Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs

T. Lynn Eudey
Joshua D. Kerr
Bruce E. Trumbo
California State University East Bay

Journal of Statistics Education Volume 18, Number 1 (2010),
www.amstat.org/publications/jse/v18n1/eudey.pdf

Key Words: Blocking; chi-squared test; goodness-of-fit; nonparametric test; paired data; pedagogy; S-language; *T*-test; two-sample.

Abstract

Null distributions of permutation tests for two-sample, paired, and block designs are simulated using the R statistical programming language. For each design and type of data, permutation tests are compared with standard normal-theory and nonparametric tests. These examples (often using real data) provide for classroom discussion use of metrics that are appropriate for the data. Simple programs in R are provided and explained briefly. Suggestions are provided for use of permutation tests and R in teaching statistics courses for upper-division and first year graduate students.

1. Introduction

Important issues in the practical application of hypothesis testing are to understand the purpose and nature of the data, to measure the size of an effect, and to determine whether the effect is statistically significant. We believe that a consideration of various permutation tests provides a useful framework for discussing these issues. The computer software package R (R Core Development Team, 2008) is useful in finding very close approximations to exact distributions. Normal approximations (or other distributional approximations) are then less important. We provide small bits of R code so that the reader can follow the logic of the programming. The remainder of the programs can be found in an appendix and on the JSE website.

Permutation tests are often taught as part of a statistics curriculum, particularly in nonparametric courses (Tarpey, Acuna Cobb, De Veaux, 2002). [It is not our intent here to show details of standard nonparametric tests or to provide a comprehensive introduction to permutation tests. Coakley, 1996, gives an extensive bibliography of nonparametric texts and monographs, including Capéraà and Van Cutsem, 1988, and Sprent, 1993. An additional reference of interest is chapter 10 of Stapleton, 2008. Good, 2000, provides a practical guide to permutation tests. Additionally, Moore, 2007 and Utts and Heckard, 2006 provide introductions to nonparametric tests in supplemental chapters of their basic textbooks. Hesterberg, Monaghan, Moore, Clipson and Epstein 2003 provide a supplemental chapter on the web which uses S-Plus®: http://bcs.whfreeman.com/pbs/cat_160/PBS18.pdf] Callaert (1999) states: "Students in an applied statistics course offering some nonparametric methods are often (subconsciously) restricted in modeling their research problems by what they have learned from the *T*-test. When moving from parametric to nonparametric models, they do not have a good idea of the variety and richness of general location models." Permutation tests can also provide insight to the general concepts of hypothesis testing and allow us to use different metrics. Wood (2005) suggests the use of simulation models for "deriving bootstrap confidence intervals, and simulating various probability distributions." Wood points out that the use of a simulation model provides an alternative approach to that of "deriving probabilities and making statistical inferences" that is more accessible to the student. Aberson, Berger, Healy and Romero (2002) describe an interactive web-based tutorial for illustrating statistical power. Thus, the use of repeated sampling to illustrate statistical concepts is not new to the classroom. And permutation tests provide a useful alternative when the distributional assumptions for parametric tests are under question.

Nonparametric tests are often permutation tests and for small samples the exact P-values can be found in many references (Hollander and Wolfe (1999) for example). For large samples the traditional approach is to use asymptotic parametric distributions to find approximate P-values. We suggest using simulations (programmed in R) to obtain approximate P-values by taking a large random subset of all the possible permutations of the data.

We use the computer software package R (R Core Development Team, 2008) to perform the simulations and provide P-values. R and S-Plus® are very similar derivatives of the S programming language. Authors (Verzani, 2008, Hodgess, 2004) in this journal advocate the use of R in both introductory and advanced courses. As Verzani points out: "Why R? There are many good reasons. R is an open-source project, hence free for the students and the institution to use. R is multi-platform (Windows, Mac OS X, and Linux). It has excellent graphics. It has an extensive collection of add-on packages, such as pmg."

In this article, we present the use of simulation (re-sampling) to develop approximate distributions for permutation test statistics using a variety of data (categorical, ordinal and continuous). Several metrics for testing are explored with the take-home message that the metrics for "center" and "dispersion," and consequently test statistics, should fit the nature of the data. We use *metric* here and distinguish metric from a *test statistic* in that the metric is a raw measure of a population characteristic that has not been standardized (examples are median as a metric for the center and range as a metric for dispersion). Permutation tests can be built on the distribution of the metric itself (without standardization). We estimate the permutation distribution of the metric by re-sampling.

The objectives of this article are to explore permutation tests as a viable alternative to the parametric location tests, to explore the performance of the different metrics, and to acquaint the student with the use of R. Several suggested student exercises for each section to illustrate these concepts are provided in Appendix A. We have included some bits of code in the text and representative R programs (with color coded alternative code in comment lines) in Appendix B to illustrate the structure of R and how to perform simulations in R. Full R code for all the figures can be found on the JSE website at: http://www.amstat.org/publications/jse/v18n1/Permutation/Welcome.html

Traditionally, a drawback to the use of permutation tests has been the difficulty of determining the permutation distribution of the metric under the null hypothesis. In a few very simple situations, it is possible to use elementary combinatorial methods to find the permutation distribution, or at least some of its key tail probabilities. For moderate or large sample sizes it may be possible to obtain sufficiently accurate approximations to tail probabilities using asymptotic methods. Nowadays, it is common to simulate the permutation distribution using a statistical package. In the simulation approach, one does not attempt to enumerate all possible permutations, but randomly to sample enough of them to get a serviceable approximation to the permutation distribution. Here we use R to do the sampling, summarize the results, and find a close approximation to the true P-value. This idea is not new, Dwass (1957) points out "The main point is that instead of basing our decision on $\binom{m+n}{n}$ permutations of the observations, we can base it on a smaller number of permutations and the power of the modified test will be 'close' to that of the most powerful nonparametric test." R makes this process readily available to the instructor and the student. Jöckel (1986) explores the Pitman efficiencies of this approach, thus justifying its use.

In a beginning upper-division statistics course where it is possible to project a computer screen, an instructor might briefly discuss a particular permutation test, show what is being simulated by our corresponding program, demonstrate a run of the program, and discuss the output. In a more advanced class, students might be asked to make slight changes in our programs to get results for a different dataset or to use a different metric. Nowadays, R is so widely used in industry and research that familiarity with R can be an important skill in the job market. Because R is free software readily downloaded from the web (www.R-project.org), students should have no trouble obtaining R for use on personal computers.

Primers in the use of R are widely available and it is not our purpose here to provide a detailed introduction to its use. [See especially the current version of Venables and Smith: *An Introduction to R,* http://cran.r-project.org/doc/manuals/R-intro.pdf, Verzani: *Simple R: Using R for Introductory Statistics*, http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf, and the books by Dalgaard (2002) and Rizzo (2008).] However, because all of our permutations are performed with the random function `sample` in R, we begin by showing briefly how this function works. We say this is a random function because it uses a pseudo-random number generator to sample a (possibly) different result each time the function is used. Outcomes from the state-of-the-art generator in R are, for practical purposes, not distinguishable from random samples. As we use it, the `sample` function has three arguments: the first is a vector designating the population, the second is the number of items sampled from that population, and the third is required only when sampling is done with replacement (as in Section 3). When sampling is done without replacement (the default) and the number sampled is the same as the population size, the result is a permutation of the population. For example, there are 3! = 6 permutations of the three numbers in the vector (1, 2, 3), and each time we use the command `sample(1:3, 3)` we get—at random—one of these six permutations. We illustrate with four iterations, all of which happened to give different results:

```
> sample(1:3, 3)
[1] 3 1 2

> sample(1:3, 3)
[1] 3 2 1

> sample(1:3, 3)
[1] 1 2 3

> sample(1:3, 3)
[1] 1 3 2
```

In Section 2 we consider three different data types in a block design, and some of the various metrics that can be used to measure the effect in question. We also illustrate permutation tests for these situations along with results of some more familiar normal theory and nonparametric tests. In Section 3 we consider the same issues, restricting the discussion to paired data and discussing how to think about outliers. In Section 4 we look at permutation tests for two-sample data.

2. Block Designs: Judging a Food Product

Suppose we wish to compare $g = 3$ brands of chocolate pudding (individually packaged single portions, ready to eat). Each of $n = 6$ randomly chosen professional judges independently tastes all three recipes (data in section 2 are from "Taster's Choice" column of San Francisco Chronicle, Sept. 1994).

2.1. Scenario 1: Each judge picks a favorite

Each judge assigns the code 1 to his or her favorite brand, and 0 to the other two. The matrix of results is shown below, where each entry is either 0 or 1 and each column has exactly one 1 and where we denote the sum of the i^{th} row as X_i, for $i = 1, 2, 3$.

Brand	Taster:	1	2	3	4	5	6	X = Obs. pref. count
A		0	1	1	1	1	1	5
B		1	0	0	0	0	0	1
C		0	0	0	0	0	0	0

Our null hypothesis is that the three brands are equally appealing. That is, each brand has probability $p_i = p = 1/3$ of being chosen as favorite by any one of the judges. Then, by multinomial theory, the expected number of judges preferring each brand is $E = np = 2$. Are our observed results consistent with this null hypothesis?

We need a numerical scale or metric to express how well the data agree with our null hypothesis, and then a way to decide whether agreement is good or poor. The traditional metric used here is Pearson's chi-squared statistic $Q = \Sigma_i (X_i - E)^2/E$. According to this statistic, $Q = 0$ indicates perfect agreement with the null hypothesis and a sufficiently large value of Q indicates such poor agreement that the null hypothesis should be rejected. For our data, $Q = (9 + 1 + 4)/2 = 7$. As a nonparametric test, this is equivalent to *Cochran's test* which is a special case of the Friedman test for randomized block designs when the responses are binary.

Asymptotically, for sufficiently large n, the distribution of Q is well approximated by the chi-squared distribution with $\nu = 2$ degrees of freedom, CHISQ(2). This approximation gives a P-value of 0.03. That is, only 3% of the probability in this distribution lies at or above $Q = 7$. Based on this we might be tempted to reject the null hypothesis at the 5% level. Unfortunately, the true distribution of Q agrees rather poorly with CHISQ(2) for only $n = 6$ judges. This does not mean that the statistic Q is a bad way to measure agreement between the observed and expected preference counts. It does mean that we cannot use CHISQ(2) to make a valid decision based on our observed value of Q.

In this simple case, it is possible (although a little tedious) to use elementary combinatorial methods to find the exact distribution of Q for our problem. It turns out that the possible values of Q are 0, 1, 3, 4, 7, and 12, with $P\{Q = 7\} = 0.0494$ and $P\{Q = 12\} = 0.0041$. [For example, there are $3^6 = 729$ possible data matrices of which only 3 give $Q = 12$, the ones with sums 6, 0, 0 in some order, so $P\{Q = 12\} = 3/729 = 0.0041$.] Thus the exact P-value for our data is $P\{Q \geq 7\} = 0.0494 + 0.0041 = 0.0535$, and we cannot quite reject the null hypothesis at the 5% level.

However, in only slightly different problems it can be very tedious, extremely difficult, or practically impossible to find the exact distribution of the test statistic Q. A modern solution to this difficulty is to simulate at random a very large number of data matrices from among the 729 possible ones, find the value of Q for each, and use the results to approximate the distribution of Q. The first program in Appendix B does this simulation for our data [Figure 1(a)]. Note here that an "exact" P-value is the probability that the statistic is as extreme as the observed value using the true distribution of the statistic given that the null hypothesis is true. An "approximate" P-value is the probability that the statistic is as extreme as the observed value using a distribution that is approximating the true distribution of the statistic under the null hypothesis.

In this program we denote the observed $g \times n$ data matrix as `OBS`. A permutation of this matrix is made by permuting the 0's and 1's of the rows within each of the $n = 6$ columns in turn, essentially determining at random for each judge (column of the matrix) which brand is his or her favorite (which of the $g = 3$ possible positions has code 1). The result is the permuted matrix `PRM`. The inner loop of the program does this permutation. The brackets `[,j]` indicate the j^{th} column of `PRM`. For example, when the first column is permuted the population of the sample function is (0, 1, 0) and the permuted first column is equally likely to become (1, 0, 0), (0, 1, 0), or (0, 0, 1).

```
PRM = OBS
for (j in 1:n)
        {
        PRM[,j] = sample(PRM[,j], 3)
        }
```

The outer loop of the program generates 10,000 permutations in this way and computes the chi-squared statistic Q_{prm} (`q.prm`) for each. Finally, we have a sample of 10,000 simulated values of Q_{prm} from the permutation distribution, with which to compare the observed value Q_{obs} (`q.obs`) = 7. These results are plotted in the histogram of Figure 1(a), where the curve is the (obviously ill-fitting) density of the distribution CHISQ(2) and the small dots show exact probabilities. The results can also be tallied as shown below. Of the 10,000 simulated values Q_{prm} about 5.6% are 7 or greater, so the (approximate) P-value of the permutation test is 0.056, which agrees well with the exact P-value 0.054. By using more than 10,000 iterations one could get an approximation arbitrarily close to the exact distribution.

```
           x          0       1        3        4        7       12
Exact P{Q=x}      .1235   .4938    .2058    .1235    .0494    .0041
Simulated         .1257   .4944    .2047    .1188    .0523    .0041
```

The first line of the program sets the seed for the pseudorandom number generator. If you use the same seed and software we did, you should get exactly the results shown here. If you omit the first line (or "comment it out" by putting the symbol # at the beginning of the line), you will get your own simulation, which will be slightly different from ours. By making several simulation runs (without specifying a seed), each with $m = 10{,}000$ iterations, you will see that it is reasonable to expect approximations somewhat larger than 0.05 for this P-value. Five additional runs yielded P-values between 0.052 and 0.057. More precisely, the 95% margin of simulation error for $P\{Q \geq 7\}$ can be computed (using the normal distribution for the 95% confidence and the binomial (10000, 0.054) standard error of the simulation) as approximately as $1.96[(.054)(.946)/10{,}000]^{1/2} = 0.004$.

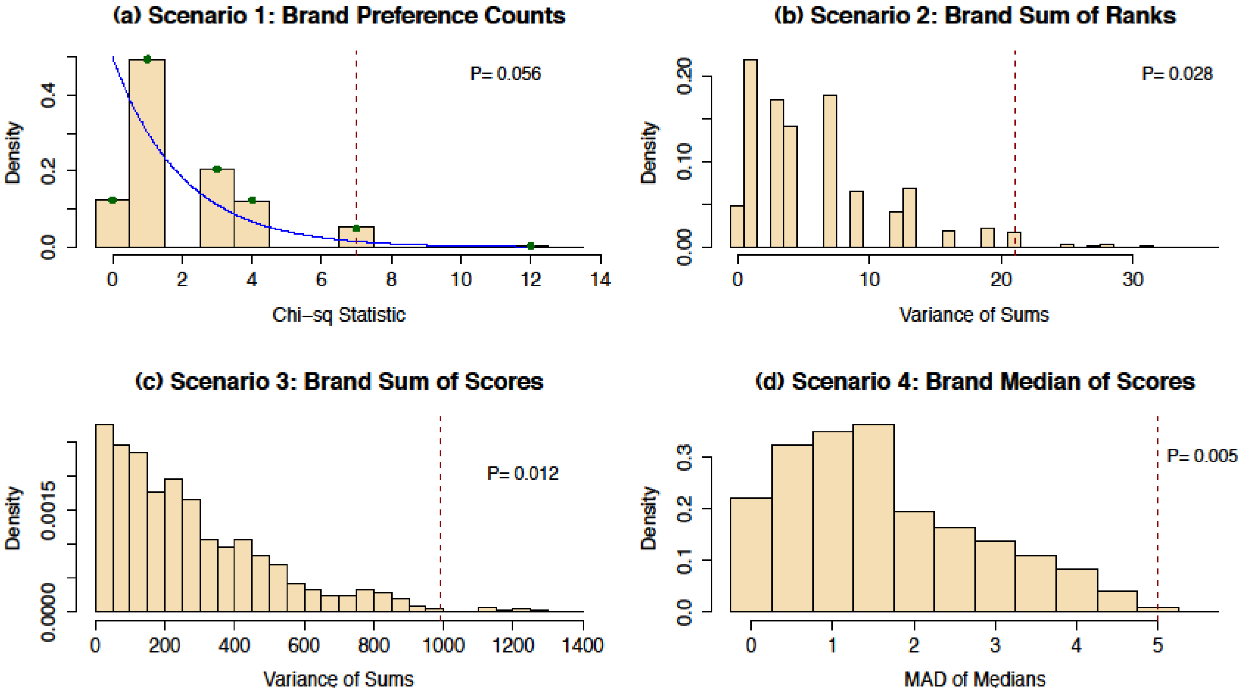

Figure 1. Four simulated permutation distributions for taste tests in Sec. 1 and Sec. 2: different data types and metrics.

2.2. Scenario 2. Each judge ranks the brands

Next, suppose we require the judges to provide a little more information. Specifically, each judge ranks the three recipes: 1 for least liked, 2 for next best, 3 for best. In this case the entries in the data matrix must be 1, 2, and 3 (in some order) down each column, and again we denote the row totals as X_i. Now the data matrix might be as follows:

Recipe	Taster:	1	2	3	4	5	6	X = Obs. sum of ranks
A		1	3	3	3	3	3	16
B		3	2	2	2	2	2	13
C		2	1	1	1	1	1	7

Under the null hypothesis that the three brands are equally preferable, each judge has an equal chance of making one of the 3! = 6 possible assignments of the ranks 1, 2, and 3 to the three brands. So, on average, the X_i ought to be equal, each with expected value 12. The next issue is to decide what metric to use to measure how similar or how different the X_is are from one another. One reasonable metric is the variance, V, of the X_i. The observed variance (of the sums 16, 13 and 7) is $V = 21$. If all tasters have exactly the same opinions, we would have $V = 0$; we will reject the null hypothesis for large values of V.

Is $V = 21$ a sufficiently large value that we should reject the null hypothesis? Combinatorial methods for finding the exact distribution of the variance seem out of reach. However, if we simulate the permutation distribution of the variance in much the same way as we did before, we can compare it to the observed variance and find a useful P-value. The required program is a very slight modification of the one used in Scenario 1 (see Appendix B for code for Figure 1 (b)). The numerical result is that about 3% of the 10,000 simulated variances equal or exceed 21, so the P-value is about 0.0304. In Figure 1(b) this corresponds to the area under the histogram and to the right of the vertical line. Thus the increased information provided by the judges in this scenario is enough to allow us to reject, at the 5% level, the null hypothesis that all brands are equally preferred. An equivalent metric is $U = \Sigma X_i^2$ because $V = (U - 36^2/3)/2 = (U - 432)/2 = 21$. For the dimensions of this design, the quantity subtracted in the numerator is always 432. Thus there is a monotone relationship between U and V, and a permutation test using either metric would give exactly the same P-value.

We mention that this situation is essentially the *Friedman test* for data in a block design with $g = 3$ groups and $n = 6$ blocks, where data within each block are ranked separately. (Two suitably comprehensive references are Sprent (1993) and Higgins (2004).) As here, the data in each block may already be ranks or, as in Scenario 3, they may be scores that can be ranked. For moderately large n, the Friedman test statistic (which is also a monotone function of U, and hence of our metric V) is distributed approximately CHISQ(2), so the P-value is about 0.0302. [With `OBS` specified as in Appendix B, the Friedman test can be performed using the R statement `friedman.test(t(OBS))`, where `t` indicates the transpose of our data matrix.] Texts on nonparametric statistics, such as Hollander and Wolfe (1999) provide tables with exact P-values for some nonparametric tests, for these data the Friedman test statistic is $S = 7$ and the P-value found in their Table A.15 is 0.029. A permutation test using any reasonable metric avoids the necessity of making tables for small designs or finding asymptotic approximations for larger ones. Also, permutation tests handle ties (up to some breakdown point where there is too little information) with no need for adjustments.

2.3. Scenario 3. Each judge scores each brand

Finally, assume that the each judge is asked to make notes on several aspects of each brand. Perhaps judges are asked to rate whether the pudding tastes good, whether it looks appealing, whether it feels smooth in the mouth, and whether it has an authentic chocolate taste. Points are given for each aspect and added—with a maximum possible score of 20 for each brand by each judge. Now the data matrix might look like this:

Recipe	Taster:	1	2	3	4	5	6	X = Obs. sum of scores
A		7	15	20	16	15	18	91
B		16	6	4	10	13	11	60
C		8	0	0	9	5	6	28

If we believe the judges follow the scoring rules carefully and that the scale is defined appropriately, we may be willing to take the judges' total scores as numerical values on an interval scale. Also, it is possible that these scores might be nearly normally distributed because each score is the sum of several sub-scales reflecting various aspect of pudding quality. Outliers seem unlikely because each score is constrained to lie between 0 and 20. Assuming normality, the data can be analyzed according to a two-way ANOVA complete block design with $g = 3$ brands or recipes (levels of a fixed effect) and $n = 6$ tasters (considered as random blocks) and no interaction effect. Then, by a standard ANOVA computation, the observed value of the test statistic is $F = 165.39/21.32 = 7.76$. Under the null hypothesis that all brands have equal population means, this statistic is distributed as F(2, 10), so the P-value is 0.00925.

However, because we are asking the judges to provide numerical scores for subjective judgments there may be some doubt whether these data should be treated as strictly numerical, and thus doubt about using the *F*-statistic as a metric. For example, it seems difficult to imagine that the observed extreme scores 0 and 20 are warranted. Do these judges really feel they taste, respectively, very nearly the worst and very nearly the best possible chocolate pudding that can ever be made? Additionally, these are subjective ranks; one judge's 15 may be another judge's 17.

If we choose to use ranks instead of numerical scores, we can rank the scores for each judge and use Friedman's test as described above (P-value 0.0302). An alternative rank-based procedure would be to rank all 18 scores (assigning rank 1.5 to each of the scores tied at 0), and perform an ANOVA on the ranks. This is called a *rank-transform test*. The resulting P-value is 0.0115. Clearly the ranks are not normal, so the *F* statistic does not have exactly an F distribution and this P-value must be viewed as only a rough indication of the degree of significance.

Now we consider two permutation tests. In both cases the null hypothesis is that the three brands are of equal quality, and in both cases we permute separately the scores for each judge, but we use different metrics to look for possible differences among brands. First, as in Scenario 2, we sum the scores and use the variance of the sums as the test statistic. Except for the statement that provides the data, the program is the same as in Scenario 2. The P-value is 0.012, and the histogram of the simulated permutation distribution is shown in Figure 1(c). We might use this metric if we regard the scores as truly numerical.

Second, we find the median score for each of the three brands and then use as the test statistic the MAD (here the median of the absolute deviations from the middle one of the three brand medians). The modified program is shown on the website. Our simulation distribution had only eleven distinct values of the MAD, of which the observed value 5, the largest of these eleven, occurred for only 40 of the 10,000 permutations. Therefore, the P-value is about 0.004. (Five additional runs gave values between 0.0038 and 0.0046.) The relevant histogram is shown in Figure 1(d) (see JSE website for programs of Figures 1(c) and 1(d)). The MAD is a reasonable metric if we believe the data follow an ordinal scale so that higher scores reflect better quality, even if not in a strictly numerical sense. In general, this metric may be more satisfactory for larger values of *g* and *n* where there would be more distinct values of the MAD.

2.4. Summary of three scenarios

Table 1 below summarizes our tests on the pudding data, with the four simulated permutation tests shown in **bold** font. The specific type of permutation test depends on the kind of data (indication of one favorite brand, ranking of three brands, or numerical scores for three brands), how the data are summarized across judges (counting, adding, or taking the median), and how these brand summaries become a test statistic (chi-squared GOF statistic, variance, or MAD).

Table 1. Results from the pudding data with simulated permutation tests in **bold** font.

Scenario and Data Type	Test and Metric	P-Value
1. Count Favorites per Brand	Goodness-of-Fit test chi.sq. approx. (ill fitting)	0.030
1. Count Favorites per Brand	GOF test: Exact distribution of chi-sq. statistic	0.0535
1. Count Favorites per Brand	**Sim. Perm. test using chi-sq. GOF statistic**	**0.056**
2. Sum of Ranks per Brand	**Sim. Permutation test using variance of sums**	**0.030**
2. Sum of Judge-Ranks per Brand	Friedman test (approx. chi-sq. distribution)	0.0302
3. Sum of Scores per Brand	ANOVA, F-test (assumes normal data)	0.00925
3. Sum of Global-Ranks per Brand	Rank transf. ANOVA, approx. F-test	0.0125
3. Sum of Scores per Brand	**Sim. Permutation using variance of sums**	**0.012**
3. Median of Scores per Brand	**Sim. Permutation using MAD of medians**	**0.004**

The tests from Scenario 3, where we have numerical scores, have smaller P-values and hence seem to provide “stronger“ evidence that not all brands are alike. The evidence is *usefully* stronger to the extent that judges’ scores accurately express subjective impressions as numerical scores. The permutation test using the MAD of group medians happens to have the smallest P-value; it mutes somewhat the discord caused by Judge 1’s preference for Brand B and Judge 3’s strong preference for Brand A.

It would not be correct to conclude that one of these tests is more powerful than another just because it happens to have a smaller P-value for the pudding data. Determination of power requires an exact specification of an alternative (non-null) distribution. For the standard normal-theory tests (such as *F*-tests and *T*-tests) applied to normal data, the distribution theory for power is known and the power for various alternative parameter values can be computed exactly. For many nonparametric and permutation tests, simulation is used to find approximate power. We revisit the issue of power briefly in Section 3.

3. Paired Designs: Asthma Patients Breathing Clean and Contaminated Air

In this section we consider the special case of a block design in which the number of groups is $g = 2$. In this case, a complete block design is usually called a *paired design*. If data collected according to such a design are normal, the appropriate normal-theory test of a difference between the two group population means is the paired T-test. In this relatively simple setting, we discuss the effects of nonnormal data.

Increase in airway constriction during exercise was measured in each of $n = 19$ asthma patients, in pure air and also in air contaminated with 0.25ppm sulfur dioxide (SO_2). The purpose of the study is to see whether SO_2 increased the specific airway resistance (SAR). Results are shown below (Bethel, Haye, Oberzanek, George, Jimerson and Ebert,. 1989). (Link to AsthmaData.txt on the website. The data can also be found in Pagano and Gauvreau, 2000.)

Subj	1	2	3	4	5	6	7	8	9	10
Air	.82	.86	1.86	1.64	12.57	1.56	1.28	1.08	4.29	1.34
SO$_2$	.72	1.05	1.40	2.30	13.49	.62	2.41	2.32	8.19	6.33
Dif	0.10	-0.19	0.46	-0.66	-0.92	0.94	-1.13	-1.24	-3.90	4.99

Subj	11	12	13	14	15	16	17	18	19
Air	14.68	3.64	3.89	.58	9.50	.93	.49	31.04	1.66
SO$_2$	19.88	8.87	9.25	6.59	2.17	9.93	13.44	16.25	19.89
Dif	-5.20	-5.23	-5.36	-6.01	7.33	-9.00	-12.95	14.79	-18.23

Data under both conditions are strongly right-skewed with extreme outliers in the right tails. In this paired design it is sufficient to look at differences d_i, $i = 1, \ldots, 19$, obtained by subtracting the measurement with SO_2 from the measurement in pure air. Thus, the 14 negative values of d_i out of 19 indicate bad effects from the air contaminated with SO_2.

Because we cannot imagine that SO_2 contamination would be beneficial to asthma patients, we perform a left-tailed test. The paired T-test has $T = –1.687$. If the d_i were normal with 0 mean, this test statistic would have the t distribution with $\nu = 18$ degrees of freedom, and so the P-value would be 0.054, indicating that the null hypothesis of equal group means cannot quite be rejected at the 5% level. This is a counterintuitive result because of the great preponderance of subjects who suffered worse airway constriction with $S0_2$.

3.1 Traditional nonparametric tests

Briefly, here are results of three nonparametric tests that are commonly used for paired data. All of them lead to the conclusion (at the 5% level) that SO_2 is associated with increased airway constriction.

- *The sign test.* The null hypothesis is that positive and negative values of d_i are equally likely, so the number B of positive d_i has $B \sim \text{BINOM}(19, .5)$. The P-value of this test is $P\{B \leq 5\} =$

0.0318, and we reject the null hypothesis that SO_2 has no effect in favor of the alternative that it does harm. It is worthwhile observing that one could not perform such a sign test exactly at the 5% level: the obtainable P-values adjacent to 5% are at 0.0835 and 0.0318. The sign test is a simple example of a permutation test. Imagine tossing a fair coin to determine whether each d_i is positive or negative. Because we know the exact null distribution of B, using simulation to approximate this binomial distribution serves no practical purpose.

- *The Wilcoxon signed-rank test* of the null hypothesis that median of the paired differences is 0 against the alternative that the median is less than 0. Under the null hypothesis of this test, the d_is have a continuous distribution symmetrically distributed about 0. The P-value of this test, based on a reasonable normal approximation, is 0.018. [With the obvious specifications of `Air` and `SO2`, one can obtain this result using the R code `wilcox.test(Air, SO2, alt="less", pair=T)`. In practice, the data may be arranged in increasing order of $|d_i|$ to facilitate hand computation of the Wilcoxon statistic.]

- *The rank-transform test.* This approximate test involves ranking all 38 observations and performing a paired T-test on the differences in ranks. Pretending that, under the null hypothesis, the resulting t-statistic follows the t distribution with $\nu = 18$, we obtain the P-value 0.011.

3.2. Permutation tests using simulation. A permutation test for paired data is based on the null hypothesis that either of the two values observed within a pair could equally likely have been observed for the first group. In terms of differences, this amounts to saying that the sign of any of the differences d_i might equally well have been positive or negative. Accordingly, the R code to make each permutation samples $n = 19$ values at random *with* replacement from the values –1 and 1, and (element wise) multiplies the resulting vector of 19 "signs" by the vector of the 19 observed d_i (denoted `Dif` in the program). Various metrics could be used to summarize the sign-permuted differences (`perm`). We consider the T-statistic, mean, median, and 10% trimmed mean.

Specifically, when the metric is the *mean*, the inner loop of the program, shown in full in Appendix B (Figure 2(a)), is as follows:

```
for (i in 1:m)
    {
    perm = sample(c(-1,1), n, repl=T) * Dif
    mean.perm[i] = mean(perm)
    }
```

Note: In computing the 10% trimmed mean, the largest and smallest 10% of the n observations are deleted before the mean is taken. The number $0.1n$ is rounded down to the nearest integer, so for $n = 19$ the mean of the middle 17 observations is taken. The 50% trimmed mean is interpreted to be the median, a fact we use to simplify the first program in Appendix B.

Figure 2 shows histograms of the simulated permutation distributions of the mean and the three other metrics mentioned above. Simulated P-values are shown in Table 2 below.

Table 2: P-values from traditional tests and simulated P-values from the metrics of Figure 2

Test and Metric	**P-value**
T-test	0.054
Sign test	0.0318
Wilcoxon signed rank test	0.018
Rank transform test (global ranks, differences of pairs)	0.011
Sim. Permutation: *T*-stat. of sign-permuted d_i	0.054
Sim. Permutation: Mean of sign-permuted d_i	0.054
Sim. Permutation: Median of sign-permuted d_i	0.017
Sim, Perm.: 10% Trimmed mean of sign-permuted d_i	0.038

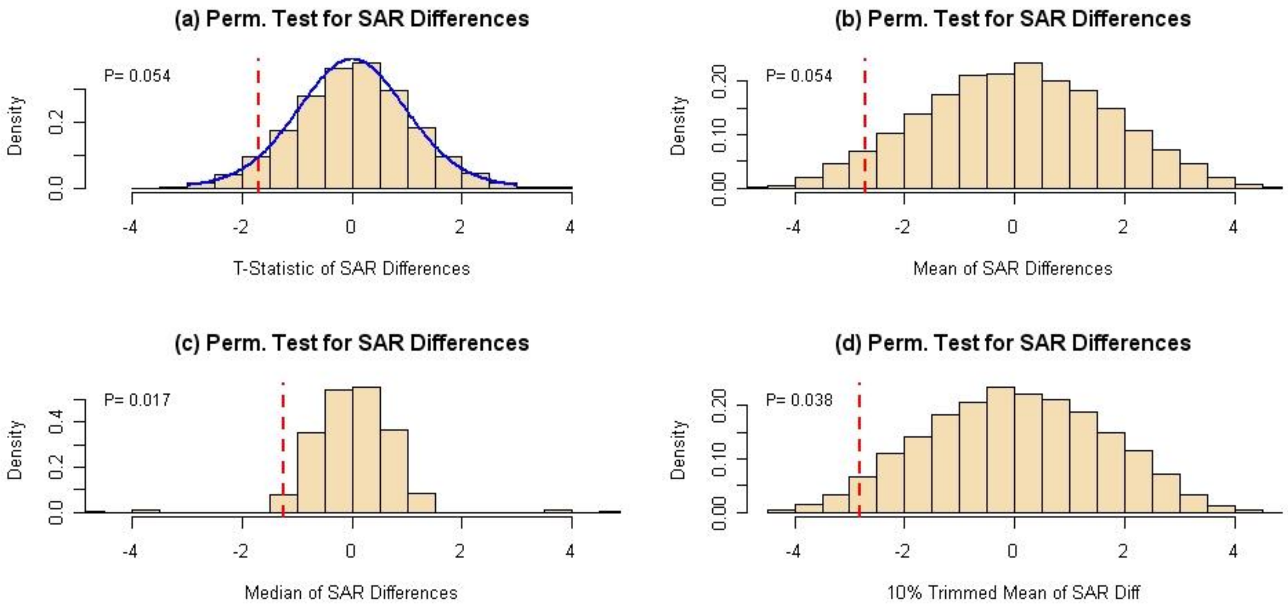

Figure 2. Histograms of simulated permutation distributions for the paired SAR data of Section 3. Vertical dotted lines show the observed value for the metric used. The density curve of T(18) is shown in panel (a).

3.3. Which test is “best”?

The answer lies in the meaning of the measurements. In particular, we have to understand the practical importance of the largest and smallest values among the original SAR measurements. For example, at one extreme the largest SAR values might indicate frightening or life-threatening episodes for the asthma patients, and at the other extreme they may be transient glitches in the equipment used to make the measurements. Looking at the smallest SAR values, we may wonder whether either of the two SAR values for Patient 1 represents clinically significant restriction. Also, we may ask whether any of the eight patients with $|d_i| < 2$ actually notices a difference in his or her breathing, or whether such a difference may be of clinical importance even if not noticed by the patient. These are questions to be asked by an alert statistical consultant and answered by experts in pulmonary medicine. (In a totally different context, if the measurements were monthly costs for two comparably sized branches of a chain store, then one would almost certainly not choose a metric that suppresses the influence of large outliers.) The goal is to match the metric to the meaning of the data.

An advantage of simulated permutation tests is that they allow a choice of metric without requiring theoretical derivation of the resulting null distribution. For example, one may wonder whether the traditional paired *T*-test fails to reject because the data are too far from normal for the *T*-statistic to have anywhere near a t distribution. However, when we use the *T*-statistic as the metric in a permutation test, understanding that thereby we are attaching considerable importance to outliers, there is no doubt that the P-value accurately reflects the evidence as measured by the *T*-statistic. For our data this P-value is nearly equal to that of the traditional *T*-test, so this is an instance that illustrates the legendary robustness of the traditional *T*-test.

It turns out that the *T*-statistic and the mean are equivalent metrics for our permutation tests for paired data. (The simulated P-values are exactly the same because we used the same seed and so sampled the same collection of permutations for both.) Algebraically, it is not difficult to show this is a consequence of the invariance of $\Sigma_i\, d_i^2$ under permutation of the signs of the d_i, so that the value of the *T*-statistic for permutations of the signs of the d_i is a monotone function of the mean $\bar{d}$ (or of the sum $\Sigma_i\, d_i$). Graphically, this is illustrated in panel (a) of Figure 3 where each dot represents a different one of the 10,000 permutations (code in Appendix B). By contrast, panels (b) and (c) indicate that the mean, median, and trimmed mean are three fundamentally different metrics. See the program located on the website.

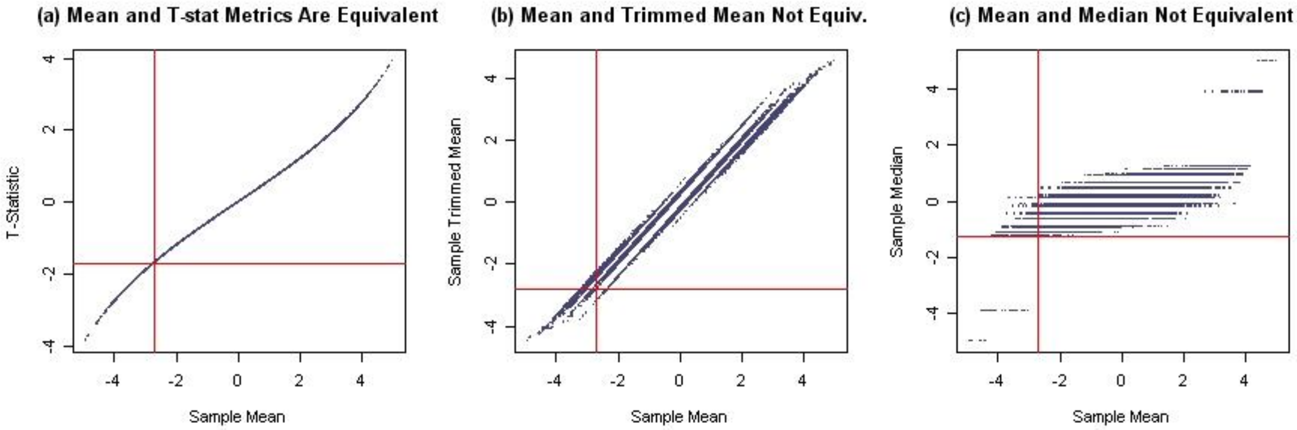

Figure 3. For permutation tests on the paired data of Section 3, the mean, median, and trimmed mean are fundamentally different metrics, but the mean and the *T*-statistic yield the same P-value. Lines show observed values. Each point results from a random permutation of the 19 observed SAR differences. Red lines indicate the rejection regions; in the lower left hand corner both tests reject the null.

3.4. Rejection probabilities for multi-test decision procedures

With so many tests available, a naive (or unprincipled) "statistician" may be tempted to "shop around" until he or she finds a test that rejects the null hypothesis and then report only the result of that one test. This is misleading because the *overall* significance level for such a shopping spree is larger than the significance level for any one of the tests tried. In general, simulation is required to find the true significance level of such a multi-test procedure. We illustrate this by simulating the rejection probability for a combination of a sign test and a *T*-test in the case where we have 13 observations from a normal population and we want to test the null hypothesis that the data are centered at $\mu = 0$ against the alternative that the mean is negative. The R code for the results shown in Figure 4(a) and Figure 4(d) is in Appendix B.

- *Significance level.* We simulated 100,000 normal samples of size $n = 13$ from the standard normal distribution NORM(0, 1). (Here, we are using 100,000 samples for higher accuracy.) If we reject when the *T*-statistic is –1.7823 or smaller, then the significance level (probability of falsely rejecting the null hypothesis) is 5%. For the sign test, if we reject when the number of positive results B is 3 or smaller then the significance level is 0.0461 = 4.61%. (We choose $n = 13$ because this sample size provides a one-sided test with significance level near 5%.) Simulation shows that the probability of rejecting with at least one of these two tests is P(Reject H_0 with either test $| \mu = 0) \approx 0.07$. Thus the two-test decision procedure has level 7%, not the intended 5%. The top two panels of Figure 4 plot summary values of each of the first 5000 samples of these 100,000. Even though each test has significance level around 5%, these graphs show how differently the tests can treat the data from individual samples in rejecting, or failing to reject, the null hypothesis.

- *Power.* The power (probability of correctly rejecting the null hypothesis) is a function of the alternative value of μ. We illustrate for the specific value $\mu = –1/2$. The simulation procedure is the same as for the significance level, except now the samples are drawn from NORM(–1/2, 1), so that the population is centered at $\mu = -1/2$ and the probability is 0.3085 that any one observation is positive. Also, against this alternative, the power of the T-test is 0.52 (from a computation in R using the non-central t distribution: `se = 1/sqrt(13); pt(-1.7823, 12, ncp=-.5/se)` and the power of the sign test is 0.39 (`pbinom(3, 13, 0.3085)`). However, P(Reject H_0 with either test $| \mu = -1/2$) = 0.56, which is only a little larger than for the T-test alone. The bottom two panels of Figure 4 illustrate this situation. By selecting different values of μ one could simulate the power of this two-test decision procedure against other alternatives.

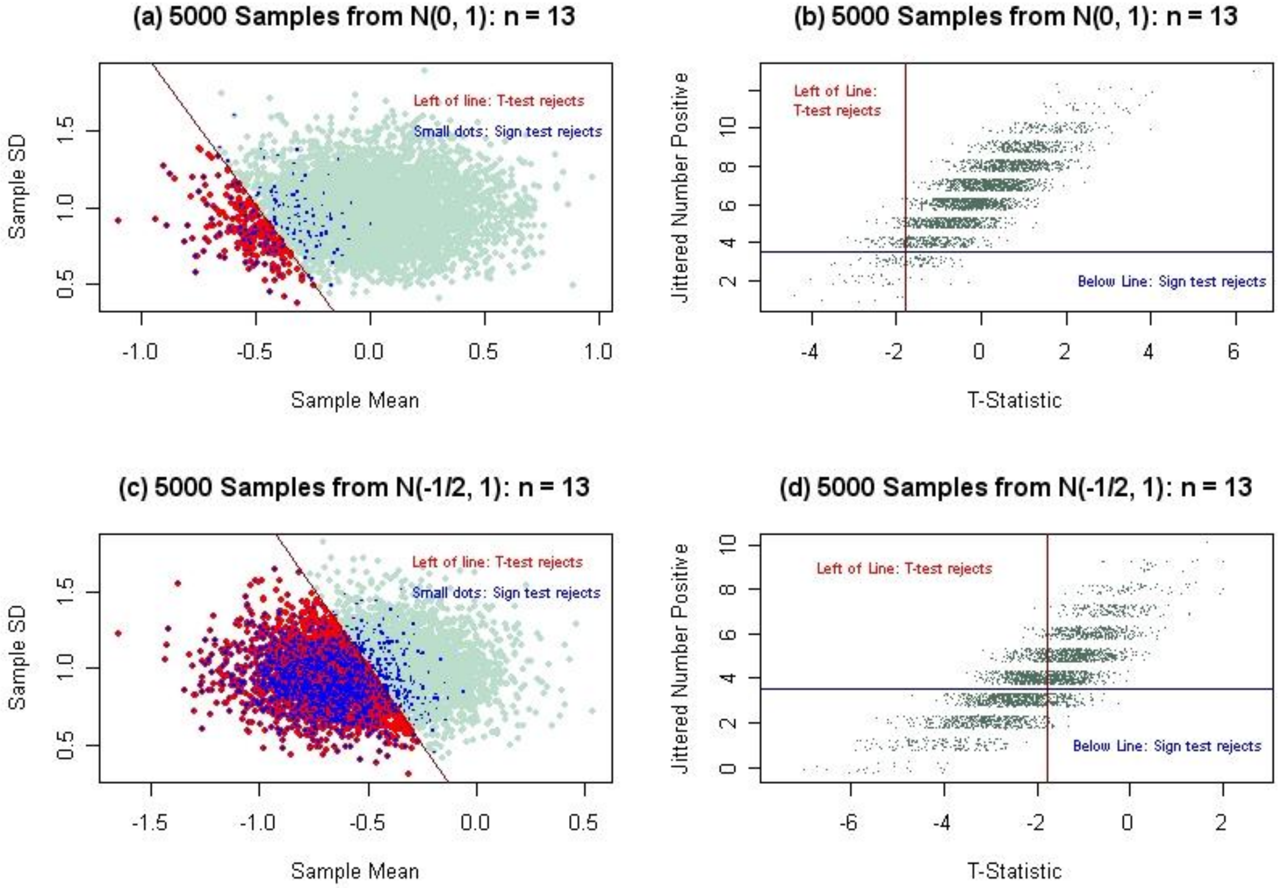

Figure 4. Illustration of significance level (top panels) and power (bottom panels) of a *T*-test and a sign test performed on the same data. Each dot represents a different simulated normal sample of size 13. (See section 3.4.)

4. Two-Sample Designs: Corn Yields and the *Challenger* Disaster

Now we turn to two simple examples in which the purpose is to compare two independent samples, of sizes n_1 and n_2, respectively, which are chosen from two possibly different populations. If we want to determine whether the means μ_1 and μ_2 of two normal populations are equal, then traditional tests are the pooled *T*-test (if the population variances are thought to be nearly equal) and the Welch separate-variances *T*-test (if we do not assume equal variances). We note here that since the advent of computer packages the Welch test is often used without checking the equality of variance. However, beginning statistics texts still often use the pooled *T*-test and hence, we include both tests here.

For two continuous populations of the same shape, but possibly differing as to their location (implying different medians η_1 and η_2 as well as different means), the Mann-Whitney-Wilcoxon (MWW) rank sum test is often used. Nonparametric tests are not free of assumptions. For the MWW test the same-shape assumption implies equal dispersion (whether measured in terms of variance or otherwise). The continuity assumption implies that ties occur only because continuous data must be rounded to some extent, and the ideal is that ties are rare or nonexistent. There are many other nonparametric tests for two-sample designs, but we do not discuss them here.

4.1. Small samples: Corn yield data

When samples are very small, it is not feasible to judge whether the data come from a normal population. In an experiment to study the effect of certain kind of weed on corn yield, data on yields were obtained from four plots that were free of weeds and from four plots with heavy coincident weed growth. Data as reported in a text by Moore and McCabe (1999) are as shown below. (The one very low yield in the second sample was noted and verified as correct.)

```
Weed Free:   166.7,   172.2,   165.0,   176.9

Many Weeds:  162.8,   142.4,   162.8,   162.4
```

We consider two-sided tests. The pooled *T*-test has $T = 2.19$, so if the assumptions are met, this statistic would have the t distribution with $\nu = 6$ degrees of freedom, and P-value 0.071 against the two-sided alternative. In this balanced case (where $n_1 = n_2$), the separate-variances test must also have $T = 2.19$, but now $\nu = 4.6$ and the P-value would be 0.085. The MWW test has P-value 0.0286 as explained below.

A permutation test is based on the 8! permutations of the $n_1 + n_2 = 8$ observations, in which the first four observations are taken as coming from the first sample and the last four from the second sample. Then according to some metric we summarize the information in the two permuted samples and compare them. The test statistic might be the difference in the two means. Then all that is required is to notice that all of the yields in the first sample are larger than any of those in the second. There are $C(8, 4) = 8! / (4!)^2 = 70$ possible combinations of four observations from among eight, of which one of the two most extreme combinations has occurred—with the four smallest values in the many-weed group. Another equally extreme combination would have been to see the four largest values in this group. Thus our two-sided P-value is 2/70 = 0.0286. (In this case where a most-extreme combination occurs, the P-value of our permutation test is the same as for the MWW test.)

The inner loop of the program to simulate this permutation test is as follows. See the R code for Figure 5(a) in Appendix B.

```
for (i in 1:m)
{
perm = sample(x, n1+n2)
dfmn.prm[i] = mean(perm[1:n1])- mean(perm[(n1+1):(n1+n2)])
}
```

The simulated P-value from one run is 0.0288. The margin of simulation error based on 100,000 iterations is about 0.001, so this result agrees quite well with the exact P-value obtained above by combinatorial methods. As was the case with the paired *T*-test, the balanced case, the pooled *T*-statistic is an equivalent metric to the difference in means. That is, either metric will produce exactly the same simulated P-value for the same set of 100,000 permutations.

However, for our data the exact P-value 2/70 differs substantially from the P-value of either *T*-test. In panel (a) of Figure 5 we plot the histogram of the simulated permutation distribution of the *T*-statistic to show how poorly it agrees with density curve of the t distribution with $\nu = 6$ degrees of freedom. (This value of ν is for the pooled test; the separate-variances test potentially has a different value of ν for each permutation, so it would be difficult to know which t distribution to show.) This is a case in which the *T*-test works very poorly—because of the small sample size and the outlier. The lumpy appearance of the histogram mainly reflects the granularity of the permutation distribution. Because of the small sample sizes this distribution takes only about 50 distinct values. However, even when sample sizes are larger, outliers can cause a multimodal permutation distribution, as the outliers fall randomly into either the first permuted sample or the second.

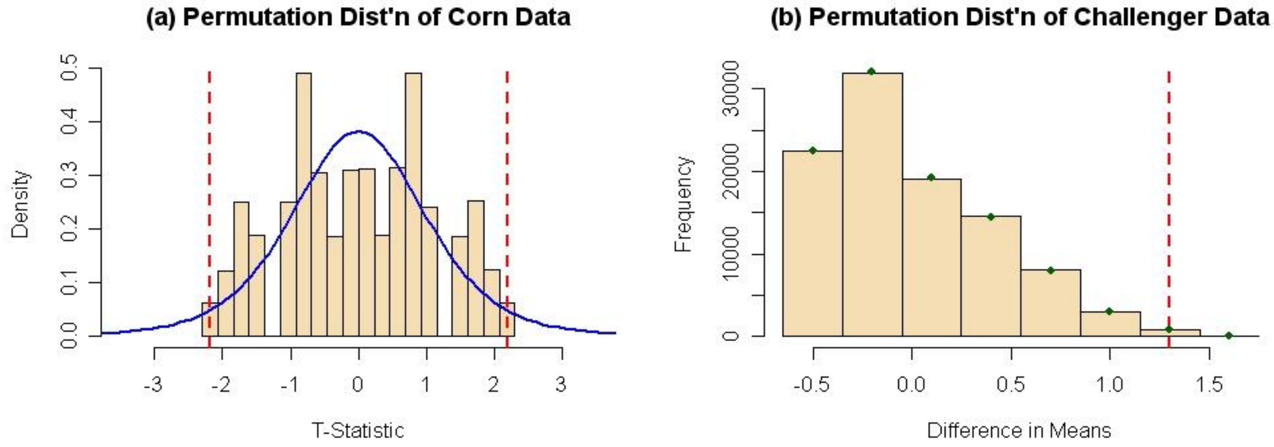

Figure 5. Histograms of simulated permutation distributions for two-sample tests in Section 4. (a) For the corn data, the permutation distribution does not match the t density with $\nu = 6$ (superimposed curve). (b) For the *Challenger* data, the simulated permutation distribution matches the exact permutation distribution (dots) very well.

4.2. Samples with many ties: Challenger data

In 1986 the space shuttle *Challenger* exploded after being launched at a temperature of 29°F, the lowest temperature at which any shuttle had been launched. A subsequent investigation found that the explosion was caused by fuel leakage. Historically, non-catastrophic fuel seal problems due to the failure of O-rings had been noted. Most of these problems had occurred at low temperatures. The text by Ramsey and Shafer (2002) reports the following data on number of O-ring failures on each of 24 earlier launches, broken out by launch temperatures below 65°F (which we call Cold) and above 65°F (Warm) from data shown in Feynman (1988).

Cold	1, 1, 1, 3
Warm	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2

We want to see if the number of O-ring failures is negatively associated with temperature. Here we use a one-sided test because the issue of O-ring failures at colder temperatures had apparently been discussed and dismissed as inconclusive before *Challenger* was launched. The MWW test can be adjusted for ties by assigning an average rank to those values that tie, and finding the P-value through permutation of the ranks works well. However, the normal approximation used for the MWW test no longer applies without adjustments to the formulas for expected value and variance of the MWW statistic. Consequently, we discuss an alternative methodology here. Although there are no outliers here, the data do not seem to fit a normal distribution at all. Nevertheless, we note that the statistic for the pooled T-test is $T = 3.8876$, giving a P-value of 0.0004 according to the t distribution with $\nu = 22$. (The Welch separate-variances test has P-value 0.038.)

Now we illustrate the permutation test based on the difference of the two sample means, for which the observed value is 1.3. Because there are so many ties, it turns out that the permutation distribution takes only eight distinct values, whose probabilities can be found with some effort using elementary combinatorial methods [see Ramsey and Schafer (2002)]. Shown in the table below are these eight values, their exact probabilities from combinatorics, and their simulated probabilities based on a run of a program similar to the one used in Section 4.1. (The pooled T-statistic is an equivalent metric, but in this unbalanced case, with n_2 much larger than n_1, the Welch T-statistic is a quite different metric.)

Difference of Means	-0.05	-0.2	0.1	0.4	0.7	1	1.3	1.6
Exact Probability	0.224	0.320	0.192	0.144	0.080	0.030	0.009	0.001
Sim. Probability	0.225	0.320	0.191	0.145	0.081	0.029	0.009	0.001

Thus the P-value of the permutation test is 0.010. In retrospect, these data show a clear association between cold weather and an increased risk of fuel leakage. Panel (b) of Figure 5 (R code in the Appendix B) shows the histogram of the simulated frequencies and small dots show the corresponding expected frequencies based on the exact probabilities. The simulated P-value from this run of 100,000 iterations is 0.0095 ± 0.0006.

5. Conclusion

Most textbooks stress the use of nonparametric tests for data sets that are too small to adequately test the assumptions of parametric tests and for use with data sets not meeting the normality assumption of the *T*-test. A commonly used graphical method for assessing normality is to create a boxplot and look for outliers. However, this method is not very sensitive, nor very specific. Hoaglin, Iglewicz and Tukey (1986) show that approximately 30% of samples of size *n* = 19 sampled from normal distributions will have outliers as defined by the 1.5 IQR rule. We emphasize that nonparametric tests should also be used when the data are not inherently numeric in nature, for example when using subjective ratings of a judge or judges.

We present a variety of data from categorical (picking a favorite), ordinal (ranks and ratings), and continuous data (asthma example, corn data and Challenger data), including skewed data (asthma data) and a small data set (corn data), in order to illustrate the importance of choosing a test metric appropriate to the nature of the data. This is done in the setting of approximating P-values for permutation tests by using simulations programmed in R. Our examples can be presented easily in a classroom and provide motivation for a discussion of appropriate measures of central tendency and dispersion. Permutation tests provide valuable alternative tests when the assumptions of the standard parametric tests are not met, either because of outliers or skewness, or because the data are not truly numeric in nature. Our suggested code is flexible enough for the students to alter for the suggested exercises (and for different data sets) and thus provides interactive illustrations of fundamental statistical concepts.

Appendix A: Suggested Exercises for Students

For Section 2:

1. The scores for the first five judges in Scenario 3 are reported in Saekel (1994), but one of the authors, a self-proclaimed expert on chocolate pudding, served as the sixth judge. The scores in Scenarios 1 and 2 are derived from those given in Scenario 3. The addition of a sixth judge slightly simplified the exposition of Scenario 1. Analyze the original data for the first five judges: (a) according to Scenario 1, (b) according to Scenario 2, and (c) according to Scenario 3.

2. In Section 2.3, the ANOVA model permits an *F*-test test for a "Judge" effect. (a) Verify the *F*-test for differences among Brands, and then perform and interpret the *F*-test for variation among judges. (b) Repeat part (a) using the rank-transformed data. (c) Why is a test for a Judge effect not included in the procedure for the Friedman test on Brands?

For Section 3:

1. Ignore the first ten asthma patients as showing "too little difference to matter." If $5 < |d_i| < 10$, then rate the patient as ±1, where the sign matches that of d_i. Similarly, if $|d_i| > 10$, then rate the patient as ±2. (a) Do a sign test on the resulting data. Do you find a significant difference at the 5% level between the two groups? (b) Report results of a simulated permutation test on these ratings, using the mean rating as the metric.

2. Refer to Section 3.3. Complete the details of the algebraic proof that, within a simulated permutation test, the one-sample *T*-statistic is a monotone increasing function of the sample mean.

3. As in Section 3.4, use simulation find the power against the alternative $\mu = -3/4$ for both tests, separately and in combination. (Answer: For the *T*-test, power is 0.817.)

4. Derive the equation of the slanted line in panels (a) and (c) of Figure 4 from a statement that says when a *T*-test on a sample of size 13 rejects at the 5% level.

For Section 4:

1. Perform the MWW test for the corn yield data. With suitable specifications of data vectors `Free` and `Many`, use the following R code: `wilcox.test(Free, Many, exact=T).` In the data vector `Many`, change 162.8 to 162.7 and repeat.

2. Consider the permutation test on the *Challenger* data. (a) The largest value of the difference *D* between group means occurs when the permutation leads to values 1, 1, 2, and 3 in the Cold group. Show that the probability of this is 10 / 10,626 = 0.001. (b) The next largest value of this statistic occurs when the Cold group has values 1, 1, 1, and 3 (as actually observed). This same value also occurs when the Cold group has values 0, 1, 2, and 3. Show that P(*D* = 1.3) = (10 + 85)/10,626.

3. Run the program for Figure 5(a), using the option shown for the Welch *T*-statistic. How do you know this is not the same metric as the pooled *T*-statistic for the *Challenger* data?

Appendix B: R Programs

```
####################################   FIGURE 1(a)   ####################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                         #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 1(a) Simulated permutation test for GOF statistic                      #
# (Some embellishments for in graphs for publication omitted.)                           #
##########################################################################################

# Specify constants, input data, summarize data
n = 6;  g = 3
OBS = matrix(c(0,1,1,1,1,1,
               1,0,0,0,0,0,
               0,0,0,0,0,0), byrow=T, nrow=g)  # g x n data matrix
x.obs = rowSums(OBS); x.obs                    # g-vector of sums
q.obs = sum((x.obs - 2)^2 / 2);  q.obs         # observed GOF statistic

# Simulate permutation distribution
set.seed(308)                               # omit for different simulation
m = 10000                                   # number of iterations
q.prm = numeric(m)                          # m-vector of 0's
for (i in 1:m)                              # begin outer loop
     {
     PRM = OBS
     for (j in 1:n)                           # begin inner loop
         {
         PRM[,j] = sample(PRM[,j], 3)      # permute each column
         }                                 # end inner loop
     x.prm = rowSums(PRM)                # g-vector of sums
     q.prm[i] = sum((x.prm - 2)^2 / 2)   # GOF stat. of permuted matrix
     }                                   # end outer loop

# Graphical display of results
cut = (0:(max(q.prm)+2)) - .5                  # for "nice" hist. intervals
hist(q.prm, breaks = cut, prob=T, col="wheat") # draw histogram
xx = seq(0, max(q.prm), len = 200)             # x-values for density curve
lines(xx, dchisq(xx, 2), col="blue")           # draw density curve
text(10, .45, paste("P=", round(mean(q.prm >= q.obs), 3)), cex=.9, pos=4)    # p-value
abline(v=q.obs, lty="dashed", col="darkred")   # line at obs. GOF stat
x = c(0, 1, 3, 4, 7, 12)                       # possible values of GOF stat
exact = c(30, 120, 50, 30, 12, 1)/243          # exact GOF probabilities
points(x, exact, pch=19, col="darkgreen")      # exact probability points

# Printed display of results
length(unique(q.prm))                          # no. of distinct values of sim GOF stat
summary(as.factor(q.prm))/m                    # tally GOF values (divided by m)
summary(q.prm)                                 # compare mean with E[CHISQ(df=2)] = 2
var(q.prm)                                     # compare variance with V[CHISQ(df=2)] = 4
mean(q.prm >= q.obs)                           # P-value of sim. perm. test
p.apx = 1 - pchisq(q.obs, 2); p.apx            # compare with chi-sq. approx.

# Bold maroon lines show, for comparison, chi-sq. approximation of GOF statistic.
# Italic dark green lines show, for comparison, exact probabilities of GOF statistic.


# -------------------------------------------------------------------------------------------------
# Printed output for seed shown
# > x.obs = rowSums(OBS); x.obs                    # g-vector of sums
# [1] 5 1 0
# > q.obs = sum((x.obs - 2)^2 / 2);  q.obs         # observed GOF statistic
# [1] 7
# ...
# > length(unique(q.prm))                          # values of sim GOF stat
# [1] 6
# > summary(as.factor(q.prm))/m                    # tally GOF values (divided by m)
#      0      1      3      4      7     12
# 0.1257 0.4944 0.2047 0.1188 0.0523 0.0041
# > summary(q.prm)                                 # compare mean with E[CHISQ(df=2)] = 2
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#   0.000   1.000   1.000   1.999   3.000  12.000
# > var(q.prm)                                     # compare variance with V[CHISQ(df=2)] = 4
# [1] 3.394938
# > mean(q.prm >= q.obs)                           # P-value of sim. perm. test
# [1] 0.0564
# > p.apx = 1 - pchisq(q.obs, 2); p.apx           # compare with chi-sq. approx.
# [1] 0.03019738


###################################   FIGURE 1(b)   ######################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                         #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# (Some embellishments in graphs for publication omitted.)                               #
##########################################################################################

# Specify constants, input data, summarize data
n = 6;  g = 3
# Ranks within columns for Figure 1(b)
OBS = matrix(c(1, 3, 3, 3, 3, 3,
               3, 2, 2, 2, 2, 2,
               2, 1, 1, 1, 1, 1), byrow=T, nrow=g)

row.obs = rowSums(OBS)
statsta.obs = var(row.obs)
row.obs;  statsta.obs

# Simulate permutation distribution
set.seed(308)
m = 10000
stat.prm = numeric(m)

for (i in 1:m)
    {
    PRM = OBS
    for (j in 1:n)
        {
        PRM[,j] = sample(PRM[,j], 3)
        }
    row.prm = rowSums(PRM)
    stat.prm[i] = var(row.prm)
    }

# Display of graphical results
cut = (0:(max(stat.prm)+1)) - .5
hist(stat.prm, breaks = cut, prob=T, col="wheat")
abline(v=statsta.obs, lty="dashed", col="darkred")
text(30, .2, paste("P=", round(mean(stat.prm >= statsta.obs), 3)), cex=.9, pos=4)   # p-value
# Display of numerical results
length(unique(stat.prm))                # granularity: no. of unique sim. values of perm. stat.
mean(stat.prm >= statsta.obs)           # P-value of simulated permutation test



# ------------------------------------------------------------------------------------------
# Printed output for seed shown

## Fig. 1(b)
# > row.obs;  statsta.obs
# [1] 16 13  7
# [1] 21
# ...
# > length(unique(stat.prm))
# [1] 16
# > mean(stat.prm >= statsta.obs)
# [1] 0.0304


##################################   FIGURE 2(a)   #######################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                         #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 2, panel (a).  Code in purple mark where changes need to be done to    #
#     get panels (b-d)                                                                   #
# Use full screen to display graph in graphics window of R.                              #
##########################################################################################
```

```
# Input and manipulate data
# (Optionally perform paired t-test or Wilcoxon signed-rank test)
Air = c(.82, .86, 1.86, 1.64, 12.57, 1.56, 1.28, 1.08, 4.29, 1.34,
        14.68, 3.64, 3.89, .58, 9.50, .93, 0.49, 31.04, 1.66)
SO2 = c(.72, 1.05, 1.40, 2.30, 13.49, .62, 2.41, 2.32, 8.19, 6.33,
        19.88,8.87,9.25, 6.59, 2.17, 9.93,13.44,16.25,19.89)
Dif = Air - SO2
n = length(Dif)
# t.test(Air, SO2, pair=T, alte="less")      # paired t-test (omit first # to activate)
# wilcox.test(Air, SO2, pair=T, alte="less") # Wilcoxon SR test (omit first # to activate)

# Simulate permutation distribution
set.seed(1234)                               # Omit seed for a different simulation
m = 10000;  tsta.prm = mean.prm = medn.prm = trmn.prm = numeric(m)
tsta.obs = sqrt(n)*mean(Dif)/sd(Dif)                # t-stat metric

for (i in 1:m)
 {
 perm <- sample(c(-1,1), n, repl=T)*Dif
 tsta.prm[i] = sqrt(n)*mean(perm)/sd(perm)          # t-stat metric
}
tsta.pv = round(mean(tsta.prm <= tsta.obs) ,3)      # Simulated P-value (t-stat)

# Display graphical results (P-values in text on each graph)

  hist(tsta.prm, prob=T, xlim=c(-4.5,4.5), col="wheat",
  main="(a) Perm. Test for SAR Differences", xlab="T-Statistic of SAR Differences")
  yc =.9 *max(hist(tsta.prm, prob=T, plot=F)$density)  # y-coordinate for text
  text(-4.7, yc, paste("P=",tsta.pv), cex=.9, pos=4)   # See ?text for parameters
  xx = seq(-3,3,len=200);  lines(xx,dt(xx,n-1), lwd=2, col="blue")
  abline(v=tsta.obs, lwd=2, col="red", lty="dashed")


######################################   FIGURE 3(a)   ###################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                         #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 3, panels (a), (b) and (c)                                             #
# In R, use full width of graphics window to display result.                             #
##########################################################################################

# Input and manipulate data
Air = c(.82, .86, 1.86, 1.64, 12.57, 1.56, 1.28, 1.08, 4.29, 1.34,
        14.68, 3.64, 3.89, .58, 9.50, .93, 0.49, 31.04, 1.66)
SO2 = c(.72, 1.05, 1.40, 2.30, 13.49, .62, 2.41, 2.32, 8.19, 6.33,
        19.88,8.87,9.25, 6.59, 2.17, 9.93,13.44,16.25,19.89)
Dif = Air - SO2;  n = length(Dif)
mean.obs = mean(Dif);           tsta.obs = sqrt(n)*mean(Dif)/sd(Dif)

# Simulate permutation distribution
set.seed(1234)
m = 10000;  mean.prm = tsta.prm = trmn.prm = medn.prm = numeric(m)

for (i in 1:m)
    {
    perm = sample(c(-1,1), n, repl=T)*Dif
    mean.prm[i] = mean(perm)
    tsta.prm[i] = sqrt(n)*mean(perm)/sd(perm)
    }

# Display graphical results (P-values in text on each graph)
       plot(mean.prm, tsta.prm, pch=".", cex=2, col="#444477",
         main="(a) Mean and T-stat Metrics Are Equivalent",
         xlab="Sample Mean", ylab="T-Statistic")
       abline(h=tsta.obs, col="red")
       abline(v=mean.obs, col="red")


# Print P-values:
# Rationale: The comparision (tsta.prm <= tsta.obs) yields a 'logical' m-vector of T's and F's.
#    When arithmetic is performed on a logical vector, T is taken to be 1, and F to be 0.
#    The 'mean' of a vector of 0's and 1's is the proportion of 1's.
```

```
mean(tsta.prm <= tsta.obs)              # simulated P-value with t-stat as metric
mean(mean.prm <= mean.obs)              # simulated P-value with mean as metric

# ------------------------------------------------------------------------------------------
# Printed output for seed shown
> mean(tsta.prm <= tsta.obs)            # simulated P-value with t-stat as metric
[1] 0.0538
> mean(mean.prm <= mean.obs)      # simulated P-value with mean as metric
[1] 0.0538
```

```
######################################   FIGURE 4(a,d)   ##################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                          #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 4(a) and 4(d).  Significance level (mu=0) and Power (mu=-1/2)          #
# Except for mu, (a) is similar to (c), (b) similar to (d).                               #
# Plot 'main' titles and all 'text' statements added for publication.                    #
#   Also, coordinates for 'text' statements are ad hoc for each graph.                    #
#   In R graphics window, use full screen for proper display of 'text' statements.       #
###########################################################################################

mu = 0  # Null hypothesis true: P(Reject) = Significance level
m = 5000;  b = xbar = std = t = numeric(m)
set.seed(508)
for (i in 1:m) {
     x = rnorm(13, mu, 1)
     b[i] = sum(x > 0)
     xbar[i] = mean(x);  std[i] = sd(x)
     t[i] = sqrt(13)*mean(x)/sd(x) }

par(mfrow=c(2,1))  # puts 2 graphs into 1-page figure

        plot(xbar, std, pch=20, col="#BBDDCC",
          main="(a) 5000 Samples from N(0, 1): n = 13", xlab="Sample Mean", ylab="Sample SD")
        condt = (t < qt(.05, 12))
        points(xbar[condt], std[condt], pch=20, col="red")
        condb = (b < qbinom(.05, 13, .5))
        points(xbar[condb], std[condb], pch=".", cex=2, col="blue")
        abline(a=0, b=sqrt(13)/qt(.05,12), col="darkred")
        text(.14, 1.7, "Left of line: T-test rejects", cex=.7, pos=4, col="red")
        text(.14, 1.5, "Small dots: Sign test rejects", cex=.7, pos=4, col="blue")

mean(condb); mean(condt); mean(condb | condt)  # printed rejection probabilities

mu = -1/2  # Null hypothesis false: P(Reject) = Power against the alternative mu = -1/2
b = xbar = std = t = numeric(m)
for (i in 1:m) {
     x = rnorm(13, mu, 1)
     b[i] = sum(x > 0)
     xbar[i] = mean(x);  std[i] = sd(x)
     t[i] = sqrt(13)*mean(x)/sd(x) }

        plot(t, b+runif(m,-.3,.3), pch=".", main="(d) 5000 Samples from N(-1/2, 1): n = 13",
                xlab="T-Statistic", ylab="Jittered Number Positive", col="#507060")
        abline(v = qt(.05, 12), col="darkred")
        abline(h = 3.5, col="darkblue")
        text(-7, 9, "Left of Line: T-test rejects", cex=.7, pos=4, col="red")
        text(-1.5, 1, "Below Line: Sign test rejects", cex=.7, pos=4, col="blue")

mean(condb); mean(condt); mean(condb | condt)  # printed rejection probabilities

par(mfrow=c(1,1))  # return to normal graphics configuration
```

```
###################################   FIGURE 5(a)   ####################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                          #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 5(a).                                                                   #
# In R, use full width of graphics window to display result.                             #
###########################################################################################

## Corn data
```

```
x1 = c(166.7,   172.2,   165.0,   176.9)
x2 = c(162.8,   142.4,   162.8,   162.4)
x = c(x1, x2);  n1 = length(x1);  n2 = length(x2)
t.test(x1, x2, var.eq=T)                                # pooled
t.test(x1, x2)                                          # Welch

se = sqrt((1/n1+1/n2)*((n1-1)*var(x1)+(n2-1)*var(x2))/(n1+n2-2))  # pooled
tsta.obs = dfmn.obs/se; tsta.obs

# Simulate permutation distribution
set.seed(123)
m = 100000                  # For speed over accuracy, use 10000
tsta.prm = numeric(m)

for (i in 1:m)
       {
       perm = sample(x, n1+n2)
       p1 = perm[1:n1]; p2 = perm[(n1+1):(n1+n2)]
       num = mean(p1) - mean(p2)
       se = sqrt((1/n1+1/n2)*((n1-1)*var(p1)+(n2-1)*var(p2))/(n1+n2-2))  #pooled
       tsta.prm[i] = num/se
       }

# Display graphical results
#par(mfrow=c(1,3))

       cut = seq(-2.30, 2.30, length=21)
       hist(tsta.prm, breaks=cut, prob=T, col="wheat", xlim=c(-3.5, 3.5),
        xlab="T-Statistic", main = "(a) Permutation Dist'n of Corn Data")
       abline(v=c(-tsta.obs, tsta.obs), lwd=2, col="red", lty="dashed")
       tt = seq(-4, 4, len=100)
       lines(tt, dt(tt, n1+n2-2), lwd=2, col="blue")


# Print 2-sided P-values
mean((tsta.prm >= tsta.obs)|(tsta.prm <= -tsta.obs)) # metric: pooled-t
2/choose(8, 4)                                       # exact (for corn data)



####################################   FIGURE 5(b)   #####################################
# L. Eudey, J. Kerr, B. Trumbo (September 2009):                                         #
# Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs #
# Code for Figure 5(b).                                                                  #
# Use full screen to display graph in graphics window of R.                              #
##########################################################################################

## Challenger data
x1 = Cold = c(1, 1, 1, 3)
x2 = Warm = c(0, 0, 0, 0, 0,  0, 0, 0, 0, 0,  0, 0, 0, 0, 0,  0, 0, 1, 1, 2)
x = c(x1, x2);  n1 = length(x1); n2 = length(x2)

dfmn.obs = mean(x1) - mean(x2); dfmn.obs
# Simulate permutation distribution
set.seed(508)
m = 100000                                  # For speed over accuracy, use 10000
dfmn.prm = numeric(m)
for (i in 1:m)
       {
       perm = sample(x, n1+n2);  p1 = perm[1:n1]; p2 = perm[(n1+1):(n1+n2)]
       num = mean(p1) - mean(p2);  dfmn.prm[i] = num
       }

# Display graphical results
```

```
       xv = seq(-.5, 1.6, by=.3);  cut = c(xv, 1.9) -.15
       hist(dfmn.prm, breaks=cut, col="wheat", xlab="Diff. in Means",
        main = "(b) Perm. Dist'n of Challenger Data")    # Frequency histogram
       abline(v=dfmn.obs, lwd=2, col="red", lty="dashed")
       rs = m*round(c(2380, 3400, 2040, 1530, 855, 316, 95, 10)/choose(24,4), 3)
       points(xv, rs, pch=19, col="darkgreen")                # Expected Frequencies

# Print right-sided P-values
mean(dfmn.prm >= dfmn.obs) # metric: diff of means
mean(tsta.prm >= tsta.obs) # metric: t-statistic
105/choose(24, 4)                                      # exact (for Challenger data)
```

Acknowledgements

The authors would like to thank the editors (Dr. William Notz, Dr. John Gabrosek), the anonymous associate editor and anonymous referees for their valuable comments, edits and assistance in publishing this manuscript.

References

Aberson, C.L., Berger, D.E., Healy, M.R. and Romero, V. L. (2002), “An Interactive Tutorial for Teaching Statistical Power,” *Journal of Statistics Education*, 10(3). http://www.amstat.org/publications/jse/v10n3/aberson.html

Bethel, R.A., Haye, W.H., Oberzanek, G., George, D.T., Jimerson, D.C. and Ebert, M.H., (1989), “Effect of 0.25 ppm Sulphur Dioxide on Airway Resistance in Freely Breathing, Heavily Exercising Asthmatic Subjects,” *American Review of Respiratory Disease*, 10, pp. 659-661 (Asthma data).

Callaert, H. (1999), “Nonparametric Hypotheses for the Two-Sample Location Problem,” *Journal of Statistics Education* [Online] 7(2). http://www.amstat.org/publications/jse/secure/v7n2/callaert.cfm

Capéraà, P., Van Cutsem, B. (1988), *Méthodes et Modéles les en Statistique NonParamétrique: Exposé Fondamental*. Dunod (Paris), Chapters IV, V and VI

Coakley, C. W. (1996), “Suggestions for Your Nonparametric Statistics Course,” *Journal of Statistics Education* [Online] 4(2). http://www.amstat.org/publications/jse/v4n2/coakley.html

Dalgaard, P. (2002), *Introductory Statistics with R*, Springer, New York

Dwass, M. (1957). “Modified randomization tests for nonparametric hypotheses,” *Annals of Mathematical Statistics*, 28, pp. 181–187

Feynman, R.P. (1988), *Who Cares What People Think?*, W.W. Norton, New York, p. 86 (Challenger data)

Good, P. (2000), *Permutation Tests: A Practical Guide to Resampling Methods and Testing Hypotheses,* 2nd Edition, Springer, New York

Hesterberg, T., Monaghan, S., Moore, D.S., Clipson, A., Epstein, R. (2003), *Bootstrap Methods and Permutations Tests: Companion Chapter 18 to The Practice of Business Statistics*, W.H. Freeman and Co., New York, http://bcs.whfreeman.com/pbs/cat_160/PBS18.pdf

Higgins, J.J. (2004), *Introduction to Modern Nonparametric Statistics,* Duxbury Advanced Series, Thomson-Brooks/Cole, Pacific Grove, CA

Hoaglin, D.C., Iglewicz, B., Tukey, J.W. (1986), “Performance of Some Resistant Rules for Outlier Labeling,” *Journal of the American Statistical Association*, 81(396), pp. 991-999

Hodgess, E.M. (2004), “A Computer Evolution in Teaching Undergraduate Time Series,” *Journal of Statistics Education* [Online] 12(3). http://www.amstat.org/publications/jse/v12n3/hodgess.html

Hollander, M. and Wolfe, D.A. (1999), *Nonparametric Statistical Methods,* 2nd Edition, Wiley, New York

Jöckel, K.-H. (1986). “Finite sample properties and asymptotic efficiency of Monte Carlo tests.” *Annals of Statistics,* 14(1)**,** pp. 336–347

Moore, D.S. (2007), *Basic Practice of Statistics*, 4th Edition, W.H. Freeman and Co., New York, Chapter 26 on CD-ROM.

Moore, D.S. and McCabe, G.P. (1999), *Introduction to the Practice of Statistics*, 3rd Edition, W.H. Freeman and Co., New York, Ch. 14 (CD-ROM) problem 14.4: Corn data collected by Phillips, S., Purdure University)

Pagano, M. and Gauvreau, K. (2000), *Principles of Biostatistics*, 2nd Edition, Duxbury Brooks/Cole, Pacific Grove, Chapter 13, problem 8, p. 318

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.R-project.org. ISBN 3-900051- 07-0

Ramsey, F.L. and Schafer, D.W. (2002), *The Statistical Sleuth,* Duxbury, p. 97 (Challenger data)

Rizzo, M.L. (2008), *Statistical Computing with R*, Chapman & Hall/CRC, Boca Raton

S-Plus®, Insightful Corporation, Seattle, WA

Saekel, K. (1994), *Taster's Choice Column*, San Francisco Chronicle, September 14, 1994, Food Section, p. 2 (Chocolate pudding data)

Sprent, P. (1993), *Applied Nonparametric Statistical Methods*, 2nd Edition, Chapman & Hall, London

Stapleton, J.H. (2008), *Models for Probability and Statistical Inference: Theory and Applications.* Wiley, Chapter 10

Tarpey, T., Acuna, C., Cobb, G., De Veaux, R. (2002), "Curriculum Guidelines for Bachelor of Arts Degrees in Statistical Science," *Journal of Statistics Education* [Online] 10(2). http://www.amstat.org/publications/jse/v10n2/tarpey.html

Utts, J.M. and Heckard, R.F. (2006), *Mind on Statistics*, 3rd Edition, Duxbury, Supplemental Topic 2 on CD-ROM

Venables, W.N., Smith, D.M. and the R Development Core Team (2008), *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics, Version 2.7.0* (2008-04-22). http://cran.r-project.org/doc/manuals/R-intro.pdf

Verzani, J. (2001-02), "Simple R: Using R for Introductory Statistics," http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

Verzani, J. (2008), "Using R in Introductory Statistics Courses with the pmg Graphical User Interface," *Journal of Statistics Education*, [Online] 16(1). http://www.amstat.org/publications/jse/v16n1/verzani.html

Wood, M. (2005), "The Role of Simulation Approaches in Statistics," *Journal of Statistics Education* 13(3). http://www.amstat.org/publications/jse/v13n3/wood.html

---

Lynn Eudey
Assistant Professor
Department of Statistics and Biostatistics
California State University East Bay
Hayward, CA 94542
e-mail: lynn.eudey@csueastbay.edu

Joshua Kerr
Assistant Professor
Department of Statistics and Biostatistics
California State University East Bay
Hayward, CA 94542
e-mail: josh.kerr@csueastbay.edu

Bruce Trumbo
Professor Emeritus of Statistics and Mathematics
Department of Statistics and Biostatistics
California State University East Bay
Hayward, CA 94542
e-mail: lynn.eudey@csueastbay.edu

Volume 18 (2010) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications