Probability Sampling Methods

- Sampling Process via a **known randomised mechanism.** The probability of selection **may not be the same** throughout all units of the sampling frame. Element of **chance** in selection process **eliminates biases** associated with selection.

-**Simple Random Sampling**: A sample of size n is chosen from the sampling frame such that every unit has an equal chance to be selected, through RNG. Advantage: Good Representation, Disadvantage: Non-response, time consuming, accessibility of info

-**Systematic Sampling**:The xth unit is chosen from every n/k units •where x,k are chosen integers and n is the size of the sampling frame. k selection interval. Advantage: Simple Disadvantage: Not good representation

-**Stratified Random Sampling**: The population is divided into groups (strata) and SRS is applied to each strata to form the sample. Ex: Sample count during GE. Advantage: Good representation Disadvantage: Need info about sampling frame and strata.

-**Cluster Sampling**:The population is divided into similar clusters and a fixed number of clusters are chosen using SRS. Advantage: less tedious, time-consuming, costly. Disadv: High variability if clusters are dissimilar, req larger sample size to achieve low margin of error.

Non-Probability Sampling

- **Convenience sampling**: Subjects are chosen based on proximity and availability (Mall surveys)
- **Volunteer sampling**: Subjects volunteer themselves into a sample (Online Polls)

Criteria for generalisability

0.Sampling frame ≥ population (Include people that used to be in population, duplicate, etc)
1.Probability sampling method implemented (selection bias ↓)
2.Large sample size (variability and random error ↓)
3.Minimise non-response

Types of Variables

Categorical: Variables that take on mutually exclusive categories (eg colours of cars)
Numerical: Variables with numerical values where arithmetic can be performed meaningful (mass)

Variable Sub-types

Ordinal: Categorical variables where there is some natural ordering (eg feeling on a scale of 1-5)
Nominal: Categorical variable where there is no intrinsic ordering (eg pet ownership in SG)
Discrete: Numerical variable with gaps in the set of possible numbers (eg no of members in fam, 3.75 doesnt exist)
Continuous: Numerical variable that can be all values in a given range Random: Numerical variable with probabilities assigned to each value (eg range of time from 0-5s, all possible values have meaningful intepretation)

Study Design

Experimental study: The independent variable is intentionally manipulated to observe its effect on the dependent variable (change x to see change in y)
Observational study: Individuals are observed and variables are measured without any manipulation

Blinding

Single blinding is achieved when subjects do not know what group yhey belong to
Double blinding is achieved when neither the subjects nor the assessors •are aware of the assignment

Research Targets

Population: Entire group we wish to know something about
Sample: A proportion of the population selected in the study
Sampling frame: “Source Material” from which sample is drawn
Census: An attempt to reach out to the entire population of interest

Basic Rule of Rates

rate(A | B) ≤ rate(A) ≤ rate(A | NB) or vice versa. This means: 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[rate(A |B) + rate(A | NB)] If rate(A | B) = rate(A | NB), rate(A) = rate(A | B) = rate(A | NB)

Probability, Sensitivity and Specificity

Probability in Independent Event
For independent events A and B: P(A) = P(A | B) P(A) × P(B) = P(A ∩ B)
Sensitivity and Specificity
Sensitivity = P(Test Positive | Individual is infected) Specificity = P(Test Negative | Individual is not infected)

Correlation Coefficient

Measure of the **linear association** between two variables

-1 ≤ r ≤ 1 •0 to ± 0.3 = weak, ± 0.3 to ± 0.7 = moderate, ± 0.7 to ± 1 = strong

Removing outliers can increase, decrease, or cause no change to r

r is not affected by interchanging the x and y variables r=Cov(X,Y)/SDx*SDy. Cov(X,-Y)=Cov(Y,X)

r is not affected by adding a number to all values of a variable. (eg y=2x, if +10 to allx,curve move right)

r is not affected by multiplying a number to all values of a variable

Outliers

An outlier is an observation that falls well above or below the overall bulk of the data . A general rule is that outliers should not be removed unnecessarily x is an outlier if x > Q3 + 1.5·IQR or x < Q1 - 1.5·IQR.

Left skewed curve --> Peak on the right. Mean < Median < Mode

Confounders

A third variable that is associated with both the independent and dependent variables. When a confounder is present, segregate the data by the confounding variable. This method is called slicing

Simpson's Paradox

A phenomenon in which a trend appears in more than half of the groups of data but changes when the groups are combined

Symmetric Association

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)

Establishing Association

Positive Assoc. between A and B (Negative flip sign)

Rate (A|B) > Rate (A|NB)

Rate (B|A) > Rate (B|NA)

Rate (NA|NB) > Rate (NA|B)

Rate (NB|NA) > Rate (NB|A)

Confidence Intervals

Confidence interval is a range of values likely to contain a population parameter based on a certain degree of confidence We are 95% confident that the population parameter lies within the confidence interval Another interpretation is that 95% of the researchers who repeat the experiment will have intervals that contain the population parameter It is a common mistake to say that there is 95% chance that the population parameter lies within the confidence interv-alProperties of Confidence Intervals The larger the sample size, the smaller the random error and •narrower the confidence interval The higher the confidence level, the wider the confidence interval. pop'n mean = xbar +- t* x (s/root n)

pop'n proportion= p +- z x root(p$(1-p)$/n)

Normal Distribution

- **The null hypothesis** asserts the stand of no effect, meaning that the variances in the sample are not inherent in the population and occured by random chance when choosing sample

-**The alternative hypothesis** is what we wish to confirm and pit against the null hypothesis Through hypothesis testing, we wish to reject the null hypothesis in favour of the alternative hypothesis

-If p-value ≥ SL, do not reject null hypothesis

t test and chi square

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.	Data required for the test is the count for the categories of a categorical variable.
Data used should be acquired via random sampling.	Data used should be acquired via random sampling.

Ecological and Atomistic Data

-**Ecological Fallacy** deduces the inferences on correlation about individuals based on aggregated data (country with high average income, assumes indiv is wealthy)

-**Atomistic Fallacy** generalise the correlation based on indiv towards the aggregate level correlation

(eg one person with high education makes more money, means higher education in country will lead to higher national income)