1 Design of Studies

Definition 1.1 *Exposure/Treatment variable (x): Variable participants are exposed to.*

Definition 1.2 *Response/Outcome variable (y): Response to elicit from exposure/treatment*

Definition 1.3 *Control Group: Participants without treatment or exposure, serves as a comparison*

1.1 Controlled Experiment

Expose different groups of people different levels of exposure variable, compare results.
Randomised Control Assignment makes the groups more similar, less likely to be affected by confounders

1.2 Observational Study

Placed into groups based on exposure they already have, observe different results.
NO randomised assignment; cannot ask a non-smoker to smoke.

1.3 (Double-)Blinding

Participants or Experimenters knowing groups skew results
Placebo can be used for double-blinding

1.4 Confounders

Definition 1.4 *Confounder: Third factor associated with both exposure and response variables.*

Slicing: Sub-divide sample, study relation between exposure and response in each subgroup

1.5 Rate & Association

$rate(A|B) > rate(A|\sim B)$ or
$rate(B|A) > rate(B|\sim A)$
show **positive assocation between A and B**

$rate(B) = rate(B|A) * rate(A) + rate(B|\sim A) * rate(\sim A)$
Assocation != Causation

Definition 1.5 *Yule-Simpson's Paradox: Trends in several groups disappear when groups are combined.*

2 Association

Relationship between 2 variables; **Scatter Diagram**

2.1 Correlation Coefficient (r)

- Strength of relationship
- Not affected by addition/multiplication, except **multiplcation of negative numbers**
- Regression line of y on x: Can calculate estimated y when x = ?
- $|r| \to 1$: Strong linear association
- Removal of data affects r

2.1.1 Limitations

- Causality Unclear
- Outliers distort Correlation
- **Non-Linear Causation**

2.2 Ecological Correlation

Correlation based on group averages (Tends to overstate strength of association in individuals)

Ecological Fallacy	Applying findings for group averages on individuals
Atomistics Fallacy	Applying findings for individual averages on groups

2.3 Range of data

Definition 2.1 *Attenuation Effect: r understates strength of association if range is restricted from initial studied range.*

Outside of studied range: predicting y based on x is inaccurate since **association may change**

3 Sampling

Definition 3.1 *Population: Collection of Units*
Census: Measurement taken for entire population

Definition 3.2 *Sample: Portion of Population*
A good sample allows results to extend to the population

Definition 3.3 *Sampling Frame: List of sampling units to identifies all units in population. A good sampling frame has good coverage and is up-to-date.*

3.1 Sampling Methods

3.1.1 Probability Sampling Plans

Simple Random	Pure RNG
Systematic	Select unit every k interval. Is SRS if order of units is random
Stratified	Divide population into groups, SRS within every group
Cluster	Divide population into groups, choose random group, sample all in group
Multi-stage	Divide population into groups, SRS groups and within groups

Limitation: Practical Considerations

3.1.2 Non-P Sampling Plans

Volunteer	Participants volunteer to join May not require sampling frame
Convenience	Just pick :)
Judgement	Interviewer's discretion to choose participants
Quota	Stratified but non-P selection

Limitation: Selection Bias

3.1.3 Difficulties in Sampling

Imperfect Sampling Frame	Unwanted units included, increase cost of study to find and remove
Non-response	No one respond :(

3.2 Estimating Sample Parameter

$$Estimate = Parameter + RandomError + Bias$$

3.2.1 Bias

Selection Bias	Exclude kind of person from sample. Caused by: 1) Imperfect samp frame 2) Non-P sampling
Non-response Bias	Distorts response; non-respondents may be different than respondents

3.2.2 Random Bias

Definition 3.4 *X% Confidence Interval: Range of values that we are X% confident our population parameter lies in.*

Alternative Definition: *If ∞ samples are taken, 95% of CIs would contain the PP.*

CI is NOT: *"P(PP lies in this range)"*

Larger sample size, $\downarrow$ random error
$\Rightarrow$ More certain where PP is, $\downarrow$ width of CI

99% CI is wider than 95% CI, $\because$ accomodate for more samples to have the PP within its range

4 Risks and Odds

$Risk(A|B) = Rate(A|B)$
Risk Ratio= $\frac{Risk(A|Group1)}{Risk(A|Group2)}$
$Odds(A|B) = \frac{risk}{1-risk}$
Odds Ratio= $\frac{Odds(A|Group1)}{Odds(A|Group2)}$

4.1 Cohort Studies

- Samples from exposure groups, monitored for period of time
- Estimates Population RR and OR

4.2 Case-Control Studies

- Samples from outcome (disease) groups
- Then compare $rate(A|Disease)$ & $(A|NoDisease)$
- Ensures there is representation from all groups; Good for rarer diseases
- Estimates (only) OR, NOT RR

5 Probabilities

5.1 Independent Events

$P(A|B) = P(A|\sim B) = P(A)$

5.2 Mutually Exclusive

$P(A \cap B) = 0$
$P(A \cup B) = P(A) + P(B) + P(A \cap B) = P(A) + P(B)$

5.3 Disease/ Rare Events

Base Rate	$P(Disease)$
Sensitivity	$P(Positive\|Disease)$
Specificity	$P(Negative\|NoDisease)$

6 Hypothesis Testing

Definition 6.1 *Null Hypothesis H_0: Old Belief*
Alternative Hypothesis: New Belief

From an observation: Do we reject null hypothesis or accept new one?

Definition 6.2 *p-value: P(outcome equal to or more extreme than observed outcome) assuming H_0 is true*

Big p-value: more likely that observation occured by chance, null hypothesis likely remains true.

Small p-value: Unlikely for observed to occur by change, unlikely null hypothesis is true

- p-value $\leq$ Level of statistical significance $\rightarrow$ **Reject null-hypothesis**
- Increasing sample size $\rightarrow$ p-value $\downarrow$