Journal of Modern Applied Statistical Methods

Volume 8 | Issue 2 Article 2

11-1-2009

Analysis of MultiFactor Experimental Designs

Phillip I. Good
Information Research, Huntington Beach, CA, drgood@statcourse.com

Journal of Modern Applied Statistical Methods
November, 2009, Vol. 8, No. 2, 355-359

1538 – 9472/09/$95.00

INVITED ARTICLE

Analysis of MultiFactor Experimental Designs

Phillip Good
Information Research
Huntington Beach, CA.

In the one-factor case, Good and Lunneborg (2006) showed that the permutation test is superior to the analysis of variance. In the multi-factor case, simulations reveal the reverse is true. The analysis of variance is remarkably robust against departures from normality including instances in which data is drawn from mixtures of normal distributions or from Weibull distributions. The traditional permutation test based on all rearrangements of the data labels is not exact and is more powerful that the analysis of variance only for 2xC designs or when there is only a single significant effect. Permutation tests restricted to synchronized permutations are exact, but lack power.

Key words: analysis of variance, permutation tests, synchronized permutations, exact tests, robust tests, two-way experimental designs.

Introduction

Tests of hypotheses in a multifactor analysis of variance (ANOVA) are not independent of one another and may not be most powerful. These tests are derived in two steps: First, the between-cell sum of squares is resolved into orthogonal components. Next, to obtain p- values, the orthogonal components are divided by the within-cell sum of squares. As they share a common denominator, the test statistics of main effects and interactions are *not* independent of one another. On the plus side, Jagers (1980) showed that if the residual errors in the linear model are independent and identically distributed, then the distribution of the resultant ratios is closely approximated by an F-distribution even if the residual errors are *not* normally distributed. As a result, ANOVA p-values are almost exact.

But are ANOVA tests the most powerful? In the one-way design (the one-factor case), Good and Lunneborg (2005) found that tests whose p-values are based on the permutation distribution of the F-statistic rather than the F-distribution are both exact and more

Phillip Good is a statistical consultant. He authored numerous books that include *Introduction to Statistics via Resampling Methods and R/S-PLUS* and *Common Errors in Statistics (and How to Avoid Them)*. Email: drgood@statcourse.com.

powerful than the analysis of variance when samples are taken from non-normal distributions. For example, when the data in a four-sample, one-factor comparison are drawn from mixtures of normal distributions, 50% N(δ, 1) and 50% N(1+δ, 1), in an unbalanced design with 2, 3, 3, and 4 observations per cell, the permutation test was more powerful at the 10% level, a power of 86% against a shift in means of two units compared to 65% for the analysis of variance.

Unfortunately, the permutation test for interaction in a two-factor experimental design based on the set of all possible rearrangements among the cells is not exact. The residual errors are not exchangeable, nor are the p-values of such permutation tests for main effects and interactions independent of one another. Here is why:

Suppose the observations satisfy a linear model, $X_{ijm} = \mu + s_i + r_j + (sr)_{ij} + \varepsilon_{ijm}$ where the residual errors $\{\varepsilon_{ijm}\}$ are independent and identically distributed. To test the hypothesis of no interaction, first eliminate row and column effects by subtracting the row and column means from the original observations. That is, set

$$X'_{ijk} = X_{ijk} - \bar{X}_{i\_} - \bar{X}_{.j.} + \bar{X}_{\_\_},$$

where by adding the grand mean $\bar{X}_{\_}$, ensure the overall sum will be zero. Recall that

$$X'_{ijk} = X_{ijk} - \bar{X}_{i\_} - \bar{X}_{.j.} + \bar{X}_{\_}$$

or, in terms of the original linear model, that

$$X'_{ijk} = \varepsilon_{ijk} - \bar{\varepsilon}_{i\_} - \bar{\varepsilon}_{.j.} + \bar{\varepsilon}_{\_}.$$

However, this means that two residuals in the same row such as X'_{i11} and X'_{i23} will be correlated while residuals taken from different rows and columns will not be. Thus, the residuals are not exchangeable, a necessary requirement for tests based on a permutation distribution to be exact and independent of one another (see, for example, Good, 2002).

An alternative approach, first advanced by Salmaso and later published by Pesarin (2001) and Good (2002), is to restrict the permutation set to synchronized permutations in which, for example, an exchange between rows in one column is duplicated by exchanges between the same rows in all the other columns so as to preserve the exchangeability of the residuals.

The purpose of this article is to compare the power of ANOVA tests with those of permutation tests (synchronized and unsynchronized) when applied to two-factor experimental designs.

Methodology

Observations were drawn from one of the following three distributions:

1. Normal.
2. Weibull, because such distributions arise in reliability and survival analysis and cannot be readily transformed to normal distributions. A shape parameter of 1.5 was specified.
3. Contaminated normal, both because such mixtures of distributions are common in practice and because they cannot be readily transformed to normal distributions. In line with findings in an earlier article in this series, Good and Lunneborg (2006), we focused on the worst case distribution, a mixture of 70% N(0, 1) and 30% N(2, 2) observations.

Designs with the following effects were studied:

a)

+δ 0

+δ 0

b)

+δ 0

0 +δ

c)

+δ 0 … -δ

+δ 0 … - δ

d)

+δ 0 … 0 -δ

-δ 0 … 0 +δ

e)

0 +δ 0

0 +δ 0

0 +δ 0

f)

0 +δ 0

0 0 0

+δ 0 0

g)

+δ 0 0

0 +δ 0

0 0 +δ

h)

0 +δ 0 0 -δ

0 +δ 0 0 –δ

0 +δ 0 0 –δ

0 +δ 0 0 –δ

i)

1 δ 1 1 δ

1 δ 1 1 δ

1 δ 1 1 δ

1 δ 1 1 δ

To compare the results of the three methodologies, 1,000 data sets were generated at random for each design and each alternative (δ =0, 1, or 2). p-values for the permutation tests were obtained by Monte Carlo means using a minimum of 400 random (synchronized or unsynchronized) permutations per data set. The alpha level was set at 10%. (The exception being the 2x2 designs with 3 observations per cell where the highly discrete nature of the synchronized permutation distribution forced adoption of an 11.2% level.)

The simulations were programmed in R. Test results for the analysis of variance were derived using the anres() function. R code for the permutation tests and the data generators is posted at: http://statcourse.com/AnovPower.txt.

Results

Summary

In line with Jager's (1980) theoretical results, the analysis of variance (ANOV) applied to RxC experimental designs was found to yield almost exact tests even when data are drawn from mixtures of normal populations or from a Weibull distribution. This result holds whether the design is balanced or unbalanced. Of course, because the ANOVA tests for main effects and interaction share a common denominator - the within sum of squares - the resultant p-values are positively correlated. Thus a real non-zero main effect may be obscured by the presence of a spuriously significant interaction.

Although tests based on synchronized permutations are both exact and independent of one another, there are so few synchronized permutations with small samples that these tests lack power. For example, in a 2x2 design with 3 observations per cell, there are only 9 distinct values of each of the test statistics.

Fortunately, tests based on the entire set of permutations, unsynchronized as well as synchronized, prove to be almost exact. Moreover, these permutation tests for main effects and interaction are negatively correlated. The result is an increase in power if only one effect is present, but a loss in power if there are multiple effects. These permutation tests are more powerful than ANOVA tests when the data are drawn from mixtures of normal populations or from a Weibull distribution. They are as powerful, even with data drawn from normal distributions, with samples of $n \geq 5$ per cell.

2xK Design

In a 2x2 design with 3 observations per cell, restricting the permutation distribution to synchronized permutations means there are only 9 distinct values of each of the test statistics. The resultant tests lack power as do the tests based on synchronized permutations for 2x5 designs with as many as five observations per cell. For example, in a 2x4 design with four observations per cell, the synchronized permutation test had a power of 53% against a shift of two units when the data were drawn from a contaminated normal, while the power of the equivalent ANOVA test was 61%. As a result of these

negative findings, synchronized permutation tests were eliminated from further consideration.

In a balanced 2x2 design with 5 observations per cell, the powers of the ANOVA test and the traditional permutation test against a normal are equivalent. Against a contaminated normal or Weibull alternative, the permutation test is fractionally better. With only 3 observations per cell and a Weibull alternative with a doubling of scale, the permutation test is again fractionally superior.

In an unbalanced 2x2 design with 5 observations in each cell of the first column, and 3 observations in each cell of the second column, against a normal with a column effect of one unit (design a), ANOVA is markedly inferior with a power of 60% versus a power of 70% for the permutation test. Against a Weibull alternative with a doubling of the scale factor, the power of the ANOVA is 56%, while that of the permutation test is 71%. Noteworthy in this latter instance is that although there is no interaction term in design a, spurious interaction was recorded 18% of the time by the analysis of variance and 13% by permutation methods.

In a 2x5 design of form c with 3 observations per cell, the permutation test is several percentage points more powerful than ANOVA against both normal and contaminated normal alternatives.

3x3 Designs

When row, column, and interactions are all present as in design f, ANOVA is more powerful than the permutation test by several percentage points for all effects against both normal and contaminated normal alternatives. (See Table 1a, b.)

Table1a: Normal Alternative $\delta = 1$, 3 Observations Per Cell, Design f

Row-Column Interaction	
ANOVA	Permutation
187	139
178	138
344	316

Table 1b: Contaminated Normal Alternative $\delta = 2$, 3 Observations Per Cell, Design f

Row-Column Interaction	
ANOVA	Permutation
150	114
169	137
336	318

However, when a pure column effect (design e) or a pure interaction (design g) exists, the permutation test is superior to the analysis of variance by several percentage points. See, for example, Table 2.

Table 2: Contaminated Normal Alternative $\delta = 2$, 3 Observations Per Cell, Design g

Row-Column Interaction	
ANOVA	Permutation
115	70
108	70
461	529

4x5 Designs

The power against balanced designs of type h with four observations per cell of permutation and ANOVA tests are equivalent when the data is drawn from a normal distribution. The power of the permutation test is fractionally superior when the data is drawn from a mixed-normal distribution. Likewise, with a design of type i, the permutation test is several percentage points superior when the data is drawn from a Weibull distribution and the design is balanced. Synchronized permutations fared worst of all, their power being several percentage points below that provided by the analysis of variance.

When the design is unbalanced as in

4	4	4	4	4
4	4	4	4	4
2	3	4	5	3
2	3	4	5	3

the analysis of variance has the advantage in power over the permutation tests by several percentage points.

Discussion

Apart from 2xC designs, there appears to be little advantage to performing alternatives to the standard analysis of variance. The permutation tests are more powerful if only a single effect is present, but how often can this be guaranteed? Even with 2xC designs, the results reported here will be of little practical value until and unless permutation methods are incorporated in standard commercial packages. Wheeler suggests in a personal communication that if a package possesses a macro-language, a vector permutation command and an ANOVA routine, a permutation test for the multi-factor design can be readily assembled as follows:

1. Use the ANOVA command applied to the original data set to generate the sums of squares used in the denominators of the tests of the various effects.
2. Set up a loop and perform the following steps repeatedly:
    a. Rearrange the data.
    b. Use the ANOVA command applied to the rearranged data set to generate the sums of squares used in the denominators of the tests of the various effects.
    c. Compare these sums with the sums for the original data set.
3. Record the p-values as the percentage of rearrangements in which the new sum equaled or exceeded the value of the original.

References

David, F. N., & Johnson, N. L. (1951). The effect of non-normality on the power function of the f-test in the analysis of variance. *Biometrika*, 38, 43-57.

Good, P. (2002). Extensions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, *1*, 243-247.

Good, P. (2005). *Permutation, parametric, and bootstrap tests of hypotheses* (*3rd Ed.*). NY: Springer.

Good, P., & Lunneborg, C. E. (2005). Limitations of the analysis of variance. I: The one-way design. *Journal of Modern Applied Statistical Methods*, *5*(*1*), 41-43.

Jagers, P. (1980). Invariance in the linear model—an argument for chi-square and f in nonnormal situations. *Mathematische Operationsforschung und Statistik*, *11*, 455-464.

Pesarin, F. (2001). *Multivariate permutation tests*. NY: Wiley.

Salmaso, L. (2003). Synchronized permutation tests in 2^k factorial designs. *Communications in Statistics - Theory and Methods*, *32*, 1419-1438.