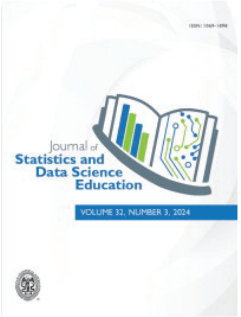


Taylor & Francis
Taylor & Francis Group

Journal of Statistics and Data Science Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ujse21

Simulation-Based Inference: Random Sampling vs. Random Assignment? What Instructors Should Know

Beth Chance, Karen McGaughey, Sophia Chung, Alex Goodman, Soma Roy & Nathan Tintle

To cite this article: Beth Chance, Karen McGaughey, Sophia Chung, Alex Goodman, Soma Roy & Nathan Tintle (22 May 2024): Simulation-Based Inference: Random Sampling vs. Random Assignment? What Instructors Should Know, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2024.2333736

To link to this article: https://doi.org/10.1080/26939169.2024.2333736

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2024, VOL. 00, NO. 0, 1–10
https://doi.org/10.1080/26939169.2024.2333736

Taylor & Francis
Taylor & Francis Group

Teaching Simulation-Based Inference: Random Sampling vs. Random Assignment? What Instructors Should Know

Beth Chance[a], Karen McGaughey[a], Sophia Chung[a], Alex Goodman[a], Soma Roy[a], and Nathan Tintle[b,c]

[a]Department of Statistics, California Polytechnic State University, San Luis Obispo, CA; [b]Superior Statistical Research, Sioux Center, IA; [c]University of Illinois – Chicago, Chicago, IL

ABSTRACT

"Simulation-based inference" is often considered a pedagogical strategy for helping students develop inferential reasoning, for example, giving them a visual and concrete reference for deciding whether the observed statistic is unlikely to happen by chance alone when the null hypothesis is true. In this article, we highlight for teachers some implications of different simulation strategies when analyzing two variables. In particular, does it matter whether the simulation models random sampling or random assignment? We present examples from comparing two means and simple linear regression, highlighting the impact on the standard deviation of the null distribution. We also highlight some possible extensions that simulation-based inference easily allows. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2023
Accepted February 2024

KEYWORDS
Permutation test; Randomization distribution; Standard error; Theory-based inference

1. Introduction

The term "Simulation-based inference" (SBI) has come to indicate curricula that focus on using tactile and computer-based simulations before, or instead of, "normal-based" and "theory-based" approaches, to help students understand the process of statistical inference. For example, for a one-proportion test of significance students can be asked to examine a distribution of "could-have been" values of the statistic, obtained from a process that models the null hypothesis and the randomness in the study design. For a 50/50 process, a tactile simulation could have every student in class toss a coin *n* times, and then crowd-source their simulated sample proportions of heads in a dotplot on the board to see where the observed sample proportion from the study falls in this distribution. This method for assessing the strength of evidence against the null hypothesis is arguably more intuitive than using the binomial or normal distributions alone (Cobb 2007; Rossman and Chance 2014).

With two variables, the question of how to best design the simulation is debatable. For example, with two categorical variables, one could design the simulation to model random sampling by taking random samples from two different populations with the same population proportion. Or one could design the simulation to model random assignment by shuffling the response outcomes (e.g., writing "success" and "failure" on index cards) and reassigning them to the explanatory variable groups. An advantage to using the second approach is that reshuffling the response values models the process of "breaking the association" between the explanatory and response variables which provides students with a concrete illustration of what could happen (for any choice of statistic) "by chance alone." This method of reshuffling can be applied to all two-variable situations in the introductory course (e.g., testing two proportions, two means, simple linear regression).

So what's the problem with using random shuffling? Nothing—If the study data arise from a randomized experiment, the simulation helps reinforce the role of the study design on the analysis and scope of conclusions. However, if the study is not a randomized experiment but does involve random sampling, then the first simulation approach is more appropriate. The choice of simulation model used can impact the analysis, primarily through estimation of the standard error of the statistic. Our goal in this article is to explore which types of studies/datasets are most impacted by the way the simulation is carried out, with an eye toward implications on teaching practice. The intention of this article is not to debate which simulation approach is more valid in practice (e.g., Lock Morgan 2017) or even which simulation approach is more intuitive for students (e.g., Chance et al. 2022). Rather, we hope to help instructors understand the differences in the simulation strategies to better inform their own decisions of how to adapt a simulation approach for their classes and to better respond to student questions that may arise.

2. Example 1: Comparing Group Means

2.1. Dung Beetle Study

An example in *Intermediate Statistical Investigations* (see Appendix) examines whether a view of the night sky influences

CONTACT Beth Chance bchance@calpoly.edu Department of Statistics, California Polytechnic State University, San Luis Obispo, CA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/ujse.

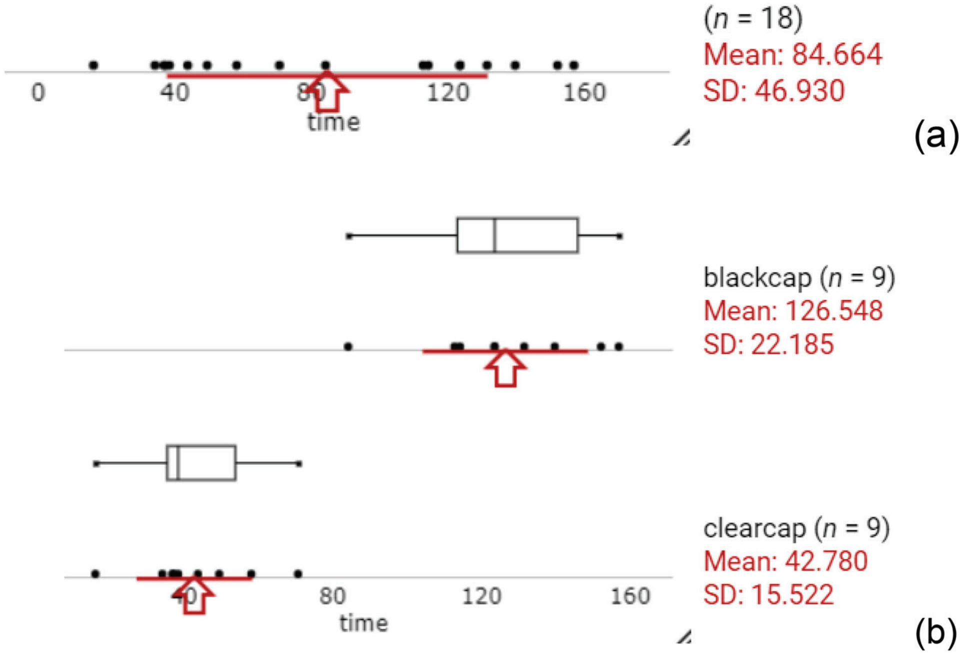


Figure 1. Distributions of (a) travel times of 18 beetles overall and (b) separated into two treatment groups. Notice the large difference between the overall variability (SD = 46.93 sec) and the within group variability (22.19 and 15.52 sec).

the speed at which dung beetles can navigate a dung ball to the edge of a circular platform (away from center). Nocturnal African dung beetles (*Scarabaeus satyrus*) were randomly assigned to wear either a clear cap or a black cap, the latter obscuring their view of a moonless but starry night sky. Consider the output in Figure 1 (slightly modified from the actual research study to have equal sample sizes for the initial discussion):

There is clearly a large difference in the mean time to reach the platform edge between the two experimental groups ($\bar{y}_{black} - \bar{y}_{clear} = 83.8$ sec). In fact, a *p*-value is really not necessary here as there is no overlap in the distributions of the times between the two groups and the observed group split is the most extreme possible out of the C(18,9) = 48,620 possible assignments. Regardless, we find this study an engaging context for discussing a randomization test: how large a difference in the experimental group means would we expect to see if these 18 times were randomly redistributed to two groups of 9.

To illustrate this process, we begin with a tactile simulation:

- Each student (or pair of students) is given 18 index cards to match the 18 beetles in the study.
- Students write the 18 observed responses on the 18 cards. We tell students this step models the idea that the beetle travel time is “fixed” and not impacted by which experimental group the beetles are assigned to. In other words, assuming the null hypothesis of “no experimental effect” is true.
- Students shuffle their 18 cards and deal them out to two groups of 9, computing $\bar{y}_{black} - \bar{y}_{clear}$ for their random reassignment.
- Each student (pair) adds their shuffled difference in means to a dotplot on the board in the classroom, creating a null distribution of could-have-been differences in means just by random chance alone.
- Students then determine how unusual it is to find a difference in means as extreme as the observed difference in means of 83.8 sec.

We feel this activity, which results in each student (pair) having a dot in the null distribution, is very helpful in students’ understanding of the simulation process and what the *p*-value measures. After the tactile simulation, we then have students use an applet to generate thousands of random reshuffles of the data. Below, we show the results of 10,000 shuffles from the Rossman/Chance Comparing Groups applet.

Students see that 83.8 rarely occurs in 10,000 random shuffles, if at all, yielding a very small *p*-value. For more information and advice on teaching with simulation-based inference see for example Case and Jacobbe (2018), Chance et al. (2022), Chance, Chung, and Tintle (2022); and the “SBI blog” (*https://www.causeweb.org/sbi/*).

In our introductory course, we then follow the randomization test analysis with a comparison to the *t*-test. We teach students to calculate the *t*-statistic using $SE\left(\bar{y}_1 - \bar{y}_2\right)$. One option for this computation uses the following formula to estimate the standard deviation of the difference in sample means when the population variances are unknown,

$$SE\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{22.185^2}{9} + \frac{15.522^2}{9}} = 9.025 \tag{1}$$

But wait a minute, this is nowhere near the simulated randomization distribution’s standard deviation (e.g., SD = 22.005 in Figure 2)!

Because the process of random shuffling pools all of the data values together, assuming they arise from the same population distribution, and then redistributes them, perhaps we should

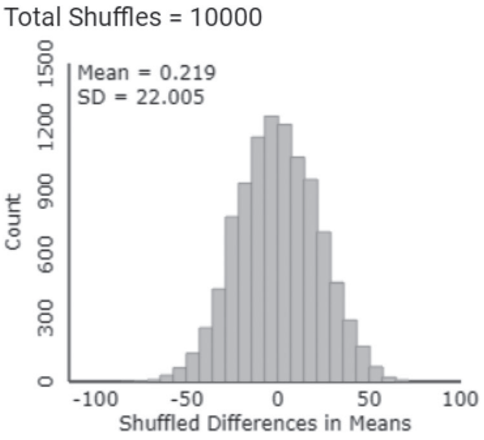


Figure 2. Simulated randomization distribution for the difference in group means from shuffling the 18 times back to two groups (Dung Beetles data). The standard deviation of the difference in means from these 10,000 shuffles is 22.005 sec.

use the pooled standard error instead to match the simulation results?

$$\text{Pooled } SE\left(\bar{y}_1 - \bar{y}_2\right) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 19.15\sqrt{\frac{2}{9}} = 9.027 \quad (2)$$

But this value is practically the same as the unpooled value because the two groups have similar standard deviations and equal sample sizes. Here is where we need to remember that both of these formulas assume *independent random samples from large populations.*

2.2. Random Sampling

The output in Figure 3 shows the simulation results when taking 10,000 independent random samples from two populations (salaries of the Western conference NBA players and salaries of Eastern conference NBA players for the 2020–2021 season, downloaded from *http://www.espn.com/nba/salaries* July 2021) using the Rossman/Chance Sampling from Two Populations applet.

The original standard deviation formula (see (1)), now using the population standard deviations, works very well in predicting the standard deviation of the sampling distribution (compared to 4.09 in Figure 3):

$$SD\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{9.20^2}{9} + \frac{8.45^2}{9}} = 4.16 \quad (3)$$

Note that the population standard deviations (9.20 and 8.45) are found by dividing by N (263 and 301) rather than N-1 because we know μ for the entire population in each conference. We could improve this SD calculation by considering the population sizes and using a finite population correction factor for each sample mean:

$$SD\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{\sigma_1^2(N_1 - n_1)}{n_1\,(N_1 - 1)} + \frac{\sigma_2^2\,(N_2 - n_2)}{n_2\,(N_2 - 1)}} \quad (4)$$

$$= \sqrt{\frac{9.20^2(263 - 9)}{9(263 - 1)} + \frac{8.45^2(301 - 9)}{9(301 - 1)}} = 4.10$$

The difference between these two standard deviations, 4.16 and 4.10, is not large because the population sizes are more than 10 times the sample sizes, satisfying the "10% rule" so that the population correction factors are approximately equal to one.

2.3. Exact Randomization Distribution

The simulation-based approach approximates the permutation test that examines the distribution from all possible random shuffles of the observed results. With a categorical response variable, the hypergeometric distribution models the distribution of the number of successes, out of the M total successes in the sample, that end up among the n observations in "Group A." The standard deviation of the hypergeometric distribution,

$$\sqrt{\frac{N - n}{(N - 1)} \times n \times \frac{M}{N} \times \frac{N - M}{N}} \quad (5)$$

illustrates the source of the "finite population correction factor" $(N - n)/(N - 1)$ that accounts for the lack of dependence when we select n observations from N.

In fact, population size is something we should also consider with the randomization distribution in Example 1. (With quantitative data, we need to know more than the number of successes in group A, we need to also the resulting difference in means for each possible random assignment.) Because we always use the same 18 beetles, they are the population. By splitting that population in half for our two samples (essentially sampling 9 beetles from the population of 18 beetles for group 1), we are clearly violating the 10% rule. But even more important, by simply putting the other 9 beetles into group 2, we are also violating the "independent samples" rule. For example, if the 9 fastest beetles are randomly assigned to the clear cap, then we know the 9 slowest beetles went to the black cap and the mean of that group is predetermined. In particular, the correlation coefficient between the two groups means is −1. With correlated samples, we should use the formula below for the standard deviation of the difference in means:

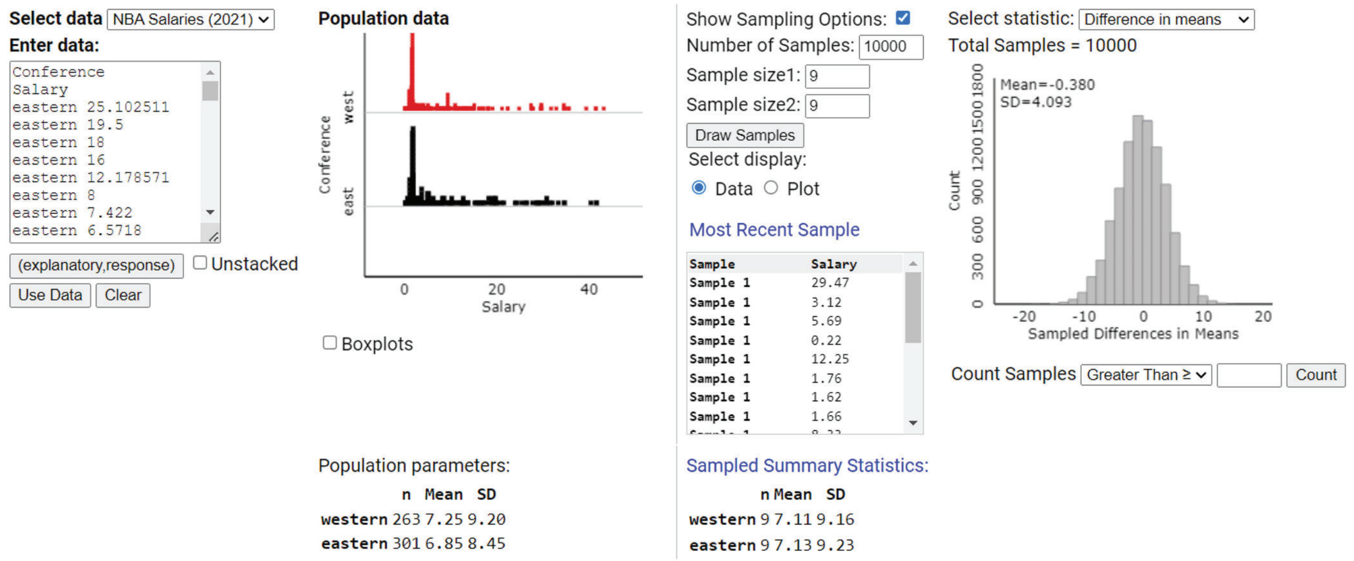
$$SD\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2 \times \text{cov}\left(\bar{Y}_1, \bar{Y}_2\right)} \quad (6)$$

So now applying

- the relationship between correlation and covariance,
- the finite population correction factor, and
- the distinction between "population" and "sample" standard deviations

we can determine the expected standard deviation for the difference in group means for the randomization distribution:

$$SD\left(\bar{y}_1 - \bar{y}_2\right) = \sqrt{\frac{\sigma_1^2(N_1 - n_1)}{n_1\,(N_1 - 1)} + \frac{\sigma_2^2\,(N_2 - n_2)}{n_2\,(N_2 - 1)} - 2 \times (-1) \times \sqrt{\frac{\sigma_1^2\,(N_1 - n_1)}{n_1\,(N_1 - 1)}\,\frac{\sigma_2^2\,(N_2 - n_2)}{n_2\,(N_2 - 1)}}} \quad (7)$$



Select data: NBA Salaries (2021)

Enter data:

```
Conference
Salary
eastern 25.102511
eastern 19.5
eastern 18
eastern 16
eastern 12.178571
eastern 8
eastern 7.422
eastern 6.5718
```

☐ Unstacked

☐ Boxplots

Population parameters:

```
        n  Mean  SD
western 263 7.25 9.20
eastern 301 6.85 8.45
```

Show Sampling Options: ☑

Number of Samples: 10000

Sample size1: 9

Sample size2: 9

Select display: ● Data ○ Plot

Most Recent Sample

```
Sample      Salary
Sample 1    29.47
Sample 1    3.12
Sample 1    5.69
Sample 1    0.22
Sample 1    12.25
Sample 1    1.76
Sample 1    1.62
Sample 1    1.66
```

Sampled Summary Statistics:

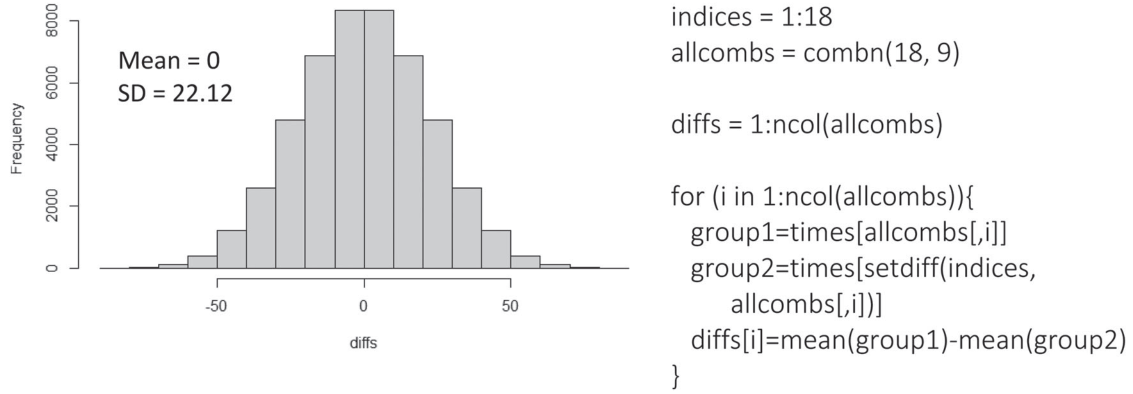
```
        n Mean  SD
western 9 7.11 9.16
eastern 9 7.13 9.23
```

Select statistic: Difference in means

Total Samples = 10000

Count Samples: Greater Than ≥

Figure 3. Population distributions and simulated sampling distribution of difference in sample means for $n_1 = n_2 = 9$. The standard deviation of the difference in means is 4.09 (thousand dollars).



```
indices = 1:18
allcombs = combn(18, 9)

diffs = 1:ncol(allcombs)

for (i in 1:ncol(allcombs)){
  group1=times[allcombs[,i]]
  group2=times[setdiff(indices,
      allcombs[,i])]
  diffs[i]=mean(group1)-mean(group2)
}
```

Figure 4. Exact randomization distribution for Dung Beetle data. The standard deviation of the statistic is 22.12 sec.

A similar derivation can be found in Chapter 27, footnote 11 of Freedman et al. (2007).

With a bit of algebra, and noting that $N_1 = N_2$ and $n_1 = n_2 = \frac{N_1}{2}$ and $\sigma_1 = \sigma_2 = \sigma$ and $s = \sigma\sqrt{\frac{N}{N-1}}$, this simplifies to

$$SD\left(\bar{y}_1 - \bar{y}_2\right) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}!}, \tag{8}$$

where the exclamation point stands for amazement, not factorial. For the Dung Beetle data,

$$SD\left(\bar{y}_1 - \bar{y}_2\right) = 46.93\sqrt{\frac{2}{9}} = 22.12, \tag{9}$$

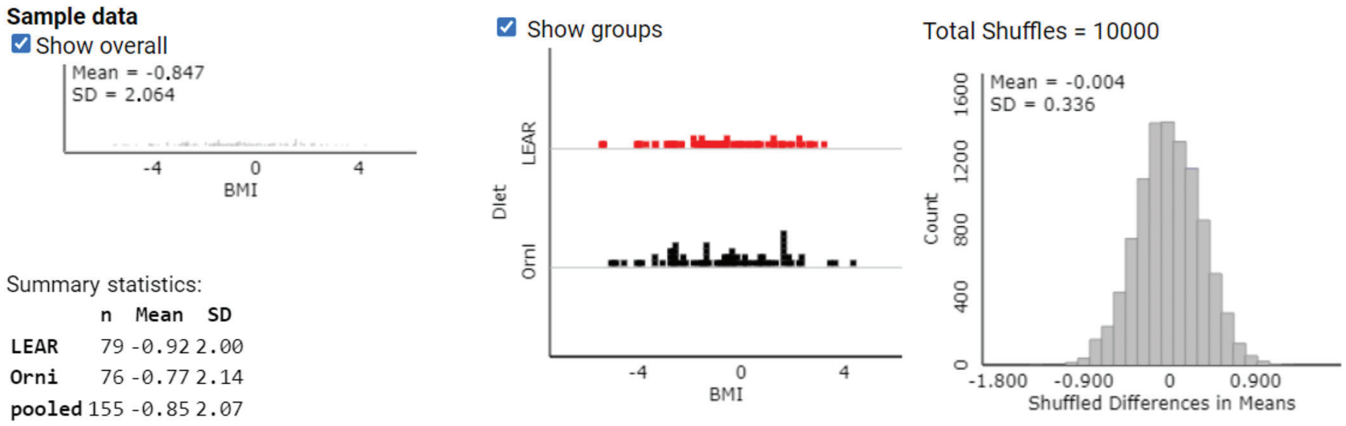
much more in agreement with the standard deviation obtained in the simulation, 22.005.

How do we know this is the "right" answer? With a permutation test, we can find the *exact randomization* distribution, that is, the difference in means for all C(18,9) = 48,620 possible random assignments. The histogram in Figure 4 (produced using the "combn()" function in R, from the combinate package, assuming "times" contains the response variable values) shows the distribution of all possible differences in group means applied to the Dung Beetle data. (See also Kaiser and Lacy 2009.)

In the exact randomization distribution for the Dung Beetle data, the standard deviation of the difference in means (22.12) is what we calculated taking into account the finite populations and the correlation between the data values split between the two groups. So to calculate the standard deviation of the randomization distribution (which assumes no genuine difference between the two groups), we should use the simplified (8) which uses the standard deviation of the response overall, rather than the traditional SE formula, (1), which uses the within group standard deviations and often results in a much smaller value.

2.4. Assuming the Null Hypothesis is True

In the Dung Beetle data, the evidence against the null hypothesis was very strong. That is, the variability in the times explained by the group membership (black vs. clear cap) was extremely large. This leads to a very large difference between the value of the traditional SE for the difference in means, (1), and the randomization SE for the difference in means. When the evidence against the null hypothesis is much weaker, the traditional SE (1) will not look as different from the standard deviation of the randomization distribution. Figure 5 shows output from a study comparing the mean weight loss on the LEARN diet to



Sample data

☑ Show overall

Mean = -0.847
SD = 2.064

Summary statistics:

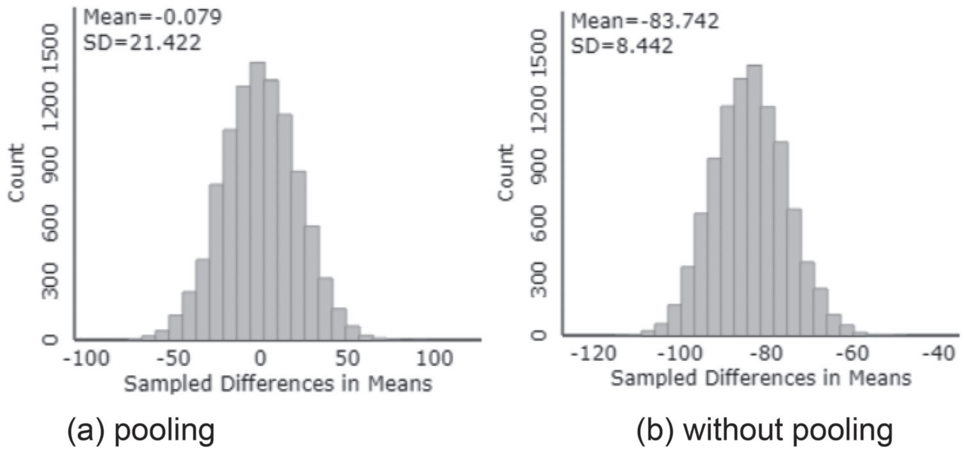
	n	Mean	SD
LEAR	79	-0.92	2.00
Orni	76	-0.77	2.14
pooled	155	-0.85	2.07

☑ Show groups

Total Shuffles = 10000

Mean = -0.004
SD = 0.336

Figure 5. Comparing changes in BMI between two diets after one year. The standard deviation of the statistic is 0.336 kg.



(a) pooling

(b) without pooling

Figure 6. Bootstrap distributions for the Dung Beetle data with and without pooling the groups together. The standard deviations of the difference in means are 21.422 sec with pooling and 8.442 sec without pooling.

that on the Ornish diet (data from Gardner et al. 2007). Because the overall standard deviation (2.064 kg) is similar to the pooled standard deviation (2.07 kg) due to the large amount of overlap in the two experimental groups (small R^2), the standard deviation of the randomization distribution using (8) is similar to the traditional SE using (1): $\sqrt{\frac{2.00^2}{79}+\frac{2.14^2}{76}} \approx 0.333$ versus $2.064\sqrt{\frac{1}{79}+\frac{1}{76}} \approx 0.332$

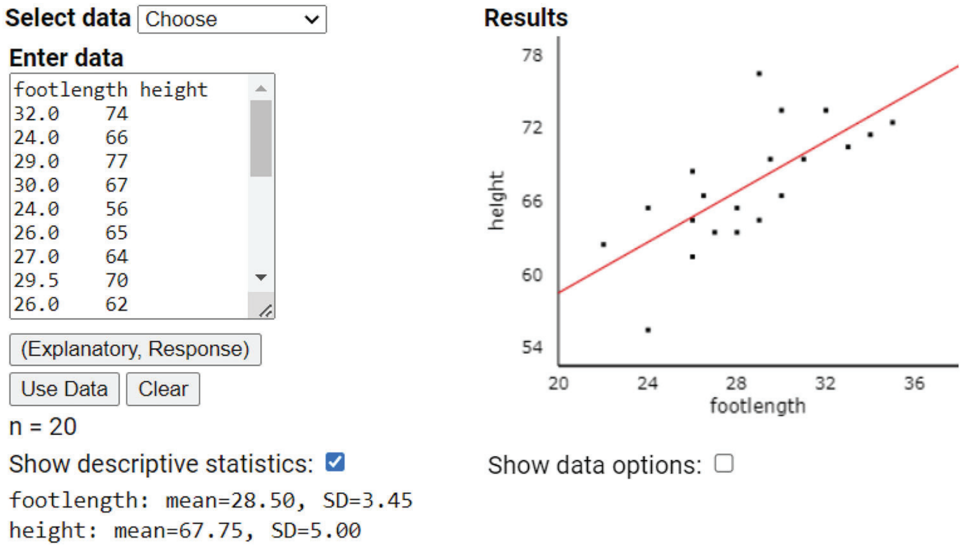
The randomization distribution, assuming the null hypothesis is true (so all variation is natural variation), can result in a much larger SD of the differences in means than if the natural variation only consists of within group variation (i.e., variation in the residuals). In other words, this difference in the values from these two methods of estimating the standard deviation of the difference in means is larger the stronger the underlying relationship between the response and explanatory variables.

The overall idea makes some intuitive sense: the randomization distribution assumes the null hypothesis is true. If the null hypothesis is actually not true, the overall standard deviation of the data is much larger than the within group standard deviations, leading to a much larger standard deviation of the difference in means in the randomization distribution than the traditional SE formula produces. A similar distinction is seen when comparing two proportions where the formula for $SE\left(\hat{p}_1 - \hat{p}_2\right)$ differs depending on whether we assume the null hypothesis is true, that is, "pooled" (in the test statistic) or "unpooled" (used for confidence intervals). The same distinction also arises with bootstrapping—do we pool the groups together first or resample within each group? In the latter case, the individual sample means will have much less sample-to-sample variation (e.g., a time of 120 sec would *only* be observable in the black cap group rather than being equally likely to be in either experimental group) and will result in smaller variability in the difference in means statistic. For the Dung Beetles data using the Comparing Groups—Bootstrapping applet, Figure 6 shows that if we pool the samples together first, the simulated standard deviation is around 21.4 (similar to random assignment), but if we don't pool the samples together first, the simulated standard deviation is around 8.44 (more similar to the traditional formula (1)).

3. Example 2: Simple Linear Regression

Generalizing from comparing groups on a quantitative response to a linear regression model, the same distinction appears. Figure 7 shows a dataset of heights (inches) and foot lengths (cm) for a sample of 20 college students (default data in the Rossman/Chance Two Quantitative variable applet, see Figure 7).

Figure 8 shows a randomization distribution of slopes using the applet: each y-value is randomly reassigned to the observed x-values. (The lines all pass through $(\bar{x}, \bar{y})$ because neither mean



Select data Choose

Enter data

```
footlength height
32.0    74
24.0    66
29.0    77
30.0    67
24.0    56
26.0    65
27.0    64
29.5    70
26.0    62
```

(Explanatory, Response)

Use Data Clear

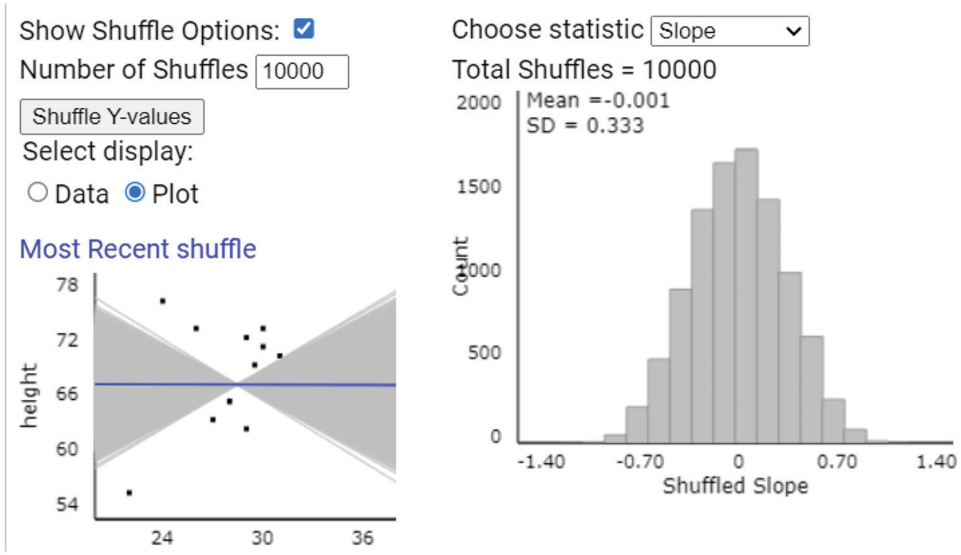
n = 20

Show descriptive statistics: ☑

```
footlength: mean=28.50, SD=3.45
height: mean=67.75, SD=5.00
```

Show data options: □

Figure 7. Sample data for heights (inches) and footlengths (cm) for 20 college students. Regression line: *predicted height* = 38.3 + 1.03*footlength*, *residual SE* = 3.61 inches.



Show Shuffle Options: ☑

Number of Shuffles 10000

Shuffle Y-values

Select display:

○ Data ● Plot

Choose statistic Slope

Total Shuffles = 10000

Mean =-0.001
SD = 0.333

Figure 8. Simulated regression lines from reshuffling heights to foot sizes and simulated randomization distribution of slopes. The standard deviation of the slope is 0.333 in.

changes when we shuffle the ordering of the values for the response variable.)

After viewing the randomization distribution, we want students to be able to interpret the standard deviation of the shuffled slopes, for example, 0.333 in Figure 8, and we then have them compare this value to the traditional regression output:

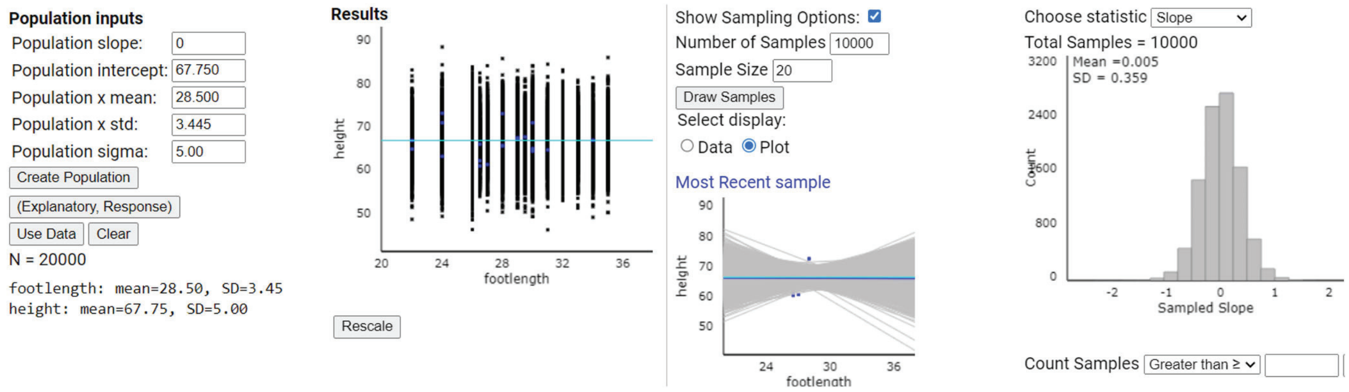
Term	Coeff	SE	*t*-stat	*p*-value
Intercept	38.30	6.905	5.55	<0.0001
footlength	1.03	0.241	4.29	0.0004

But again, we were initially troubled that the usual formula for the standard error of the slope,

$$SE\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{\sqrt{(n-1) \times Var\,(X)}} = \frac{3.61}{\sqrt{19 \times 3.45}} = 0.241 \tag{10}$$

produces a value that is not that similar to the randomization result.

If we want to model random sampling instead, one approach is to construct a bivariate population to sample from, with similar characteristics as the sample. To pick a value for "Population x std," we can use the sample standard deviation of foot lengths (3.45 cm in Figure 7), but what about the standard deviation of the heights? If there is no association between the two variables, then we would expect similar variation about the horizontal regression line as we saw in the sample (e.g., $s_y = 5.00$ in). But if there is an association, then the unexplained variation about the regression line should be more similar to the residual standard error (3.61 in Figure 7). For example, we wouldn't want to assign a 24 cm foot length to *any height* between 56 and 77 (the min and max in the dataset) like random shuffling would, instead a 24-cm foot length should correspond to lower height values ($\approx \hat{y} \pm 2 \times 3.61$). Predictably, the choice of "Population sigma" has a large impact on the variation of the slopes. In Figure 9, when the population sigma is set to 5, the standard deviation in the simulated slopes is 0.359. In Figure 10, when the population



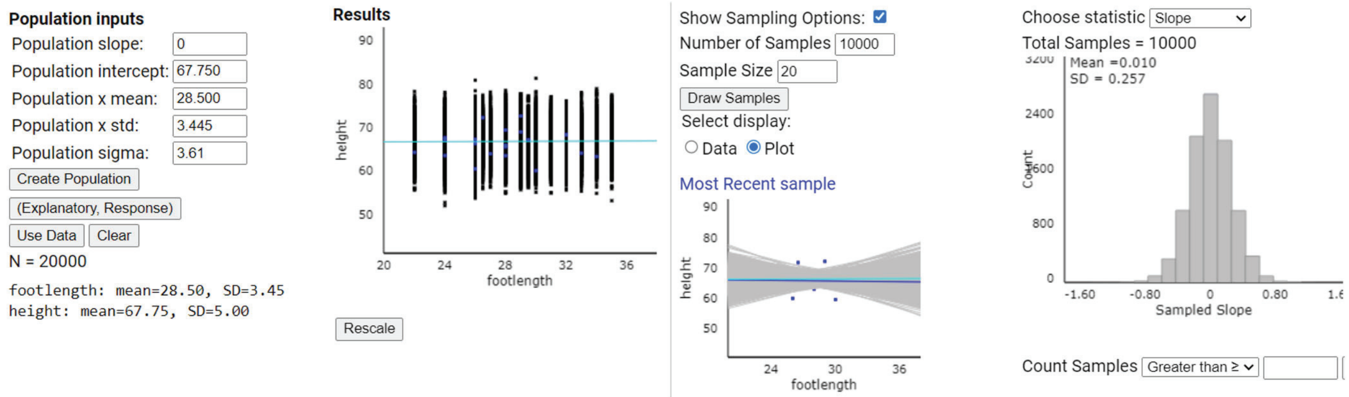
Population inputs
Population slope: 0
Population intercept: 67.750
Population x mean: 28.500
Population x std: 3.445
Population sigma: 5.00
Create Population
(Explanatory, Response)
Use Data Clear
N = 20000

```
footlength: mean=28.50, SD=3.45
height: mean=67.75, SD=5.00
```

Show Sampling Options: ☑
Number of Samples 10000
Sample Size 20
Draw Samples
Select display:
○ Data ● Plot

Choose statistic Slope
Total Samples = 10000
Mean =0.005
SD = 0.359
Count Samples Greater than ≥

Figure 9. Constructed hypothetical population with no association but matching the descriptive statistics for the original sample, using $\sigma = 5$. Resulting regression lines and sampling distribution of slopes are displayed to the right. The standard deviation of the slope is 0.359 in.



Population inputs
Population slope: 0
Population intercept: 67.750
Population x mean: 28.500
Population x std: 3.445
Population sigma: 3.61
Create Population
(Explanatory, Response)
Use Data Clear
N = 20000

```
footlength: mean=28.50, SD=3.45
height: mean=67.75, SD=5.00
```

Show Sampling Options: ☑
Number of Samples 10000
Sample Size 20
Draw Samples
Select display:
○ Data ● Plot

Choose statistic Slope
Total Samples = 10000
Mean =0.010
SD = 0.257
Count Samples Greater than ≥

Figure 10. Constructed hypothetical population with no association but matching the descriptive statistics for the original sample, using $\sigma = 3.61$. Resulting regression lines and sampling distribution of slopes are displayed to the right. The standard deviation of the slope is 0.257 in.

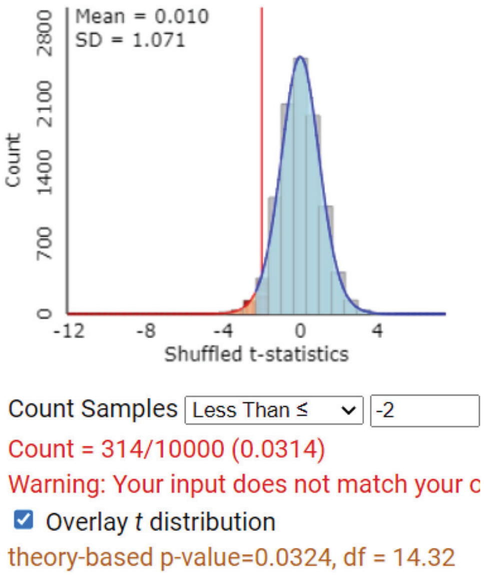
sigma is set to 3.61, the standard deviation in the simulated slopes is 0.257. Note: In the simulations in Figures 9 and 10, we take the x-values to be fixed at the observed values, yet another simulation-design decision.

One interesting observation is that the regression lines from random sampling do not all pass through the same point because the $\bar{x}$ and $\bar{y}$ values change with each sample. As we would expect, the smaller we make the population sigma (unexplained variation about the population regression line), the smaller the standard deviation in the sample slopes. When we use the observed residual standard error, the simulation result is much more similar to the usual formula ((10), 0.241).

If the residual standard error is closer to the standard deviation of the responses themselves (smaller R^2), the SE for the slope will be more similar between these approaches. The moral being, don't expect the regression output SE to match the simulation results unless you use a dataset with a weaker association.

4. Extensions

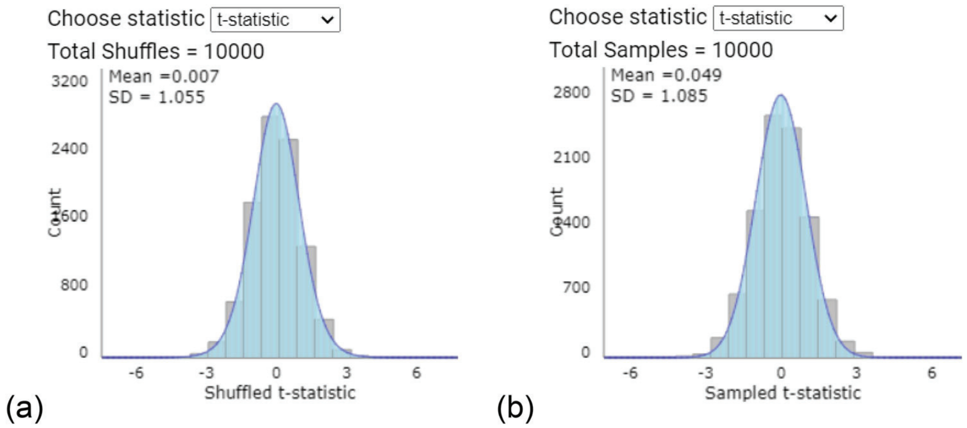
In our introductory algebra-based statistics courses, we don't have this debate with our students on which simulation approach to use. Instead, we emphasize that the approaches make slightly different assumptions, including whether or not the null hypothesis is true. We do briefly discuss with students the distinction between the simulated randomization distribution and the exact randomization distribution (showing them the code and output) to reinforce that using 1000 or 10,000 repetitions in the simulation is quite adequate for understanding the behavior of the distribution (see also Chance and Rossman 2023). We do emphasize these distinctions a bit



Mean = 0.010
SD = 1.071

Count Samples Less Than ≤ -2
Count = 314/10000 (0.0314)
Warning: Your input does not match your c
☑ Overlay *t* distribution
theory-based p-value=0.0324, df = 14.32

Figure 11. Comparison of randomization distribution of t-statistics using the "wrong" standard error and the theoretical t-distribution.


Choose statistic t-statistic
Total Shuffles = 10000
Mean =0.007
SD = 1.055

Choose statistic t-statistic
Total Samples = 10000
Mean =0.049
SD = 1.085

Figure 12. The same *t*-distribution can be used to predict (a) the randomization distribution and (b) the sampling distribution of the standardized regression slope.

more with our mathematics and statistics majors and in our intermediate statistics course, emphasizing why the randomization distributions tend to have larger standard deviations than formula (1) would predict. We have given the task of simplifying formula (7) as an optional assignment for our math/stat majors and a few did voluntarily take on the challenge, demonstrating curiosity in the ideas and different learning preferences. We also wonder whether showing students the impact of a nonzero covariance more often, and how that simplifies to the SD of the differences in a matched pairs design, would be beneficial. In an intermediate course, you could also make the connections with bootstrapping residuals. Below we describe a few other extensions that you could take with introductory students.

4.1. *t*-statistics

In Section 2 we focused on comparing the standard deviation of the difference in means depending on whether a simulation models random assignment versus random sampling. What about the randomization distribution of the *t*-statistic? It turns out that if we use the unpooled *t*-statistic for each shuffle (with the "wrong" standard error from (1) for the difference in means), this randomization distribution is still well-modeled by the *t*-distribution. When the simulated difference in means is smaller, the *t*-statistic will divide by a larger SE, but when the simulated difference in means is larger, the *t*-statistic will divide by a smaller SE. The resulting distribution of randomization *t*-statistics ends up behaving more like the heavy-tailed *t*-distribution than if we use the same (larger) SE (9) for all the differences. For the Dung Beetle data, Figure 11 compares the simulated and theoretical probabilities below $t = -2$. We can see that these two probabilities are quite similar. (In Figure 11 we used $t = -2$, rather than the observed *t*-statistic for the Dung Beetle data of $t = -9.28$ which gives a small *p*-value either way.) The moral is that we can use the same *t*-test for both random sampling and random assignment.

Similarly, Figure 12 shows that the theoretical *t*-distribution is a reasonable approximation for both the randomization distribution and the sampling distribution of regression slopes for the height/foot dataset in Example 2.

Another interesting observation from Figure 3 (NBA Salaries) is how bell-shaped and symmetric the sampling distribution is even though the population distributions are quite skewed and the sample sizes are small ($n = 9$). Using such visual simulations allows quick confirmation that the *t*-distribution works rather well when the population shapes are similar and the sample sizes are the same, but less well if the sample sizes differ, even when they are of more moderate sizes. Figure 13 shows the sampling distribution after changing the sample sizes to 15 and 30 and the resulting skewness in the distribution of *t*-statistics.

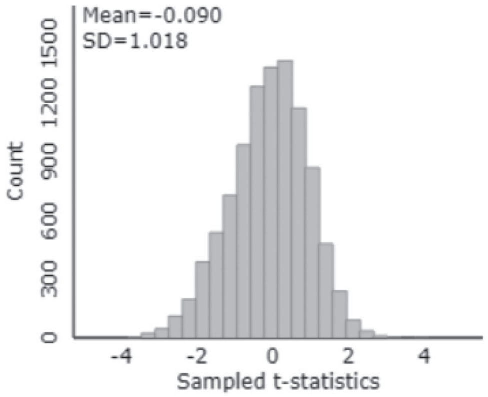


Figure 13. Simulated sampling distribution for NBA data using sample sizes of $n_1 = 15$ and $n_2 = 30$. The sample sizes are larger than for Figure 3 but there is more skewness in the distribution of t-statistics due to the unequal sample sizes.

4.2. Modeling an Alternative Hypothesis

Both random sampling and random assignment simulations also give students the ability to model the case when a null hypothesis is not true and then examine the resulting distribution of the statistic. For example, we can specify a hypothesized difference of 50 sec ($\mu_{clear} - \mu_{black} = -50$) and adjust for this hypothesis before random shuffling. Figure 14 illustrates the process of adding 50 to all times in the clear cap group (a) to (b), then mixing which values are shuffled to each group (b) to (c), akin to randomizing residuals, and then then subtracting 50 sec from the time for each beetle assigned to the clear group (c) to (d). See the supplementary materials for an animated gif dynamically visualizing this process from the Comparing Group applet.

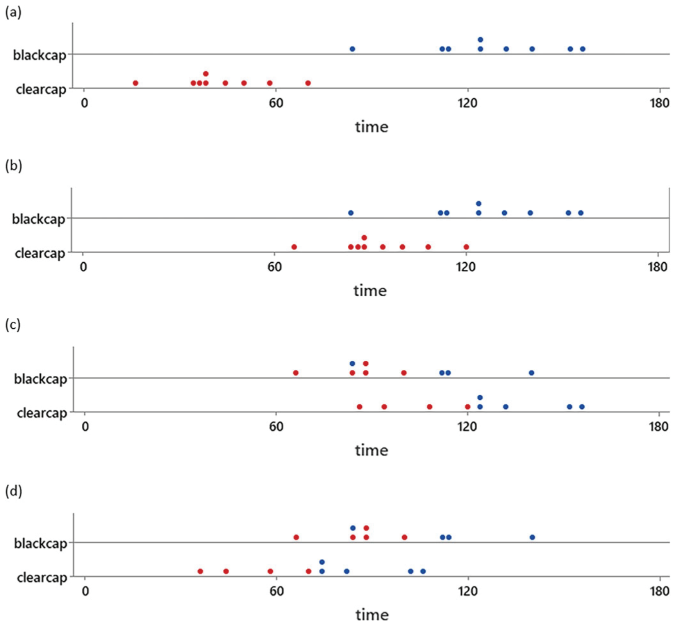
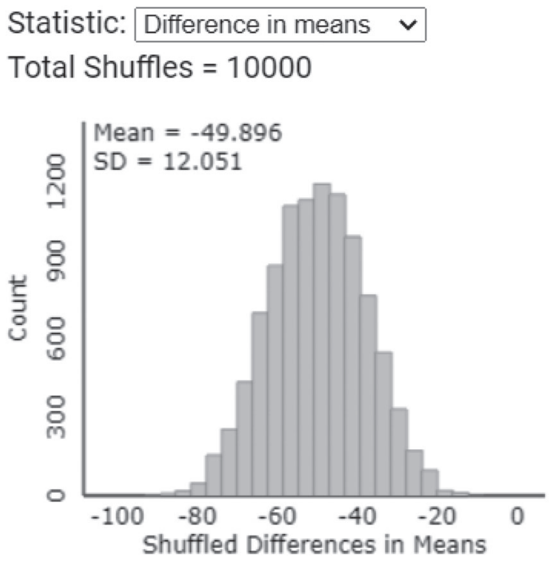


Figure 14. Sequence showing the adjustment for treatment effects before randomization. The difference in means is calculated each time, building up a randomization distribution matching the hypothesized effect.


Statistic: Difference in means

Total Shuffles = 10000

Mean = -49.896
SD = 12.051

Figure 15. A randomization distribution assuming $y = \mu + (-50 \textit{ if clear cap}) + \epsilon$, found by adjusting for the clear cap effect and then randomizing the residuals. The standard deviation of the difference in means is 12.051 sec.

The mean of the resulting randomization distribution in Figure 15 is now −50 as modeled, but also note the smaller SD of the randomization distribution as we have accounted for a large source of variation in the times, so the "unexplained variation" is more similar to the pooled within-group variability.

Simulating under alternative hypotheses is useful for exploring "non-central" distributions as well as power calculations. We do sometimes ask students to carry out this simulation in homework, asking them to explain what the applet is doing and to explain why the mean of the simulation distribution has changed and how that impacts the p-value (though we don't usually ask about the change in the standard deviation).

Students can also use the simulations to "invert the test" to develop an interval of plausible values for the underlying difference in treatment means and compare the results to the t-interval. Another possible extension is to have students use technology to examine the coverage of the t-interval procedure for different standard deviation formulas and different cases of the underlying treatment difference.

5. Summary

When using a theory-based approach for testing two population means, we typically use the unpooled t-statistic formula, regardless of whether the study involved random sampling or random assignment. Then we emphasize the point that students need to consider the random sampling versus random assignment distinction when drawing their final conclusions (e.g., generalizability, causation). When using simulation-based inference, changing the method of simulation depending on the type of study, can help reinforce this distinction for students.

However, if you use simulation-based approaches to introduce students to the reasoning of statistical significance when comparing groups (including in simple linear regression), keep in mind that random shuffling makes different assumptions than the traditional theory-based formulas for the standard error of the statistic (e.g., independence between groups, use of overall vs. within group variability). Because of this, the simulation-based standard error may not match that obtained from the traditional formula. Our advice in order to avoid students being distracted by this difference, is to start with a dataset with a weaker association between the response and explanatory variables. This means the SD of the statistic determined from the randomization process will be similar to the calculated value of the standard error of the statistic. This then allows students to provide a reasonable interpretation of the standard error by describing the simulation process and the "variation due to random chance alone." Then, once you switch to the standardized statistic, the distinction between random sampling and random assignment no longer matters in the analysis (but is still critical to determining appropriate conclusions for the study).

Appendix

A.1. Introduction to Exploration 1.2 in Tintle et al. (2020)

Dacke, Baird, Byrne, Scholtz, and Warrant ("Dung Beetles Use the Milky Way for Orientation," *Current Biology*, 23, 2013) report on several experiments they ran to document whether nocturnal African dung beetles (*Scarabaeus satyrus*) use stars in the night sky to navigate. In one of their studies, beetles were placed on top of a dung ball at the center of a circular wooden platform (10 cm in diameter) and the researchers timed how long it took each beetle to reach the edge of the platform (another way of determining how straight a path was taken). Some of the beetles were given a small, black cardboard "cap" which obscured their view of the sky (up) but not of the edge of the platform (out), while others were given a transparent cap. On a moonless, starry night beetles wearing the transparent cap took an average of 40.1 sec to reach the edge, compared to an average time of 124.5 sec for beetles wearing the black cardboard cap.

Supplementary Materials

In the supplement, we provide an animated gif illustrating the visualization possible in the Comparing Groups applet (*https://www.rossmanchance.*

com/applets/2021/anovashuffle/AnovaShuffle.htm) to show how the randomization simulation reflects the non-zero hypothesized difference in the underlying treatment means.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by National Science Foundation (grant number: NSF-DUE-2235355).

References

Case, C., and Jacobbe, T. (2018), "A Framework to Characterize Student Difficulties in Learning Inference from a Simulation-Based Approach," *Statistics Education Research Journal*, 17, 9–29.

Chance, B., and Rossman, A. (2023), "Investigating Statistical Concepts, Applications, and Methods," available at *http://www.rossmanchance.com/iscam3/*

Chance, B., Chung, S., and Tintle, N. (2022), "Use of Small-Scale Classroom Experiments to Inform Statistics (SBI) Pedagogy in Tertiary Classrooms," in *Proceedings of 11th International Conference on Teaching Statistics (ICOTS-11)*. Available at *http://iase-web.org/icots/11/proceedings/pdfs/ICOTS11_163_CHANCE.pdf?1669865530*

Chance, B., Tintle, N., Reynolds, S., Patel, A., Chan, K., and Leader, S. (2022), "Student Performance in Curricula Centered on Simulation-Based Inference," *Statistics Education Research Journal*, 21, 4. *https://iase-web.org/ojs/SERJ/article/view/6*

Cobb, G. (2007), "The Introductory Statistics Course: A Ptolemaic Curriculum?," *Technology Innovations in Statistics Education*, 1. DOI:10.5070/T511000028

Freedman, D., Pisani, R., and Purves, R. (2007). *Statistics* (4th ed.), New York: W. W. Norton & Company.

Gardner, C. D., Kiazand, A., Alhassan, S., Kim, S., Stafford, R. S., Balise, R. R., Kraemer, H. C., and King, A. C. (2007), "Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors among Overweight Premenopausal Women: The a to Z Weight Loss Study: A Randomized Trial," *JAMA*, 297, 969–977. DOI:10.1001/jama.297.9.969.

Kaiser, J., and Lacy, M. (2009), "A General-Purpose Method for Two-Group Randomization Test," *The Stata Journal: Promoting Communications on Statistics and Stata*, 9, 70–85.

Lock Morgan, K. (2017), "Reallocating and Resampling: A Comparison for Inference," available at *https://doi.org/10.48550/arXiv.1708.02102*

Rossman, A., and Chance, B. (2014), "Using Simulation-Based Inference for Learning Introductory Statistics," *WIREs Computational Statistics*, 6, 211–221. DOI:10.1002/wics.1302

Tintle, N., Chance, B., Cobb, G., Rossman, A., Roy, S., Swanson, T., and VanderStoep, J. (2020), *Introduction to Statistical Investigations* (2nd ed.), New York: Wiley.