Permutation Tests for Experimental Data

Charles A. Holt[*] **Sean P. Sullivan**[†]

November 2021

This paper surveys the use of nonparametric permutation tests for analyzing experimental data. The permutation approach, which involves randomizing or permuting features of the observed data, is a flexible and convenient way to draw statistical inferences in many common settings. It is particularly valuable when few independent observations are available, as is often the case for controlled experiments in economics and other social sciences. When viewed as a framework, the permutation method constitutes a comprehensive approach to statistical inference. In two-treatment testing, permutation concepts underlie popular rank-based tests, like the Wilcoxon and Mann-Whitney tests. But permutation reasoning is not limited to ordinal contexts. Analogous tests are easily constructed for the permutation of continuous measurements, and we argue that these non-ranked alternatives should often be preferred when working with continuous data. Permutation tests can also be used with multiple treatments, with ordered hypothesized effects, and with complex data structures, such as hypothesis testing in the presence of nuisance variables. Drawing examples from the experimental literature, this paper illustrates how permutation testing solves common data analysis challenges. Our aim is to help experimenters move beyond the handful of overused tests in play today, and to show how permutation testing constitutes a general framework for conducting statistical inference with experimental data.

[*] University of Virginia, Department of Economics, Charlottesville, VA 22903; cah2k@virginia.edu.

[†] University of Iowa College of Law, Iowa City, IA 52241; sean-sullivan@uiowa.edu. We wish to thank Dan Gingerich, Sophia Lochner, John Pepper, Erica Sprott, and members of the University of Virginia Quantitative Collaborative for comments on an early draft. Please do not quote or cite the current version without permission.

Table of Contents

1 Introduction and Motivation

In economics and other social sciences, data from laboratory and field experiments present two common challenges for statistical inference. The first is interdependence in the data. Markets and other group interactions can create dependence relationships between observational units. The second challenge is small sample sizes. The costs of recruiting and incentivizing subjects to participate in research experiments often drives experimenters to work with relatively small numbers of subjects. These costs were magnified during the recent Covid-19 lockdowns that required interactive experiments to be run online, with a significant fraction of Zoom meeting sessions being interrupted by subjects leaving the meeting or experiencing connectivity issues. The number of observations can also be limited in natural experiments, especially where there is little exogenous geographic dispersion of treatment conditions.[1] While experimenters have never let these obstacles stand in the way of useful research, neither have they grappled as seriously as one might hope with the question of how to conduct statistical inference in light of these complexities of experimental data.

These days, a common but conservative approach to addressing the interdependence problem is to perform statistical inference on a summary measure of behavior that can plausibly be interpreted as independent within the overall design of the experiment. To illustrate, suppose an experiment assigns subjects to 8 sessions, with each session involving 10 replications of a simulated market. A common approach to addressing interdependence within the repeated measurements for each session would be to compute a single average efficiency measure for each session, yielding a final sample of 8 independent observations for the experiment as a whole. The argument for aggregating so much of the data is not that economists cannot make progress with models of lower-level individual interactions. Economics is replete with such tools. The problem is that the assumptions used to motivate sophisticated empirical models can

[1] For example, Kagel and Roth (2000) summarize a comparison of different "clearinghouse" methods of matching medical residents with hospitals in the United Kingdom. Two locations used a "priority product" method that tended to fail; two locations used a "deferred acceptance" method that, in theory, was "stable" with respect to bilateral deviations from assigned matches; and one location switched from one method to the other. Minor procedural and geographic details differentiated the locations and the match values and costs of making early matches were unobserved, so the authors conducted a laboratory experiment in which each of the alternative matching methods was used in three laboratory sessions with carefully crafted parallel conditions.

lack credibility in an experimental context, especially when things like rationality and foresight assumptions are the very things being tested—not assumed—in the study.

Averaging lower-level observations helps to mitigate dependence problems but only exacerbates the second challenge: small sample sizes. When the experimenter is limited to few data points—perhaps six independent observations for an entire study—common statistical tests fail to provide a credible basis for inference. Small sample sizes make it difficult to assess the distributional conditions that many tests require to justify null distributions of the test statistic. Even more so, small sample sizes preclude tests that rely on limit theorems to motivate their null distributions. To address these problems, experimenters turn to nonparametric tests that sacrifice statistical power in exchange for validity under a wide range of distributional conditions.

The use of nonparametric tests is now common in experimental research, but the selection of tests often seems to be driven more by familiarity than by the properties of the tests themselves. This mirrors how experimental methods are taught. If you ask a colleague how they introduce nonparametric testing in their graduate classes, the response will probably be that that they instruct by example, presenting specific applications from papers as they arise. This approach has the advantage of introducing students to tests that are appropriate for common data patterns. But it has the disadvantage of obscuring relationships between different tests as well as the tradeoffs between them. Little is gained by directing students to textbooks for these additional details. Traditional statistics texts cover a wide array of techniques, beginning (and, for busy graduate students, ending) with tests of limited relevance to the numerical, multi-dimensional data encountered in many experiments.

From an experimenter's perspective, a better resource is something like Sidney Siegel's 1956 classic: *Non-parametric Statistics for the Behavioral Sciences*.[2] Siegel's presentations are clear, insightful, and laden with intuition. Even better, Siegel draws examples from behavioral psychology and economics experiments, so his presentations of statistical methods build upon and inform experimental design skills. Much of Siegel's text has stood the test of time. In the 60 years since the book was first published, however, advances in computing power have added some important new capabilities to the economist's toolkit. The approaches these capabilities enable now deserve an equally accessible introduction to the field.

[2] Siegel's impact on experimental methodology cannot be overstated. Indeed, the annual Economic Science Association prize for the best experimental economics dissertation is still called the *Sidney Siegel Award*.

This paper follows Siegel's lead in emphasizing intuition and relevant examples while introducing a family of permutation tests that can be used to solve experimental data analysis challenges. The central idea behind these tests is to take seriously the experimental design that generated the data, relying on randomization and the null hypothesis of no treatment effect to construct statistical tests customized to individual applications. Common nonparametric tests like the Mann-Whitney and Wilcoxon tests are special cases of the approach we describe, derived by applying permutation methods to the ranks of experimental measurements.

Beyond our intuitive introduction to permutation testing, we also offer two comments on the relative attractiveness of different permutation tests for experimental data. First, rank-based tests are overused today. Currently obscure tests based on unranked data often present more intuitive and possibly more powerful bases for statistical inference. Second, an underappreciated property of all permutation methods is the ability to tailor these tests to specific data problems. Interdependence, secondary nuisance variables, and other strata in the sample data can be easily and intuitively incorporated into permutation testing methods.

This paper illustrates these and other properties of permutation tests in the analysis of experimental data. We begin in Part 2 with an introduction to $k = 2$ sample permutation testing for independent samples, comparing the now common Mann-Whitney test to a simpler and potentially more powerful permutation test that experimenters could be using instead. Part 3 does the same for data composed of matched pairs of observations. Parts 4-6 generalize the two-sample treatment to cases involving $k > 2$ samples. Finally, Part 7 briefly discusses the use of permutation methods for linear effects models in correlation and multiple regression analysis.

2 Permutation Tests for $k = 2$ Independent Samples

The most fundamental statistical test for experimental work is the comparison of averages between unmatched samples. This situation is typical of experiments that draw subjects from a common pool and that expose each subject, or subject group, to a single treatment in the design. When contrasting measurements collected under one treatment, $\{x_1, \dots, x_n\} \sim F_x$, against those collected under another treatment $\{y_1, \dots, y_m\} \sim F_y$, the null hypothesis of no treatment effect corresponds to a situation in which measurements from the both samples are independent and identically distributed (iid) draws from the same underlying distribution: $F_x = F_y = F$.

Appropriate alternative hypotheses must be derived from context and theory, as illustrated in examples below. For simplicity, we confine our discussion to "shift" models in which distributions are assumed to differ in a measure of central tendency if at all. This assumption will often be plausible in the experimental context—especially where the only difference between observations is random assignment to a particular treatment—but should not be ignored. Many popular test statistics have power to detect not only differences in central tendency but also differences in distribution shape and variability. The assumption of a shift model justifies attributing rejection of the null hypothesis to a difference in central tendency.[3]

To keep things concrete, consider a simple experiment motivated by a change in the way license plates were auctioned in Shanghai. To curtail traffic congestion and raise revenue for subways and other infrastructure projects, the city of Shanghai had for many years sold a limited number of licenses plates each month. Prior to 2008, these auctions used a pay-as-bid format: *Q* plates were sold to the *Q* highest bidders, with winners paying their bid amounts. This procedure is usually referred to as a *discriminatory* auction, to distinguish it from the *uniform price* auction procedure in which the *Q* highest bidders win, and all pay the same market-clearing price (the highest rejected bid). In 2008, a new "*Shanghai auction"* was implemented. Bidders submitted initial sealed bids as before, with provisional winners being revealed at that point. Then they entered a limited-time bid-revision phase, in which each bidder could make up to two bid changes, provided these fell within a narrow range (about $100) above the lowest accepted bid at that point in time. This new auction procedure was widely understood as an effort to reduce the politically embarrassing high sales prices, which reached the price of a new economy car, or even higher in some cases.

Auction prices did initially decline after the rule change, but since the number of licenses being auctioned was also doubled during this time, it is hard to determine how much of the price change owed to the new rules as opposed to the new quantities. License prices also climbed back to original levels within a few years, but the Chinese economy was growing at roughly 8% per

[3] When distributions differ in not just location but also variance and shape, permutation tests with alternative hypothesis limited to locational shift may fail to control the probability of Type I error. This behavior is discussed and illustrated by Boik (1987), Romano (1990), Hayes (2000), and others. One solution to this problem is to adopt a more general alternative hypothesis. Other solutions attempt to control Type I error rates through modification of tests or testing procedures. Examples include Neuhäuser & Manly (2004) and Chung and Romano (2016).

annum, again making it difficult to isolate an effect. In short, observational data provide an imperfect view of how the new rules affected bids and auction outcomes.

Liao and Holt (2016) sought to answer this problem by measuring the price effect of the Shanghai auction rules while holding fixed the quantity of available plates, macroeconomic growth, and other confounding factors. To do this, they conducted a laboratory experiment in which each session consisted of 12 bidders who competed to buy 6 plates in a sequence of auctions. Bidders' profits were the difference between their private values (drawn randomly from a common distribution prior to each auction) and the price paid for a license plate. Each "revenue" figure in Table 1 represents an average of the auction revenues collected in each of the 10 auctions in a session.[4] The three numbers in the top row are average revenues in sessions with uniform price auctions; the three numbers in the second row are average revenues in sessions with discriminatory (pay-as-bid) auctions, as used prior to 2008; and the three numbers in the third row are average revenues in sessions with Shanghai auction rules. The column on the right shows average auction revenues by treatment.

Table 1. Auction Revenues by Session[a]

Auction Mechanism	Session Revenues			Mean
Uniform Price Sealed Bid	74.6	76.6	82.2	78.5
Discriminatory Sealed Bid	74.9	73.6	80.5	76.3
Shanghai Auction	57.1	54.9	53.6	55.2

[a] Liao and Holt (2016).

For investigating how average revenue differs between the discriminatory sealed bid treatment (sample x) and the Shanghai auction treatment (sample y), an obvious test statistic is the difference in sample average revenues:

$$T = \bar{x} - \bar{y} \tag{1}$$

Here, the difference $T_{obs} = 76.3 - 55.2 = 21.1$ is suggestive of the anticipated revenue reduction effect. Statistical inference, however, requires comparing this test statistic to a

[4] Revenues are expressed as a percentage of the maximum buyer surplus: i.e., the area under a demand array constructed from buyer valuations to the left of a vertical line representing the supply of licenses being auctioned.

sampling distribution, and with only three observations in each treatment, the usual assumptions required for normal-theory testing are hard to defend.[5]

At the cost of some statistical power, permutation tests provide a credible basis for statistical inference in this and similar experimental settings. Instead of assuming a specific distribution for the test statistic, the general strategy of permutation testing is to compute the null distribution of the test statistic on a case-by-case basis, using only the observed data and an understanding of the data generating process to motivate the test.

2.1 Permuting Measured Observations: The Pitman Permutation Test

Appropriate permutation strategies for constructing the null distribution of a test statistic can be inferred from knowledge of the experimental design and what the null hypothesis would mean for counterfactual sample draws. For example, under the null hypothesis that average auction revenues were identical under the Shanghai and discriminatory auctions, average revenues observed in any of these sessions of the experiment represent independent draws from a common average-revenue distribution. This means that each of these observed revenue draws would be just as likely to have been assigned to the Shanghai-auction treatment as the discriminatory-auction treatment: so every permutation of the data between these treatments has an equal ex ante probability of having been observed. With 3 observations in each of two independent samples, there are $\binom{6}{3} = 20$ equally probable ways that these data could have been assigned to the two treatments if the null hypothesis were true. The null distribution of the test statistic (difference in treatment average revenues) can be constructed by computing the value that the test statistic assumes for each of these 20 permutations of the sample data, as shown in the right-hand column in Table 2.

[5] While, in many settings, researchers rely on limit theorems to describe the asymptotic distribution of a test statistic with even moderate sample sizes, asymptotic arguments are difficult to accept for a combined sample size of 6 observations.

Table 2. Computing the Null Distribution of the Test Statistic

Permutation Number (i)	Discriminatory Auction Revenue (x)	Shanghai Auction Revenue (y)	Test Statistic $T_i = \bar{x}_i - \bar{y}_i$
1	53.6 54.9 57.1	74.9 73.6 80.5	-21.133
2	53.6 54.9 73.6	74.9 80.5 57.1	-10.133
3	53.6 54.9 74.9	73.6 80.5 57.1	-9.2667
4	53.6 54.9 80.5	74.9 73.6 57.1	-5.533
5	53.6 57.1 73.6	74.9 80.5 54.9	-8.667
6	53.6 57.1 74.9	73.6 80.5 54.9	-7.8
7	53.6 57.1 80.5	74.9 73.6 54.9	-4.067
8	53.6 73.6 74.9	80.5 57.1 54.9	3.2
9	53.6 73.6 80.5	74.9 57.1 54.9	6.933
10	53.6 74.9 80.5	73.6 57.1 54.9	7.8
11	54.9 57.1 73.6	74.9 80.5 53.6	-7.8
12 … 19	These rows are treatment reversals of rows 2-9, so the corresponding test statistics switch signs, as happened in rows 10 and 11.		
20 (observed)	73.6 74.9 80.5	57.1 54.9 53.6	21.133

Once these values are computed, hypothesis testing is straightforward. In conventional hypothesis testing, the p-value for a two-sided test represents the probability of drawing a value of the test statistic, T, at least as extreme as the observed value, T_{obs}, provided that the null hypothesis was true:

$$\text{two-sided } p\text{-value} = P_{H_0}(|T| \geq |T_{obs}|) \tag{2}$$

This probability is usually calculated by assuming that the test statistic follows a known distribution. Permutation testing takes the same approach, but instead of assuming a distribution for the test statistic, it constructs the empirical null distribution of the test statistic from the sample data. Here, $T_{obs} = 21.1$. There are only two ways that these data could be rearranged to result in a test statistic as extreme or more extreme than T_{obs}. These are the permutations in the first and last row of Table 2. Since the null hypothesis implies that all permutations are equally probable, the p-value for a two-sided test of the null hypothesis is $2/20 = 0.1$. More generally, the two-sided p-value for this permutation test is the proportion of all test-statistic values greater than or equal to the observed value among all $\binom{m}{n}$ ways of permuting the data between the two samples:

$$\text{Pitman permutation test, two-sided } p\text{-value} = \frac{\sum_{i=1}^{\binom{m}{n}} 1\left(|T_i| \geq |T_{obs}|\right)}{\binom{m}{n}} \tag{3}$$

where T_i is the value of the test statistic for the *i*th permutation. The "as or more extreme" aspect is captured by using the absolute value of the revenue difference, and the $1(\cdot)$ in the formula is the indicator function with a value 1 if its argument is logically true and 0 otherwise.

One-sided versions of this permutation test are computed similarly; the only difference is that the numerator consists of signed values of the test statistic greater than (or less than) the observed value of the test statistic. For example, since the Shanghai auction rules were ostensibly designed to reduce average revenues, a reasonable alternative hypothesis is that average revenue would be lower in a Shanghai auction than in a discriminatory auction. The *p*-value for this one-sided test is the probability of observing a signed value of the test statistic greater than or equal to the observed value under the null. Reviewing Table 2, only the observed permutation in the bottom row meets this criterion, so the one-sided *p*-value is $1/20 = 0.05$.

In our experience, students often find the methodology of permutation testing more intuitive than standard normal theory. Some find it so appealing that they never look back. Even so, it is instructive to consider the similarities and differences between this permutation procedure and the familiar Student's two-sample *t*-test.

First, unlike the *t* or normal distributions, the null distribution of this permutation test can be highly discrete. In the above two-sided example, the *p*-value of 0.1 is actually the strongest rejection of the null hypothesis that the test supports for these sample sizes. Intuitively, no configuration of the data can be more extreme than the case where all 3 observations in one sample are greater than all 3 in the other sample, so the *p*-value for this permutation test could never fall below $2/20 = 0.1$, no matter how extreme the difference is between the two samples.

Second, while having a less discrete null distribution can allow Student's *t*-test to reach lower *p*-values (0.003 for the two-sided *t*-test versus 0.1 in the above two-sided version of the permutation test), it does so at the cost of assuming a specific distribution for the test statistic. This distributional assumption is not innocuous. Inaccurate distributional assumptions can invalidate a parametric test. The permutation test imposes no distributional assumptions in the sense of requiring the sample data to come from any particular population distribution.[6] While the permutation test is not equally powerful for all possible distributions, it remains valid in many practical settings.

Third, the similarity of this permutation test and Student's *t* test is not superficial. Both tests are based on the same test statistic (a function of a difference of sample averages). As sample sizes become large, the efficiency of the above one- and two-sided permutation tests also converges with the analogous *t*-tests (Hoeffding 1952; Miller 1997). The two-independent-sample permutation test is like a distributionally robust version of the conventional *t* test.

If everything in this section seems intuitive and straightforward, the reader might wonder why this type of permutation testing is not more common in the literature. Indeed, while the Pitman permutation subcaption of this section is a nod to one of the earliest proponents of this form of permutation testing (Pitman 1937a), few experimenters would even recognize that name today, much less the statistical test we associate with it.[7] The explanation is that, while the theory of permutation-based inference has been understood for more than 75 years, computing power has only recently made this type of case-by-case construction of the null distribution practical in common applications (Berry, Johnston, and Mielke 2019: ch. 2).

To be sure, for large enough samples, the computational burden of permutation testing still becomes prohibitive. But this is of little practical importance. For one thing, numerical simulation methods carry the idea of permutation testing past the computational horizon. An approximate permutation test simply draws a large number of randomly permuted samples (e.g. 500,000 or more) from the observed data and computes the proportion of these draws that yield a test statistic as large or larger than what was observed (either in one direction for a one-tailed test or in both directions for a two-tailed test).[8] For another thing, normal-theory testing becomes defensible as sample size grows, meaning that the practical need for permutation testing fades

[6] As noted previously, we assume a shift model throughout this paper. That distributional assumption is imposed to ensure validity for the locational difference alternatives that are the focus of our examples.

[7] As for most topics in statistics, R.A. Fisher also has strong claim to name recognition for this approach. See, e.g., Fisher (1936). Manly (2007: p. 113) provides an interesting discussion of philosophical contrasts between Pitman's and Fisher's permutation arguments. Somewhat a reflection of the weight of each scholar's work, but mainly for expositional clarity, we refer to unranked permutation testing in the two-independent-sample context as a Pitman permutation test, and to unranked permutation testing in the matched-sample context as a Fisher permutation test, to be discussed in the next section. Miller (1997: pp. 27, 53) adopts this same convention.

[8] If simulations are used, it is advisable to run several large simulations to be sure that the resulting *p* value proportions are not affected in terms of the number of trailing digits being reported.

just as computational challenges begin to appear. With no remaining technological obstacles to excuse its unfamiliarity, the Pitman permutation test deserves a more prominent role in the experimenter's toolkit than it commands today.

2.2 Permuting Ranked Observations: The Mann-Whitney Test

While few experimenters are currently familiar with the Pitman permutation test, related tests devised by Wilcoxon (1945) and Mann and Whitney (1947) are constantly employed. The Wilcoxon and Mann-Whitney approaches describe different but equivalent tests and are sometimes refenced jointly as the Wilcoxon-Mann-Whitney test. For ease of exposition, we refer to both tests as the Mann-Whitney test, which also helps to distinguish this permutation strategy from the Wilcoxon Signed Rank test (discussed in relation to matched-pairs samples in Section 3.2). Just as lack of familiarity with the Pitman permutation test owes to historic and now outdated computational difficulties, the popularity of the Mann-Whitney test owes to the inertia of computational shortcuts that are of little importance today.

The Mann-Whitney procedure presented in most non-parametric books involves ranking all sample data (both samples combined) and replacing each sample's measured value by ordinal ranks in the combined sample. Both the Wilcoxon and Mann-Whitney versions of the test then compute special test statistics with computationally convenient null distributions (Siegel 1956; Miller 1997; Gibbons and Chakraborti 2003). The specific definition of the test statistic is not important for present purposes. The key point to note is that the approach is equivalent to running a permutation test on the ranked data instead of the original values.

Specifically, the Mann-Whitney test is a two independent-sample permutation test (covered in Section 2.1) in which the data being permuted and compared in average difference are not the original data but the ordinal ranks of each data point in the combined sample (Siegel, 1956: p. 155). For example, instead of permuting observed average revenues in the Shanghai auction experiment, the Mann-Whitney test would permute the ranked values of each observation in the Shanghai auction and discriminatory auction samples. The Shanghai auction revenues of 57.1, 54.9, and 53.6, are the three lowest revenues, with ranks of 3, 2, and 1 respectively; the discriminatory auction revenues of 74.9, 73.6 and 80.5 are the three highest revenues, with ranks of 5, 4, and 6, respectively. The Mann-Whitney test is thus equivalent to running the Pitman permutation test, covered in the previous section, with these rank-values substituted in place of the actual observations.

Actually, it makes no difference whether one works in ranks or level data in this particular application. Since the ranked versions of the samples, $x = \{3, 2, 1\}$ and $y = \{5, 4, 6\}$, are more extreme than any permutation except their compete reverse, the p-value for the two-sided test is the same in ranks as it is in the original data: $2/20 = 0.1$. This is a special case, however, and permutation tests based on ranks are not generally the same as those based on observed values.

To illustrate the potential difference of these tests, consider an experiment reported by Bohr, Holt, and Schubert (2019), involving asset market performance with saving decisions over a simulated lifetime. Subjects, in this experiment, were permitted to buy and sell asset "shares" that paid dividends each period. The dividends and interest paid on cash induced a flat fundamental share value (present value of future dividends) of $20 per share. In six sessions of the experiment (the "private-savings" treatment), subjects traded assets while also deciding how much of their incomes to save for low-income "retirement' years. In another 6 sessions (the "government-savings" treatment), a fixed portion of each subject's income was instead simply withheld by the "government" for retirement years. As a result of the difference in savings policy, subjects in the private-savings treatment carried cash stocks that were about twice as large as those in the government-savings treatment. Peak price data are provided for both treatments in Table 3, below.

Table 3. Peak Price Data and Ranks for Asset Shares[a]

Treatment	Session Peak Prices						Mean
Private Savings	42	36	53	61.5	38.5	70	50.2
Government Savings	42.5	21.25	30	26	43	38	33.5

[a] Bohr, Holt, and Schubert (2019).

While major price bubbles were observed in most sessions of the experiment, an interesting research question is whether peak asset prices were greater under the cash-rich private-savings treatment than under the government-savings treatment. The data are qualitatively consistent with this hypothesis but there is overlap between the samples. When permuting the measured data under the Pitman test, there are $\binom{12}{6} = 924$ possible permutations of the 12 observations across treatments, 36 of which yield a treatment difference at least as extreme as the observed value, for a two-sided *p*-value of about 0.039. By contrast, when permuting the data in ranks under the Mann-Whitney test, there are fully 86 permutations in which the treatment difference (in average ranks) is at least as extreme as the observed value, yielding a two-sided *p*-value of about 0.093. In this example, the rank-based Mann-Whitney test barely supports rejection of the null hypothesis at the 10% level, while the Pitman test's more complete use of the sample information allows for rejection at less than the 5% level.[9]

If the only difference between the Mann-Whitney and Pitman permutation tests is that the Mann-Whitney tests drops information from the measured data, then why is the Mann-Whitney test so popular? It once enjoyed the important advantage of having a null distribution that could be pre-computed and printed in the form of critical-value tables in the back of statistics tests. But modern computing power makes this practice all but irrelevant. The rank-based test is still appropriate when the measured data are ordinal in nature.[10] And, because the rank conversion suppresses distortions caused by outliers, the Mann-Whitney test may be preferable to the Pitman test when outliers caused by measurement errors, procedural issues, or other artificial influences are believed to be a concern.

In most cases, however, the Pitman test presents the more compelling option. The Pitman test uses more of the information contained in the sample data and is more sensitive to differences between measured observations. In a Monte Carlo comparison of the Pitman test against the parametric Student's t-test and Mann-Whitney test, Moir (1998) observes the Pitman test to perform as well as or better than either of the other tests in most applications.[11] Where measured data are more than ordinal (prices, auction revenues, market efficiencies), and where policy interest concerns the magnitudes of treatment effects, it is hard to justify eschewing the Pitman test for an alternative that is relatively insensitive to these aspects of the data. Especially

[9] The sharper result obtained with the Pitman test owes to the fact that the "reversals" from the general trend (peaks of 43, 42.5 and 38 under the government-savings treatment) are only slightly larger than some of the more modest peaks under the private-savings treatment (42, 38.5 and 36), whereas the largest peak prices under the private-savings treatment (70, 61.5, and 53) greatly exceed most of the government-savings observations.

[10] For example, measures of individual characteristics, e.g. risk aversion or type-A personality, are typically considered to be ordinal in the absence of precise preference models that may observed responses to questions.

[11] Specifically, Moir (1998) finds the Pitman test (in that paper referred to as the "ER means test") to perform about as well as the t-test when the underlying distribution is close to normal and to outperform the *t*-test in some non-normal settings. The Mann-Whitney test underperformed both the Pitman test and *t*-test in most settings considered.

when extreme observations are believed to reflect factors of importance to the study, it would be a mistake to suppress the information by using a rank transformation.

3 Permutation Tests for $k = 2$ Matched Samples

An important distinction when analyzing experimental data concerns the difference between *within-subjects* designs and *between-subjects* designs. In a between-subjects design, each person or group is exposed to a single experimental treatment. This produces samples of independent observations, the focus of the previous section. A within-subjects design exposes each person or group to multiple treatments in sequence. Exposure to more than one treatment has the potential drawback that behavior induced by one treatment may carry over to another treatment—a design bias known as a *sequence effect*. But within-subjects designs have the advantage of collecting data in a way that allows each person or group to serve as its own control group—a potentially valuable property when trying to study a treatment effect in the presence of substantial individual heterogeneity.

In the two-treatment context, within-subjects designs produce samples consisting of matched pairs of observations. For example, if measurements are taken when n subjects are exposed to a control condition of an experiment, $(x_1, \dots, x_n) \sim F_x$, and measurements are repeated when the same n subjects are exposed to a treatment condition of the experiment, $(y_1, \dots, y_n) \sim F_y$, then the difference vector $(d_1 = x_1 - y_1, \dots, d_n = x_n - y_n) \sim F_{x-y}$ reflects how exposure to the treatment has changed the measured outcome *within* each subject in the experiment. It is often convenient to work directly from this difference vector when testing hypotheses in a within-subject design. The null hypothesis of no treatment effect, $F_x = F_y$, equates to a difference distribution, F_{x-y}, that is symmetric about 0, such that differences between the control and treatment are explained by random noise alone. The alternative hypothesis of treatment distributions differing in central tendency equates to a difference distribution F_{x-y} with non-zero central value.

A helpful example of a within-subjects design is an experiment created to study how prices respond to changes in the number of sellers and market power in a posted-price oligopoly. Davis and Holt (1994) assigned each of 12 sessions of an experiment to two of three treatments. Six of the sessions entailed 30 periods of price competition followed by 30 periods of competition under a redistribution of production capacity that created or reduced market power, holding the number of sellers fixed.[12] The other six sessions entailed 30 periods of price competition followed by 30 periods of competition with the addition or removal of 2 sellers from the market, holding market power fixed.[13]

Table 4 shows observed price measures for this experiment. The numbers in the table are average prices over the final 15 replications of a market treatment—that is, the later replications in which strategies and behavior have had time to reach steady states. Asterisks on session labels denote those sessions in which subjects were first exposed to the topmost of the two treatments. Orthogonal treatment assignment was intended to mitigate the potential design bias caused by any sequence effects in the experiment.

Table 4. Average Prices for Different Oligopoly Markets[a]

Treatment / Session	S1*	S2	S3*	S4	S5*	S6	S7*	S8	S9*	S10	S11*	S12
5 sellers no market power	329	308	341	410	310	397						
5 sellers market power	407	468	430	455	397	441	415	471	392	401	392	512
3 sellers market power							425	470	408	436	424	517

[a] Davis and Holt (1994).
* The top treatment was the first exposure in this session.

The researchers sought to answer two related questions. First, does the market power created by redistribution of production capacity to larger sellers confer pricing power? Second, even holding market power constant, is there a pure *numbers effect* in which fewer competitors means greater pricing power? In both cases, the data appear to reflect a treatment effect, but with only 6 matched-pair data points for each comparison, tests based on assumed distributions are again hard to defend. A permutation approach is more persuasive.

[12] Specifically, in the Power design, the mixed strategy Nash equilibrium price distributions are above the competitive price, and in the No-Power design the Nash equilibrium price equals the competitive price.

[13] Constant market power is achieved by structuring demand and cost conditions such that the symmetric mixed strategies of sellers in setting prices yields the same predicted price distributions for each treatment.

3.1 Permuting Measured Observations: The Fisher Permutation Test

As before, the appropriate permutation strategy for constructing the null distribution of a test statistic can be inferred from knowledge of the experimental design and what the null hypothesis would mean for counterfactual sample draws. Here, the null hypothesis of no treatment effect implies that the vector-difference of matched pairs should have zero mean, suggesting the use of the average difference as an appropriate test statistic:[14]

$$T = \bar{d} \tag{4}$$

As noted above, the null hypothesis implies that the difference distribution F_{x-y} should be symmetric around zero. This motivates a simple permutation strategy for constructing the null distribution of the test statistic.

Since every difference is equally likely to be either positively or negatively signed under the null, the null distribution of the test statistic can be computed by permuting the signs of the sample differences. Equivalently, and perhaps more intuitively, since F_x and F_y are equal under the null hypothesis, the observed values in every matched pair are equally likely to have been assigned to opposite treatments under the null—which corresponds to simply swapping the sign of their difference. This strategy of permuting signs is often attributed to Fisher (1935). For lack of a better name, we refer to it as *the Fisher permutation test.*

For a sample of n matched pairs, there are 2^n ways that the signs of all matched pairs could be permuted under the null. Taking the *p*-value to be the probability of seeing a value of the (average difference) test statistic *T* as extreme or more extreme than the observed value, T_{obs}, a permutation *p*-value for match-pairs data is as follows:

$$\text{Fisher permutation test, two-sided } p\text{-value} = \frac{\sum_{i=1}^{2^n} 1(|T_i| \geq |T_{obs}|)}{2^n} \tag{5}$$

where $1(\cdot)$ is again the indicator function.[15]

To illustrate, consider the toy case of 3 matched pairs with difference vector *d* = (−9, −6, 3) for an average difference of $T_{obs} = -4$. There are $2^3 = 8$ ways to permute the signs of these

[14] Of course, $\bar{d} = \bar{x} - \bar{y}$, so the test statistic could equivalently be expressed as in equation (1). We adopt this form of the test statistic because it better matches the intuition of the matched-pairs permutation strategy.

[15] Absolute values are needed for a 2-tailed test when the test statistic is a signed value.

differences. Of these 8 permutations, four yield an average difference equal to, or more extreme than, -4:

- an outcome more extreme in the negative direction: (−9, −6, −3);
- an outcome more extreme in the positive direction: (9, 6, 3);
- the observed outcome: (−9, −6, 3); and
- an equally extreme outcome in the positive direction: (9, 6, −3).

The permutation p value for this 2-tailed matched-pairs test would thus be $4/8 = 0.5$.

If the preceding illustration is too abstract, application to the Davis and Holt experiment may be more concrete. The first question the experiment sought to answer is whether increased market power confers pricing power in this setting. For the 6 matched pairs in sessions S1–S6 of Table 4, the average price is higher for the market power treatment in every case. That is, the sample-difference vector has only positive signs. Since there are $2^6 = 64$ possible permutations of the signs of these matched pairs under the null—only two of which result in a test statistic as-or-more-extreme than the observed test statistic—the two-sided p-value can be quickly intuited to be $2/64 = 0.031$. Actually, since economic theory predicts a positive sign for this treatment effect, it is arguably more appropriate to use a one-sided p-value in this setting. The one-sided p-value corresponding to a test of the alternative hypothesis that prices are higher with greater market power seeks only those permutations with test statistic values strictly greater than or equal to the observed value. Since no alternative permutation of the signs of these data provides an average value larger than what was observed, the one-sided p-value is $1/64 = 0.016$. Market power leads to statistically significant pricing power in this setting. This example is, however, misleading in its simplicity. Since all differences were of the same sign, the magnitudes of differences did not need to be considered in calculating the 1/64 probability for each tail. Next, we consider a more realistic example with some overlap.

The second question the experiment sought to address is whether a pure *numbers effect* gives smaller number of competitors greater pricing power, even when holding technological market power constant. For the 6 matched pairs in sessions S7–S12 of Table 4, the average price is greater under the 3-seller treatment in all but one case. Focusing on the one-sided test, there are 2 possible permutations that would yield a value of the average-difference test statistic greater than or equal to the observed value:

- the observed matched pairs, reflected in Table 4; and

- the observed matched pairs, but with 470 and 471 reversed in column S8.

The one-sided p-value is thus $2/64 = 0.031$, indicating a significant numbers effect; the corresponding two-sided p-value would be $4/64 = 0.063$.

It is worth noting that the previous conclusion would *not* have been reached if the testing framework had not taken account of the matched pairs design. If the 3-seller power and 5-seller power samples on the right side of Table 4 had simply been treated as independent samples, a Pitman permutation test would give a one-sided p-value of 0.272. The intuitive explanation for this is that different groups of subjects can differ quite a bit in terms of competitiveness. Note that the most collusive outcome (3 sellers in session S12) is from the same subject group that produced the most collusive outcome with 5 sellers. Using each group as its own control helps to mitigate the effects of subject heterogeneity and thus helps to identify a treatment effect that would otherwise be difficult to distinguish from noise in the data.

3.2 Permuting Ranked Observations: The Wilcoxon Signed-Rank Test

Just as the ability to perform permutation tests on the measured data is often overlooked in experimental analysis of independent samples, so is it overlooked in the matched-pairs context. By far, the most commonly used test in the small-sample matched-pairs setting is the Wilcoxon (1945) signed-rank test. The name of this test reflects the peculiar transformation that it applies to the measured data. All matched-pair differences are ranked from smallest to largest in absolute values, and then these ranks are assigned the signs of the original difference data. To illustrate, the vector of sample treatment differences $d = (-6, 4, 0, -3)$ becomes $SR_d = (-4, 3, 1, -2)$ after applying the signed-rank transformation.

As was the case for the Mann-Whitney test, the original motivation for using the signed-rank transformation was primarily to enable reliance on a test statistic for which a precompiled null distribution could be provided in printed form. The specifics of the relevant test statistic and its distribution are interesting but are not belabored here.[16] It is sufficient to note that the Wilcoxon signed-rank test is a Fisher permutation test conducted on the signed ranks of the sample difference vector as opposed to the measured values of the difference vector.

[16] For additional background on the signed-rank test, see Wilcoxon (1945) and any introductory text on nonparametric statistics (e.g., Siegel 1956; Miller 1997; Gibbons and Chakraborti 2003).

How do the Fisher permutation test and the Wilcoxon signed-rank test compare? For the Davis and Holt experiment, both approaches yield the same p-values. While, in the past, the Wilcoxon test would have enjoyed a large computational advantage over the Fisher test, improvements in computing power have eroded this difference to the point that it should no longer dictate the choice between these tests. Fisher's permutation test has strong intuitive appeal as the permutation analog of Student's one-sample t-test (Miller 1997). Monte Carlo evidence demonstrates the superior power of the Fisher permutation test over the Wilcoxon signed rank test for a variety of sample sizes and distributions (Kempthorne and Doerfler 1969). This is an intuitive finding, since a permutation test with unranked data uses more information from the sample—information about magnitudes. In our opinion, the Fisher permutation test should be the experimenter's default choice when differences between observed measurements reflect important outcomes, with the Wilcoxon test limited to situations in which data are ordinal as measured or outliers are believed to be complicating analysis.

To illustrate the last point, suppose that the first session of the 5 seller no market power treatment had yielded an observed price of 700, rather than 329. Since the Fisher permutation test is based on measured values instead of ranks, the magnitude of this observation affects the test statistic quite a bit. When considering the one-sided alternative, there are now 26 possible sign permutations that yield average differences greater than or equal to the observed difference, for a p-value of $26/64 = 0.41$. The magnitude of the high-price observation has less effect when converted to a signed rank in the Wilcoxon test: when considering the one-sided alternative, there are only 14 possible sign permutations that would yield an average difference of signed ranks greater than or equal to the observed value, for a p-value of $14/64 = 0.22$. The Fisher test is more sensitive to the existence of large observations like this. That is a strength of the test when these extreme observations reflect reality, but a potential weakness of the test if extreme observations are artificial outliers.

4 Permutation Tests for $k > 2$ Independent and Unordered Samples

The permutation tests discussed thus far have involved comparisons of two treatments, but many experiments involve more than two treatments. Multiple pairwise comparisons can always be conducted in the multiple treatment situation, but the interpretation of these results is beset by complications. For one thing, it can be difficult to interpret results when some tests justify rejection of the null hypothesis while others do not. For another thing, the simultaneous assessment of multiple tests usually should be accompanied by multiple-comparison adjustments to control for error inflation in this procedure.

Simultaneous tests designed to detect locational differences among a set of more than two samples present an attractive alternative. Two common examples of this approach are tests of multiple unordered treatment effects, and tests of treatment effects with monotone predicted intensity. This section takes up the case of unordered treatment effects.

4.1 Permuting Measured Observations: The *F* Test

The test of multiple unordered treatment effects is a simple generalization of the test of locational difference between two independent samples. In the two-sample setting, samples x and y are compared to see whether the difference in something like their means is statistically distinguishable from zero. In the more general setting, $k > 2$ samples $x_1, \dots, x_k$ with sizes $n_1, \dots, n_k$ are compared simultaneously to see whether differences in *any* of their means are statistically distinguishable from zero. In a basic statistics textbook, this procedure would fall under the title of one-factor analysis of variance (ANOVA) testing, and statistical inference would most likely be based on the *F* statistic and its comparison to the *F* distribution:

$$F = \frac{(k-1)^{-1}\sum_{j=1}^{k} n_j\left(\overline{x_j} - \bar{x}\right)^2}{(N-k)^{-1}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{i,j} - \overline{x_j}\right)^2} \tag{6}$$

where $x_{i,j}$ denotes the *i*th observation in sample x_j, $\overline{x_j}$ and n_j denote the mean and number of observations in sample x_j, $\bar{x}$ denotes the grand mean when all samples are pooled together, and *N* is the total number of observations. This test is known to be relatively robust against non-normality (Miller 1997). However, the type of data often encountered in experiments pushes the boundaries of what could plausibly justify the assumption of a specific, parametric null distribution. As elsewhere, permutation testing provides a more credible basis for inference.

Applications of permutation testing in one-way ANOVA go back nearly as far as the use of permutation testing in the two-sample setting (Pitman 1938). The reason is that the two-sample permutation process generalizes in an intuitive way to the higher order setting. To illustrate, suppose the reporters of the Shanghai Action experiment (Table 1 on p. 5) were interested in testing the null hypothesis of no difference in treatment effect between any of the three auction mechanisms, against the alternative hypothesis of a locational difference between at least two of the mechanisms.[17] The observed data yield an *F* statistic of 44.85. All that remains is to construct the null distribution and to see how this value compares with it.

As before, the null distribution can be derived from knowledge of the experimental design and the implications of the null hypothesis. If the null of no treatment effect were true, then all observations in the experiment would be equally likely to have been assigned to any of the treatments. If there are $k > 2$ treatments with $n = \sum n_k$ total observations, then there are $\binom{n}{n_1}$ ways that observations could have been assigned to the first treatment. For each of these, there are $\binom{n-n_1}{n_2}$ ways that observations could have been assigned to the second treatment; for each of these, there are $\binom{n-n_1-n_2}{n_3}$ ways that observations could have been assigned to the third treatment; and so on, for a total of $n!/(n_1! \times \ldots \times n_k!)$ equally likely permutations of the data under the null. The *p*-value for a permutation-based test of the null hypothesis of equality of all samples comes from computing the *F* statistic for every permutation of the data and counting the proportion of these *F* statistics that are greater than or equal to the observed value:

$$\text{permutation } F\text{test}, p\text{-value} \ = \frac{\sum_{i=1}^{N} 1(F_i \geq F_{obs})}{N} \tag{7}$$

where $N = n!/(n_1! \times \ldots \times n_k!)$. For the Shanghai Action experiment, three groups of three observations yields $N = \ 9!/(3! \times 3! \times 3!) = 1{,}680$ possible permutations of the observed data. As noted above, the observed value of the test statistic is 44.85. And since 36 of the 1,680 possible permutations yield *F* statistics greater than or equal to 44.85, the *p*-value for a permutation *F* test of the null hypothesis is $36/1{,}680 = 0.021$. Notice how this compares with the weaker *p*-values of $2/20 = 0.1$ obtained under pairwise comparison of the Shanghai and discriminatory auction, or Shanghai and uniform price auctions using the Pitman permutation test discussed in Section 2.1.

4.2 Permuting Ranked Observations: The Kruskal-Wallis Test

The above description of the permutation *F* test compares to the familiar Kruskal-Wallis test—the current default choice of most experimenters for a nonparametric test of locational

[17] Again, we assume a shift model in which the distributions differ in location alone if they differ at all.

difference among k samples. Like the Wilcoxon and Mann-Whitney tests, the Kruskal-Wallis test statistic is a function of ranked observations:

$$H = (N-1)\frac{\sum_{j=1}^{k} n_j\left(\overline{r_j} - \bar{r}\right)^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(r_{i,j} - \overline{r_j}\right)^2} \tag{8}$$

where $r_{i,j}$ denotes the rank (among all observed values) of the *i*th observation in sample x_j, $\overline{r_j}$ and n_j denote the mean rank and number of observations in sample x_j, $\bar{r}$ denotes the average of all ranks, and *N* is the total number of observations.

How does the Kruskal-Wallis Test compare to the permutation *F* test? Consistent with our discussion in the two-sample context, rank-based tests make sense where the sample data are actually ordinal, or where the experimenter has reason to be concerned about serious outliers not connected to fundamentals of the subject being investigated. Outside of these special cases, it is hard to justify destroying sample information by the rank transformation.

There is, however, one important way in which the Kruskal-Wallis test is *less* attractive than rank-based tests in the two-sample context. Unlike its two-sample analogs, the Kruskal Wallis test does not have an easily computed null distribution. Implementations of the Kruskal Wallis test in statistics software typically address this deficiency by substituting approximations to the null distribution of *H,* but these approximations have poor accuracy for small sample sizes (Meyer and Seaman 2013). A solution is to construct the exact null distribution by permutation for small samples, but then the Kruskal-Wallis test does not even exhibit a modest computation advantage over the permutation *F* test—and is even harder to defend.

To illustrate how these considerations play out in an actual example, return to the Shanghai Auction experiment and the null hypothesis of no difference in treatment effect between any of the auction mechanisms. Recall that the permutation *F* test rejected the null hypothesis with a *p*-value of 0.021. An exact *p*-value for the Kruskal-Wallis test can be computed by following the same process, only replacing the *F* statistic with the *H* statistic. For the Shanghai Auction data, the observed value of the test statistic is $H = 5.6$. Since 84 of the 1,680 possible permutations yield *H* statistics greater than or equal to 5.6, the exact Kruskal-Wallis *p*-value is 0.05.[18]

[18] For comparison, the *p*-value of a typical chi-square approximation to the null distribution is 0.061.

5 Permutation Tests for $k > 2$ Independent and Ordered Samples

In the previous discussion, the direction of the alternative hypothesis was left unspecified. The null hypothesis of no treatment effect was compared to the agnostic alterative that at least two of the treatments differed from each other in central tendency. This alternative will often be appropriate, but sometimes experiments are designed so that treatments differ in intensity along a single dimension: e.g., the group size or the incentive to defect from a cooperative outcome in a social dilemma. In such cases, a directional alternative hypothesis may be more appropriate. In the case of increasing treatment intensity, for example, an alternative hypothesis of monotonically increasing treatment effect may be of primary interest.

While something like the all-purpose permutation *F* test is sensitive to the presence of ordered treatment effects—and thus a valid test even when an ordered treatment effect is expected—more powerful tests may be constructed to address these effects (Miller 1997). Much like the difference between one-sided and two-sided tests in the two-treatment context, specialized tests of ordered treatment effects are preferable when the underlying theory suggests this alternative hypothesis.

Where the relationship between treatments and hypothesized effects is approximately linear, correlation and regression methods provide an attractive basis for inference. We discuss the use of permutation arguments for inference in these models of association in Section 7. Here, we consider situations where linearity cannot be assumed. Linear models may not apply, for example, when treatments differ by broad, qualitative distinctions, such as when subjects are categorized into bins like risk averse, risk neutral, or risk seeking. More generally, situations may arise in which experimenters expect to see an ordered treatment effects but cannot be confident about more that the ordinal sequence in the relationship.

5.1 Permuting Ranked Observations: The Jonckheere-Terpstra Test

Consider a volunteer's dilemma game reported by Goeree, Holt, and Smith (2017). This experiment randomly matched subjects into groups, with group-size treatments of 2, 3, 6, 9, and 12 players. Each subject in a group was tasked with simultaneously choosing whether to incur a cost *C* to provide a public good to the group. The value of the public good was $V > C$ for each subject if provided. If no one volunteered to provide the public good, then the outcome was a low payoff of *L*, with $V - C > L$. The dilemma, in this game, is that each player prefers to provide the public good if no one else will do so but prefers not to incur the cost of volunteering if another player would do so. With simultaneous choices, the symmetric Nash equilibrium involves randomized volunteering, with the equilibrium probability of volunteering a decreasing and nonlinear function of the group size, g. Observed rates of volunteering for different group size treatments are reproduced in Table 5.

Table 5. Average Volunteer Rates by Group Size[a]

	$g = 12$	$g = 9$	$g = 6$	$g = 3$	$g = 2$
Session Average	0.188	0.194	0.28, 0.20, 0.31	0.42, 0.38, 0.39	0.55, 0.51, 0.49

[a] Goeree, Holt, and Smith (2017).

Since theory predicts that the rate of volunteering will grow as the group size shrinks, tests of a group-size treatment effects may compare the null hypothesis of no treatment effect against the alternative hypothesis that the average volunteer rate grows as one moves to the right across the columns (treatments) of Table 5. The most commonly used nonparametric test for ordered treatment effects is the Jonckheere-Terpstra test (Jonckheere, 1954; Terpstra, 1952).[19] When treatment categories are ordered so that the predicted effect increase from left to right (as in Table 5), the test statistic J is the sum of all "binary wins" in the predicted direction. In other words, J is the total number of larger observations in columns to the right of each observation:

$$J = \sum_{s=1}^{k-1} \sum_{t=s+1}^{k} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} 1(x_{i,s} < x_{j,t}) \tag{9}$$

where $x_{i,s}$ is the ith observation in the sample data from ordered treatment s, $x_{j,t}$ is the jth observation in the sample data from ordered treatment $t > s$, $1(\cdot)$ is the indicator function, and n_s, and n_t are respective sample sizes. Notice that the first summation, indexed by s, is over all

[19] A notable and early example of the use if this test is the classic Smith (1964) paper on "The Effect of Market Organization on Competitive Equilibrium." The comparison was between three price-setting institutions. The predicted order was that prices in markets with sequential seller offers would be below those in a double auction with both buyer bids and seller offers, which would be below those with only buyers making bids. A more recent example is Eckel and Füllbrunn (2015). There, the Jonckheere-Terpstra test helped to establish that price bubbles in asset market experiments with declining fundamental values are more frequent in groups with all male subjects, as compared with mixed or all-female groups. An example from political science can be found in the Siebert et al. (2013) experiment, in which the ordered treatments were probabilities that an aggressor in a political conflict would prevail if its demands were rejected. These win probabilities ranged from low to high: 0.2, 0.4, 0.6, and 0.8.

columns to the left of the final column, *k*. The second summation, indexed by *t*, is over all columns to the right of *s*, up to and including column *k*. The final two summations are over all pairs of observations in columns *s* and *t*, which are used to obtain a "less-than" count via the indicator function. For example, the volunteer rate of 0.188 for groups of size 12 in Table 5 is smaller than all 10 numbers to its right, so the first term in the sum for *J* would be a 10; the volunteer rate of 0.194 for groups of size 9 is smaller than all 9 numbers to its right, so the second term in the sum for *J* is 9; and so on.

For large sample sizes, approximate null distributions for *J* are available. For small samples like this, the researcher is left to search for a precomputed null distribution or to generate one via permutation. By the logic used in Section 4.1, the null hypothesis of no treatment effect implies that there are $n!/(n_1! \times \ldots \times n_k!)$ equally likely permutations of the observed sample data. The permutation *p*-value for this test is thus the same as for the *F* test described in equation (7) but with J_i and J_{obs} substituted in place of F_i and F_{obs}. Here, the observed value of the test statistic is $J_{obs} = 10 + 9 + (3 \times 6) + (3 \times 3) = 46$. There are $N = 11!/(3!)^3 = 184{,}800$ possible permutations of the data, of which the observed value of the test statistic is the largest value, so the *p*-value for the Jonckheere-Terpstra test is $1/184{,}800 < 0.001$.

The Jonckheere-Terpstra test is an attractive option when sample data are measured as ordinal values. When sample data are more than ordinal, however, this test operates like other rank-based tests in discarding potentially important information. The "binary win" (<) comparison makes only rank-order use of the differences between measured observations. For the data in Table 5, this flattening of the sample data is not an obstacle to strong rejection of the null hypothesis. In other contexts, it may be helpful to consider tests that are more sensitive to the magnitudes of cross-category differences.

5.2 Permuting Measured Observations: A Directional Difference Test

How might a magnitude-sensitive version of the Jonckheere-Terpstra test be constructed? A simple approach would be to replace the "binary win" count with a sum of differences, which could be either positive or negative.[20] As before, the treatment vectors of observations are listed as columns of a table that is ordered, left to right, in increasing order of predicted effect under the

[20] Another alternative is to construct a permutation test based on the ranks of binary differences (Shan, Young, and Kang, 2014).

alternative hypothesis. Let the test statistic D be the sum of all differences between each observation and all observations in columns to the left in the ordered array:

$$D = \sum_{s=1}^{k-1} \sum_{t=s+1}^{k} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} (x_{j,t} - x_{i,s}) \tag{10}$$

where all terms are defined as above. For the volunteer's dilemma data, the observed test statistic is $D = 8.19$, which is the most extreme value of the test statistic among 184,800 possible permutations of the data. The extremity of these data makes the choice between the Jonckheere-Terpstra test and the direction difference test academic. But the tests are not the same.

For example, consider the Traveler's Dilemma experiment reported in Capra et al. (1999). Subjects were randomly paired and made simultaneous "claims," subject to the following rules. First, each subject's claim had to fall within the range from 80 to 200. Second, after both claims were made, each subject would earn the smaller of the two claims, minus a penalty of *R* if that subject's claim was the larger claim and plus a reward of *R* if that subject's claim was the smaller one. No penalty or reward was applied if the two claims were equal. Since each subject has a unilateral incentive to undercut any claim greater than the minimum in this game, the unique Nash equilibrium is for both players to make the minimum claim, regardless of the size of the incentive parameter*, R*. This is counterintuitive, as common sense suggests that claims should fall with increases in the size of the penalty (for being high) and reward (for being low).

An experiment was conducted with session groups of 9-12 subjects, who were randomly matched into pairs for 10 rounds. Table 6 summarizes the treatments and data averages. A different value of *R* was applied to each session, as shown in the top row of Table 6. Average claims are shown in the second row. The most salient feature of the average claim data is the sensitivity to the size of the incentive parameter, *R*, which sharply contradicts the Nash prediction of no effect. Indeed, the data appear more consistent with the intuition that average claims vary inversely with the size of the incentive parameter.

Table 6. Session Claim Averages for a Traveler's Dilemma Game[a]

Incentive Parameter	*R* = 80	*R* = 50	*R* = 25	*R* = 20	*R* = 10	*R* = 5
Average Claim	81	85	138	119	186	195

[a] Capra et al. (1999).

Recall that the Jonckheere-Terpstra test statistic is the sum, over all observations, of the counts of larger observations in treatments with greater predicted value. Adding up the number of larger observations to the right of each column of Table 6, we get $J_{obs} = 5 + 4 + 2 + 2 + 1 = 14$. Upon inspection, there are 6 possible permutations of the claims data that would yield a J value at least as large as J_{obs}.[21] With $6! = 720$ possible permutations, the one-sided p-value for the Jonckheere-Terpstra test is thus $6/720 = 0.0083$. Next, consider the directional difference test statistic defined in equation (10). The sum of differences between every observation and all observations in columns to the right equals $D_{obs} = 854$. Not all reversals get the same weight in the directional difference test since the magnitude differences matter. Of the $6! = 720$ possible permutations, only 5 result in test statistics greater than or equal to D_{obs}, yielding a one-sided p-value for the directional difference test of $5/720 = 0.0069$.

This example contains two lessons. First, just as in previous comparisons, the availability of an intuitive test operating on the data as measured demands an explanation for preferring another test that discards information contained in the sample. Unless a precomputed null distribution is available, calculating an exact p-value for the Jonckheere-Terpstra test involves the same permutation approach as the directional difference test—so computation efficiency is not an answer. The rank-based approach may be attractive when the sample data are ordinal as measured. We conjecture that the rank-based approach could also have superior properties when the samples contain large outliers, though we are aware of no Monte Carlo analysis or formal proof to that effect. Outside of these limited circumstances, the directional difference test seems the better choice.

Second, there is potentially great research value in designing experiments with the high degree of treatment variation used in this example. Suppose that, instead of collecting one data point from each of six treatments, the researchers had followed the more conservative approach of collecting three data points for each of two treatments. The strongest possible rejection by a permutation test in the two-treatment context would be at a p-value of $1/20 = 0.05$, assuming no reversals between the treatments. Here, with the same number of data points spread across six treatments, the p-value is an order of magnitude smaller.

[21] The observed data in the table have one reversal between the middle two categories. There are 4 other permutations that yield reversals in adjacent treatments, each with a J value that is also 14. Finally, the case of no reversals would yield a J value of 15.

Several generalizations of the Jonckheere-Terpstra test that have been used for physical systems in which too much of a treatment (e.g., a drug) may have a negative effect. The "umbrella test" is for the case where the alternative hypothesis involves a "hill-shaped" data pattern as treatment intensity is increased (Mack and Wolfe, 1981). This test essentially involves combining two directional tests, one for data to the left of the mode, and a reversed test for data to the right.[22] Analogous test statistics could be devised for the directional difference test. This illustrates one of the most intriguing properties of permutation testing. As Pearson (1937) observed at the dawn of this methodology, by decoupling computation of the null distribution from the choice of test statistic, the permutation method leaves the experimenter free to select whatever test statistic is most sensitive to hypothesized relationship in a given setting.

6 Permutation Tests for $k > 2$ Dependent Samples

Just like the two-independent-sample permutation procedures, the analogous matched-sample procedures generalize in an intuitive way to higher order settings with k treatments per observational unit. In classical statistics texts, the study of treatment effects when the same subjects are exposed to multiple treatments in sequence is presented under the heading of two-way analysis of variance. In the permutation context, we find it easier to conceptualize the data generating process in terms of stratification. The following presents an intuitive guide to stratified permutation testing, first for a simple two-treatment context with an additional nuisance variable, and then in the more general case of multiple sample comparisons.

6.1 Stratified Permutation Tests for $k = 2$ Treatments with Nuisance Variables

A common problem in experimental data analysis is dealing with procedural differences in data groupings that are unrelated to the differences of interest. These procedural groupings are essentially nuisance variables: secondary treatments that would ideally be held constant when evaluating the effects of the primary treatments. For example, suppose an experiment involves two market treatments with each treatment run using subjects from two different subject pools. If the experimenter has reason to believe the subject pools are interchangeable, then observations

[22] Another variation would be to weight each of the directional differences in (10) by the corresponding difference in treatment intensities. This would diminish the impact of observed reversals when treatment intensities are close, as for the $R = 20$ and $R = 25$ treatments that resulted in the reversal in the middle columns of Table 6.

can be pooled by treatment and a simple two-sample test can be employed. But, if subject pools cannot be assumed to be interchangeable, things become more difficult. The experimenter can run separate tests for the treatment effect within each subject pool. But this approach has some important downsides, as illustrated below.

A stratified permutation approach is often a more attractive option. To explain what we mean, let the primary treatments be indexed by j, and let the secondary groupings be indexed by g. Thus, x_{ijg} denotes experimental observation i taken when treatment j is applied to subjects from group g. The idea behind stratified permutation testing is to construct the null sampling distribution by permuting the primary treatment labels *within* groups but not *between* groups. This captures the null hypothesis—that observed measurements are equally likely to be seen under any treatment—without imposing the additional assumption that observed measurements are equally likely to be seen under any secondary grouping. The stratified permutation procedure essentially allows the experimenter to test the null hypothesis of no treatment effect while holding the nuisance variable constant.

As a concrete example, consider the two-treatment asset market experiment reported by Holt, Porzio, and Song (2017). One treatment, applied to 14 sessions, involved a 25-period trading sequence, and the other treatment, applied to 10 sessions, involved a 15-period trading sequence. These markets were blocked on subject gender: half of the sessions in each treatment were female-only and half were male-only. In all treatments, the fundamental (present) value of asset shares was constant, at \$28, for all periods.[23] Price bubbles, with peaks well above \$28, were observed in all sessions. Table 7 shows peak asset prices for female-only sessions (top row) and male-only sessions (bottom row); the longer 25-period market sessions are on the left, while the shorter 15-period market sessions are on the right. The sessions run with only 15 periods afforded subjects less opportunity to accumulate large cash balances, resulting in smaller cash-asset-value ratios. This difference in cash-asset values motivated a question whether the peak asset prices were also lower in the shorter markets. For purposes of testing the null hypothesis of

[23] All-male and all-female markets were run in the same room at the same time to obscure gender sorting, but otherwise the markets were independent. The dividend and interest structures of the markets were all identical (with identical sequences of dividend realizations), and all markets had the same "flat" fundamental value of \$28 that equated the expected dividend return to the known interest rate paid on cash. (Holt, Porzio, and Song, 2017).

no trading-length treatment effect, trading-length treatments are the control variable of primary interest; gender groupings are a nuisance variable.

Table 7. Peak Prices by Market with Gender Sorting[a]

Market Pool	25-Period Markets	15-Period Markets	Mean
Female Only	87 95 61 177.5 75.5 37 152	66 36 58 64 42	79.3
Male Only	55 48 68 85 65 56.5 50	50 70 45 43 53	57.4
Mean	79.5	52.7	68.3

[a] Holt, Porzio, and Song (2017)

In testing the primary treatment effect despite the secondary nuisance variable, a stratified permutation test permutes peak price observations across session-length treatments (the columns of Table 7), but not across gender-groups (the rows of Table 7). A one-sided permutation test of the null of no treatment effect against the alternative of higher price peaks in longer markets thus considers the observed treatment difference in the bottom row of the table, $T_{obs} = 79.5 - 52.7 = 26.8$. This observed difference is compared to the null distribution of this statistic under the constrained set of permutations in which observations are moved between market-length treatments, but not across gender labels. There are $\binom{12}{5} = 792$ ways that treatment labels could be assigned to the numbers in the top row. For each of these top row permutations, there are another 792 ways that treatment labels could be reassigned in the bottom row. Thus, there are 627,264 total permutations to consider. Of these, there are only 6,259 that involve a treatment effect greater than or equal to the observed value, yielding a one-sided p-value of about 0.01.[24]

How does this compare to the alternative approach of conducting two separate permutation tests, one for each gender group? Conducting separate Pitman tests for each gender grouping yields one-sided p-values of 0.12 for the male-only group, and 0.028 for the female-only group. This illustrates the previously discussed problems with the multiple-comparison approach. Can the experimenter credibly conclude, in this multiple-comparison exercise, that the null hypothesis of no treatment effect is strongly rejected? Moreover, if the experimenter really was seeking to test every combination of treatment effect and gender grouping simultaneously, then

[24] Conversely, an analogous test that stratifies on the number of periods can be used to show that the null hypothesis of no gender effect on peak prices cannot be rejected (in this flat fundamental value setting).

the testing procedure should be constructed in a way that controls the family-wise error rate of these tests (the probability of falsely rejecting at least one null hypothesis among the two tests). While a detailed discussion of multiple-comparison adjustments is beyond the scope of this paper, it is helpful to note that Bonferroni-adjusted *p*-values—one way to control the family-wise error rate—are 0.24 for the male-only group and 0.056 for the female-only group.[25] These adjusted *p*-values are obviously greater than the 0.01 value that results from the joint test based on stratification.

Stratified permutation testing provides expositional simplicity as well as more statistical power in this setting. It does so by the simple act of tailoring the permutation strategy to the underlying randomization of the experiment's design. It bears emphasis that the null hypothesis for the stratified permutation test—that observations are drawn from the same distribution *within strata*—does not restrict observations to share a common distribution *across strata*. This leaves room for distributional differences between strata. Treatment effects should be expected to be the same across strata, though, or else separate tests would be appropriate.[26]

Before moving on, note that this stratification process generalizes easily to situations with multiple nuisance variables. A simple illustration is provided by Comeig, et al. (2017), who report an experiment designed to study risk appetite as a function of framing ("downside risk" vs "upside risk"), payoff scale, and subject gender. The experiment involved 256 subjects, half male and half female, each tasked with making a choice between a risky lottery and a safe lottery.[27] Half of the subjects were exposed to treatments in which the risky option was presented as downside risk (a small probability of a low payoff); the other half were presented the risky option as upside risk (a small probability of a large payoff). If small probabilities tend to be "over-weighted," as Prospect Theory predicts, then subjects would tend to shy away from the downside risk of a low payoff and to be attracted to the upside risk of a low probability of a high payoff. In every pairwise choice, both experimental lotteries were scaled such that the expected

[25] For a simple introduction to multiple-comparison problems, and adjustments, see Miller (1997: p. 75). For an example application of multiple-comparison adjustments of statistical tests using experimental data, see Holt et al. (2012).

[26] Put another way, the maintained assumption is that the shift model applies within each strata and that all shifts are of the same magnitude across strata.

[27] The paper contains data for 10 choice pairs, with one selected at random ex post for payment. Here, we restrict attention to a single pair of upside or downside risk choices that was used for the treatments in which each subject only made a single decision.

payoff from the risky lottery exceeded the payoff from the safe lottery by the same fixed amount. Finally, these treatments were blocked on payoff scale, with half of subjects presented payoffs five times larger than the other half. Of the 32 male and 32 female subjects exposed to each combination of risk profile and payoff scale, the number of subjects choosing the risky option is presented in Table 8, below.

Table 8. Risky Option Choice Proportions by Treatment[a]

Gender	Payoff Scale	Downside Risk	Upside Risk
Male	1x	25/32 = 78%	30/32 = 94%
Male	5x	17/32 = 53%	28/32 = 88%
Female	1x	19/32 = 59%	29/32 = 91%
Female	5x	4/32 = 13%	27/32 = 84%

[a] Comeig, et al. (2017) single choice data.

While these data may be used to explore various hypotheses, perhaps the most interesting prediction is a greater willingness to take upside risks than downside risks, even when the expected payoff and standard deviation of the safe and risky options are the same. In testing this hypothesis, both gender and payoff scale are nuisance variables.

A stratified permutation test of the effect of risk type on lottery choice would involve permuting risk-type labels across each of the 256 lottery choices, subject to the constraint that labels are never reassigned in ways that cross any of the strata in the different rows of the table. With two crossed nuisance variables, this equates to tracking four separate strata during the permutation process: male/1x, male/5x, female/1x, and female/5x. Otherwise, the procedure is the same as the one-nuisance-parameter case.

Given the large number of possible permutations for a sample of this size, an exact permutation test would be computationally costly. An approximate permutation test can be conducted by randomly sampling 1,000,000 or more possible permutations for purposes of constructing the null distribution. Here, we find that the number of random permutations resulting in risk-choice differences as or more extreme than that observed in the reported data implies a p-value of less than 0.001. Similar tests, omitted here, could be used to evaluate the effects of payoff scale or subject gender.

One of the most attractive properties of this stratified permutation approach is the intuitive nature of randomizing across only the dimension of the data at focus. Indeed, the stratification tests we present in this section are really just generalizations of the matched-pairs permutation strategy. In the matched-pairs context, each pair of observations is treated as its own stratum, while in the more general stratified permutation testing, multiple observations may fall within each stratum. Different strata may even have different numbers of observations. As a general strategy for conducting statistical inference in the presence of nuisance variables, stratified permutation testing strikes an attractive balance of analytical flexibility and ease of presentation.

6.2 Stratified Permutation Tests for $k > 2$ Treatments with Nuisance Variables

Recall that matched-pair samples arise in experimental designs in which subjects are exposed to two different treatment conditions in sequence. The matched-pair sample is a special case of a more general design in which a subject, or group of subjects, is exposed to $k > 2$ different treatments. In the language of classical statistics, the resulting data invite two-way analysis of variance: the experimenter may be primarily interested in studying the different effects of the k treatments, but analysis should also account for dependence relationships among the multiple measurements taken from subjects or subject groups.

A particularly clear example of this data structure is provided by Ma, Noussair, and Renneboog (2019), who report a laboratory experiment designed to solicit relative valuation of paintings of different color components. In a laboratory setting, the authors showed each of 465 unique subjects a set of 6 constructed Rothko-like paintings, each with a different primary color: blue, red, green, yellow, etc. Subjects were then provided an opportunity to purchase each painting under a bidding process that incentivized the revelation of private values for each painting.[28] Results indicated substantial differences in average valuation by color, with bids for red and blue paintings exceeding the average bid by about 17-19 percent.

With each subject viewing and bidding on multiple paintings, subject heterogeneity could be pronounced in this design. A subject who particularly liked art, or who happened to need a painting for decoration purposes, might bid higher on all 6 colors than would other subjects. The large sample size in this experiment unlocks various options for statistical inference but suppose, for sake of argument, that the experimenters wanted to employ a non-parametric approach. If the experiment had only compared two colors, then the matched-pairs tests discussed in Section

[28] Bids were solicited using the Becker, DeGroot, Marschak (1964) method.

could be used to study the treatment effect of interest while controlling for individual-specific heterogeneity. With 6 color treatments, a more general within-subject test is required.

That test is easily constructed by applying stratified permutation methods to the standard two-way ANOVA statistics. For the null hypothesis of no treatment effect between any of the 6 colors, two-way ANOVA with both color and subject factors provides an *F* statistic of 15.61 for the painting color factor. To generate an empirical null distribution for this test statistic, we can permute the color labels associated with each subject's bids. In other words, possible subject heterogeneity is controlled by limiting permutations to occur only within the strata of measurements taken from a given subject. Intuitively, this is a generalization of the label swapping permutations performed when studying matched-pairs samples.

This stratification constraint greatly compresses the permutation space, but with 6 color-print observations per subject and 465 unique subjects in the experiment, even the restricted permutation space entails $(6!)^{465}$ possible permutations of the sample data—far too many to exhaust computationally. Instead, we compute an approximate p-value by randomly sampling 9,999 within-strata permutations and computing the associated color-factor *F* statistic for each of them. This approximate test leads to strong rejection of the null hypothesis of equality of all color treatments, with no random permutation of the sample data leading to a larger value of the *F* statistic than the observed data: an approximate p-value of $1/10{,}000 < 0.001$. Having rejected the null hypothesis of equality of the 6 color treatments, further pairwise comparisons could be conducted along the lines described in Section 3.

The rank-based analogue of the test just described is the Friedman test (Friedman 1937, 1939, 1940). In case you are interested, especially if you are an economist, this is *Milton* Friedman who won a Nobel Prize in Economics many years later for other work. The title of the paper coveys the main motivation for the test: "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," and he argues that for socio-economic data, normality is "likely to be the exception than the rule." This paper was written before World War II, and more importantly, at a time when computers were just dream machines in the minds of people like John von Neumann. Friedman (1937, p. 675) notes that computations using ranks are "less arduous… requiring but a fraction of the time" and that this computational advantage will be enhanced for "those large scale collections of social and economic data which have become increasingly frequent in recent years…." Times have changed, and like other rank-based tests, exact computation of p-values in the Friedman test requires the same type of permutation that is required to construct empirical null distributions in the data-as-measured ANOVA approach. Like other rank-based tests, the present-day advantages of the Friedman test over direct permutation of the measured data are narrow and limited.

7 Permutation Tests for Linear Relationships

A final class of statistical tests arises in situations where the experimenter has reason to expect a linear, approximately linear, or linearizable relationship between the treatments and measured outcomes of interest. Often, this involves the desire to conduct statistical inference around a measure of correlation. Multiple regression models are another important area where a defensible basis for inference may be needed and lacking. In both cases, permutation methods provide a basis for small-sample statistical inference.

7.1 Tests of Correlation Coefficients

The need to test correlations arises with some frequency in experimental settings. The data in a correlation study consist of n pairs of observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. The null hypothesis of no correlation corresponds to a situation in which there is no monotonic relationship between the x and y values.[29] Stochastic independence of the x and y samples is a sufficient condition for absence of correlation. As explained in every basic course on applied statistics, the standard Pearson correlation coefficient r is a function of the products of deviations of the sample observations from their respective sample means:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

In conventional hypothesis testing, the observed value of the correlation coefficient, r_{obs}, would be compared to the sampling distribution of the correlation coefficient—something that would either be inferred from assumptions about the distributions of the x and y values under the null hypothesis, or approximated via reference to a limit theorem for large sample sizes. Just as

[29] For example, a symmetric "umbrella" data pattern of y values with x on the horizontal axis would yield a measured correlation of 0 (analogous to a flat least-squares regression line), despite the clear hill-shaped association between x and y values.

for the previously discussed locational tests, these standard approaches are difficult to defend for the small samples typical of economics experiments.

Fortunately, a permutation test can be easily constructed around a correlation coefficient. If there were no association between the variables, as the null hypothesis insists, then observed covariation would reflect exchangeable error in the experimental design. Because each observed value of the measured outcome would be equally likely to have been observed under every value of the treatment variable, the null distribution of the correlation test statistic can be constructed by computing the value of r under each of the $n!$ ways that one of the two variables could be reordered, holding the order of the other variable constant. For a two-sided test, the permutation p-value is thus computed as follows:

$$\text{permutation test of correlation, two-sided } p\text{-value } = \frac{\sum_{i=1}^{n!} 1(|r_i| \leq |r_{obs}|)}{n!} \qquad (11)$$

An opportunity for illustration is the Capra et al. (1999) experiment, described in Table 6. As a reminder, this experiment concerned a Traveler's Dilemma game with an unintuitive Nash equilibrium prediction that average claims should not depend on the size of a penalty/reward parameter, *R*. Intuitively, one would instead expect to see negative correlation between these variables. The observed value of the correlation between penalties and average claims is, in fact, negative at $r_{obs} = -0.876$. To test whether that observation is statistically significant, the null distribution of the correlation coefficient can be computed under all possible permutations of the observed claims data. Out of $6! = 720$ possible permutations of the average claim data, only 3 result in a correlation coefficient at least as small as the observed value. Thus, a one-sided p-value for a permutation test of negative correlation against the null is $3/720 = 0.004$, a significant rejection of the Nash hypothesis, despite the limited sample size. This p value is lower than the value of $5/720 = 0.007$ obtained from using the Directional Difference Test in Section 5.2. The intuition behind this sensitivity is that the "xy products" that appear in the calculation of the correlation coefficient are determined by the magnitudes of both the data point and the treatment measure for that pair. Hence, the reversal of average claims in the middle two columns of Table 6 has a smaller impact because the *R* treatment values for those columns, 20 and 25, are relatively close together.

As should be obvious by this point, nothing about the permutation strategy just described depends on the particular choice of the Pearson correlation coefficient as the test statistic. The Pearson statistic made sense for the theoretically continuous data under investigation, but other experimental contexts could motivate the use of other statistics. In an “ink bomb” risk aversion experiment, for example, a subject is shown 12 boxes and allowed to check as many boxes as desired, understanding that each box checked earns the subject $1 unless and until a randomly hidden ink bomb is encountered—in which case nothing is earned. This setup can be used to elicit risk aversion, but the mapping from number-of-boxes-checked to some measure of risk aversion depends on the measure of risk aversion used (e.g., constant relative risk aversion) and is nonlinear in any event. In this application, it might make sense to rank subjects by the number of boxes they choose to check and by some other proxy for risk aversion, such as amount saved for retirement. In this ordinal context, a rank-based measure of association—such as Kendall’s τ or Spearman’s ρ statistic—would likely be a more defensible correlation concept. The same approach described above could be used to arrive at a permutation *p*-value. The only difference would be which measure of association was used as the test statistic.

7.2 Tests of Linear Regression Models

Moving beyond correlation analysis, permutation methods can also be used to conduct statistical inference for linear models. There are, however, significant limitations to the use of permutation inference in this setting. Obvious permutation strategies are forthcoming only for a few special regression models. In most cases, the experimenter will face a choice of different permutation strategies, and effort may be required to identify the appropriate strategy for the application. Standard *parametric* permutation tests do not exhibit these difficulties and, for even moderate sample sizes, may be robust enough to deviations from parametric assumptions to support credible inference (Kennedy 1995). For small sample sizes, however, permutation tests will still constitute a more reliable basis for inference and should be preferred.

Starting with one of the lucky special cases for permutation testing, consider the following bivariate data generating process:

$$\boldsymbol{y} = \alpha + \delta \boldsymbol{z} + \boldsymbol{\epsilon} \qquad (12)$$

for $\boldsymbol{\epsilon}$ a mean-zero error term unrelated to the value of the regressor $\boldsymbol{z}$. Suppose interest is in testing the null hypothesis $\delta = 0$. Under the null hypothesis, all variation in the elements of $\boldsymbol{y}$ is attributable to random error, $\boldsymbol{y_0} = \alpha + \boldsymbol{\epsilon}$, so every element of in the $\boldsymbol{y}$ vector is equally likely to have been paired with every element of the $\boldsymbol{z}$ vector. This suggests a simple permutation strategy: compare the observed *t* statistic for the least squares estimate of δ against the set of *t* statistics calculated under all possible permutations of the order of elements in the $\boldsymbol{y}$ vector while holding fixed the order of elements in the $\boldsymbol{z}$ vector.[30] This should look familiar. It is a simple application of the correlation-coefficient permutation strategy described in Section 7.1.

Now consider the more general case of multiple regressors with nuisance variables. Specifically, suppose the data generating process has the following form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\delta} + \boldsymbol{\epsilon} \tag{13}$$

for $\boldsymbol{\epsilon}$ an error term as before, $\boldsymbol{X}$ a matrix of nuisance variables which may include a constant term, and $\boldsymbol{Z}$ a matrix of regressors of interest. Under the null hypothesis that $\boldsymbol{\delta} = \boldsymbol{0}$, each element of the response vector is now more than random error: $\boldsymbol{y_0} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. This makes the simple permutation strategy of reordering $\boldsymbol{y}$ generally indefensible, as variation in the $\boldsymbol{y}$ vector is partly attributable to the influence of the nuisance variables in $\boldsymbol{X}$. One might think that this could be solved by either subtracting $\boldsymbol{X\beta}$ from both sides of the equation or by permuting the rows of the $\boldsymbol{X}$ matrix in lockstep with the rows of the $\boldsymbol{y}$ vector but neither of these strategies is very attractive. The first requires knowledge of $\boldsymbol{\beta}$, which is unavailable in most interesting cases. The second fails to preserve collinearity between $\boldsymbol{X}$ and $\boldsymbol{Z}$.

In fact, while many permutation strategies have been suggested for the multiple regression context, there remains no generally accepted permutation approach for this problem. A full survey of the literature is beyond the scope of this paper but helpful surveys are provided by Kennedy (1995), Kennedy and Cade (1996), Manly (2007: ch. 8), Anderson and Robinson (2001), and Winkler et al. (2014). To illustrate one intuitive option that has surfaced in the literature, consider the following permutation strategy due to Freedman and Lane (1983), as articulated by Kennedy (1995):

[30] In the one-regressor context, the value of the parameter estimate is also a suitable test statistic (instead of the *t* statistic). This does not hold true for more complicated models.

1. Fit the full model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\delta} + \boldsymbol{\epsilon}$ and compute the observed *F* statistic for testing the null hypothesis that $\boldsymbol{\delta} = \mathbf{0}$; where the null hypothesis restricts only one parameter, the relevant *t* statistic is an equivalent test statistic.
2. Fit the reduced model, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ and use this model to compute a reduced-model prediction vector $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and a reduced-model residual vector $\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$.[31]
3. Permute the order of the reduced-model residual vector $\boldsymbol{r}$ and add each permutation of the residual vector to the reduced-model prediction vector $\widehat{\boldsymbol{y}}$ to generate a new permutation of the $\boldsymbol{y}$ vector. For the observed data, this reconstructs the observed $\boldsymbol{y}$ vector. For all other permutations, it constructs a new $\boldsymbol{y_p}$ vector in which only the variation *not explained* by the nuisance variables is being permuted.
4. For each such permutation, fit the full model $\boldsymbol{y_p} = \boldsymbol{X\beta} + \boldsymbol{Z\delta} + \boldsymbol{\epsilon}$ and compute the *F* statistic for testing the null hypothesis that $\boldsymbol{\delta} = \mathbf{0}$.
5. Compare the observed value of the *F* statistic from step 1 to the permutation distribution to compute a *p*-value for this test.

As a concrete illustration, consider data from an asset market experiment reported by Harper et al. (2021). This experiment comprised 12 sessions, each with different groups of 9 subjects who traded asset shares that paid dividends in a sequence of periods and were then redeemed for an unannounced final redemption value. The underlying fundamental value of each share was not publicly known—since it was based on dividends, interest paid on cash, and an unannounced final redemption value—but a variable number of traders (1, 3, 6, or 9) were "insiders" who were informed of the final redemption value. The strong form of the efficient markets hypothesis implies that asset prices will summarize all information, private and public. Some support for this prediction has been reported for experiments with trade for a *nondurable* asset that only pays a dividend in a single period. Harper et al. (2021) sought to determine whether the anticipated speculative bubbles for a durable asset would blur or negate any effects of insider information. The hypothesis being tested was that an increase in the degree of publicness of information (with a higher proportion of insiders) would reduce deviations from the fundamental value, motivating a one-sided test.

[31] We abuse notation for simplicity in this section. Obviously, the unobserved error vector ($\boldsymbol{\epsilon}$) and parameter values will differ between one specification of the model and the next.

Three sessions were conducted for each of these insider treatments, as shown in the leftmost column of Table 9. Because cognitive scores are one of the prominent factors that have been reported to diminish the magnitudes of asset price bubbles in multiple laboratory experiments (Holt, 2019, p. 446), subjects were given a 3-question cognitive response test, and the average score for the traders in the session is given in the middle column. This nuisance variable is continuous, which precludes a stratification approach. Peak deviations from fundamental value are reported in the rightmost column. Most sessions resulted in robust price bubbles.

To first illustrate the comparatively simple permutation process for bivariate regression, consider a simple linear regression of peak deviation ($\boldsymbol{y}$) on number of insiders ($\boldsymbol{z}$). Fitting the model $\boldsymbol{y} = \alpha + \delta\boldsymbol{z} + \boldsymbol{\epsilon}$ via OLS yields the following parameter estimates and standard errors:

$$\begin{array}{cc} \hat{\alpha} = 22.506 & \hat{\delta} = -1.269 \\ (4.881) & (0.866) \end{array}$$

The *t* statistic for the estimate of δ is $t_{obs} = -1.269/0.866 = -1.465.$ Under the null hypothesis that $\delta = 0$ (the number of insiders has no effect), all variation in the $\boldsymbol{y}$ vector is attributable to exchangeable error. Therefore, the null distribution of the *t* statistic can be computed by permuting the order of the $\boldsymbol{y}$ vector and recalculating the value of the *t* statistic at each permutation. With 12 observations, the full set of all 12! permutations numbers in the hundreds of millions, so we instead randomly shuffle the y values 99,999 times to construct an approximate null distribution. Of the 100,000 total permutations considered in this manner, 8,505 yield *t* statistics that are at least as small as t_{obs}, resulting in a one-sided test p-value of 0.085.

Table 9. Peak Price Deviations from Fundamental Value in Asset Markets with 9 Traders and a Variable Number of Insiders Who Know the Final Share Redemption Value[a]

Number of Insiders ($\boldsymbol{z}$)	Average Cognitive Score ($\boldsymbol{x}$)	Peak Deviation ($\boldsymbol{y}$)
1	1.889	\$18.20
1	1.556	\$17.00
1	1.444	\$34.37
3	1.333	\$7.00
3	1.444	\$30.00
3	1.444	\$19.37
6	1.333	\$6.62
6	1.444	\$8.00
6	1.667	\$13.86
9	1.444	\$23.20
9	2.111	\$17.00
9	1.222	\$3.09

[a] Harper, et al. (2021).

Next, consider the more complicated linear model that allows for both average cognitive score ($\boldsymbol{x}$) and number of insiders ($\boldsymbol{z}$) to influence peak deviations ($\boldsymbol{y}$). Fitting the full model $\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \delta\boldsymbol{z} + \boldsymbol{\epsilon}$ via OLS yields the following parameter estimates and standard errors:

$$\begin{array}{ccc} \hat{\alpha} = 9.626 & \hat{\beta} = -8.464 & \hat{\delta} = -1.280 \\ (17.683) & (11.149) & (0.885) \end{array}$$

For purposes of testing the joint null hypothesis that $\beta = \delta = 0$, the approach discussed above could be repeated with the *F* statistic for that hypothesis substituted in place of the *t* statistic. For purposes of testing the less restrictive null that $\delta = 0$, however, a more involved permutation scheme is needed. The challenge is to find a way of approximating the stratified permutation process for a nuisance variable that does not actually divide observations into discrete strata.

Following the Freedman and Lane (1983) procedure outlined above, we can start by noting the observed *t* statistic value of $t_{obs} = -1.280/0.885 = -1.446$ for the estimate of δ. Fitting the reduced model $\boldsymbol{y} = \alpha + \beta\boldsymbol{x} + \boldsymbol{\epsilon}$ via OLS then yields the two components needed to construct approximate permutations of the $\boldsymbol{y}$ vector: (1) a vector of predicted values from this reduced regression and (2) a vector of residuals from the reduced regression. To construct an approximate null distribution, we shuffle the residuals vector 99,999 times, each time adding it to the predicted value vector to form a new permutation of the $\boldsymbol{y}$ vector, fitting that new $\boldsymbol{y}$ vector to the full model, and recording the *t* statistic associated with the estimated value of δ. Of the 100,000 total permutations considered in this manner, 8,672 yielded *t* statistics smaller than the observed value, resulting in a one-sided test *p*-value of 0.086. Accounting for the influence of the nuisance variable makes little difference in this particular case but could be of great importance in others.

8 Conclusion

Our objective in this paper is to survey and illustrate the use of permutation-based tests for conducting statistical inference with experimental data. Two themes stand out. First, permutation tests are best understood not a collection of related procedures but as a unified framework for conducting statistical inference when working with experimental data. Second, permutation tests that operate on measured data will generally constitute a more intuitive and defensible basis for inference than permutation tests based on rank-transformed data. Both themes are reflected in Table 10, which organizes the various tests that we discussed.

The rows of Table 10 illustrate the common framework that underlies all permutation tests. Starting from the experimental design and the null hypothesis to be tested, the researcher first observes the appropriate permutation strategy. Then an appropriate test statistic is selected, and statistical inference is conducted by comparing the observed value of the test statistic against the empirical null distribution of that test statistic under the permutation strategy. Every permutation test is an application of this common process.

The columns of Table 10 illustrate the opportunity cost of relying on familiar rank-based tests. Every rank-based permutation test is an application of a more general permutation test to rank-transformed values of the measured data. When the measured data are more than ordinal, the rank transformation discards information contained in the sample data. While this may have important consequences for power in particular applications, we rest our critique on the more basic point that the rank transformation is unintuitive and unnecessary in most cases. Use of rank-based permutations tests may be justified by properties of the experimental design or the measured data but should not be the uncritical default option that it is today.

Table 10. Overview of Methodology

	Feature of Data Subject to Permutation	
	Measured Observations	Rank-Transformed Observations
$k = 2$ independent samples $\binom{m}{n}$ permutations	**Pittman Permutation Test** difference in sample averages	**Mann-Whitney Test** difference sample averages for ranked data
$k = 2$ paired samples 2^n permutations	**Fisher Permutation Test** sum of differences	**Wilcoxon Test** sum of signed ranks of differences
$k > 2$ independent samples unordered $\frac{n!}{n_1! \times \ldots \times n_k!}$ permutations	**Permutation *F* Test** relative variance in measured values	**Kruskal-Wallace Test** relative variance in ranked data
$k > 2$ independent samples, ordered $\frac{n!}{n_1! \times \ldots \times n_k!}$ permutations	**Directional Difference Test** sum of directional differences	**Jonckheere-Terpstra Test** sum of directional “binary wins”
$k = 2$ dependent samples, nuisance variables restricted permutations	**Stratified Permutation Test** sum of sample differences across strata	**Stratified Permutation Test** sum of sample differences in average ranks within strata
$k > 2$ dependent samples, nuisance variables restricted permutations	**Stratified Permutation *F* Test** relative variance in measured values	**Friedman Test** relative variance in ranked data
Linear relationships $n!$ Permutations	**Correlation or Regression** Pearson’s correlation coefficient, OLS estimators	**Correlation Test** rank-based statistic like Kendall’s τ or Spearman’s ρ

Table 10 also illustrates how the nature of the statistical test used for either type of data, ordinal or ranked, depends on the treatment structure of the experiment. The flip side of this observation is the importance of designing an experiment to shine a bright light on the questions of interest. Experimentalists should consider going beyond the standard treatment-and-control framework. One alternative is the use of a wide range of intensity-based treatments to generate *k* ordered samples, which can yield a surprisingly large gain in sensitivity to treatment effects even with very small sample sizes, as was illustrated in the traveler’s dilemma example. In addition, the stratified permutation tests shown in the fifth row of Table 10 offer two important advantages. The first is that the combination of data analysis from different strata avoids multiple-test correction issues and can afford greater sensitivity to treatment effects by using all the data in a single test. The second advantage is that permutations within separate strata permit a coherent analysis of richer experiment designs by focusing on treatment variations, one at a time, while controlling for nuisance variable differences between strata. Indeed, the presence of nuisance variables (from procedural variations or secondary treatments) is more the rule than the exception for experiments in social sciences. The stratification solution used can be applied more generally. For example, ordered directional tests in the fourth row can be adapted to control for nuisance variables by using stratification. In this case, the test statistic would be constructed as a sum of order-based binary win comparisons across different strata.

The approach to permutation testing that we describe in this paper moves fluidly between experimental design and data analysis—selecting and even customizing tests to fit the specific design choices and research goals of particular experiments. Although software for constructing these types of permutation tests is increasingly accessible and convenient,[32] we acknowledge that this approach requires more effort than simply reaching for a familiar canned statistical test. We think the game is worth the candle. Our hope is that this paper inspires a richer use of the range of permutation tests available to experimental researchers, and a stronger and more efficient use of data we collect.

[32] The R scripts used to perform all tests described in this paper are available as an online appendix. The publicly available V*e*conlab experiment website also allows users to perform standard permutation tests with independent samples, matched samples, and a range of possible clusters for stratified permutation testing: http://veconlab.econ.virginia.edu\rand\rand.php. As of the time of writing, convenient building blocks for conducting permutation tests can be found in many widely accessible programs and scripting languages, including R, Python, Matlab, Stata, and Excel.

References

Anderson, Marti J. and John Robinson (2001) "Permutation Tests for Linear Models," *Australian & New Zealand Journal of Statistics*, 43(1), 75–88.

Anderson, Marti J. and Cajo J. F. ter Braak (2003) "Permutation Tests for Multi-factorial Analysis of Variance," *Journal of Statistical Computation and Simulation*, 73(2), 85–113.

Becker Gordon M., Morris H. DeGroot, and Jacob Marschak (1964) "Measuring Utility by a Single-Response Sequential Method," *Behavioral Science* 9(3), 226-232.

Berry, Kenneth J., Janis E. Johnston, and Paul W. Mielke, Jr. (2019) *A Primer of Permutation Statistical Methods*, Springer.

Bohr, Clement E., Charles A. Holt, and Alexandra V. Schubert (2019) "Assisted Saving for Retirement: An Experimental Investigation," *European Economic Review*, 119, 42-54.

Boik, Robert J (1987) "The Fisher-Pitman Permutation Test: A Non-robust Alternative to the Normal Theory F test when Variances are Heterogeneous," *British Journal of Mathematical & Statistical Psychology*, 40(1), 26-42.

Capra, C. Monica, Rosario Gomez, Jacob Goeree, and Charles Holt (1999) "Anomalous Behavior in a Traveler's Dilemma," *American Economic Review*, 89(3), 678-690.

Chung, EunYi and Joseph P. Romano (2016) "Asymptotically Valid and Exact Permutation Tests Based on Two-Sample U-Statistics," *Journal of Statistical Planning and Inference*, 168, 97–105.

Comeig, Irene, Charles A. Holt, and Ainoah Jaramillo (2017) "Dealing with Risk: Gender, Stakes, and Probability Effects," Discussion Paper, University of Virginia.

Davis, Douglas D. and Charles A. Holt (1994) "Market Power and Mergers in Laboratory Markets with Posted Prices," *The RAND Journal of Economics*, 25(3), 467-487.

Eckel, Catherine C. and Sascha C. Füllbrunn (2015) "Thar SHE Blows? Gender, Competition, and Bubbles in Experimental Asset Markets," *American Economic Review*, 105(2), 906-920.

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

Friedman, Milton (1937) "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, 32, 675-701.

Friedman, Milton (1939) "A Correction," *Journal of the American Statistical Association*, 34, 109.

Friedman, Milton (1940) "A Comparison of Alternative Tests of Significance for the Problem of *m* Rankings," *Annals of Mathematical Statistics*, 11, 86-92.

Freedman, David and David Lane (1983) "A Nonstochastic Interpretation of Reported Significance Levels," *Journal of Business & Economic Statistics*, 1, 292-298

Gibbons, Jean D. and S. Chakraborti (2003) *Nonparametric Statistical Inference*, New York: Marcel Dekker.

Goeree, Jacob K., Charles A. Holt, and Angela M. Smith (2017) "An Experimental Examination of the Volunteer's Dilemma," *Games and Economic Behavior*, 102(C), 305–315.

Harper, Daniel Q., Charles A. Holt, and Maggie Isaacson (2021) "The Efficient Market Hypothesis in Experimental Asset Markets: Private Information, Public Information, and Bubbles," Discussion Paper, University of Virginia, presented at the 2021 North American Economic Science Association Meeting in Tucson, Arizona.

Hayes, Andrew F. (2000) "Randomization Tests and the Equality of Variance Assumption When Comparing Group Means," *Animal Behaviour*, 59, 653–656.

Holt, Charles A. (2019) *Markets, Games, and Strategic Behavior: An Introduction to Experimental Economics*, Princeton NJ, Princeton University Press.

Holt, Charles A., Cathleen A. Johnson, Courtney A. Mallow, and Sean P. Sullivan (2012) “Water Externalities: Tragedy of the Common Canal,” *Southern Economic Journal*, 78(4), 1142-1162.

Holt, Charles A., Megan Porzio, and Michelle Song (2017) “Price Bubbles, Gender, and Expectations in Experimental Asset Markets,” *European Economic Review*, 100, 72-94.

Hoeffding, Wassily (1952) “The Large-Sample Power of Tests Based on Permutations of Observations,” *Annals of Mathematical Statistics*, 23(2), 169-192.

Jonckheere, A. R. (1954) "A Distribution-free *k*-sample Test against Ordered Alternatives," *Biometrika*, 41, 133–145.

Kagel, John H. and Alvin E. Roth (2000) “The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment,” *Quarterly Journal of Economics*, 115, 201-237.

Kempthorne, Oscar and T. E. Doerfler (1969) “The Behaviour of Some Significance Tests Under Experimental Randomization,” *Biometrika*, 56(2), 231-248.

Kennedy, Peter E. (1995) “Randomization Tests in Econometrics,” *Journal of Business & Economic Statistics*, 13(1), 85-94.

Kennedy, Peter E. and Brain S. Cade (1996) “Randomization Tests for Multiple Regression,” *Communications in Statistics - Simulation and Computation*, 25(4), 923-936.

Kruskal, W. H. and W. A. Wallis (1952) “Use of Ranks in One Criterion Variance Analysis,” *Annals of Mathematical Statistics*, 47, 583-621.

Liao, Evan Zuofu and Charles A. Holt (2015) "The Pursuit of Revenue Reduction: An Experimental Analysis of the Shanghai License Plate Auction," Discussion Paper, University of Virginia.

Ma, Marshall X., Charles N. Noussair, and Luc Renneboog (2019) "Colors, Emotions, and the Auction Value of Paintings," CentER Discussion Paper No. 2019-006.

Mack, G. A. and D. A. Wolfe (1981) "K-sample Rank Tests for Umbrella Alternatives," *Journal of the American Statistical Association,* 76, 175-181.

Mann, H. B. and D. R. Whitney (1947). "On a Test of whether One of Two Random Variables is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, 18(1), 50–60.

Manly, Bryan F.J. (2007) *Randomization, Bootstrap and Monte Carlo Methods in Biology* (3rd ed.), Boca Raton: Chapman & Hall/CRC.

Meyer, J. Patrick and Michael A. Seaman (2013) "A Comparison of the Exact Kruskal-Wallis Distribution to Asymptotic Approximations for All Sample Sizes up to 105," *The Journal of Experimental Education*, 81(2), 139-156.

Miller, Rupert G. (1997) *Beyond ANOVA: Basics of Applied Statistics*, Boca Raton: Chapman & Hall/CRC.

Moir, Robert (1998) "A Monte Carlo Analysis of the Fisher Randomization Technique: Reviving Randomization for Experimental Economists," *Experimental Economics*, 1(1), 87-100.

Neuhäuser, Markus and Bryan F. J. Manly (2004) "The Fisher-Pitman Permutation Test When Testing for Differences in Mean and Variance," *Psychological Reports*, 94, 189-194.

Page, E. B. (1963) "Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks," *Journal of the American Statistical Association*, 58(301), 216-230.

Pearson, E. S. (1937) "Some Aspects of the Problem of Randomization," *Biometrika*, 29, 53-64.

Pearson, Karl (1895) "Notes on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London*, 58, 240–242.

Pitman, E. J. G. (1937a) "Significance Tests Which May Be Applied to Samples from Any Populations," *Supplement to the Journal of the Royal Statistical Society*, 4(1), 119–130.

Pitman, E. J. G. (1937b) "Significance Tests Which May Be Applied to Samples from Any Populations. II. The Correlation Coefficient Test," *Supplement to the Journal of the Royal Statistical Society*, 4(2), 225–232.

Pitman, E. J. G. (1938) "Significance Tests Which May Be Applied to Samples from Any Populations III. The Analysis of Variance Test," *Biometrika*, 29(3), 322–335.

Romano, Joseph P. (1990) "On the Behavior of Randomization Tests Without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85(411), 686-692.

Shan G, Young D. and L. Kang (2014) "A New Powerful Nonparametric Rank Test for Ordered Alternative Problem" *PLoS ONE* 9(11), e112924. doi:10.1371/journal.pone.0112924

Siebert, K., D. Clark, C. Holt, T. Nordstrom, and W. Reed (2013) "An Experimental Analysis of Asymmetric Power in Conflict Bargaining," *Games and Economic Behavior*, 2013, 4(3), 375-397.

Siegel, Sidney (1956) *Non-parametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.

Smith, Vernon L. (1964) "The Effect of Market Organization on Competitive Equilibrium," *Quarterly Journal of Economics,* 78, 181-201.

Starmer, Chris and Robert Sugden (1991) “Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation,” *American Economic Review*, 81(4), 971-978.

Terpstra, T. J. (1952) “The Asymptotic Normality and Consistency of Kendall’s Test against Trend, when Ties are Present in One Ranking,” *Indagationes Mathematicae*, 14, 327-333.

van de Wiel, M. A., A. D. Bucchianico, and P. van der Laan (1999) “Symbolic Computation and Exact Distributions of Nonparametric Test Statistics,” *The Statistician*, 48(4), 507-516.

Wilcoxon, Frank (1945) “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, 1(6), 80–83.

Winkler, Anderson M., Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols (2014) “Permutation Inference for the General Linear Model,” *NeuroImage*, 92, 381–397.