GEA 1000 CheatSheet AY23/24 SEM 2
NUS CEG Lee Hao Zhe

Chapter 1 Getting Data

Types of **research questions:**

Making an **estimate** about the population: What is the average number of hours that students study each week?

Test a claim about the population Does the majority of students qualify for student loans?

Compare 2 sub-populations/Investigate a **relationship** between 2 variables in the population: In university X, do female students have a higher GPA score than male students?

Sampling methods **(Probability sampling)**

- The selection process is via a known/randomized mechanism in which every unit in the population has a non-zero and known probability of being selected. It eliminates biases associated with human selection.

Simple random sampling: Units are selected randomly from the sampling frame by a random number generator Sample results do not change haphazardly from sample to sample and variability is due to chance.

Advantages: Good representation

Disadvantages: Subject to non-response. Possible limited access of information as the selected individuals may be located in different geographical location.

Systematic sampling: A method of selecting units from a list through the application of selection interval K, so that every Kth unit on the list, following a random start (0 – k), is included in the sample.

Advantages: More straightforward and simpler selection process than the SRS. Do not need to know the exact population size at the planning stage.

Disadvantages: May not be good representation if sampling list is non-random.

Stratified sampling: Population is divided into strata where each stratum are similar in nature but size may vary across strata and then we randomly obtain a sample from each group.

Advantages: Able to get a representative sample from every stratum

Disadvantages: Quite complicated and time-consuming. Need information about sampling frame and stratum, which can be hard to define.

Cluster sampling: Population is first divided into clusters. Then we take a random selection of clusters from all clusters and include all units in the chosen clusters to comprise our sample.

Advantages: Less tedious, costly, and time-consuming as opposed to other sampling methods (e.g. stratified sampling)

Disadvantages: High variability due to dissimilar clusters or small numbers of clusters.

Sampling methods - **Non-Probability sampling**

Convenience sampling: Researchers use the subjects most easily available to participate in the research study. **Selection bias** occurs as certain members of the population may not be included in convenience sampling. **Non-response bias** occurs as individuals asked to participate in the study may opt out of the study due to inconvenience faced.

Volunteer sampling: Researcher actively seeks volunteers to participate in the study. **Selection bias** might occur as the researcher could pick volunteers more likely to respond in a certain desirable way. Volunteers sought by the researcher may be unwilling to participate in the study due to inconvenience hence causing **non-response bias.**

Criteria for generalizability: Good sampling frame that is equal to or larger than the population. Adopt a **probability-based** sampling method to minimise selection bias. Have a **large sample size** to reduce the variability or random errors in the sample. **Minimise the non-response rate**.

Categorical: categories or label values. The categories
or labels are mutually exclusive, meaning that an observation cannot be placed in two different categories or given two different labels at the same time.

Ordinal → some natural ordering and numbers can represent the ordering.

Nominal variable is a categorial variable where there is no intrinsic ordering.

Numerical: numerical values and we are able to perform arithmetic operations like adding and taking average.

Discrete → has gaps in the set of possible numbers taken on by the variable.

Continuous → can take on all possible numerical values in a given range or interval.

Central Tendencies

Mean/Median: Adding/Multiplying a constant value to all the data points adds/multiply the mean by that constant value.

Mode: The value of the variable that appears the most frequently,
can take on both numerical and categorical values. Not very useful when values are unique.

Standard Deviation: Measure of spread around the mean

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

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

Properties: standard deviation S_x is always non-negative. $\mathbf{S_x = 0}$ is when the data points are all identical. In this case, the variance is zero and so is the standard deviation. STDEV shares the same unit as the numerical variable x. **Adding** a constant c does **not change** the standard deviation.

Multiplying all the data points by a constant, standard deviation being multiplied by **|c|.** → **|c| * S_x**

Coefficient of variation = **STDev / Mean**

IQR: The **difference** between the third and first quartile. A small IQR values means that the middle 50% of the data values have a narrow spread and vice versa.

It is always **non-negative.** + to all points do not change IQR. Multiply will multiply IQR by

Experimental study: Intentionally manipulated one variable to provide evidence for cause and effect between 2 variables. **Random assignment** to ensure the treatment and control groups are similar in aspect when the number of subjects is large. Can have different sizes if the size of the groups are quite large.

Blinding: Prevent bias. A placebo can help make the blinding effect and this is done to prevent the **subject's own beliefs** about the treatment from affecting the results (Can have **psychological** effects too). Blinding of **assessors** can also take place to prevent bias due to certain **pre-conceived notions** about the treatment/control group. Can provide evidence of a **cause-and-effect** relationship.

Observational Study: Observe without direct manipulation. Subjects assign themselves into either the treatment or control group. Can provide evidence of association.

Chapter 2 Categorical Data Analysis / RATES

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

Symmetry rules:

rate(A | B) > rate(A | NB) ⇔ rate(B | A) > rate(B | NA).
rate(A | B) < rate(A | NB) ⇔ rate(B | A) < rate(B | NA).
rate(A | B) = rate(A | NB) ⇔ rate(B | A) = rate(B | NA).

Basic rules on rates:

Overall rate(A) will **always lie** between rate(A | B) and rate(A | NB).
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 x [rate(A |B) + rate(A | NB)].
If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

Simpsons Paradox: Trend appears in **more than half** of the groups of data but **disappears or reverses** when the groups are combined. There is definitely a **confounding variable present**. However, the existence of a confounder does not necessarily lead to us observing Simpson's Paradox.

Confounder: A **third** variable that is associated with both the independent and dependent variables whose relationship we are investigating. To overcome confounders, **slicing** of data (further break down of data to smaller subgroups). **Random assignment**, it causes the distribution between the two groups to be similar by the law of probability (not always possible).

Chapter 3 Dealing with Numerical Data

Histograms: Presents a graphical display of a distribution + quick/easy to grasp. Useful for large data sets. Avoid large (bad display of variability) or small (cannot tell shape) bin width.

Shape of a distribution: **peaks and skewness**. Unimodal/Multimodal = 1/multiple distinct peaks.

Left skewed is lean right, **symmetric**, **right skewed** is lean left. **Centre** of a distribution**:** In a symmetrical distribution, the mean, the median and the mode will be very close to each other at the peak of the distribution.

In a **left** skewed distribution, **mean < median < mode.**

In a **right** skewed distribution, **mean > median > mode. Spread** of a distribution: range (min – max) and standard deviation. The **higher the variability**, the **wider the range** in which the data is being spread across.

Outliers: May be reasons for strong skew. Mean can be pulled hard in the direction of skew that it becomes useless. Value is strictly greater than Q3 + 1.5 x IQR or strictly less than Q1 – 1.5 x IQR.

Boxplots: Skewed right if lower half (median – min) has less variability than the upper half (max – median) and vice versa. (Boxplot and histogram for **univariate** EDA)

Bivariate EDA: Deterministic relationship: One numerical value or quantitative variable can be used to determine another. **Statistical/non-deterministic relationship:** Only association between the 2 variables.

Correlation coefficient: It is a measure of the linear association, ranging from -1 to 1. R>0 positive association, R<0 negative association, R=0 no association (just mean no **LINEAR** relationship). ***Removing outliers** may not change R at all as the points may just lie on the straight line or the graph is symmetrical* 0-0.3=weak, 0.3-0.7=moderate,0.7-1=strong.

First, convert each data point into its standard unit

$SU_X = \frac{X - average(X)}{s_x}$ and $SU_Y = \frac{Y - average(Y)}{s_y}$ where s_X is the standard deviation of X and s_Y is the standard deviation of Y

Second, r value is just the average of the product of X and Y in standard units.

R **is immune** to interchange axis, When with -ve, for 1 variable, R change sign but if both variables then stay same (cancel out).

Ecological Correlation is grouping data points together and find the R at the aggregated group level rather than individual units. For example instead of taking everyone in the world as data points, you group them by country and find R from the groups.

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

Linear Regression: Gradient = Sy/Sx * R

If qn give you linear regression statistics and you require to find y = mx + b, if you manage to find b, you can sub y average and x average in to get b. **REMEMBER** the y value you **predict** is the "**predicted average** value of y". It will **not always** be true. You cannot predict values that are out of range (LOOK out for range in Qn). Valid if abs(R) > 0.7

Chapter 4 Statistical Inference

Sample space: All outcomes, **Event:** Subset of sample space.

Sample statistic = population parameter + random error + bias. Aim to **minimise** bias via **good sampling frame, probability sampling and 100% response rate.**

Mutually exclusive: $P(E \cup F) = P(E) + P(F)$ and $P(E \cap F) = 0$.

Independent events: P (A and B) = P (A) × P (B) and P (A or B) = P (A) + P (B) − P (A and B). Independent events **do not imply** mutual exclusive. Vice versa.

Uniform Probability: Equal probability for all outcomes.

Conditional Probability: If $P(E \mid F) = 0$, either $P(E \cap F) = 0$ or $P(F) = P(E) = 0$.

Sensitivity is true positive rate while specificity is true negative rate. P(Test + | Positive) & P(Test - | Negative)

$P(A \mid B) \neq P(B \mid A)$ in general, this is **prosecutor's fallacy**. Unless $P(A) == P(B)$.

Independence ⇔ No association ⇔ $\text{Rate}(A) = \text{Rate}(A \mid B)$ ⇔ $P(A) = P(A \mid B)$.

One would have committed **conjunction fallacy** if one believes that $P(A \cap B) > P(A)$ or $P(A \cap B) > P(B)$. It should be <= instead.

The base rate fallacy is 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.

Discrete random variables: Points in the plot (x-axis) separated by gaps hence the values are discrete. Probability is **adding up the probability of individual** discrete values.

Continuous random variable: Any continuous random variable Y can be visualised with a density curve on the standard x and y axis in which a curve can be viewed as continuous series of points. **Probability is area under curve.**

Confidence Interval: Provides a range of values that we are reasonably certain that the population parameter lies in. **95% C.I** means **95 out of 100 sample** will contain population parameter. Must eliminate bias to **use C.I accurately.**

100% response rate means **sample parameter = population parameter**

C.I for Population Mean = sample mean plus/minus "t-value" from t-distribution * STDev / sqrt n

C.I for Population Proportion = sample proportion (p) plus/minus "z-value" from normal-distribution * sqrt[(p* (1 – p)) / n]

C.I is a way to **quantify random** error. Increase in C.I = Wider interval. At the same C.I level, larger sample size = smaller interval.

Hypothesis Testing: The p-value is the probability of obtaining results at least as favourable or as extreme to the alternative hypothesis as the collected data/observed, computed based on the **assumption** that the null hypothesis is true.

Outcomes is what is observed in the actual data and **consider events** that are as **extreme or more extreme** than what's observed when calculating p-value.

Statistical inference method used to decide if the data from a random sample is sufficient to support a particular hypothesis about a population.

Step 1: Identify the question and state the null hypothesis and alternative hypothesis. How these hypotheses are stated depends on the context of the question and our aim.

Step 2: Next, set the significance level of our test. The significance level is often set at 5%, although others like 1% or 10% are also used frequently.

Step 3: Using our sample, we find the relevant sample statistic.

Step 4: With the sample statistic and the hypothesis, we can calculate the p-value

Step 5: We then make a conclusion of the hypothesis test. What the conclusion turns out to be depends on the p-value calculated and the significance level set for the test.

If p-value < significance level, **can reject** null, in favour of alternate.

If p-value > significance level, **inconclusive** 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 (less than 30).	Data required for the test is the count for the categories of a categorical variable.
Data should be acquired via random sampling.	Data should be acquired via random sampling.