- **Describe data** → Descriptive statistics
- **Make decisions about populations** → Inferential statistics
- **Population**: entire group of interest
- **Sample**: subset of the population used for analysis

Descriptive Statistics

- Used to **summarize data you already have**.
- Mean
- Median
- Standard deviation
- Graphs (histograms, boxplots)

Inferential Statistics

- Used to **draw conclusions about a population** using a sample.
- Confidence intervals
- Hypothesis testing
- Regression
- ANOVA
- **Uncertainty is always involved**

A. Qualitative (Categorical) Data

1. Frequency Table

Purpose: Count how often each category appears.

Example (iMac data)

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Calculation: Relative Frequency

$$\text{Relative Frequency} = \frac{\text{Category Frequency}}{\text{Total}}$$

$$\frac{85}{500} = 0.17$$

Pie Chart

- **Purpose**: Show **proportions of a whole**
- Few categories
- Focus on percentages
- Avoid for comparisons across groups

Bar Chart

- **Purpose**: Compare **counts** across categories
- Categories on x-axis
- Frequencies on y-axis
- Bars **do not touch**

B. Quantitative Data

Stem-and-Leaf Plot

- **Purpose**: Show distribution shape + exact values
- **How it works**:
- Stem = tens digit
- Leaf = ones digit
- 37 → stem 3, leaf 7
- Best for **small to medium datasets**

Histogram

- **Purpose**: Show shape of large datasets
- **Steps**:

1. Create class intervals
2. Count frequencies per interval
3. Draw bars (touching)

Summarizing Distributions & Bivariate Data

1. Mean (Average)

$$\bar{x} = \frac{\sum x}{n}$$

Used when:

- Distribution is symmetric
- No extreme outliers

2. Median

Middle value after sorting data.

Preferred when:

- Data is skewed
- Outliers exist

3. Variance & Standard Deviation

Sample Standard Deviation

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

Step-by-step logic:

1. Subtract mean
2. Square deviations
3. Sum squares
4. Divide by $n-1$
5. Square root

4. Scatter Plot (Bivariate Data)

Purpose: Show relationship between X and Y

Patterns:

- Positive association
- Negative association
- No association

5. Pearson Correlation (r)

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

- Range: −1 to +1
- Strength + direction
- **No causation**

2. Box Plot & Outliers

Steps

1. Sort data
2. Find:
   - Minimum
   - Q1 (25%)
   - Median (Q2)
   - Q3 (75%)
   - Maximum
3. Compute **H-spread**

$$H = Q3 - Q1$$

4. Compute **Step**

$$\text{Step} = 1.5 \times H$$

5. Compute fences:

$$\text{Lower Fence} = Q1 - 1.5H$$

$$\text{Upper Fence} = Q3 + 1.5H$$

6. Trimean

$$\text{Trimean} = \frac{Q1 + 2Q2 + Q3}{4}$$

---

7. Trimmed Mean

1. Sort data
2. Remove % from each tail
3. Compute mean of remaining data

Geometric Mean

$$GM = \sqrt[n]{\prod X}$$

Used for **growth rates only**

Probability

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total outcomes}}$$

$$P(A^c) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Used when events overlap.

4. Multiplication Rule (Independent Events)

$$P(A \cap B) = P(A) \times P(B)$$

16. Permutations

$$nP_r = \frac{n!}{(n-r)!}$$

17. Combinations

$$nC_r = \frac{n!}{r!(n-r)!}$$

- **Permutations (Order Matters):** Use when the specific sequence, rank, or position changes the outcome.
    - *Keywords:* Arrange, list, schedule, rank, code.
    - *Examples:* Assigning 1st/2nd place, picking a PIN, or arranging books on a shelf.
- **Combinations (Order Doesn't Matter):** Use when you are selecting a group or subset where the arrangement is irrelevant.
    - *Keywords:* Select, choose, committee, set, group.
    - *Examples:* Picking 3 friends for a movie, choosing pizza toppings, or drawing a hand of cards.

18. Binomial Distribution

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

Mean:

$$\mu = np$$

Variance:

$$\sigma^2 = np(1-p)$$

Poisson Distribution

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$$

20. Hypergeometric Distribution

$$P(X = x) = \frac{kCx(N-k)C(n-x)}{NCn}$$

Normal Distribution & Estimation

Properties of Normal Distribution

- Symmetric
- Mean = Median = Mode

- Bell-shaped

2. Z-Score

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

Interpretation:

- How many SDs from mean

Example:

- z = 1.5 → 1.5 SD above mean

3. Confidence Interval (Mean)

$$\bar{x} \pm z \cdot \frac{s}{\sqrt{n}}$$

Steps:

1. Compute mean
2. Compute standard error
3. Multiply by critical z
4. Add/subtract

Logic of Hypothesis Testing

1. Hypotheses

- **H_0**: No effect / no difference
- **H_1**: Effect exists

2. p-Value

- Probability of observing result **assuming H_0 is true**
- If p < α → reject H_0

3. Errors

Type	Meaning
Type I	Reject true H_0
Type II	Fail to reject false H_0

Regression

1. Regression Equation

$$Y' = bX + A$$

Where:

- b = slope
- A = intercept

2. Slope Calculation

$$b = r\frac{s_y}{s_x}$$

- **r** → Pearson correlation

- s_Y → standard deviation of Y values
- s_X → standard deviation of X values

3. Intercept

$$A = \bar{Y} - b\bar{X}$$

- Y and x bar- mean

4. Error of Prediction

$$Y - Y'$$

Used to compute:

- SSE (error)
- Model accuracy

TOTAL SUM OF SQUARES (SSY)

a

$$SSY = \sum (Y - \bar{Y})^2$$

PARTITIONING SSY

$$SSY = SSY' + SSE$$

PROPORTION OF VARIANCE EXPLAINED

a

$$\frac{SSY'}{SSY}$$

ONE-WAY ANOVA

- Compare **means of ≥ 3 groups**

GRAND MEAN

$$\bar{X} = \frac{\sum X}{n}$$

GROUP MEANS

$$\bar{X}_j = \frac{\sum X_j}{n_j}$$

BETWEEN-GROUPS SS (SSB)

mula

$$SSB = \sum n_j (\bar{X}_j - \bar{X})^2$$

WITHIN-GROUPS SS (SSW)

mula

$$SSW = \sum (X_{ij} - \bar{X}_j)^2$$

TOTAL SS

$$SST = SSB + SSW$$

Degrees of Freedom

Source	df
Between	k−1
Within	n−k
Total	n−1

MEAN SQUARES

$$MSB = \frac{SSB}{df_B}$$
$$MSW = \frac{SSW}{df_W}$$

F-STATISTIC

$$F = \frac{MSB}{MSW}$$

TWO-WAY ANOVA

Purpose

Test:

1. **Factor A effect**
2. **Factor B effect**
3. **Interaction effect**

FACTOR SS

$$SS_A = \sum n(\bar{X}_A - \bar{X})^2$$

WITHIN SS

$$SS_W = \sum (X - \bar{X}_{cell})^2$$

INTERACTION SS

$$SS_{AB} = SS_T - SS_A - SS_B - SS_W$$

Degrees of Freedom

Source	df
A	a−1
B	b−1
A×B	(a−1)(b−1)
Within	n−ab
Total	n−1

F-RATIOS

$$F_A = \frac{MS_A}{MS_W}$$
$$F_B = \frac{MS_B}{MS_W}$$
$$F_{AB} = \frac{MS_{AB}}{MS_W}$$

CHI-SQUARE (χ^2)

- **Data is categorical**
- **Table of frequencies given**

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

O = values directly from the table

$$E = \frac{(\text{Row Total})(\text{Column Total})}{\text{Grand Total}}$$

Create calculation table

Cell	O	E	O − E	$(O-E)^2$	$(O-E)^2 / E$
1					
2					

Calculate chi-square value using formula above
Calculate degrees of freedom:

$$df = (r-1)(c-1)$$

- **r = number of rows**
- **c = number of columns**

Option A: Critical Value

$$\chi^2_{\text{calc}} > \chi^2_{\text{critical}} \Rightarrow \text{Reject } H_0$$

Option B: p-value

$$p < \alpha \Rightarrow \text{Reject } H_0$$

For no columns and rows:

Formula

$$E = \frac{\text{Total Observations}}{\text{Number of Categories}}$$

(only if proportions are equal)

$$df = k - 1$$

Where:

- k = number of categories