Journal of Data Science **7**(2009), 459-468

Two-by-two ANOVA: Global and Graphical Comparisons Based on an Extension of the Shift Function

Rand R. Wilcox
University of Southern California

Abstract: When comparing two independent groups, the shift function compares all of the quantiles in a manner that controls the probability of at least one Type I error, assuming random sampling only. Moreover, it provides a much more detailed sense of how groups compare, versus using a single measure of location, and the associated plot of the data can yield valuable insights. This note examines the small-sample properties of an extension of the shift function where the goal is to compare the distributions of two specified linear sums of the random variables under study, with an emphasis on a two-by-two design. A very simple method controls the probability of a Type I error. Moreover, very little power is lost versus comparing means when sampling is from normal distributions with equal variances.

Key words: Distribution-free techniques, effect size, interactions, nonparametric methods, quantile estimation, two-way ANOVA.

1. Introduction

The Doksum and Sievers (1976) shift function provides a global and detailed description of how the distributions, corresponding to two independent random variables, compare. It does this via confidence intervals for the differences between all of the quantiles, and the resulting plot can reveal information that is completely missed when attention is restricted to a single measure of location. The method is distribution free in the sense that the simultaneous probability coverage over all quantiles can be determined exactly assuming random sampling only. Moreover, in the event sampling is from normal distributions that differ in location only, its power compares reasonably well with Student's t. And under general conditions it can have higher power than any method based on a single measure of location simply because it is sensitive to broader range of features associated with the distributions under study.

The primary goal of this note is to consider an extension of the shift function to a two-by-two ANOVA design. Let X_1 and X_2 be the random variables associated with the first level of the first factor. And let X_3 and X_4 be the random variables associated with the second level of the first factor. Consider $Y_1 = (X_1 + X_2)/2$ and $Y_2 = (X_3 + X_4)/2$,and let F_1 and F_2 be the corresponding distributions. Then an approach to understanding how the levels of the first factor differ is to investigate in detail how F_1 compares to F_2. Of course, an analog of interactions could be studied as well where now $Y_1 = (X_1 - X_2)/2$ and $Y_2 = (X_3 - X_4)/2$.

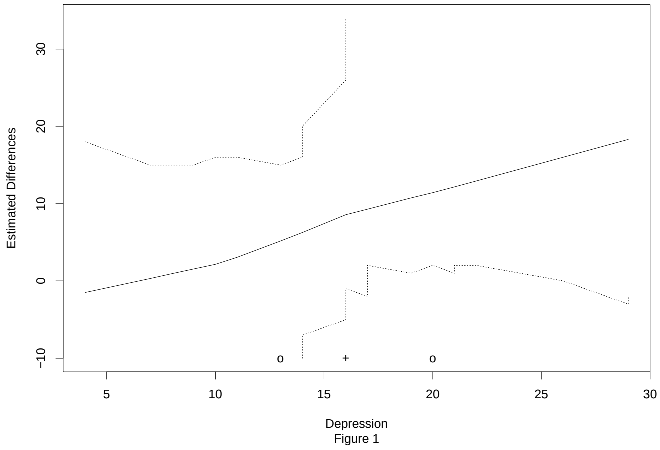

Figure 1: A plot of the shift function for the depression data.

Before continuing, an illustration of the shift function, as it is currently used, might help motivate this paper. Victoroff *et al.* (2006) were generally interested in both psychological and physiological measures related to terrorism. In one portion of the study, which dealt with measures of depression among 14-year-old boys living in Gaza, the goal was to compare two groups of individuals: those who had a family member wounded or killed by Israelis, and those who had not. The solid line in Figure 1 shows the estimated difference between the quantiles, where the x-axis is the measure of depression for the first group (a family member has not been killed or wounded). The circle indicates the sample median for the first group and the upper and lower quartiles are marked by a +. The dashed lines form an approximate .95 confidence band for the difference of all quantiles. (The exact simultaneous probability coverage can be determined, assuming random sampling only, and is .98.) This graph indicates that when comparing individuals with relatively low measures of depression, there is relatively little difference, but as we move toward situations where measures of depression are higher, the difference between the groups becomes more pronounced. (Note that if the groups differ in location only, the shift function would be a straight horizontal line.) In particular, the hypothesis of equal quantiles is rejected for the quantiles extending from .69 to .90.

It seems evident that the Doksum-Sievers method can make a practical difference versus using a single measure of location to compare groups. Consequently, it is only natural to consider how the method might be applied when dealing with a two-way design.

2. Review of the Doksum-Sievers Shift Function

This section reviews the technical details associated with the shift function that will be needed here. Let G_1 and G_2 be the distributions associated with the independent random variables X_1 and X_2, respectively. Let x_{1q} and x_{2q} be the corresponding q-th quantiles. The shift function is

$$\Delta(x_{1q}) = x_{2q} - x_{1q}.$$

Let $\hat{G}_j(x)$ be the usual empirical distribution associated with the j-th group, based on a random sample of size n_j, and let

$$T(\hat{G}_1, \hat{G}_2) = M^{1/2}\text{sup}_x|\hat{G}_1(x) - \hat{G}_2(x)|$$

be the two-sample Kolmogorov-Smirnov test statistic, where $M = n_1n_2/N$ and $N = n_1 + n_2$. Consider the null hypothesis

$$H_0 : G_1(x) - G_2(x) = 0,\ \forall x,$$

and when H_0 is true, suppose c satisfies

$$P(T(\hat{G}_1, \hat{G}_2) \leq c) = 1 - \alpha.$$

Doskum and Sievers (1976, p. 423) show that, because $T(\hat{G}_1, \hat{G}_2) \leq c$ is equivalent to

$$\hat{G}_1 - \frac{c}{\sqrt{M}} \leq \hat{G}_2 \leq \hat{G}_1 + \frac{c}{\sqrt{M}},$$

a level $1 - \alpha$ simultaneous distribution-free confidence band for the $\Delta(x)$ is given by

$$[G_2^{-1}\{G_1(x) - \frac{c}{\sqrt{M}} - x\}, G_2^{-I}\{G_1(x) + \frac{c}{\sqrt{M}} - x,)$$

where $G_2^{-1}(u) = \inf\{G_2(x) \geq u\}$ and $G_2^{-I}(u) = \sup\{G_2(x) \geq u\}$. (Also see Switzer, 1976.) This band is called an S band.

Given c, the exact value of $P(T(\hat{G}_1, \hat{G}_2) \le c)$ can be determined using the recursive method in Kim and Jennrich (1973) when tied values occur with probability zero. When tied values can occur, results in Schroër and Trenkler (1995) can be used to determine $P(T(\hat{G}_1, \hat{G}_2) \le c)$.

Following Doksum and Sievers, the method can be implemented as follows. Let X_{ij} $(i = 1, \ldots, n_j)$ be a random sample of size n_j from the jth group. Denote the order statistics for the jth group by $X_{(1)j} \le \cdots \le X_{(n_j)j}$. For convenience, let $X_{(0)j} = -\infty$ and $X_{(n_j+1)j} = \infty$. For any x satisfying $X_{(i)1} \le x < X_{(i+1)1}$, let

$$k_* = \left\langle n_1 \left(\frac{i}{n_2} - \frac{c}{\sqrt{M}} \right) \right\rangle,$$

where the notation $< x >$ means to round up to the nearest integer. Let

$$k^* = \left[n_1 \left(\frac{i}{n_2} + \frac{c}{\sqrt{M}} \right) \right],$$

where $[x]$ means that x is rounded down to the nearest integer. Then a level $1-\alpha$ simultaneous, distribution-free confidence band for $\Delta(x)$ $(-\infty < x < \infty)$ is

$$[X_{(k_*)2} - x, X_{(k^*+1)2} - x), \tag{2.1}$$

where $X_{(k_*)2} = -\infty$ if $k_* < 0$ and $X_{(k^*)2} = \infty$ if $k^* \ge n_2 + 1$.

3. Description of the Proposed Extension

The focus is on $J = 4$ independent groups, but comments regarding the more general case where $J > 4$ will be made. Let X_j be some random variable associated with the jth group $(j = 1, \ldots, 4)$. Let

$$Y_1 = X_1 + X_2,$$

$$Y_2 = X_3 + X_4,$$

and let F_k be the distribution of Y_k $(k = 1,\ 2)$. One goal might be to test

$$H_0 : F_1(y) = F_2(y),\ \forall y,$$

versus $F_1(y) \ne F_2(y)$ for some y. But if this null hypothesis is rejected, perhaps a more important goal is to determine where the distributions differ and by how much using an obvious extension of the the shift function:

$$\Delta(y_{1q}) = y_{2q} - y_{1q}.$$

Let X_{ij} $(i = 1, \ldots, n_j)$ be a random sample of size n_j from the jth group, let

$$B_{ik1} = X_{i1} + X_{k2},$$

and

$$B_{ik2} = X_{i3} + X_{k4}.$$

For convenience, the $n_1 n_2$ B_{ik1} values are written as $V_{\ell 1}$ $(\ell = 1, \ldots, n_1 n_2)$ and $V_{\ell 2}$ is defined in the same manner using the B_{ik2} values. Also set $N_1 = n_1 n_2$, $N_2 = n_3 n_4$ and $M_w = N_1 N_2/(N_1 + N_2)$. Let $\hat{F}_j$ be the empirical distribution associated with the $V_{\ell j}$ values and let

$$W = M_w^{1/2} \text{sup} |\hat{F}_1(v) - \hat{F}_2(v)|.$$

Then an extension of the shift function to the situation at hand is achieved if now c can be determined such that

$$P(W \leq c) = 1 - \alpha. \tag{3.1}$$

A fundamental obstacle to determining $P(W \leq c)$ is that among the N_1 $V_{\ell 1}$ variables, some are dependent, and of course the same is true for the $V_{\ell 2}$ variables. Consequently, the recursive algorithm derived by Kim and Jennrich (1973) for determining $P(T(\hat{G}_1, \hat{G}_2) \leq c)$ does not extend to the situation at hand. And if this issue is ignored, it is readily verified (via simulations) that the actual probability coverage can differ substantially from the nominal level. But there is a simple solution: Use simulations to determine c so that under normality, equation (2) is approximately true. (This is readily done with modern computers and software for accomplishing this goal is described and illustrated in section 5.)

Note that $T(\hat{G}_1, \hat{G}_2)$ is invariant under order-preserving transformations of the data, but this is not quite the case when using W. So the strategy is to to generate n_j values from a standard normal distribution for the jth group, compute W, and repeat this say I times. Letting $W_{(1)} \leq \cdots \leq W_{(I)}$ be the resulting W values written in ascending order, choose c to be $W_{(d)}$, where $d = (1-\alpha)I$ rounded to the nearest integer. Here $I = 1000$ is used. Simulations reported in section 4 indicate that for non-normal distributions that represent a seemingly extreme departure from normality, the actual probability of a Type I error remains fairly close to the nominal level. The same is true for the discrete distributions to be considered.

3.1 The case $J > 4$

In principle, the method just described can be extended to $J > 4$ groups, where primarily for convenience, J even is assumed. Let $K = J/2$. consider, for example,

$$B_1 = X_1 + \cdots + X_K$$

and

$$B_2 = X_{K+1} + \cdots + X_J.$$

It is evident that this approach becomes impractical because estimating the distribution of B_1, for example, would require computing $n_1 n_2 \ldots, n_K$ terms, which soon becomes too large. To get at least an approximate solution, one possibility would be to use a sample of size L of the $n_1 < n_2 < \cdots < n_K$ combinations and estimate the distribution of B_1 for each of the L_1 combinations, and of course the same could be done when dealing with B_2. Here, $L_1 = n_{(K-1)}n_{(K)}$ is used, where $n_{(1)}, \leq n_{(K)}$ and L_2 is defined likewise.

4. Some simulation results

If the goal is to have the probability of at least one Type I error equal to .05, say, when applying the shift function, under general conditions this cannot be achieved exactly because the distribution of the test statistic is discrete. That is, the exact probability of at least one Type I error can be determined, given a critical value c, but choosing c so that the probability of at least one Type I error equal to .05, say, is impossible. So one goal here is to check on how close the actual probability of at least one Type I error is to .05 using the approximation of the critical value previously described. Because W is not quite invariant under monotone increasing transformations of the data, another goal is to check the level of the test when sampling from some non-normal distributions. Yet another goal is to determine how much power is lost versus comparing means under normality and homoscedasticity.

First consider the case $n_1 = n_2 = n_3 = n_4 = 10$ and suppose the goal is to test the hypothesis of no interactions. (The same results are obtained when dealing with main effects.) To begin, observations are sampled from standard normal distributions. Then based on simulations with 2000 iterations, the actual probability of at least one Type I error is .046. Increasing the sample sizes to 20, the estimate is now .051.

As a check on the effects of sampling from non-normal distributions, observations were generated from g-and-h distributions (Hoaglin, 1985). If Z has a standard normal distribution, then

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

has a g-and-h distribution where g and h are parameters that determine the first four moments. In addition to normal distributions ($g = h = 0$), simulations were run with a symmetric heavy-tailed distribution ($h = .5$, $g = 0$), an asymmetric distribution with relatively light tails ($h = .5$, $g = 0$), and an asymmetric distribution with heavy tails ($g = h = .5$). Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) of these four distributions. Note that $h = .5$ represents a seemingly extreme departure from normality because now kurtosis is not even defined. And with $g = h = .5$, skewness is not defined. Here, equal sample sizes of 10, 20 and 100 were used, plus a situation where the first two groups have sample sizes of 10 and the other two have sample sizes of 100. The estimated probability of a Type I error, among all conditions considered, ranged between .046 and .058. The highest estimate corresponds to where unequal sample sizes are used and sampling is from an asymmetric distribution with heavy tails ($g = h = .5$). With equal sample sizes the largest estimate was .056.

Table 1: 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	1.75	8.9
0.5	0.5	—	—

A related issue is how tied values affect the probability of a Type I error. To get at least some indication, observations were generated from the binomial distribution

$$\begin{pmatrix} 10 \\ x \end{pmatrix} .3^x .7^{10-x}.$$

With $n_1 = n_2 = n_3 = n_4 = 10$, now the probability of at least one Type I error was estimated to be .041.

As for power, consider again the situation where data are generated from a normal distribution with $\delta = .8$ added to every observation in the first group. The usual ANOVA F test has power .44 when testing at the .05 level. The heteroscedastic method for means derived by Johansen (1980) has power .42, and the extension of the shift function has power .41. Let Φ denote a standard normal cumulative distribution function. If instead sampling is from the contaminated normal distribution

$$.9\Phi(x) + .1\Phi(x/10),$$

which has heavy tails, and if now $\delta = 2$, power for the ANOVA F, Johansen's method and the shift function is now .35, .50 and .93, respectively.

Now consider the case of a two-by-four design where the goal is to test the hypothesis of no main effects for the first factor. In this case, only a subset of all possible combinations of the data are used and an issue is whether this has a major impact, under normality, on power versus ANOVA methods based on means. If now $\delta = 1.2$, with all sample sizes equal to 20 and again testing at the .05 level, power for the ANOVA F, Johansen's method and the shift function is now .5, .5 and .42, respectively. For a two-by-five design, now power is .38, .38 and .32. For this latter case, if sampling is from the contaminated normal, power is now .07, .08 and .17. Increasing δ to 2, the power estimates are .12, .13, and .37.

5. More Comments and Illustrations

R software for applying the method in this paper is available from the author. For convenience functions specifically designed for a two-by-two design are included. They are the functions Aband, intended to examine main effects for the first factor, Bband, which deals with main effects for the second factor, and iband, which deals with interactions. For the more general case where the goal is to compare any two linear contrasts, the function sintcon can be used.

Note that the extensions of the shift function provide a graphical check on the nature of any differences among the groups. If groups differ in location only, the resulting plots should yield two straight, horizontal lines. If the groups differ in both location and scale, again we get straight lines, but now they are not horizontal. And if the groups differ in terms of skewness, the plotted lines will be curved.

Also note that when testing the hypothesis of no interaction, the resulting plot created by the function iband provides no information about the extent to which the interaction is disordinal. One way of dealing with this, when dealing with a two-by-two design, is to plot two shift functions. The first might be the plot associated with the two groups corresponding to the first level of the first factor, and the other would be the plot associated with the second level of the first factor. If there is no disordinal interaction, the plots should be identical. The function disband, also available from the author, can be used to create this plot.

To illustrate some of these features, first consider an artificial situation where $n = 80$ for each of four groups (a two-by-two design), the first group has a lognormal distribution, shifted to have a median of zero, two groups have a standard normal distribution, and the fourth has a normal distribution with mean 0 and standard deviation 2. The upper-left panel of Figure 2 shows the resulting plot when examining interactions (with the R function iband). The upper-right panel is the output from disband using the same data. The two shift functions differ, suggesting a disordinal interaction, and the graph reflects the additional fact that the magnitude of the differences increases as we move from low to high values. The lower-left panel is based on data stemming from a study dealing with weight gain associated with four types of diet. The two factors were source of protein (beef versus cereal) and amount of protein (high versus low). Shown is a shift function for the difference between the first two levels versus the second two (using the function iband). An interaction is found, and because the plot is approximately a straight horizontal line, it suggests that the interaction consists essentially of a change in location only. But this plot does not reveal any information about the extent to which the interaction is disordinal. To deal with this issue, a shift function for both high versus low protein when the source of protein is beef, as well as a shift function of high versus low when the source of protein is cereal, is shown in the lower-right panel (using the function disband). As can be seen, the two shift functions differ, consistent with a disordinal interaction, and the nature of the differences appears to be primarily shifts in location.

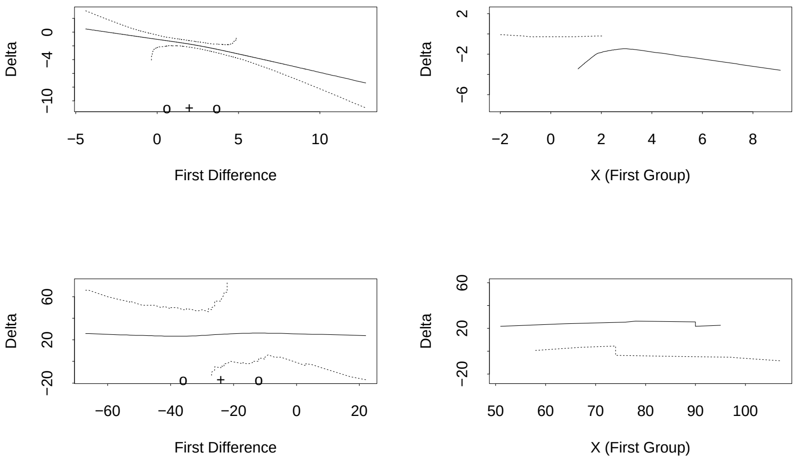

Figure 2

Figure 2: Plots based on the R functions iband and disband that provide details about interactions.

6. Concluding Remarks

As a final note, it would be of interest to extend the method in this paper to dependent groups. Perhaps some extension of the method in Lombard (2005) could be used, but this remains to be determined.

References

Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: graphical comparisons of two populations. *Biometrika* **63**, 421-434.

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

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika* **67**, 85-93.

Kim, P. J. and Jennrich, R. I. (1973). Tables of the exact sampling distribution of the two-sample Kolmogorov-Smirnov criterion, D_{mn}, $m \leq n$. In *Selected Tables in Mathematical Statistics* (Edited by H. L. Harter and D. B. Owen), Vol. I. American Mathematical Society.

Lombard, F. (2005). Nonparametric confidence bands for a quantile comparison function. *Technometrics* **47**, 364-369.

Schroër, G. and Trenkler, D. (1995). Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. *Computational Statistics and Data Analysis* **20**, 185-202.

Switzer, P. (1976). Confidence procedures for two-sample problems. *Biometrika* **63**, 13-25.

Victoroff, J., Quota, S., Celinska, B., Abu-Safieh, R. Y., Adelman, J., Stern, N., Wilcox, R. and Sapolsky, R. (2006). Biopsychological correlates of sympathy for terrorism among refugee boys in Gaza. Paper presented at the Annual Meeting of the American Psychological Society.

Received October 1, 2007; accepted February 5, 2008.

Rand R. Wilcox
Dept of Psychology
University of Southern California
Los Angeles, CA 90089-1061, USA
rwilcox@usc.edu