GEA1000

AY24/25 Sem 1

by ngmh

1. Getting Data

Exploratory Data Analysis

- Population: Entire group (of individuals or objects) that we wish to know something about.
- Research Question: Question that investigates characteristic of a population.
  - Make an estimate about the population.
  - Test a claim about the population.
  - Compare two sub-populations / investigate a relationship between two variables in the population.

Sampling

- Population of Interest: Group which we have interest in drawing conclusion on.
- Population Parameter: Numerical fact about a population.
- Census: Attempt to reach out to the entire population of interest.
- Sample: Proportion of population selected in the study.
- Estimate: Inference about the population parameter based on information obtained from sample.
- Sampling Frame: List from which the sample was obtained.
- Generalisability: Application of results from sample to general population. Should strive to:
  - Have a sampling frame large than or equal to the population.
  - Adopt probability based sampling to minimise selection bias.
  - Have a large sample size to reduce variability or random errors.
  - Minimise non-response rate.
- Biasness: Affects generalisability.
  - Selection Bias: Imperfect sampling frame, resulting in exclusion of units.
  - Non-Response Bias: Participant's unwillingness, results in exclusion of information.
- Probability Sampling: Select units using a randomised mechanism.
  - Simple Random Sampling: Every unit has equal chance to be chosen, without replacement. Good representation, but time consuming.
  - Systematic Sampling: Apply a fixed selection interval k, then choose every kth unit from a fixed offset. Simple selection process, but could under-represent population.
  - Stratified Sampling: Divide sampling frame into groups (strata). Each stratum should have similar characteristics. Apply simple random sampling to each stratum. Good representation, but might be hard to define stratum.
  - Cluster Sampling: Divide sampling frame into clusters. Choose entire clusters using simple random sampling. Less time-consuming, but could see high variability if clusters are not similar.
- Non-Probability Sampling: Selection without randomisation.
  - Convenience Sampling: Choose subjects most easily available. Introduces selection bias, might suffer from non-response bias as well.
  - Volunteer Sampling: Subjects volunteer, could be unrepresentative.

Variables

- Variable: Attribute that can be measured or labeled.
- Data Set: Collection of individuals and variables.
- Independent Variable: Subjected to manipulation.
- Dependent Variable: Hypothesised to change due to manipulation of independent variables.
- Types of Variables:
  - Categorical: Take on categories or label values. Ordinal variables have natural ordering, as opposed to Nominal variables.
  - Numerical: Take on numerical values, can meaningfully perform mathematical operations. Discrete variables have gaps in the set of possible values, as opposed to continuous.

Summary Statistics

- Allow us to perform numerical or quantitative comparisons between groups of data.
- Measure central tendencies (mean / median / mode), or spread (variance / standard deviation).
  - Mean: $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$. Average value of numerical variable. Can be calculated on subgroups and combined through weighted average.
  - Proportion: Consider fractions, rather than absolute values.
  - Sample Variance: $Var = \frac{(x_1-\bar{x})^2 + \ldots (x_n-\bar{x})^2}{n-1}$. $s_x = \sqrt{Var}$. Dividing by $n-1$ is to account for correction in a sample v/s population.
  - Coefficient of Variation: $cv = \frac{s_x}{\bar{x}}$. Used to quantify spread of data relative to mean, has no units.
  - Median: Middle number in sorted list of data points. Mean of middle 2 numbers if even size. Median of sample must lie between lowest and highest subgroups.
  - Q_1: 25th percentile, Q_3: 75th percentile. Split data points in lower and upper half, then find median. If odd number, exclude middle in halves.
  - Interquartile Range: $IQR = Q_3 - Q_1$.
  - Mode: Value that appears the most often.

Study Design

Experimental Study

- Manipulate independent variable to observe possible effects on dependent variable, to find evidence for a cause-and-effect relationship.
- Separate Sample into Treatment Group v/s Control Group
- Random Assignment: Use chance to allocate subjects into treatment and control groups. By law of probability, if the number of subjects is large, subjects in both groups will tend to be similar in all aspects.
- Assignment could lead to overstating or understanding of actual effects, due to bias.
- Placebo: Inactive substance or intervention that looks the same as actual treatment. Can cause plcaebo effect.
- Subject Blinding: Give control group placebo so that subjects do not know which group they are in.
- Assessor Blinding: Do not let assessors know which group the subjects they are grading are from.

Observational Study

- Observes individuals and measures variables of interest, without direct manipulation of variables, might not provided covincing evidence for cause-and-effect.
- Can be used when ethical issues are present.
- Describe as Exposure and Non-Exposure group instead, where subjects self-assign.

2. Categorical Data Analysis

Rates

- Proportion of Sample.
- Marginal Rate: Rate relating to one variable.
- Conditional Rate: A given $B = rate(A|B) = \frac{rate(A \cap B)}{rate(B)}$.
- Joint Rate: Rate of two variables, e.g. $rate(A \cap B)$.
- Symmetry Rule:
  - $rate(A|B) > rate(A|NB) \leftrightarrow rate(B|A) > rate(B|NA)$.
  - $rate(A|B) < rate(A|NB) \leftrightarrow rate(B|A) < rate(B|NA)$.
  - $rate(A|B) = rate(A|NB) \leftrightarrow rate(B|A) = rate(B|NA)$.
- Basic Rule:
  - $rate(A)$ lies between $rate(A|B)$ and $rate(A|NB)$.
  - The closer $rate(B)$ is to 1, the closer $rate(A)$ is to $rate(A|NB)$.
  - If $rate(B) = 50\%$, then $rate(A) = \frac{1}{2}[rate(A|B) + rate(A|NB)]$.
  - If $rate(A|B) = rate(A|NB)$ then $rate(A) = rate(A|B) = rate(A|NB)$.

Association

- Association $\neq$ Causation.
- Positive Association: $rate(A|B) > rate(A|NB)$, $rate(B|A) > rate(B|NA)$, $rate(NA|NB) > rate(NA|B)$, $rate(NB|NA) > rate(NB|A)$.
- Negative Association: Opposites of all above.
- There are other things we can compare to show association between A and B such as NA and NB.

Simpson's Paradox

- Trend in strictly more than half of subgroups disappears or reverses when groups are combined.
- Implies that there is a confounder present.

Confounders

- External variable associated with the two variables being investigated.
- Managed by segregating data by the confounding variable.
- Measure and collect data on additional variables that might be relevant.
- Can perform randomised assignment to remove association between treatment and confounder.

3. Dealing with Numerical Data

Univariate EDA

Describing Distributions

- Histogram: Sort data points into ranges or bins.
- Shape: Peaks and Skewness.
  - Unimodal distribution has 1 distinct peak, compared to multimodal with more than 1 peak.
  - Symmetrical: Left and right halves of the distribution are approximately mirror images.
  - Skewed: Peak shifted to side. Right Skewed: Tail on right, Left Skewed: Tail on left.
  - Central Tendency: Mean, Median, Mode. Left Skewed: $mean < median < mode$, Right Skewed: $mode < median < mean$ in general.
  - Spread: Standard Deviation and Range.
  - Outlier: Falls well above of below most data points. Should not be removed unnecessarily.

Box Plots

- Use 5 Number Summary: Minimum, Q_1, Median, Q_3, and Maximum.
- Outlier: $> Q_3 + 1.5 \times IQR$, or $< Q_1 - 1.5 \times IQR$.
- Draw box from Q_1 to Q_3.
- Draw vertical line at median.
- Draw lines (whiskers) from Q_1 and Q_3 to the most extreme data points which are not outliers.
- Mark outliers with dots or asterisks.

Bivariate EDA

- Statistical Relationship: Non-deterministic. Given one variable, can find average value of another variable, unlike deterministic where we can find exact value.
- Plot using scatter plot.
- Compare level of linear association using correlation coefficient.
- Fit best fit line or curve by performing regression analysis.

Describing Bivariate Relationships

- Direction: Positive relationship: Both increase at the same time, Negative relationship: Both change in opposite directions.
- Form: Shape of scatter plot, e.g. linear or non-linear (e.g. quadratic or exponential).
- Strength: How closely data follows form of relationship.

Correlation Coefficient

- Measure of linear association.
- Ranges from -1 to 1.
- Sign tells us about direction of association.
- Magnitude tells us about strength of association. Weak: 0 to 0.3, Moderate: 0.3 to 0.7, Strong: 0.7 to 1.
- Computation:
  - Find mean and SD of x and y.
  - Convert into standard units, using $\frac{x-\bar{x}}{s_x}$.
  - Compute xy for each for each data point.
  - Then $r = \frac{\sum xy}{n-1}$.
  - r is not affected by interchanging axes, or adding / multiplying all data points by a constant.

Fallacies

- Ecological: Using ecological / aggregate level correlation to conclude about individual level correlation.
- Atomistic: Using individual level correlation to conclude about ecological / aggregate level correlation.

Linear Regression

- Fit data points to linear model, then use this model to predict data points.
- Least Squares: Find line that minimises sum of squared error between y values.
- Regression line will always pass through averages for data set.
- Line has the form $Y = mX + b$, where $m = \frac{s_Y}{s_X} r$.
- Regression cannot be applied to data outside of the range.
- Exponential relationships: In on both sides to model it as a linear relationship.

4. Statistical Inference

Probability

- Mathematical means to reason about uncertainty.
- Outcomes: Set of possibilities of probability experiment.
- Must be repeatable and allowing for listing of all possible outcomes.
- Sample Space: Collection of all possible outcomes of a probability experiment.
- Event: Sub-collection of sample space.
- An outcome is an event, but an event is not necessarily an outcome.
- Probability of event is the total probability that outcome of experiment is an element of event.
- Rules of Probability
  - $0 \leq P(E) \leq 1$ where E is an event.
  - $P(S) = 1$ where S is the entire sample space.
  - If E and F are mutually exclusive, $P(E \cup F) = P(E) + P(F)$.
- Uniform Probability: Every outcome has an equal probability in the sample space.

Conditional Probability and Independence

- If E and F cannot happen together, then $P(E|F) = 0$. This is also true if $P(F)$ = 0.
- For two independent events, $P(A \cap B) = P(A) \cdot P(B)$. The probability of one happening does not affect the probability of the other.
- Two events are conditionally independent given a 3rd event if $P(A \cap B|C) = P(A|C) \cdot P(B|C)$.
- Law of Total Probability: If E, F, G are events from a sample space where E and F are mutually exclusive, and $E \cup F = S$, then $P(G) = P(G|E) \cdot P(E) + P(G|F) \cdot P(F)$.

Fallacies

- Prosecutor's Fallacy: Wrongly assuming that $P(A|B) = P(B|A)$.
- Conjunction Fallacy: Believing that $P(A \cap B) > P(A)$ or $P(A \cap B) > P(B)$. In reality, the opposite is true.
- Base Rate Fallacy: Information about rate of occurence (base rate information) is ignored or not given appropriate weight. Example: Given low chance of falsely detecting drunk drivers, always accurately detecting drunk drivers, and a low chance of actual drunk driving, assuming that the probability a person is drunk given a positive result is high.

Medical Testing

- True Positive Rate / Sensitivity: $P(TestPostive|Infected)$.
- True Negative Rate / Specificity: $P(TestNegative|NotInfected)$.
- Use contingency table to tabulate.

Random Variables

- Numerical variable with probabilities assigned to each possible value.
- Can be discrete or continuous.
- Continuous random variables can be plotted with density curves. Probability of a range of values is then the integral / area under the graph.

Statistical Inference

- Drawing inferences or conclusions about the population in question.
- Sample Statistic = Population Parameter + Bias + Random Error. Ideally, Sample Statistic should be as close to Population Parameter as possible.
- Fundamental Rule: Available data can be used to make inferences about a much larger group if the data can be considered to be representative with regards to the question of interest.
- Good sampling methods and practices can reduce bias to an insignificant level.

Confidence Interval

- Range of values that is likely to contain a population parameter based on a certain degree of confidence (confidence level).
- Population Proportion: Proportion of population fulfilling a certain criteria. To construct the confidence interval, use $p^* \pm z^* \times \sqrt{\frac{p^*(1-p^*)}{n}}$, where p^* is sample proportion, z^* is z-value from a standard normal distribution, and n is the sample size. The $\pm$ part is the margin of error.
- We are $x\%$ confident that the population parameter lies within the confidence interval. If many simple random samples of the same size are taken, and a confidence interval is constructed for each of them, $x\%$ of the confidence intervals constructed would contain the population parameter.
- Confidence intervals cannot give us an exact value. Uncertainty arises from sampling, not value of population parameter.
- Smaller sample size: Larger random error, wider confidence interval.
- When the confidence level is higher, the confidence interval is wider.
- Population Mean: $\bar{x} \pm t^* \times \frac{s}{\sqrt{n}}$, where $\bar{x}$ is sample mean, t^* is t-value from t-distribution, s is sample standard deviation, and n is sample size.

Hypothesis Testing

- Hypothesis Test: Statistical inference method to decide if data from a random sample is sufficient to support a particular hypothesis about a population.
- Steps:
  - Identify question, state null and alternative hypotheses.
  - Set significance level; a measurement of threshold for determining if deviation can be explained by chance.
  - Find the relevant sample statistic.
  - Calculate the p-value.
  - Make a conclusion based on the p-value and significance level.
- 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.
- When p-value < significance level, we have sufficient evidence to reject the null hypothesis in favour of the alternative hypothesis.
- When p-value $\geq$ significance level, we have insufficient evidence to reject the null hypothesis. The hypothesis is inconclusive. This does not mean that we accept the null hypothesis.

Hypothesis Test for Population Proportion

- Null Hypothesis: $H_0 : p = 0.5$.
- Alternative Hypothesis: $H_1 : p < 0.5$.

Hypothesis Test for Sample Mean

- Null Hypothesis: $H_0 : \mu = 69$.
- Alternative Hypothesis: $H_1 : \mu > 69$.

Hypothesis Test for Association

- Use a chi-square test.
- Null Hypothesis: There is no association between A and B at the population level.
- Alternative Hypothesis: There is an association between A and B at the population level.

pls give me A+ dash thanks