Permutation Tests for More Than Two Samples

Ferry Butar Butar, Ph.D. †

Abstract

An *F* statistic is a classical test for the analysis of variance where the underlying distribution is a normal. For unspecified distributions, the permutation test is easy to utilize both for the original observations and ranks. If one knows a priori the direction of the alternative hypothesis then one can easily use a Resampling Stats to obtain a p-value. This test is easy to implement and the significance level is exact when calculating all possible permutations. The approximate significance level can be used when the numbers of permutations are very large.

Introduction

There are many fields do empirical and extensive use of experimentation, for example, the research in education, agricultural, medicine, engineering, industry and psychology. Statistical methods can increase the efficiency of these experiments and strengthens the conclusions so obtained. In order to the experimenter needs to use the simple statistical techniques, he needs to design the data as simple as possible.

One-way analysis of variance is one of the techniques to use if one wants to test that there is no difference between treatment 1, treatment 2, up to treatment *k*. If the underlying distribution of the observations is normal, then one usually uses an *F* statistic. Most practitioners hardly check whether their data come from a normal distribution, but they still force to use an *F* statistic. In order to use an *F* statistic, there are some assumptions need to satisfy such as the random sample for *k* treatments are mutually independent, a random sample comes from a normal distribution, and the variance for each of the treatments are equal and constant. Note that we observed many graduate students (or doctoral students) from education major here at our university use an *F* statistic even though their data are qualitative.

The permutation test does not need all of the assumptions above. Under the null hypothesis, there is no difference among k treatment effects. Thus any observations from one treatment can be permuted to any other treatments. This is called the data are exchangeable. For example, suppose there are two classes in statistics. Class A consists of 10 students and class B consists of 8 students. Then under null hypothesis that there is no difference between these two classes, any students in class A can sit in class B or vice versa. In permutation test you can relax to choose your own statistic and use it to perform the analysis. You can use your original data set as a basis for permutation test or you can use the ranks of the data. See (Good, 1994; Higgins, 2004; Hollander and Wolf, 1999).

Testing Hypothesis and Data Layout

Suppose observations are randomly selected from population with cumulative distribution function (cdf) $F_1(x)$, $F_2(x),...,F_k(x)$. The null hypothesis to be tested is equality of distribution, that is

$$H_0: F_1(x) = F_2(x) =...= F_k(x)$$

and the alternative hypothesis is given by

$$H_a: F_i(x) \leq F_j(x) \text{ or } F_i(x) \geq F_j(x)$$

with strict inequality hold for at least one x. The alternative of hypothesis can be thought as the shift locations of parameters $\mu_1, \mu_2,...,\mu_k$ not all equal so that

$$H_a: F_i(x) = F(x - \mu_i).$$

This shift can be modeled as the analysis of variance as

$$X_{ij} = \mu_i + \varepsilon_{ij},$$

where X_{ij} is the j^{th} response variable $j=1,...,n_i$ for the i^{th} treatment $i=1,...,k$ and μ_i are means of treatment i, and ε_{ij}'s are independent and identically distributed random variable with distribution $F(\varepsilon)$. The following table (I) is the layout of the observation.

Table I
One-Way Data Layout

Treatments	*Observations*	*Sample sizes*	*Means*	*Variances*
1	$X_{11}, X_{12},..., X_{1j},..., X_{1n_1}$	n_1	$\overline{X}_1$	S_1^2
2	$X_{12}, X_{22},..., X_{2j},..., X_{2n_2}$	n_2	$\overline{X}_2$	S_2^2
⋮	⋮	⋮	⋮	⋮
i	$X_{i1}, X_{i2},..., X_{ij},...,X_{in_i}$	n_i	$\overline{X}_i$	S_i^2
⋮	⋮	⋮	⋮	⋮
k	$X_{k1}, X_{k2},..., X_{kj},..., X_{kn_k}$	n_k	$\overline{X}_k$	S_k^2

Test Statistic

The t-test statistic is a classical test for comparing two populations under normality assumptions. And a classical test for comparing several populations (treatments) is an F test statistic. Based on Analysis of Variance (ANOVA)

techniques, we can partition the total sum of squares in terms of sum squares between treatments and sum squares within treatments, that is (see Kuehl, 1994)

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{X}_i-\bar{X}\right)^2$$

$$=\sum_{i=1}^{k}(n_i-1)S_i^{\,2}+\sum_{i=1}^{k}n_i\left(\bar{X}_i-\bar{X}\right)^2$$

$$=SSE+SST,$$

where $\bar{X}=N^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}$, $N=\sum_{i=1}^{k}n_{i,}$ $\bar{X}_i=n_i^{-1}\sum_{j=1}^{n_i}X_{ij.}$

The mean squares for treatment is

$$MST=SST/(k-1)=(k-1)^{-1}\sum_{i=1}^{k}n_i\left(\bar{X}_i-\bar{X}\right)^2$$

and the mean squares for error is

$$MSE=SSE/(N-k)=(N-k)^{-1}\sum_{i=1}^{k}(n_i-1)S_i^{\,2}.$$

A test statistic is given by F = MST/MSE. If the distribution of the observation is normal and the variance is constant, the test statistic *F* has an *F* distribution with degrees of freedom *k-1* for the numerator and *N-k* for the denominator. Using this distribution one can calculate its p-value.

In practice, we may not know or not assume that the observations come from a normal distribution. Since observations are independent and identically distributed, hence they are exchangeable, so we can employ the permutation distribution of *F*.

Permutation Tests

R. A Fisher (1935) was the first to introduce the idea of permuting data among treatments as a way of statistical inference. Since then a considerable theoreticians have developed this idea.

In the context of k-sample problems (k>2), the permutation tests is that all *k* samples were randomly selected from the same population and randomly assigned to label *1, 2, ...k*. Calculate the test statistic for the original data set, called it TS_{obs}. Then examine all possible permutation of the *N* observations among *k* treatments where there are n_i observations in treatment *i*, *i=1,2,...,k*. All possible permutations are $\frac{N!}{n_1!n_2!...n_k!}$. Calculate the statistic for each permutations. Now obtain the p-value as the fraction of statistic from all permutations calculated above that are greater than or equal to TS_{obs}. This p-value is exact proportion since we calculate from all possible permutations and

usually called as significance of the test. (See Butar Butar and Jae-Wan Park, 2007). If it is not possible to consider all possible permutations, then we randomly select sample of *R* permutations. In practice a random sample of 1000-2000 is usually sufficient to approximate a p-value. This procedure does not require analytical derivation of test statistic under the null hypothesis. There is a relaxation in choosing the test statistic. With this relaxation, this permutation test has advantageous over the parametric test (Hollander and Wolfe, 1999.

The proposed Formula

We would like to find a simple test statistic, easy to implement but powerful result. Recall that the total sum of squares (TSS) is $TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}\right)^2$. Note that TSS = SST + SSE. For every permutation, total sum of squares remains the same value, let's call this as a constant, C. Thus the F statistic can be written as

$$F = \frac{SST/(k-1)}{(C-SST)/(N-k)}.$$

Since F is an increasing function of SST, hence the test statistic can reduce based on SST and is equivalent to a test based on F statistic. Note that sum of squares treatment can be simplified to

$$SST = \sum_{i=1}^{k} n_i\left(\overline{X}_i - \overline{X}\right)^2 = \sum_{i=1}^{k} n_i \overline{X}_i^{\,2} - N\overline{X}^2,$$

which shorten the calculations. Also, since $\overline{X}$ has the same value for every permutation of the observations, we can use our test statistic which reduces to

$$SSX = \sum_{i=1}^{k} n_i \overline{X}_i^{\,2},$$

This test is again easy to implement on the Resampling Stats software (2000).

Example

Data were collected on student teachers relative to their use of certain teaching strategies that had been presented to them in preservice education. In 1978 there were 6 teachers who did not learn to use the strategies and they were used as a control group. There were 9 in 1979, 9 in 1980, and 10 in 1981 students teachers who had learned to use strategies. The investigator recorded the average number of strategies used per week by each of the student teachers during student teaching assignments. The hypothesis is the number of strategies used by the students were different among the years.

Table II
Average Number of Different Strategies Used

1978	1979	1980	1981
6.88	7.25	10.85	7.29
5.40	10.50	7.43	14.38
16.00	8.43	6.71	6.00
9.80	8.63	7.60	5.00
7.63	8.63	9.71	5.38
5.00	7.00	5.57	14.14
	11.13	8.71	9.25
	7.25	5.86	5.71
	10.38	7.20	7.35
			10.75

The permutation *F* and *SSX* test are implemented in Resampling Stats. The original *F* observed data is .2335. Based on 10000 randomly selected permutations, the p-value are 0.8783 and 0.8774, respectively. The 80th, 85th, 90th, 95th, 97.5th and 99th percentiles of the permutations were found to be 1.5978, 1.8481, 2.1947, 2.8445, 3.5365, 4.469, respectively. The statistic is not significant for any level.

The Kruskal-Wallis Test Statistic

The Kruskal-Wallis test use ranks instead of the original observations. Similar to the table 1, let R_{ij} be the rank of observation X_{ij}. Let $N=n_1+n_2+...+n_k$. To determine the ranks of all the observations we combine the sample sizes in k treatments. If all observations are distinct, let 1 be the rank of the smallest observation, 2 be the rank of the second smallest observation, and N be the rank of the largest observation. The sum of the rank is N (N+1)/2. The Kruskal-Wallis test statistic is similar to the *F* test statistic and is given by

$$KW = (S_R{}^2)^{-1} \sum_{i=1}^{k} n_i \left(\overline{R}_i - \overline{R}\right)^2, \qquad (1)$$

where $S_R{}^2$ is the sample variance of the combined ranks. Note that $\overline{R} = N^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} R_{ij} = \frac{N+1}{2}$, and

$$
\begin{aligned}
S_R{}^2 &= (N-1)^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(R_{ij}-\bar{R}\right)^2 \\
&= (N-1)^{-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R_{ij}{}^2 - \frac{N(N+1)^2}{4}\right) \\
&= \frac{N(N+1)(2N+1)}{6(N-1)} - \frac{N(N+1)^2}{4(N-1)} \\
&= \frac{N(N+1)}{12(N-1)}\left(2(2N+1)-3(N+1)\right) \\
&= \frac{N(N+1)}{12} \qquad (2)
\end{aligned}
$$

Thus, using (2), equation (1) becomes

$$KW = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\left(\bar{R}_i - \frac{N+1}{2}\right)^2 \qquad (3)$$

We can find a p-value for the KW statistic from the permutation distribution of this statistic. This can be done which is similar to the *F* statistic, except that we use rank instead of the original observations. Note that for large sample sizes, $\sum_{i=1}^{k} n_i\left(\bar{R}_i - \frac{N+1}{2}\right)^2$ can be approximated by χ^2 distribution with *k-1* degrees of freedom. We can also simplify $\sum_{i=1}^{k} n_i\left(\bar{R}_i - \frac{N+1}{2}\right)^2$ becomes $\sum_{i=1}^{k} n_i \bar{R}_i{}^2 - \frac{N(N+1)^2}{4}$. Thus, the proposed test statistic is given by

$$KW = \sum_{i=1}^{k} n_i \bar{R}_i{}^2$$

If two or more observations have same values or ties, then we use the average ranks of ties and assigned the average rank to each of the tie observations. For example, the original observations are: 43 32 32 56. Since two values has the same number 32, assign 1 and 2 ranking to them and average those ranks and assign to each of 32, thus final rankings are: 3, 1.5, 1.5, 4.

Comparison between F and Permutation Distributions

Consider four treatments generated from normal distribution. Observations for treatment 1, 2, 3, and 4 were generated from normal distribution with mean

10, 20, 30 and 40 with the same standard deviation of 10. The following table is the observations generated from the above statement.

Table III
Samples from Normal population, μ_1=10, μ_2 =20, μ_3 = 30, μ_4 =40, σ=10

Treatment 1	14.627	6.1787	10.503	28.41	9.8445
Treatment 2	27.325	23.93	28.238	30.352	24.856
Treatment 3	26.433	59.702	38.337	23.775	39.261
Treatment 4	39.682	39.113	66.574	25.496	40.416

Using the Resampling Stats software, we randomly selected 10000 permutations and obtain the permutation distribution. The following table compares the percentiles of the F distribution with 3 degrees of freedom for numerator and 16 degrees of freedom for denominator to the permutation distributions.

Table IV
Comparison of Permutation Percentiles to F-distribution Percentiles

Percentile	80	85	90	95	97.5	99
Permutation Distribution	1.72	2.02	2.43	3.21	4.02	4.98
F-distribution	1.74	2.03	2.46	3.24	4.08	5.29

From the table which has only five observations per treatment, we can conclude that there is not much different between the values of F distributions and the permutation distributions.

Ordered Alternatives Hypothesis

It is possible researchers have prior belief that the observations from treatment 1 tend to be smaller than observations from treatment 2, and so on. For example quality or quantity of materials, severity of disease, drug dosage levels, and temperature. The Kruskal-Wallis statistic does not employ any such prior information regarding an alternative hypothesis. The Jonckheere-Terpstra statistic is more powerful than the Kruskal-Wallis statistic when the treatments can be ordered a priori in particular direction. Let θ_i be the population median for the i^{th} treatment. The null hypothesis can be written as

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$$

versus the alternative hypothesis

$$H1: \theta_1 \le \theta_2 \le ... \le \theta_k ,$$

where at least one of the alternatives is strict inequality. Note that the ordering of the hypothesis must be specified before the observations are collected. In terms of the cdf's, let $F_i(x)$ be the cdf of treatment *i*, then the alternative hypothesis can be written as

$$H_1 : F_1(x) \ge F_2(x) \ge ... \ge F_k(x).$$

Let the observations from treatment 1 be denoted $X_{11},X_{12},...,X_{1n1}$ and the observations from treatment 2 be denoted $X_{21},X_{22},...,X_{2n2}$. If the data have no ties, then any given observations are either strictly less than or strictly greater than any other observation. The *Mann-Whitney* statistic , denoted U_{ij} is defined as

$$U_{ij} = number\ of\ \ pairs(X_{1i,} X_{2i})\ \ for\ which\ X_{1i} < X_{2i}.$$

The Jonckheere-Terpstra statistic, *JT*, is given by

$$JT = \sum_{i<j} U_{ij} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}.$$

To compute a p-value from the permutations test, first obtain the JT_{obs} from the observed data. Second, if possible examine all possible permutations of the data among the *k* treatments. If this not feasible examine a randomly selected subset of the permutations. In practice no more than 2000 is enough for a simulation. Record a 1 if a statistic you just found greater or equal to the JT_{obs}. Last, the upper-tail p-value is the total number of 1's divided by number of permutations. Note that one can also use the *Wilcoxon rank-sum* statistic for calculating U_{ij}.

Example

The following data from the National Transportation Safety Administration are left femur loads on driver-side crash dummies for automobiles in various weight classes. We want to test whether the left femur loads on driver-side are increasing order of vehicle weight classification from the weight of 1700 lb to weight of 3700 lb. Data source: http://lib.stat.cmu.edu/DASL/. The permutation test based on Wilcoxon rank sum test was applied to the JT statistic to the above table. The value of the JT statistic based on the original data was 1355.5. Based on 5000 randomly selected permutations, the 80^{th}, 85^{th}, 90^{th}, 97.5^{th}, and 99^{th} percentiles of the permutation distribution found to be 1351.5, 1364.5, 1381.5, 1407.5, 1431.5, 1452.3, respectively. Therefore, the statistic is significant only at the 20% level. An approximate permutation p-value is 0.188.

Table V
Vehicle Weight Classification

1700 lb	2300 lb	2800 lb	3200 lb	3700 lb
574	791	865	998	1154
976	1146	775	1049	541
789	394	729	736	406
805	767	1721	782	1529
361	1385	1113	730	1132
529	1021	820	742	767
	2073	1613	1219	1224
	803	1404	705	314
	1263	1201	1260	1278
	1016	205	611	
	1101	1380	1350	
	945	580	1650	
	139	1803	1143	

Conclusion

The percentile for a permutation distribution is comparable to the *F* distribution in one-way analysis of variance. This permutation test can be used as an approximate to an *F* test especially when the sample size is large. One can simplify the formula to reduce the calculation time. The Kruskal-Wallis is the test which is similar to an *F* test with the ranks used instead of the original observations. If one knows the order of alternative before collecting the data, then the more powerful test is the Jonckheere-Terpstra test. The Resampling Stats can be use to perform all of the methods above.

† *Ferry Butar Butar, Ph.D.*, Sam Houston State University, USA

References

Butar Butar, F, and Jea-Wan Park (2007). Permutation tests for comparing two populations, Technical report, Sam Houston State University.

Conover, W. J. (1999). Practical Nonparametric Statistics, third ed., John Wiley & Sons, New York.

Good, P. (1994). Permutation Tests: A practical guide to resampling methods for testing hypothesis, Springer, New York.

Hicks, C.R., Turner, K.V. (1999). Fundamental Concepts in the Design of Experiments, Fifth ed., Oxford University Press, New York.

Higgins, J.J. (2004). Introduction to Modern Nonparametric Statistics, Duxbury.

Hollander, M., and Wolfe, D. A. (1999). Nonparametric Statistical Methods, Second ed., Wiley, New York.

Johnson, R. A., and Wichern, D.W. (2000). Applied Multivariate Statistical Analysis, fourth ed., Prentice Hall, New Jersey.
Kuehl, R. O. (1994). Statistical Principles of Research Design and Analysis, Duxbury.
Resampling Stats User's Guide. (2000). Resampling Stats, Inc., Vol 5.0.2, Arlington, VA.
Siegel, S. and Castellan, N.J. (1998). Nonparametric Statistics for the Behavioral Sciences, second ed. McGraw Hill, Boston.