Journal of Modern Applied Statistical Methods

Volume 4 | Issue 1 Article 2

5-1-2005

Within By Within ANOVA Based On Medians

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

DOI: 10.22237/jmasm/1114905720

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No. 1, 2-10

1538 – 9472/05/$95.00

INVITED ARTICLE
Within By Within ANOVA Based On Medians

Rand R. Wilcox
Department of Psychology
University of Southern California, Los Angeles

This article considers a *J* by *K* ANOVA design where all *JK* groups are dependent and where groups are to be compared based on medians. Two general approaches are considered. The first is based on an omnibus test for no main effects and no interactions and the other tests each member of a collection of relevant linear contrasts. Based on an earlier paper dealing with multiple comparisons, an obvious speculation is that a particular bootstrap method should be used. One of the main points here is that, in general, this is not the case for the problem at hand. The second main result is that, in terms of Type I errors, the second approach, where multiple hypotheses are tested based on relevant linear contrasts, performs about as well or better than the omnibus method, and in some cases it offers a distinct advantage.

Keywords: Repeated measures designs, robust methods, kernel density estimators, bootstrap methods, linear contrasts, multiple comparisons, familywise error rate

Introduction

Consider a J by K ANOVA design where all JK groups are dependent. Let θ_{jk} (j =1,...J; k =1,...K) represent the (population) medians corresponding to these JK groups. This article is concerned with two strategies for dealing with main effects and interactions. The first is to perform an omnibus test for no main effects and no interactions by testing

$$Ho: C\theta = 0, \tag{1}$$

Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology at the University of Southern California, Los Angeles.

where θ is a column vector containing the *JK* elements θ_{jk} ,and C is an ℓ by *JK* matrix (having rank ℓ) that reflects the null hypothesis of interest. (The first *K* elements of θ are $\theta_{11}, \ldots, \theta_{1K}$, the next *K* elements are $\theta_{21}, \ldots, \theta_{2K}$, and so forth.) The second approach uses a collection of linear contrasts, rather than a single omnibus test, and now the goal is to control the probability of at least one Type I error.

A search of the literature indicates that there are very few results on comparing the medians of dependent groups using a direct estimate of the medians of the marginal distributions, and there are no results for the situation at hand. In an earlier article (Wilcox, 2004), two methods were considered for performing all pairwise comparisons among a collection of dependent groups. The first uses an estimate of the appropriate standard error stemming from the influence function of a single order statistic. The second method uses the usual sample median in conjunction with a bootstrap estimate of the standard error. The bootstrap method performed quite well in simulations in terms of controlling the probability of at least one Type I error.

Recently, Dawson, Schell, Rissling and Wilcox (2004) dealt with an applied study where a two-way ANOVA design was used with all JK groups dependent. An issue is whether the results in Wilcox (2004) extend to this two-way design. One of the main results here is that the answer is no. The other main result deals with the choice between an omnibus test versus performing multiple comparisons where each hypothesis corresponding to a collection of relevant linear contrasts is to be tested. It is found that simply ignoring the omnibus test, and performing the relevant multiple comparisons, has practical value.

Some Preliminaries

For convenience, momentarily consider a single random sample $X_1,\ldots,X_n$ and for any q, $0<q<1$, suppose the q^{th} quantile, x_q, is estimated with $X_{(m)}$, where $m=[qn+.5]$ and [.] is the greatest integer function. Then, ignoring an error term, which goes to zero as $n \to \infty$,

$$X_{(m)} = x_q + \frac{1}{n}\sum IF_q(X_i), \qquad (2)$$

where

$$IF_q(x) = \begin{cases} \dfrac{q-1}{f(x_q)}, & \text{if } x < x_q \\ 0, & \text{if } x = x_q \\ \dfrac{q}{f(x_q)}, & \text{if } x > x_q, \end{cases}$$

(Bahadur, 1966; also see Staudte & Sheather, 1990).

Now consider the situation where sampling is from a bivariate distribution. Let X_{ik} (i=1,...,n; k=1, 2) be a random sample of n vectors. Let $X_{(1)k} \le \cdots \le X_{(n)k}$ be the observations associated with k^{th} variable written in ascending order. Two estimates of the population median are relevant here. The first is

$$\hat{\theta}_k = X_{(m)k},$$

where again $m=[.5n+.5]$, and the other is $\hat{\theta}_j = M_k$, the usual sample median based on $X_{1k},\ldots,X_{nk}$.

Although the focus is on estimating the median with q=.5, the results given here apply to any q, $0<q<1$. Let f_k be the marginal density of the k^{th} variable and let

$$V_1 = (q-1)^2 P(X_1 \le x_{q1}, X_2 \le x_{q2}),$$
$$V_2 = q(q-1)P(X_1 \le x_{q1}, X_2 > x_{q2}),$$
$$V_3 = q(q-1)P(X_1 > x_{q1}, X_2 \le x_{q2}),$$

and

$$V_4 = q^2 P(X_1 > x_{q1}, X_2 > x_{q2}),$$

where x_{q1} and x_{q2} are the q^{th} quantiles corresponding to the first and second marginal distributions, respectively. Then for the general case where $m=[qn+.5]$, a straightforward derivation based on equation (2) yields an expression for the covariance between $X_{(m)1}$ and $X_{(m)2}$:

$$\tau_{12}^2 = \frac{V_1+V_2+V_3+V_4}{nf_1(x_{q1})\, f_2(x_{q2})}. \qquad (3)$$

Also, (2) yields a well-known expression for the squared standard error of $X_{(m)1}$, namely,

$$\tau_{11}^2 = \frac{1}{n}\frac{q(1-q)}{f^2_1(x_{q1})}.$$

Using (3) to estimate τ_{12}^2 requires an estimate of the marginal densities. Here, a variation of an adaptive kernel density estimator is used (e.g., Silverman, 1986), which is based in part on an initial estimate obtained via a so-called expected frequency curve (e.g., Wilcox, 2005; cf. Davies & Kovac, 2004). To elaborate, let MAD_k be the median absolute deviation associated with the kth marginal distribution, which is the median of the values $|X_{1k} - M_k|, \ldots, |X_{nk} - M_k|$. For some constant κ to be determined, the point x is said to be close to X_{ik} if

$$|X_{ik} - x| \le \kappa \times \frac{MAD_k}{.6745}.$$

Under normality, $\mathrm{MADN}_k = MAD_k/.6745$ estimates the standard deviation, in which case x is close to X_{ik} if x is within κ standard deviations of X_{ik}. Let

$$N_k(x) = \{i : |X_{ik} - x| \le \kappa \times MADN_k\}.$$

That is, $N_k(x)$ indexes the set of all X_{ik} values that are close to x. Then an initial estimate of $f_k(x)$ is taken to be

$$\tilde{f}_k(x) = \frac{1}{2\kappa MADN_k} \sum I_{i \in N_k(x)},$$

where I is the indicator function. Here, κ=.8 is used.

The adaptive kernel density estimate is computed as follows. Let

$$\log g = \frac{1}{n} \sum \log \tilde{f}_\kappa(X_i)$$

and

$$\lambda_i = (\tilde{f}_k(X_{ik})/g)^{-a},$$

where a is a sensitivity parameter satisfying $0 \le a \le 1$. Based on comments by Silverman (1986), a=.5 is used. Then the adaptive kernel estimate of f_k is taken to be

$$\tilde{f}_\kappa(x) = \frac{1}{n} \sum \frac{1}{h\lambda_i} K\left\{h^{-1}\lambda_i^{-1}(x - X_i)\right\},$$

where

$$\begin{aligned} K(t) &= \frac{3}{4}\left(1 - \frac{1}{5}t^2\right)/\sqrt{5}, \quad |t| < \sqrt{5} \\ &= 0, \quad \text{otherwise,} \end{aligned}$$

is the Epanechnikov kernel, and following Silverman (1986, p. 47 – 48), the span is

$$h = 1.06 \frac{A}{n^{1/5}},$$

$$A = min(s, IQR/1.34),$$

and where s is the standard deviation and IQR is the interquartile range based on $X_{1k}, \ldots, X_{nk}$.

Here, IQR is estimated via the ideal fourths. Let $\ell = [(n/4) + (5/12)]$. That is, ℓ is $(n/4)+(5/12)$ rounded down to the nearest integer. Let

$$h = \frac{n}{4} + \frac{5}{12} - l.$$

Then the estimate of the .25 quantile is given by

$$q_1 = (1-h)X_{(\ell)} + hX_{(\ell+1)}. \qquad (4)$$

Letting $\ell' = n - \ell + 1$, the estimate of the upper quartile, is

$$q_2 = (1-h)X_{(\ell')} + hX_{(\ell-1)} \qquad (5)$$

and the estimate of the interquartile range is

$$IQR = q_2 - q_1.$$

All that remains is estimating V_1, V_2, V_3 and V_4. An estimate of V_1 is obtained once an estimate of $P(X_1 \le x_{q1}, X_2 \le x_{q2})$ is available. The obvious estimate of this last quantity, and the one used here, is the proportion of times these inequalities are true among the sample of observations. That is, let A_{i12}=1 if simultaneously $X_{i1} \le X_{(m)1}$ and $X_{i2} \le X_{(m)2}$, otherwise A_{i12}=0. Then an estimate of V_1 is simply

$$\hat{V}_1 = \frac{(q-1)^2}{n} \sum A_{i12}.$$

Estimates of V_2, V_3 and V_4 are obtained in a similar manner. The resulting estimate of the covariance between $X_{(m)1}$ and $X_{(m)2}$ is labeled $\hat{\tau}^2_{12}$. Of course, the squared standard error of $X_{(m)1}$ can be estimated in a similar fashion and is labeled $\hat{\tau}^2_{11}$.

An alternative approach is to use a bootstrap method, a possible appeal of which is that the usual sample median can be used when *n* is even. Generate a bootstrap sample by resampling with replacement *n* pairs of values from X_{ik} yielding X^*_{ik} (i=1,...,n; k=1, 2). For fixed k, let M^*_k be the usual sample median based on the bootstrap sample and corresponding to the kth marginal distribution. Repeat this B times yielding M^*_{bk}, b=1,...,B. Then an estimate of the covariance between M_1 and M_2 is

$$\hat{\xi}_{12} = \frac{1}{B-1} \sum (M^*_{i1} - \bar{M}_1)(M^*_{i2} - \bar{M}_2),$$

where $$\bar{M}_k = \sum M^*_{bk} / B.$$

Methodology

Now consider the more general case of a *J* by *K* design and suppose (1) is to be tested. Based on the results in the previous section, two test statistics are considered. The first estimates the population medians with a single order statistic, $X_{(m)}$, and the second uses the usual sample median, *M*.

Let X_{ijk} be a random sample of n_j vectors of observations from the jth group ($i = 1,\ldots,n_j$; j=1,...,J; k=1,...,K).

Let $\hat{\theta}_{jk} = X_{(m)jk}$ be the estimate of the median for the jth level of first factor and the kth level of the second. Then a test statistic for (1) can be developed along the lines used to derive the test statistic based on trimmed means, which is described in Wilcox (2003, section 11.9).

For convenience, let $\hat{\Theta}' = (\hat{\theta}_{11},\ldots,\hat{\theta}_{JK})$. For fixed j, k and ℓ, $k \ne \ell$, let $v_{jk\ell}$ be the estimated covariance between θ_{jk} and $\theta_{j\ell}$. That is, $v_{jk\ell}$ is computed in the same manner as $\hat{\tau}^2_{12}$, only now use the data X_{ijk} and $X_{ij l}$, $i = 1,\ldots,n_j$. When $k=\ell$, $v_{jk\ell}$ is the estimated squared standard error of $\hat{\theta}_{jk}$. Let $\boldsymbol{V}$ be the K by K matrix where the element in the Kth row and ℓth column is given by $v_{jk\ell}$. The test statistic is

$$Q = \Theta' \mathbf{C}' (\mathbf{C}\mathbf{V}\mathbf{C}')^{-1} \mathbf{C}\Theta. \qquad (6)$$

As is well known, the usual choices for $\boldsymbol{C}$ for main effects for Factor A, main effects for Factor B, and for interactions are $\boldsymbol{C}=\boldsymbol{C}_J \otimes \boldsymbol{j}'_K$, $\mathbf{C} = \mathbf{j}'_J \otimes \mathbf{C}_K$ and $\boldsymbol{C}=\boldsymbol{C}_J \otimes \boldsymbol{C}_K$, respectively, where $\boldsymbol{C}_J$ is a J-1 by J matrix having the form

$$\begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & -1 & 0 & \ldots & 0 \\ & & & \vdots & & \\ 0 & 0 & \ldots & 0 & 1 & -1 \end{pmatrix},$$

$\boldsymbol{j}_J'$ is a $1\times J$ matrix of ones and $\otimes$ is the (right) Kronecker product.

There remains the problem of approximating the null distribution of Q. Based on results in Wilcox (2003, chapter 11) when comparing groups using a 20% trimmed mean, an obvious speculation is that Q has, approximately, an F distribution with ν_1 and ν_2 degrees of freedom. For main effects for Factor A, main effects for Factor B, and for interactions, ν_1 is equal to J-1, K-1 and (J-1)(K-1), respectively. As for ν_2, it is estimated based on the data, but an analog of this method for medians was not quite satisfactory in simulations; the actual probability of a Type I error was too far below the nominal level. A better approach was simply to take $\nu_2 = \infty$, which will be assumed henceforth. This will be called method A.

An alternative approach is to proceed exactly as in method A, only estimate the .5 quantiles with the usual sample median and replace $\boldsymbol{V}_j$ with the bootstrap estimate described in section 2. (Here, B=100 is used.) This will be called method B.

An Approached Based on Linear Contrasts

Another approach to analyzing the two-way ANOVA design under consideration is to test hypotheses about a collection of linear contrasts appropriate for studying main effects and interactions. Consider, for example,

$$\hat{\Psi}_j = \sum \hat{\theta}_{jk}.$$

$j = 1, \ldots, J$. Then when dealing with main effects for Factor A, one could perform all pairwise comparisons among the Ψ_j. This is for every $j < j'$,

$$H_0 : \Psi_j = \Psi_{j'}.$$

There is the problem of controlling the probability of at least one Type I error among the $(J^2 - J)/2$ hypotheses to be tested, and here this is done with a method derived by Rom (1990). Interactions can be studied by testing hypotheses about all of the relevant $(J^2 - J)(K^2 - K)/4$ tetrad differences, and of course, main effects for Factor B can be handled in a similar manner.

For convenience, attention is focused on Factor A (the first factor). Here, Ψ_j is simply estimated with

$$\hat{\Psi}_j = \sum \hat{\theta}_{jk}.$$

Writing

$$\hat{\Psi}_j - \hat{\Psi}_{j'} = \sum \sum c_{jk} \hat{\theta}_{jk}$$

for appropriately chosen contrast coefficents c_{jk}, then of course an estimate of the squared standard error of $\hat{\Psi}_j - \hat{\Psi}_{j'}$ is

$$\hat{n}^2 = \sum \sum c_{jk} \hat{\tau}_{jk},$$

Based on results in Wilcox (2004), the null distribution of T is approximated with a Student's T distribution with $n-1$ degrees of freedom.

To elaborate on controlling the probability of at least one Type I error with Rom's method, and still focusing on Factor A, let $D = (J^2 - J)/2$ be the number of hypotheses to be tested and let $P_1, \ldots P_D$ be the corresponding p-values. Put the p-values in descending order yielding $P_{[1]} \geq P_{[2]} \geq \ldots P_{[D]}$.

Proceed as follows:

1. Set $\ell = 1$.

2. If $P_{[\ell]} \le d_\ell$, where d_ℓ is read from Table 1, stop and reject all *D* hypotheses; otherwise, go to step 3 (If $\ell > 10$, use $d_\ell = \alpha / \ell$).

3. Increment ℓ by 1. If $P_{[\ell]} \le d_{\ell,}$, stop and reject all hypotheses having a significance level less than or equal d_ℓ.

4. If $P_{[\ell]} > d_\ell$, repeat step 3.

5. Continue until a significant result is obtained or all *D* hypotheses have been tested.

A Simulation Study

Simulations were used to study the small-sample properties of the methods just described. Vectors of observations were generated from multivariate normal distributions having a common correlation, ρ. To study the effect of non-normality, observations were transformed to various g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If *Z* has a standard normal distribution, then

$$W = \begin{cases} \dfrac{\exp(gZ) - 1}{g} \exp\left(hZ^2/2\right), & \text{if g > 0} \\ Z\exp(hZ^2/2) & \text{if g > 0} \end{cases}$$

has a g-and-h distribution where *g* and *h* are parameters that determine the first four moments. The four distributions used here were the standard normal (*g* = *h* =0.0), a symmetric heavy-tailed distribution (*h* = 0.5, *g* = 0.0), an asymmetric distribution with relatively light tails (*h* = 0.0, *g* = 0.5), and an asymmetric distribution with heavy tails (*g* = *h* = 0.5). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution considered. For *h* = .5, the third and fourth moments are not defined and so no values for the skewness and kurtosis are reported. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 1: Critical values, $d_{\ell,}$, for Rom's method.

ℓ	α = .05	α = .01
1	.05000	.01000
2	.02500	.00500
3	.01690	.00334
4	.01270	.00251
5	.01020	.00201
6	.00851	.00167
7	.00730	.00143
8	.00639	.00126
9	.00568	.00112
10	.00511	.00101

Table 2: Some properties of the g-and-h distribution.

g	h	(κ_1)	(κ_2)
0.0	0.0	0.00	3.0
0.0	0.5	0.00	—
0.5	0.0	1.81	8.9
0.5	0.5	—	—

Table 3: Estimated probability of a Type I error, $J = K = 2$, $n = 20$, $\alpha = .05$

			Method A		Method B		Method C	
g	h	ρ	Factor A	Inter	Factor A	Inter	Factor A	Inter
0.0	0.0	0.0	.074	.068	.046	.050	.051	.052
0.0	0.0	0.8	.072	.073	.032	.036	.048	.048
0.0	0.5	0.0	.046	.045	.048	.053	.025	.027
0.0	0.5	0.8	.049	.036	.047	.038	.026	.027
0.5	0.0	0.0	.045	.053	.045	.044	.045	.049
0.5	0.0	0.8	.044	.024	.047	.029	.043	.048
0.5	0.5	0.0	.030	.038	.030	.038	.021	.020
0.5	0.5	0.8	.019	.027	.032	.015	.023	.024

Table 4: Estimated Type I error rates using Methods A and C, $J = 2$, $K = 3$, $n = 20$, $\alpha = .05$

			Method A			Method C		
g	h	ρ	Factor A	Factor B	Inter	Factor A	Factor B	Inter
0.0	0.0	0.0	.047	.036	.043	.059	.044	.049
0.0	0.0	0.8	.062	.021	.023	.056	.057	.047
0.0	0.5	0.0	.034	.023	.026	.026	.018	.019
0.0	0.5	0.8	.038	.012	.015	.031	.023	.025
0.5	0.0	0.0	.040	.032	.039	.053	.040	.045
0.5	0.0	0.8	.055	.020	.016	.052	.047	.050
0.5	0.5	0.0	.027	.017	.023	.024	.015	.019
0.5	0.5	0.8	.035	.010	.010	.025	.024	.023

Simulations were run for the case $J = K = 2$ with $n = 20$. (Simulations also were run with $n = 100$ and 200 as a partial check on the software.) Table 3 shows the estimated probability of a Type I error when $\rho = 0$ or .8 when testing Factor A and the hypothesis of no interaction with method A. For brevity, results for Factor B are not shown because they are essentially the same as for Factor A, which should be the case. The estimates are based on 1,000 replications. (From Robey & Barcikowski, 1992, 1,000 replications is sufficient from a power point of view. More specifically, if we test the hypothesis that the actual Type I error rate is .05, and if we want power to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required).

As is evident, method A does a reasonable job of controlling the probability of a Type I error, the main difficulty being that when sampling from a very heavy-tailed distribution, the estimated probability of a Type I error can drop below .025. Switching to method B does not correct this problem. Generally, when using method B the estimated probability of a Type I error was approximately the same or smaller than the estimates shown in Table 3. For example, under normality with $\rho = .8$, the estimates corresponding to Factor A and the hypothesis of no interaction were .035 and .011, respectively. As for method C it performs well with the possible appeal that the estimate never drops below .02, unlike method B.

Table 4 reports results for methods A and C when $J = 2$ and $K = 3$. Both methods avoid Type I error probabilities well above the nominal level. Both methods have estimates that drop below .02, but in general method C seems a bit more satisfactory.

When $J = 3$ and $K = 5$, method A deteriorates even more when dealing with Factor B and interactions, with estimated Type I error probabilities typically below .01. (One exception is normality with $\rho = 0$; the estimates were .020 and .023.) All indications are that method C does better at providing actual Type I error probabilities close to the nominal level. For example, under normality with $\rho = .8$, method A has estimated Type I error probabilities equal to .044, .006 and .001 for Factors A, B and interactions, respectively. For method C, the estimates were .057, .042 and .068.

Conclusion

In summary, the bootstrap version of method A (method B) does not seem to have any practical value based on the criterion of controlling the probability of a Type I error. This is in contrast to the situations considered in Wilcox (2004) where pairwise multiple comparisons among J dependent groups were considered. A possible appeal of method B is that it uses the usual sample median when n is even rather than a single order statistic, but at the cost of risking actual Type I error probabilities well below the nominal level.

Methods A, B and C perform well in terms of avoiding Type I error probabilities well above the nominal level, but methods A and B become too conservative in certain situations where method C continues to perform reasonably well. It seems that applied researchers rarely have interest in an omnibus hypothesis only; the goal is to know which levels of the factor differ. Because the linear contrasts can be tested in a manner that controls FWE, all indications are that method C is the best method for routine use. Finally, S-PLUS and R functions are available from the author for applying method C. Please ask for the function mwwmcp.

References

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics, 37,* 577–580.

Davies, P. L., & Kovac, A. (2004). Densities, spectral densities and modality. *Annals of Statistics, 32,* 1093–1136.

Dawson, M., Schell, A., Rissling, A., & Wilcox, R. R. (2004). Evaluative learning and awareness of stimulus contingencies. Unpublished technical report, Dept of Psychology, University of Southern California.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes.* (p. 461–515). New York: Wiley.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283–288.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77,* 663–666.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.

Staudte, R. G., & Sheather, S. J. (1990). *Robust Estimation and Testing.* New York: Wiley.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques.* San Diego, CA: Academic Press.

Wilcox, R. R. (2004). Pairwise comparisons of dependent groups based on medians. *Computational Statistics & Data Analysis,* submitted.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypot Testing,* 2nd Ed. San Diego, CA: Academic Press