Quiz 2 554 CS

Sample size computation in proportion tests

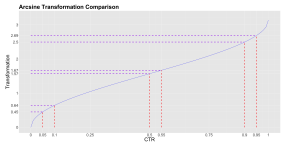
- Earlier was test with mean duration (continouous response) -> use t-test
  - Earlier test: "we want to see if changeing website design from"X" to "Y" will increase the average time spent on the website"
- We use **z-test** for **proportion tests**
  - proportion means value between 0 and 1
- E.g. see the click through rate (CTR) of a website between control "X" and treatment "Y"
  - CTR $\in [0, 1]$

Effect size h

- No need σ nor μ for proportion test
- Need to know the effect size h (difference in proportions)
  - Use `ES.h(data_control, data_treatment)` to calculate
- Effect size has an **arcsine transformation**:

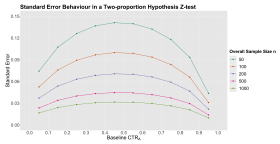
$$h = 2\arcsin(\sqrt{p_Y}) - 2\arcsin(\sqrt{p_X})$$

- This means:
  - Smaller transformed effect (y-axis) in the middle
  - Larger transformed effect (y-axis) at the ends



- This behaviour is related to the **statistic's standard error** of the two-proportion z-test
  - $\delta = p_Y - p_X$
  - $H_0 : \delta < some\_value$, $H_a : \delta \geq some\_value$

$$Z = \frac{\hat{\delta}_{\text{CTR}} - some\_value}{\sqrt{\frac{\hat{\text{CTR}}_A(1-\hat{\text{CTR}}_A)}{n/2} + \frac{\hat{\text{CTR}}_B(1-\hat{\text{CTR}}_B)}{n/2}}} = \frac{\hat{\delta}_{\text{CTR}} - some\_value}{\text{SE}\left(\hat{\delta}_{\text{CTR}}\right)}$$



More error in the middle, less error at the ends. Smaller sample size = more error

Using function `pwr.2p.test`

```
CTR_effect_size = ES.h(data_control, data_treatment)
pwr.2p.test(
  h = CTR_effect_size,
  sig.level = alpha,
  power = pow,
  alternative = "greater"
)
```

We can see that we need more sample size when:

- δ is smaller
- CTR_X is closer to 0.5, more error in the middle

Early stopping in A/B Testing

- **Peeking**: looking at the data before the experiment is over/ through the experiment
- Still compute overall sample size to get n_{max}
- If at some peek, **updated test statistic** is significant, we can stop the experiment

- **Aggresive Peeking**: look at the data after every new experimental unit
  - If test $P_Y > P_X$ (and in fact it is true), aggressive peeking will improve power of the test if:
    - $P_Y > 1.5 * P_X$
    - Otherwise, it gives a concerning power decrease
  - If test $P_Y = P_X$,
    - the proportion of replicates where $z_{test} > z_{1-\alpha}$ correspond to type I error rate
    - will lead to **inflating the type I error rate**

Observational Studies

- We **do not** have control of the variable of interest.
- **Without randomization, life is harder**. Strategy includes:
  1. Recording potential confounders
  2. Use confounders as part of the analysis
  3. Tempering the causal strength in light of inherent challenges in (i) and (ii)

Example: Pharmaco-epidemiology

- Response Y: binary indicator of disease state. (1: disease, 0: no disease)
- X: binary indicator of behavior (1: type A, 0: type B)
  - Type A: more agressive personality, competitive, etc
  - Type B: more laid back personality, relaxed, etc
- Confounders C_j (e.g. age, sex, BMI, cholesterol level, etc.)

- Since it is binary => binary logistic regression
  - Use log-odds $logit(p) = \log(\frac{p}{1-p})$
  - odds-ratio: $\text{OR} = \frac{\frac{n_{X=1,Y=1}/n}{n_{X=1,Y=0}/n}}{\frac{n_{X=0,Y=1}/n}{n_{X=0,Y=0}/n}} = \frac{\frac{n_{X=1,Y=1}}{n_{X=1,Y=0}}}{\frac{n_{X=0,Y=1}}{n_{X=0,Y=0}}} = \frac{n_{X=1,Y=1} \times n_{X=0,Y=0}}{n_{X=1,Y=0} \times n_{X=0,Y=1}}$
    - OR = 1: X does not effect Y
    - OR > 1: X increases the odds of Y
    - OR < 1: X decreases the odds of Y
  - $\text{SE} = \sqrt{\frac{1}{n_{X=1,Y=1}} + \frac{1}{n_{X=1,Y=0}} + \frac{1}{n_{X=0,Y=1}} + \frac{1}{n_{X=0,Y=0}}}$
- Can also just use binary logistic regression in R

Adding confounders

- Turn a continous confounder (e.g. age) into discrete categories and add to the model
  - `data |> mutate(age_bins = cut(age, breaks = c(min(age), quantile(age, (1:3) / 4), max(age)), include.lowest = TRUE)`
- By making **stratum-specific inference** with multiple confounders, we aim to infer causality between X and Y
  - However, there will be few observations in each strata (not enough data)
  - **Solution**: use binomial logistic regression (**Overall Model-Based Inference**) with interaction terms

```
glm(Y ~ X * C1 * C2, data = data, family = binomial) |>
  tidy(conf.int = 0.95)
```

recall: odds ratio is $exp(\beta)$ where β is the coefficient/ estimate

Assumptions for Causal Model-based Inference (binary logistic regression)

1. **Simple/ smooth structure in how the Y-specific log-OR varies across the strata**
   - Check using ANOVA comparing the simple model (all additive terms) and the complex model (with interaction terms of all confounders with each other)
   - `complex_model <- glm(Y ~ X + C1 * C2, data = data, family = binomial)`
   - `anova(simple_model, complex_model, test = "LRT")`
2. **Strength of (X, Y) association is constant across the strata** (i.e. no interaction between X and C)
   - complex model: all simple terms + double interactions of X and confounders
   - `complex_model_c1 <- glm(Y ~ X + C1 + C2 + X:C1, data = data, family = binomial)`
   - `complex_model_c2 <- glm(Y ~ X + C1 + C2 + X:C2, data = data, family = binomial)`
   - Compare all models with simple model using ANOVA
3. **No unmeasured confounders** (All confounders are included in the model)
   - Add new model with unmeasured confounder
   - `new_model <- glm(Y ~ X + C1 + C2 + CU, data = data, family = binomial)` where CU is the unmeasured confounder

Stratified Analysis vs model-based inference

	Stratified Analysis	Model-based Inference
Description	Separate analysis for each stratum	Single analysis with interaction terms
Pros	gives intuitive way to see relationships in homogenous groups	Allows simultaneous adjustment for all confounders
	Helps identify effect modification	Utilizes entire dataset -> leads to more power and efficiency
Cons	Can have small or no samples in some strata	Assumes correct specification of the regression model, difficult in complex relationships/ non-linear relationships
	Can lead to multiple comparison issues	Vulnerable to multicollinearity

Sampling Schemes in Observational Data

- Must Explore how to select our sample **before** executing the study
- Analysis must also account for the sampling scheme

Sampling Assesment via the Ground Truth

- **Ground Truth**: checking the results of the study in terms of its estimation accuracy against the **real-world**
- Ground truth is hard to get/ not available in many cases
  - Frequentist paradigm: we **do not** have access to true population parameters
  - Need to reply on sample to estimate the population parameters
  - It is better than purely simulated synthetic data because it includes realistic complexities

    - Has variations across the variables of interest and confounders

Simulation: Proxy for Ground Truth

- **Definition**: simulated dataset derived from a previously collected represntative sample for which some inference was made on the population of interest
- **Benefits**:
    - Can be used to evaluate bias, variablity, and power of the study
    - Can be used to evaluate the sampling scheme
- A better simulation technique than Monte Carlo
- **Core Idea**:
    - **Use a relevant sample dataset**: assumes previous sample data is representative of the population
    - **Fit a model** with `Y ~ X + C_j` (where C_j are a determined set of confounders)
    - **Simulate a proxy ground truth** by generating new data from the model
- Still use the 3 assumptions for the causal model-based inference
- Steps:
    1. Get a dataset and generate multiple rows of data (break continuous variables into quartile bins)
    2. Fit a model with the data
    3. Simulate a proxy ground truth by generating new data from the model

Sampling Schemes using Proxy Ground Truth

- Three Sampling Schemes:
    1. **Cross-Sectional Sampling**
    2. **Case-Control Sampling**
    3. **Cohort Sampling**
- These schemes will imply different **temporalities**

Cross-Sectional (CS) Sampling Scheme

- **Contemporaneous**: all data is collected at the same time
    - Grab a simple random sample of size n from the population
    - Similar to an instantaneous snapshot of all study variables
- **Ideal** in **early research stages** to get a sense of the data
    - Fast to run

Case-Control (CC) Sampling Scheme

- **Retrospective**: data is collected after the event of interest has occurred
    - Sample into cases `Y=1` and controls `Y=0` (Equal sample sizes for both)
    - Then ask the subjects "have you been exposed to `X` in the past?"
- **Ideal** where outcome `Y=1` is rare
    - Not have have a lot of cases of `Y=1` so recruit a lot of patients with `Y=1` for study
- It will **oversample** $Y = 1$ cases and **undersample** $Y = 0$ controls
    - Leads to a second statistical inquiry: "Is it a winning strategy to oversample cases?"
        - Do **modified Power Analysis** using Case:Control ratio
        - If ratio = 1, then control = case
        - If ratio > 1, then case > control
        - If ratio < 1, then case < control
    - According to SE behaviors, in populations when $Y = 1$ is rare, get **more precise** estimates by oversampling cases and undersampling controls

Cohort (CO) Sampling Scheme

- **Prospective**: data is collected over time
    - Sample into exposed `X=1` and unexposed `X=0` (Equal sample sizes for both)
    - Then follow the subjects over time to see if they develop the disease `Y=1`
- `Y` is assumed as the recorded outcome at the end of the study
- **Ideal** for when exposure `X=1` is rare
    - Not have have a lot of cases of `X=1` so recruit a lot of patients with `X=1` for study

Matched Case-Control Studies

- All our previous sampling schemes used a **binary logistic regression** model
- Recall we checked that CC is the best compared to the other 2 designs (CS and CO) in terms of power when $Y = 1$ is rare

Alternative Data Collection

- Using CC is acceptable because:
    - We assumed $Y|X, C_1, \ldots, C_p$
    - Create artificial population by "cloning subjects"
        - This is done using the estimated model from (previous reperesentative sample) + induced random noise
        - AKA Proxy Ground Truth
- CC is useful when $Y = 1$ is rare and sampling is costly

CC Matching

- Need to have artificial population to be representative of the true population
- Then will sample to get a sample size of n
- Record the confounders of interest as **strata**
- Sample $n/2$ cases then $n/2$ controls
    - Keep Case:Control ratio = 1
    - Match **exactly** on the confounder to the case counterpart
        - e.g. case has confounder $C_1 = 1$, then control must have $C_1 = 1$
- **Important**: Cannot fit Binary Logistic Regression model since we have matched pairs
    - Can get a sparse data problem
    - Use **McNemar's Test** instead
- CC-matched will show a smaller average bias compared to CC-unmatched
    - Power is the same once n increases

	Control $X = 0$	**Control** $X = 1$
Case $X = 0$	n_{00}	n_{01}
Case $X = 1$	n_{10}	n_{11}

- n_{00} and n_{11} are the **concordant pairs**
- n_{01} and n_{10} are the **discordant pairs**
- Estimator of the **odds ratio** is based on discordant pairs: $OR = \frac{n_{10}}{n_{01}}$
    - $OR = 1$ implies no association
    - $OR > 1$ implies positive association
    - $OR < 1$ implies negative association

McNemar's Test

- $H_0 : log(OR) = 0$, $H_a : log(OR) \neq 0$
- $log(OR) = logn_{10} - logn_{01}$
- It is approximately normally distributed with an SE of:
    - $SE = \sqrt{\frac{1}{n_{10}} + \frac{1}{n_{01}}}$
- Test statistic: $Z = \frac{log(OR)}{SE}$

Ordinal Regressors

- The numerical confounding strata we have been using are ordinal

Example

- Y is continuous, X is ordinal

1. Need to convert the variable X to ordinal

```
data$X_ord <- ordered(data$X, levels=c("low", "medium", "high"))
```

2. Fit a **one-way analysis of variance (ANOVA)** model

- R uses **polynomial contrasts** in ordered type factors
    - Elements of vectors sum to 0
    - Roughly, if have `k` levels, then `k-1` polynomials
    - E.g. `l=4`, we will have linear, quadratic, cubic contrasts
    - Use `contr.poly(l)` to get the contrasts when `l` levels
        - Gives design matrix for the contrasts
        - Cols sum to 0
        - Rows are the contrasts and are orthogonal

```
OLS <- lm(Y ~ X_ord, data=data)
OLS |> model.matrix()
```

3. Hypothesis test:
    - H_0: there is no GIVEN trend in the ordered data
    - H_1: there is a GIVEN trend in the ordered data
    - GIVEN will be replaced with linear, quadratic, cubic, etc

Successive Difference Constrasts

- Alternative to make inferential interpretations more straightforward
- Want to answer whether differences exist between ordered levels

```
options(contrasts = c("contr.treatment", "contr.sdif"))
OLS_sdif <- lm(Y ~ X_ord, data=data) |> tidy()
```

- Interpretation is very straightforward