# DATA ANALYTICS

*DIPLOMA WALLAH*

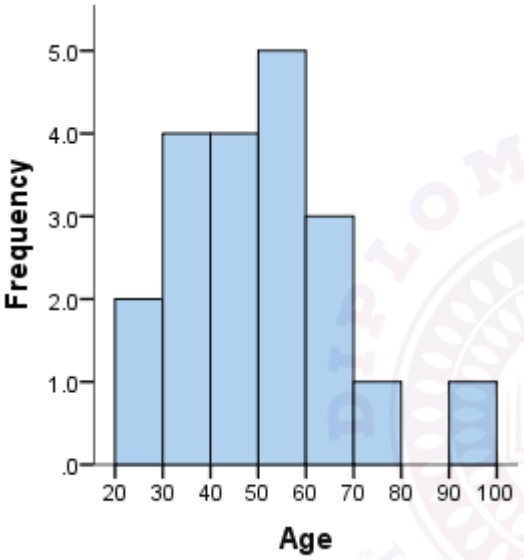
***OPEN ELECTIVE***

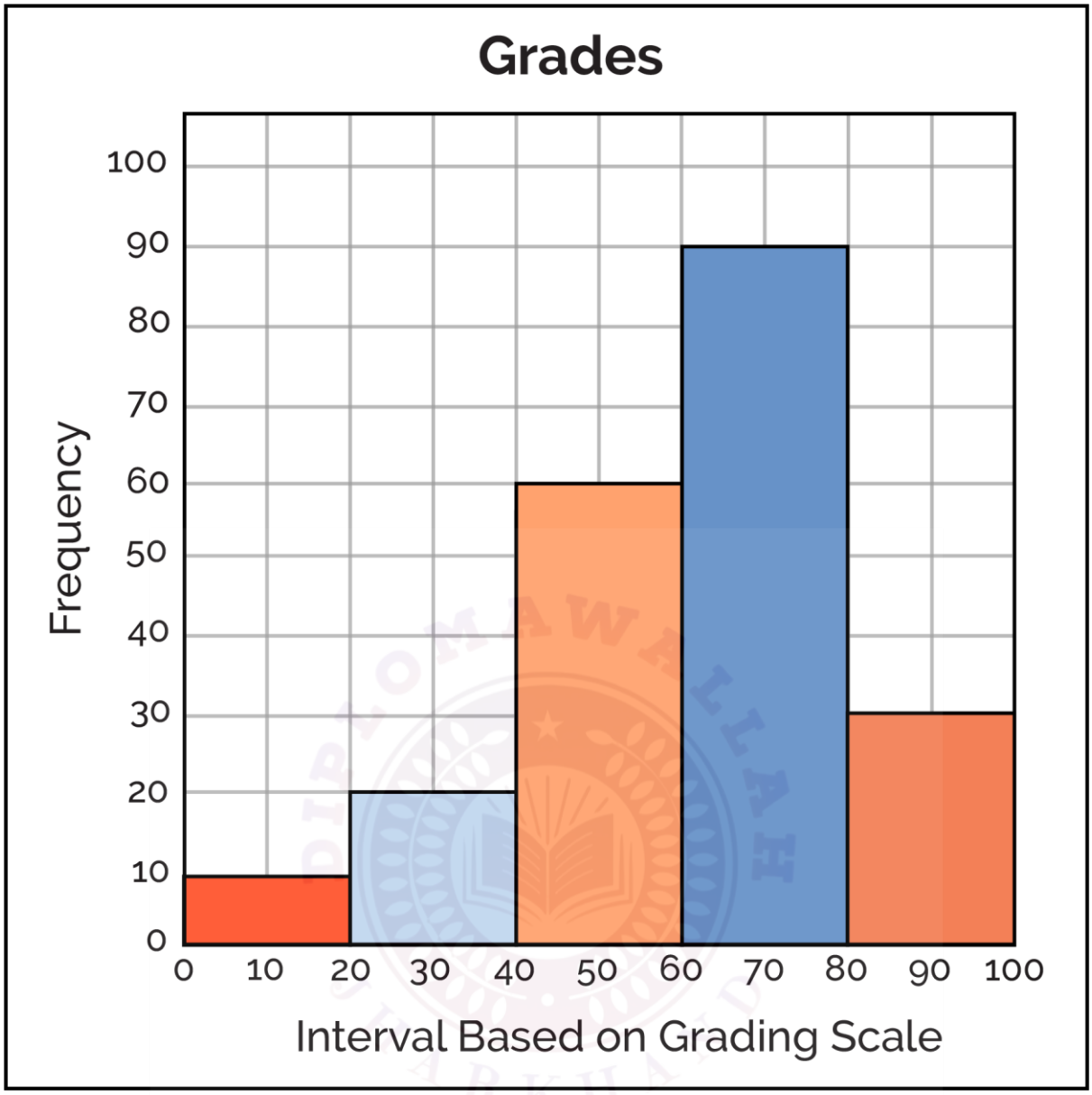
***Jharkhand University Of Technology (JUT)***

## UNIT - II STATISTICAL ANALYSIS

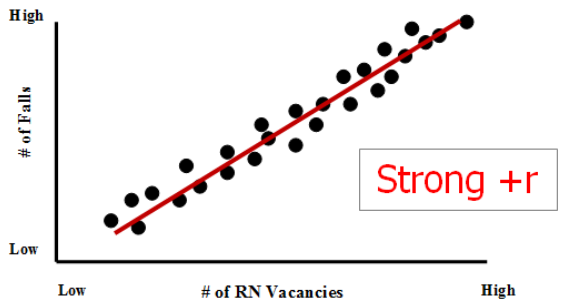
2.1 Graphical Techniques, Box Plot, Skewness & Kurtosis, Descriptive Statistics

Graphical Techniques

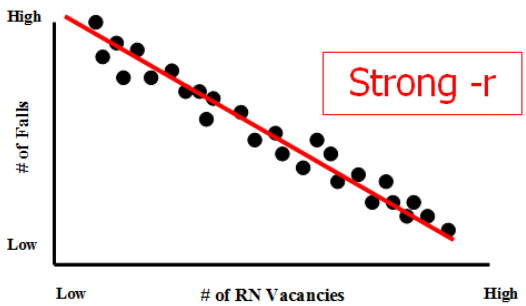




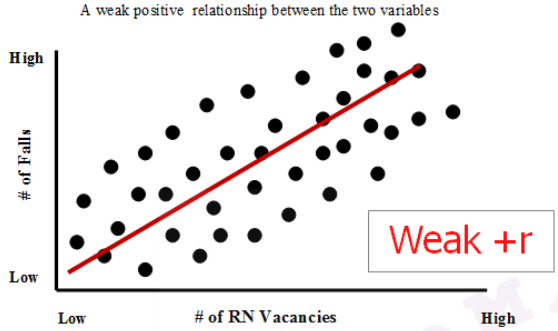
A strong positive relationship between the two variables



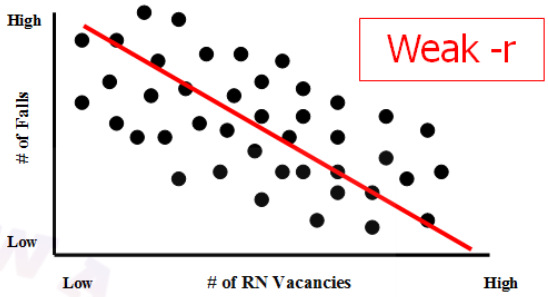

A strong negative relationship between the two variables

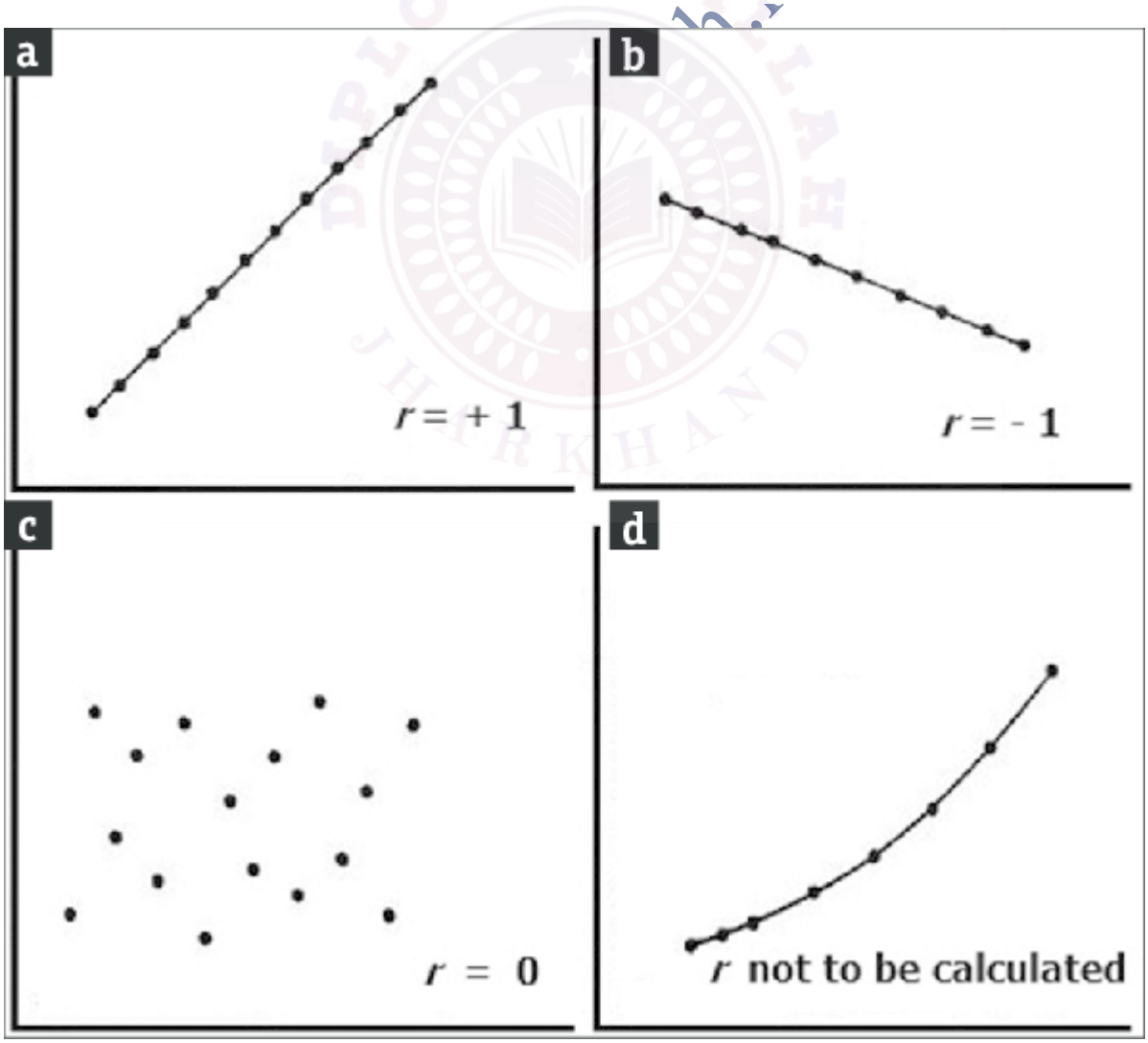


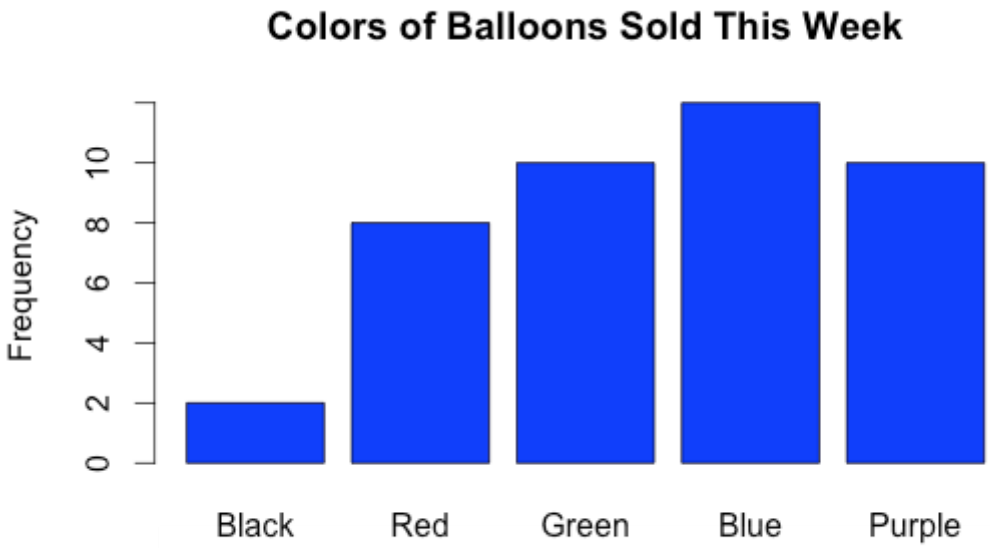

A weak positive relationship between the two variables

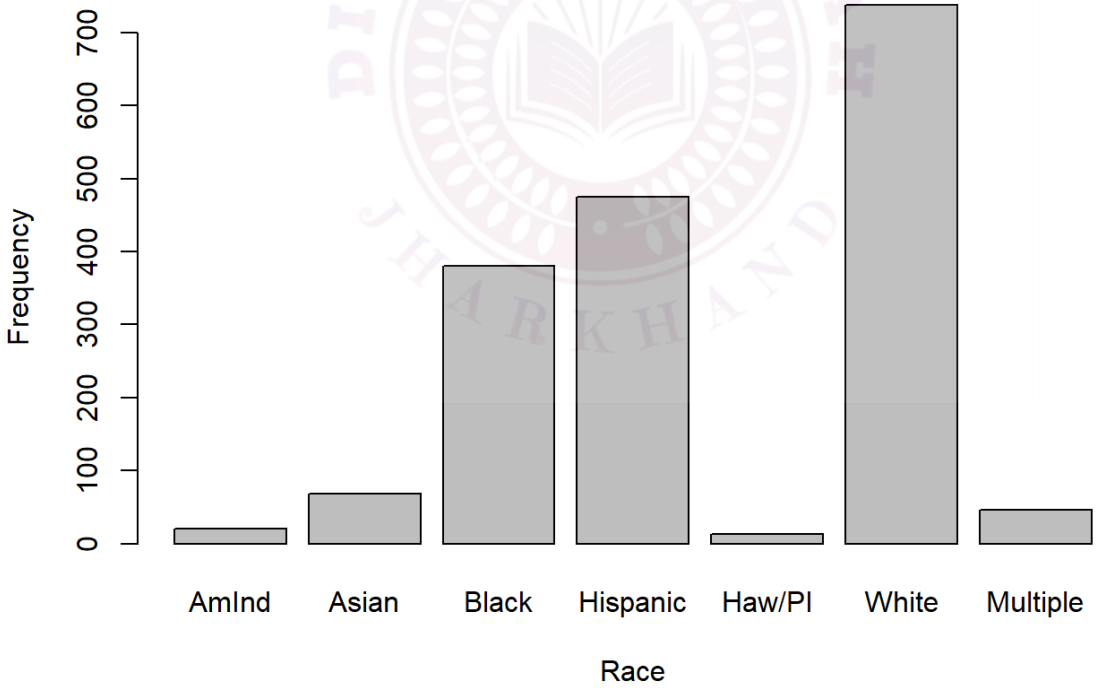



A weak negative relationship between the two variables






What it is: Graphical techniques are visual methods (plots, charts) used to summarise and explore data. Instead of just numbers, you *see* the shape, spread, outliers, and patterns.

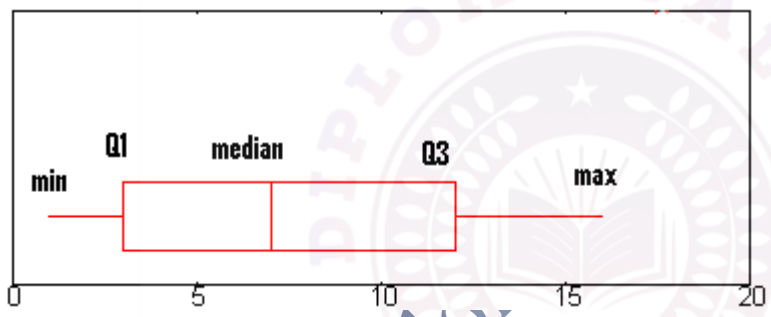
Why it matters: In your exams and your work, you'll often need to pick a suitable graph, interpret what the graph is telling you, and say what conclusions you can (or

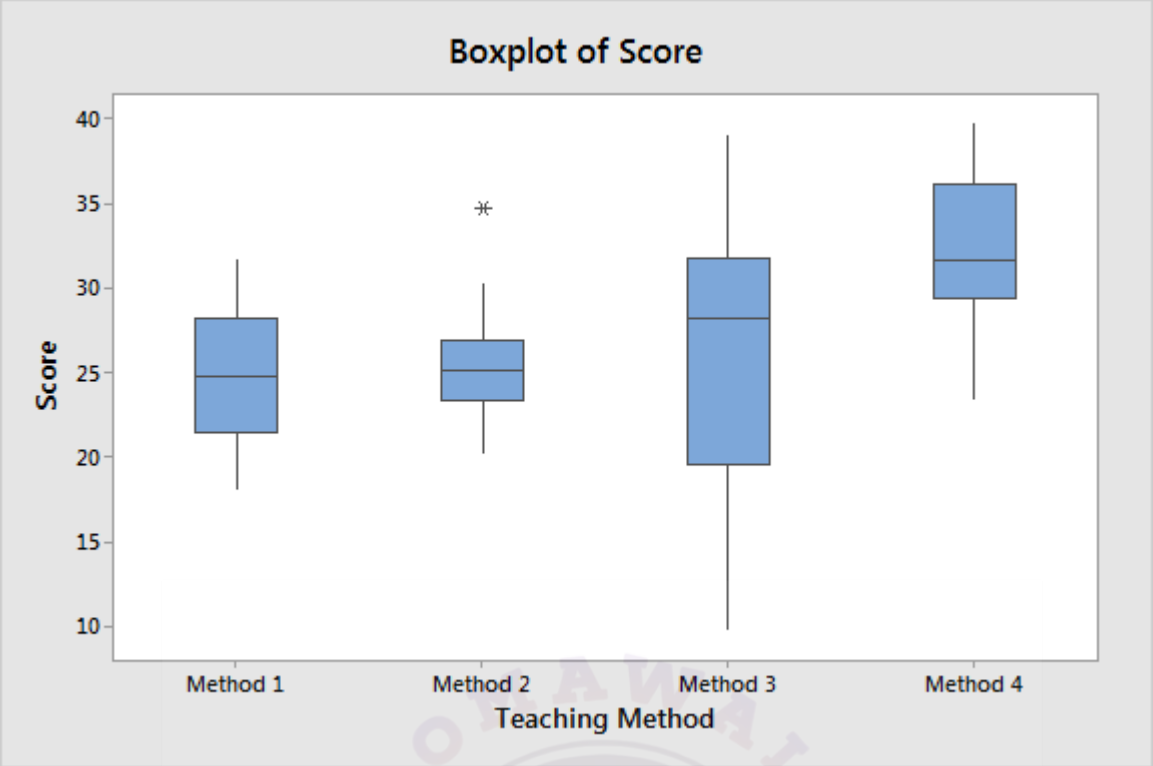
can't) draw.
Key types:

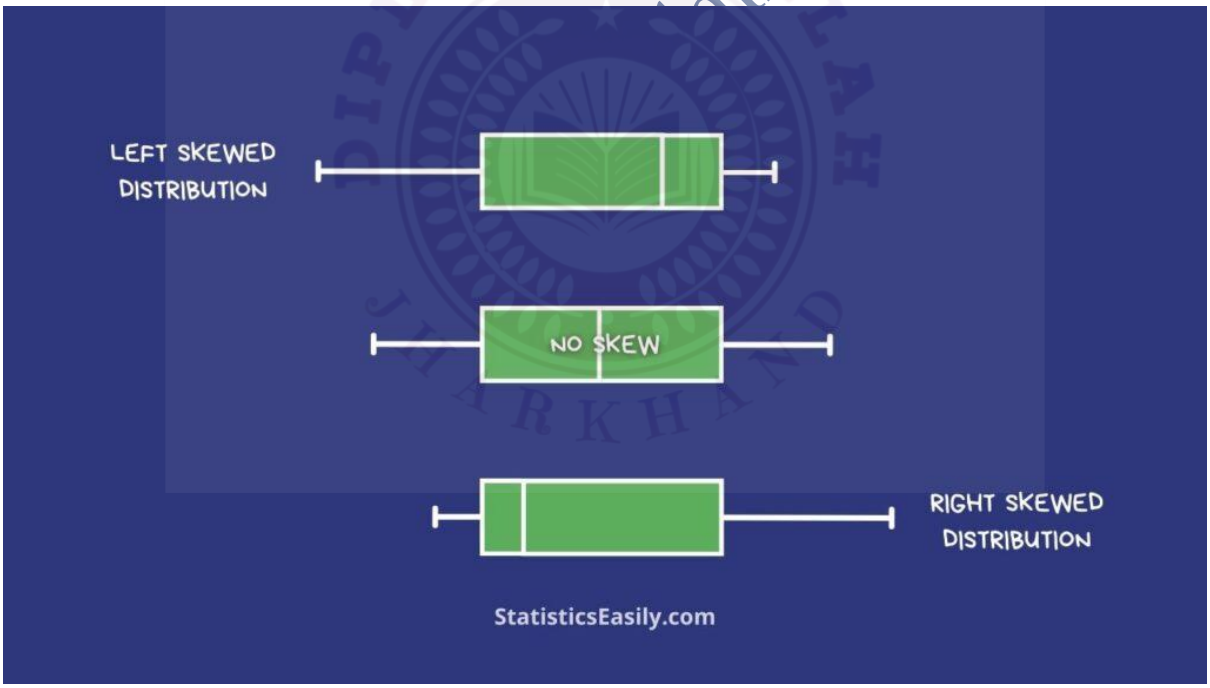
- Histogram: for a single numeric variable, shows frequency distribution.
- Bar chart: for categorical variables, shows counts or proportions.
- Scatter plot: for two numeric variables, to visualise relationship.
- Box plot: summarises distribution using quartiles (see next section).
  Tips: When interpreting a graph, always mention: what the axes represent, the shape of distribution (e.g., symmetric/skewed), spread, existence of outliers, any clusters or gaps.

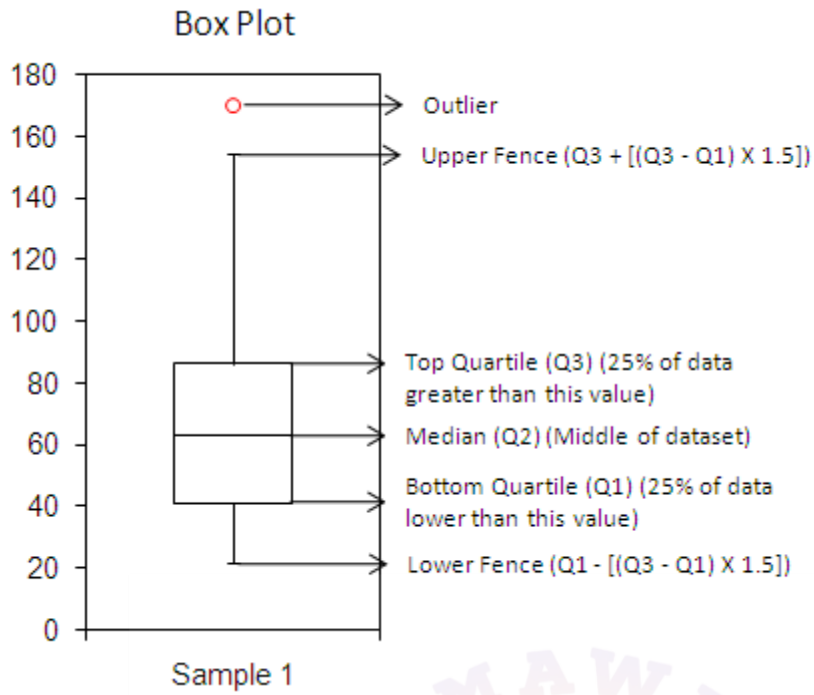
---

Box Plot





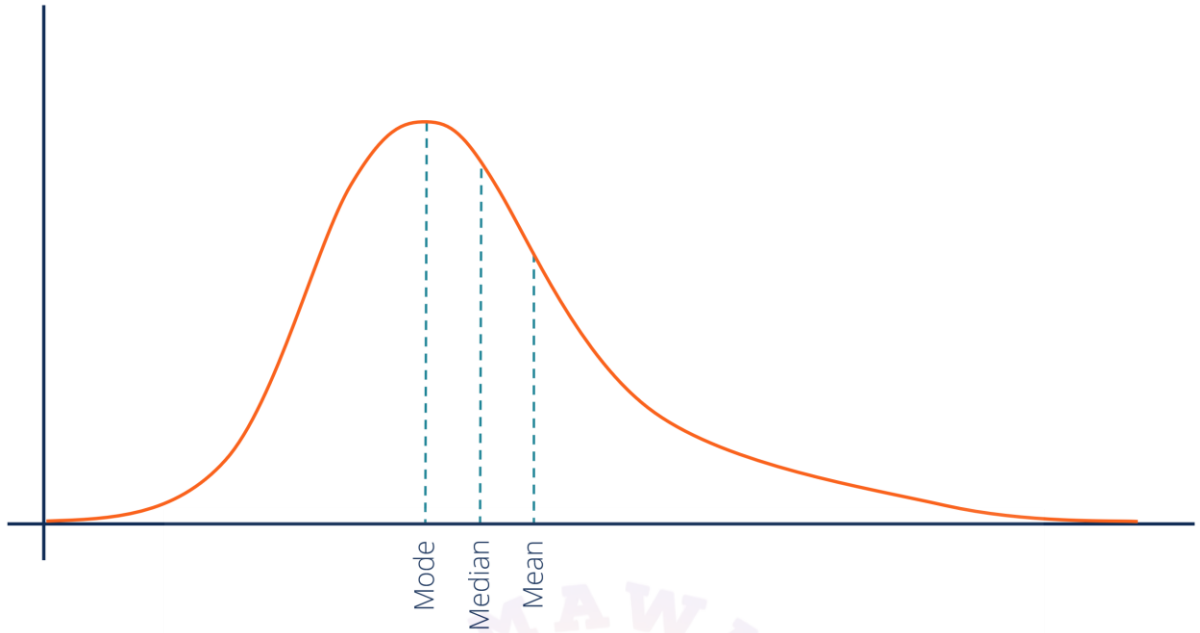


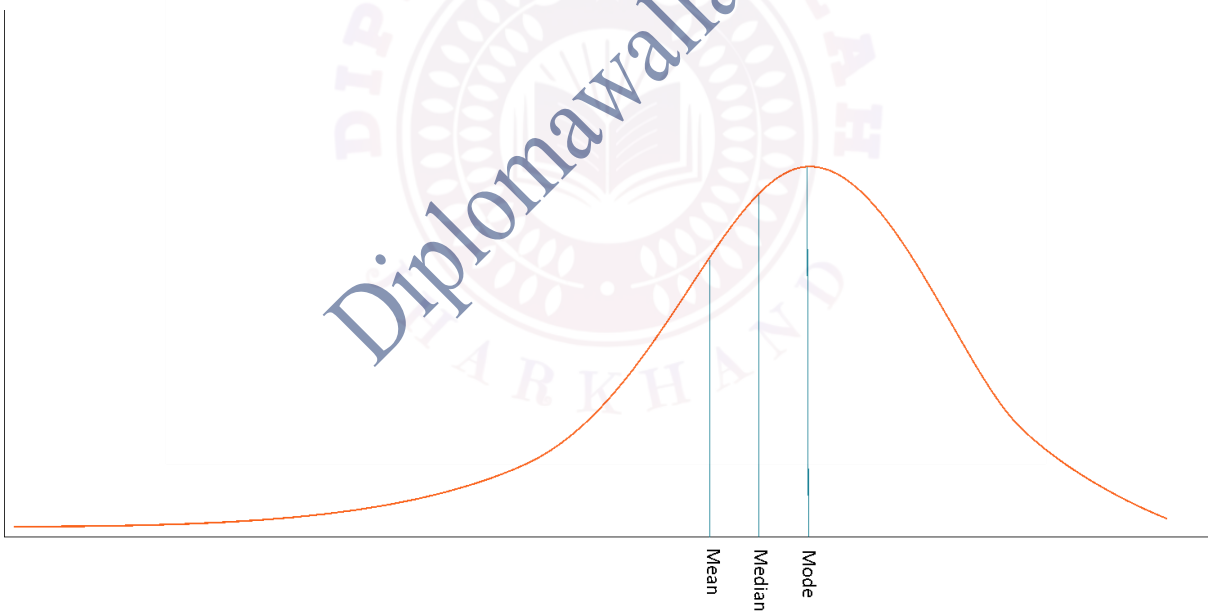

Definition: A box plot (aka box-and-whisker plot) visually summarises a dataset by showing its five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. (Simply Psychology)
How to interpret it:

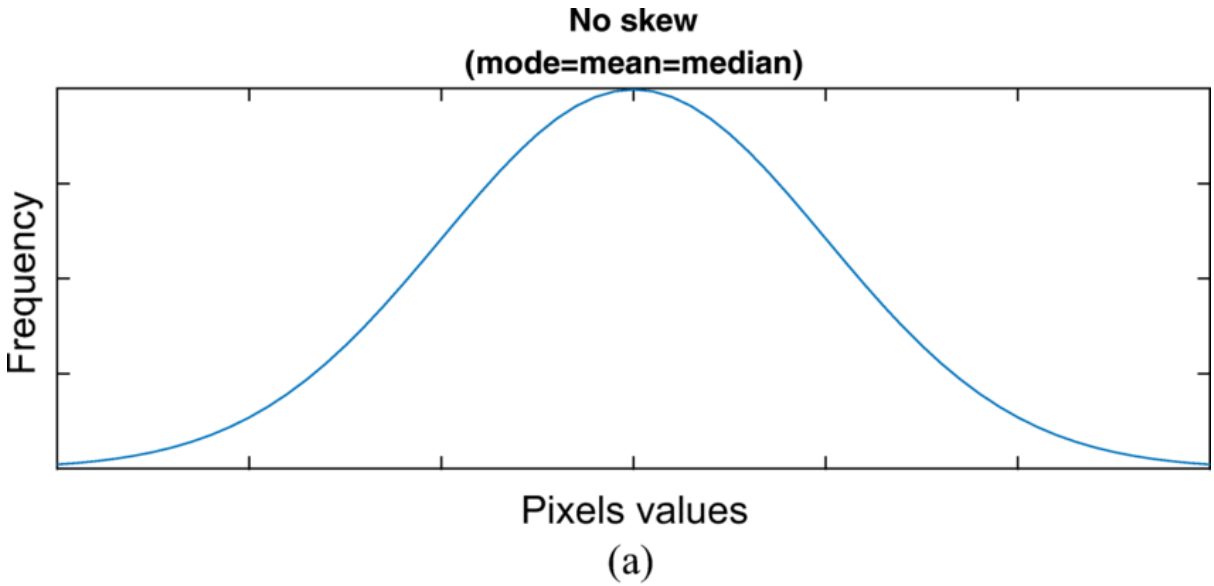
- The "box" spans Q1 to Q3 — that covers the middle 50 % of data.
- The line inside the box is the median (50th percentile).
- Whiskers extend to minimum and maximum or sometimes to 1.5×IQR beyond Q1/Q3; points beyond those are often treated as outliers. (GeeksforGeeks)
- Where the median lies inside the box and the lengths of whiskers tell you skewness: e.g., if the median is closer to the bottom of the box and the upper whisker is long → positive skew.
  Why it matters (exam style):
- You might be given data and asked to draw a box plot.
- Or shown a box plot and asked to interpret: e.g., "Comment on spread, median, skewness and outliers."
  Exam tip answer:

"From the box plot, the median is at ~ 50, the IQR is from ~ 35 to ~ 65, the upper whisker is longer than the lower → the distribution appears positively skewed. There is one data point beyond the upper whisker which is an outlier."

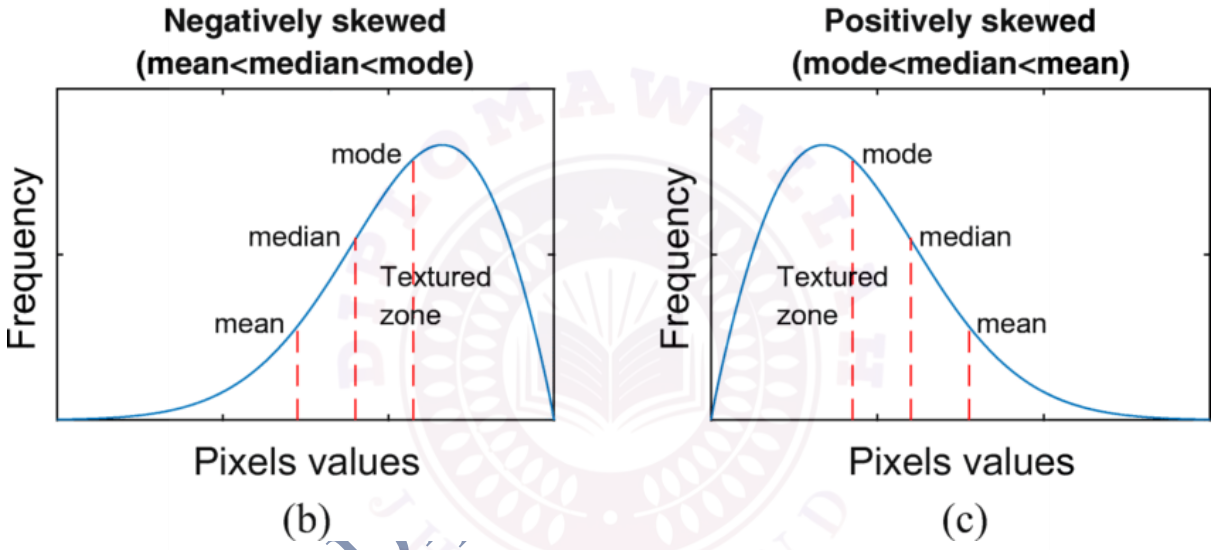
Skewness & Kurtosis



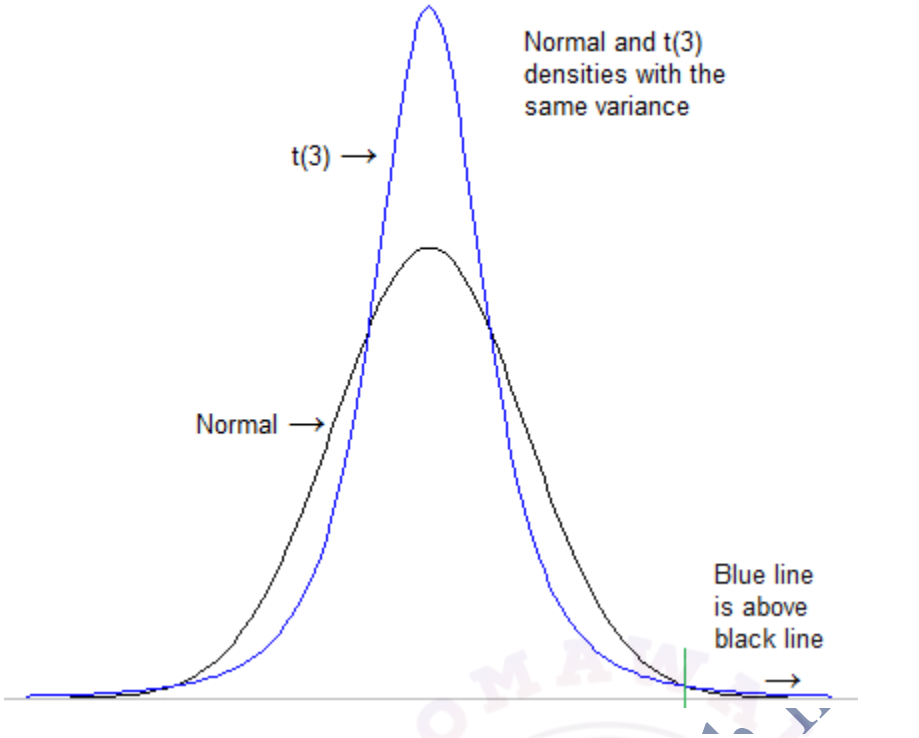



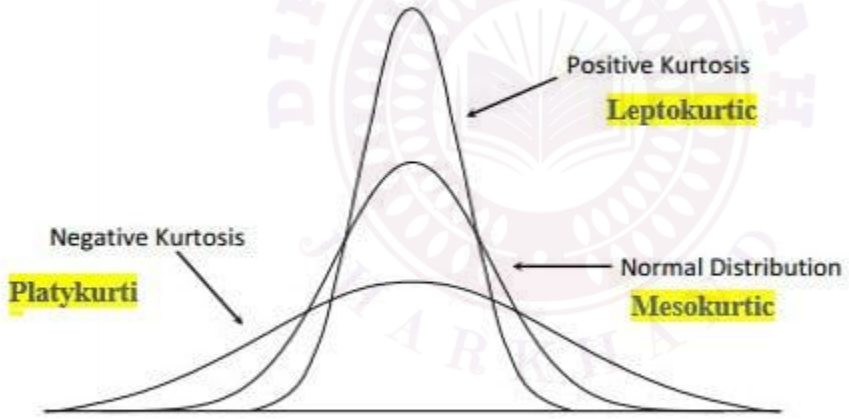



(a)




(b) (c)



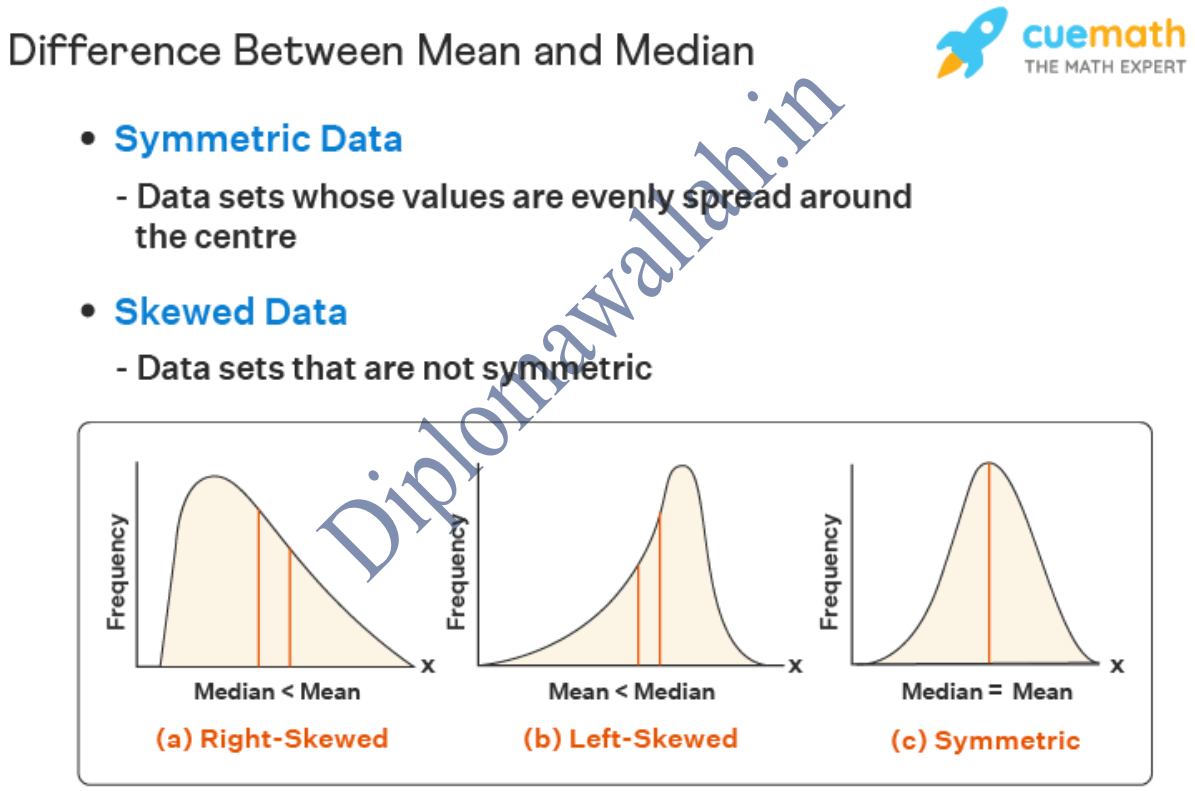

Skewness:

- It measures the *asymmetry* of a distribution. (NIST ITL)
- If skewness > 0 → right (positive) skew (tail on right, a few large values). If skewness < 0 → left (negative) skew.
- Why it matters: Skewness affects which "centre" measure (mean vs median) is appropriate and whether normal-based methods are valid.
  Kurtosis:
- It tells you about how heavy or light the tails of a distribution are ("tailedness"). (LearnVern)

- If kurtosis is high (heavier tails) → more extreme values/outliers; if lower → lighter tails.
  Exam style:
- Define both clearly.
- Given skewness and kurtosis values, interpret them: e.g., "Skewness = +1.2 → distribution is positively skewed."
- Explain implications: "Because distribution is heavily skewed and has heavy tails (kurtosis high), the mean may not be a good measure of centre and parametric tests assuming normality may not hold."

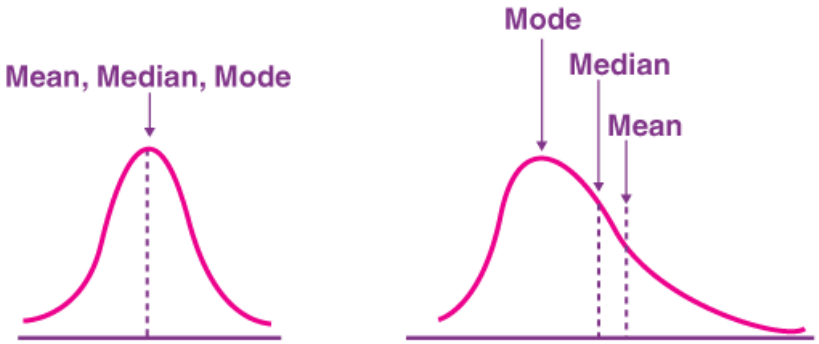
---

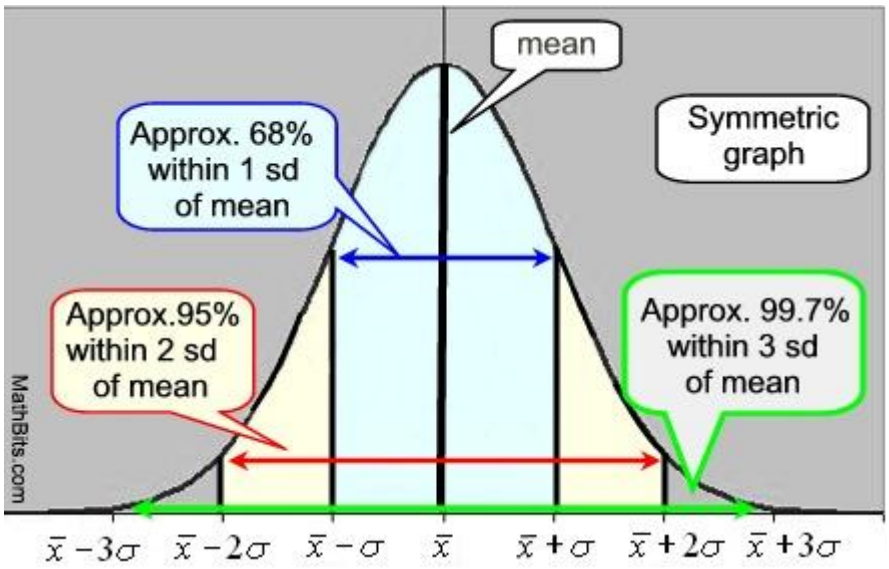
Descriptive Statistics

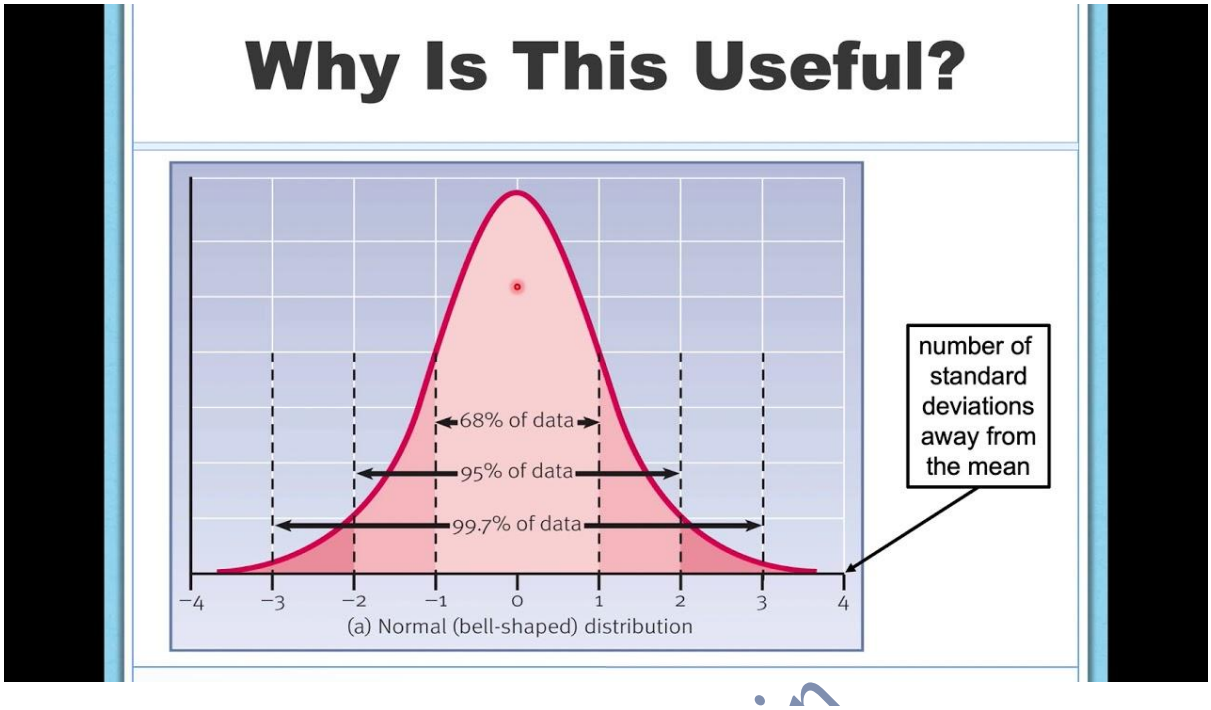


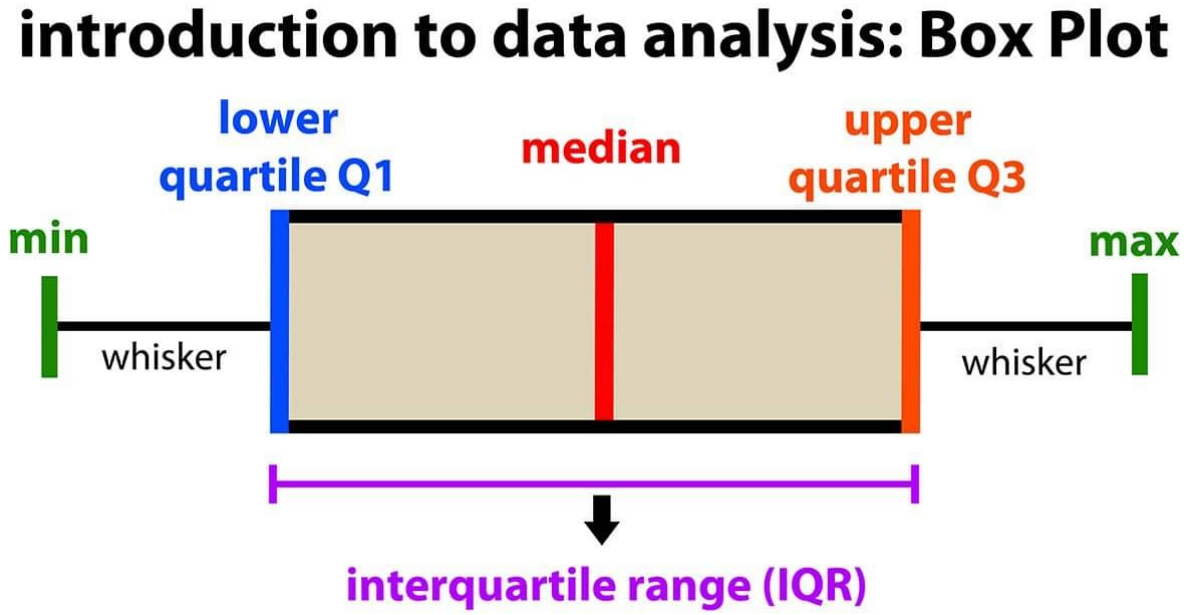
BYJU'S The Learning App

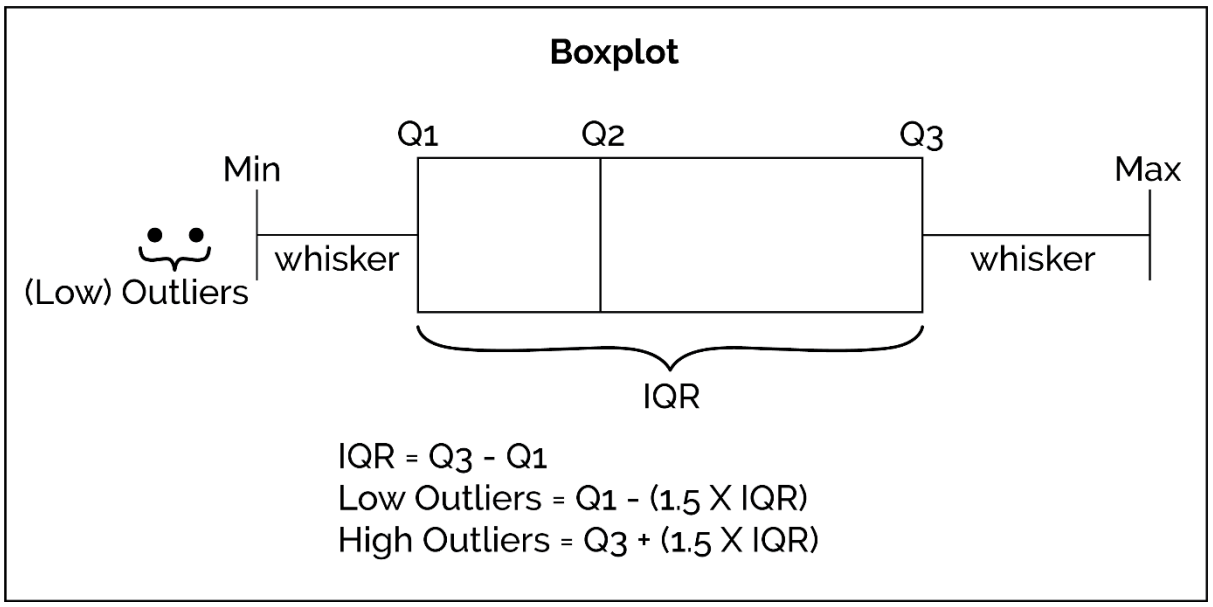
# Measures of Central Tendency, Mean, Median & Mode




Why Is This Useful?
68% of data
95% of data
99.7% of data
number of standard deviations away from the mean
−4 −3 −2 −1 0 1 2 3 4
(a) Normal (bell-shaped) distribution


introduction to data analysis: Box Plot
lower quartile Q1
median
upper quartile Q3
min
max
whisker
whisker
interquartile range (IQR)



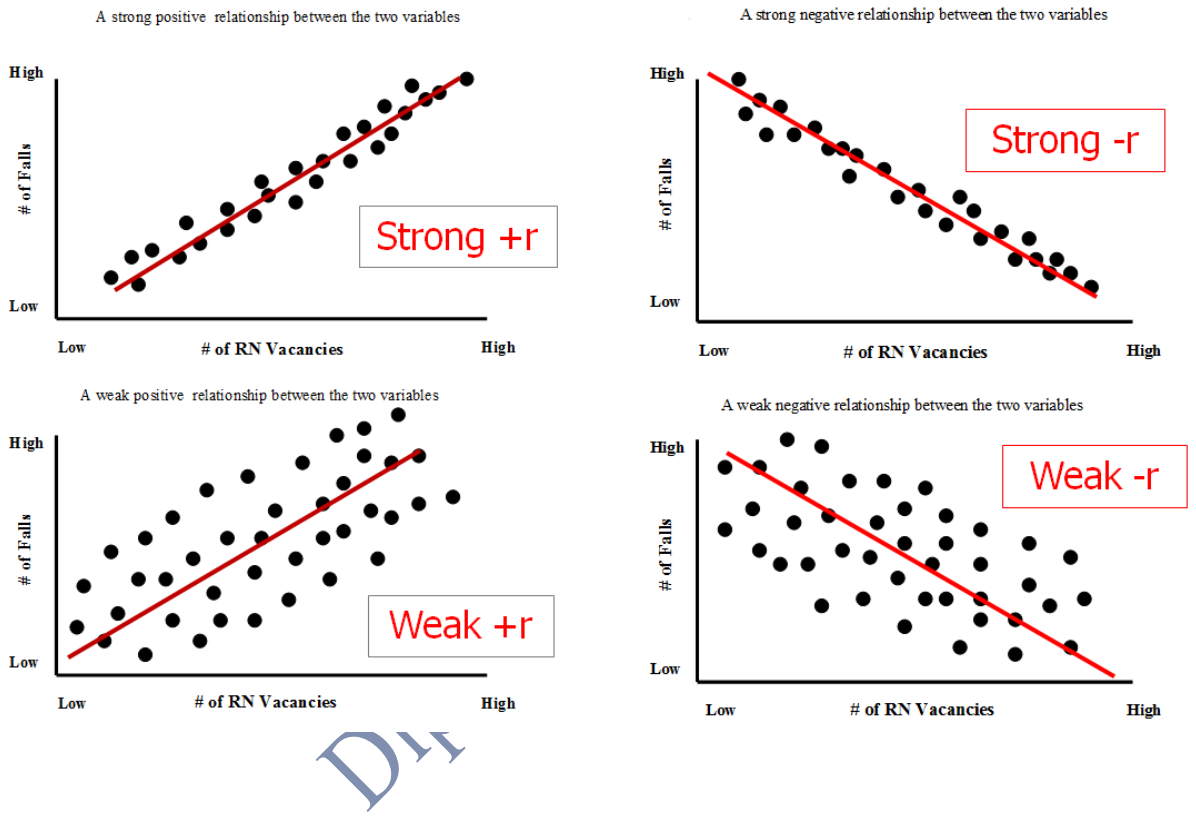

Definition: Descriptive statistics are summary measures that describe and simplify the main characteristics of a dataset — like its centre, spread, and shape. (Investopedia)
Key measures:

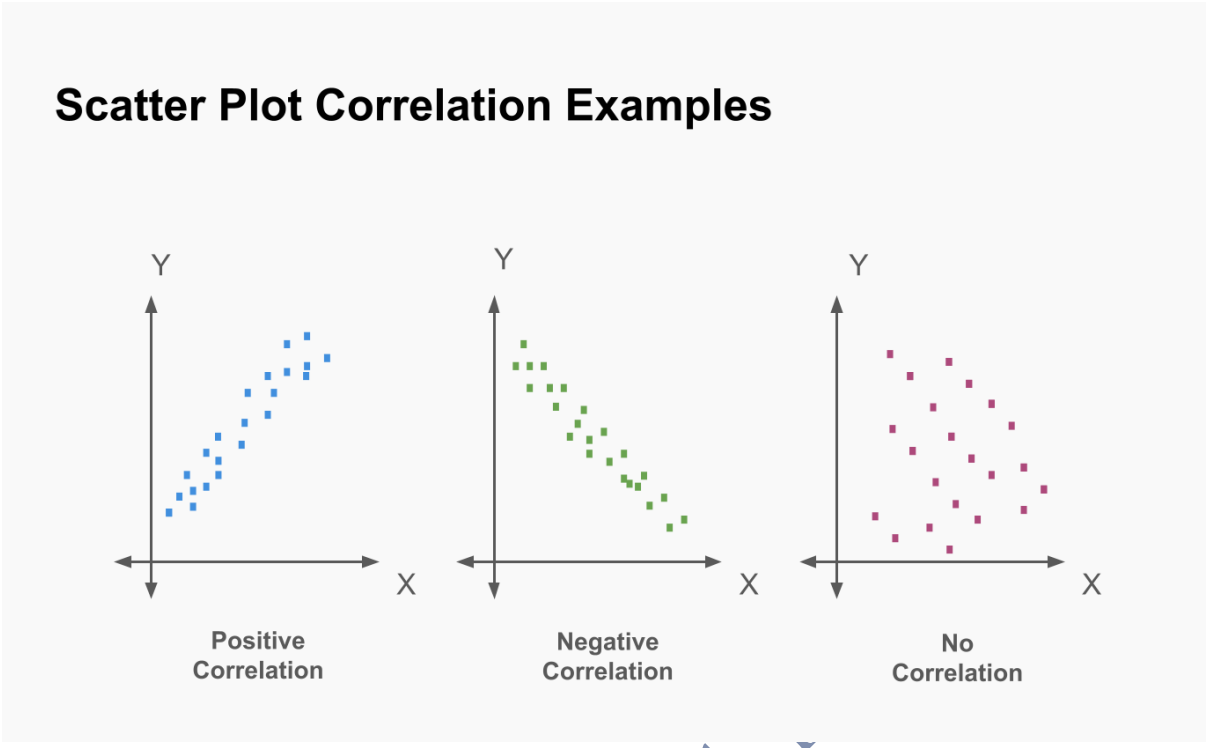
- Measures of central tendency: Mean, median, mode.
- Measures of dispersion (spread): Range, inter-quartile range (IQR), variance, standard deviation.
- Measures of shape: Skewness, kurtosis (already discussed).
  Why they matter: Before you dive into deeper inferential statistics (tests, modelling), you need to understand your data: where is it centred, how spread out, any skewness or outliers.
  Exam style:
- When asked "Which measure of centre is better when data are skewed?" → answer: median (because mean is sensitive to outliers/skew).
- When asked "Interpret the descriptive output: mean=45, SD=10, median=42" → note that mean > median suggests positive skew, the spread is moderate (SD 10), etc.
  Tips for calculation/interpretation:
- Know the formulas (or at least conceptually): e.g., standard deviation is the square root of average squared deviation from mean.
- Include interpretation in real terms: e.g., "With mean = 72 and SD = 8, about 68% of values lie between 64 and 80 if distribution is roughly normal."
- Always mention whether assumptions (normality, no extreme outliers) apply when using mean/SD.

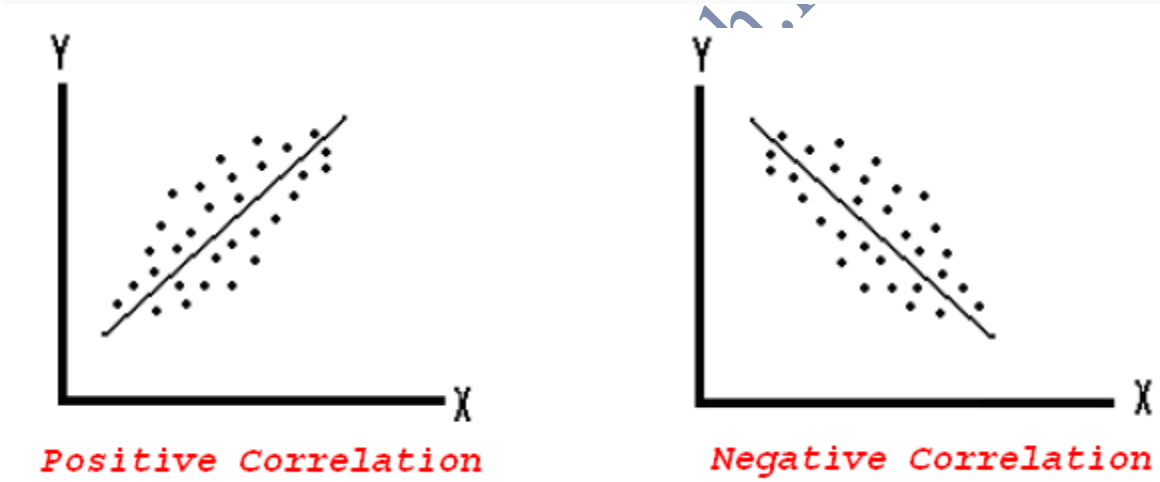
Great — let's continue with Section 2.2: Correlation and Regression, Data Cleaning, explained in detailed yet easy, exam-friendly language, with images.

## 2.2 Correlation and Regression, Data Cleaning

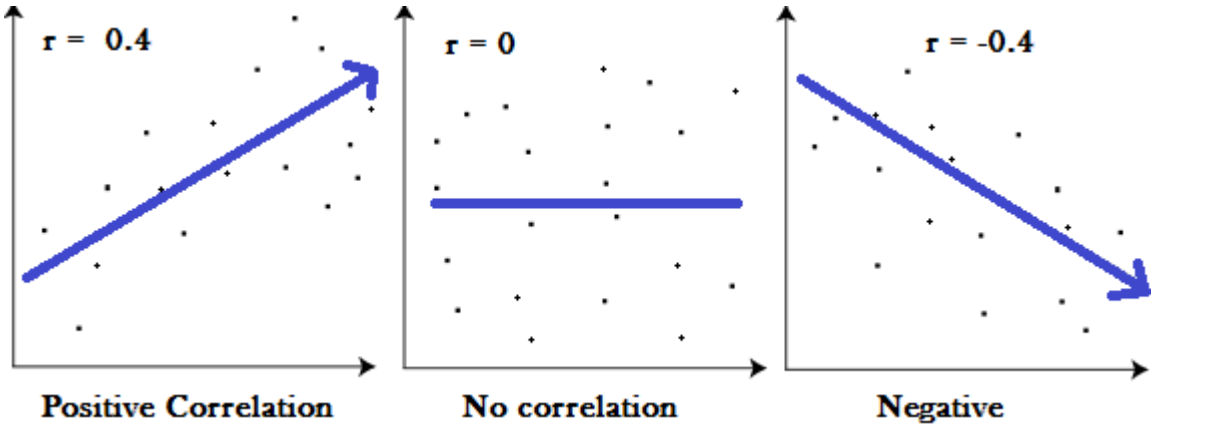
### Correlation

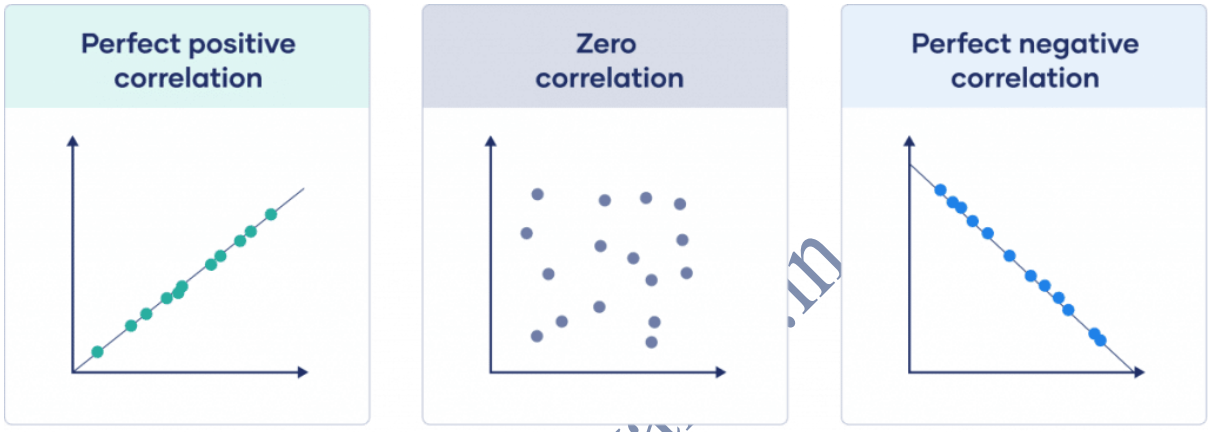



Scatter Plot Correlation Examples
Y
X
Positive Correlation
Y
X
Negative Correlation
Y
X
No Correlation


Y
X
Positive Correlation
Y
X
Negative Correlation


0
Positive Relationship
0
Negative Relationship
0
No Relationship





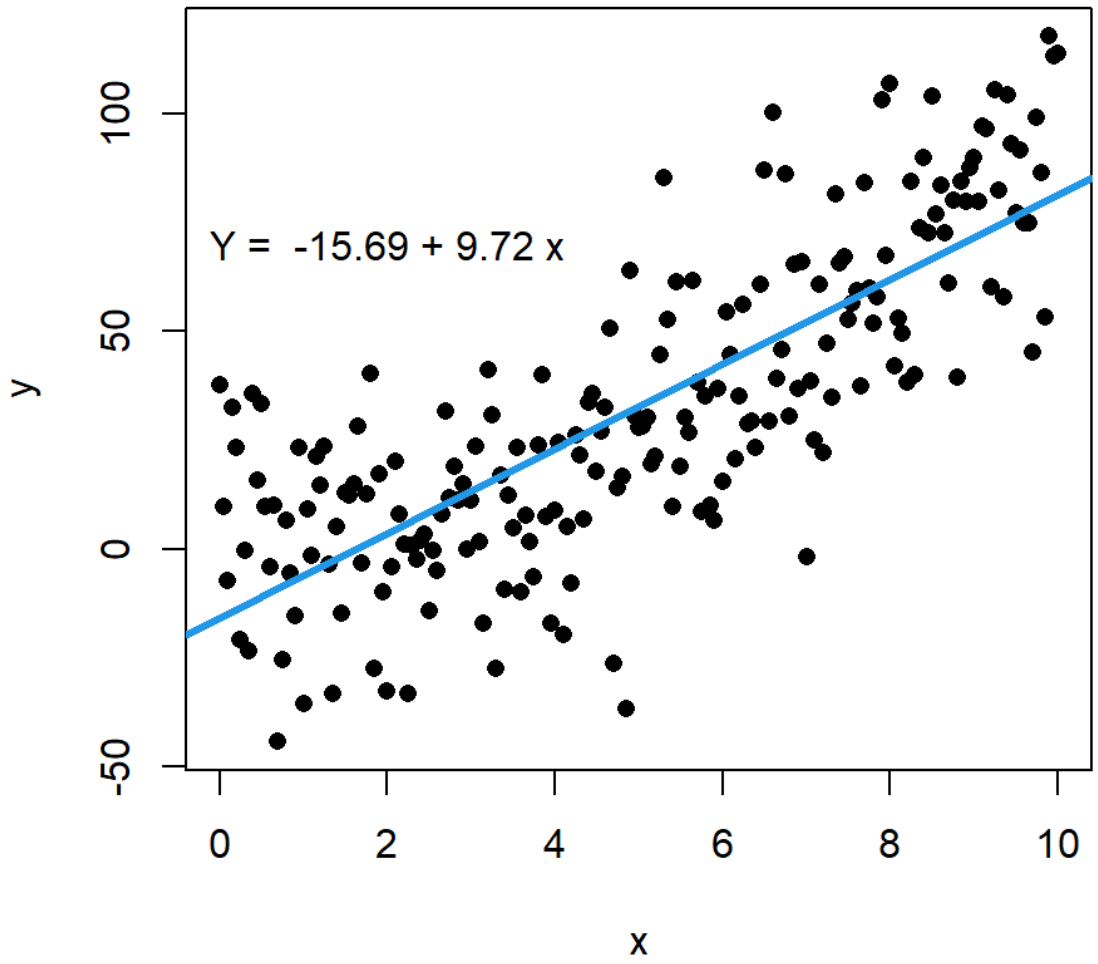

What it is:
Correlation is a statistical measure that expresses the extent to which two variables are linearly related — i.e., when one changes, the other tends to change in a certain way. (JMP)
Why it matters:
In your exams you might be asked: "What is the correlation coefficient?", "Interpret a given r value", or "Why is correlation not the same as causation?"
Key points & how to interpret:

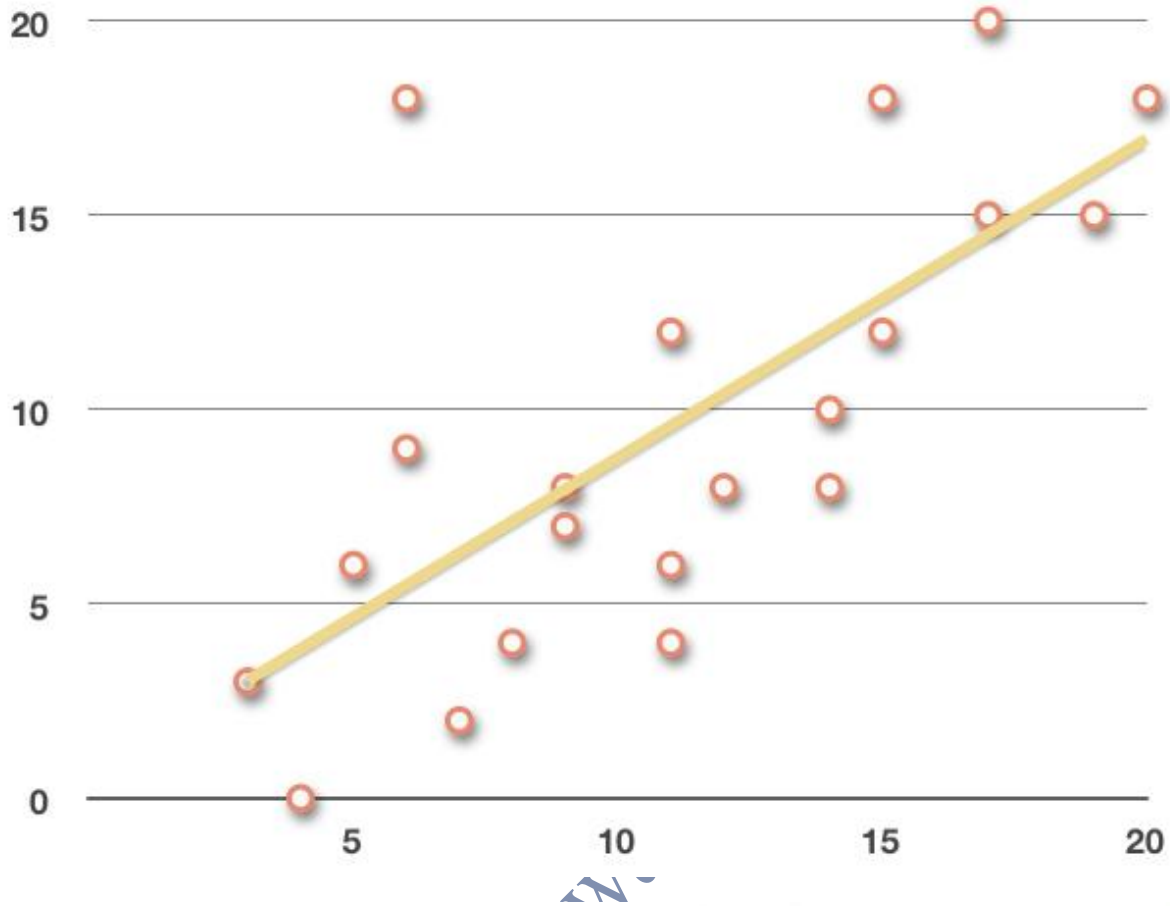
- Correlation coefficient (r) ranges between –1 and +1. (BYJU'S)
- (r = +1) implies perfect positive linear relationship; (r = –1) perfect negative linear; (r = 0) implies no linear relationship.
- Positive (r): as X increases, Y tends to increase. Negative (r): as X increases, Y tends to decrease.
- *Important*: Correlation does not imply cause-and-effect. There might be a lurking third variable, or the relationship might be non-linear (so linear correlation is misleading). (PMC)
  Exam style answer:

"The correlation coefficient r = 0.85 indicates a strong positive linear relationship between X and Y: when X increases, Y tends to increase. However, this does *not* prove that X causes Y — other factors might influence both."
What to include in your notes:

- Definition of correlation
- The formula concept: (r = \dfrac{\text{cov}(X,Y)}{s_X s_Y}) (covariance divided by product of standard deviations) (Statistics LibreTexts)
- Conditions/assumptions: both variables numeric, linear relationship, absence of high outliers or extreme leverage points.
- Many types of correlation (Pearson for linear, Spearman for rank-order) but for your syllabus Pearson is typical.
- Limitations: only linear relationships, sensitive to outliers, doesn't prove causation.

Regression (Simple Linear Regression)

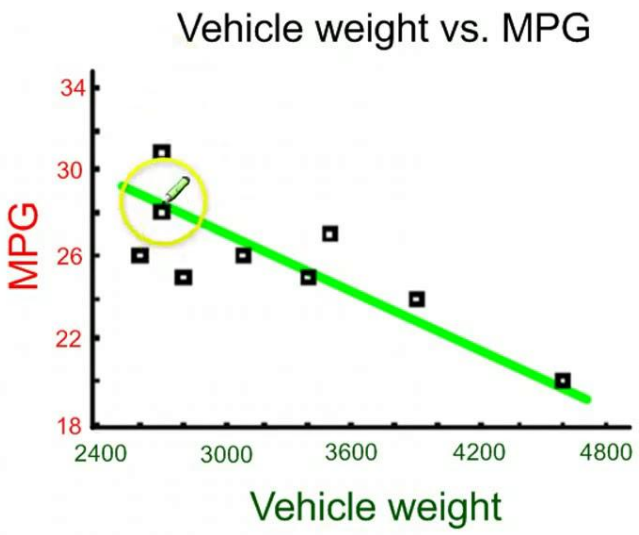

Y = -15.69 + 9.72 x
y
x
100
50
0
-50
0
2
4
6
8
10

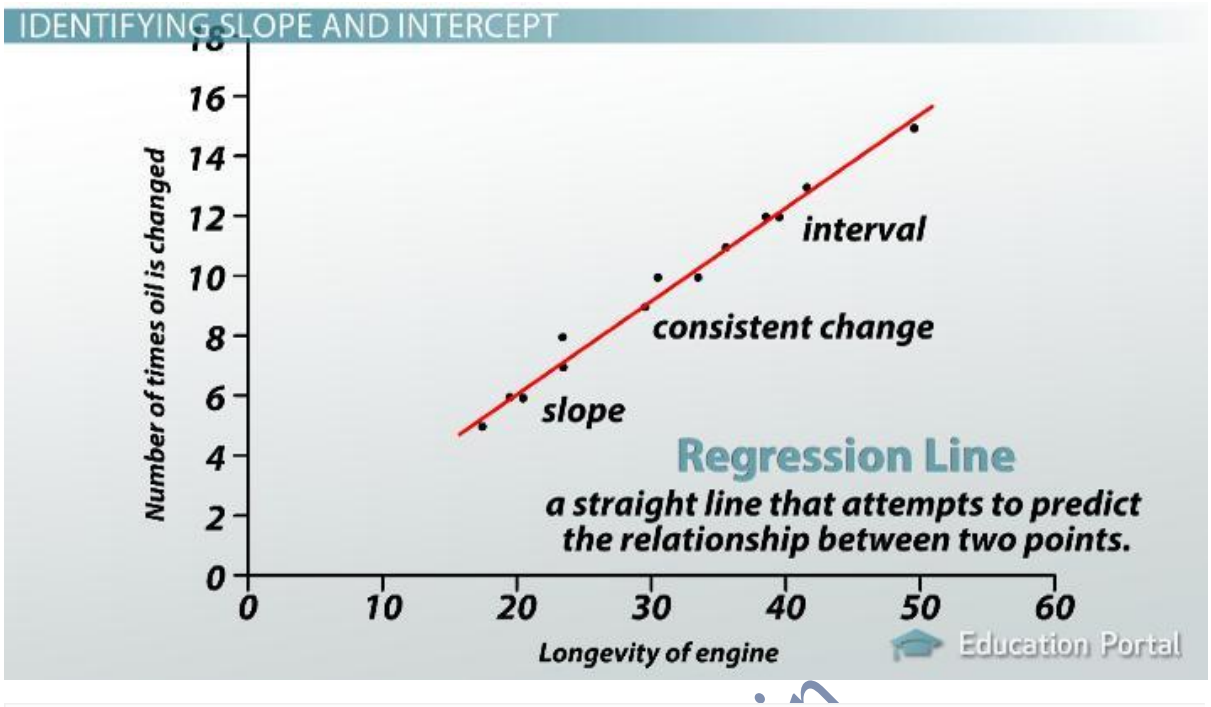


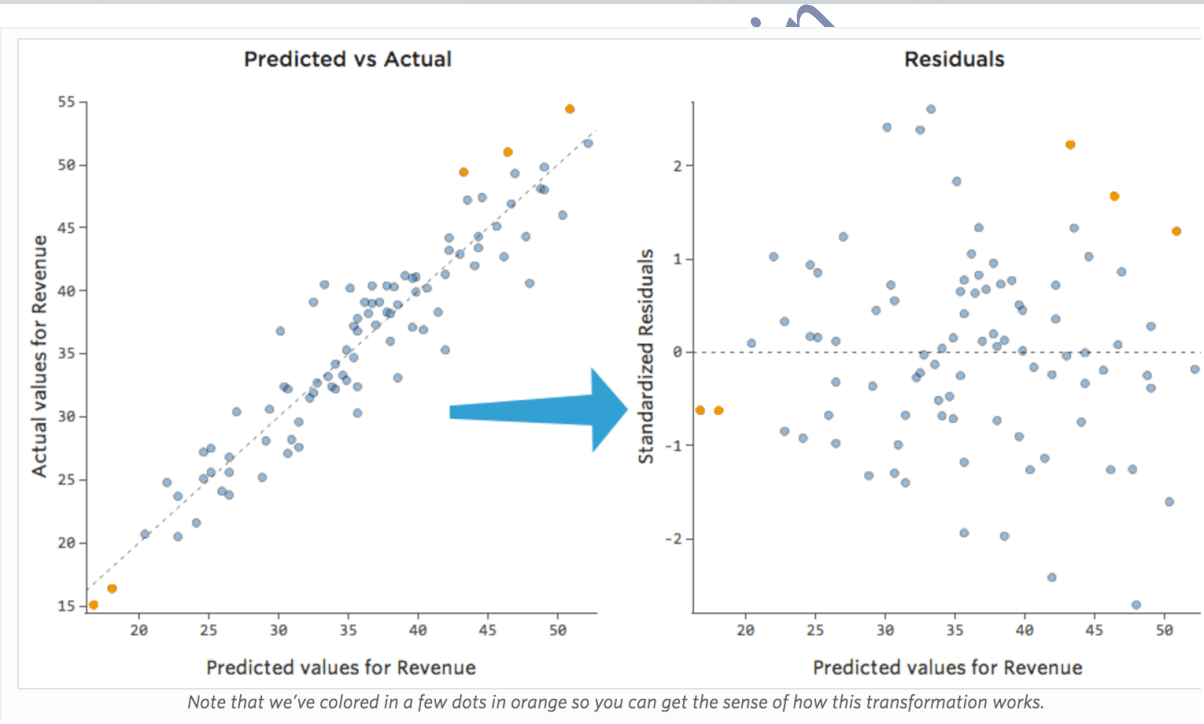

Sketch the [regression line]:

[line of best fit]:

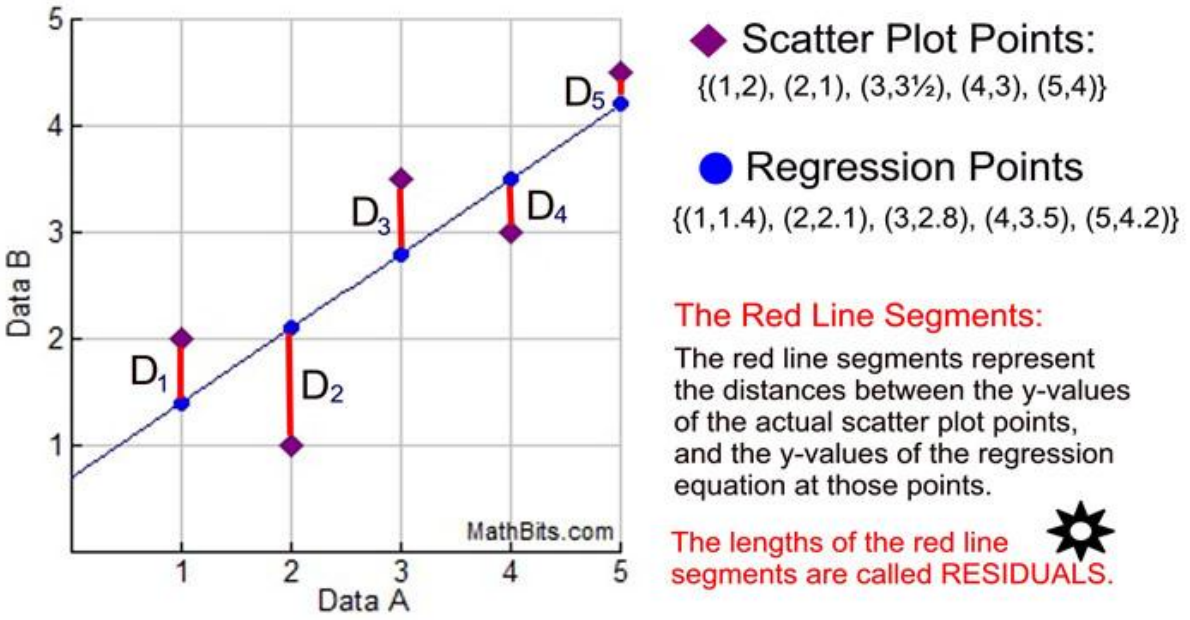
[least squares line]






*Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.*



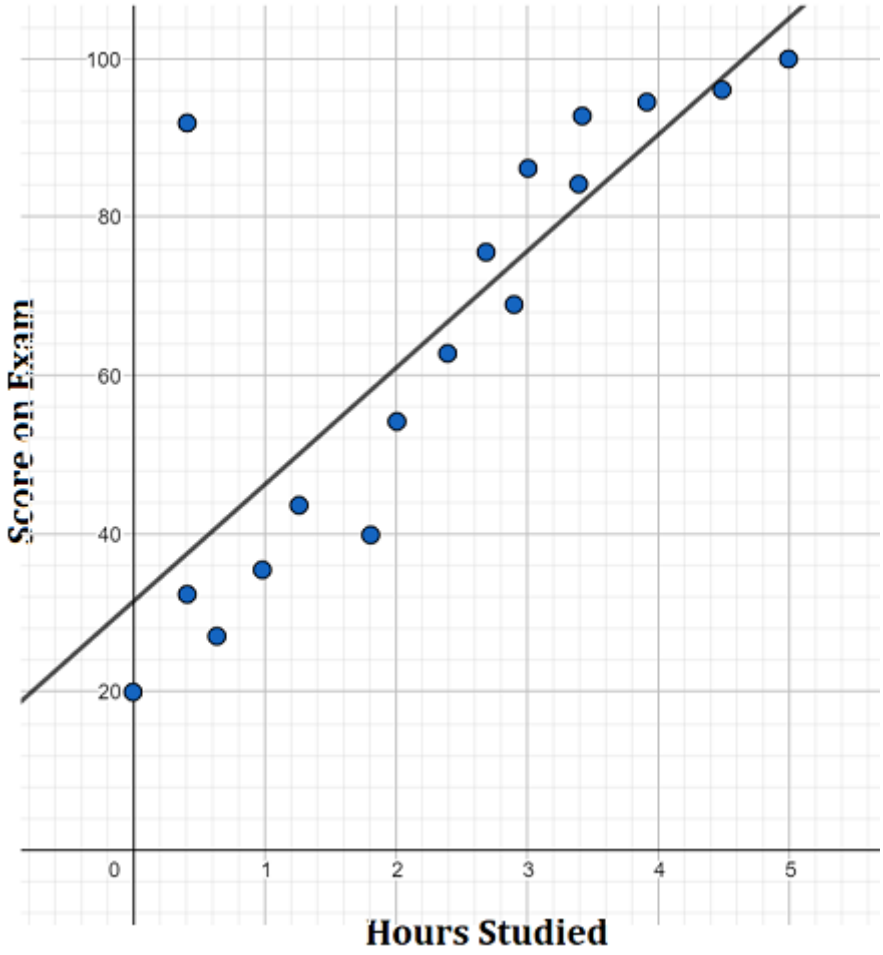

What it is:
Regression is the technique of modelling the relationship between a dependent (response) variable Y and one (or more) independent (predictor) variable(s) X. In simple linear regression, the model is:
[
\hat Y = \beta_0 + \beta_1 X
]
Here, (\beta_0) is the intercept (value of Y when X=0), (\beta_1) the slope (change in Y when X increases by one unit). (University of Oklahoma)
Why it matters:
In your syllabus you will often see: "Given data, fit regression line", "Interpret slope and intercept", "Check assumptions", "Use regression for prediction".
Key points & how to interpret:

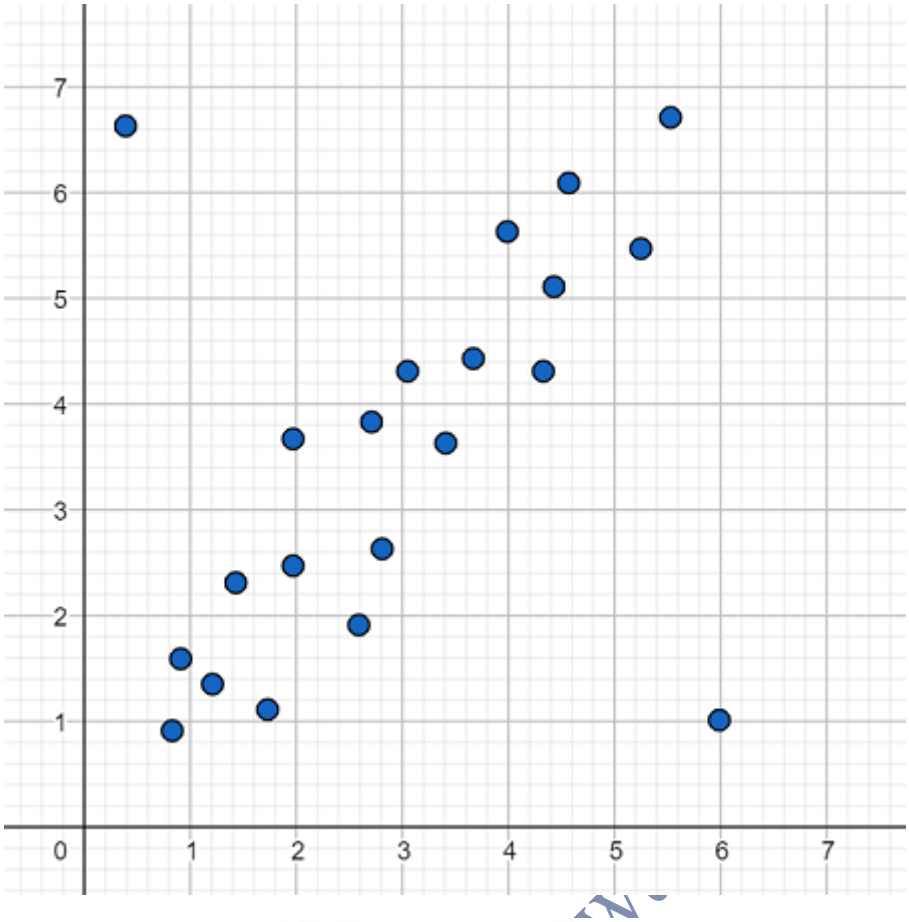
- The slope (\beta_1) tells: for each 1 unit increase in X, the predicted Y changes by (\beta_1) units (on average).
- The intercept (\beta_0) may not always have meaningful interpretation depending on X=0 being in range.
- Goodness-of-fit: $R^2$ statistic tells how much of variation in Y is explained by X (for simple regression, (R^2 = r^2)).
- Important to check assumptions: linear relationship, residuals roughly normally distributed, constant variance (homoscedasticity), independence of observations.
  Exam style answer:

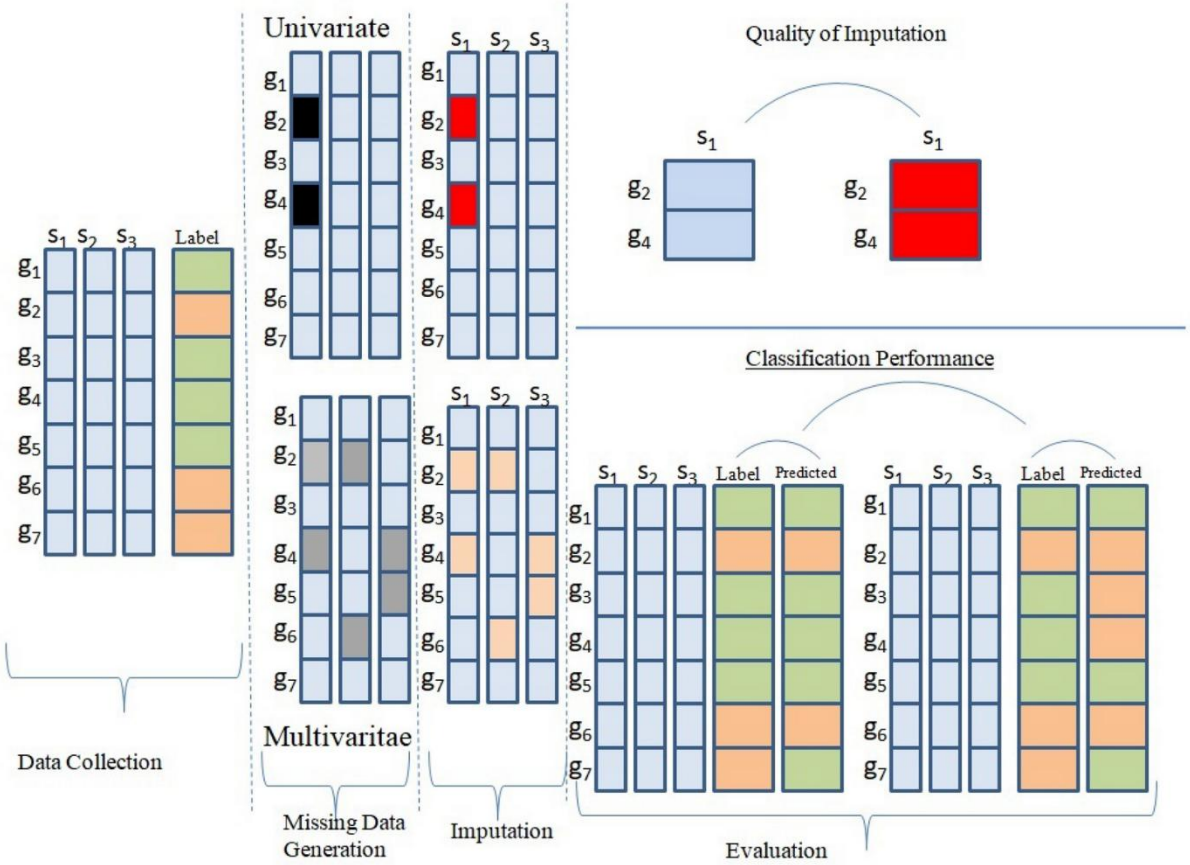
"The regression equation (\hat Y = 20 + 5X) means: when X = 0, predicted Y = 20. For each increase of 1 in X, Y is predicted to increase by 5 units. If we have X = 8, then predicted Y = 20 + 5×8 = 60."

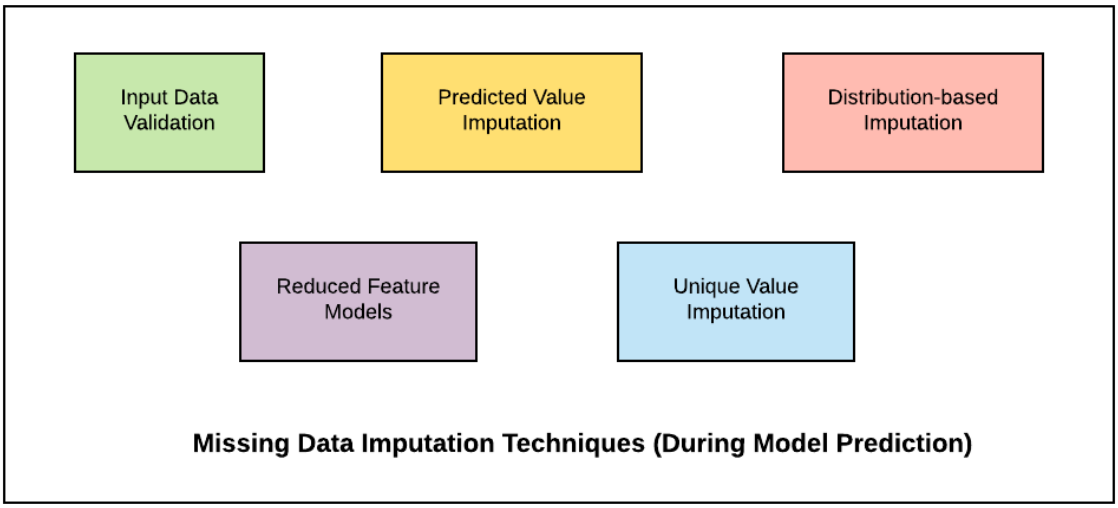
- Then interpret in context: "If X represents hours studied and Y represents test score, then each additional hour studied increases predicted score by 5 points, starting from baseline of 20 when zero hours."
  What to include in your notes:
- Regression vs correlation (correlation measures strength; regression builds predictive equation) (Statistics LibreTexts)
- Formula of regression line, how to calculate coefficients (least squares method) – though full derivation may be beyond syllabus.
- Interpretation of slope & intercept.
- Residuals: difference between observed and predicted Y; residual analysis helps check model validity.
- Predictions & caution: Extrapolation beyond data range is risky; causal inference requires design, not just regression.
- Example question: "Fit a regression line given X & Y values; interpret coefficients; predict Y when X=…"
- Summary of key assumptions and consequences of violation.

Data Cleaning




0
1
2
3
4
5
6
7


Univariate
Multivaritae
Quality of Imputation
Classification Performance
Label
Predicted
Data Collection
Missing Data Generation
Imputation
Evaluation




What it is:
Data cleaning (also called data preparation) involves detecting and fixing or removing errors, inconsistencies, missing values, duplicates, outliers, incorrect formats in your dataset before you apply statistical methods.
Why it matters:
If you apply correlation/regression (or any statistical method) on dirty data, results may be misleading (garbage in → garbage out). In an exam you might be asked: "Why is data cleaning important?" "What steps would you take before running regression?"
Key steps & how to do:

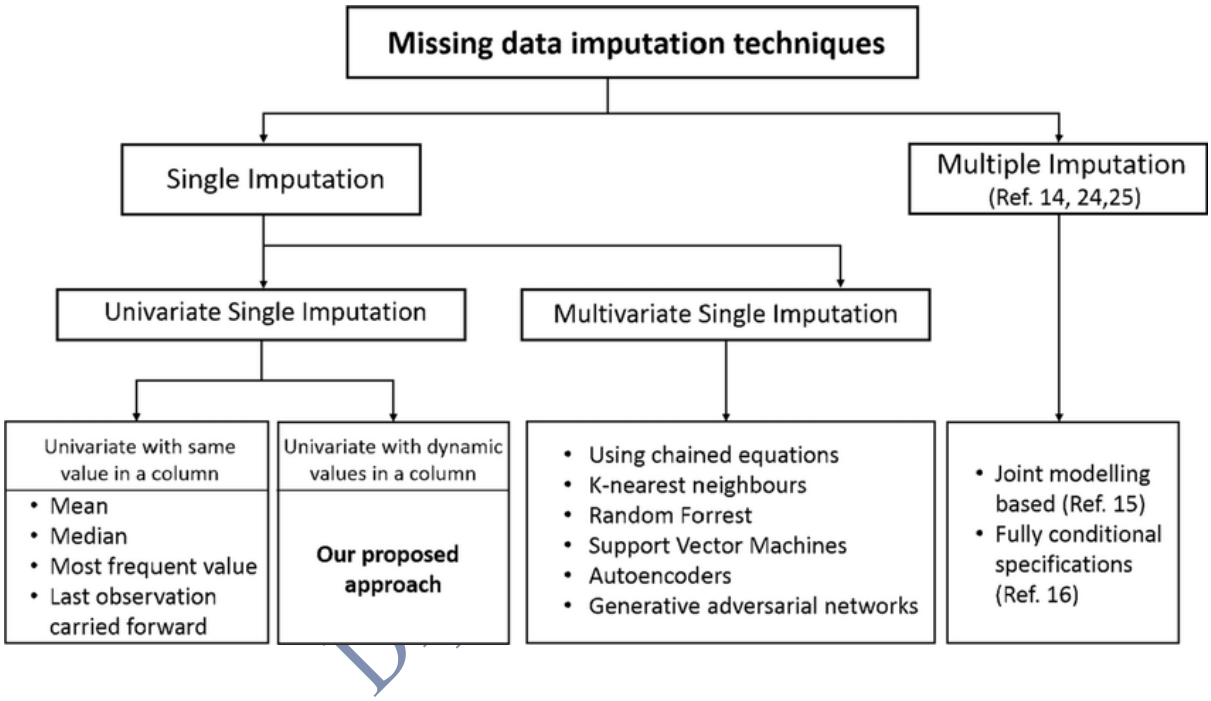
- Missing values: Decide whether to omit records, impute missing values (with mean/median, regression imputation, KNN, multiple imputation).
- Outliers/influential points: Use box plots, scatter plots to identify extreme values; decide if they are errors (typos) or legitimate extremes; maybe remove or transform.
- Data type/format issues: Ensure variables are measured consistently (units, scales), categories coded properly.
- Duplicates: Remove exact duplicates if they represent redundant entries.
- Consistency checks: E.g., ages negative? Scores > max? Fix or remove.
  Exam style answer:

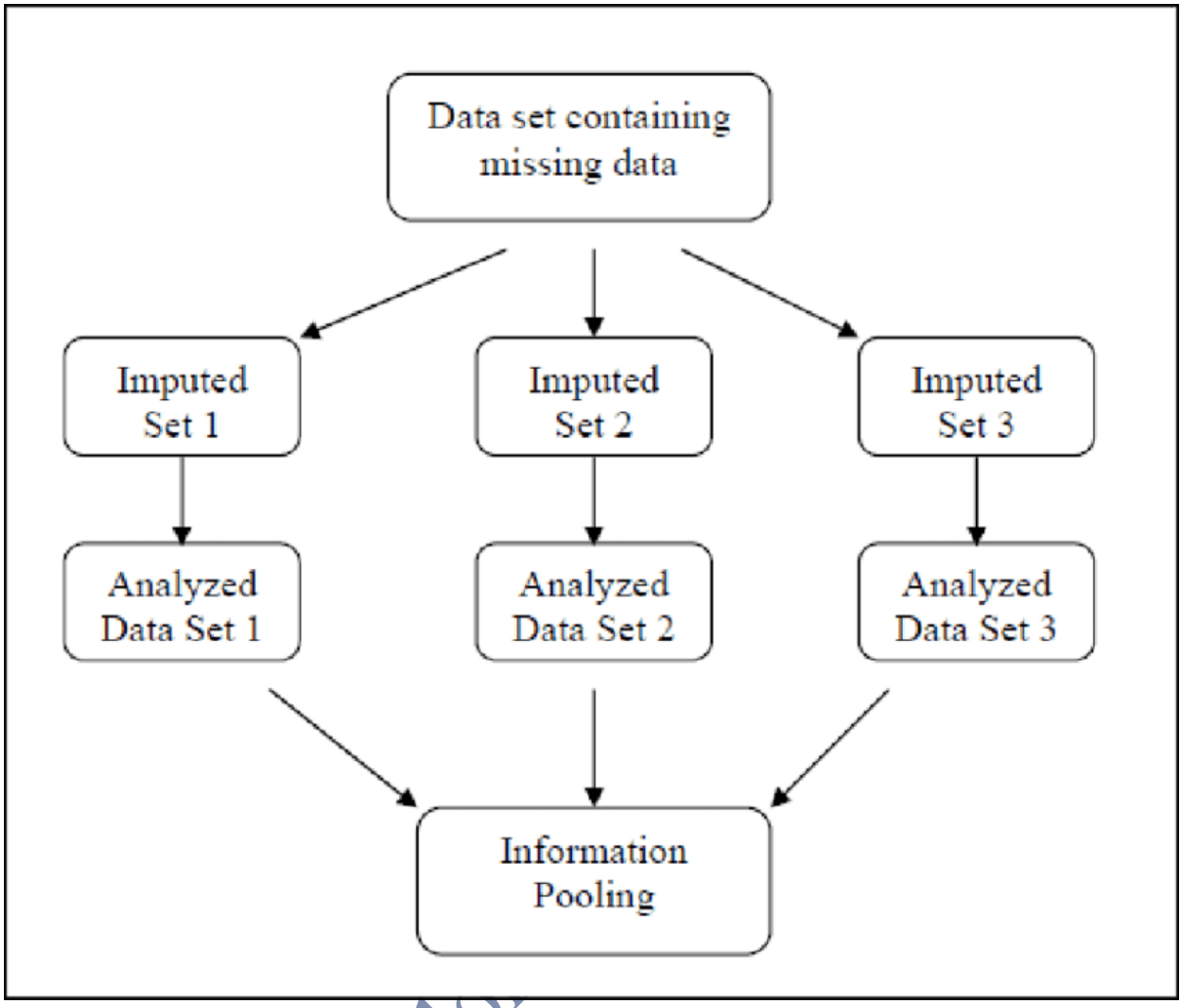
"Before fitting a regression model I would inspect for missing values and outliers. I would create summary statistics and box plots to check for extreme values; if a data point lies far from others I would investigate if it's a data-entry error. I would also check that variables are coded correctly (e.g., numeric rather than text), and that units are consistent. Cleaning reduces bias and improves reliability of the model fit."
What to include in your notes:

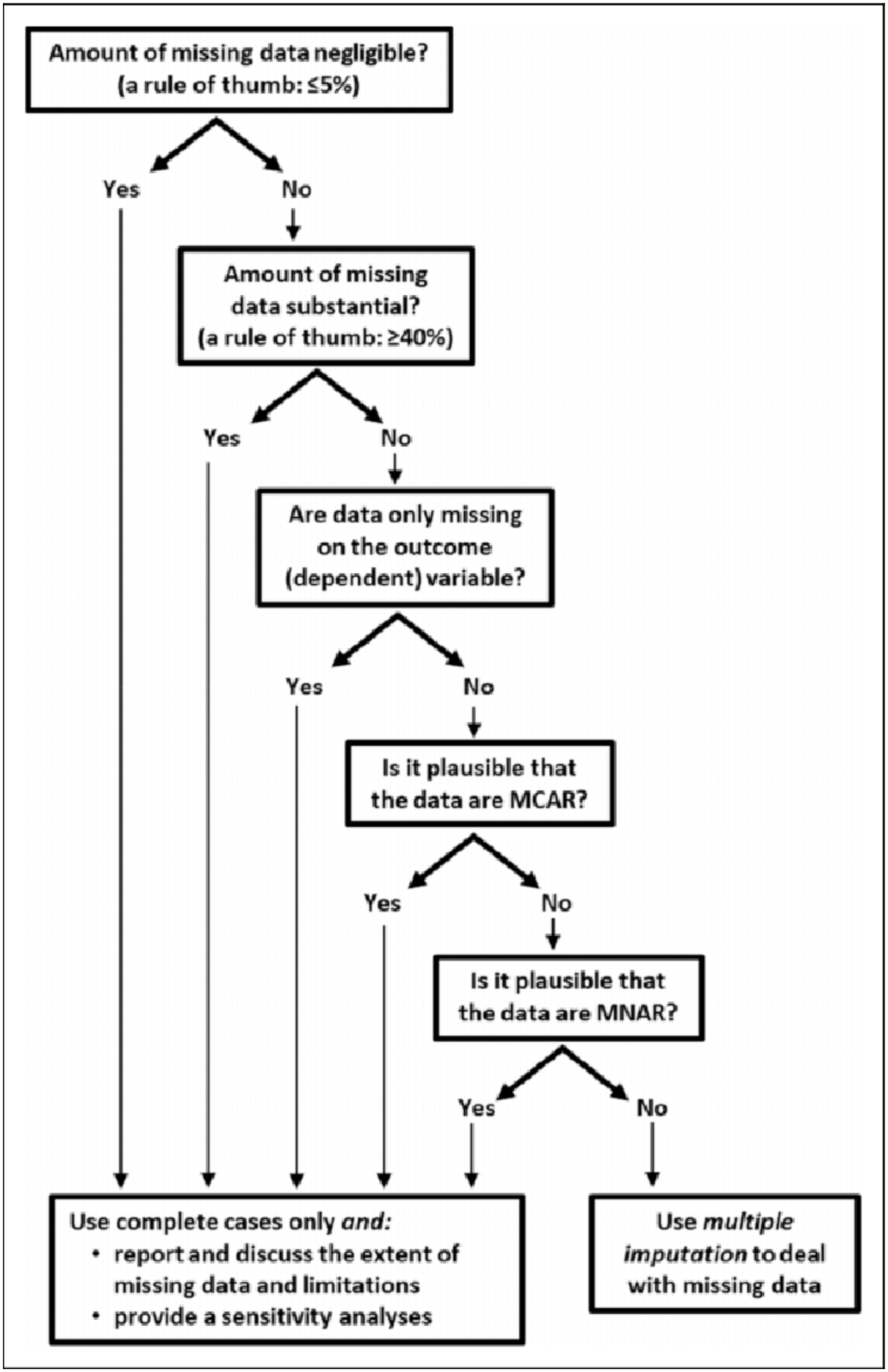
- Definition of data cleaning and why it matters.
- List of typical tasks (missing values, outliers, duplicates, formatting).
- Links between cleaning and later stages (esp. regression).
- Example of how an outlier can distort correlation/regression (e.g., one extreme value can inflate correlation).
- Reminder: data cleaning is essential, not optional.

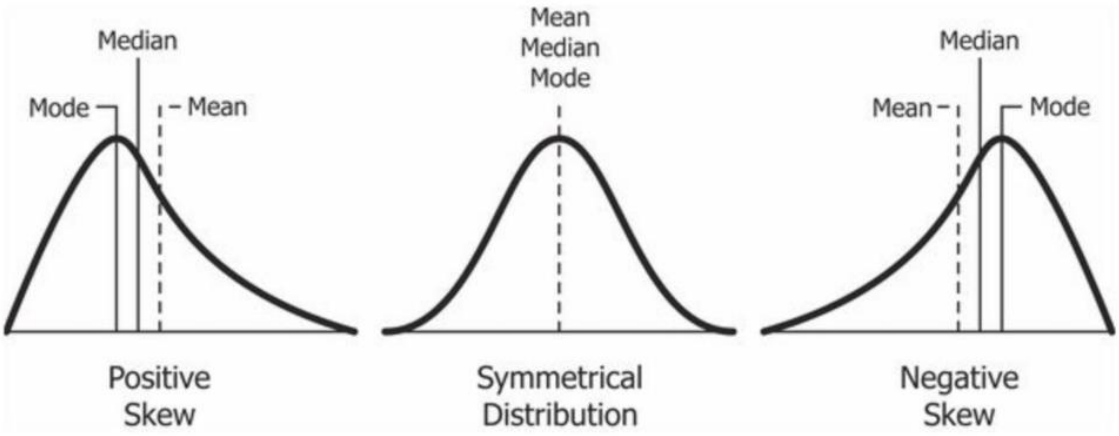
---

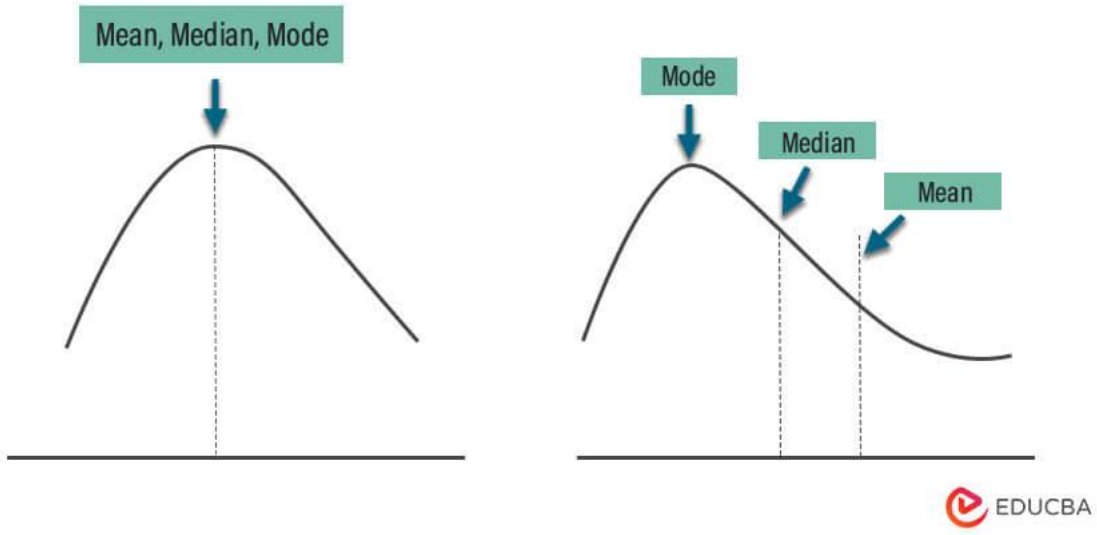
## 2.3 Imputation Techniques




Data set containing missing data
Imputed Set 1
Imputed Set 2
Imputed Set 3
Analyzed Data Set 1
Analyzed Data Set 2
Analyzed Data Set 3
Information Pooling


Amount of missing data negligible? (a rule of thumb: ≤5%)
Yes
No
Amount of missing data substantial? (a rule of thumb: ≥40%)
Yes
No
Are data only missing on the outcome (dependent) variable?
Yes
No
Is it plausible that the data are MCAR?
Yes
No
Is it plausible that the data are MNAR?
Yes
No
Use complete cases only and:
• report and discuss the extent of missing data and limitations
• provide a sensitivity analyses
Use multiple imputation to deal with missing data




What it is

Imputation is the process of replacing missing values in a dataset with estimated values so that you can perform analyses on a "complete" data set rather than deleting many records. (Wikipedia)

Why it matters

If you have missing data and you ignore the problem (just remove records with missing values), you risk: reducing your sample size, introducing bias (if the missingness is not random), and messing up your statistical power. Imputation helps keep data usable. (Simplilearn.com)

Key methods & how to interpret/explain

- Mean/median/mode imputation: Replace missing numeric values with the variable's mean or median; or categorical with mode. Simple, but reduces variability and may bias results. (The Analysis Factor)

- Regression imputation: Use other variables to predict missing ones, e.g., fit a regression model.
- K-Nearest Neighbours imputation (KNN): Use similar observations (neighbors) to impute missing values.
- Multiple imputation: Create several imputed datasets, analyse each, then combine results — accounts for uncertainty from imputing. (PMC)

Exam style answer

"If missing values exist in a dataset, I would first determine the mechanism of missingness (MCAR, MAR, MNAR). If the data are missing completely at random (MCAR), I might use mean imputation for simplicity; but if missingness is related to other variables (MAR), I would prefer multiple imputation because it better preserves variability and reduces bias."

Tips for your notes

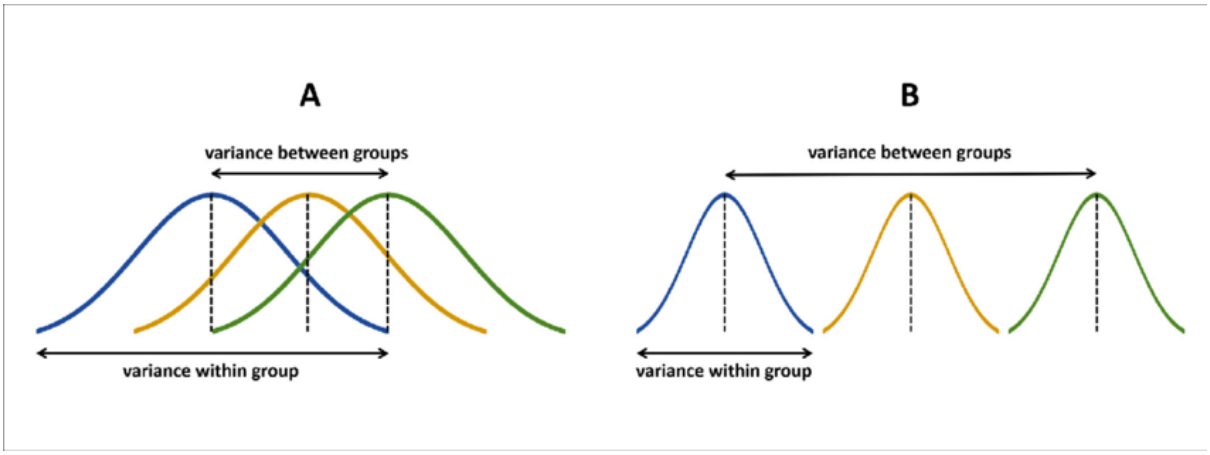
- Define imputation clearly.
- List 2-3 methods with pros/cons.
- Mention why choice of method matters depending on missingness mechanism.
- If asked: "Explain one imputation technique and its limitation" — choose something like mean imputation and say "it reduces variance, treats all missing as identical, may bias estimates".

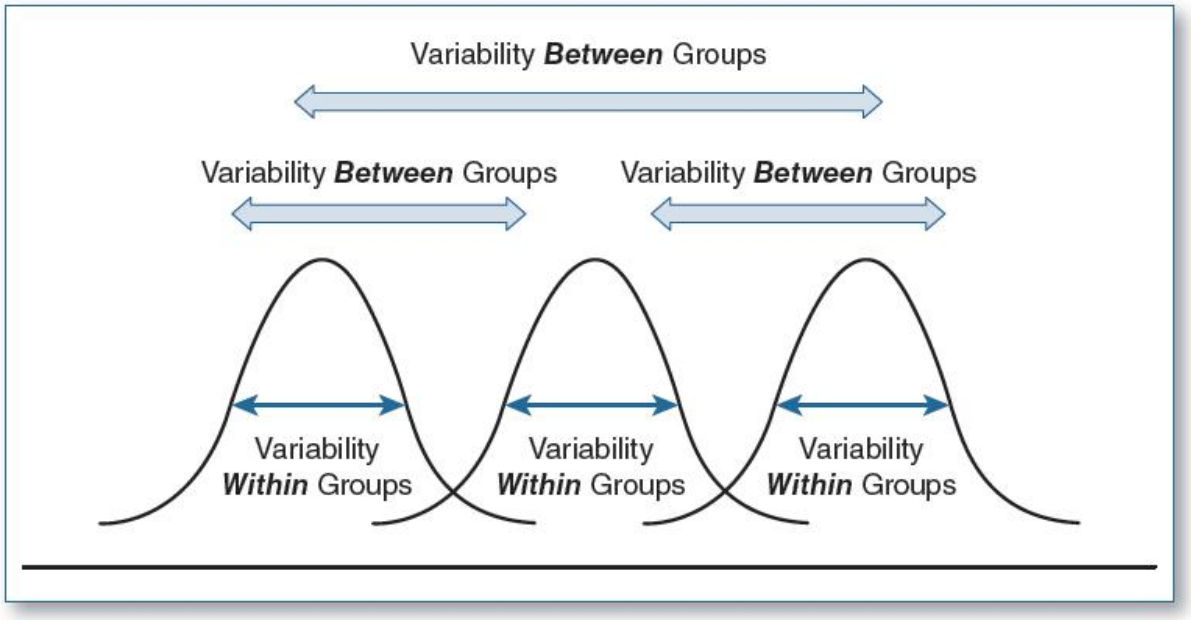


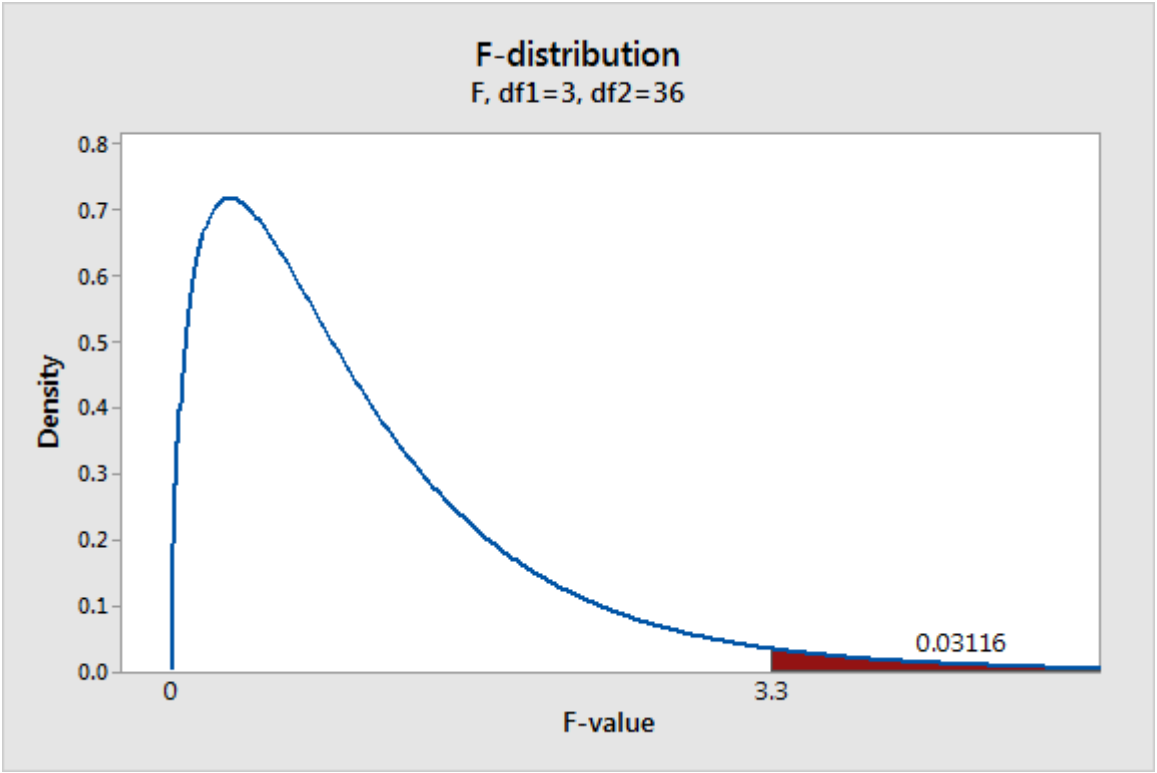
---

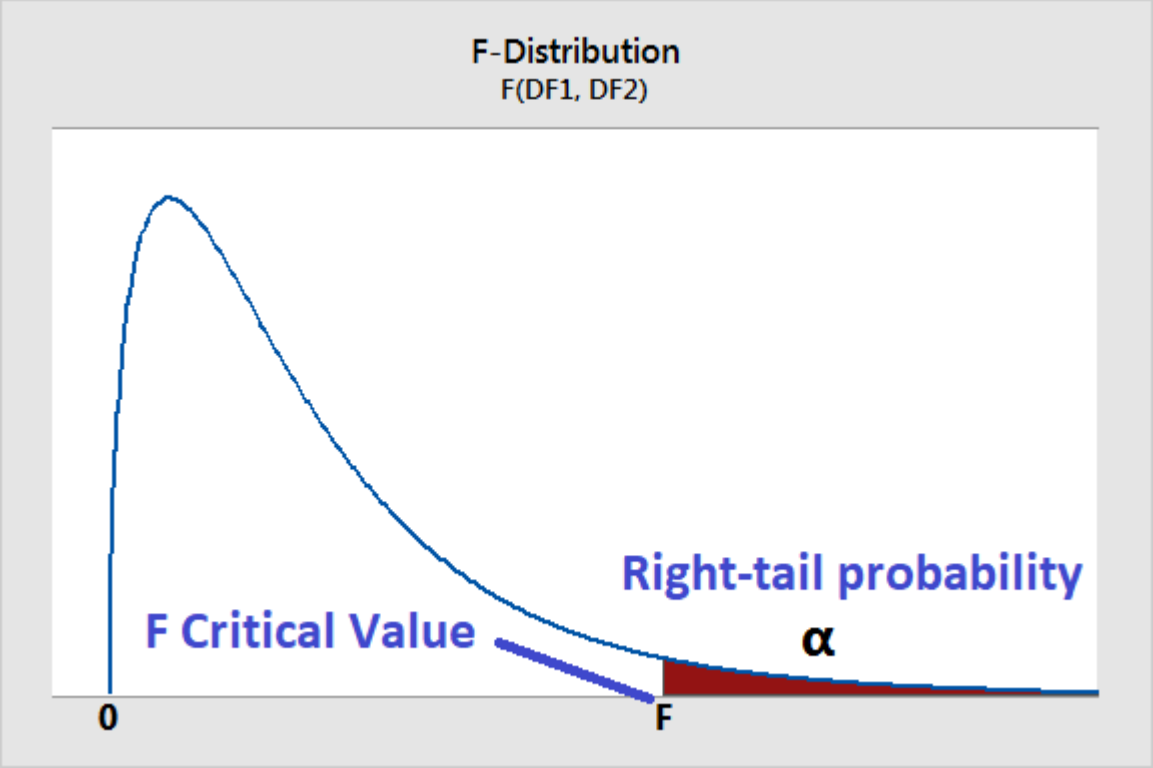
2.4 ANOVA and Chi-Square

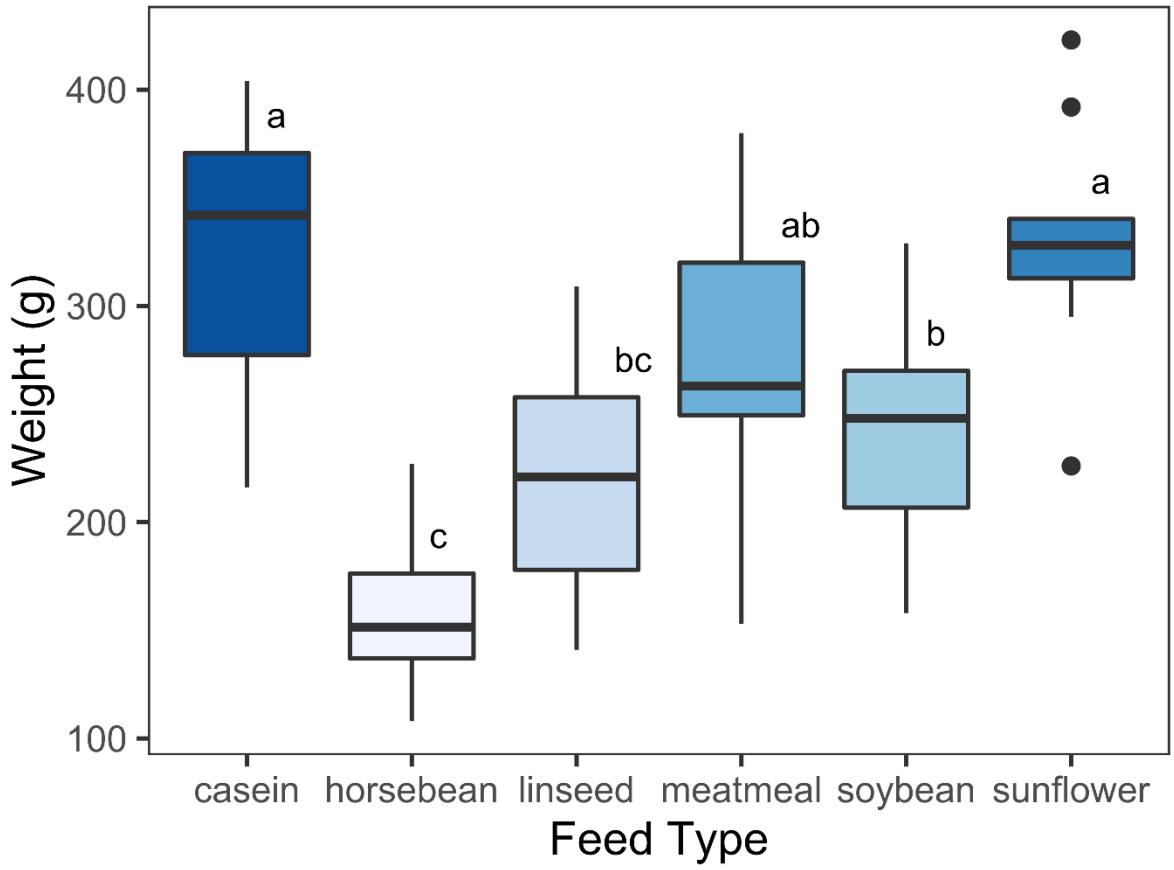
ANOVA (Analysis of Variance)

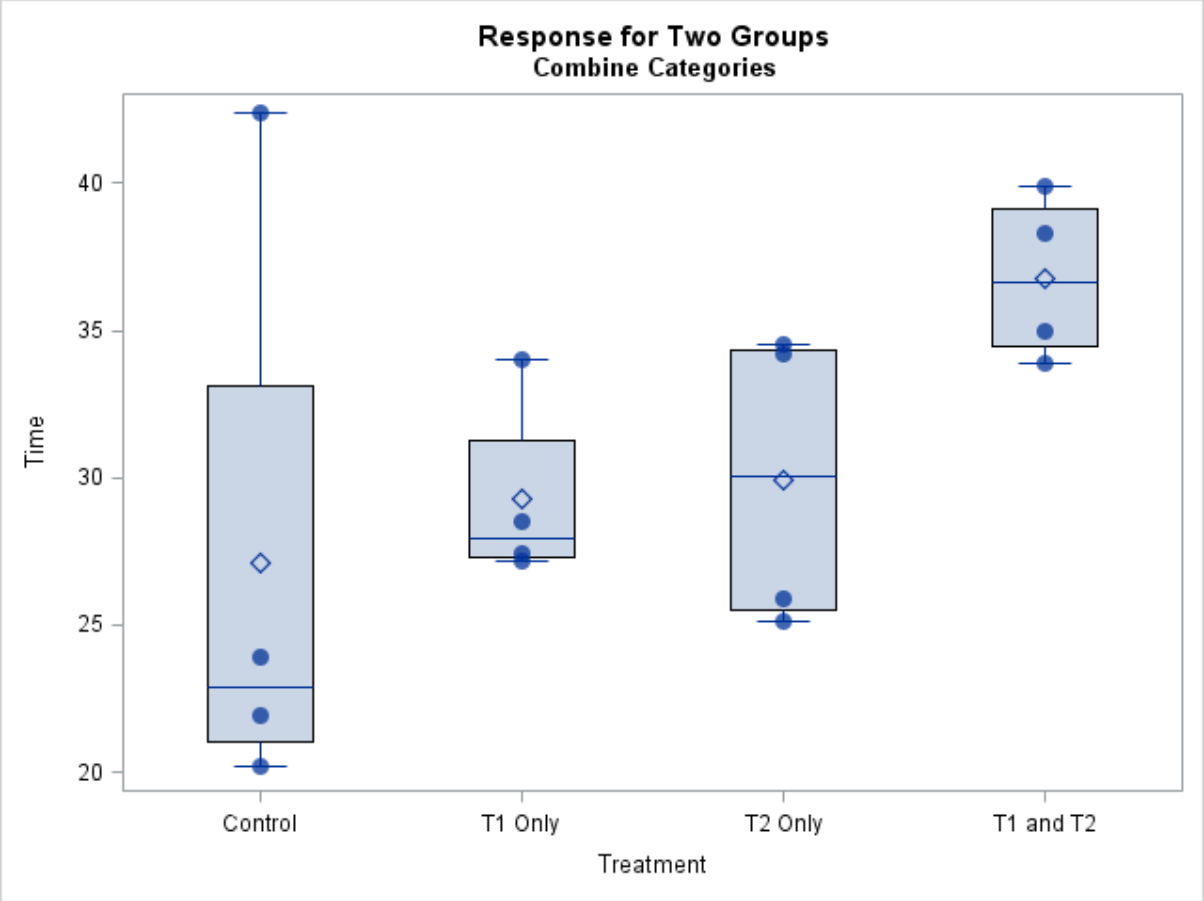



Variability Between Groups
Variability Between Groups
Variability Between Groups
Variability Within Groups
Variability Within Groups
Variability Within Groups


F-distribution
F, df1=3, df2=36
Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
3.3
0.03116
F-value


F-Distribution
F(DF1, DF2)
Right-tail probability
F Critical Value
α
0
F


Weight (g)
400
300
200
100
a
c
bc
ab
b
a
casein
horsebean
linseed
meatmeal
soybean
sunflower
Feed Type

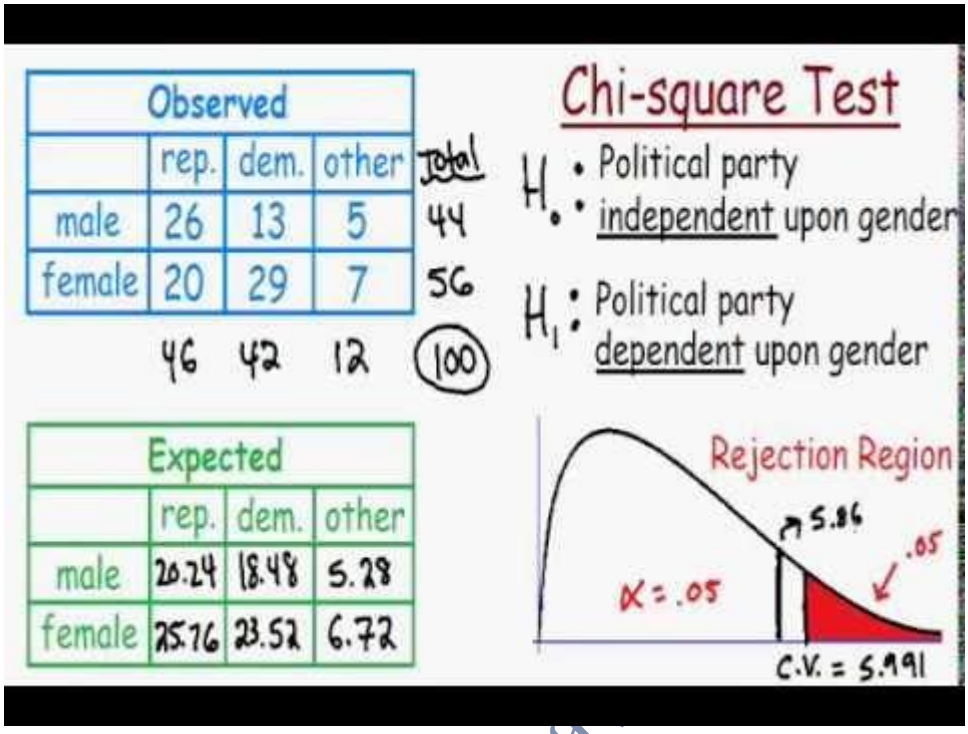



What it is:
ANOVA is a statistical procedure used when you want to compare the means of three or more groups to see if at least one group mean differs from the others. (Dutton Institute)
Why it matters:
In exams you may be given data for, say, three teaching methods and student scores, and asked: "Use one-way ANOVA to test whether teaching method affects score." You need to know how to set up hypotheses, compute F-statistic (conceptually), and interpret result.
How to interpret/explain:

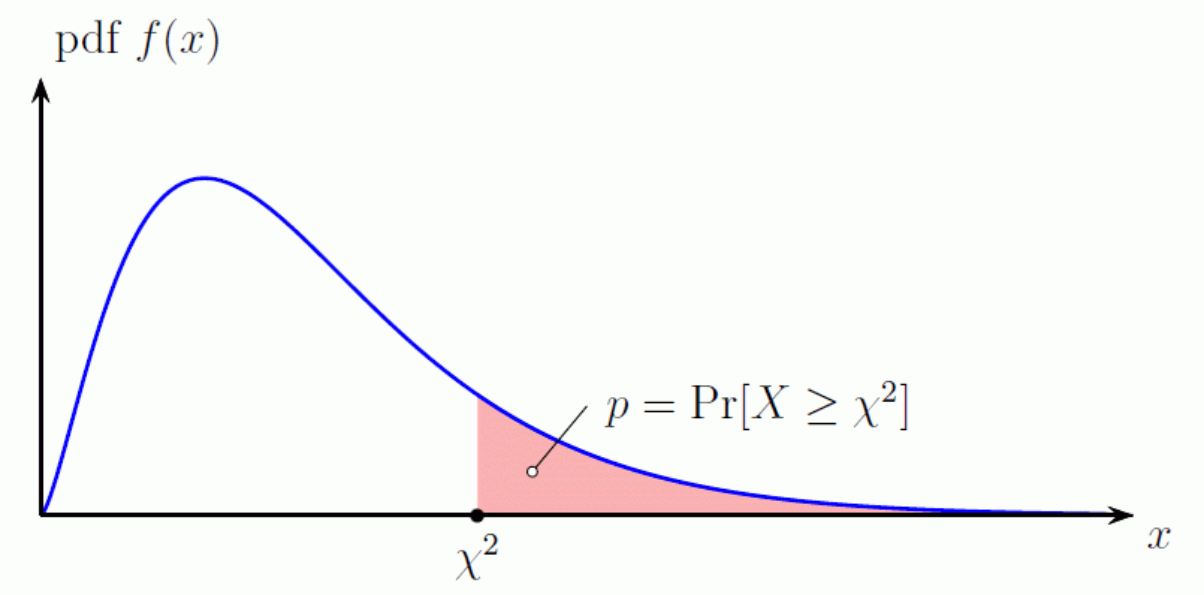
- Null hypothesis (H_0): All group means equal ($\mu_1 = \mu_2 = \mu_3 = \ldots$).
- Alternative (H_a): At least one mean differs.
- F = (Between-group variance) / (Within-group variance).
- If F is large (or p-value < α) → reject (H_0): means are not all equal.
  Assumptions: observations independent; residuals roughly normal; homogeneity of variances.
  Exam style answer:

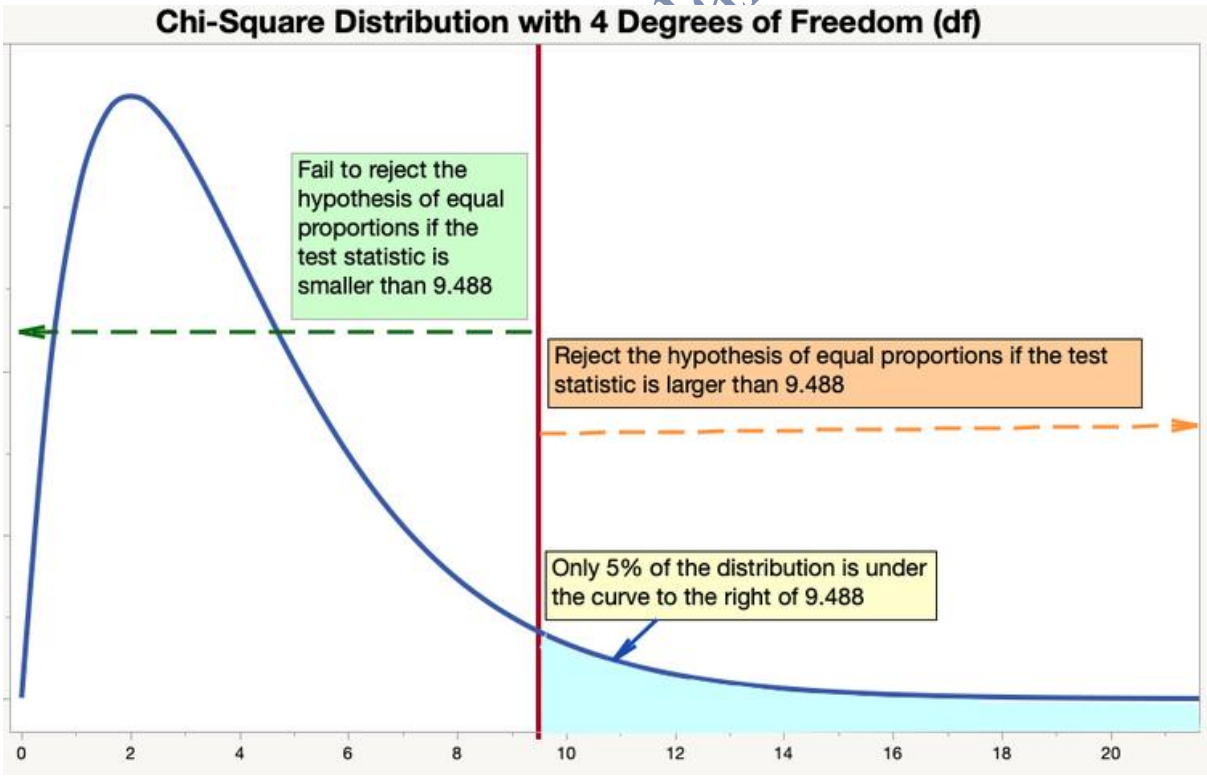
"At α = 0.05, we computed F = 6.12 with p = 0.004. Because p < 0.05, we reject $H_0$ and conclude there is a significant difference between the group means. To find which groups differ, a post-hoc test (e.g., Tukey) is needed."

Chi-Square Test




| | | Sport Preference | | | |
|---|---|---|---|---|---|
| | | Archery | Boxing | Cycling | |
| Gender | Female | 35 | 15 | 50 | 100 |
| | Male | 10 | 30 | 60 | 100 |
| | | 45 | 45 | 110 | 200 |




What it is:
The chi-square ($\chi^2$) test is used with categorical data. Two major uses:

1. Test of independence: are two categorical variables independent? (Scribbr)
2. Goodness-of-fit: does observed distribution match expected?
   Why it matters:

If you have categorical variables (e.g., gender and preference), exam might ask: "Perform $\chi^2$ test of independence and interpret result."
How to interpret/explain:

- Null (H_0): variables are independent (or observed distribution equals expected).
- Alternative (H_a): not independent (or observed ≠ expected).
- Formula: (\chi^2 = \sum \frac{(O - E)^2}{E}) where O = observed freq, E = expected freq. (Simplilearn.com)
- Degrees of freedom for test of independence: (rows-1)×(columns-1). (Hugo Portfolio Theme)
- If $\chi^2$ > critical value or p < α → reject (H_0).
When to use which test:
- Use ANOVA when dependent variable is numeric, independent is categorical (3+ groups). (Statology)
- Use $\chi^2$ when both variables are categorical.
Exam style answer:

"In the contingency table (2×3), df = (2-1)*(3-1) = 2. The computed $\chi^2 = 5.92$ with p = 0.051. At α = 0.05 we fail to reject $H_0$, so there is no significant association between the variables."

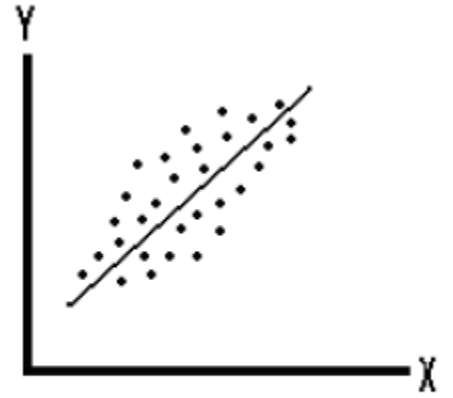
Tips for your notes

- Include clear hypotheses, formulas, interpretations.
- Provide a small example for each test.
- Mention assumptions and when each test is appropriate.
- Use a short table comparing ANOVA vs $\chi^2$ (data type, question asked, what the test checks).



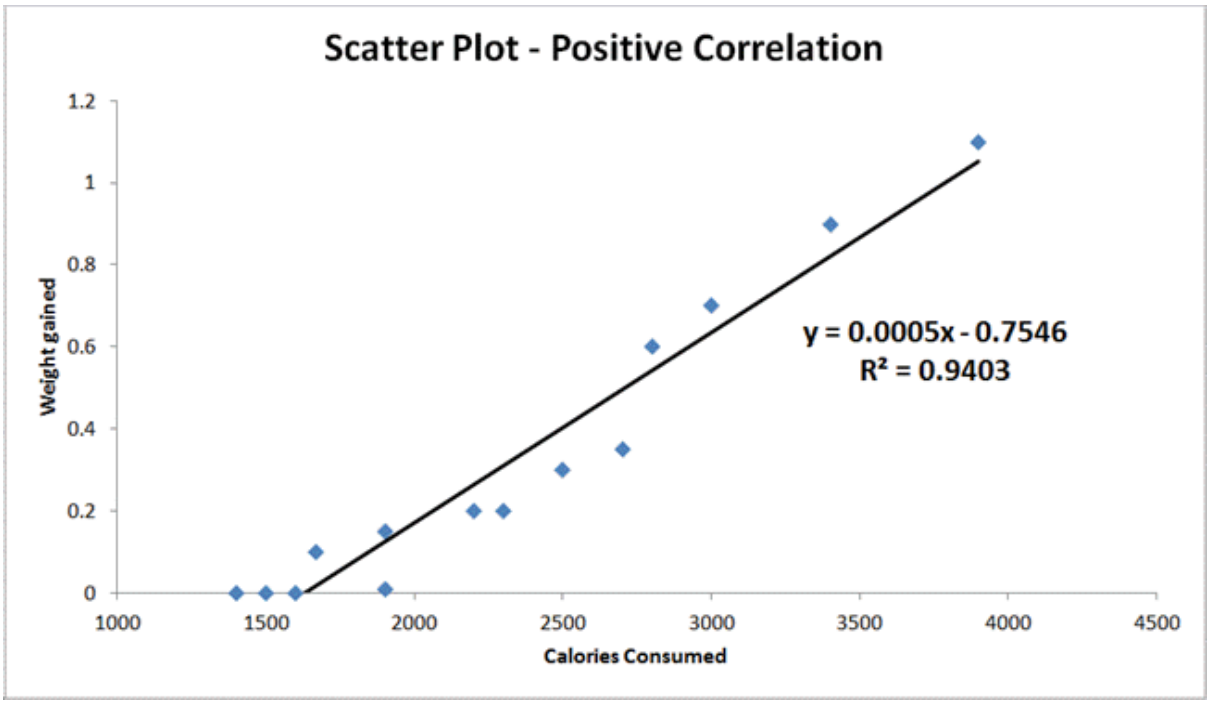
---

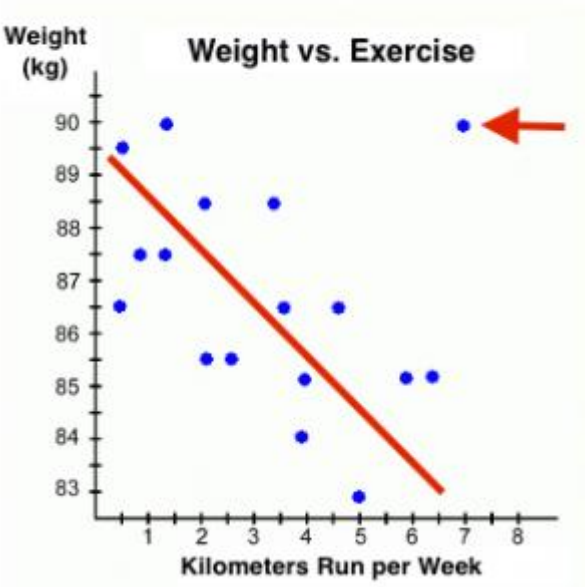
2.5 Scatter Diagram

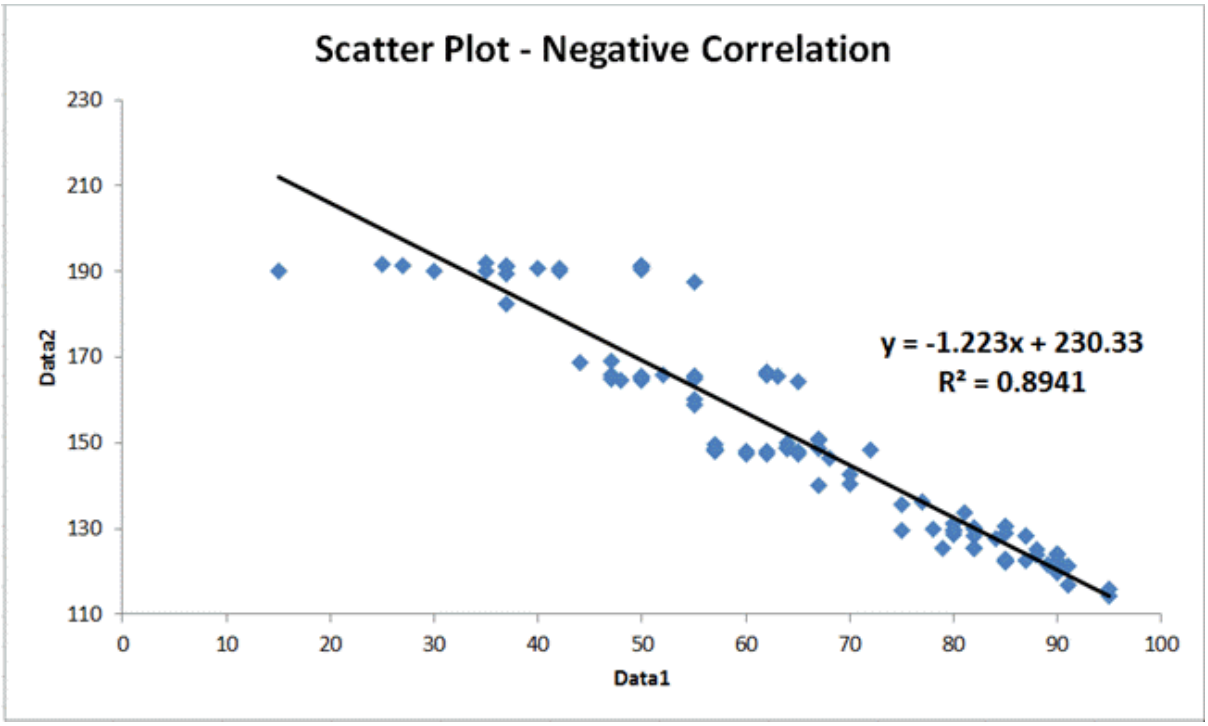


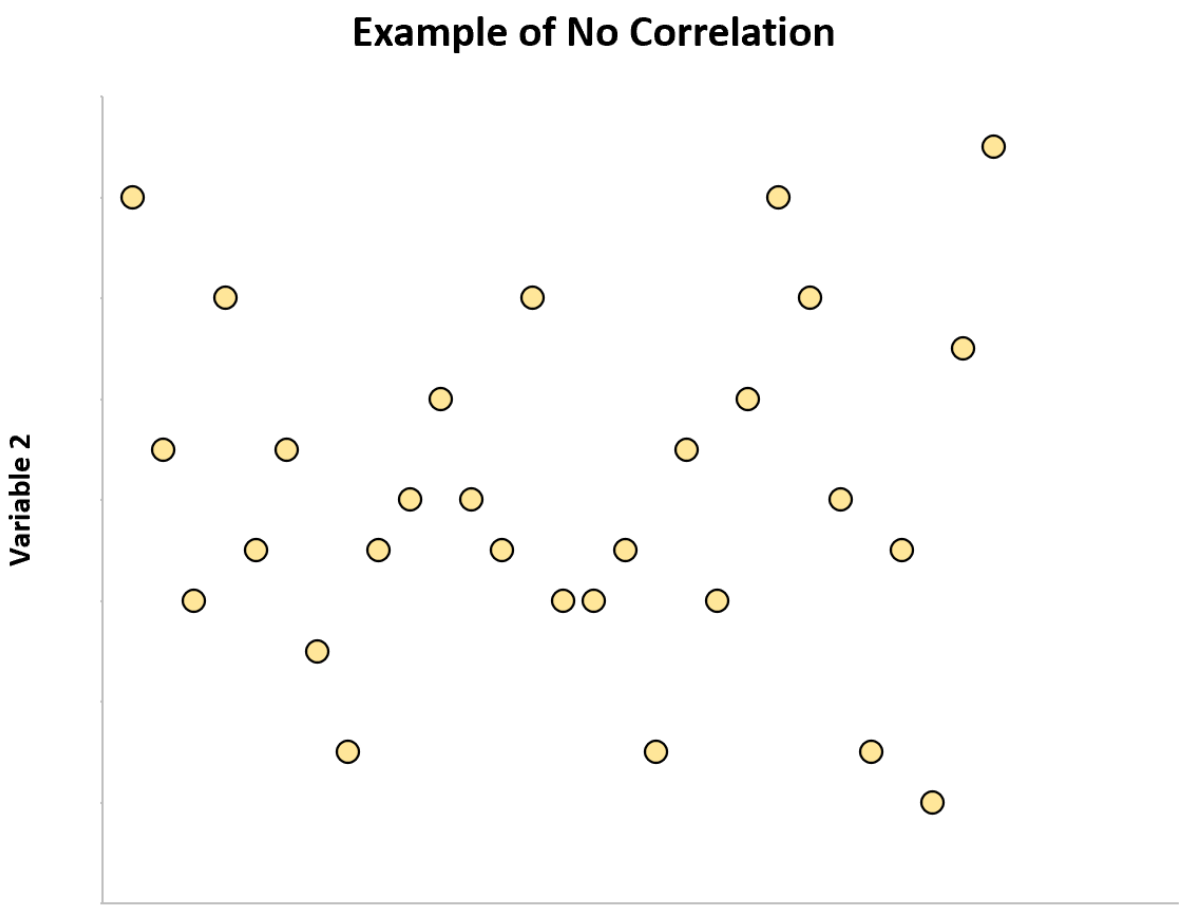

Positive Correlation

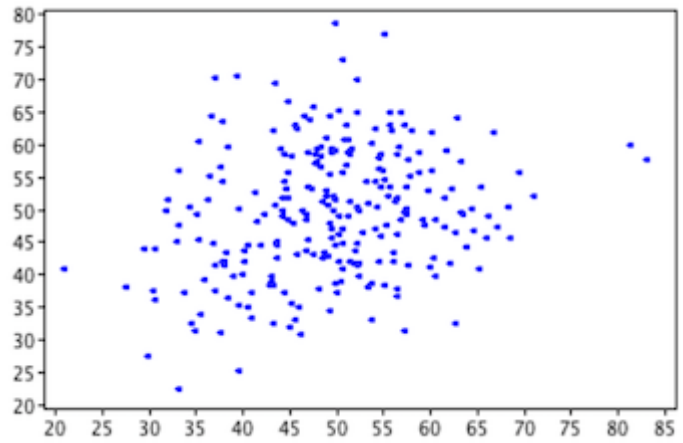

Negative Correlation




Weight vs. Exercise
Weight (kg)
90
89
88
87
86
85
84
83
1
2
3
4
5
6
7
8
Kilometers Run per Week


Scatter Plot - Negative Correlation
Data2
230
210
190
170
150
130
110
0
10
20
30
40
50
60
70
80
90
100
Data1
y = -1.223x + 230.33
R² = 0.8941




What it is:
A diagram that plots one numeric variable on the X-axis and another on the Y-axis, each point representing one observation. You can use it to visualise relationship between the two variables.
Why it matters:
Before doing correlation/regression, you should inspect the scatter plot to understand the form (linear or non), direction, strength, presence of

outliers/clusters. In exam you might be asked to "Draw a scatter diagram for these data" or "Interpret this scatter diagram."
How to interpret/explain:

- Direction: upward slope → positive association; downward → negative; no slope → weak/no linear relation.
- Strength: points tightly clustered around an imaginary line → strong; widely scattered → weak.
- Form: Straight line fit possible → linear; curved → non-linear relationship; no pattern → no relation.
- Outliers: Points far from cluster; may distort correlation/regression.
  Example exam answer:

"The scatter diagram shows a clear upward trend: as X increases, Y tends to increase (positive relationship). The points are fairly close to a straight line, indicating a strong linear association. One point lies far away from the cluster (outlier) which may influence a regression line; hence I would check for influence before modelling."

Tips for your notes

- Mention what each axis represents.
- Show small examples (positive, negative, none).
- Explain next steps: "If scatter suggests linear relationship, then compute correlation and possibly fit regression; if form not linear or outliers present, consider non-parametric or transformation."
- Include in exam tips: Always label axes, no lines connecting points, comment on trend, clusters, outliers.



---

2.6 Estimation and Hypothesis Testing

Estimation

What it is:
Estimating population parameters (like mean μ, proportion p) from sample statistics (like sample mean (\bar x), sample proportion (\hat p)). Two types:

- Point estimate (single value).
- Interval estimate (confidence interval) which gives range with a stated level of confidence.
  Why it matters:
  Exams may ask: "Compute a 95 % confidence interval for the mean", "What is

point estimate vs interval estimate?", or "Interpret this confidence interval."
How to explain:

- If (\bar x = 50), sample size n = 100, σ known =10 → 95% CI: (50 ± (1.96×10/√100) = 50 ±1.96).
- Interpretation: "We are 95% confident the true population mean lies between 48.04 and 51.96."
- Emphasise: "Confidence" refers to method, not saying that the parameter has 95% probability in that interval (once interval calculated, parameter either is or isn't).

Hypothesis Testing

What it is:
A process for deciding whether sample data provide enough evidence to reject a claim (null hypothesis) about a population.
Why it matters:
Exam questions often: "State $H_0$ and $H_1$", "Calculate test statistic", "Make decision at α = 0.05", "Interpret result."
How to explain:

- Step 1: State hypotheses. e.g., (H_0: \mu = 100), (H_a: \mu ≠ 100).
- Step 2: Choose significance level (α).
- Step 3: Compute test statistic (z, t, etc).
- Step 4: Find critical value(s) or p-value.
- Step 5: Compare: if p < α (or test stat beyond critical) → reject $H_0$; else fail to reject.
- Step 6: Interpret in context: "At the 5 % level, there is sufficient evidence to conclude…" or "there isn't sufficient evidence…"
- Also mention Type I error (rejecting true $H_0$) and Type II error (failing to reject false $H_0$).

Tips for your notes

- Provide formulas for common tests (z, t).
- Provide interpretation examples.
- Emphasise assumptions: e.g., for z-test known σ; t-test unknown σ & normal/sufficient sample size.
- Highlight difference between "fail to reject" and "accept" — you never accept $H_0$, just fail to reject.

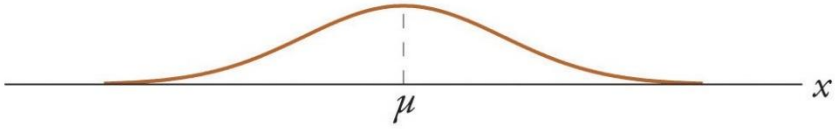
- Use simple real-world example: "Testing if the average time to complete a task is 30 minutes."

---

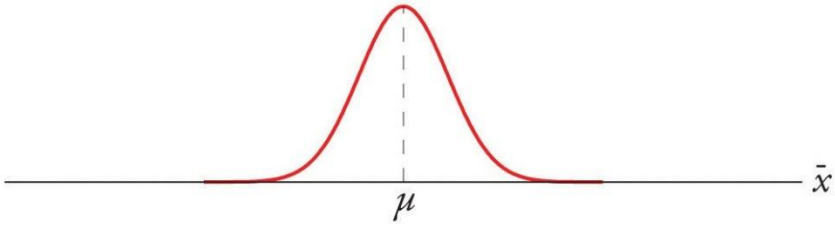
## 2.7 Sampling Distributions & Counting

### Sampling Distributions

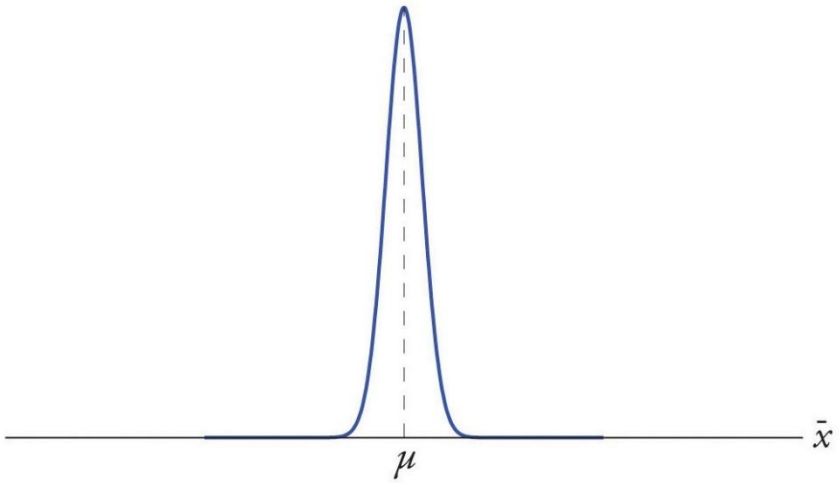
Population distribution



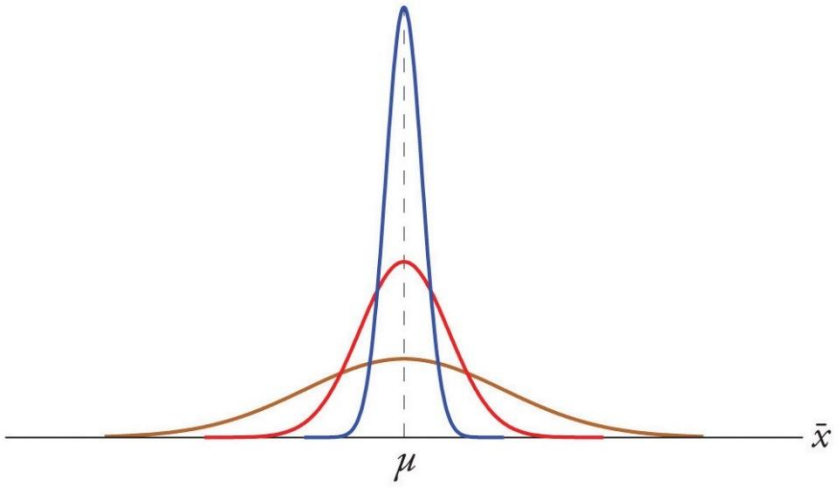

Sampling distribution of $\overline{X}$ with $n = 5$

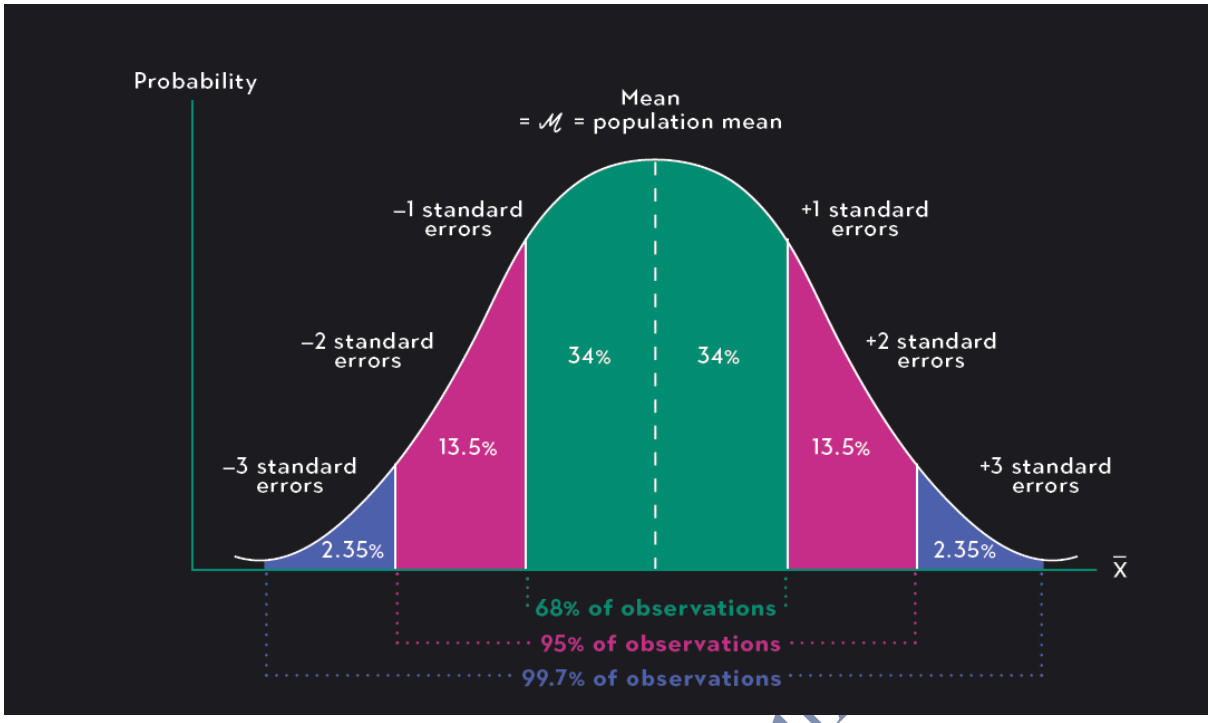


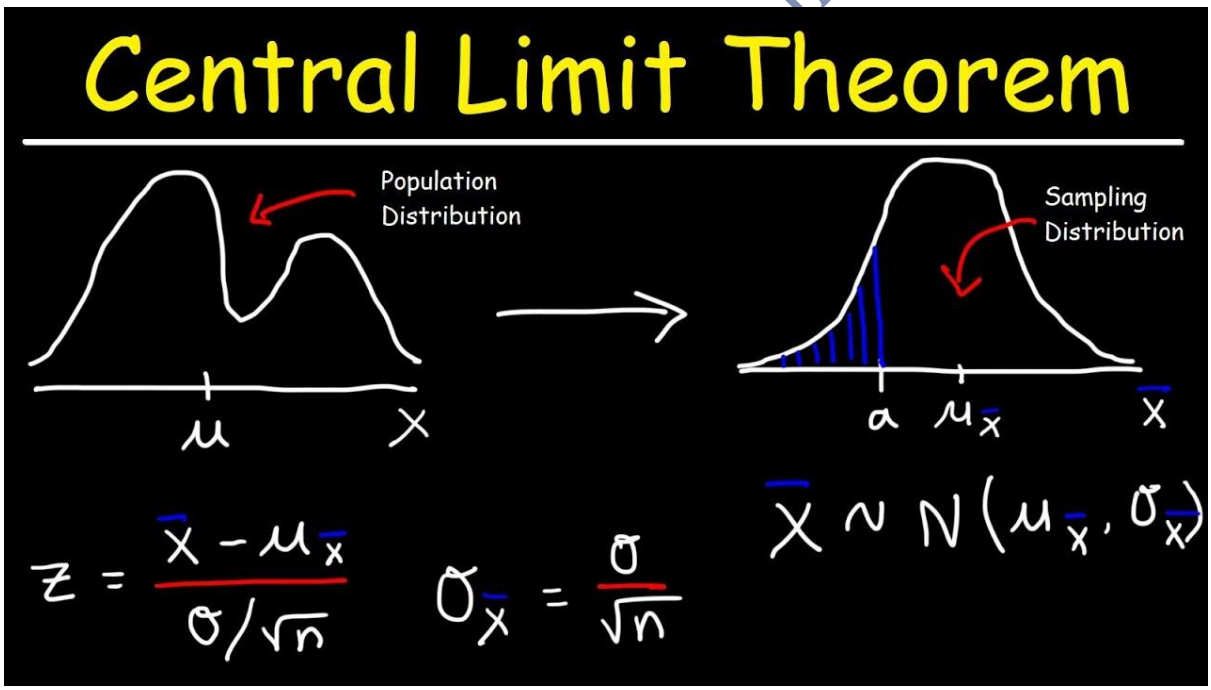

Sampling distribution of $\overline{X}$ with $n = 30$

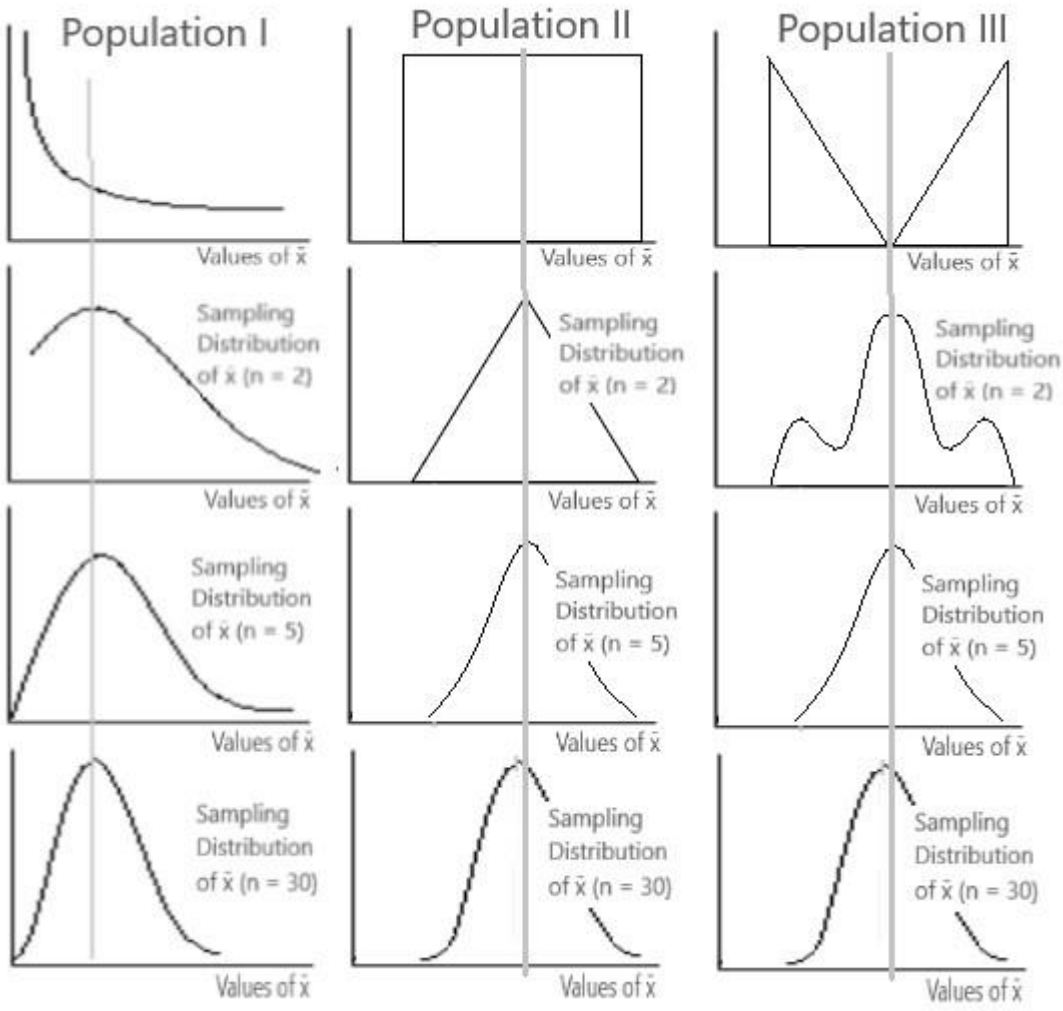


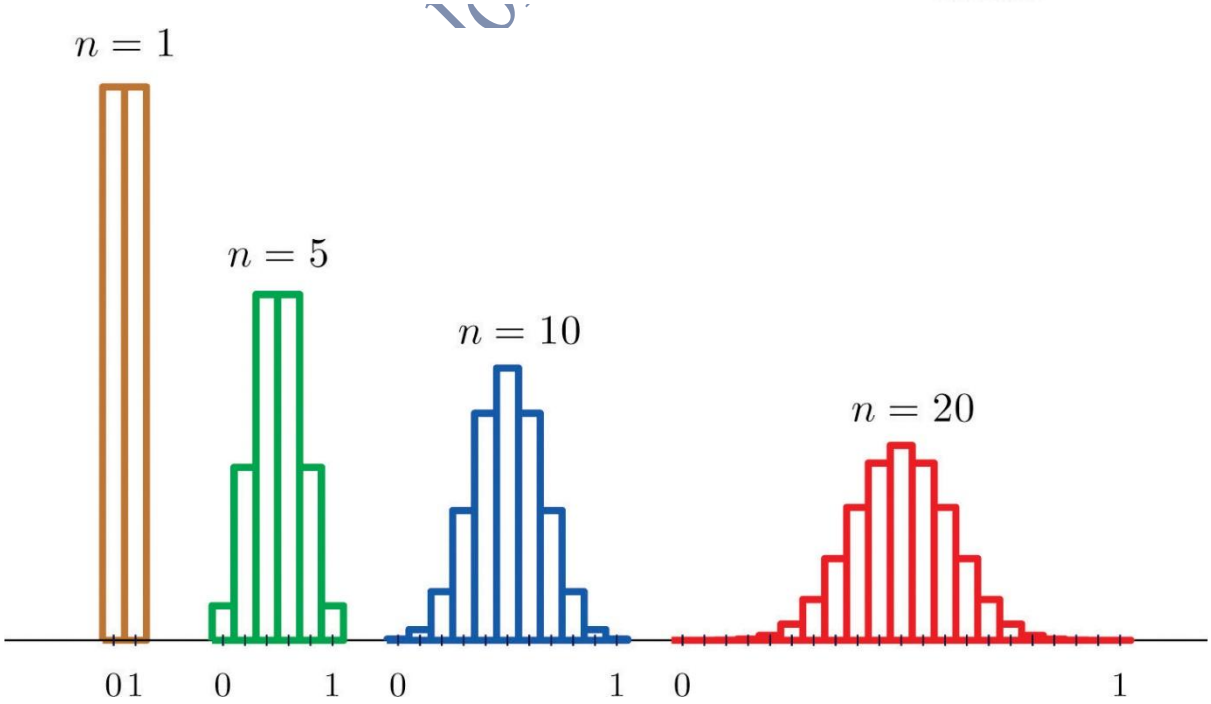

Distributions superimposed




Probability
Mean
= 𝓜 = population mean
–1 standard errors
+1 standard errors
–2 standard errors
+2 standard errors
–3 standard errors
+3 standard errors
34%
34%
13.5%
13.5%
2.35%
2.35%
x̄
68% of observations
95% of observations
99.7% of observations


Central Limit Theorem
Population Distribution
Sampling Distribution
μ
X
a
μx̄
X̄
z = (x̄ − μx̄) / (σ/√n)
σx̄ = σ/√n
X̄ ~ N(μx̄, σx̄)


Population I
Population II
Population III
Values of x̄
Sampling Distribution of x̄ (n = 2)
Sampling Distribution of x̄ (n = 5)
Sampling Distribution of x̄ (n = 30)


n = 1
n = 5
n = 10
n = 20
0 1
0
1
0
1
0
1

What it is:
A sampling distribution is the probability distribution of a statistic (e.g., sample mean (\bar x)) when you take many samples of the same size from a population. (Investopedia)
Why it matters:
It underpins inferential statistics: confidence intervals and hypothesis tests rely on the idea of repeated sampling and how the statistic behaves across samples.
How to explain:

- If population has mean μ and standard deviation σ, and you take samples of size n and compute (\bar x), then the sampling distribution of (\bar x) has mean μ and standard deviation (standard error) = σ/√n.
- Central Limit Theorem (CLT): For large n, the sampling distribution of (\bar x) is approximately normal, regardless of population distribution (provided some conditions). (ssc.wisc.edu)
  Example exam answer:

"If population mean μ = 50 and σ = 10, and we take samples of size n = 25, then the mean of the sample means is 50 and standard error = 10/√25 = 2. Thus about 95% of sample means will lie within 50 ± (1.96×2) = 46.08 to 53.92."

Counting (Combinatorics)

What it is:
Counting techniques (factorials, permutations, combinations) help determine how many possible outcomes exist in situations, which is necessary for probability calculations.
Why it matters:
In probability sections, you may need to answer: "How many ways can you choose 3 students from 10?" → combination formula; or "How many ways to arrange 5 books?" → permutations.
Key formulas:

- Factorial: (n! = n×(n-1)×…×1).
- Combination: (\displaystyle \binom{n}{r} = \frac{n!}{r!(n-r)!}) (select r items from n, order not important).
- Permutation: (\displaystyle P(n,r) = \frac{n!}{(n-r)!}) (order matters).

Tips for your notes

- Provide short table of formulas.
- Show simple example: e.g., "Number of ways to select 2 from 4 = 4C2 = 6."
- Emphasise why counting is required before computing probability: you need denominator (total possible outcomes) and numerator (favourable outcomes).

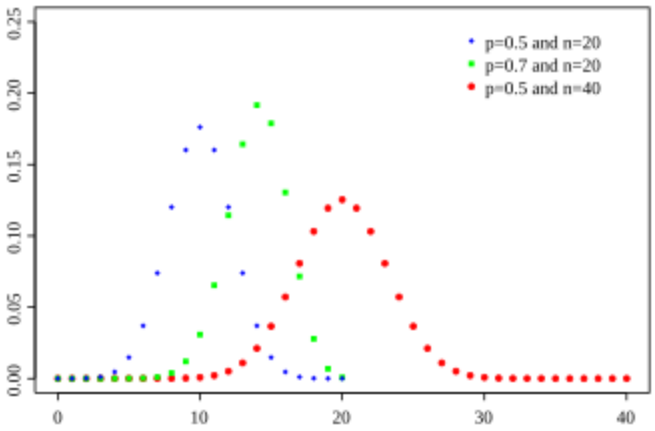
## 2.8 Probability & Probability Distributions

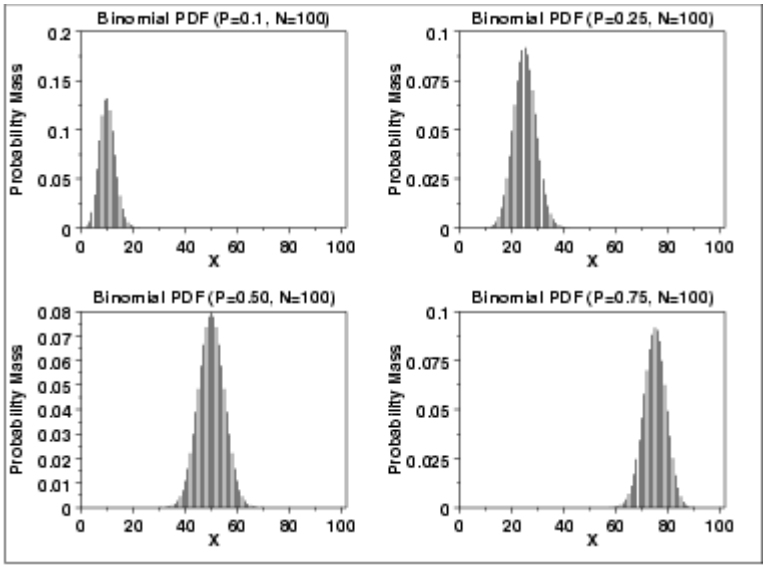
### Probability Basics

What it is:
Probability quantifies how likely an event is to occur (value between 0 and 1).
Why it matters:
The foundation for later parts of the unit: you can't do distributions, hypothesis tests or sampling theory without understanding probability.
Key rules to include in your notes/exams:

- Complement rule: (P(A^c) = 1 − P(A)).
- Addition rule (for non-overlapping events): (P(A ∪ B) = P(A) + P(B)). If events not mutually exclusive: (P(A ∪ B) = P(A) + P(B) − P(A ∩ B)).
- Multiplication rule for independent events: (P(A ∩ B) = P(A) × P(B)).
- Conditional probability: (P(A | B) = \dfrac{P(A ∩ B)}{P(B)}).

### Probability Distributions








What it is:
A probability distribution describes how probabilities are assigned over possible outcomes of a random variable. Two broad types:

- Discrete distributions (variable takes countable values): e.g., binomial, Poisson.
- Continuous distributions (variable takes any value in a range): e.g., normal, $\chi^2$ (chi-square), t, F. (Scribbr)
  Why it matters:
  Understanding distributions is essential for inference: you'll need to know e.g., "If X ~ Normal(μ, $\sigma^2$), then what is P(X > a)?" Or: "Which distribution to use for variance test?"
  How to explain:
- For discrete: list values and their probabilities (PMF).
- For continuous: specify PDF and note that probability for exact single value is zero; we use intervals.
- Important properties: mean, variance, shape.

Tips for your notes

- Include common distributions with their parameters & conditions: e.g., Normal(μ, $\sigma^2$), Binomial(n, p), Chi-square(df).
- Include when to use which distribution: e.g., $\chi^2$ for variances or goodness-of-fit, t for small sample mean testing.

- Explain link between distributions and inference: sampling distributions, test statistics often follow these distributions under $H_0$.
- Provide small example: "If X ~ Binomial(10, 0.3), what is P(X = 4) = 10C4 × (0.3)^4 × (0.7)^6."

---

✅ Final Review & Exam Tips

- For *each topic*, always include: definition, why it matters, how it works (or how to interpret it), and exam style answer/example.
- Use "real-world" context when you write your answers: e.g., "student scores", "manufacturing defect counts", "survey responses" etc.
- Always mention assumptions (especially for tests like ANOVA, regression) and limitations (e.g., imputation, outliers, non-normality).
- When you see a graph or table in exam: first state what it shows, then interpret meaning, then note any caution or further steps.
- Use clear key formulas in your notes and know when to use which formula.
- Time your practice: write short paragraphs (2-4 sentences) summarising each topic, as you might in an exam.

Diploma Wallah

**Made with 💗 by Sangam**