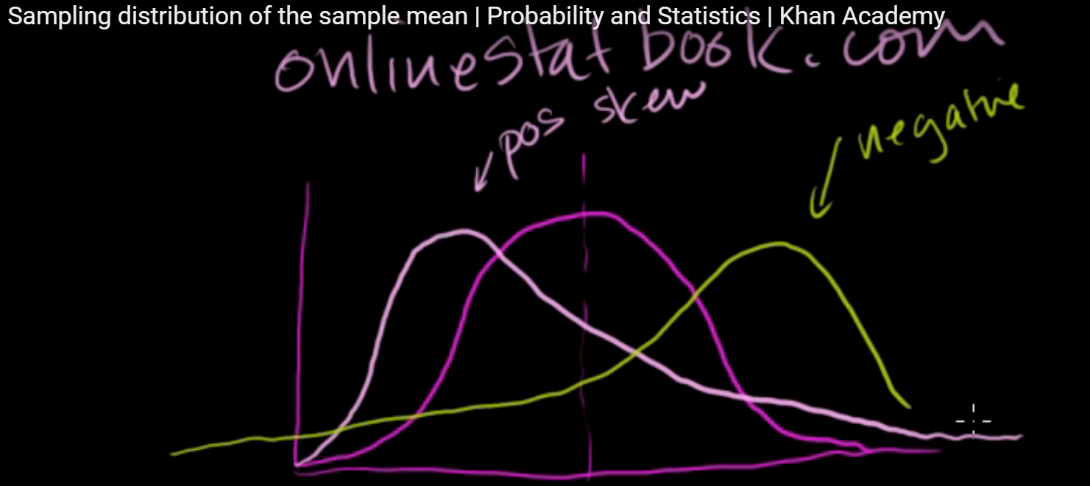
(1) Measures of Central Tendency

(2) Measures of Dispersion

(3) Skewness

(4) kurtosis

(5) Using graphical(?) techniques



Machine generated alternative text:
And maybe in future videos we'll 
explore that in more detail, 

Ex1 : Make a statement on the listener of audio tutorial

**Measure of central tendency**

- First collect all the info about audience, what is your passion, what is knowledge level, etc.

**Understand the dispersion (failure)**

Ex 1: How frequently you fail to learn the concepts. i.e. understand the deviation of the particular person.

Ex 2: How often the vehicle will breakdown or need to take to workshop **[Measures of Dispersion]**

**Skewness or kurtosis?**

- is the vehicle suitable only for ladies, kids, highways (i.e. not normally distributed). this is called [**Skewness or** kurtosis**?** ]

**Outlier or anomalies**

- Everything is good but design is bad (outlier). All good but price is high (outlier)

something is stopping us to buy and we call this as **outlier or anomalies**.

**(10) Measures of Central Tendency**

- Mean, Medium and Mode - used to understand the nature of the particular dataset.

Ex 1: is a student topper, average, or poor (requires the average marks). Based on the number, entire result changes.

(i) **Mean** : is nothing but average.

(ii) **Medium** : when the data is uneven or contains some higher/lower values. We need to go with medium.

Arrange data into ascending/descending order and pick the middle.

(iii) **Mode** : highest repetition.

e.g. 1,2,2,3,4,5,5,5 ==> mode of data set is 5

Ex 1: Go to a showroom to buy a vehicle. Mileage 20 KM/L but in reality it gives only 10 or 12 KM.

But showroom guy says that it’s the highway mileage this is called **outlier**.

i.e. whenever we understand the particular dataset, remove the outlier and make the statement. Always mean value get

affected due to the outliers.

**What to do ?** If we have many outlier, arrange the data in a sequential order and pick the central value (median) .

In the same dataset, if there is a repetition go for the mode.

Ex : Most of the time people are speaking about iPhone(how many times). So I will buy iPhone.

**(11) Measures of Dispersion**

**if the data is completely different, even if you have the good ML algorithm, the accuracy will be low.**

Variance and Standard Deviation are two measures of dispersion within a data set.

(i)Variance

- the units of each and every measure in the square.

- if an employee comes to office 9 am, 11 am, 5 am, 6 am. Here too much variance.

here we need to study the variance.

* the variance measures the average degree to which each point differs from the mean. The greater the variance, the larger the overall data range.
* average of the squared differences from the Mean

(ii) Standard Deviation

- Square root of variance.

- any value which is deviating from the mean is called standard deviation.

e.g. every day an employee comes to office at 9 but one day he comes at 10. so deviation is 1 hour.

i.e. 1 hour deviated from the average.

Machine generated alternative text:
Population Variance 
Population Standard Deviation 

Machine generated alternative text:
Sample Variance 
Sample Standard Deviation 
s 

* *Because of squaring, the variance is no longer in the same unit of measurement as the original data.*

*So taking the square root covert the value back to original unit.*

* *Why n-1 ? Sample is the subset of population and we need to make sure the calculated variance is unbiased.*

*Since the sample may not have all the data points, it's safe to divide by n-1 to give a slightly large(unbiased) standard*

*deviation [2].*

(iii) Range

**(12-13) Normality, Skewness and Kurtosis**

Do not make the statement after looking the single side of the story. Look at all the direction and make the statement.

Otherwise, the solution will be given for only the single sider not multi sider. You can't spare an error while collecting data.

*(i) Normality*

- Normality or what is normally distributed data ?

Mean == Medium == Mode

i.e. there is no outliers, no typo errors and data is completely equally distributed.

Ex : person comes with very small pay scale, other small group may take big pay scale.

Collect end to end data and perform analysis.

(ii) Skewness

* You may collect skewed data. Ex : You may collect data only from your location and do the analysis.

So u will get the features of property from your location only.

(iii) Kurtosis

- you may collect kurtosis data.

**(14) Story telling of EDA**

- Assume in a manufacturing company, you need to identify the number of defects caused by the machinery A(prob statement.)

- Will u take one month or two month or one year data, Or which technique will apply

- Collected data should be normally distributed, there shouldn't be any skewness or Kurtosis.

e.g. If not (e.g. only 6 months data) , defects rate is too high and this will make a wrong statement.

Instead, if the machinery is 4 years old, you need to collect all the data, analyze and make the statement.

- the way you collect data, based on that *measures of central tendency changes*.

- e.g. you got cheated in Flipkart 3-5 times, will u buy again? No. next time, you will enquire many times.

In order to buy an apartment, need input from multiple people. you need to talk to many people and understand the

failures.

- For this problem, this is the correct and sufficient data.

- Check for mean, mode, medium , typo error , etc.

- Get more proper understanding about the data.

**(15) Data Science is a Random Variable**

- Random Variable is nothing but changes w.r.t time.

e.g. today I got sale of 10 K next day I am not sure I will get 10 K (it changes w.r.t time )

this month my expenditure 20K next month I am not sure (it may go up or down )

- Changes with respect to time must be solved in Data Science.

- Accuracy of the model also a random variables (it changes with respect to time. First 90%, when new data comes 70%, etc )

- Today you use decision tree and other day use another alg (random variable)

if there is randomness in the statement, need to deal with probability distribution.

Data science is all about random variables (and accuracy changes over the period of time)

**(16) Divisions in Random Variables**

*(i) Discrete Random Variables*

- We **can't split** the data (e.g. date of birth, designation).

*(ii) Continuous Random Variables*

- We **can split** the data . Data can be measured (e.g. experience, salary, expenses - it changes w.r.t time)

**(17) Is Randomness Root of Probability Distribution**

in order to understand the randomness of the data, we go with probability distribution . We tried to understand the distribution of the datapoint. e.g. Tossing a coin (on and average 20 mins - 16 times it gives tail and 4 times head). From this, there are more changes to happen tail.

E.g. There is a job profile. There are too much randomness and unable to make statement. 10 people placed java, 5 people placed

in .Net and 50 people placed in ML. So just pick ML as it has more success count.

- So randomness of the data is the root of probability.

e.g. today is cloudy weather and what would be the weather tomorrow. (random)

**(18) Division in probability Distribution**

(i) Discrete probability Distribution

- in order to understand the distribution of the data related to job. What is the probability to get a job ==> known as discrete probability distribution.

- in histogram, discrete variable on the X axis and probability value (ranging 0-1) on the Y axis.

(ii) Continuous probability Distribution

what is the probability to get this much amount of sale, profit, risk , call it as continuous probability distribution.

- in histogram, continuous variable will stay on the X axis and probability value (ranging 0-1) stay in Y axis.

Based on the data type, nature of data , it could be either one.

Difference between probability distribution and normal distribution: In probability Distribution, X axis similar but Y axis changes between probability and normality. When you are looking at the normality, you will get the frequencies of the data point but when you look at the probability, you will get the probability values. That's the only difference between discrete/continuous probability probability distribution versus normal distribution.

**(19-20) Sample (and its properties) and Population**

- Ex. Planning to buy a car. You are OK with the price. First you will give test drive then check the features.

If you want to buy iPhone10, you need to compare the features and you may need same iPhone10 samples as well.

if you need to learn the Car, you should learn about Car not Bike.

- one of the main property between sample and population is => sample is small in size but population is big in size.

This is the only difference (size) and rest everything is same.

- Sample has equal properties and characteristics of population but inline of size they are small.

If you need to check the rice is cooked, you don't need to eat the entire rice instead some sample. Sample has same properties

like population. If the size differ no issues but properties same.

- In ML or statistics, trying same example with different Dataset. Different type of data is possible.

* Sample must be equal to wrt characteristics/properties (not in size).
* e.g. Data Scientist PayScale is a problem -> you need to collect samples from different location, private/public sector, different age, etc.

**(21-22) Various Errors of Sample and Random Sample**

- Bias: For doing analysis on sample, If you collect the sample from the location where you stay (e.g. only from Bangalore),

Collected data is not normally distributed and it will have details about only Bangalore employees. This analysis is not good.

Another one is dependency. i.e. shouldn't collect the data from same/similar people.

Further, don't select the data which is in some order (ascending or descending). Algorithm should pick all the data randomly and random sampling is the solution. Make sure there is no dependency between the data points you collected (e.g. one friend data and another friend data - both are independent)

Random sampling is the remedy for the normal sampling (randomly distributed data not sequentially or normally distributed).

**(23) Prescriptive vs Predictive vs Statistical Analysis**

*Statistical Inference/Analysis* - In order to make a statement on the population, take a sample , understand the sample and based on the understanding, make the statement on the population.

Ex 2. Go to mobile shop. Select the phones and compare the features. Understand the features and increase the confidence level and make a statement. In statistical inference, not learning but analyzing and making statement.

**From the sample, understand the sample and make a statement about the population is known as Statistical Inference.**

*Predictive Analytics:* Taking the sample and learning all the properties present in the sample and making the statement.

Ex 1: listening the audio. It's not for just analyzing but learn the content of it. Based on the learning from the audio book,

we can make a statement.

From the sample, learn the entire features, based on the learning of the sample make a statement is known as Predictive Analysis.

*Prescriptive analysis:* is deal with the purpose. Why are u learning ML?

(24-25) Central Limit Theorem

-if you can solve everything using equations, you don't need to go for theorem. So there is a limitation and unable to do some calculation. So we go for theorem. Central Limit Theorem: Always make the statement **sum of means == mean of the population**

When the population size is too high, you can go for central limit theorem.

When I can know the population size too high or not ? When we are unable to calculate the mean of the population, we can go for central limit theorem. i.e. LHS = RHS ( accept the theorem).

Ex 1: Across three different branches of single company, turnover is 10 Billion dollar but there is no single database.

In order to make a statement about population (company), take , Sum of sales in Singapore, sum of the sale in Canada and sum of the sale in US.

Ex 2: pay scale of Data Scientist. Each company will have their own databases. Average of sample mean.

We must rely on Central Limit Theorem.

Purpose, when we are unable to make any statement on the population, use the Central Limit Theorem.

Based on the average of sample accuracy, we can make the statement and it could be the final accuracy of my model.

[3] When the sample size grows, the graph or sample means will look more like normal distribution.

An essential component of the Central Limit Theorem is that the average of your sample means will be the population mean. In other words, add up the means from all of your samples, find the average and that average will be your actual population mean. Similarly, if you find the average of all of the standard deviations in your sample, you’ll find the actual standard deviation for your population. It’s a pretty useful phenomenon that can help accurately predict characteristics of a population.

**Sum of Sample Mean = Mean of population**

**(26-27) What/Why Correlation**

Ex : If you intend to buy a car but it’s a sports car so mileage is less. You can make the statement.

If you eat more, you will gain more galleries and increase the weight. You can make the statement.

Since there is a correlation, we are able to make the statement. If the one value is given, we are able to give another.

*Correlation* : The relationship between two parameters. If sale increase, then profit increase.

But, if there is no correlation, we can't make statement.

*Problems with Correlation: one festival season, expense more. Next festival season, expense is less, next festival average. So there is no good pattern. Now we can't make statement.*

*(i) Strong Positive Correlation*

*- as sale increase, dam profit increases*

*(ii) Moderate Positive Correlation*

*- there is an increase in sale, there is a moderate increase in profit.*

*(iii) Strong Negative Correlation*

*- there is an increase in sale, there is a dam decrease in profit.*

*(iv) Moderate Negative Correlation*

*- there is an increase in sale, there is a moderate decrease in profit.*

*(v) No Correlation*

*- you increase or increase , there is no effect in profit.*

*(vi) Curvy Linear relationship*

*- as sale increase, profit is decreasing then increasing*

**Scatter Plot (graphical tool-but not a reliable tool always) is used to understand the correlation of the datapoint.**

**There must be a threshold value to decide the correlation and this is called correlation coefficient.**

*The p-value is the probability that the null hypothesis is true. That's it.*

Ex : Test the marketing campaign generates more revenue.

*Null hypothesis:* no change in the revenue after the result of marketing campaign

*Alternate hypothesis :* there is a change in revenue after the result of marketing campaign

- if the value of p is 0.25 then 25% probability that there is no increase/decrease in revenue upon the result of marketing campaign.

- i.e. lower p value gives more confident to the alternate hypothesis is true.

i.e. marketing campaign cause an increase/decrease in revenue.

In order to accept the test result, p value must be low. Acceptable p value is 0.05 and few fields requires p value under 0.01.

**(28) How Correlation is diluted**

- Some parameters are diluting the co-relationship of dataset.

Example: if 1K sale => 100 profit, 2 K sale => 200 profit….5K Sale => 500 profit === > Good case

1K sale => 100 profit, 2 K sale => 2K profit (typo), 4K sale => null …8K sale -> 7K profit(outlier) ==> Bad Case

Before making the statement for 9K sale, check the three different parameters(typo, null, outlier) in the correlation.

These three parameters (typo, null, outlier) are diluting the correlationship and these should be removed before

making the statement.

**Multi co-linearity** - Ex 1: if we have three parameters : Sale, Expenses and Profit

when Sale increase then profit increase. when expense increase then profit increase.

So when any one of Sale or Expense increase then profit increase

we can safely choose one parameter (remove multi co-linearity) and make the statement.

**(29) Limitations of Scatter Over Correlation Coefficient**

- Usually when you look at the data on scatter plot, we can't judge the correlationship and its bit tricky.

- So we need to work on the correlation coefficient (we call it as R value)

- R value helps us to understand the data is scattered which type of correlation (strong positive, etc.).

- If R value is > 0.85, strong positive correlation

- If R value ranging between 0.65 and 0.85, moderate positive correlation

- If R value > -0.85, strong negative correlation

- If R value ranging between -0.65 and -0.85, moderate negative correlation

- If R value ranging between -0.65 and 0.65, NO correlation

*Summary:**Scatter plotter is limited when the data is huge. So, first remove the typo, null and outliers. Then look for correlation coefficient. Based on the threshold make the statement on strong/moderate/negative/no correlation.*

**(30-31) Limitation of Forecasting, Regression and Classification**

- *Forecasting*: is based on the time series, trend, seasonality we can make the statement.

Ex : temperature is 30 o C at 7 am, 31 oC at 9 am. In this case, we can predict the temperature at 11 am.

- Not all the forecasting is depend on time but depend on other various parameters.

Ex: Sales (not depend on time). Depend on various parameter like advertising, marketing ,etc.

- *Regression*: is based on correlation (e.g. weather and humidity, weather and wind ==> temperature)

from the independent variable predict the dependent variable.

- *Classification*:

- Ex 2: no relationship between age and salary (so there is a limitation), no relationship between location and salary

In this case, go with classification. Logic based solution.

*- Time based => forecasting*

*- Pattern Based => Regression*

*- Logic Based => Classification*

**(32) Hypothesis Testing**

What is Hypothesis analysis : All about intermediate analysis before taking decision.

Ex 1: Planning for a career change-

- There is a limitation and you can't decide immediately. You will search in google and do some research.

- Based on the intermediate analysis you will decide to go for a new job or stay in the present job.

**Types**

**(i) Null Hypothesis : always stay with no change. If you are a .Net developer. are you willing to change to ML? No change**

**(ii) Alternative Hypothesis : Planning to go for new skillset and it requires sufficient amount of documentation.**

**E.g. Are you planning to go for an Adventure movie?**

**If you stay at home => null hypothesis.**

**If you choose an alternative to go for a movie => alternative hypothesis (and for this you need additional**

**documents - e.g. movie rating, good and bad, etc.)**

**Errors is Hypothesis :**

*Error 1*: You were convinced by friends and you watched the movie: After watching movie, it was not good and you

didn't like. So Null is true. Call it as Alpha Risk or Type1 error

i.e. when NULL is true you went for alternative. It's called Alpha risk or Type1 error.

*Error 2*: You were not convinced by friends and you didn't watched the movie: But the movie was good one.

i.e. when alternative is true, you stayed with NULL. This is called Type2 or Beta Errors

***Causes of Error :***

Lack of proper understanding or availability of data.

-------------------------------------------

(1) Z-Score

- z-score is a number tells the number of standard deviations from the mean value.

i.e. it’s a measures of how many standard deviations below or above the population mean.

- z-score also known as a standard score and it can placed on a normal distribution curve. Further, it can range from -3 standard deviations up to +3 standard deviations.

- z-score is a way to compare results from a test to "normal" population. i.e. experiment of survey will have thousands of possible results and many seems to be meaningless.

e.g. z-score can tell us the weight of the person compared to the average population.

Machine generated alternative text:
13.57. 
13.9. 
950/0 

*Formula:*

Machine generated alternative text:
Formula to find population mean 
Ex 
n 
Formula to find population standard deviation 
n 
Formula to find the z-score 
z score 

Formula to find the z-score when we have multiple samples:

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image007.png

*Example Problem:*

Machine generated alternative text:
The test scores for a class are normally distributed. The test scores for a class are normally distributed. 
Given: and 0=10 
Given: and 0=10 
What is the probability a student scored above@What is the probability a student scored above 60 
83.31 % 
0-7S 
10 

*Problem to find out z-score when sample is given:*

Machine generated alternative text:
Sample problem: In general, the mean height of women is 65" with a standard deviation of 3.5". What is the 
probability of finding a random sample of 50 women with a mean height of 70", assuming the heights are normally 
distributed? 

Machine generated alternative text:
= (70 - 65) / (3.5"/50) = 5/0.495 = 10.1 
The key here is that we're dealing with a sampling distribution of means, so we know we have to include the 
standard error in the formula. We also know that 99% of values fall within 3 standard deviations from the mean in 
a normal probability distribution (see 68 95 99.7 rule). Therefore, there's Less than 1% probability that any sample 
of women will have a mean height of 70" 

(2) t-score

Machine generated alternative text:
The t score formula is: 

When to use t-score?

* Has a sample size below 30,
* Has an unknown population standard deviation.

Machine generated alternative text:
DO you know the population 
standard deviation, o, ? 
Yes 
Is the sample size 
aoove 30? 
Yes 
No 
use the z-score 
No 
use the t-score 
use the t-score 
• Replace s in the t-score formula with c 

Machine generated alternative text:
The average test score of a population is 75. A sample 
of 9 students are randomly selected. The standard 
deviation for the sample is 10. What is the probability 
the average score for the sample is above 80? 
n: q s: 10 
T Value 
= 1.5 
Degrees of Freedom (df) = 8 
Calculate 
Two tailed P- Value = 0.172003 
One tailed P- Value = 0.036002 
pc k 780): 
: .oto 

*When to use z-score and when to use t-score? Examples*

Machine generated alternative text:
The average test score for an entire school is 75 The average test score for an entire school is 75. 
The standard deviation of a random sample of 
with a standard deviation of 10. What IS the 
probability that a random sample of 5 students students is p. What is the probabilty the 
scored above 80? 
Conditions for usinq t 
1. 
2. n<30 
The average test score for an entire school is 75. 
The sgndatd-dgua.t.lm of a ranüG-GöiÖÖf9 
students is ID. What is the probabilty the 
average test score for the sample is above 80? 
average test score for the sample is above 80? 
Conditions for usinq t 
l.c is unknown 
Conditions for usinq t 
l.c is unknown l/ 
n: 

**ANOVA (Analysis of Variance)**

* Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of three or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

Machine generated alternative text:
e w a esc eme o 
ace we reall 

* *Types of ANOVA:*
* Sample Problem:

Assume three medical treatments where applied on patients with similar diseases. Upon receiving the test results, we need to analyze that which treatment took least time to cure the patients in the best among them. What if some of these patients had already been partially cured, or if any other medication was already working on them?

In order to make a confident and reliable decision, we will need evidence to support our approach and ANOVA can help in this area.

* ANOVA compares the samples on the basis of their means and show that how different these samples are from one another.
* we can use the t-test for the same purpose but it can't give reliable result if three are more than 2 samples (i.e. it will have compounded effect on the error rate of the result).
* Terminologies

*Grand Mean: Mean is an average of a range of values.*

*sample means*

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image016.png

*. Grand mean mean of sample means or mean of all observation combined, irrespective of the sample.*

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image017.png

*Hypothesis: In the medication example, there are two cases are possible - (i) either medications will have effect on patients or it won't. These statements are called hypothesis and it should be tested by experiment or observation. Null hypothesis of ANOVA - valid when all the sample means are equal or they don't have any significant difference. Alternative hypothesis of ANOVA - At least one of the sample mean is different from the rest of the sample means.*

Machine generated alternative text:
Null hypothesis 
Alternate hypothesis 

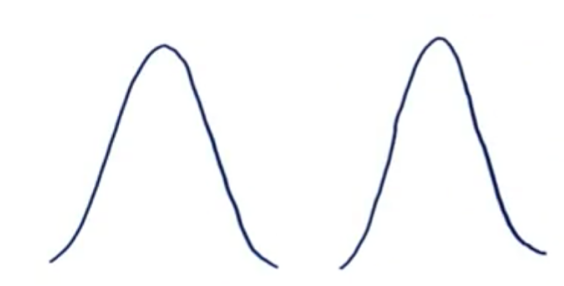
* Between Group Variability

(i) Below two samples distribution are overlapping. So their individual mean will not differ by a great margin.

Hence the difference between individual and grand mean won't be significant enough.



(ii) From the below two samples distributions, samples differ from each other by a big margin. So their individual means also would differ. The difference between the individual and grand means would also be significant. This variability between groups called **Between-group variability**.



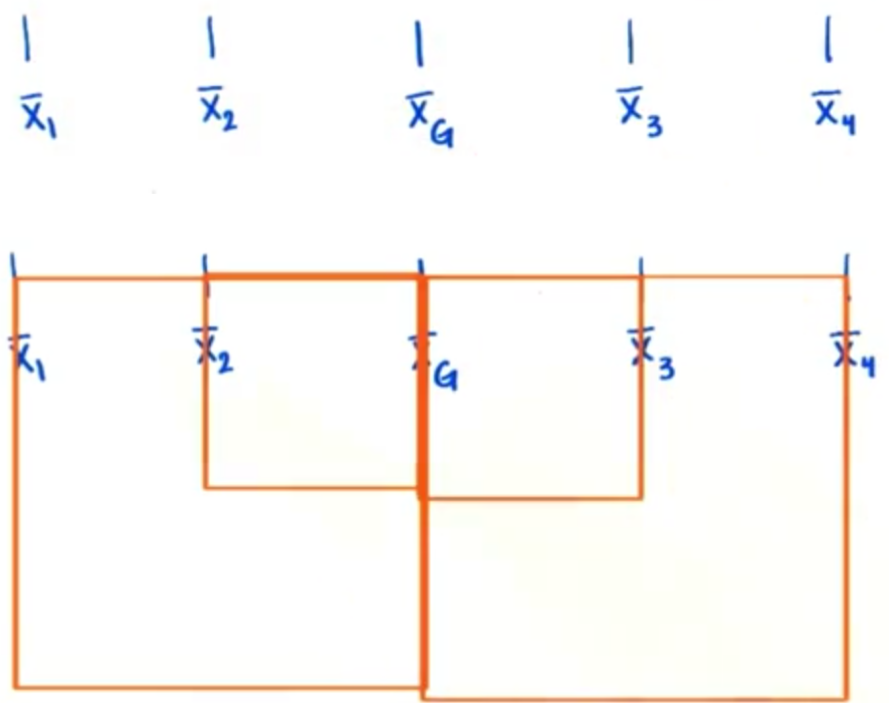
More samples

* Machine generated alternative text:
  Little discrimination 
  Discrimination between Two Groups, 
  but not the third 
  Some Discrimination 
  Large Discrimination 

Steps:

(i) First compute the mean and group mean of the samples.

Given the sample means and Grand mean, we can calculate it as:



(ii) Weigh each squared deviation by size of the sample. i.e. deviation is given greater weight if its from a larger sample. So each squared deviation is multiplied by each sample size and add them up. This is called the **sum-of-squares for between-group variability**

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image023.png

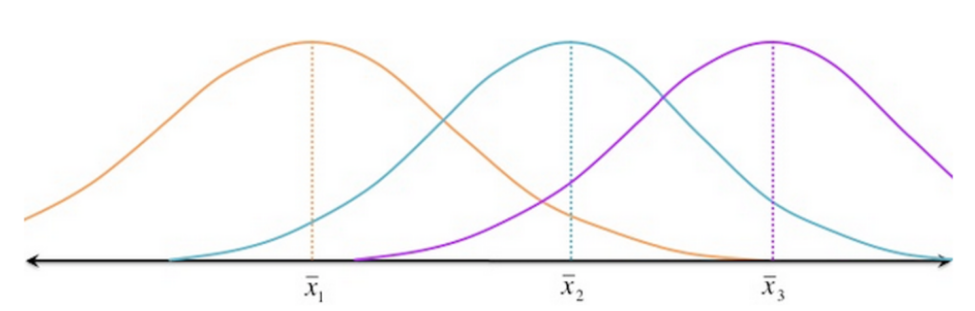
Machine generated alternative text:
SSbetween = ni(T1 — + n2(T2 — iä)2 + n3(G — + .... 

(iii) Measure the **variance between the groups or Measures between groups** (recall sample standard deviation). i.e. sum of each squared deviation and divide it by the degrees of freedom. Degrees of freedom => number of samples (K) - 1

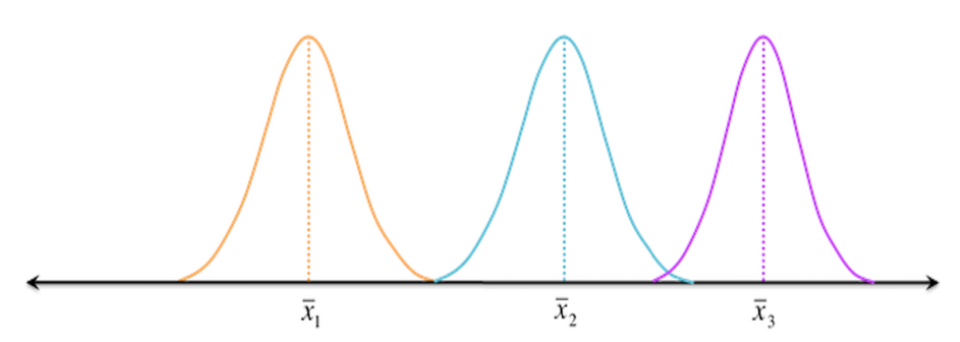
Machine generated alternative text:
+ - + - + ....nk(i+ 
MSbetween = 
k-l 

**Within Group Variability**

In the below three sample distributions, variability of each sample is increased but their distribution overlap and they become part of a big population.



In the another below distribution, same three samples but less variability. Although the means of the samples are similar to the samples in the above image, they belong to different populations.



Such variations within a sample are denoted by **Within-group variation**. It refers to variations caused by differences within individual groups. We can measure the within-group variability using **sum of squares for within-group variability**.

Machine generated alternative text:
= E(xv -3)2 
Note: is the ith value from the first sample, Xi2 is the ith value from the second sample, and so on all 
the way to Xik, the ith value from the kth sample. Xij is therefore the ith value from thejth sample. 
Each squared 
deviation of each value 
from its sample mean 
With within-group variability, is the sum of each squared deviation of each value from its respective 
sample mean (the total area of all the squares in the figure above) MS 
• is the average-sized square. 

we then divide the sum of squared deviations by the **degrees of freedom**

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image029.png

to find a less-biased estimator for the average squared deviation (essentially, the average-sized square from the figure above). Again, this quotient is called the mean square, but for within-group variability:

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image030.png

. This time, the degrees of freedom is the sum of the sample sizes (N) minus the number of samples (k).

Machine generated alternative text:
dfwithin = (ni — 1) + 1) + (m — 1) = m + m + n3 + .... + m — k(l) = N —k 
M Swithin E(Xij ¯ 

F-Statistic:

is statistic which measure the means of different samples are significantly different or not. This is called F-Ratio.

Lower the F-Ratio, more similar are the sample means. In that case, we can't reject the null hypothesis.

Machine generated alternative text:
F = Between group variability I Within group variability 

The F-statistic calculated here is compared with the F-critical value for making the conclusion. In terms of medication example, if the value of the calculated F-statistic is more than the F-critical value, then we can reject the null hypothesis and can say that the treatment had a significant effect.

Machine generated alternative text:
Between-Group 
W riance 
VariatEe 
Within-group variance is larger, and the 
between-group variance smaller, so F will 
be smaller (reflecting the likely-hood of 
no significant differences between these 
3 sample means) 

Unlike the z and t-distributions, the F-distribution does not have any negative values because between and within-group variability are always positive due to squaring each deviation.

Machine generated alternative text:
Shaded Area = alpha 

**One Way ANOVA**

Example Problem : A recent study claims that using music in a class enhances the concentration and consequently helps students absorb more information. As a teacher, your first reaction would be skepticism.

What if it affected the results of the students in a negative way? Or, what kind of music would be a good choice for this? Consider everything and prove that actually it works.

* To figure this out, implement it on a smaller group randomly selected 10 students (same age) from three different classes.

Each classroom has different environment for the students to study. Classroom A - constant music being played in the background, Classroom B- variable music being played , Classroom C - regular class with no music playing. After a month, we conducted a test for all the three groups and collected their test score.

Machine generated alternative text:
Test scores of students (out of 10) 
Class A (constant 
sound) 
Class B (variable 
sound) 
ClassC no sound 
Now, we will calculate the means and the Grand mean. 
= 7,ß2 = 4, = 4.3 & g = 5.1. 
So, in our case, 
Class A (constant 
sound) 
Class B (variable 
sound) 
Class C no sound 
Test scores of students (out of 10) 
10 
10 
Mean 
4.3 
5.1 

- in the above table, mean square of students from Group A is definitely greather than the other two groups. So the treatment must be helpful. But there is also few best students in the classroom A, which resulted in better test score and this lead few questions:

(i) how to decide that these three groups performed differently because of the different situations (not merely by chance)?

(ii) how different are these three samples from each other =?

(iii) what is the probability of group A students performing so differently than the other two groups?

To answer all these questions, first we will calculate the F-statistic which can be expressed as the ratio of Between Group variability and Within Group Variability.

Machine generated alternative text:
SSbetween = 
SSwithin = 
54.6 
90.1 
dfbetween 2 
27 
dfwithin = 
MS between = 
27.3 
MSwithin 3.33 
F = 8.18 

Let’s complete the ANOVA test for our example with

C:\894B6B05\BC36E13B-C99D-4CC2-831C-2C0CCA0C3597_files\image037.png

 = 0.05.

Limitations:

A one-way ANOVA tells us that at least two groups are different from each other. *But it won’t tell us which groups are different*

Machine generated alternative text:
Anova Formula 
Anova is a statistical test which analyzes variance It is helpful in making comparison of two or more means which enables a researcher to draw various results and predictions 
about two or more sets of data Anova test includes one-way anova, two-way anova or multiple anova depending upon the type and arrangement of the data One-way anova has 
the following test statistics. 
MST 
MSE 
Where, 
F = Anova Coefficient 
MST Mean sum of squares due to treatment 
MSE = Mean sum of squares due to error 
Formula for MST is given below: 
SST 
MST = 
SST 
Where, 
SST - Sum of squares due to treatment 
Total number of populations 
p 
n - Total number of samples in a population 
Formula for MSE is given below: 
SSE 
MSE = 
SSE = - 1)S2 
Where, 
SSE = Sum of squares due to error 
S - Standard deviation of the samples 
N - Total number of observations. 

Additional References:

[1] <http://statisticslectures.com/topics/statistics/>

[2] <https://www.quora.com/What-does-N-1-mean-in-a-sample-variance>

[3] <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/central-limit-theorem-definition-examples/#less>

[4] P value article

<http://www.eecs.qmul.ac.uk/~norman/blog_articles/p_values.pdf>

<https://dzone.com/articles/what-is-p-value-in-layman-terms>

[5]z-score

<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/>

[6] t-score

<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/t-distribution/t-score-formula/>

Hypothesis Testing

<https://www.analyticsvidhya.com/blog/2015/09/hypothesis-testing-explained/>

ANOVA

<https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/>

ANOVA sample problems

<https://formulas.tutorvista.com/math/anova-formula.html>

<https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/>

Onlinestatebook.com

<https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/sampling-distribution-of-the-sample-mean>

<http://onlinestatbook.com/stat_sim/sampling_dist/index.html>