GEA1000 Quant. Reasoning with Data

AY 24/25 Sem 2 — github/omgeta

1. Studying Data

Population is a group of interest we have interest in. Population parameter is a numerical fact of the population. Census is a study of the complete 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 sampling**.
ii. **Systematic sampling**: k^{th} subject of each size r component chosen.
iii. **Stratified sampling**: sampling from each strata, a subdivision of similar characteristic.
iv. **Cluster sampling**: sampling all members of sampled clusters, a natural subdivision.

Non-probability sampling:

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

Study types:

i. **Experimental study**: observe dependent variable after direct manipulation of independent variable. Random treatment and control groups are similar. Shows cause-effect relationship.
ii. **Observational study**: observe variable of interest without manipulation of variables. Shows association, not necessarily cause-effect.

Generalizability depends on sampling frame size $\geq$ population, probability sampling, large sample size and little bias.

2. Categorical Data Analysis

Categorical variables are ordinal (naturally ordered) or nominal (no natural order).

Rates

When variables A, B are not associated:

i. $rate\,(A \mid B) = rate\,(A \mid B')$

When variables A, B are associated:

i. $rate\,(A \mid B) > rate\,(A \mid B')$ and $rate\,(A' \mid B') > rate\,(A' \mid B)$ (+ve)
ii. $rate\,(A \mid B) < rate\,(A \mid B')$ 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 an association, in order to remove the confounder.

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 1^{st}, 2^{st}, 3^{rd} quarter element of variable x.

Sample variance, Var, of variable x is given by:

$$\text{Var} = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

Standard derivation, s_x, of variable x is given by:

$$s_x = \sqrt{\text{Var}}$$

Coefficient of variance, $\frac{s_x}{x}$, measures variance 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. $x > Q_3 + 1.5 \times IQR$
ii. $x < Q_1 - 1.5 \times IQR$

Univariate EDA

Histograms

Histograms show data distribution, are better at greater frequencies and represent data points better. 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 minimum, Q_1, median, Q_3 and maximum.

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:

$$r = \frac{\text{Pop. covariance}}{\text{Pop. SD}_x \times \text{Pop. SD}_y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

*unaffected by swapping x, y or adding/scaling by constant

Direction, form and magnitude can be summarized by r:

i. $r > 0$ (+ve direction)
ii. $r < 0$ (−ve direction)
iii. $r = 0$ (Non-linear form)
iv. $0 < |r| < 0.3$ (Weak association)
v. $0.3 < |r| < 0.7$ (Moderate association)
vi. $0.7 < |r| < 1$ (Strong association)

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 line for predicting variable Y given X (and not vice versa) is given by:

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

4. Statistical Inference

Probability of event E in sample space S, $P(E)$, is given by:

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 have special results:

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 have special results:

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

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

Fallacies

Distribution fallacies:

i. **Ecological fallacy**: when we use aggregate level correlation to draw conclusions on individual data.
ii. **Atomistic fallacy**: when we use individual level correlation to draw conclusions on group data.

Probability fallacies:

i. **Conjunction fallacy**: probability of two events occurring together is always less than of either event occurring alone.
ii. **Base rate fallacy**: ignoring the base rate of an event when calculating its probability.

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 likely to contain a population parameter at a certain 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 tests can be used for population proportion, population mean, and association, given a null hypothesis H_0, alternative hypothesis H_1, and significance value a.
For hypothesis test on association, we take:

i. H_0 there is no association
ii. 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 $< a$ and not reject H_0 (not implying truth) when p-value $\ge a$