GEA1000 Cheat Sheet

Brians Tjipto

Research Targets

- Population: Entire group of interest.
- Sample: Proportion selected for study.
- Sampling frame: Source material for sampling.
- Census: Reaching the entire population.

Major Biases

- Selection bias refers to the researcher's biased selection of participants
- Non-response bias refers to participants' non-participation in the research

Probability Sampling Methods

- Simple random sampling: A sample of size n is chosen from the sampling frame such that every unit has an equal chance to be selected.
- Systematic sampling: The x^{th} unit is chosen from every n/k units where x, k are chosen integers and n is the size of the sampling frame.
- Stratified sampling: The population is divided into groups (strata) and SRS is applied to each strata to form the sample.
- Cluster sampling: The population is divided into clusters and a fixed number of clusters are chosen using SRS.

Non-Probability Sampling Methods

- Convenience sampling: Based on availability.
- Volunteer sampling: Participants volunteer.

Generalizability Criteria

1. Sampling frame $\geq$ population.
2. Probability sampling method implemented (selection bias $\downarrow$)
3. Large sample size (variability and random error $\downarrow$)
4. Minimize non-response rate.

Variable Types

- Categorical: Variables that take on mutually exclusive categories.
- Numerical: Variables with numerical values where arithmetic can be performed meaningfully.

Variable Sub-Types

- Ordinal: Categorical variables where there is some natural ordering.
- Nominal: Categorical variable where there is no intrinsic ordering.
- Discrete: Numerical variable with gaps in the set of possible numbers.
- Continuous: Numerical variable that can be all values in a given range.
- Random: Numerical variable with probabilities assigned to each value.

Properties of Mean ($\bar{x}$) and Median (r)

- Adding c to all data points: $\bar{x}$ to $\bar{x} + c$, $r + c$.
- Multiplying c to all data points: $\bar{x}$ to $c\bar{x}$, cr.

Properties of Standard Deviation (s_x) and IQR

- s_x, IQR are > 0 unless identical data.
- Adding c to all data point doesn't change s_x and IQR.
- Multiplying c all data points changes s_x to $|c|\, s_x$ and IQR to $|c|\, IQR$.

Study Designs

- Experimental study: The independent variable is intentionally manipulated to observe its effect on the dependent variable.
- Observational study: Individuals are observed and variables are measured without any manipulation.

Blinding

- Single blinding is achieved when subjects do not know what group they belong to.
- Double blinding is achieved when neither the subjects nor the assessors are aware of the assignment.

$$\text{Simple Variance, Var} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}$$

$$\text{Standard Deviation, } s_x = \sqrt{Var}$$

$$\text{coefficient of variation} = \frac{s_x}{\bar{x}}$$

Positive	Negative
rate(A|B) > rate(A|NB)	rate(A|B) < rate(A|NB)
rate(B|A) > rate(B|NA)	rate(B|A) < rate(B|NA)
rate(NA|NB) > rate(NA|B)	rate(NA|NB) < rate(NA|B)
rate(NB|NA) > rate(NB|A)	rate(NB|NA) < rate(NB|A)

Symmetry Rules

$$rate(A|B) > rate(A|NB) \iff rate(B|A) > rate(B|NA)$$
$$rate(A|B) < rate(A|NB) \iff rate(B|A) < rate(B|NA)$$
$$rate(A|B) = rate(A|NB) \iff rate(B|A) = rate(B|NA)$$

Basic Rule on Rates

- $rate(A\,|\,B) \leq rate(A) \leq rate(A\,|\,NB)$ or vice versa.
- The closer $rate(B)$ is to 100%, the closer $rate(A)$ is to $rate(A|B)$
- If $rate(B) = 50\%$, then $rate(A) = 0.5[rate(A\,|\,B) + rate(A\,|\,NB)]$
- If $rate(A\,|\,B) = rate(A\,|\,NB)$, then $rate(A) = rate(A\,|\,B) = rate(A\,|\,NB)$

Simpson's Paradox

A phenomenon where a trend appears in more than half of the groups but changes when the groups are combined.

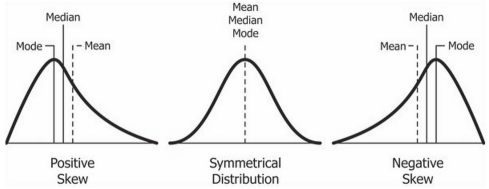
Confounders

- A third variable associated with both the independent and dependent variables.
- When a confounder is present, segregate the data by the confounding variable. This method is called slicing.

Outliers

- An outlier is an observation that falls well above or below the overall bulk of the data.
- A general rule is that outliers should not be removed unnecessarily.
- x is an outlier if $x > Q3 + 1.5 \cdot IQR$ or $x < Q1 - 1.5 \cdot IQR$.

Bell curve



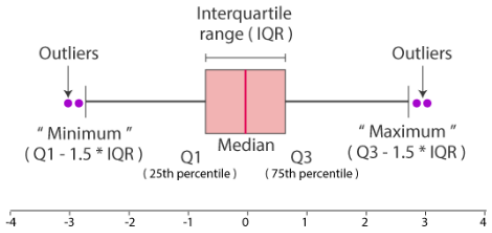
- Symmetic: Mean = Median = Mode
- Left-Skewed: Mean < Median < Mode
- Right-Skewed: Mean > Median > Mode

Analyzing Histograms

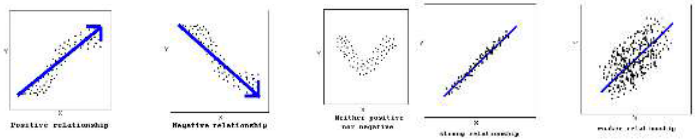
- Peaks show the mode.
- More spread indicates higher variability.

Analyzing Box Plots

- Center is the median.
- Whiskers are smallest and largest non-outlier values.
- Skewness: compare $max - median$ and $median - min$.



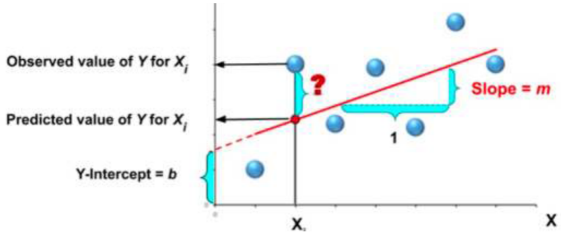
Correlation Coefficient



- Measure of linear association between two variables, $-1 \leq r \leq 1$.
- 0 to ± 0.3 = weak, ± 0.3 to ± 0.7 = moderate, ± 0.7 to ± 1 = strong.
- Removing outliers can increase, decrease, or cause no change to r.

Properties of r

- r is not affected by interchanging the x and y variables.
- r is not affected by adding or multiplying a constant to all values of a variable.
- *Association $\neq$ Causation: r value indicates a statistical relationship only.*



$$m = \frac{s_Y}{s_X} r$$

Method of Least Squares

- Fits a line through data points by minimizing the sum of squared errors or the distance between the observed value and predicted outcome is the error (e):
- Error Sum of Squares: $e_1^2 + e_2^2 + \ldots + e_n^2$

Conditional Probability

- $P(E|F) = \frac{P(E \cap F)}{P(F)}$

Probability Rules

- For an event E: $0 \leq P(E) \leq 1$.
- For a sample space S: $P(S) = 1$.
- Sum of square errors $e_1^2 + e_2^2 + \ldots + e_n^2$.
- For mutually exclusive events E and F: $P(E \cup F) = P(E) + P(F)$.

Probability in Independent Events

- For independent events A and B: $P(A) = P(A|B)$ and $P(A) \times P(B) = P(A \cap B)$.

Sensitivity and Specificity

- Sensitivity: $P(\text{Test Positive} \,|\, \text{Individual is infected})$.
- Specificity: $P(\text{Test Negative} \,|\, \text{Individual is not infected})$.

Law of Total Probability

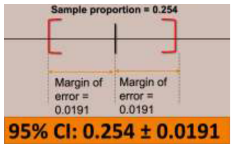
- Formally, the law of total probability states that if E, F, and G are events from the same sample space S such that:
  1. E and F are mutually exclusive
  2. $E \cup F = S$
- Then, $\mathrm{P(G) = P(G \,|\, E) \times P(E) + P(G \,|\, F) \times P(F)}$

Normal Distribution

- Continuous random variables denoted as $N(\mu, \sigma^2)$ where μ: mean, σ^2: variance.
- Density curve is bell-shaped and symmetric about the mean.
- For normal distribution: mean = median = mode.

Confidence Intervals

- A confidence interval is a range of values likely to contain a population parameter based on a certain degree of confidence.



We are 95% confident that the population parameter lies within the confidence interval

Another interpretation is that 95% of the researchers who repeat the experiment will have intervals that contain the population parameter

It is a common mistake to say that there is 95% chance that the population parameter lies within the confidence interval

- Larger sample size → smaller random error → narrower confidence intervals.
- Higher confidence level → wider confidence interval.

Null and Alternative Hypotheses

- **Null Hypothesis (H_0)**: Asserts no effect; observed variances occurred by random chance.
- **Alternative Hypothesis (H_a)**: The hypothesis to confirm, opposed to H_0.
- Hypothesis testing aims to reject H_0 in favor of H_a.

Significance Level α

- Specifies how convincing the evidence must be before rejecting H_0 ($0 \leq \alpha \leq 1$).
- Lower significance level α → greater evidence required.

p-Value

- The probability of obtaining a test result at least as extreme as the result observed, assuming the H_0 is true
- Alternatively, the probability of observing a test result that favours the H_α at least as much as what is observed, assuming the H_0 is true
- If $p \geq \alpha$, do not reject H_0.
- If $p < \alpha$, reject H_0.
- *Never accept H_0 or reject H_a.*

One-sample t-test	Chi-squared test
Mainly used to test difference between sample mean and a known or hypothesised mean.	Mainly used to test for association between two categorical variables.
Population distribution should be approximately normal if sample size is small.	Data required for the test is the count for the categories of a categorical variable.
Data used should be acquired via random sampling.	Data used should be acquired via random sampling.

Found: https://github.com/brianstm/NUS.git