Inference: Neyman's Repeated Sampling

STA 320
Design and Analysis of Causal Studies
Dr. Kari Lock Morgan and Dr. Fan Li
Department of Statistical Science
Duke University

Office Hours

- My Monday office hours will be 12-1pm for the next 3 weeks (2/10, 2/17, 2/24), not 3-4pm
- Wednesdays: 3-4pm
- Fridays: 1-3pm

R

- R code corresponding to all the problems from last class is available
- There are many different ways to code each problem – this is just an example
- For more information on any of it (for loops, subsetting data, handling NAs, etc.) see this R guide I wrote for intro stat
- You will have to do your own coding for homework and your project (you can talk, but do not share code)

HW 2

- Because the due date for HW 2 has gotten pushed back a week (now due Monday, 2/10), the next hw was dropped and instead I've added some problems to HW corresponding to today's class
- If you already downloaded it, make sure to look at the updated version

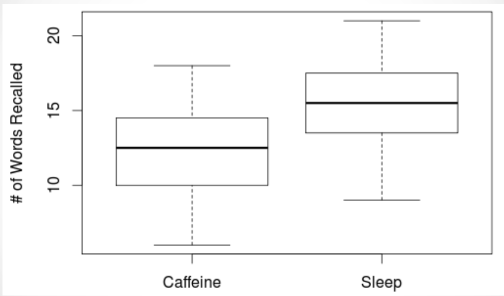
Causal Inference

Sleep or Caffeine?

- Is sleep or caffeine better for memory?
- 24 adults were given a list of words to memorize, then randomly divided into two groups
- During a break one group took a nap for an hour and a half, while the other group stayed awake and then took a caffeine pill after an hour
- Y: number of words recalled

Mednick S., Cai D., Kanady J., and Drummond S., "Comparing the benefits of caffeine, naps and placebo on verbal, motor and perceptual memory", Behavioural Brain Research, 2008; 193: 79-86.

Sleep or Caffeine



Jerzy Neyman

1894 – 1981

Fisher and Neyman

- At the same time Fisher was developing his framework for inference, Neyman was developing his own framework…
- Fisher: more focused on testing
  - is there a difference?
  - p-values
- Neyman: more focused on estimation
  - average treatment effect
  - unbiased estimators
  - confidence intervals

Sleep or Caffeine

- Fisher: Is there any difference between napping or staying awake and consuming caffeine, regarding number of words recalled?
- Neyman: On average, how many more words are recalled if a person naps rather than stays awake and consumes caffeine?

Neyman's Plan for Inference

1. Define the estimand
2. Look for an unbiased estimator of the estimand
3. Calculate the true sampling variance of the estimator
4. Look for an unbiased estimator of the true sampling variance of the estimator
5. Assume approximate normality to obtain p-value and confidence interval

Slide adapted from Cassandra Pattanayak, Harvard

11

Finite Sample vs Super Population

- Finite sample inference:
  - Only concerned with units in the sample
  - Only source of randomness is random assignment to treatment groups
  - (Fisher exact p-values)
- Super population inference:
  - Extend inferences to greater population
  - Two sources of randomness: random sampling, random assignment
  - "repeated sampling"
- We'll first explore finite sample inference…

Estimand

- Neyman was primarily interested in estimating the average treatment effect
- In the finite sample setting, this is defined as

$$\tau \equiv \overline{Y(1)} - \overline{Y(0)} \equiv \frac{\sum_{i=1}^{N} Y_i(1)}{N} - \frac{\sum_{i=1}^{N} Y_i(0)}{N}$$

Estimator

- A natural estimator is the difference in observed sample means:

$$\hat{\tau} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs}$$

$$\equiv \frac{\sum_{i=1}^{N} W_i Y_i(1)}{N_T} - \frac{\sum_{i=1}^{N} (1-W_i) Y_i(0)}{N_C}$$

Sleep vs Caffeine

- Estimand: the average word recall for all 24 people if they had napped – average word recall for all 24 people if they had caffeine
- Estimator:

$$\hat{\tau} \equiv \overline{Y}_S^{obs} - \overline{Y}_C^{obs} = 15.25 - 12.25 = 3$$

- (Sleep – Caffeine)

Unbiased

- An estimator is unbiased is the average of the estimator computed over all assignment vectors (W) will equal the estimand
- The estimator is unbiased if

$$E\left(\hat{\tau}\right) = \tau$$

Unbiased

For completely randomized experiments,

$$\hat{\tau} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs} \equiv \frac{\sum_{i=1}^{N} W_i Y_i(1)}{N_T} - \frac{\sum_{i=1}^{N} (1-W_i) Y_i(0)}{N_C}$$

is an unbiased estimator for

$$\tau \equiv \overline{Y(1)} - \overline{Y(0)} \equiv \frac{\sum_{i=1}^{N} Y_i(1)}{N} - \frac{\sum_{i=1}^{N} Y_i(0)}{N}$$

Neyman's Inference (Finite Sample)

1. Define the estimand: $\boxed{\tau \equiv \overline{Y(1)} - \overline{Y(0)}}$
2. unbiased estimator of the estimand:

$$\boxed{\hat{\tau} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs}}$$

3. Calculate the true sampling variance of the estimator

Slide adapted from Cassandra Pattanayak, Harvard

18

True Variance over W

$$\mathrm{var}(\overline{Y}_T^{obs} - \overline{Y}_C^{obs}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C} - \frac{S_{TC}^2}{N}$$

$$S_T^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \overline{Y}(1)\right)^2$$

$$S_C^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \overline{Y}(0)\right)^2$$

Sample variance of potential outcomes under treatment and control, *for all units.*

$$S_{TC}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left((Y_i(1) - Y_i(0)) - (\overline{Y}(1) - \overline{Y}(0))\right)^2$$

For the derivation of this, see Chapter 6.

Extra Term

$$S_{TC}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left((Y_i(1) - Y_i(0)) - (\overline{Y}(1) - \overline{Y}(0))\right)^2$$

- Always positive
- Equal to zero if the treatment effect is constant for all i
- Related to the correlation between Y(0) and Y(1), (perfectly correlated if constant treatment effect)

Neyman's Inference (Finite Sample)

1. Define the estimand: $\boxed{\tau \equiv \overline{Y(1)} - \overline{Y(0)}}$
2. unbiased estimator of the estimand:
$$\boxed{\hat{\tau} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs}}$$
3. true sampling variance of the estimator
$$\boxed{\mathrm{var}(\hat{\tau}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C} - \frac{S_{TC}^2}{N}}$$
4. Look for an unbiased estimator of the true sampling variance of the estimator

(IMPOSSIBLE!)

Slide adapted from Cassandra Pattanayak, Harvard

21

Estimator of Variance (of estimator)

$$\widehat{\mathrm{var}}(\hat{\tau}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}$$

(look familiar???)

$$s_T^2 = \frac{1}{N_T - 1}\sum_{i=1}^{N} W_i\left(Y_i(1) - \overline{Y}_T^{obs}\right)^2$$

$$s_C^2 = \frac{1}{N_C - 1}\sum_{i=1}^{N}(1 - W_i)\left(Y_i(0) - \overline{Y}_C^{obs}\right)^2$$

Sample variances of observed outcomes under treatment and control

Estimator of Variance

- This is the standard variance estimate used in the familiar t-test
- For finite samples, this is may be an *overestimate* of the true variance
- Resulting inferences may be too conservative (confidence intervals will be too wide, p-values too large)

Sleep vs Caffeine

$$\widehat{\mathrm{var}}(\hat{\tau}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}$$
$$= \frac{3.3^2}{12} + \frac{3.5^2}{12}$$
$$= 1.958$$

Neyman's Inference (Finite Sample)

1. Define the estimand: $\boxed{\tau \equiv \overline{Y}(1) - \overline{Y}(0)}$
2. unbiased estimator of the estimand:
$$\boxed{\hat{\tau} \equiv \overline{Y}_T^{obs} - \overline{Y}_C^{obs}}$$
3. true sampling variance of the estimator
$$\boxed{\mathrm{var}(\hat{\tau}) = \frac{S_T^2}{N_T} + \frac{S_C^2}{N_C} - \frac{S_{TC}^2}{N}}$$
4. unbiased estimator of the true sampling variance of the estimator
(IMPOSSIBLE!) Overestimate: $\boxed{\widehat{\mathrm{var}}(\hat{\tau}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}}$
5. Assume approximate normality to obtain p-value and confidence interval

Slide adapted from Cassandra Pattanayak, Harvard

25

Central Limit Theorem

- Neyman's inference relies on the central limit theorem: sample sizes must be large enough for the distribution of the estimator to be approximately normal
- Depends on sample size AND distribution of the outcome (need larger N if highly skewed, outliers, or rare binary events)

Confidence Intervals

$$\hat{\tau} \pm z^* \sqrt{\widehat{\mathrm{var}}(\hat{\tau})}$$

$$\overline{Y}_T^{obs} - \overline{Y}_C^{obs} \pm z^* \sqrt{\frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}}$$

- z* (or *t**) is the value leaving the desired percentage in between –z* and z* in the standard normal distribution
- (Confidence intervals due to Neyman!)

Confidence Intervals

For finite sample inference:

- Intervals may be too wide
- Inference may be too conservative
- A 95% interval will contain the estimand *at least* 95% of the time

Sleep vs Caffeine

$$\overline{Y}_T^{obs} - \overline{Y}_C^{obs} \pm z^* \sqrt{\frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}}$$

```
> qt(.975, df=11)
[1] 2.200985
```

$$15.25 - 12.25 \pm 2.2\sqrt{\frac{3.3^2}{12} + \frac{3.5^2}{12}}$$

95% CI: (-0.08, 6.08)

Confidence Intervals - Fisher

- You can also get confidence intervals from inverting the Fisher randomization test
- Rather than assuming no treatment effect, assume a constant treatment effect, c, and do a randomization test
- The 95% confidence interval is all values of c that would not be rejected at the 5% significance level

Hypothesis Testing

- Fisher: sharp null hypothesis of no treatment effect for any unit

$$H_0 : Y_i(1) = Y_i(0) \text{ for all } i$$

- Neyman: null hypothesis of no treatment effect on average

$$H_0 : \bar{Y}(1) = \bar{Y}(0)$$

Hypothesis Testing

- Fisher: compare any test statistic to empirical randomization distribution
- Neyman: compare t-statistic to normal or t distribution (relies on large n)

$$t = \frac{\hat{\tau}}{\widehat{\mathrm{var}}(\hat{\tau})} = \frac{\bar{Y}_T^{obs} - \bar{Y}_C^{obs}}{\sqrt{\frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}}}$$

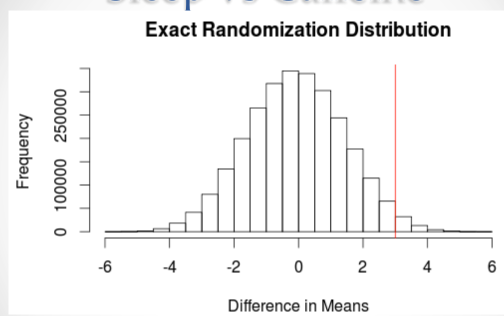
(Neyman's approach is the familiar t-test)

Sleep vs Caffeine

$$t = \frac{\bar{Y}_S^{obs} - \bar{Y}_C^{obs}}{\sqrt{\frac{s_S^2}{N_S} + \frac{s_C^2}{N_C}}} = \frac{15.25 - 12.25}{\sqrt{\frac{3.3^2}{12} + \frac{3.5^2}{12}}} = 2.14$$

```
> pt(2.14, df=11, lower.tail=FALSE)
[1] 0.02780265
```

Sleep vs Caffeine



Exact p-value = 0.0252

Super Population

- Suppose we also want to consider random sampling from the population (in addition to random assignment)
- How do things change?

Neyman Inference (Super Population)

1. Define the estimand: $\boxed{\mathbb{E}\left(Y_i(1) - Y_i(0)\right)}$
2. unbiased estimator of the estimand:

$$\boxed{\hat{\tau} \equiv \bar{Y}_T^{obs} - \bar{Y}_C^{obs}}$$

3. true sampling variance of the estimator

$$\boxed{\mathrm{var}(\hat{\tau}) = \frac{\sigma_T^2}{N_T} + \frac{\sigma_C^2}{N_C}}$$

4. unbiased estimator of the true sampling variance of the estimator

$$\boxed{\widehat{\mathrm{var}}(\hat{\tau}) = \frac{s_T^2}{N_T} + \frac{s_C^2}{N_C}}$$

5. Assume approximate normality to obtain p-value and confidence interval

Slide adapted from Cassandra Pattanayak, Harvard

36

Super Population

- Neyman's results (and therefore all the familiar t-based inference you are used to) are considering both random sampling from the population and random assignment

Fisher vs Neyman

Fisher	Neyman
Goal: testing	Goal: estimation
Considers only random assignment	Considers random assignment and random sampling
H_0: no treatment effect	H_0: average treatment effect = 0
Works for any test statistic	Difference in means
Exact distribution	Approximate, relies on large n
Works for any known assignment mechanism	Only derived for common designs

To Do

- Read Ch 6
- HW 2 due Monday 2/10