A Closer Look at Testing the "No-Treatment-Effect" Hypothesis in a Comparative Experiment[1]

Joseph B. Lang

Department of Statistics and Actuarial Science
University of Iowa, Iowa City, IA USA

Abstract

Standard tests of the "no-treatment-effect" hypothesis for a comparative experiment include permutation tests, the Wilcoxon rank sum test, two-sample t tests, and Fisher-type randomization tests. Practitioners are aware that these procedures test different no-effect hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. Borrowing from the rich literatures on causality and finite-population sampling theory, this paper develops a modeling framework that affords answers to several important questions, including: exactly what hypothesis is being tested?, what model assumptions are being made?, and are there other, perhaps better, approaches to testing a no-effect hypothesis? The framework lends itself to clear descriptions of three main inference approaches: science-based, randomization-based, and selection-based. It also promotes careful consideration of model assumptions and targets of inference, and highlights the importance of randomization. Along the way, Fisher-type randomization tests are compared to permutation tests and a less well known Neyman-type randomization test. A small-scale simulation study compares the operating characteristics of the Neyman-type randomization test to those of the other more familiar tests.

Keywords and Phrases: Causal effects; Completely randomized design; Finite-population sampling theory; Fisher vs. Neyman; Fisher's exact test; Horvitz-Thompson estimator; Non-measurable probability sample; Permutation tests; Potential variables; Randomization-based inference; Randomization tests; Science-based inference; Selection-based inference

1 Introduction

We begin with a simple example of a comparative experiment. Researchers are interested in determining whether cell phone use while driving has an impact on reaction times. Toward this end, 64 University of Utah student volunteers were enlisted to take part in a randomized comparative experiment (Strayer and Johnston 2001). Of the 64 students, 32 were randomized to treatment 1 (operate a driving simulator while using a cell phone) and 32 were randomized to treatment 2 (operate a driving simulator without a cell phone). For a summary description of the data and of the way the two treatments were actually administered, see Agresti and Franklin (2007:446). In the driving simulation, each student encountered several red lights at random times. Each student's response was the average time required to stop when a red light was detected. The 64 responses, in milliseconds, are recorded in Table 1.

[1] 1/25/12 This research was supported in part by NSF grant SES-1059955.

Table 1. Reaction Times (milliseconds).

Cell Phone:	636 623 615 672 601 600 542 554 543 520 609 559 595 565 573 554
	626 501 574 468 578 560 525 647 456 688 679 960 558 482 527 536
Control:	557 572 457 489 532 506 648 485 610 444 626 626 426 585 487 436
	642 476 586 565 617 528 578 472 485 539 523 479 535 603 512 449
Generically...	
Treatment 1:	$y_{1,1}, y_{1,2}, \ldots, y_{1,32}$
Treatment 2:	$y_{2,1}, y_{2,2}, \ldots, y_{2,32}$

Is there a cell phone use effect? Generically, is there a *treatment effect*?

Standard tests of the "no-treatment-effect" hypothesis include permutation tests (Pitman 1937, 1938), the Wilcoxon rank sum test (Wilcoxon 1945), two-sample t tests (cf. Welch 1938), and Fisher-type randomization tests (Fisher 1935). Most practitioners are aware that these procedures test different "no-effect" hypotheses and are based on different modeling assumptions. However, this awareness is not always, or even usually, accompanied by a clear understanding or appreciation of these differences. This paper looks at each of these testing approaches and addresses the all important questions, exactly what hypothesis is being tested? and what model assumptions are being made? Along the way, we will have to confront several other questions such as, how is the definition of *treatment effect* operationalized?, what is the actual target of inference?, what is the role of randomization?, and are there other, perhaps better, approaches to testing a no-effect hypothesis?

To address these questions, we draw on ideas from the rich literature on causal analysis. In particular, we employ the useful concept of "potential variables." Although the idea of potential variables can be traced back to Neyman (1923), Rubin, begining with a series of papers on causal models in the 1970's (see Rubin 2010 and references therein) is usually credited with more explicitly stating the potential variable model and extending it to both randomized and non-randomized design settings, with or without covariates (see Rubin's causal model, Holland 1986). Between Neyman and Rubin, potential variables were used by relatively few authors; Welch (1937), Kempthorne (1952, 1955), and Cox (1958) were among the notable early proponents. Around the time of and after Rubin, many more authors made important contributions to the potential variables literature. See for example, Copas (1973), Holland (1986), Greenland (1991, 2000), Gadbury (2001), and the references therein.

To be clear, it is not the goal of this paper to summarize the vast literature on potential variables and causal modeling. (To this end, see Paul R. Rosenbaum's very informative website and references therein, `www-stat.wharton.upenn/ rosenbap/downloadTalks.htm`.) Instead, the first goal is to exploit the benefits of hindsight to develop a modeling framework that supports clear descriptions and comparisons of the different testing approaches, and promotes careful consideration of the model assumptions and targets of inference. This modeling framework and associated notation draws clear distinctions between realizations and random variables, and between observed and unobserved data. It accommodates both treatment assignment and sampling from populations, and clearly differentiates between the two. Although the proposed model lends itself to generalizations in many directions (e.g. more than two treatments, restricted randomization, etc.), to simplify exposition, we will focus on the two-treatment comparative experiment setting. This restriction allows us to more directly highlight the useful features of the proposed modeling framework.

The second goal of this paper is to address the question of availability of other testing approaches, besides the four common ones mentioned above. Toward this end, we revisit ideas introduced in Neyman (1923). Using the model structure introduced herein, we describe a less well known Neyman-type randomization test, which is qualitatively different than the Fisher-type randomization test (cf. Welch 1937, Rubin 1990, 2004, 2010).[2] The Neyman-type randomization test, which uses a less restrictive "no-effect" hypothesis than Fisher's, is based on a test statistic with the common form, (estimator minus estimand)/(standard error of estimator). Neyman, with an eye on interval estimation rather than testing, derived the standard error with respect to a randomization distribution using tools from finite-population sampling theory. In retrospect, Neyman's derivation approach is hardly surprising given that he "may be said to have initiated the modern theory of survey sampling" (Lehmann 1994) in his landmark paper of 1934 (Neyman 1934). Compared to Fisher-type randomization tests, the Neyman tests do have their advantages and disadvantages. One disadvantage is that Neyman tests are approximate, whereas Fisher tests are exact. An advantage is that the Neyman test can be more powerful than the Fisher test (see section 10 below). Another advantage is that, unlike the Fisher-type randomization test, the Neyman version can be used to test hypotheses about a population when units are randomly sampled from the population and then randomized to treatment levels.

The third and final goal of this paper is to compare the operating characteristics of the five tests: the permutation test, the Wilcoxon rank sum test, the two-sample t-test, the Fisher-type randomization test, and the Neyman-type randomization test. The penultimate section of the paper includes a small-scale simulation study of the size and power of these five tests. Based on these comparisons, we make tentative recommendations on which test to use in different settings.

The remainder of this paper is organized as follows: Section 2 introduces potential variables and recasts the data in Table 1 within this framework. Section 3 introduces a sequential data generation model that explicitly accommodates both random sampling and randomization. The components in the three-level sequential model are identified as the "science," the "sampling," and the "randomization." This model, along with a useful component-selection notation, leads to an explicit identification of the observed data and the three main targets of inference. Section 4 gives candidate definitions of treatment effects that are based on potential variables, along with corresponding no-treatment-effect hypotheses. An overview of the three main inference approaches—science-based, selection-based, and randomization-based—is given in Section 5 and candidate test statistics corresponding to these inference approaches are given in Section 6. Sections 7-9 describe each of the three inference approaches in more detail. These sections give specific examples of testing procedures, some of them well known and some of them less well known. Section 10 carries out an analysis of the cell phone data and includes a small scale simulation study of the operating

[2] Readers with an interest in history are encouraged to read Neyman (1935), along with the discussions, to see how Neyman and Fisher publicly aired their differences of opinions on testing in randomized design settings.

characteristics of the different testing approaches discussed herein. Finally, Section 11 includes a brief discussion.

2 What Might Have Been: The Potential Variables Viewpoint

Going back to Neyman (1923) and following the lead of Welch (1937), Kempthorne (1955), Cox (1958), and Rubin (e.g. 2005), we will view the data as observed values of a sample of "potential values."

Consider a population $\underline{P}$ of N units that are, without loss of generality, identified by the numbers 1 through N; in symbols, $\underline{P} = (1, \ldots, N)$. Let $Y_{t.i}$ be the response for unit i when exposed to treatment t, where $i = 1, \ldots, N$ and $t = 1, 2$. The response variables $Y_{1.i}$ and $Y_{2.i}$ are called potential variables for reasons made clear in the next paragraph.

The introduction of these potential variables leads to intuitively appealing definitions of treatment effects that are based on head-to-head comparisons of $Y_{1.i}$ and $Y_{2.i}$. There is a catch, however. Although there is the potential to observe either $Y_{1.i}$ or $Y_{2.i}$, unfortunately, it is not possible to observe both. Strictly speaking, it is not possible to observe the values of both potential variables because the same subject cannot be simultaneously exposed to both treatments. To the potential variable advocates, this is the "fundamental problem of causal analysis" (Holland, 1986). As an example, if we observe the value of $Y_{2.i}$, then the value of $Y_{1.i}$, and hence the difference $Y_{1.i} - Y_{2.i}$, cannot be observed. In this case, the unobserved value of $Y_{1.i}$ is relegated to fantasy, the value is "what might have been" had unit i been exposed to treatment 1 rather than treatment 2.

The data in Table 1 can be viewed as observed values of a sample of the potential variable values. Specifically, a sample $\underline{s}$ of size $n = 64$ is taken, without replacement, from the population $\underline{P}$. That is, $\underline{s} = (s_1, \ldots, s_n)$, where $s_j \in \underline{P}$ and $s_j \neq s_{j'}$. One of the two treatments will be assigned to each of the units in the sample $\underline{s}$. For the example, treatment 1 was assigned to 32 and treatment 2 was assigned to 32 of the 64 sampled units.

Let $y_{t.s_j}$ be the response value for sampled subject s_j when exposed to treatment t. That is, $y_{t.s_j}$ is a realization of $Y_{t.s_j}$. Of course, for each subject s_j, only one of the realizations, $y_{1.s_j}$ or $y_{2.s_j}$, will be observed. From a potential variables viewpoint, the original data in Table 1 can be viewed as follows:

Table 2.

Treatment 1:	$\xcancel{y_{1}}_{.s_1}, \xcancel{y_{1}}_{.s_2}, y_{1.s_3}, \ldots, \xcancel{y_{1}}_{.s_{63}}, y_{1.s_{64}}$	Only the 32 non-×'ed out values are observed.
Treatment 2:	$y_{2.s_1}, y_{2.s_2}, \xcancel{y_{2}}_{.s_3}, \ldots, y_{2.s_{63}}, \xcancel{y_{2}}_{.s_{64}}$	Only the 32 non-×'ed out values are observed.

Here, $\underline{s} = (s_1, \ldots, s_{64})$ is a sample from some population $\underline{P} = (1, \ldots, N)$, $N \geq 64$.

3 Data-Generation Models and Inference Goals

Let $\underline{Y} = (Y_{1.1}, \ldots, Y_{1.N}, Y_{2.1}, \ldots, Y_{2.N})$ be the vector of potential variables for the population $\underline{P}$ and $\underline{y} = (y_{1.1}, y_{1.2}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N})$ be the corresponding vector of realizations. We will use this notational convention–upper case letters for random variables and lower case letters for realizations–throughout the paper.

To simplify and to highlight vector component identification, we introduce dot '.' operations and a component-selection bracket '[ ]' notation that is similar to the matrix syntax used in computer languages such as R. Let $\underline{x}$ and $\underline{w}$ be m-dimensional vectors and let k be a scalar. Define

$$\underline{x}.\underline{w} = (x_1.w_1, \ldots, x_m.w_m) \quad \text{and} \quad k.\underline{x} = (k.x_1, \ldots, k.x_m).$$

Consider an m-dimensional vector $\underline{x}$ with components identified by subscripts $a_1, \ldots, a_m$. That is, $\underline{x} = (x_{a_1}, \ldots, x_{a_m})$. Provided $\underline{b} = (b_1, \ldots, b_q)$ has components $b_i \in \{a_1, \ldots, a_m\}$, for each $i = 1, \ldots, q$, the vector $\underline{x}[\underline{b}]$ is defined as $\underline{x}[\underline{b}] = \underline{x}[b_1, \ldots, b_q] = (x_{b_1}, \ldots, x_{b_q})$.

As an example, $\underline{y} = (y_{1.1}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N})$ can be expressed as $\underline{y} = \underline{y}[1.\underline{P}, 2.\underline{P}]$. Similarly $\underline{y}[1.\underline{s}] = (y_{1.s_1}, \ldots, y_{1.s_n})$ and $\underline{y}[\underline{t}.\underline{s}] = (y_{t_1.s_1}, \ldots, y_{t_n.s_n})$. We will also use a notation for averages: As examples,

$$\overline{Y}[t.\underline{P}] = N^{-1}\sum_{i=1}^{N} \underline{Y}[t.i], \qquad \overline{y}[t.\underline{P}] = N^{-1}\sum_{i=1}^{N} \underline{y}[t.i], \qquad \text{and} \qquad \overline{y}[t.\underline{s}] = n^{-1}\sum_{j=1}^{n} \underline{y}[t.s_j].$$

The data-generation models we consider in this paper are based on the following sequential generations:

$$\begin{array}{lll} \underline{y} \leftarrow \underline{Y} & & \text{Here, } \underline{y} = (y_{1.1}, \ldots, y_{1.N}, y_{2.1}, \ldots, y_{2.N}) \\ \underline{s} \leftarrow \underline{S} \mid (\underline{Y} = \underline{y}) & & \text{Here, } \underline{s} = (s_1, \ldots, s_n),\ s_j \in \underline{P},\ \ s_j \neq s_{j'} \\ \underline{t} \leftarrow \underline{T} \mid (\underline{Y} = \underline{y},\ \underline{S} = \underline{s}) & & \text{Here, } \underline{t} = (t_1, \ldots, t_n),\ t_j \in \{1, 2\}. \end{array} \tag{1}$$

The left arrow "←" is read, "is a realization of." The sequencing in (1) is not required to correspond to the temporal sequencing of data generation. It is meant only to be a device for specifying the joint distribution of $(\underline{Y}, \underline{S}, \underline{T})$. For a related discussion, see Rubin (2010, between equations (4) and (5)).

In words, the $2N$ potential deviates in $\underline{y}$ are realized, at least in theory. There are two deviates for each unit in the population–one deviate for each of the two hypothetical parallel worlds corresponding to the two treatments. We sample n subjects $\underline{s}$ from the population. The sampling may depend on potential deviates $\underline{y}$; this dependence often stems from selecting on covariates that are statistically related to the potential variables (see Rubin 2010, between equations (4) and (5)). Finally, we assign treatment levels $\underline{t}$ to units in the sample; that is, we choose which of the two parallel worlds we will observe for each unit in the sample. The treatment assignment may depend on the potential deviates $\underline{y}$ and/or the sampled units $\underline{s}$. However, when mechanical or physical randomization (cf. Fisher 1935, Kempthorne 1955) is used, the treatment assignment can be made to be independent of the potential deviates.

Borrowing from Rubin (2005), we will refer to the potential variables $\underline{Y}$ and values $\underline{y}$ as the "science," to differentiate them from the "selection" variables $(\underline{S}, \underline{T})$ and values $(\underline{s}, \underline{t})$. The science portion describes how things behave in the two parallel worlds and the selection portion determines how we go about observing this behavior. Owing to the sampling and treatment assignment (the selection), we do not observe the entire vector of potential deviates $\underline{y}$ (the science). Indeed, the "fundamental problem of causal inference" rules out the possibility of fully observing the $2N$-dimensional data vector $\underline{y}$. Instead we observe only the n-dimensional sub-vector

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}].$$

The inference goal of this paper can be stated succinctly as follows...

Inference Goal. Use the observed data $\underline{y}[\underline{t}.\underline{s}]$ from a comparative experiment to reduce uncertainty about one of the three targets: the vector $\underline{y}[1.\underline{s}, 2.\underline{s}]$, the vector $\underline{y}[1.\underline{P}, 2.\underline{P}]$, or the distribution of $\underline{Y}$.

4 Treatment Effects and "No-Treatment-Effect" Hypotheses

4.1 Treatment Effects

We began this paper with the question of whether there was a treatment effect. Of course, this begs another question: What exactly is a "treatment effect"?

In a comparative experiment, a treatment effect can be viewed as some measure of the difference between the response ($\underline{Y}$) distribution or response values ($\underline{y}$) for treatment level 1 and the response distribution or response values for treatment level 2. The potential variables viewpoint lends itself to intuitively-appealing candidate definitions of such treatment effects (cf. Cox 1958, Rubin 1990, 2005, 2010). Some of the candidates considered in this paper are:

Realized Unit-Specific Effects:	$\underline{y}[1.s_j] - \underline{y}[2.s_j], j = 1, \ldots, n$ or $\underline{y}[1.i] - \underline{y}[2.i], i = 1, \ldots, N$
Expected Unit-Specific Effects:	$E(\underline{Y}[1.i]) - E(\underline{Y}[2.i]), i = 1, \ldots, N$
Realized Aggregate Effects:	$\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$ or $\overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}]$
Expected Aggregate Effects:	$E(\overline{Y}[1.\underline{P}]) - E(\overline{Y}[2.\underline{P}])$

To take one example, the realized unit-specific treatment effect $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ is simply the difference between unit s_j's responses under two scenarios or two parallel worlds–in one world the unit is exposed to treatment 1 and in the other world the unit is exposed to treatment 2.

Of course, treatment effects need not be defined in terms of simple differences, arithmetic averages, or means of distributions. As an example of another expected unit-specific effect, consider $median(\underline{Y}[1.i]) - median(\underline{Y}[2.i])$. As another example, if $\underline{Y}[t.i] \sim F_{t.i}$ (cdf) then a general expected treatment effect has the functional form $\delta(F_{1.i}, F_{2.i})$, where $\delta()$ is some distance measure. Other examples, not considered in this paper, include realized unit-specific effects, such as $(\underline{y}[2.s_j] - \underline{y}[1.s_j])/\underline{y}[1.s_j]$, and realized aggregate effects, such as $\|\underline{y}[1.\underline{s}] - \underline{y}[2.\underline{s}]\|$ or $var(\underline{y}[1.\underline{s}]) - var(\underline{y}[2.\underline{s}])$ or $\dfrac{\overline{y}[2.\underline{s}] - \overline{y}[1.\underline{s}]}{\overline{y}[1.\underline{s}]}$, etc.

Unfortunately, none of the treatment effects mentioned above is observable. The expected effects cannot be observed because the distribution of $\underline{Y}$ is not completely known. The realized effects cannot be observed because, by the fundamental problem of causal inference, only one of the realizations, for example, either $\underline{y}[1.s_j]$ or $\underline{y}[2.s_j]$, can be observed. Fortunately, this does not preclude unbiased estimation of these unobservable treatment effects, as we point out below.

In the potential-variables causal literature, the treatment effects defined above would be considered causal effects provided certain assumptions hold (e.g. Rubin 1990, 2005, 2010). To avoid the ongoing debate about the nature of causality, we will refrain from referring to treatment effects as causal effects.

4.2 "No-Treatment-Effect" Hypotheses

Corresponding to each treatment effect definition, there is a "no-treatment-effect" hypothesis. As examples,

$$H_0^U : \quad \underline{Y}[1.i] = \underline{Y}[2.i], \text{with probability 1}, \quad i = 1, \ldots, N.$$

$$H_0^{EU} : \quad \underline{Y}[1.i] \ \sim \ \underline{Y}[2.i], \quad i = 1, \ldots, N. \qquad \text{Herein, "}\sim\text{" means "distributed as."}$$

$$H_0^{EU.1} : \quad E(\underline{Y}[1.i]) = E(\underline{Y}[2.i]), \quad i = 1, \ldots, N.$$

$$H_0^{EU.2} : \quad median(\underline{Y}[1.i]) = median(\underline{Y}[2.i]), \ i = 1, \ldots, N.$$

$$H_0^{RU} : \quad \underline{y}[1.i] = \underline{y}[2.i], \quad i = 1, \ldots, N.$$

$$H_0^{RA} : \quad \overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}].$$

$$H_0^{RU.s} : \quad \underline{y}[1.s_j] = \underline{y}[2.s_j], \quad j = 1, \ldots, n.$$

$$H_0^{RA.s} : \quad \overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}].$$

The indentations are used to denote nesting. For example, both H_0^{EU} and H_0^{RU} are implied by H_0^U. Similarly, $H_0^{RA.s}$ is implied by $H_0^{RU.s}$. The superscripts remind us of the type of treatment effect used in the hypothesis. For example, the hypothesis H_0^{EU} uses Expected Unit-specific effects, and $H_0^{RA.s}$ uses Realized Aggregate (over $s$ample $\underline{s}$) effects.

5 Inference Approaches

The $(\underline{y}, \underline{s}, \underline{t})$ components in the observed data $\underline{y}[\underline{t}.\underline{s}]$ are viewed as outcomes of the sequential generations of (1). The complete, but only partially observed, data $\underline{y}$ is a realization of the $2N$-dimensional vector of potential variables $\underline{Y}$.

As stated previously, the inference goal is to use the observed data $\underline{y}[\underline{t}.\underline{s}]$ to reduce uncertainty about one of three targets: the distribution of $\underline{Y}$, the vector $\underline{y}[1.\underline{P}, 2.\underline{P}]$, or the vector $\underline{y}[1.\underline{s}, 2.\underline{s}]$. The choice of inference approach depends on which of these targets we are interested in and it depends on what assumptions we can reasonably make about the joint distribution of $(\underline{Y}, \underline{S}, \underline{T})$, where $\underline{Y}$ is the "science" variable and $(\underline{S}, \underline{T})$ are the "selection" variables. More specifically, $\underline{S}$ is the "sampling" variable and $\underline{T}$ is the treatment "randomization" variable. In this paper, we consider three candidate inference approaches.

Science-Based Inference. With the science-based approach, we condition on the selection (only $\underline{Y}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t}) \quad \sim \quad \underline{Y}[\underline{t}.\underline{s}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t})$$

to carry out inferences about the distribution of $\underline{Y}$. (The discussion section describes more general inferences.)

No assumptions about the $(\underline{S}, \underline{T})$ distribution are made, except that sampling is done without replacement so that $\underline{S}$ generates a sample made up of distinct units. Science-based inference is simplified when the selection is carried out independently of the science, in symbols, $(\underline{S}, \underline{T}) \perp \underline{Y}$. In this special setting, the observed data $\underline{y}[\underline{t}.\underline{s}]$, which is a realization of the *select* random variables $\underline{Y}[\underline{T}.\underline{S}]$, also can be viewed as a realization of the *selected* random variables $\underline{Y}[\underline{t}.\underline{s}]$. It follows that we need only model the [unconditional] distribution of the science $\underline{Y}$.

Selection-Based Inference. With the selection-based approach, we condition on the science (only $(\underline{S}, \underline{T})$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}) \quad \sim \quad \underline{y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y})$$

to carry out inferences about $\underline{y}[1.\underline{P}, 2.\underline{P}]$.

No assumptions about the $\underline{Y}$ distribution are made. Selection-based inference is simplified when the selection is carried out independently of the science, that is, $(\underline{S}, \underline{T}) \perp \underline{Y}$. In this case, we need only specify the [unconditional] distribution of the selection $(\underline{S}, \underline{T})$.

Randomization-Based Inference. With the randomization-based approach, we condition on both the science and the sample (only $\underline{T}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \quad \sim \quad \underline{y}[\underline{T}.\underline{s}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s})$$

to carry out inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

No assumptions about the $(\underline{S}, \underline{Y})$ distribution are made. In particular, the sampling is allowed to depend on the response/science values. Randomization-based inference is simplified when the randomization is conditionally independent of the science, that is, $\underline{T} \;\perp\; \underline{Y} \mid \underline{S}$. In this case, we need only specify the distribution of $\underline{T} | (\underline{S} = \underline{s})$.

These three inference approaches are described in separate sections below.

6 Test Statistics

With the exception of the Wilcoxon rank sum statistic (denoted $W(\underline{X}) = W^*(\underline{R})$ below), all the other test statistics considered in this paper are based on the following difference functions:

$$\begin{aligned} D(\underline{y}, \underline{s}, \underline{t}; \underline{w}) \quad &\equiv \; n^{-1} \textstyle\sum_{j=1}^{n} \displaystyle \frac{\underline{y}[1.s_j] \mathbb{1}(\underline{t}.\underline{s} \ni 1.s_j)}{\underline{w}[1.s_j]} \; - \; n^{-1} \sum_{j=1}^{n} \frac{\underline{y}[2.s_j] \mathbb{1}(\underline{t}.\underline{s} \ni 2.s_j)}{\underline{w}[2.s_j]}. \\ D(\underline{y}, \underline{s}, \underline{t}; \underline{w}, \underline{P}) \; &\equiv \; N^{-1} \textstyle\sum_{i=1}^{N} \displaystyle \frac{\underline{y}[1.i] \mathbb{1}(\underline{t}.\underline{s} \ni 1.i)}{\underline{w}[1.i]} \; - \; N^{-1} \sum_{i=1}^{N} \frac{\underline{y}[2.i] \mathbb{1}(\underline{t}.\underline{s} \ni 2.i)}{\underline{w}[2.i]}. \end{aligned} \tag{2}$$

Here $\mathbb{1}(\cdot)$ is the indicator function and candidate definitions of the $\underline{w}$ components are

$$\begin{array}{lll} \underline{w}[t.s_j] & = \underline{p}_{\underline{s}}[t.s_j] & \equiv n^{-1}n_t, \quad j=1,\ldots,n. \\ \underline{w}[t.s_j] & = \underline{\pi}_{\underline{s}}[t.s_j] & \equiv E(\mathbb{1}(\underline{T}.\underline{S} \ni t.s_j) \mid \underline{S}=\underline{s}) = P(\underline{T}.\underline{S} \ni t.s_j | \underline{S}=\underline{s}), \; j=1,\ldots,n. \\ \underline{w}[t.i] & = \underline{\pi}_{\underline{P}}[t.i] & \equiv E(\mathbb{1}(\underline{T}.\underline{S} \ni t.i)) = P(\underline{T}.\underline{S} \ni t.i), \quad i=1,\ldots,N. \end{array}$$

Sections 8 and 9 below explain that the latter two $\underline{w}$ components are first-order inclusion probabilities (cf Särndal et al. 1992), using language from finite-population sampling theory.

For convenience, let $\underline{x} \equiv \underline{y}[\underline{t}.\underline{s}]$ and $\underline{X} \equiv \underline{Y}[\underline{t}.\underline{s}]$. That is, $\underline{X}$ is the selected random variable, not the select random variable $\underline{Y}[\underline{T}.\underline{S}]$. In general, $\underline{x}$ is a realization of $\underline{X}|(\underline{S}=\underline{s},\underline{T}=\underline{t})$, but unless $(\underline{S},\underline{T}) \perp \underline{Y}$, it is NOT a realization of $\underline{X}$.

We will make use of the difference functions in (2) to define the following test statistics...

Science-Based Test Statistics:

$$\begin{array}{ll} D(\underline{X}) & \equiv D(\underline{Y},\underline{s},\underline{t};\underline{p}_{\underline{s}}), \quad T(\underline{X}) \equiv \dfrac{D(\underline{X})}{SE(D(\underline{X}))}, \quad T_p(\underline{X}) \equiv \dfrac{D(\underline{X})}{SE_p(D(\underline{X}))}, \quad \text{and} \\ W(\underline{X}) & \equiv \sum_{j=1}^{n} \; R_j \mathbb{1}(t_j=1) \equiv W^*(\underline{R}), \\ & \quad \text{where} \;\; R_j = rank(X_j), \text{ the rank of } X_j \text{ out of the } n \text{ components in } \underline{X}. \end{array} \tag{3}$$

Randomization-Based Test Statistics:

$$D(\underline{T}) \equiv D(\underline{y},\underline{s},\underline{T};\underline{\pi}_{\underline{s}}) \qquad \text{and} \qquad Z(\underline{T}) \equiv \frac{D(\underline{T})}{SE(D(\underline{T}))}. \tag{4}$$

Selection-Based Test Statistics:

$$D(\underline{S},\underline{T}) \equiv D(\underline{y},\underline{S},\underline{T};\underline{\pi}_{\underline{P}},\underline{P}) \quad \text{and} \quad Z(\underline{S},\underline{T}) \equiv \frac{D(\underline{S},\underline{T})}{SE(D(\underline{S},\underline{T}))}. \tag{5}$$

The standard error used in the statistic T is defined as $SE(D(\underline{X})) = \sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}$, where $\widehat{\sigma}_t^2 = (n_t-1)^{-1}\sum_{j:t_j=t}\left(\underline{Y}[t.s_j] - n_t^{-1}\sum_{j:t_j=t}\underline{Y}[t.s_j]\right)^2$, which is the commonly used sample variance based on $\underline{Y}[t.s_j]$ for units s_j with $t_j=t$. The standard error used in the statistic T_p is defined as $SE_p(D(\underline{X})) = \sqrt{\widehat{\sigma}^2(\frac{1}{n_1}+\frac{1}{n_2})}$, where $\widehat{\sigma}^2 = \dfrac{(n_1-1)\widehat{\sigma}_1^2 + (n_2-1)\widehat{\sigma}_2^2}{n_1+n_2-2}$ is the commonly-used pooled estimator. The standard errors, $SE(D(\underline{T}))$ and $SE(D(\underline{S},\underline{T}))$, are based on the randomization distribution $\underline{T}|(\underline{S}=\underline{s})$ and the selection distribution $(\underline{S},\underline{T})$, respectively. These standard errors are computed using sampling theory in a manner related to Neyman's approach (Neyman 1923, see also Rubin 1990 and Gadbury 2001).

Although somewhat disguised, $D(\underline{X})$ is simply the difference between the two sample averages, the "$\overline{Y}_1 - \overline{Y}_2$" of textbooks, and $T(\underline{X})$ and $T_p(\underline{X})$ are the commonly-used two-sample t test statistics. In textbook symbols,

$$T(\underline{X}) = \text{“}\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\text{”} \quad \text{and} \quad T_p(\underline{X}) = \text{“}\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_p^2(\frac{1}{n_1}+\frac{1}{n_2})}}\text{”}.$$

In contrast, $D(\underline{T})$ is not generally the simple difference between two sample averages unless the selection probability $\underline{\pi}_{\underline{s}}[t.s_j] = P(\underline{T}.\underline{S} \ni t.s_j | \underline{S}=\underline{s}) = n_t/n$. Similarly, $D(\underline{S},\underline{T})$ is not generally the simple difference between two sample averages.

7 Science-Based Inference

With the science-based approach, we condition on the selection (only $\underline{Y}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t}) \quad \sim \quad \underline{Y}[\underline{t}.\underline{s}] \mid (\underline{S} = \underline{s}, \underline{T} = \underline{t})$$

to carry out inferences about the distribution of $\underline{Y}$.

We will consider the following candidate assumptions.
$A_1 : (\underline{S}, \underline{T}) \;\perp\; \underline{Y}$;
$A_2 : \underline{Y}[1.i, 2.i],\ i = 1, \ldots, N,$ are independent; $A_3 : \underline{Y}[t.i] \;\sim\; F_t, \quad i = 1, \ldots, N,\ t = 1, 2$;
$A_4 : F_t \;\in\; \{\text{continuous cdfs}\}$; $A_5 : F_t \;\in\; \{N(\mu_t, \sigma_t^2) \text{ cdfs}\}$; $A_6 : F_t \;\in\; \{N(\mu_t, \sigma^2) \text{ cdfs}\}$;
$A_7 : F_t \;\in\; \{\text{cdfs with mean and variance } (\mu_t, \sigma_t^2)\}$.

By exchangeability arguments, the independence and identically distributed Assumptions A_2 and A_3 are not as restrictive as they may initially appear, because for example, the units in $\underline{P}$ are arbitrarily assigned identifiers. Indeed, by exchangeability arguments, it is often reasonable to assume the stronger condition that the bivariate vectors $\underline{Y}[1.i, 2.i]$ are IID. Generally, the more tenuous assumptions are A_1, that the selection is carried out independently of the science, and assumptions A_4–A_7, that the model for the distribution of $\underline{Y}$ is correctly specified.

Assumption A_1 is equivalent to the two assumptions, $\underline{T} \perp \underline{Y} | \underline{S}$ and $\underline{S} \perp \underline{Y}$. When mechanical randomization is used to assign treatments to the sampled units, the first assumption can be made tenable. However, the reasonableness of the second assumption, that the sampling variable $\underline{S}$ is independent of $\underline{Y}$, is often questionable in practice. For example, with haphazard or convenience sampling, rather than probability sampling, it often turns out that $\underline{S}$ and $\underline{Y}$ are not independent. The dependence typically stems from sampling on the basis of covariates that are related to $\underline{Y}$.

Under the simplifying (but often untenable!) assumption A_1, we have that $\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{t}.\underline{s}]$. This implies that we need only correctly specify the [unconditional] distribution of $\underline{Y}$. We need not specify a more complicated model for $\underline{Y} | (\underline{S} = \underline{s}, \underline{T} = \underline{t})$. In this simplified setting, it is convenient to use the notation introduced above,

$$\underline{x} \equiv \; \underline{y}[\underline{t}.\underline{s}] \quad \text{and} \quad \underline{X} \equiv \underline{Y}[\underline{t}.\underline{s}].$$

We will also make use of the following permutation set notations,

$$\Pi_n = \{\text{permutations of } \{1, \ldots, n\}\} \quad \text{and} \quad \Pi(\underline{x}) = \{\underline{x}[\underline{\pi}] : \; \underline{\pi} \in \Pi_n\}.$$

7.1 Permutation Test

Consider the no-treatment-effect hypothesis

$$H_0^{EU} : \;\; \underline{Y}[1.i] \;\sim\; \underline{Y}[2.i], \;\; i = 1, \ldots, N$$

and assumptions $A_1 : (\underline{S}, \underline{T}) \perp \underline{Y}$; $A_2 : \underline{Y}[1.i, 2.i]$ are independent; and $A_3 : \underline{Y}[t.i] \;\sim\; F_t$. When $H_0 \equiv (A_1, A_2, A_3, H_0^{EU})$ holds, we have the following:

- H_0^{EU} can be expressed as $F_1 = F_2$.
- $X_j \;\; IID \;\sim\; F_1, \;\; j = 1, \ldots, n.$
- $\underline{x} \;\leftarrow\; \underline{X}|(\underline{X} \in \Pi(\underline{x})) \;\sim\; \underline{X}^{(c)}$, where $P_{H_0}(\underline{X}^{(c)} = \underline{x}') = \dfrac{\sum_{\underline{\pi} \in \Pi_n} \; \mathbb{1}(\underline{x}[\underline{\pi}] = \underline{x}')}{n!} \mathbb{1}(\underline{x}' \in \Pi(\underline{x}))$.
- $cpval_1 D(\underline{x}) = P_{H_0}(D(\underline{X}^{(c)}) \geq D(\underline{x})) = P_{H_0}(\underline{X}^{(c)} \in \{\underline{x}' \in \Pi(\underline{x}) : \;\; D(\underline{x}') \geq D(\underline{x})\})$ is a computable one-sided conditional p-value. Here, $D(\underline{x})$ is the observed value of $D(\underline{X})$, of (3).
- $cpvalD(\underline{x}) = P_{H_0}(|D(\underline{X}^{(c)})| \;\geq\; |D(\underline{x})|)$ is a computable two-sided p-value.
- The test, reject H_0 if and only if $(cpvalD(\underline{X}^{(c)}) \leq \alpha)$ is observed, has size $\leq \alpha$.

This conditional test, which is aptly called a permutation test, is based on ideas originating in Pitman (1937, 1938), cf. Ernst (2004). That $cpval_1 D(\underline{x})$ and $cpvalD(\underline{x})$ are computable follows because the distribution of $\underline{X}^{(c)}$ is known and $D(\underline{x}')$, defined in (3), can be re-expressed as

$$D(\underline{x}') = n_1^{-1} \sum_{j:t_j=1} x_j' \;\; - \;\; n_2^{-1} \sum_{j:t_j=2} x_j' \; .$$

If $cpvalD(\underline{x}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\leq \alpha$) event has occurred or H_0 is false. If we reject H_0 but A_1, A_2, and A_3 are assumed to be true, then we reject H_0^{EU}; that is, we have statistical evidence that $F_1 \neq F_2$.

This permutation test based on $D(\underline{X})$ is tailored to detect location shift differences between F_1 and F_2. To detect other differences between F_1 and F_2, such as scale differences, an alternative to $D(\underline{X})$ should be used. In theory, any alternative test statistic can be used.

7.2 Wilcoxon Rank Sum Test

Consider the no-treatment-effect hypothesis

$$H_0^{EU} : \;\; \underline{Y}[1.i] \;\sim\; \underline{Y}[2.i], \;\; i = 1, \ldots, N$$

and assumptions A_1–A_3 and $A_4 : \; F_t \in \{\text{continuous cdfs}\}$.
When $H_0 \equiv (A_1, A_2, A_3, A_4, H_0^{EU})$ holds, we have the following:

- H_0^{EU} can be expressed as $F_1 = F_2$.
- $X_j \;\; IID \;\sim\; F_1$, a continuous cdf, $\quad j = 1, \ldots, n.$
- $W(\underline{X}) = W^*(\underline{R})$ of (3) has a known distribution because $P_{H_0}(\underline{R} = \underline{r}') = \dfrac{\mathbb{1}(\underline{r}' \in \Pi_n)}{n!}$.
- $pval_1 W(\underline{x}) = P_{H_0}(W(\underline{X}) \geq W(\underline{x})) = P_{H_0}(\underline{R} \in \{\underline{r}' \in \Pi_n : \;\; W^*(\underline{r}') \geq W^*(\underline{r})\})$ is a computable one-sided p-value. Here, $\underline{r} = rank(\underline{x})$.
- $pvalW(\underline{x}) = 2 \;\min\{P_{H_0}(W(\underline{X}) \geq W(\underline{x})), \;\; P_{H_0}(W(\underline{X}) \leq W(\underline{x}))\}$ is a computable two-sided p-value.
- The test, Reject H_0 if and only if $(pvalW(\underline{X}) \leq \alpha)$ is observed, has size $\leq \alpha$.

If $pvalW(\underline{x}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\leq \alpha$) event has occurred or H_0 is false. If we reject H_0 but A_1, A_2, A_3, and A_4 are assumed to be true, then we reject H_0^{EU}; that is, we have statistical evidence that $F_1 \neq F_2$.

The Wilcoxon rank sum test is tailored to detect location shift differences between F_1 and F_2. It may not be very powerful when the shapes of F_1 and F_2 are quite different. Indeed, the test can alternatively be viewed as a test of $H_0 = (A_1, A_2, A_3, A_4, L, H_0^{EU.2})$, where $L : F_1(u) = F_2(u + \Delta)$ and $H_0^{EU.2}$: $median(Y[1.i]) = median(Y[2.i]), i = 1, \ldots, N$, see Hollander and Wolfe (1973:67) for such a formulation. Under this H_0, $H_0^{EU.2}$ is equivalent to $\Delta = 0$ and rejection of H_0, when $A_1 - A_4$ and L are assumed true, implies there is statistical evidence that $\Delta \neq 0$, i.e. $median(F_1) \neq median(F_2)$.

7.3 Two-Sample t Test (Welch's Approximation)

Consider the no-treatment-effect hypothesis

$$H_0^{EU.1} : \quad E(\underline{Y}[1.i]) = E(\underline{Y}[2.i]), \quad i = 1, \ldots, N$$

and assumptions A_1–A_3 and $A_5 :\ F_t \in \{\ N(\mu_t, \sigma_t^2) \text{ cdfs } \}$.
When $H_0 \equiv (A_1, A_2, A_3, A_5, H_0^{EU.1})$ holds, we have the following:

- $H_0^{EU.1}$ can be expressed as $\mu_1 = \mu_2$.
- $X_j\ indep \ \sim\ N(\mu_1, \sigma_{t_j}^2), \quad j = 1, \ldots, n.$
- $D(\underline{X}) \ \sim\ N(0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}).$
- $T(\underline{X}) \ \sim\ approx\ \ t(r)$, where r is Welch's (1938) approximate degrees of freedom.
- $pval_1T(\underline{x}) = P_{H_0}(T(\underline{X}) \geq T(\underline{x})) \ \approx\ P(t(r) \ \geq T(\underline{x})) \equiv apval_1T(\underline{x})$.
- $pvalT(\underline{x}) = P_{H_0}(|T(\underline{X})| \geq |T(\underline{x})|) \ \approx\ 2P(t(r) \ \geq |T(\underline{x})|) \equiv apvalT(\underline{x})$ is an approximate two-sided p-value.
- The test, Reject H_0 if and only if $(apvalT(\underline{X}) \leq \alpha)$ is observed, has size $\approx \alpha$.

If $apvalT(\underline{x}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\approx \alpha$) event has occurred or H_0 is false. If we reject H_0 but A_1, A_2, A_3, and A_5 are assumed to be true, then we reject $H_0^{EU.1}$; that is, we have statistical evidence that $\mu_1 \neq \mu_2$.

Practitioners are well aware that this same t test can be used to test the less restrictive $H_0 = (A_1, A_2, A_3, A_7, H_0^{EU.1})$, where $A_7 :\ F_t \in \{\text{cdfs with mean and variance } (\mu_t, \sigma_t^2)\}$; that is, we do not assume Normality. By the central limit theorem and Slutsky's theorem, the size will still be approximately α. However, the reasonableness of the approximation depends in a complicated way on the unknown F_t's and the sample sizes n_1 and n_2. Practically speaking, the sample sizes must be relatively large and the distributions must not be too skewed.

A clarification is in order here. In the hypothesis $\mu_1 = \mu_2$, the means μ_1 and μ_2 are commonly referred to as "population" means. We think it more appropriate to call them "process" or "probability" means, or expected values. On the one hand, the average $\overline{y}[1.\underline{P}]$ *is* a population mean–it is an average of all the $\underline{y}[1.i]$ values for units i in finite population $\underline{P}$. On the other hand, μ_1 is not usually an average of values for units in some finite population. Rather, μ_1 is the expected value, or probability mean, of $\underline{Y}[1.i]$ under the IID model for the science. As an empirical interpretation, if the random experiment generating the science were repeated over and over again, the long-run average of the $\underline{Y}[1.i]$ values would be μ_1.

7.4 Two-Sample t Test (Pooled)

Consider the no-treatment-effect hypothesis

$$H_0^{EU.1}: \quad E(\underline{Y}[1.i]) = E(\underline{Y}[2.i]), \quad i = 1, \dots, N$$

and assumptions A_1–A_3 and $A_6: \ F_t \in \{ \ N(\mu_t, \sigma^2) \text{ cdfs } \}$.
When $H_0 \equiv (A_1, A_2, A_3, A_6, H_0^{EU.1})$ holds, we have the following:

- $H_0^{EU.1}$ can be expressed as $\mu_1 = \mu_2$.
- $X_j \ IID \ \sim \ N(\mu_1, \sigma^2), \quad j = 1, \dots, n.$
- $D(\underline{X}) \ \sim \ N(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})).$
- $T_p(\underline{X}) \ \sim \ t(\nu)$, where $\nu = n_1 + n_2 - 2$.
- $pval_1 T_p(\underline{x}) = P_{H_0}(T_p(\underline{X}) \geq T_p(\underline{x})) \ = \ P(t(\nu) \ \geq T_p(\underline{x}))$ is computable.
- $pval T_p(\underline{x}) = P_{H_0}(|T_p(\underline{X})| \geq |T_p(\underline{x})|) = 2P(t(\nu) \geq |T_p(\underline{x})|)$ is a computable two-sided p-value.
- The test, Reject H_0 if and only if $(pval T_p(\underline{X}) \leq \alpha)$ is observed, has size equal to α.

If $pval T_p(\underline{x}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $= \alpha$) event has occurred or H_0 is false. If we reject H_0 but A_1, A_2, A_3, and A_6 are assumed to be true, then we reject $H_0^{EU.1}$; that is, we have statistical evidence that $\mu_1 \neq \mu_2$.

8 Randomization-Based Inference

With the randomization-based approach, we condition on both the science and the sample
(only $\underline{T}$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \ \leftarrow \ \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s}) \quad \sim \quad \underline{y}[\underline{T}.\underline{s}] \mid (\underline{Y} = \underline{y}, \underline{S} = \underline{s})$$

to carry out inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$.

We will consider the candidate assumptions
$B_1: \ \underline{T} \ \perp \ \underline{Y} \mid \underline{S}$
B_2 : The distribution of $\underline{T} | (\underline{S} = \underline{s})$ is completely known and satisfies...

$$P(\underline{T}.\underline{S} \ni t.s_j \mid \underline{S} = \underline{s}) \ > \ 0, \quad j = 1, \dots, n, \quad t = 1, 2.$$
$$P(\underline{T}.\underline{S} \ni t.s_j, \ \underline{T}.\underline{S} \ni t'.s_{j'} \mid \underline{S} = \underline{s}) \ > \ 0, \quad \text{unless } t \neq t' \text{ and } \ j = j'.$$

In words, B_1 implies that the randomization (i.e. treatment assignment) is conditionally independent of the science, given the sample. The use of mechanical randomization makes this assumption tenable.

In B_2, the probabilities are called first- and second-order inclusion probabilities for the random sample, namely $\underline{T}.\underline{S}|(\underline{S} = \underline{s})$, taken from $(1.\underline{s}, 2.\underline{s})$. Assumption B_2 imposes constraints on these inclusion probabilities. The positive first-order inclusion probabilities imply that "proper" randomization is used to assign treatments, that is, each unit in the sample has a positive probability of receiving either treatment; we say that the comparative experiment has a completely randomized design. Put another way, $\underline{T}.\underline{S}|(\underline{S} = \underline{s})$ is a *probability* sample from $(1.\underline{s}, 2.\underline{s})$. Because the same unit cannot be assigned different treatments, the second-order inclusion probabilities with $t \neq t'$ and $j = j'$ are 0. This implies that the probability sample is non-measurable, to use language from sampling theory (cf. Särndal et al. 1992:32-33). This non-measurability has implications for the estimation of the variance of the statistic $D(\underline{T})$, as Neyman was fully aware of in 1923.

Under B_1 and B_2, the statistic $D(\underline{T})$ defined in (4) is an unbiased Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952, Särndal et al. 1992:43) of the estimand $\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$. In symbols,

$$E(D(\underline{T}) \mid \underline{S} = \underline{s}) \quad = \quad \overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]. \tag{6}$$

This unbiased estimator owes its existence to the fact that proper randomization is used, whereby each unit in the sample has a positive probability of receiving either treatment. Without proper randomization, such an estimator generally would not exist. These facts give an operational explanation as to why randomization is important for estimating treatment effects.

The variance, $var(D(\underline{T}) \mid \underline{S} = \underline{s})$, can be computed using sampling theory as described in Särndal et al. (1992). However, finding a reasonable estimator of this variance is more difficult because of the 0 second-order inclusion probabilities. Toward this end, Neyman (1923) derived a reasonable estimator of a tight upper bound for the variance under simplifying assumptions on the inclusion probabilities (Rubin 1990, Gadbury 2001, see Copas 1973 for a related result). It is useful to note that the variance attains this upper bound when $\underline{y}[1.s_j] = \underline{y}[2.s_j] + constant,\ j = 1, \ldots, n$. In this paper, we use the Neyman-type estimator of variance. The square root of this estimator, the standard error, will be denoted simply $SE(D(\underline{T}))$.

Under B_1 and B_2 and conditional on $(\underline{S} = \underline{s})$, central limit theorems can be used to argue that quite generally

$$\frac{D(\underline{T}) - (\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}])}{SE(D(\underline{T}))} \quad \sim \quad approx \quad N(0, 1).$$

The approximation generally improves as the number of support points in $\underline{T}|(\underline{S} = \underline{s})$ increases. However, when the differences $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable, the unit variance in the approximation can be a substantial over estimate (see Gadbury, 2001) and when $\underline{y}[1.s_j] - \underline{y}[2.s_j] = constant$ the unit variance can be a slight under estimate when the sample sizes are small (based on observations from the simulation study carried out for this paper).

8.1 Fisher-Type Randomization Test: What Would Fisher Do?

Fisher tacitly (Welch 1937, Rubin 1990, 2005) used the no-unit-specific-effects hypothesis in this setting. That is, it can be presumed that to Fisher, the no-treatment-effect hypothesis had the form:

$$H_0^{RU.s}: \quad \underline{y}[1.s_j] = \underline{y}[2.s_j], \quad j = 1, \ldots, n.$$

When $H_0 = (B_1, B_2, H_0^{RU.s})$ holds, we have the following:

- $E(D(\underline{T}) \mid \underline{S} = \underline{s}) = 0.$
- The one-sided p-value

$$pval_1 D(\underline{t}) = P_{H_0}(D(\underline{T}) \geq D(\underline{t}) \mid \underline{S} = \underline{s}) = P_{H_0}(\underline{T} \in \ \{\underline{t}' : \ D(\underline{t}') \geq D(\underline{t})\} | \underline{S} = \underline{s})$$

  is computable because $D(\underline{t}')$ is computable for every $\underline{t}'$. This follows because, under $H_0^{RU.s}$, $\underline{y}[1.\underline{s}, 2.\underline{s}]$ is computable given the observed data $\underline{y}[\underline{t}.\underline{s}]$.
- $pvalD(\underline{t}) = P_{H_0}(|D(\underline{T})| \geq |D(\underline{t})| \mid \underline{S} = \underline{s})$ is a computable two-sided p-value.
- The test, reject H_0 if and only if $(pvalD(\underline{T}) \leq \alpha)$ is observed, has size $\leq \alpha$.

This test is called a Fisher-type randomization test because it is based on the randomization approach and it was introduced by Fisher (1935).

If $pvalD(\underline{t}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\leq \alpha$) event has occurred or H_0 is false. If we reject H_0 but B_1 and B_2 are assumed to be true, then we reject $H_0^{RU.s}$; that is, we have statistical evidence that for at least one unit s_j, $\underline{y}[1.s_j] \neq \underline{y}[2.s_j]$.

This Fisher-type randomization test based on $D(\underline{T})$ is tailored to detect differences between $\overline{y}[1.\underline{s}]$ and $\overline{y}[2.\underline{s}]$. To detect other differences, such as scale differences between the $\underline{y}[1.\underline{s}]$ and $\underline{y}[2.\underline{s}]$, an alternative to $D(\underline{T})$ should be used.

Attractive features of this Fisher-type randomization test include the following: it has size guaranteed to be no larger than α, it is valid when the sampling depends on the science ($\underline{S} \not\perp \underline{Y}$); it does not require a model for the science $\underline{Y}$; and it does not require an estimate of the variance, $var(D(\underline{T})|\underline{S} = \underline{s})$.

Randomization vs. Permutation P-values: It is clear that this Fisher-type randomization test is conceptually very different from the science-based permutation test. Indeed, as a rule, the randomization p-value $pvalD(\underline{t})$ is numerically different than the permutation conditional p-value $cpvalD(\underline{x})$. There is an exception to this rule. Consider the special case uniform randomization distribution,

$$P(\underline{T} = \underline{t}' | \underline{S} = \underline{s}) = \frac{n_1! n_2!}{n!} \mathbb{1}(\underline{t}' \in \mathcal{T}), \tag{7}$$

where $\mathcal{T}$ is the set of all possible treatment assignments such that n_1 units are assigned treatment 1 and n_2 are assigned treatment 2. In this case, $cpvalD(\underline{x}) = pvalD(\underline{t})$. It is this identity that often leads practitioners to incorrectly conclude that the permutation test is identical to the randomization test. See Ernst (2004) for an interesting discussion.

8.2 Neyman-Type Randomization Test: What Would Neyman Do?

Compared to Fisher, Neyman was apparently more interested in detecting non-zero treatment effects of the aggregate variety, especially $\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$. He apparently found it less practically useful to detect unit-specific effects if the average effect was 0. For this reason, Neyman used the no-average-effect hypothesis (cf. Welch 1937). That is, he viewed the no-treatment-effect hypothesis as

$$H_0^{RA.s}:\ \overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}].$$

Because $H_0^{RA.s} \supset H_0^{RU.s}$, Neyman's approach focused on a narrower set of alternatives than Fisher, thereby opening up the possibility of finding a test with higher power than the Fisher-type randomization test, at least for alternatives of practical (in Neyman's view) interest.

When $H_0 = (B_1, B_2, H_0^{RA.s})$ holds, we have the following:

- $E(D(\underline{T}) \mid \underline{S} = \underline{s}) = 0$.
- Conditional on $(\underline{S} = \underline{s})$, $\quad Z(\underline{T}) = \dfrac{D(\underline{T})}{SE(D(\underline{T}))} \quad \sim \quad approx\ N(0,1)$.
- $pval_1 Z(\underline{t}) = P_{H_0}(Z(\underline{T}) \geq Z(\underline{t}) \mid \underline{S} = \underline{s}) \approx\ P(N(0,1)\ \geq Z(\underline{t})) \equiv apval_1 Z(\underline{t})$. The approximate p-value $apval_1 Z(\underline{t})$ is called a Neyman-type randomization one-sided p-value.
- $pvalZ(\underline{t}) = P_{H_0}(|Z(\underline{T})| \geq |Z(\underline{t})| \mid \underline{S} = \underline{s}) \approx 2P(N(0,1) \geq |Z(\underline{t})|)\ \equiv apvalZ(\underline{t})$. The approximate p-value $apvalZ(\underline{t})$ is a Neyman-type randomization two-sided p-value.
- The test, reject H_0 if and only if $(apvalZ(\underline{T}) \leq \alpha)$ is observed, has size $\approx \alpha$.

This test is called a Neyman-type randomization test because it is based on the randomization approach and ideas in Neyman (1923).

If $apvalZ(\underline{t}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\approx \alpha$) event has occurred or H_0 is false. If we reject H_0 but B_1 and B_2 are assumed to be true, then we reject $H_0^{RA.s}$; that is, we have statistical evidence that $\overline{y}[1.\underline{s}] \neq \overline{y}[2.\underline{s}]$.

Why not use $D(\underline{T})$ rather than $Z(\underline{T})$? It is tempting to think that Neyman would approach the testing problem in Fisher-like fashion and compute a p-value defined as

$$P_{H_0}(D(\underline{T}) \geq D(\underline{t}) \mid \underline{S} = \underline{s}) = P_{H_0}(\underline{T} \in \{\underline{t}':\ D(\underline{t}') \geq D(\underline{t})\}).$$

However, Neyman would have recognized that the set $\{\underline{t}' : D(\underline{t}') \geq D(\underline{t})\}$, and hence the p-value, canNOT be computed under $H_0^{RA.s}$. This computational problem stems from the fact that $\underline{y}[1.\underline{s}, 2.\underline{s}]$ is only partially observed and, unlike under Fisher's more restrictive $H_0^{RU.s}$, is not determined by $\underline{y}[\underline{t}.\underline{s}]$ under the no-average-effect hypothesis $H_0^{RA.s}$. It follows that $D(\underline{t}')$ cannot be computed for any $\underline{t}'$ not equal to the observed $\underline{t}$. Hence, Fisher's p-value approach is not available for testing Neyman's no-average-effect hypothesis.

Unlike the Fisher-type randomization test of $H_0^{RU.s}$, the size of the Neyman test of $H_0^{RA.s}$ is not guaranteed to be less than or equal to α; it is only approximately size α. For smaller n_1 and n_2 and when the more restrictive hypothesis $H_0^{RU.s}$ holds, the Neyman-type randomization test tends to be anti-conservative, with size a bit larger than the nominal α. This follows because the Neyman-type estimator of the variance tends to slightly under-estimate the true variance in this case. For moderate n_1 and n_2 the approximation is usually reasonable provided $D(\underline{T})$ has enough support points with respect to the $\underline{T}|(\underline{S} = \underline{s})$ distribution. We empirically explore this approximation below.

9 Selection-Based Inference

With the selection-based approach, we condition on the science (only $(\underline{S}, \underline{T})$ is random) and use

$$\underline{y}[\underline{t}.\underline{s}] \;\leftarrow\; \underline{Y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y}) \quad \sim \quad \underline{y}[\underline{T}.\underline{S}] \mid (\underline{Y} = \underline{y})$$

to carry out inferences about $\underline{y}[1.\underline{P}, 2.\underline{P}]$.

We will consider the candidate assumptions
C_1 : $(\underline{S}, \underline{T}) \;\perp\; \underline{Y}$
C_2 : The distribution of $(\underline{S}, \underline{T})$ is completely known and satisfies...

$$P(\underline{T}.\underline{S} \ni t.i) \;>\; 0, \quad i = 1, \ldots, N, \quad t = 1, 2.$$
$$P(\underline{T}.\underline{S} \ni t.i, \; \underline{T}.\underline{S} \ni t'.i') \;>\; 0, \quad \text{unless } t \neq t' \text{ and } \; i = i'.$$

In words, C_1 implies that the selection is independent of the science. As discussed in the science-based Section 7, this assumption is not usually tenable in practice because the sampling and science are often dependent. This dependence typically stems from sampling on the basis of covariates that are related to $\underline{Y}$.

As discussed in the randomization-based section, assumption C_2 imposes constraints on first- and second-order inclusion probabilities. In this case, the random sample $\underline{T}.\underline{S}$ is taken from $(1.\underline{P}, 2.\underline{P})$. The assumption implies that each of the $2N$ elements in $(1.\underline{P}, 2.\underline{P})$ has a positive probability of being selected. Thus, the random sample is a probability sample. The 0 second-order inclusion probabilities implies that the probability sample is non-measurable.

Under C_1 and C_2, the statistic $D(\underline{S}, \underline{T})$ defined in (5) is an unbiased Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952, Särndal et al. 1992:43) of the estimand $\overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}]$. In symbols,

$$E(D(\underline{S}, \underline{T})) \quad = \quad \overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}].$$

Just as with $var(D(\underline{T})|\underline{S} = \underline{s})$ in the randomization approach, the variance, $var(D(\underline{S}, \underline{T}))$, can be computed and estimated using sampling theory. The variance estimation, however, is subject to the same problems as in the randomization approach because of the non-measurability of probability sample $\underline{T}.\underline{S}$. Suffice it to say that a reasonable Neyman-type estimator exists. Denote the square root of this estimator by $SE(D(\underline{S}, \underline{T}))$.

Under C_1 and C_2, and using the same arguments as in the randomization approach, we have that quite generally

$$\frac{D(\underline{S}, \underline{T}) - (\overline{y}[1.\underline{P}] - \overline{y}[2.\underline{P}])}{SE(D(\underline{S}, \underline{T}))} \quad \sim \quad approx \quad N(0, 1).$$

The approximation generally improves as the number of support points in $\underline{T}.\underline{S}$ increases. However, when the differences $\underline{y}[1.i] - \underline{y}[2.i]$ are highly variable, the unit variance in the approximation can be a substantial over estimate (see Gadbury, 2001).

9.1 Fisher-Type Selection Test: What Would Fisher Do?

It can be presumed that to Fisher, the no-treatment-effect hypothesis in this setting had the form:

$$H_0^{RU} : \quad \underline{y}[1.i] = \underline{y}[2.i], \quad i = 1, \ldots, N.$$

When $H_0 = (C_1, C_2, H_0^{RU})$ holds, we have the following:

- $E(D(\underline{S}, \underline{T})) = 0.$
- $pval_1 D(\underline{s}, \underline{t}) = P_{H_0}(D(\underline{S}, \underline{T}) \geq D(\underline{s}, \underline{t})) = P_{H_0}((\underline{S}, \underline{T}) \in \{(\underline{s}', \underline{t}') : \ D(\underline{s}', \underline{t}') \geq D(\underline{s}, \underline{t})\})$ is **NOT** computable because $D(\underline{s}', \underline{t}')$ is not computable for any $\underline{s}' \neq \underline{s}$. This follows because for $\underline{s}' \neq \underline{s}$, there is an s_j' such that both $\underline{y}[1.s_j']$ and $\underline{y}[2.s_j']$ are unobserved and hence not computable even under H_0^{RU}.

It follows that a Fisher-type selection test is **not** available in this selection-based setting. Fisher would have to condition on the sample and be content using the randomization-based approach to draw inferences about $\underline{y}[1.\underline{s}, 2.\underline{s}]$, rather than $\underline{y}[1.\underline{P}, 2.\underline{P}]$.

9.2 Neyman-Type Selection Test: What Would Neyman Do?

In analogy to the randomization setting, Neyman would use the no-average-effect hypothesis:

$$H_0^{RA} : \quad \overline{y}[1.\underline{P}] = \overline{y}[2.\underline{P}].$$

When $H_0 = (C_1, C_2, H_0^{RA})$ holds, we have the following:

- $E(D(\underline{S}, \underline{T})) = 0.$
- $Z(\underline{S}, \underline{T}) = \dfrac{D(\underline{S}, \underline{T})}{SE(D(\underline{S}, \underline{T}))} \quad \sim \quad approx\ N(0, 1).$
- $pval_1 Z(\underline{s}, \underline{t}) = P_{H_0}(Z(\underline{S}, \underline{T}) \geq Z(\underline{s}, \underline{t})) \approx P(N(0, 1) \geq Z(\underline{s}, \underline{t})) \equiv apval_1 Z(\underline{s}, \underline{t})$. The approximate p-value $apval_1 Z(\underline{s}, \underline{t})$ is called a Neyman-type selection one-sided p-value.
- $pvalZ(\underline{s}, \underline{t}) = P_{H_0}(|Z(\underline{S}, \underline{T})| \geq |Z(\underline{s}, \underline{t})|) \approx 2P(N(0, 1) \geq |Z(\underline{s}, \underline{t})|) \ \equiv apvalZ(\underline{s}, \underline{t})$
  The approximate p-value $apvalZ(\underline{s}, \underline{t})$ is a Neyman-type selection two-sided p-value.
- The test, Reject H_0 if and only if $(apvalZ(\underline{S}, \underline{T}) \leq \alpha)$ is observed, has size $\approx \alpha$.

This test will be called a Neyman-type selection test because it is a selection-based approach that is based on the ideas in Neyman (1923).

If $apvalZ(\underline{s}, \underline{t}) \leq \alpha$, we reject H_0, but of course either H_0 is true and a rare (probability $\approx \alpha$) event has occurred or H_0 is false. If we reject H_0 but C_1 and C_2 are assumed to be true, then we reject H_0^{RA}; that is, we have statistical evidence that $\overline{y}[1.\underline{P}] \neq \overline{y}[2.\underline{P}]$.

Just as in the randomization setting, the size of the Neyman test of H_0^{RA} is not guaranteed to be less than or equal to α; it is only approximately size α. Remarks regarding the approximation in this selection setting are analogous to those given at the end of Section 8.2, in the randomization setting.

10 Empirical Investigations

10.1 Cell Phone Use Example (Revisited)

The science variable $\underline{Y}[t.i]$ is defined as the reaction time for the i^{th} unit in population $\underline{P}$ when exposed to treatment t. Inference about the science $\underline{Y}$ distribution will be difficult to describe because the sample of 64 students was not taken from any well-defined population $\underline{P}$. For any substantively interesting population, for example, $\underline{P}$ = licensed drivers in Utah, the assumption that $\underline{S} \perp \underline{Y}$ is untenable given the haphazard nature of the sample selection. The untenability of $\underline{S} \perp \underline{Y}$ also implies that it will be difficult to carry out inferences about the population values $\underline{y}[1.\underline{P}, 2.\underline{P}]$ for any substantively interesting population $\underline{P}$. For these reasons, it makes sense to focus on inferences about the 128 potential values in $\underline{y}[1.\underline{s}, 2.\underline{s}]$. That is, it is arguably better to use randomization-based inference for this example.

We assume that the randomization was carried out mechanically so that $\underline{T} \perp \underline{Y}|\underline{S}$ and we assume that the distribution of $\underline{T}|(\underline{S} = \underline{s})$ is uniform in the sense of (7); that is, conditions B_1 and B_2 of Section 8 are assumed to hold. We will use the Fisher-type randomization test to test the no-treatment-effect hypothesis $H_0^{RU.s}$: $\underline{y}[1.s_j] = \underline{y}[2.s_j]$, $j = 1, \ldots, 64$ and the Neyman-type randomization test to test the no-treatment-effect hypothesis $H_0^{RA.s}$: $\overline{y}[1.\underline{s}] = \overline{y}[2.\underline{s}]$.

For these data, we observe

$$D(\underline{t}) = 51.59, \quad Z(\underline{t}) = \frac{51.59}{19.30} = 2.67, \quad pvalD(\underline{t}) = 0.0074, \text{ and } \quad apvalZ(\underline{t}) = 0.0075.$$

Because the Fisher-type randomization p-value $pvalD(\underline{t}) = 0.0074$ is small, we have sufficient evidence to reject $H_0^{RU.s}$; there is statistical evidence that $\underline{y}[1.s_j] \neq \underline{y}[2.s_j]$ for at least one subject in the sample of 64. Because the Neyman-type randomization p-value $apvalZ(\underline{t}) = 0.0075$ is small, we have sufficient evidence to reject $H_0^{RA.s}$; there is statistical evidence that $\overline{y}[1.\underline{s}] \neq \overline{y}[2.\underline{s}]$. In fact, because $D(\underline{t}) = 51.59$ is a Horvitz-Thompson unbiased estimate of $\overline{y}[1.\underline{s}] - \overline{y}[2.\underline{s}]$, see (6), the Neyman test gives statistical evidence that the reaction time values are higher on average when cell phones are used, at least for this sample of 64. In other words, there is statistical evidence of a treatment effect.

For completeness and for comparison purposes, we also give the values of the other commonly used p-values, viz., permutation, Wilcoxon, t(Welch), and t(pooled):

$$cpvalD(\underline{x}) = 0.0074, \quad pvalW(\underline{x}) = 0.0184, \quad apvalT(\underline{x}) = 0.0110, \quad \text{and} \quad pvalT_p(\underline{x}) = 0.0107,$$

Strictly speaking, these are only applicable for science-based inference, so they are of questionable utility for this example. As noted above, because the randomization distribution is uniform, the permutation p-value $cpvalD(\underline{x})$ is numerically (but not conceptually!) identical to the Fisher-type randomization p-value $pvalD(\underline{t})$.

All computations were carried out in R. The author has written code to compute the Neyman-type randomization p-value. The Fisher-type randomization and permutation p-values were approximated using Monte-Carlo estimation (here we used 10^6 simulations) as carried out in `twot.permutation {DAAG}`. The Wilcoxon p-value was computed using `wilcox.test {stats}`. Note that when there are ties, as there are in this example, `wilcox.test` only reports approximate p-values.

10.2 A Small-Scale Simulation Study

This section empirically compares the operating characteristics of the different tests considered in this paper, under a variety of scenarios. All computations were carried out in R, with p-values computed as described at the end of the previous sub-section. The simulated data are generated according to models of the form:

$$\begin{array}{lll} \underline{y}[1.i] & \leftarrow & \underline{Y}[1.i] \;\; IID \;\sim\; [scenario], \\ \underline{y}[2.i] & \leftarrow & \underline{Y}[2.i] \;\;\sim\; [scenario], \;\; i=1,\ldots,N \\ \underline{s} & \leftarrow & \underline{S}|(\underline{Y}=\underline{y}) \;\sim\; P(\underline{S}=\underline{P} \mid \underline{Y}=\underline{y})=1 \\ \underline{t} & \leftarrow & \underline{T}|(\underline{Y}=\underline{y},\underline{S}=\underline{s}) \;\sim\; P(\underline{T}=\underline{t}'|\underline{Y}=\underline{y},\underline{S}=\underline{s}) = \dfrac{n_1!n_2!}{n!}\mathbb{1}(\underline{t}' \in \mathcal{T}) \end{array} \tag{8}$$

where $\mathcal{T}$ is the set of all possible treatment assignments such that n_1 units receive treatment 1 and n_2 receive treatment 2. Looking back at the science-based assumptions of Section 7, we see that A_1 holds, but none of A_2–A_7 are guaranteed to hold. Both the randomization-based assumptions B_1 and B_2 of Section 8 hold, as do both the selection-based assumptions C_1 and C_2 of Section 9. A more extensive simulation would also investigate scenarios where more of the assumptions do not hold.

For data-generation models of the form (8), we have that (i) the randomization- and selection-based approaches are identical because the sample $\underline{S}$ is taken to be equal to the population $\underline{P}$, which also implies that $n = N$; and (ii) the permutation and Fisher-type randomization p-values are numerically (not conceptually!) identical because the randomization distribution is uniform over the set of all possible treatment assignments.

Although the permutation-, Wilcoxon-, and t-tests are science-based approaches, we will estimate their operating characteristics for both the science and randomization (here, randomization=selection) distributions. Similarly, the Fisher- and Neyman-type randomization tests are randomization-based approaches, but we report their operating characteristics for both the science and the randomization distributions. In the tables below, the rows labeled "Randomization" give Monte Carlo estimates of the power of the tests over the distribution $\underline{T}|(\underline{Y}=\underline{y},\underline{S}=\underline{s})$. The rows labeled "Science" give Monte Carlo estimates of the power of the tests over the distribution $\underline{Y}|(\underline{S}=\underline{s},\underline{T}=\underline{t})$. In all cases, the nominal size is set at $\alpha = 0.05$.

Tables 3 - 6 about here.

The simulation results in Tables 3–6 give us a glimpse at the operating characteristics of the tests for a variety of scenarios, labeled "Sc.#." The following summary focuses on comparisons between the Fisher- and Neyman-type randomization tests, but the table entries afford broader comparisons.

For small n_1, n_2, when $\underline{y}[1.s_j] - \underline{y}[2.s_j] = constant$, the Neyman-type randomization test tends to be just a bit anti-conservative for testing $H_0^{RA.s}$; that is, the actual size appears to be a little larger than the nominal size (see scenarios 1, 2, and 6 of Table 3). This anti-conservativeness presumably stems from the fact that the Neyman-type estimator of the variance, $var(D(\underline{T})|\underline{S} = \underline{s})$, tends to be slightly biased on the low side when $\underline{y}[1.s_j] - \underline{y}[2.s_j] = constant$. For larger n_1, n_2, this anti-conservativeness disappears (scenarios 1, 2, and 6 of Table 5).

When the differences $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable, the Neyman-type randomization test tends to be a bit conservative for testing $H_0^{RA.s}$, although not as conservative as the Fisher-type randomization test (scenarios 4 and 7 in Tables 3 and 5). This conservativeness presumably stems from the fact that the Neyman-type estimator of the variance, $var(D(\underline{T})|\underline{S} = \underline{s})$, tends to be biased on the high side when $\underline{y}[1.s_j] - \underline{y}[2.s_j]$ are highly variable (see Gadbury 2001).

For small n_1, n_2, the Normal approximation to the Neyman-type test statistic can be unreasonable when there are extreme outliers present (scenario 3 of Table 3). With larger n_1, n_2, the Normal approximations become more reasonable in the presence of extreme outliers (scenario 3 of Table 5).

In all of the simulation scenarios, the Neyman-type randomization test had higher power than the Fisher-type randomization test (see Tables 4 and 6), especially when n_1, n_2 are smaller (see Table 4). Of course, power comparisons are most useful when both tests have the same size. Because neither of these tests has size exactly equal to the nominal 0.05, these power comparisons should be considered carefully. In particular, in head-to-head comparisons, the Fisher test is at a disadvantage because its actual size is guaranteed to be no larger than 0.05; the Neyman test has size that is only approximately equal to, and can exceed, the nominal 0.05.

On the basis of this limited simulation study, we recommend that practitioners at least think seriously about using the Neyman-type randomization test as an alternative to the Fisher-type randomization test, especially when n_1, n_2 are moderate, say at least 10, and when there are no extreme outliers.

11 Discussion

This paper used concepts from the rich literatures on causal analysis and finite-population sampling theory to clear up some of the confusion that exists about tests of the no-treatment-effect hypothesis in the comparative experiment setting. Our approach lends itself to explicit specifications of the candidate no-treatment-effects hypotheses and targets of inference. We clearly distinguished between three main inference approaches: science-based, randomization-based, and selection-based. The commonly-used permutation test, Wilcoxon rank sum test, and two-sample t tests are examples of science-based approaches. Examples of randomization-based approaches, include the commonly-used Fisher-type randomization test and the less commonly-used Neyman-type randomization test.

We also described a Neyman-type selection test. A small-scale empirical comparison of these different tests was carried out. On the basis of the simulation results, we recommend that practitioners consider using the Neyman-type randomization test in certain scenarios.

In our description of the science-based approach, we focused on testing hypotheses about the distribution of $\underline{Y}$. More generally, the science-based approach can be used to both estimate, or test hypotheses about, characteristics of the distribution of $\underline{Y}$ *and* predict/estimate the unobserved values $\underline{y}[-\underline{t}.\underline{s}]$. Here, $\underline{y}[-A]$ is the collection of all $2N$ components of $\underline{y}$ excluding those with subscripts in the set A. A look back at the assumptions A_1–A_7 shows that we did not have to specify a model for the joint distribution of $\underline{Y}$ to carry out a test of no treatment effect. We only assumed independence across units and modeled the marginal distributions of $\underline{Y}[1.i]$ and $\underline{Y}[2.i]$. In contrast, the prediction of unobserved values generally requires a model for the joint distribution of $\underline{Y}$, equivalently, a model for $(\underline{Y}[\underline{t}.\underline{s}],\ \underline{Y}[-\underline{t}.\underline{s}])$, the "$(Y_{obs}, Y_{mis})$" of Rubin (e.g. 2005). Rubin advocates using a Bayesian approach to science-based prediction of $\underline{y}[-\underline{t}.\underline{s}]$.

This paper restricted attention to inferences about one population or sample, under two scenarios corresponding to two treatments. Owing to randomization, we were able to compare these two treatment scenarios; for example, see equation (6). Comparing two populations of distinct units is a qualitatively different inference problem. However, similar notations and model structures can be used to study this problem as well. Interestingly, in this two population setting, Fisher-type randomization tests, as described herein, are generally not applicable. In contrast, the other tests described in this paper, including the Neyman-type selection test, are applicable.

The notation and model structure introduced in this paper can be directly applied in more general settings where there are more than two treatments being compared. There are extensions in other directions. For example, rather than testing hypotheses, the ideas introduced in this paper could be used to derive useful confidence intervals. More work in this direction will be forthcoming.

In the binary response, comparative experiment setting, *Fisher's exact test* for 2×2 tables (see Agresti, 2002:91) is equivalent to the Fisher-type randomization test of $H_0^{RU.s}$ when $\underline{T} \perp \underline{Y}|\underline{S}$ and $\underline{T}|(\underline{S} = \underline{s})$ has a uniform distribution as in (7); recall that $H_0^{RU.s}$ states that the binary response values satisfy $\underline{y}[1.s_j] = \underline{y}[2.s_j]$, $j = 1, \ldots, n$. Fisher's exact test is also equivalent to the permutation test of H_0^{EU} when $(\underline{S}, \underline{T}) \perp \underline{Y}$ and $\underline{Y}[t.i]$ *indep* $\sim$ $bin(1, \pi_t)$; here H_0^{EU} is equivalent to $\pi_1 = \pi_2$. In fact, in the simulation (scenarios 6 and 7 of Tables 3 and 5, and scenario 6 of Tables 4 and 6), because of the uniform randomization distribution, we were able to use the R code for Fisher's exact test, `fisher.test {stat}`, to compute the exact values of the Fisher-type randomization and permutation p-values. On a related note, we point out that the Neyman-type randomization test is also available for testing the no-treatment-effect hypothesis $H_0^{RA.s} : \bar{y}[1.\underline{s}] = \bar{y}[2.\underline{s}]$ in 2×2 tables. This paper's simulation results suggest that when the randomization distribution is uniform as in (7), this Neyman-type randomization test for 2×2 tables may be somewhat more powerful than Fisher's exact test.

12 References

Agresti, A. (2002). *Categorical Data Analysis,* 2nd edition, New York: John Wiley and Sons, Inc.

Agresti, A and Franklin, C. (2007). *Statistics: The Art and Science of Learning from Data*, Pearson/Prentice Hall: Upper Saddle River, New Jersey.

Copas, J.B. (1973). "Randomization Models for the Matched and Unmatched 2×2 Tables," *Biometrika*, Vol. 60, No. 3, 467-476.

Cox, D.R. (1958). *The Planning of Experiments*, New York: John Wiley and Sons, Inc.

Ernst, Michael D. (2004). "Permutation Methods: A Basis for Exact Inference," *Statistical Science*, Vol. 19, No. 4, 676-685

Fisher, R. A. (1935). *The Design of Experiments*, Edinburgh:Oliver Boyd.

Gadbury, G.L. (2001). "Randomization Inference and Bias of Standard Errors," *The American Statistician,* Vol. 55. No. 4., 310-313.

Greenland, S. (1991). "On the Logical Justification of Conditional Tests for Two-by-Two Contingency Tables," *The American Statistician,* Vol. 45, No. 3, 248-251.

Greenland, S. (2000), "Causal Analysis in the Health Sciences," *J. Amer. Statist. Assoc.*, Vol. 95, No. 449, 286-289

Holland, P.W. (1986). "Statistics and Causal Inference," *J. Amer. Statist. Assoc.*, Vol. 81, No. 396, 945-968.

Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistical Methods*, New York: John Wiley and Sons.

Horvitz, D.G. and Thompson, D.J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe", *J. Amer. Statist. Assoc.*, Vol. 47, 663-685.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*, New York: John Wiley and Sons.

Kempthorne, O. (1955). "The Randomization Theory of Experimental Inference," *J. Amer. Stat. Assoc.*, Vol. 50, No. 271, 946-967.

Lehmann (1994). "Jerzy Neyman, 1894-1981: A Biographical Memoir," in *Biographical Memoirs, Vol. 63,* edited by Office of the Home Secretary, National Academy of Sciences, Washington D.C.: National Academies Press.

Neyman, J. (1923). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Roczniki Nauk Rolniczych Tom X* [in Polish]; English translation of excerpts by D.M. Dabrowska and T.P. Speed (1990), *Statistical Science*, Vol. 5, No. 4, 463-472.

Neyman, J. (1934). "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Sampling (with discussion)," *J. Roy. Statist. Soc.*, Vol. 97, No. 4, 558-625.

Neyman, J. with cooperation of K. Iwaskiewicz and St. Kolodziejczyk (1935). "Statistical Problems in Agricultural Experimentation (with discussion)," *Suppl. J. Roy. Statist. Soc.,* Vol. 2, No. 2, 107-180.

Pitman, E.J.G. (1937). "Significance Tests which can be Applied to Samples from any Populations," *Suppl. J. Roy. Statist. Soc.*, Vol. 4, No. 1, 119-130.

Pitman, E.J.G. (1938). Significance Tests which can be Applied to Samples from any Populations. III. The Analysis of Variance Test," *Biometrika*, Vol. 29, 322-335.

Rosenbaum, P.R. (retrieved 1/13/12). URL `www-stat.wharton.upenn/∼rosenbap/downloadTalks.htm`

Rubin, D.B. (1990). "[On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, Vol. 5, No. 4, 472-480.

Rubin, D.B. (2004). "Teaching Statistical Inference for Causal Effects in Experiments and Observational Studies," *J. Educ. and Behav. Statist.*, Vol. 29, No. 3, 343-367.

Rubin, D.B. (2005). "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *J. Amer. Statist. Assoc.*, Vol. 100, No. 469, 322-331.

Rubin, D.B. (2010). "Reflections Stimulated by the Comments of Shadish (2010) and West and Thoemmes (2010)," *Psychological Methods*, Vol. 15, No. 1, 38-46. doi: 10.1037/a0018537

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.

Strayer, D.L. and Johnston, W.A. (2001). "Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone," *Psychological Science*, Vol. 12, No. 6, pp. 462-466.

Welch, B.L. (1937). "On the z-Test in Randomized Blocks and Latin Squares," *Biometrika*, Vol. 29, No. 1/2, 21-52.

Welch, B.L. (1938). "The Significance of the Difference Between Two Means when the Population Variances are Unequal," *Biometrika*, Vol. 29, No. 3/4, 350-62.

Wilcoxon F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, Vol. 1, No. 6, 80-83.

Table 3. Monte Carlo Estimates of Power when $n_1 = n_2 = 10$, Nominal Size=5%.

n_1=n_2 = 10	Permutation[a]	Wilcoxon	t(Welch)	t(Pooled)	Fisher[a]	Neyman	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ N(10, 2^2)$						Sc.1
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \ldots, 20.$						
Randomization	4.6	3.6	4.7	4.7	4.6	6.5	
Science	4.3	3.4	4.2	4.3	4.3	6.9	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ Gamma(shape = 1, scale = 5)$						Sc.2
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \ldots, 20.$						
Randomization	5.0	4.9	4.1	4.6	5.0	7.4	
Science	4.0	4.1	3.2	3.5	4.0	7.7	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ 0.9U(0, 20) + 0.1U(200, 201)$, "mixture of uniforms"						Sc.3
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \ldots, 20.$						
Randomization[b]	4.6	3.9	0.0	0.0	4.6	0.0	
Science	3.8	3.5	1.1	1.8	3.8	11.2	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ N(10, 2^2)$						Sc.4
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + E_i - \overline{E},\quad E_i\ IID \sim\ N(0, 3^2),\quad i = 1, \ldots, 20.$						
Randomization	1.5	1.9	1.7	1.8	1.5	3.3	
Science	2.7	2.0	2.5	2.6	2.7	4.2	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ Gamma(shape = 1, scale = 5)$						Sc.5
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i] - \overline{Y}[1.\underline{P}],\quad i = 1, \ldots, 20.$						
Randomization	4.8	6.8	4.3	4.4	4.8	7.6	
Science	4.0	7.4	3.6	3.7	4.0	7.4	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim bin(1, 0.28)$						Sc.6
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \ldots, 20.$						
Randomization[c]	0.0	NA	9.1	9.1	0.0	9.1	
Science	2.1	NA	4.5	4.5	2.1	11.3	
H_0^{EU}, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim^d bin(1, 0.28)$						Sc.7
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i]\ IID \sim^d bin(1, 0.28),\ corr(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.37,\quad i = 1, \ldots, 20.$						
Randomization[e]	0.4	NA[f]	1.6	1.6	0.4	4.6	
Science	0.2	NA	1.0	1.0	0.2	3.7	

Table entries give the percent of times out of 1000 the simulated data gave a p-value $\leq$ 5%.

All indented hypotheses are also true, see Section 4.2. For example, in row 1, H_0^U is true. It follows that all the other hypotheses in Section 4.2 are also true.

[a] For this simulation, the permutation and Fisher-type randomization test results are numerically identical.

[b] The fixed $\underline{y}$ includes one large observation from the $U(200, 201)$ distribution.

[c] The fixed $\underline{y}[1.\underline{P}] = 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0 = \underline{y}[2.\underline{P}]$.

[d] This is an approximation because the $\underline{Y}$ values are adjusted to satisfy $H_0^{RA.s}$.

[e] The fixed $\underline{y}[1.\underline{P}]$ =1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0, $\underline{y}[2.\underline{P}] = 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 0$.

[f] Because of the many ties in the binomial case, the Wilcoxon test as described herein is not applicable.

Table 4. Monte Carlo Estimates of Power when $n_1 = n_2 = 10$, Nominal Size=5%.

n_1=n_2 = 10	Permutation[a]	Wilcoxon	t(Welch)	t(Pooled)	Fisher[a]	Neyman	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10,2^2)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 2, \quad i = 1,\ldots,20.$						Sc.1
Randomization	52.7	49.3	51.3	52.5	52.7	59.9	
Science	55.9	51.6	55.5	56.1	55.9	62.7	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10,2^2)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 2 + E_i - \overline{E}, \quad E_i\ IID \sim N(0,3^2), \quad i = 1,\ldots,20.$						Sc.2
Randomization	26.2	23.6	24.1	25.7	26.2	35.6	
Science	28.6	23.8	26.1	27.2	28.6	36.6	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10,2^2)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.2\underline{Y}[1.i], \quad i = 1,\ldots,20.$						Sc.3
Randomization	34.7	27.1	34.8	35.3	34.7	43.0	
Science	48.4	43.0	47.5	48.4	48.4	57.2	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim Gamma(shape = 1, scale = 5)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i], \quad i = 1,\ldots,20.$						Sc.4
Randomization	19.2	12.9	16.1	18.7	19.2	28.6	
Science	30.2	23.7	23.4	26.0	30.2	38.5	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim Gamma(shape = 1, scale = 5)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 3\underline{Y}[1.i] + E_i, E_i\ IID \sim N(0,5^2), \quad i = 1,\ldots,20.$						Sc.5
Randomization	45.7	28.6	40.2	45.3	45.7	65.5	
Science	49.2	38.1	39.8	44.4	49.2	63.5	
$H_0^{EU.1}$, $H_0^{RA.s}$ false	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim bin(1,0.28)$ $\underline{y}[2.i] \leftarrow \underline{Y}[2.i]\ IID \sim bin(1,0.71),\ corr(\underline{Y}[1.i],\underline{Y}[2.i]) = 0.29, \quad i = 1,\ldots,20.$						Sc.6
Randomization[b]	18.9	NA	37.3	37.3	18.9	37.4	
Science	29.6	NA	48.0	48.0	29.6	50.3	

Table entries give the percent of times out of 1000 the simulated data gave a p-value $\leq$ 5%.

[a] For this simulation, the permutation and Fisher-type randomization test results are numerically identical.

[b] The fixed $\underline{y}[1.\underline{P}] =$ 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 1 0 1 0, $\underline{y}[2.\underline{P}] =$ 0 1 1 1 1 1 1 0 1 0 1 1 0 0 1 1 1 0 1 1.

Table 5. Monte Carlo Estimates of Power when $n_1 = n_2 = 50$, Nominal Size=5%.

n_1=n_2 = 50	Permutation[a]	Wilcoxon	t(Welch)	t(Pooled)	Fisher[a]	Neyman	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ N(10, 2^2)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \dots, 100.$						Sc.1
Randomization	4.0	4.0	4.0	4.0	4.0	4.4	
Science	4.7	4.8	4.8	4.8	4.7	5.5	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ Gamma(shape = 1, scale = 5)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \dots, 100.$						Sc.2
Randomization	4.9	5.0	4.8	4.8	4.9	5.4	
Science	4.1	3.9	3.9	3.9	4.1	4.8	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ 0.9U(0, 20) + 0.1U(200, 201)$, "mixture of uniforms"						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \dots, 100.$						Sc.3
Randomization[b]	4.2	6.5	4.3	4.5	4.2	8.6	
Science	5.3	5.4	5.3	5.3	5.3	6.6	
$H_0^{EU.1},\ H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ N(10, 2^2)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + E_i - \overline{E},\quad E_i\ IID \sim\ N(0, 3^2),\quad i = 1, \dots, 100.$						Sc.4
Randomization	2.5	3.1	2.4	2.4	2.5	3.4	
Science	3.0	4.6	2.9	3.2	3.0	3.9	
$H_0^{UE.1},\ H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID\ \sim\ Gamma(shape = 1, scale = 5)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 2\underline{Y}[1.i] - \overline{Y}[1, \underline{P}],\quad i = 1, \dots, 100.$						Sc.5
Randomization	4.6	42.5	4.4	4.4	4.6	6.1	
Science	2.8	35.1	2.8	2.8	2.8	5.0	
H_0^U	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim bin(1, 0.28)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i],\quad i = 1, \dots, 100.$						Sc.6
Randomization[c]	2.2	NA	5.5	5.5	2.2	5.5	
Science	3.5	NA	5.0	5.0	3.5	5.9	
$H_0^{EU},\ H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim^d bin(1, 0.28)$						
true	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i]\ IID \sim^d bin(1, 0.28),\ corr(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.37,\quad i = 1, \dots, 100.$						Sc.7
Randomization[e]	0.5	NA	0.8	0.8	0.5	1.0	
Science	1.6	NA	2.8	2.8	1.6	3.5	

Table entries give the percent of times out of 1000 the simulated data gave a p-value $\leq$ 5%.

[a] For this simulation, the permutation and Fisher-type randomization test results are numerically identical.

[b] The fixed $\underline{y}$ includes 7 large observations from the $U(200, 201)$ distribution.

[c] The fixed $\underline{y}[1.\underline{P}] = \underline{y}[2.\underline{P}]$ with $\bar{y}[1.\underline{P}] = \bar{y}[2.\underline{P}] = 32/100$.

[d] This is an approximation because the $\underline{Y}$ values are adjusted to satisfy $H_0^{RA.s}$.

[e] The fixed $\underline{y}$ is such that $\underline{y}[1.\underline{P}] \neq \underline{y}[2.\underline{P}],\quad \bar{y}[1.\underline{P}] = \bar{y}[2.\underline{P}] = 33/100$, and $corr(\underline{y}[1.\underline{P}], \underline{y}[2.\underline{P}]) = 0.186$.

Table 6. Monte Carlo Estimates of Power when $n_1 = n_2 = 50$, Nominal Size=5%.

n_1=n_2 = 50	Permutation[a]	Wilcoxon	t(Welch)	t(Pooled)	Fisher[a]	Neyman	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10, 2^2)$						Sc.1
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 1, \quad i = 1, \ldots, 100.$						
Randomization	80.9	76.4	80.4	80.4	80.9	81.3	
Science	69.5	67.7	69.9	69.9	69.5	70.4	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10, 2^2)$						Sc.2
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = \underline{Y}[1.i] + 1 + E_i - \overline{E}, \quad E_i\ IID \sim N(0, 3^2), \ i = 1, \ldots, 100.$						
Randomization	36.3	31.4	36.2	36.4	36.3	42.7	
Science	37.9	36.3	37.5	38.0	37.9	42.8	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim N(10, 2^2)$						Sc.3
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.1\underline{Y}[1.i], \quad i = 1, \ldots, 100.$						
Randomization	70.5	68.6	71.0	71.1	70.5	72.1	
Science	66.6	63.9	65.7	65.7	66.6	67.4	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim Gamma(shape = 1, scale = 5)$						Sc.4
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.5\underline{Y}[1.i], \quad i = 1, \ldots, 100.$						
Randomization	46.6	39.0	46.2	46.4	46.6	49.5	
Science	49.2	40.6	48.0	48.2	49.2	51.8	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim Gamma(shape = 1, scale = 5)$						Sc.5
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i] = 1.5\underline{Y}[1.i] + E_i, E_i\ IID \sim N(0, 5^2), \quad i = 1, \ldots, 100.$						
Randomization	41.0	35.2	40.4	40.5	41.0	44.3	
Science	39.2	30.7	38.9	39.0	39.2	44.2	
$H_0^{EU.1}$, $H_0^{RA.s}$	$\underline{y}[1.i] \leftarrow \underline{Y}[1.i]\ IID \sim bin(1, 0.28)$						Sc.6
false	$\underline{y}[2.i] \leftarrow \underline{Y}[2.i]\ IID \sim bin(1, 0.50),\ corr(\underline{Y}[1.i], \underline{Y}[2.i]) = 0.36, \quad i = 1, \ldots, 100.$						
Randomization[b]	48.8	NA	58.8	58.8	48.8	60.3	
Science	51.1	NA	60.1	60.1	51.1	60.3	

Table entries give the percent of times out of 1000 the simulated data gave a p-value $\leq$ 5%.

[a] For this simulation, the permutation and Fisher-type randomization test results are numerically identical.

[b] The fixed $\underline{y}$ is such that $\underline{y}[1.\underline{P}] \neq \underline{y}[2.\underline{P}]$, $\overline{y}[1.\underline{P}] = 24/100$, $\overline{y}[2.\underline{P}] = 45/100$, and $corr(\underline{y}[1.\underline{P}], \underline{y}[2.\underline{P}]) = 0.386$.