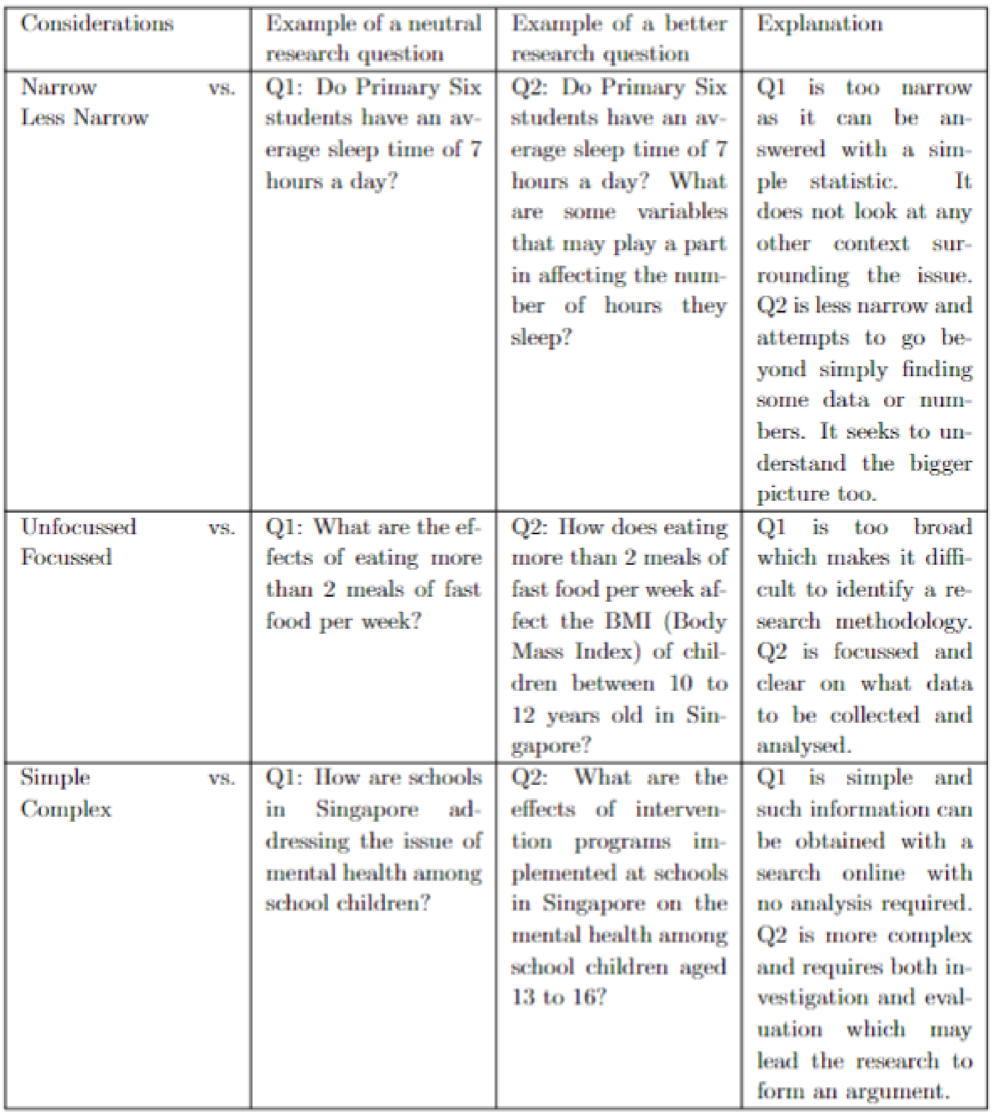
Chapter 1 Getting Data

1.1 Exploratory Data Analysis (EDA)

Definition: Exploratory Data Analysis (EDA) is a systematic process of examining a dataset to identify patterns, summarize variables, and visualize data.



Considerations	Example of a neutral research question	Example of a better research question	Explanation
Narrow vs. Less Narrow	Q1: Do Primary Six students have an average sleep time of 7 hours a day?	Q2: Do Primary Six students have an average sleep time of 7 hours a day? What are some variables that may play a part in affecting the number of hours they sleep?	Q1 is too narrow as it can be answered with a simple statistic. It does not look at any other context surrounding the issue. Q2 is less narrow and attempts to go beyond simply finding some data or numbers. It seeks to understand the bigger picture too.
Unfocussed vs. Focussed	Q1: What are the effects of eating more than 2 meals of fast food per week?	Q2: How does eating more than 2 meals of fast food per week affect the BMI (Body Mass Index) of children between 10 to 12 years old in Singapore?	Q1 is too broad which makes it difficult to identify a research methodology. Q2 is focussed and clear on what data to be collected and analysed.
Simple vs. Complex	Q1: How are schools in Singapore addressing the issue of mental health among school children?	Q2: What are the effects of intervention programs implemented at schools in Singapore on the mental health among school children aged 13 to 16?	Q1 is simple and such information can be obtained with a search online with no analysis required. Q2 is more complex and requires both investigation and evaluation which may lead the research to form an argument.

Steps in EDA:

1. Generate research questions about the data.
2. Explore answers using visualization tools and statistical modeling (e.g., regression).
3. Reflect: Does the data answer our research questions?
4. Refine questions or generate new ones for further exploration.

Key Example:
From an article discussing trends in Singapore marriages and divorces during COVID-19, one might ask:

- What kind of data supports this conclusion?
- Is the conclusion valid?

1.2 Sampling

Definitions:

1. Population: The entire group of interest (e.g., all university students).
2. Sample: A subset of the population used for analysis.
3. Population Parameter: Numerical facts about the population (e.g., mean, median).
4. Census: Data collection from every member of the population (often costly and time-intensive).
5. Estimate: Inference about the population parameter derived from the sample.
6. Sampling Frame: The list from which the sample is drawn.

Bias in Sampling:

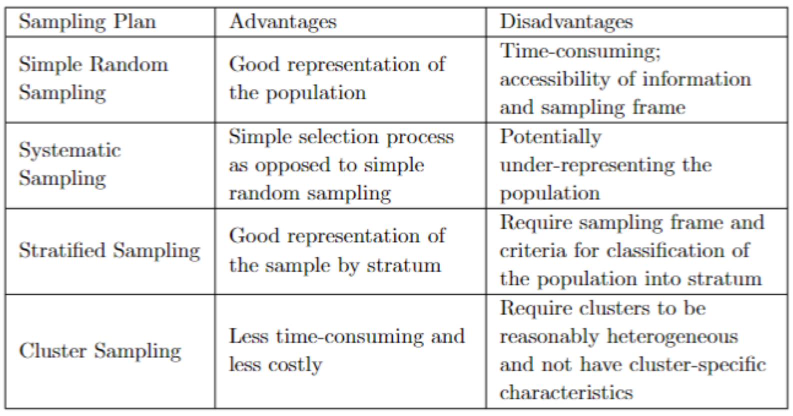
- Selection Bias: When parts of the population are systematically excluded.
- Non-Response Bias: When selected individuals do not participate, skewing the results.

Example:

- Selection Bias: Sampling only engineering students for a university-wide study excludes students from other faculties.
- Non-Response Bias: Students may avoid surveys about financial assistance due to privacy concerns.

Sampling Methods:

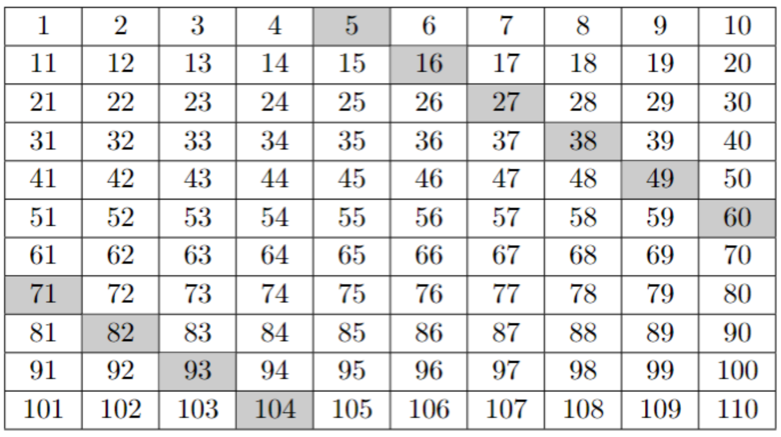
1. Probability Sampling: Every unit has a known, non-zero chance of selection.
   - Simple Random Sampling (SRS): Each unit has an equal chance (e.g., lucky draw tickets).
   - Systematic Sampling: Select every k^{th} individual after a random start.
   - Stratified Sampling: Divide the population into strata (e.g., by gender) and sample from each.
   - Cluster Sampling: Divide into clusters (e.g., schools) and sample entire clusters.
2. Non-Probability Sampling: Selection is not random (prone to bias).
   - Convenience Sampling: Survey those easiest to access (e.g., mall shoppers).
   - Volunteer Sampling: Participants self-select, often leading to skewed results.



Sampling Plan	Advantages	Disadvantages
Simple Random Sampling	Good representation of the population	Time-consuming; accessibility of information and sampling frame
Systematic Sampling	Simple selection process as opposed to simple random sampling	Potentially under-representing the population
Stratified Sampling	Good representation of the sample by stratum	Require sampling frame and criteria for classification of the population into stratum
Cluster Sampling	Less time-consuming and less costly	Require clusters to be reasonably heterogeneous and not have cluster-specific characteristics

Example - Systematic Sampling

Suppose we know there are 110 sampling units in the population and we would like to select a sample with 10 units. Imagine the sampling units are numbered from 1 to 110



1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110

Since $p = 110$, and $n = 10$, we can select one unit from every $k = \frac{110}{10} = 11$, so we randomly select a number from 1 to 11, which will start off the sampling process, and skip after every k, so if we selected 5 to start, we would have $\{5, 16, 27, 38, 49, 60, 71, 82, 93, 104\}$.

Generalizability Criteria:

- Sampling frame should cover the entire population.
- Use probability-based sampling.
- Large sample size reduces random error.
- Minimize non-response.

1.3 Variables and Summary Statistics

Types of Variables:

1. **Categorical**:
   - **Nominal**: Categories without order (e.g., gender).
   - **Ordinal**: Ordered categories (e.g., happiness scale).
2. **Numerical**:
   - **Discrete**: Countable values (e.g., number of modules taken).
   - **Continuous**: Any value in a range (e.g., height).

Independent vs. Dependent Variables:

- **Independent**: Manipulated to observe effects.
- **Dependent**: Measured for changes.

Example:

- Independent: Time spent gaming.
- Dependent: Exam scores.

1.4 Summary Statistics - Mean

Mean ($\bar{x}$):

The average of a dataset:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Properties:

1. Adding a constant c to all values increases the mean by c.
2. Multiplying all values by c scales the mean by c.

1.5 Variance and Standard Deviation

Definitions:

1. **Variance**:

$$\text{Var} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

2. **Standard Deviation (s_x)**:

$$s_x = \sqrt{\text{Var}}$$

Example

The highest temperature recorded per month:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
30.1	31.1	31.8	32.1	31.9	32.6	33.0	32.4	32.0	32.5	31.3	29.6

Mean $\bar{x}$:

$$\frac{30.1 + 31.1 + 31.8 + 32.1 + 31.9 + 32.6 + 33.0 + 32.4 + 32.0 + 32.5 + 31.3 + 29.6}{12} = 31.7$$

Variance (Var) :

$$\frac{1}{11}\left((30.1 - 31.7)^2 + (31.1 - 31.7)^2 + \cdots + (31.3 - 31.7)^2 + (29.631.7)^2\right) \approx 1.038$$

Standard Deviation (s_x):

$$s_x = \sqrt{Var} \approx \sqrt{1.038} \approx 1.019$$

Key Points:

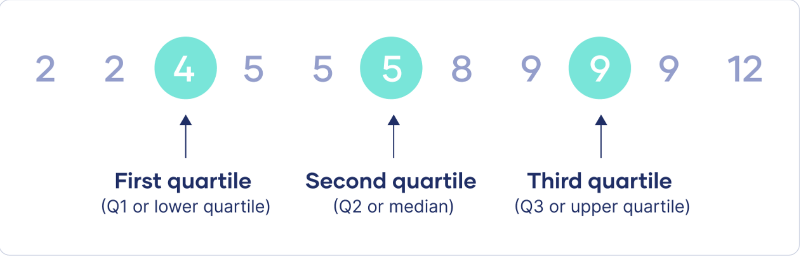
- Variance uses squared differences to avoid cancellation of positive and negative deviations.
- Standard deviation shares the same units as the original data.

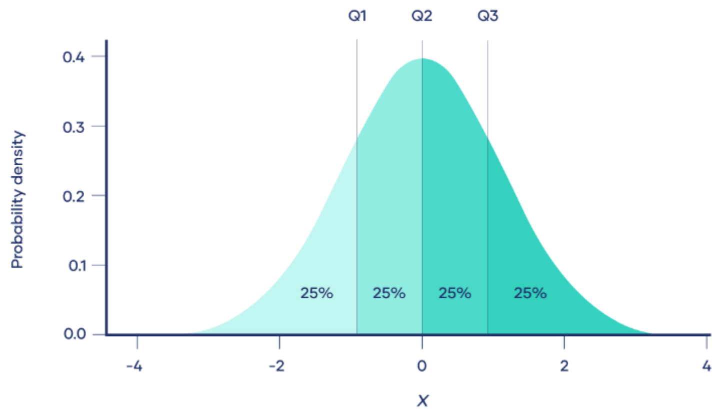
1.6 Median, Quartiles, and IQR

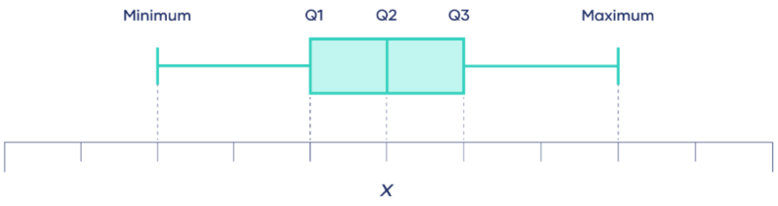
Definitions:

- **Median**: Middle value in ordered data.
- **Quartiles**:
  - $Q1$: 25th percentile.
  - $Q3$: 75th percentile.
- **IQR**:

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







Properties:

- Adding a constant c affects $Q1$ and $Q3$ but not IQR.
- Multiplying by c scales $Q1$, $Q3$, and IQR by $|c|$.

1.7 Study Designs

Types:

1. **Experimental Studies**:
   - Manipulate an independent variable.
   - Use **random assignment** to control bias.
   - Example: Testing a drug's effectiveness with treatment and control groups.
2. **Observational Studies**:
   - Observe variables without manipulation.
   - Less definitive for causation due to potential confounding factors.

Techniques to Reduce Bias:

- **Blinding**: Subjects (and sometimes researchers) do not know their group.
- **Placebo**: Control group receives an inert treatment to account for psychological effects.