1. Getting Data

1.1 Exploratory Data Analysis (EDA)

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.

1.1.1 Research Questions

	Descriptive	**Inferential**	**Comparative**
Definition	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 – These questions aim to test specific hypotheses or claims regarding the population, often through statistical testing.	To compare two sub-populations or to investigate a relationship between two variables in the population – These questions explore relationships or comparisons, such as examining differences between groups or assessing the association between variables.
Example	What is the average number of hours university students in Singapore study each week?	Do more than 50% of university students in Singapore prefer online classes over in-person classes?	Is there a difference in academic performance between university students who work part-time and those who do not?

1.1.2 EDA

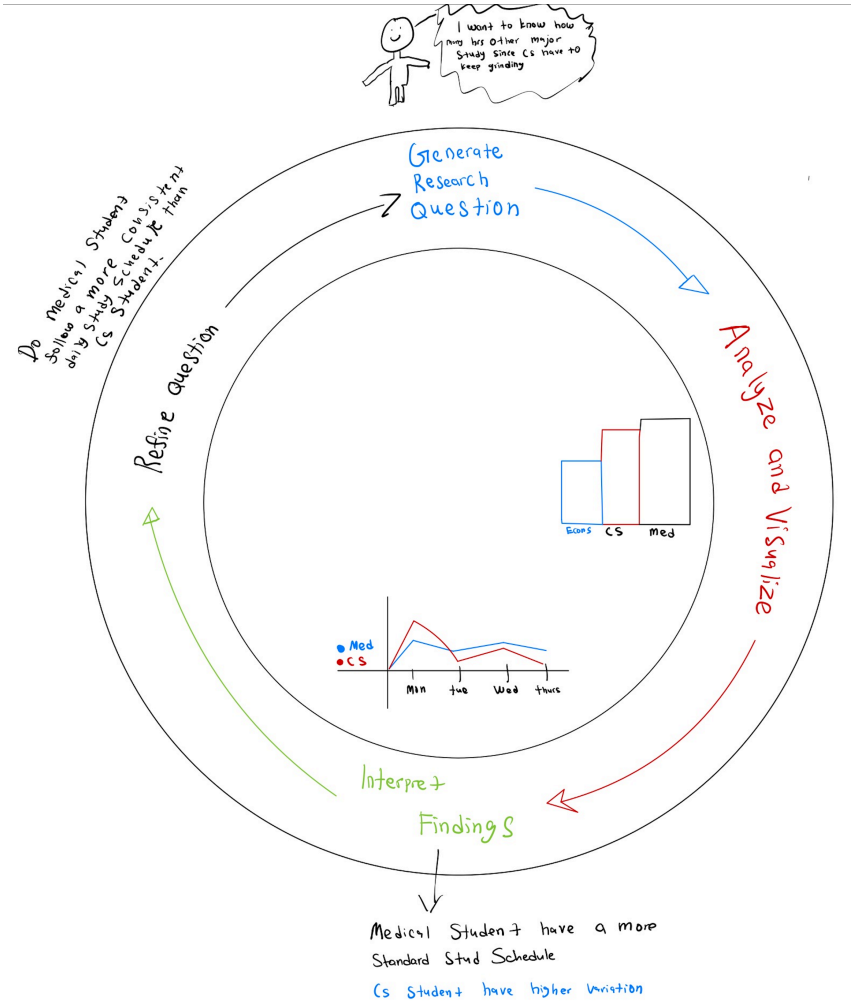


Figure 1: Example of EDA

1. Generate Research Question: Start with a broad question, e.g., “How many hours do other majors study?”

2. Analyze and Visualize: Collect data, then use charts (e.g., bar and line graphs) to see study patterns across majors. This shows, for example, that medical students study consistently, while CS students vary more.

3. Interpret Findings: Review the visualizations and insights. Medical students have a consistent study routine, whereas CS students have fluctuating study hours.

4. Refine Question: Based on the interpretation, refine the question to focus on consistency: “Do medical students follow a more consistent daily study schedule than CS students?”

1.2 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.

Example

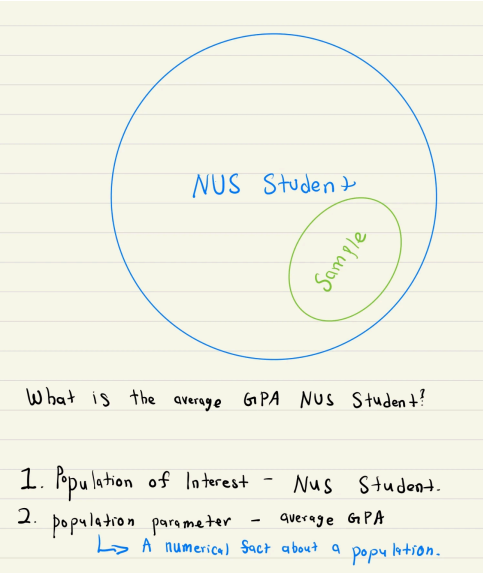


Figure 2: Example of population interest and parameter

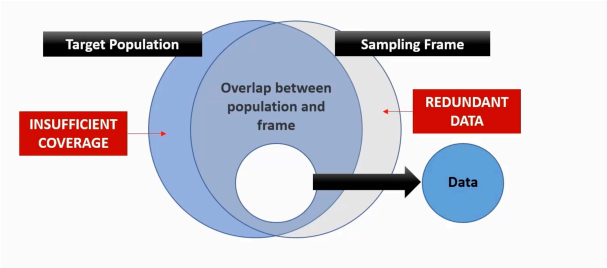


Figure 3: Sampling Frame

- Think of your **target population** as all **fish in a lake**.
- The **sampling frame** is the **net** you use to catch them.
  - If your net has **holes** (insufficient coverage), some fish escape.
  - If you catch **extra things** like weeds (redundant data), they **mess up your study**.
  - The best result? A **perfect overlap** where your net catches exactly the right fish.

1.2.1 Example

I want to find out the number of people who wear spectacles in Computer Science (CS). My target population P is all CS students, and my sampling frame S is all Singaporeans. Since not all CS students are covered in the frame (some might be missing), this is insufficient coverage.

Additionally, my sampling frame contains redundant data (non-CS students), introducing irrelevant data points. These issues could lead to biased results, reduced generalizability, missed insights, and potential misinterpretation. To accurately select a sample, I should take the intersection of the target population P and the sampling frame S, represented as: $N = S \cap P$ where N is the sample of CS students within the Singaporean population. With careful sampling, N can be generalized to the target population of CS students. (More on ***generalizability***)

1.2.2 Census vs Sample

Definition: A census is indeed an attempt to gather data from the entire population, but it doesn't always have to be fully completed

Aspect	Census	Sample
Definition	Data collection from the entire population	Data collection from a subset of the population
Scope	Entire population	Representative subset
Cost	High cost	Lower cost
Time Requirement	Time-consuming	Faster
Accuracy	High accuracy with minimal sampling error	Subject to sampling error
Feasibility	Less feasible for large populations	Highly feasible for large populations
Use Cases	National population data, employee census	Political polls, product testing, customer surveys

When to use Census or Sample?

1. **Census:**
   - Use a census when accuracy is critical and resources are available, or when the population is small enough to make a complete survey feasible.
   - Examples: National census for government planning, employee census in a small company.
2. **Sample:**
   - Use a sample when the population is large, and there are time or budget constraints. Proper sampling techniques can provide accurate estimates without the need for a full census.
   - Examples: Political polls, customer feedback surveys, and scientific studies.

Why Census is preferred over sample

People prefer sampling over a census because it is generally cheaper, faster, more feasible for large populations, and allows for better data quality and control. Sampling provides reliable insights without the massive resources and logistical challenges that a census demands, making it the preferred method in many research and data collection scenarios.

1.3 Bias

1.3.1 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.

1.3.1.1 Causes of selection bias

The reason selection bias happens is due to 4 reasons: sample frame issues, self-selections, exclusions criteria, and convenience sampling.

Sampling frame issue: When the sampling frame doesn't cover the entire population, certain segments are left out. For example, surveying only on-campus students about cafeteria food excludes commuters, who may have different opinions. This incomplete sampling frame results in biased findings, as it doesn't reflect the views of all students.

Self-Selection: When participants opt into a study voluntarily, those with strong interest in the topic are overrepresented. For instance, a survey about exercise is likely to attract fitness enthusiasts, while those uninterested may ignore it. This self-selection bias skews the results, capturing primarily those with a vested interest.

Exclusion Criteria: Setting exclusion criteria can unintentionally omit important perspectives. For example, a job satisfaction study that excludes part-time workers assumes their experiences are irrelevant, but part-time employees may have valuable insights. This exclusion creates bias by not representing the entire workforce.

Convenience Sampling: Relying on easily accessible participants often limits the diversity of views. For example, interviewing only park-goers about a new park excludes those who dislike parks, likely leading to overly positive feedback. This convenience sampling doesn't capture a full range of opinions, resulting in a biased sample.

1.3.1.2 Volunteer Bias

Volunteer Bias happens when the people who choose to participate in a study are different in some important way from those who don't choose to participate. This can lead to biased results because the sample might overrepresent certain traits or opinions.

You might be wondering if it sounds similar to self-selection. **Volunteer Bias** is a type of Self-Selection Bias and is specific to **interest** in the topic. **Self-selection bias** is more general and covers any reason someone might choose to join a study on their own, free time, convenience and personal beliefs.

Imagine you want to find out how many people enjoy going to the gym, so you put out a survey for anyone interested to respond. Who do you think is most likely to answer?. The people who answer this already enjoy the gym or feel strongly about fitness. These people are more likely

to volunteer because they have a personal interest in the topic. People who don't like the gym or feel indifferent about fitness are less likely to take the time to respond.

To reduce Volunteer Bias, researchers can use random sampling to select participants, encourage participation by offering incentives to all respondents, or limit open-participation studies in favour of directly reaching out to a diverse sample. By addressing Volunteer Bias, studies can achieve more representative results and avoid skewed findings.

1.3.1.3 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.

1.3.2 Non-Response Bias

Non-Response Bias happens when individuals don't participate in a study for reasons such as lack of interest, inconvenience, or sensitivity. This can lead to a sample that doesn't accurately represent the entire population, so researchers often use follow-ups, incentives, and confidentiality assurances to reduce Non-Response Bias.

Imagine a survey on mental health conducted within a company. Employees who feel uncomfortable sharing sensitive information about their mental health might choose not to respond. Others may not participate because they aren't interested in the topic, or they may find the survey format inconvenient. As a result, the responses are likely to reflect the views of a limited group, potentially overrepresenting those who feel safe discussing their mental health and underrepresenting those who may need more support.

1.4 Sampling Methods

In this section, we will be discussing the different sampling methods to reduce bias but they don't eliminate bias entirely

1.4.1 Probability Sampling

Probability-Based Sampling is a type of sampling method in which each member of the population has a known, non-zero chance of being selected for the sample. This approach is foundational for achieving a representative sample, as it reduces bias and allows for the generalization of results to the entire population.

1.4.1.1 Advantage and Disadvantage

Sampling Plan	Advantages	Disadvantages
Simple Random Sample	Provides a good representation of the population.	It can be time-consuming and requires accessibility to complete information about the population.
Systematic Sample	Offers a simpler selection process compared to simple random sampling.	May potentially underrepresent certain parts of the population if there is a periodic pattern.
Stratified Random Sample	Ensures good representation of the sample by each stratum, or subgroup, within the population.	Requires a detailed sampling frame and criteria for classifying the population into appropriate strata.
Cluster Random Sample	Less time-consuming and more cost-effective.	Requires a larger sample size to achieve a low margin of error, as clusters may not fully represent the diversity within the entire population.

1.4.1.2 Simple Random Sampling

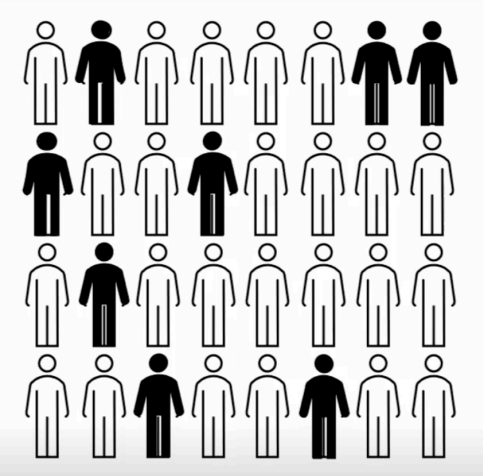


Figure 4: Simple Random Sampling

In simple random sampling, every member of the population has an equal chance of being selected. Selection is typically done using random number generators or other randomizing techniques. It minimizes selection bias and ensures each individual has an equal opportunity to be chosen.

Imagine you have a list of 1,000 employees and want to survey 100 of them. By assigning each employee a unique number and using a random number generator, you select 100 employees at random.

The advantage is it gives a good representation but the disadvantage is can reach everyone that was selected.

1.4.1.3 Systematic Sampling

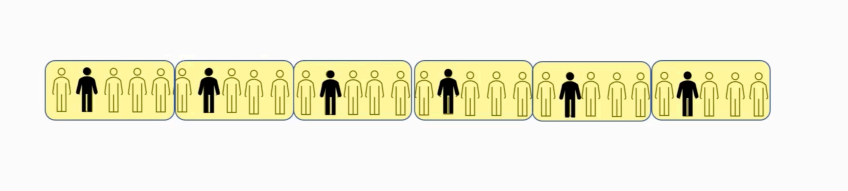


Figure 5: Systematic Sampling

In systematic sampling, you select every nth individual from a list after randomly choosing a starting point. This method spreads the sample across the population in an orderly way.

$$k = \frac{N}{n}$$

N is the total population size,
n is the desired sample size.

1. Randomly Select a Starting Point: Choose a random starting point between 1 and k
2. Select Every k Individual: From the starting point, select every n individual until the desired sample size is reached.

If you have a list of 1,000 people and need a sample of 100, you could randomly select a starting point (e.g., the 2th person) and then pick every 10th person after that. So 2, 12, 22, 32 ….

Easier to implement than random sampling, especially with large populations; spreads the sample evenly across the list but might not be a good representation because if the list has a repeating pattern that aligns with the interval, it could introduce periodicity bias.

1.4.1.4 Stratified Sampling

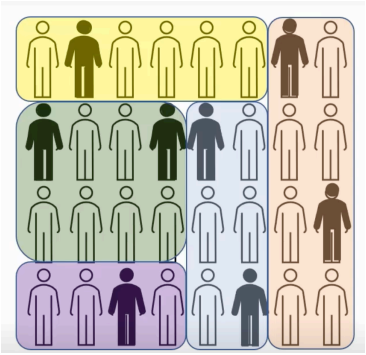


Figure 6: Stratified Sampling

The population is divided into subgroups (strata) based on certain characteristics (e.g., age, gender, income), and then a random sample is drawn from each subgroup. This ensures the representation of all important subgroups.

If a company has 60% male and 40% female employees, you could divide the population by gender and randomly sample from each subgroup proportionally, maintaining the 60/40 split in your sample, so 6 male and 4 female.

Helps ensure all significant subgroups are represented, reducing sampling frame bias (A type of selection bias) but requires knowledge of the population's subgroups and can be time-consuming to implement.

1.4.1.5 Cluster

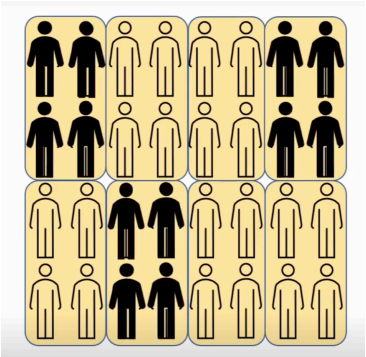


Figure 7: Cluster Sampling

The population is divided into clusters (often based on geographic or natural groupings), and then a random sample of clusters is selected. All individuals within these clusters are surveyed.

A researcher wants to study student behavior in a district, so they randomly select five schools (clusters) and survey all students within those schools. Adventatge and disadvanatage.

1. **Single-Stage Cluster Sampling:**
   - You randomly select a subset of the clusters (school districts) and then survey everyone within those selected clusters.
   - For instance, you could randomly choose 2 of the 5 districts and then survey all students within those 2 districts, assuming that gives you the 50 people needed.
2. **Two-Stage Cluster Sampling:**
   - First, you select clusters (in this case, all 5 school districts).
   - Then, you randomly select individuals from within each selected cluster.
   - Here, you would randomly choose 10 individuals from each of the 5 districts, totaling 50 people.

Reduces costs and logistical effort, especially with geographically spread populations but can increase sampling error if the clusters themselves are not representative of the entire population.

1.4.2 Non-Probability Sampling

Non-probability sampling refers to sampling methods where participants are selected based on factors other than random chance. This approach doesn't provide every individual in the population with a known or equal chance of being chosen. Non-probability sampling is often

used for exploratory research, quick data collection, or when a representative sample is not necessary. However, it comes with limitations in terms of generalizability and introduces various types of biases. Two common non-probability sampling methods are convenience sampling and volunteer sampling.

1.4.2.1 Convenience Sampling

In convenience sampling, participants are chosen based on accessibility and proximity to the researcher, rather than a structured or randomized process. This method is quick, easy, and cost-effective but can introduce significant selection bias and non-response bias.

A researcher standing outside a coffee shop to survey customers about their coffee preferences is using convenience sampling. This sample might not represent all coffee drinkers, as it excludes those who don't frequent that coffee shop or go at different times.

1.4.2.2 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 where people are asked to participate via an open call (e.g., online surveys or polls). Volunteer sampling can introduce both selection bias and non-response bias.

A study on gym habits that allows people to sign up for participation might attract primarily fitness enthusiasts, resulting in an overrepresentation of people who are already committed to working out regularly.

1.5 Variables

A variable is an attribute that can be measured or labelled like age, phone number, name, race, gender and etc

1.5.1 Independent

An independent variable is a variable that the researcher changes or controls in an experiment to observe its effect on the dependent variable. It is considered the **"cause"** in an experiment because changes in this variable are hypothesized to lead to changes in the dependent variable.

1.5.2 Dependent Variables

A dependent variable is the variable being tested and measured in an experiment.It's considered the **"effect"** or outcome that depends on the independent variable. It is the response or outcome that is observed and recorded, as it is ***hypothesized*** to change when the independent variable is manipulated.

Research Question	**Variables**
Does the amount of study time affect exam scores?	**Independent variable**: Amount of study time
	Dependent variable: Exam score

1.5.3 Types of Variables

1.5.3.1 Categorical Variable

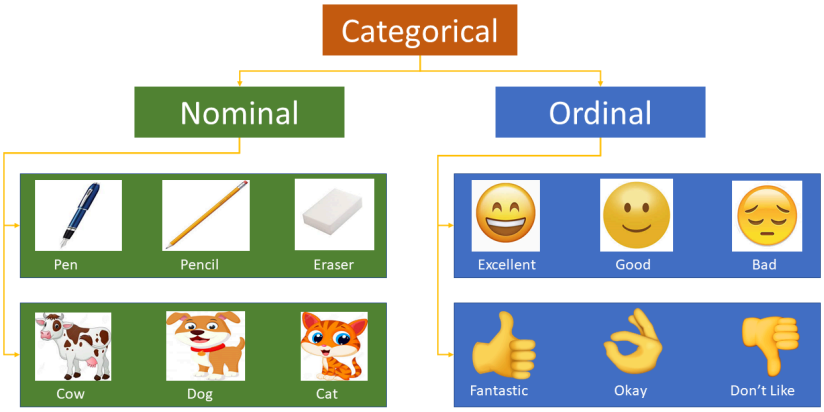


Figure 8: Examples of Categorical Variable

A categorical variable is a type of variable that represents data which can be divided into **distinct groups or categories**. These categories are often labels or names that describe qualitative attributes **rather than numerical values**.

1.5.3.1.1 Nominal

A label or name distinct categories without any order or ranking. In other words, nominal variables represent groups or classifications where each category is unique and there is no logical order to the categories.

Key Characteristics:

1. No Order or Ranking: The categories in nominal variables have no meaningful sequence. One category is not "higher" or "better" than another.
2. Discrete Categories: Each category is separate and distinct, often represented by labels or names.
3. Qualitative Data: Nominal variables deal with non-numeric data, where values are used to identify groups or types.

Example: Male, Female, Non-binary. There is no ranking or order among genders in this variable.

Nominal variables are useful for identifying and grouping data by type. They help researchers organize data into categories that can be easily counted and compared, even though no mathematical operations like addition or subtraction can be performed.

Nominal data is often analyzed using **frequency counts** to see how many observations fall into each category. This data can be visualized with **bar charts or pie charts** to show proportions or comparisons between groups. Statistical tests, like the **chi-square test**, are commonly used with nominal data to analyze relationships between categories.

Visualization Type	Nominal	Ordinal
Bar Chart	✓	✓
Stacked Bar Chart		✓
Grouped Bar Chart	✓	
Pie Chart / Donut Chart	✓	✓
Ordered Pie Chart		✓
Stacked Area Chart		✓
Heatmap		✓
Mosaic Plot	✓	
Frequency Table	✓	✓

1.5.3.1.2 Ordinal

It has a meaningful order or ranking, but the intervals between the categories are not necessarily equal or meaningful. In other words, while ordinal variables show a relative positioning of categories, they don't tell us the exact distance between them.

They are useful when data needs to be grouped by order but don't require precise measurements. They're often used in surveys and scales, such as rating levels of agreement or satisfaction. These variables allow for comparisons like "higher" or "lower" but don't permit meaningful mathematical operations like addition or subtraction.

Example: Satisfaction levels indicate an order of preference, but we cannot measure the exact difference in satisfaction between each level.

In data analysis, ordinal variables are often summarized with **frequency tables** or visualized with bar charts. Statistical tests for ordinal data are typically non-parametric, such as the **Mann-Whitney U test or Kruskal-Wallis** test, which don't assume equal intervals between categories.

1.5.3.2 Numerical Variable

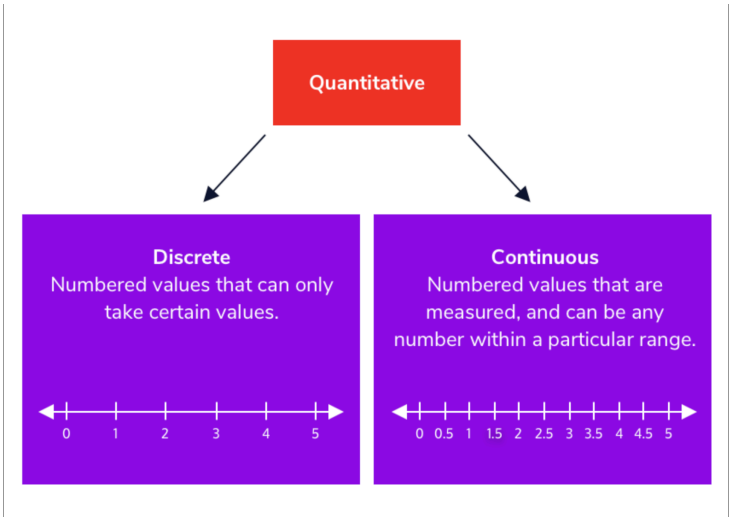


Figure 9: Examples of Numerical Variable

1.5.3.2.1 Discrete

A discrete variable is a type of quantitative variable where the possible values form a set of separate numbers with gaps in between. Discrete variables take on a finite or countable number of values. For **example** number of students in a class (e.g., 20, 21, 22).

Discrete variables are ofter whole numbers and they don't have any intermediate values between the numbers.

1.5.3.2.2 Continuous

A continuous variable is a type of quantitative variable where the possible values form an unbroken sequence, with no gaps between values. Continuous variables can take on an infinite number of values within a given range. For **example** time taken to complete a task (e.g., 12.5 seconds, 13.7 seconds).

Visualization Type	Discrete	Continuous
Bar Chart	✓	
Histogram	✓	✓
Line Graph		✓
Scatter Plot	✓	✓
Box Plot		✓
Frequency Table	✓	
Dot Plot	✓	✓
Violin Plot		✓
Density Plot		✓

1.6 Mean

1.6.1 Summary Statistics

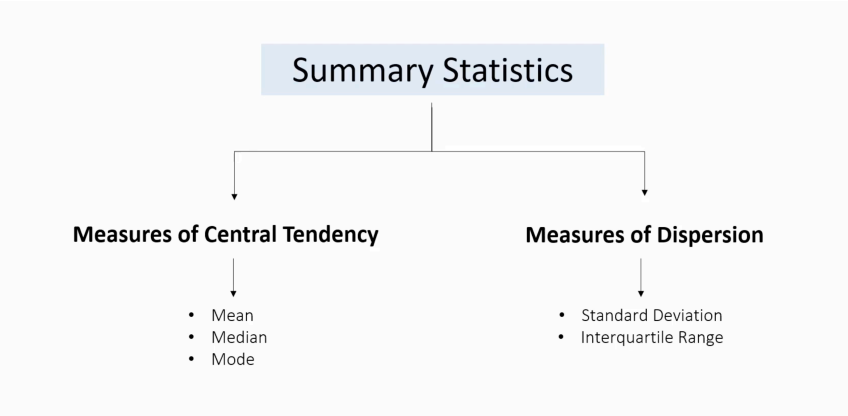


Figure 10: Summary Statistics

1.6.2 Definition and properties of mean

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

Mean (or average) is a measure of central tendency, representing the sum of all values in a dataset divided by the total number of values. It gives a “central” value that summarizes the dataset.

If a **constant c is added** to every value in the dataset, the mean will also increase by c. For **example**, if the original mean is 10 and you add 5 to each data point, the new mean will be 15. **Formula:** New Mean = Old Mean + c

If every value in the dataset is multiplied by a constant k, the mean will also be multiplied by k. For **example**, if the original mean is 10 and you multiply each data point by 3, the new mean will be 30. **Formula**: New Mean = Old Mean × k

Knowing the Mean Does Not Reveal the Distribution: The mean gives the central value of a dataset but does not describe how the data points are spread around it.

1.6.3 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

School	Number of Students	Mean
School A	349	32.21
School B	46	30.72
Overall	395	32.04

Applying the formula

$$\frac{(349 * 32.21) + (46 * 30.72)}{395} = 32.04(2d.p)$$

The weighted average will fall between the subgroup means and will be closer to the mean of the subgroup with a larger weight (or higher proportion).

1.6.4 How a proprtion is a mean

Can we say that the new drug is more effective since 200 attacks is better then 300 right? **Wrong**

We need to get the rate:
Rate(Asthma attack for new drug) 0.4 > Rate(Asthma attack for existing drug) 0.3

1.7 Standard deviations, Medians, Inter-quartile ranges

1.7.1 Sample variance and standard deviation

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

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

The deviation comes form the variance Standard deviation tells us tge spread of the data about the mean Formula of variance Formulae of deviation The intuition behind squaring the formula

- Variance: Measures the average squared deviation from the mean.
- Standard Deviation: The square root of variance, showing data spread in the original units.
- Squaring deviations prevents negatives from canceling positives, emphasizes larger deviations, and provides a standard measure for spread.

1.7.2 Properties of standard deviations

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).

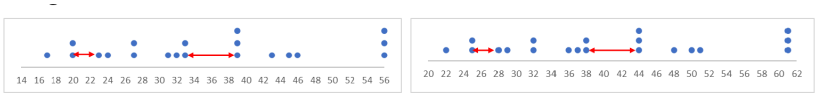


Figure 11: Adding constant to standard deviations

When you add a constant to each data point, it shifts all values up or down by that constant, but the distances between the data points remain the same. Therefore, the spread of data around the mean (standard deviation) does not change.

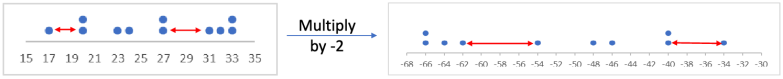


Figure 12: Multiplying constant to standard deviations

When you multiply each data point by a constant, the distances between the data points are scaled by that constant, which also scales the standard deviation.

New Formula

$$|s| * k$$

Coefficient of Variation (CV)

The Coefficient of Variation (CV) is a way to compare the spread of data relative to its mean, especially useful when comparing the variability of datasets with different units or scales.

Formula

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

1. Interpretation: A higher CV indicates greater variability relative to the mean, while a lower CV indicates less variability.

2. 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.

1.7.3 Example of SD

Reporting test scores or heights, where you want the variability in the same units (e.g., if the average height is 170 cm with an SD of 10 cm, you know most people are around 160–180 cm).

We can use the empirical rule **68-95-99.7** = (1SD, 2SD, 3SD) anyone outside this is a outlier

1.7.4 Definitions and Properties of median

The median is a measure of central tendency that represents the middle value in a dataset when it is ordered. It divides the dataset into two equal halves, with **50% of the values lying below** it and **50% above** it. It's less affected by the outlier

Steps

1. **Arrange the Data:** First, sort the data values in ascending or descending order
2. If the Number of Observations is Odd:
   - The median is the **middle value** in the ordered list
   - Example: In the dataset [3, 5, 7], the median is 5 (the middle value)
3. If the Number of Observations is Even:
   - The median is the **average of the two middle values**.
   - Example: In the dataset [2, 4, 6, 8], the median is $\frac{4+6}{2} = 5$

Adding a Constant Value: When you add a constant value to every data point, the median will increase by that constant as well. Median + K constant

Multiplying All Data Points by a Constant Value: When you multiply each data point by a constant, the median will also be multiplied by that constant. Median x K constant

1.7.5 Relationship between means and medians

When distributions of points are roughly symmetric, the mean and median will be quite close to each other.

1.7.6 Quartiles and interquartile range

- **Q1:** The 25th percentile, or the first quartile, is the value below which 25% of the data falls.
- **Q2:** The 50th percentile, also known as the median, divides the data into two halves.
- **Q3:** The 75th percentile, or third quartile, is the value below which 75% of the data falls.

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

$$\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}$$

A small IQR indicates that the middle 50% of the data points are closely clustered around the median.

A large IQR indicates that the middle 50% of the data points are more spread out from the median.

Adding a Constant: No change. Adding a constant shifts all data points equally, so the distance between Q1 and Q3 (IQR) remains the same.

Multiplying by a Constant: IQR is multiplied by that constant. This scales the spread of the data proportionally.

1.7.7 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 Designs

1.8.1 Experimental Studies

Research Question: Does drinking coffee daily help students achieve better scores?

Coffee	No coffee
Drink exactly one cup of coffee every day for one month	*Not drink any coffee for one month*
Treatment Group	**Control Group**

Figure 13: Treatment and Control Group

Controlled Experimental to provide the evidence that there is cause and effect relationship between two variables. So we muniplate the indepenanct variable like coffee to see if the score increase. **The treatment group and the control can have varying sizes.**

Why can't we have one big group?

The reason is simple if we don't have a base to compare would anyone belive the result hence we have 2 different groups

1.8.1.1 Random Assignment

In research, **"random"** has a strict meaning related to an impartial chance mechanism. Random assignment ensures that each participant has an equal chance of being placed in either the treatment or control group, helping to minimize the impact of other variables. The size of the random does not need to be the same. Can be 60 40.

We aim to make sure that the **independent variable** (e.g., coffee consumption) is the only factor influencing the **dependent variable** (e.g., test scores). Random assignment helps achieve this by **distributing extraneous variables** (such as motivation, study habits, or prior knowledge) evenly across both groups. This reduces the likelihood that these factors will influence results in one group more than the other.

Example: Imagine we want to study whether **drinking coffee daily improves students' test scores**. If we observe that everyone in the coffee group scores higher on average, does that necessarily mean coffee caused this improvement? **Not necessarily**. If we didn't use random assignment, it could be that the coffee group just happened to have more naturally motivated students, which could explain their better scores.

By **randomly assigning** students to the coffee (treatment) and no-coffee (control) groups, we ensure that **motivation, study habits, and other variables** are likely similar in both groups. This way, if the coffee group still scores better, we can be more confident that **coffee—not other variables—is contributing to the difference**.

1.8.1.2 Blinding

Blinding is a technique used in research to prevent bias and ensure more reliable results. It's designed to keep participants, researchers, or assessors unaware of who is receiving the treatment and who is not. This helps to avoid any psychological or observational biases that could influence the results.

Types of Blinding

1. **Single-Blind Study:**
   - In a single-blind study, only the participants do not know whether they are receiving the treatment or the placebo.
   - This prevents participant bias, where a participant's expectations might influence their response to the treatment.
2. **Double-Blind Study:**
   - In a double-blind study, both the participants and the assessors are unaware of who is receiving the treatment or the placebo.

- This prevents observer or assessor bias, where assessors' knowledge of the treatment could influence their interpretation of the results.
- Double-blind studies are considered the gold standard for clinical trials because they reduce bias on multiple levels.

Example of Blinding

- Treatment Group: Receives the actual drug.
- Control Group: Receives a placebo (a substance with no therapeutic effect, like a pill made of sugar or saline water) to mimic the appearance of the drug.

This placebo ensures that any psychological effects of **"thinking they are receiving treatment"** apply equally to both groups, so any real effects are due to the drug itself and not the belief in receiving treatment.

Placebo bias can happen so it's essential that the placebo resembles the treatment as closely as possible to prevent any clues that might influence participants' behavior or beliefs. This can be done by matching them in appearance, taste, and delivery method

1.8.2 Observational Studies

An observational study is a type of research conducted when it would be too dangerous, unethical, or impractical to perform an experimental study. In these studies, researchers observe and collect data without manipulating any variables.

Groups in Observational Studies

1. **Treatment Group (Exposed)**: This group consists of participants who have been exposed to the condition or factor being studied (e.g., smokers in a study on smoking and lung health).
2. **Control Group (Non-Exposed)**: This group consists of participants who have not been exposed to the condition (e.g., non-smokers in the same study).

Key Characteristics

- **No Manipulation:** Unlike in experimental studies, researchers do not manipulate the independent variable. They simply observe existing conditions and behaviors.
- **Natural Settings:** Observational studies often take place in natural settings, where participants live their normal lives without intervention from the researchers.

Example

Suppose researchers want to study the effect of air pollution on respiratory health. Assigning people to live in polluted areas would be unethical. Instead, they can conduct an observational study:

- Treatment Group (Exposed): People living in high-pollution areas.
- Control Group (Non-Exposed): People living in low-pollution areas.

Researchers can then observe differences in respiratory health between these groups without actively controlling or manipulating where people live.

1.8.3 Experimental Vs Observational Study

Feature	Experimental Study	Observational Study
Purpose	To establish a cause-and-effect relationship between variables.	To explore associations or correlations between variables.
Manipulation of Variables	Yes – Independent variable is manipulated by the researcher.	No – Variables are observed as they naturally occur without manipulation.
Groups	Divided into treatment and control groups.	Divided into exposure (treatment) and non-exposure (control) groups.
Random Assignment	Yes – Participants are randomly assigned to groups to control for confounding variables.	No – Participants are observed in their natural settings without random assignment.
Control of Confounding Factors	Easier to control for confounding variables through randomization and controlled conditions.	Difficult to control, as groups may differ in other variables beyond exposure.
Type of Data Collected	Generally experimental data (outcomes resulting from the manipulation of variables).	Generally observational data (outcomes observed without interference).
Bias and Confounding	Lower risk of bias and confounding due to controlled conditions.	Higher risk of bias and confounding, as uncontrolled factors may influence outcomes.
Examples	Clinical trials, lab experiments, A/B testing.	Cohort studies, case-control studies, surveys, and observational field studies.
Establishing Causation	Yes – Can establish causation due to controlled manipulation of variables.	No – Can only suggest association, not causation.

2. Categorical Data

We have covered categorical variables.

2.1 Rates

Formula

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

Where:

Numerator: Represents the "event" or quantity you're measuring.

Denominator: Represents the total context in which the event occurs.

Case Study:

Size of Stone	Gender of Patient	Treatment Type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

In this section, we will explore the kidney stone treatment case study to illustrate the application of the PPDAC cycle (Problem, Plan, Data, Analysis, Conclusion).

In the Problem stage, we aim to determine the effectiveness of different kidney stone treatments. Since treatment outcomes vary and there is a risk of failure, it's crucial to assess which treatment offers a higher success rate.

For the Analysis phase, we can use the marginal rate to obtain an overall success rate, providing a broad understanding of treatment effectiveness across all patients. This helps us make an initial assessment of whether a treatment is generally reliable, setting the foundation for more detailed analysis if needed.

2.1.1 Marginal rates

Marginal rate is the overall success or failure rate of something across the entire data set, without looking at specific subgroups. It gives a general picture of how often a particular outcome occurs in the whole population being studied.

Variable	**Count**	**Rate**	**Percentage**
Success	831	$\text{rate}(\text{Success}) = \frac{831}{1050} = 0.791$	79.1%
Failure	219	$\text{rate}(\text{Success}) = \frac{219}{1050} = 0.209$	20.9%
Total	1050	1.000	100%

For example, in the kidney treatment study, the marginal rate tells us the total success rate (79.1%) and failure rate (20.9%) of the treatment for all patients combined, regardless of other factors like gender or treatment type.

2.1.2 Conditional rates

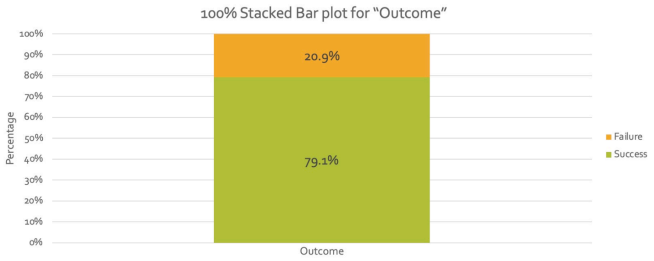


Figure 14: Marginal Rate Stacked Graph

Now that we know the overall success rate for kidney stone treatment is higher than the failure rate, as illustrated in the figure, we can conclude that treatment is generally effective. However, this doesn't answer the original question: "Which treatment is better?". We can find the better treatment by using **conditional rate**.

Treatment	**Success**	**Failure**	**Row Total**
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

What is the Success given Treatment X

$$\text{rate}(\text{Success}|X) = \frac{542}{700} = 0.774$$

What is the Success given Treatment Y

$$\text{rate}(\text{Success}|Y) = \frac{289}{350} = 0.825$$

Formula:

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

2.1.3 Joint Rates

Joint Rates refer to the probability or proportion of an outcome occurring for a specific **combination of two categorical variables**, such as treatment type and outcome in the kidney stone case study.

Treatment	Stone Size	Success	Failure	Total
X	Large	300	100	400
X	Small	242	58	300
Y	Large	120	30	150
Y	Small	169	31	200

To determine the success rates of Treatments X and Y in relation to stone size, we can use joint rates. This approach allows us to analyze the effectiveness of each treatment specifically for large and small stones, providing a clearer comparison within each category.

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

2.2 Associations

Association indicates a relationship or pattern between variables but does not imply causation. It shows that the presence (or level) of one variable is related to the likelihood of an outcome in another variable, but it doesn't mean that one variable directly causes the other.

Knowing associations helps us identify patterns, make predictions, assess risks (Knowing smoking causes cancer), and form hypotheses. It guides interventions, directs resources effectively, and serves as a foundation for deeper investigations, even when causation isn't established.

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

A = Success	NA = Failure	B = Treatment X	NB = Treatment Y

Using the previous example of the kidney stone case study. $\text{Rate}(\text{A}|\text{B}) = \text{rate}(\text{Success}|\text{X}) = \frac{542}{700} = 0.774$. $\text{Rate}(\text{A}|\text{B}) = \text{rate}(\text{Success}|\text{Y}) = \frac{289}{350} = 0.826$.

Since $\text{rate}(A|B) < \text{rate}(A|\text{NB})$ then we can say that the success of the treatment is negatively associated with treatment X.

2.2.1 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:

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

2.2.2 The basic rules on rates

The overall rate(A) will always lie between $\text{rate}(A \mid B)$ and $\text{rate}(A \mid \text{NB})$.

2.2.3 Consequences of the basic rule of rates

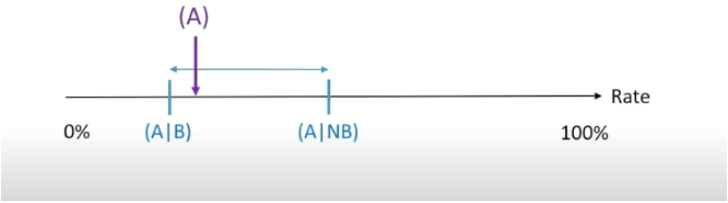


Figure 15: Marginal Rate Stacked Graph

Consequence 1: The closer $\text{rate}(B)$ is to 100%, the closer $\text{rate}(A)$ is to $\text{rate}(A \mid \text{NB})$

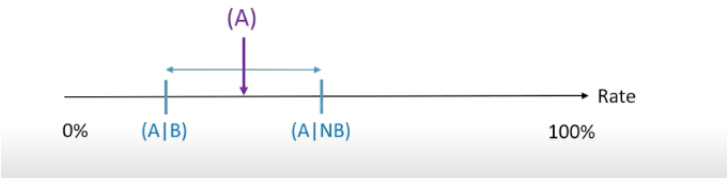


Figure 16: Marginal Rate Stacked Graph

Consequence 2: If $\text{rate}(B) = 50\%$, then $\text{rate}(A) = \frac{1}{2}[\text{rate}(A \mid B) + \text{rate}(A \mid \text{NB})]$

Figure 17: Marginal Rate Stacked Graph

Consequence 3: If $\text{rate}(A \mid B) = \text{rate}(A \mid \text{NB}), \text{then } \text{rate}(A) = \text{rate}(A \mid B) = \text{rate}(A \mid \text{Not } B)$

2.3 Simpsons Paradox

Simpson's Paradox is a phenomenon in statistics where a trend or effect observed in different groups of **data reverses when the groups are combined**. This can lead to misleading conclusions if data is not properly segmented or if key variables are not accounted for. In simple terms, the paradox occurs when the relationship between **variables changes direction when a third variable**, known as a confounding variable, is introduced.

Previously, we concluded that Treatment Y appeared better because it was positively associated with success when we considered the overall rates. However, we haven't yet taken into

account an important variable: the size of the kidney stone. The stone size could play a crucial role in determining the effectiveness of each treatment. Let's take a closer look at this additional variable to see if it changes our initial conclusion.

Large Stone

Treatment	Success	Failure	Total
X	381	145	526
Y	55	25	80
Total	436	170	606

$$\text{rate}(\text{Success}|\text{X}) = \frac{381}{526} = 0.724 > \text{rate}(\text{Success}|\text{Y}) = \frac{55}{80} = 0.688$$

Now we are observing that treatment X is positively associated with success in treating large stones.

Small Stone

Treatment	Success	Failure	Total
X	161	13	174
Y	234	36	270
Total	395	49	444

$$\text{rate}(\text{Success}|\text{X}) = \frac{161}{174} = 0.925 > \text{rate}(\text{Success}|\text{Y}) = \frac{234}{270} = 0.867$$

Now we are observing that treatment X is positively associated with success in treating small stones also.

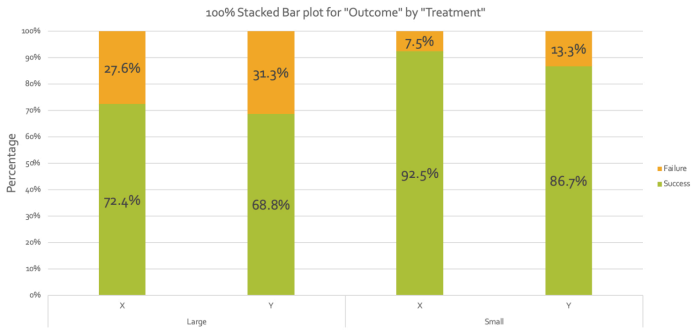


Figure 18: Sliced Bar Graph showing the Kidney Stones

	Large Stones			Small Stones			Total (Large + Small)		
Treat-ment	**Suc-cess**	**Total Trt**	**R(succ)**	**Suc-cess**	**Total Trt**	**R(succ)**	**Suc-cess**	**Total Trt**	**R(Suc-cess)**
X	381	526	72.4%	161	174	92.5%	540	700	77.4%
Y	55	80	68.8%	234	270	86.7%	289	350	82.6%

Let's take a look at the table to figure out why when individual Treatment X is better but combined Treatment Y is better. From the **Basic rule of rates** we know that the rate is between $\text{rate}(A \mid B)$ and $\text{rate}(A \mid \text{NB})$. Since the $\text{rate}(B) = \frac{526}{700}$ is higher it's closer to $\text{rate}(A|B)$ which is 72.4. Apply the same rule we can see the same for why Treatment Y has a higher percentage.

The reason Simpson's paradox happens is due to a confounder. if Simpson's paradox exists ➪ confounder exist. Confounder presence does not mean Simpson's paradox exists.

2.4 Confounder

A confounder (or confounding variable) is an external variable that influences both the independent variable and the dependent variable in a study, potentially leading to a misleading or biased association between them. In other words, a confounder creates a false impression of causation or exaggerates the association between variables.

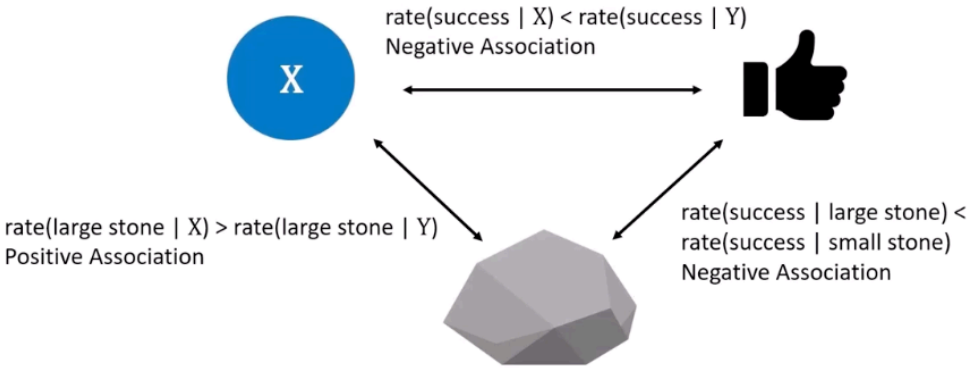


Figure 19: Prove of confounder

In the kidney stone case, **stone size** acts as a confounder. To identify if a variable is a confounder, we can examine its association with both the outcome and the other variable of interest. A variable is considered a confounder only if it is associated with both the outcome (treatment success or failure) and the treatment type (X or Y).

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.

To summarize we can do random assignments but it is not always possible due to ethical reasons.

3. Dealing with 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

A histogram is a graphical representation of the distribution of numerical data. It organizes data into **bins (or intervals)** and displays the **frequency (count)** of data points falling into each bin.

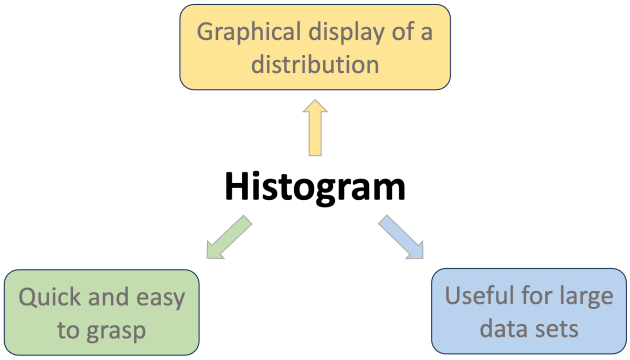


Figure 20: Benefits of Histogram

Let's take a look at Histogram.

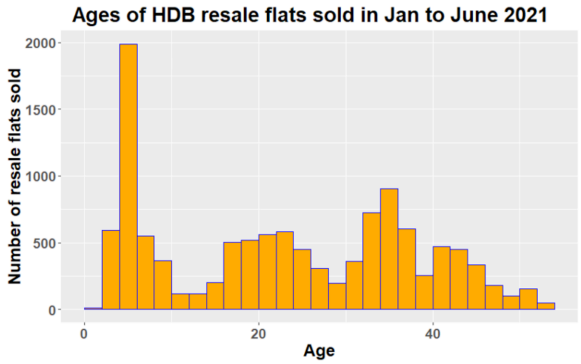


Figure 21: Example of Histogram

3.2.1 Bins

One of the benefits of having a Histogram is that you can sort them into bins like it been done in the above figure.

Age	Frequency
2	9
3	8
4	583
5	1105
6	884
7	295
8	255
9	219
10	147

→

Bins	Frequency
0-2	**9**
2-4	**591** = 8 + 583
4-6	**1989** = 1105 + 884
6-8	**550** = 295 + 255
8-10	**336** = 219 + 147

Figure 22: Converting into Bins

You can see above that it's been converted for individual ages into a bin form where the ages range from a certain number. Like 3-4 is combined together to form the range 2 < x <= 4

Sturges' Rule:

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

Where:

k: The number of Bins

n: Number of Data points

Freedman-Diaconis Rule:

$$\text{Bin Width} = 2 * \frac{\text{IQR}}{\sqrt[3]{n}}$$

Where:

Bin Width: The width of each bin in the histogram

IQR: Interquartile Range (difference between Q3 and Q1)

n: Total number of observations (data points).

Got general advice

3.2.2 Describing Distributions

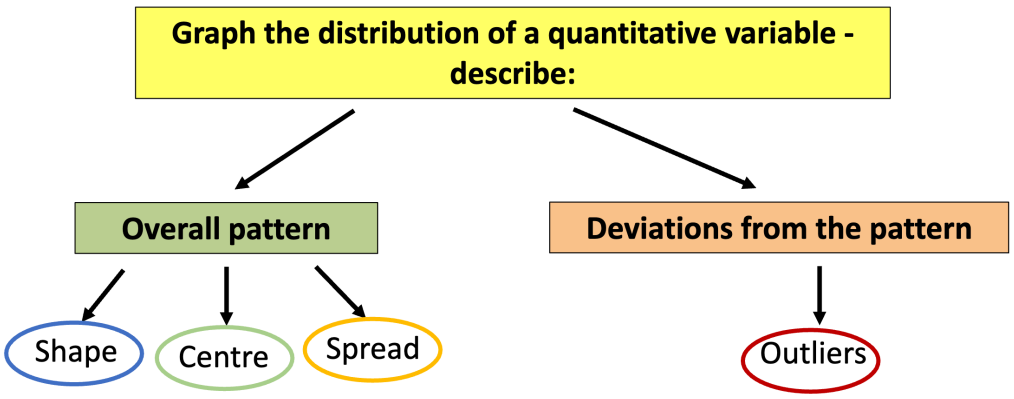


Figure 23: Different types of Distributions

3.2.2.1 Shape

The shape of the can be split into 2 components peaks and skewness.

Peak Three types of pick, Unimodel, Bimodal, Multimodal.

Unimodel

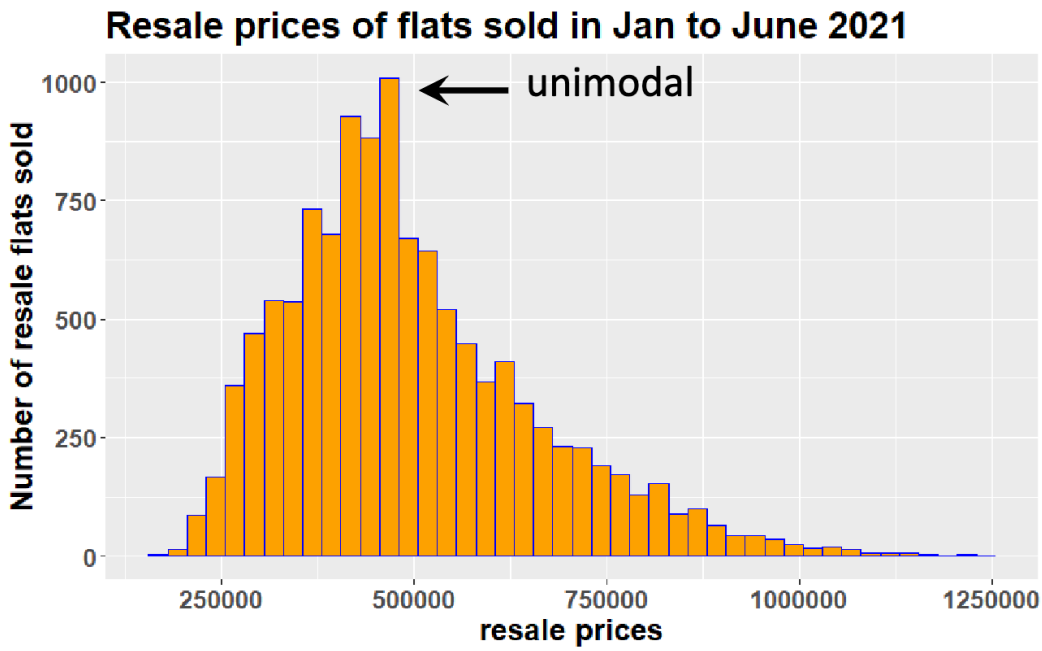


Figure 24: Unimodal

- Look for a bar where all the other values are formed around
- A single tall bar or a cluster of bars in the middle or at any position along the x-axis.
- Forms a bell-shaped curve, often resembling a normal distribution.
- Right-skewed (tail to the right) or left-skewed (tail to the left).

Bimodal

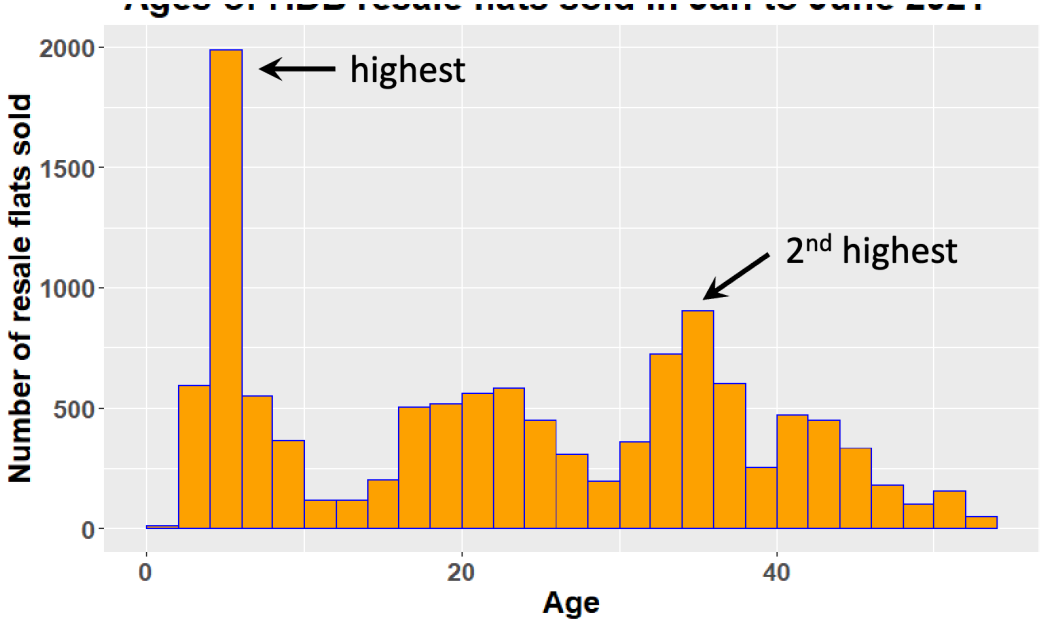


Figure 25: Bimodal

- Just look at the two cluster with two peak values
- Look for two prominent peaks that are clearly separated.
- Ensure the peaks are not just random fluctuations in the data but indicate meaningful groupings.
- If there are two distinct modes, the data is bimodal.

Multimodal

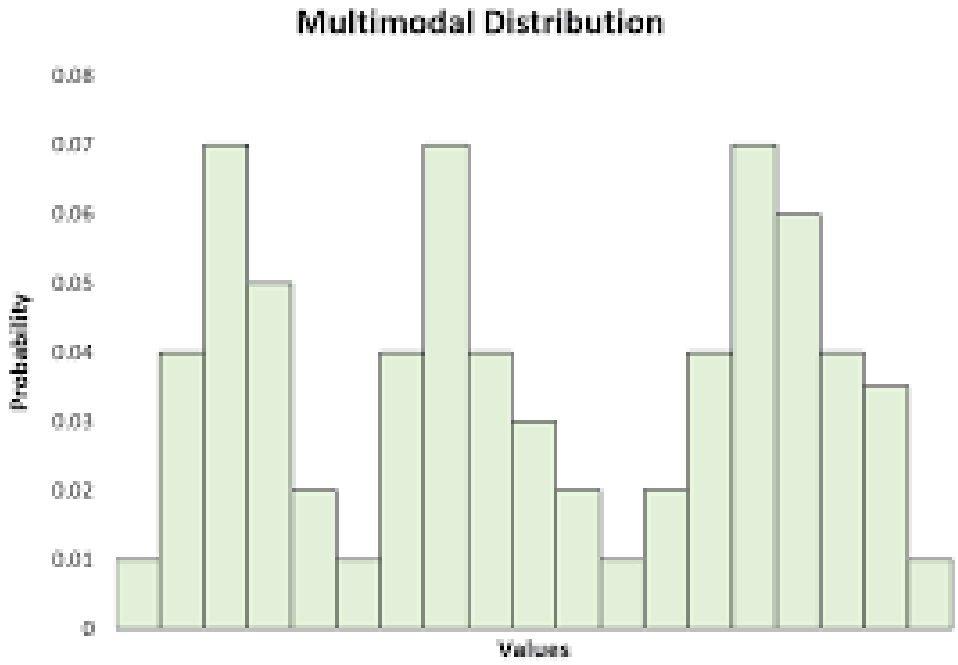


Figure 26: Multimodal

- There are three or more clusters and look for three or more peak values
- if the dataset has three or more modes, it's multimodal.

Skewed

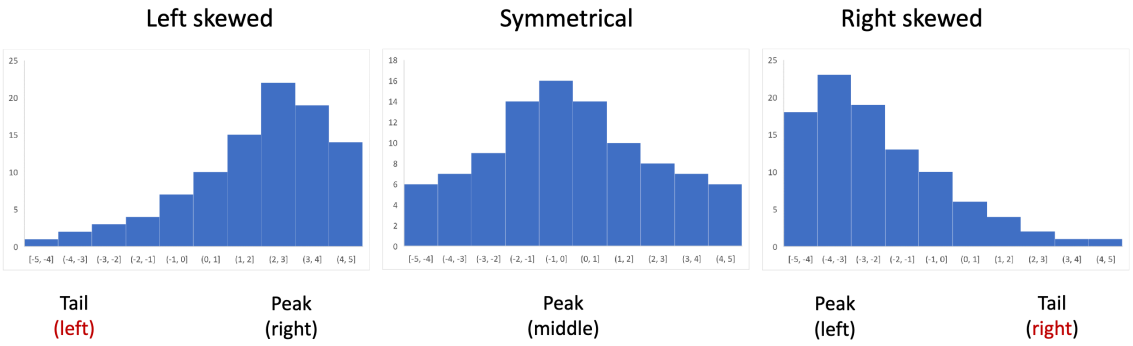


Figure 27: Different types of Skew

Symmetrical

- The left and right sides of the histogram are mirror images.
- The data is evenly distributed around the central peak.
- Mean ≈ Median ≈ Mode

Left Skewed

- This is also called the positive skew
- The tail extends to the right, and the bulk of the data is concentrated on the left.
- The peak is closer to the lower values.
- Mean > Median > Mode

Right Skewed

- This is also called the negative skew
- The tail extends to the left, and the bulk of the data is concentrated on the right.
- The peak is closer to the higher values.
- Mean < Median < Mode.

Normal Distribution

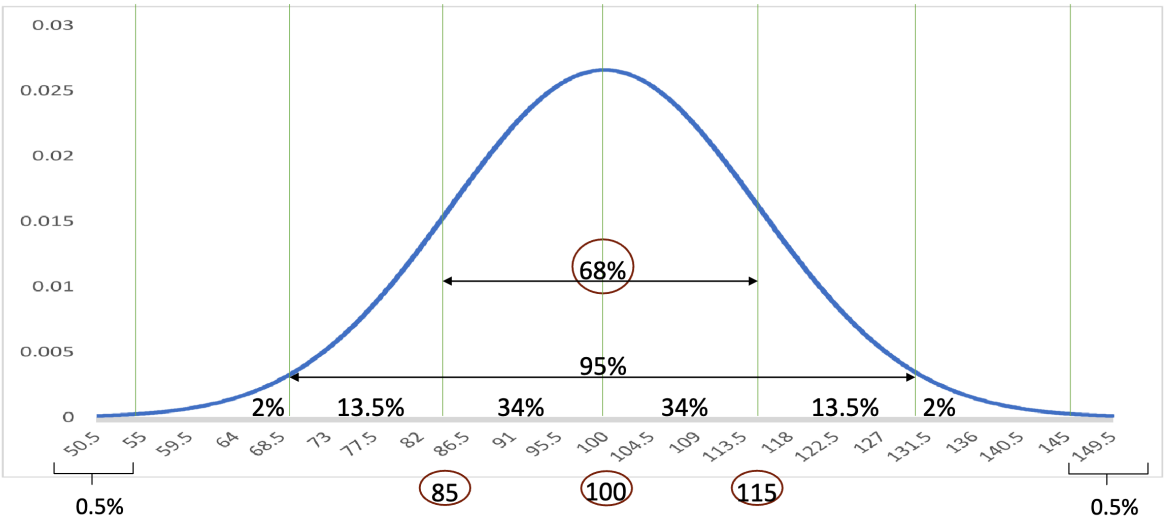


Figure 28: Normal Distribution

1. **Symmetry:**
   - In a normal distribution, the data is symmetrically distributed around the mean.
   - In a boxplot, the median line (Q2) is typically centered within the box, and the whiskers are of similar length on both sides, showing symmetry.
2. **Quartiles and Spread:**
   - The middle 68% of data (as shown between −1 and +1 standard deviations) corresponds to data concentrated around the mean in the boxplot.
   - In a boxplot, this middle spread is represented by the interquartile range (IQR), the distance between Q1 (25th percentile) and Q3 (75th percentile).
3. **Outliers:**
   - Points outside ±2 or ±3 standard deviations in the normal curve are rare (5% or 0.5% probability) and often classified as outliers.
   - In a boxplot, these extreme values are represented as individual points outside the whiskers.
4. **Range:**
   - The range in a boxplot (min to max) captures 100% of the data, similar to how the tails of the normal distribution theoretically extend indefinitely.

3.2.2.2 Centre

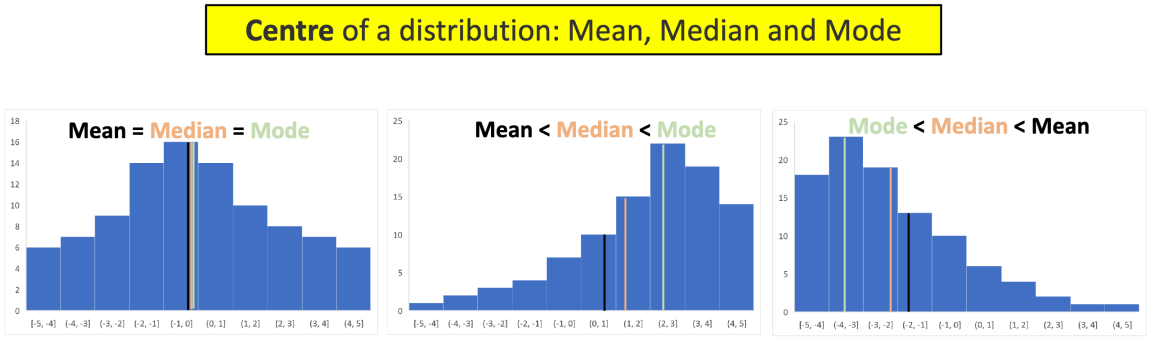


Figure 29: Centre Distribution

3.2.2.3 Spread

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

Standard Deviation

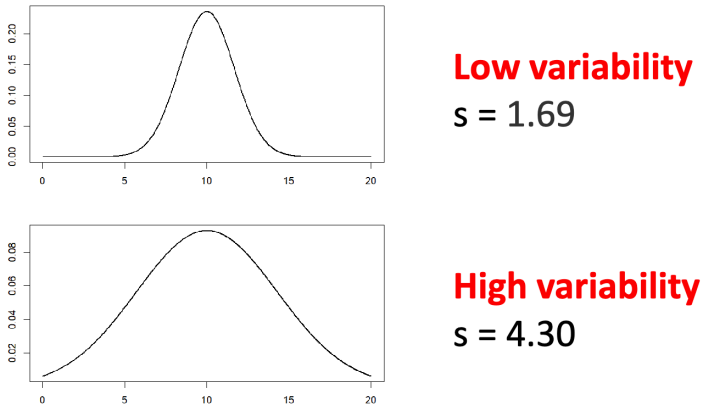


Figure 30: Spread of Data

- **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.

3.2.2.4 Outliers

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300
Mean = 5.45	Mean = 30
Median = 5	Median = 5.5
Mode = 5	Mode = 5
Standard Deviation = 1.04	Standard Deviation = 85.03

In a histogram, outliers show up as isolated bars far from the main distribution. They can significantly affect how you interpret the spread, central tendency, and variability of the data.

Note: may be good practice to repeat the analysis with and without the outliers

3.2.3 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

3.3 Boxplots

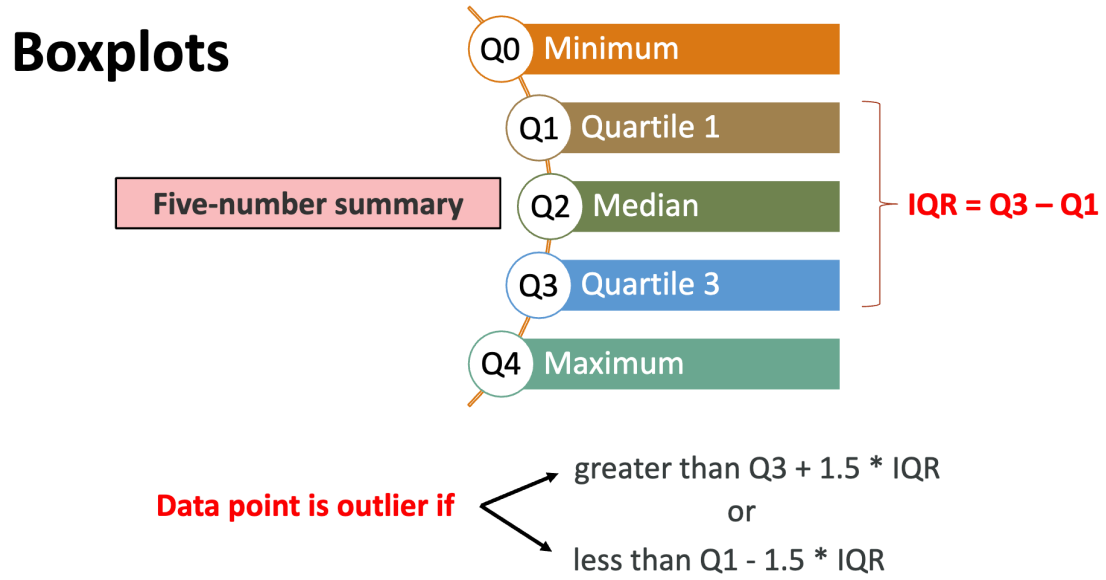


Figure 31: Five-number summary

How to construct a box plot

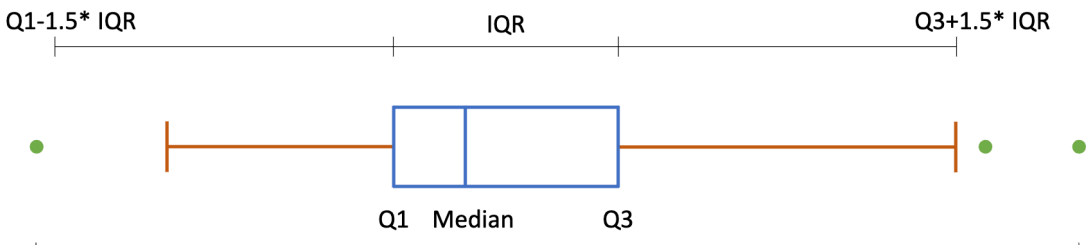


Figure 32: Boxplot

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

Note: When there is a X mark it indicates **mean**

3.3.1 Boxplots 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.

3.4 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.

Bivariate is not a **deterministic relationship:**

1. **Deterministic Relationship:**
   - A deterministic relationship means that one variable completely predicts the other with no uncertainty.
   - An **example** of the relationship between Celsius = Fahrenheit

2. **Bivariate Relationship:**
   - A deterministic relationship means that one variable completely predicts the other with no uncertainty.
   - 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.
   - **Example**: Height and weight are related (positively correlated), but knowing someone's height doesn't precisely predict their weight due to other influencing factors.

You can do bivariate data analysis in three ways:

Scatter Plots	Correlation Coefficients	Regression Analysis
Get an idea of the pattern	Check for a linear relationship	Fit a line or curve to the data

3.5 Scatter Plot

A scatter plot is a graphical representation of the relationship between two variables:

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

We can use scatter plot to visualize:

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.

We can interpret the scatter plots in the following ways:

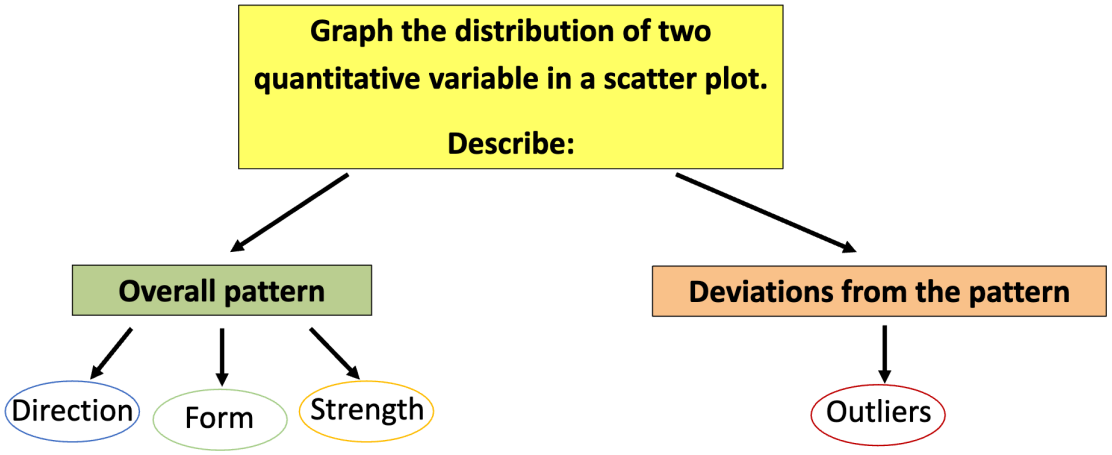


Figure 33: Interpreting Scatter Plots

Direction

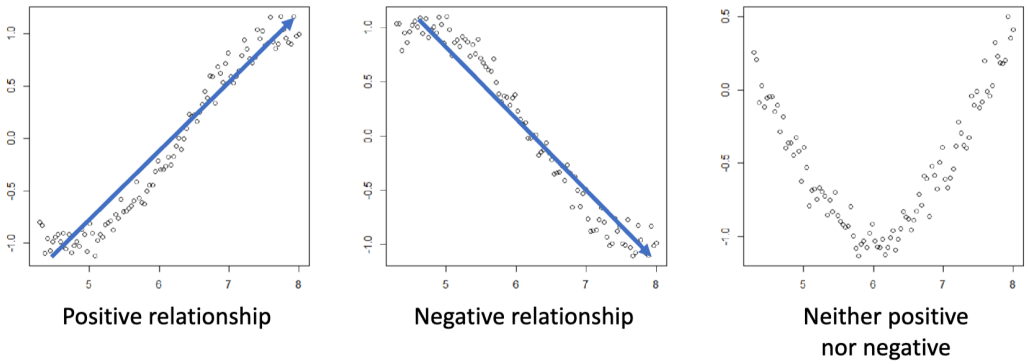


Figure 34: Directions

Form

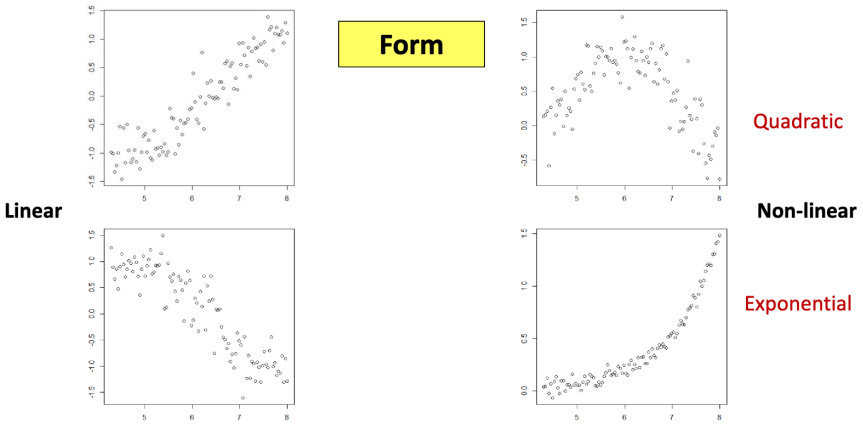


Figure 35: Forms

Strength

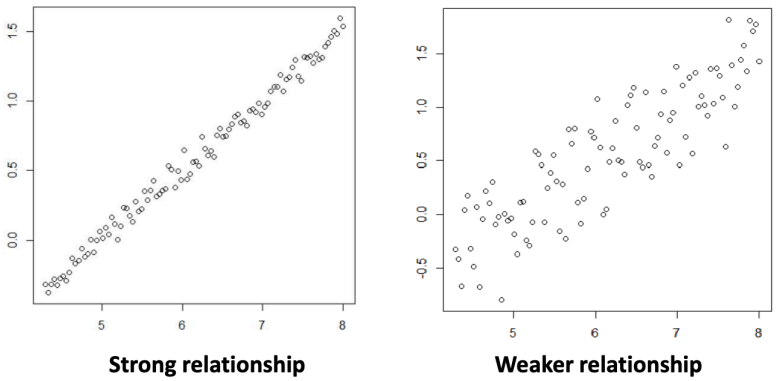


Figure 36: Forms

- 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.

3.6 Correlation Coefficient

Correlation Coefficient is used to measure the **strength** and **direction** of linear relationship. The range of the r is −1 to 1.

- $r > 0$ -> 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)

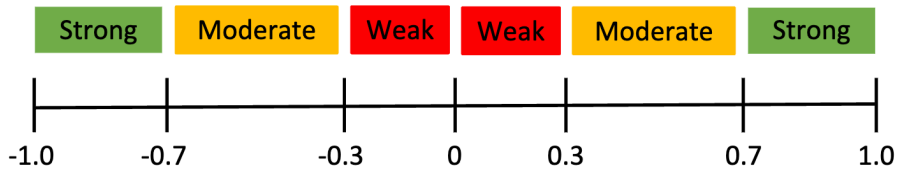


Figure 37: Forms

Perfect Linear Association

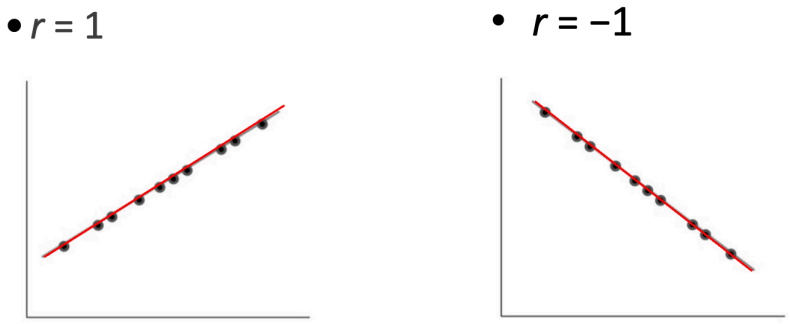
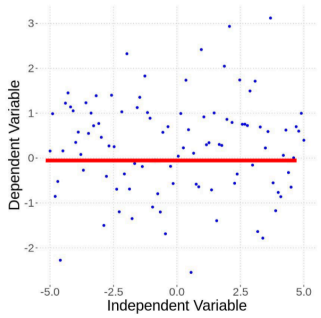


Figure 38: Forms

No Linear Association

- No Linear or Non-Linear Relationship



- Quadratic Relationship

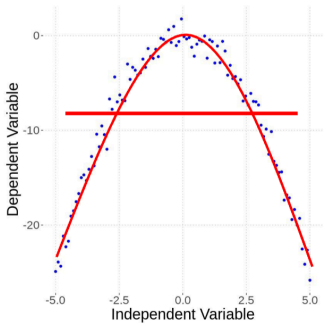


Figure 39: Forms

Calculate Correlation Coefficient r

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

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

Example

X	Y
9	41
4	17
5	28
10	50
6	39
3	26
7	30
2	6
8	4
1	10

Average of X	5.5
Average Y	25.1
Standard deviation of X	3.03
Standard deviation of Y	15.65

Finding SU. Just the first instance

$$X = \frac{9 - 5.5}{3.03} = 1.16$$

$$Y = \frac{41 - 25.1}{15.65} = 1.02$$

X (Standard Unit)	Y (Standard Unit)	Product
1.16	1.02	1.17
−0.50	−0.52	0.26
−0.17	0.19	−0.03
1.49	1.59	2.36
0.17	0.89	0.15
−0.83	0.06	−0.05
0.50	0.31	0.15
−1.16	−1.22	1.41
0.83	−1.35	−1.11
−1.49	−0.96	1.43

Note: if you wondering why 1.17 and not 1.16 there since it's rounded it's like this

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

3.7 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

Outlier When we observe and see data that fall far from the main cluster points it will affect our correlation coefficient by either increasing or decreasing them.

3.8 Linear Regression

So the question you have in my mind we know about correlation but how is it useful it gives us the **predictive power.** Linear regression help us do that

Formula:

$$Y = r * \frac{S_y}{S_x} + b$$

where: b: intercept of the regression line (value of Y when X = 0)

What if we have a non-linear (exponential) relationship using linear regression by transforming the variables into logarithmic scale

Formula for non-linear

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

1. For each data point (x,y), compute (x, ln y)
2. Find a linear regression line.

$$\ln(y) = m * x + b$$

Where:

$\ln(y)$: The natural logarithm of the dependent variable y	x: The independent variable	m: The slope of the line, which represents how much $\ln(y)$changes for a one-unit increase in t	b: The y-intercept, which is the value of $\ln(y)$ when t = 0

Formula to find Covariance $\mathrm{Covariance}(\ln\ \mathrm{y},\ \mathrm{x}) = \frac{\sum_{i=i}^{n}(X_i - X_m)(Y_i - Y_m)}{n}$

Formula to find m:

$$m = \frac{\mathrm{Covariance}(\ln\ \mathrm{y},\ \mathrm{x})}{\mathrm{Variance}(\mathrm{x})}$$

3. Since with have the value of ln(m) and ln(b), we can get

$$m = e^{\ln(m)}\ \text{and}\ b = e^{\ln(b)}$$

4. You can then write down the exponential equation relating y and t. **Example**:

1.

$$\ln y = 4.287 + 0.066x$$

2.

$$\ln(m) = 4.287\ \text{and}\ln(b) = 0.066$$

3.

$$m = e^{4.287}\ \text{and}\ b = e^{0.066}$$

4.

$$y = \mathrm{cm}^x = e^{4.287}e^{0.066}x$$

3.9 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 policy-making 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.

3.9.1 Ecological Fallacy

The ecological fallacy occurs when inferences about individual behavior or characteristics are drawn from data aggregated at a group or population level. It assumes that relationships observed for groups hold true for individuals within those groups, which can lead to incorrect conclusions.

Examples:

1. **Group-Level Correlation**
   - 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.
2. **Health Studies**
   - A study might show that regions with higher air pollution have higher asthma rates.
   - Ecological fallacy: Assuming every individual exposed to air pollution in those regions develops asthma.

3.9.2 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

Statistical inference exists because it is **often impractical or impossible to gather complete data** from an entire population. Instead, we use sample data a subset of the population to **draw conclusions and make predictions** about the whole population. This process involves assessing the likelihood that findings from the sample can be generalised to the population as a whole.

4.1 Introduction to Probability

There is a concept called **probability experiment** which is the basis for making generalizations from samples to populations. **Probability experiment** is simply an activity or event where you don't know what will happen for sure, but you do know all the possible things that could happen.

An example of probability space:

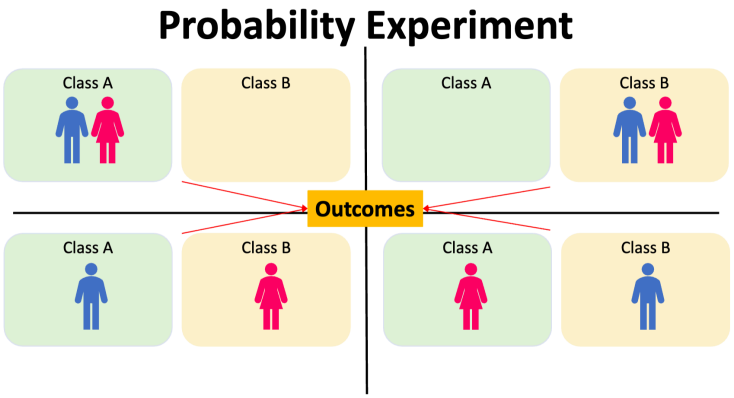


Figure 40: Probability Space

Let's take this example where we have Andrea and Izen how likely are they to be enrolled in the same class? The above figure shows **all the possible outcomes**.

- 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.

4.1.1 Rules of probabilities

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 $P(E \cup F)$ = $P(E) + P(F)$

Note: Finite must all add to 1

4.1.2 Uniform Probabilities

Uniform Probability is a concept in probability theory where all outcomes in a sample space are equally likely to occur. This is the concept used for random sampling.

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

4.2 Conditional Probability

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

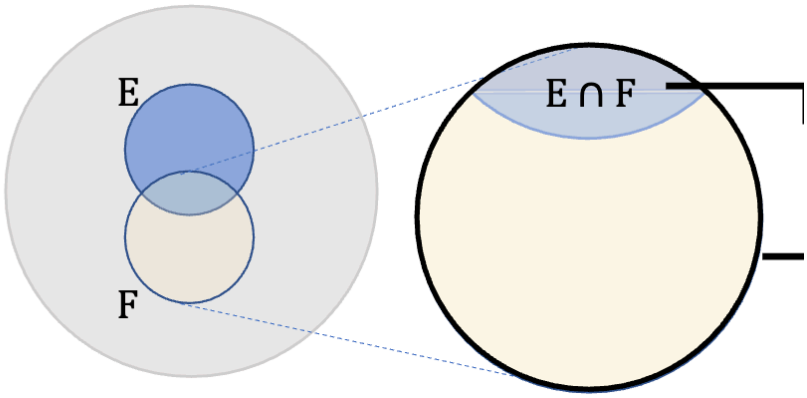


Figure 41: Conditional Probability

You read this has a Probability of E given F. You have to view the figure to get a general idea of how this works.

4.3 Does $P(A|B) = \text{rate}(A|B)$?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\text{rate}(A \cap B)}{\text{rate(B)}} = \left(\frac{\frac{\text{Size of } A \cap B}{\text{Size of Sampling Frame}}}{\frac{\text{Size of B}}{\text{Size of Sampling Frame}}} \right) = \frac{\text{Size of } \ A \cap B}{\text{Size of B}} = \text{rate}(A|B)$$

Yes, it does. It provides a mathematical basis for treating rate(A|B) as equivalent to P(A|B) in statistical inference.

4.4 Independent between two events

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

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

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

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

- Helps verify independence between events.
- Provides a mathematical foundation for analyzing probabilistic systems or scenarios.

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

4.5 Prosecutor's Fallacy

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

Unless

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

Assuming independence when it does not exist: if evidence (B) is independent of guilt (A) then $P(A \cap B) = P(A) * P(B)$

4.6 Law of Total Probability

The Law of Total Probability is a fundamental rule that allows us to compute the probability of an event A by considering all possible ways that A can occur, based on a partition of the sample space.

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

Can be written as

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \times P(B_i)$$

Example:

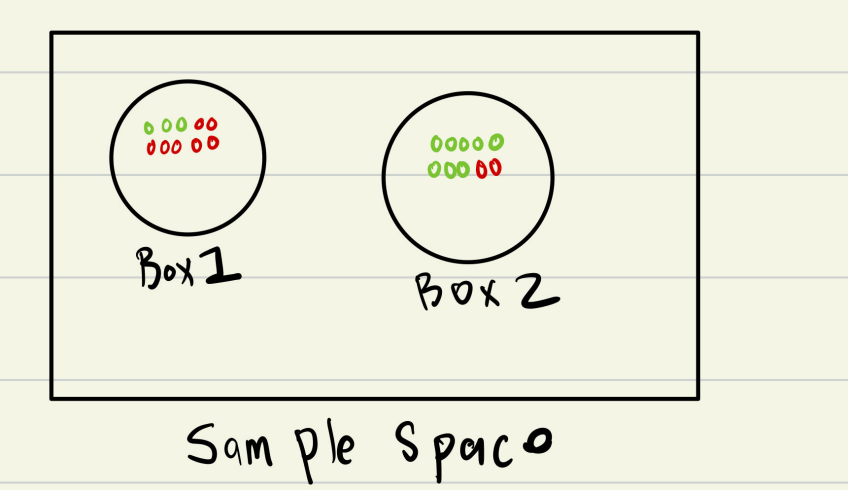


Figure 42: Marble in 2 boxes

What would I need to do to find the $P(\text{Green Marble})$ using the law of Total Probability?

$$P(\text{Green Marble}) = P(\text{Green Marble}|\text{Box 1}) * P(\text{Box 1}) + P(\text{Green Marble}|\text{Box 2}) * P(\text{Box 2})$$

$$P(\text{Green Marble}) = 0.3 \times 0.5 \times +0.8 \times 0.5 = 0.55$$

4.7 Conjunction Fallacy and Base Rate Fallacy

4.7.1 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

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

4.7.2 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.

Example:

- Rise in the number of drink-driving cases
- Breathalyser developed to detect drivers who drive after drinking
- Breathalyser falsely detects 5% of sober drivers as being drunk
- Breathalyser never fails to detect a truly drunk person
- 1 in 1000 drivers drives after drinking

We have to include the base rate in this case it is the **1 in 1000 drivers.**

Status	Positive Test (+)	Negative Test (-)	Total
Drunk	100	0	100
Sober	4,995	94,905	99,900
Total	5,095	94,905	100,000

$$P(\text{Drunk}| +) = \frac{100}{5095} = 0.019627 \approx 2\%$$

Bayes Theorem Using this only if data is not explicitly tabulated

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

4.8 Random Variable

A random variable is a variable that takes on numerical values determined by the outcome of a random experiment. It provides a way to map outcomes of the experiment to numbers, enabling the use of mathematical tools to analyze and understand probabilities.

Two types of Random Variable:

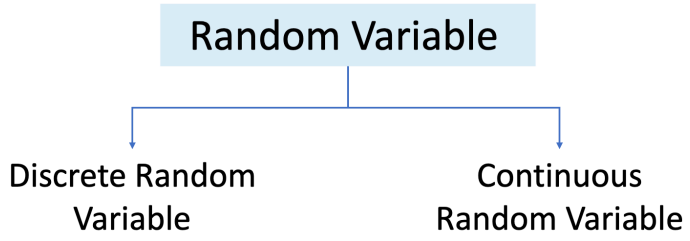


Figure 43: Types of Random Variable

1. Discrete Random Variable: Takes on a finite or countable number of distinct values (e.g., the number of heads in a coin toss).
2. Continuous Random Variable: Takes on an infinite number of values within a range or interval (e.g., the time it takes for an event to occur).

4.9 Introduction to Statistical inference

When a census is available, it is possible to conclude the entire, but we use **statistical inference** since that's not possible.

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

Fundamental Rule for using Data for inference: Available data can be used to make inferences about a much larger group if the data can be considered to be representative with regard to the question of interest

How to eliminate bias from Sample statistics?

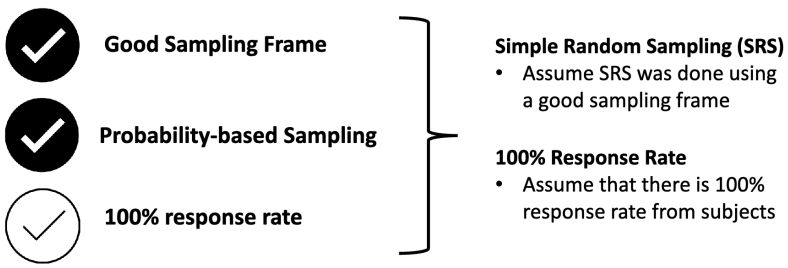


Figure 44: Eliminating Bias

- **Bias** arises when the sample data is not representative of the population. This can occur due to poor sampling methods or non-responses.
- **To eliminate bias:**
  - Use a probability sampling method: This ensures every member of the population has a known, non-zero chance of being selected. Examples include simple random sampling and stratified sampling.
  - Achieve a 100% response rate: This ensures that all selected participants provide their data, preventing systematic exclusion of certain groups.

4.10 Confidence Interval

Confidence intervals address random error by providing a range of values where the true population parameter is likely to exist. They help quantify the uncertainty inherent in using sample data to make inferences about a population.

here are **two types of confidence intervals** commonly used:

1. **Population Proportion:**
   - Refers to the fraction or percentage of the population that exhibits a specific characteristic.
   - Example: The proportion of females in Singapore.
2. **Population Mean:**
   - Refers to the average value of a characteristic across the population.
   - Example: The average math score for students in a school.

4.10.1 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

4.10.2 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

4.10.3 Example: A Survey on Coffee Consumption

1. A researcher surveys **500 people** in a city to understand coffee-drinking habits.
2. They collect two types of data:
   - **Proportion Data**: 300 out of 500 people (60%) say they drink coffee daily.
   - **Mean Data**: The daily coffee consumption for these 500 people is recorded, and the **sample mean** is 2.3 cups, with a **sample standard deviation** of 0.8 cups.

The researcher wants to:
1. Calculate a **95% confidence interval** for the **proportion** of coffee drinkers in the city.
2. Calculate a **95% confidence interval** for the **mean** number of cups consumed daily by the population.

Population Proportion:

$$\frac{300}{500} \pm 1.96 \times \sqrt{\frac{\left(\frac{300}{500}\right)\left(1-\frac{300}{500}\right)}{500}} = 0.0219$$

Population Mean:

$$2.3 \pm 1.96 \times \frac{0.8}{\sqrt{500}} = 0.0701$$

4.10.4 Properties of Confidence Intervals

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.

4.11 Hypothesis Testing

Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on sample data. It involves testing an assumption (or hypothesis) about a population parameter.

The Significance Level (α) is $\alpha = 1 - \text{Confidence Level}$. So if the confidence level is 95%, then the Significance Level is 0.05.

Hypothesis testing can be interpreted through confidence intervals. If the confidence interval does not contain the value specified in the null hypothesis, reject H_0.

Steps for hypothesis testing:

1. Determine the null and the alternative hypotheses
2. Set the significance level (typically 5%)
3. Find the relevant sample statistic
4. Calculate the p-value
5. Conclusion of hypothesis test

Types of Hypothesis Test:

Hypothesis Test for Population Proportion	Hypothesis Test for Population Mean (t-test)	Hypothesis Test for association (chi-squared test for association)

4.11.1 P-value

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**.

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.

4.11.2 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.

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

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

4.11.3 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.

4.11.4 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.

$$H_0 : \text{The two variables are independent}$$

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