

Methods: Mind the Gap
Webinar Series

Randomization-Based Inference in Prevention Research

Presented by:

Michael Proschan, Ph.D.
National Institute of Allergy and Infectious Disease (NIAID)



Hail Randomization

- Randomization is what separates clinical trials from all other studies
- Key property: Randomization eliminates bias in who receives treatment
 - In observational studies, people who choose an experimental intervention may be different (e.g., more health conscious) than those who don't
 - In a randomized trial, a coin flip decides

Analyze as You Randomize Principle

- Randomization can also be the sole basis for deciding whether the intervention works
- This *randomization-based inference* is sometimes called the *analyze as you randomize principle*
- Invented by Fisher (1935) in connection with the “lady tasting tea” experiment

“Tea” Test

- A lady claims she can tell whether cream or tea was added first to a cup
- Experiment: add tea first to 4 cups, cream first to 4
T T T T C C C C
- Present cups in random order

“Tea” Test

- A lady claims she can tell whether cream or tea was added first to a cup
- Experiment: add tea first to 4 cups, cream first to 4
C T T C C T T C
- Present cups to the lady in random order
- The lady guesses which cups had cream added first
- Count the number correct among the cream first cups
- Suppose she correctly guesses 2,4,7,8 had cream added first

“Tea” Test

○ her guess

Cup: 1 2 3 4 5 6 7 8

The null hypothesis is she is guessing (guess would have been the same regardless of random order)



“Tea” Test

○ her guess
T tea first
C cream first

Cup:	1	2	3	4	5	6	7	8
Truth:	T	C	T	C	T	T	C	C

The null hypothesis is she is guessing (guess would have been the same regardless of random order)



“Tea” Test

○ her guess
T tea first
C cream first

Cup:	1	2	3	4	5	6	7	8
Truth:	T	C	T	C	T	T	C	C
		✓		✓			✓	✓

The null hypothesis is she is guessing (guess would have been the same regardless of random order)

Number correct among cream first cups: 4



“Tea” Test

- How can we obtain a reference distribution of number correct under the null hypothesis that she is completely guessing?
 - Consider all possible orderings of the cups

“Tea” Test

○ her guess
T tea first
C cream first

Cup:	1	2	3	4	5	6	7	8
What-if:	T	T	T	T	C	C	C	C
							✓	✓

The null hypothesis is she is guessing (guess would have been the same regardless of random order)

Number correct among cream first cups: 2



“Tea” Test

○ her guess
T tea first
C cream first

Cup:	1	2	3	4	5	6	7	8
What-if:	C	C	T	T	C	C	T	T

✓

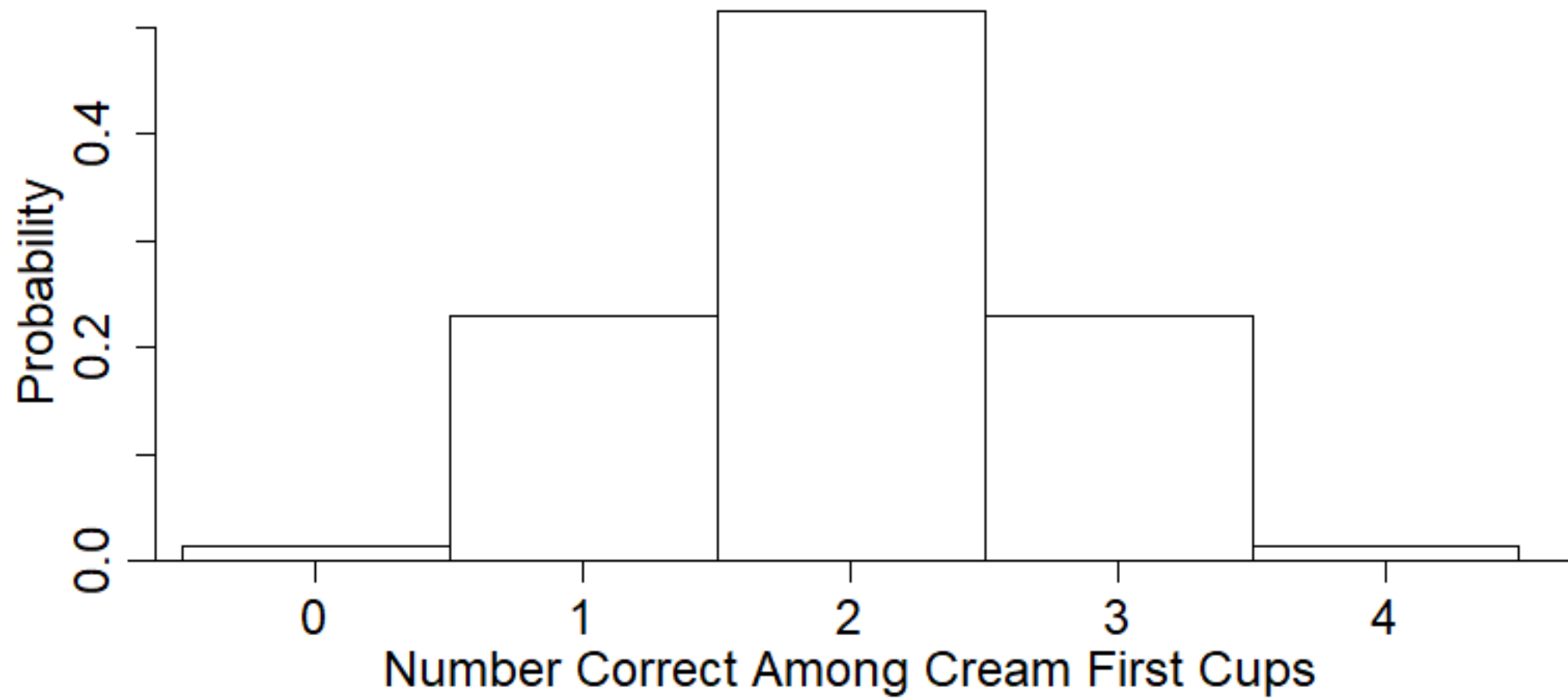
The null hypothesis is she is guessing (guess would have been the same regardless of random order)

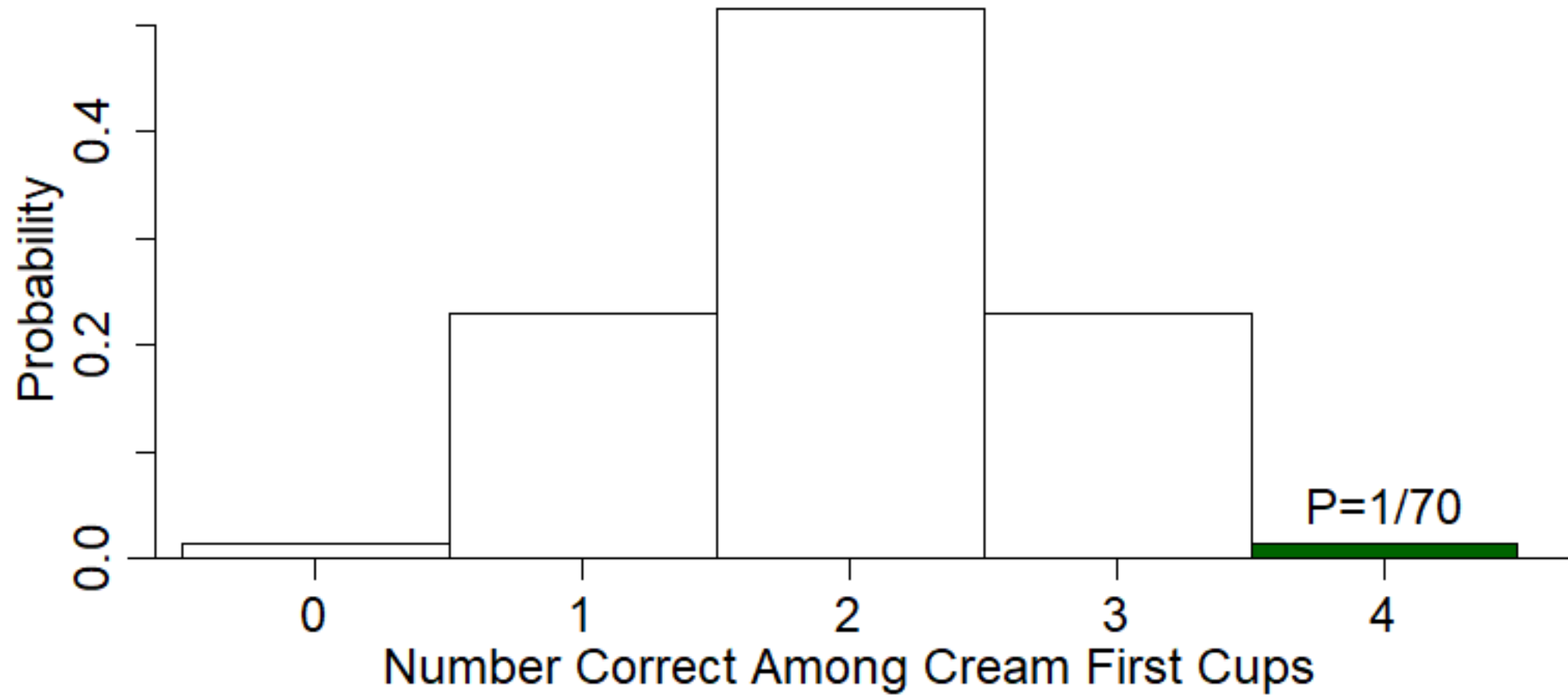
Number correct among cream first cups: 1



“Tea” Test

- Do for all possible orders to get a null reference distribution of number correct
- P-value: proportion of randomizations producing at least as many correct guesses as she got among cream first cups
- Called a *randomization test* (AKA *re-randomization test* or *permutation test*)
- She got all 4 cream first cups right





How Is This Connected to Clinical Trials?

- T and C are now treatment and control instead of tea or cream first
- Data are now dead or alive instead of the lady's guesses
- Randomize patients to T or C using *random permuted blocks*:
- For a block of size 8, have 4 cards with “T” (Treatment) and 4 with “C” (Control)

T T T T C C C C

- Randomly permute the order and assign next 8 people

How Is This Connected to Clinical Trials?

- T and C are now treatment and control instead of tea or cream first
- Data are now dead or alive instead of the lady's guesses
- Randomize patients to T or C using *random permuted blocks*:
- For a block of size 8, have 4 cards with “T” (Treatment) and 4 with “C” (Control)

C T T C C T T C

- Randomly permute the order and assign next 8 people

Clinical Trial

○ dead
T treatment
C control

People	1	2	3	4	5	6	7	8
Truth:	T	C	T	C	T	T	C	C
		✓		✓			✓	✓

The null hypothesis is people who were going to die would have died, no matter what they were randomized to

Number dead among control-group: 4



Clinical Trial

○ dead
T treatment
C control

People	1	2	3	4	5	6	7	8
What if:	T	T	T	T	C	C	C	C
							✓	✓

The null hypothesis is people who were going to die would have died, no matter what they were randomized to

Number dead among control-group: 2



Clinical Trial

○ dead
T treatment
C control

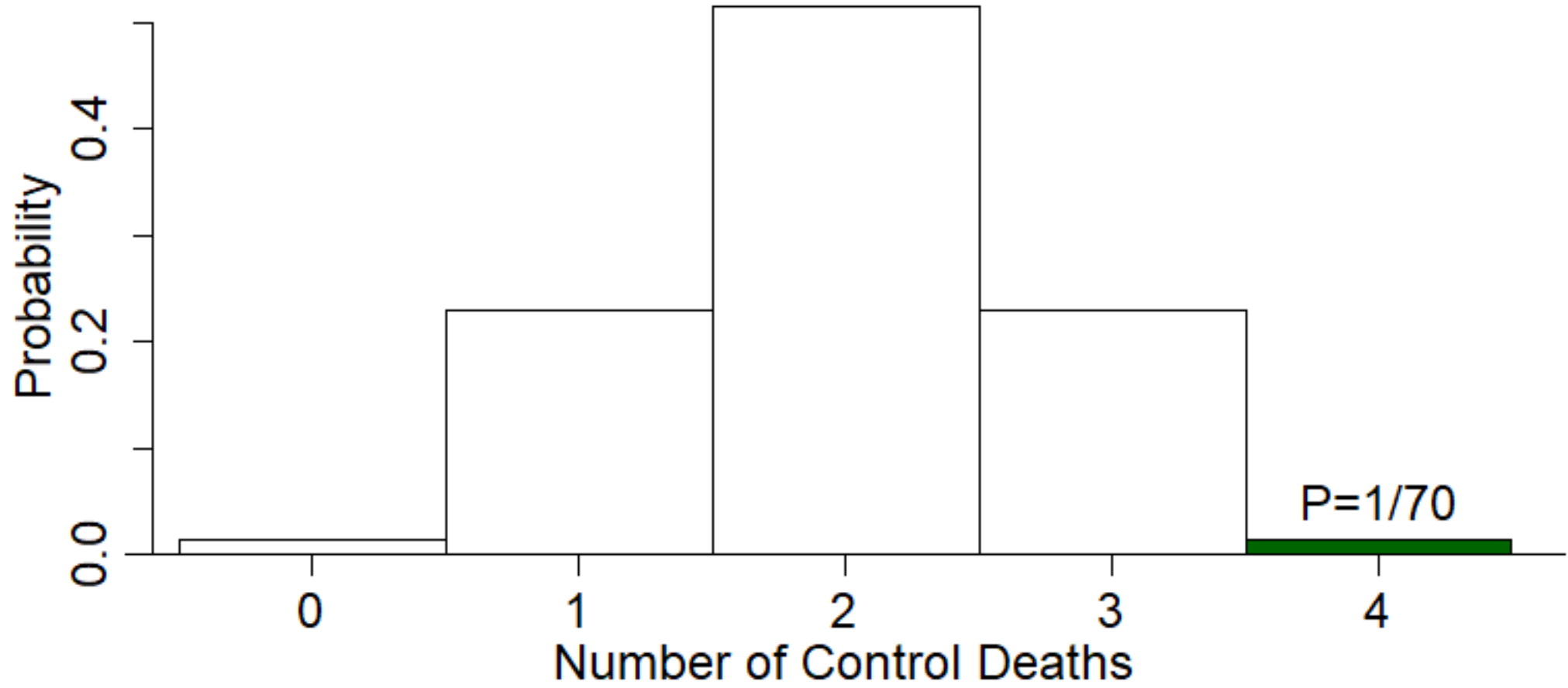
People	1	2	3	4	5	6	7	8
What if:	C	C	T	T	C	C	T	T

✓

The null hypothesis is people who were going to die would have died, no matter what they were randomized to

Number dead among control-group: 1





Fisher's Exact Test

- In this example, the event does not have to be death (any binary outcome)
- Example had 4 deaths, but method works regardless of the number of deaths
- Data remain fixed; only the treatment labels are random
- This is called Fisher's exact test

Randomization Test

1. Treat the data as fixed (nonrandom) numbers and compute the observed treatment effect
2. Re-randomize using the same method used in the trial
3. Compute the treatment effect pretending the re-randomized assignments were the real assignments
4. Repeat steps 1-3 a huge number of times
5. P-value: proportion of re-randomized trials with an intervention effect at least as extreme as observed effect

Randomization Test

- Premise: if intervention has no effect, then observed results are equally plausible regardless of intervention and control labels
- Randomization test works with different data types (continuous, categorical, and so on)
- Provides a valid test with **NO** assumptions

Randomization Test

- Most statistical methods assume participants are randomly sampled from the population of interest
- How often does this happen in clinical trials?
- Never!
 - We don't have a list of all people who have a disease
 - Even if we had a list and sampled randomly from it, not everyone would consent
- Even though we never randomly sample, we always randomize

Randomization Test

- A randomization test tests the strong null hypothesis that the intervention has no effect on anyone
 - Important caveat because it could come out statistically significant if treatment has an effect on the variance, but not the mean
 - Even more important caveat if randomization tests are used and an unplanned change in design occurs before breaking randomization
- A randomization test can be used with any randomization method

Randomization Test

- Randomization methods
 - Simple (fair coin flip for each patient)
 - Permuted block (randomly permuting blocks with half T, half C)
 - Stratified (permuted blocks separately within subgroups)
 - Covariate-adaptive (CAR), AKA *minimization* or *dynamic allocation*
 - Use covariate info on next patient to see which assignment maximizes balance; favor that assignment in randomization
 - Response-adaptive (RAR)
 - Use outcome data to change probabilities to favor treatments performing better
 - Lead us not into the temptation of RAR! (Proschan and Evans, 2020)

Covariate Adaptive Randomization

- E.g., suppose have factors gender and hypertension status, & so far:

Gender (G)		Hypertension (H)		
	M	F	Yes	No
T	10	3	8	5
C	8	3	6	5

Imbalance=G imbalance + H imbalance

Covariate-Adaptive Randomization (CAR)

Next patient is **male, non hypertensive**

	Gender (G)		Hypertension (H)	
	M	F	Yes	No
T	10	3	8	5
C	8	3	6	5

$I = G \text{ imbalance} + H \text{ imbalance}$

If next patient is T, imbalance = $11 - 8 + 6 - 5 = 4$

If next patient is C, imbalance = $10 - 9 + 6 - 5 = 2$

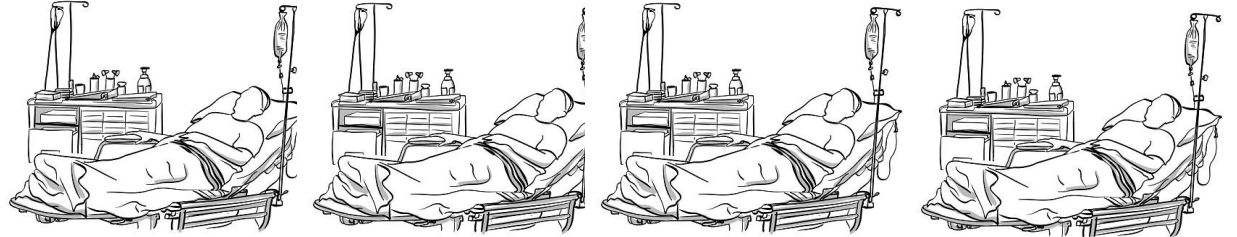
More balance if we assign to the control, so flip an unfair coin with $P(C) = 2/3$

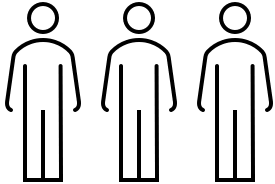
Response-Adaptive Randomization (RAR)

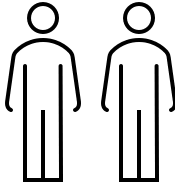
Intervention 



 Control



Intervention 

 Control

COMMIT

- Randomization tests can also use in a paired setting such as the Community Intervention Trial for Smoking Cessation (COMMIT investigators, 1995)
 - Pair-matched 22 communities based on geographic location, size, and sociodemographic factors
 - Randomized one member of each pair to no intervention, and the other to an intervention to help smokers quit
 - Primary outcome in each community was the proportion of $\approx$ 550 heavy smokers aged 25-64 who quit smoking

COMMIT

- The data in one of the pairs was

	I	C	$D=I-C$
Pair 1	0.204	0.249	-0.045

- If intervention has no effect, the data should be equally plausible if we switch I and C labels:

	C	I	$D=I-C$
Pair 1	0.204	0.249	+0.045

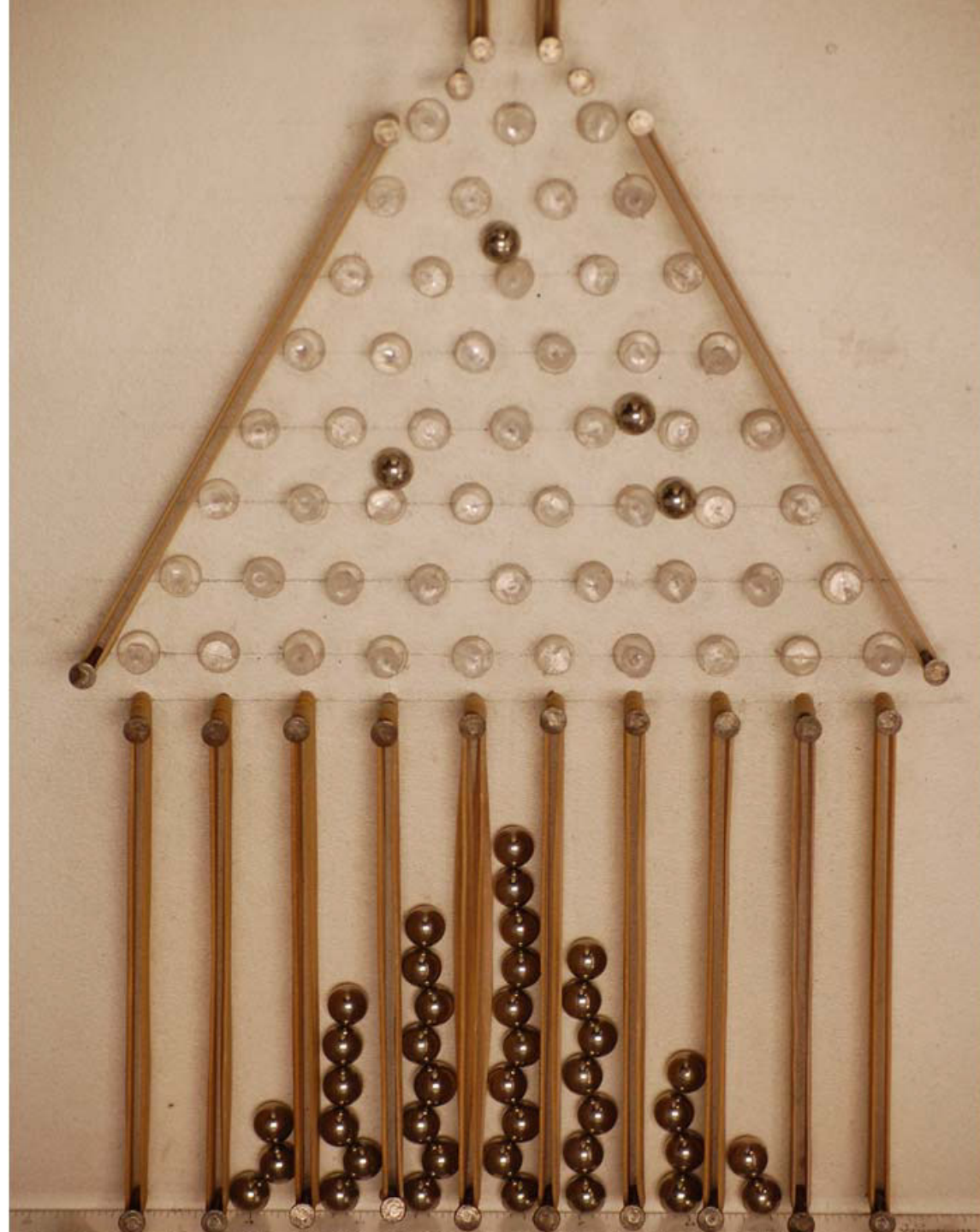
COMMIT

- The randomization distribution for this pair is -0.045 or $+0.045$ with probability $\frac{1}{2}$ each
- Similarly, each pairwise difference is equally likely to be $+d$ or $-d$
- The null reference distribution is obtained by computing the mean of the 11 paired differences for every possible re-labeling
- P-value was $p=0.686$
- P-value for t-test was $p=0.685$
- What a coincidence!

Spoiler Alert

- This was **NOT** a coincidence!
- For traditional randomization methods, when sample sizes are large, the randomization and t-tests give nearly the same answer (in COMMIT, 11 was large enough!)
- Works because the test statistic is like a sum of + or – deflections of roughly equal size
- Think about a quincunx, AKA Galton board ($\approx$ \$35 at Amazon)

Approximately

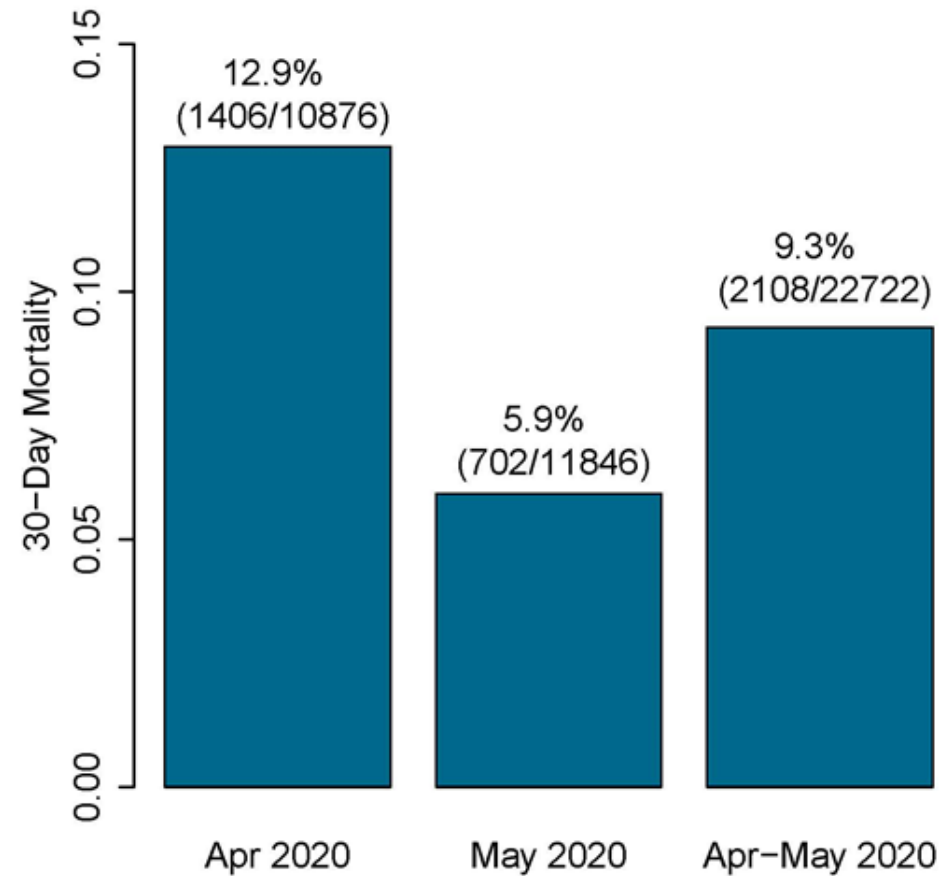


Normal

Temporal Trends

- A randomization test protects against temporal trends
- In trials in emerging infectious diseases, baseline health can change dramatically over time because of
 - Evolution of the virus
 - Better background care
 - Introduction of vaccines

COVID-19



Asch et al. (2021) JAMA Intern Med. 181, 471-478

Temporal Trends

- A bad randomization method can be subject to bias from temporal trends
- Example of a bad randomization method:
 - Use 1:4 randomization to T or C in first half of trial, then use 4:1 in the second half
- Suppose there is a temporal trend from 1st to 2nd half
- Most controls were in first half, most treated patients were in second half, so treatment effect is confounded by time

Temporal Trends

- Chance of a false positive can be elevated with some standard analysis methods
- A randomization test automatically handles temporal trends
 - Temporal trends will occur in the re-randomized trials as well
 - To reach statistical significance, the observed effect has to be large enough to be distinguishable from a possible temporal trend
- Note: this randomization method was foolish, but similar things can happen with response adaptive randomization (RAR) and did happen in one trial using covariate adaptive randomization (CAR)

Eye-Opening Experience for Minimization

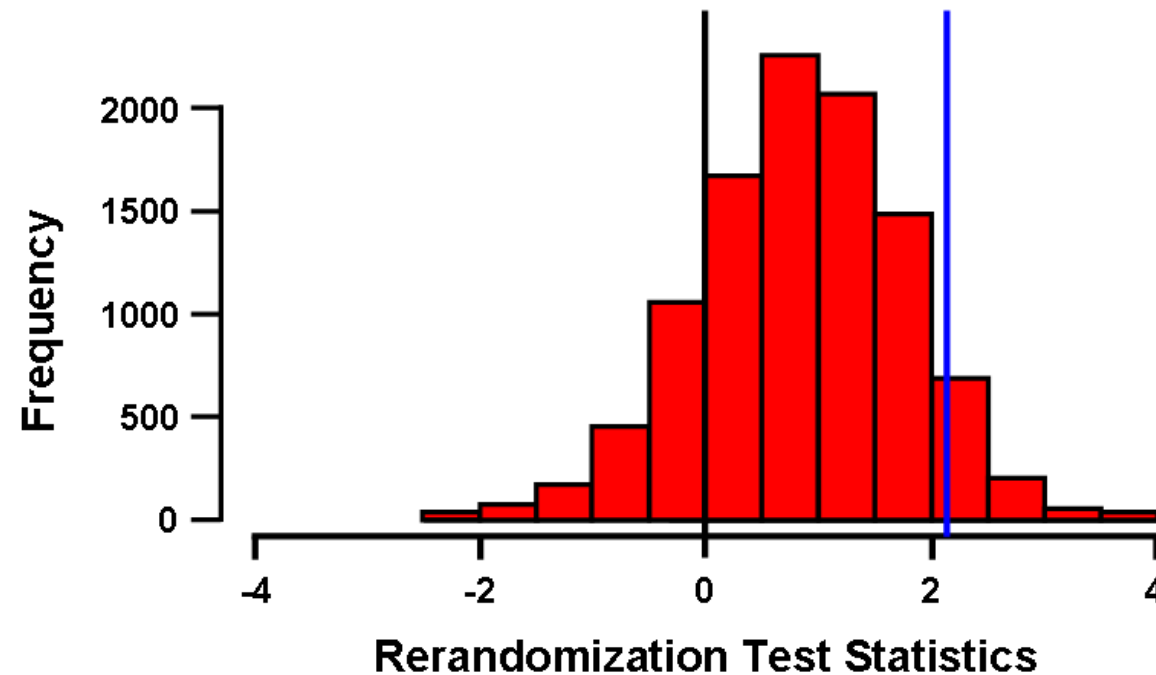
- October 2008 FDA Advisory Committee meeting on Pompe disease
 - Very rare, debilitating neuromuscular disease
 - Infant onset, juvenile onset, adult onset
 - Infant onset is most deadly, adult onset is still bad
 - Patients often progress to wheelchair dependence, ventilator, and death

Eye-Opening Experience for Minimization

- Genzyme conducted Late Onset Treatment Study (LOTS)
 - 90 patients with late onset Pompe disease
 - Primary outcome: 6 minute walk test
 - 2:1 allocation to drug/placebo using CAR
 - Site
 - BL 6 minute walk ($\leq 300\text{m}$, $> 300\text{m}$)
 - Forced vital capacity ($\leq 55\%$ pred., $> 55\%$ pred.)
- One analysis: randomization test

Problem with Application of Rerandomization Test in Analysis of 6MWT

- Distribution of 6MWT ANCOVA test statistics



ANCOVA $p=.035$

Randomization $p=.06$

Eye-Opening Experience for Minimization

- Notice the mean of randomization distribution is NOT 0
 - It is 0 for standard randomization methods
- The nonzero mean tells us that usual analysis methods might be prone to confounding if there are temporal trends
- A randomization test protects against the confounding

Eye-Opening Experience for Minimization

- For more details on LOTS trial, see Van der Ploeg et al. (2010)
- For more details about statistical problems minimization caused, see: Proschan, Brittain, and Kammerman (2011).
- For the fix for CAR with unequal allocation, see Kuznetsova and Tymofyeyev (2012).

One Question Quiz

- True or false: A t-test allows generalization to the population, but a randomization test applies only to participants in your trial

- Hint



gg125811848 www.gograph.com

Explanation of Quiz Answer

- The statement is false because generalization is **always** subjective, regardless of what statistical test is used
- With large sample sizes, the t-test and randomization test are nearly identical, so how could only one of them be generalizeable?

SUMMARY

- Randomization provides a basis for analysis called the analyze as you randomize principle
 - Fix data at observed values and re-generate treatment labels to get the randomization distribution of test statistic
- Easy to explain to nonstatisticians
- For large sample sizes and **standard randomization methods**, it is virtually the same as usual t-test or test of proportions

SUMMARY

- Provides a valid test with essentially no assumptions
- Caveat: it tests a strong null hypothesis of no effect of intervention on anyone
 - The observed data would have happened irrespective of the randomizations
- Can be used with any randomization, but don't use just *any* randomization
 - Avoid response-adaptive randomization

References

- Community intervention trial for smoking cessation (COMMIT) (1995). *American Journal of Public Health* **85**, 193–200.
- Fisher (1935). *The Design of Experiments* (5th ed.). Hafner Publishing, New York.
- Kuznetsova and Tymofyeyev, (2012). *Statistics in Medicine* 31, 701–723.
- Proschan, Brittain, and Kammerman (2011). *Biometrics* **67**, 1135–1141.
- Proschan and Evans (2020). *Clinical Infectious Diseases*, **71**, 3002–3004.
- Van der Ploeg et al. (2010). *NEJM* **362**, 1396–1406, 2010).