Statistics:

No. 1
Date . .

→ Statistics is a method of interpreting, analysing and summarising the data

Data: Raw facts or pieces of information

Example:
1) Age of students in classroom
2) Weight of student in classroom

Rajan Devkota
r.devkota.98@gmail.com

Statistics

Descriptive Statistics
→ It consist of organizing and summarizing the data
→ Entire EDA is descriptive statistics

-Descriptive Data with Graph
(a) Bargraph (b) Pie chart
(c) bar graph (d) Candlestick
(e) distribution

Method:
1) Measure of central tendency:
a) mean b) Median c) Mode.

Inferential Data
→ It consists of collecting sample data and making correlation about population data using some experiment

Making Correlation → hypothesis testing

Method:
a) Z-test
b) t-test
c) P-test
d) chi-square
e) F-test

No.

Date . .

Descriptive stats and Inferential stats:

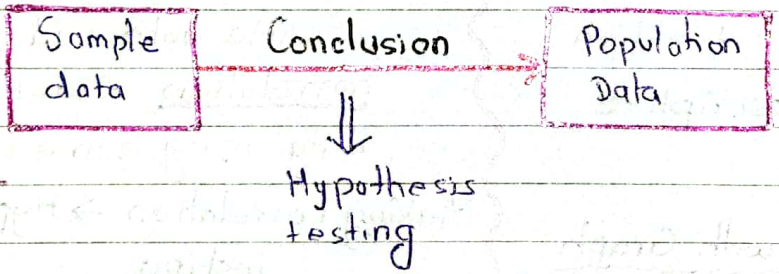
Lets say, there are 20 classroom in an university and you have collected the age of student and weight of student from one classroom.

Descriptive – What is the average age of student in the classroom?

Inferential – ~~What~~ Are the average age and weight of students in the classroom less than overage age and weight of student in the university?

Sampling Techniques

Sampling means selecting the group that you will actually collect data from, in your research.



Types :-

1) Simple Random Sampling :-

Every member of your population N has an equal chance of being selected for your sample.

Example:

Census Bureau follow a random selection of individual inhabitants of the united states for a year, asking detailed question about their lives in order to draw conclusion about the whole population of US.

No.

Date

How to perform Simple Random Sampling?

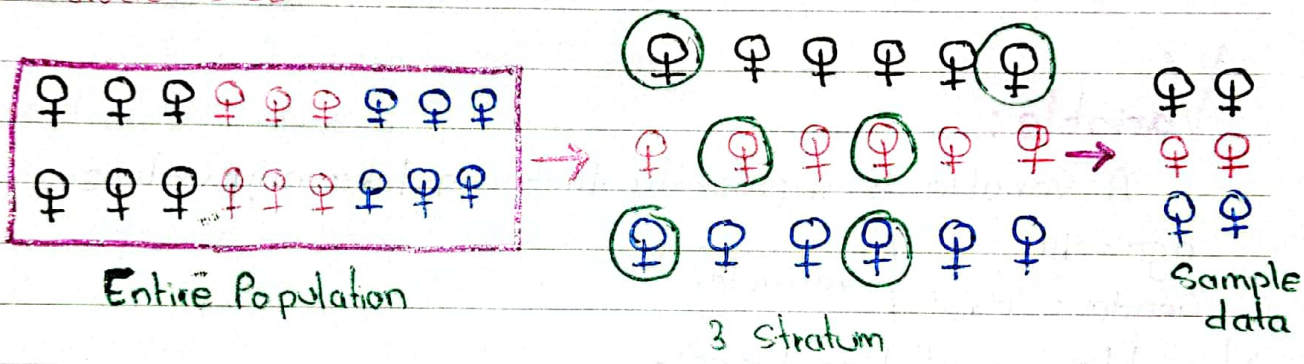
→ Steps

1. Define the population (N).
2. Decide on the sample size (n).
3. Randomly select your sample
4. Collect data from your sample

2) Stratified Sampling

Stratified sampling is a random sampling method of dividing the population into various subgroup, strata, cluster, and drawing a random sample from each.

Each subgroup or stratum consist of items that have common characteristics.



→ Heterogeneous population is segregated into various homogeneous subgroup or strata and sample is extracted from each.

How it works?

Suppose the population of town has to be divided into 3 categories

A	below 18 year	4100
B	18 - 60 years	3500
C	above 60	2400
		10,000

If sample size (n) = 2000
That is 20% of total population (N).

Then:

Sample of A: $4100 \times 0.2 \Rightarrow 820$

Sample of B: $3500 \times 0.2 \Rightarrow 700$

Sample of C: $2400 \times 0.2 \Rightarrow 480$

No.

Date

3. Systematic Sampling

Systematic sampling is a type of probability sampling method in which sample members from a larger population are selected according to a random starting point but with a fixed, periodic interval.

* Selecting every n^{th} individual out of population (N).

4. Convenience Sampling

Only those who are interested in the survey will only participate.

Data Science Survey → Interest DS Enthusiast

Variable:

A variable is a property that can take any values.

age = 14
Gender = 'Male' } variable.

Variable are of two types:-

① Quantitative

→ Measured numerically
→ Can perform mathematical Operation
⇒ Age, weight, height, temp etc.

Quantitative

Discrete Variable
Eg: Whole number → Fixed
Eg: No. of children

Continuous Variable
Eg: Continuous → Decimal value
Eg: height, weight

No.

Date . .

② Qualitative

→ Categorical Variable

→ Based on some characteristics, they can be grouped together

Eg: Gender, Types of flower

Graphical Analysis of Descriptive Data:

1) Histogram

→ a graphical representation of a grouped frequency distribution with continuous class

→ It is 2-dimensional figure.

ⓐ Sort the numbers.

ⓑ Bins → No. of groups

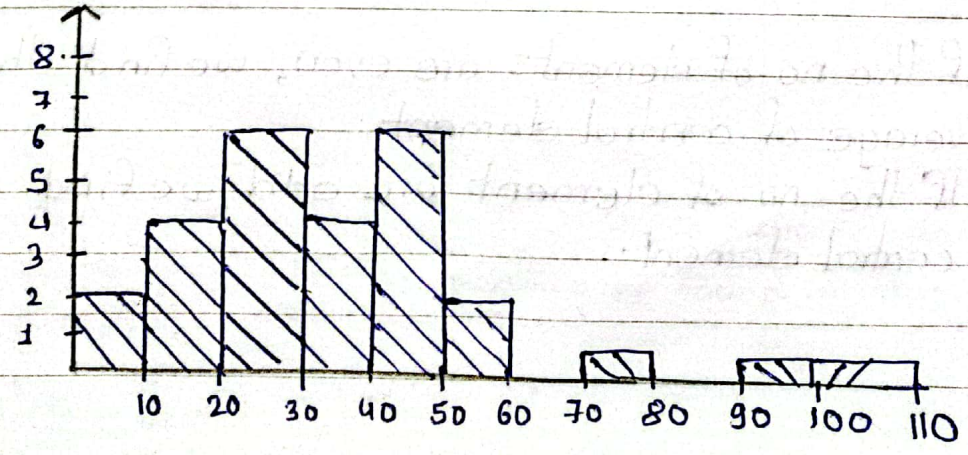
ⓒ Bins size → Size of bins

Example:

Ages = {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100} → Already sorted.

Suppose bins be 10.

Then bins size = $\frac{\text{Max} - \text{Min}}{\text{bins}} = \frac{100 - 0}{10} = 10$



No.

Date

pdf: probability density function ⟶ continuous dataset
pmf : probability mass function ⟶ discrete dataset

2) Measure of central tendency:

1) Mean
2) Median
3) Mode

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position

1) Mean

Population (N)

$$\text{Population mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$$

Sample (n)

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

2) Median

step to find out median:-

1) Sort the number
2) find the central number

Note:-

1) If the no. of elements are even, we find the average of central element
2) If the no. of element are odd, we find the central element.

No.

Date

3) Mode

- Most frequently occuring element.

Measure of Dispersion

1) Variance (σ^2)

→ The variance is the sum of squared deviation from the mean.

Population Variance

* Denoted by - σ^2

$$\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample Variance

* Denoted by - s^2

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

$n-1$ → degree of freedom

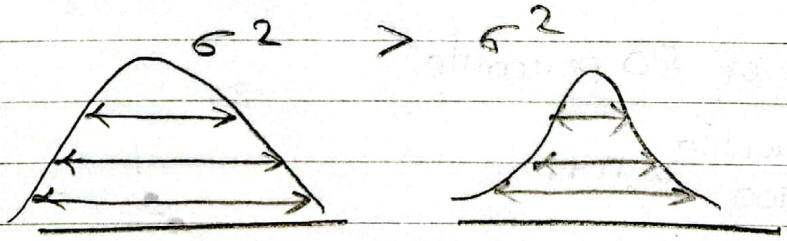
When variance ↑↑ → Spread ↑↑

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

$\mu = 3$.

$$\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

$$= 2$$



Why n-1 in sample variance?

→ Bassel's correction refers to the 'n-1'. This is the correction made to correct for the fact that these sample statistics tend to underestimate the actual parameter found in the population.

No.

Date

2) Standard deviation (σ)

$$\sigma = \frac{\sum_{i=0}^{N} X_i - \mu}{N} \quad \text{or} \quad S = \frac{\sum_{i=0}^{n} X_i - \bar{X}}{n-1}$$

or $S.D = \sqrt{\text{variance}}$

3) Percentile and Quartiles

Percentile:

A percentile is a value below which a certain percentage of observation lie.

99 percentile → It means the person has got better marks than 99% of the entire student.

Dataset: 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10, 11, 12.

What is the percentile rank of 10?

Soln

$$\text{Percentile rank of } x = \frac{\#\text{no. of value below } x}{n}$$

$$= \frac{16}{20} = 80 \text{ percentile}$$

Reverse, What is the value of 80 percentile?

$$\text{Value} = \frac{\text{Percentile}}{100} \times n+1$$

$$= \frac{80}{100} \times (20+1)$$

$$= 16.8 \rightarrow 17^{th} \text{ index}$$

$$\Downarrow$$

$$10$$

No.

Date

4> 5 number Summary

1> Minimum
2> First Quartile (Q_1) → 25 percentile
3> Median
4> Third Quartile (Q_3) → 75 percentile
5> Maximum

Lower fence: $Q_1 - 1.5(IQR)$
Higher fence: $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, ~~are more~~ showing that data near the mean are more frequent in occurrence then data far from the mean.

- In graphical form, it appears as a "bell curve"
- In a normal distribution, it has μ and SD.

when → $\mu = 0$, $S.D = 1$ & skew = 0 & Kurtosis = 3

Then, it become Standard Normal Distribution.

No.

Date . .

Emperical Rule of Normal Distribution

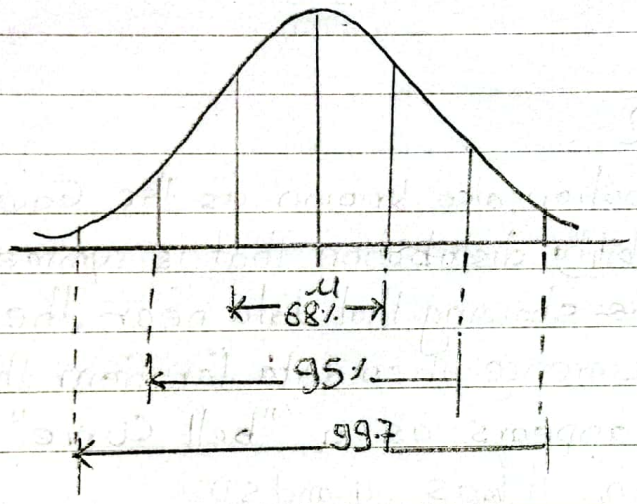
The emperical rule (also called as 68-95-99.7 rule) is a guideline for how data is distributed in a normal distribution.

The rule states that (approximately) :-

a) 68% of the data points will fall within one standard deviation of the mean.

b) 95% of the data points will fall within two standard deviation of the mean.

c) 99.7% of the datapoints will fall within third standard deviation of the mean.



Standard Normal Distribution

$X \sim$ Gaussian Distribution (μ, σ)

$\Downarrow$

$y \sim SND(\mu = 0, \sigma = 1)$

If a continuous random variable X follow normal distribution with parameter mean (μ) & variance (σ^2) then $z = \frac{X - \mu}{\sigma}$

No.

Date

follow standard normal distribution with mean zero and variance one if and only if its probability density function is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Z-score : $\frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow \frac{x_i - \mu}{\sigma}$

Standard Error ← $\frac{\sigma}{\sqrt{n}}$ → $n = 1$ since we apply for each data.

Example:

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

$\bar{x} = 3$,

$\sigma = 1.45$

Value lies bet^n -3 to 3.

$\mu = 0$, $\sigma = 1$

$x[1] = \frac{1-3}{1.45} = -1.414$

$x[2] = \frac{2-3}{1.45} = -0.707$

$x[3] = 0$

$x[4] = \frac{4-3}{1.45} = 0.707$

$x[5] = \frac{5-3}{1.45} = 1.414.$

$\Rightarrow \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why should we have scaled data?

(Years) Age	(kg) weight	(cm) Height
24	70	160
90	91	183

In ML algorithm → some maths are involved. When we have big data and they have high difference, they need more computation to calculate.

No.

Date

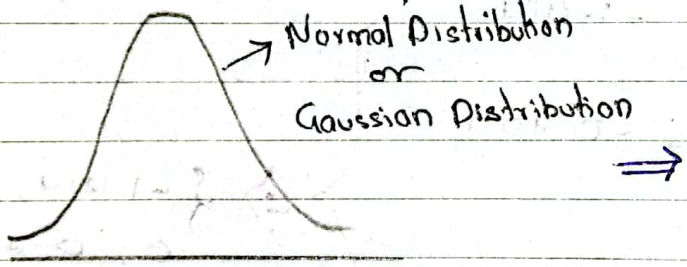
In Normalization, we can give the range.
like between [-1, 1], [0, -5], it can be anything.

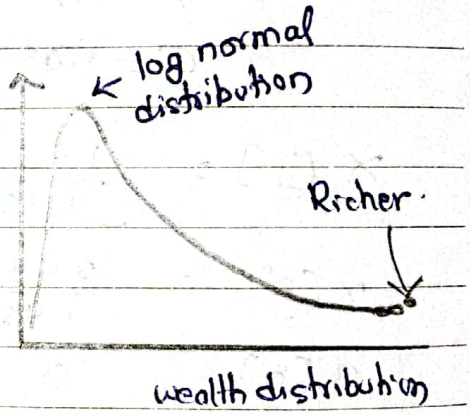
1) **MinMax Scaler** → Normalization techniques
[0, -1]

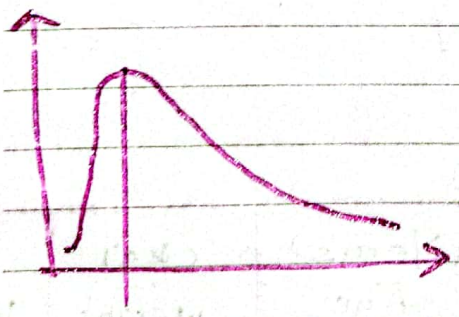
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- preserve shape of data ↑
- Sensitive to outliers ↓
- It is extensively used in deep learning. ↑

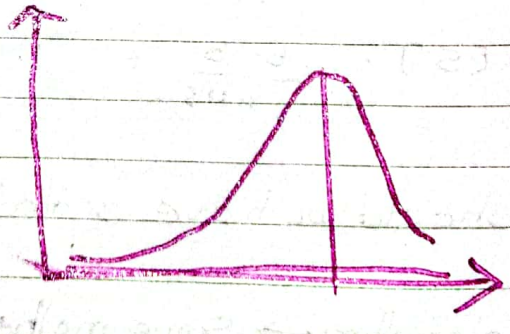
* Log Normal Distribution







- mean will ~~be~~ Higher



mean will be lower

No.

Date . .

If we have distribution which follows log normal distribution, then,

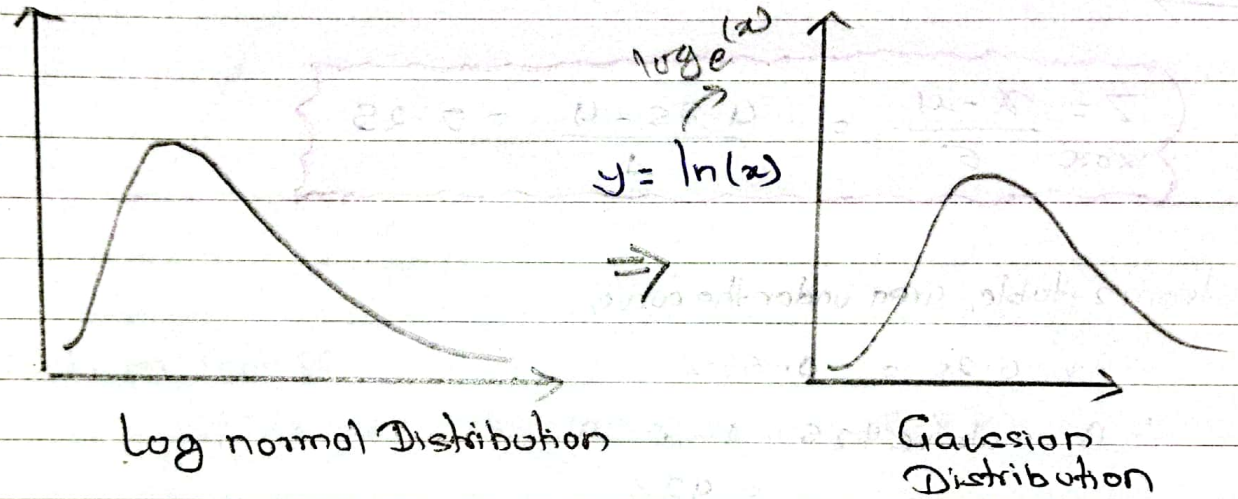
$X \sim$ log Normal Distribution

$\Downarrow$

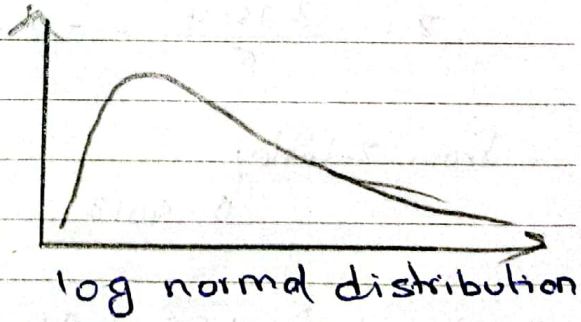
$y = \ln(x)$

$\Downarrow$

$y \sim$ Gaussion Distribution



$\Downarrow$ $x = \exp(y)$



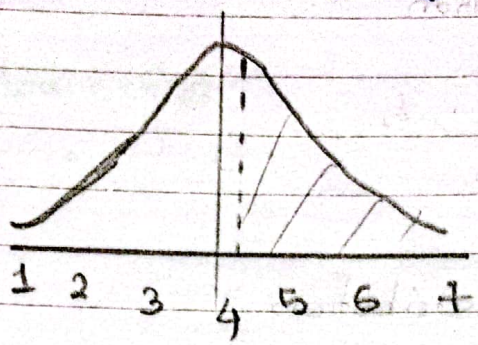
No.

Date

Q. $x = \{1, 2, 3, 4, 5, 6, 7\}$

let $\mu = 4$

$\sigma = 1$



Q: What is the percentage of probability score that falls above 4.25?

$\rightarrow$

$$\text{Z score} = \frac{x - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

From z-table, area under the curve,

$0.25 = 0.59$

$\therefore$ Area above 4.25 = $1 - 0.59$

$= 41\%$

Q. Area below 3.75.

$$z = \frac{3.75 - 4}{\sigma} = -0.25$$

From z-table;

0.4013

Q. Area bet^n 4.75 & 3.75

Area under 4.75

$z_{score} = \frac{4.75 - 4}{1} = 0.75$

Area: $0.75 \Rightarrow 0.7734$

Area under 3.75

$z_{score} = \frac{3.75 - 4}{1}$

Area: $1.75 \Rightarrow 0.9599$

$\therefore$ Area bet^n = $0.9599 - 0.2734 = 18.65\%$

No.

Date . .

Q. In india, the average IQ is 100 with a standard deviation of 15. What is the percentage of population would you expect to have an IQ. ~~lower~~

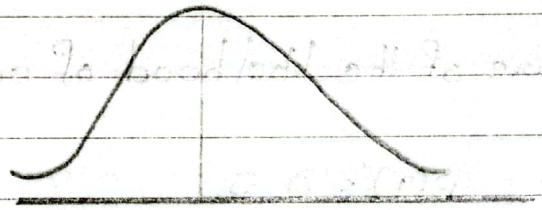
a) lower than 85

b) higher than 85

c) Between 85 and 100

Soln,

Assume, it is gaussian distribution.



a) lower than 85

$$\text{z score} = \frac{85 - 100}{15} = -1 \Rightarrow 15.87$$

b) higher than 85

$$1 - 15.87 = 84.13$$

c) Between 85 and 100

$$\text{z score, 100} = 0 \Rightarrow 50\%$$

$$\text{z score, 85} = -1 \Rightarrow 15.87$$

$$\text{Betn} = 50 - 15.87$$
$$= 34.13$$