1 Study Design

Aims of study design: draw unbiased conclusions and provide estimates.

Bias is a lack of accuracy.

Study units, cases, subjects : units where data are obtained. One observation per unit.

Response/ Dependent variable : variables of interest that are dependent

Explanatory/ Independent variables : variables used to explain and predict the response variables

Confounding variables : other dependent variables not interested, related to both exp. and resp. vars. We wish to control or eliminate them.

1.1 Validity and biases

A biased study includes measures not accounted that influences the response variable.

To minimize biases

Comparison comparing with a control group, placebo, current best (current standard treat.), natural groups

Control restrictions limits, protocols being systematic, blinding (single and double), eliminate or holding constant confounding variables

Randomization Randomize representative subjects, randomize subject assigned to subject groups

Confounding factors (lurking variables) are related to treatment group, causation with outcome. It is the reason that correlation does not imply causation.

To reduce confounding

Distribution Random subject groups, fair distribution of characteristics

Randomization of treatment order (practice eff.)

Restriction on experimenters

Blocking small mini-experiments of common characteristics

Block when you can, randomize otherwise.

1.2 Precision

A precise study has close and more confident estimates with small error.

To maximize precision

Blocking Divide study units into blocks of similar characteristics. Randomize treatment within block. (Stratification for observational studies)

Replication increasing total subjects sampled, or repeating measures within study group. NOT repeatability or reliability. Linked with degree of freedom. More replicates = Higher df = More complicated models

Balance equally sized study groups (minimize SE with same total samp. size)

Matched pair, twins study are extreme levels of blocking.

1.3 Study types

Observation studies have data collected through observations. Subjects decide the group they are in. Cannot generate a causal link. Evidence by observation.

Observation studies problems

Selection bias Surveys may be biased, or self selection of subjects

Reporting bias groups are biased in responding/ reporting (diag. bias)

Question wording

Confounding factors not accounted

Designed experiments have the experimenter deliberately impose treatment to study groups. The experimenter decides the group subjects are in. Can prove causation. Evidence by design.

Designed experiments can better randomize, block, to reduce biases and confounding variables.

A completely randomize design has no matching, usually done with mechanical or computer randomizers.

2 Exploratory Data Analysis

Used to: Discover important data features, Improve understanding of underlying population, Transform data into information.

Steps

1. Display sample data (graphs)
2. Summarize distribution of sample data
3. Describe stats, graph information, and summary
4. Conjecture about the population

2.1 Variable types

Hierarchy of information (Least ⟶ Most info)

Categorical nominal Groups

Categorical ordinal Ordered groups

Numerical discrete Ordered, scale component

Numerical continuous Most informative

Questions we can ask

Categorical Category, mode, association

Numeric Mean, variance, min, max, median, outliers

Distribution features

Shape Symmetrical, skewed, (right tail is positive skewness) or unimodal, multimodal

Location mean, median

Spread variance, IQR

Unusual outliers, groupings

Relationships correlations (for numericals), associations (for categorical)

To describe data points, consider its shape, center, spread and outliers.

2.2 Graphical displays

N	C	Displays
1	0	dotplot, histogram, boxplot
0	1	Table (with percentages) or barchart
2	0	Scatterplot, boxplot of differences
1	1	Comparative dotplot, boxplot
0	2	Continecy tables (with %) or comparative bar charts
3	0	Surface Plot
2	1	Grouped Scatterplot
1	2	Interaction plot
0	3	Cross tabulation or comparative bar chart

2.3 Categorical

Figures using tables or barplots.

Simpson's paradox is a phenomenon where the trend exhibited within each group changes when combining all groups. Caused by imbalanced group sizes, where most frequently sampled values get over-proportionally weighted.

2.4 Numerical

Figures using dotplots for individual data, boxplot to summarize, comparative boxplot for comparisons, histograms for large dataset.

Figures

Histogram Divide range into bins, height of each bin is the number of data points within the range

Boxplot Five number summary of quartiles. Outliers are 1.5IQR above Q_1 or below Q_2 and represented by crosses

Location

Mean The statistical average

$$\mu = \bar{x} = \frac{1}{x}\sum x_i$$

Order statistic The observations sorted by value. $x_{(i)}$ is the ith ordered statistics. Can infer quartiles

Quartiles Lower quartile is $x_{(\frac{n+1}{4})}$, upper quartile is $x_{(\frac{3(n+1)}{4})}$. Median is $x_{(\frac{n+1}{2})}$. Use linear interpolation if the index is fractional.

Spread

Range Difference between largest and smallest value. Sensitive to outliers

IQR Middle 50% of data:

$$IQR = Q_3 - Q_1$$

Used when dataset is skewed

Std Dev Measures the consistency that observations are to the mean, the expected deviation of observation to mean. The sample standard deviation is

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \times \sum (x_i - \bar{x})^2}$$

with the $n-1$ accounting for degree of freedom for samples

Outliers Check if: a legit data value, entry mistake, or belonging in another population group. Quantitatively, outlier if observation is 1.5 IQR from Q_1 or Q_3.

Chebyshev's inequality states that for most distributions, at least 75% of data points are within 2sd. For normal distributions, the 68-95-99.7 rule. Squaring std dev to get variance.

Use medians and IQR when dataset is skewed.

2.5 Several numerical

Figures using scatterplots or dotplots on differences for data pairs.

Correlation is the strength (magnitude) and direction (sign) of a linear relationship between two random variables:

$$r = \frac{1}{n-1}\sum \frac{x-\bar{x}}{s_x}\frac{y-\bar{y}}{s_y}$$

Attributes of r

- From -1 to 1, with -1 and 1 indicating perfect correlation
- Positive indicating positive correlation
- 0 implies no linear association
- affected by outliers and unitless