Finite Sample vs. Superpopulation

Survey

Imaginary superpopulation - infinite
Finite population of interest - 1, 2, ..., N
Sample that we collect - 1, 2, ..., n

Sample mean \( \overline{Y} \)
Variance of \( \overline{Y} = \frac{s^2}{n} \left( \frac{N-n}{N} \right) \) where \( s^2 \) = variance of \( Y_1, ..., Y_N \)

Why superpopulation?
- Approximately for large N
- Allows us to use probability models

Model-Based Inference for Randomized Experiments

Key idea is that we can map \( Y_i(0), Y_i(1), W_i \) into \( Y_i^{\text{obs}} \) and \( Y_i^{\text{mis}} \) (e.g., \( Y_i^{\text{obs}} = (Y_i(1) \text{ if } W_i = 1 \) and \( Y_i(0) \text{ if } W_i = 0 ) \).

Then we can define the estimand in terms of missing and observed data

\[
\tau = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))
= \frac{1}{N} \sum_{i=1}^{N} (W_i Y_i^{\text{obs}} + (1 - W_i) Y_i^{\text{mis}}) - (W_i Y_i^{\text{mis}} + (1 - W_i) Y_i^{\text{obs}})
= \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) (Y_i^{\text{obs}} - Y_i^{\text{mis}})
\]

We then perform inference by building an imputation model for \( Y_i^{\text{mis}} \rightarrow Y_i^{\text{mis}} \) then

\[
\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) (Y_i^{\text{obs}} - Y_i^{\text{mis}})
\]

Imputation Strategies
- Mean imputation

\[
\hat{Y}_i^{\text{mis}} = \begin{cases} 
Y_i^{\text{obs}} & \text{if } W_i = 0 \\
\bar{Y}_c^{\text{obs}} & \text{if } W_i = 1
\end{cases}
\]

Reasonable estimate because it’s a randomized study
No uncertainty for estimate

- Discrete imputation (bootstrap)

\[
\hat{Y}_i^{\text{mis}} = \begin{cases} 
\text{random draw with replacement from treatment} & \text{if } W_i = 0 \\
\text{random draw with replacement from control} & \text{if } W_i = 1
\end{cases}
\]
Reasonable estimate (expected value of $\hat{\tau} = \hat{\tau}_{\text{mean imput}}$
Some uncertainty (via simulation, calculate all possible imputations)

- Full model
  \[ Y(1) \sim N(\mu_T, \sigma_T^2) \]
  \[ Y(0) \sim N(\mu_c, \sigma_c^2) \]
  Observed data to fit model

Bayesian Approach

Set up: Model $f(Y(1), Y(0)|\theta)$

- Prior distribution $P(\theta)$
- Assignment Mechanism $f(W|Y(1), Y(0), \theta)$ - we know this in randomized experiment

Computational Approach

- Distribution of missing data given observed data
  $f(Y_{mis}|Y_{obs}, W, \theta)$
- Posterior distribution of $\theta$
  $P(\theta|Y_{obs}, W)$
- Posterior predictive distribution of missing values
  $f(Y_{mis}|Y_{obs}, W)$ (average over $\theta$)
- Posterior predictive distribution of $\tau$

The above completes our discussion of randomized experiments. Next we consider observational studies where causal inference is more complicated.

Observational Studies

Focus on studies “regular assignment mechanism”

- Individualistic $P_i(X, Y(0), Y(1)) = P_{W_i = 1}(X, Y(0), Y(1)) = q(X_i, Y_i(0), Z_i(1))$
- Probabilistic $0 < q(X_i, Y_i(0), Y_i(1)) < 1$
- Unconfounded $q(X_i, Y_i(0), Y_i(1)) = q(X_i)$
  $W_i \perp Y_i(0), Y_i(1)|X_i$