Causal Inference & Causal Learning

Randomized Experiment

Xiaolei Lin

School of Data Science
Fudan University

March 15, 2024

Goal of this week

- Randomized trials Designs (Bernoulli, Complete, Block, Paired)
- Design-Based vs Model-Based Inference
- Fisher's Exact P-Values for Completely Randomized Experiments
- Neyman's approach to Model-Based Inference

Notation

- Treatment vector, for subject $i = 1, ..., n$

$$\boldsymbol{A} = (A_1, ..., A_n)$$

- Potential outcome

$$\boldsymbol{Y}^1, \boldsymbol{Y}^0$$

- Covariates $\boldsymbol{C}$

Motivation

- For estimating causal effects, we want treatment groups that are similar regarding covariates
- A big theme of the course: create covariate balance across treatment groups
- Easiest way to accomplish this: randomized experiments

Randomized Experiment

- The assignment mechanism is random, known, and controlled by the researcher
- Because the treatments are randomly assigned, the joint distribution of all observed C and unobserved U pretreatment confounders is identical:

$$P(\boldsymbol{C}, \boldsymbol{U} \mid A = 1) = P(\boldsymbol{C}, \boldsymbol{U} \mid A = 0)$$

- Because the treatments are randomly assigned, treatment assignment is statistically independent of C and U

$$\{\boldsymbol{C}, \boldsymbol{U}\} \perp \boldsymbol{A} \rightarrow \{\boldsymbol{Y}^1, \boldsymbol{Y}^0\} \perp \boldsymbol{A}$$

- For classical randomized experiments, the assignment mechanism is individualistic, probabilistic, unconfounded and controlled by design

Propensity score

- The propensity score at $A = a$ is the average unit assignment probability for units with $A_i = a$:

$$e(a) = \frac{1}{n_a} \sum_{A_i = a} p_i(\boldsymbol{A}, \boldsymbol{Y}^1, \boldsymbol{Y}^0)$$

- Assuming individualistic and random assignment, the propensity score is just the probability of units with $A = a$ getting the active treatment.

- Assuming individualistic and unconfounded assignment given covariates, the propensity score is just the probability of units with $A = a$ getting the active treatment (or exposure) given covariates.

$$e(a) = \frac{1}{n_a} \sum_{A_i = a} p_i(\boldsymbol{A}, \boldsymbol{Y}^1, \boldsymbol{Y}^0 \mid \boldsymbol{C}) = \frac{1}{n_a} \sum_{A_i = a} p_i(\boldsymbol{A} \mid \boldsymbol{C})$$

Randomized Experiments

We'll cover four types of classical randomized experiments:

- Bernoulli randomized experiment
- Completely randomized experiment
- Stratified randomized experiment
- Paired randomized experiment
- Increasingly restrictive regarding possible assignment vectors

Bernoulli

- In a Bernoulli experiment, the treatment for each unit is determined by a coin flip
- Treatment assignments for units are independent (Usually, $e(a) = \frac{1}{2}$)
  - $e(a) = \frac{1}{2}$ maximizes precision
  - why might $e(a)$ differ from $\frac{1}{2}$?
- $e(a)$ can depend on covariates (rare)
- Any assignment vector, $\boldsymbol{A}$, is possible

Possible assignment vectors

Bernoulli 2^N

i = 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
i = 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
i = 3	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
i = 4	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Why might this not be a good design?

Completely Randomized

- In a completely randomized experiment, sample sizes for each treatment group are fixed in advance
- $N_1 =$ size of treatment group
- $N_0 =$ size of control group
- Often $N_1 = N_0$, but not always
- $e(a) = N_1/(N_1 + N_0)$
- Group sizes are the only restriction

Possible assignment vectors

Bernoulli 2^N

i = 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
i = 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
i = 3	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
i = 4	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Completely randomized experiment $\binom{N}{N_1}$

i = 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
i = 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
i = 3	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
i = 4	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Stratified

- In a stratified randomized experiment, units are partitioned into blocks or strata that are similar with respect to one or more covariates
- Units are completely randomized within each block/strata
- Ensures balance for important covariate(s)
- Also called blocking
- Advice: "block what you can, randomize what you cannot"

Possible assignment vectors

Bernoulli 2^N

i = 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
i = 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
i = 3	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
i = 4	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Completely randomized experiment $\binom{N}{N_1}$

i = 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
i = 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
i = 3	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
i = 4	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Stratified randomized experiment

female 1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1
female 2	0	0	0	1	0	0	1	0	0	1	1	1	0	1	1	1
male 1	0	0	1	0	0	1	0	0	1	1	0	1	1	0	1	1
male 2	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	1

Paired

- In a paired randomized experiment, units are first matched into pairs of similar units
- Within each pair, randomize which unit is treated
- Special case of blocking
- Goal: improve covariate balance and increase precision
- Also called matched pairs experiments

Paired

How to pair?

- same person at different points in time
- pairs with closest values of covariates
- twin studies

Comparison of designs

- Enforcing positivity
- Ensure that there are enough treated and control units under each assignment
- Ensure balance

Case study

Does drinking a sports drink (e.g. Maidong) make you run faster, as opposed to just drinking water?

How would you design an experiment with each of the following designs?

- Bernoulli?
- Completely randomized?
- Stratified?
- Paired?

Randomization inference vs model-based inference

- Randomization as the "reason basis for inference" (Fisher)
- Randomness comes from the physical act of randomization, which then can be used to make statistical inference
- Also called design-based inference
- Advantage: design justifies analysis, analysis is model free
- Contrast this with model-based inference, which assumes a distribution for potential outcomes
- Advantage of model-based inference: flexibility, allows for more complexity in the assignment mechanisms

Fisher vs Neyman

Sir Ronald A. Fisher (1890-1962)

Jerzy Neyman (1894-1981)

Case study: Diet Cola and Calcium

- Does drinking diet cola leach calcium from the body?
- 16 healthy women aged 18-40 were randomly assigned to drink 24 ounces of either diet cola or water
- urine was collected after 3 hours, and calcium excreted was measured (in mg)
- Is there a significant difference?

Case study: Diet Cola and Calcium

Drink	Calcium Excreted
Diet cola	50
Diet cola	62
Diet cola	48
Diet cola	55
Diet cola	58
Diet cola	61
Diet cola	58
Diet cola	56
Water	48
Water	46
Water	54
Water	45
Water	53
Water	46
Water	53
Water	48

Brief review of hypothesis testing

1. Choose a null hypothesis:
   - $H_0 : \tau = 0$
   - No average causal effect
   - Claim we would like to reject
2. Choose a test statistic
   - $T_i = \bar{Y}^1_{obs} - \bar{Y}^0_{obs}$
3. Determine the distribution of the test statistic under the null.
   - Statistical thought experiment: if we knew the truth, what data should we expect?
4. Calculate the probability of the test statistics under the null.
   - What is this called? p-value

Sharp null hypothesis of no effect

$$H_0 : \tau_i = Y_i^0 - Y_i^1 = 0$$

- Different than no average treatment effect, which does not imply the sharp null.
- Take a simple example with two units: $\tau_1 = 1$ and $\tau_2 = -1$, here, $\tau = 0$ but the sharp null is violated.
- This null hypothesis formally links the observed data to all potential outcomes.

Comparison to the average null

- Sharp null allows us to say that $Y_i^0 = Y_i^1 \rightarrow$ impute all potential outcomes.
- Average null only allows us to say that $E(Y_i^0) = E(Y_i^1) \rightarrow$ tells us nothing about the individual causal effects.
- We are looking for evidence against the sharp null (i.e. we are testing whether we can reject this null hypothesis, that there is low probability that sample arose from a population in which $Y_i^0 = Y_i^1$)
- Stochastic version of "proof by contradiction."

Test statistic

A test statistic, T, can be any function of:

- the observed outcomes, Y
- treatment assignment vector, A
- the covariates, C

The test statistic must be a scalar (one number), for example:

- Difference in means
- Regression coeffcients
- Rank statistics

Want a test statistic with high statistical power: has large values when the null is false, unlikely large values when then null is true.

Case study: Diet Cola and Calcium

Difference in sample means between treatment group (diet cola drinkers) and control group (water drinkers)

$$T_{obs} = \frac{\sum_{i=1}^{n} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{n} (1 - A_i) Y_i^0}{N_0} = 6.875$$

Is a difference of 6.875 mg more extreme than we would have observed just by random chance, if there were no difference between diet cola and water regarding calcium excretion?

p-value

- T: A random variable (i.e. the value that the test statistics takes for each exposure assignment when the sharp null is true and the assignment mechanism is completely at random)
- T_{obs} : the observed test statistic computed in the actual experiment
- The p-value is the probability that T is as extreme as T_{obs} , if the null is true
- GOAL: Compare T_{obs} to the distribution of T under the null hypothesis, to see how extreme T_{obs} is
- SO: Need distribution of T under the null

Sir Ronald A. Fisher (1890-1962)

Randomness

- In Fisher's framework, the only randomness is the treatment assignment A
- The potential outcomes are considered fixed, only observed outcome is random
- The distribution of T arises from the different possibilities for A
- For a completely randomized experiment, there are $\binom{N}{N_1}$ possibilities for A

Sharp Null Hypothesis

- Fisher adopts the sharp null hypothesis if there is no treatment effect:

$$H_0 : \tau_i = Y_i^1 - Y_i^0 = 0$$

- Advantage of Fisher's sharp null: under the null, all potential outcomes are "known"!

- EXAMPLE: There is NO EFFECT of drinking diet cola (as compared to water) regarding calcium excretion. So, for each person in the study, their amount of calcium excreted would be the same, whether they drank diet cola or water.

Randomization Distribution

- The randomization distribution is the distribution of the test statistic T assuming the null is true, over all possible assignment vectors A
- For each possible assignment vector, compute T (keeping observed Y fixed, because we are assuming the null)
- The randomization distribution gives us exactly the distribution of T, assuming the sharp null hypothesis is true

Case study: Diet Cola and Calcium

There are in total $\binom{16}{8} = 12,870$ different possible assignment vectors

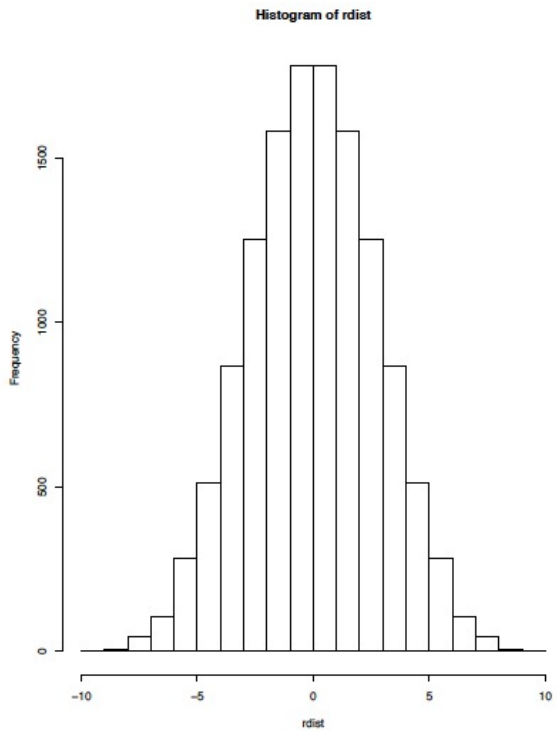
```
> A <- c(rep(1,8), rep(0,8))
> Y <- c(50,62,48,55,58,61,58,56,48,46,54,45,53,46,53,48)
> Abold <- genperms(A,maxiter = 12870)
> Abold[, 1:6]
   [,1] [,2] [,3] [,4] [,5] [,6]
1     1    1    1    1    1    1
2     1    1    1    1    1    1
3     1    1    1    1    1    1
4     1    1    1    1    1    1
5     1    1    1    1    1    1
6     1    1    1    1    1    1
7     1    1    1    1    1    1
8     1    0    0    0    0    0
9     0    1    0    0    0    0
10    0    0    1    0    0    0
11    0    0    0    1    0    0
12    0    0    0    0    1    0
13    0    0    0    0    0    1
14    0    0    0    0    0    0
15    0    0    0    0    0    0
16    0    0    0    0    0    0
> dim(Abold)
[1]    16 12870
```

For subject 1, what is the potential value of the outcome under the sharp null in the random assignments 1 to 6? For subject 13?

NOTE, in practice: we also fix the value of the outcome and reshuffling the exposure assignment.

Case study: Diet Cola and Calcium

For each of these, calculate T, the difference in sample means, keeping the values for calcium excretion fixed




```
> rdist <- rep(NA, times = ncol(Abold))
> for (i in 1:ncol(Abold)) {
+   A_tilde <- Abold[, i]
+   rdist[i] <- mean(Y[A_tilde == 1]) -
mean(Y[A_tilde == 0])
+ }
> hist(rdist)
```

Case study: Diet Cola and Calcium

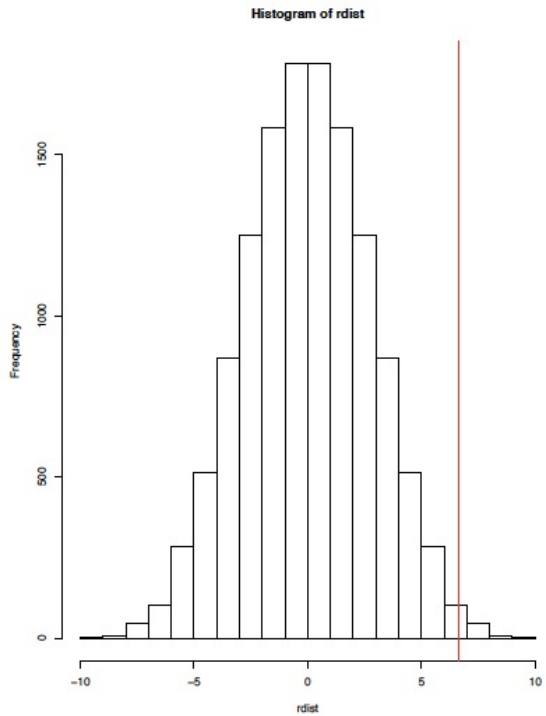
- From the randomization distribution, computing the p-value is straightforward
- The exact p-value is the proportion of test statistics in the randomization distribution that are as extreme as T_{obs}

$$P(T(\boldsymbol{a}, \boldsymbol{Y}) \geq T(\boldsymbol{A}, \boldsymbol{Y}) \mid \tau = 0) = \sum_{\boldsymbol{a}} I(T(\boldsymbol{a}, \boldsymbol{Y}) \geq T(\boldsymbol{A}, \boldsymbol{Y}))/K$$

- where $K = \binom{N}{N_1}$ in classical experiments
- This is exact p-value because there are no distributional assumptions - we are using the exact distribution of T

Case study: Diet Cola and Calcium

Exact p-value = 0.005




```
> # p-value
> pval <- mean(rdist >= T_stat)
> pval
[1] 0.005283605
> quant <- quantile(rdist,probs = 1-pval)
> hist(rdist)
> abline(v = quant,col="red")
```

Case study: Diet Cola and Calcium

- This approach is completely nonparametric - no model specified in terms of a set of unknown parameters
- We don't model the distribution of potential outcomes (they are considered fixed)
- No modeling assumptions or assumptions about the distribution of the potential outcomes

Case study: Diet Cola and Calcium

- For larger samples, the number of possible assignment vectors $\binom{N}{N_1}$ gets very large
  - $N = 6$ and $N_1 = 3$, 20 assignment vectors
  - $N = 10$ and $N_1 = 5$, 252 assignment vectors
  - $N = 100$ and $N_1 = 50$, 1.0089134×10^{29} assignment vectors

- Enumerating every possible assignment vector becomes computationally dfficult

- It's often easier to simulate many (10,000? 100,000?) random assignments

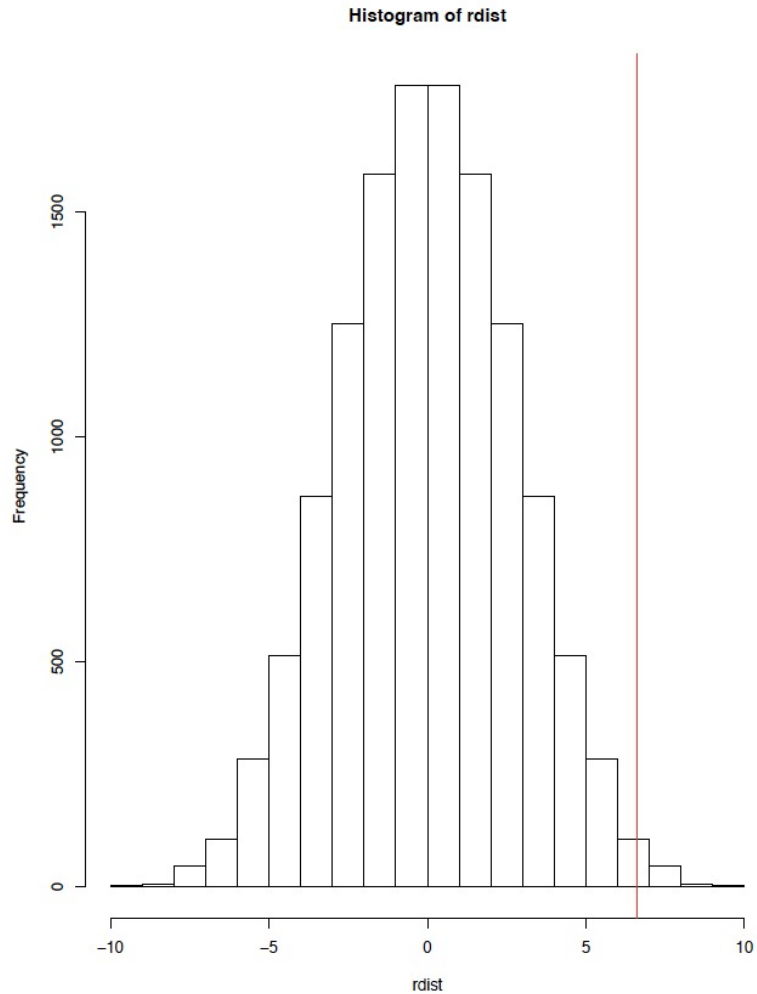
  1. take K samples from the treatment assignment space.
  2. calculate the randomization distribution in the K samples.
  3. tests no longer exact, but bias is under you're control! (increase K)

Approximate p-value

- Repeatedly randomize units to treatments, and calculate test statistic keeping the observed Y fixed
- If the number of simulations is large enough, this randomization distribution will look very much like the exact distribution of T
- Note: estimated p-values will differ slightly from simulation to simulation. This is okay!
- The more simulations, the closer this approximate p-value will be to the exact p-value

Case study: Diet Cola and Calcium

Approximate randomization distribution (approximate p-value = 0.004)



Other test statistics

- The difference in means is great when effects are:
  1. constant and additive
  2. few outliers in the data
- When outliers are present, there is more variation in the randomization distribution
- What about alternative test statistics?

Other test statistics

To further protect against outliers, we can use the differences in quantiles or rank as a test statistics

- We could estimate the difference in quantiles at any point in the distribution: (the median, 0.25 quantile or the 0.75 quantile).
- We could rank the outcomes (higher values of Y_i are assigned higher ranks) and compare the average rank of the treated and control groups.

Confidence intervals via test inversion

- CIs usually justified using Normal distributions and approximations.
- Can calculate CIs here using the duality of tests and CIs:
  - A $100(1-\alpha\%)$ confidence interval is equivalent to the set of null hypotheses that would not be rejected at the α significance level.
  - the 95% CI: find all values of τ such that the null hypothesis is not rejected at the α level
- Operationally:
  1. Choose grid across space of τ: $-0.9, -0.8, ..., 0.8, 0.9$
  2. For each value, use RI to test sharp null at α level.
  3. Collect all values that you cannot reject as the 95% CI.

Notes on RI CIs

- CIs are correct, but might have overcoverage.
- With RI, p-values are discrete and depend on N and N_1.
- If the p-value of 0.05 falls "between" two of these discrete points, a 95% CI will cover the true value more than 95% of the time.

Point estimates

Is it possible to get point estimates? Not really the point of RI, but still possible:

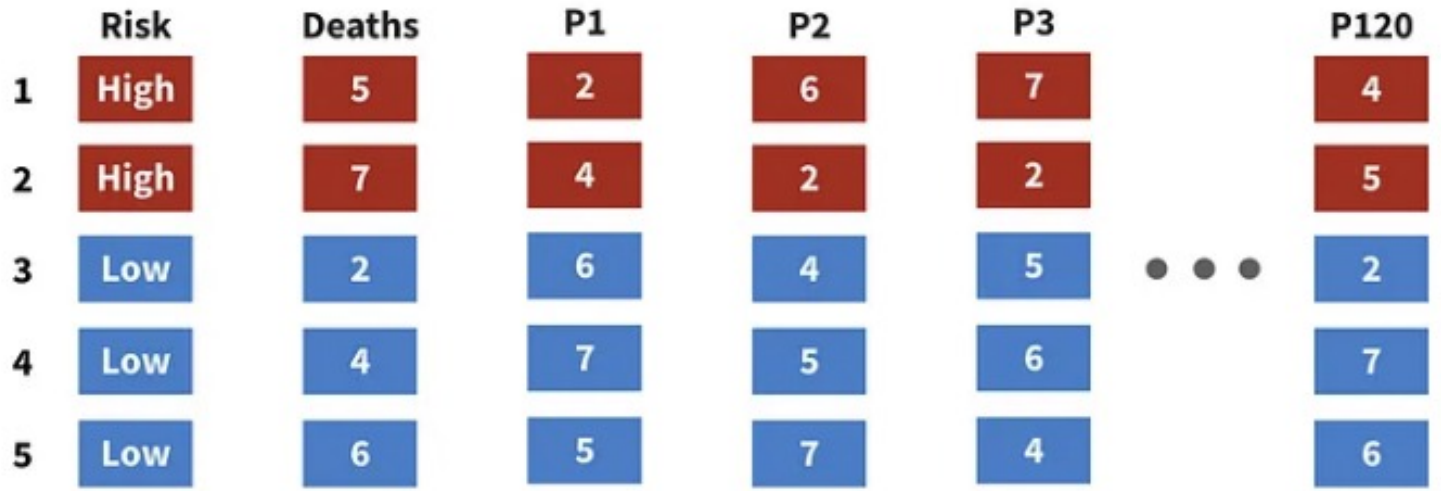
1. Create a grid of possible sharp null hypotheses.
2. Calculate p-values for each sharp null.
3. Pick the value that is "least surprising" under the null.

Usually this means selecting the value with the highest p-value.

Summary

- Physical randomization of treatment assignment as a reason basis for inference
- Inference over repeated (hypothetical) randomization has advantages:
  1. design-based, assumption-free inference
  2. ability to incorporate complex randomization scheme
- Inference over repeated (hypothetical) randomization has limitations:
  1. sharp null hypothesis, constant treatment effect
  2. sample inference rather than population inference
  3. not easy to derive estimates of the effects as the focus is on testing rather than estimation
- Model-based estimation inference under Neyman coming next

Permutation test



	Risk	Deaths	P1	P2	P3	P120
1	High	5	2	6	7	4
2	High	7	4	2	2	5
3	Low	2	6	4	5	2
4	Low	4	7	5	6	7
5	Low	6	5	7	4	6

- 5 observations, two variables of interest: risk and death
- Permute variable death without replacement
  1. Calculate the median of the observed data (the Deaths column).
  2. For each permutation, calculate the median.
  3. Determine the proportion of permutation medians that are more extreme than our observed median.
  4. That proportion is our p-value.

Jerzy Neyman (1894-1981)

Neyman's perspective

Neyman's basic questions were the following:

1. What would the average outcome be if all units were exposed to the active treatment, Y^1?
2. How did that compare to the average outcome if all units were exposed to the control treatment, Y^0?
3. Most importantly, what is the difference between these averages, the average causal effect

What were the basic questions investigated by Fisher? Do they differ?

Neyman's perspective

Neyman's basic concerns were the following:

1. Recover an unbiased estimator for the average causal effect

2. Construct an interval estimator for the causal estimand based on an unbiased estimator for the sampling variance of the average treatment effect estimator

What were the basic concerns investigated for Fisher? Do they differ?

Case study: Sleep or Caffeine

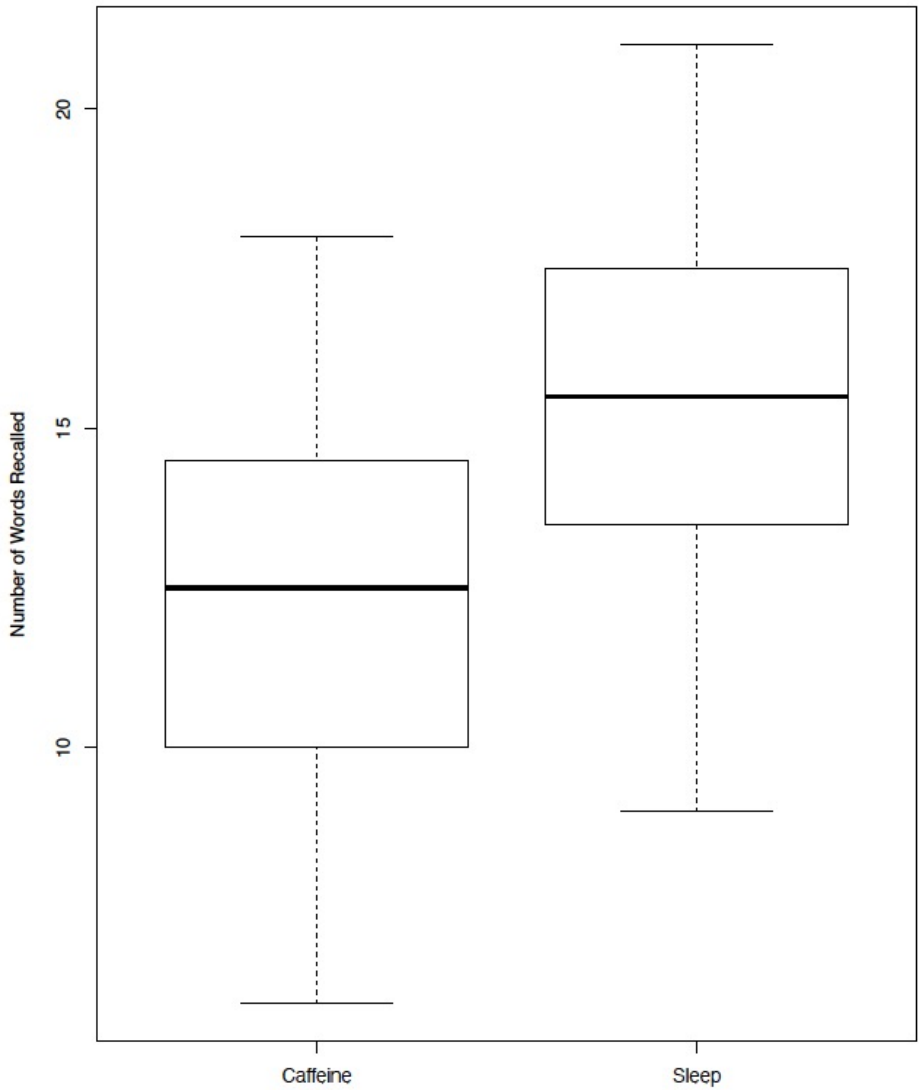
Is sleep or caffeine better for memory?

- 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

Case study: 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?

Case study: Sleep or 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

Finite Sample vs Super Population

Finite sample inference:

1. Only concerned with units in the sample
2. Only source of randomness is random assignment to treatment groups (Fisher exact p-values)

Super population inference:

1. Extend inferences to greater population
2. Two sources of randomness: random sampling, random assignment
3. "repeated sampling"

We'll first explore finite sample inference

Estimand

Neyman was primarily interested in estimating the average causal effect. In the finite sampling setting, this is defined as

$$SACE = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum_{i=1}^{N} Y_i^1}{N} - \frac{\sum_{i=1}^{N} Y_i^0}{N}$$

Note: this can also be called SATE, sample average treatment effect, in the literature.

Estimator

A natural estimator is the difference in observed sample means (Why?)

$$\hat{SACE} = \bar{Y}^1_{obs} - \bar{Y}^0_{obs} = \frac{\sum_{i=1}^{N} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{N} (1 - A_i) Y_i^0}{N_0}$$

Case study: 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?

Case study: Sleep or 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 (let $A = 1$ denote sleep and $A = 0$ denote caffeine):

$$\hat{SACE} = \bar{Y}^1_{obs} - \bar{Y}^0_{obs} = 15.25 - 12.25 = 3$$

Unbiasedness

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

$$\hat{SACE} = SACE$$

For completely randomized experiments,

$$\hat{SACE} = \bar{Y}^1_{obs} - \bar{Y}^0_{obs} = \frac{\sum_{i=1}^{N} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{N} (1 - A_i) Y_i^0}{N_0}$$

is an unbiased estimator for

$$SACE = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum_{i=1}^{N} Y_i^1}{N} - \frac{\sum_{i=1}^{N} Y_i^0}{N}$$

Neyman's inference (finite sample)

1. Define the estimand

$$SACE = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum_{i=1}^{N} Y_i^1}{N} - \frac{\sum_{i=1}^{N} Y_i^0}{N}$$

2. unbiased estimator of the estimand:

$$\hat{SACE} = \bar{Y}_{obs}^1 - \bar{Y}_{obs}^0 = \frac{\sum_{i=1}^{N} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{N} (1 - A_i) Y_i^0}{N_0}$$

3. Calculate the true sampling variance of the estimator

True Sampling Variance of the Estimator

$$var(\bar{Y}^1_{obs} - \bar{Y}^0_{obs}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}$$

where

$$S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^1 - \bar{Y}^1)^2$$

$$S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^0 - \bar{Y}^0)^2$$

$$S_{10}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[(Y_i^1 - Y_i^0) - (\bar{Y}^1 - \bar{Y}^0)]^2$$

For derivations of this, see Chapter 6 of Imbens and Rubin.

Extra term

$$S_{10}^2 = \frac{1}{N-1}\sum_{i=1}^{N}[(Y_i^1 - Y_i^0) - (\bar{Y}^1 - \bar{Y}^0)]^2$$

- 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

$$SACE = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum_{i=1}^{N} Y_i^1}{N} - \frac{\sum_{i=1}^{N} Y_i^0}{N}$$

2. unbiased estimator of the estimand:

$$\hat{SACE} = \bar{Y}_{obs}^1 - \bar{Y}_{obs}^0 = \frac{\sum_{i=1}^{N} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{N} (1 - A_i) Y_i^0}{N_0}$$

3. True sampling variance of the estimator

$$var(\hat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}$$

4. Look for an unbiased estimator of the true sampling variance of the estimator...impossible!

Estimator of Variance

$$\hat{var}(\hat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

$$S_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N} A_i (Y_i^1 - \bar{Y}_{obs}^1)^2$$

$$S_0^2 = \frac{1}{N_0 - 1} \sum_{i=1}^{N} (1 - A_i)(Y_i^0 - \bar{Y}_{obs}^0)^2$$

These are the sample variances of observed outcomes under $A = 1$ and $A = 0$

Estimator of Variance

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

Case study: Sleep or Caffeine

$$\hat{var}(\hat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = \frac{3.3^2}{12} + \frac{3.5^2}{12} = 1.958$$

Neyman's inference (finite sample)

1. Define the estimand

$$SACE = \bar{Y}^1 - \bar{Y}^0 = \frac{\sum_{i=1}^{N} Y_i^1}{N} - \frac{\sum_{i=1}^{N} Y_i^0}{N}$$

2. unbiased estimator of the estimand:

$$\hat{SACE} = \bar{Y}_{obs}^1 - \bar{Y}_{obs}^0 = \frac{\sum_{i=1}^{N} A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^{N} (1 - A_i) Y_i^0}{N_0}$$

3. True sampling variance of the estimator

$$var(\hat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}$$

4. Estimator of the true sampling variance of the estimator is an overestimate:

$$\hat{var}(\hat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

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

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

$$\hat{SACE} +/- z^* \sqrt{\hat{var}(\hat{SACE})}$$

$$\bar{Y}^1_{obs} - \bar{Y}^0_{obs} +/- z^* \sqrt{\frac{s_1^2}{N_1} + \frac{S_0^2}{N_0}}$$

z^* is the value leaving the desired percentage in between $-z^*$ and z^* in the standard normal distribution

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

Case study: Sleep or Caffeine

$$\bar{Y}^1_{obs} - \bar{Y}^0_{obs} + / - z^* \sqrt{\frac{s_1^2}{N_1} + \frac{S_0^2}{N_0}}$$

$$qt(0.975, df = 11) = 2.2$$

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

95% CI: (-0.86, 6.08)

Confidence Intervals

For finite sample inference:

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

Hypothesis Testing

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

$$H_0 : Y_i^1 = Y_i^0$$

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)

Neyman's approach is the familiar t-test

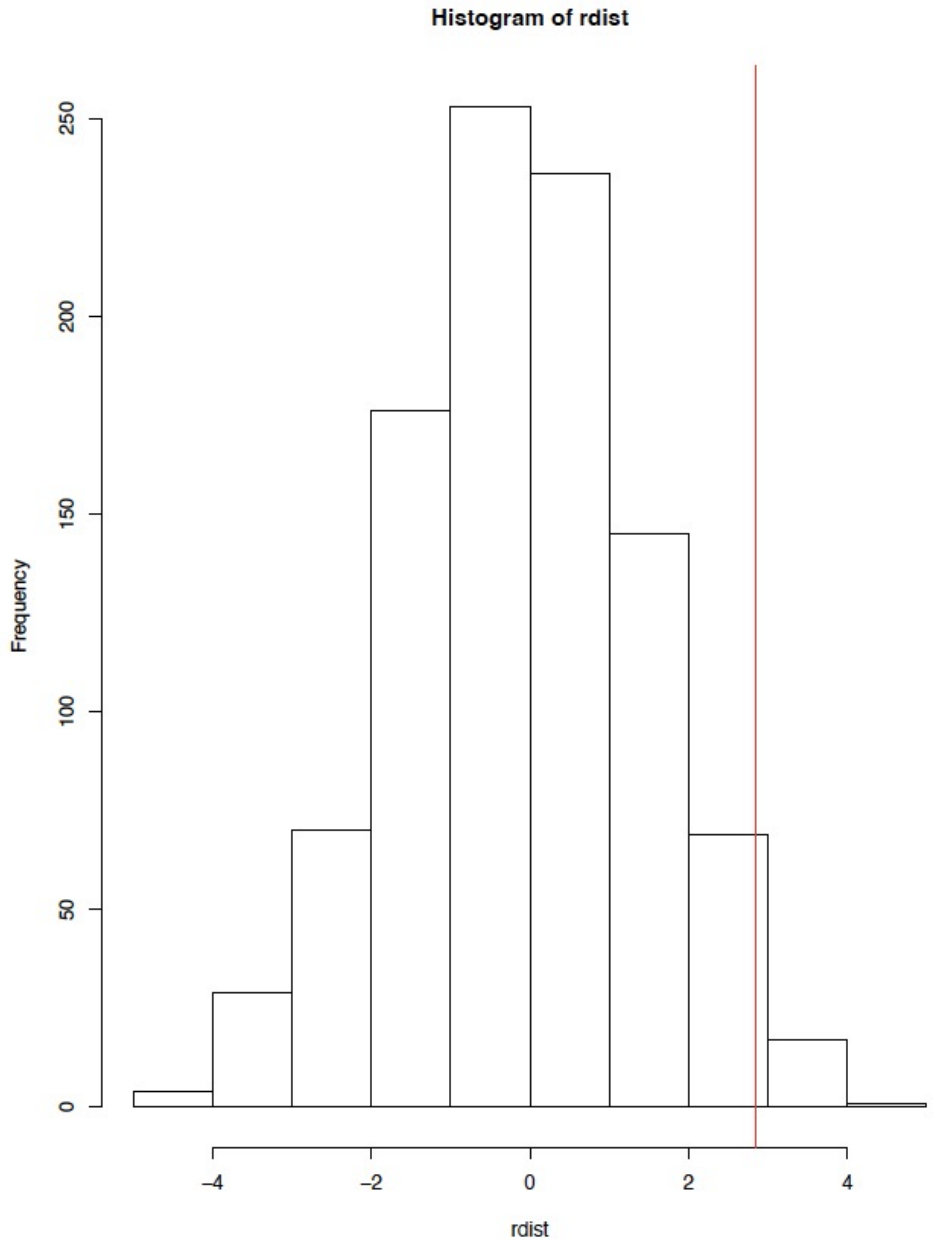
$$t = \frac{\hat{SACE}}{\sqrt{\hat{var}(\hat{SACE})}} = \frac{\bar{Y}^1_{obs} - \bar{Y}^0_{obs}}{\sqrt{\frac{s_1^2}{N_1} + \frac{S_0^2}{N_0}}}$$

Case study: Sleep or Caffeine

$$t = \frac{\hat{SACE}}{\sqrt{\hat{var}(\hat{SACE})}} = \frac{\bar{Y}^1_{obs} - \bar{Y}^0_{obs}}{\sqrt{\frac{s_1^2}{N_1} + \frac{S_0^2}{N_0}}} = \frac{15.25 - 12.25}{\sqrt{\frac{3.3^2}{12} + \frac{3.5^2}{12}}} = 2.16$$

$$pt(2.16, df = 11, lower.tail = F) = 0.0278$$

Case study: Sleep or Caffeine



Population Average Causal Effect

Suppose we also want to consider random sampling from the population (in addition to random assignment)

How do things change?

Neyman's inference (super population)

1. Define the estimand

$$PACE = E(Y_i^1 - Y_i^0)$$

2. unbiased estimator of the estimand:

$$\hat{PACE} = \bar{Y}_{obs}^1 - \bar{Y}_{obs}^0 = \frac{\sum_{i=1}^N A_i Y_i^1}{N_1} - \frac{\sum_{i=1}^N (1 - A_i) Y_i^0}{N_0}$$

3. True sampling variance of the estimator

$$var(\hat{PACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_{10}^2}{N}$$

4. Estimator of the true sampling variance of the estimator is an overestimate:

$$\hat{var}(\hat{PACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$$

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

Population Average Causal Effect

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

- Goal: testing
- Considers only random assignment
- H_0: no treatment effect
- Works for any test statistic
- Exact distribution
- Works for any known assignment mechanism

Fisher vs Neyman

Neyman

- Goal: estimation
- Considers random assignment and random sampling
- H_0: average treatment effect $=0$
- Difference in means
- Approximate, relies on large N
- Only derived for common designs