GEA1000
AY24/25 SEM 1
github/SelwynAng

1 Getting Data

1.1 Exploratory Data Analysis

- **Population:** Entire group (of individuals or objects) that we wish to know something about.
- **Research Question:** Seeks to investigate some characteristic of a population.
  1. To make an estimate about the population
  2. To test a claim about the population
  3. To compare 2 subpopulations, to investigate a relationship between 2 variables in the population
- **Exploratory Data Analysis:** Systematic process where we explore a data set and its variables and come up with summary statistics as well as plots
  1. Generate research questions about the data
  2. Search for answers to research questions using data visualization tools → Can perform data modelling too.
  3. Ask ourselves to what extent does the data we have, answer the questions we are interested in
  4. Refine our existing questions or generate new questions about the data before going back to the data for further exploration

1.2 Sampling

- **Population of Interest:** Group in which we have interest in drawing conclusions on in a study
- **Population Parameter:** Numerical fact about a population (Eg. Average height *(population parameter)* of all P6 students in a particular primary school *(population)*.)
- **Census:** An attempt to reach out the entire population of interest (However, it has a high cost, takes a long time to complete, may not be able to achieve 100% response rate)
- **Sample:** Proportion of population selected in the study
- **Estimate:** Inference about the population parameter based on the information obtained from a sample
- **Sampling Frame:** List from which sample was obtained
- **Generalizability:** Ability to generalize findings from a sample to the population
  1. Sampling frame must be equal or greater than the population of interest
  2. Probability-based sampling to minimize selection bias
  3. Large sample size to reduce variability or random errors in sample
  4. Minimize non-response rate
- **Selection bias:** Associated with the researcher's biased selection of units into the sample, caused by imperfect sampling frame or non-probability sampling
- **Non-response bias:** Associated with participants' non-disclosure or non-participation in the research study → Exclusion of information from this group (Non-response bias may occur regardless of whether sampling method is probabilistic or non-probabilistic in nature)
- **Probability Sampling:** Sampling scheme such that the selection process is done via a known randomized mechanism (Every unit in the sampling frame has a known non-zero probability of being selected, but the probability of being selected does not have to be the same for all units)
  1. Simple Random Sampling (SRS): Every set of n units has an equal chance to be the sample actually selected
  2. Systematic Sampling: Selecting units from a list by applying a selection interval k & a random starting point from the 1st interval
     - Simpler sampling process than SRS as we do not need to know how many sampling units there are exactly
     - If listing is not random, but instead contains some inherent grouping/ordering of units → Possible that a sample produced by systematic sampling may not be representative of population (Sample could have selection bias)
  3. Stratified Sampling: Sampling frame is divided into groups called strata, each stratum is similar in characteristics, but size of each stratum does not need to be the same → Apply SRS to each stratum (Eg. Taking a SRS of the voters at each polling station (stratum) and then computing the weighted average of the overall vote count, based on size of each stratum)
  4. Cluster Sampling: Sampling frame is divided into clusters → Fixed number of clusters are then selected using SRS → All units from selected clusters are then included in the overall sample

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

- **Non-probability Sampling:** Selection of units is not done by randomization
  1. Convenience Sampling: Researcher chooses subjects to form a sample among those that are most easily available to participate in the study (can introduce selection bias & non-response bias)
  2. Volunteer Sampling: Subjects volunteer themselves into a sample (Sample contains subjects who have strong opinion on research question than rest of population)

1.3 Variables and Summary Statistics

- **Variable:** Attribute that can be measured or labelled
  - Independent Variable: Variable subjected to manipulation in a study
  - Dependent Variable: Variable hypothesized to change depending on how the independent variable is manipulated in the study (Does not mean the dependent variable must change)
- **Data Set:** Collection of individuals and variables pertaining to the individuals
- **Categorical Variable:** Variable that take on categories or label values, which are mutually exclusive
  - Ordinal Variable: Some natural ordering and numbers can be used to represent the ordering
  - Nominal Variable: No intrinsic ordering
- **Numerical Variable:** Variable that take on numerical values, can perform arithmetic operations
  - Discrete Numerical Variable: There are gaps in the set of possible numbers taken on by the variable
  - Continuous Numerical Variable: Can take on all possible numerical values in a given range or interval

1.4 Summary Statistics - Mean

- **Mean:** Average value of a numerical value x ($\overline{x}$)
- **Properties of Mean:**
  1. Adding a constant value c to all data points changes the mean by that constant value
  2. Multiplying a constant value of c to all data points will result in the mean being changed by same factor of c
  3. Mean does not tell us the distribution of the data
  4. Overall mean can be computed using the weighted average of 2 subgroup means (Eg. Group A has X subjects & I mean, Group B has Y subjects & J mean → Overall Mean $= \frac{X}{X+Y} * I + \frac{Y}{X+Y} * J$)
  5. Overall mean will always be between the smallest & largest means among all the subgroups

1.5 Summary Statistics - Variance and Standard Deviation

- **Sample Variance,** $Var = \frac{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}$
- **Standard Deviation,** $S_x = \sqrt{var}$
- **Properties of Standard Deviation:**
  1. s_x is always non-negative, when $s_x = 0$: all data points are identical
  2. Adding a constant c to all data points does not change the standard deviation
  3. Multiplying all the data points by a constant c results in the standard deviation being multiplied by $|c|$
  4. Even though mean is X and standard deviation is Y, it does not imply that the largest value is $X + Y$
- **Coefficient of Variation:** Quantifies the degree of spread relative to the mean ($\frac{s_x}{\overline{x}}$)

1.6 Summary Statistics - Median, Quartiles, IQR, Mode

- **Median:** Middle value of the variable after arranging the values of the data set in ascending or descending order (if there are 2 middle values, we take the average of the 2 middle values as the median)
- **Properties of Median:**
  1. When constant c is added to every data point in a data set, median increases by c too
  2. When constant c is multiplied to all data points, median is multiplied by c too
  3. Even when there more than 2 subgroups, overall median will always be between the lowest median and highest median among all the subgroups
- **Percentiles:** Median is 50th percentile, Q_1 is 25th percentile, Q_3 is 75th percentile
  - Divide data set into lower half & upper half → 1st quartile is median of lower half, 3rd quartile is median of upper half
  - When data set has odd number of data points, we do not include the median in both the lower and upper halves
- **IQR:** $Q_3 - Q_1$
- **Properties of IQR:**
  1. IQR is always non-negative
  2. Adding constant c to all data points results in no change in IQR
  3. Multiplying all data points by a constant c results in IQR being multiplied by $|c|$
- **Mode:** Numerical value that appears the most often in the data for numerical variable, category that has the highest occurrence in the data for categorical variable

1.7 Study Designs

- **Experimental Study:** Intentionally manipulate 1 variable to observe whether it has an effect on another variable (provide evidence for a cause and effect relationship)
  - To establish cause-and-effect relationship between 2 variables, we need to make sure that the independent variable is the only factor that impacts the dependent variable
  - Random assignment is impartial procedure that uses chance to allocate subjects into treatment & control groups
  - If the number of subjects is large, by law of probability, the subjects in the treatment & control groups will tend to be similar in all aspects
  - If we make it known to the control group that they are indeed the control group, this could lead to bias
  - Placebo: Inactive substance or other intervention that looks the same as, and is given the same way as, an active drug or treatment being tested (given to control group)
  - Single Blinding: Subjects do not know which group they belong to
  - Double Blinding: Subjects & assessors are blinded about the assignment
- **Observational Study:** Observes individuals and measures the variables of interest, usually without any direct/deliberate manipulation of the variables by the researchers (do not provide convincing evidence of a cause-and-effect relationship between 2 variables, normally used to circumvent ethical issues in experimental studies)

2 Categorical Data Analysis

2.1 Rates

- **Analyzing 2 categorical variables using a table:** By convention, dependent variable is placed on the columns of the table, while independent variable is placed on the rows
- **Types of Rates:**
  1. Marginal: Eg. $r(A)$ (relates to 1 categorical variable each time)
  2. Conditional: Eg. $r(A|B)$ (set a condition)
  3. Joint: Eg. $r(A\&B)$ (total is the baseline)

2.2 Association

- **Types of Association:**
  1. $rate(A|B) = rate(A|NB)$: Rate of A is not affected by presence or absence of B
  2. $rate(A|B) > rate(A|NB)$: Positive association between A and B (Presence of A when B is present is stronger compared to when B is absent)
  3. $rate(A|B) < rate(A|NB)$: Negative association between A and B (Presence of A when B is present is weaker compared to when B is absent)
- **More associations:**

Establishing association	
Positive association between A and B: (any of the following)	Negative association between A and B: (any of the following)
$rate(A \mid B) > rate(A \mid NB)$	$rate(A \mid B) < rate(A \mid NB)$
$rate(B \mid A) > rate(B \mid NA)$	$rate(B \mid A) < rate(B \mid NA)$
$rate(NA \mid NB) > rate(NA \mid B)$	$rate(NA \mid NB) < rate(NA \mid B)$
$rate(NB \mid NA) > rate(NB \mid A)$	$rate(NB \mid NA) < rate(NB \mid A)$

2.3 Rules on Rates

- **Symmetry Rules on Rates:**
  1. Symmetry Rule Part 1: $rate(A|B) > rate(A|NB) \leftrightarrow rate(B|A) > rate(B|NA)$
  2. Symmetry Rule Part 2: $rate(A|B) < rate(A|NB) \leftrightarrow rate(B|A) < rate(B|NA)$
  3. Symmetry Rule Part 3: $rate(A|B) = rate(A|NB) \leftrightarrow rate(B|A) = rate(B|NA)$
- **Basic rule on rates:**
  1. Basic rule on rates: Overall $rate(A)$ will always lie between $rate(A|B)$ and $rate(A|NB)$
  2. Consequence 1: Closer $rate(B)$ is to 100% → closer $rate(A)$ is to $rate(A|B)$
  3. Consequence 2: $rate(B)$ = 50% → $rate(A) = 0.5(rate(A|B) + rate(A|NB))$
  4. Consequence 3: $rate(A|B) = rate(A|NB) \rightarrow rate(A) = rate(A|B) = rate(A|NB)$

2.4 Simpson Paradox

- **Definition:** Phenomenon in which a trend appears in more than half of the groups of data but disappears (variables are no longer associated) or reverses when the groups are combined
- Note: In examples where there more than 2 subgroups, Simpson Paradox is observed as long as a majority of individual subgroup rates show the opposite trend to the overall rate
- **Relationship between SP & Confounding variables:**
  1. When Simpson Paradox is observed → There is definitely a confounding variable present (3rd variable that is associated with the 2 variables whose relationship we are investigating)
  2. Existence of confounder $\not\Rightarrow$ Observation of Simpson Paradox

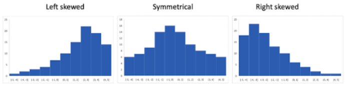
2.5 Confounders

- **Definition:** Confounder is a 3rd variable that is associated with both the independent and dependent variables whose relationship we are investigating
- Note: A confounding variable is associated with both independent & dependent variables, so removing 1 of the associations is enough to remove the confounding variable
- An alternative approach to address potential confounders is to rely on random assignment (However, random assignment is not possible all the time)
- Only slicing can be done to control for confounder in Observational studies. Random assignment is only suitable for Experimental studies.

3 Dealing with Numerical Data

3.1 Univariate Exploratory Data Analysis

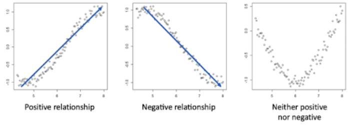
- **Distribution:** An orientation of data points, broken down by their observed number or frequency of occurrence
- **Histograms:** A histogram is a graphical representation that organizes data points into ranges or bins (Left-end point of interval is excluded, aka. $bin\_interval = (a, b])$
  1. Unimodal (1 distinct peak)



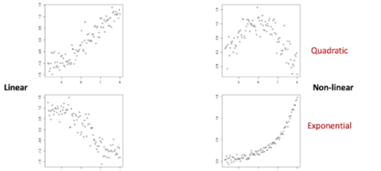
  2. Bimodal (2 distinct peaks)
  3. Multimodal (more than 1 distinct peaks)
- **Note about histograms:** Histogram shows distribution of numerical variable across a number line, while bar graphs show different categories of categorical variable
- **Symmetrical Distribution - Bell Curve:** Normal Distribution
- **Central Tendency - Mean, Median, Mode:**
  - Symmetrical distribution: Mean, median, mode very close to each other near peak of distribution
  - Left skewed distribution: Usually (but not always) have mean < median < mode
  - Right skewed distribution: Usually (but not always) have mode < median < mean
- **Spread - Standard deviation and range:**
  - Standard Deviation: Measure of variability around central Tendency
  - Range: Difference between largest and smallest data points in the distribution
- **Outlier:** Observation that falls well above or below overall bulk of the data
  - Outliers should not be removed unnecessarily as they do tell us something about the behavior of the variable and prompt us to investigate further why such extreme values can happen
  - Mean is most affected by removal of outlier, while median and mode either remains the same or only change slightly (Median and mode are robust statistics)
  - Data point is considered an outlier if it is greater than $Q_3 + 1.5 * IQR$ or less than $Q_1 - 1.5 * IQR$
- **Box-plots for Univariate EDA:**
  - To construct box-plots, we need the five number summary (Minimum, Q1, Median, Q3, Maximum)
  - Upper half of data have greater variability than lower half → Distribution is right-skewed
- **Histograms VS Box-plots**
  1. Histogram gives better sense of shape of distribution of variable, compared to box-plots
  2. To compare distributions of different data sets, putting different box-plots side by side is more illustrative than using histograms
  3. Box-plots do a better job than histograms at identifying and indicating outliers
  4. Number of data points we have in a data set is better shown in a histogram than in a box-plot (2 distributions with very different number of data points can have almost identical box-plots)

3.2 Bivariate Exploratory Data Analysis

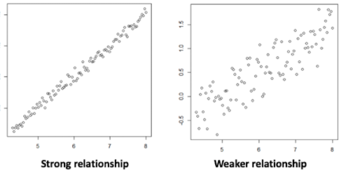
- **Scatter Plot:** Gives us an idea of the pattern formed by data between 2 variables in question
- **Correlation Coefficient:** Quantifies the level of linear association between 2 variables
- **Direction of Relationship:** Either positive (increase in 1 of the variables is associated with an increase in the other variable), negative (increase in 1 of the variables is associated with decrease in the other) or neither



- **Form of Relationship:** Describes general shape of scatter plot (linear or non-linear)

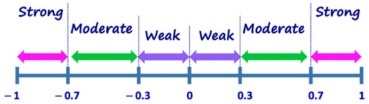


- **Strength of Relationship:** Indicates how closely the data follow the form of the relationship



3.3 Correlation Coefficient

- Measures the linear association between 2 variables, denoted by r, ranges $[-1, 1]$, summarizes the direction and strength of linear association between 2 variables
- $r > 0 \rightarrow$ Association is positive, $r < 0 \rightarrow$ Association is negative, $r = 1$ or $r = -1 \rightarrow$ Perfect association, $r = 0 \rightarrow$ No linear association
- Magnitude of r tells us the strength of linear association between 2 numerical variables



- No linear association between variables does not necessarily mean no association between variables (Eg. Quadratic relationship)
- When a straight line joining all data points is actually a straight horizontal or vertical line $\rightarrow r = 0$ and there is no association between 2 variables
- Computing Correlation Coefficient (assuming bivariate data (x, y)):
  1. Compute mean and standard deviation of x and y
  2. Convert each value of x and y into standard units ($\frac{x - \overline{x}}{s_x}$) and ($\frac{y - \overline{y}}{s_y}$)
  3. Compute the product xy in their standard units for each data point
  4. Sum the products xy over all data points, then divide the sum by $n - 1$, where n is the number of data points → Result is r
- 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 positive number to all values of a variable
- Association is not causation (we can only conclude a statistical relationship between x and y, not a causal relationship)
- r does not tell us anything about non-linear association
- Outliers can affect r significantly
- Fallacies
  - Ecological correlation is computed based on aggregates rather than on individuals
  - In general, when association for both individuals and aggregates are in the same direction, ecological correlation based on aggregates will typically overstate the strength of association in individuals

Fallacy	Using	To conclude
Ecological	Ecological correlation (aggregate level)	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation (aggregate level)

3.4 Linear Regression

- **Equation of straight line:** $y = mX + b$
- **Sum of Squares of Errors:** $e_1^2 + e_2^2 + \ldots + e_n^2$ where $e_i = y_{pred} - y_{actual}$
- **Notes:**
  1. Least squares regression line obtained from a set of observed data points will always pass through $(\overline{x}, \overline{y})$
  2. Regression line to predict y based on x cannot be used to predict x based on y
  3. Given $Y = mX + b \rightarrow m = \frac{s_Y}{s_X} r$
  4. We should not use the regression line to make prediction outside the independent variable's range in the data set
  5. Transform into linear relationship: $y = cb^t == ln(y) = ln(c) + tln(b)$

4 Statistical Inference

4.1 Probability

- **Sample Space:** Collection of all possible outcomes of a probability experiment
- **Event:** Sub-collection of the sample space
- **Rules of Probability:**
  1. $P(E)$ is a number between 0 and 1
  2. $P(S)$ is 1
  3. If E and F are mutually exclusive events $\rightarrow P(E \cup F) = P(E) + P(F)$
- **Uniform Probability:** Way of assigning probabilities to outcomes such that equal probability is assigned to every outcome in the finite sample space

4.2 Conditional Probability and Independence

- **Conditional Probability:** $P(E|F) = \frac{P(E \cap F)}{P(F)}$
- **Prosecutor Fallacy:** Mistake of confusing $P(A|B)$ as $P(B|A)$ (Unless $P(A) == P(B)$)
- **Independence:** A and B are independent $\rightarrow P(A) = P(A|B) == P(A) * P(B) = P(A \cap B)$ (A and B are independent events whenever A and B are not associated with each other)
- **Conditionally Independence:** A and B are conditionally independent given an event C if $P(A \cap B|C) = P(A|C) * P(B|C)$

4.2.1 Conjunction Fallacy, Base Rate Fallacy, Random Variables

- **Law of Total Probability:** If E, F, G are events from the same sample space S such that (1): E and F are mutually exclusive and (2): $E \cup F = S \rightarrow P(G) = P(G|E) * P(E) + P(G|F) * P(F)$
- **Conjunction Fallacy:** One would have committed Conjunction Fallacy if one believes that $P(A \cap B) > P(A)$ or $P(A \cap B) > P(B)$ (Chances of 2 things happening together is higher than the chance of 1 of those things happening alone) → What is actually true is $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$
- **Base Rate Fallacy:** A decision-making error in which information about the rate of occurrence of some trait in a population, called the base rate information, is ignored or not given appropriate weight
- **True Positive Rate:** $P(TestPositive|IndividualIsInfected)$ (Sensitivity of test)
- **True Negative Rate:** $P(TestNegative|IndividualIsNotInfected)$ (Specificity of test)
- **Random Variable:** Numerical variable with probabilities assigned to each of the possible numerical values taken by the numerical variable
  1. Discrete random variable
  2. Continuous random variable

4.3 Statistical Inference and Confidence Intervals

- **Statistical Inference:** Use of samples to draw inferences or conclusions about the population in question
- **Sample Statistic:** Sample Statistic = Population parameter + bias + random error (By adopting good sampling methods, we can reduce selection bias. Having high response rate will minimize non-response bias)
- **Confidence Interval:** Range of values that is likely to contain a population parameter based on a degree of confidence
  - Confidence Interval for population proportion: $p^* \pm z^* * \sqrt{\frac{p^*(1-p^*)}{n}}$, where p^* is sample proportion, z^* is z-value from standard normal distribution, n is sample size
  - Confidence Interval for population mean: $\overline{x} \pm t^* * \frac{s}{\sqrt{n}}$, where $\overline{x}$ is sample mean, t^* is ”t-value” from t-distribution, s is sample SD, n is sample size
  - Margin of error: Directly impacts the width of the confidence interval
  - ”95% confident” means that if many simple random samples of the same size are taken, and a confidence interval is constructed for each of them, then about 95% of the confidence intervals constructed would contain the population parameter
  - Either the population parameter is in the interval or is not (Wrong to say there is a 95% chance that it is in the interval, there is no probabilistic element)
  - Properties of Confidence Intervals
    1. Larger the sample size, the smaller the random error $\rightarrow$ Results in a narrower confidence interval
    2. Higher the confidence level at which the confidence interval is constructed $\rightarrow$ Wider the confidence interval

4.4 Hypothesis Testing

- **Definition:** Statistical inference method used to decide if the data from a random sample is sufficient to support a particular hypothesis about a population
- **5 steps of hypothesis test:**
  1. Identify the questions, state the null hypothesis and alternative hypothesis
  2. Set significance level of test
  3. Using sample, find the relevant sample statistic
  4. With sample statistic and hypothesis, calculate p-value
  5. Make conclusion of hypothesis test (dependent on p-value and significance level of test)
- **p-value:** Probability of obtaining a result as extreme or more extreme than our observation in the direction of the alternative hypothesis, assuming the null hypothesis is true
  - p-value $<$ significance level: Sufficient evidence to reject null hypothesis in favor of the alternative hypothesis
  - p-value $\geq$ significance level: Insufficient evidence to reject the null hypothesis. The hypothesis test is inconclusive. This does not mean that we accept the null hypothesis
- **Hypothesis test for population proportion/mean:** H_0: population parameter = null value, H_1: population parameter $<$ null value or H_1: population parameter $>$ null value
- **Hypothesis test for association:** H_0: No association, H_1: There is an association

5 End-of-chapter Questions Pointers:

5.1 Chapter 1

- Control group can simply be not receiving treatment, receiving a placebo, or receiving an existing treatment (with known success rate)
- Observational studies generally have control groups (subjects are self-assigned to different treatment & control groups)

5.2 Chapter 2

- Given $Rate(Male) = 0.4$, $Rate(Male|MilkTea) = 0.6$, $Rate(Male|FruitTea) = 0.3$, $Rate(Male)$ is closer to $Rate(Male|FruitTea)$ $\rightarrow$ Closer $Rate(FruitTea)$ is to 100%
- Since $Rate(Email) = 0.45$, $Rate(Email|Response) = 0.5$, $Rate(Email|Non-response) = 0.4$, then $Rate(Response) = 0.5$

5.3 Chapter 3

- Associations are not necessarily transitive
- Given that correlation coefficients for males and females is positive each, it is possible for correlation coefficient for both combined is negative
- Predicted value based on regression line is NOT equal to actual/exact value

5.4 Chapter 4

- By the interpretation of confidence intervals via repeated sampling, we conclude that about 95% of the samples will contain the population parameter within their respective confidence intervals.
- Any confidence interval constructed from any sample, regardless of the significance level, may or may not contain the population parameter
- Hypothesis tests regarding statements about the population should be conducted on probability-based samples only (not census)