GEA1000 Summary

AY23/24 Sem 1

Original from github.com/gerteck

1. Data Collection

Biasness

- **Selection Bias**: Associated with Researcher's Biased selection of units. Imperfect sampling frame (units excluded). Caused by non-probability sampling.
- **Non-Reponse Bias**: Associated with Participants' non participation, or non-disclosure of (sensitive) information.

Probability Sampling

Four types. Every unit has a known non-zero probability of being selected (need not be same). Element of chance to eliminate bias. Randomized mechanism.

- **SRS: Simple Random**: All units selected randomly without replacement, with equal chance. Subject to non-response.
- **Systematic Sampling**: Apply some selection interval k and random starting point from the first interval. List should be random.
- **Stratified Sampling**: (some units of all groups) Divided into strata based off similar nature, size may vary. SRS to each strata.
- **Cluster Sampling**: (whole cluster of only certain clusters): Divide into clusters. Fixed number of clusters chosen using SRS, which all units are used.

Sampling Plan	Advantages	Disadvantages
Simple Random Sample	Good Representation of the Population	Time-consuming; accessibility of information
Systematic Sample	Simpler selection process as opposed to Simple Random Sampling	Potentially under-representing the population
Stratified Random Sample	Good Representation of Sample by Stratum	Require Sampling Frame and criteria for classification of population into stratum
Cluster Random Sample	Less time-consuming and less costly	Require larger sample size in order to achieve low margin of error

Non Probability Sampling

Selection not done by randomisation but by human discretion. Broad Types include: (Non mutually exclusive) Quota, Convenience, Judgement, Volunteer Samplings.

- **Convenience Sampling**: Subjects most easily available to participate, e.g. Mall surveys
- **Volunteer Sampling**: Self-selected sample, biased and non representative.

Approach + Generalizability Criteria

- Choose Sampling frame. (Larger than or equal to target population, members of target pop must not be left out.
- Sample from Sampling frame (Decide if Probability Sampling in sample frame is feasible.)
- Remove unwanted Units.
- **Generalizability Criteria**: Good sampling frame that covers target population, probability based sampling (Need to be used to minimise selection bias), large sample size (Helps to reduce variability of data, reduce error amount in sample estimate, Minimal non-response rate.

Categorical Variables:

Either category or label values (mutually exclusive, variable cannot be placed in two different categories)
Ordinal Variables: Natural ordering, numbers represent order (e.g. Happiness)
Nominal Variables: No intrinsic (e.g. Eye colour)

Numerical Variables:

Discrete Variables: Possible values form a set of numbers with "gaps" e.g. Number of siblings
Continuous Variables: Can take on all possible values in an interval e.g. Time.

Summary Statistics for Numerical Variables

Central Tendency Measures: Mean, Median, Mode
Mean: Adding constant value to all changes mean by that value. Multiplying all changes mean similarly. **Dispersion Measures**: Standard deviation, Inter-quartile Range
Standard deviation: distance between each point and the mean. Measure of data distribution/spread.

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

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

Coefficent of Variation:

The concept of *coefficient of variation* is often used to quantify the degree of spread *relative* to the mean. The formula is

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

Median: Middle value of (ascending/descending ordered) data set. Overall median will always be between lowest and highest median amongst all subgroups.
Quartile 1: 25th percentile value,
Quartile 3: 75th Percentile value. IQR: Q3-Q1.
Mode: Value that appears the most often.

Remark 1.6.10 For a numerical variable, we can always use the mean and standard deviation as a pair of summary statistics to describe the central tendency as well as the dispersion and spread of the data. Similarly, the median and IQR can also be used. Which choice is more appropriate? There is no clear cut answer but very often, the choice depends on the distribution of the data. Generally speaking, the median and IQR is preferred if the distribution of the data is not symmetrical or when there are outliers.

Experimental Study

Controlled experiment, manipulate independent variable to observe effect on dependent variable. Goal is to provide evidence for cause-effect relationship. Make sure independent variable is the only factor, through random assignment. (Uses probability to allocate subjects into treatment and control groups) By law of probability, subjects will tend to be similar in all aspects.
Placebo: Inactive substance, likely caused by the psychology of believing.
Double Blinding: Patients and researchers both unaware of grouping.

Observational Study

Used when there are ethical issues. Observes individuals and measures variable of interest, without direct manipulation of variables. Does not provide convincing evidence of cause-effect relationship, and only Association.

2. Categorical Data

Joint Rate: Chance of an event occurring out of all the possible outcomes:
Conditional Rate: Based on a given condition (X) , in which rate of success/failure is found.

$$Rate(Success|X)$$

Association: Positive / Negative Association: If there is no association, we write that

$$rate(A|B) = rate(A|NB)$$

Four comparisons are mathematically equivalent:

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 Rule on Rates:

Symmetry Rule Part 1:

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

Symmetry Rule Part 2:

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

Symmetry Rule Part 3:

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

Basic Rule on Rates: The overall $rate(A)$ will always lie between $rate(A|B)$ and $rate(A|NB)$

- **Simpson's Paradox**: is a phenomenon in which a trend appears in more than half of the groups of data but disappears or reverses when the groups are combined. Here, "disappears" means the two variables in question (say A and B) are no longer associated. Rate of A given B is now equal to rate of A given not B.
- **Confounder**: A confounder is a third variable that is associated with both the independent and dependent variables whose relationship is being investigated. (Can be positive or negative association.)
  They can be addressed by the **splicing of data** according to the confounding variable or by **randomized assignment** (general solution across all confounders).
- **Observation of the Simpson's paradox** implies that there is definitely a (third) confounding variable present. However, existence of confounder does not necessarily lead to Simpson's paradox, nor does lack of observation imply lack of confounder.

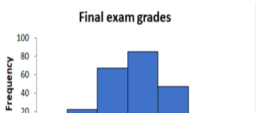
3. Numerical Data

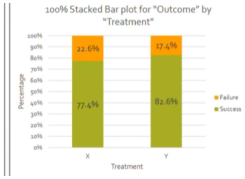
Univariate EDA

Exploratory Data Analysis of Univariate (one variable) numerical data: Consider **Distribution, Histograms, Boxplots**.
Describing Distributions (Overall Pattern + Deviations): Focus on shape, centre and spread of distribution, and outliers. Can be in the form of (mode) multimodal distribution (local maxima), unimodal, (Standard Variation, range of distribution) low variability vs. high variability, and outliers.
Median and Mode are robust statistics - Outliers have little to no effect on these values. (e.g. median salary)

Histograms

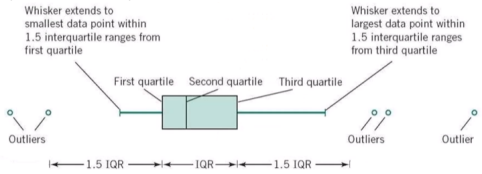
- Graphical representation that organises data points into ranges/bins. Useful for large data sets.
- **Histogram vs. Bar Graph**: A histogram shows the distribution of a numerical variable across a number line, but a bar graph makes comparisons across categories of a variable. Orderings of bar in histogram cannot be changed, unlike bar graph. No gaps between bars in a histogram.





Boxplots

- **Five Number Summary**: Minimum, Q1 (25th), Median (Q2) , Q3 (75th), Maximum.
- **Outliers:** Greater than $Q3 + 1.5 * IQR$ or smaller than $Q1 - 1.5 * IQR$.



- **Understanding boxplots: Shape, and Spread.**
  Shape: left-skewed vs right-skewed (variability of data on lower and upper half respectively).
  Centre: Described by Median. Cross represents mean. We can compare the relative positions of the median and mean from the boxplot.
  Spread: IQR gives us idea of the spread for the middle 50% of the data set, used to measure across different distributions.

Boxplots vs. Histograms:
Histogram: Better sense of shape of distribution of a variable. Boxplot: Better identifies and indicates outliers.
Bottom line: Used together to complement each other.

Bivariate EDA

Focus on relationship between two variables in a population.

- **Deterministic Relationship**: Value of one variable can be determined exactly from the other. (e.g. Conversion of units of measurement, $m \Leftrightarrow ft$, $^\circ C \Leftrightarrow ^\circ F$.)
- **Association (Non-Deterministic)** Statistical relation, given one variable value, we can describe average value of the other variable.
- Consider scatterplots (idea of pattern), correlation coefficients (check for linear relation) and regression analysis (fitting line or curve to data).

Scatter Plots

Direction, Form, Strength and Outliers.

- Direction: Positive / Negative relationship or neither (curved).
- Form: General shape, classify as linear or non-linear.
- Strength: How closely data follows form.

Fallacies

- **Atomistic Fallacy**: Using individual level correlation to conclude ecological (aggregate level) correlation
- **Ecological Fallacy**: Using ecological (aggregate level) correlation to conclude individual level correlation

Correlation Coefficient, r

Correlation coefficient between two numerical values, r, is a **measure of linear association** between them. Always ranges between -1 and 1.

- **Sign and Magnitude of r**: Tells us about the direction of the linear association. If $r > 0$, association is positive, when one increases the other tends to increase as well. $r < 0$, association is negative, increase in one variable leads to decrease of the other. If $r = 1$ or $r = -1$, there is perfect positive/negative association. When $r = 0$, there is no linear association. Magnitude of r tells us the strength of the linear association. Approx: (0 - 0.3 weak, 0.3 - 0.7 moderate, 0.7 - 1 strong)
- **Calculation of r**:

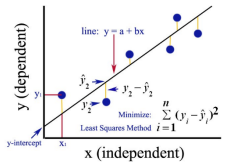
$$\mathrm{r} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- **Properties of r**: r is not affected by adding a number to all values of a variable, or by multiplying a positive number to all values of a variable.
- **Limitations of r**: Association is not causation. r does not give indication of non-linear association. Outliers can affect the correlation coefficient r significantly.
- **Standard Unit**: $\frac{x-\bar{x}}{s_x}$

Linear Regression

If we believe that two variables are linearly associated, we may model relationship by fitting a straight line to the observed data, known as linear regression.

- **The slope of the line** is the amount of change in Y when the value of X increases by 1.
- **Finding Regression Line: Method of least squares**: Fit the line to minimize the square of error terms. Hence, two regression lines are different and not interchangeable.



- **Slope vs. Correlation Coefficient** Slope of regression line and correlation coefficient related by:

$$m = \frac{s_y}{s_x} r$$

where s_y is the standard deviation for y and s_x is the standard deviation for x.

- Important to remember that correlation coefficient is not necessarily equal to gradient of the regression line.
- Extrapolation: **Prediction beyond the observed range is dangerous (Not advisable)**
- **Linear Regression on Non-Linear Models**: Model relationship indirectly (e.g. property of log) to form a linear relation.

4. Statistical Inference

Statistical Inference is the use of samples to draw inferences or conclusions about population in question.

Probability

Probability as a mathematical means to reason about uncertainty.

- **Sample Space**: Collection of all possible outcomes of a probability experiment.
- **Event**: Subcollection of the sample space is an event.
- **Rules of Probability**: Probability of an event E, $P(E)$, is between 0 and 1 inclusive. Probability of entire sample space $P(S)$ is 1.
- If E and F are mutually exclusive events, then the probability of E union F is equal to the sum of the probabilities of E and F. That is,

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

- **Uniform Probability and Rates**: Way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space. Relevant in random sampling.

Conditional Probability and Independence

Conditional Probability is written using the notatoin $P(E|F)$ and read as "probability of E given F".

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

- **Mutually Exclusive Events**: No overlap between E and F, meaning not simultaneously possible. Then, $P(E \cap F) = 0$. If an event F itself cannot occur, then by convention $P(E \cap F)$ is also equal 0.
- **Law of Total Probability**: If E, F, G are events from sample space S, and (i) E and F are mutually exclusive and (ii) $E \cup F = S$, $P(G) = P(G|E) \times P(E) + P(G|F) \times P(F)$
- Analogy between Probability and Sampling:

Random sampling	Corresponds to	Probability experiment
Sampling frame	Corresponds to	Sample space
A subgroup A of the sampling frame	Corresponds to	An event A of the sample space
The rate of A, rate(A)	Corresponds to	The probability of A, $P(A)$

- **Conditional Probabilities**: equivalent to conditional rate:

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

- **Independent Events:** For independent events A and B, the probability of A is the same as the probability of A given B.

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

If we express conditional probability $P(A|B)$ as: $\frac{P(A \cap B)}{P(B)}$

then A and B being dependent means that

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

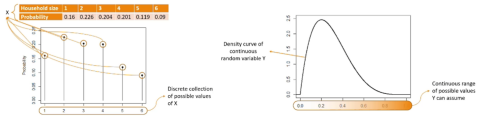
which is an equivalent definition for two independent events.

- **Independence as non-association**: A and B are independent event whenever A and B are not associated with each other.
- Independent Probability Experiments: E.g. Coin toss, where one instance is independent of the other.

Random Variables

A random variable is a numerical variable with **probabilities assigned to each of the possible numerical values** taken by the numerical variable. Conceived as mathematical way to model data distribution.

- May be **Discrete** or **Continuous** Random Variables. **Visualisation:** (respectively)

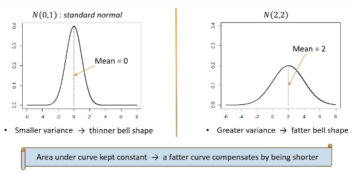


- For discrete rv, sum of probabilities assigned to each outcome must equals 1. For continuous rv, area under density curve is always equal to 1.

Normal Distributions

A class of continuous random variables. $N(x, y)$. (bell curve god)

- Normal Distributions only differ by means and variances. (mean x, variance y).
- **Common Properties**: Bell-shaped curve, Peak of curve occurs at the mean, Curve is symmetrical about the mean. (Mean = Mode = Median).



Confidence Intervals

Using a sample statistic to estimate the population parameter is subjected to inaccuracies (bias / random error).

- A **Confidence Interval** is a **range of values** that is likely to contain a population parameter based on a certain degree of confidence. This degree of confidence is known as the **confidence level** and is usually expressed as a percentage (%).
- To construct **confidence intervals** for population proportion:

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

where:
p^* = sample proportion
z^* = "z-value" from standard normal distribution (table)
n = sample size

- To construct **confidence intervals** for population mean μ:

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

where:
μ = sample mean
t^* = "t-value" from t-distribution (table)
s = sample standard deviation
n = sample size

- **Interpreting Confidence Interval:** Two parts: **Confidence Level** (e.g. 95%) and **Interval** (e.g. 0.254 ± 0.0191 [margin of error]) This means: we are 95% confident that the population parameter (e.g. mean) lies within the confidence interval. **Idea of confidence level**: 95 of 100 SRS of same size will contain an unknown population parameter. (** Not ~~95% chance~~, chances are in sampling procedure)
- **Properties of CI**: The larger the sample size, the smaller the random error, narrower CI. The higher the confidence level, the wider the CI. CI is way to quantify random error.

Hypothesis Testing

1. Null and Alternative Hypothesis.

- Null hypothesis usually asserts stand of no effect / difference. Alternative is what we wish to confirm and pit against null hypothesis. (Mutually exclusive) e.g.

Null Hypothesis $H_\circ$: $P(H) = 0.5$
Alt. Hypothesis: H_1: $P(H) > 0.5$

2. Collect data and determine test statistic.

- Testing usually involves some **random variable**, and its probability distribution. (e.g. coin, vaccine safety)

3. Set level of significance and compute p-value.

- **Significance level**: How convincing evidence must be to reject $H_\circ$
- The lower the S.L., the greater the evidence needed. Commonly used is 0.05 level, or 5% level of Sig, or 0.1 (10%), or 0.01 (1%).
- **p-value**: Probability of obtaining test result at least as extreme as result observed, assuming null hypothesis is true.
  Also the **probability of observing test result that favours alternative hypothesis** at least as much as observed in current sample, assuming null hypo is true.
  Small p-value: **Less likely** to observe a test result that is at least as extreme as observation in sample if H_0 is true.

4. Compare p-value and level of significance.

- Hence, we **reject null hypothesis** in favor of alternate if

p-value < significance

(logically it is very unlikely)

- However, if

p-value > significance

We **do not reject the null hypothesis**
~~(cannot accept, does not mean $H_\circ$ is true)~~ (we don't know if observation is due to chance, inconclusive)

- We only carry out hypothesis test with sample data. When given population data, all can be determined.

Common Hypothesis Tests: One-sample t-test and Chi-squared test:

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.

Fallacies

- **Prosecutor's Fallacy**: Thinking $P(A|B) = P(B|A)$
- **Conjunction Fallacy**: Thinking $P(A \cap B) > P(A)$
- **Base Rate Fallacy**: Using solely facts to conclude, and ignoring base rate of disease/trait of population

Sensitivity & Specificity

Using COVID tests as example,
Sensitivity: P(Positive|Has COVID)
Specificity: P(Negative|Does not have COVID)