Randomization Tests in the Context of Trial Disruptions

Michael Proschan

NIAID

Pre-Quiz: True or False?

1. Randomization tests do not allow generalization to the population, whereas t-tests do

2. If I look at blinded data and see a bimodal distribution, that will unblind me and:
   a. cause alpha inflation
   b. have other negative consequences

Blinded Adaptations

- Blinded adaptations have been used for a long time; e.g., to modify sample size
  - Binary outcomes: use pre-specified treatment effect & interim estimate of overall event probability (Gould, 1992)
  - Continuous outcomes: use pre-specified treatment effect & interim "lumped" variance (Gould & Shih, 1992)
- Such adaptations are **pre-planned**
- Trial disruptions are **not** pre-planned

True Fable 1

- Once upon a time, investigators chose a primary endpoint for a double-blinded tuberculosis trial
  - Pretend it was the number of lesions on the lung
- Examining scans blinded to treatment assignment, they noticed it was not measurable (not in sense of Royden!)
  - Changed the primary endpoint to something that could be measured
    - Pretend it was the volume of lesions
- A randomization test allows investigators to live happily ever after!

True Fable 2

- Some infectious disease trials eliminate early deaths–some patients are too sick to be helped
- PREVAIL II Ebola treatment trial (Davey et al, 2016)
  - Liberia, Sierra Leone, Guinea
  - Randomized Ebola patients to standard care versus standard care plus ZMapp, a triple monoclonal antibody cocktail
  - Primary outcome: 28-day mortality
  - Ended early because epidemic ended
- Suggestion: Eliminate early deaths
- Yuck! What if treatment kills patients?

True Fable 2

- Jim Neaton suggested using principal stratification:
  - Use logistic regression to identify a linear combination L of baseline predictors of early death
  - Stratify analysis by P(early death) and focus on low risk stratum
- Determine L using all patients (blinded)
  - Use stepwise regression, etc.
  - Then apply a randomization test (Fisher's exact test)
- We lived happily ever after

True Fable 3

- The Adaptive COVID-19 Trial (ACTT-1) (Beigel et al, 2020):
  - Hospitalized patients with COVID-19
  - Remdesivir+standard care vs placebo+standard care
  - Primary outcome: time to recovery
  - Very little information about COVID-19 before trial, so best endpoint was unclear

- We originally considered proportional odds model on 8-point ordinal outcome at day 15
  - 1=out of hospital, 8=dead
  - Model assumes treatment to control odds ratio of a score of *s* or better is same for s=1,2,…,7

True Fable 3

- Problem: we do not know what day to choose (no prior experience with COVID-19)

- We changed to time to recovery before looking at any outcome data
  - We lived happily ever after

- What if we had used proportional odds model but determined day based on blinded data?
  - Blinded look won't tell us where the treatment effect is except in extreme situations
    - E.g. no one (or everyone) recovers by day 15 ⇨ day 15 is bad!
  - Without such an extreme situation, we would NOT have lived happily ever after!

Caveats

- Note: caveats of Martin Posch's talk apply
- E.g., in TB example of changing endpoint, correct conclusion is that treatment affected at least one outcome variable you considered
- Still, judgment is required
  - Given that the original outcome could not be measured, it seems reasonable to conclude that treatment affected volume of lesions

Caveats

- Even though randomization tests do not inflate alpha under the strong null hypothesis of no effect on any variable you looked at, it can lead to overestimation or underestimation of treatment effect
  - Under H_1, it can cause unblinding (e.g., bimodal distribution when you know it should be unimodal under H_0)
  - Unblinded investigators could lead to differential background treatment in the two arms

Conditioning on Ancillary Statistics

- Conditioning on ancillary statistics to get the distribution of a statistic is a generally accepted principle in inference

- Example: Compare treatment & control means of a continuous outcome; assume
  - Data are normally distributed
  - Common variance 1

- Suppose you flip a fair coin to determine the sample size
  - Heads: use 10 people per arm
  - Tails: use 1,000 per arm

- The coin is heads, so you use 10 per arm

Conditioning on Ancillary Statistics

- No one would compute the variance of $\bar{Y}_T - \bar{Y}_C$ by taking into account that the coin **could** have been tails, so *n* **could** have been 1,000!

- We would all condition on the coin being heads and *n* being 10

- Randomization tests follow a similar principle:
  - Without treatment labels, the data give almost no information about the treatment effect
  - Condition on the data!

Thesis

- Randomization tests are known to be asymptotically the same as t-tests under reasonable assumptions

- If randomization tests are valid after looking at data, then t-tests should also be

Randomization Tests: Asymptotically Like t-Tests?

- Consider a paired setting like the Community Intervention Trial for Smoking Cessation (COMMIT)
  - Pair-matched and randomized 22 communities to a 4-year smoking cessation intervention or not
  - Primary endpoint: paired differences $D_1,\ldots,D_{11}$ of quit rates among 550 pre-selected heavy smokers

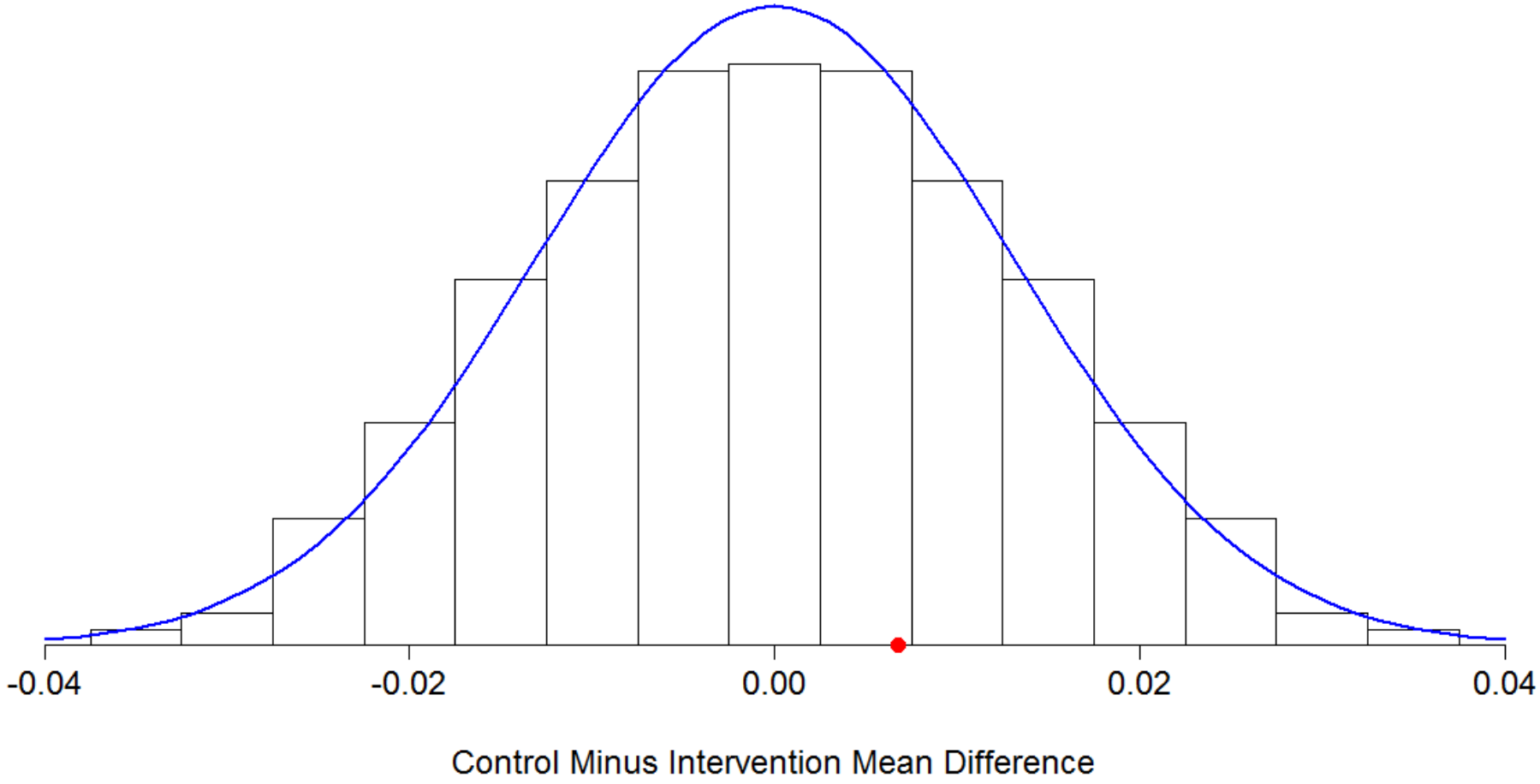
- A randomization test permutes labels within pairs
- Each observed difference is $\pm d_i$ with probability 1/2

Randomization Tests: Asymptotically Like t-Tests?

- For each randomization, compute $\bar{d}$
- Resulting distribution of the 2^{11} values $\bar{d}_1, \bar{d}_2, \ldots$ is the randomization distribution

- Results of 1-tailed paired t-test and randomization test are remarkably similar
  - Paired t-test p-value: 0.685
  - Randomization p-value: 0.686

COMMIT TRIAL



Randomization distribution: histogram
Normal approximation: blue superimposed curve

Randomization Tests: Asymptotically Like t-Tests?

- COMMIT had only 11 pairs!
- Normal approximation often kicks in fairly quickly
- Belies the myth that you can generalize with the t-test, but not with the randomization test
  - They are essentially the same test for moderately large sample sizes!
  - Whether you can generalize depends on judgment, not which test you use

Randomization Tests: Asymptotically Like t-Tests?

- But there are circumstances when the randomization and t-tests are not close
- Suppose there is an outlier
- Subtract 10 from D_{11} in COMMIT (disregard the fact that D is between -1 and 1)

$$T = \frac{\sum D_i}{\sqrt{11s^2}} = -1.00 \ (p = 0.17)$$

- Randomization p-value: 0.30

Randomization Distribution is Bimodal When There Is An Extreme Outlier

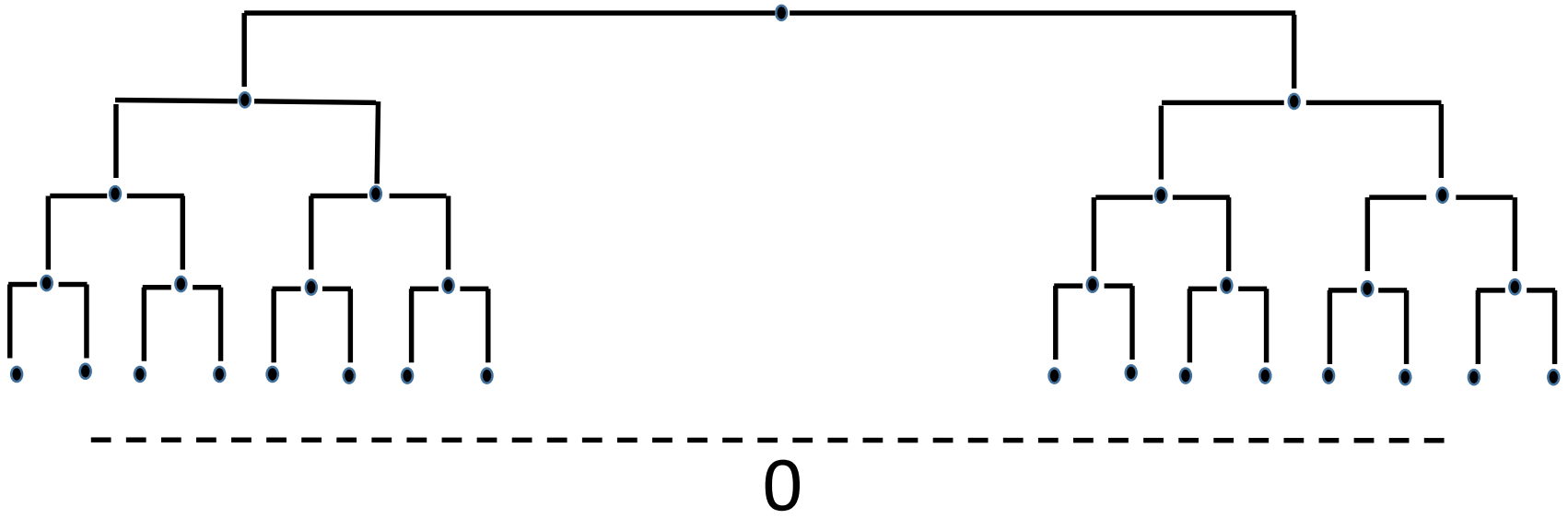


Illustration with 4 Paired Differences

Randomization Tests: Asymptotically Like t-Tests?

- Also dissimilar results if treatment effect is huge

- Subtract 10 from **each** paired difference (impossible, again, but suspend reality)

$$T = \frac{\sum D_i}{\sqrt{11s^2}} = -737.95 \ (p < 2.2 \times 10^{-16})$$

- But randomization test p-value: $1/2^{11}=0.0005$

Randomization Tests: Asymptotically Like t-Tests?

- Incidentally, normal approximation **IS** still accurate if you use the randomization variance

$$\hat{\sigma}^2 = \left(\frac{1}{n}\right)\sum D_i^2 \text{ instead of } s^2 = \left(\frac{1}{n-1}\right)\sum (D_i - \overline{D})^2$$

$$Z = \frac{\sum D_i}{\sqrt{11\hat{\sigma}^2}} = \text{Z=-3.317 (p=0.0005)}$$

compared to randomization p-value $1/2^{11}$=0.0005

Randomization Tests: Asymptotically Like t-Tests?

- See also Proschan and Shaw (2016), page 163, exercise 10 regarding the effect of an outlier on t-tests and permutation tests

A Beautiful Connection

- Beautiful connection between randomization test and t-test
- A randomization test fixes $D_1^2, \ldots, D_n^2$ and uses the null conditional distribution of $\sum D_i$ given $D_1^2, \ldots, D_n^2$ to compute a p-value
- Asymptotically, the randomization distribution is normal and depends on $D_1^2, \ldots, D_n^2$ only through $\sum D_i^2$

$$\frac{\sum D_i}{\sqrt{\sum D_i^2}} \Big| \sum D_i^2 \approx N(0,1) \quad \text{i.e.,} \quad \frac{\sum D_i}{\sqrt{\sum D_i^2}} \approx \text{indep of } \sqrt{\sum D_i^2}$$

A Beautiful Connection

- Suggests that for iid N(μ,σ^2) data $D_1,\ldots,D_n$, the modified t-statistic

$$\frac{\sum D_i}{\sqrt{\sum D_i^2}}$$

is independent of $\sum D_i^2$ under null that μ=0

- Seems hard to believe, but it is true!

A Beautiful Connection

- Follows from Basu's theorem (the most beautiful theorem in statistics) because under $N(0,\sigma^2)$,

$$\frac{\sum D_i}{\sqrt{\sum D_i^2}}$$

is ancillary and $\sum D_i^2$ is complete and sufficient

- Useful with adaptive t-tests (Proschan, Glimm and Posch, 2014)

A Beautiful Connection

- The ideas of sufficient, complete, and ancillary statistics can also be used in two-sample settings to show that only a randomization test can control the type 1 error rate regardless of the true common continuous distribution of data (Lehmann, 1959)

Conclusions

- Trial disruptions can happen
- Randomization tests can save you if adaptations were made based on blinded data
- Because randomization tests are valid and are asymptotically equivalent to t-tests, t-tests are approximately valid after blinded changes if sample sizes are large (but use the randomization variance)

Answers to Pre-Test

1. Randomization tests do not allow generalization to the population, whereas t-tests do

FALSE

1. If I look at blinded data and see a bimodal distribution, that will unblind me and can:
   a. cause alpha inflation FALSE
   b. have other negative consequences TRUE

References

- Beigel et al. (2020). **NEJM** DOI: 10.1056/NEJMoa2007764
- Davey (2016) et al (2016). *NEJM* **375**, 1448-1456
- Gould (1992). *Stat. in Med.* **11**, 55-66.
- Gould & Shih (1992). *Commun. in Stat.* **21**, 2833-2853.
- Lehmann, E.L. (1959). Testing Statistical Hypotheses. John Wiley & Sons
- Posch & Proschan (2012). *Stat. in Med.* **31**, 4146-4153.

References (continued)

- Proschan, Glimm & Posch (2014). *Statistics in Medicine* **33**, 4734-4742.
- Proschan and Shaw (2016). *Essentials of Probability Theory for Statisticians*. Chapman & Hall.
- Wald & Wolfowitz (1944). *Ann. Math. Statist.* **15**, 358-372.