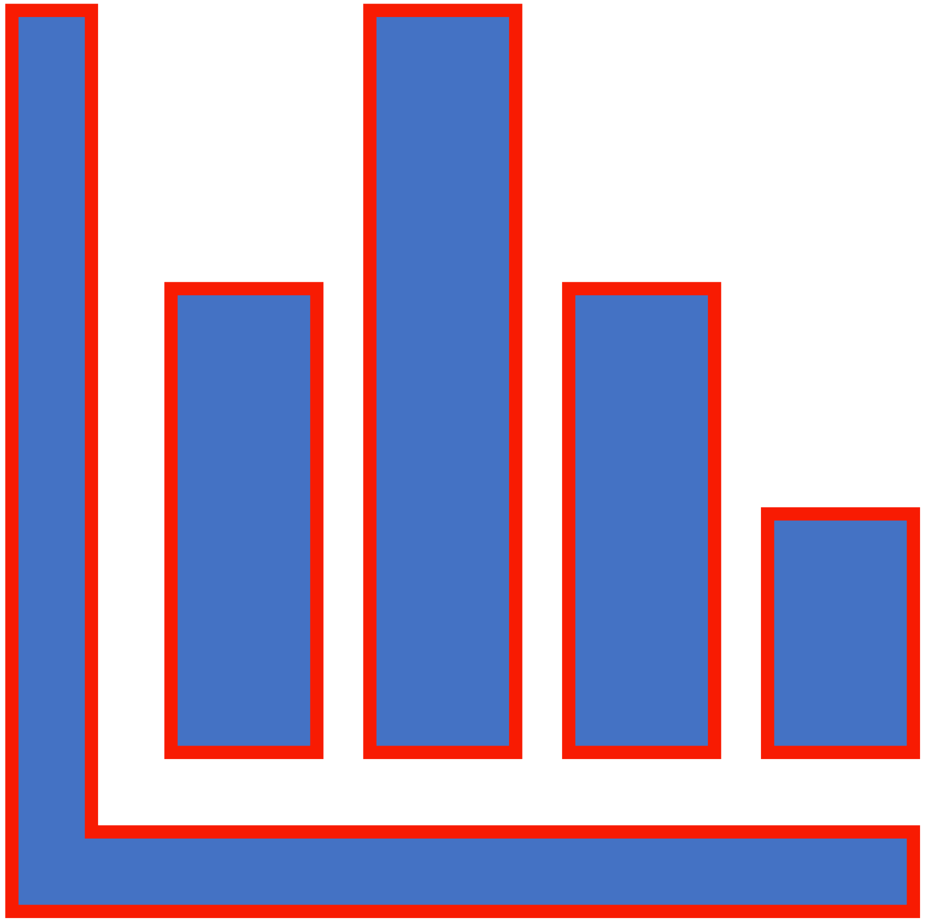
Statistics for Data Science



Introduction to Statistics

Science of learning from data.

Methodical data collection.

Employ correct data analysis.

Presenting analysis effectively.

Importance

- Helps in avoiding getting biased samples
- Prevent over-generalization
- Wrong causality.
- Identify Incorrect **Analysis.**
- Can be applied **to any domain**

Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write: HG WELLS(1903)

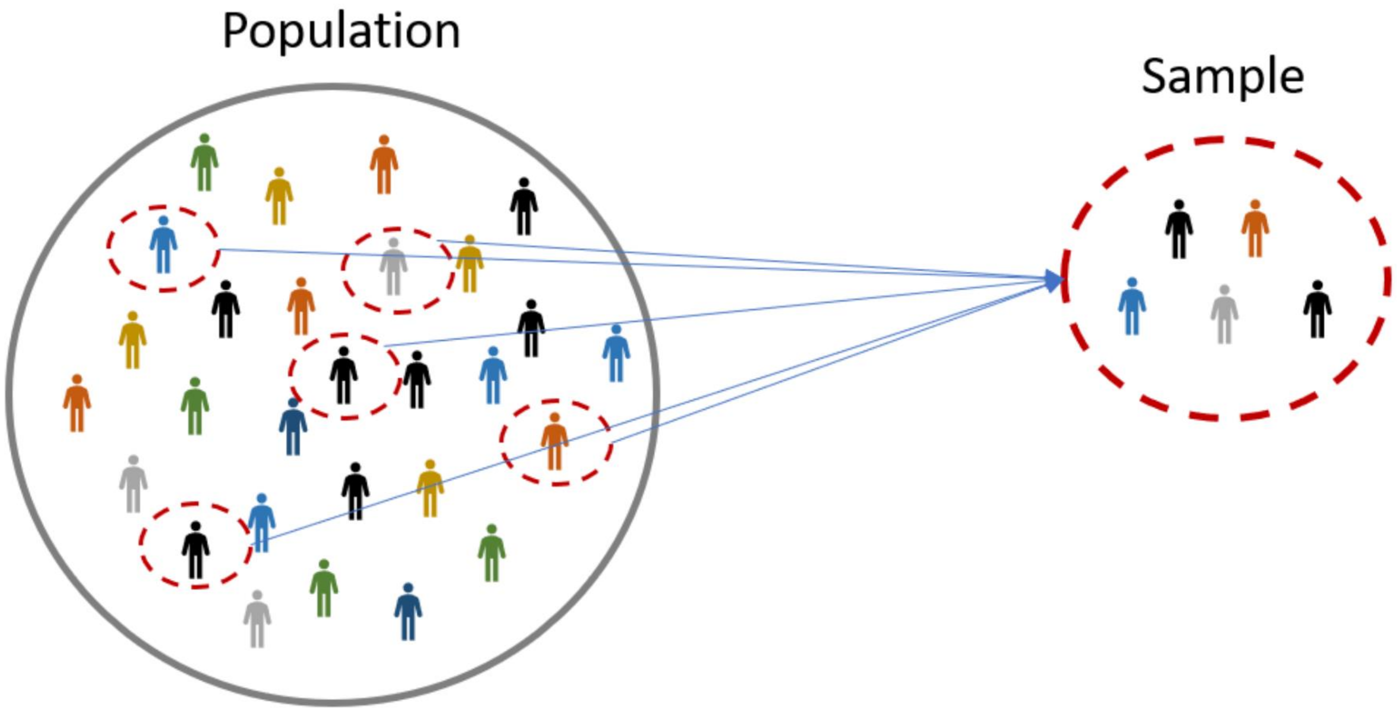
Stages of Statistical Analysis

- Data Gathering
- Data Understanding
- Analysis and Interpretation
- Data Presentation

Statistical Analysis provides a way to extract information from data on Objective basis rather than relying on personal Experience)

1. Data Gathering: Extracting Data

Population and Sample



Samples are used to make inferences about **populations**. Samples are easier to collect data from because they are practical, cost-effective, convenient and manageable.

Parameter vs Statistic

Parameter

Statistic

A **parameter** is a number describing a whole population (e.g., population mean), while a **statistic** is a number describing a sample (e.g., sample mean).

Parameter vs Statistic

Sample statistic	Population parameter
Proportion of 2000 randomly sampled participants that support the Farm Laws bill.	Proportion of all Indian residents that support the Farm Laws bill.
Median income of 500 Data Scientists in Chennai and Delhi.	Median income of all Data Scientists in India.
Standard deviation of weights of apples from one farm.	Standard deviation of weights of all apples in a region.
Mean screen time of 3000 high school students in India.	Mean screen time of all high school students in India.

Parameter vs Statistic

Parameter

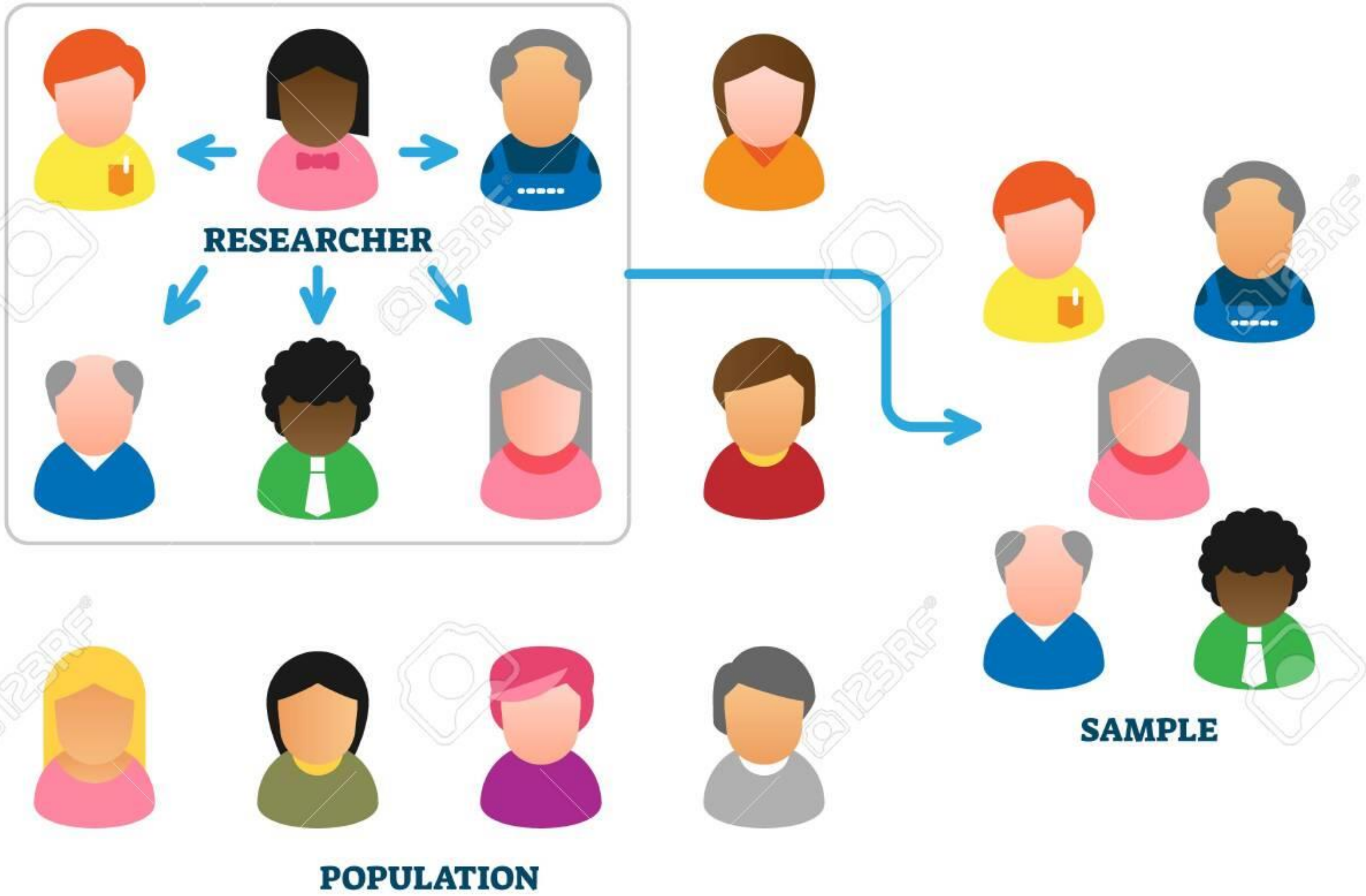
Statistic

A **parameter** is a number describing a whole population (e.g., population mean), while a **statistic** is a number describing a sample (e.g., sample mean).

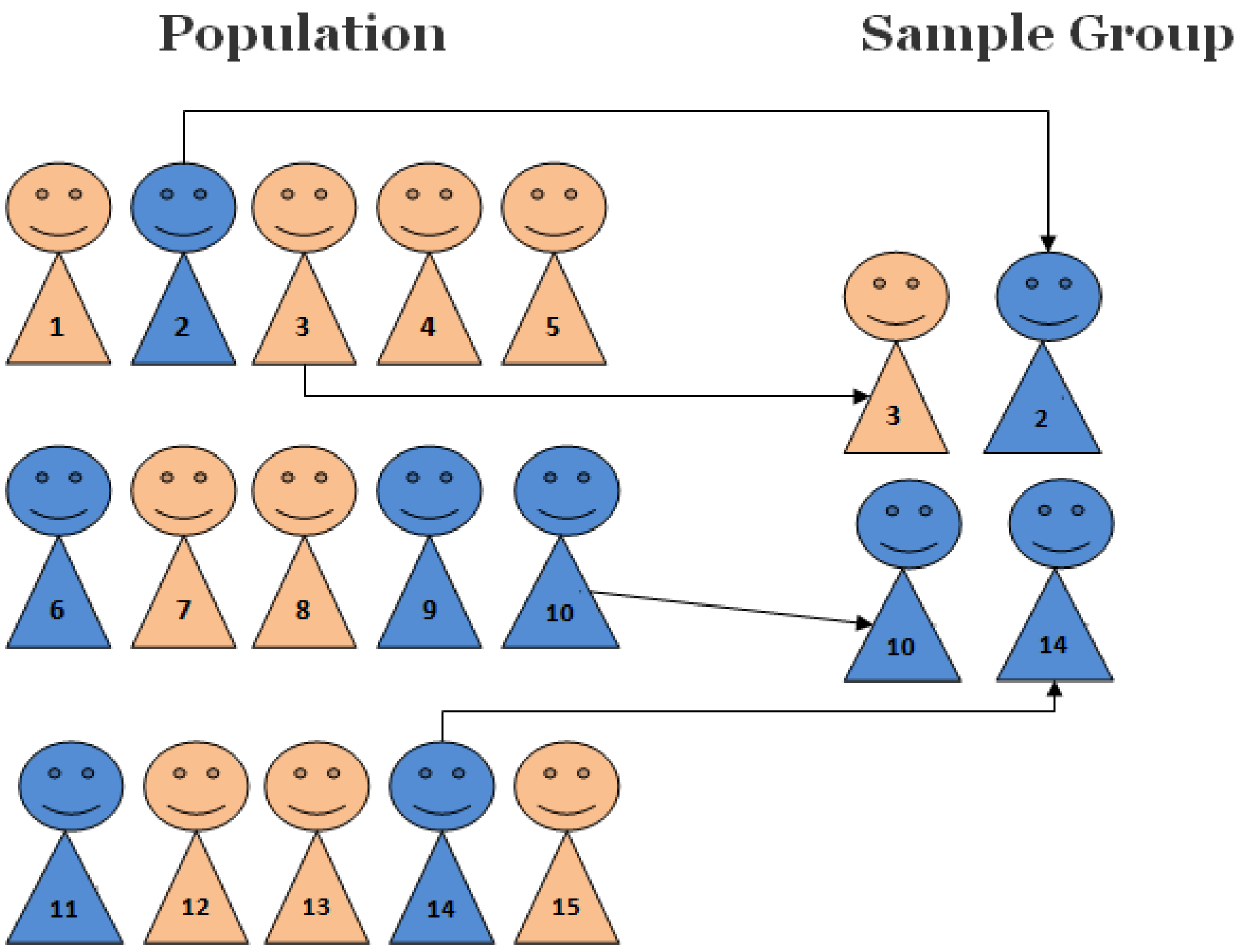
Data Gathering: Sampling Techniques

- Convenient Sampling
- Random Sampling
- Systematic Random Sampling
- Stratified Sampling
- Cluster Sampling

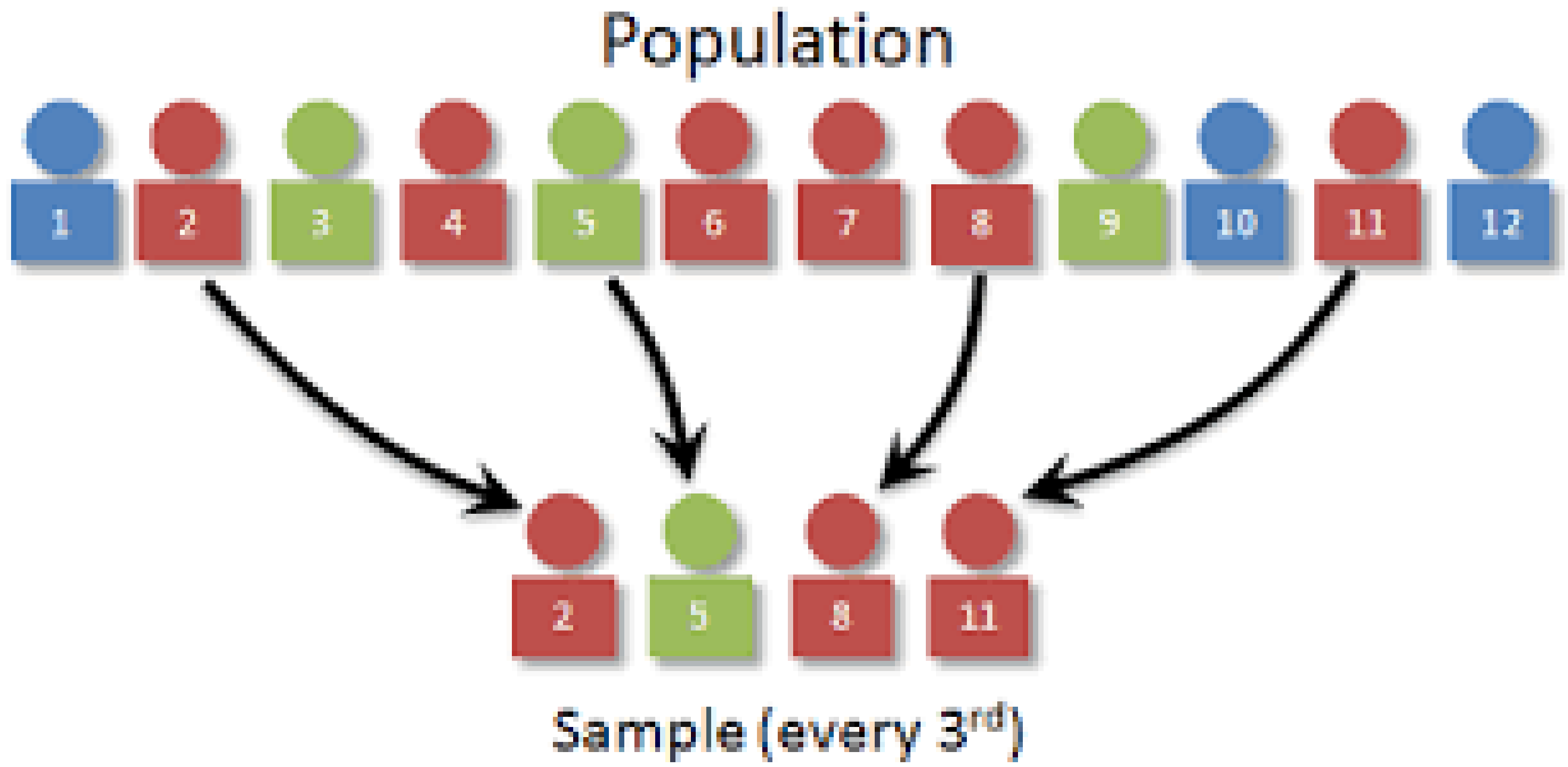
CONVENIENCE SAMPLING

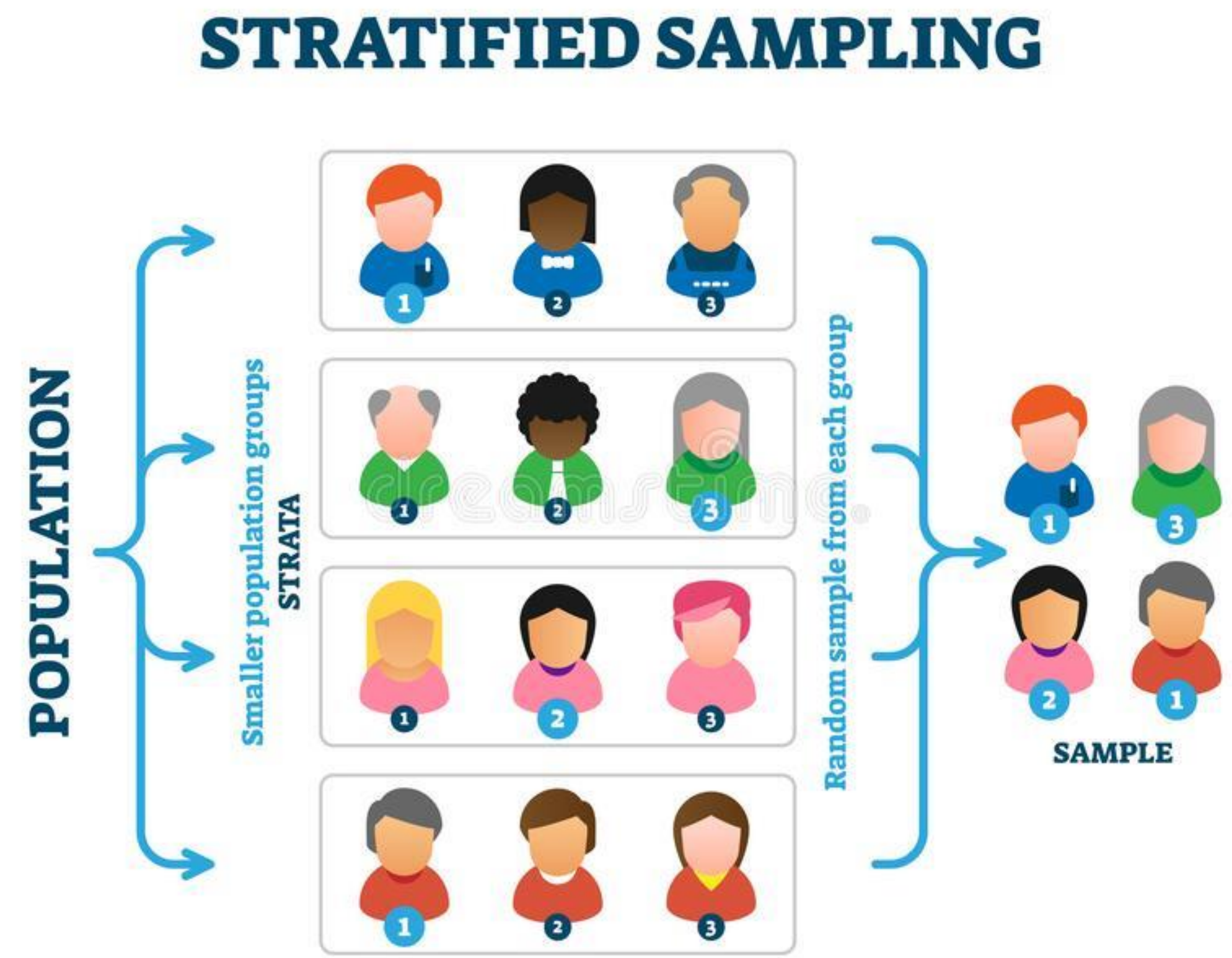


Random Sampling

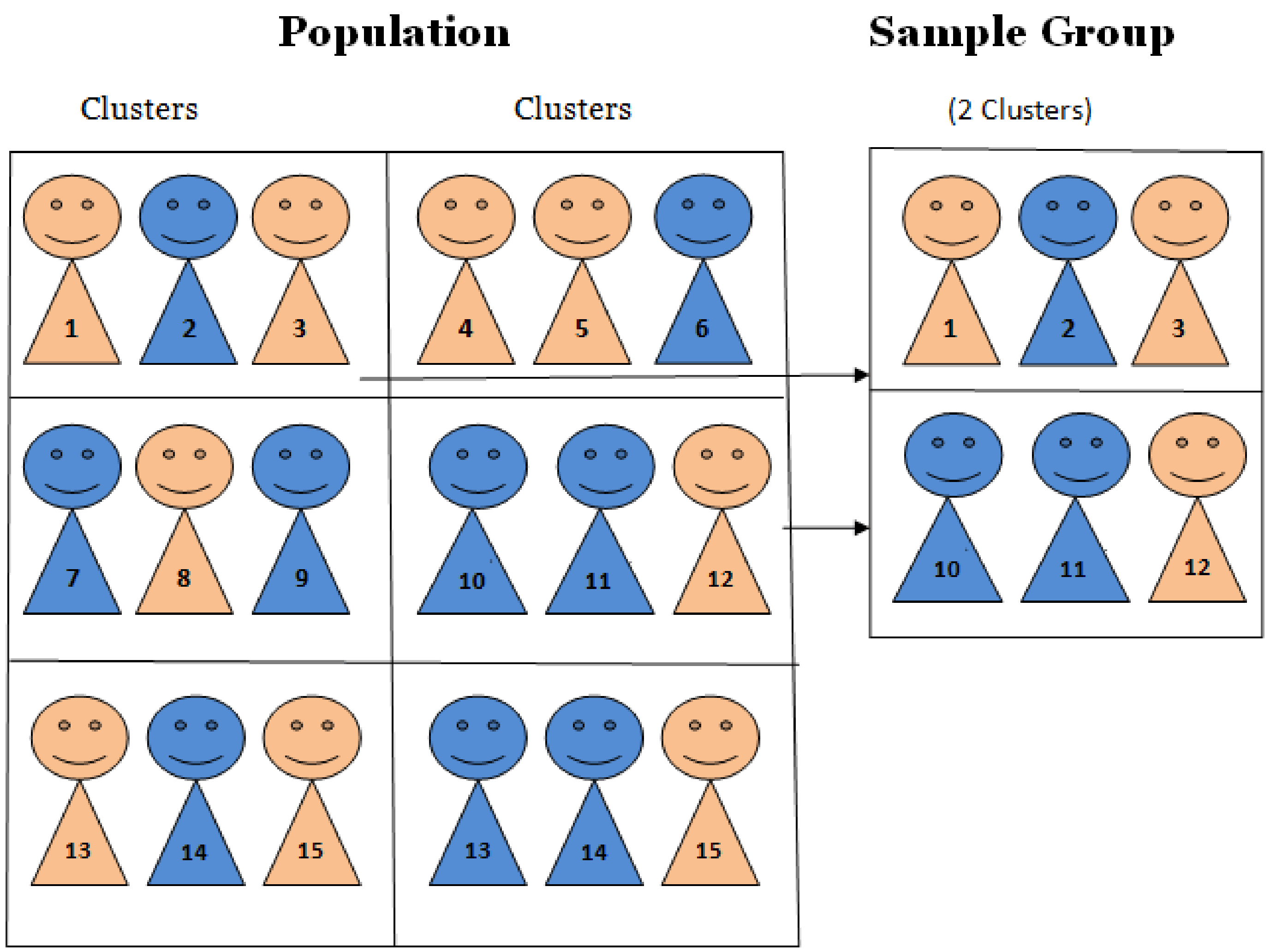


Systematic Random Sampling




STRATIFIED SAMPLING
POPULATION
Smaller population groups
STRATA
Random sample from each group
SAMPLE

Cluster Sampling



2. Data Understanding: Variables and Entities

Data Understanding: Variables

Dependent

Independent

number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	dept	salary
2	157	3	0	1	0	sales	low
5	262	6	0	1	0	sales	medium
7	272	4	0	1	0	sales	medium
5	223	5	0	1	0	sales	low
2	159	3	0	1	0	sales	low

Variables: represents
a characteristic of an Entity

- Explanatory (predictor or independent)
- Response (outcome or dependent)

number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	dept	salary
2	157	3	0	1	0	sales	low
5	262	6	0	1	0	sales	medium
7	272	4	0	1	0	sales	medium
5	223	5	0	1	0	sales	low
2	159	3	0	1	0	sales	low

Variables: Quantitative vs Qualitative

- Quantitative - Numerical data. Eg. weight, temperature, number_project
- Qualitative - Non-numerical data. Eg. dept, salary

Types of Quantitative Variables

Continuous - Numerical values.

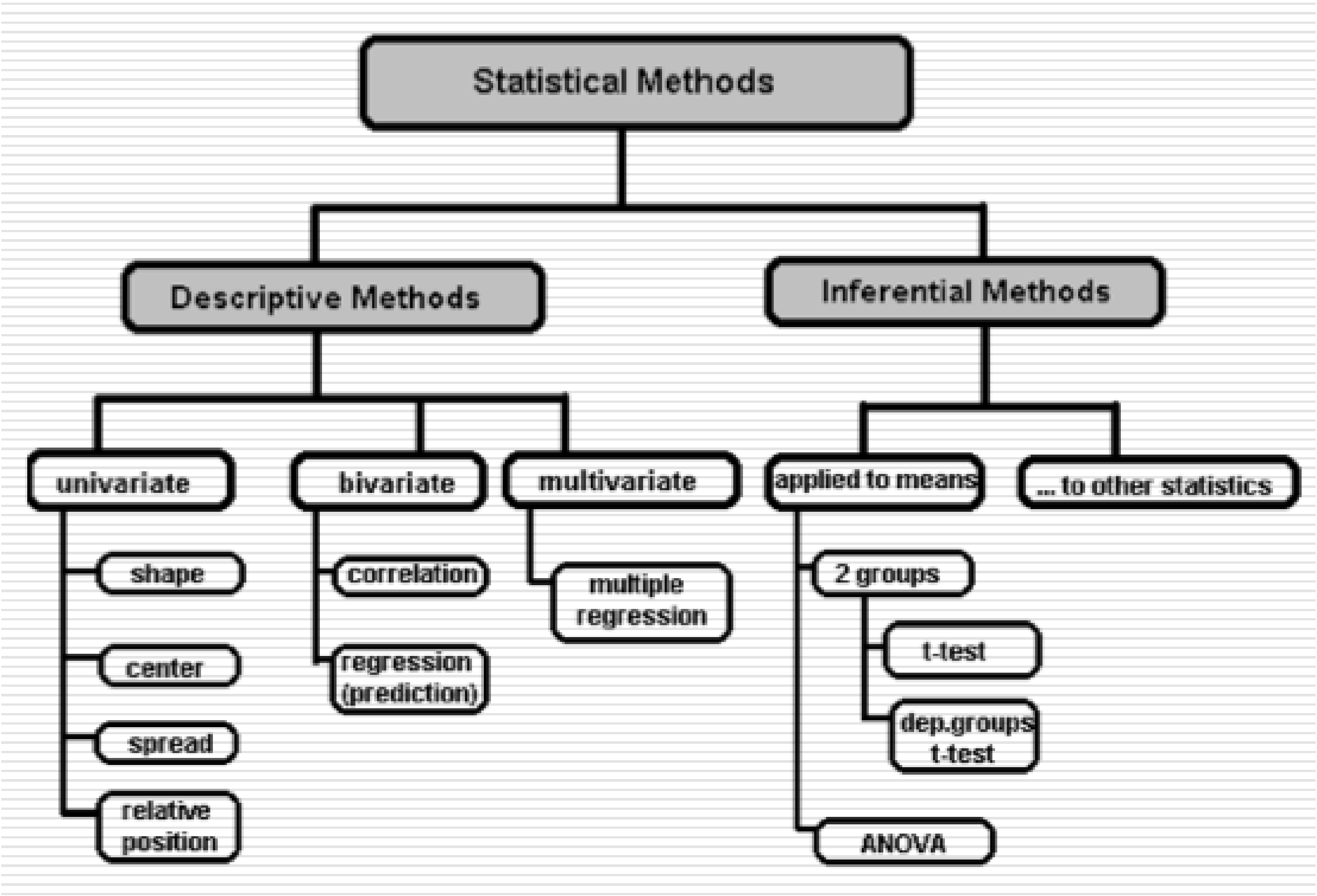
Discrete - Count of presence a Characteristics

Types of Qualitative/Categorical Variables

Nominal: Ex - **dept (sales, RD etc.)**

Ordinal: Ex. Salary**(low, medium, high**), Binary(Yes , No)

3. Data Analysis: Describing Data through Statistics



Taxonomy of Statistics

Types of Statistical Analysis

INFERENTIAL STATISTICS - DRAW CONCLUSIONS FROM THE SAMPLE & GENERALIZE FOR ENTIRE POPULATION. COMMON TOOLS - HYPOTHESIS TESTING, CONFIDENCE INTERVALS, REGRESSION ANALYSIS

DESCRIPTIVE STATISTICS - DESCRIBES DATA. COMMON TOOLS - CENTRAL TENDENCY, DATA DISTRIBUTION, SKEWNESS

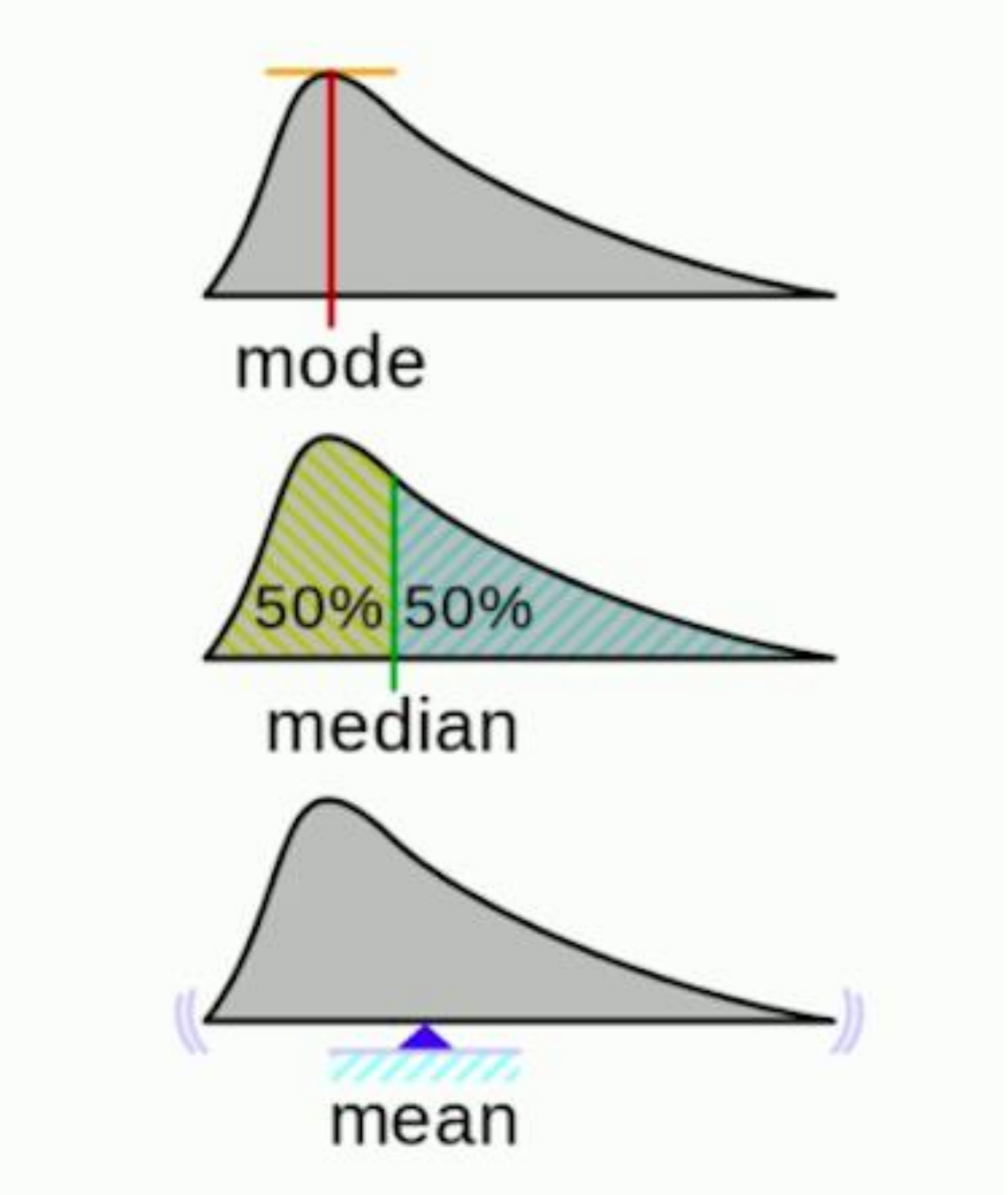
Measure of Central Tendency

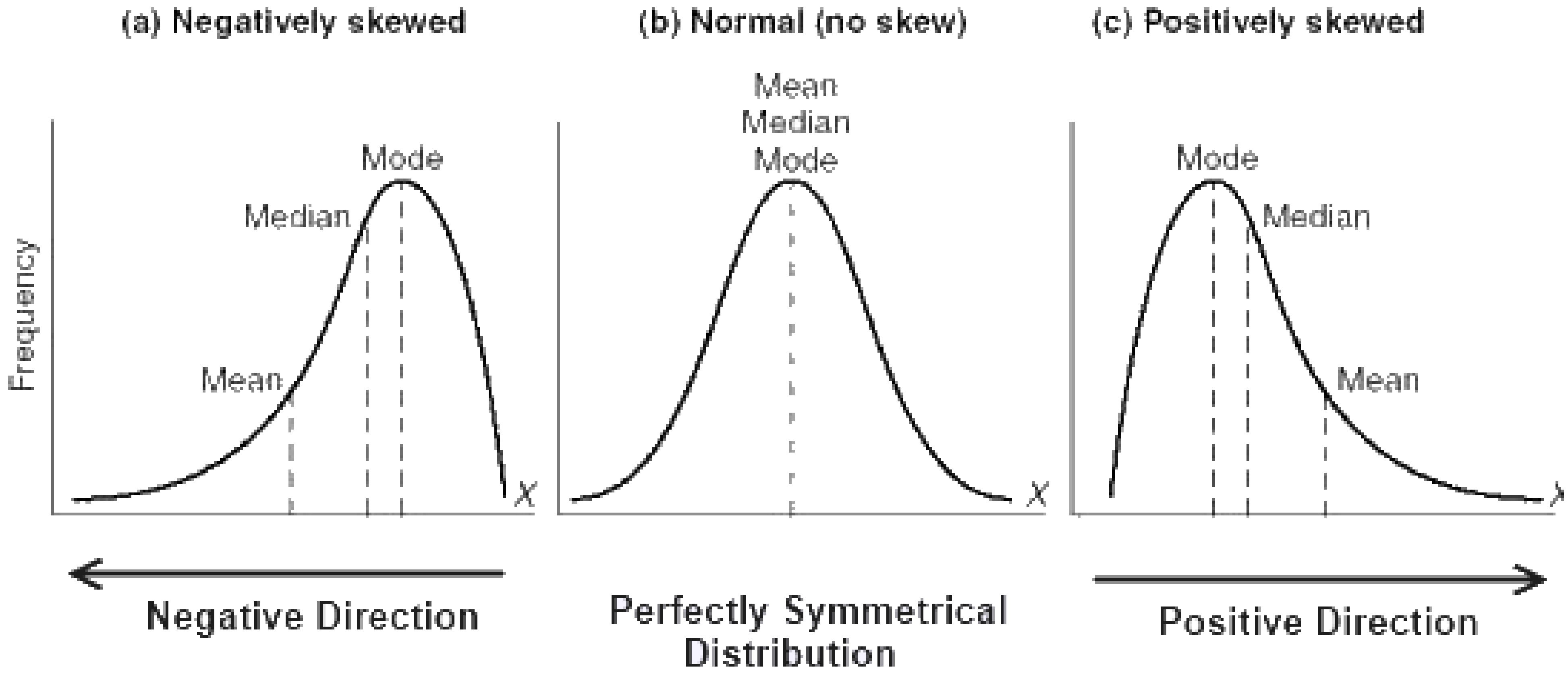
Mean - Average of data, suited for continuous data with no outliers

Median **- Middle value of ordered data, suited for continuous data with** outliers

Mode **- Most occuring data, suited for categorical data (both nominal and ordinal)**

Mode
Vs
Median
Vs
Mean





Mean Vs Median Vs Mode

QnA
Quiz

Session 2

Measure of Variance

RANGE

INTERQUARTILE RANGE

VARIANCE

STANDARD DEVIATION

Range: In statistics, the range of a set of data is the difference between the largest and smallest values.

AGES OF STUDENTS

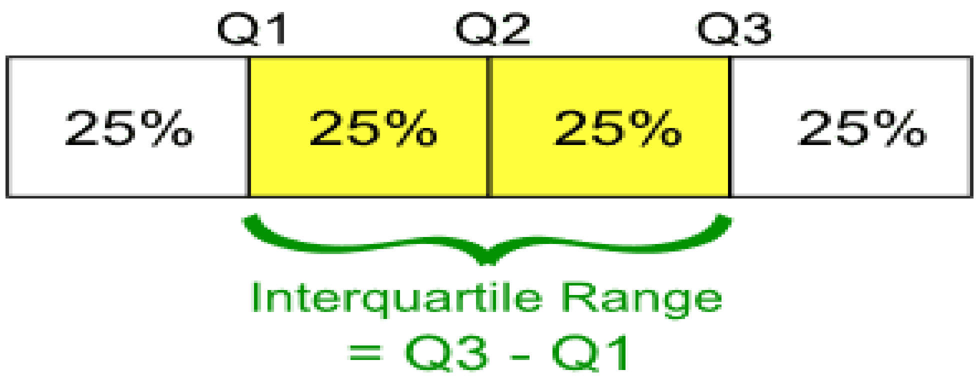
13,13,14,14,14,15,15,15,15,16,16,16

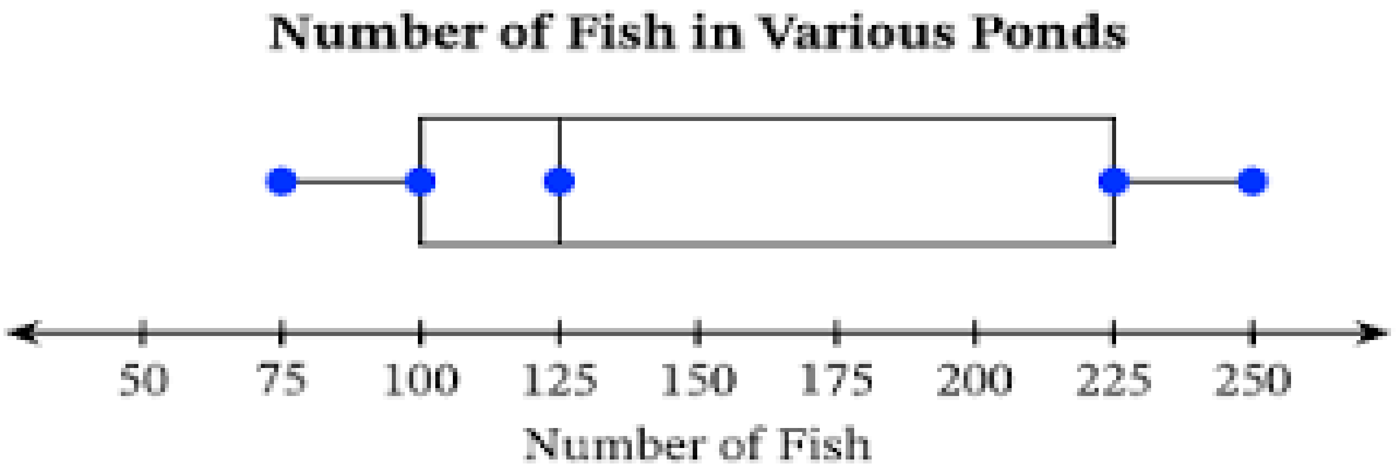
Range = highest − lowest

= 16 − 13

Range = 3

Interquartile Range: The **interquartile range** is a measure of where the "middle fifty" is in a data set.





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

Observation(x)	μ	x- μ	(x- μ)^2
105	101	4	16
100		-1	1
102		1	1
95		-6	36
100		-1	1
98		-3	9
107		6	36

Variance: The Variance is defined as the average of the squared differences from the Mean

Standard Deviation: it is the **square root** of the **Variance**

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

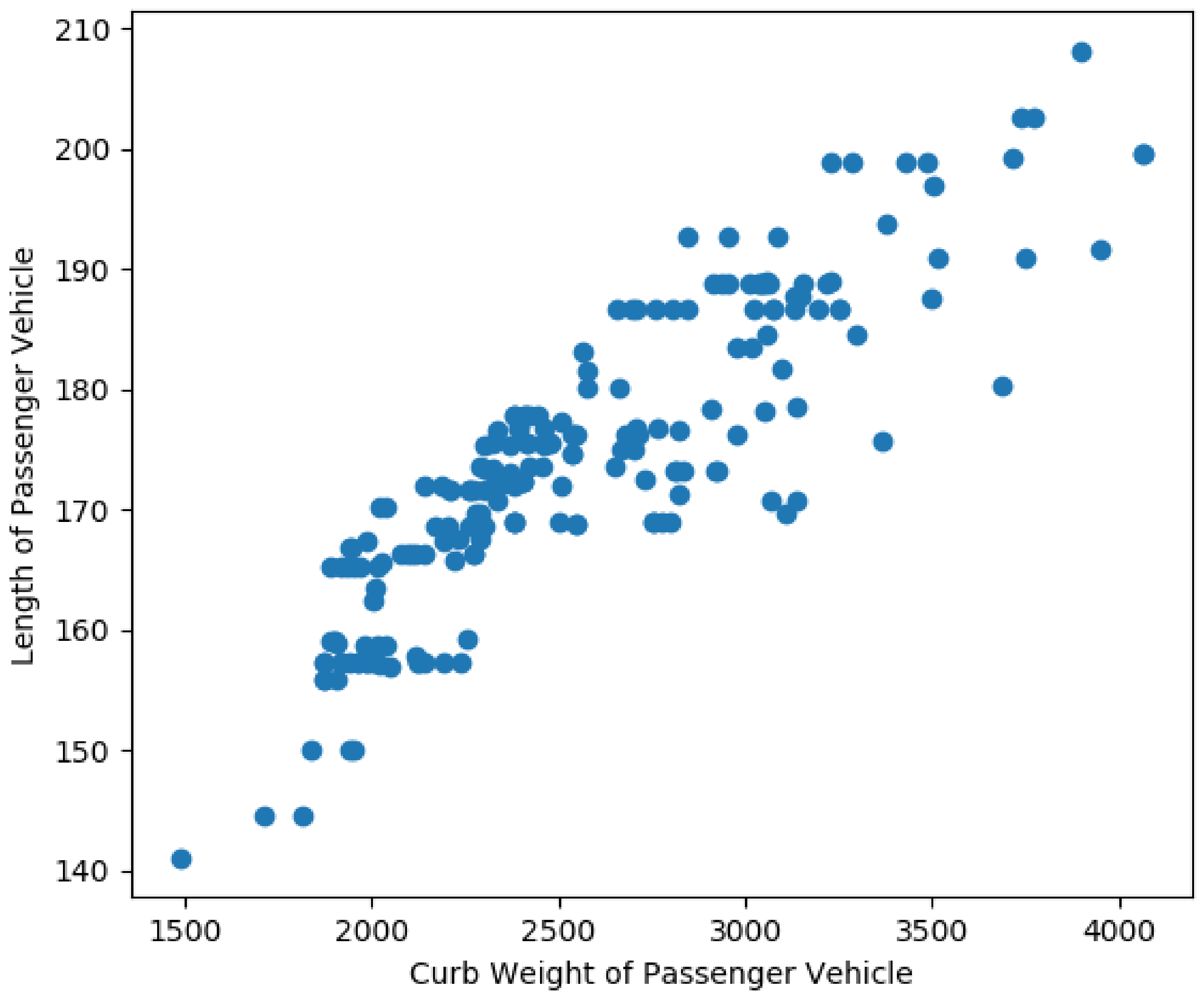
Variance and Standard Deviation: Comparative Analysis

BASIS FOR COMPARISON	VARIANCE	STANDARD DEVIATION
Meaning	Variance is a numerical value that describes the variability of observations from its arithmetic mean.	Standard deviation is a measure of dispersion of observations within a data set.
What is it?	It is the average of squared deviations.	It is the root mean square deviation.
Labelled as	Sigma-squared (σ^2)	Sigma (σ)
Expressed in	Squared units	Same units as the values in the set of data.
Indicates	How far individuals in a group are spread out.	How much observations of a data set differs from its mean.

Quiz

Q: If all the observations in a data set are identical, then what will be the value of Standard Deviation and Variance?

Relationship between Variables

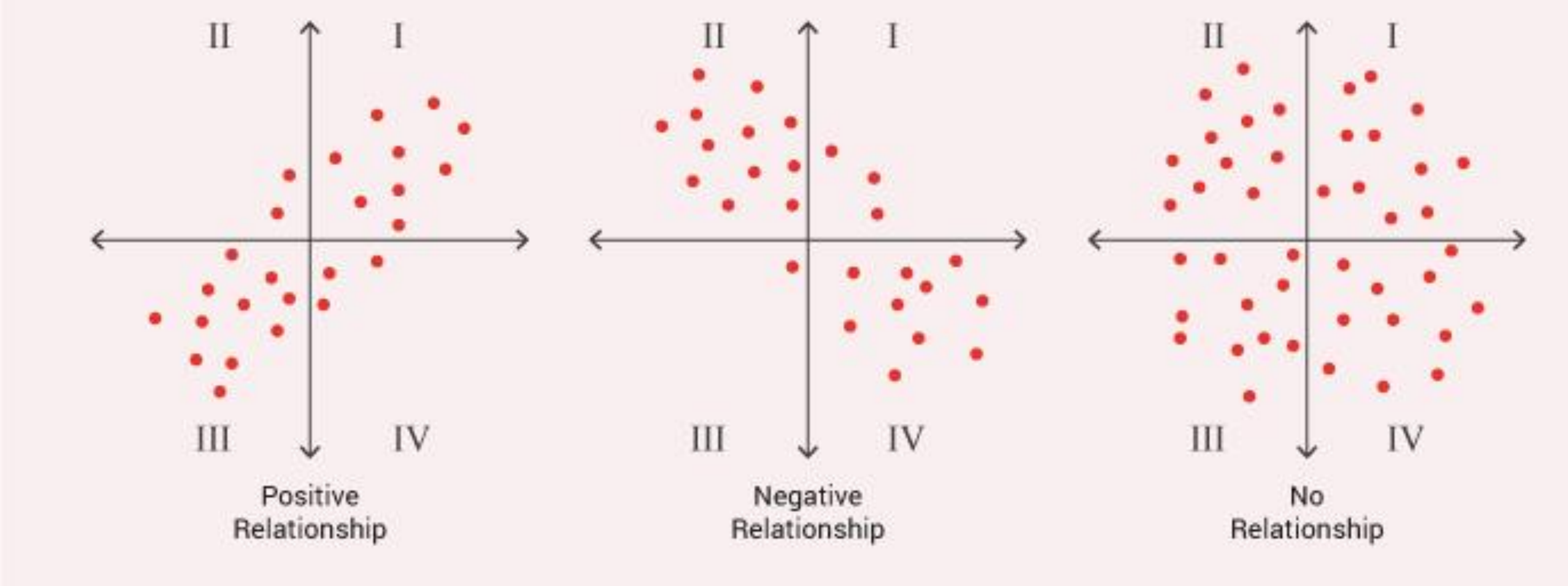


Relationship diagram: Weight vs Length of Passenger Vehicle

Covariance is a measure of how much two variables vary together.

It's similar to Variance, but where variance tells you how a *single* variable varies, co-variance tells you how **two** variables vary together.

$$\sigma_{XY} = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n}$$

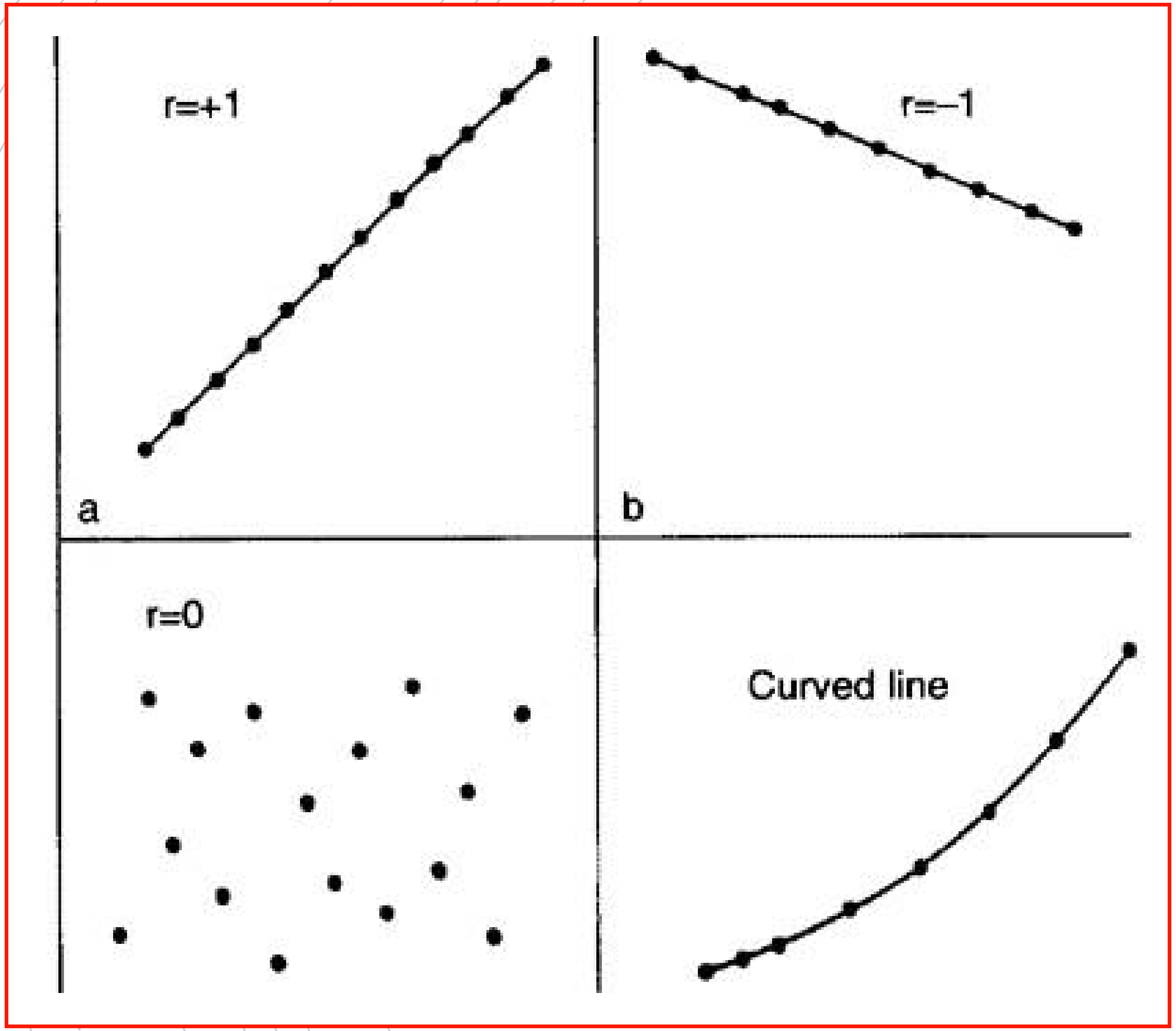


Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together.

Range of Correlation is between –1 to 1 where magnitude implies strength of relationship.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

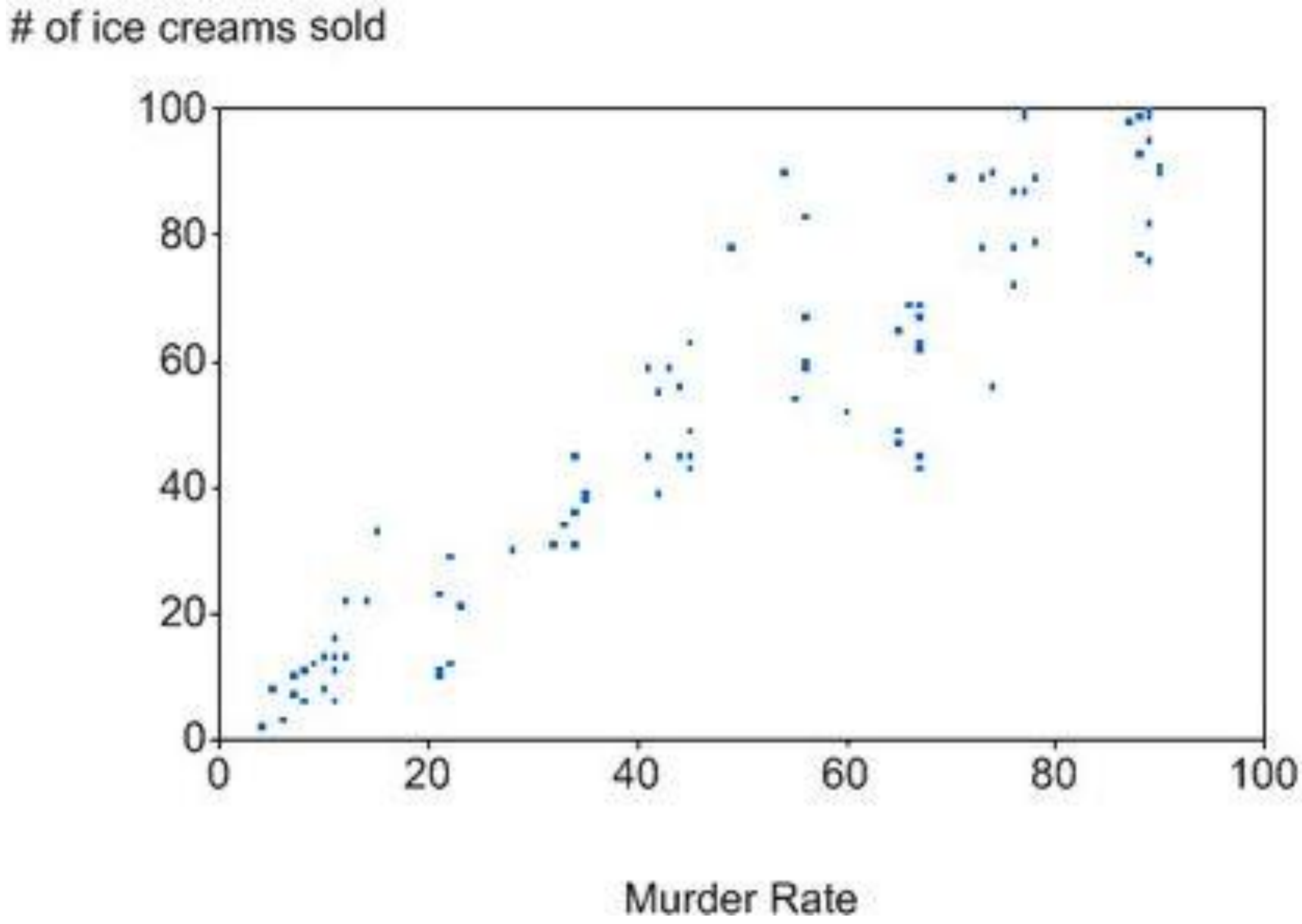
$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$



Correlation: Correlation is a normalized version of Covariance. It's a measure of linear assocITation between two Variables.

- Correlation value between 0, 1 mean positive correlation I.e both variables increase or decrease together.
- 0 correlation means no relationship
- Value between –1 to 0 means negative relationship I.e One variable increase while other variable decreases and vice versa

NOTE: Correlation does not imply causation i.e. High correlation does not mean one causes other.



r = 0.88, does not mean Ice cream sales is causing the death of people.

Probability and Disributions

Probability of Single Event:

$$\text{Probability of an outcome} = \frac{\text{Number of Outcome}}{\text{Total number of equally likely outcome}}$$

Probability of Two Independent Events:

- **P(A AND B) = P(A) * P(B)**
  - Probability of heads on tossing of two coins P(A) * P(B) = ½ * ½ = ¼

- **P(A OR B) = P(A) + P(B) - P(A AND B)**
  - Probability of head in 1st flip or probability of head in 2nd flip or both ½ + ½ - ¼ = ¾

Conditional Probability:

- Probability of an event given the other event has occurred.
- P(B|A) - Probability of event B given A has happened
  - P(A AND B) = P(A) * P(B|A)

- Probability of drawing 2 aces = P(drawing one ace from deck) * P(drawing one ace given already one ace is pulled out)
  - Probability of drawing 2 aces = 4/52 * 3/51

Bayes' theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Or the extended alternative:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|\bar{A}) * P(\bar{A})}$$

Where Ā must be understand as not-A

LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

POSTERIOR

The probability of "A" being True, given "B" is True

MARGINALIZATION

The probability "B" being True.

Bayes Theorem:

Example:

- Dangerous fires are rare (1%)
- but smoke is fairly common (10%),
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(\text{Fire}|\text{Smoke}) = \frac{P(\text{Fire})\, P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})}$$

$$= \frac{1\% \times 90\%}{10\%}$$

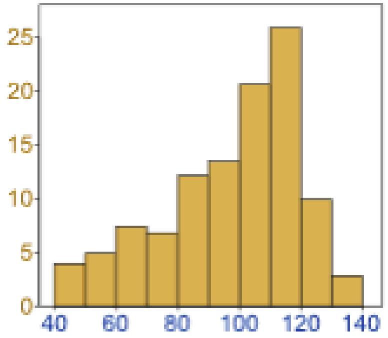
$$= 9\%$$

Distributions

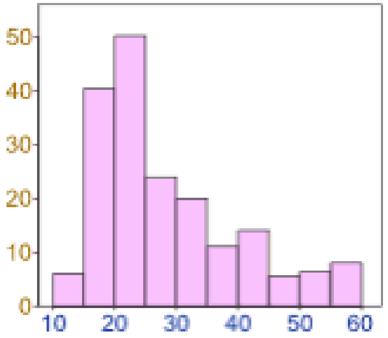
Data Distribution

Data can be "distributed" (spread out) in different ways.

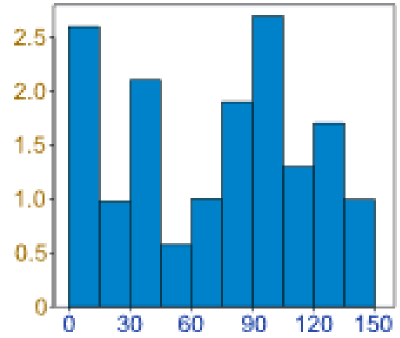
It can be spread out more on the left



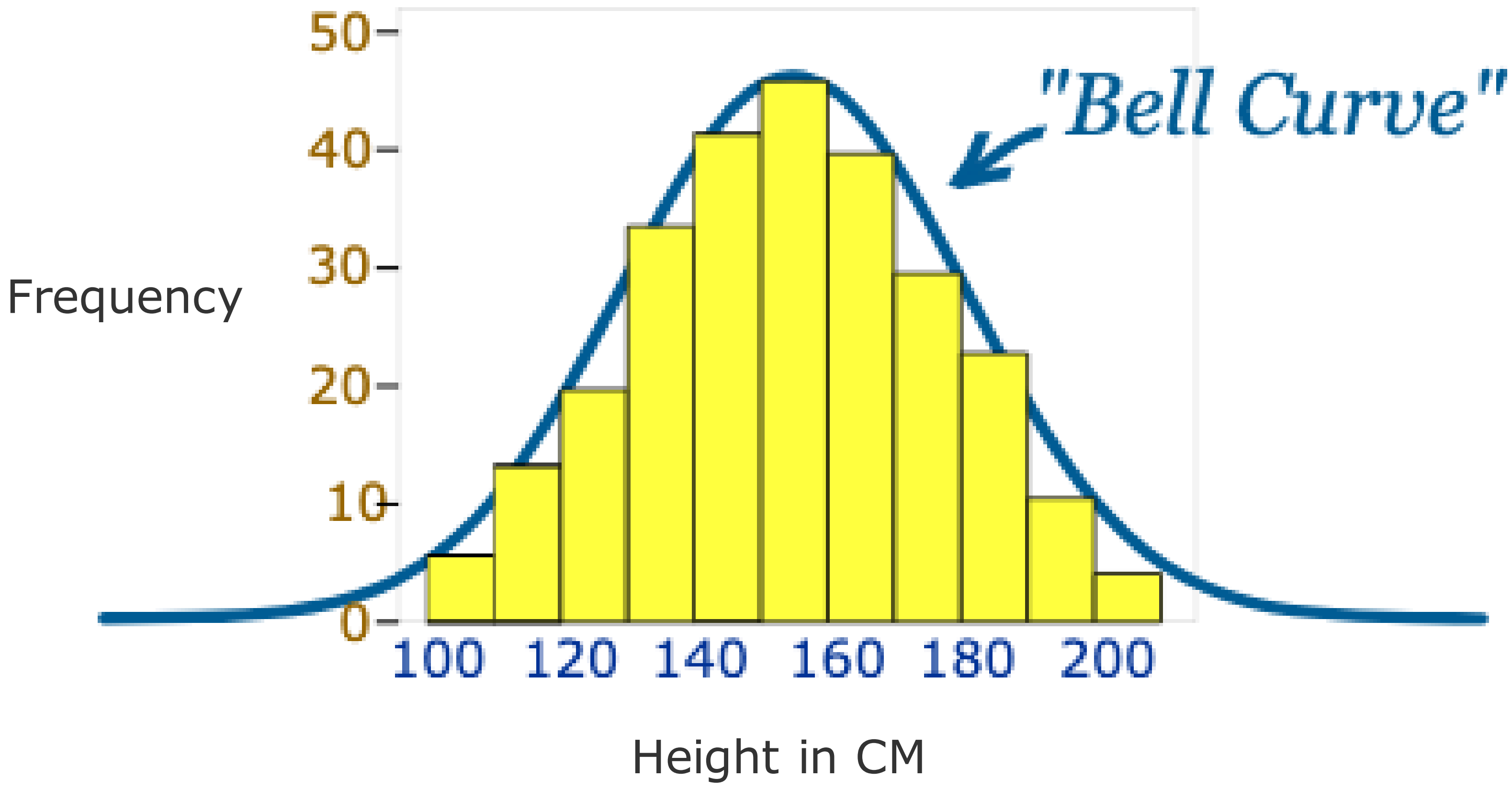
Or more on the right



Or it can be all jumbled up

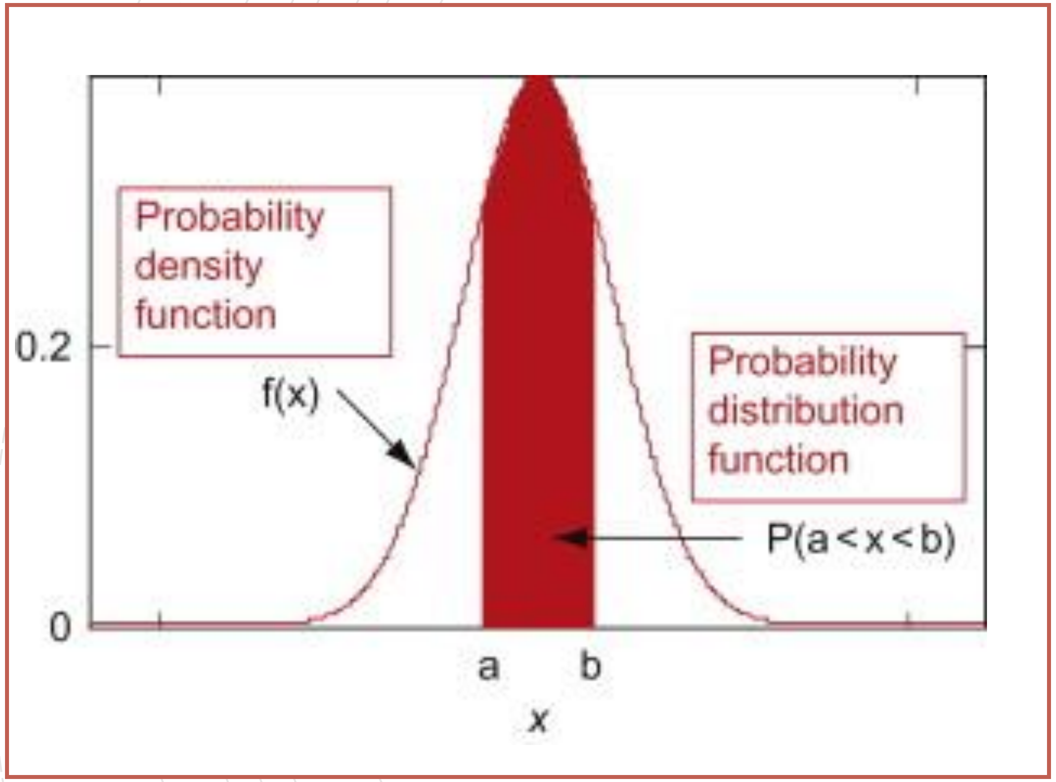


But there are many cases where the data tends to be around a central value with no or very little bias to left or right, and it gets close to a "Normal Distribution" like this:



This distribution has some really interesting properties which we will discuss in upcoming slides.

Probability distribution Function:



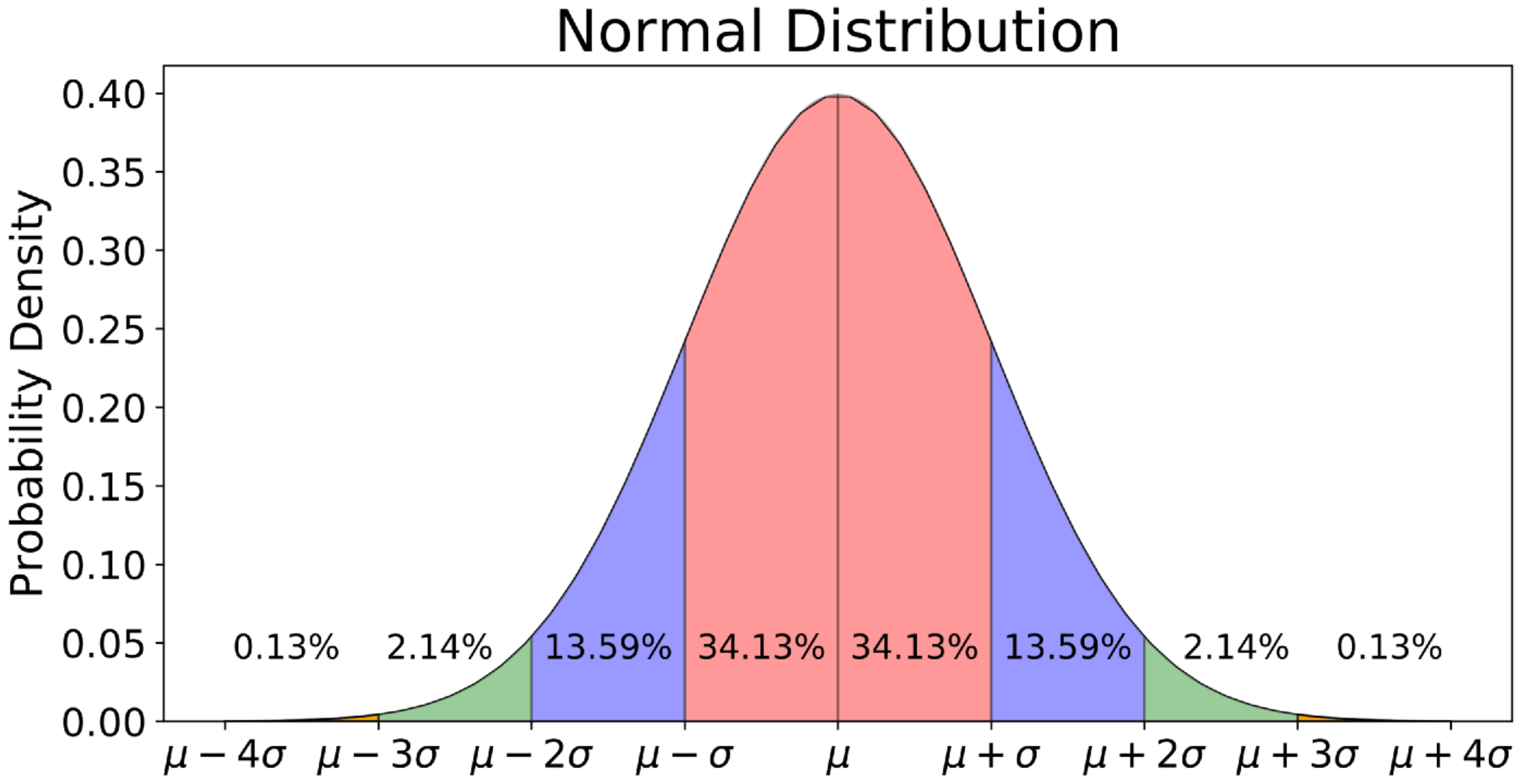
- A function describing the likelihood of obtaining possible values that a random variable can assume.
- PDF is used to specify the probability of the random variable falling *within a particular range of values*, as opposed to taking on any one value.
- This probability is given by the integral of this variable's PDF over that range

Normal Distribution

A normal distribution, sometimes called the bell curve, is a distribution that is used to represent real valued continuous distributions very often.

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \mu &= \text{ Mean} \\ \sigma &= \text{ Standard Deviation} \\ \pi &\approx 3.14159\cdots \\ e &\approx 2.71828\cdots \end{aligned}$$



- The bell curve/Normal Distribution is symmetrical.
- Half of the data will fall to the left of the **mean**; half will fall to the right.

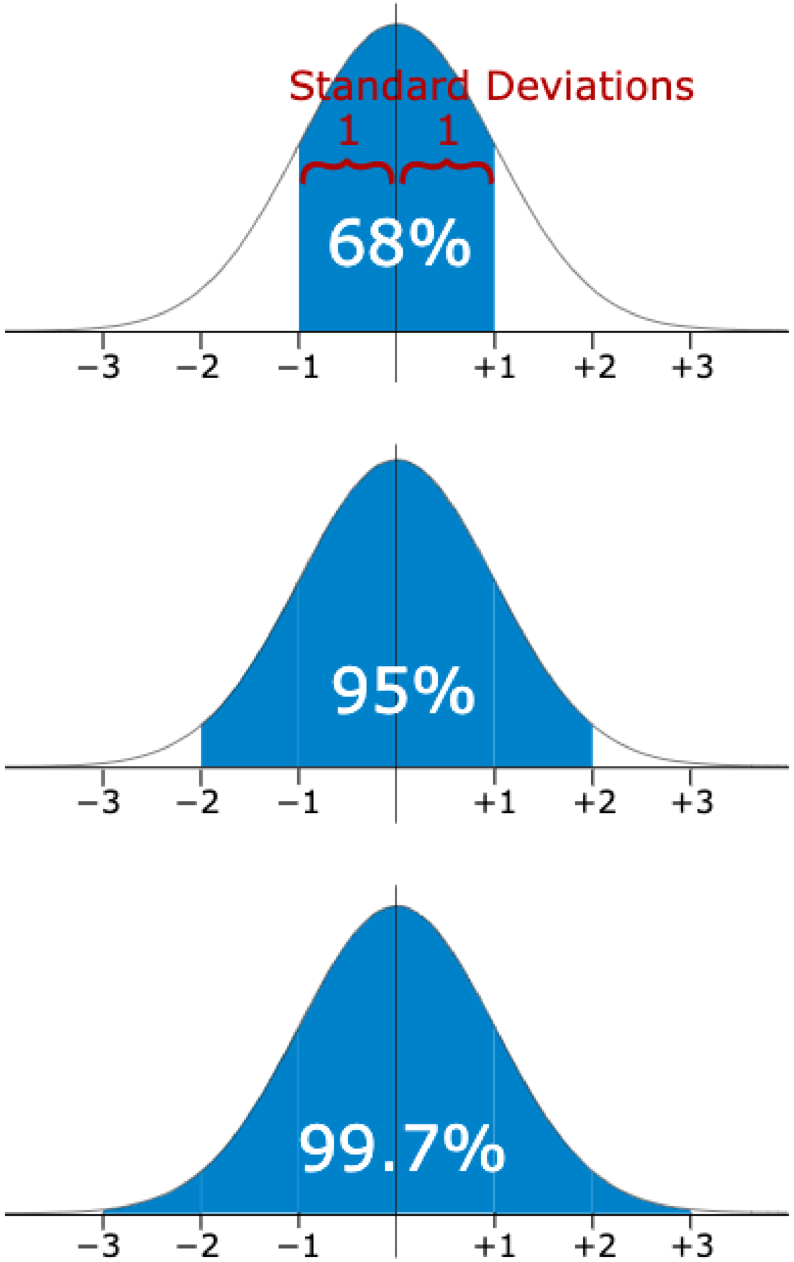
Normal Distribution

Properties of Normal Distributions:

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e., around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

A **standard normal distribution** is an extension of normal distribution with a **mean of 0 and a standard deviation of 1.**

Standard Deviation of Normal Distributions:



68% of values are within
1 standard deviation of the mean

95% of values are within
2 standard deviations of the mean

99.7% of values are within
3 standard deviations of the mean

Normal Distribution

Examples of Normal Distributions:

- Marks of Students in Tests
- Rainfall
- Salary of Employees
- Height of People
- IQ Scores

Quiz

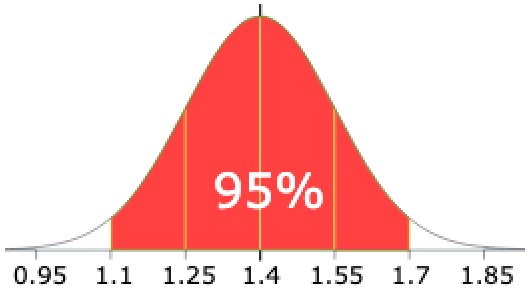
Q: 95% of students at of a class scored between are between **20 marks** and **80 marks** in a test. Assuming this data is **normally distributed** can you calculate the mean and standard deviation?

Solution:

Step 1: The mean is halfway between 20 and 80:

$$\text{Mean} = (20 + 80) / 2 = \mathbf{50}$$

Step 2: 95% is two standard deviation either side of mean so total 4 deviations:



$$4 \text{ std} = (80-20)$$

$$1 \text{ std} = (80-20)/4$$

$$\mathbf{Std = 15}$$

Standard Score: The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "z-score".

Q: One student score 95 marks. What will be his Z-score:

Ans: To convert a value to a Standard Score ("z-score"):

- first subtract the mean: 95-50 = 45
- then divide by the Standard Deviation: 45/15 = 3

$$z = \frac{x - \mu}{\sigma}$$

Q: The NEXA Tea Company pack tea in bags marked as **250 g**.
A large number of packs of tea were weighed and the mean and standard deviation were calculated as **255 g and 2.5 g** respectively.
Assuming this data is normally distributed, what percentage of packs are underweight?

Q: Students pass a test if they score 50% or more.
The marks of a large number of students were sampled and the **mean and standard deviation** were calculated as **42% and 8%** respectively.
Assuming this data is **normally distributed**, what percentage of students pass the test?

Q: The mean June midday temperature in Chennai is **36°C and the standard deviation is 3°C**
Assuming this data is normally distributed, how many days in June would you expect the midday temperature to be between **39°C and 42°C**?

Normality Test

In statistics, **normality tests** are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

- D'Agostino's K-squared test,
- Jarque–Bera test,
- Anderson–Darling test,
- Cramér–von Mises criterion,
- Kolmogorov–Smirnov test
- Lilliefors test
- Shapiro–Wilk test,
- Pearson's chi-squared test

Statistical inference

Central Limit Theorem
and
Hypothesis Testing

Statistical inference

- **Statistical inference** is the process of using data analysis to deduce properties of an underlying distribution of probability.
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Sampling distribution

The S**ampling distribution** of a statistic is the distribution of that statistic

For example:

- Consider a normal population with mean X and standard deviation sigma.
- Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample.
- This statistic is called the sample mean.
- The distribution of these means, or averages, is called the "sampling distribution of the sample mean".
- The standard deviation of the sampling distribution of a statistic is referred to as the standard error of that quantity.

$$\mathrm{SD}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

- The **Central Limit Theorem** states that the **sampling distribution of the sample means** approaches a normal distribution as the sample size gets larger - *no matter what the shape of the population distribution*.

- This fact holds especially true for sample sizes over 30.

- The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

Hypothesis testing

- **Hypothesis testing** is an act in statistics whereby an analyst **tests** an assumption regarding a population parameter.
- **Hypothesis testing** is used to assess the plausibility of a **hypothesis** by using **sample** data.
- The **null hypothesis** is the one to be tested and the **alternative** is everything else.
- For **example**:
    - **Null hypothesis:** The mean data scientist salary is 80,000 INR PM.
    - **Alternative hypothesis**: The mean data scientist salary is not 80,000 INR

6 steps of hypothesis testing

- Step 1: Specify the Null Hypothesis.
- Step 2: Specify the Alternative Hypothesis.
- Step 3: Set the Significance Level
- Step 4: Calculate the Test Statistic and Corresponding P-Value.
- Step 5: Drawing a **Conclusion**.

Let's understand these steps with example in Jupyter: