**Statistics**

Definition: Statistics is the science of collecting, organizing, and analyzing data.

**Data:**

Facts of pieces of information that can be measured.

Ex. Age of students in class. => {21,22, 23, 22, 25, 24, 26}

**Type of Stat**

**1. Descriptive:** It consists of organizing and summarizing data.

1.1. measure of central tendency [ mean, median, mode]

1.2. measure of [Variance, STD]

1.3. Different types of Distribution of data. eg. histogram, pdf, pmf

2. Inferential: It consists of using data you have measured to form a conclusion.

2.1 Z test

2.2 T test

2.3 CHI square

2.4 Hypothesis testing, H0, H1, p values, significance value

2.5 ANOVA /F test

eg. We fetch sample from population.

**Population** : The group of you are interested in studying

**Sample**: a subset of population.

Simple random sampling: Every member of the population (N) has an equal chance of being selected for your sample (n)

Stratified Sampling: Where the population (N) is split into non-overlapping groups (Strata)

e.g. Gender – Male , Female

Age – (0-10), (10-20)

Profession ()

Systematic Sampling: From Population (N) each nth individual

Convenience Sampling: Only those people who satisfy requirements/criteria.

**Variable:** A variable is a property that can take on any value.

E.g. Height –

Weight -

type of Variable/ Data:

1. Quantitative : (Numerics)

1.1. Discrete : (Whole number) eg. # bank account

1.2. Continuous : (Any value) e.g. height weight

2. Qualitative: (Categorical)

2.1. Nominal : eg. Gender, Blood group, pincode

2.2. Ordinal : Customer feedback like good, bad, ver bad

2.3. Interval: order is matters, value matters, natural Zero is not present,

2.4. Ratio: {Assignment}

4. Scale of measurement

4.1 nominal scale data

4.2 Ordinal scale data

4.3 Interval scale data

4.4 Ratio scale data

4.1 nominal scale data

> Qualitative/ categorical

> eg. Gender, colors,

> Order does not matter

Eg.

4.2 Ordinal scale data

> Ranking is important

> Order matter

> Difference can not be measured

Eg. 1-> best, 2->good, 3-> bad

4.3 Interval scale data

> The order matter.

> Different can be measured

> Ratio can not be measured

> No True “0” starting point.

Eg.

4.4 Ratio scale data

> Order matter

> Differences are measurable including ratio

> it contains the “0” starting point.

Eg. Student marks in class

90,60,30,75,40,50

Order -> ASC = 30,40,50,60,75,90

Diff -> 40-30 = 10

Ratio-> 90/30 = 3:1

Contain 0 point -> always start with 0

**Frequency distribution:**

**Bar graph 🡪** for discrete data

**Histograph 🡪** for Continuous data

ASS

1. Length of diff river –

Length is a ratio scale variable. A ratio scale variable is a variable that has all of the properties of an ordinal scale variable, plus it has a true zero point. This means that we can perform all mathematical operations on ratio scale data, including multiplication and division.

1. Favorite food brand on gender.

ORdinal

1. Marital status

Topic name

1. Measure of central Tendency
   1. Mean
   2. Median
   3. Mode
2. Median of dispersion
3. Gaussian Distribution
4. Z Score
5. Standard Normal Distribution

**1. Measure of central Tendency:**

Refers to the measure used to determine the center of the distribution of data.

* 1. Mean:
* Population 🡪 Sum(all number) / total number, Formula=?
* Sample 🡪 Sum(all number) / total number – 1, Formula=?
  1. Median : specifically for handling outliers (i.e. out of league number)
* Sort all numbers before the operation
* Then check for outlier
* Remove outlier
* Then if the total number is even then take sum(2 number) / 2
* Or total number is odd then take the central number
  1. Mode / Average: Highest Frequency

For Categorical missing data - we use mode

Check numeric value then check outlier using boxplot technic.

**EDA stands for Exploratory Data Analysis.** It is an approach in data analysis and statistics used to summarize and visualize data to better understand its main characteristics patterns and relationships

1. **Median of dispersion:** dispersion means spread of data.
   1. **Variance:** 
      1. **Population Variance:**

* + 1. **Sample Variance:**

QUE. Why do we divide n-1?

ANS🡪 The sample variance is divided by n-1 so that we can create an **unbiased estimator** of the population variance.

Eg. {1,2,3,4,5}

* 1. **Standard derivation:** : Root of Population variance.

SD means kay: when calculate karun aapan center of curve kadhto but curve cha spred kiti aahe he understand sathi SD kadhto.

e.g. mue = 2.83, jr 1.81 asel tr SD ha root of asel i.e. root(1.81) == 1.345

------

To find outlier

**Percentiles and Quartiles:**

**Percentiles:** It is a value below which a certain percentage of observation lie.

Que. What value exists at percentile ranking of 25%?

Value = (Percentile/100) \* (n+1) = (25/100) \* (20+1) = 5.25 i.e. 5 It is index position. Value at index 5 is 3.

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

**Five number Summery**

1. **Minimum**
2. **First Quartile (Q1)**
3. **Median**
4. **Third Quartile (Q3)**
5. **Maximum**

**Remove the outlier.**

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

We must define the range which shows in between numbers are valid and out of this range will be considered Outlier. **[Lower fence <= valid number >= Higher fence]**

**Q1** = 25% = (25/100) \* (19+1) = 5 # here 5 is index in datapoint so value is at 5 index is 3

**Q3** = 75% = (75/100) \* (19+1) = 15 # here 15 is index in datapoint so value is at 15 index is 7

**Interquartile Range (IQR) = Q3-Q1** = 7 -3 = 4

**Lower Fence = Q1 – 1.5 (IQR)** = 3 - 1.5 (4) = -3

**Higher Fence = Q3 + 1.5 (IQR)** = 7 + 1.5 (4) = 13

[-3 to 13] will be valid numbers.

**Minimum = 1 (In datapoint), Q1 = 3, Median = 5, Q3 = 7, Max = 9(In datapoint)**

Median

Q1

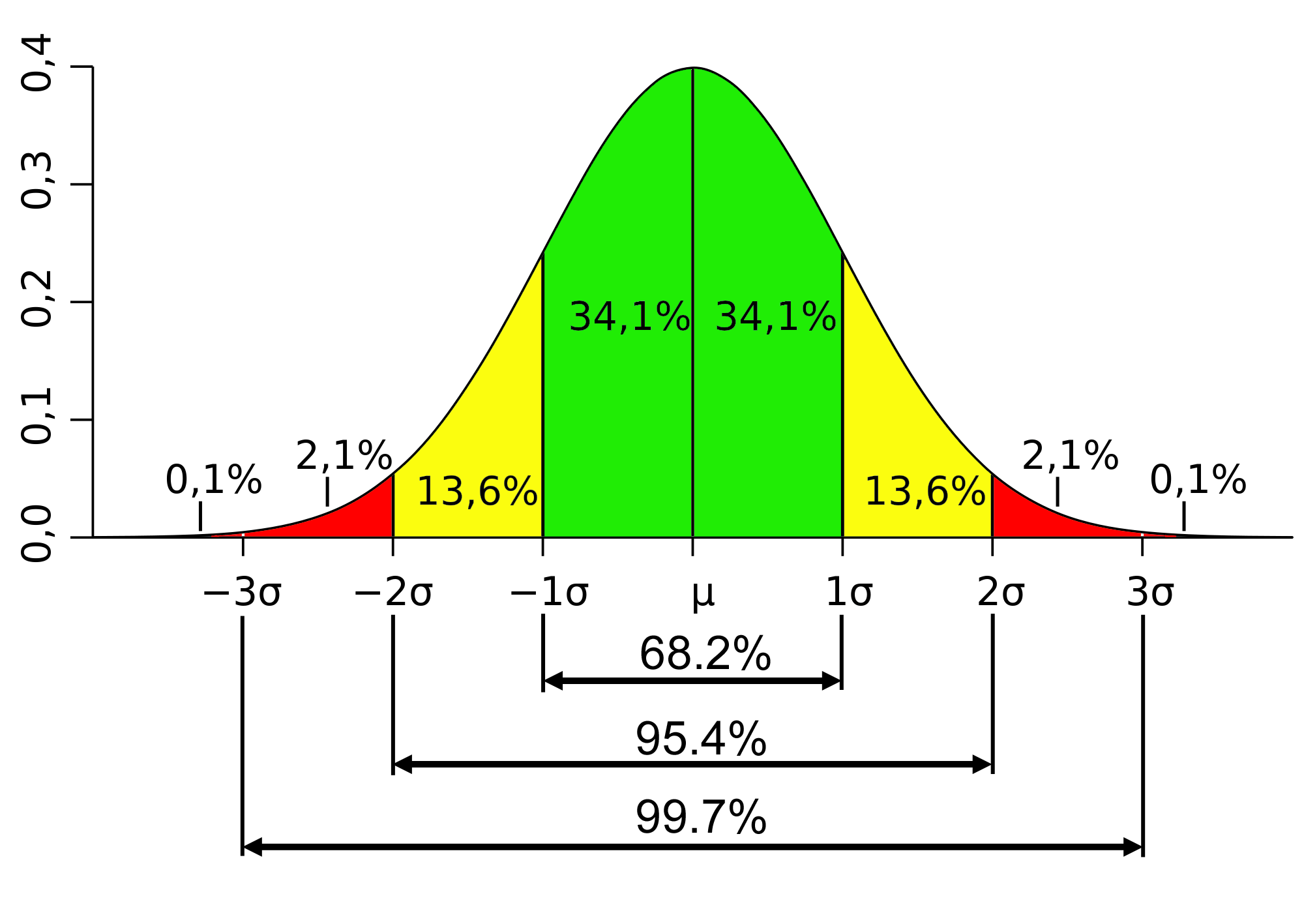
Q3

MAX

MIN

Video = 1 hr:25 min

1. **Distribution: To get idea of the dataset.** 
   1. **Gaussian / Normal Distribution:** Symmetrical curve to other side too.



**Emperical Formula:** 68-95-99.7 % rule, 68% of entire data i.e., 1st standard deviation contains 68% of all data,

Z-Score – will help you to find out how much Standard deviation is away from means value.

Z- Score = (xi – mue) / standard deviation =

If mue = 4., standard deviation= 1

{1,2,3,4,5,6,7} 🡺 Z-Score => {-3,-2,-1,0,1,2,3}

1. Right z table
2. Left

Standard normal distribution (mue==0, standard deviation = 1) 🡪 ideal condition.

**Standardization** – process of trying to convert distribution in standard normal distribution where property is mue is 0 and Standard deviation is 1

**Normalization** where convert all value in between 0 to 1.

MinMax scaler used to normalization.

* 1. **kf**

Random variable: processs of mapping the output of a random process or experiments to a number

Eg. Tossing a coin

X = 0 if head

1 if tail

Rolling a dice

Measure temp for next day

SETs

A – {1,2,3,4,5,6,7,8}

B –{3,4,5,6,7}

Operation:

1. Intersection – A common B
2. Union – All element
3. Difference - A-B
4. Subset –
5. Superset

Histogram and skewness

Histogram 🡺 Frquency

Ages = {10,12,14,18,24,26,30,35,36,37,40,41,42,4,50,51}

50/10 = 5 bin size

So histogram’s x axis will be like 10,15,20,25,30,35,40,45,50

Y axis will be count of ages

Normal distribution / Gaussion where Mean == mode == median

**Skewness**

No skew : Mean == mode == median at center , Q3-Q2~

Right skewness / positive skewed : Mean > mode > median , boxplot Q3 -Q2 >= Q2 – Q1

Left skewness / negative :

Covariance and correlation:

Que. Difference in Covariance and vareince

Cov(x,y) -> if x increase ,y increase then its’ +ve covarince

If x decrease ,y decrease then its’ -ve covarince

Advantage:

1. It helps to find relationship bet X and Y.

Disadvantage:

1. Covarince doesnot have specfific limit value. Which cause the non conclusion
2. To deal with that we use pearson correaltion coefficient [-1 to 1]
   1. The more the value towards +1 the more +ve corre
   2. The more the value towards -1 the more -ve corre
3. Spearman rank correlation : [-1 to 1]

Formuala

Where to use:

Probability Distribution function / Density function (PDF)/ PMF

1. Probaboility Density function (PDF) – to calculate area covered under the curve.
   1. Use for Continuous Random variable:
2. Probability Mass function (PMF):

2.1 use for discrete data

1. Cumulative Distribution function (CDF): Addition of the

i.e dice probability 1/6 for each number so in cumulative 1/6+1/6+ 1/6+1/6+1/6 + 1/6, = 6/6 = 1

Here it is too difficult to understand but CDF is connected to PDF and PMF

Type of probability Distribution:

1. Normal / Gaussian Distribution: we need to create PDF

* Type of continuous probability distribution for a real-values random variable.
* Emperial Formula – 68 – 95 – 99.7 % rule

1. Bernoulli Distribution: to create use (PMF) , outcomes in binary only, p=1-q, q=1-p, if dice throw 10 times so getting head 2 time the prob is 0.2 i.e. P, tail will be 0.8 i.e. q
2. Uniform distribution Distribution: to create use (PMF)
3. Poisson Distribution: to create use (PMF)

* Discrete random variable
* Lambda - # of event
* Describe the event occurring in fixed time interval.
* E.g. # people visiting per hour.

1. Log normal Distribution: to create use (PDF)
2. Binomial Distribution: to create use (PMF)

* Every experiment’s output is binary.
* There experiment will performed n trial
* Discrete random variable

27-09-2023

Z-score and Z-table

What % of data is falling above 4.5?

Wt % of data is falling below 2.5?

Z\_score = (2.5/1) – 4 = -1.5

Area under the curve (<= 2.5) =

Que . In india, the average IQ is 100, with a standard deviation of 15. Whar is % of the population would you axcept to have an IQ lower than 85:

ANS 🡺 mue = 100, stan deviation=15, xi = 85

z-score = (xi-mue) / stan deviation

= -1

Area under the curve (>=85) = 1-0.15866 = 0.84134

**Central limit theorem:**

Sample data che means calculate karun plot karta tevha to normal/gausiaon distribution asel

28-09-2023

Confidence interval –

Estimate – It is an observed numerical value used to estimate an unknown population parameter

Point estimate – a single numerical value used to estimate an unknown population parameter

Hypothesis and Hypothesis Testing –

Hypothesis Testing Mechanism—

1. Null Hypothesis : Assumption that you are beginning with i.e. Person is not guilty
2. Alternate Hypothesis (H): Opposite of null Hypothesis i.e. person is guilty
3. Experiment – Different test to provr alternate hypothesis.

P- Value

It is number, calculated from a statistical test,, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

Significance value-

29-09-2023

T stat and T test

1. Null hypothesis
2. Alternative
3. Degree of freedom
4. Decision rule
5. Calculate T test statistic (using t test table)
6. Conclusion

Que. what is the significance level commonly used in hypothesis testing

The significance level is the probability of rejecting the null hypothesis when it the null hypothesis is true and is denoted by α . The 5% significance level is a common choice for statistical test.

Que .

Ans 🡪 Student's t-distribution, also known as the t-distribution, is a probability distribution that is used in statistics for making inferences about the population mean when the sample size is small or when the population standard deviation is unknown.

Que. T test and z test when to use?

z-test is used to test a Null Hypothesis if the population variance is known, or if the sample size is larger than 30, for an unknown population variance. A t-test is used when the sample size is less than 30 and the population variance is unknown.

Or

Do you know spopulation standard deviation?

Yes then is sample size above 30?

Yes use Ztest

No use Yest and DOF

No then T test

Type 1 and Type 2 errors

Reality : Null hypothesis is true or null hypo is false

Decision : Null hypothesis is true or null hypo is false

Outcome 1: we reject the null hypo, when in reality if is false Which GOOD.

Outcome 2: We reject the null hypo, when in reality it is true – Type 1 Error

Outcome 3: We retain the null hypo, when in reality it is false – Type 2 error

Outcome 4: We retain the null hypo, when in reality it is true – GOOD

Confidence interval : We construct income

30-09-2023

Bayes therom

1. Independent
2. Dependent

Chi Square test

: for goodness of fit test claim about population proportion. {Categorical variable}

Defination 🡪 It is a non parametric test that is performed on categorical {original, normal} data.

**F distribution / F -ratio**

* 2 Degree of freedom, d1, d2 > 0
* DOF= n-1 so it needs to be min. 2

**F test / variance ration test**

4-10-2023

analysys of variance (ANOVA )

it is stat method used to compare the means of 2 or mre group

Type – Factors (variable)

Levels

Assumption in ANOVA

1. **Normality of sampling distribution of means**

The distribution of sample means is normally distributed.

1. **Absence of outlier:** Outlying score need to be removed from dataset.
2. **Homogeneously of variance:** Each one of the populations has same variance **OR**

Population variance in different levels of each independent variables are equal

1. **Sample are independent and random.**

Types:

1. One way anova : one factor with atleast 2 levels, these levels are independent.
2. Repeated measure anova : one factor with atleast 2 levels, levels are dependents
3. Factorial anova: 2 or more factors (each of which 2 levels), levels can be independent od dependent.

**Portion of variance in Anova:**