On the justification of sampling-based statistics in randomised experiments

Jan Vanhove*

September 2026

Abstract

When popular statistical tools such as the *t*-test are taught, the narrative is one in which the data are sampled randomly from some population. Real studies employing random sampling are, however, much rarer than the ubiquity of these tools would suggest, and students may wonder what justifies their use in the absence of random sampling or even of a clear notion of what the population might be. Reassuringly, some of these tools can be understood in an altogether more common setting: experiments in which units are randomly assigned to conditions. This perspective clarifies why familiar procedures such as the *t*-test tend to perform reasonably well in randomised experiments, but also underscores a difference in the scope of inference afforded by random sampling versus random assignment. Furthermore, this perspective foregrounds flexible randomisation-based methods, whose validity follows from the study design and which, as I will suggest, represent a natural starting point for teaching statistical inference to non-mathematicians.

*University of Fribourg, Department of Multilingualism, Rue de Rome 1, 1700 Fribourg, Switzerland. jan.vanhove@unifr.ch. https://janhove.github.io. https://orcid.org/0000-0002-4607-4836

1 Introduction

Much of statistical practice and teaching is built on the assumption that the data are sampled randomly from some population—witness the omnipresent bell curves illustrating the workings of the central limit theorem in introductions to statistics or the widespread use of t-tests and the like in research papers. But research projects in which the units[1] are randomly sampled are few and far between. Far more common are studies in which the units, however obtained, are *assigned* randomly to conditions (e.g., in between-subjects experiments) or to sequences (e.g., in within-subjects experiments using counterbalancing). If my own experience, both as a student and as a lecturer, is anything to go by, this disconnect is mightily confusing: what justifies the use of inferential machinery that was introduced as a way to learn about properties of populations on the basis of random samples in the absence of such random samples—and more often than not in the absence of a clear notion of what exactly the populations might be?

Students may feel some unease about this disconnect, though it may be tempered by the continued more or less successful application of statistical methods derived from sampling theory. My first goal is to explain why such methods perform reasonably well in practice. To this end, I will contrast two approaches to statistical inference: one in which techniques are derived from sampling theory, which I will call the population model, and one in which techniques are derived from what is known about how the units were assigned to experimental conditions, which I will call the randomisation model. We will see that random assignment justifies a form of statistical inference even in the absence of random sampling and that, moreover, there are reassuring connections between techniques rooted in the population model and ones derived from the randomisation model. While the ideas that I will present are all "well known" (as mathematicians like to say), they do not seem to be part and parcel of the way statistics is typically taught and discussed in fields such as psychology and applied linguistics. My hope is this discussion may help readers be clearer-eyed about the scope of and justification for inferences that can be drawn in the absence of random sampling.

Along the way, we will encounter some less commonly used inferential procedures, in particular randomisation testing. My second goal is to outline

[1] In psychology, the units are typically the study's participants. But units could also be schools, words, or anything else that the researchers want to make claims about.

Table 1: Summary statistics for the fictitious experiment. The variances are sample variances, i.e., using Bessel's correction.

Condition	n	Mean	Variance
Without alcohol	4	5.00	0.193
With alcohol	5	5.82	0.502

a few reasons why it makes sense to adopt this technique when teaching statistical inference to non-mathematicians.

2 Two models of statistical inference

Two models of statistical inference can be distinguished: the population model and the randomisation model. In both, the notion of chance or randomness plays a crucial role: on the one hand, randomness prevents us from drawing conclusions about the data with absolute certainty; on the other hand, we can make assumptions about the randomness in the data in order to glean some probabilistic insights from the data. But whereas in the population model, this randomness is thought of as the result of random sampling, the randomisation model assumes that it has been induced by random assignment.

To introduce some notation and a few basic concepts, we will examine how the same randomised experiment can be analysed under each model. For ease of exposition, the experiment is a fictitious one, but I will turn to a real-life example in a later section.

The fictitious setting is this. We want to test the idea that moderate alcohol intake increases verbal fluency in a second language (L2). To that end, we recruit nine students (L1 German, L2 English) and randomly assign them such that four participate in the control condition and five in the experimental condition. Nine participants is obviously a low number, but it keeps things tractable. Participants in the experimental condition drink one pint of ordinary beer; those in the control condition drink one pint of alcohol-free beer. Afterwards, they watch a video clip and relate what happens in it in their L2. Verbal fluency is operationalised as the average number of syllables uttered per second during the first minute of the description. Figure 1 shows the results; Table 1 the summary statistics that we will need in what follows.

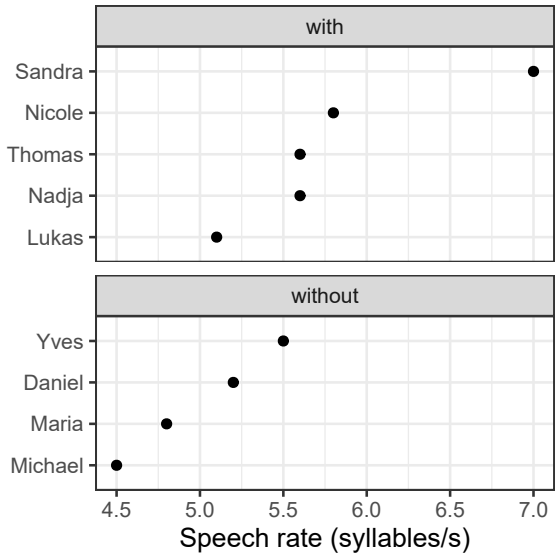

Figure 1: Individual results of a fictitious randomised experiment.

2.1 The population model

In the population model, the observations in Figure 1 are treated as having been obtained through random sampling. For Student's t-test for independent samples, for instance, the observations in the 'without alcohol' and 'with alcohol' conditions are treated as simple random samples from populations P_0 and P_1, respectively. Inference is achieved by assuming that population P_0 follows a Normal distribution with mean μ_0 and variance σ^2 and population P_1 a Normal distribution with mean $\mu_1 = \mu_0 + \delta$ and the same variance σ^2. If these assumptions are well-founded, Student's t-test provides an exact test for the null hypothesis that $\delta = \delta_0$ for some specified number δ_0, where typically $\delta_0 = 0$.[2]

In practice, these assumptions may be relaxed by appealing to the large-sample behaviour of sample means (e.g., Lindeberg's central limit theorem) or

[2]Let the random variable P denote the p-value that the test results in. A test is said to be *exact* if, under the null hypothesis, $\mathbb{P}(P \leq \alpha) \leq \alpha$ for all $\alpha \in (0, 1)$; it is not necessary for equality (i.e., $\mathbb{P}(P \leq \alpha) = \alpha$) to hold. An exact test may *conservative*, meaning that $\mathbb{P}(P \leq \alpha)$ tends to be appreciably less than α under the null hypothesis. Tests may also be *approximate*, meaning that $\mathbb{P}(P \leq \alpha) \approx \alpha$ under the null hypothesis. Such approximate equality is usually demonstrated by investigating how the test behaves as the amount of data increases asymptotically.

using procedures with different or less restrictive assumptions (e.g., Welch's t-test or bootstrapping). The details are not too important; what matters here is that statistical inference in the population model is motivated by treating the observations or the error terms as having been sampled randomly from populations and that it concerns properties of these populations.[3]

It is easy to set up computer simulations in which the data are sampled from some population. But at the coalface, most research, like in our example, is done with volunteers or students scraping together credits. And even if there is some population that these participants can be considered to be a sample from, their recruitment did not involve any mathematically well-defined sampling procedure. In such settings, the population model, while often practically useful, is a fiction.

2.2 The randomisation model

The population model of statistical inference contrasts with the randomisation model. In the population model, the randomness in the results is assumed to be caused by random sampling; in the randomisation model, it is assumed to be caused by randomly assigning the units in the study to different conditions. Whereas random sampling is exceedingly rare in the social sciences, random assignment is quite common. And if random assignment was used, the randomisation model results in valid hypothesis tests that do not make any unverifiable distributional assumptions.

In our example, $m = 9$ units are randomly assigned to two conditions in such a way that $k = 4$ are assigned to the first and $\ell = 5$ are assigned to the second. Under the randomisation model, for each unit $i = 1, \ldots, m$, there exist two potential outcomes $y_i(0)$ and $y_i(1)$, only one of which is observed:

- either $y_i(0)$ if the unit was assigned to the first condition;
- or $y_i(1)$ if the unit was assigned to the second condition.

These potential outcomes are treated as fixed, though possibly unknown, quantities—not as random variables having some distribution. Table 2 shows the observed potential outcomes in our fictitious experiment.

What *is* random is which of the two quantities we end up observing.

[3]The populations do not have be interpreted literally. Sometimes, they are better understood as hypothetical 'superpopulations' of possible observations. Regardless of whether the populations are real or merely conceptual, the techniques used to draw conclusions about them are derived from the assumption that the observations constitute random samples.

Table 2: The observed potential outcomes in the fictitious experiment.

i	Name	$y_i(0)$	$y_i(1)$
1	Daniel	5.2	
2	Lukas		5.1
3	Maria	4.8	
4	Michael	4.5	
5	Nadja		5.6
6	Nicole		5.8
7	Sandra		7.0
8	Thomas		5.6
9	Yves	5.5	

Specifically, let B_i be the random variable that indicates which condition the i-th unit is assigned to. Then the observed outcome for the i-th participant is

$$Y_i = y_i(B_i) = \begin{cases} y_i(0) & \text{if } B_i = 0, \\ y_i(1) & \text{if } B_i = 1. \end{cases}$$

Crucially, the researchers control the randomisation, so they *know* what the distribution of the assignment vector $\boldsymbol{B} = (B_1, \ldots, B_m)$ is. In the present example, the assignment vector was picked uniformly at random from the set of vectors with length $k + \ell = m$ containing exactly k zeroes and ℓ ones. Formally,

$$\boldsymbol{B} \sim \text{Uniform}(\mathcal{B}_{k,\ell}),$$

where

$$\mathcal{B}_{k,\ell} := \{\boldsymbol{b} \in \{0,1\}^{k+\ell} : \sum_{i=1}^{k+\ell} b_i = \ell\}.$$

If a different randomisation scheme was applied (e.g., simple randomisation or blocked randomisation), the distribution of $\boldsymbol{B}$ will be different, too.

In the randomisation model, it is these or similar assumptions that motivate inferential procedures. Unlike assumptions about population distributions, these assumptions are typically known to be correct as the researchers themselves implemented the randomisation. The next section discusses how statistical inference is achieved in the randomisation model.

Table 3: The potential outcomes in the fictitious experiment under the sharp null hypothesis.

i	Name	$y_i(0)$	$y_i(1)$
1	Daniel	5.2	5.2
2	Lukas	5.1	5.1
3	Maria	4.8	4.8
4	Michael	4.5	4.5
5	Nadja	5.6	5.6
6	Nicole	5.8	5.8
7	Sandra	7.0	7.0
8	Thomas	5.6	5.6
9	Yves	5.5	5.5

3 Randomisation-based inference

Statistical inference in the randomisation model has a long pedigree with two of the founding fathers of statistical, R. A. Fisher and J. S. Neyman, making vital contributions. Perhaps unsurprisingly given the legendary acrimony between them, both statisticians went about solving the inferential puzzle in rather different ways.

3.1 Fisher's sharp null hypothesis

Fisher's approach starts by formulating the so-called sharp null hypothesis that there is no differential treatment effect whatsoever (Fisher 1937). Formally,

$$H_{0F} : y_i(1) = y_i(0) \text{ for } i = 1, \ldots, m.$$

Under this null, the values for $y_i(0)$ and $y_i(1)$ can immediately be reconstructed from the observed outcomes: $y_i(0) = y_i(1) = Y_i$ for $i = 1, \ldots, m$; see Table 3.

An exact p-value testing H_{0F} can now be computed as follows.

1. Decide on a test statistic $T(\boldsymbol{y}, \boldsymbol{b})$.
2. Compute the test statistic for the data at hand: $T_{\text{obs}} := T(\boldsymbol{Y}, \boldsymbol{B})$.
3. For a left-sided p-value, compute the probability that $T(\boldsymbol{Y}, \tilde{\boldsymbol{B}}) \leq T_{\text{obs}}$, where $\tilde{\boldsymbol{B}}$ is independently drawn from the distribution of $\boldsymbol{B}$. This can be computed exactly by going through all the possible assignment vectors $\boldsymbol{b}$ and counting how often $T(\boldsymbol{Y}, \boldsymbol{b}) \leq T_{\text{obs}}$. For a right-sided p-value,

instead compute the probability that $T(\boldsymbol{Y}, \tilde{\boldsymbol{B}}) \geq T_{\text{obs}}$. A two-sided p-value can be computed as twice the smaller of the left-sided and the right-sided p-value.

The resulting p-value is valid regardless of the test statistic chosen in the first step (see appendix). Ding (2024) and Hand (2026) discuss a handful of the countless possibilities for useful test statistics.[4] One natural choice for our present example is to use the mean difference, defined as

$$t_{\text{MD}} := \hat{y}(1) - \hat{y}(0),$$

where $\hat{y}(j)$ represents the mean observed outcome among the units with $B_i = j, j = 0, 1$.

Another natural alternative is to use the studentised mean difference. This is defined as

$$t_{\text{SMD}} = \frac{\hat{y}(1) - \hat{y}(0)}{\sqrt{\widehat{V}}},$$

where

$$\widehat{V} := \frac{\widehat{S}^2(0)}{k} + \frac{\widehat{S}^2(1)}{\ell}.$$

In this latter formula, $\widehat{S}^2(j)$ is the usual sample variance of the observed outcomes among the units with $B_i = j, j = 0, 1$. This definition of the studentised mean difference coincides with the test statistic used in Welch's t-test.

Figure 2 shows how both of these test statistics are distributed under the sharp null hypothesis based on all 126 possible assignment vectors. Of the 126 assignment vectors, five resulted in a mean difference of at least 0.82 (i.e., about 4%) and three in a studentised mean difference at least as large as the one actually observed (i.e., about 2.4%). Consequently, the right-sided p-values are 0.040 and 0.024 for the mean difference and studentised mean difference, respectively. R code for reproducing these is available from https://github.com/janhove/justification.

The procedure outlined above is known as randomisation testing. Two major advantages of randomisation testing for researchers are (1) that it is flexible in the sense that it provides exact significance tests for *any* test statistic and (2) that its validity is derived solely from the study's design. I will argue in

[4] But, of course, trying out different test statistics until you find one that yields $p < 0.05$ is an excellent way of invalidating the procedure.

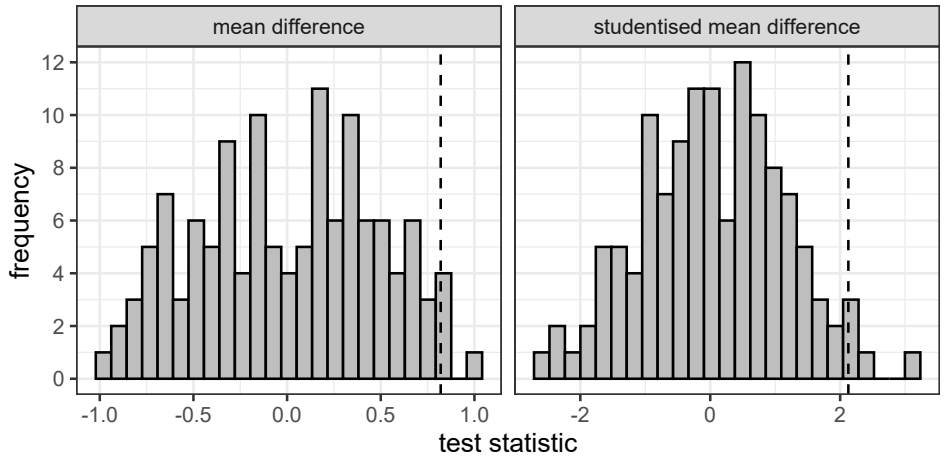

Figure 2: The distributions of the mean difference and of the studentised mean difference in the fictitious experiment under the sharp null hypothesis. The dashed vertical line highlights the test statistic for the data that were actually observed.

Section 7 that, moreover, that randomisation testing provides an attractive starting point for teaching statistical inference to students in fields making extensive use of randomised experiments.

That said, an important drawback to randomisation testing concerns the construction of confidence sets. In principle, confidence sets can be constructed as follows. First, define the sharp null hypothesis of a constant shift δ:

$$H_{0F}(\delta) : y_i(1) = y_i(0) + \delta \text{ for } i = 1, \ldots, m.$$

Under this null, too, all potential outcomes can be reconstructed from the observed ones, and hence, $H_{0F}(\delta_0)$ can be tested for any δ_0 by means of a randomisation test. Thanks to the duality of confidence sets and significance tests, a $(1 - \alpha)$% confidence set for δ can now be constructed as the set of all δ_0 for which the randomisation test of $H_{0F}(\delta_0)$ results in a p-value no larger than α, for $\alpha \in (0, 1)$.

One problem with this approach is that the data may be incompatible with any constant shift, resulting in an empty confidence set. Neyman's approach offers a more attractive solution in this respect.

3.2 Neyman's weak null hypothesis

A rather different approach rooted in the randomisation model was developed by Spława-Neyman (1990, originally published in Polish in 1923). We define the mean potential outcomes (both the observed and unobserved ones) as

$$\bar{y}(j) := \frac{1}{m} \sum_{i=1}^{m} y_i(j)$$

for $j = 0, 1$. The parameter that Neyman wished to estimate is

$$\tau := \bar{y}(1) - \bar{y}(0),$$

which is also commonly referred to as the sample average treatment effect (SATE) in the causal inference literature. An unbiased estimator of τ is

$$\widehat{\tau} := \hat{y}(1) - \hat{y}(0),$$

but what else can be deduced about the distribution of this estimator under the random assignment scheme?

Neyman was able to show that the variance of $\widehat{\tau}$ is conservatively estimated by $\widehat{V}$, which we defined in the previous section. He further showed that, under mild regularity conditions, the quantity

$$T := \frac{\widehat{\tau} - \tau}{\sqrt{\text{Var}(\widehat{\tau})}}$$

has a standard Normal distribution at asymptote; see Ding (2024, Theorem 4.2). What this means is that, other things equal, the distribution of the quantity T under random assignment resembles a standard Normal distribution ever more closely for ever larger k, ℓ. Combining both facts, an asymptotically conservative two-sided $(1 - \alpha)\%$ confidence interval for τ may be constructed as

$$[\widehat{\tau} \pm z_{\alpha/2} \sqrt{\widehat{V}}],$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard Normal, for $\alpha \in (0, 1)$.

The duality of confidence intervals and significance tests further allows us to define an asymptotically conservative test for what is called the weak

(Neymanian) null hypothesis:

$$H_{0N} : \tau = 0.$$

The weak null hypothesis does not assume that the individual potential outcomes are identical in both conditions, but that, whatever the differences between these potential outcomes are, they cancel out on average. This null hypothesis is weak in the sense that the veracity of the sharp (Fisherian) null hypothesis implies the veracity of the weak (Neymanian) one, but not vice versa.

The asymptotically conservative test involves computing T and comparing it against a standard Normal distribution. In our present example, we obtain $\widehat{\tau} = 0.82$, $T = 2.13$, yielding the two-sided 95% confidence interval for τ of $[0.06, 1.58]$ and the right-sided p-value for the weak null hypothesis of $p = 0.017$.

Compared to randomisation tests of the sharp null, the Neymanian approach offers little flexibility in terms of the choice of the test statistic.[5] Further, we are only given asymptotic guarantees. With small sample sizes, the asymptotics may not yet have kicked in sufficiently for the results to be taken at face value. However, the *average* treatment effect among the units represented (i.e., τ) arguably represents a more sensible target of inference than the purported *constant* treatment effect δ that randomisation test-based confidence sets would be concerned with.

4 A real-life example

4.1 Data

Our running example is unrealistically small, so let us see how three different methods stack up in a real-life example. The data for this example stem from a replication of Oppenheimer & Monin (2009) by Klein et al. (2014): participants were randomly assigned to one of two conditions. In one condition, they were asked to imagine that they walk into a casino and see a man rolling three dice, all of which come up 6; in the other condition, they were asked to imagine that only two out of the three dice came up 6. They then estimated how many times the man had been rolling the three dice before they showed up. Klein et al. (2014) replicated this and several other

[5] However, it is possible to account for covariates in the Neymanian approach, too; see Lin (2013).

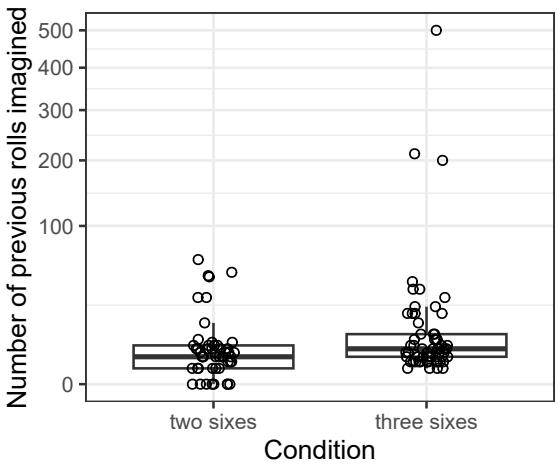

Figure 3: The raw outcomes in the Brasília sample in Klein et al.'s replication. The y-axis is square-root transformed, but the values analysed are not.

Table 4: Summary statistics the *gambler's fallacy* replication (raw data). The variances are sample variances, i.e., using Bessel's correction.

Condition	n	Mean	Variance
Two sixes	51	8.14	199
Three sixes	63	22.21	5052

studies in several convenience samples. The data analysed in this example are those from their Brasília sample.

While Oppenheimer & Monin (2009) analysed the raw estimates and did not exclude any outliers, Klein et al. (2014) square-root transformed the data and excluded responses farther than three standard deviations from the mean. In the analyses below, I consider both the raw and the square-root transformed data; I did not exclude outlying values, however.

4.2 Raw data

Figure 3 shows the distribution of the particpants' raw estimates. The y-axis is square-root-transformed, but the data analysed are not. The extreme skew in the data is evident even with the square-root-transformed y-axis. Table 4 shows the summary statistics.

Welch's t-test is often recommended as the default choice for comparing the means of two independent samples (Delacre, Lakens & Leys 2017; Ruxton 2006). When we apply this technique rooted in random sampling theory to the Brasília data, we obtain $t(68) = 1.5$, for a right-sided p-value of 0.065 and a two-sided 95% confidence interval for the mean difference of $[-4.2, 32.4]$.

For the randomisation test, we use the studentised mean difference as the test statistic, which coincides with the test statistic used in Welch's t-test. Two remarks are in order first, however.

First, Klein et al.'s replication seems to have used simple (Bernoulli) randomisation rather than complete randomisation. To account for this, we could use the set $\{0, 1\}^{51+63}$ as the source of the assignment vectors instead of the set $\mathcal{B}_{51,63}$. However, it is easily shown that an exact test for fixed group sizes remains exact when treating random group sizes as fixed [Lydersen, Fagerland & Laake (2009), p. 1165; also see the appendix], so I will follow this approach.

Second, the set $\mathcal{B}_{51,63}$ contains about $8 \cdot 10^{32}$ assignment vectors, which is too many for exhaustive randomisation testing. Instead, we sample $M = 19999$ assignment vectors independently and uniformly at random from $\mathcal{B}_{51,63}$ and add to this sample the assignment vector actually obtained in the experiment. A valid right-sided p-value can then be computed as

$$p_r = \frac{\#\text{ assignment vectors yielding } T(\boldsymbol{Y}, \boldsymbol{b}) \geq T_{\text{obs}}}{M + 1},$$

see the appendix.

Figure 4 shows the resulting randomisation distribution of the studentised mean difference; it looks anything but bell-shaped. The right-sided p-value is 0.083.

The asymptotic Neyman confidence interval for τ was derived assuming complete randomisation, but can also be shown to be correct asymptotically for experiments using simple randomisation. The 95% confidence interval it yields is $[-3.9, 32.0]$, with a right-sided p-value of 0.062.

4.3 Transformed data

Let us now turn to the square-root transformed data, the summary statistics for which are shown in Table 5. Welch's approach yields $t(98.5) = 2.01$, for a right-sided p-value of 0.024 and a 95% confidence interval of $[0.01, 2.05]$. A

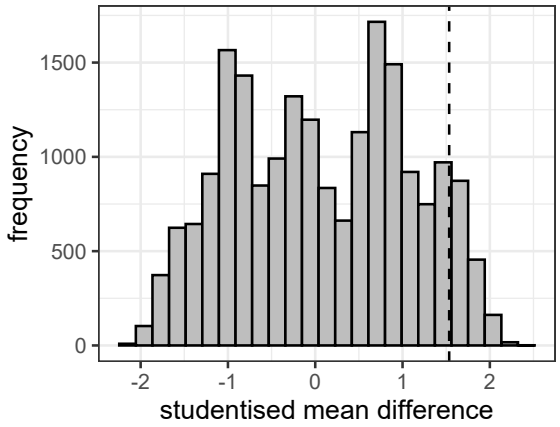

Figure 4: The randomisation distribution for the studentised mean difference in the *gambler's fallacy* replication study (raw data). The dashed vertical line highlights the test statistic for the data that were actually observed.

Table 5: Summary statistics the *gambler's fallacy* replication (square-root transformed data). The variances are sample variances, i.e., using Bessel's correction.

Condition	n	Mean	Variance
Two sixes	51	2.17	3.5
Three sixes	63	3.20	12.2

randomisation test with $M = 19999$ using the studentised mean difference yields a right-sided p-value of 0.026; the randomisation distribution shown in Figure 5 looks considerably less wonky than the one for the raw data. Finally, Neyman's approach yields a right-sided p-value of 0.022 and a 95% confidence interval of $[0.03, 2.03]$.

4.4 Discussion

4.4.1 Similarities between the approaches

Despite the differences in the null hypotheses being tested—equality of population means in the Welch t-tests, the sharp null hypothesis in the randomisation tests, and the weak null hypothesis in the Neyman approach—the inferential results are rather similar across the three methods. This is especially the case for the square-root transformed data, for which the

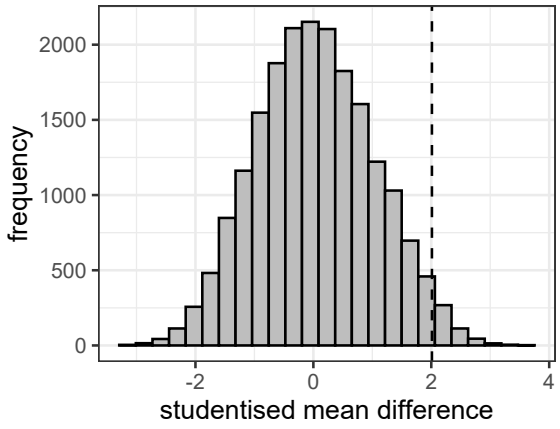

Figure 5: The randomisation distribution for the studentised mean difference in the *gambler's fallacy* replication study (square-root transformed data). The dashed vertical line highlights the test statistic for the data that were actually observed.

randomisation distribution (Figure 5) already hints at connections between randomisation testing and approaches appealing to Normal distributions (i.e., the Welch and Neyman approaches). For the raw data, the differences between these techniques are more pronounced, and the randomisation distribution (Figure 4) is correspondingly more wonky.

4.4.2 Transformations change some hypotheses

It is, however, worth thinking about what the square-root transformation does and does not change. In the Welch approach, inference now concerns the difference between the means of the square-root transformed populations from which the data were putatively randomly sampled and not, say, the difference between the square roots of the means from these population. Similarly, in the Neyman approach, inference now concerns the quantity

$$\tilde{\tau} := \frac{1}{m} \sum_{i=1}^{m} \left(\sqrt{y_i(1)} - \sqrt{y_i(0)} \right)$$

rather than

$$\tau = \frac{1}{m} \sum_{i=1}^{m} \left(y_i(1) - y_i(0) \right).$$

The quantity $\tilde{\tau}$ is neither directly related to τ nor to, say, the sum of the square-roots between the individual potential outcomes. In particular, $\tau = 0$ does not imply $\widetilde{\tau} = 0$, nor vice versa. Researchers, then, need to be aware that such transformations change the hypotheses being tested in the Welch and Neyman approaches.

In contrast, when using randomisation testing, the transformation merely amounts to a change in the test statistic being used; it does not change the sharp null hypothesis being tested. It is this change in test statistic that results in a difference between the p-values obtained for the raw and transformed data, not a change in the hypothesis being tested. The randomisation test itself is valid regardless of the test statistic used. This underscores the flexibility of hypothesis testing: researchers may freely choose the test statistic that they anticipate would put difference between the conditions in boldest relief.

5 Connections between the different approaches

The example in the previous section hints at a few important connections between the different approaches covered, despite differences in their derivation.

First, Welch's t-test, which is derived from a random sampling perspective, and Neyman's approach, which is derived from a random assignment framework, use the same test statistic, namely the studentised mean difference

$$\frac{\hat{y}(1) - \hat{y}(0)}{\sqrt{\frac{\widehat{S}^2(0)}{k} + \frac{\widehat{S}^2(1)}{\ell}}}.$$

In the Neyman approach, the standard Normal distribution is used as a conservative limiting distribution of this test statistic in the sense that the test statistic is Normally distributed asymptotically, but with a variance that is at most 1. In the Welch approach, sampling theory yields a t-distribution with $\nu \leq k + \ell - 2$ degrees of freedom as the approximate law of this statistic. The $t(\nu)$ distribution is more heavy-tailed than the standard Normal, with the difference becoming smaller for larger ν. As a result, despite the difference in derivation between these two techniques, the numerical results will differ only slightly between them, with the p-values obtained with the Welch test being somewhat larger and the confidence intervals somewhat wider than those obtained with the Neyman approach. The Welch t-test can, then, be

considered a more conservative version of Neyman's method, which is itself asymptotically conservative.

Second, Ding (2017) showed that a randomisation test using the studentised mean difference as the test statistic is an asymptotically conservative test of the weak null H_{0N} as well as being an exact test of the sharp null H_{0F} (also see Ding & Dasgupta 2017). Importantly, however, this first property relies on the studentisation and does not hold for every test statistic. Ding (2024:sec. 8.1) sketches the intuition behind the need for studentisation for this property to hold.

Conversely, the sharp null hypothesis implies the weak one. By contraposition, a willingness to reject the weak null logically implies a willingness to reject the sharp null.[6] Thus, in practice, the Neyman test of the weak null (and hence also the Welch test) also tests the sharp null, and in sufficiently large samples, the randomisation test of the sharp null serves as a test of the weak null, provided it uses the studentised mean difference as the test statistic.

A third connection between the population and randomisation models is provided by permutation testing (Ernst 2004). Mechanically, permutation tests and randomisation tests are identical. In both cases, the statistician computes a test statistic for the observed data and compares it to the values obtained by recomputing the statistic after repeatedly reallocating treatment labels. Conceptually, however, the two procedures derive their validity from different assumptions. In the randomisation model, validity follows from knowledge of the assignment mechanism and the potential outcomes framework. In the population model, validity is justified by the exchangeability of the observations under the null hypothesis. Thus, the same computational procedure receives two different theoretical justifications. Writing about now-popular techniques such as the t- and F-tests, Fisher (1936) remarked that

> the statistician does not carry out this very simple and very tedious process [i.e., permutation testing, JV], but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method. (p. 59)

From this perspective, these popular techniques are mathematically conve-

[6] As Ding (2017) demonstrated, however, this does not mean that a significant Neyman test implies that the corresponding randomisation test will be significant.

nient approximations to procedures that, historically, were computationally prohibitive.

In consequence, while it is arguably confusing to be taught sampling-based techniques and then asked to apply these in the absence of sampling, the connections between these two fields of application are considerable. These connections are reassuring and explain why reanalysing randomised experiments using randomisation-based techniques will often yield conclusions similar to those obtained using familiar textbook methods. The connections, however, may also be opaque to both students and researchers. At least from a pedagogical perspective, therefore, it may make sense to make more explicit the logic of statistical inference in the context of randomised experiments.

6 The scope of inference

In spite of the connections between the population and the randomisation models, it is sensible to keep them mentally separate as they differ in what they allow us to infer.

Under the population model, conclusions are drawn about populations on the basis of samples. If students are often taught statistical methods in this framework and later encounter the same methods in the analysis of randomised experiments, they could be forgiven for believing that these methods afford statistical generalisation to whatever population the units may be considered to have been sampled from.

The randomisation model makes clear, however, that this is not the case, as it only references the units included in the experiment. In the absence of some well-defined sampling scheme, the only units about which purely statistical conclusions can be drawn are the ones in the experiment. This is true both in the Fisherian and the Neymanian framework. From this perspective, the fact that procedures derived from the population model often perform similarly to ones derived from the randomisation model is a mathematical nicety that may muddle conceptual clarity.

This, of course, does not mean that generalisations beyond the units in the experiment is impossible. But such generalisations are to be derived from logical and substantive arguments rather than from purely statistical ones.

7 Pedagogical excursion

Berger et al. (2002), among others, have argued that randomisation tests should be the tool of choice for analysing randomised experiments. I am sympathetic to their arguments, but the point I want to make here is that randomisation tests could be more valuable still in teaching statistical inference (also see Tintle et al. 2012 for some empirical evidence). I see a few reasons for this.

First, randomisation tests clearly connect a key aspect of the research design (i.e., random assignment) with the analysis. In doing so, they underscore that restrictions to the randomisation, as in cluster-randomised experiments or blocked designs, should be taken into account in the analysis.

Second, they are entirely mechanical and make no reference to mathematical idealisations such as Normal or t-distributions. Their mechanical nature also makes it straightforward to visualise how they work: for small examples, lecturers can iterate through all assignment vectors and illustrate how the randomisation distribution of the test statistic of choice comes about.

Third, they do not rely on unverifiable assumptions about, say, the normality of the parent distributions or whether the samples are large enough for the central limit theorem to have kicked in sufficiently.

Fourth, the proofs of the classical central limit theorem and the Gosset–Fisher theorem, which concerns the validity of t-testing under random sampling from a Normal distribution, rely on advanced mathematical concepts that students and researchers outside of mathematical fields are unlikely to be familiar with; the proofs of the more general Lindeberg central limit theorem or of the validity of, say, bootstrapping are more demanding still. By contrast, the proof of the validity of exhaustive randomisation testing relies on much more elementary arguments, and two other properties that I relied on in the real-life example are only slightly more involved. I provide these three proofs in the appendix; I think that, with some guidance from the instructors, they are accessible to students in non-mathematical fields.

What is pedagogically useful about proving a result is not merely that it establishes correctness. Proofs make explicit which assumptions the result relies on and which features are unimportant. For instance, the proof of the validity of randomisation testing does not depend on the distribution of the data or on whether the group sizes are balanced. Further, students do not have to accept these results on authority nor do they have to rely

solely on examples and simulations that necessarily are limited to particular cases. Moreover, such proofs also make it clear what they are *not* about—in the present case, inference to a larger population than the units under investigation, for instance.

None of this implies that asymptotic results such as the central limit theorem or procedures such as Welch's t-test should not be taught in one way or another. But randomisation testing arguably provides a more natural starting introduction to statistical inference in the context of randomised experiments than do these sampling-based procedures.

8 Conclusion

Sampling-based techniques are commonly used to analyse randomised experiments despite the absence of any well-defined sampling scheme. The fact that these techniques tend to work reasonably well even under these circumstances can be understood by their connections to purely randomisation-based techniques. Distinguishing between the randomisation and population models also helps us being clearer-eyed about the scope of inference: random assignment justifies statistical conclusions about the units represented in the experiment, whereas broader generalisations require substantive rather than purely statistical arguments.

9 Appendix: Some proofs

The first and most central proof shows that exhaustive randomisation testing is exact for *any* test statistic. The most complex stochastic argument it relies on is that

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B),$$

which instructors can justify by drawing a Venn diagram.

Theorem 9.1 (Exactness of randomisation testing)**.** *Under the sharp null hypothesis, exhaustive randomisation testing results in a p-value P such that $\mathbb{P}(P \leq \alpha) \leq \alpha$ for each $\alpha \in (0, 1)$.*

Proof. Assume there are M possible assignment vectors $\boldsymbol{b}$, with each having probability $1/M$ of being selected. Under the sharp null hypothesis, we can compute the test statistic $T(\boldsymbol{Y}, \boldsymbol{b})$ for each assignment vector $\boldsymbol{b}$. Sort these test statistics in descending order, breaking ties randomly if necessary:

$$t_1 \geq t_2 \geq \cdots \geq t_{M-1} \geq t_M.$$

After this sorting, the right-sided p-value that the i-th randomisation would have resulted in (call it p_i) can be seen to be at least $i/M, i = 1, \ldots, M$. (It is possible that $p_i > i/M$ if $t_{i+1} = t_i$.)

For any $\alpha \in (0, 1)$, now compute $\kappa := \lfloor \alpha M \rfloor$. ($\lfloor \cdot \rfloor$ rounds down the number to the nearest integer.) We were equally likely to have generated each of the M possible assignment vectors. Hence, the probability that we generated an assignment vector resulting in a right-sided p-value P_r no larger than α is

$$\begin{aligned}
\mathbb{P}(P_r \leq \alpha) &= \frac{\#\text{ assignments with } p_i \leq \alpha}{M} \\
&\leq \frac{\#\text{ assignments with } i/M \leq \alpha}{M} \\
&= \frac{\#\text{ assignments with } i \leq \kappa}{M} \\
&= \frac{\kappa}{M} \\
&\leq \alpha.
\end{aligned}$$

(The inequality on the second line follows from $p_i \geq i/M$; the equality on the third line follows from the definition of κ.)

To demonstrate validity for left-sided p-values, sort the test statistics in ascending order and proceed analogously.

Finally, let P_ℓ and P_r denote the left- and right-sided p-values, respectively, and define the two-sided p-value as

$$P := 2 \cdot \min\{P_\ell, P_r\}.$$

Then, for all $\alpha \in (0, 1)$,

$$\begin{aligned}
\mathbb{P}(P \leq \alpha) &= \mathbb{P}(2 \cdot \min\{P_\ell, P_r\} \leq \alpha) \\
&= \mathbb{P}(P_\ell \leq \alpha/2 \text{ or } P_r \leq \alpha/2) \\
&= \mathbb{P}(P_\ell \leq \alpha/2) + \mathbb{P}(P_r \leq \alpha/2) - \mathbb{P}(P_\ell \leq \alpha/2 \text{ and } P_r \leq \alpha/2) \\
&\leq \mathbb{P}(P_\ell \leq \alpha/2) + \mathbb{P}(P_r \leq \alpha/2) \\
&\leq \alpha/2 + \alpha/2 \\
&= \alpha.
\end{aligned}$$

(The inequality on the penultimate line follows from what we have already proved for P_ℓ and P_r.) Hence, the two-sided p-value obtained by randomisation testing is exact. □

The second proof shows that conditioning on observed random quantities preserves exactness. This justifies using the randomisation test for fixed group sizes k, ℓ when the group sizes are random quantities K, L, e.g., as when using simple randomisation or when the total size of the experiment was not fixed in advance. The most complex stochastic argument used is the law of total probability, which can be illustrated using a tree diagram.

Theorem 9.2 (Conditioning preserves exactness)**.** *Suppose that a test is exact regardless of which value a random quantity C takes. That is, under the null hypothesis,*

$$\mathbb{P}(P \leq \alpha \text{ if } C = c) \leq \alpha$$

for any possible value c of C and every $\alpha \in (0, 1)$. Then the test is also exact when C is ignored, that is,

$$\mathbb{P}(P \leq \alpha) \leq \alpha$$

for every $\alpha \in (0, 1)$.

Proof. Let $\mathcal{C}$ be the set of the possible values of C. By the law of total

probability, we have

$$\begin{aligned}\mathbb{P}(P \le \alpha) &= \sum_{c \in C} \underbrace{\mathbb{P}(P \le \alpha \text{ if } C = c)}_{\le \alpha} \mathbb{P}(C = c) \\ &\le \alpha \underbrace{\sum_{c \in C} \mathbb{P}(C = c)}_{=1}.\end{aligned}$$

□

Exhaustive randomisation testing is infeasible for larger experiments, but Monte Carlo randomisation testing is also exact. The proof is a bit more complicated, relying as it does on the notion of exchangeability.

Theorem 9.3 (Exactness of Monte Carlo randomisation testing)**.** *Suppose that the sharp null hypothesis holds. Let $\mathcal{B}$ be the distribution from which the assignment vector was drawn. Generate M assignment vectors by independent draws from $\mathcal{B}$. Add to this selection the assignment vector used in the experiment. Then*

$$P_r = \frac{R}{M+1},$$

where R is the number of assignment vectors resulting in test statistic at least as large as the one observed in the experiment, is an exact right-sided p-value.

Similarly,

$$P_\ell = \frac{L}{M+1},$$

where L is the number of assignment vectors resulting in test statistic at most as large as the one observed in the experiment, is an exact left-sided p-value, and $P = 2 \cdot \min\{P_\ell, P_r\}$ is an exact two-sided p-value.

Proof. Compute the test statistic $T(\boldsymbol{Y}, \boldsymbol{b})$ for each of the $M+1$ assignment vectors $\boldsymbol{b}$ and sort them in descending order, breaking ties randomly when necessary:

$$t_1 \ge t_2 \ge \cdots \ge t_M \ge t_{M+1}.$$

Since the $M+1$ assignment vectors were drawn independently from $\mathcal{B}$, these test statistics are exchangeable. In particular, the test statistic generated by the assignment vector actually used in the experiment occurs with equal probability in each of the $M+1$ positions. Hence, letting $\widetilde{R}$ be the rank of

the observed test statistic in this ordering, we have $\mathbb{P}(\widetilde{R} = r) = 1/(M+1)$ for each $r = 1, \ldots, M+1$.

For every $\alpha \in (0,1)$, write $\kappa = \lfloor (M+1)\alpha \rfloor$. Further note that $\widetilde{R} \leq R$. ($\widetilde{R} < R$ may occur if the test statistic actually obtained is identical to another one.) For the right-sided p-value P_r, we now obtain

$$\begin{aligned}
\mathbb{P}(P_r \leq \alpha) &= \mathbb{P}\left(\frac{R}{M+1} \leq \alpha\right) \\
&= \mathbb{P}(R \leq (M+1)\alpha) \\
&= \mathbb{P}(R \leq \kappa) \\
&\leq \mathbb{P}(\widetilde{R} \leq \kappa) \\
&= \sum_{r=1}^{\kappa} \mathbb{P}(\widetilde{R} = r) \\
&= \sum_{r=1}^{\kappa} \frac{1}{M+1} \\
&= \frac{\kappa}{M+1} \\
&= \frac{\lfloor (M+1)\alpha \rfloor}{M+1} \\
&\leq \alpha.
\end{aligned}$$

(If R is as least as large as $\widetilde{R}$, then the probability that R is no larger than κ must be at most as large as the probability that $\widetilde{R}$ is no larger than κ.)

Sort the test statistics in ascending order for the left-sided p-value, and refer to the proof of Theorem 9.1 for the two-sided p-value. □

References

Berger, Vance W., Clifford E. Lunneborg, Michael D. Ernst & Jonathan G. Levine. 2002. Parametric analysis in randomized clinical trials. *Journal of Modern Applied Statistical Methods* 1(1). 74–82. doi:10.22237/jmasm/1020255120.

Delacre, Marie, Daniël Lakens & Christophe Leys. 2017. Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology* 30(1). 92–101. doi:10.5334/irsp.82.

Ding, Peng. 2017. A paradox from randomization-based causal inference. *Statistical Science* 32(3). 331–345. doi:10.1214/16-STS571.

Ding, Peng. 2024. *A first course in causal inference*. Boca Raton, FL: CRC Press. doi:https://doi.org/10.1201/9781003484080.

Ding, Peng & Tirthankar Dasgupta. 2017. A randomization-based perspective on analysis of variance: A test statistic robust to treatment effect heterogeneity. *Biometrika* 105(1). 45–56. doi:10.1093/biomet/asx059.

Ernst, Michael D. 2004. Permutation methods: A basis for exact inference. *Statistical Science* 19. 676–685.

Fisher, R. A. 1937. *The design of experiments*. 2nd ed. Edinburgh: Oliver; Boyd.

Fisher, Ronald Aylmer. 1936. "The coefficient of racial likeness" and the future of craniometry. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 66. 57–63. doi:10.2307/2844116.

Hand, David J. 2026. *What's the question? Deciding what you really want to know*. CRC Press. doi:10.1201/9781003726821.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology* 45(3). 142–152. doi:10.1027/1864-9335/a000178.

Lin, Winston. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7(1). 295–318. doi:10.1214/12-AOAS583.

Lydersen, Stian, Morten W. Fagerland & Petter Laake. 2009. Recommended tests for association in 2×2 tables. *Statistics in Medicine*.

Oppenheimer, Daniel M & Benoît Monin. 2009. The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making* 4(5). 326–334. doi:10.1017/S1930297500001170.

Ruxton, Graeme D. 2006. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral*

Ecology 17. 688–690. doi:10.1093/beheco/ark016.

Spława-Neyman, Jerzy. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5(4). 465–480. doi:10.1214/ss/1177012031.

Tintle, Nathan L., Kylie Topliff, Jill VanderStoep, Vicki-Lynn Holmes & Todd Swanson. 2012. Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research* 11(1). 21–40.