GEA1000 Quantitative Reasoning with Data
AY23/24, Y1S1
Notes

Sim Ray En Ryan

December 3, 2025

Contents

1 Data

1.1 Research Questions (RQ)

Research questions typically aim to:

- **Estimate** something in the population.
- **Test** a claim in the population.
- **Investigate** a relationship between two variables in a population.

1.2 Variables

- **Independent vs. Dependent vs. Controlled** variables.
- **Categorical** variables:
  - **Ordinal** (categories have a natural ordering).
  - **Nominal** (categories do not have a natural order).
- **Numerical** variables:
  - **Discrete** (countable values).
  - **Continuous** (any value within a range).

1.3 Charts

Common charts for data visualization include:

- **Scatter Plots**.
- **Histograms**.
- **Box Plots**.

2 Sampling and Experimentation

2.1 Sampling

Sampling involves studying a **proportion of the population**.

- **Population**: The entire group of interest.
- **Population Parameter**: A numerical fact about a population.
- **Census**: Data collected from everybody in the population (not a sample method).

2.1.1 Probability Sampling

Methods where every unit has a known chance of being selected:

- **Simple Random**: All units have the same chance of selection.
- **Systematic**: Selecting every nth unit after the first r units.
- **Stratified Random**: Splitting the population into groups based on similar characteristics, then performing simple random sampling within each group.
- **Cluster**: Splitting the population into clusters, then performing simple random sampling of the clusters.

2.1.2 Non-Probability Sampling

Methods that do not use random selection:

- **Convenience**.
- **Volunteer**.

2.2 Generalisability and Bias

2.2.1 Generalisability

Generalisability (how well sample results apply to the population) is improved by:

- A larger **Sampling Frame**.
- A **Probabilistic Sampling Method**.
- A larger **Sample Size**.

2.2.2 Biases

- **Selection Bias**: Arises from an imperfect sample frame.
- **Non-response Bias**: Occurs due to inconvenience or unwillingness to disclose information.

2.3 Experimentation

Experiments are classified as **Observational vs. Controlled**.

- Subjects are typically divided into a **Treatment** group (receiving exposure) and a **Control** group (non-exposure).
- **Random Assignment** is used to ensure groups are comparable.
- **Blinding** (using a placebo) is used to mitigate bias.

3 Exploratory Data Analysis (EDA)

3.1 Univariate EDA

Analysis focused on a single variable.

3.1.1 Measures of Centrality and Shape

- **Mean vs. Median vs. Mode**.
- **Skewness** is determined based on the relative positions of the mean, median, and mode.
- The distribution's **Shape** includes observation of peaks and skewness.

3.1.2 Measures of Spread

- **Variance and Standard Deviation (s_x)**: $s_x = \sqrt{\text{Variance}}$.
- **Coefficient of Variation**: s_x/mean.
- **Interquartile Range (IQR)**: $Q3 - Q1$.

3.1.3 Box Plots and Outliers

A **Box Plot** summarizes the five-number summary:

- **Min (Q_0), Q_1, Median (Q_2), Q_3, Max (Q_5)**.
- $\text{IQR} = Q3 - Q1$.
- An **outlier** is a data point greater than $Q3 + 1.5 \times \text{IQR}$ or lower than $Q1 - 1.5 \times \text{IQR}$.
- Outliers should generally **not be removed** because they tell a story about the variable.

3.2 Bivariate EDA

Analysis focused on the **Association** between two variables.

- **Correlation (r)**: Measures the **sign** and **magnitude** of the linear relationship (e.g., $1, 0.7, 0.3, 0$).
- **Correlation $\neq$ Causation**.
- **Linear Regression**: Fits a straight line that will pass through the average of x and y. This method **cannot predict outside the range** of the observed data (extrapolation).

4 Rates and Association

4.1 Rates

- **Rates** are basically the probability that something happens.
- **Conditional Rates** correspond to conditional probability.
- **Joint Rate** refers to the probability of two events occurring together.

4.2 Association between Events

Association between events A and B can be **Positive, Negative, or No Association**.

4.2.1 Types of Association

- **No Association**: $\text{rate}(A|B) = \text{rate}(A|\text{NB})$.
- **Positive Association**: $\text{rate}(A|B) > \text{rate}(A|\text{NB})$. This implies $\text{rate}(B|A) > \text{rate}(B|\text{NA})$, $\text{rate}(\neg A|\neg B) > \text{rate}(\neg A|B)$, and $\text{rate}(\neg B|\neg A) > \text{rate}(\neg B|A)$.
- **Negative Association**: $\text{rate}(A|B) < \text{rate}(A|\text{NB})$, and so on.

4.2.2 Rules on Rates

- **Symmetry Rule**: $\text{rate}(A|B) > \text{rate}(A|\text{NB}) \iff \text{rate}(B|A) > \text{rate}(B|\text{NA})$.
- **Basic Rule**: The overall $\text{rate}(A)$ must be between $\text{rate}(A|B)$ and $\text{rate}(A|\text{NB})$.
  - The closer $\text{rate}(B)$ is to 100%, the closer $\text{rate}(A)$ is to $\text{rate}(A|B)$.
  - If $\text{rate}(B)$ is 50%, $\text{rate}(A)$ is the average of the two conditional rates.
  - If B is independent from A, $\text{rate}(A) = \text{rate}(A|B) = \text{rate}(A|\text{NB})$.

4.3 Simpson's Paradox and Confounders

- **Simpson's Paradox**: Occurs when a trend appears in more than half the groups of data, but disappears or reverses when the groups are combined.
- It may happen because the sample sizes of each categorical variable group are very different.
- The existence of Simpson's Paradox implies the existence of a **Confounder**.
- **Confounder**: A third variable associated with **both** the independent and dependent variable.
- To remove confounders, modify the sample so that the third variable is no longer associated with one of the primary variables.

5 Probability and Testing

5.1 Probability

- **Mutually Exclusive Events** (E and F): $P(E \text{ or } F) = P(E) + P(F)$.
- **Conditional Probability and Independence** are key concepts.
- **Prosecutor's Fallacy**: The error of assuming $P(A|B) = P(B|A)$.

5.2 Statistical Testing

Statistical inference involves:

- **Random Variables**.
- **Confidence Intervals**.
- **Hypothesis Testing**.