The Stata Journal (2007)
7, Number 3, pp. 402–412

An exact and a Monte Carlo proposal to the Fisher–Pitman permutation tests for paired replicates and for independent samples

Johannes Kaiser
Laboratory for Experimental Economics
University of Bonn
Bonn, Germany
johannes.kaiser@uni-bonn.de

Abstract. This article concerns the nonparametric Fisher–Pitman tests for paired replicates and independent samples. After outlining the theory of exact tests, I derive Monte Carlo simulations for both of them. Simulations can be useful if one deals with many observations because of the complexity of the algorithms in regard to sample sizes. The tests are designed to be a more powerful alternative to the Wilcoxon signed-rank test and the Wilcoxon–Mann–Whitney rank-sum test if the observations are given on at least an interval scale. The results gained by Monte Carlo versions of the tests are accurate enough in comparison to the exact versions. Finally, I give examples for using both supplemented tests.

Keywords: st0134, permtest1, permtest2, nonparametric tests, Monte Carlo, permutation tests

1 Introduction

In behavioral sciences, frequently used statistical tools are regression analysis and nonparametric tests like Spearman's rank correlation, the McNemar change test, the Fisher exact test, the Kruskal–Wallis one-way analysis of variance, the Wilcoxon signed-rank test, and the Wilcoxon–Mann–Whitney rank-sum test. This article deals with two tests that can replace the last two tests mentioned if the observations are given at least on an interval scale. These tests are the Fisher–Pitman test for paired replicates and the Fisher–Pitman permutation test for independent samples (see Fisher 1935 and Pitman 1937), also referred to as randomization tests. Why are the permutation tests more powerful than the respective Wilcoxon tests? Siegel and Castellan (1988) compare both Wilcoxon and permutation tests with the appropriate parametric test. They find that the asymptotic power efficiencies of both the Wilcoxon signed-rank and rank-sum tests compared with the respective parametric t test are only 95.5%, whereas both permutation tests display power efficiencies of 100%.

Here I outline two algorithms for the well-known permutation tests: one for paired replicates and one for two independent samples. Both algorithms are complex in regard to sample size. Thus p-values are time consuming to compute even for moderate sample sizes. After I outline the exact algorithms, I show a Monte Carlo simulation approach

© 2007 StataCorp LP

to approximate p-values. Later, I give an example of the supplemented implementations of both tests before concluding the case for the permutation tests.

2 The tests

Below, I outline the exact algorithms first. Per Siegel and Castellan (1988), I limit the description of the details to the extent necessary for specifying the respective algorithm. You can find instructions on how to carry out the tests in the cited book. After deriving two algorithms for the exact case, I show a method to facilitate Monte Carlo simulations.

2.1 One exact algorithm for each permutation test

The permutation test for paired replicates assumes as the null hypothesis that paired observations of an outcome under two different conditions are randomly assigned to the two conditions for each subject. Below, I summarize the rationale of the test before deriving an algorithm to compute the significance levels.

Let X_i specify the interval-scaled outcome under the first condition for a subject $i \in \{1, \ldots, n\}$ and Y_i the outcome for the same subject under the second condition. Let then $d_i = X_i - Y_i$ be the difference of the outcomes under the first and the second condition. If H_0 were true, a positive and a negative sign for d_i would be equally likely. Because the size of the regarded sample is n, there are 2^n possibilities for the distribution of a positive or a negative sign among all differences in d_i, which would be all equally likely if H_0 were true.

For each possibility, one can calculate the sum of the differences, $\sum d_i$, and compare it with the $\sum d_i$ actually observed (the critical value). The relation of the number of all theoretically possible sums that are less than or equal to the critical value of all theoretically possible sums, 2^n, is equal to the lower-tailed p-value; the relation of the number of all theoretically possible sums that are greater than or equal to the critical value of 2^n is equal to the upper-tailed p-value. The two-tailed p-value is the minimum of 1 and twice the value of the upper-tailed and the lower-tailed p-values.

The supplemented algorithm that facilitates the necessary computations uses binary counting to derive all possible $\sum d_i$. In particular, it performs the following steps:

- Let X_i and Y_i contain the observed values of subject i in a sample of n independent observations.
- Create the differences $d_i = X_i - Y_i$.
- Compute the critical value $c = \sum d_i$.
- Create an n-rows sign vector $S^\top = (-1, -1, \ldots, -1)$.
- Let $l = 0$ and $u = 0$.

- Repeat the following steps:
  1. For every $j \in \{1, \ldots, n\}$:
     a. If $s_j = -1$, set $s_j = 1$ and end this loop.
     b. Set $s_j = -1$.
  2. Compute the test statistic $a = \sum_{i=1}^{n} s_i \times |d_i|$.
  3. If $a \leq c$, increase l by one.
  4. If $a \geq c$, increase u by one.
  5. If $s_i = (1, 1, \ldots, 1)$, end this loop.
- The upper-tailed p-value equals $p_{\text{upper}} = \frac{u}{2^n}$.
- The lower-tailed p-value equals $p_{\text{lower}} = \frac{l}{2^n}$.
- The two-tailed p-value equals $p_{\text{two}} = \min(1, 2 \times p_{\text{upper}}, 2 \times p_{\text{lower}})$.

Since this test considers not only the signs but also the size of the difference of the observations, it accounts for more of the data than the Wilcoxon signed-rank test.

The Fisher–Pitman permutation test for independent samples is a powerful alternative to the Wilcoxon–Mann–Whitney rank-sum test. It tests the difference between the means of two independent samples. Let X_i contain the interval-scaled outcome of a subject i among m subjects in the first group and Y_j the outcome of a subject j among n subjects in the second group. The null hypothesis states that there is no difference in the mean of the population from which X_i is drawn to the mean of the population from which Y_i is drawn, i.e., that all of the $m+n$ observations may be considered to be from the same population. If H_0 yielded true, it would be equally likely that an observed value occurs in X or in Y. This scenario creates $\binom{m+n}{n}$ equally likely possibilities of distributing all observed values among X and Y.

For each possibility, one can calculate the difference of the sums of both theoretically possible samples $\sum X_i - \sum Y_j$ and compare it with the same measure of the observed values. The latter one is the critical value for the test. The relation of the number of all theoretically possible sums that are less than or equal to the critical value to all theoretically possible sums, $\binom{m+n}{n}$, is equal to the lower-tailed p-value; the relation of the number of all theoretically possible sums that are greater than or equal to the critical value to $\binom{m+n}{n}$ is equal to the upper-tailed p-value. The two-tailed p-value is the minimum of 1 and twice the value of the upper-tailed and the lower-tailed p-value.

The supplemented algorithm performs the necessary computations as described here:

- Let X_i contain the observed value of an individual, i, in the first group of m independent observations and Y_j contain the observed value of another individual, j, in the second group of n independent observations.
- Let Z be the concatenation of X and Y. Thus, $X_i = Z_i \ \forall i \in \{1, \ldots, m\}$ and $Y_j = Z_{m+j} \ \forall j \in \{1, \ldots, n\}$.

- Compute the critical value $c = \sum_{i=1}^{m} Z_i - \sum_{j=m+1}^{m+n} Z_j$.
- Let $l = 0$ and $u = 0$.
- Create a $(m+n) \times \binom{m+n}{n}$ matrix M that contains in its columns all possibilities to distribute m times the number 1 and n times the number -1 in the $m+n$ rows. This is done by using Mata's `cvpermute()` function (see [M-5] **cvpermute( )**).
- For every $e \in \left(1, \ldots, \binom{m+n}{n}\right)$:
  1. Calculate the test statistic $a = \sum_{i=1}^{m+n} M_{ie} \times Z_i$.
  2. If $a \leq c$, increase l by one.
  3. If $a \geq c$, increase u by one.
- The upper-tailed p-value equals $p_{\text{upper}} = \frac{u}{\binom{m+n}{n}}$.
- The lower-tailed p-value equals $p_{\text{lower}} = \frac{l}{\binom{m+n}{n}}$.
- The two-tailed p-value equals $p_{\text{two}} = \min(1, 2 \times p_{\text{upper}}, 2 \times p_{\text{lower}})$.

Just like the Fisher–Pitman permutation test for paired replicates, this test is superior to the respective Wilcoxon tests if the observed values are given on at least an interval scale.

The realization of either test turns out to be time consuming: for the permutation test for paired replicates, the outer loop has a complexity of $O(2^n)$. The permutation test for independent samples can be similarly intensive in computation: the complexity totals to $O(\binom{m+n}{n})$. Below, I draw a method to approximate the significance levels by using Monte Carlo simulations.

2.2 One Monte Carlo–based algorithm for each permutation test

Monte Carlo simulations are an appropriate device to reduce complexity while setting aside accuracy only to a small extent. Instead of the test statistic's being computed for the complete set of the sign vectors, the test statistic is calculated only for randomly drawn sign vectors (with the possibility of repetition). The p-value equals the ratio of the number of sign vectors for which the test statistic is less than or equal to (or greater than or equal to for a right-tailed test) the critical value to the total number of sign vectors drawn. Of course, this approach is less accurate and leads to an error term in the p-values, but with a sufficiently high k the error term influences only their fourth or fifth decimal place.

In detail, the Monte Carlo–based algorithm for the Fisher–Pitman permutation test for paired replicates looks as follows:

- Let X_i and Y_i contain the observed values of subject i in a sample of n independent observations.
- Let k be the number of simulation runs to facilitate.
- Create the differences $d_i = X_i - Y_i$.
- Compute the critical value $c = \sum d_i$.
- Let $l = 0$ and $u = 0$.
- Repeat the following steps:
  1. For every $j \in \{1, \ldots, k\}$:
  2. Create a sign vector $S^\top = (s_1, \ldots, s_n)$ by setting $s_i = 1 - 2R \; \forall i \in \{1, \ldots, n\}$ with $R \sim$ Bernoulli $(p = 0.5)$.
  3. Compute the test statistic $a = \sum_{i=1}^{n} s_i \times |d_i|$.
  4. If $a \leq c$, increase l by one.
  5. If $a \geq c$, increase u by one.
- The upper-tailed p-value equals $p_{\text{upper}} = \frac{u}{k}$.
- The lower-tailed p-value equals $p_{\text{lower}} = \frac{l}{k}$.
- The two-tailed p-value equals $p_{\text{two}} = \min(1, 2 \times p_{\text{upper}}, 2 \times p_{\text{lower}})$.

Limiting the number of investigated sign vectors to k drastically reduces the overall computational effort. By default, the supplemented test carries out a total of $k = 2 \times 10^5$ runs of the simulation.

How is the permutation test for independent samples carried out as a Monte Carlo simulation? The algorithm is conducted as follows:

- Let X_i contain the observed value of an individual i in the first group of m independent observations and Y_j contain the observed value of another individual j in the second group of n independent observations.
- Let Z be the concatenation of X and Y. Thus, $X_i = Z_i \; \forall i \in \{1, \ldots, m\}$ and $Y_j = Z_{m+j} \; \forall j \in \{1, \ldots, n\}$.
- Compute the critical value $c = \sum_{i=1}^{m} Z_i - \sum_{j=m+1}^{m+n} Z_j$.
- Let $l = 0$ and $u = 0$.
- Create a sign vector $S^\top = (s_1, \ldots, s_{m+n})$ with $s_i = 1$ if $1 \leq i <= m$ and $s_i = -1$ if $m < i \leq m + n$.
- For every $i \in (1, \ldots, k)$:

1. Shuffle the sign vector S. This is done by using Mata's `_jumble()` function (see [M-5] **sort( )**).
2. Calculate the test statistic $a = \sum_{i=1}^{m+n} S_i \times Z_i$.
3. If $a \leq c$, increase l by one.
4. If $a \geq c$, increase u by one.

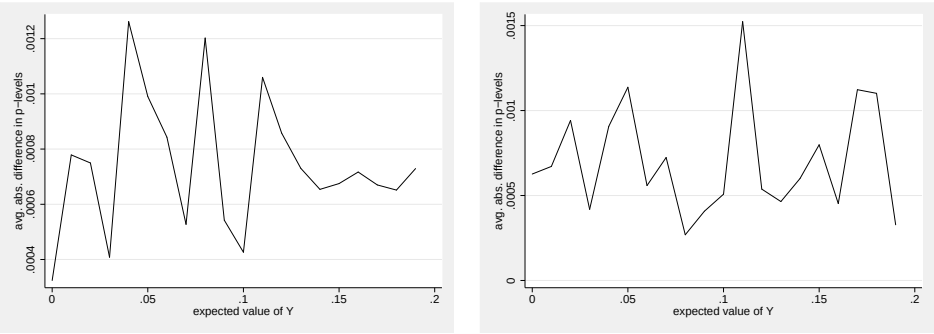
- The upper-tailed p-value equals $p_{\text{upper}} = \frac{u}{k}$.
- The lower-tailed p-value equals $p_{\text{lower}} = \frac{l}{k}$.
- The two-tailed p-value equals $p_{\text{two}} = \min(1, 2 \times p_{\text{upper}}, 2 \times p_{\text{lower}})$.

Just as for the permutation test for paired replicates, the supplemented test carries out a total of $k = 2 \times 10^5$ runs of the simulation by default.

3 Comparison of exact and Monte Carlo results

To get an idea on the size of the difference in the p-values given by the exact and the Monte Carlo versions of the tests, I conduct a simulation study.

For the test for paired replicates, one draws two samples, x and y, with the same sample size, $n = 12$, with specific underlying distributions $X \sim N(\mu_x, 1)$ and $Y \sim N(\mu_y, 1)$. Lower-tailed p-values are calculated for both exact and Monte Carlo versions of the test, and the absolute difference between the p-values is stored. This process is repeated c times for each μ_x and μ_y in question. For two independent samples, one proceeds analogously with a combined sample size of 12. Figure 1 displays the average absolute differences in p-values for $c = 5$, $\mu_x = 0$, and $\mu_y = i \times 0.01 \; \forall i \in \mathbb{Z} \wedge 0 \leq i < 20$.



Paired replicates Independent samples

Figure 1: Mean absolute difference in p-values of exact and Monte Carlo tests for random samples of $N(0, 1)$ against $N(\mu_y, 1)$

The differences in p-values are negligibly small, that is, smaller than 0.001 (paired replicates, significant at $p = 0.0016$, one-tailed Wilcoxon signed-rank test; independent samples, significant at $p = 0.0004$, same test). The difference in means of the underlying distributions also appears not to be correlated with the simulation error (paired replicates, Bravais–Pearson's product moment correlation coefficient $r = -0.0145$; independent samples, $r = 0.0142$). Thus the Monte Carlo versions of the tests seem to be accurate enough for the investigated distributions.

4 Usage

Below are Stata implementations of both tests. This section deals with a command synopsis of both tests and illustrates both in the context of a study.

Two new commands are available: `permtest1` executes the Fisher–Pitman permutation test for paired replicates (one-sample case), and `permtest2` executes the Fisher–Pitman permutation test for independent samples (two-sample case).

4.1 Fisher–Pitman permutation test for paired replicates

Syntax

[`by` *varlist*`:`] `permtest1` *varname*`=`*exp* [*if*] [*in*] [`,` `runs(`*integer*`)` `exact` `simulate`]

where *varname* specifies the variable to test and *exp* specifies the expression to test the variable against. *exp* may be a constant, another variable, or any other expression.

Options

`runs(`*integer*`)` specifies the number of Monte Carlo simulation runs to perform. It defaults to 2×10^5.

`exact` forces the calculation of exact significance levels. Specifying this option may increase run time even with moderate sample sizes.

`simulate` forces the estimation of significance levels by using Monte Carlo simulations. This method is less accurate but also less time consuming (see sec. 2.2). By default, the test uses Monte Carlo simulations automatically if the sample size exceeds 13.

The options `exact` and `simulate` may not be specified at the same time, and the `runs(`*integer*`)` option makes sense only with Monte Carlo simulations.

Saved results

`permtest1` saves the following in `r()`:

Scalars

`r(criticalValue)`	critical value
`r(zero)`	number of zeros in the difference vector
`r(negative)`	number of negative values in the difference vector
`r(positive)`	number of positive values in the difference vector
`r(runs)`	number of simulation runs conducted
`r(mode)`	`1` if the exact test, `2` if Monte Carlo simulations were used
`r(N)`	sample size
`r(twotail)`	two-tailed p-value
`r(uppertail)`	upper-tailed p-value
`r(lowertail)`	lower-tailed p-value

4.2 Fisher–Pitman permutation test for independent samples

Syntax

[`by` *varlist*`:`] `permtest2` *varname* [*if*] [*in*], `by(`*varname*`)` [`runs(`*integer*`)` `exact` `simulate`]

where *varname* specifies the variable to test.

Options

`by(`*varname*`)` is required and specifies the grouping variable. It must be numeric, and there must be exactly two different groups in the specified sample.

`runs(`*integer*`)` specifies the number of Monte Carlo simulation runs to perform. It defaults to 2×10^5.

`exact` forces the calculation of exact significance levels. Specifying this option may increase run time even with moderate sample sizes.

`simulate` forces the estimation of significance levels by using Monte Carlo simulations. This method is less accurate but also less time consuming (see sec. 2.2). By default, the test uses Monte Carlo simulations automatically if the sample size exceeds 15.

The options `exact` and `simulate` may not be specified at the same time, and the `runs(`*integer*`)` option makes sense only with Monte Carlo simulations.

(*Continued on next page*)

Saved results

`permtest2` saves the following in `r()`:

Scalars

`r(criticalValue)`	critical value
`r(n1)`	number of observations in the first group
`r(n2)`	number of observations in the second group
`r(runs)`	number of simulation runs conducted
`r(mode)`	`1` if the exact test, `2` if Monte Carlo simulations were used
`r(N)`	sample size
`r(twotail)`	two-tailed p-value
`r(uppertail)`	upper-tailed p-value
`r(lowertail)`	lower-tailed p-value

4.3 Example using both tests

Consider the following (fictional) setting: six high school and six graduate students are asked independently of each other to collect receipts that show money spent on cinema visits and on music CDs over 3 months. The receipts are then collected and totaled for each student. The students are classified by the age group they belong to: a student aged 15–18 years belongs to the age group 1; a student aged 22–25 years belongs to the age group 2. Here is the dataset:

age_group	expd_cinema	expd_music
1	65.22	68.02
1	72.13	83.77
1	58.69	55.96
1	66.72	90.13
1	64.38	70.54
1	81.29	82.43
2	45.08	55.15
2	60.09	61.12
2	33.22	39.75
2	59.67	57.09
2	18.39	26.88
2	22.82	33.64

We can use the Fisher–Pitman permutation test for paired replicates to determine the statistical significance, if any, of a subject's spending more money for movies or music.

```
. use permtest_example
. permtest1 expd_cinema = expd_music
Fisher-Pitman permutation test for paired replicates
```

difference vector	expd_cinema-expd_music
observations	12
- positive	2
- negative	10
- zero	0
critical value	-76.77999305725098

```
mode of operation:  exact (complete permutation)
Test of hypothesis Ho: expd_cinema>=expd_music : p = .00415039
Test of hypothesis Ho: expd_cinema<=expd_music : p = .99609375
Test of hypothesis Ho: expd_cinema==expd_music : p = .00830078
```

How can we interpret this result? The output states that the probability of falsely rejecting the null hypothesis that the expenditure for cinema visits is greater than or equal to the expenditure for music CDs is only 0.415%—this probability can be considered highly significant.

We can use the same dataset to demonstrate the permutation test for independent samples. Suppose that one wants to know if the expenditures for music CDs differ significantly between age groups. We can use the Fisher–Pitman permutation test for independent samples to investigate this research question.

```
. permtest2 expd_music, by(age_group)
Fisher-Pitman permutation test for two independent samples
```

age_group	obs	mean	std.dev.
1	6	75.141665	12.586086
2	6	45.605	14.084001
combined	12	60.373333	20.00247

```
mode of operation:  exact (complete permutation)
Test of hypothesis Ho: expd_music(age_group==1) >= expd_music(age_group==2) :
> p=.9978355 (one-tailed)
Test of hypothesis Ho: expd_music(age_group==1) <= expd_music(age_group==2) :
> p=.00324675 (one-tailed)
Test of hypothesis Ho: expd_music(age_group==1) == expd_music(age_group==2) :
> p=.00649351 (two-tailed)
```

The `permtest2` command claims that the probability of falsely rejecting the null hypothesis that the expenditure for music CDs in the first age group is lower than or equal to that for music CDs in the second age group is only 0.325%. Just as for the first case, this probability can be regarded highly significant.

5 Conclusion

In this article, I outlined the rationale of two powerful nonparametric tests. The Fisher–Pitman permutation test for paired replicates provides p-values for the difference in means of two outcomes of one subject in a sample, whereas the Fisher–Pitman permutation test for independent samples provides p-values for the difference in means of two independent groups. The complexity of the underlying algorithms requires approximating the p-values if the sample size is large; thus, we use a Monte Carlo simulation approach. A comparative simulation study demonstrates that the difference in p-values of exact and Monte Carlo approaches is small enough. I explained the usage of the supplemented code and then gave an example using a fictitious dataset.

There is still some room for future work. On one hand, calculating confidence intervals from both tests could be possible. One could extend the provided code to perform this task. On the other hand, the number of simulation runs to perform is set to a fixed value and can be modified manually. Doing so should not be necessary. One could derive a statistical proof for the number of permutations to test in a Monte Carlo situation to reduce the error term in p-values to a fixed minimum given the sample size. Nevertheless, the tests in their present design provide the scientist with a serviceable tool to investigate significance levels of differences in means.

6 References

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

Pitman, E. J. G. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society* 4: 119–130.

Siegel, S., and N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. 2nd ed. New York: McGraw–Hill.

About the author

Johannes Kaiser holds a German Diplom, the equivalent of a bachelor's plus a master's degree, in business engineering. His key interests include experimental macroeconomics, behavioral and experimental finance, and nonparametric statistics. Currently, he is employed as a research assistant and Ph.D. student of Reinhard Selten at the Laboratory for Experimental Economics, University of Bonn, Bonn, Germany.