

Statistics

1ST TUTORIAL

- *Statistics is the science of collecting, organizing and analyzing data.*
- *Data – Facts and pieces of info.*
- *Statistics is of 2 types –*
 - 1. Descriptive stats – Organizing and summarizing of data.*
 - a. Measure of Central Tendency (Mean, Median, Mode)*
 - b. Measure of Dispersion (Variance, Std. Deviation)*
 - c. Diff types of distribution of data.*
 - d. Histogram, pdf, pmf etc.*
 - 2. Inferential stats – Drawing Conclusions from sample data for Population Data.*
 - a. Z- test*
 - b. T- test*
 - c. CHI sq. test*
 - d. ANOVA test*
 - e. Hypothesis Testing, H_0, H_1 , p value, significance value.*

Example – Age in some class of 20 students are given.

Descriptive stat ques. – Avg Age in the class.

Inferential stat ques. – Are the of students in classroom similar to the ages of all students in university?

2ND TUTORIAL

- *Population(N) – The Group we are interested in studying.*
- *Sample(n) – Subset of the Population.*
- *Goal of the sampling is to create a sample that is representation of entire population.*
- *Best example is Exit Poll of elections or ht/wt of some state etc.*

- *4 Types of Sampling Techniques –*
 1. *Simple Random Sampling – Every member of the population(N) has equal chance of being selected for sample(n).*





 2. *Stratified Sampling – Population(N) is divided into non overlapping subsets based on some characterstics and then randomly selecting from each group or strata for sample(n).*

 3. *Systematic Sampling – Starting from particular point, selecting every kth individual for sample(n).*

 4. *Clustered/Convenience Sampling – Selecting a whole subset from population(N) for sample(n).*
Ex – For example doing a survey related to data analyst requires people of data analyst expertise.

Note – We use diff sampling techniques reqd to our domain.

3RD TUTORIAL

- *Variable – A variable is a property that can take on many values.*
- *Types of Variable –*
 1. *Quantitative Variable –*
 -  *Discrete Variable*
 -  *Continuous Variable*
 2. *Qualitative/ Categorical Variable –*
 -  *Nominal – categories with no intrinsic order or ranking. Ex – Gender, color etc.*
 -  *Ordinal – categories with intrinsic order or ranking but intervals between them can be irregular. Ex – Educational level, Ranks in Police etc.*
 3. *Binary (Dichotomous) Variable – Categorical variables with only 2 possible outcomes. Ex – Yes/No, True/False, Success/Failure etc.*
 4. *Interval Variable - Numerical data with meaningful intervals, don't have true zero point. Ex – IQ scores, temperature etc.*
 5. *Ratio Variable – Numerical data with meaningful intervals, have true zero point allowing comparisons of absolute magnitude. Ex – ht, wt, age etc.*

4TH TUTORIAL

- *Measure of Central Tendency - Refers to the measure used to determine the center of the distribution of data.*
 1. *Mean – Average (easily affected by outliers)*

Population(N)

$$\mu = \sum_{i=1}^N \frac{x_i}{N}$$

Sample(n)

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

2. *Median – Arranging elements in ascending order and picking the middle element (in case of odd no of elements) or picking the avg of middle of 2 elements (in case of even no. of elements).*
3. *Mode – Most Frequent Element. Used in categorical data mostly.*

5TH TUTORIAL

- *Measure of Dispersion - Describes how spread out or scattered the values in a data set are.*
 - *Variance – A measure of how much a value in dataset spread out from their mean value.*

Population(N)

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

μ = Mean of Population data.

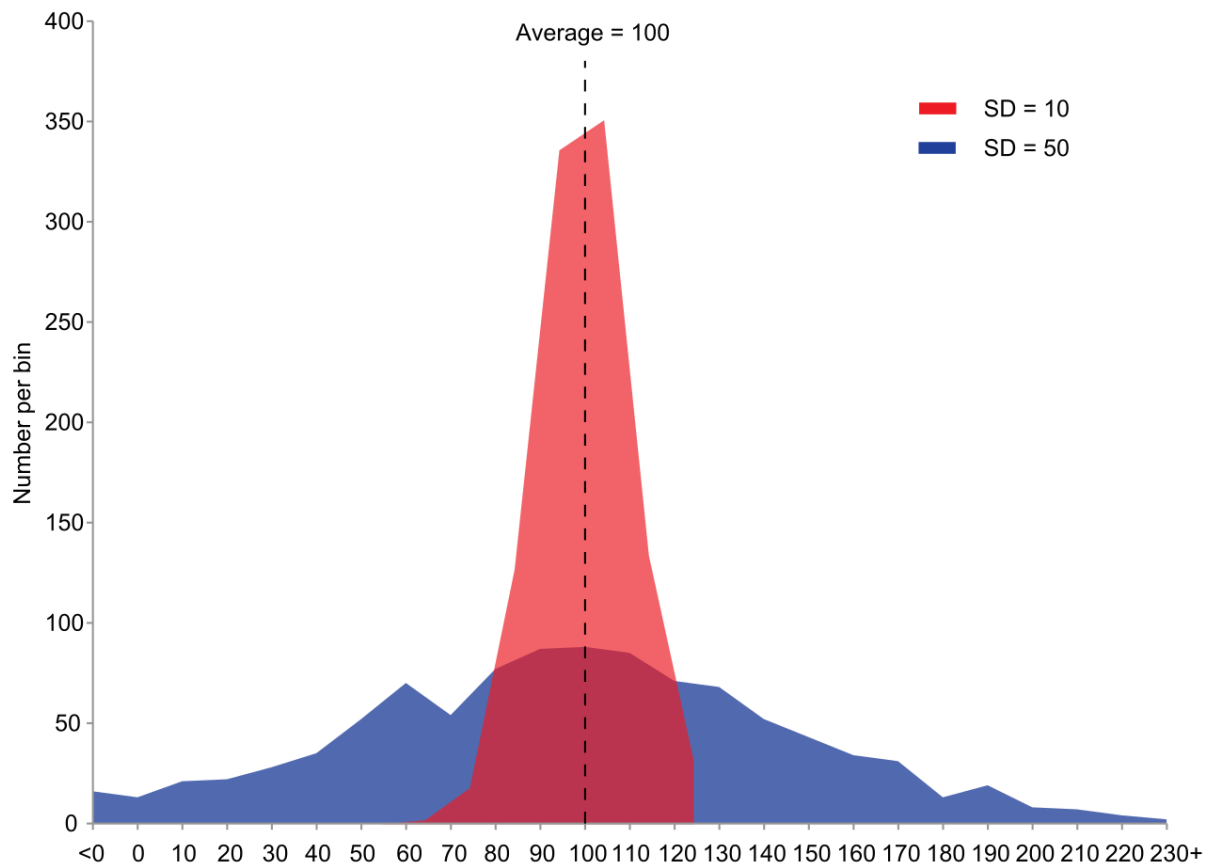
Sample(n)

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

\bar{x} = Mean of Sample data.

$n-1$ = Bessel's correction or Degree of Freedom.


- Higher the value of variance, dispersion of graph will be more and peak will spread and vice versa...



Eg: $X = \{1, 2, 2, 3, 4, 5\}$

x	\bar{x}	$(x - \bar{x})$	$(x - \bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
$\bar{x} = 2.83$			10.84


$\frac{1+2+2+3+4+5}{6} = \frac{17}{6}$



x	\bar{x}	$(x - \bar{x})$	$(x - \bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
$\bar{x} = 2.83$			10.84

$S^2 = \frac{10.84}{5} = 2.168$ { Spread of the data }

↓
Sample Variance



- *Standard Deviation – preferred over variance because it is in same units as the dataset, so it is easier to compare values.*

Population(N)

$$\sigma = \sqrt{\text{variance}}$$

Sample(n)

$$s = \sqrt{\text{variance}}$$

6TH TUTORIAL

- *Percentile – A percentile is a value below which a certain percentage of observations lie.
Ex - 95 percentile means that the person has got better marks than 95% of the entire class.*

Example Dataset – 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

$$N = 20$$

What's the percentile of 10?

$$\begin{aligned}\text{Percentile of } 10 &= \frac{\text{No.of values before } 10 * 100}{\text{Total no of values}(n)} \\ &= \frac{16 * 100}{20} = 80\text{percentile}\end{aligned}$$

This 80percentile indicates than 10 is greater than 80% of the values.

What value exists at percentile rank of 25?

$\text{Index} = \frac{\text{Percentile}}{100} * (n + 1)$
--

$$\text{Index} = \frac{25}{100} * 21 = 5.25$$

This 5.25 is the index where the 25 percentile lies.

Since 5.25 is decimal no., we will take the avg of 5th and 6th element i.e. 5, Therefore at 25 percentile 5 lies.

Or we can also say that 5 is greater than 25% of values.

- Quartiles – Divides the dataset into 4th equal parts with help of 3 quartiles Q1, Q2, Q3.
- $Q1 = 25$ percentile data, formula = $\frac{n+1}{4}$
- $Q2 = 50$ percentile data, formula = $\frac{n+1}{2}$
- $Q3 = 75$ percentile data, formula = $\frac{3(n+1)}{4}$
- IQR (Inter Quartile Range) = $Q3 - Q1 = 50\%$ of data lies here.

Note – The median lies at Q2.

7th TUTORIAL

- Five Number Summary and Box Plot
 - Minimum
 - First Quartile (25%) Q1
 - Median (Q2)
 - Third Quartile (75%) Q3
 - Maximum

$$\text{Minimum} = Q1 - 1.5(IQR)$$

$$\text{Maximum} = Q3 + 1.5(IQR)$$

Removing the Outliers

{ 1, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 }

Lower Fence ← Higher Fence

$Q1 = \frac{25}{100} \times (19+1) = \frac{25}{100} \times 20 = 5 \Rightarrow \text{index} \Rightarrow 3$

$Q3 = \frac{75}{100} \times 20 = 15 \Rightarrow \text{index} \Rightarrow 7$

$IQR = Q3 - Q1$

Lower Fence = $Q1 - 1.5(IQR)$

Higher Fence = $Q3 + 1.5(IQR)$

$Q1 = 3$

$Q3 = 7$

$IQR = 4$

Lower Fence = 1.5

Higher Fence = 11.5


Outliers: 7, 27

Lower Fence = $Q1 - 1.5(IQR)$
 Higher Fence = $Q3 + 1.5(IQR)$

$Q3 = \frac{75}{100} \times 20 = 15 \Rightarrow \text{index}$
 $Q3 = 7$

$IQR = Q3 - Q1 = 7 - 3 = 4$

$LF = 3 - 1.5(4) = 3 - 6 = -3$ $[-3 \leftrightarrow 13]$
 $HF = 7 + 1.5(4) = 7 + 6 = 13$



Removing the Outliers

$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$


$Q1 = \frac{25}{100} \times (19+1) = \frac{25}{100} \times 20 = 5 \Rightarrow \text{index} \Rightarrow 3$
 $IQR = Q3 - Q1$

$Q3 = \frac{75}{100} \times 20 = 15 \Rightarrow \text{index}$
 $Q3 = 7$

$IQR = Q3 - Q1 = 7 - 3 = 4$

Lower Fence \leftrightarrow Higher Fence

Outlier \rightarrow Box plot



Summary Box Plot

- Minimum = 1
- First Quartile (25%) $Q1 = 3$
- Median = 5
- Third Quartile (75%) $Q3 = 7$
- Maximum = 9

\Rightarrow Box plot \rightarrow outlier detect


Removing the Outliers

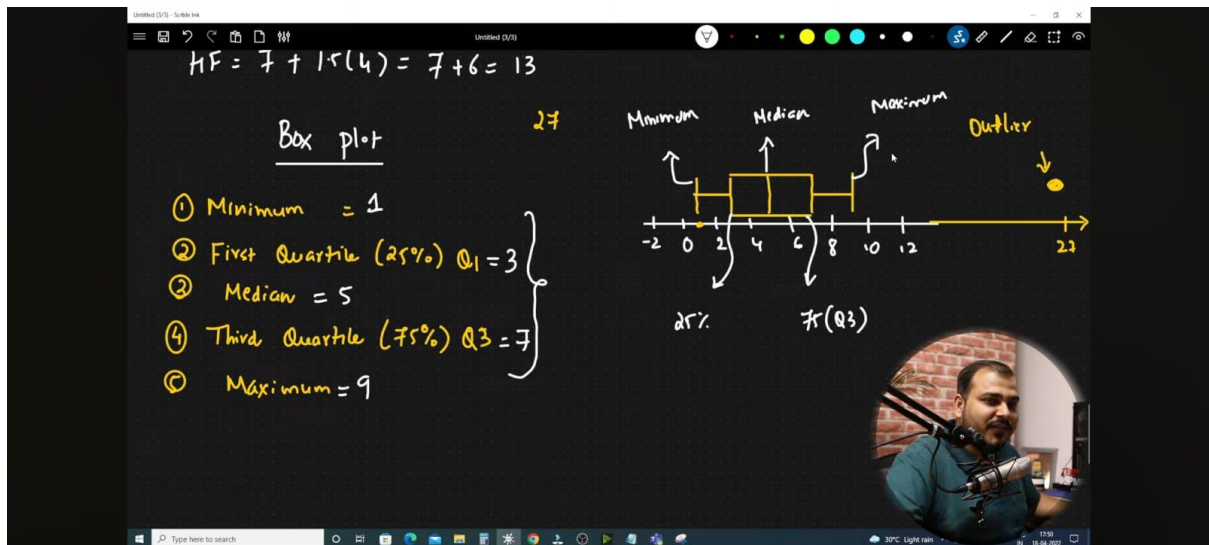
$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

$Q1 = 3$

Lower Fence \leftrightarrow Higher Fence

Outlier \rightarrow B.





8TH TUTORIAL

- *Z-Score – A no. that tells how far the data point is from the mean in terms of std.*
- *A good technique to find outliers.*
- *If $z > 3$ or $z < (-3)$, the data pt. is considered as a potential outlier.*

$$Z = \frac{x - \mu}{\sigma}$$

```

]: # dataset = [11,10,12,14,12,15,14,13,15,102,12,14,17,19,107,10,13,12,14,12,108,12,11,14,13,15,10,15,12,10,14,13,15,10]

]: outliers = []

def detect_outliers(data):
    threshold = 3
    mean = np.mean(data)
    std = np.std(data)

    for i in data:
        z_score = (i-mean)/std

        if z_score > threshold:
            outliers.append(i)

    return outliers

]: detect_outliers(dataset)

]: [102, 107, 108]

```

Finding Outliers By IQR

Finding Outlier by IQR

1. Sort The Data
2. Calculate Q1(25%) & Q3(75%)
3. $IQR = Q3 - Q1$
4. Lower Fence = $Q1 - 1.5(IQR)$
5. High Fence = $Q3 + 1.5(IQR)$

```
dataset = np.sort(dataset)
print("Dataset : ", dataset)

Q1,Q3 = np.percentile(dataset,[25,75])
print("\nQ1 : ", Q1,"\nQ3 : " , Q3)

IQR = Q3 - Q1
print("\nIQR : ", IQR)

Higher_Fence = Q3 + 1.5*(IQR)
Lower_Fence = Q1 - 1.5*(IQR)
print("\nLower Fence : ", Lower_Fence , "\nHigher Fence : ", Higher_Fence)
```

```
Dataset : [ 10  10  10  10  10  11  11  12  12  12  12  12  12  12  13  13  13  13
 14  14  14  14  14  15  15  15  15  15  17  19 102 107 108]
```

```
Q1 : 12.0
```

```
Q3 : 15.0
```

```
IQR : 3.0
```

```
Lower Fence : 7.5
```

```
Higher Fence : 19.5
```

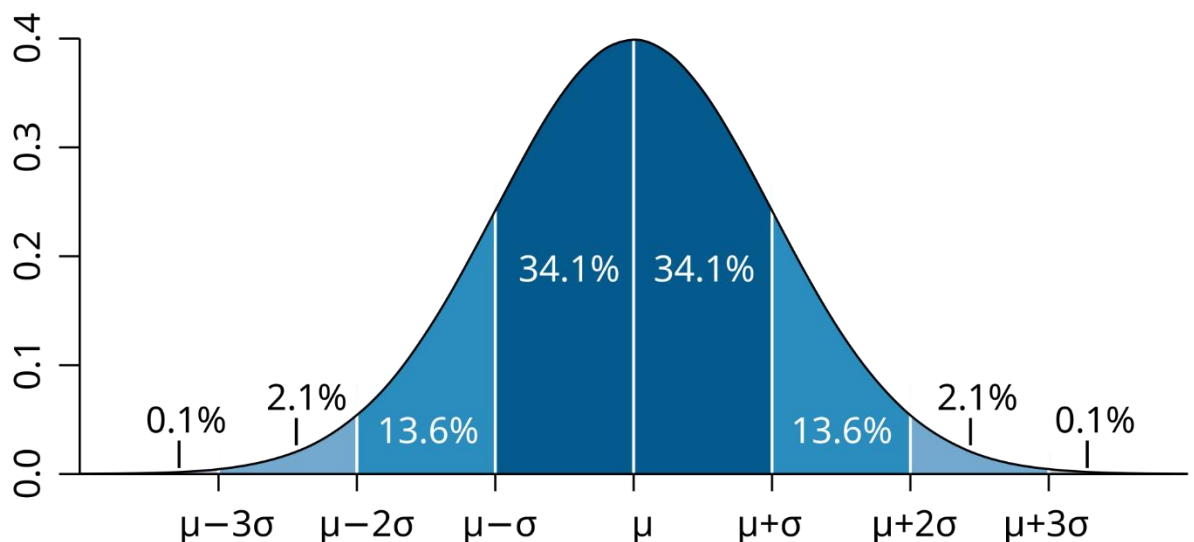
```
outliers = []
for i in dataset:
    if i > Higher_Fence or i < Lower_Fence:
        outliers.append(i)

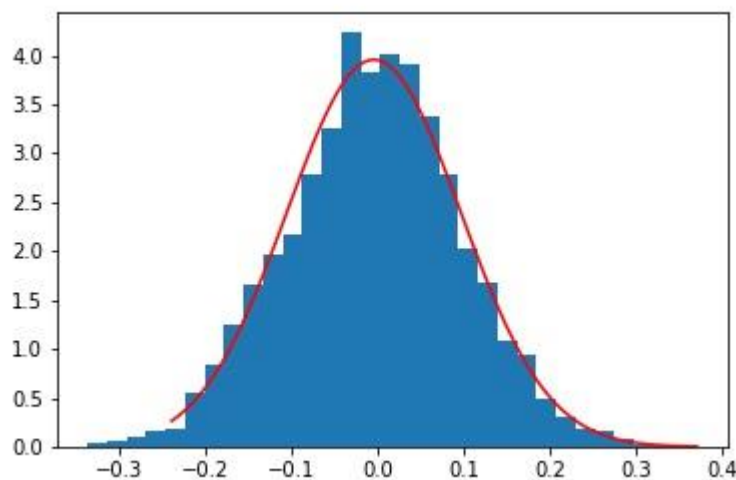
print(outliers)
```

```
[np.int64(102), np.int64(107), np.int64(108)]
```

9TH TUTORIAL

- *Normal/Gaussian Distribution – A continuous probability distribution that is symmetric and bell-shaped.*
- *Mean, Median, Mode all coincide in Normal Distribution at centre of distribution.*
- **68-95-99.7 Rule:** *This rule describes how data is distributed in a normal distribution:*
 - *About 68% of the data falls within one standard deviation of the mean.*
 - *About 95% of the data falls within two standard deviations of the mean.*
 - *About 99.7% of the data falls within three standard deviations of the mean.*
- *This **68-95-99.7 Rule** is also known as **Empirical Rule** or **3 σ Rule**. (3 sigma)*
- *Rest 0.3% data are outliers.*
- *Ex – Ht, Wt, of population*





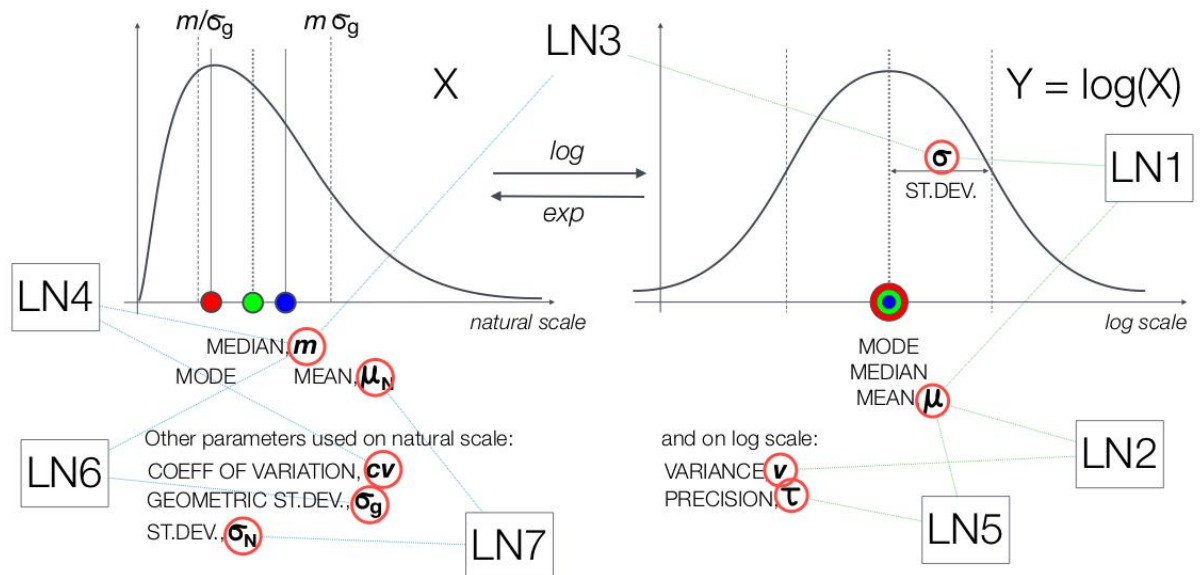
Note – Smoothing the Histogram creates this kind of graph, this kind of graph is called “Probability Density Function” (pdf).

10TH TUTORIAL

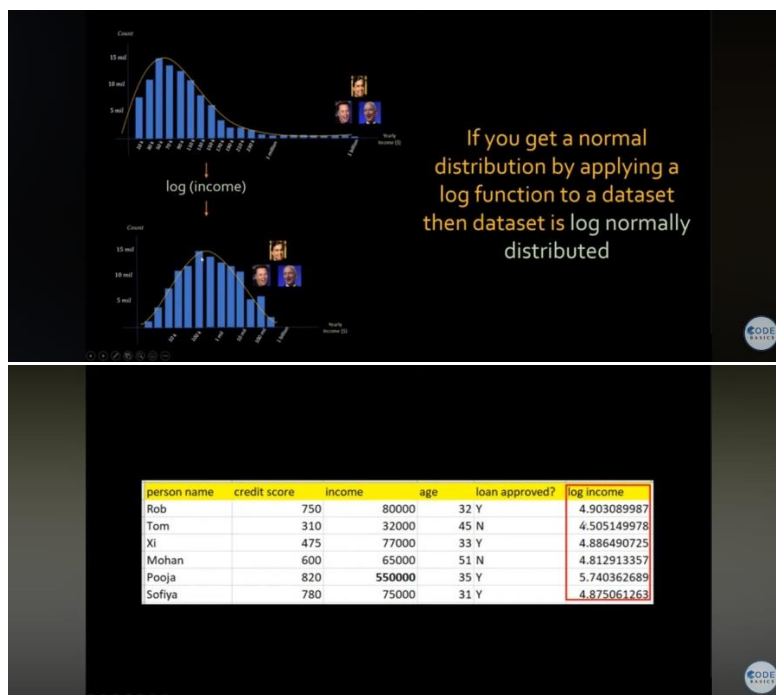
- *Central Limit Theorem - It states that, under certain conditions, the distribution of the sample mean (or sum) of a large number of independent, identically distributed (i.i.d.) random variables will approximate a normal distribution, regardless of the original distribution of the population.*
- *A sample size of 30 or more is considered sufficient for the CLT to hold.*
- *Ex – Imagine a dataset of thousands of data where it shows income of people or ht of people or something else, the histogram made of this data can be normal distribution, skewed or uniform but when we repeatedly take multiple samples of size 30 or more and calculate the means of these multiple samples, then create a histogram of these means, it will approximate a normal distribution.*

11TH TUTORIAL

- **Log Normal Distribution** – A dataset(x) follows a log-normal distribution if the logarithm of the data (y created by using the natural logarithm, $\log_e x$) follows a normal distribution.

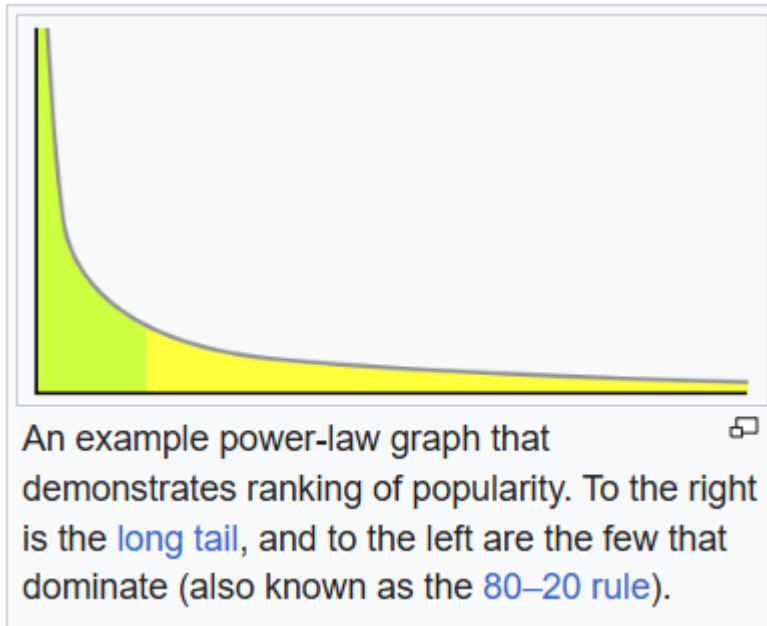


- **Ex – Wealth Distribution, Days spending in Hospital by Patients etc.**



12TH TUTORIAL

- *Power Law Distribution - A type of probability distribution that is characterized by a heavy tail. It is often used to describe phenomena where a small number of events are very common, while a large number of events are rare but possible.*

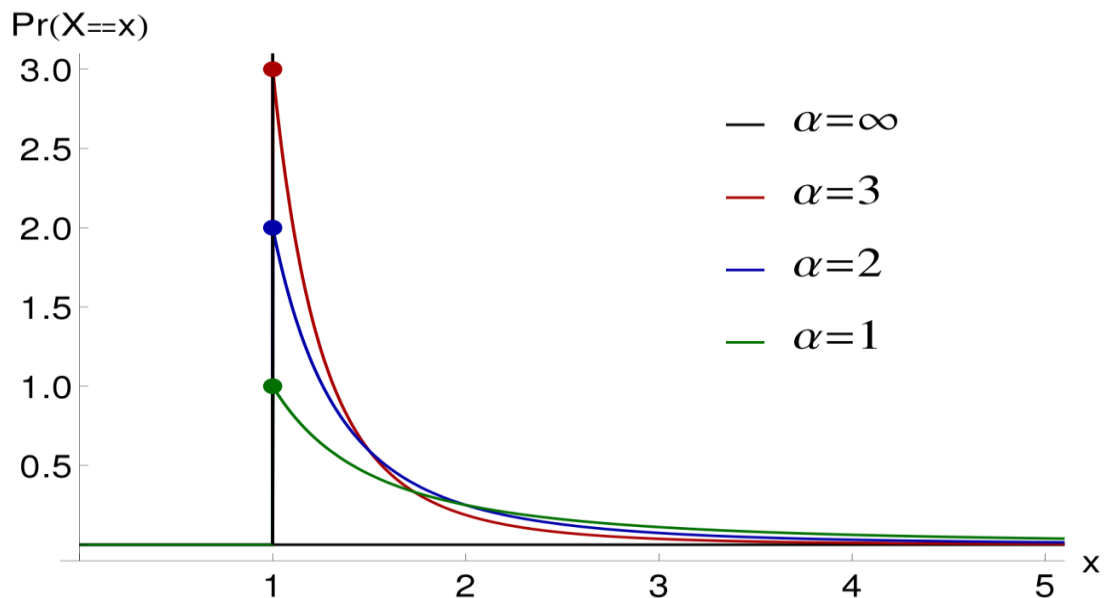


- *Ex – In IPL, a team mostly wins matches with contribution of 2-3 players performance, in real estate, most properties are owned by few rich people, 80% of oil is owned by 20 % of rich nation, 80% of wealth is owned by 20% of people etc.*
- *It can have 90-10,80-20 or even 95-5 distribution.*
- *Through Box Cox Transform we transform Power Law Distribution to Normal Distribution.*

13TH TUTORIAL

- *Pareto Distribution – An example of Power Law Distribution, a non-Gaussian Distribution.*

- The 80-20 rule is directly associated with Pareto Distribution.
- While it can also represent similar relationships like 70-30 or 90-10 but the basic principle (few have much and many have little) is always same.



- The value of α when continuously increases, the graph gradually becomes more and more straight.

14TH TUTORIAL

- Hypothesis Testing – A form of Statistical inference that uses sample data to draw conclusions about population parameter or a population probability distribution.
- Null Hypothesis (H_0) – Baseline, Default, assumption
- Alternate Hypothesis (H_1) – Logically opposite of Null Hypothesis.
- Ex – Tossing Coin is Fair or Not
 - H_0 – Coin is fair.
 - H_1 – Coin isn't fair.

- *Now the experiment will be done to support either Null or Alternate Hypothesis.*
- *For example, tossing a coin 100 times, if we get 50 H and 50 T, then coin is fair and Null Hypothesis is correct.*
- *But we can get 60-40 (H-T) or even 70-30 (H-T) or vice versa so too we can think of it as Null Hypothesis.*
- *The interval of this 60-40 or 70-30 is called Confidence Interval.*

15TH TUTORIAL

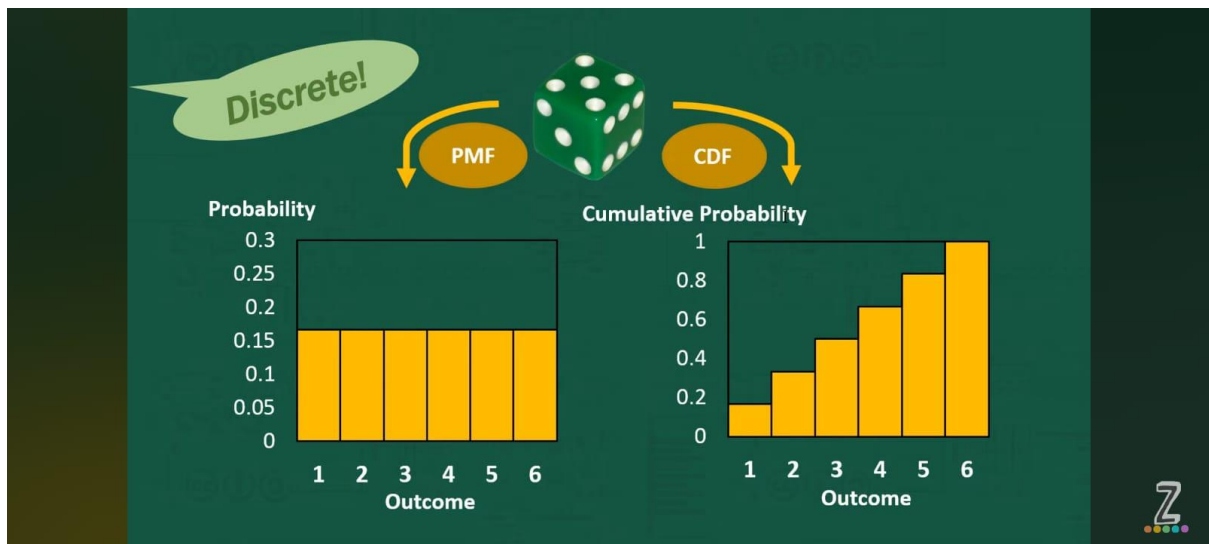
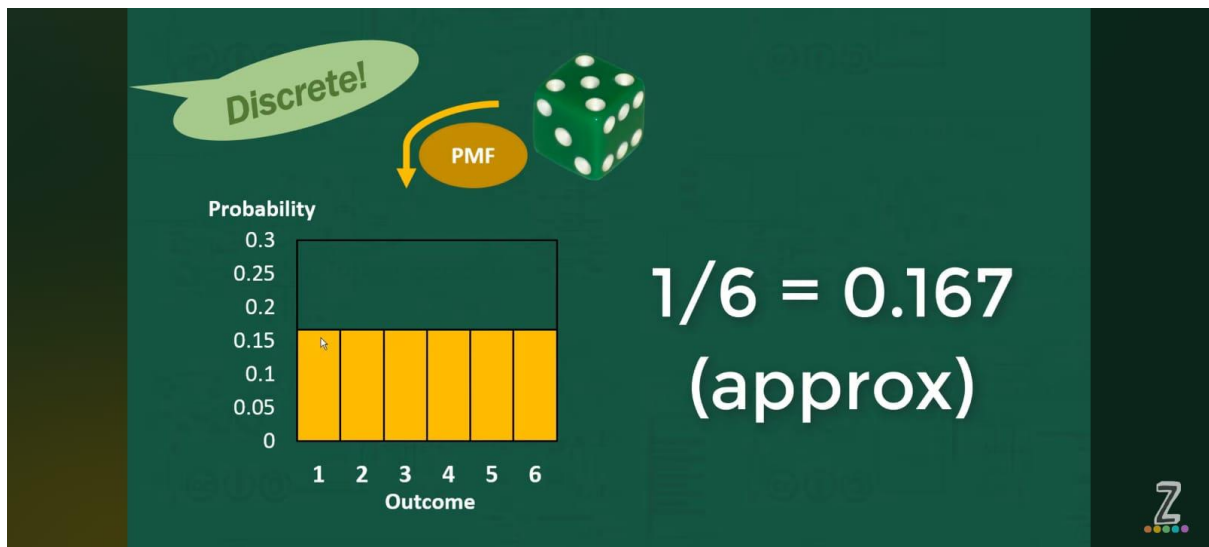
- *It was about implementing practical knowledge in Jupyter notebook.*
- *Refer to [Statistics-Practical-Implementation/Statistics Practical.ipynb at main · Sooraj1411/Statistics-Practical-Implementation \(github.com\)](#)*

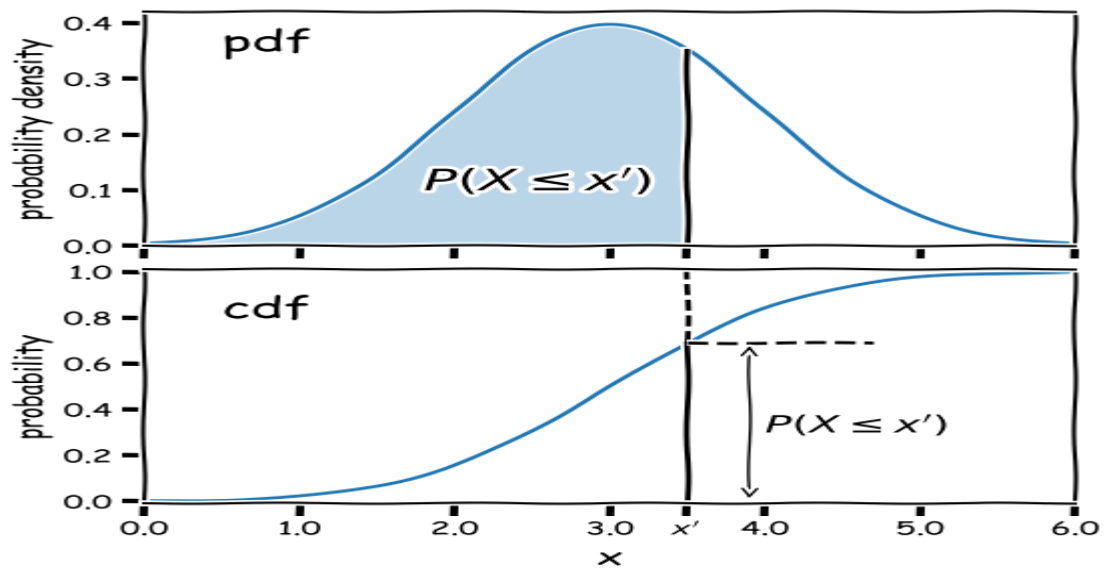
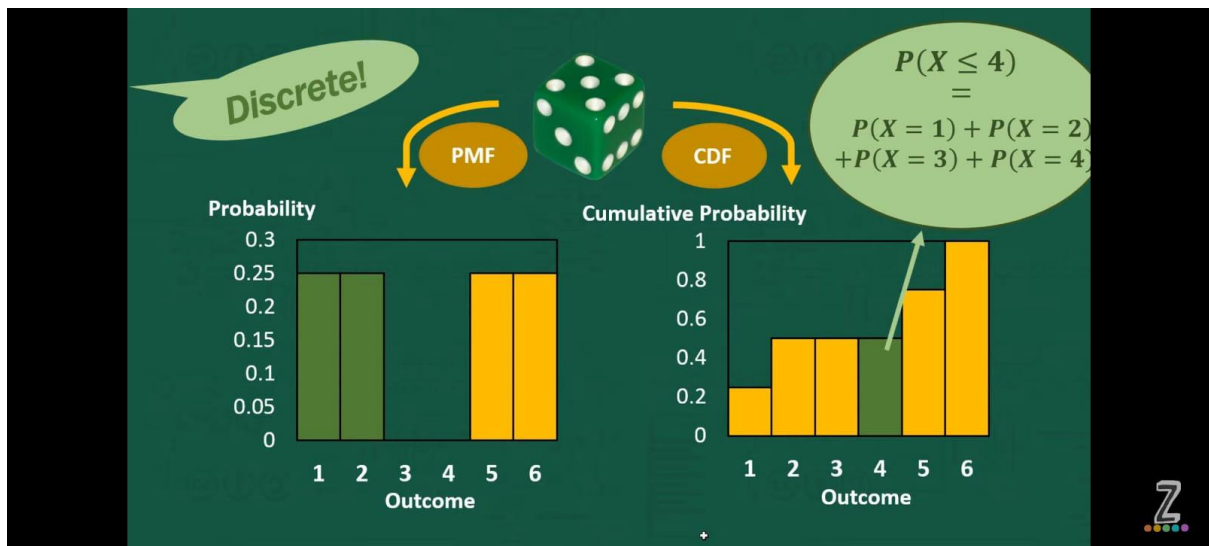
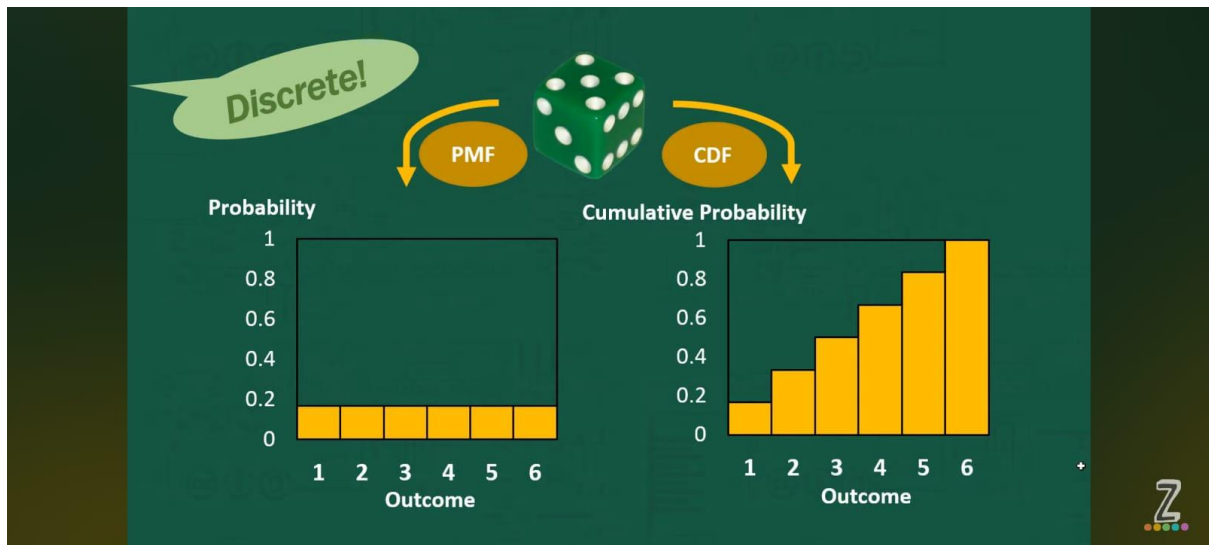
16TH TUTORIAL

- *Probability Distribution Function - The Term can sometimes be used more broadly to refer to either a probability mass function (for discrete variables) or a probability density function (for continuous variables).*
- *Are of 2 Types –*
 1. *Probability Density Func (pdf) – for continuous random variables.*
 2. *Probability mass function (pmf) – for discrete random variables.*

Note - A **random variable** is a variable that represents a numerical outcome of a random process or experiment. It is called "random" because it can take on different values based on the outcome of the experiment, and the specific value it takes is determined by chance.

Ex – The no. of head when tossing 3 coins, the ht. of person (where the ht. can vary between a definite range) etc.





- *Cumulative Distribution Func* - represents the probability that a random variable takes a value less than or equal to a specific point.
- *Examples – (1.) For a die roll, the CDF at $x=3$ gives the probability that the die shows a number less than or equal to 3, which is $\frac{3}{6} = 0.5$ (50%).*
(2.) If the CDF at $x=170$ cm is 0.7, it means there is a 70% chance that an adult's height is 170 cm or less.
(3.) If the CDF at $x=20$ degrees is 0.8, it indicates that there is an 80% probability that the temperature will be 20 degrees or less.

Relationship between CDF and PDF:

- *To get the CDF from the PDF, you integrate the PDF.*
- *To get the PDF from the CDF, you differentiate the CDF.*

Mathematically:

- **$CDF(F(x)) = \int PDF(f(t)) dt$ (from $-\infty$ to x)**
- **$PDF(f(x)) = d(CDF(F(x))) / dx$**

17th TUTORIAL

- *If we have population's std and our sample data(n) > 30, then we will do Z-test otherwise we will go with T-test.*
- *Z-test – Compares sample mean with population mean.*
- *2 Types of Z test –*
 1. *One Sample Z test – Having single mean and compare it to population mean.*

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

For example – College students mean comparing with university students.

2. Two Sample Z test – Comparing 2 diff means from 2 different samples.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

For example – males & females.

Q. Average of ht. of all residents in a city is 168 cm with a population std of 3.9 (σ) . A doctor believes the mean to be diff. He measured the ht. of 36 individuals and found the avg. to be 169.5.

a) State Null and Alternate Hypothesis.

b) At a 95% CI, is there enough evidence to reject the Null Hypothesis.

Solution ->

Given,

$$\mu = 168 \text{ cm} \quad \sigma = 3.9 \quad \bar{x} = 169.5 \text{ cm} \quad n = 36$$

a) Null Hypothesis (H_0) - $\mu = 168 \text{ cm}$

Alternate Hypothesis (H_1) - $\mu \neq 168 \text{ cm}$

It will be a 2 Tail Test because the value can either be greater than 168 or smaller.

$$CI = 95\% = 0.95, \quad \alpha = 1 - CI = 0.5$$

So, when graph is made the 2.5% values lies at both ends.

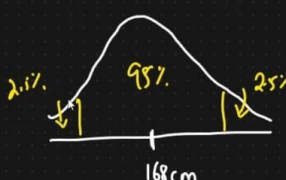

hypothesis

Ans) $\mu = 168\text{cm}$ $\sigma = 3.9$ $n = 36$ $\bar{x} = 169.5\text{cm}$

a) Null Hypothesis H_0 $\mu = 168\text{cm}$

b) Alternate Hypothesis H_1 $\mu \neq 168\text{cm}$ { 2 Tail Test }

c) $C.I = 0.95 \Rightarrow 95\%$ $\alpha = 1 - C.I = 1 - 0.95 = 0.05$

The 95% area is our acceptance boundary and rest 2.5% at both ends are our rejection area.

The whole area of graph = 1

Rejection area at one end = 0.025 (at both ends it will be 0.5)

$$1 - 0.025 = 0.9750$$

This 0.9750 is the acceptance area (it includes 1 one rejection area and excludes other rejection area.).

Now we will check the Z table from [Z-table - Statistics By Jim](#)

It is giving value of 1.96 from Z table.

Now if my Z test value falls between -1.96 and +1.96, then we fail to reject the Null Hypothesis.

Statistical Analysis :

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ (inserting values from above)}$$

$$Z \approx 2.31$$

Conclusions – 2.31 is out of range of (-1.96, +1.96)

So, the Null Hypothesis is rejected.

One more thing, 2.31, the value of Z test is in +ve so it can be concluded that the real mean of the population will be greater than 168 cm and it will lie in the right side of rejection area.

18TH TUTORIAL

- *If population std of population isn't given then use T-test instead of Z-test or even if it is given and the sample size is below 30, then use T-test instead of Z-test.*

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

Q. In the population the avg. IQ is 100. A team of researchers want to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140 within std of 20. Did the medication affect intelligence. CI = 95%

Solution ->

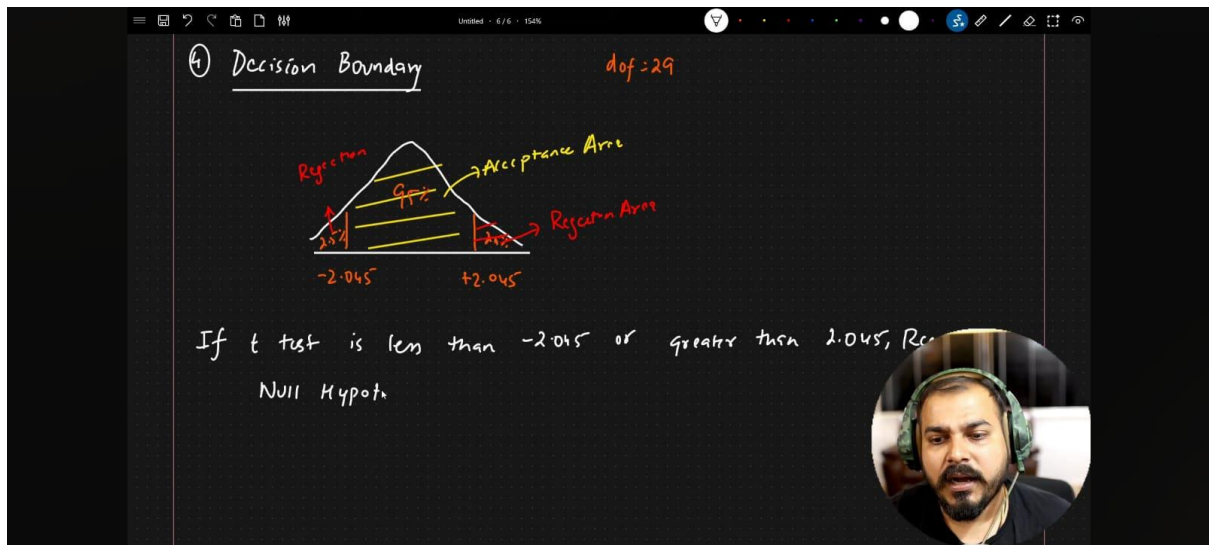
$$\mu = 100, n = 30, s = 20, \bar{x} = 140, CI = 95\%$$

$$\alpha(\text{significance value}) = 1 - 0.95 = 0.05$$

Null Hypothesis (H_0): $\mu = 100$

Alternate Hypothesis (H_1): $\mu \neq 100$ (2 Tail Test)

Degree of Freedom = $n - 1 = 29$



Now we will go to T- Table to search for decision boundary where for 2 Tail Test, $dof = 29$ on [T Table - T Table \(tdistributiontable.com\)](http://tdistributiontable.com)

And the Decision boundary will be ± 2.045

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \left(\frac{140 - 100}{20} \right) * \sqrt{30} = 10.95$$

Conclusion- If t is less than or greater than ± 2.045 then we will reject the Null Hypothesis.

Since $t = 10.45 > +2.045$, we reject the Null Hypothesis.

And since the value of t is positive, it means the medication used by researchers is giving positive effect and helping in increasing the IQ.

19TH TUTORIAL

- 2 Types of Error in Hypothesis Testing.

Actual Predicted \	REJECT H_0	ACCEPT H_0
H_0 True	Type I Error (FALSE POSITIVE)	CORRECT (TRUE POSITIVE)
H_0 False	CORRECT (TRUE NEGATIVE)	TYPE II ERROR (FALSE NEGATIVE)

This is "Confusion Matrix"

- *Type I Error (False Positive) – When Null Hypothesis is CORRECT but we REJECT it.*
- *Type II Error (False Negative) – When Null Hypothesis isn't CORRECT but we ACCEPT it.*
- *Ex of Type I ERROR – When a person has disease but reports say he doesn't*
- *Ex of Type II ERROR – When a person doesn't has disease but report says he does.*

20TH TUTORIAL

SOME KEYWORDS AND THEIR DEFINITION:

1. Estimation – Drawing Conclusion for population based on sample data.

i. Point Estimation – process of giving a single best value as your guess for the unknown population parameter.

a) Properties of Pt. Estimation are Consistent (larger the sample data, more accurate the estimation will be), Unbiased.

b) Drawback of Pt. Estimation are Reliability, not enough evidence.

ii. Interval Estimation – Provides intervals in which population parameter falls, more accurate than Pt. Estimation.

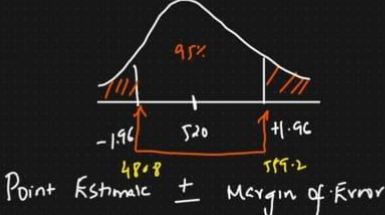
a. Confidence Interval – Interval of Values that is computed from sample data that is likely to contain true population parameter.

Pt. estimation \pm Margin of error

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

① On the Verbal Section of CAT Exam, the standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct the 95% C.I. around the mean?

Ans) $\bar{x} = 520$ $\sigma = 100$ $n = 25$ $C.I. = 0.95$ $\alpha = 0.05$



Point Estimate + Margin of Error

Lower C.I. = $\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 520 - (1.96) \frac{100}{5} = 480.8$

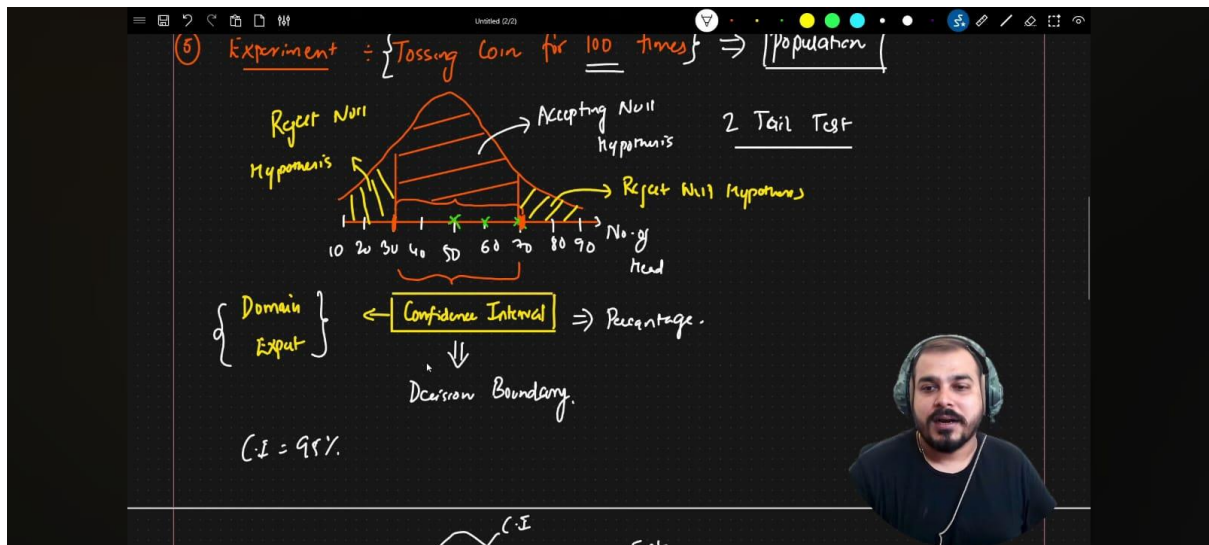
Higher C.I. = $\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 520 + (1.96) \frac{100}{5} = 559.2$

Conclusion

I am 95% confident the mean CAT score lies between

480.8 and 559.2

2. *Population Parameter* – Numerical values that describe the characteristic of whole population. Usually unknown, fixed value.
3. *Sample Statistics* – Numerical value that describes the characteristic of sample. Due to based on sample it can vary from sample to sample.



4. Level Of Significance (α) – a predetermined threshold. Acts as a decision boundary which helps to decide if we have enough evidence to reject or accept the Null Hypothesis.

$$\alpha = 1 - CI$$

21ST TUTORIAL

- Chi Square Test – This test claims about population proportion {categorical variables}-> [ordinal & nominal].
- It is a non-parametric test that is performed on categorical data.
- The graph is always RIGHT-SKEWED.
- Two Types of Chi Square Test –
 - Chi Square Goodness of Fit Test
Used to compare if there is a significance diff. between single category's expected distribution and actual calculated distribution.
 - Chi Square Test for Independence
Used to compare or see if there is any significance diff. between 2 categories.

- *Chi Sq. Goodness for Fitness Test Ex –*

Q. In 2010 census of city, the weight of individuals in small city were as follows –

<i><50 KG</i>	<i>50-75 KG</i>	<i>>75 KG</i>
<i>20%</i>	<i>30%</i>	<i>50%</i>

In 2020, weight of 500 individuals were sampled. Below are results.

<i><50 KG</i>	<i>50-75 KG</i>	<i>>75 KG</i>
<i>140</i>	<i>160</i>	<i>200</i>

Using $\alpha = 0.05$, would you conclude the population diff. of weights has changed in last decade?

Solution-> 2010 Expected

<i><50 KG</i>	<i>50-75 KG</i>	<i>>75 KG</i>
<i>20%</i>	<i>30%</i>	<i>50%</i>

For n = 500 Observation in 2020

<i><50 KG</i>	<i>50-75 KG</i>	<i>>75 KG</i>
<i>140</i>	<i>160</i>	<i>200</i>

For n = 500 Expected data in 2020

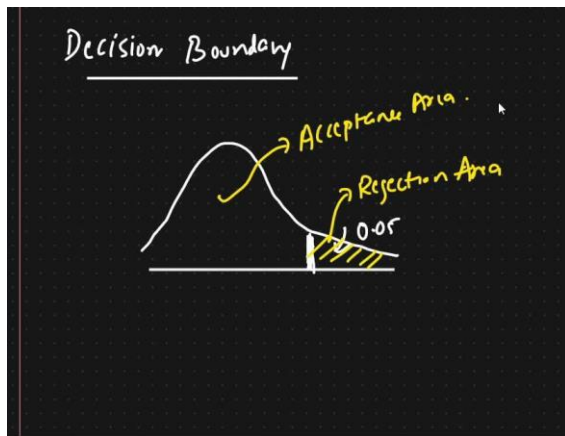
<i><50 KG</i>	<i>50-75 KG</i>	<i>>75 KG</i>
<i>20%*500=</i> <i>100</i>	<i>30%*500=</i> <i>150</i>	<i>50%*500=</i> <i>250</i>

Null Hypothesis (H_0) – The data meets the expectation.

Alternate Hypothesis (H_1) – The data doesn't meet the expectation.

$$\alpha = 0.05 \quad , \text{ CI} = 95\%$$

$$\text{dof} = k-1 = 3-1 = 2 \text{ (k = no. of categories)}$$



For finding the threshold, check Chi Sq. Table on [Chi Square Table & Chi Square Calculator](#)

For this the threshold will be 5.991.

Now if χ^2 is greater than 5.991, we reject the Null Hypothesis or if χ^2 is less than 5.991, we accept Null Hypothesis.

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} = \frac{40^2}{100} + \frac{10^2}{150} + \frac{(-50)^2}{250} = 26.66$$

Now since χ^2 is greater than 5.991, we reject Null Hypothesis.

Conclusion – There is a population diff. and it has increased compared to census in 2010.

- *Chi Sq. Test for Independence Ex –*

Q. Association between males & females to watch a certain film.

<div> <div>Watch</div> <div>Gender</div> </div>	Yes	No	Row Total
Male	140	44	184
Female	178	38	216
Col. Total	318	82	400

Null Hypothesis (H_0) – There is no relation between gender and the film to watch.

Alternate Hypothesis (H_1) – There is a relation between gender and the film to watch.

$\alpha = 0.05$,

dof for Chi Sq. Indep. Test = (row count -1)(col. count -1)*
 $= (2-1)*(2-1) = 1$

Note – for Chi Sq. Test for Independence, we have to find out the expected values through a formula

$$\text{Expected freq.} = \frac{\text{Row total} * \text{Col. total}}{\text{Grand Total}}$$

M/F	Y/N	Obs.	Exp.
M	Y	140	146
M	N	44	37
F	Y	178	171
F	N	38	44

$$\text{Expected males who will watch} = \frac{184 \cdot 318}{400} \approx 146$$

$$\text{Expected males who won't watch} = \frac{184 \cdot 82}{400} \approx 37$$

$$\text{Expected females who will watch} = \frac{216 \cdot 318}{400} \approx 171$$

$$\text{Expected females who won't watch} = \frac{216 \cdot 82}{400} \approx 44$$

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(-6)^2}{146} + \frac{7^2}{37} + \frac{7^2}{171} + \frac{(-6)^2}{44} = 2.65$$

For $\alpha = 0.05$ and $dof = 1$, the threshold will be 3.84

Since the value of χ^2 is less than 3.84, we accept the Null Hypothesis.

Conclusion – The preferences of both genders to watch a film doesn't depend on their gender.

22nd TUTORIAL

- *Analysis Of Variance (ANOVA) – A statistical method used to compare the means of three or more groups to see if at least one group mean is significantly different from the others.*
- *F-distribution or F table is used (Right-skewed).*
- *Key Terms in ANOVA –*
 - › *Factors (variables) - A factor is an independent variable that is being tested in an ANOVA. It represents a categorical variable.*
 - › *Levels - Levels refer to the different categories or values that a factor can take.*
 - › *Ex - If you are studying the effect of teaching methods on student performance,*

*where the teaching method is the factor, and you have three different methods (lecture, group study, online learning), these three methods would be the **levels** of the **factor** "teaching method."*

- *Assumptions in ANOVA test –*

- › *The distribution of sample mean is normally distributed.*
- › *The outliers won't be present in dataset.*
- › *Homogeneity of Variance – Each of population has same variance*

$$(\sigma_1^2 = \sigma_2^2 = \sigma_3^2)$$

- › *Population variable in diff. levels of each independent variable should be equal.*
- › *Samples should be random & independent.*

- *Types of ANOVA –*

- › *One Way ANOVA – One factor with at least 2 levels, these levels are independent.*

Ex – Dr want to test a new medication to dec. headache. They split the participants in 3 categories (10 mg, 20mg, 30mg). The dosage of medications are independent of each other.

- › *Repeated Measure ANOVA - One factor with at least 2 levels, these levels are dependent.*

Ex – A man is running 3 days consecutively, what he run on 2nd or 3rd day will depend on the previous days. If he runs more on 1st day, he will automatically run less on 2nd day due of the loss of energy and so on at 3rd day too.

- › *Factorial ANOVA – Two or more factors (each of which with at least 2 levels), levels can be dependent or independent.*

Ex - Running -> Factor

	<i>Day 1</i>	<i>Day 2</i>	<i>Day 3</i>
<i>Male</i>	8	5	6
<i>Female</i>	6	5	4

- *Hypothesis Testing in ANOVA*

Null Hypothesis (H_0) - $\mu_1 = \mu_2 = \mu_3 = \dots \dots \mu_k$

Alternate Hypothesis (H_1) – At least one of mean will be diff.

$$F = \frac{\text{Variance between Sample}}{\text{Variance within Sample}}$$

23rd TUTORIAL

Q. Dr want to test a new medication to dec. headache. They split the participants in 3 categories (15mg, 30mg, 45mg). Later on the Dr asked patients to rate the headache between [1-10]. Are there any diff. between 3 conditions with $\alpha = 0.05$?

<i>15 mg</i>	<i>30 mg</i>	<i>45 mg</i>
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

Null Hypothesis (H_0) - $\mu_{15} = \mu_{30} = \mu_{45}$

Alternate Hypothesis (H_1) – At least one of mean will be diff.

$N = 21$ (Total records), $a = 3$ (total variable), $n = 7$ (record in each factor)

$$df_{\text{between}} (df_1) = a-1 = 3-1 = 2$$

$$df_{\text{within}} (df_2) = N-a = 21 - 3 = 18$$

$$df_{\text{total}} = N-1 = 21-1 = 20 \quad (df_{\text{between}} + df_{\text{within}} = 20)$$

For $\alpha = 0.05$ and $(df_1, df_2) = (2, 18)$, check threshold value (critical value) from F-Table on [F Distribution Table \(statology.org\)](http://statology.org)

The critical value for this is 3.5546.

Decision Rule – If f-test is greater than critical value, we reject the Null Hypothesis or vice versa.

	SS(sum of sq.)	df	MS(mean sq.)	F
Between	98.67	2	49.34	
Within	10.29	18	0.54	
Total	108.96	20		

$$MS = SS/df$$

$$SS_{\text{between}} = \sum \frac{(\sum a_i)^2}{n} - \frac{T^2}{N}$$

$\sum a_i$ = sum of all values of 15mg, 30mg, 45mg

T^2 = sum of all values of 15mg, 30mg, 45mg completely

$$SS_{\text{between}} = \frac{57^2 + 47^2 + 21^2}{7} - \left[\frac{57+47+21}{21} \right]^2 = 98.67$$

$$SS_{within} = \sum y^2 - \sum \frac{(\sum a_i)^2}{n}$$

$\sum y^2$ = sum of all values with their sq.

$$SS_{within} = 853 - \frac{57^2 + 47^2 + 21^2}{7} = 10.29$$

$$F = \frac{\text{Variance between Sample}}{\text{Variance within Sample}} = \frac{MS_{between}}{MS_{within}} = \frac{49.34}{0.54} = 86.56$$

Since, the value of F is more than critical value ($86.56 > 3.5546$), we reject the Null Hypothesis. And hence, there is a clear diff. with 0.95 CI.

24TH TUTORIAL

- *F Distribution – Also known as Snedecor's F distribution or the Fisher-Snedecor distribution. A continuous probability distribution that's helps in comparing the variance of 2 or more samples.*
- *Graph – Right Skewed.*
- *Parameter – There are 2 df (degree of freedom) d_1 , d_2 and both are greater than 0.*
- *Support – $x \in (0, +\infty)$.*

$$\bullet \chi = \frac{s_1/d_1}{s_2/d_2}$$

- *This formula can be used in ANOVA test or F Test.*

25TH TUTORIAL

Q. The following data shows the np. Of bulbs produced daily for some days by workers A and B.

Can we consider based on data that B is more stable and efficient with $\alpha = 0.05$.

A	B
40	39
30	38
38	41
41	33
38	32
35	39
	40
	34

Null Hypothesis – Variance of A = Variance of B

Alternate Hypothesis – Not Equal to & worker B is more stable.

Mean of A = 37

Sample Variance of A (S1) = 16

Dof1 = n-1 = 6-1 = 5

Mean of B = 37

Sample Variance of B(S2) = 12

Dof2 = n-1 = 8-1 = 7

$$\chi = \frac{s_1/d_1}{s_2/d_2} = \frac{16/5}{12/7} = 1.86$$

Now for $\alpha = 0.05$, dof1 = 5, dof2 = 7 check the critical value from f table which is 3.9715

Since $1.86 < \text{critical value}$ so we fail to reject the Null Hypothesis.

Therefore, There is no clear evidence to support that worker B is more efficient.