Multiple Comparison Methods for Data Review of Census of Agriculture Press Releases

Richard Griffiths, Bureau of the Census
3321 William Johnston Lane #12, Dumfries, VA 22026

KEY WORDS: Multiple comparison, comparisonwise error rate, experimentwise error rate, Tukey-Kramer

1. INTRODUCTION

Press releases issued by the Census Bureau must be reviewed in accordance with the Bureau's standards for presentation of errors in data. Specifically, our concern is to verify the statistical validity of comparisons made in census of agriculture press releases.

Often, this verification process takes the form of ranking k populations. For example, we may be interested in publishing the leading state in value of agricultural products sold. Or, we may be interested in reporting the leading crop in sales in a particular state.

The method used to perform statistical tests of hypotheses for these comparisons for the 1987 Census of Agriculture was the Least Significant Difference (LSD) method. The reason for our reevaluation of this method along with the presentation of alternative methods and an empirical study of these methods is given in this paper.

2. THE LSD METHOD

Use of the LSD method requires each pairwise comparison to be performed with level of significance α. α is the comparisonwise error rate; that is, the probability of making a type I error on an individual comparison.

The LSD method ignores the fact that over all comparisons in a set of comparisons necessary to rank k populations, the probability of making at least one type I error -- the experimentwise error rate -- is greater than α.

For the LSD method, the experimentwise error rate is

$$\epsilon = 1-(1-\alpha)^c \qquad (1)$$

where $c=\binom{k}{2}$ is the total number of pairwise comparisons made in ranking a set of k populations. Eq.(1) assumes all tests are independent. Šidák (1967) has shown eq.(1) to be an upper bound for normally distributed random variables.

For example, if we want to rank four populations, the experimentwise error rate, using a .10 comparisonwise error rate and assuming independence of tests, would be

$$1-(1-.10)^6=.469 \quad .$$

Although an upper bound, this is a far greater probability of type I error than we intended.

3. FOUR MULTIPLE COMPARISON METHODS

It is generally accepted that if we want to control the experimentwise error rate we should employ multiple comparison methodology. In this paper, we examine four multiple comparison methods that control the experimentwise error rate: the Scheffé, Bonferroni, Šidák and Tukey-Kramer methods. There are many other multiple comparison methods, but these are the four most applicable to our press release data review situation.

In general, for each pairwise comparison, a test is performed based on the test statistic

$$t=\frac{\hat{Y}_i-\hat{Y}_j}{se(\hat{Y}_i-\hat{Y}_j)} \qquad (2)$$

where $\hat{Y}_i$, $\hat{Y}_j$ are the estimated totals (or means) for populations i and j, respectively, and $se(\hat{Y}_i-\hat{Y}_j)$ is the standard error of the difference. In census of agriculture testing procedures, we force t to be positive by specifying $\hat{Y}_i$ to be the larger of the two operands. This necessitates the use of a two-sided test. Thus, all tests considered in this paper are two-sided.

Assuming normality of the underlying distribution, t has a t-distribution. With the large sample sizes, ranging from the hundreds to tens of thousands, encountered in census of agriculture applications, t is assumed to have a standard normal distribution.

All four multiple comparison methods work the same way: for each comparison, a decision is reached by comparing t to a critical value. It is on the basis of the critical values that the four methods differ.

The critical value of Scheffé's method, based on the critical value for the overall F-test in an analysis of variance, is

$$\sqrt{(k-1)\cdot F_{\epsilon,k-1,dfe}}$$

where k = the number of populations to be compared, dfe = degrees of freedom for error, ϵ = the predetermined experimentwise error rate.

Application of Scheffé's method proceeds as follows. If Y_i and Y_j are being compared, with $\hat{Y}_i > \hat{Y}_j$, then Y_i is said to be greater than Y_j if $\frac{\hat{Y}_i - \hat{Y}_j}{se(\hat{Y}_i-\hat{Y}_j)}$ is greater than $\sqrt{(k-1)\cdot F_{\epsilon,k-1,dfe}}$. This test may be performed for all pairwise combinations of items with an experimentwise error rate less than or equal to ϵ.

Scheffé's method goes one step further, though. It allows all possible contrasts to be tested. So, not only may items be compared pairwise, but also all other contrasts of items may be performed with guaranteed experimentwise error rate ϵ. Pairwise comparisons are a subset of this larger group. Consequently, the comparisonwise error rate must be very small to account for all possible contrasts while maintaining the experimentwise error rate at ϵ. This fact makes Scheffé's method a little conservative.

Bonferroni's method has critical value

$$t_{\frac{\epsilon}{2},dfe}$$

where the comparisonwise error rate, α, is set equal to

$$\frac{\epsilon}{\frac{k(k-1)}{2}} = \frac{\epsilon}{c}$$ to ensure an experimentwise error rate of less than or equal to ϵ; c = k(k-1)/2 = number of comparisons to be performed; dfe = degrees of freedom.

For all pairwise comparisons, Y_i vs. Y_j, with $\hat{Y}_i > \hat{Y}_j$, Y_i is said to be greater than Y_j if $\frac{\hat{Y}_i - \hat{Y}_j}{se(\hat{Y}_i - \hat{Y}_j)}$ is greater than

$$t_{\frac{\epsilon}{2},dfe}$$.

By Bonferroni's inequality, it is seen that using ϵ/c as the level of significance for each test yields an experimentwise error rate of less than or equal to ϵ. Bonferroni's method produces critical values smaller than Scheffé's method.

Šidák's method is based on a comparisonwise error rate of $1-(1-\epsilon)^{1/c}$. This comparisonwise error rate is derived in such a way that it maintains the experimentwise error rate at or below ϵ for a set of c tests. If all tests are independent, the experimentwise error rate equals ϵ.

Šidák's critical value for each pairwise comparison, for a two-sided test, is

$$t_{\frac{\alpha}{2},dfe}$$

where $\alpha = 1-(1-\epsilon)^{1/c}$ is the comparisonwise error rate.

Šidák's comparisonwise error rate turns out to be slightly greater than Bonferroni's. Thus, the critical value used in this test is slightly lower than Bonferroni's.

The Tukey-Kramer method is based on a critical region of the studentized range distribution, denoted by q.

Using this method, with $\hat{Y}_i > \hat{Y}_j$, Y_i is said to be greater than Y_j if $\frac{\hat{Y}_i - \hat{Y}_j}{se(\hat{Y}_i - \hat{Y}_j)}$ is greater than $\frac{q_{\epsilon,k,dfe}}{\sqrt{2}}$. For values of the q-distribution, see Harter (1960).

The Tukey-Kramer method produces critical values smaller than those of the other three previously-mentioned multiple comparison methods. It also controls the experimentwise error rate at approximately ϵ (see Dunnett (1980)).

4. POWER CONCERNS

At the core of the debate over whether to use the LSD method or a multiple comparison method are the issues of experimentwise error rate and power. In sections 2 and 3, we talked about the experimentwise error rate. In this section, we discuss power.

The power of a test is defined as the probability of rejecting a false null hypothesis. In general, the approach of the statistician is a conservative one. In this case, it is to guard against a type I error; i.e. to control the experimentwise error rate. This allows us to assert with a high degree of confidence (say, $1-\alpha = .9$) that there is truly a population difference if we find significant evidence of this in our sample data. However, the power of the test suffers. Thus, if we do not find significant evidence of a population difference, we have very little confidence stating there is no difference. We usually just say we have not found enough evidence to indicate a population difference.

If, however, we know the power of a test is .9, and our test provided a nonsignificant result, we feel a good degree of confidence stating there is no population difference. Basically, there is a 90% chance that if a population difference exists, we will find it; i.e. reject the null hypothesis. Thus, if we don't find evidence of a difference, we feel fairly confident stating there is no population difference. So, in comparing multiple comparison methods, we judge not only by the method's ability to control the experimentwise error rate, but also by the power of the method.

In an application such as ours, usually very little work is done to determine the power of a test. Most likely, we can determine α or ϵ, but have no idea of the power. Most of this is due to the fact that determining the power of a test depends, in large part, on the true population value which is unknown. However, with a little knowledge of our data, we may be able to form a general idea of the power of a test for a meaningful region of population values.

When we determine the null hypothesis to be equality of the two population values and the alternative to be inequality, this supposes that irrespective of the size of the true difference in the population values, this difference is of interest. Usually, however, there is a region of indifference. That is, there is a set of values for which the population difference is so small that we are, in general, not interested in the difference. For example, let's say we are interested in the sales of four particular crops in the state of Kansas. For this example, we want only to prove the null hypothesis wrong if the sales of a crop is at least $1,000,000 more than another crop. So, if the difference in sales is actually less than $1,000,000, we would say this difference is rather meaningless relative to the magnitude of our estimates, and we are thus not very concerned with the power of the test in this region. We can, however, determine the power of the test if this difference is at least $1,000,000. Determining the power of the test when the difference is $1,000,000 will give us the minimum power of the test over the population values of interest.

Assuming the true population standard deviation is 700,000 for the estimated sales of the four crops of interest and assuming a comparisonwise error rate of .10 for the LSD method and an experimentwise error rate of .10 for the four multiple comparison methods, allows us to construct Table 1.

Table 1 provides us with the power of the LSD method and the four multiple comparison methods for different choices of indifference region for this example. If, in fact, we are not concerned with differences of less than $1,000,000, we can see from the table that the probability the LSD method will reject a null hypothesis is at least .2611 when the difference is in the region of interest. For the Tukey-Kramer method, this probability is at least .1003. Thus, we have gained an idea of the power of the methods for this particular example. Perhaps, in this case, with the power as low as it is, we would want to find some way to increase the power on the region of interest.

While each situation will have its' own particular indifference region, the concept of indifference region can be a useful tool for evaluating the power of the testing method.

Table 1
Power on an Individual Test, by Indifference Region, for an Example with k=4, σ= 700,000.

	comparisonwise error rate	experimentwise error rate	Power on an Individual Test		
			Indifference Region		
			$500,000	$1,000,000	$5,000,000
LSD	.10	.47*	.1271	.2611	.9997
Scheffé	.012	.10	.0233	.0681	.9946
Bonferroni	.016	.10	.0294	.0823	.9960
Šidák	.017	.10	.0307	.0853	.9962
Tukey-Kramer	.022	.10	.0375	.1003	.9971

*Upper bound

5. EMPIRICAL STUDY

To highlight the differences among the four multiple comparison methods, an empirical study of census of agriculture data was conducted. The purpose of the study was to quantify the power differences among the four multiple comparison methods and examine the applicability of these methods to data review of census of agriculture press releases.

Data from the 1987 Census of Agriculture were used. These data consisted of estimates and coefficients of variation (CVs) for items that would commonly be found in a census of agriculture press release.

The study was naturally divided into two parts: 1) within state comparisons and 2) across state comparisons. Within state comparisons were comparisons for which the main concern was the ranking of different items within one state. For example, the ranking of the sales of different types of crops in Kansas was a within state comparison. Across state comparisons were comparisons for which states were ranked for one item. For example, the ranking of the 50 states on farm production expenditures was an across state comparison.

All tests necessary to perform the rankings were conducted and the total number of significant results for each set of tests, i.e. each ranking, was observed.

Now, before we get to the actual results of the empirical study, we need to note some adjustments applied to the multiple comparison methods.

An assumption made in the discussions of previous sections, and particularly in the use of eq.(2) as the test statistic, is that of homogeneity of variances for the k populations being investigated.

The test statistic actually used in the empirical study was

$$t = \frac{\hat{Y}_i - \hat{Y}_j}{\sqrt{se^2(\hat{Y}_i) + se^2(\hat{Y}_j)}} \quad . \tag{3}$$

Eq.(3) assumes heterogeneous variances.

With heterogeneous variances, the pooled estimate of variance is not appropriate and, as a result, t no longer has a t-distribution. This problem results from the fact that the denominator no longer has a chi-square distribution. An adjustment to the degrees of freedom, proposed by Satterthwaite (1946), however, allows us to approximate the distribution of the denominator by a chi-square distribution with the adjusted degrees of freedom. This adjustment allows the distribution of t to be approximated by the t-distribution.

Satterthwaite's adjusted degrees of freedom is

$$dfe' = \frac{[V(\hat{Y}_i) + V(\hat{Y}_j)]^2}{\frac{[V(\hat{Y}_i)]^2}{r_i} + \frac{[V(\hat{Y}_j)]^2}{r_j}} \tag{4}$$

where $V(\hat{Y}_i)$ and $V(\hat{Y}_j)$ denote the variances of $\hat{Y}_i$ and $\hat{Y}_j$, respectively, r_i and r_j, the sample sizes minus 1 from populations i and j, respectively, denote the degrees of freedom for $V(\hat{Y}_i)$ and $V(\hat{Y}_j)$. Dunnett (1980) notes that dfe' is always between the degrees of freedom for $V(\hat{Y}_i)$ and $V(\hat{Y}_j)$.

With r_i and r_j sufficiently large, as in our case, dfe' should be sufficiently large to assume the approximate distribution of t is standard normal.

So, the empirical study was conducted assuming that t had an approximate standard normal distribution and thus the Scheffé, Bonferroni and Šidák methods are used unaltered from the discussions in section 3.

The Tukey-Kramer method is founded on the distribution of the studentized range statistic

$$q = \frac{\hat{Y}_{(1)} - \hat{Y}_{(k)}}{se(\hat{Y}_{(1)} - \hat{Y}_{(k)})}$$

where $\hat{Y}_{(1)}$ and $\hat{Y}_{(k)}$ denote, respectively, the largest and smallest of the k estimates and $se(\hat{Y}_{(1)} - \hat{Y}_{(k)})$ is estimated by a pooled estimate of standard error.

If there is heterogeneity of the variances of the k populations, the statistic

$$q = \frac{\hat{Y}_{(1)} - \hat{Y}_{(k)}}{\sqrt{se^2(\hat{Y}_{(1)}) + se^2(\hat{Y}_{(k)})}} \tag{5}$$

no longer has a studentized range distribution and the foundation of the Tukey-Kramer method is weakened. Games and Howell (1976), though, proposed using

Table 2
Cumulative Totals of Number of Tests Rejected for the Across State Comparison of Total Net Cash Income.

	Scheffé	Bonferroni	Šidák	Tukey-Kramer	Number of tests performed
Number of tests rejected	490	756	760	770	1225

Satterthwaite's adjustment for the degrees of freedom in this case. Using the adjusted dfe given by eq.(4), eq.(5) has an approximate studentized range distribution with dfe' degrees of freedom.

Again, since our sample sizes are large, the studentized range distribution with dfe' degrees of freedom is closely approximated by a studentized range distribution with ∞ degrees of freedom. Thus, the critical value of the Tukey-Kramer method used in the empirical study remains

$q_{\epsilon,k,\infty}/\sqrt{2}$.

6. RESULTS

As stated earlier, the method that has the most power and controls the experimentwise error rate is the preferred method. While it is obvious that the Tukey-Kramer method is the most powerful of the four methods, the use of actual press release data helps to quantify how much more powerful the Tukey-Kramer method is in the given setting, that of the review of census of agriculture press releases. It also sheds some light on the usefulness of multiple comparison methods for this application.

The results of a comparison of all 50 states on total net cash income is provided in Table 2. This table provides the number of tests found significant when ranking was performed for the 50 states on this particular item.

The Tukey-Kramer method rejected ten more tests than Šidák's method. As expected, Scheffé's method was quite conservative, rejecting far fewer tests than the other three methods. The Bonferroni and Šidák methods were relatively close to each other in number of tests rejected.

Since the empirical study was intended to provide a quantification of the differences among the four multiple comparison methods, let's look at the results from Table 2 in terms of the power of each method for a hypothetical example. This example should give us an idea of the power of each of the four multiple comparison methods relative to the others.

Suppose of the 1225 tests conducted, 300 null hypotheses were true; i.e. 300 tests were performed for equal totals. Since the experimentwise error rate was set at .10 for all four methods, it is probable that none of the multiple comparison methods made a type I error. This means that of the remaining 925 false null hypotheses, the Scheffé method rejected 490 or 53.0% (see Table 3), the Bonferroni method rejected 81.7%, the Šidák method 82.2% and the Tukey-Kramer method 83.2%. These percentages are then the power of the methods over this set of tests under the assumption of 300 true null hypotheses. Thus, the Tukey-Kramer method was 1.0% more powerful than Šidák's method, 1.5% more powerful than Bonferroni's and 30.2% more powerful than Scheffé's method under the assumption.

Results similar to these in terms of the power of the four multiple comparison methods were found for some of the other results. However, something else quite interesting emerged.

In Table 4, we find the results for the within state comparisons of three separate items each with several subcategories. This table provides the number of comparisons found significant when ranking was conducted, within each of ten states, for the subcategories of these three different items. The results depicted in this table are remarkable for the lack of appreciable differences exhibited by the four methods. All rejected fairly the same number of tests.

The interesting thing to note here is that all four methods rejected almost all the tests for all three items. The total number of tests conducted was 910 for the rankings of sales of different crops. Scheffé's method rejected 99.3% of all tests and the Tukey-Kramer method 99.5%. Similar were the results for the livestock sales and farm production expenses categories.

For the livestock sales category, there was a total of 150 tests. All four methods rejected all of the tests. For the farm production expenses category, there were 1050 tests performed. Even though the Tukey-Kramer method

Table 3
Power of the Four Multiple Comparison Methods Assuming 300 of the 1225 Tests Conducted had True H_0.

	Scheffé	Bonferroni	Šidák	Tukey-Kramer
number of false H_0 rejected under assumption of 925 false H_0	490	756	760	770
POWER = number of false null hypotheses rejected divided by the number of false null hypotheses	53.0%	81.7%	82.2%	83.2%

Table 4
Cumulative Totals of Number of Tests Rejected for Within State Comparisons of Subcategories of Certain Items Within Each of Ten States (with Percentage of Total Number of Tests Rejected in Parentheses).

Item	Scheffé	Bonferroni	Šidák	Tukey-Kramer	Number of tests performed
sales of crops	904 (99.34%)	905 (99.45%)	905 (99.45%)	905 (99.45%)	910
sales of livestock	150 (100%)	150 (100%)	150 (100%)	150 (100%)	150
farm production expenses	1012 (96.38%)	1021 (97.24%)	1021 (97.24%)	1023 (97.43%)	1050

rejected 11 more tests than Scheffé's method, the Scheffé method still rejected 96.4% of all the tests.

The results displayed in Table 4 were not atypical. In many cases, a large proportion of the tests conducted were rejected by all four multiple comparison methods.

This large number of significant results may indicate that most of the population totals were different. To illustrate this point, let's look at an example.

Suppose we conduct a study in which we rank the 50 states for a particular item. And suppose the true population totals for the 50 states can be divided into two groups. In group 1 are 25 (which 25 is unknown) states which have equal population totals,

$$Y_1 = Y_2 = Y_3 = \ldots = Y_{25} = Y \ .$$

In group 2 are the other 25 states which have unequal population totals,

$$Y_i \neq Y_j \neq Y, \text{ for } i \neq j = 26, \ldots, 50$$

For the totals in group 1, the equal totals, there are $\binom{25}{2} = 300$ pairwise comparisons. Assuming the experimentwise error rate is .10, it is probable none of the 300 tests of hypothesis on these equal totals will be significant.

Let's also assume that all tests of hypothesis for pairwise comparisons of group 2 totals, the unequal totals, are significant; i.e. we commit no type II errors -- power = 100%. Thus, we know that since there are 300 pairwise comparisons for group 2, these 300 tests are significant.

And let's assume that the other 625 tests between totals from the two groups are all significant, resulting in another 625 rejected tests; again, we commit no type II errors. This means that assuming 100% power on the tests of unequal population totals and an experimentwise error rate of .10 on the tests of equal population totals, 925 tests are found significant.

Thus, even though 50% of the population totals are actually different, and the power of our tests is a perfect 100%, in this example, only 75.5% (925) of the 1225 tests are rejected. This percentage is somewhat smaller than the percentage of tests rejected in the study this paper is concerned with. This percentage is 90% or above for many items. Thus, we would expect that far more than 50% of the population totals in this study are actually different.

In light of these results, a word of caution may be in order. While multiple comparison methods provide a way of controlling the experimentwise error rate, they are also a less powerful method than, say, the LSD method. In general, the approach to testing hypotheses is a conservative one: to protect against a type I error at the expense of allowing a greater chance of a type II error; hence, the use of multiple comparison methodology. However, if most population totals are different, i.e. most H_0 are actually false, there are very few tests for which we need to worry about a type I error. Apparently, this is the case for most items under study in this paper.

This situation highlights the fact that it is important to evaluate the power of the test as well as the experimentwise error rate.

7. CONCLUSION

The empirical study verified the obvious. The Tukey-Kramer method is the most powerful of the four multiple comparison methods and also controls the experimentwise error rate at a specified value very well.

However, we also learned from the empirical study it is important to know the data with which we work. If there is an indication that many of the population parameters being estimated are different, perhaps results of previous censuses of agriculture would suggest this, then we might want to consider adding a little power to our testing procedure by tolerating a higher experimentwise error rate. This is one option. Or, perhaps we could use an approach that takes advantage of any a priori knowledge we may have. Further research needs to be conducted for this option. Keep in mind, though, that this is not a license to ignore the experimentwise error rate; it still must be accounted for.

The caution given by the results of the empirical study is if few of the original null hypotheses are true, there are few tests for which we need to worry about making a type I error. The fact that we may want to be a little more concerned with the power of the tests, however, should not be viewed as a recommendation to use the LSD method. While the LSD method is more powerful than multiple comparison methods, the fact that it ignores the experimentwise error rate renders it unacceptable for our application. We need to use a multiple comparison method in order to control the experimentwise error rate. However, the empirical study forces us to realize that both the power and experimentwise error rate are important quantities.

For comparisons in which there is a need to control the experimentwise error rate, and this applies to most situations as well as the census of agriculture press release situation, the Tukey-Kramer multiple comparison method is clearly the choice of the four methods presented in this paper.

ACKNOWLEDGMENTS

The author would like to thank all those who reviewed this paper, especially Inez Chen of the Agriculture Division of the Bureau of the Census for her insightful and valuable recommendations.

REFERENCES

Bain, Lee J. and Engelhardt, Max (1987), Introduction to Probability and Mathematical Statistics, Duxbury Press, Boston.

Boardman, Thomas J. and Moffit, Donald R. (1971), "Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures", Biometrics, 27, pp. 738-744.

Cohen, Jacob (1977), Statistical Power Analysis for the Behavioral Sciences, Academic Press, New York.

Dunnett, Charles W. (1980),"Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case", Journal of the American Statistical Association, 75, pp. 789-795.

Games, Paul A. (1971), "Multiple Comparison of Means", American Educational Research Journal, vol.8, no.3, pp. 531-565.

Games, Paul A. and John F. Howell (1976), "Pairwise Multiple Comparison Procedures with Unequal n's and/or Variances: A Monte Carlo Study", Journal of EducationalStatistics, 1, pp. 113-126.

Graybill, Franklin A. (1976), Theory and Application of the Linear Model, Duxbury Press, Boston.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), Sample Survey Methods and Theory, vol. 2, John Wiley and Sons, New York.

Harter, Leon H. (1960), "Tables of Range and Studentized Range", Annals of Mathematical Statistics, 31, pp. 1122-1147.

SAS/STAT User's Guide Release 6.03 Edition (1988), pp. 593-599, SAS Institute Inc, Cary, NC.

Satterthwaite, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components", Biometric Bulletin, 2, pp. 110-114.

Saville, D.J. (1990), "Multiple Comparison Procedures: The Practical Solution", The American Statistician, vol.44, no.2, pp. 174-180.

Scheffé, Henry (1959), The Analysis of Variance, John Wiley & Sons, New York.

Šidák, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions", Journal of the American Statistical Association, 62, pp. 626-633.

Snedecor, G.W. and Cochran, W.G. (1980), Statistical Methods (seventh edition), Iowa State University Press, Ames, Iowa.

Winer, B.J. (1971), Statistical Principles in Experimental Design, McGraw-Hill Series in Psychology, New York.