DOMS | Page No.
Date / /

Introduction to statistics

Def:- Statistis is the science of collecting, organizing and analyzing data

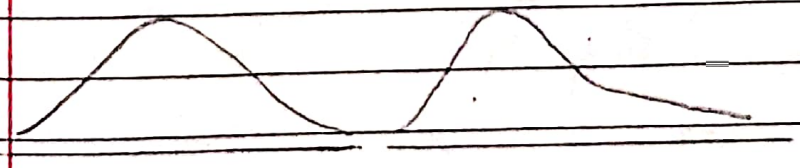
Data:- "facts or pieces of information"
Eg:- Height of student in classroom
{ 175cm, 160cm, 190cm --- }

Types of statistics

1) Descriptive statistics

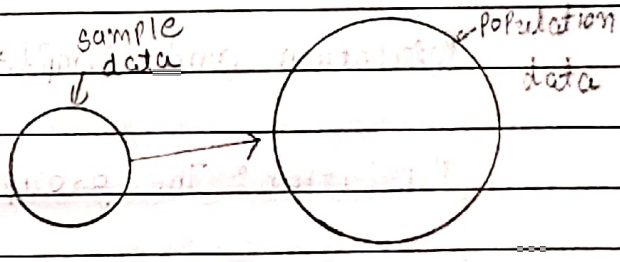
→ It consist of organizing and summarizing data

- Mesure of central Tendency [Mean, Median, Mode]
- Mesure of Dispersion [variance, Std]
- Different type of distribution of data
  Ex:- Histogram, Pdf, Pmf



2) Inferential statistics

→ It consist of using data you have masured to form conclussion



- Z-test } Hypothesis Testing
- T-test } He, H, p value
- CHI Square } significance value

Ex Lets say there are 20 classes at your college and you have collected the heights of student in the class. Heights are recorded [175cm, 180cm, 140cm, 140cm, 135cm, 160cm, 135cm, 190cm]

Descriptive Question

"What is the average height of the entire classroom

$$\frac{175+180+140+135+160+135+190}{8} =$$

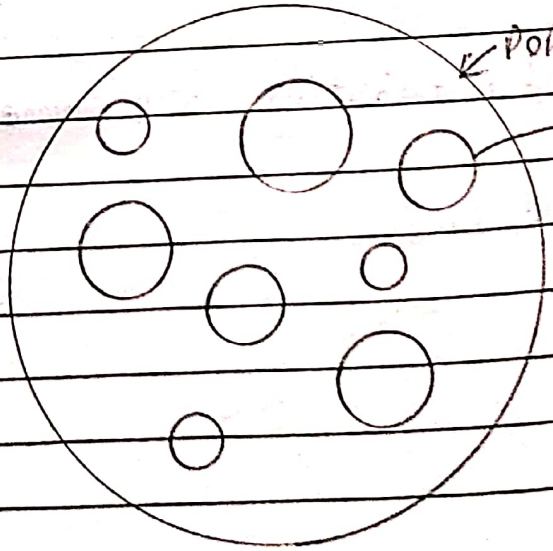
Interential Question (sample)

"Are the height of the students in classroom similar to what you expect in the entire college (↑ Population)

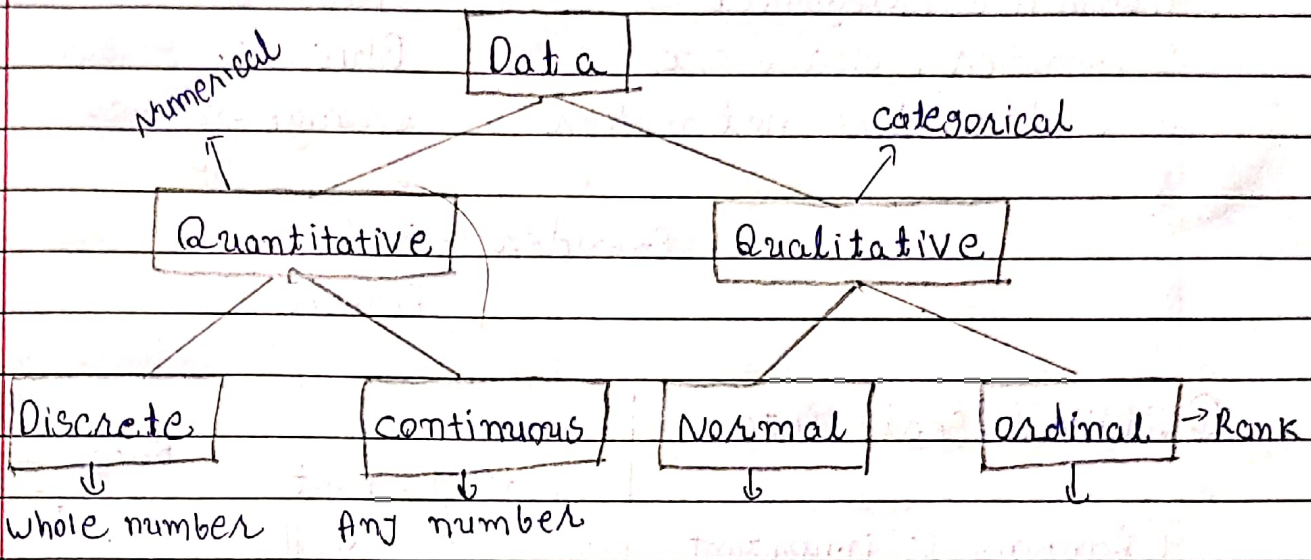
Population and sample data

Population: The group you are interested in studying

Sample: A subset of population



Types of data



No of bank account
No. of childsen in family

Height
Weight
Speed

Gender
Blood Group

Customes feedback
Best, Good, Bad

Scale of measurement

1. Nominal Scale Data
2. Ordinal Scale Data
3. Interval Scale Data
4. Ratio Scale Data

DOMS | Page No.
Date / /

① Nominal Scale Data

1. Qualitive / Categorical
2. Gender, color ∈ E.x.
3. Order does not matter

Eg:- Favourite color.
Red → 5 — 50%.
Blue → 3 — 30%.
Orange → 2 — 20%.
10

Gender :- Male
Female

② Ordinal Scale Data

1. Ranking is important
2. Order matter
3. Difference cannot be measurd

Eg:- 1 → Best
2 → Good
3 → Bad

Race
1st
2nd
3rd

③ Interval Scale Data

1. The order matter
2. Difference can be measured
3. Ratio cannot be measured
4. No such "0" starting point

Eg:- Temprature variable
30 F
60 F
90 F
120 F

0F (not starting point)
-30f

④ Ratio Scale Data

1. The order matter
2. Difference are measurable (including ratio)
3. contain a "0" starting point

Eg:- Student masks in class
0, 90, 30, 75, 40
Ass = 0, 30, 40, 75, 90
40 - 30 = 10
50 - 30 = 20

Ratio = $\frac{90}{30}$ = 3:1

Measure of Central Tendency

1. Mean
2. Median
3. Mode

1 Mean :- average of given numbers

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

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

Eg $x = \{1, 1, 2, 2, 3, 3, 4\}$

$\mu = \frac{1+1+2+2+3+3+4}{7} = \frac{16}{7}$

$\bar{x} = \frac{1+1+2+2+3+3+4}{7} = \frac{16}{7}$

2 Median :- Most frequent value in the sample test

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

1. Sort the Random variable $\{1, 2, 2, 3, 4, 5\}$
2. No. of element $\Rightarrow 6$
3. if count is even

$\{1, 2, \boxed{2, 3}, 4, 5\}$

$\text{Median} = \frac{2+3}{2} = 2.5$

if count is odd

$\{1, 2, 2, \boxed{3}, 4, 5, 6\}$

Median = 3

3 Mode :- Most frequent value in sample test

$\{2, 1, 1, 1, 4, 4, 5, 6, 6\}$

Mode = 1

DOMS
Page No.
Date / /

Measure of Dispersion [Spread of data]

1. Variance
2. Standard deviation

1. Variance

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

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

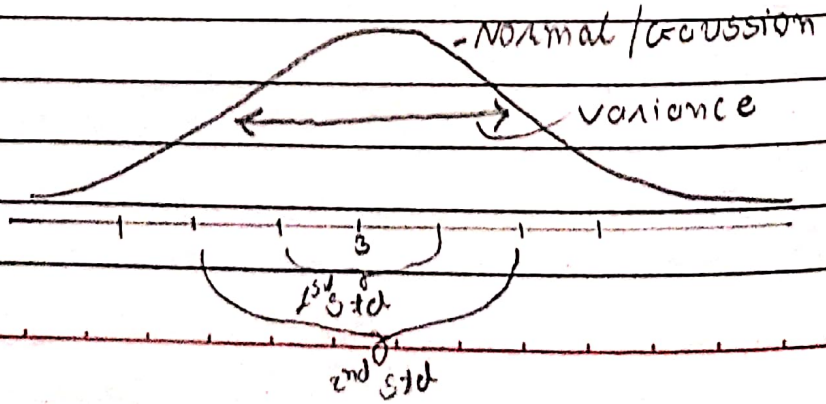
Why we divide Sample Variance by n-1?

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

2. Standard deviation

Population ~~Variance~~ standard deviation $\sigma = \sqrt{\text{variance}}$

Sample standard deviation $std = \sqrt{s^2}$



Random Variable

$x + 5 = 7$ $\quad x = 2$

$8 = y + x \quad y = 6$

Random variable is a process of mapping the output of a random process or experiments to a number.

Eg :- Tossing a coin $\quad y$ = Sum of rolling dice 1 time

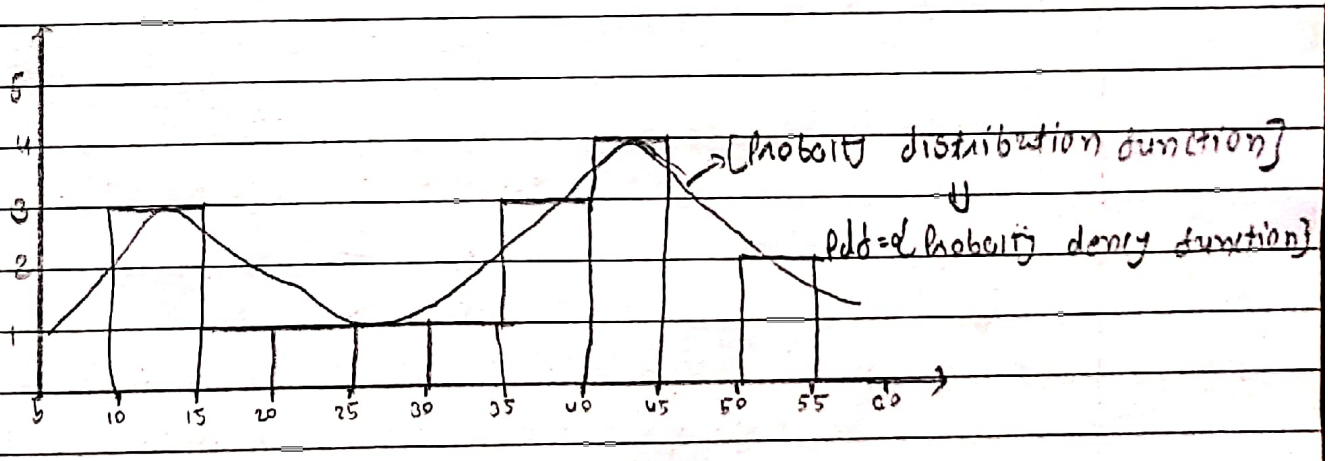
$x = \begin{cases} 0 \text{ if Head} \\ 1 \text{ if Tails} \end{cases}$

Histogram and skewness → [Frequency]

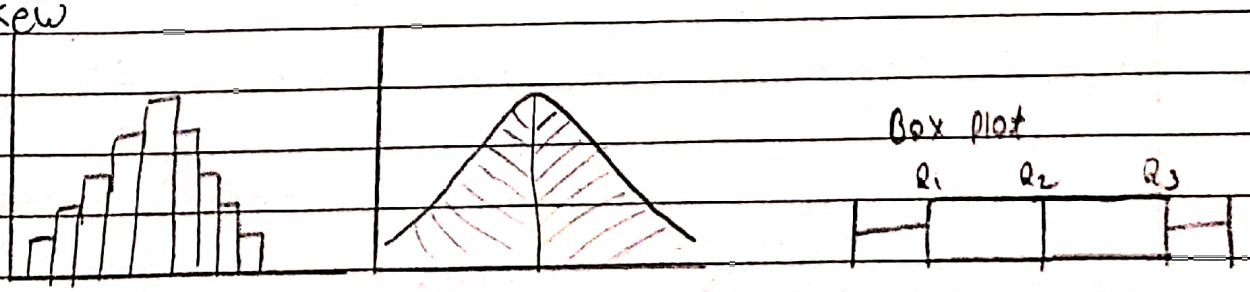
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

$\frac{50}{10} = 5$ → bin size {no of bins = 10}

$\frac{50}{20} = 2.5$ → bin size {no of bins = 20}

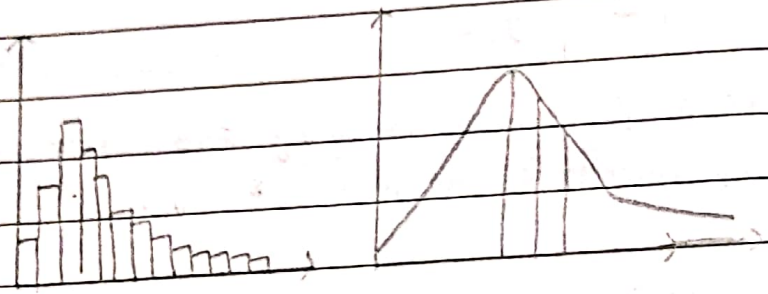


1. Skew

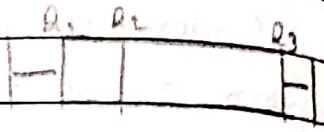


The mean, mode and median all are perfectly at the center

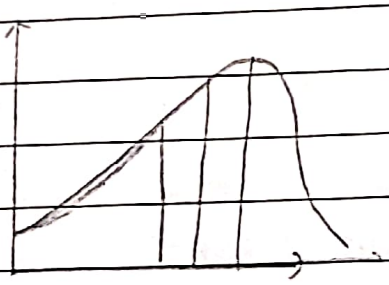
2 Right skewed



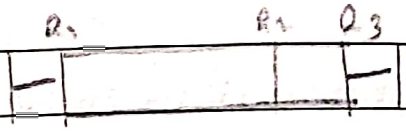
Box plot



3 Left skewed

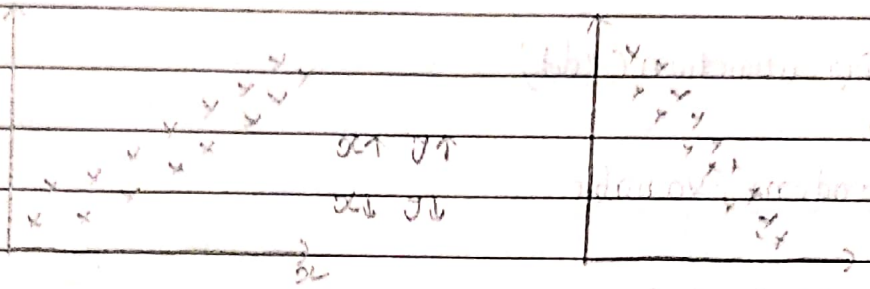


Box plot



DOMS | Page No.
Date / /

Covariance & Correlation



Covariance

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Advantages	Disadvantages
Relation between x and y +ve or -ve value	Covariance does not have a specific limit value

Pearson correlation coefficient [-1 to 1]

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

The more value towards +1 the more +ve correlation it is (x, y)

The more value towards -1 the more -ve correlation it is (x, y)

Spearman Rank correlation [-1 to 1]

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) + \sigma(R(y))}$$

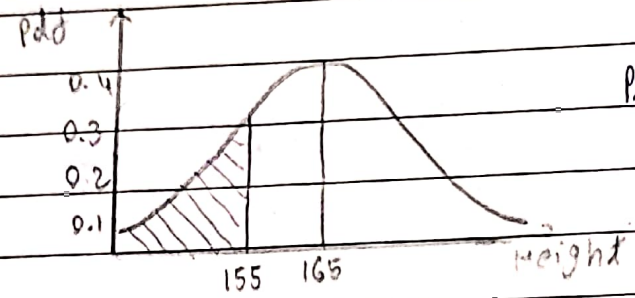
DOMS | Page No.
Date / /

Probability Distribution Function

1. Probability Density function (pdf)

Continuous Random Variable

Eg: Height of Student in classroom

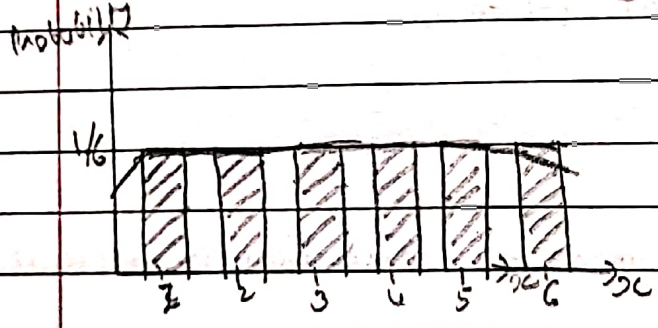


$Pr(x \leq 155)$ Area under the curve

2. Probability Mass function (pmf)

Discreate Random variable

Eg: Rolling a Dice {1, 2, 3, 4, 5, 6}



$Pr(1) = 1/6$

$Pr(2) = 1/6$

$Pr(x \leq 4) = Pr(1) + Pr(2) + Pr(3) + Pr(4)$

$= 1/6 + 1/6 + 1/6 + 1/6$

$= 4/3$

3 Comulative Distribution function (cdf)