1 Chapter 1: Getting Data

Definitions:
1. A population is the entire group (of individuals or objects) that we wish to know something about.
2. A research question is usually one that seeks to investigate some characteristic of a population.
3. Exploratory Data Analysis (EDA) is a systematic process where we explore a dataset and its variables and come up with summary statistics as well as plots. EDA is usually done iteratively until we find useful information that helps us answer the questions we have about the data set.

1.1 Research Question

- To make an estimate about the population – These questions seek to estimate certain characteristics or values within a population, such as averages or proportions.
- To test a claim about the population
- To compare two sub-populations or to investigate a relationship between two variables in the population

1.2 Selection Bias

Selection bias happens when certain individuals or groups within a population have a higher or lower probability of being included in the sample than others, resulting in an unrepresentative sample.

Causes of Selection Bias:
- Sampling Frame Issues: The sampling frame doesn't represent the entire population, excluding certain groups. Example: Surveying only on-campus students about cafeteria food excludes commuters, leading to biased findings.
- Self-Selection: Participants with strong interest in the topic voluntarily join, overrepresenting certain views. Example: A fitness survey attracts fitness enthusiasts, excluding less interested individuals.
- Exclusion Criteria: Setting criteria omits key perspectives unintentionally. Example: A job satisfaction study excluding part-time workers misses their valuable insights.
- Convenience Sampling: Relying on easily accessible participants limits diversity. Example: Interviewing park-goers about a new park excludes those who dislike parks, skewing results.

1.3 Impact of Selection Bias

Reduced Generalizability: If the sample isn't representative, findings can't be generalized accurately to the broader population.
Inaccurate Conclusions: Selection bias can lead to misleading results, as the sample does not reflect true population characteristics.
Distorted Correlations and Relationships: The relationships between variables might appear stronger or weaker than they actually are, leading to false associations.

1.4 Sampling Frame

Definitions:
1. A population of interest refers to a group in which we have interest in drawing conclusions on in a study.
2. A population parameter is a numerical fact about a population.
3. A census is an **attempt** to reach out to the entire population of interest.

The **sampling is taken** from $N = S \cap P$

1.4.1 Sampling Method

1.4.1.1 Probability Sampling

Probability sampling is a sampling method where each member of a population has a **known, non-zero** chance of being selected.

There can be 50 men in one group and 20 men in another. It's fine as long as the size is accounted for

1. **Simple Random Sampling:**
   - Everyone has an equal chance of being selected.
   - Randomly picking 100 employees from a list of 1,000.
2. **Systematic Sampling:**
   - Select every k-th person after a random start.
   - Start at person 2 and pick every 10th person (e.g., 2, 12, 22...).
   - $k = \frac{\text{Total Size}}{\text{sample size}}$
3. **Stratified Sampling:**
   - Divide the population into subgroups (strata) and sample proportionally. Like gender, age and etc
   - Then some number of people at random is picked from each and every strata
   - Population is 60% male and 40% female; sample keeps the same ratio.
4. **Cluster Sampling:**
   - Divide population into clusters (e.g., schools, regions) and sample all or part of them.
     - **Single-Stage:** Pick clusters and survey everyone.
     - **Two-Stage:** Pick clusters, then randomly sample from within.
   - Randomly choose 2 schools and survey all students in those schools.

Advantage and Disadvantage

Method	Best for	Advantage	Disadvantage
Simple Random	Small, accessible populations	Reduces bias, good representation.	Time-consuming, needs full data.
Systematic	Large, ordered populations	Easy, spreads sample evenly.	Risk of periodic bias.
Stratified	Populations with subgroups	Guarantees subgroup representation.	Needs detailed subgroup data.
Cluster	Geographically spread groups	Cost-efficient for large areas.	May increase sampling error.

1.4.1.2 Non-probability Sampling

Convenience Sampling: They are chosen based on accessibility and proximity to the researcher, rather than a structured or randomized process.

Volunteer Sampling: Volunteer sampling occurs when participants opt in or volunteer to participate in a study, often because they have a particular interest in the topic. This method is commonly used in surveys. **If census** was attempted then it won't be considered non-probability

1.5 Variables

- An attribute that can be measured (e.g., age, gender, race).
- **Independent Variable:** The cause in an experiment, controlled by the researcher.
- **Dependent Variable:** The effect being measured.

1.5.1 Types of variables

Categorical Variables
1. Nominal: Categories with no order or ranking (gender)
2. Ordinal: Ordered categories with unequal intervals. (Rating)

Numerical
1. Discrete: Countable values, gaps between numbers. (Number of students) Can be decimal or fraction has long as it's possible to have gaps
2. Continuous: Infinite values within a range. (Time taken ippt)

1.6 Mean

$$X = \frac{\sum_{i=0}^{n} x_i}{n}$$

Properties:
Add constant: Old Mean + c
Multiply by constant: Old Mean × k
Note:
- The mean gives the central value of a dataset but does not describe how the data points are spread around it.
- A mean is 0 does not mean the SD is also

1.6.1 Overall means vs subgroup means

$$X_w = \frac{\sum(w_i * X_i)}{\sum_{W_i}}$$

1. X_i represents the mean of each subgroup
2. W_i is the weight (or size) of each subgroup
3. $\sum_{W_i}$ is the sum of all weights

1.7 Standard Deviation, Median, Inter-quartile

Range: Range = Highest value - Lowest value
Formula for Sample Variance

$$s^2 = \frac{\sum X_i - M}{n - 1}$$

Where:

X_i represents each data point | M is the mean of the sample | n is the number of data points in the sample

Formula for Sample Deviation

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

It's on it's on unit so 10cm. Standard Deviations is never a negative number. If the standard deviation is **zero**, it means all **data points are identical** (no spread around the mean).
Properties
Add constant: **No change**
Multiply by constant: $|s| * c$

Coefficient of Variation (CV)

$$\frac{s}{M} * 100\%$$

Interpretation: A higher CV indicates greater variability relative to the mean, while a lower CV indicates less variability.
Use Case: CV is particularly useful for comparing the spread of datasets that have different units or means, as it standardizes the measure of variability.

1.7.1 Median

- Odd: Take the middle number
- Even Take the to numbers in the middle and average it out

Properties
Add constant: Median + K constant
Multiply by constant: Median x K constant

1.7.2 IQR

Q1 Position = $\frac{n+1}{4}$ | Q3 Position = $\frac{3*(n+1)}{4}$

- IQR = Q3 - Q1
- Q3 ≥ Q1
- Q1 can be found by taking the lower half of the data if **odd, exclude the median**, even, directly split the data into two halves. Take middle of the bottom half
- Same for Q3 but take the top half

Properties
Add constant: **No change**
Multiply by constant: $\text{IQR} * c$

1.7.3 Mode

The mode is the value that appears most frequently in a dataset. Unlike median and mean **Mode** can be used for categorical variable.

1.8 Study Design

Note: The treatment group and the control can have varying sizes.
1. **Experimental Studies:**
   - **Purpose:** Establish a cause-and-effect relationship between variables.
   - **What Happens:** Researchers manipulate the independent variable and measure its effect on the dependent variable.
   - **Groups**:
     - Treatment Group: Receives the intervention (e.g., coffee).
     - Control Group: Does not receive the intervention (e.g., no coffee).
   - **Settings:**
     - **Random Assignment** ensures that extraneous variables (e.g., motivation, prior knowledge)(confonder) are evenly distributed.
     - Blinding minimizes bias:
       - **Single-Blind:** Participants don't know their group.
       - **Double-Blind:** Both participants and researchers are unaware of group assignments.
2. **Observational Studies**
   - **Purpose:** Explore associations or correlations between variables when experiments are unethical, impractical, or dangerous.
   - **What Happens:** Researchers observe and collect data without manipulating variables.
   - **Groups:**
     - **Exposed (Treatment Group):** Naturally exposed to a condition (e.g., smokers).
     - **Non-Exposed (Control Group):** Not exposed to the condition (e.g., non-smokers).
   - **Settings:**
     - No random assignment.
     - Conducted in natural settings.
     - Higher risk of bias and confounding factors

1.9 Generalizability

Sampling Frame ≥ Population

- Complete Coverage: A complete sampling frame includes all members of the target population, avoiding coverage bias. Missing parts of the population in the frame leads to unrepresentative results.
- Reduced Bias: A comprehensive sampling frame ensures all subgroups within the population are represented, reducing sampling bias. A smaller frame increases the risk of excluding certain groups.
- Avoiding Missed Insights: Incomplete frames exclude important population segments, leading to missed data and insights. A well-covered frame ensures a diverse and comprehensive analysis.
- Accurate Representation: A larger frame enables confident sampling that reflects the population's diversity and characteristics, improving reliability and generalizability.

2 Categorical-Data

2.1 Rates

Formula

$$\text{Rate} = \frac{\text{Number of Events}}{\text{Total Time Or Quantity}}$$

Conditional Rate:

$$\text{rate}(\text{Success}|X) = \frac{\text{Number of Successful Treatments with X}}{\text{Total Patients who Received Treatment X}}$$

Joint Rates

$$\text{rate}(\text{Success with Treatment X and Large Stone}) = \frac{300}{1050} = 0.29$$

2.2 Assoications

No association	Positive Association	Negative Association
$\text{rate}(\text{A}\|\text{B}) = \text{rate}(\text{A}\|\text{NB})$	$\text{rate}(\text{A}\|\text{B}) > \text{rate}(\text{A}\|\text{NB})$	$\text{rate}(\text{A}\|\text{B}) < \text{rate}(\text{A}\|\text{NB})$

2.3 Two Rules on Rates

2.3.1 Symmetry Rule

Symmetry Rule 1:

$$\text{rate}(A|B) > \text{rate}(A|\text{NB}) \Leftrightarrow \text{rate}(B|A) > \text{rate}(B|\text{NA})$$

Symmetry Rule 2:

$$\text{rate}(A|B) < \text{rate}(A|\text{NB}) \Leftrightarrow \text{rate}(B|A) < \text{rate}(B|\text{NA})$$

Symmetry Rule 3:

$$\text{rate}(A|B) = \text{rate}(A|\text{NB}) \Leftrightarrow \text{rate}(B|A) = \text{rate}(B|\text{NA})$$

2.3.2 The basic rules on rates

The overall rate(A) will always lie between $\text{rate}(A \mid B)$ and $\text{rate}(A \mid \text{NB})$.
Consequence 1:
The closer $\text{rate}(B)$ is to 100%, the closer $\text{rate}(A)$ is to $\text{rate}(A \mid B)$
Consequence 2:
If $\text{rate}(B) = 50\%$, then $\text{rate}(A) = \frac{1}{2}[\text{rate}(A \mid B) + \text{rate}(A \mid \text{NB})]$
Consequence 3:
$\text{rate}(A \mid B) = \text{rate}(A \mid \text{NB})$, then $\text{rate}(A) = \text{rate}(A \mid B) = \text{rate}(A \mid \text{Not } B)$

2.4 Simpson's Paradox

In simple terms, the paradox occurs when the relationship between **variables changes direction when a third variable**, known as a confounding variable, is introduced.
1. Break them into smaller group
2. Check if pattern is reversed
3. Check if there is any kind confounder

2.5 Confounder

A confounder is an external variable that affects both the independent (exposure) and dependent (outcome) variables in a study, creating a spurious or misleading association between them.
Characteristics of a Confounder:
1. Associated with the Exposure: The confounder is related to the independent variable (exposure) without being a consequence of it.
2. Associated with the Outcome: The confounder is related to the dependent variable (outcome).
3. Not on the Causal Pathway: The confounder does not lie in the direct pathway from the exposure to the outcome.

Identifying a Confounder
1. Check if it is related to the exposure.
2. Check if it affects the outcome.
3. Ensure it is not caused by the exposure (i.e., it is not part of the causal pathway).

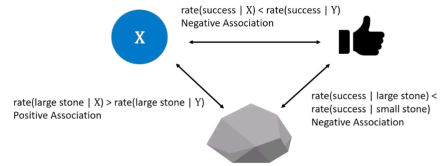


Figure 1: Prove of confounder

Controlling for Confounders in Study Designs
1. **Randomization (Experimental Studies**
   - In randomized controlled trials (RCTs), participants are randomly assigned to treatment groups, which helps distribute confounding variables evenly across these groups.
   - Randomization minimizes the likelihood that confounders will systematically differ between groups, thus isolating the effect of the treatment on the outcome.
2. **Stratification (Observational Studies):**
   - Stratification involves dividing the data into subgroups based on the confounder. For example, in the kidney stone case, stratifying by stone size (large vs. small) allows comparison of treatment effectiveness within each subgroup.
   - This method reveals the true association within each level of the confounder, helping to understand how the confounder affects the outcome.

3 Numerical Data

3.1 Univariate

Univariate uses a single variable at a time. It **doesn't care about the interactions between variables**. It focuses on Distribution (the spread and shape of the data), central tendency, and Dispersion.

3.2 Histogram

Bins
- range 2 < x <= 4

Sturges' Rule:

$$k = \lceil \log(n) + 1 \rceil$$

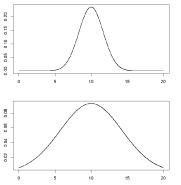
Where:

k: The number of Bins | n: Number of Data points

Shape and Centre

Shape	Description	Mean vs. Median	Example
Symmetrical (Normal)	Bell-shaped, evenly distributed	Mean ≈ Median ≈ Mode	Heights, test scores
Right-Skewed	Long tail on the right	Mean > Median > Mode	Income distribution
Left-Skewed	Long tail on the left	Mean < Median < Mode	Age at retirement
Uniform	Equal bar heights	No central peak	Rolling a fair die
Unimodal	A single distinct tall bar	mean, median, and mode are approximately the same	
Bimodal	Two distinct peaks	Depends on the modes	Test scores for two groups
Multimodal	More than two peaks	Depends on the modes	Seasonal sales data
Exponential	Steep drop-off	Mean > Median	Waiting times

Spread
Range: is the difference between the highest and lowest values in the dataset.
Standard Deviation



Low variability
s = 1.69

High variability
s = 4.30

Figure 2: Spread of Data

- **Low variability**
  - When data has low variability, most values are tightly packed near the center of the distribution.
  - This means the median, which represents the middle value, will likely be close to most data points.
- **High variability**
  - When there is high variability, it means the data points are more spread out from the center (mean or median)
  - The median might not represent the dataset well because values are spread far from the center.

3.2.1 Outlier

Suppose the outlier is removed from the data set, The removal will cause the standard deviation to decrease, The removal will cause the range to change. The mean will decrease or increase depending on the outlier's position. The IQR can increase, remain the same or decrease

3.3 Histogram Vs Bar Graph

Feature	Histogram	Bar Graph
Data Type	Continuous	Categorical
Bars Touch?	Yes	No
X-Axis	Numerical Ranges	Categories
Purpose	Data Distribution	Comparison of Groups
Examples	Exam Scores, Heights	Fruits, Countries

3.4 Boxplot

How to construct a box plot

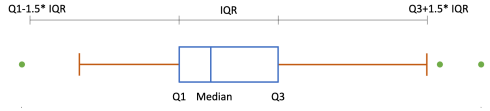


Figure 3: Boxplot

1. Draw a box from Q1 to Q3
2. Draw a vertical line in the box where the median is
3. Extend a line from Q1 to the smallest value that is not an outlier, and from Q3 to the largest value that is not an outlier. These are called whiskers.
4. Indicate outliers with dots
5. Outlier: Greater than $\mathrm{Q3} + 1.5 \times \mathrm{IQR}$
6. Outlier: Lesser than $\mathrm{Q1} - 1.5 \times \mathrm{IQR}$

Note: When there is an X mark it indicates **mean Note:** We can apply transformation directly to the value

3.5 Boxplot vs Histogram

Aspect	Histogram	Boxplot
Shape	Displays the frequency distribution of data, showing the overall shape (e.g., symmetric, skewed, or multimodal).	Summarizes distribution using median, quartiles, and spread without explicitly showing the shape.
Comparison of Data	Suitable for analyzing frequency distribution over ranges of values.	Useful for side-by-side comparison of medians, ranges, and interquartile ranges (IQR) of datasets.
Outliers	Harder to identify, as outliers may appear as sparse or isolated bins.	Clearly shows outliers as individual points beyond the whiskers.
Number of Points	Shows approximate distribution of data across intervals but not the exact number of total points.	Indicates summary statistics but doesn't display the total count or individual data points.

3.6 Bivariate

Bivariate: If you're analyzing the trend or relationship between time (months) and resale prices, then it's bivariate because you're studying how one variable affects or correlates with another. It does not have a **deterministic relationship.** This relationship is often **statistical** rather than deterministic, meaning that one variable can influence the other, but there may still be some degree of randomness or variation.

3.7 Scatter Plot

1. **Independent Variable:** This variable is the one you control or consider as the cause, and it is plotted on the x-axis (h-axis).
2. **Dependent Variable:** This variable is the outcome or the effect, and it is plotted on the y-axis (v-axis).

We can use scatter plot

1. **Relationships:** How two variables are associated (e.g., positive, negative, or no correlation).
2. **Patterns or Trends:** Such as linearity, clusters, or outliers in the data.
3. If a SD is 0 then the relationship between the variables cannot be analyzed meaningfully.

Notes:

- Check if the value is within range

We can interpret the scatter plots in the following ways:

Direction

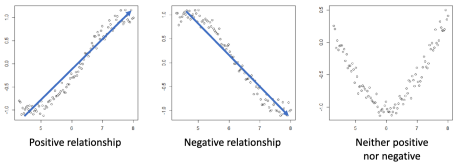


Figure 4: Directions

Form

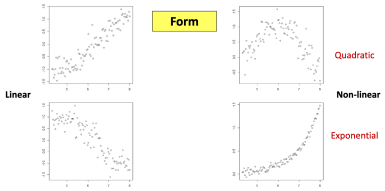


Figure 5: Forms

Strength

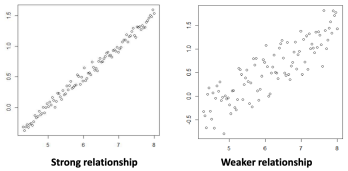


Figure 6: Forms

- When they are close to each it indicates a strong linear relationship
- When they are spread out (scattered) it indicates a weak relationship

Outliers

1. **Deviates from the Trend:**
   - If most of the points follow a clear pattern (e.g., linear or nonlinear), an outlier lies far away from this pattern.
2. **Significant Distance:**
   - Outliers are points that are unusually distant from the bulk of the data in either the horizontal (x-axis) or vertical (y-axis) direction, or both.
3. **Impact on Analysis:**
   - Outliers can distort the trend, such as the slope in a linear regression line, or impact the correlation coefficient, making it weaker or stronger than it should be.

3.8 Correlation Coefficient

• $r > 0$ -> Positive Linear Association • $r < 0$ -> Negative Linear Association • $r = 1$ -> Perfect Positive Linear Association • $r = -1$ -> Perfect Negative Linear Association • $r = 0$ -> No Linear Association (Just no linear association does not mean not relation)	• Stong: $-1 \leq x \leq -0.7$ • Stong: $0.7 \leq x \leq 1$ • Moderate: $-0.7 < x \leq -0.3$ • Moderate: $0.3 < x \leq 0.7$ • Weak: $-0.3 < x \leq 0.3$

Calculate Correlation Coefficient r

$$\mathrm{SU}_z = \frac{X - \text{average}(X)}{S_x} \text{ and } \mathrm{SU}_y = \frac{Y - \text{average}(Y)}{S_y}$$

$$r = \frac{\sum(\mathrm{SU}_x * \mathrm{SU}_y)}{n-1}$$

Properties of r

- Interchanging does not affect the value of R
- Adding a number to all the values does not change R
- Multiplying a positive number to all values of a variable does not change R

Limitations of Correlation Coefficient

- Correlation does not imply causation
- A third variable might affect the outcome
- r only measures linear association between two variables
- Always look at the scatter plot not only at the r value

3.9 Linear Regression

- We can use the x-axis to predict the y-axis
- The values calculated should be within the cloud

$$Y = \mathrm{m}x + c$$

$$m = r * \frac{S_y}{S_x}$$

$$c = \text{meanY} - \mathrm{m}(\text{slope}) \times \text{meanX}$$

$$\text{New Slope} = \frac{\text{old slope} \times \text{SDx}}{\text{SDy}}$$

Formula for non-linear

$$y = \mathrm{cm}^x = \ln(y) = \ln(m) + x\ln(b)$$

3.10 Ecological Correlation

Ecological correlation analyzes relationships at a group or aggregate level, rather than at an individual level. It is often used in public health, sociology, environmental science, and policymaking when individual-level data is unavailable or unnecessary.

Key Benefits:

- Provides population-level insights for trends and patterns.
- Aids in policymaking and large-scale decision-making.
- Simplifies analysis with aggregate data, saving time and cost.
- Useful for hypothesis generation for further individual-level studies.
- Applicable in environmental and social studies where variables naturally exist at group levels.

Limitations:

- Risk of ecological fallacy: assuming group-level relationships apply to individuals.
- Can miss nuances within groups, leading to overgeneralization.

Group-Level Correlation Example

- A country with a high average income may also have a high average life expectancy.
- Ecological fallacy: Concluding that every individual in that country with a high income has a high life expectancy.

3.11 Atomistic Fallacy

The atomistic fallacy is the opposite of the ecological fallacy. It occurs when conclusions about a group are incorrectly drawn from data or relationships observed at the individual level. In essence, it assumes that individual-level relationships apply to the broader group or population, which can lead to misleading conclusions.

4 Statistical Inference

- Definition **Sample Space:** All the possible outcomes
- Definition **Event:** A specific event that's a subset of Sample space

All comes are events but not all events are outcomes.

4.1 Probability

- Definition **Sample Space:** All the possible outcomes
- Definition **Event:** A specific event that's a subset of Sample space

All comes are events but not all events are outcomes.

Rules of Probability

1. $0 \leq P(E) \leq 1$ for each event E
2. $P(S) = 1$ if S is the entire Sample Space
3. If E and F are non-overlapping(Mutually Exclusive) events, then $P(E \cup F) = P(E) + P(F)$

Uniform Probabilities

$$\text{Every outcome has the same probability} = \frac{1}{\text{size of sample space}}$$

Conditional Probability

$$P(E|F) = \frac{P(E \cap F)}{P(F)}, P(F) \neq 0$$

Independence between two events

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

Conditional Independence $P(A \cap B \mid C) = P(A|C) * P(B|C)$

Sensitivity and Specificity

Sensitivity (True Positive Rate):

The ability of a test to correctly identify individuals who do have the condition (true positives). $P(+ \mid \text{Drunk})$

Specificity (True Negative Rate):

The ability of a test to correctly identify individuals who do not have the condition (true negatives). $P(- \mid \text{Sober})$

Mutually Exclusive: Two events are mutually exclusive if they cannot occur at the same time. $P(A \cap B) = 0$ or P(A or B) = P(A)+P(B)

Independent Events: Two events are independent if the occurrence of one does not affect the probability of the other. $P(A \cap B) = P(A) \times P(B)$

4.2 Prosecutor's Fallacy

$$P(A|B) \neq P(B|A)$$

Unless

$$P(A) = P(B) \rightarrow P(A|B) = P(B|A)$$

4.3 Law of Total Probability

$$P(A) = P(A|B) * P(B_1) + P(A|B_2) + P(B_2) + \ldots + P(A|B_n) * P(B_n)$$

4.4 Conjunction Fallacy and Base Rate Fallacy

Conjunction Fallacy

The Conjunction Fallacy occurs when people assume that the probability of two events occurring together (a conjunction) is more likely than the probability of one event occurring on its own, which violates the rules of probability.

Wrong interpretation

$$P(A \cap B) > P(A) \text{ and } P(A \cap B) > P(B)$$

This is the correct interpretation

$$P(A \cap B) \leq P(A) \text{ and } P(A \cap B) \leq P(B)$$

Base Rate Fallacy The Base Rate Fallacy occurs when people ignore or undervalue the base rate (prior probability) of an event in favor of specific information (evidence or likelihood), leading to incorrect conclusions.

Bayes Theorem Using this only if data is not explicitly tabulated

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

4.5 Statistical Inference

$$\text{Sample statistic} = \text{population parameter} + \text{bias} + \text{random error}$$

How to eliminate bias from Sample statistics?

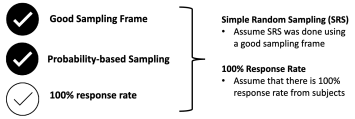


Figure 7: Eliminating Bias

4.6 Confidence Interval

We conclude that about 95% of the samples will contain the population parameter (in this case the population unemployment rate) within their respective confidence intervals.

Note:

- Any confidence interval constructed from any sample, re- gardless of the significance level, may or may not contain the population parameter.
- Every confidence interval for a population mean constructed using a sample will contain the corresponding sample mean,

Population Proportion

Formula

$$P^* \pm z \times \sqrt{\frac{P^*(1-P^*)}{n}}$$

Where:

P^*: The sample proportion ($\frac{\text{number of successes}}{\text{sample size}}$) z: The critical value from the z-distribution n: The sample size

Z-Values for Confidence Levels

Confidence Level (%)	z-Value
90%	1.645
95	1.96
99%	2.576
99.9%	3.291

Population Mean

Formula

$$\overline{x} \pm z \times \frac{\sigma}{\sqrt{n}}$$

Where:

$\overline{x}$: Sample mean z: Z-critical value (based on confidence level) σ: Population standard deviation n: Sample Size

1. **Confidence Level:**
   - The confidence level determines how certain we are that the interval captures the true population parameter.
   - **Higher Confidence Level:**
     - Larger error margin (wider interval).
     - More likely to include the true population parameter.
     - Example: A 99% confidence level will have a wider range than 95%.
   - **Lower Confidence Level:**
     - Smaller error margin (narrower interval).
     - Less likely to include the true population parameter.
2. **Sample Size:**
   - The sample size affects the width of the confidence interval, assuming the confidence level remains the same.
   - **Larger Sample Size:**
     - Smaller interval (more precise estimate).
     - Reduces variability (random error) in the sample statistic.
     - Example: Surveying 1,000 people gives a narrower interval than surveying 100 people.
   - **Smaller Sample Size:**
     - Larger interval (less precise estimate).
     - Greater variability due to limited data.

4.7 Hypothesis Testing

(α) is $\alpha = 1 - \text{Confidence Level}$

- A p-value is a statistical measure that helps determine the strength of the evidence against a null hypothesis (H_0). It tells us the probability of obtaining the observed data, or something more extreme, if the null hypothesis is true. It uses **conditional probability**.
- For p-value computation, we focus on the probability of outcomes that are as favorable or more favorable to the alternative hypothesis

1. **Reject Null Hypothesis:**
   - **Condition**: $\text{p-value} < \text{significance level}$ (e.g., 0.05)
   - **Implication:**
     - You have enough evidence to reject the null hypothesis (H_0)
     - This suggests that the alternative hypothesis (H_1), is likely true
2. **Do Not Reject Null Hypothesis:**
   - **Condition**: $\text{p-value} \geq \text{significance level}$
   - **Implication:**
     - You do not have sufficient evidence to reject the null hypothesis.
     - This does not mean the null hypothesis is true—just that the data does not provide strong enough evidence against it.
   - Why accept H_0?
     - Hypothesis tests do not prove hypotheses; they test for evidence against the null.
     - A failure to reject H_0 does not confirm H_0; it could simply mean insufficient data or power in the test.

Hypothesis test for population

Tests whether a population proportion (p) is equal to a specific value or if there is a significant difference. For **example** the proportion of people who favour a policy, the proportion of defective items in a batch, etc.

$$\text{Null Hypothesis: } H_0 = p = p_0$$

$$\text{Alternate Hypothesis: } H_a = p \neq p_0, p > p_0 \text{ or } p < p_0$$

Hypothesis Test for Population Mean (t-test)

Tests whether a population mean (μ) is equal to a specific value or if there is a significant difference. For **example** average test scores, average weight, or average time spent on an activity.

$$\text{Null Hypothesis: } H_0 = \mu = \mu_0$$

$$\text{Alternate Hypothesis: } H_a = \mu \neq \mu_0, \mu > \mu_0 \text{ or } \mu < \mu_0$$

Types

- One-sample t-test: Compares the sample mean to a known population mean.
- Two-sample t-test: Compares the means of two independent groups.
- Paired t-test: Compares means before and after a treatment within the same group.

Hypothesis Test for association (chi-squared test for association)

Tests whether there is an association (relationship) between two categorical variables. For **example** examining if gender and product preference are related, or if education level and voting behaviour are associated.

$$H_0 : \text{The two variables are independent no association}$$

$$H_a : \text{The two variables are not independent Association}$$