GEA1000 Cheatsheet AY 22/23

Research Targets
Population: Entire group we wish to know something about
Sample: A proportion of the population selected in the study
Sampling frame: "Source Material" from which sample is drawn
Census: An attempt to reach out to the entire population of interest

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 xth 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: Subjects are chosen based on ease of availability
Volunteer sampling: Subjects volunteer themselves into a sample

Generalisability Criteria
1. Sampling frame ≥ population
2. Probability sampling method implemented (selection bias ↓)
3. Large sample size (variability and random error ↓)
4. Minimise 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)
1. Adding c to all data points changes $\bar{x}$ to $\bar{x} + c$ and r to r + c
2. Multiplying c to all data points changes $\bar{x}$ to $c\bar{x}$ and r to cr

Properties of Standard Deviation and IQR
1. sx and IQR are positive and 0 only when all data points are identical
2. Adding c to all data points does not change sx and IQR
3. Multiplying c to all data points changes sx to |c|sx 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

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

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

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

Establishing association	
Positive association between A and B: (any of the following)	Negative association between A and B: (any of the following)
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
1. rate(A | B) > rate(A | NB) ⟺ rate(B | A) > rate(B | NA)
2. rate(A | B) < rate(A | NB) ⟺ rate(B | A) < rate(B | NA)
3. rate(A | B) = rate(A | NB) ⟺ rate(B | A) = rate(B | NA)

Base Rule on Rates
The overall rate(A) will always lie between rate(A | B) and rate(A | NB).

Consequence 1:
The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B).

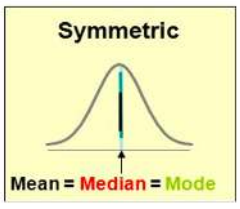
Consequence 2:
If rate(B) = 50%, then rate(A) = $\frac{1}{2}$[rate(A |B) + rate(A | NB)].

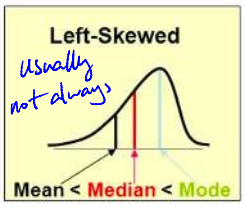
Consequence 3:
If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

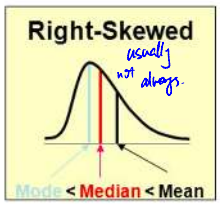
Simpson's Paradox
A phenomenon in which a trend appears in more than half of the groups of data but changes when the groups are combined

Confounders
- A third variable that is 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·IQR or x < Q1 - 1.5·IQR

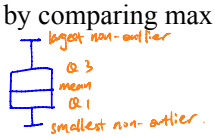




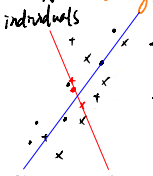


Analysing Histograms
- The peak/s show the mode
- A more spread out histogram shows higher variability

Analysing Box Plots
- The center is the median
- The smallest and largest non-outlier values are the whiskers
- Skewness can be observed by comparing max - median and median - min



An ecological correlation is computed based on aggregates rather than individuals



← the red colour + × are the means of subgroups + × (Aggregate pts)

* It can be seen that association for individuals and aggregates may not be in same direction (if in same direction, ecological tend to overstate the strength in individual)

Fallacy	Use	to conclude
Ecological	Ecological correlation	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation

Correlation Coefficient

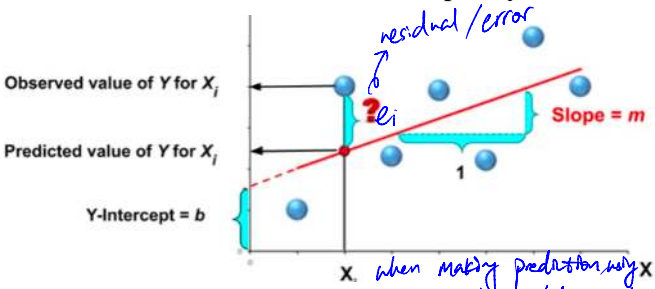
- Measure of the linear association between two variables
- $-1 \leq r \leq 1$
- 0 to ± 0.3 = weak, ± 0.3 to ± 0.7 = moderate, ± 0.7 to ± 1 = strong
- Removing outliers can increase, decrease, or cause no change to r

Properties of r

$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\bar{x}}{s_x}\cdot\frac{y-\bar{y}}{s_y}\right)$ ⇒ if x and y are evenly distributed, r = 0

- r is not affected by interchanging the x and y variables
- r is not affected by adding a number to all values of a variable
- r is not affected by multiplying a number to all values of a variable

Association ≠ Causation, r value is a purely statistical relationship



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

when making prediction using regression line, we can only conclude on the average, and within the range of independent variable.

The Method of Least Squares

- For a line fit through a set of data points, the distance between the observed value and predicted outcome is the error (e)
- The sum of squares of errors is given by $e1^2 + e2^2 + \ldots + en^2$
- Our target is to minimize the sum of squares of errors
- Say we are finding linear relation between A and B, the regression line formed when A is independent and B dependent is Not interchangable with A as dependent and B as independent.

Probability Rules

1. For an event E, $0 \leq P(E) \leq 1$
2. For a sample space S, $P(S) = 1$
3. The sum of squares of errors is given by $e1^2 + e2^2 + \ldots + en^2$
4. For mutually exclusive events E and F, $P(E \cup F) = P(E) + P(F)$

Conditional Probability

prosecutor fallacy: mistaken P(A|B) with P(B|A)

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

Conditionally independent: $P(A \cap B \mid C) = P(A|C) \times P(B|C)$

Probability in Independent Events

For independent events A and B:
$P(A) = P(A \mid B)$
$P(A) \times P(B) = P(A \cap B)$

Conjunction Fallacy
Believing P(2 things happening together) > P(one of the things happens)

Sensitivity and Specificity

Base-rate fallacy
decision making error in which base rate of some traits in population is ignored / not given proper weight.

True positive
Sensitivity = P(Test Positive | Individual is infected)
Specificity = P(Test Negative | Individual is not infected)
True negative

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; and

(2) $E \cup F = S$.

Then,

$$P(G) = P(G \mid E) \times P(E) + P(G \mid F) \times P(F).$$

Sample statistic = population parameter + bias + random error

Construct confidence interval for population proportion:

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

p^* = sample proportion
z^* = z-value for n-distribution
n = sample size

Confidence interval for population mean μ.

$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$

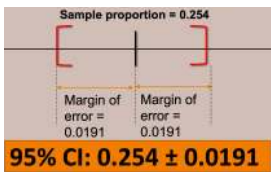
event: sub collection of sample space.
outcome: one result in the sample space

Normal Distribution

- A class of continuous random variables denoted as N(x,y) where x = mean and y = variance
- The density curve is usually bell-shaped
- The peak of the curve occurs at the mean and the curve is symmetrical about the mean
- For a 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

Properties of Confidence Intervals

- The larger the sample size, the smaller the random error and narrower the confidence interval
- The higher the confidence level, the wider the confidence interval

Null and Alternative Hypothesis

- The null hypothesis asserts the stand of no effect, meaning that the variances in the sample are not inherent in the population and occured by random chance when choosing sample
- The alternative hypothesis is what we wish to confirm and pit against the null hypothesis
- Through hypothesis testing, we wish to reject the null hypothesis in favour of the alternative hypothesis

Significance Level

- How convincing the evidence should be before rejecting the null hypothesis. ($0 \leq SL \leq 1$)
- The lower the significance level, the greater the evidence required

p-Value

- The probability of obtaining a test result at least as extreme as the result observed, assuming the null hypothesis is true
- Alternatively, the probability of observing a test result that favours the alternative hypothesis at least as much as what is observed, assuming the null hypothesis is true
- If p-value ≥ SL, do not reject null hypothesis
- If p-value < reject the null hypothesis

Never accept null hypothesis or reject alternate hypothesis

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.