GEA1000 Quant. Reasoning with Data

AY 24/25 Sem 2 — github/omgeta

1. Studying Data

Population is the entire group of interest.
Population parameter is a population's numerical fact.
Census is an attempted survey of full population.

Sample is a subset of a population from a sampling frame.
Sample statistic is a numeric fact of the sample.
Estimates infer pop. parameters from sample statistics.

Selection bias is caused by flawed sampling frame or non-probability sampling. Non-response bias is caused by systematic exclusion of subjects by unwillingness.

Probability sampling:

i. **Simple random**.
ii. **Systematic**: k^{th} subject of each size-r component.
iii. **Stratified**: Divide into strata sharing similar characteristic, then SRS within each stratum.
iv. **Cluster**: Divide into natural clusters, then SRS including all subjects within selected clusters.

Non-probability sampling:

i. **Convenience sampling**: subjects chosen by convenience; selection bias.
ii. **Volunteer sampling**: self-selected sample, usually with subjects off strong opinions; selection bias.

Study types:

i. **Experimental study**: observe dependent variable after direct manipulation of independent variable.
Random treatment and control groups are similar.
Placebo: fake treatment to control psych. effects.
Blinding: Hiding treatment assignment from subjects (single) or also researchers (double).
Shows cause-effect relationship.
ii. **Observational study**: observe variable of interest without manipulation of variables.
Shows association, not necessarily cause-effect.

Generalizability: frame size $\geq$ population, probability sampling, large sample size and minimal bias.

2. Categorical Data Analysis

Categorical variables are ordinal (naturally ordered w/o discrete gap) or nominal (no natural order).

Rates

Association Rules:

i. $rate\,(A \mid B) = rate\,(A \mid B')$ (none)
ii. $rate\,(A \mid B) > rate\,(A \mid B')$ and $rate\,(A' \mid B') > rate\,(A' \mid B)$ (+ve)
iii. $rate\,(A \mid B) < rate\,(A \mid B')$ and $rate\,(A' \mid B') < rate\,(A' \mid B)$ (−ve)

Symmetry Rules:

i. $rate\,(A \mid B) > rate\,(A \mid B') \iff rate\,(B \mid A) > rate\,(B \mid A')$
ii. $rate\,(A \mid B) < rate\,(A \mid B') \iff rate\,(B \mid A) < rate\,(B \mid A')$
iii. $rate\,(A \mid B) = rate\,(A \mid B') \iff rate\,(B \mid A) = rate\,(B \mid A')$

Basic Rule on Rates:

i. $rate\,(A)$ lies between $rate\,(A \mid B)$ and $rate\,(A \mid B')$
ii. As $rate\,(B) \to 100\%$, $rate\,(A) \to rate\,(A \mid B)$
iii. $rate\,(B) = 50\% \implies rate\,(A) = \frac{1}{2}[rate\,(A \mid B) + rate\,(A \mid B')]$
iv. $rate\,(A \mid B) = rate\,(A \mid B') \implies rate\,(A) = rate\,(A \mid B) = rate\,(A \mid B')$

Simpson's Paradox

Simpson's paradox is the observation that a trend appearing in majority of the groups of the data disappears/reverses when the groups are combined.

Confounders

Confounder is a third variable associated with both the independent and dependent variable being investigated. Randomised assignment can help to remove associations, removing the confounder in experimental studies.

3. Numerical Data Analysis

Numerical variables are discrete or continuous.

Summary Statistics

Mean, $\overline{x}$, is the average of variable x.
Mode is the most common element in variable x.
Q_1, Median, Q_3 are the ordered 1st, 2nd, 3rd quarter element of variable x.

Sample variance, Var, and standard deviation, s_x, of variable x are given by:

$$\begin{aligned} \text{Var} &= \frac{\sum(x_i - \overline{x})^2}{n-1} \\ s_x &= \sqrt{\text{Var}} \end{aligned}$$

Coefficient of variance, $\frac{s_x}{\overline{x}}$, measures spread relative to mean between different variables and has no units.

Median with $IQR = Q_3 - Q_1$ is preferred for asymmetrical distributions or when there are outliers.

Outliers satisfy one of the conditions:

i. $x > Q_3 + 1.5 \times IQR$ (> upper whisker)
ii. $x < Q_1 - 1.5 \times IQR$ (< lower whisker)

Univariate EDA

Histograms

Histograms show shape/distribution of data, are better at greater frequencies and show number of data points.
Distributions with n peaks are called n-modal.

Unimodal distribution shapes can be:

i. Symmetrical (mean = mode = median)
ii. Left-skewed (mean < mode < median)
iii. Right-skewed (mean > mode > median)

Bell distributions are symmetrical with spread:

i. 68% of data within 1 S.D.
ii. 95% of data within 2 S.D.

Boxplots

Boxplots side-by-side help compare distributions of different data sets, and are better to identify outliers. They consist of Q_1, median, Q_3 and whiskers.

Boxplot shapes can be:

i. Symmetrical (Q_1, Q_3 equidistant to median)
ii. Left-skewed (Q_1 closer to median)
iii. Right-skewed (Q_3 closer to median)

Boxplot spread for middle 50% is given by IQR.

Bivariate EDA

Deterministic relationships determine exactly a variable given the value of the other variable.
Association is a statistical relation describing average value of a variable given the value of the other variables

Correlation coefficient, r, is given by:

$$r = \frac{1}{n}\sum\left(\frac{x_i-\overline{x}}{SD_x}\cdot\frac{y_i-\overline{y}}{SD_y}\right) = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\cdot\sum(y_i-\overline{y})^2}}$$

*unaffected by swapping x, y or adding/scaling by +ve c

Direction, form and magnitude can be summarized by r:

i. $r > 0$ (+ve association)
ii. $r < 0$ (−ve association)
iii. $r = 0$ (No linear association)
iv. $0 < |r| < 0.3$ (Weak association)
v. $0.3 < |r| < 0.7$ (Moderate association)
vi. $0.7 < |r| < 1$ (Strong association)

Outliers can increase/decrease strength of correlation.

Linear Regression

Linear regression between variables believed to be linearly associated predicts the average value of the dependent variable given the independent variable.

Least squares regression for Y given X (not vice versa) is:

$$Y = mX + b, \quad m = \frac{s_Y}{s_X}r$$

4. Statistical Inference

Probability of event E in sample space S, $P(E)$, has:

i. $P(E) = \frac{|E|}{|S|}$, where $0 \le P(E) \le 1$
ii. $P(E') = 1 - P(E)$ (Complement)

Conditional probability of B given A is given by:

$$P(B \mid A) = \frac{P(A\cap B)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)}$$

Mutually exclusive events A, B satisfy:

i. $P(A\cap B) = 0$ (Intersection)
ii. $P(A\cup B) = P(A) + P(B)$ (Union)
iii. $A \cup B = S$ (Total probability)
$\implies P(C) = P(C \mid A)P(A) + P(C \mid B)P(B)$

Independent events A, B satisfy:

i. $P(A\cap B) = P(A)\cdot P(B)$ (Intersection)
ii. $P(A \mid B) = P(A)$ (Conditional)

Conditionally independent events A, B given C satisfy:

i. $P(A\cap B \mid C) = P(A \mid C)\cdot P(B \mid C)$ (Intersection)

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Fallacies

Ecological correlation fallacies:

i. **Ecological**: using ecological correlation (aggregate-level) to conclude individual correlation.
ii. **Atomistic**: using individual correlation to conclude ecological correlation (aggregate-level).

Probability fallacies:

i. **Prosecutor's**: mistaking $P(A \mid B)$ for $P(B \mid A)$
ii. **Conjunction**: thinking $P(A\cap B) > P(A)$ or $P(B)$
iii. **Base rate**: ignoring the base rate of an event when calculating likelihood (e.g. thinking +ve disease test means you have it, ignoring rarity of disease).

Relation between sample statistic and population parameter is given by:

Sample statistic = pop. parameter + bias + random error

Confidence Intervals

Confidence interval is a range of values from a sample such that, under repeated sampling, the proportion of intervals containing the true population parameter matches the desired confidence level.

Given a sample proportion p^* and sample size n, confidence interval for population proportion is given by:

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

where z^* is the z-value for desired confidence level.

Given a sample mean $\overline{x}$, sample SD s_x and sample size n, confidence interval for population mean is given by:

$$\overline{x} \pm t^* \times \frac{s_x}{\sqrt{n}}$$

where t^* is the t-value for desired confidence level.

Hypothesis Testing

Hypothesis test can be used for population proportion, mean and association, given a null hypothesis H_0, a alternative hypothesis H_1, and a significance value α.

For Chi-squared test of association, we take:

H_0: there is no association
H_1: there is an association.

p-value can be defined as:

i. Probability of obtaining a sample statistic as extreme or more extreme than the observed statistic, assuming H_0 is true.
ii. Smallest level of significance at which H_0 is rejected, assuming H_0 is true

where we reject H_0 in favour of H_1 when p-value $< \alpha$
or not reject H_0 (doesn't imply H_0 true) when p-value $\ge \alpha$