Basic Statistics

1. population and Samples.

Research Question: Stress levels in teenagers.

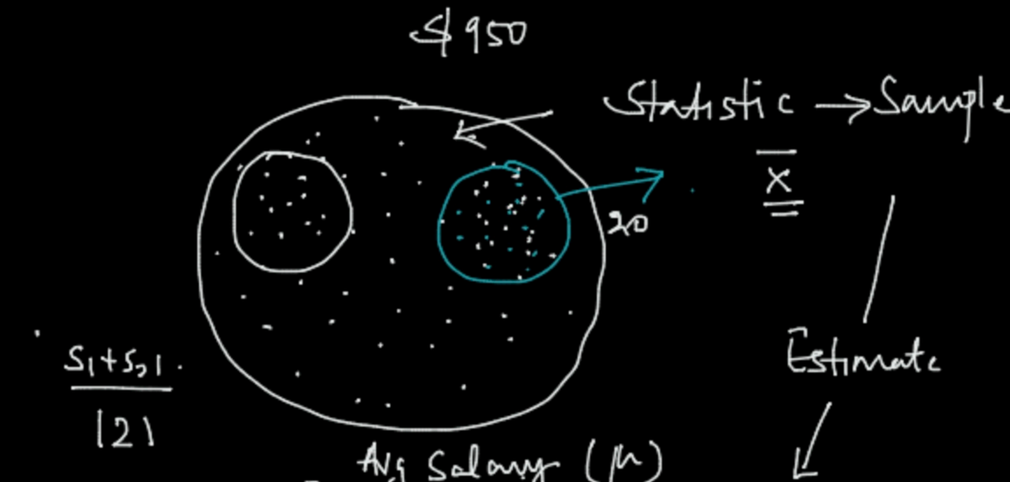
Population: All teenagers in the world

Sample: 200 teenagers.

Quality of biscuits made in a factory.

Population: All biscuits made in the factory

Sample: 100 biscuits



"Statistic" serves as an estimate for the "Parameter"

Good Sample:

1. Representative → Represent the entire population
2. Random → Every datapoint should have equal chance of being selected

Zoom Workplace Meeting View Edit Window Help

Food Habits of Indians

Population: All Indians

Sample: 1000 → Chennai.

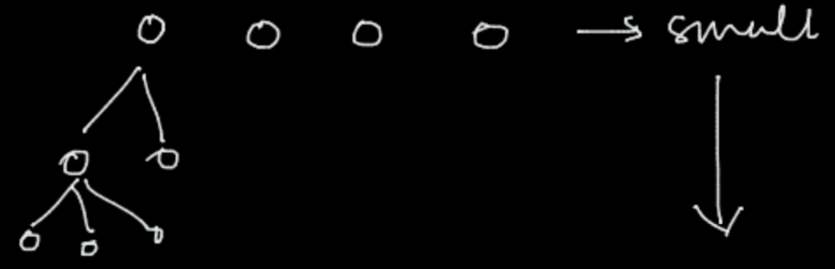
N S E W C → Strata

200 200 200 200 200

1. Simple Random Sampling (SRS)
   - Blind folded selection
   - When no info abt the population

Estonia → SRS

2. Stratified Sampling
   - Some info avlbl abt the population



3. Snowball Sampling
   - Sensitive info
   -

2 / 2

Types of Statistics

Descriptive Statistics

Measures of

1. Frequency ✓
   — Counts, Frequency / proportions
2. Central tendency — Most likely Value
   — Mean, Median, Mode
3. Dispersion / spread.
   — Range, Standard deviation / Variance
4. position
   — Percentiles, Deciles, quartiles
5. Shapes
   — Skewness, Kurtosis

Inferential Statistics

— Draw some conclusions
— Hypothesis testing

	Age	Course	Gender	Country
1			M	—
2			M	—
3			F	—

n = 100 Avg.

10:45 am

1. Arithmetic Mean

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

$n = 5$

Median ( )

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + x_4 + x_5)$$

Marks (x)	
x_1	45
x_2	70
x_3	54
x_4	66
x_5	75
	310

$\bar{x} =$ 62

outliers affect the mean

↓

Extreme Values

8 – DW — 81

1 – CEO — ~~45~~ 100

1 – C70 — 80

2

93
90
97
95

→ 5

Avg ⇒ 93.75 ✓

Median → 93.5 ✓

380 / 5 = 76% →

64 + 100 + 80 → ~~15 lakh~~ 244 / 10

→ 25L

2. Median — unaffected by outliers

— arrange no. in order

— take central value

5 90 93 95 97

5 75 90 93 95 97

91.5

50% 50%

3. Mode → Most Frequent Value

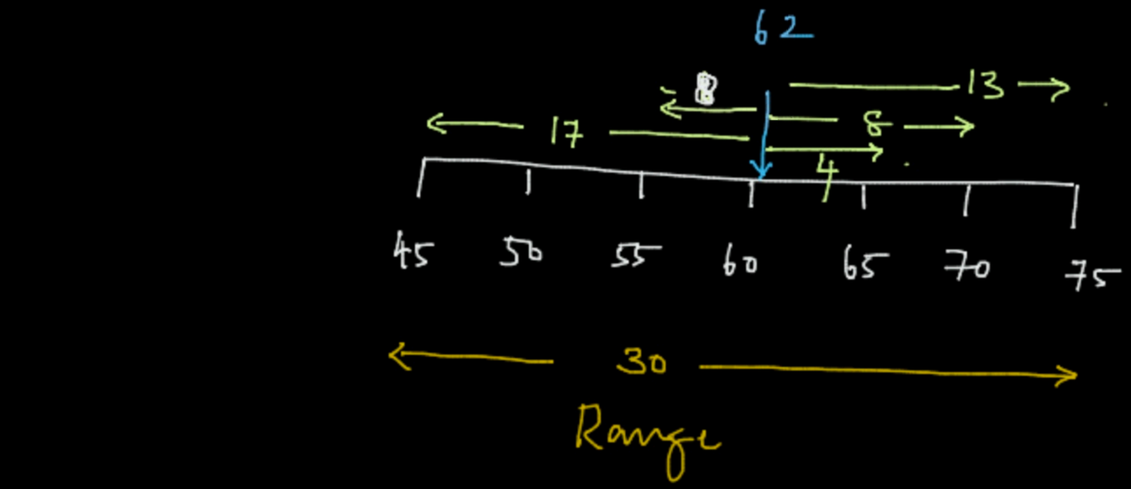
70 M ←
30 F

As	30
Am	30
En	30
La	15

1, 2, 2, 3, 4, 5 → 2.

160 → 37

$2/5 = 0.4$ X



Marks (x)	$(x-\bar{x})$	$(x-\bar{x})^2$
45	-17	289
70	+8	64
54	-8	64
66	+4	16
75	+13	169
310		602

$= 120.4$

$\bar{x} = 62$

Measures of dispersion / spread.

Class 1	Class 2
50	50
$\sigma = 5$	$\sigma = 20$

a. Range:

$= \text{Max} - \text{Min}$

$= 75 - 45$

$= 30$

(b) Variance (σ^2)

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

(c) Standard Deviation — on an avg how far are the datapoints from the central value.

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

$\sigma = 11$

Population

No. of Records N

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

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

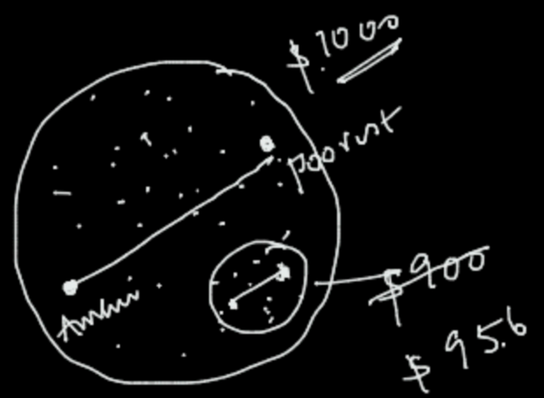
Sample

n.

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

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

↳ Bessel's Correction



Sample Variance underestimates population Variance

$n-1$ → Degrees of Freedom

→ the no. of datapoints that are free to vary given a Constraint

Constraint

Total = 100 ✓

$n = 10$

$df = 9$

Total = 9

x_1 — 2 ✓

x_2 — 4 ✓

x_3

P_1 — 0

P_2 — 13

P_3 :

P_9 — 20 → 85

P_{10} — x

Area	Bed	Price
1500	2	50 ✓
1800	2	60 ✓
2200	2	75 ✓
2500	2	83 ✓

↳ Var = 0

Variance → Information

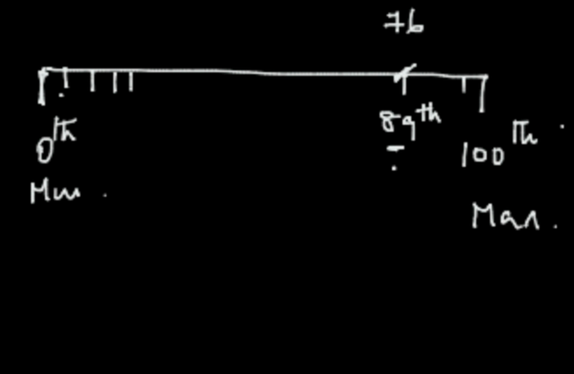
← Measures of position →

→ 65 %. → 90th percentile

85% → 40th percentile

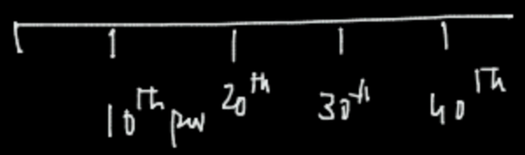
↓

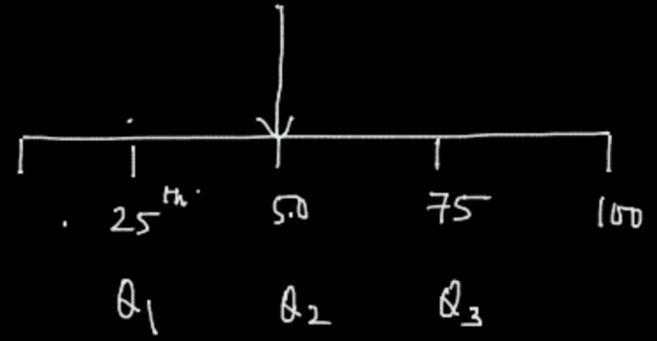
⋮



Quartiles → 25th percentile → Q_1

Deciles → 10 pw



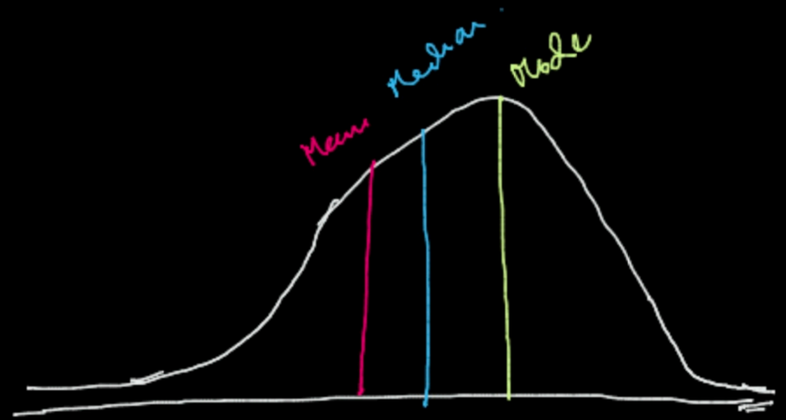


QuaNtile → Percentile
→ Decile
→ QuaRtile.

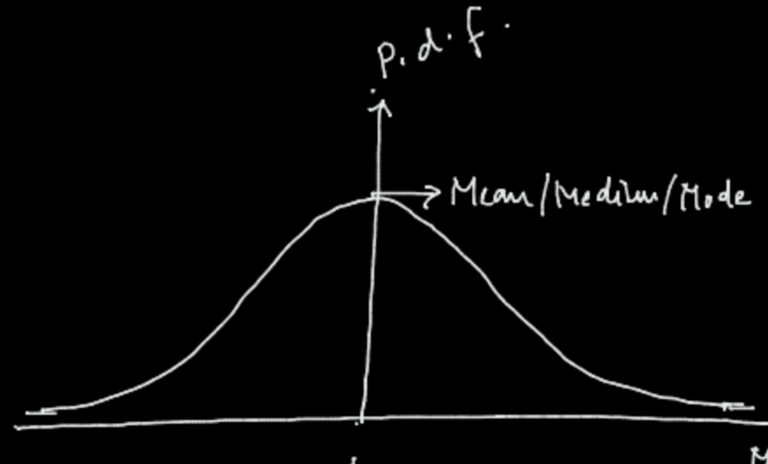
← Measures of Shape →

(a) Skewness

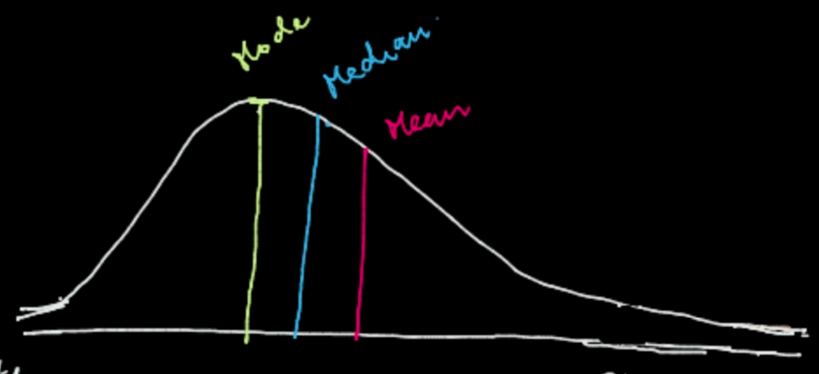
— How Symmetric is the data distr.



Left Skew (-ve)

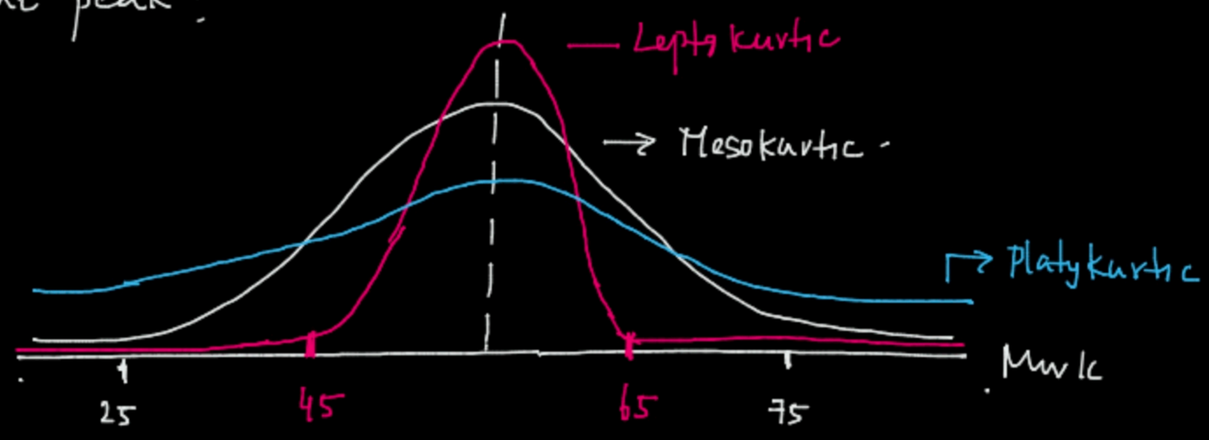


No Skew.



Right Skew (+ve)

(b) Kurtosis — How sharp is the peak?



Datatypes

Numeric

Discrete

— Counted

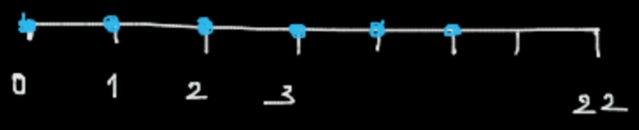
Eg: No. of students

No. of cars

No. of Children

— Fractions don't make sense

— Only certain values are possible.



— Finite prob. of getting specific value

Continuous

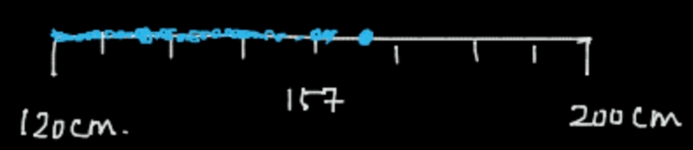
— Measured

Eg. Height, Wt.

150.7.5cm

85.3 kg

— Fractions are meaningful.



— Any value is possible within a range

— Prob. of finding a particular value = 0

Categorical

— Categorize the data.

Ex: Gender, Race, Religion

— only a label.

M/F. B/G.

0 | 1

← Probabilities →

Random Expts: Result changes

Sample space → [All possible outcomes of a random expt.]

Toss a coin → [H, T]

Roll a dice = [1, 2, 3, 4, 5, 6]

$$P(X) = \frac{\text{No. of Fav. events}}{\text{\# Total possible outcomes}}$$

$P(3) = 1/6$.

$P(7) = \frac{1}{2}$

Draw a card → [13 H, 13 C, 13 S, 13 D]

$P(\text{Heart}) \rightarrow \frac{13}{52} = 1/4$.

$P(X = 150) \rightarrow$ [120, ... ∞ ..... , 200]

$\rightarrow \frac{10}{\infty} \rightarrow 0$.

$P(151 < x < 152) \rightarrow$ Finite

Random Variable:

Variable used to hold the result of a Random Experiment.

- discrete Random Variable
- Continuous Random Variable

X → hold the result of - tossing a coin

X → prob. of Finding a person whose height = 150 cms.

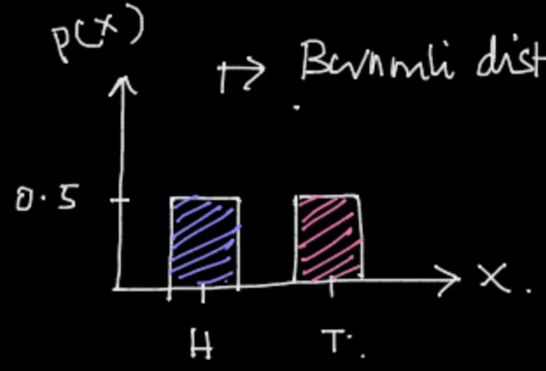
1. Bernauli
2. Binomial
3. Uniform

1. Normal Dist.
2. Z- dist (or) std Normal dist.
3. t- dist.

Probability distribution of a random Variable

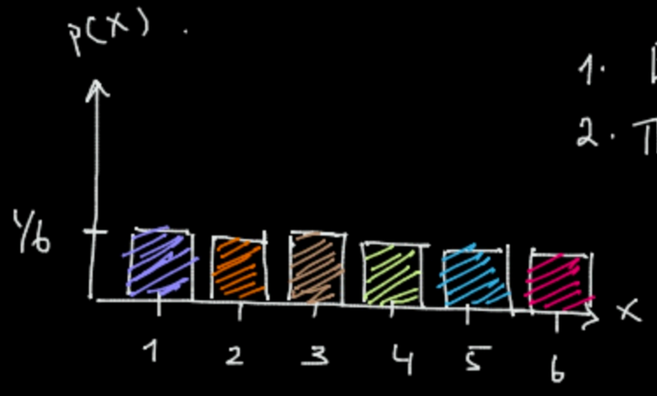
— The prob. associated with all possible values The random Variable can hold.

$X \rightarrow [H, T]$



↦ Bernoulli dist

$X \rightarrow [1, 2, 3, 4, 5, 6]$



(3) Bernoulli

1. Expt performed only once.
2. There should only be 2 possible outcomes.

1. Uniform Distribution →

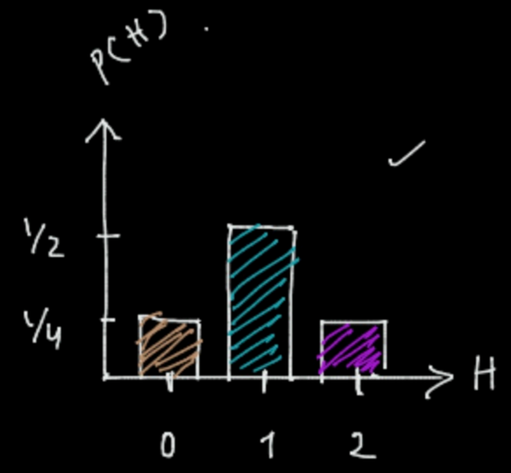
— Every outcome has equal prob. of occuring

2. Binomial Distribution

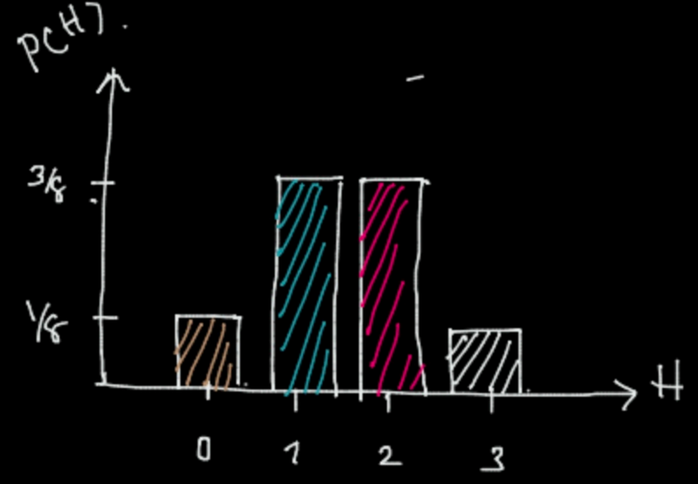
1. Expt should be repeated more than once.
2. There should only be 2 possible outcomes

Toss a coin twice

$$\rightarrow \begin{bmatrix} H & H \\ H & T \\ T & H \\ T & T \end{bmatrix}$$



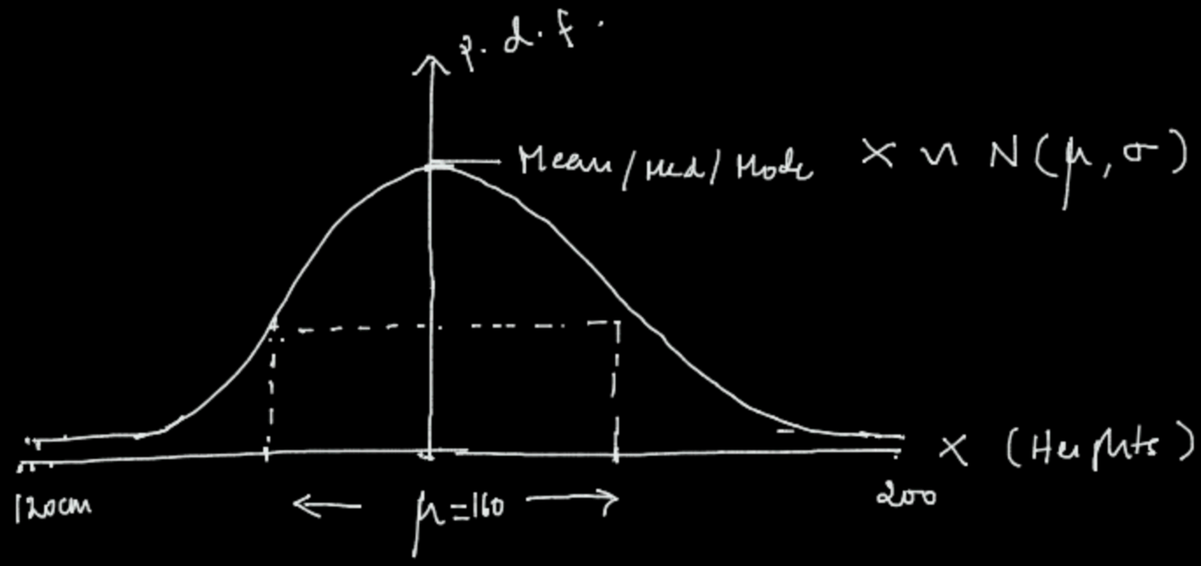
$$\begin{bmatrix} T & T & T \\ H & H & H \\ H & H & T \\ H & T & H \\ H & T & T \\ T & H & H \\ T & H & T \\ T & T & H \end{bmatrix}$$



(a) **Normal distribution**

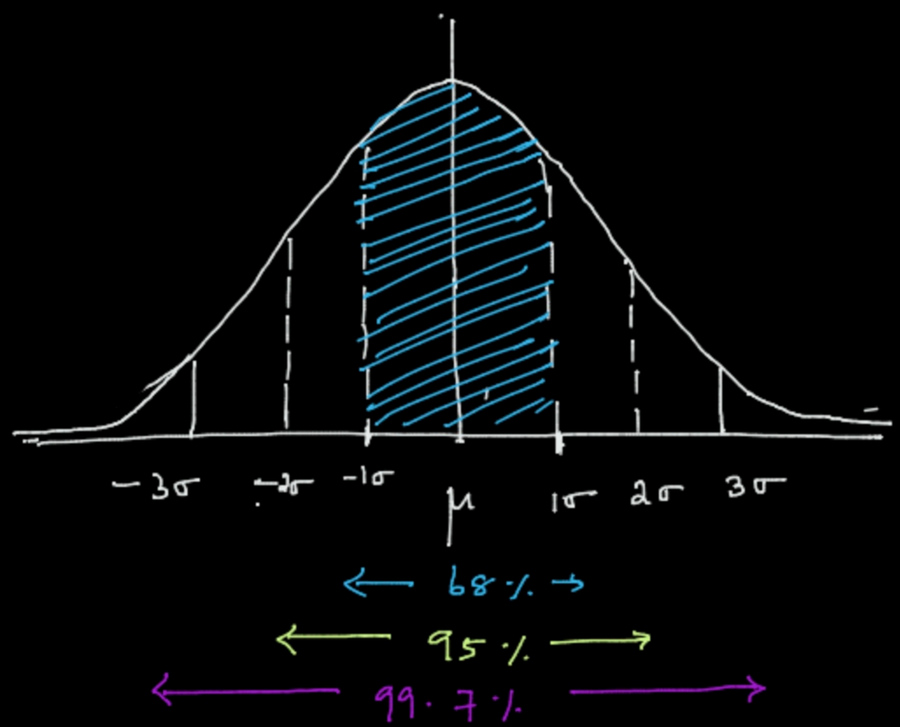
← Continuous distributions →

1. Bell shaped
2. Symmetric
3. $p.d.f = \frac{1}{\sigma\sqrt{2\pi}}\left[\exp\left[\frac{x-\mu}{\sigma}\right]^2\right]$



4. Mean = Median = Mode

5. Empirical Rule

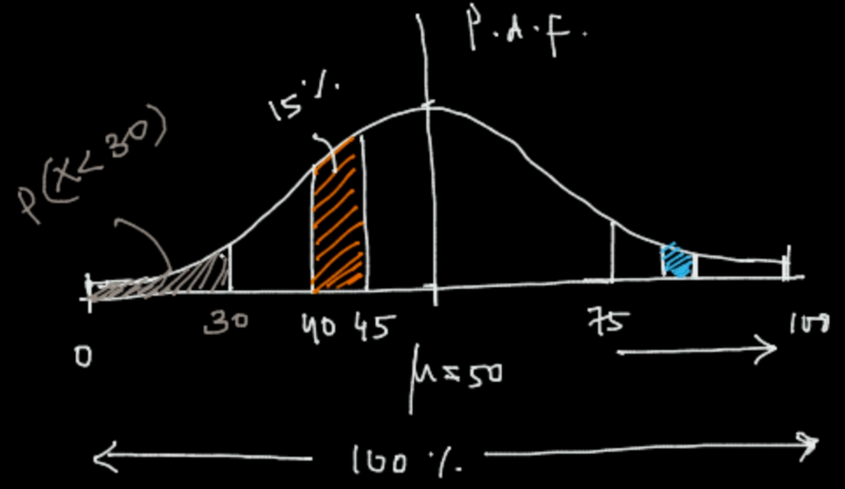


$\bar{(x)}$ Heights

$\mu = 160$

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

$\sigma = 5$



P.d.f

Area under the curve ⇒ Probability

0 → 100.

Total Area = 100 %.

1. P(40 < x < 45) → 15%.

15 %. of students have marks both 40 & 45

P(X<30)

x > 75.

N(60,5)

μ = 60

σ = 5

(x)

→ | 45 | 70 | 65 | ⋮ | ⋮ |

μ = 55

σ = 7

68 %. → 55 & 65

95 %. → 50 & 70

99 % → 45 & 75

Standardize.

$$Z_x = \left(\frac{X - \mu}{\sigma}\right)$$

Hawls x	$x-\mu$	$Z_x = \frac{x-\mu}{\sigma}$
45	− 17	− 1.54
70	+ 8	+ 0.72
54	− 8	− 0.72
66	+ 4	+ 0.36
75	+ 13	+ 1.18
	$\mu = 62$	$\mu = 0$

$\sigma = 11$.

→

Z_x
− 1.54
0.72
− 0.72
0.36
+ 1.18

Z-distribution / Std. Normal distribution

$\mu - 3\sigma$, $\mu + 3\sigma$ → 99.7 %.

$[-3, +3]$ → 99.7 %.

Scaling → Bring all features to same scale

1. Normal dist.

$\mu = 0$; $\sigma = 1$.