

Statistics → 25ⁿ



Dfn:

It is the science of collecting,
organizing and analyzing
data



↳ Machine Learning



Data

Age = 23 }

→ 6 ft }

Data = "facts or pieces of information"

Statistics



Descriptive Stats

Inferential Statistics

① Consists of organizing
and summarizing data

② Consists of using data
you have measured to form
conclusions/inferences

There are 20 classes at a university and you have collected
the ages of students in one class

Ages = { 20, 21, 18, 35, 26, 24, 22, 21 }

↳ Histogram



Descriptive : What is the most common age of student in
your class? { 21 } → mode

Inferential Stats : Are the ages of the students in this classroom

Similar to what you expect in a normal class at this university? → Conclusion?

3rd day

Class 1
↳ 100
↳ 10%

University? ?

- ① Z test
- ② t test
- ③ Chi-square test
- ④ ANOVA

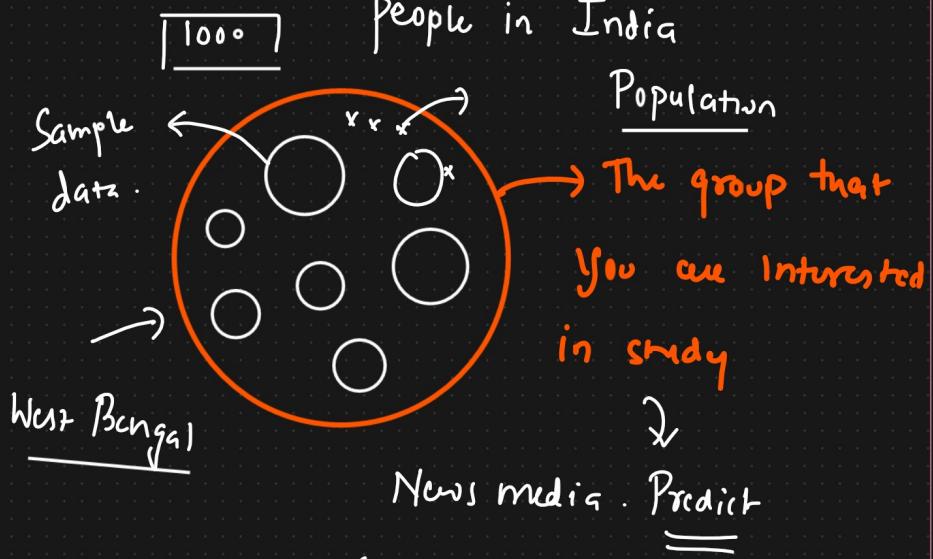
① Population And Sample

Surveys → Sampling.

EXIT POLL Results

↓
Fails

Population (N)



Sample (n)

① Simple Random Sampling {Randomly} Selection

② Stratified Sampling → Stratified → Layers → Groups

Population (N)

Non overlapping groups

Random

Sampling

Male

Female



National Housing Bank

RBI

Sunny }

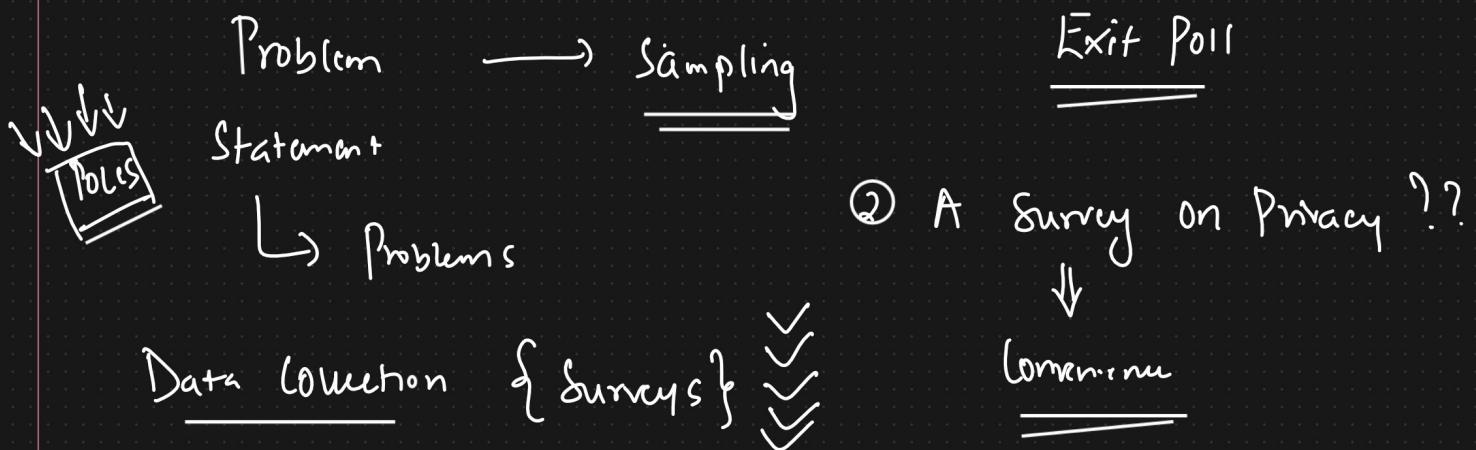
| > 18 years |

③ Systematic Sampling : Every n^{th} individual from the population (N) is placed in the sample (n)

Mail → 6th person → A specific survey }

④ Convenience Sampling = { Voluntary Response Sampling }

Survey → 1000 people → NO → Interested → fill ↴



① Statistics

② Data

③ Types of Statistics { 53 min }

④ Population & Sample

⑤ Sampling Technique

⑥ Variable : Variable is a property that can take on many values.

of 24,

25

26

27

100 }

Eg: Age = 23

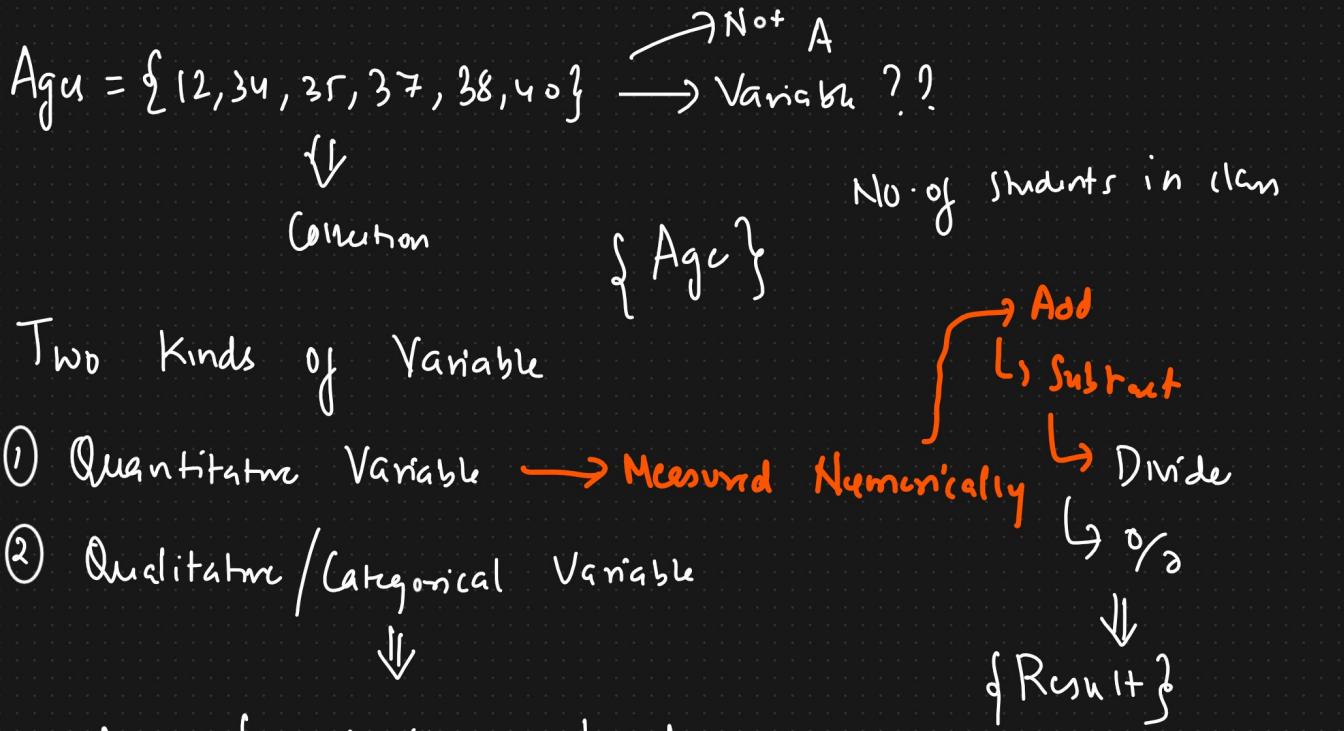
Gender of M

Weight = 180

F }

170

100 }



Allows for classification based on
Some characteristics

Gender $\{ M, F \}$

Categorical

1, 2, 3, 4, 5

Department $\{ \}$

Eg: No. of bank account, No. of

④ Quantitative Variables

\rightarrow Discrete variable Children

\rightarrow Continuous variable

\Downarrow

Eg: 2.5 inches, 6.2 feet, 172.5 pounds

Song length ??

Married Not married Widowed.

Marital Status ?? \rightarrow

Histograms \div frequency chart

$\{ 12, 24, 24, 26, 28, 34, 36, 38, 55, 60, 62, 75, 85, 89, 90,$

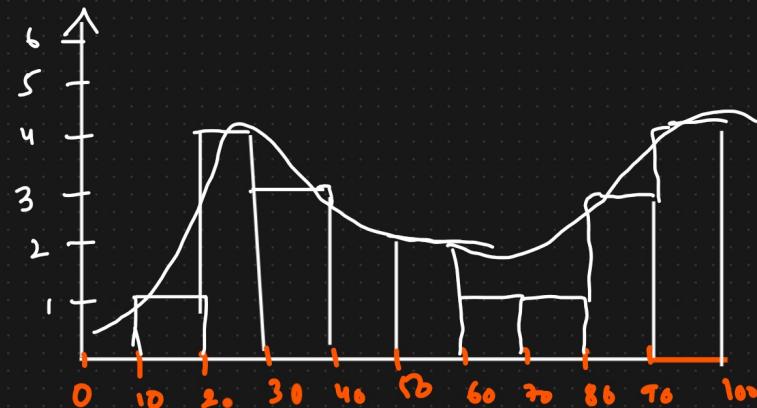
frequency

$92, 93, 94, 100 \}$

Histogram

Prb

bins = 10



Smoothed

Histograms

$\downarrow \downarrow \downarrow$
Pdf

Descriptive Statistics

① Measure of Central Tendency \rightarrow Population (N)
 \rightarrow Sample (n)

② Measure of Variance \rightarrow Population
 \rightarrow Sample

③ Arithmetic Mean: {Average} \rightarrow outlier

$x = \{1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 100\}$ \rightarrow Distribution

$$\text{Mean} : \sum_{i=1}^N x_i$$

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

$$= \frac{1+1+2+2+2+3+3+4+5+5+100}{11}$$

$$\left. \begin{array}{l} 2.8 \\ \downarrow \\ 1163 \end{array} \right\}$$

$$= \underline{\underline{2.8}} \quad \frac{128}{11} = 11.6 \overline{3}$$

① Median → Outliers → Stop $\{IQR\}$

$$x = \{2, 1, 4, 5, 3, 2\} \Rightarrow 1, 2, \boxed{2, 3}, 4, 5 \quad \text{Even}$$

\Downarrow Sort

$$\{1, 2, 2, \boxed{3}, 4, 5, 100\} \quad \text{Median} = \boxed{3}$$

$$2.5 \approx 3$$

$$\{2, 1, 7, 5, 3, 2\} \Rightarrow 1, 2, 2, \boxed{3}, 5, 7, 100 \rightarrow 2.5$$

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

④ Mode ÷ $\{2, 1, 7, 5, 3, 2\} \rightarrow$ Max no. of occ of elements

$$\hookrightarrow \boxed{2}$$

Female $\{2, 2, 3, 3, 1, 5\} \rightarrow 2 \frac{4}{3}$

$$\Downarrow$$

Male

NAN Mode {Categorical variable}

Female

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

$$\mu \gg \bar{x}$$

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

or

$$\mu \approx \bar{x}$$

④ Measure of Dispersion

Data set A = {11, 11, 12, 12, 12, 13, 13, 14, 14, 14} ✓

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = 12.6$$

Data set B = {1, 1, 2, 2, 3, 4, 4, 31, 32, 42} ✓

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = 12.6$$

① Variance { Spread of Data } { 1, 2, 3, 4, 5 }

Population Variance

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

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

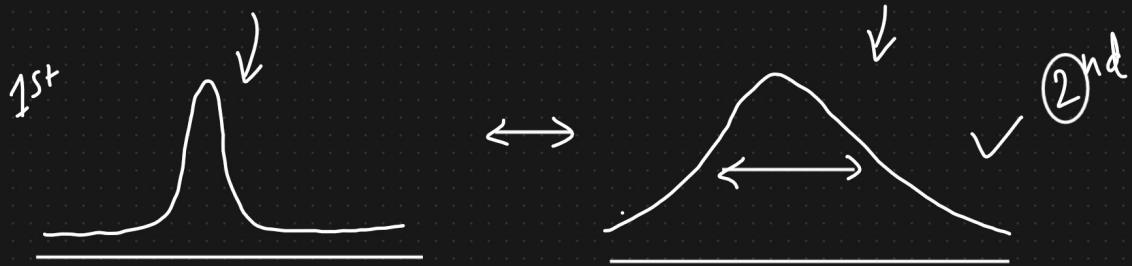
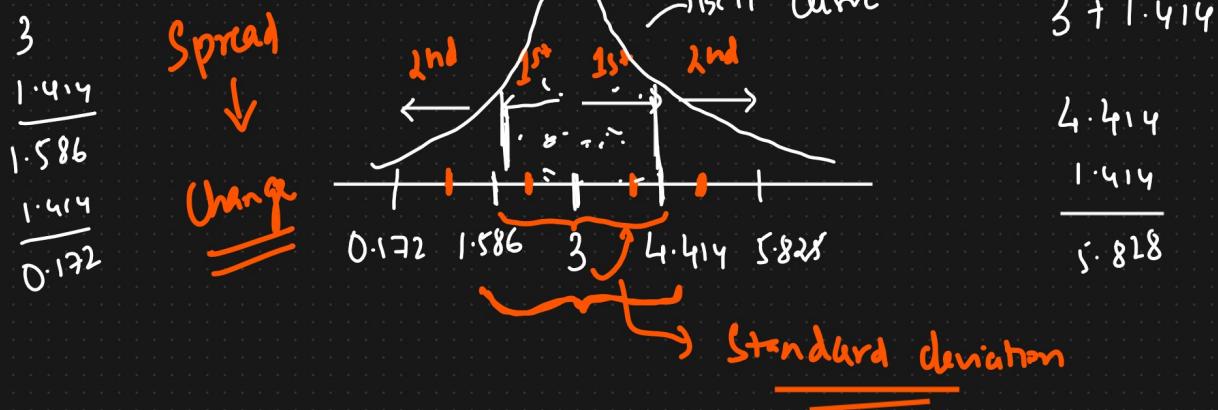
$$= \frac{4 + 1 + 0 + 1 + 4}{5} = \boxed{2}$$

$$\boxed{\mu = 3}$$

Population Standard Deviation

$$\sigma = \sqrt{\text{Variance}} = \sqrt{2} = 1.414$$

Variance



Sample Variance

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

why??

$$s = \sqrt{s^2}$$