Florida Journal of Educational Research Fall 1984, Vol. 26, No. 1, Pp. 5-20

Some Comments on "The Unit of Analysis: Group Means Versus Individual Observations"

R. Clifford Blair
University of South Florida
and
J. J. Higgins
Kansas State University

ABSTRACT. Hopkins (1982) has criticized the use of means as the unit of analysis in situations where intact groups (e.g. classes) rather than individuals have been randomly assigned to various treatment conditions. Instead Hopkins advocates the use of certain ANOVA models which, insofar as tests for treatment effects are concerned, yield results that are equivalent to those that would be obtained if class means were employed as the unit of analysis. This paper points out that, because of the nonrobustness of the sample mean as an estimator of location, use of the class mean as the unit of analysis or of the ANOVA models advocated by Hopkins can lead to larger than necessary Type II error rates in tests of significance for treatment effects. This paper also shows how, in the nonnormal population situation, use of summary statistics other than the mean (e.g. members of the family of trimmed means) can lead to significant increases in the power of tests for treatment effects. It is also suggested here that the pooling options offered by Hopkins should be viewed with caution.

Thanks are due to Judy Wilkerson for her helpful comments on an earlier version of this paper.

Hopkins (1982) has provided a useful and correct discussion regarding the analysis of data collected in situations where intact groups (e.g. classes) rather than individuals have been randomly assigned to various treatment conditions. In particular, Hopkins criticizes the commonly accepted practice of using group means as the analytic unit and proposes that certain ANOVA models be used instead. Hopkins summarizes the salient points of his paper in the following statement:

> This paper has shown that the common recommendation to use group means where there may be nonindependence among observational units is unnecessary, unduly restrictive, impoverishes the analysis, limits the questions that can be addressed in a study, and does not insure that the relevant independence assumption has been met. When random factors are properly identified and included in the analysis, the results for all common effects (F_g and critical F_s) are identical in balanced ANOVA designs, regardless of the observational unit employed. The use of individual observations, however, also allows other interesting questions pertaining to interaction and generalizability to be explored. The question of the proper observational unit (or unit of analysis) is answered directly, correctly, and implicitly when the proper statistical model is employed. (p.17)

While we believe that the issues raised by Hopkins are important and that they may constitute a legitimate basis for the use of the ANOVA models he advocates, we also believe that other analytic strategies will be more appropriate in certain research contexts. The purpose of this paper then is to show why the techniques advocated by Hopkins are not always the techniques of choice and to introduce to the research community the other analytic strategies mentioned above.

A key element in the argument put forth by Hopkins is the fact that the ANOVA models he advocates yield an F test for methods (treatment effects) that is identical to the one that would be obtained if group means were used in the analysis. But herein lies a cause for concern for, as Andrews et al. (1972) have pointed out (p.7), "The arithmetic mean is a simple, well understood estimate of location. However, it is highly non-robust being very sensitive to extreme outliers." In reporting the results of a large scale Monte Carlo study of some 68 estimators of location these same authors state (p.239) under the heading "Which was the worst estimator in the study" the following: "If there is any clear candidate for such an overall statement, it is the arithmetic mean, long celebrated because of its many 'optimality properties' and its revered use in applications." Under the heading "What results appear to be of special interest for applied statisticians?" these authors state (p.240) "The arithmetic mean, in its strict mathematical sense, is 'out'. The mean combined with any reasonable rejection procedure, however, can survive, though not very well."

These statements concerning the lack of robustness of the sample mean as an estimator of population location bear directly on the problem at hand since this lack of robustness is manifested in a marked increase in the variance of the mean whenever various nonnormal population shapes are encountered. It should be noted that even modest departures from normality can induce these variance inflations. Factors such as tail weight of the sampled population, skew and/or the presence of even a few extreme observations can multiply the variance of the mean by an uncomfortably large factor.

Table 1 illustrates the problem insofar as tail weight is concerned. Distributions considered in this table range from the light-tailed normal distribution, under which the mean is optimal for the problem considered here, to the extremely heavy-tailed Cauchy distribution. Intermediate to these are the Laplace (or double exponential) distribution, the t curve with three degrees of freedom and a particular contaminated distribution. (For details of this latter distribution see Andrews et al. [1972, p.78 and exhibit 5-76]). The numerical entries in this table represent variances of the mean and several trimmed means (discussed below) when samples are taken from the

TABLE 1

Variances (multipled by n=20)(2) Of The Mean And Several Trimmed Means When Sampling Is From Selected Population Distributions.

	Distribution				
Estimator	Normal	Laplace	t3	Contam-inated	Cauchy
$\overline{X}_0$ (mean)	1.000	2.100	3.138	26.220	12548.000
$\overline{X}_5$	1.022	1.770	1.883	14.930	24.000
$\overline{X}_{10}$	1.056	1.600	1.683	6.710	7.300
$\overline{X}_{15}$	1.098	1.480	1.605	3.280	4.600
$\overline{X}_{25}$(mid-mean)	1.199	1.330	1.591	2.180	3.100
$\overline{X}_{50}$(median)	1.498	1.370	1.817	2.480	2.900

The data in this table, as well as in Table 2, are based on results obtained by Andrews et al. (1972). These authors multiplied all variances by the sample size in order to facilitate comparisons across sample sizes.

indicated populations. As this table illustrates, the variance of the sample mean multiplies rapidly as tail weight increases and soon becomes many times greater than the variance calculated under normal theory.

An important consequence of this nonrobustness to tail weight, as well as to other factors, is that statistical tests that employ group means as the unit of analysis or ANOVA models that produce equivalent results may be handicapped by larger than necessary error terms. That is to say variance inflations of the sample mean will usually produce increased Type II error rates in tests for treatment effects. Fortunately, there are no mathematical or statistical imperatives to force use of the mean as the unit of analysis (or ANOVA models that produce equivalent results) when testing for treatment effects. One might wish, therefore, to employ a summary statistic that is less sensitive to departures from normality than is the sample mean. Again we are fortunate in that the statistical literature is replete with discussions of just such robust estimators of location and related issues. (See for example Andrews et al. [1972], Bickel [1965], Bickel and Hodges [1967], Birnbaum and Laska [1967], Chernoff, Gastwirth and Johns [1967], Crow and Siddiqui [1967], Elashoff and Elashoff [1978]. Filliben [1969], Gastwirth [1966], Gastwirth and Cohen [1970], Hampel [1968], Hoaglin [1971], Hodges and Lehmann [1963], Hogg [1967,1974,1979], Huber [1964], Leone, Jayachanchan and Eisenstat [1967], Mosteller [1947], and Siddiqui and Raghunandanan [1967].)

Of particular interest in this regard are the findings of Andrews et al. (1972). In this large scale study the authors compared certain properties of some 67 robust estimators of location with those of the mean. As has been alluded to previously, the mean did not fair at all well in this study. Other estimators, however, did perform well, maintaining reasonable stability across a variety of population shapes. Some of the more effective estimators belong to the families of "M" estimators, "L" estimators and "adaptive" estimators. While some of these estimators are fairly complex, others are quite simple. (FORTRAN programs are available for computing all of these statistics.)

Among the simple but fairly effective robust estimators of location is the family of trimmed means. Because of their simplicity and familiarity, we will focus on various of the trimmed means in the illustrations that follow. It should be noted, however, that other estimators will probably be more effective in most situations.

The trimmed mean is defined (Elashoff and Elashoff, 1978) quite simply as:

$$\bar{X}g = \frac{1}{(n - 2g)} \qquad (X_{(g+1)} + \ldots X_{(n-g)})$$

with the g largest and g smallest observations being discarded or "trimmed". This statistic is often expressed as $\bar{X}_p$ with p being the percentage of observations trimmed from each end. Thus $\bar{X}_0$ is the mean, $\bar{X}_{25}$ is termed the "midmean" and $\bar{X}50$ is the familar median.

As Table 1 shows, trimmed means are not as efficient under normal theory assumptions as is the mean. However, the loss of efficiency in this situation is usually fairly small while gains in efficiency in the nonnormal situation may be quite large. As a result, the researcher who is fairly confident that extreme observations may occur in the analytic problem might choose to trim classroom data rather severely...say twenty-five percent, in order to maximize gains in efficiency. On the other hand, researchers who are less informed as to the nature of their data may choose to trim more modestly...say five or ten percent. This latter strategy allows the researcher to minimize losses in the situation where sampled populations closely approximate the normal curve while still retaining the potential for sizable gains in efficiency if the populations deviate from normality.

The reader should note that the above comments are made not primarily as guides to analysis strategies, but rather to emphasize the fact that a variety of analytic strategies is available even within the limited scope of the family of trimmed means. The researcher need not therefore depend *solely* on the nonrobust mean as the unit of analysis nor on models that yield equivalent results. Further insights as to how use of estimators other than the mean can influence tests of significance can be gained by considering the following.

Suppose we are interested in a balanced one-way ANOVA model with T treatments, C classrooms per treatment and n students per classroom. We assume that treatments and students are randomly assigned to classrooms. The mathematical model for an individual's score is given by

$$Y_{ijk} = \mu + \alpha_i + b_{ij} + w_{ijk}$$

where μ is the overall mean, α_i is the effect of the ith treatment, b_{ij} is the between classroom error, and w_{ijk} is the within classroom error. It is assumed that the b_{ij}'s are independent and for simplicity of the discussion it will be assumed that the w_{ijk}'s are also independent. It should be pointed out however that tests for treatment effects that employ a classroom measure such as the mean or median do not depend on this latter assumption. The mathematical model for the classroom mean is

$$\overline{Y}_{ij} = \mu + \alpha_i + e_{ij}$$

where

$$e_{ij} = b_{ij} + \frac{w_{ijk}}{n}$$

As Hopkins has pointed out, both the individual's score model and the classroom means model provide the same test of hypothesis for treatment effects, that is

$$H_o: \alpha_1 = \alpha_2 = \ldots = \alpha_T$$

$$H_a: \text{not all } \alpha_i\text{'s are equal.}$$

The variance of experimental error for the classroom means model is given by

$$\sigma^2_e = \sigma^2_b + \frac{\sigma^2_w}{n}$$

where σ^2_b and σ^2_w are the between-classroom and within-classroom contributions to the variance of the sample mean under the assumption of independence of the w_{ijk}'s. If these terms are not independent, one would simply add terms involving the within-classroom correlations to the expression.

The above expression allows some insights as to how the power of the F test for treatment effects might be increased or decreased. One obvious method to increase the power of the test is to increase the number of students per classroom. This strategy would not increase the degrees of freedom for the test as one might suppose (this would be accomplished by adding classrooms to the model), but rather reduces the error variance by reducing the magnitude of $\frac{\sigma^2_w}{n}$. One should note that this method of reducing error variance has a lower bound of σ^2_b. The amount of increase in power depends on the relative magnitudes of σ^2_b and σ^2_w. If σ^2_w is relatively large as compared to σ^2_b, gains in power may be substantial. Otherwise, they may be relatively small.

Another method of reducing experimental error is through the use of robust estimators. As with the number of students per classroom, use of robust estimators may help by reducing the within-classroom component of experimental error. The amount of reduction to be obtained via this method depends upon the within-classroom distribution of the dependent variable. If this measure is <u>perfectly</u> normally distributed, then use of the mean is optimal. Use of a robust estimator in this situation will result in an increase in experimental error, though available evidence (Andrews et al. 1972) indicates that this increase will usually be quite modest. On the other hand, when the within-classroom distribution is not normal, robust estimators may be quite helpful.

As an example, let us suppose that the within-classroom measure takes the form of the moderately heavy-tailed Laplace distribution with mean A and variance σ^2_w. The distribution is defined (Johnson & Kotz, 1970) as

$$f(x) = (\sigma_w \sqrt{2})^{-1} \exp [- \sqrt{2} \mid x = A \mid / \sigma_W]$$

Because of its familiarity and well known properties, let us choose the median (i.e. the 50 percent trimmed mean) as our summary statistic for the classroom data. For the asymptotic case the within classroom contribution to the variance for the classroom median is in this situation approximately

$$\frac{.25}{n} \quad \frac{1}{f^2(A)} = \frac{.5\,\sigma^2_w}{n}$$

(Rao, 1973). Thus for large samples the within-classroom component of the variance of the median is approximately half that of the mean. Simulations for samples of size 20 in Andrews et al. (1972) yield a value of .65 for the ratio of the variance of the median to that of the mean. Hence, assuming a classroom size of 20 and a ratio of variances of .65, we obtain an experimental error variance for the median of

$$\sigma^2_e \text{ (median)} = \sigma^2_b + \frac{.65\,\sigma^2_w}{20}$$

Thus we have a ratio of error variances, median to mean, given by

$$\frac{\sigma^2_e \text{ (median)}}{\sigma^2_e \text{ (mean)}} = \frac{20\,\sigma^2_b + .65\,\sigma^2_w}{20\,\sigma^2_b + \sigma^2_w}$$

In the event that the ratio of σ^2_w to σ^2_b is small, then the reduction in experimental error brought about by the use of the classroom median rather than the mean will be small. On the other hand, when the ratio is large the reduction will be more substantial. In any event, regardless of the particular ratio of σ^2_w to σ^2_b, it would take approximately 31 students in each classroom if one were to choose the class mean in order to obtain the same experimental error variance as would be obtained with 20 students per classroom, if the class median were chosen instead.

As to the assumption of population normality which underlies the F test, the within-classroom distribution of the median is approximately normal in most circumstances as is the distribution of many other estimators of location. Therefore, if the between classroom error b_{ij} can be assumed to be approximately normally distributed then the normality assumption of the F test will for all practical purposes be met.

Table 2 gives values of $\sigma^2 e$ (estimator)/$\sigma^2 e$ (mean) for four trimmed means assuming five different values of $\sigma^2 w/\sigma^2 b$ under five symmetric distributions. While this table shows how experimental error may be reduced under several symmetric theoretical distributions, it does not do the same for nonsymmetric or more importantly, for actual data distributions generated in the context of educational inquiries. This information seems not to be available at the present, although studies are being planned that should provide much of this information. Since it is known that even a few extreme observations can greatly destabilize the mean, it is expected that robust estimators will prove their worth when actual population data sets are examined. Incidently, Table 2 also shows that the midmean would have been preferable to the median for the example problem outlined above. The midmean, as well as those means more lightly trimmed, produces only small increments in the error variance in the normal situation while producing substantial reductions in error variance in the nonnormal case. This is a fortunate characteristic of many robust estimators of location. The median on the other hand produces somewhat larger increments in error variance in the normal case while producing reductions in the nonnormal situation that are not quite as substantial as those produced by the midmean.

To this point we have argued that the power of tests for treatment effects may be increased significantly by summarizing classroom data with statistics other than the mean. This argument extends directly to the models advocated by Hopkins. At this point the question might be raised as to whether or not the pooling options offered by Hopkins might not more than offset power advantages gained through the use of robust estimators. When exercised, certain of these pooling options can lead to tests of significance for treatment effects that are very similar to tests that employ individual observations as the unit of analysis. The increment in degrees of freedom thus

TABLE 2

σ^{2}e (estimator/ σ^{2}e (mean) For Selected Trimmed Mean Estimators And Population Distributions For N=20.

		$\sigma^{2}{}_{w}/\ \sigma^{2}{}_{b}$				
Estimator	Distribution	2	5	10	20	30
$\bar{X}_{5}$	normal	1.00	1.00	1.01	1.01	1.01
	Laplace	.99	.97	.95	.92	.91
	t3	.96	.92	.87	.80	.76
	Contaminated	.96	.91	.86	.78	.74
	Cauchy	.91	.80	.67	.50	.40
$\bar{X}_{10}$	normal	1.01	1.01	1.02	1.03	1.03
	Laplace	.98	.95	.92	.88	.86
	t3	.96	.91	.85	.77	.72
	Contaminated	.93	.85	.75	.63	.55
	Cauchy	.91	.80	.67	.50	.40
$\bar{X}_{25}$	normal	1.02	1.04	1.07	1.10	1.12
	Laplace	.97	.93	.88	.82	.78
	t3	.96	.90	.84	.75	.70
	Contaminated	.92	.82	.69	.54	.45
	Cauchy	.91	.80	.67	.50	.40
$\bar{X}_{50}$	normal	1.06	1.10	1.17	1.25	1.30
	Laplace	.97	.93	.88	.83	.79
	t3	.96	.92	.86	.79	.75
	Contaminated	.92	.82	.70	.55	.46
	Cauchy	.91	.80	.67	.50	.40

obtained would lead to a more powerful test of significance.

Hopkins is careful to point out that before pooling one must establish that the hypothesis $\sigma^2 b = o$ is tenable. In regards to this Hopkins states (p.13), "Of course, these hypotheses must be tested with good power--which suggests that perhaps α should often be relaxed to .20 or .25, especially if the degrees of freedom for the error term are not large."

One notes that the suggested alpha levels would result in a researcher not employing the pooling option in some 20 to 25 percent of the situations where it was appropriate. Even more important, however, is the fact that it is always risky to use failure to reject a null hypothesis as evidence that the null hypothesis is true, since one cannot, in most circumstances, establish the Type II error rate for the test of significance. Establishing α at such a high level does not ensure good power, it merely ensures better power than would have been attained if the test had been carried out at some more traditional level of significance.

This is important since, for example, if $\sigma^2 b$ is non zero but relatively small, the test for this effect may not have sufficiently high probability of detecting its presence. This in turn might lead the researcher to test for treatment effects by means of a model that assumes $\sigma^2 b$ to be zero. It is important to note that even small amounts of $\sigma^2 b$ can greatly distort the Type I error rate of this latter model when testing for treatment effects. Thus, while $\sigma^2 b$ may be too small to detect with high probability, it may still be large enough to distort the Type I error rate of the test.

Glendening (1977) used computer similations to study the advisability of using tests of significance to determine whether or not to pool in situations similar to those considered here. Commenting on the results of this study Glendening and Porter (1976) state

> Because our research showed it to be very important (in inferential situations) to have independent units of analysis, one might ask, is it feasible to use a conditional testing procedure by first doing an initial test of

> independence, testing the equality of the expected mean squares between and within classes, and then on the basis of that test choosing unit of analysis (student or classroom) for the primary test of treatment effects? This study showed clearly, both analytically and empirically that the answer to this question is no.

When one combines the inherent difficulties associated with attempts to ensure a minimal Type II error rate in hypothesis testing with the results obtained by Glendening (1977), the prudent course would suggest a great deal of caution on the part of a researcher who plans to employ the pooling option outlined by Hopkins.

We do not wish to suggest that the methods discussed here (i.e. use of robust estimators) should supplant those discussed by Hopkins in all circumstances. Many additional insights can be gained through the analytic procedures he advocates. Indeed, if our choices were limited to only two analytic strategies, use of the class mean or the models discussed by Hopkins, we would choose the latter for the reasons outlined in that paper. However, experience has taught us that educational data does not always choose to distribute itself in the smooth, regular manner depicted by the normal curve. (See Blair [1981] for examples.) This being the case, and since the nonrobustness of the mean as an estimator of location has been well established, researchers should be aware of the fact that more robust procedures are available for the problem at hand. These techniques are particularly useful when, as is often the case in research/evaluation studies, primary interest centers around tests for treatment effects.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). Robust estimates of location survey and advances. Princeton, NJ: Princeton University Press.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." Review of Educational Research, 51, 499-507.

Bickel, P. J. (1965). On some robust estimates of location. Annals of Mathematical Statistics, 36, 847-858.

Bickel, J. P. & Hodges, J. L. (1967). The asymptotic theory of Galton's test and a related simple estimate of location. Annals of Mathematical Statistics, 38, 73-89.

Birnbaum, A. & Laska, E. (1967). Optimal robustness: A general method with applications to linear estimates of location. Journal of The American Statistical Association, 62, 1230-1240.

Chernoff, H., Gastwirth, J. & Johns, M. V. (1967). Asympototic distributions of linear combinations of functions of order statistics with applications to estimation. Annals of Mathematical Statistics, 38, 52-72.

Crow, E. & Siddiqui, M. (1967). Robust estimation of location. Journal of The American Statistical Association, 62, 353-389.

Elashoff, J. D. & Elashoff, R. M. (1978). Effects of errors in statistical assumptions. In W. H. Kruskal & J. M. Tanur (Eds.), International encyclopedia of statistics (Vol. 1). New York: The Free Press.

Filliben, J. J. (1969). Simple and robust linear estimation of the location parameter of a symmetric distribution. Unpublished doctoral dissertation, Princeton University.

Gastwirth, J. (1966). On robust procedures. Journal of The American Statistical Association, 61, 929-948.

Gastwirth, J. & Cohen, M. (1970). Small sample behaviors of some robust linear estimates of location. Journal of The American Statistical Association, 65, 946-973.

Glendening, L. K. (1977). Operationally defining the assumption of independence and choosing the appropriate unit of analysis. Unpublished doctoral dissertation, Michigan State University.

Glendening, L. K. & Porter, A. C. (1976, October). Independendence and selecting the appropriate unit of analysis. Paper presented at the Conference on Aggregating Data In Educational Research, Stanford, CA.

Hampel, F. (1968). Contributions to the theory of robust estimation. Unpublished doctoral dissertation, Berkeley.

Hoaglin, D. (1971). Optimal invariant estimation of location for three distributions and the invariant efficiency of some other estimators. Unpublished doctoral dissertation, Princeton University.

Hodges, J. L. & Lehmann, E. L. (1963). Estimates of location based on rank tests. Annals of Mathematical Statistics, 34, 598-611.

Hogg, R. V. (1967). Some observations on robust estimation. Journal of The American Statistical Association, 62, 1179-1186.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. Journal of The American Statistical Association, 69, 909-927.

Hogg, R. V. (1979). Statistical robustness: One view of its use in applications today. The American Statistician, 33, 108-115.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. American Educational Research Journal, 19, 5-18.

Huber. P. J. (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics, 35, 73-101.

Johnson, N. L. & Kotz, S. (1970). Continuous univariate distributions. New York: John Wiley and Sons.

Leone, F., Jayachanchan, T. & Eisenstat, S. (1967). A study of robust estimators. Technometrics, 9, 652-660.

Mosteller, F. (1947). On some useful "inefficient" statistics. Annals of Mathematical Statistics, 17, 377-408.

Rao, C. R. (1973). Linear statistical inference and its applications. (2nd ed.). New York: John Wiley and Sons.

Siddiqui, M. M. & Raghunandanan, K. (1967). Asymptotically robust estimators of location. Journal of The American Statistical Association, 62, 950-953.

AUTHORS

R. CLIFFORD BLAIR, Associate Professor, Department of Educational Measurement and Research, FAO 293, University of South Florida, Tampa, Florida 33620.

J. J. HIGGINS, Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506.